Optimization for Data Science

Lecture Notes, FS 21

Bernd Gärtner, ETH
Martin Jaggi, EPFL

February 20, 2021

Contents

Chapter 1

Theory of Convex Functions

Contents

This chapter develops the basic theory of convex functions that we will need later. Much of the material is also covered in other courses, so we will refer to the literature for standard material and focus more on material that we feel is less standard (but important in our context).

1.1 Mathematical Background

1.1.1 Notation

For vectors in $\mathbb{R}^d$, we use bold font, and for their coordinates normal font, e.g. $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. $\mathbf{x}_1, \mathbf{x}_2, \ldots$ denotes a sequence of vectors. Vectors are considered as column vectors, unless they are explicitly transposed. So $\mathbf{x}$ is a column vector, and $\mathbf{x}^\top$, its transpose, is a row vector. $\mathbf{x}^\top\mathbf{y}$ is the scalar product $\sum_{i=1}^d x_i y_i$ of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

$\|\mathbf{x}\|$ denotes the Euclidean norm (ℓ_2-norm or 2-norm) of vector $\mathbf{x}$,

$$\|\mathbf{x}\|^2 = \mathbf{x}^\top\mathbf{x} = \sum_{i=1}^d x_i^2.$$

We also use

$$\mathbb{N} = \{1, 2, \ldots\} \text{ and } \mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$$

to denote the natural and non-negative real numbers, respectively. We are freely using basic notions and material from linear algebra and analysis, such as open and closed sets, vector spaces, matrices, continuity, convergence, limits, triangle inequality, among others.

1.1.2 The Cauchy-Schwarz inequality

Lemma 1.1 (Cauchy-Schwarz inequality)**.** *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. *Then*

$$|\mathbf{u}^\top\mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| \, .$$

The inequality holds beyond the Euclidean norm; all we need is an inner product, and a norm induced by it. But here, we only discuss the Euclidean case.

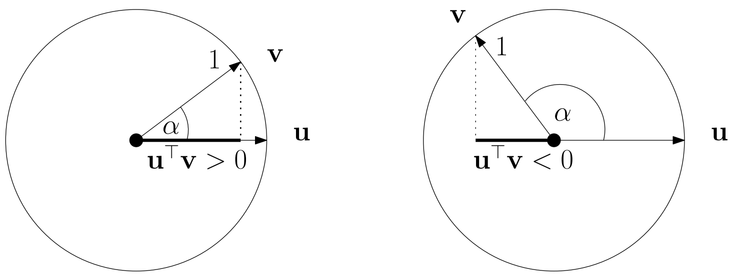
For nonzero vectors, the Cauchy-Schwarz inequality is equivalent to

$$-1 \leq \frac{\mathbf{u}^\top\mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \leq 1,$$

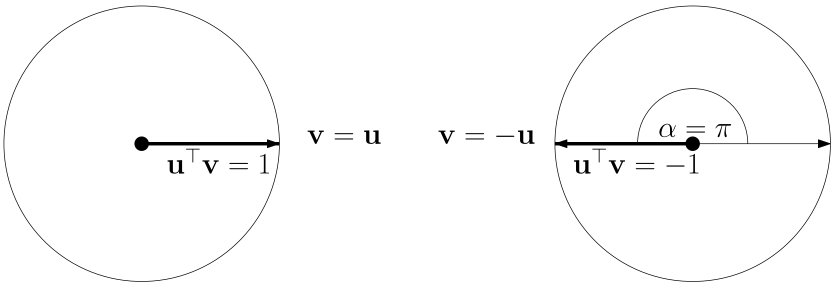
and this fraction can be used to define the angle α between $\mathbf{u}$ and $\mathbf{v}$:

$$\cos(\alpha) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|},$$

where $\alpha \in [0, \pi]$. The following shows the situation for two unit vectors ($\|\mathbf{u}\| = \|\mathbf{v}\| = 1$): The scalar product $\mathbf{u}^\top \mathbf{v}$ is the length of the projection of $\mathbf{v}$ onto $\mathbf{u}$ (which is considered to be negative when $\alpha > \pi/2$). This is just the highschool definition of the cosine.



Hence, equality in Cauchy-Schwarz is obtained if $\alpha = 0$ ($\mathbf{u}$ and $\mathbf{v}$ point into the same direction), or if $\alpha = \pi$ ($\mathbf{u}$ and $\mathbf{v}$ point into opposite directions):


$\mathbf{u}^\top \mathbf{v} = 1$
$\mathbf{v} = \mathbf{u}$
$\mathbf{v} = -\mathbf{u}$
$\alpha = \pi$
$\mathbf{u}^\top \mathbf{v} = -1$

Fix $\mathbf{u} \neq \mathbf{0}$. We see that the vector $\mathbf{v}$ maximizing the scalar product $\mathbf{u}^\top \mathbf{v}$ among all vectors $\mathbf{v}$ of some fixed length is a positive multiple of $\mathbf{u}$, while the scalar product is minimized by a negative multiple of $\mathbf{u}$.

Proof of the Cauchy-Schwarz inequality. There are many proof, but the authors particularly like this one: define the quadratic function

$$f(x) = \sum_{i=1}^{d} (u_i x + v_i)^2 = \left(\sum_{i=1}^{d} u_i^2 \right) x^2 + \left(2 \sum_{i=1}^{d} u_i v_i \right) x + \left(\sum_{i=1}^{d} u_i^2 \right) =: ax^2 + bx + c.$$

We know that $f(x) = ax^2 + bx + c = 0$ has the two solutions

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

This is known as the *Mitternachtsformel* in German-speaking countries, as you are supposed to know it even when you are asleep at midnight.

As by definition, $f(x) \geq 0$ for all x, $f(x) = 0$ has at most one real solution, and this is equivalent to having *discriminant* $b^2 - 4ac \leq 0$. Plugging in the definitions of a, b, c, we get

$$b^2 - 4ac = \left(2\sum_{i=1}^{d} u_i v_i\right)^2 - 4\left(\sum_{i=1}^{d} u_i^2\right)\left(\sum_{i=1}^{d} u_i^2\right) = 4(\mathbf{u}^\top \mathbf{v})^2 - 4\left\|\mathbf{u}\right\|^2 \left\|\mathbf{v}\right\|^2 \leq 0.$$

Dividing by 4 and taking square roots yields the Cauchy-Schwarz inequality.

1.1.3 The spectral norm

Definition 1.2 (Spectral norm). *Let A be an $(m \times d)$-matrix. Then*

$$\|A\| := \max_{\mathbf{v} \in \mathbb{R}^d, \mathbf{v} \neq 0} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

is the 2-norm (or spectral norm) of A.

In words, the spectral norm is the largest factor by which a vector can be stretched in length under the mapping $\mathbf{v} \rightarrow A\mathbf{v}$. Note that as a simple consequence,

$$\|A\mathbf{v}\| \leq \|A\|\|\mathbf{v}\|$$

for all $\mathbf{v}$.

It is good to remind ourselves what a norm is, and why the spectral norm is actually a norm. We need that it is absolutely homegeneous: $\|\lambda A\| = |\lambda|\|A\|$ which follows from the fact that the Euclidean norm is absolutely homegeneous. Then we need the triangle inequality: $\|A + B\| \leq \|A\| + \|B\|$ for two matrices of the same dimensions. Again, this follows from the triangle inequality for the Euclidean norm. Finally, we need that $\|A\| = 0$ implies $A = 0$. Which is true, since for any nonzero matrix A, there is a vector $\mathbf{v}$ such that $A\mathbf{v}$ and hence the Euclidean norm of $A\mathbf{v}$ is nonzero.

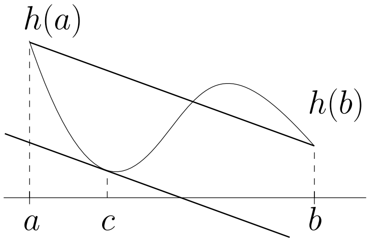
1.1.4 The mean value theorem

We also recall the *mean value theorem* that we will frequently need:

Theorem 1.3 (Mean value theorem). *Let $a < b$ be real numbers, and let $h : [a, b] \to \mathbb{R}$ be a continuous function that is differentiable on (a, b); we denote the derivative by h'. Then there exists $c \in (a, b)$ such that*

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

Geometrically, this means the following: We can interpret the value $(h(b) - h(a))/(b - a)$ as the slope of the line through the two points $(a, h(a))$ and $(b, h(b))$. Then the mean value theorem says that between a and b, we find a tangent to the graph of h that has the same slope:



1.1.5 The fundamental theorem of calculus

If a function h is *continuously* differentiable in an interval $[a, b]$, we have another way of expressing $h(b) - h(b)$ in terms of the derivative.

Theorem 1.4 (Fundamental theorem of calculus). *Let $a < b$ be real numbers, and let $h : \mathbf{dom}(h) \to \mathbb{R}$ be a differentiable function on an open domain $\mathbf{dom}(h) \supset [a, b]$, and such that h' is continuous on $[a, b]$. Then*

$$h(b) - h(a) = \int_a^b h'(t)dt.$$

This theorem is the theoretical underpinning of typical definite integral computations in high school. For example, to evaluate $\int_2^4 x^2 dx$, we integrate x^2 (giving us $x^3/3$), and then compute

$$\int_2^4 x^2 dx = \frac{4^3}{3} - \frac{2^3}{3} = \frac{56}{3}.$$

1.1.6 Differentiability

For univariate functions $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$ with $\mathbf{dom}(f) \subseteq \mathbb{R}$, differentiability is covered in high school. We will need the concept for multivariate and vector-valued functions $f : \mathbf{dom}(f) \rightarrow \mathbb{R}^m$ with $\mathbf{dom}(f) \subseteq \mathbb{R}^d$. Mostly, we deal with the case $m = 1$: real-valued functions in d variables. As we frequently need this material, we include a refresher here.

Definition 1.5. *Let $f : \mathbf{dom}(f) \rightarrow \mathbb{R}^m$ where $\mathbf{dom}(f) \subseteq \mathbb{R}^d$. Function f is called* differentiable *at $\mathbf{x}$ in the interior of $\mathbf{dom}(f)$ if there exists an $(m \times d)$-matrix A and an error function $r : \mathbb{R}^d \rightarrow \mathbb{R}^m$ defined in some neighborhood of $\mathbf{0} \in \mathbb{R}^d$ such that for all $\mathbf{y}$ in some neighborhood of $\mathbf{x}$,*

$$f(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}),$$

where

$$\lim_{\mathbf{v} \rightarrow \mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = \mathbf{0}.$$

It then also follows that the matrix A is unique, and it is called the differential *or* Jacobian *of f at $\mathbf{x}$. We will denote it by $Df(\mathbf{x})$. More precisely, $Df(\mathbf{x})$ is the matrix of* partial derivatives *at the point $\mathbf{x}$,*

$$Df(\mathbf{x})_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x}).$$

f is called differentiable *if f is differentiable at all $\mathbf{x} \in \mathbf{dom}(f)$ (which implies that $\mathbf{dom}(f)$ is open).*

Differentiability at $\mathbf{x}$ means that in some neighborhood of $\mathbf{x}$, f is approximated by a (unique) affine function $f(\mathbf{x}) + Df(\mathbf{x})(\mathbf{y} - \mathbf{x})$, up to a sublinear error term. If $m = 1$, $Df(\mathbf{x})$ is a row vector typically denoted by $\nabla f(\mathbf{x})^\top$, where the (column) vector $\nabla f(\mathbf{x})$ is called the *gradient* of f at $\mathbf{x}$. Geometrically, this means that the graph of the affine function $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$ is a *tangent hyperplane* to the graph of f at $(\mathbf{x}, f(\mathbf{x}))$; see Figure 1.1.

It also follows easily that a differentiable function is continuous, see Exercise 1.

Let us do a simple example to illustrate the concept of differentiability. Consider the function $f(x) = x^2$. We know that its derivative is $f'(x) = 2x$.

