Contents

1 Theory of Convex Functions 2–32

2 Gradient Descent 32–53

3 Projected and Proximal Gradient Descent 53–69

4 Subgradient Descent 69–80

5 Stochastic Gradient Descent 80–88

6 Nonconvex functions 88–107

7 Newton's Method 107–118

8 Quasi-Newton Methods 118–136

9 Frank-Wolfe 136–138

10 Coordinate Descent 138–150

Chapter 1

Theory of Convex Functions

Contents

1.1 Notation . . . 3
1.2 The Cauchy-Schwarz inequality . . . 3
1.3 Convex sets . . . 5
1.4 Convex functions . . . 5
1.4.1 Differentiable functions . . . 8
1.4.2 First-order characterization of convexity . . . 12
1.4.3 Second-order characterization of convexity . . . 14
1.4.4 Operations that preserve convexity . . . 15
1.5 Minimizing convex functions . . . 15
1.5.1 Strictly convex functions . . . 17
1.5.2 Example: Least squares . . . 17
1.5.3 Constrained Minimization . . . 19
1.6 Existence of a minimizer . . . 20
1.6.1 Sublevel sets and the Weierstrass Theorem . . . 21
1.7 Examples . . . 22
1.7.1 Handwritten digit recognition . . . 22
1.7.2 Master's Admission . . . 23
1.8 Exercises . . . 29

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. 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.

1.2 The Cauchy-Schwarz inequality

As a warm-up, we explicitly want to mention, illustrate, and prove a basic result from linear algebra that we frequently need.

Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. The *Cauchy-Schwarz inequality* is

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

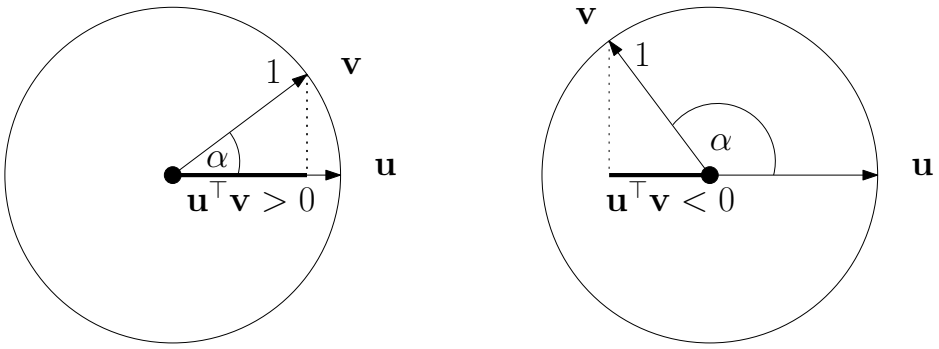
For nonzero vectors, this is equivalent to

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

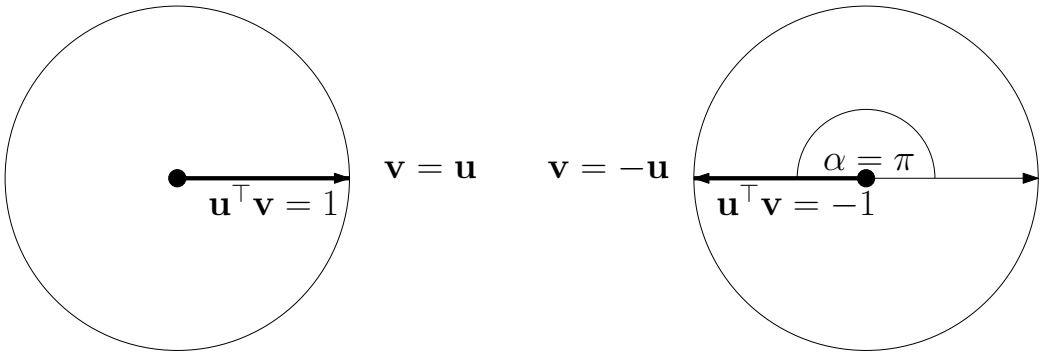
and this fraction can be used to define the angle α between $\mathbf{u}$ and $\mathbf{v}$:

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

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):



