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

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

We also use

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

We are freely using basic notions and material such as open and closed sets, vector spaces, continuity, convergence, limits, triangle inequality,...

1.2 Convex sets

Definition 1.1. *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 1.1.*

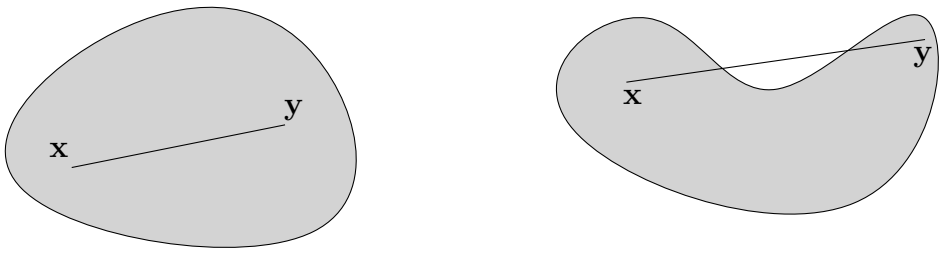

Figure 1.1: A convex set (left) and a non-convex set (right)

Observation 1.2. *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.*

1.3 Convex functions

We are considering real-valued functions $f : \mathbf{dom}(f) \to \mathbb{R}$, where $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ denotes the domain of f. 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 1.2) is the set of points above the graph,

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

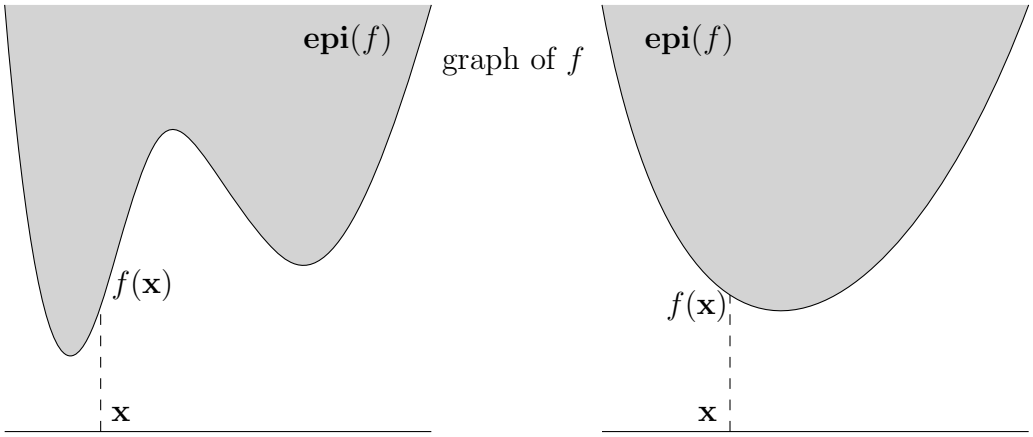

Figure 1.2: Graph and epigraph of a non-convex function (left) and a convex function (right)

Definition 1.3 ([2, 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*

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}). \tag{1.1}$$

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,

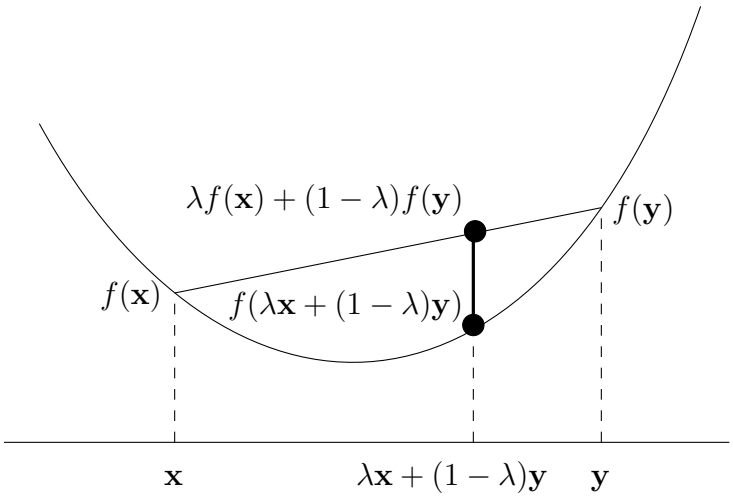

Figure 1.3: A convex function

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, (1.1) is always satisfied with equality, and line segments connecting points on the graph lie pointwise on the graph.