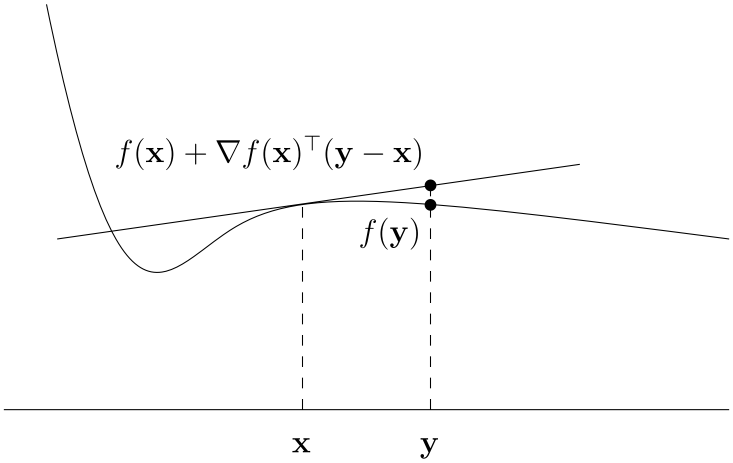

$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$
$f(\mathbf{y})$
$\mathbf{x}$
$\mathbf{y}$

Figure 1.1: If f is differentiable at $\mathbf{x}$, the graph of f is locally (around $\mathbf{x}$) approximated by a tangent hyperplane

But why? For fixed x and $y = x + v$, we compute

$$\begin{aligned} f(y) = (x+v)^2 &= x^2 + 2vx + v^2 \\ &= f(x) + 2x \cdot v + v^2 \\ &= f(x) + A(y-x) + r(y-x), \end{aligned}$$

where $A := 2x, r(y-x) = r(v) := v^2$. We have $\lim_{v \to 0} \frac{|r(v)|}{|v|} = \lim_{v \to 0} |v| = 0$. Hence, $A = 2x$ is indeed the differential (a.k.a. derivative) of f at x.

In computing differentials, the *chain rule* is particularly useful.

Lemma 1.6 (Chain rule). *Let $f : \mathbf{dom}(f) \to \mathbb{R}^m, \mathbf{dom}(f) \subseteq \mathbb{R}^d$ and $g : \mathbf{dom}(g) \to \mathbb{R}^d$. Suppose that $\mathbf{g}$ is differentiable at $\mathbf{x} \in \mathbf{dom}(g)$ and that f is differentiable at $g(\mathbf{x}) \in \mathbf{dom}(f)$. Then $f \circ g$ (the composition of f and g) is differentiable at $\mathbf{x}$, with the differential given by the matrix equation*

$$D(f \circ g)(\mathbf{x}) = Df(g(\mathbf{x}))Dg(\mathbf{x}).$$

Here is an application of the chain rule that we will use frequently. Let $f : \mathbf{dom}(f) \to \mathbb{R}^m$ be a differentiable function with (open) convex domain, and fix $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$. There is an open interval I containing $[0,1]$ such that $\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \mathbf{dom}(f)$ for all $t \in I$. Define $g : I \to \mathbb{R}^d$ by $g(t) =$

$\mathbf{x}+t(\mathbf{y}-\mathbf{x})$ and set $h = f \circ g$. Thus, $h : I \to \mathbb{R}^m$ with $h(t) = f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$, and for all $t \in I$, we have

$$h'(t) = Dh(t) = Df(g(t))Dg(t) = Df(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}). \tag{1.1}$$

1.2 Convex sets

Definition 1.7. *A set $C \subseteq \mathbb{R}^d$ is* convex *if for any two points $\mathbf{x}, \mathbf{y} \in C$, the connecting line segment is contained in C. In formulas, if for all $\lambda \in [0, 1]$, $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in C$; see Figure 1.2.*

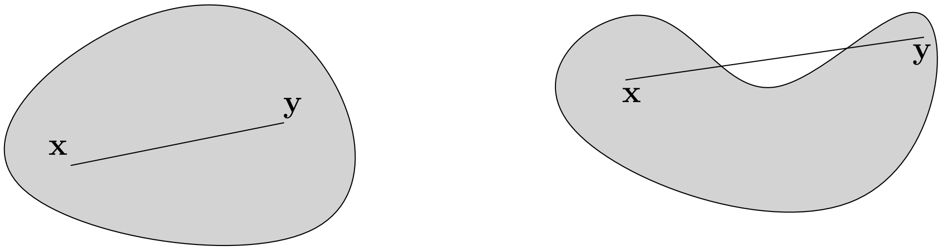


Figure 1.2: A convex set (left) and a non-convex set (right)

Observation 1.8. *Let $C_i, i \in I$ be convex sets, where I is a (possibly infinite) index set. Then $C = \bigcap_{i\in I} C_i$ is a convex set.*

For $d = 1$, convex sets are *intervals*.

1.2.1 The mean value inequality

The mean value inequality can be considered as as generalization of the mean value theorem to multivariate and vector-valued functions over convex sets (a "mean value equality" does not exist in this full generality).

To motivate it, let us consider the univariate and real-valued case first. Let $f : \mathbf{dom}(f) \to R$ be differentiable and suppose that f has bounded derivatives over an interval $X \subseteq \mathbf{dom}(f)$, meaning that for some real number B, we have $|f'(x)| \leq B$ for all $x \in X$. The mean value theorem then gives the *mean value inequality*

$$|f(y) - f(x)| = |f'(c)(y - x)| \leq B|y - x|$$

for all $x, y \in X$ and some in-between c. In other words, f is not only continuous but actually *B-Lipschitz* over X.

Vice versa, suppose that f is B-Lipschitz over a nonempty *open* interval X, then for all $c \in X$,

$$|f'(c)| = |\lim_{\delta \to 0} \frac{f(c+\delta) - f(c)}{\delta}| \leq B,$$

so f has bounded derivatives over X. Hence, over an open interval, Lipschitz functions are exactly the ones with bounded derivative. Even if the interval is not open, bounded derivatives still yield the Lipschitz property, but the other direction may fail. As a trivial example, the Lipschitz condition is always satisfied over a singleton interval $X = \{x\}$, but that does not say anything about the derivative at x. In any case, we need X to be an interval; if X has "holes", the previous arguments break down.

These considerations extend to multivariate and vector-valued functions over *convex* subsets of the domain.

Theorem 1.9. *Let $f : \mathbf{dom}(f) \to \mathbb{R}^m$ be differentiable, $X \subseteq \mathbf{dom}(f)$ a convex set, $B \in \mathbb{R}^+$. If $X \subseteq \mathbf{dom}(f)$ is nonemepty and open, the following two statements are equivalent.*

(i) f is B-Lipschitz, meaning that

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq B \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in X$$

(ii) f has differentials bounded by B (in spectral norm), meaning that

$$\|Df(\mathbf{x})\| \leq B, \quad \forall \mathbf{x} \in X.$$

Moreover, for every (not necessarily open) convex $X \subseteq \mathbf{dom}(f)$, (ii) implies (i), and this is the mean value inequality*.*

Proof. Suppose that f is B-Lipschitz over an open set X. For $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v} \to \mathbf{0}$, differentiability at $\mathbf{x} \in X$ yields for small $\mathbf{v} \in \mathbb{R}^d$ that $\mathbf{x} + \mathbf{v} \in X$ and therefore

$$B\|\mathbf{v}\| \geq \|f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x})\| = \|Df(\mathbf{x})\mathbf{v} + r(\mathbf{v})\| \geq \|Df(\mathbf{x})\mathbf{v}\| - \|r(\mathbf{v})\|,$$

where $\|r(\mathbf{v})\| \,/\, \|\mathbf{v}\| \to 0$, the first inequality uses (i), and the last is the reverse triangle inequality. Rearranging and dividing by $\|\mathbf{v}\|$, we get

$$\frac{\|Df(\mathbf{x})\mathbf{v}\|}{\|\mathbf{v}\|} \leq B + \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|}.$$

Let $\mathbf{v}^\star$ be a unit vector such that $\|Df(\mathbf{x})\| = \|Df(\mathbf{x})\mathbf{v}^\star\| \,/\, \|\mathbf{v}^\star\|$ and let $\mathbf{v} = t\mathbf{v}^\star$ for $t \to 0$. Then we further get

$$\|Df(\mathbf{x})\| \leq B + \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} \to B,$$

and $\|Df(\mathbf{x})\| \leq B$ follows, so differentials are bounded by B.

For the other direction, suppose that differentials are bounded by B over X (not necessarily open); we proceed as in [FM91].

For fixed $\mathbf{x}, \mathbf{y} \in X \subseteq \mathbf{dom}(f), \mathbf{x} \neq \mathbf{y}$, and $\mathbf{z} \in \mathbb{R}^m$ (to be determined later), we define

$$h(t) = \mathbf{z}^\top f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

over $\mathbf{dom}(h) = [0,1]$, in which case the chain rule yields

$$h'(t) = \mathbf{z}^\top Df(x + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}), \quad t \in (0,1),$$

see also (1.1). Note that $\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in X$ for $t \in [0,1]$ by convexity of X. The mean value theorem guarantees $c \in (0,1)$ such that $h'(c) = h(1) - h(0)$. Now we compute

$$\begin{aligned}
\left\|\mathbf{z}^\top (f(\mathbf{y}) - f(\mathbf{x}))\right\| &= |h(1) - h(0)| = |h'(c)| \\
&= \mathbf{z}^\top Df(x + c(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \\
&\leq \|\mathbf{z}\| \|Df(x + c(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})\| \quad \text{(Cauchy-Schwarz)} \\
&\leq \|\mathbf{z}\| \|Df(x + c(\mathbf{y} - \mathbf{x}))\| \|(\mathbf{y} - \mathbf{x})\| \quad \text{(spectral norm)} \\
&\leq B \|\mathbf{z}\| \|(\mathbf{y} - \mathbf{x})\| \quad \text{(bounded differentials)}.
\end{aligned}$$

We assume w.l.o.g. that $f(\mathbf{x}) \neq f(\mathbf{x})$, as otherwise, (i) trivially holds; now we set

$$\mathbf{z} = \frac{f(\mathbf{y}) - f(\mathbf{x})}{\|f(\mathbf{y}) - f(\mathbf{x})\|.}$$

With this, the previous inequality reduces to (i), so f is indeed B-Lipschitz over X. □

1.3 Convex functions

We are considering real-valued functions $f : \mathbf{dom}(f) \to \mathbb{R}$, $\mathbf{dom}(f) \subseteq \mathbb{R}^d$.

Definition 1.10 ([BV04, 3.1.1]). *A function $f : \mathbf{dom}(f) \to \mathbb{R}$ is* convex *if (i) $\mathbf{dom}(f)$ is convex and (ii) for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ and all $\lambda \in [0, 1]$, we have*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \tag{1.2}$$

Geometrically, the condition means that the line segment connecting the two points $(\mathbf{x}, f(\mathbf{x})), (\mathbf{y}, f(\mathbf{y})) \in \mathbb{R}^{d+1}$ lies pointwise above the graph of f; see Figure 1.3. (Whenever we say "above", we mean "above or on".) An important special case arises when $f : \mathbb{R}^d \to \mathbb{R}$ is an affine function, i.e. $f(\mathbf{x}) = \mathbf{c}^\top\mathbf{x} + c_0$ for some vector $\mathbf{c} \in \mathbb{R}^d$ and scalar $c_0 \in \mathbb{R}$. In this case, (1.2) is always satisfied with equality, and line segments connecting points on the graph lie pointwise on the graph.

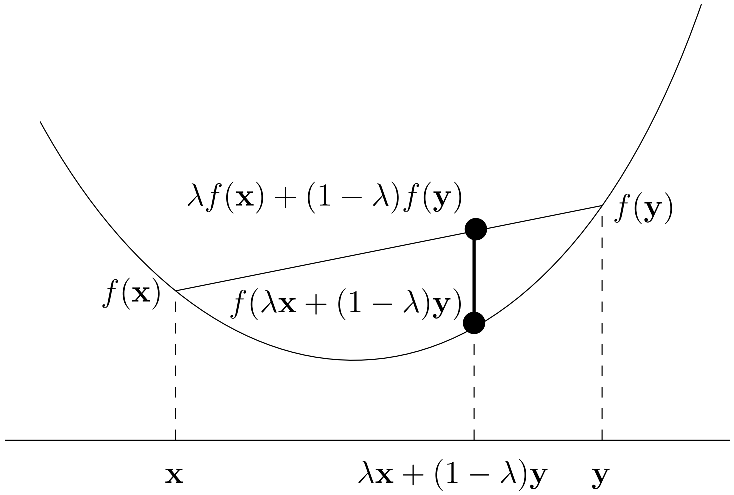

$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$
$f(\mathbf{y})$
$f(\mathbf{x})$
$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})$
$\mathbf{x}$
$\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$
$\mathbf{y}$

Figure 1.3: A convex function

While the graph of f is the set $\{(\mathbf{x}, f(\mathbf{x})) \in \mathbb{R}^{d+1} : \mathbf{x} \in \mathbf{dom}(f)\}$, the *epigraph* (Figure 1.4) is the set of points above the graph,

$$\mathbf{epi}(f) := \{(\mathbf{x}, \alpha) \in \mathbb{R}^{d+1} : \mathbf{x} \in \mathbf{dom}(f), \alpha \geq f(\mathbf{x})\}.$$