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

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

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

1.3 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.*

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.*

1.4 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,

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

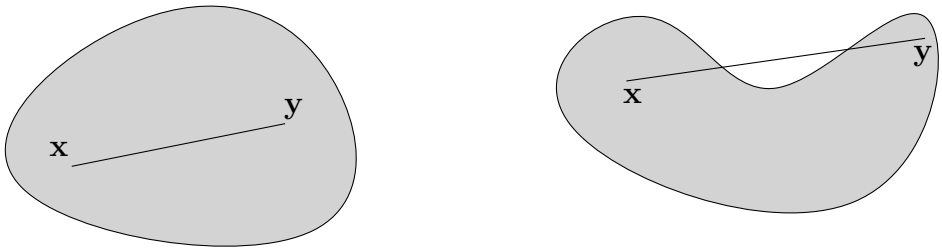


Figure 1.1: A convex set (left) and a non-convex set (right)

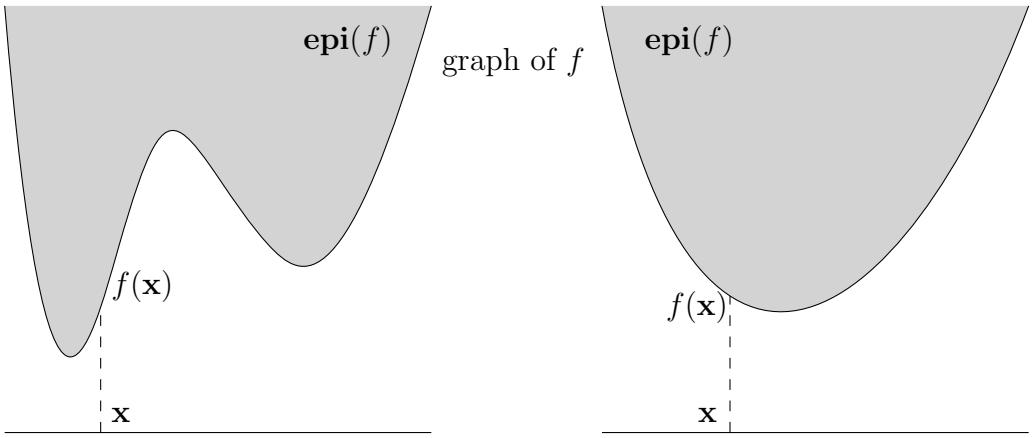


Figure 1.2: Graph and epigraph of a non-convex function (left) and a convex function (right)

Definition 1.3 ([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*

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

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, 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.

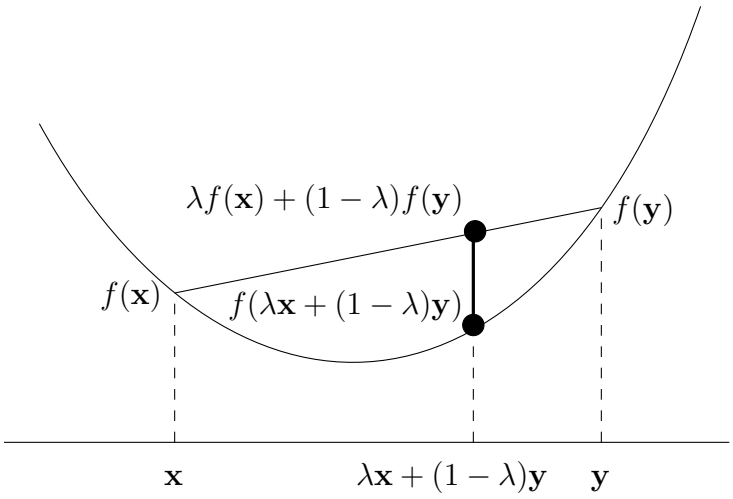


Figure 1.3: A convex function

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 $f : \mathbb{R}^d \to \mathbb{R}$ 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*

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

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 if the function under consideration is differentiable.

1.4.1 Differentiable functions

The following is standard material taught in multivariate calculus. As we frequently need it, we include a refresher here.