Observation 1.4. *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, $f(\mathbf{x}) \leq \alpha, f(\mathbf{y}) \leq \beta$, hence by convexity of f,

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

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 (1.1). □

Lemma 1.5 (Jensen's inequality)**.** *Let f be convex, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbf{dom}(f)$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^d \lambda_i = 1$. Then*

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

For $m = 2$, this is (1.1). The proof of the general case is Exercise 1.

Lemma 1.6. *Let f be convex and suppose that $\mathbf{dom}(f)$ is open. Then f is continuous.*

This is not entirely obvious (see Exercise 2), and it becomes false if we consider convex functions over general vector spaces. What saves us is that $\mathbb{R}^d$ has finite dimension.

As an example, 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 (1.1) directly is somewhat cumbersome. Next, we develop better ways to do this.

1.3.1 First-order characterization of convexity

If f is diffentiable, convexity can be characterized as follows.

Lemma 1.7 ([2, 3.1.3])**.** *Suppose that $\mathbf{dom}(f)$ is open and that f is diffentiable; in particular, the* gradient *(vector of partial derivatives)*

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

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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \tag{1.2}$$

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 1.4.

For $f(x_1, x_2) = x_1^2 + x_2^2$, we have $\nabla f(\mathbf{x}) = (2x_1, 2x_2)$, hence (1.2) boils down to

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

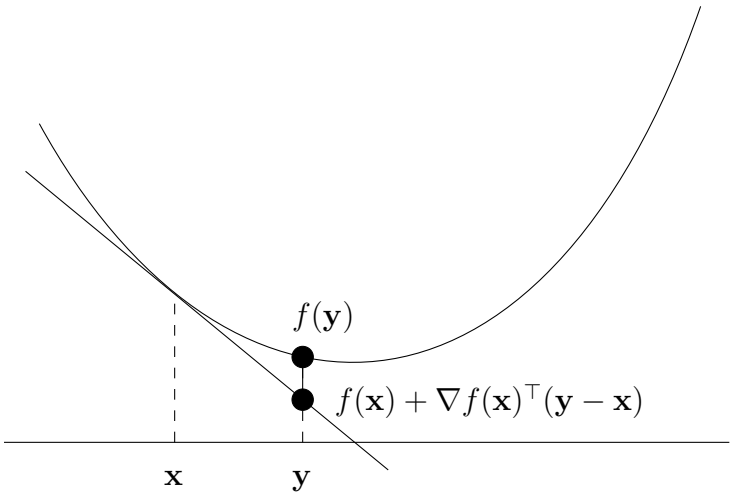

Figure 1.4: First-order characterization of convexity

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 1.5 for an example. More generally, Exercise 7 asks you to prove that the ℓ_1-norm (or 1-norm) $f(\mathbf{x}) = \|\mathbf{x}\|_1$ is convex.

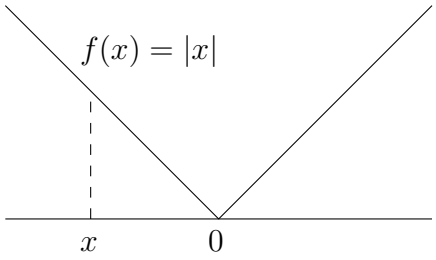

Figure 1.5: A non-differentiable convex function

1.3.2 Second-order characterization of convexity

If f is twice diffentiable, convexity can be characterized as follows.

Lemma 1.8 ([2, 3.1.4]). *Suppose that* $\mathbf{dom}(f)$ *is open and that* f *is twice differentiable; in particular, the* Hessian *(matrix of second partial derivatives)*

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

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

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{(i.e. } \nabla^2 f(\mathbf{x}) \textit{ is positive semidefinite).} \tag{1.3}$$

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

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

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

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 3. The non-squared Euclidean distance $\|\mathbf{x} - \mathbf{y}\|$ is also convex in $\mathbf{x}$, as a consequence of Lemma 1.9(ii) below and the fact that every seminorm (in particular the Euclidean norm $\|x\|$) is convex (Exercise 8). 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 1.9 (Exercise 4)**.**

(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 1.10. *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 1.11. *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 mimimum, 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 (1.1), 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 mimimum. □

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 mimimum by continuity (Lemma 1.6). 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

Lemma 1.12. *Suppose that f is convex and differentiable over an open domain $\mathbf{dom}(f)$. 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 1.7, 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: if $\mathbf{x}$ is a global minimum, then $\nabla f(\mathbf{x}) = \mathbf{0}$. This is a corollary of Lemma 1.16 below [2, 4.2.3].