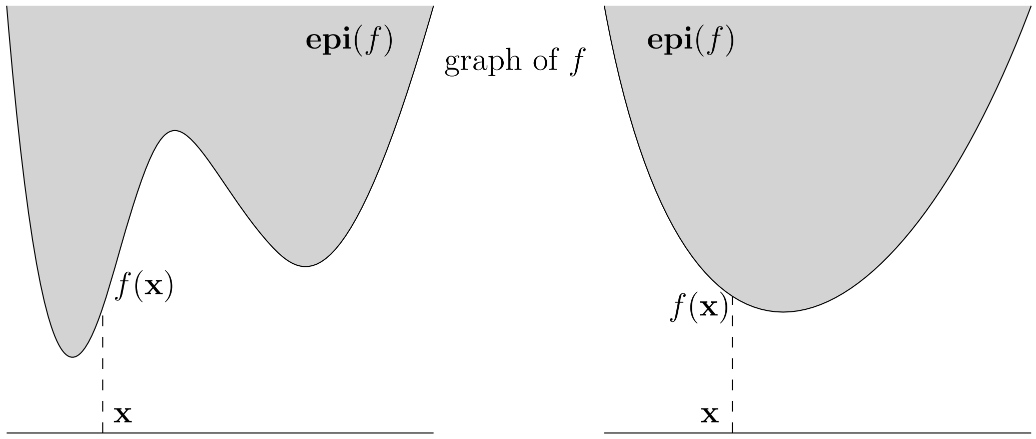


Figure 1.4: Graph and epigraph of a non-convex function (left) and a convex function (right)

Observation 1.11. *f is a convex function if and only if $\mathbf{epi}(f)$ is a convex set.*

Proof. This is easy but let us still do it to illustrate the concepts. Let f be a convex function and consider two points $(\mathbf{x}, \alpha), (\mathbf{y}, \beta) \in \mathbf{epi}(f)$, $\lambda \in [0, 1]$. This means, $f(\mathbf{x}) \le \alpha, f(\mathbf{y}) \le \beta$, hence by convexity of f,

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \le \lambda\alpha + (1-\lambda)\beta.$$

Therefore, by definition of the epigraph,

$$\lambda(\mathbf{x}, \alpha) + (1-\lambda)(\mathbf{y}, \beta) = (\lambda\mathbf{x} + (1-\lambda)\mathbf{y}, \lambda\alpha + (1-\lambda)\beta) \in \mathbf{epi}(f),$$

so $\mathbf{epi}(f)$ is a convex set. In the other direction, let $\mathbf{epi}(f)$ be a convex set and consider two points $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, $\lambda \in [0, 1]$. By convexity of $\mathbf{epi}(f)$, we have

$$\mathbf{epi}(f) \ni \lambda(\mathbf{x}, f(\mathbf{x})) + (1-\lambda)(\mathbf{y}, f(\mathbf{y})) = (\lambda\mathbf{x} + (1-\lambda)\mathbf{y}, \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})),$$

and this is just a different way of writing (1.2). □

Lemma 1.12 (Jensen's inequality)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbf{dom}(f)$, and $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^m \lambda_i = 1$. Then*

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \le \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

For $m = 2$, this is (1.2). The proof of the general case is Exercise 2.

Lemma 1.13. *Let f be convex and suppose that $\mathbf{dom}(f)$ is open. Then f is continuous.*

This is not entirely obvious (see Exercise 3) and really needs $\mathbf{dom}(f) \subseteq \mathbb{R}^d$. It becomes false if we consider convex functions over vector spaces of infinite dimension. In fact, in this case, even linear functions (which are in particular convex) may fail to be continuous.

Lemma 1.14. *There exists an (infinite dimensional) vector space V and a linear function $f : V \to \mathbb{R}$ such that f is discontinuous at all $\mathbf{v} \in V$.*

Proof. This is a classical example. Let us consider the vector space V of all univariate polynomials; the vector space operations are addition of two polynomials, and multiplication of a polynomial with a scalar. We consider a polynomial such as $3x^5 + 2x^2 + 1$ as a function $x \mapsto 3x^5 + 2x^2 + 1$ over the domain $[-1, 1]$.

The standard norm in a function space such as V is the *supremum norm* $\| \cdot \|_\infty$, defined for any bounded function $h : [-1, 1] \to \mathbb{R}$ via $\|h\|_\infty := \sup_{x \in [-1,1]} |h(x)|$. Polynomials are continuous and as such bounded over $[-1, 1]$.

We now consider the linear function $f : V \to \mathbb{R}$ defined by $f(p) = p'(0)$, the derivative of p at 0. The function f is linear, simply because the derivative is a linear operator. As $\mathbf{dom}(f)$ is the whole space V, $\mathbf{dom}(f)$ is open. We claim that f is discontinuous at 0 (the zero polynomial). Since f is linear, this implies discontinuity at every polynomial $p \in V$. To prove discontinuity at 0, we first observe that $f(0) = 0$ and then show that there are polynomials p of arbitrarily small supremum norm with $f(p) = 1$. Indeed, for $n, k \in \mathbb{N}, n > 0$, consider the polynomial

$$p_{n,k}(x) = \frac{1}{n} \sum_{i=0}^{k} (-1)^i \frac{(nx)^{2i+1}}{(2i+1)!} = \frac{1}{n} \left(nx - \frac{(nx)^3}{3!} + \frac{(nx)^5}{5!} - \cdots \pm \frac{(nx)^{2k+1}}{(2k+1)!} \right)$$

which—for any fixed n and sufficiently large k—approximates the function

$$s_n(x) = \frac{1}{n} \sin(nx) = \frac{1}{n} \sum_{i=0}^{\infty} (-1)^i \frac{(nx)^{2i+1}}{(2i+1)!}$$

up to any desired precision over the whole interval $[-1, 1]$ (Taylor's theorem with remainder). In formulas, $\|p_{n,k} - s_n\|_\infty \to 0$ as $k \to \infty$. Moreover, $\|s_n\|_\infty \to 0$ as $n \to \infty$. Using the triangle inequality, this implies that $\|p_{n,k}\| \to 0$ as $n, k \to \infty$. On the other hand, $f(p_{n,k}) = p'_{n,k}(0) = 1$ for all n, k. □

1.3.1 First-order characterization of convexity

As an example of a convex function, let us consider $f(x_1, x_2) = x_1^2 + x_2^2$. The graph of f is the *unit paraboloid* in $\mathbb{R}^3$ which looks convex. However, to verify (1.2) directly is somewhat cumbersome. Next, we develop better ways to do this if the function under consideration is differentiable.

Lemma 1.15 ([BV04, 3.1.3]). *Suppose that* $\mathbf{dom}(f)$ *is open and that* f *is differentiable; in particular, the* gradient *(vector of partial derivatives)*

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{x})\right)$$

exists at every point $\mathbf{x} \in \mathbf{dom}(f)$. *Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex and*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \tag{1.3}$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$.

Geometrically, this means that for all $\mathbf{x} \in \mathbf{dom}(f)$, the graph of f lies above its tangent hyperplane at the point $(\mathbf{x}, f(\mathbf{x}))$; see Figure 1.5.

Proof. Suppose that f is convex, meaning that for $t \in (0, 1)$,

$$f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) = f((1-t)\mathbf{x}+t\mathbf{y}) \leq (1-t)f(\mathbf{x})+tf(\mathbf{y}) = f(\mathbf{x})+t(f(\mathbf{y})-f(\mathbf{x})).$$

Dividing by t and using differentiability at $\mathbf{x}$, we get

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \frac{f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{t} \\ &= f(\mathbf{x}) + \frac{\nabla f(\mathbf{x})^\top t(\mathbf{y}-\mathbf{x}) + r(t(\mathbf{y}-\mathbf{x}))}{t} \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x}) + \frac{r(t(\mathbf{y}-\mathbf{x}))}{t}, \end{aligned}$$

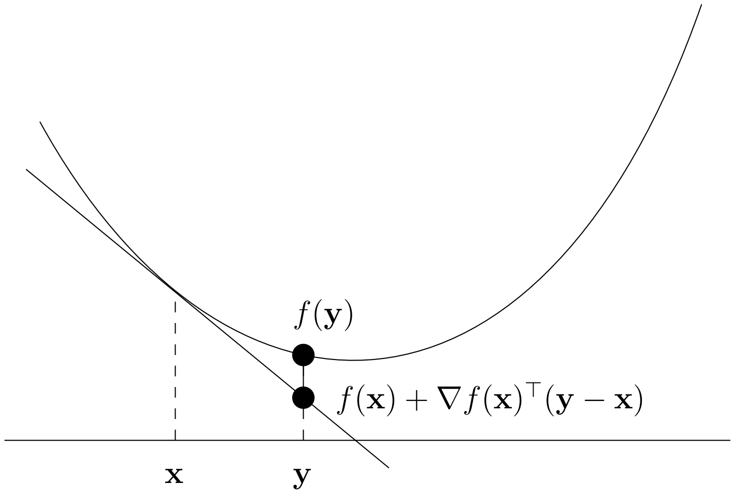

$f(\mathbf{y})$
$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$
$\mathbf{x}$
$\mathbf{y}$

Figure 1.5: First-order characterization of convexity

where the error term $r(t(\mathbf{y}-\mathbf{x}))/t$ goes to 0 as $t \to 0$. The inequality $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$ follows.

Now suppose this inequality holds for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, let $\lambda \in [0,1]$, and define $\mathbf{z} := \lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \mathbf{dom}(f)$ (by convexity of $\mathbf{dom}(f)$). Then we have

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

After multiplying the first inequality by λ and the second one by $(1-\lambda)$, the gradient terms cancel in the sum of the two inequalities, and we get

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\mathbf{z}) = f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}).$$

This is convexity. □

For $f(x_1, x_2) = x_1^2 + x_2^2$, we have $\nabla f(\mathbf{x}) = (2x_1, 2x_2)$, hence (1.3) boils down to

$$y_1^2 + y_2^2 \geq x_1^2 + x_2^2 + 2x_1(y_1 - x_1) + 2x_2(y_2 - x_2),$$

which after some rearranging of terms is equivalent to

$$(y_1 - x_1)^2 + (y_2 - x_2)^2 \geq 0,$$

hence true. There are relevant convex functions that are not differentiable, see Figure 1.6 for an example. More generally, Exercise 9 asks you to prove that the ℓ_1-norm (or 1-norm) $f(\mathbf{x}) = \|\mathbf{x}\|_1$ is convex.

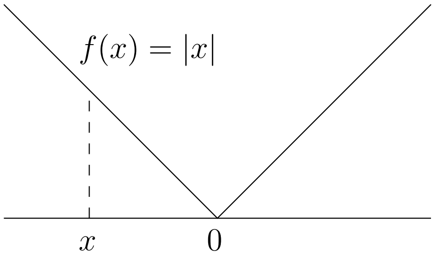


Figure 1.6: A non-differentiable convex function

There is another useful and less standard first-order characterization of convexity that we can easily derive from the standard one above.

Lemma 1.16. *Suppose that* $\mathbf{dom}(f)$ *is open and that* f *is differentiable. Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex and*

$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top(\mathbf{y} - \mathbf{x}) \geq 0 \tag{1.4}$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$.

The inequality (1.4) is known as *monotonicity of the gradient*.

Proof. If f is convex, the first-order characterization in Lemma 1.15 yields

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}), \\ f(\mathbf{x}) &\geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}), \end{aligned}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$. After adding up these two inequalities, $f(\mathbf{x}) + f(\mathbf{y})$ appears on both sides and hence cancels, so that we get

$$0 \geq \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \nabla f(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) = (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top(\mathbf{x} - \mathbf{y}).$$

Multiplying this by -1 yields (1.4).

For the other direction, suppose that monotonicty of the gradient (1.4) holds. Then we in particular have

$$(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top(t(\mathbf{y} - \mathbf{x})) \geq 0$$

for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ and $t \in (0,1)$. Dividing by t, this yields

$$(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})) \geq 0. \tag{1.5}$$

Fix $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$. For $t \in [0,1]$, let $h(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. In our case where f is real-valued, (1.1) yields $h'(t) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}), t \in (0,1)$. Hence, (1.5) can be rewritten as

$$h'(t) \geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \quad t \in (0,1).$$

By the mean value theorem, there is $c \in (0,1)$ such that $h'(c) = h(1) - h(0)$. Then

$$\begin{aligned} f(\mathbf{y}) = h(1) = h(0) + h'(c) &= f(\mathbf{x}) + h'(c) \\ &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}). \end{aligned}$$

This is the first-order characterization of convexity (Lemma 1.15). □

1.3.2 Second-order characterization of convexity

If $f : \mathbf{dom}(f) \to \mathbb{R}$ is twice differentiable (meaning that f is differentiable and the gradient function ∇f is also differentiable), convexity can be characterized as follows.