Definition 1.7. *Let $f : \mathbf{dom}(f) \to \mathbb{R}^m$ where $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ is open. Function f is called* differentiable *at $\mathbf{x} \in \mathbf{dom}(f)$ if there exists an $(m \times d)$-matrix A and an error function $r : \mathbb{R}^d \to \mathbb{R}^m$ defined around $\mathbf{0} \in \mathbb{R}^d$ 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

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

It then also follows that the matrix A is unique, and it is called the differential *or* Jacobian matrix *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 $\mathbf{x} \in \mathbf{dom}(f)$.*

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 1.4.

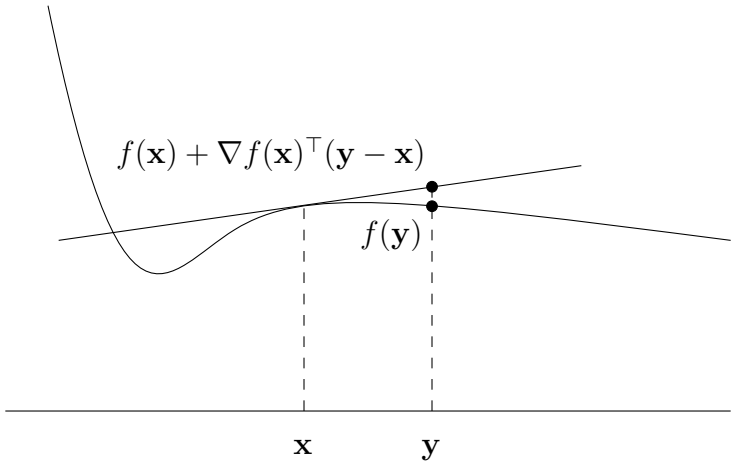


Figure 1.4: If f is differentiable at $\mathbf{x}$, the graph of f is locally (around $\mathbf{x}$) approximated by a tangent hyperplane

Let us do a simpe example to illustrate the concept of differentiability. Consider the function $f(x) = x^2$. We know that its derivative is $f'(x) = 2x$. But why? For $y = x + v$, 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 1.8 (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*

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

Let us do an example. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a differentiable function, and fix $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Now define $g : \mathbb{R} \to \mathbb{R}^d$ by $g(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$ and set $h = f \circ g$. Thus, $h(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, and we have

$$h'(t) = Dh(t) = Df(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))Dg(t) = Df(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}). \quad (1.2)$$

The following is a general result that we will later use in specific settings. As its proof also highlights some important notions and techniques, we will give it here. As a preparation, we need the concept of the *spectral norm* of a matrix.

Definition 1.9. *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 unit vector can be stretched in length under the mapping $\mathbf{v} \to A\mathbf{v}$.

Also recall that a function $f : \mathbf{dom}(f) \to \mathbb{R}^m$ is *B-Lipschitz* (or simply Lipschitz if there is a suitable B) if $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq B \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$. In particular, Lipschitz functions are continuous.

Theorem 1.10. *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 open, the following two statements are equivalent. For any convex $X \subseteq \mathbf{dom}(f)$, (ii) implies (i).*

(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, meaning that

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

Indeed, (i) might not imply (ii) if X is closed. As a trivial example, the Lipschitz condition is always satisfied over $X = \{\mathbf{0}\}$ but does not say anything about $\|Df(\mathbf{x})\|$.

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 apply the *fundamental theorem of calculus*:

$$\int_a^b h'(t)dt = h(b) - h(a), \tag{1.3}$$

where $h : \mathbf{dom}(h) \to \mathbb{R}^m$ is a univariate differentiable function, h' its componentwise derivative, $[a, b] \subseteq \mathbf{dom}(h)$ and $\int$ the componentwise integral. For fixed $\mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y}$, we apply this with

$$h(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})),$$

in which case the chain rule yields

$$h'(t) = Df(x + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}),$$

see (1.2). Note that h is well-defined since X was assumed to be convex.