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 1.13 ([2, 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*

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}). \tag{1.4}$$

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. More generally (and following up on Lemma 1.8), if the Hessian $\nabla^2 f(\mathbf{x})$ is positive definite everywhere, then f is strictly convex [2, 3.1.4]. The converse is false, though: $f(x) = x^4$ is strictly convex but has vanishing second derivative at $x = 0$.

Lemma 1.14. *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 mimima of value $f_{\min}$, and let $\mathbf{z} = \frac{1}{2}\mathbf{x}^\star + \frac{1}{2}\mathbf{y}^\star$. By (1.4),

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

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 on 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 1.6 (left).

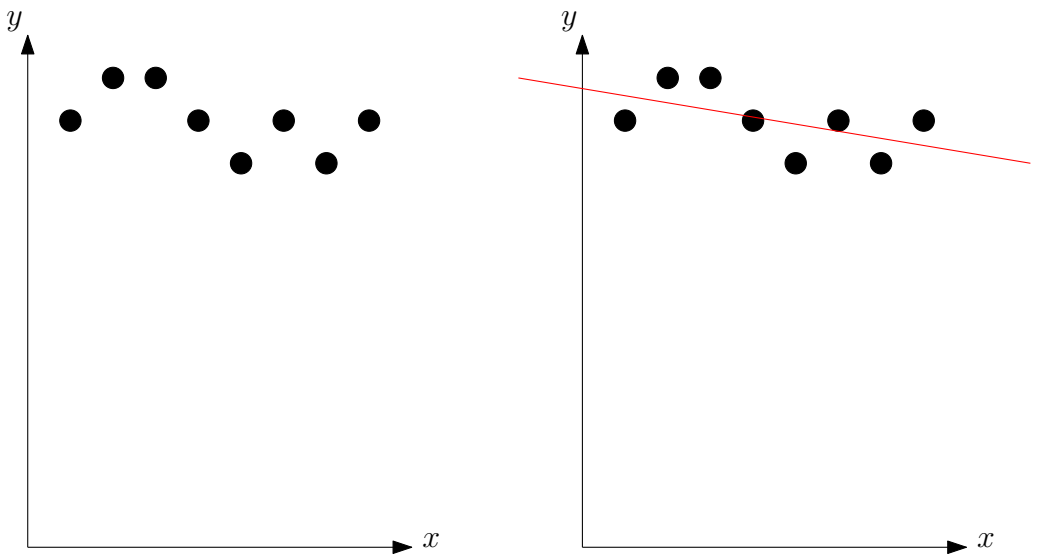

Figure 1.6: Data points in $\mathbb{R}^2$ (left) and least-squares fit (right)

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 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 1.9 that f is convex. In fact,

$$f(w_0, w_1) = 204w_1^2 + 72w_1 w_0 - 706w_1 + 8w_0^2 - 160w_0 + 804, \tag{1.5}$$

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 definite 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

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

Hence, the "optimal" line is

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

see Figure 1.6 (right).

1.4.3 Constrained Minimization

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

Definition 1.15. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and let $X \subseteq \mathbf{dom}(f)$ be a convex set. $\mathbf{x} \in X$ 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 1.16 ([2, 4.2.3])**.** *Suppose that f is convex and differentiable over an open domain $\mathbf{dom}(f)$, and let $X \subseteq \mathbf{dom}(f)$ be a convex set. $\mathbf{x}^\star \in X$ is a minimizer of f over X if and only if*

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

Geometrically, this 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 1.7. In still other words, $\mathbf{x} - \mathbf{x}^\star$ forms a non-obtuse angle with $\nabla f(\mathbf{x}^\star)$ for all $\mathbf{x} \in X$. Applying this with $X = \mathbf{dom}(f)$, we recover Lemma 1.12 and its converse [2, 4.2.3].