Lemma 1.17. *Suppose that* $\mathbf{dom}(f)$ *is open and that* f *is twice differentiable; in particular, the* Hessian *(matrix of second partial derivatives)*

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(\mathbf{x}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d}(\mathbf{x}) \end{pmatrix}$$

exists at every point $\mathbf{x} \in \mathbf{dom}(f)$ *and is symmetric. Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex, and for all* $\mathbf{x} \in \mathbf{dom}(f)$*, we have*

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{(i.e. } \nabla^2 f(\mathbf{x}) \textit{ is positive semidefinite).} \tag{1.6}$$

(A symmetric matrix M *is* positive semidefinite*, denoted by* $M \succeq \mathbf{0}$*, if* $\mathbf{x}^\top M \mathbf{x} \geq 0$ *for all* $\mathbf{x}$*, and* positive definite*, denoted by* $M \succ \mathbf{0}$*, if* $\mathbf{x}^\top M \mathbf{x} > 0$ *for all* $\mathbf{x} \neq \mathbf{0}$*.)*

The fact that the Hessians of a twice *continuously* differentiable function are symmetric is a classical result known as the Schwarz theorem [AE08, Corollary 5.5]. But symmetry in fact already holds if f is twice differentiable [Die69, (8.12.3)]. However, if f is only twice *partially* differentiable, we may get non-symmetric Hessians [AE08, Remark 5.6].

Proof. Once again, we employ our favorite univariate function $h(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, for fixed $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ and $t \in I$ where $I \supset [0, 1]$ is a suitable open interval. But this time, we also need h's second derivative. For $t \in I, \mathbf{v} := \mathbf{y} - \mathbf{x}$, we have

$$\begin{array}{rcl} h'(t) & = & \nabla f(\mathbf{x} + t\mathbf{v})^\top \mathbf{v}, \\ h''(t) & = & \mathbf{v}^\top \nabla^2 f(\mathbf{x} + t\mathbf{v})\mathbf{v}. \end{array}$$

The formula for $h'(t)$ has already been derived in the proof of Lemma 1.16, and the formula for $h''(t)$ is Exercise 10.

If f is convex, we always have $h''(0) \geq 0$, as we will show next. Given this, $\nabla^2 f(\mathbf{x}) \succeq 0$ follows for every $\mathbf{x} \in \mathbf{dom}(f)$: by openness of $\mathbf{dom}(f)$, for every $\mathbf{v} \in \mathbb{R}^d$ of sufficiently small norm, there is $\mathbf{y} \in \mathbf{dom}(f)$ such that $\mathbf{v} = \mathbf{y} - \mathbf{x}$, and then $\mathbf{v}^\top \nabla^2 f(\mathbf{x})\mathbf{v} = h''(0) \geq 0$. By scaling, this inequality extends to all $\mathbf{v} \in \mathbb{R}^d$.

To show $h''(0) \geq 0$, we observe that for all sufficiently small δ, $\mathbf{x} + \delta\mathbf{v} \in \mathbf{dom}(f)$ and hence

$$\frac{h'(\delta) - h'(0)}{\delta} = \frac{(\nabla f(\mathbf{x} + \delta\mathbf{v}) - \nabla f(\mathbf{x}))^\top \mathbf{v}}{\delta} = \frac{(\nabla f(\mathbf{x} + \delta\mathbf{v}) - \nabla f(\mathbf{x}))^\top \delta\mathbf{v}}{\delta^2} \geq 0,$$

by monotonicity of the gradient for convex f (Lemma 1.16). It follows that $h''(0) = \lim_{\delta \to 0}(h'(\delta) - h'(0))/\delta \geq 0$.

For the other direction, the mean value theorem applied to h' yields $c \in (0, 1)$ such that $h'(1) - h'(0) = h''(c)$, and spelled out, this is

$$\nabla f(\mathbf{y})^\top \mathbf{v} - \nabla f(\mathbf{x})^\top \mathbf{v} = \mathbf{v}^\top \nabla^2 f(\mathbf{x} + c\mathbf{v})\mathbf{v} \geq 0, \tag{1.7}$$

since $\nabla^2 f(\mathbf{z}) \succeq 0$ for all $\mathbf{z} \in \mathbf{dom}(f)$. Hence, we have proved monotonicity of the gradient which by Lemma 1.16 implies convexity of f. □

Geometrically, Lemma 1.17 means that the graph of f has non-negative curvature everywhere and hence "looks like a bowl". For $f(x_1, x_2) = x_1^2 + x_2^2$, we have

$$\nabla^2 f(\mathbf{x}) = \left(\begin{array}{cc} 2 & 0 \\ 0 & 2 \end{array}\right),$$

which is a positive definite matrix. In higher dimensions, the same argument can be used to show that the squared distance $d_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|^2$ to a fixed point $\mathbf{y}$ is a convex function; see Exercise 4. The non-squared Euclidean distance $\|\mathbf{x} - \mathbf{y}\|$ is also convex in $\mathbf{x}$, as a consequence of Lemma 1.18(ii) below and the fact that every seminorm (in particular the Euclidean norm $\|x\|$) is convex (Exercise 11). The squared Euclidean distance has the advantage that it is differentiable, while the Euclidean distance itself (whose graph is an "ice cream cone" for $d = 2$) is not.

1.3.3 Operations that preserve convexity

There are two important operations that preserve convexity.

Lemma 1.18 (Exercise 5)**.**

(i) Let $f_1, f_2, \ldots, f_m$ be convex functions, $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then $f := \sum_{i=1}^m \lambda_i f_i$ is convex on $\mathbf{dom}(f) := \bigcap_{i=1}^m \mathbf{dom}(f_i)$.

(ii) Let f be a convex function with $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, $g : \mathbb{R}^m \to \mathbb{R}^d$ an affine function, meaning that $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for some matrix $A \in \mathbb{R}^{d\times m}$ and some vector $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ (that maps $\mathbf{x}$ to $f(A\mathbf{x} + \mathbf{b})$) is convex on $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$.

1.4 Minimizing convex functions

The main feature that makes convex functions attractive in optimization is that every local minimum is a global one, so we cannot "get stuck" in local optima. This is quite intuitive if we think of the graph of a convex function as being bowl-shaped.

Definition 1.19. *A local minimum of $f : \mathbf{dom}(f) \to \mathbb{R}$ is a point $\mathbf{x}$ such that there exists $\varepsilon > 0$ with*

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f) \text{ satisfying } \|\mathbf{y} - \mathbf{x}\| < \varepsilon.$$

Lemma 1.20. *Let $\mathbf{x}^\star$ be a local minimum of a convex function $f : \mathbf{dom}(f) \to \mathbb{R}$. Then $\mathbf{x}^\star$ is a global minimum, meaning that*

$$f(\mathbf{x}^\star) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f).$$

Proof. Suppose there exists $\mathbf{y} \in \mathbf{dom}(f)$ such that $f(\mathbf{y}) < f(\mathbf{x}^\star)$ and define $\mathbf{y}' := \lambda\mathbf{x}^\star + (1-\lambda)\mathbf{y}$ for $\lambda \in (0,1)$. From convexity (1.2), we get that that $f(\mathbf{y}') < f(\mathbf{x}^\star)$. Choosing λ so close to 1 that $\|\mathbf{y}' - \mathbf{x}^\star\| < \varepsilon$ yields a contradiction to $\mathbf{x}^\star$ being a local minimum. □

This does not mean that a convex function always has a global minimum. Think of $f(x) = x$ as a trivial example. But also if f is bounded from below over $\mathbf{dom}(f)$, it may fail to have a global minimum ($f(x) = e^x$). To ensure the existence of a global minimum, we need additional conditions. For example, it suffices if outside some ball B, all function values are larger than some value $f(\mathbf{x}), \mathbf{x} \in B$. In this case, we can restrict f to B, without changing the smallest attainable value. And on B (which is compact), f attains a minimum by continuity (Lemma 1.13). An easy example: for $f(x_1, x_2) = x_1^2 + x_2^2$, we know that outside any ball containing $\mathbf{0}$, $f(\mathbf{x}) > f(\mathbf{0}) = 0$.

Another easy condition in the differentiable case is given by the following result.

Lemma 1.21. *Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$. Let $\mathbf{x} \in \mathbf{dom}(f)$. If $\nabla f(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is a global minimum.*

Proof. Suppose that $\nabla f(\mathbf{x}) = \mathbf{0}$. According to Lemma 1.15, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all $\mathbf{y} \in \mathbf{dom}(f)$, so $\mathbf{x}$ is a global minimum. □

The converse is also true and is a corollary of Lemma 1.27 [BV04, 4.2.3].

Lemma 1.22. *Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$. Let $\mathbf{x} \in \mathbf{dom}(f)$. If $\mathbf{x}$ is a global minimum then $\nabla f(\mathbf{x}) = \mathbf{0}$.*

1.4.1 Strictly convex functions

In general, a global minimum of a convex function is not unique (think of $f(x) = 0$ as a trivial example). However, if we forbid "flat" parts of the graph of f, a global minimum becomes unique (if it exists at all).

Definition 1.23 ([BV04, 3.1.1]). *A function $f : \mathbf{dom}(f) \to \mathbb{R}$ is* strictly convex *if (i) $\mathbf{dom}(f)$ is convex and (ii) for all $\mathbf{x} \neq \mathbf{y} \in \mathbf{dom}(f)$ and all $\lambda \in (0,1)$, we have*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}). \tag{1.8}$$

This means that the open line segment connecting $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$ is pointwise *strictly* above the graph of f. For example, $f(x) = x^2$ is strictly convex.

Lemma 1.24 ([BV04, 3.1.4]). *Suppose that $\mathbf{dom}(f)$ is open and that f is twice continuously differentiable. If the Hessian $\nabla^2 f(\mathbf{x}) \succ \mathbf{0}$ for every $x \in \mathbf{dom}(f)$ (i.e., $\mathbf{z}^\top \nabla^2 f(\mathbf{x})\mathbf{z} > 0$ for any $\mathbf{z} \neq \mathbf{0}$), then f is strictly convex.*

The converse is false, though: $f(x) = x^4$ is strictly convex but has vanishing second derivative at $x = 0$.

Lemma 1.25. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be strictly convex. Then f has at most one global minimum.*

Proof. Suppose $\mathbf{x}^\star \neq \mathbf{y}^\star$ are two global minima with $f_{\min} = f(\mathbf{x}^\star) = f(\mathbf{y}^\star)$, and let $\mathbf{z} = \frac{1}{2}\mathbf{x}^\star + \frac{1}{2}\mathbf{y}^\star$. By (1.8),

$$f(\mathbf{z}) < \frac{1}{2}f_{\min} + \frac{1}{2}f_{\min} = f_{\min},$$

a contradiction to $\mathbf{x}^\star$ and $\mathbf{y}^\star$ being global minima. □

1.4.2 Example: Least squares

Suppose we want to fit a hyperplane to a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ in $\mathbb{R}^d$, based on the hypothesis that the points actually come (approximately) from a hyperplane. A classical method for this is *least squares*. For concreteness, let us do this in $\mathbb{R}^2$. Suppose that the data points are

$$(1,10), (2,11), (3,11), (4,10), (5,9), (6,10), (7,9), (8,10),$$

Figure 1.7 (left).