Then we compute

$$\begin{aligned}
\|f(\mathbf{y}) - f(\mathbf{x})\| &= \|h(1) - h(0)\| \\
&= \left\| \int_0^1 h'(t)dt \right\| \leq \int_0^1 \|h'(t)\| \, dt \quad \text{(Exercise 46)} \\
&= \int_0^1 \|Df(x + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})\| \, dt \\
&\leq \int_0^1 \|Df(x + t(\mathbf{y} - \mathbf{x}))\| \, \|(\mathbf{y} - \mathbf{x})\| \, dt \quad \text{(spectral norm)} \\
&\leq \int_0^1 B \, \|(\mathbf{y} - \mathbf{x})\| \, dt \quad \text{(bounded differentials)} \\
&= B \, \|(\mathbf{y} - \mathbf{x})\| \, .
\end{aligned}$$

Hence, f is B-Lipschitz over X. □

1.4.2 First-order characterization of convexity

Now we come back to convex functions with image in $\mathbb{R}$. If function $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable, convexity can be characterized by an inequality involving the gradient.

Lemma 1.11 ([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

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

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 $t \in (0, 1)$,

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

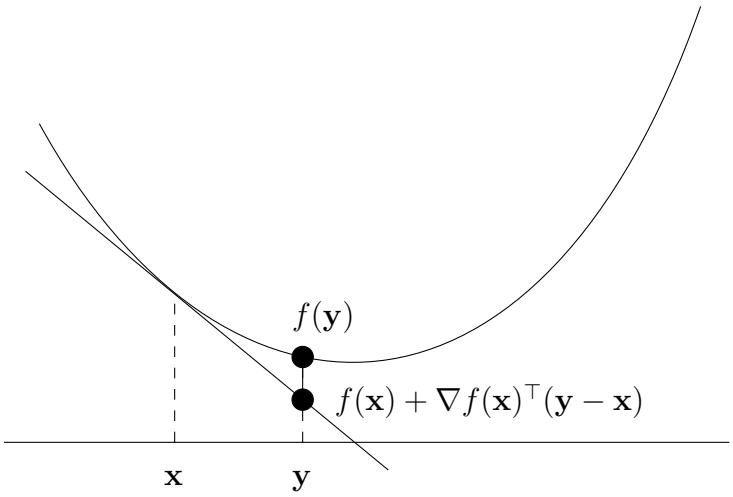


Figure 1.5: First-order characterization of convexity

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

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

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 $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ 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 (1.4) 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 7 asks you to prove that the ℓ_1-norm (or 1-norm) $f(\mathbf{x}) = \|\mathbf{x}\|_1$ is convex.

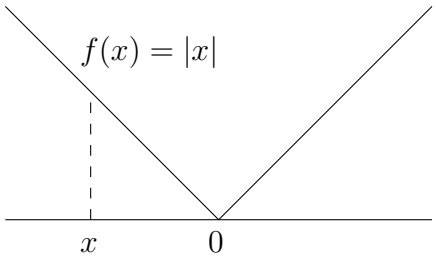


Figure 1.6: A non-differentiable convex function

1.4.3 Second-order characterization of convexity

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

Lemma 1.12 ([BV04, 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)*

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

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

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{(i.e. } \nabla^2 f(\mathbf{x}) \textit{ is positive semidefinite).} \tag{1.5}$$

(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 $\mathbf{x} \neq \mathbf{0}$.)*

Geometrically, this 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

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

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 1.13(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.

1.4.4 Operations that preserve convexity

There are two important operations that preserve convexity.

Lemma 1.13 (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 $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$.

1.5 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 1.14. *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 1.15. *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 (1.1), 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 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 1.6). 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 result.

Lemma 1.16. *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 1.11, 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 1.22 [BV04, 4.2.3].

Lemma 1.17. *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* $\nabla f(\mathbf{x}) = \mathbf{0}$.