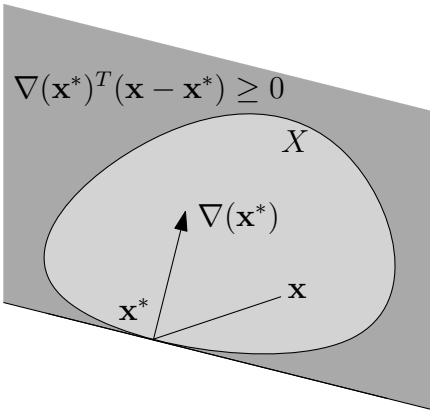

Figure 1.7: Optimality condition for constrained optimization

We typically write constrained minimization problems in the form

$$\operatorname{argmin}\{f(\mathbf{x}) : \mathbf{x} \in X\} \tag{1.6}$$

or

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X. \end{array} \tag{1.7}$$

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.

1.5.1 Sublevel sets and the Weierstrass Theorem

Definition 1.17. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$, $\gamma \in \mathbb{R}$. The set*

$$f^{\leq\gamma} := \{\mathbf{x} \in \mathbf{dom}(f) : f(\mathbf{x}) \leq \gamma\}$$

is the γ-sublevel set of f; see Figure 1.8

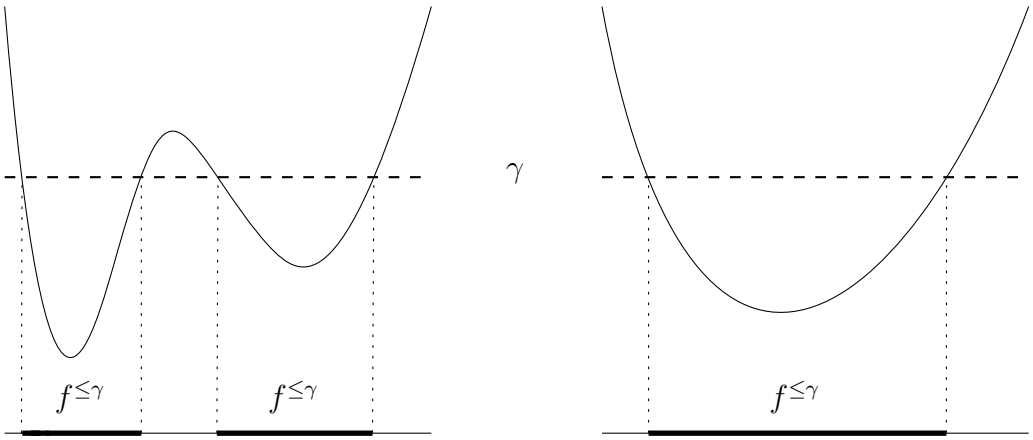

Figure 1.8: Sublevel set of a non-convex function (left) and a convex function (right)

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 (if $\mathbf{dom}(f)$ is open), sublevel sets are closed. The following (known as the Weierstrass Theorem) just formalizes an argument that we have made earlier.

Theorem 1.18. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a convex function, $\mathbf{dom}(f)$ open, and suppose there is a nonempty and bounded sublevel set $f^{\leq\gamma}$. Then f has a global minimum.*

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

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 1.9. You want to train your program with the set P, and after that use it to recognize handwritten digits in arbitrary 28×28 images.

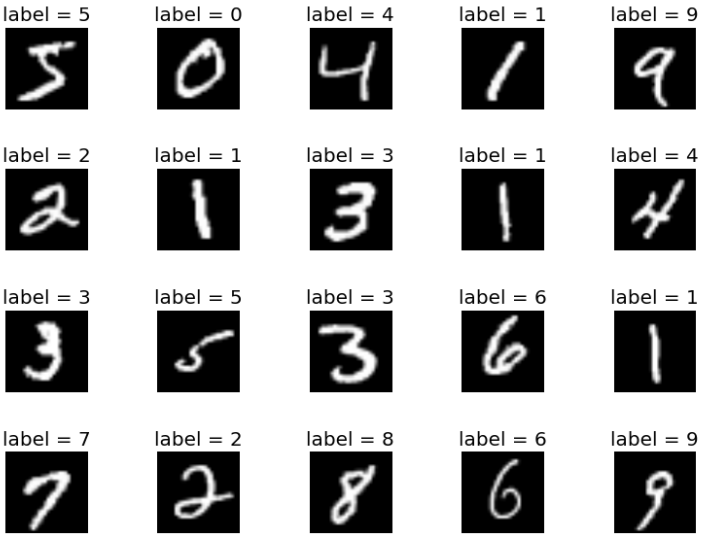

Figure 1.9: Some training images from the MNIST data set (picture from `http://corochann.com/mnist-dataset-introduction-1138.html`