Also, for simplicity (and quite appropriately in this case), let us restrict to fitting a linear model, of more formally to fit non-vertical lines of the

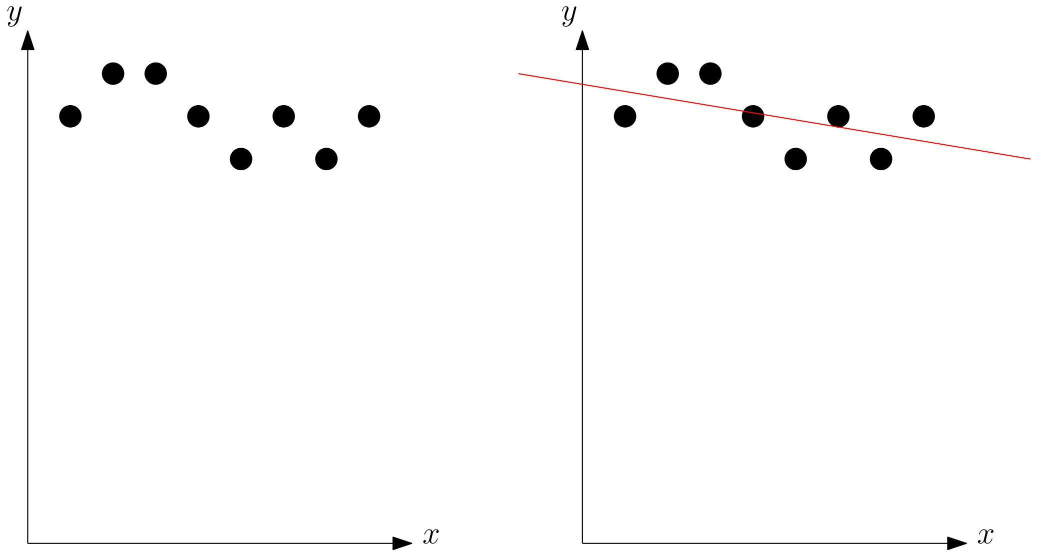


Figure 1.7: Data points in $\mathbb{R}^2$ (left) and least-squares fit (right)

form $y = w_0 + w_1 x$. If (x_i, y_i) is the i-th data point, the least squares fit chooses w_0, w_1 such that the *least squares objective*

$$f(w_0, w_1) = \sum_{i=1}^{8} (w_1 x_i + w_0 - y_i)^2$$

is minimized. It easily follows from Lemma 1.18 that f is convex. In fact,

$$f(w_0, w_1) = 204w_1^2 + 72w_1w_0 - 706w_1 + 8w_0^2 - 160w_0 + 804, \qquad (1.9)$$

so we can check convexity directly using the second order condition. We have gradient

$$\nabla f(w_0, w_1) = (72w_1 + 16w_0 - 160, 408w_1 + 72w_0 - 706)$$

and Hessian

$$\nabla^2(w_0, w_1) = \begin{pmatrix} 16 & 72 \\ 72 & 408 \end{pmatrix}.$$

A 2×2 matrix is positive semidefinite if the diagonal elements and the determinant are positive, which is the case here, so f is actually strictly convex and has a unique global minimum. To find it, we solve the linear system $\nabla f(w_0, w_1) = (0, 0)$ of two equations in two unknowns and obtain the global minimum

$$(w_0^\star, w_1^\star) = \left(\frac{43}{4}, -\frac{1}{6}\right).$$

Hence, the "optimal" line is

$$y = -\frac{1}{6}x + \frac{43}{4},$$

see Figure 1.7 (right).

1.4.3 Constrained Minimization

Frequently, we are interested in minimizing a convex function only over a subset X of its domain.

Definition 1.26. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and let $X \subseteq \mathbf{dom}(f)$ be a convex set. A point $\mathbf{x} \in X$ is a* minimizer *of f* over X *if*

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in X.$$

If f is differentiable, minimizers of f over X have a very useful characterization.

Lemma 1.27 ([BV04, 4.2.3]). *Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, and let $X \subseteq \mathbf{dom}(f)$ be a convex set. Point $\mathbf{x}^\star \in X$ is a minimizer of f over X if and only if*

$$\nabla f(\mathbf{x}^\star)^\top(\mathbf{x} - \mathbf{x}^\star) \geq 0 \quad \forall \mathbf{x} \in X.$$

Applying the this result with $X = \mathbf{dom}(f)$, we recover Lemma 1.21, and because $\mathbf{dom}(f)$ is open, its converse Lemma 1.22 follows [BV04, 4.2.3]. If X does not contain the global minimum, then Lemma 1.27 has a nice geometric interpretation. Namely, it means that X is contained in the halfspace $\{\mathbf{x} \in \mathbb{R}^d : \nabla f(\mathbf{x}^\star)^\top(\mathbf{x} - \mathbf{x}^\star) \geq 0\}$ (normal vector $\nabla f(\mathbf{x}^\star)$ pointing into the halfspace); see Figure 1.8. In still other words, $\mathbf{x} - \mathbf{x}^\star$ forms a non-obtuse angle with $\nabla f(\mathbf{x}^\star)$ for all $\mathbf{x} \in X$.

We typically write constrained minimization problems in the form

$$\operatorname{argmin}\{f(\mathbf{x}) : \mathbf{x} \in X\} \tag{1.10}$$

or

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X. \end{array} \tag{1.11}$$

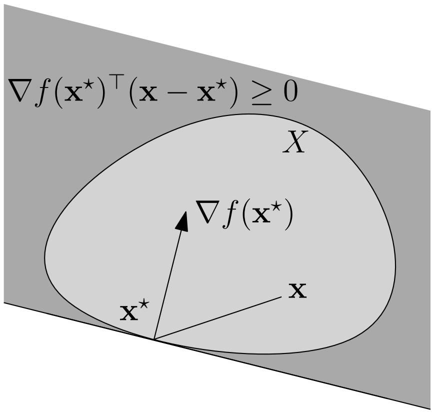

$\nabla f(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) \geq 0$
X
$\nabla f(\mathbf{x}^\star)$
$\mathbf{x}$
$\mathbf{x}^\star$

Figure 1.8: Optimality condition for constrained optimization

1.5 Existence of a minimizer

The existence of a minimizer (or a global minimum if $X = \mathbf{dom}(f)$) will be an assumption made by most minimization algorithms that we discuss later. In practice, such algorithms are being used (and often also work) if there is no minimizer. By "work", we mean in this case that they compute a point $\mathbf{x}$ such that $f(\mathbf{x})$ is close to $\inf_{\mathbf{y}\in X} f(\mathbf{y})$, assuming that the infimum is finite (as in $f(x) = e^x$). But a sound theoretical analysis usually requires the existence of a minimizer. Therefore, this section develops tools that may helps us in analyzing whether this is the case for a given convex function. To avoid technicalities, we restrict ourselves to the case $\mathbf{dom}(f) = \mathbb{R}^d$.

1.5.1 Sublevel sets and the Weierstrass Theorem

Definition 1.28. *Let $f : \mathbb{R}^d \to \mathbb{R}$, $\alpha \in \mathbb{R}$. The set*

$$f^{\leq\alpha} := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \leq \alpha\}$$

is the α-sublevel set of f; see Figure 1.9

It is easy to see from the definition that every sublevel set of a convex function is convex. Moreover, as a consequence of continuity of f, sublevel sets are closed. The following (known as the Weierstrass Theorem) just formalizes an argument that we have made earlier.

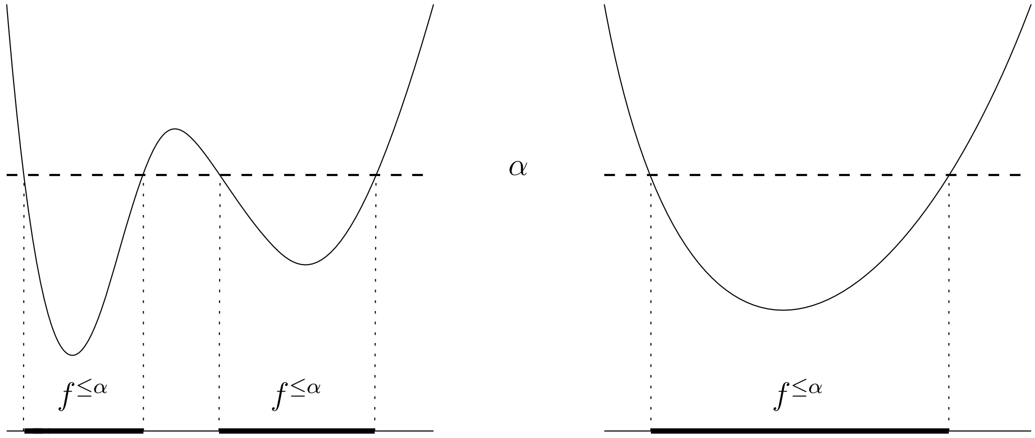


Figure 1.9: Sublevel set of a non-convex function (left) and a convex function (right)

Theorem 1.29. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and suppose there is a nonempty and bounded sublevel set $f^{\leq\alpha}$. Then f has a global minimum.*

Proof. We know that f—as a continuous function—attains a minimum over the closed and bounded (= compact) set $f^{\leq\alpha}$ at some $\mathbf{x}^\star$. This $\mathbf{x}^\star$ is also a global minimum as it has value $f(\mathbf{x}^\star) \leq \alpha$, while any $\mathbf{x} \notin f^{\leq\alpha}$ has value $f(\mathbf{x}) > \alpha \geq f(\mathbf{x}^\star)$. □

1.5.2 Recession cone and lineality space

What happens if there is no bounded sublevel set? Then f may ($f(x) = 0$) or may not ($f(x) = x, f(x) = e^x$) have a global minimum. But what cannot happen is that some nonempty sublevel sets are bounded, and others are not. Also, in the unbounded case, all nonempty sublevel sets have the same "reason" for being unbounded, namely the same *directions of recession*. Again, we assume for simplicity that $\mathbf{dom}(f) = \mathbb{R}^d$, and we are only considering unconstrained minimization. But most of the material can be adapted to general domains and to constrained optimization over a convex set $X \subseteq \mathbf{dom}(f)$. We closely follow Bertsekas [Ber05].

Recession cone and lineality space of a convex set.

Definition 1.30. *Let $C \subseteq \mathbb{R}^d$ be a convex set. Then $\mathbf{y} \in \mathbb{R}^d$ is a* direction of recession *of C if for some $\mathbf{x} \in C$ and all $\lambda \geq 0$, it holds that $\mathbf{x} + \lambda\mathbf{y} \in C$.*

This means that C is unbounded in direction $\mathbf{y}$. Whether $\mathbf{y}$ is a direction of recession only depends on C and not on the particular $\mathbf{x}$, assuming that C is closed (otherwise, it may be false).

Lemma 1.31. *Let $C \subseteq \mathbb{R}^d$ be a nonempty closed convex set, and let $\mathbf{y} \in \mathbb{R}^d$. The following statements are equivalent.*

(i) $\exists \mathbf{x} \in C : \mathbf{x} + \lambda\mathbf{y} \in C$ *for all* $\lambda \geq 0$.

(ii) $\forall \mathbf{x} \in C : \mathbf{x} + \lambda\mathbf{y} \in C$ *for all* $\lambda \geq 0$.

Proof. We need to show that (i) implies (ii), so choose $\mathbf{x} \in C, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{x} + \lambda\mathbf{y} \in C$ for all $\lambda \geq 0$. Fix $\lambda > 0$, let $\mathbf{z} = \lambda\mathbf{y}$ and $\mathbf{x}' \in C$. To get (ii), we prove that $\mathbf{x}' + \mathbf{z} \in C$. To this end, we define sequences $(\mathbf{w}_k), (\mathbf{z}_k), k \in \mathbb{N}$ via

$$\begin{aligned} \mathbf{w}_k &:= \mathbf{x} + k\mathbf{z} \in C \quad \text{(by (i))} \\ \mathbf{z}_k &:= \frac{1}{k}(\mathbf{w}_k - \mathbf{x}') = \mathbf{z} + \frac{1}{k}(\mathbf{x} - \mathbf{x}'), \end{aligned}$$

see Figure 1.10. By definition of a convex set, we have $\mathbf{x}' + \mathbf{z}_k = \frac{1}{k}\mathbf{w}_k + \left(1 - \frac{1}{k}\right)\mathbf{x}' \in C$. Moreover, $\mathbf{z}_k$ converges to $\mathbf{z}$, so $\mathbf{x}' + \mathbf{z}_k$ converges to $\mathbf{x}' + \mathbf{z} \in C$, and this is an element of C, since C is closed. □

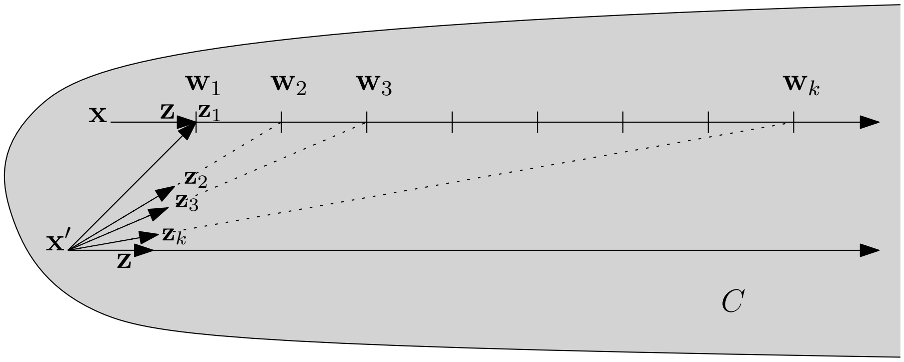


Figure 1.10: The proof of Lemma 1.31

The directions of recession of C actually form a *convex cone*, a set that is closed under taking non-negative linear combinations. This is known as the *recession cone* $R(C)$ of C; see Figure 1.11 (left).