1.5.1 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 1.18 ([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*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \tag{1.6}$$

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 1.19 ([BV04, 3.1.4]). *Suppose that $\mathbf{dom}(f)$ is open and that f is twice 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 1.20. *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 (1.6),

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

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

1.5.2 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

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

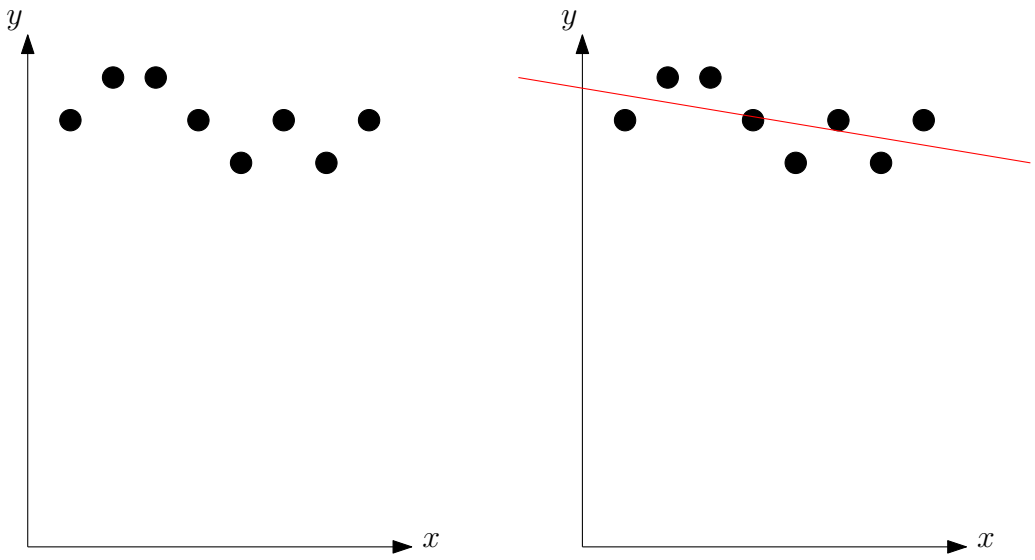


Figure 1.7: Data points in $\mathbb{R}^2$ (left) and least-squares fit (right)

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 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 1.13 that f is convex. In fact,

$$f(w_0, w_1) = 204w_1^2 + 72w_1 w_0 - 706w_1 + 8w_0^2 - 160w_0 + 804, \tag{1.7}$$

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

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

Hence, the "optimal" line is

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

see Figure 1.7 (right).

1.5.3 Constrained Minimization

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

Definition 1.21. *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 1.22 ([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 1.16, and because $\mathbf{dom}(f)$ is open, its converse Lemma 1.17 follows [BV04, 4.2.3]. If X does not contain the global minimum, then Lemma 1.22 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 $\mathbf{x} \in X$.

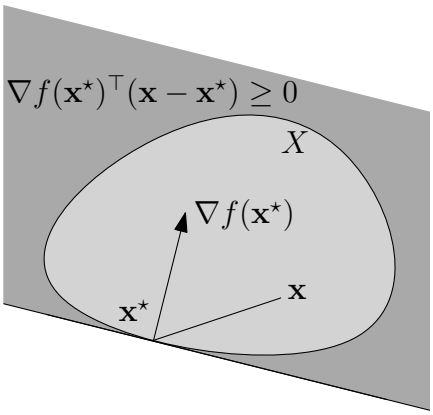


Figure 1.8: Optimality condition for constrained optimization

We typically write constrained minimization problems in the form

$$\operatorname{argmin}\{f(\mathbf{x}) : \mathbf{x} \in X\} \tag{1.8}$$

or

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X. \end{array} \tag{1.9}$$

1.6 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 $\mathbf{dom}(f) = \mathbb{R}^d$.