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 doesn't 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 probabiliy z_j. How to do this is not canonical, but here is a well-known formula that works:

$$z_j = z_j(\mathbf{y}) = \frac{e^{y_j}}{\sum_{k=0}^{9} e^{y_k}}. \tag{1.8}$$

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.

$$\ell(W) = -\sum_{\mathbf{x}\in P} \ln z_{d(\mathbf{x})}(W\mathbf{x}) = \sum_{\mathbf{x}\in P} \left(\ln \left(\sum_{j=0}^{9} e^{(W\mathbf{x})_j} \right) - (W\mathbf{x})_{d(\mathbf{x})} \right). \tag{1.9}$$

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 (1.8), probabilities are always strictly between 0 and 1, so we have $\ell(W) > 0$ for all W.

Exercise 5 asks you to prove that ℓ is convex. In Exercise 6, 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 competetive application process. Applicants are submitting various documents (GPA, TOEFL test scores, 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.[1]

[1] 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.

Data on the actual performance of admitted students is available. To keep things simple in the following example, let's 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 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

$$\mathbf{GGPA} \approx w_1 \cdot \mathbf{GPA} + w_2 \cdot \mathbf{TOEFL} + w_0. \tag{1.10}$$

However, in our scenario, the GPA scores span a range of only 0.5 while the TOEFL scores span a range of 20. The resulting least squares objective would be somewhat ugly; we already saw this in our previous example (1.5), 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.

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, $((3.52, 100), 3.92)$ is an observation (of a student failing, unfortunately).

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

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

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)

Finally, we 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, we simply need to multiply x_{ij} by $\sqrt{n/\sum_{i=1}^{d} x_{ij}^2}$ for all i, j. For our data set, the resulting normalized data are shown in Table 1.2

(right). Now the least squares objective is

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

This is minimized at

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

so if our initial hypothesis (1.10) is true, we should have

$$y_i \approx y_i^\star = 0.43 x_{i1} + 0.097 x_{i2} \tag{1.11}$$

in the normalized data. This can quickly be checked, and the results aren't 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 (1.11) and by the model $z_i^\star = 0.43 x_{i1}$ that simply ignores the second input variable

What we also see from (1.11) 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 committtee 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. If we believe that hypothesis (1.10) also holds for future applicants, we can indeed use these functions to make performance forecasts. Using z instead of y introduces a bias, though. For example, if the average TOEFL score of future applicants is above 110, z will underestimate performance by ignoring the (small) positive effect of a high TOEFL score. The advantages of z over y are (i) better interpretability (z "knows" that TOEFL is non-indicative, while y doesn't); and (ii) possibly more stable estimates (variance caused by a non-indicative input variable is removed).

The question is: how can we in general improve the quality of our forecast in the presence of non-indicative input variables such as the TOEFL score in our example? We can (as done above) just forget about a variable of weight 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. A more elegant solution has been suggested by Tibshirani in 1996 [**?**]. Instead of minimizing the least squares objective globally, it is minimized over a suitable ℓ_1-ball (ball in the 1-norm):

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{d} \|\mathbf{w}^\top \mathbf{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathbf{w}\|_1 \leq R, \end{array} \tag{1.12}$$

where $R \in \mathbb{R}_+$ is some parameter. Here $\mathbf{w} \in \mathbb{R}^d$ is the vector of optimization variables. In our case, $\mathbf{w} = (w_1, w_2)$), and if we for example

$$\begin{array}{ll} \text{minimize} & f(w_1, w_2) = 10w_1^2 + 10w_2^2 + 1.99w_1w_2 - 8.7w_1 - 2.79w_2 + 2.09 \\ \text{subject to} & |w_1| + |w_2| \leq 0.2, \end{array} \tag{1.13}$$

we obtain $\mathbf{w}^\star = (w_1^\star, w_2^\star) = (0.2, 0)$: 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 $(w_1^\star, w_2^\star) = (0.43, 0.097)$.