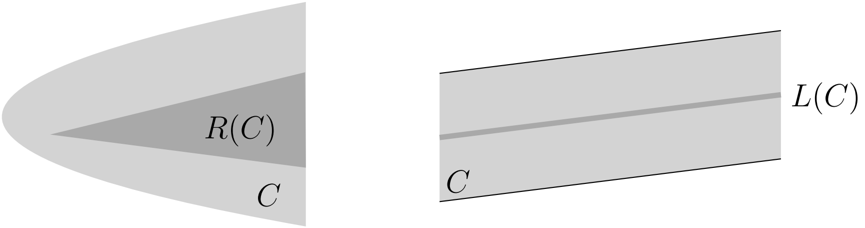


Figure 1.11: The recession cone and lineality space of a convex set

Lemma 1.32. *Let $C \subseteq \mathbb{R}^d$ be a closed convex set, and let $\mathbf{y}_1, \mathbf{y}_2$ be directions of recession of C; $\lambda_1, \lambda_2 \in \mathbb{R}_+$. Then $\mathbf{y} = \lambda_1\mathbf{y}_1 + \lambda_2\mathbf{y}_2$ is also a direction of recession of C.*

Proof. The statement is trivial if $\mathbf{y} = \mathbf{0}$. Otherwise, after scaling $\mathbf{y}$ by $1/(\lambda_1 + \lambda_2) > 0$ (which does not affect the property of being a direction of recession), we may assume that $\lambda_1 + \lambda_2 = 1$. Now, for all $\mathbf{x} \in C$ and all $\lambda \in \mathbb{R}$, we get

$$\mathbf{x} + \lambda\mathbf{y} = \mathbf{x} + \lambda_1\lambda\mathbf{y}_1 + \lambda_2\lambda\mathbf{y}_2 = \lambda_1(\underbrace{\mathbf{x} + \lambda\mathbf{y}_1}_{\in C}) + \lambda_2(\underbrace{\mathbf{x} + \lambda\mathbf{y}_2}_{\in C}) \in C.$$

□

Definition 1.33. *Let $C \subseteq \mathbb{R}^d$ be a convex set. $\mathbf{y} \in \mathbb{R}^d$ is a* direction of constancy *of C if both $\mathbf{y}$ and $-\mathbf{y}$ are directions of recession of C.*

This means that C is unbounded along the whole line spanned by $\mathbf{y}$. The directions of constancy form a linear subspace, the *lineality space $L(C)$* of C; see Figure 1.11 (right).

Lemma 1.34. *Let $C \subseteq \mathbb{R}^d$ be a closed convex set, and let $\mathbf{y}_1, \mathbf{y}_2$ be directions of constancy of C; $\lambda_1, \lambda_2 \in \mathbb{R}$. Then $\mathbf{y} = \lambda_1\mathbf{y}_1 + \lambda_2\mathbf{y}_2$ is also a direction of constancy of C.*

Proof. After replacing $\mathbf{y}_i$ with the direction of recession $-\mathbf{y}_i$ if necessary ($i = 1, 2$), we may assume that $\lambda_1, \lambda_2 \geq 0$, so $\mathbf{y}$ is a direction of recession by Lemma 1.32. The same argument works for $-\mathbf{y}$, so $\mathbf{y}$ is a direction of constancy. □

Recession cone and lineality space of a convex function. For this, we look at directions of recession of sublevel sets (which are closed and convex in our case).

Lemma 1.35. *Let $f : \mathbb{R}^d \to R$ be a convex function. Any two nonempty sublevel sets $f^{\leq \alpha}, f^{\leq \alpha'}$ have the same recession cones, i.e. $R(f^{\leq \alpha}) = R(f^{\leq \alpha'})$.*

Proof. Let $\mathbf{y}$ be a direction of recession for $f^{\leq \alpha}$, i.e. for all $\mathbf{x} \in f^{\leq \alpha}$ and all $\lambda \geq 0$, we have

$$f(\mathbf{x} + \lambda \mathbf{y}) \leq \alpha.$$

We claim that this implies the stronger bound

$$f(\mathbf{x} + \lambda \mathbf{y}) \leq f(\mathbf{x}). \tag{1.12}$$

In words, f is non-increasing along any direction of recession. Using this, the statement follows: Because $f^{\leq \alpha}$ and $f^{\leq \alpha'}$ are nonempty, there exists $\mathbf{x}' \in f^{\leq \alpha} \cap f^{\leq \alpha'}$, and then we have $f(\mathbf{x}' + \lambda \mathbf{y}) \leq f(\mathbf{x}') \leq \alpha'$, so $\mathbf{y}$ is a direction of recession for $f^{\leq \alpha'}$.

To prove (1.12), we fix λ and let $\mathbf{z} = \lambda \mathbf{y}$. With $\mathbf{w}_k := \mathbf{x} + k\mathbf{z} \in f^{\leq \alpha}$, we have

$$\mathbf{x} + \mathbf{z} = \left(1 - \frac{1}{k}\right) \mathbf{x} + \frac{1}{k} \mathbf{w}_k,$$

so convexity of f and the fact that $\mathbf{w}_k \in f^{\leq \alpha}$ yields

$$f(\mathbf{x} + \mathbf{z}) \leq \left(1 - \frac{1}{k}\right) f(\mathbf{x}) + \frac{1}{k} f(\mathbf{w}_k) \leq \left(1 - \frac{1}{k}\right) f(\mathbf{x}) + \frac{1}{k}\alpha. \tag{1.13}$$

Thus, as $k \to \infty$, the right side of (1.13) tends to $f(x)$, and therefore $f(\mathbf{x} + \mathbf{z}) \leq f(\mathbf{x})$; see Figure 1.12. □

Definition 1.36. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then $\mathbf{y} \in \mathbb{R}^d$ is a direction of recession (of constancy, respectively) of f if $\mathbf{y}$ is a direction of recession (of constancy, respectively) for some (equivalently, for every) nonempty sublevel set. The set of directions of recession of f is called the* recession cone $R(f)$ *of f. The set of directions of constancy of f is called the* lineality space $L(f)$ *of f.*

We can characterize recession cone and lineality space of f directly, without looking at sublevel sets (the proof is Exercise 6). The conditions of Lemma 1.37(ii) and Lemma 1.38(ii) finally explain the terms "direction of recession" and "direction of constancy".

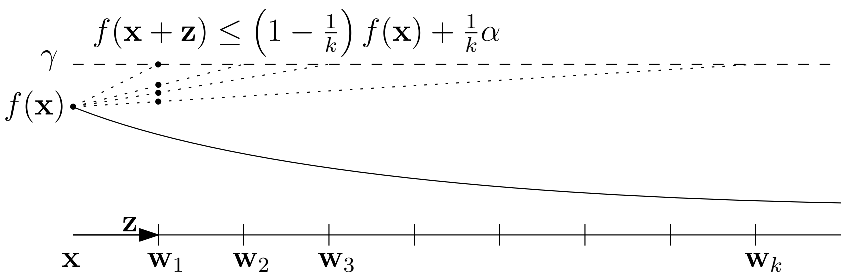


Figure 1.12: The proof of (1.12)

Lemma 1.37. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The following statements are equivalent.*

(i) $\mathbf{y} \in \mathbb{R}^d$ *is a direction of recession of f.*

(ii) $f(\mathbf{x} + \lambda\mathbf{y}) \leq f(\mathbf{x})$ *for all* $\mathbf{x} \in \mathbb{R}^d$ *and all* $\lambda \in \mathbb{R}_+$.

(iii) $(\mathbf{y}, 0)$ *is a ("horizontal") direction of recession of (the closed convex set)* $\mathbf{epi}(f)$.

Lemma 1.38. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. The following statements are equivalent.*

(i) $\mathbf{y} \in \mathbb{R}^d$ *is a direction of constancy of f.*

(ii) $f(\mathbf{x} + \lambda\mathbf{y}) = f(\mathbf{x})$ *for all* $\mathbf{x} \in \mathbb{R}^d$ *and all* $\lambda \in \mathbb{R}$.

(iii) $(\mathbf{y}, 0)$ *is a ("horizontal") direction of constancy of (the closed convex set)* $\mathbf{epi}(f)$.

1.5.3 Coercive convex functions

Definition 1.39. *A convex function f is* coercive *if its recession cone is trivial, meaning that* $\mathbf{0}$ *is its only direction of recession.*[1]

[1] The usual definition of a coercive function is that $f(\mathbf{x}) \to \infty$ whenever $\|\mathbf{x}\| \to \infty$. In the convex case, both definitions agree.

Coercivity means that along any direction, $f(\mathbf{x})$ goes to infinity. An example of a coercive convex function is $f(x_1, x_2) = x_1^2 + x_2^2$. Non-coercive functions are $f(x) = x$ and $f(x) = e^x$ (any $y \leq 0$ is a direction of recession). For a constant function $f : \mathbb{R}^d \to \mathbb{R}$, every direction $\mathbf{y}$ is a direction of recession. In general, affine functions are never coercive.

Lemma 1.40. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a coercive convex function. Then every nonempty sublevel set $f^{\leq\alpha}$ is bounded.*

This may seem obvious, as $f^{\leq\alpha}$ is bounded in every direction by coercivity. But we still need an argument that there is a global bound.

Proof. Let $f^{\leq\alpha}$ be a nonempty sublevel, and assume without loss of generality that $\mathbf{0} \in f^{\leq\alpha}$, i.e., $\alpha \geq f(\mathbf{0})$.

Let $S^{d-1} = \{\mathbf{y} \in \mathbb{R}^d : \|y\| = 1\}$ be the unit sphere. We define a function $g : S^{d-1} \to \mathbb{R}$ via

$$g(\mathbf{y}) = \max\{\lambda \geq 0 : f(\lambda\mathbf{y}) \leq \alpha\}, \quad \mathbf{y} \in S^{d-1}.$$

Since f is continuous and has no nonzero direction of recession, we know that for each $\mathbf{y} \in S^{d-1}$ the set $\{\lambda \geq 0 : f(\lambda\mathbf{y}) \leq \alpha\}$ is closed and bounded (it is actually an interval, by convexity of f), so the maximum exists and $g(\mathbf{y})$ is well-defined. We claim that g is continuous.

Let $(\mathbf{y}'_k)_{k\in\mathbb{N}}$ be a sequence of unit vectors such that $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{y} \in S^{d-1}$. We need to show that $\lim_{k\to\infty} g(\mathbf{y}') = g(\mathbf{y})$. Let us fix $\varepsilon > 0$ arbitrarily small. For $\underline{\lambda} := g(\mathbf{y}) - \varepsilon \geq 0$, we have $f(\underline{\lambda}\mathbf{y}) \leq \alpha$ (an easy consequence of convexity of f and $\alpha \geq f(\mathbf{0})$). And for $\overline{\lambda} := g(\mathbf{y}) + \varepsilon$, we get $f(\overline{\lambda}\mathbf{y}) > \alpha$ by definition of $g(\mathbf{y})$. Continuity of f then yields $\lim_{k\to\infty} f(\underline{\lambda}\mathbf{y}_k) = f(\underline{\lambda}\mathbf{y}) \leq \alpha$ and $\lim_{k\to\infty} f(\overline{\lambda}\mathbf{y}_k) = f(\overline{\lambda}\mathbf{y}) > \alpha$. Hence, for sufficiently large k, $g(\mathbf{y}_k) \in [\underline{\lambda}, \overline{\lambda}] = [g(\mathbf{y}) - \varepsilon, g(\mathbf{y}) + \varepsilon]$, and $\lim_{k\to\infty} g(\mathbf{y}') = g(\mathbf{y})$ follows.

As a continuous function, g attains a maximum $\lambda^\star$ over the compact set S^{d-1}, and this means that $f^{\leq\alpha}$ is contained in the closed $\lambda^\star$-ball around the origin. Hence, $f^{\leq\alpha}$ is bounded. □

Together with Theorem 1.29, we obtain

Theorem 1.41. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a coercive convex function. Then f has a global minimum.*

1.5.4 Weakly coercive convex functions

It turns out that we can allow nontrivial directions of recession and still guarantee a global minimum.

Definition 1.42. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function. Function f is called* weakly coercive *if its recession cone equals its lineality space, i.e. every direction of recession is a direction of constancy.*

Function $f(x) = 0$ is a trivial example of a non-coercive but weakly coercive function. A more interesting example is $f(x_1, x_2) = x_1^2$. Here, the directions of recession are all vectors of the form $\mathbf{y} = (0, x_2)$, and these are at the same time directions of constancy.