1.6.1 Sublevel sets and the Weierstrass Theorem

Definition 1.23. *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

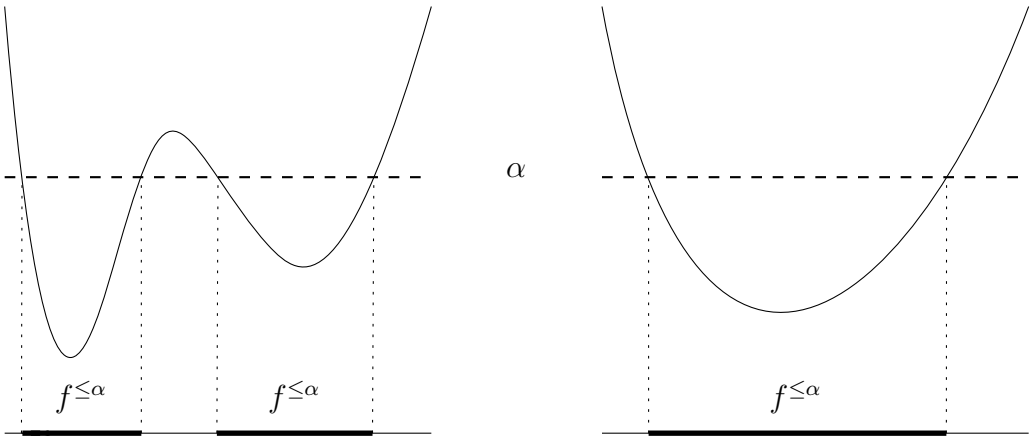


Figure 1.9: Sublevel set of a non-convex function (left) and a convex function (right)

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, sublevel sets are closed. The following (known as the Weierstrass Theorem) just formalizes an argument that we have made earlier.

Theorem 1.24. *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 $f(\mathbf{x}) > \alpha \geq f(\mathbf{x}^\star)$. □

1.7 Examples

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

1.7.1 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 1.10. You want to train your program with the set P, and after that, use it to recognize handwritten digits in arbitrary 28×28 images.

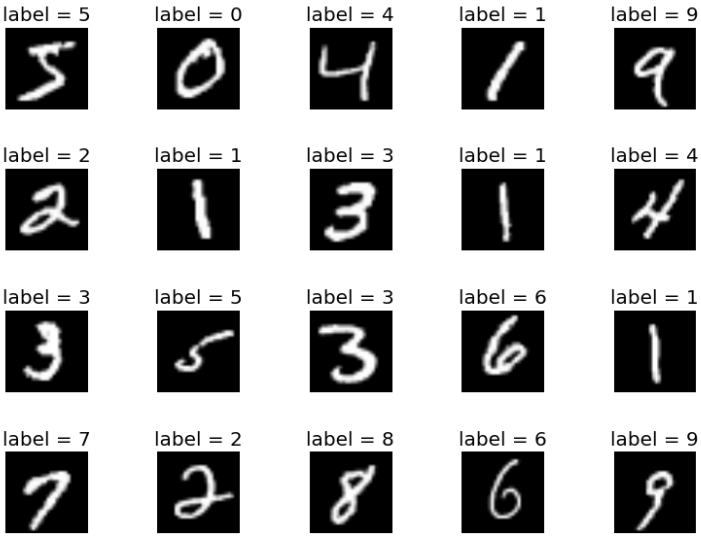


Figure 1.10: Some training images from the MNIST data set (picture from http://corochann.com/mnist-dataset-introduction-1138.html

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:

$$z_j = z_j(\mathbf{y}) = \frac{e^{y_j}}{\sum_{k=0}^{9} e^{y_k}}. \tag{1.10}$$

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.

$$\ell(W) = -\sum_{\mathbf{x}\in P} \ln\left(z_{d(\mathbf{x})}(W\mathbf{x})\right) = \sum_{\mathbf{x}\in P} \left(\ln\left(\sum_{k=0}^{9} e^{(W\mathbf{x})_k}\right) - (W\mathbf{x})_{d(\mathbf{x})} \right). \tag{1.11}$$

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 (1.10), 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.

1.7.2 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.[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 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 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 1.5.2, we are attempting a linear regression with least squares fit, i.e. we are making the hypothesis that

$$\text{GGPA} \approx w_0 + w_1 \cdot \text{GPA} + w_2 \cdot \text{TOEFL}. \tag{1.12}$$

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 (1.7), 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 9), 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 9). 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

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

This is minimized at

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

so if our initial hypothesis (1.12) is true, we should have

$$y_i \approx y_i^\star = 0.43 x_{i1} + 0.097 x_{i2} \tag{1.13}$$

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).

What we also see from (1.13) 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.