This phenomenon is not restricted to $d = 2$. The constrained minimization problem (1.12) is called the *LASSO* (least absolute shrinkage and selection operator) and has the tendency to assign weight 0 to and thus remove non-indicative input variables, where R controls how aggressive the selection is, and how much bias it potentially introduces (the smaller the value R, the more aggressive the selection, and the higher the bias).

In our example, it's easy to get an intuition why this works. Let's look at the case $R = 0.2$. The smallest value attainable in (1.13) is the smallest γ such that that the (elliptical) sublevel set $f^{\leq\gamma}$ 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 $\gamma = 0.75$, see Figure 1.10. For this value of γ, the sublevel set intersects the ℓ_1-ball exactly in one point, namely $(0.2, 0)$.

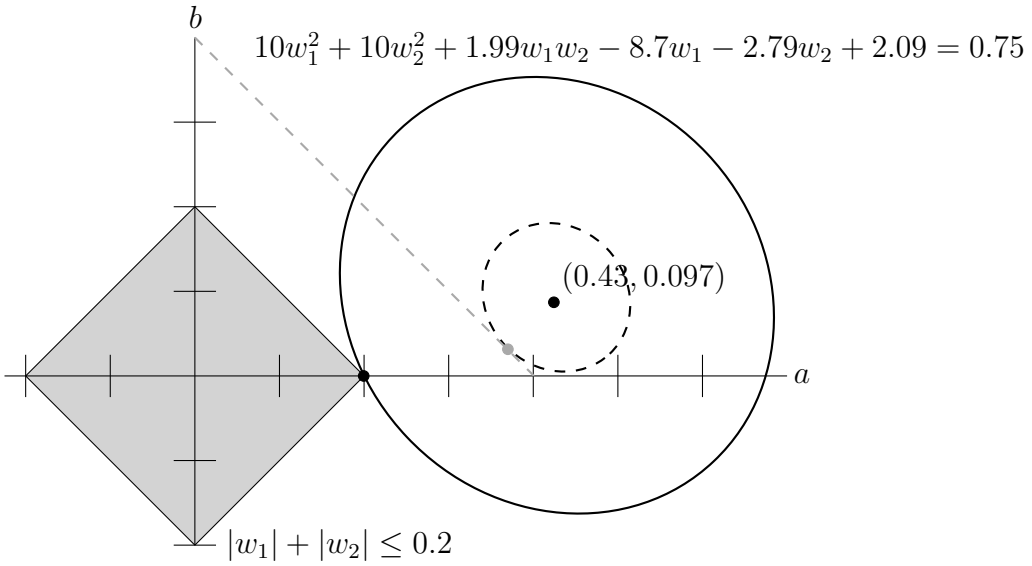

Figure 1.10: Lasso

At $(0.2, 0)$, the ellipse $\{(w_1, w_2) : f(w_1, w_2) = \gamma\}$ 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 γ 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 1.10, 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 [**?**].

1.7 Exercises

Exercise 1. *Prove Jensen's inequality (Lemma 1.5)!*

Exercise 2. *Prove that a convex funtion is continuous (Lemma 1.6)!*

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 occuring 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 $\mathbf{x} \in \mathbf{dom}(f), \varepsilon > 0$—all $\mathbf{y}$ in a sufficiently small cube around $\mathbf{x}$ satisfy $|f(\mathbf{y}) - f(\mathbf{x})| < \varepsilon$.*

Exercise 3. *Prove that the function $d_{\mathbf{y}} : \mathbb{R}^d \to R$, $\mathbf{x} \mapsto \|\mathbf{x} - \mathbf{y}\|^2$ is strictly convex for any $\mathbf{y} \in \mathbb{R}^d$. (This is essentially an exercise in computing gradients and Hessians.)*

Exercise 4. *Prove Lemma 1.9! Can (ii) be generalized to show that for two convex functions f, g, the function $f \circ g$ is convex as well?*

Exercise 5. *Consider the function ℓ defined in (1.9). Prove that ℓ is convex!*

Exercise 6. *Consider the function ℓ defined in (1.9). 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 (1.8), 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*, (1.9) 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 7. *Prove that the function* $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$ *(*ℓ_1*-norm) is convex!*

Exercise 8. *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!

Bibliography

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

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