Theorem 1.43. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a weakly coercive convex function. Then f has a global minimum.*

Proof. We know that the lineality space L of f is a linear subspace of $\mathbb{R}^d$: by Definition 1.36, L is the lineality space of every nonempty (closed and convex) sublevel set, and as such it is closed under taking linear combinations (Lemma 1.34). Let $L^\perp$ be the orthogonal complement of L. Restricted to $L^\perp$, f is coercive, as $L^\perp$ is orthogonal to any direction of constancy, equivalently to every direction of recession, since f is weakly coercive. Therefore, $L^\perp$ can contain only the trivial direction of recession. It follows that $f_{|L^\perp}$ has a global minimum $\mathbf{x}^\star \in L^\perp$ by Theorem 1.41 (which we can apply after identifying $L^\perp$ w.l.o.g. with $\mathbb{R}^m$ for some $m \leq n$). This is also a global minimum of f. To see this, let $\mathbf{z} \in \mathbb{R}^d$ and write it (uniquely) in the form $\mathbf{z} = \mathbf{x} + \mathbf{y}$ with $\mathbf{x} \in L^\perp$ and $\mathbf{y} \in L$. Then we get

$$f(\mathbf{z}) = f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) \geq f(\mathbf{x}^\star),$$

where the second equality follows from $\mathbf{y}$ being a direction of constancy; see Lemma 1.38(ii). □

1.6 Examples

In the following two sections, we give two examples of convex function minimization tasks that arise from machine learning applications.

1.6.1 Handwritten digit recognition

Suppose you want to write a program that recognizes handwritten decimal digits $0, 1, \ldots, 9$. You have a set P of grayscale images (28×28 pixels, say) that represent handwritten decimal digits, and for each image $\mathbf{x} \in P$, you know the digit $d(\mathbf{x}) \in \{0, \ldots, 9\}$ that it represents, see Figure 1.13. You want to train your program with the set P, and after that, use it to recognize handwritten digits in arbitrary 28×28 images.

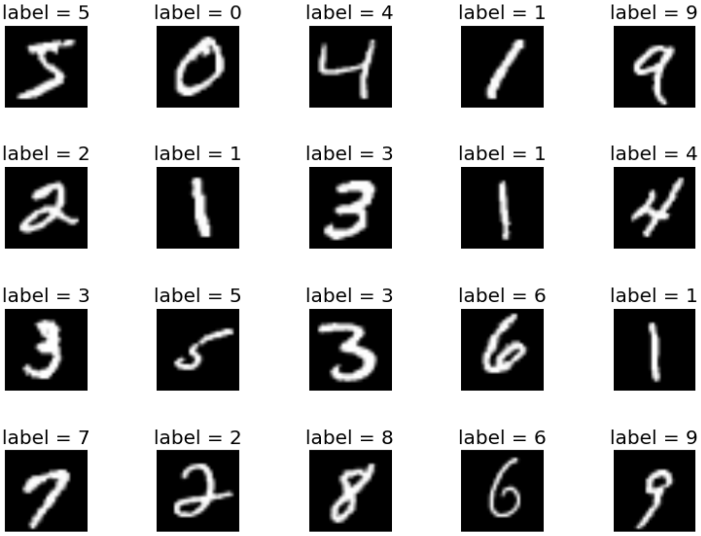


Figure 1.13: Some training images from the MNIST data set (picture from `http://corochann.com/mnist-dataset-introduction-1138.html`

The classical approach is the following. We represent an image as a *feature vector* $\mathbf{x} \in \mathbb{R}^{784}$, where x_i is the gray value of the i-th pixel (in some order). During the training phase, we compute a matrix $W \in \mathbb{R}^{10 \times 784}$ and then use the vector $\mathbf{y} = W\mathbf{x} \in \mathbb{R}^{10}$ to predict the digit seen in an arbitrary image $\mathbf{x}$. The idea is that $y_j, j = 0, \ldots, 9$ corresponds to the probability of the digit being j. This does not work directly, since the entries of $\mathbf{y}$ may be negative and generally do not sum up to 1. But we can convert $\mathbf{y}$ to a vector $\mathbf{z}$ of actual probabilities, such that a small y_j leads to a small probability z_j and a large y_j to a large probability z_j. How to do this is not canonical, but here is a well-known formula that works:

$$z_j = z_j(\mathbf{y}) = \frac{e^{y_j}}{\sum_{k=0}^{9} e^{y_k}}. \tag{1.14}$$

The classification then simply outputs digit j with probability z_j. The matrix W is chosen such that it (approximately) minimizes the classification error on the training set P. Again, it is not canonical how we measure classification error; here we use the following *loss function* to evaluate the error induced by a given matrix W.

$$\ell(W) = -\sum_{\mathbf{x}\in P} \ln\left(z_{d(\mathbf{x})}(W\mathbf{x})\right) = \sum_{\mathbf{x}\in P}\left(\ln\left(\sum_{k=0}^{9} e^{(W\mathbf{x})_k}\right) - (W\mathbf{x})_{d(\mathbf{x})}\right). \tag{1.15}$$

This function "punishes" images for which the correct digit j has low probability z_j (corresponding to a significantly negative value of $\log z_j$). In an ideal world, the correct digit would always have probability 1, resulting in $\ell(W) = 0$. But under (1.14), probabilities are always strictly between 0 and 1, so we have $\ell(W) > 0$ for all W.

Exercise 7 asks you to prove that ℓ is convex. In Exercise 8, you will characterize the situations in which ℓ has a global minimum.

1.6.2 Master's Admission

The computer science department of a well known Swiss university is admitting top international students to its MSc program, in a competitive application process. Applicants are submitting various documents (GPA, TOEFL test score, GRE test scores, reference letters,...). During the evaluation of an application, the admission committee would like to compute a (rough) forecast of the applicant's performance in the MSc program, based on the submitted documents.[2]

Data on the actual performance of students admitted in the past is available. To keep things simple in the following example, Let us base the forecast on GPA (grade point average) and TOEFL (Test of English as a Foreign Language) only. GPA scores are normalized to a scale with a

[2] Any resemblance to real departments is purely coincidental. Also, no serious department will base performance forecasts on data from 10 students, as we will do it here.

minimum of 0.0 and a maximum of 4.0, where admission starts from 3.5. TOEFL scores are on an integer scale between 0 and 120, where admission starts from 100.

Table 1.1 contains the known data. GGPA (graduation grade point average on a Swiss grading scale) is the average grade obtained by an admitted student over all courses in the MSc program. The Swiss scale goes from 1 to 6 where 1 is the lowest grade, 6 is the highest, and 4 is the lowest passing grade.

GPA	TOEFL	GGPA
3.52	100	3.92
3.66	109	4.34
3.76	113	4.80
3.74	100	4.67
3.93	100	5.52
3.88	115	5.44
3.77	115	5.04
3.66	107	4.73
3.87	106	5.03
3.84	107	5.06

Table 1.1: Data for 10 admitted students: GPA and TOEFL scores (at time of application), GGPA (at time of graduation)

As in Section 1.4.2, we are attempting a linear regression with least squares fit, i.e. we are making the hypothesis that

$$\text{GGPA} \approx w_0 + w_1 \cdot \text{GPA} + w_2 \cdot \text{TOEFL}. \tag{1.16}$$

However, in our scenario, the relevant GPA scores span a range of only 0.5 while the relevant TOEFL scores span a range of 20. The resulting least squares objective would be somewhat ugly; we already saw this in our previous example (1.9), where the data points had large second coordinate, resulting in the w_1-scale being very different from the w_2-scale. This time, we normalize first, so that w_1 und w_2 become comparable and allow us to understand the relative influences of GPA and TOEFL.

The general setting is this: we have n *inputs* $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where each vector $\mathbf{x}_i \in \mathbb{R}^d$ consists of d input variables; then we have n *outputs* $y_1, \ldots, y_n \in$

$\mathbb{R}$. Each pair $(\mathbf{x}_i, y_i)$ is an *observation*. In our case, $d = 2, n = 10$, and for example, $((3.93, 100), 5.52)$ is an observation (of a student doing very well).

With variable *weights* $w_0, \mathbf{w} = (w_1, \ldots, w_d) \in \mathbb{R}^d$, we plan to minimize the least squares objective

$$f(w_0, \mathbf{w}) = \sum_{i=1}^{n} (w_0 + \mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

We first want to assume that the inputs and outputs are *centered*, meaning that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}, \quad \frac{1}{n} \sum_{i=1}^{n} y_i = 0.$$

This can be achieved by simply subtracting the mean $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ from every input and the mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ from every output. In our example, this yields the numbers in Table 1.2 (left).

GPA	TOEFL	GGPA
-0.24	-7.2	-0.94
-0.10	1.8	-0.52
-0.01	5.8	-0.05
-0.02	-7.2	-0.18
0.17	-7.2	0.67
0.12	7.8	0.59
0.01	7.8	0.19
-0.10	-0.2	-0.12
0.11	-1.2	0.17
0.07	-0.2	0.21

GPA	TOEFL	GGPA
-2.04	-1.28	-0.94
-0.88	0.32	-0.52
-0.05	1.03	-0.05
-0.16	-1.28	-0.18
1.42	-1.28	0.67
1.02	1.39	0.59
0.06	1.39	0.19
-0.88	-0.04	-0.12
0.89	-0.21	0.17
0.62	-0.04	0.21

Table 1.2: Centered observations (left); normalized inputs (right)

After centering, the global minimum $(w_0^\star, \mathbf{w}^\star)$ of the least squares objective satisfies $w_0^\star = 0$ while $\mathbf{w}^\star$ is unaffected by centering (Exercise 12), so that we can simply omit the variable w_0 in the sequel.

Finally, we assume that all d input variables are on the same scale, meaning that

$$\frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, d.$$

To achieve this for fixed j (assuming that no variable is 0 in all inputs), we multiply all x_{ij} by $s(j) = \sqrt{n/\sum_{i=1}^n x_{ij}^2}$ (which, in the optimal solution $\mathbf{w}^*$, just multiplies $w_j^\star$ by $1/s(j)$, an argument very similar to the one in Exercise 12). For our data set, the resulting normalized data are shown in Table 1.2 (right). Now the least squares objective (after omitting w_0) is

$$\begin{aligned} f(w_1, w_2) &= \sum_{i=1}^{10}(w_1 x_{i1} + w_2 x_{i2} - y_i)^2 \\ &\approx 10w_1^2 + 10w_2^2 + 1.99w_1w_2 - 8.7w_1 - 2.79w_2 + 2.09. \end{aligned}$$

This is minimized at

$$\mathbf{w}^\star = (w_1^\star, w_2^\star) \approx (0.43, 0.097),$$

so if our initial hypothesis (1.16) is true, we should have

$$y_i \approx y_i^\star = 0.43x_{i1} + 0.097x_{i2} \tag{1.17}$$

in the normalized data. This can quickly be checked, and the results are not perfect, but not too bad, either; see Table 1.3 (ignore the last column for now).

x_{i1}	x_{i2}	y_i	$y_i^\star$	$z_i^\star$
-2.04	-1.28	-0.94	-1.00	-0.87
-0.88	0.32	-0.52	-0.35	-0.37
-0.05	1.03	-0.05	0.08	-0.02
-0.16	-1.28	-0.18	-0.19	-0.07
1.42	-1.28	0.67	0.49	0.61
1.02	1.39	0.59	0.57	0.44
0.06	1.39	0.19	0.16	0.03
-0.88	-0.04	-0.12	-0.38	-0.37
0.89	-0.21	0.17	0.36	0.38
0.62	-0.04	0.21	0.26	0.27

Table 1.3: Outputs $y_i^\star$ predicted by the linear model (1.17) and by the model $z_i^\star = 0.43x_{i1}$ that simply ignores the second input variable

What we also see from (1.17) is that the first input variable (GPA) has a much higher influence on the output (GGPA) than the second one (TOEFL).

In fact, if we drop the second one altogether, we obtain outputs $z_i^\star$ (last column in Table 1.3) that seem equivalent to the predicted outputs $y_i^\star$ within the level of noise that we have anyway.

We conclude that TOEFL scores are probably not indicative for the performance of admitted students, so the admission committee should not care too much about them. Requiring a minimum score of 100 might make sense, but whenever an applicant reaches at least this score, the actual value does not matter.

The LASSO. So far, we have computed linear functions $y = 0.43x_1 + 0.097x_2$ and $z = 0.43x_1$ that "explain" the historical data from Table 1.1. However, they are optimized to fit the historical data, not the future. We may have *overfitting*. This typyically leads to unrealiable predictions of high variance in the future. Also, ideally, we would like non-indicative variables (such as the TOEFL in our example) to actually have weight 0, so that the model "knows" the important variables and is therefore better to interpret.

The question is: how can we in general improve the quality of our forecast? There are various heuristics to identify the "important" variables' (subset selection). A very simple one is just to forget about weights close to 0 in the least squares solution. However, for this, we need to define what it means to be close to 0; and it may happen that small changes in the data lead to different variables being dropped if their weights are around the threshold. On the other end of the spectrum, there is *best subset selection* where we compute the least squares solution subject to the constraint that there are at most k nonzero weights, for some k that we believe is the right number of important variables. This is NP-hard, though.

A popular approach that in many cases improves forecasts and at the same time identifies important variables has been suggested by Tibshirani in 1996 [Tib96]. Instead of minimizing the least squares objective globally, it is minimized over a suitable ℓ_1-ball (ball in the 1-norm $\|\mathbf{w}\|_1 = \sum_{j=1}^d |w_j|$):

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \|\mathbf{w}^\top \mathbf{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathbf{w}\|_1 \leq R, \end{array} \tag{1.18}$$

where $R \in \mathbb{R}_+$ is some parameter. In our case, if we for example

$$\begin{array}{ll} \text{minimize} & f(w_1, w_2) = 10w_1^2 + 10w_2^2 + 1.99w_1w_2 - 8.7w_1 - 2.79w_2 + 2.09 \\ \text{subject to} & |w_1| + |w_2| \leq 0.2, \end{array} \tag{1.19}$$

we obtain weights $\mathbf{w}^\star = (w_1^\star, w_2^\star) = (0.2, 0)$: the non-indicative TOEFL score has disappeared automatically! For $R = 0.3$, the same happens (with $w_1^\star = 0.3$, respectively). For $R = 0.4$, the TOEFL score starts creeping back in: we get $(w_1^\star, w_2^\star) \approx (0.36, 0.036)$. For $R = 0.5$, we have $(w_1^\star, w_2^\star) \approx (0.41, 0.086)$, while for $R = 0.6$ (and all larger values of R), we recover the original solution $(w_1^\star, w_2^\star) = (0.43, 0.097)$.

It is important to understand that using the "fixed" weights (which may be significantly shrunken), we make predictions *worse* on the historical data (this must be so, since least squares was optimal for the historical data). But future predictions may benefit (a lot). To quantify this benefit, we need to make statistical assumptions about future observations; this is beyond the scope of our treatment here.

The phenomenon that adding a constraint on $\|\mathbf{w}\|_1$ tends to set weights to 0 is not restricted to $d = 2$. The constrained minimization problem (1.18) is called the *LASSO* (least absolute shrinkage and selection operator) and has the tendency to assign weights of 0 and thus to select a subset of input variables, where R controls how aggressive the selection is.

In our example, it is easy to get an intuition why this works. Let us look at the case $R = 0.2$. The smallest value attainable in (1.19) is the smallest α such that that the (elliptical) sublevel set $f^{\leq \alpha}$ of the least squares objective f still intersects the ℓ_1-ball $\{(w_1, w_2) : |w_1| + |w_2| \leq 0.2\}$. This smallest value turns out to be $\alpha = 0.75$, see Figure 1.14. For this value of α, the sublevel set intersects the ℓ_1-ball exactly in one point, namely $(0.2, 0)$.

At $(0.2, 0)$, the ellipse $\{(w_1, w_2) : f(w_1, w_2) = \alpha\}$ is "vertical enough" to just intersect the corner of the ℓ_1-ball. The reason is that the center of the ellipse is relatively close to the w_1-axis, when compared to its size. As R increases, the relevant value of α decreases, the ellipse gets smaller and less vertical around the w_1-axis; until it eventually stops intersecting the ℓ_1-ball $\{(w_1, w_2) : |w_1| + |w_2| \leq R\}$ in a corner (dashed situation in Figure 1.14, for $R = 0.4$).

Even though we have presented a toy example in this section, the background is real. The theory of admission and in particular performance forecasts has been developed in a recent PhD thesis by Zimmermann [Zim16].

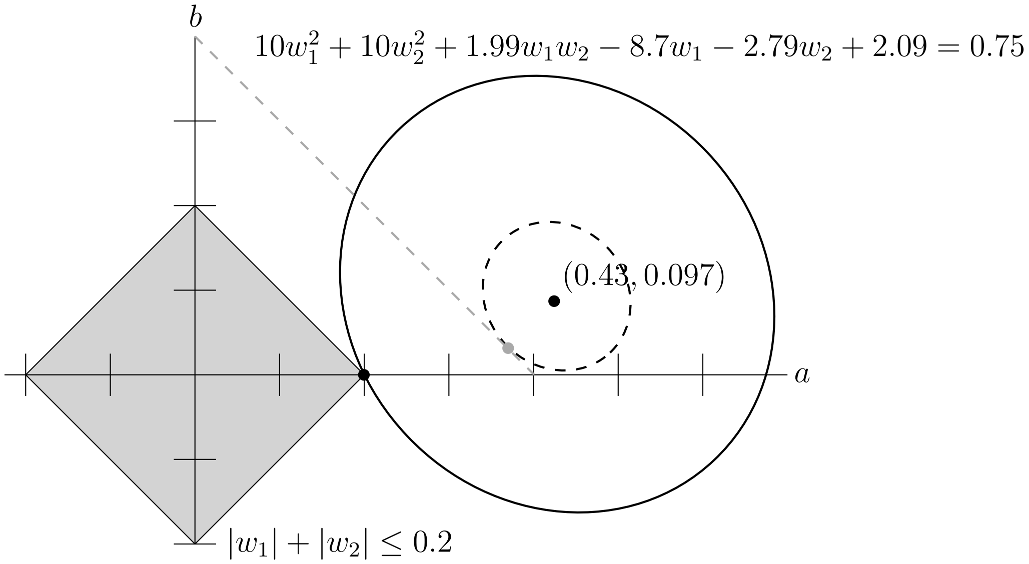

b
$10w_1^2 + 10w_2^2 + 1.99w_1w_2 - 8.7w_1 - 2.79w_2 + 2.09 = 0.75$
$(0.43, 0.097)$
a
$|w_1| + |w_2| \leq 0.2$

Figure 1.14: Lasso

1.7 Exercises

Exercise 1. *Prove that a differentiable function is continuous!*

Exercise 2. *Prove Jensen's inequality (Lemma 1.12)!*

Exercise 3. *Prove that a convex function (with* $\mathbf{dom}(f)$ *open) is continuous (Lemma 1.13)!*

Hint: *First prove that a convex function* f *is bounded on any cube* $C = [l_1, u_1] \times [l_2, u_2] \times \cdots \times [l_d, u_d] \subseteq \mathbf{dom}(f)$*, with the maximum value occurring on some corner of the cube (a point* $\mathbf{z}$ *such that* $z_i \in \{l_i, u_i\}$ *for all* i*). Then use this fact to show that—given* $\mathbf{x} \in \mathbf{dom}(f)$ *and* $\varepsilon > 0$*—all* $\mathbf{y}$ *in a sufficiently small ball around* $\mathbf{x}$ *satisfy* $|f(\mathbf{y}) - f(\mathbf{x})| < \varepsilon$*.*

Exercise 4. *Prove that the function* $d_{\mathbf{y}} : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{x} \mapsto \|\mathbf{x} - \mathbf{y}\|^2$ *is strictly convex for any* $\mathbf{y} \in \mathbb{R}^d$*. (Use Lemma 1.24.)*

Exercise 5. *Prove Lemma 1.18! Can (ii) be generalized to show that for two convex functions* f, g*, the function* $f \circ g$ *is convex as well?*

Exercise 6. *Prove Lemmata 1.37 and 1.38!*

Exercise 7. *Consider the function ℓ defined in (1.15). Prove that ℓ is convex!*

Exercise 8. *Consider the function ℓ defined in (1.15). Let us call an argument matrix W a* separator *for P if for all $\mathbf{x} \in P$,*

$$(W\mathbf{x})_{d(\mathbf{x})} = \max_{j=0}^{9} (W\mathbf{x})_j,$$

i.e. under (1.14), the correct digit has highest probability (possibly along with other digits). A separator is trivial *if for all $\mathbf{x} \in P$ and all $i, j \in \{0, \ldots, 9\}$,*

$$(W\mathbf{x})_i = (W\mathbf{x})_j.$$

For example, whenever the rows of W are pairwise identical, we obtain a trivial separator. But depending on the data, there may be other trivial separators. For example, if some pixel is black (gray value 0) in all images, arbitrarily changing the entries in the corresponding column of a trivial separator gives us another trivial separator. For a trivial separator W, (1.15) yields $\ell(W) = |P| \ln 10$.

Prove the following statement: ℓ has a global minimum if and only if all separators are trivial.

As a special case, consider the situation in which there exists a strong *(and in particular nontrivial) separator: a matrix $W^\star$ such that for all $\mathbf{x} \in P$ and all $j \neq d(\mathbf{x})$,*

$$(W^\star\mathbf{x})_{d(\mathbf{x})} > (W^\star\mathbf{x})_j,$$

i.e. the correct digit has unique highest probability. In this case, it is easy to see that $\ell(\lambda W^\star) \to_{\lambda\to\infty} 0$, so we cannot have a global minimum, as $\inf_W(\ell(W)) = 0$ is not attainable.

Exercise 9. *Prove that the function $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|$ (ℓ_1-norm) is convex!*

Exercise 10. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable. For fixed $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, consider the univariate function $h(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$ over a suitable open interval $\mathbf{dom}(h) \supseteq [0, 1]$ such that $\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \mathbf{dom}(f)$ for all $t \in \mathbf{dom}(h)$. Let us abbreviate $\mathbf{v} = \mathbf{y} - \mathbf{x}$. We already know that $h'(t) = \nabla f(\mathbf{x} + t\mathbf{v})^\top \mathbf{v}$ for $t \in \mathbf{dom}(h)$. Prove that*

$$h''(t) = \mathbf{v}^\top \nabla^2 f(\mathbf{x} + t\mathbf{v})\mathbf{v}, \quad t \in \mathbf{dom}(h).$$

Exercise 11. *A seminorm is a function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying the following two properties for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and all $\lambda \in \mathbb{R}$.*

(i) $f(\lambda \mathbf{x}) = |\lambda| f(\mathbf{x})$*,*

(ii) $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ *(triangle inequality).*

Prove that every seminorm is convex!

Exercise 12. *Suppose that we have centered observations* $(\mathbf{x}_i, y_i)$ *such that* $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}, \sum_{i=1}^n y_i = 0$*. Let* $w_0^\star, \mathbf{w}^*$ *be the global minimum of the least squares objective*

$$f(w_0, \mathbf{w}) = \sum_{i=1}^n (w_0 + \mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

Prove that $w_0^\star = 0$*. Also, suppose* $\mathbf{x}_i'$ *and* y_i' *are such that for all* i*,* $\mathbf{x}_i' = \mathbf{x}_i + \mathbf{q}$*,* $y_i' = y_i + r$*. Show that* $(w_0, \mathbf{w})$ *minimizes* f *if and only if* $(w_0 - \mathbf{w}^\top \mathbf{q} + r, \mathbf{w})$ *minimizes*

$$f'(w_o, \mathbf{w}) = \sum_{i=1}^n (w_0 + \mathbf{w}^\top \mathbf{x}_i' - y_i')^2.$$

Bibliography

[ACGH18] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *CoRR*, abs/1810.02281, 2018.

[AE08] Herbert Amann and Joachim Escher. *Analysis II*. Birkhäuser, 2008.

[Ber05] Dimitri P. Bertsekas. Lecture slides on convex analysis and optimization, 2005. `http://athenasc.com/Convex_Slides.pdf`.

[BG17] Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods. *CoRR*, abs/1712.04581, 2017.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. `https://web.stanford.edu/~boyd/cvxbook/`.

[Dav59] William C. Davidon. Variable metric method for minimization. Technical Report ANL-5990, AEC Research and Development, 1959.

[Dav91] William C. Davidon. Variable metric method for minimization. *SIAM J. Optimization*, 1(1):1–17, 1991.

[Die69] J. Dieudonneé. *Foundations of Modern Analysis*. Academic Press, 1969.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 07 2008.

[FM91] M. Furi and M. Martelli. On the mean value theorem, inequality, and inclusion. *The American Mathematical Monthly*, 98(9):840–846, 1991.

[Gol70] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[Gre70] J. Greenstadt. Variations on variable-metric methods. *Mathematics of Computation*, 24(109):1–22, 1970.

[KSJ18] Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv*, 2018.

[LW19] Ching-Pei Lee and Stephen Wright. First-order algorithms converge faster than $o(1/k)$ on convex problems. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3754–3762, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Math. Dokl.*, 27(2), 1983.

[Nes18] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, second edition, 2018.

[Noc80] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006.

[NY83] Arkady. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[Roc97] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1997.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[Vis15] Nisheeth Vishnoi. A mini-course on convex optimization (with a view toward designing fast algorithms), 2015. `https://theory.epfl.ch/vishnoi/Nisheeth-VishnoiFall2014-ConvexOptimization.pdf`.

[Zim16] Judith Zimmermann. *Information Processing for Effective and Stable Admission*. PhD thesis, ETH Zurich, 2016. .