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 (1.13) and by the model $z_i^\star = 0.43x_{i1}$ that simply ignores the second input variable

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 $\|\mathbf{w}\|_1 =$

$\sum_{j=1}^{d} |w_j|$):

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathbf{w}^\top \mathbf{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathbf{w}\|_1 \leq R, \end{array} \tag{1.14}$$

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

$$\begin{array}{ll} \text{minimize} & f(w_1, w_2) = 10w_1^2 + 10w_2^2 + 1.99w_1w_2 - 8.7w_1 - 2.79w_2 + 2.09 \\ \text{subject to} & |w_1| + |w_2| \leq 0.2, \end{array} \tag{1.15}$$

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 (1.14) 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 (1.15) 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 1.11. 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 1.11,

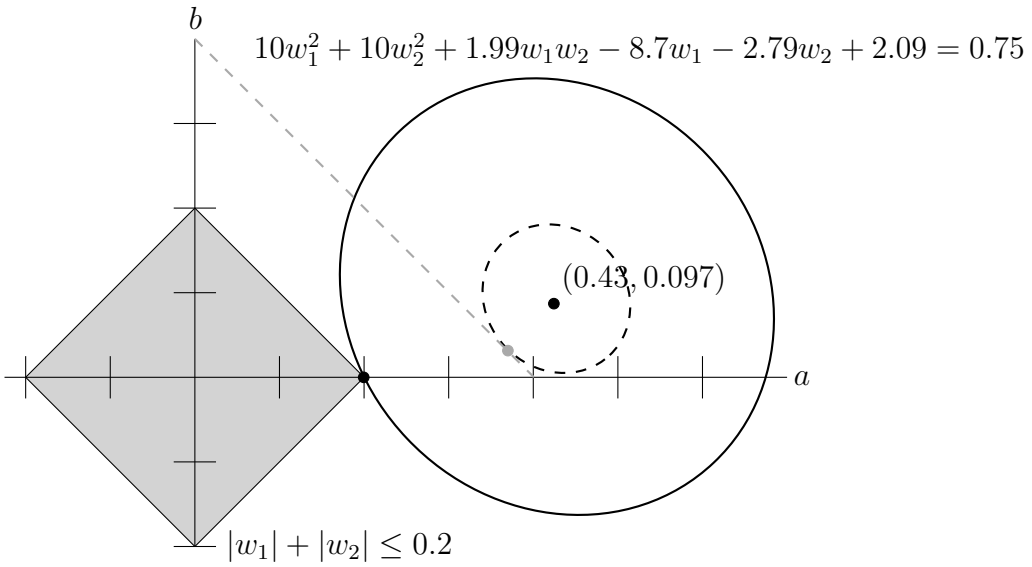


Figure 1.11: Lasso

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].

1.8 Exercises

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

Exercise 2. *Prove that a convex function (with $\mathbf{dom}(f)$ open) 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 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 3. *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 1.19.)*

Exercise 4. *Prove Lemma 1.13! 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 (1.11). Prove that ℓ is convex!*

Exercise 6. *Consider the logistic regression problem with two classes. Given a training set P consisting of datapoint and label pairs $(\mathbf{x}, y)$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, we define our loss ℓ for weight vector $\mathbf{w} \in \mathbb{R}^d$ to be*

$$\ell(\mathbf{w}) = \sum_{(\mathbf{x},y)\in P} -\ln\left(z(y\mathbf{w}^\top \mathbf{x})\right) ,$$

where $z(s) = 1/(1 + \exp(-s))$. This loss function is in fact a simplification of (1.11) when we only have two classes.

We say that the weight vector $\mathbf{w}$ is a separator *for P if for all $(\mathbf{x}, y) \in P$,*

$$y(\mathbf{w}^\top \mathbf{x}) \geq 0 .$$

A separator is said to be trivial *if for all $(\mathbf{x}, y) \in P$,*

$$y(\mathbf{w}^\top \mathbf{x}) = 0 .$$

For example $\mathbf{w} = 0$ is a trivial separator. Depending on the data P, there may be other trivial separators.

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

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!

Exercise 9. *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*

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

Chapter 2

Gradient Descent

Contents

2.1 Overview 33
2.2 The algorithm 34
2.3 Vanilla analysis 35
2.4 Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps 37
2.5 Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps 39
2.6 Acceleration for smooth convex functions: $\mathcal{O}(1/\sqrt{\varepsilon})$ steps 44
2.7 Interlude 47
2.8 Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps 48
2.9 Exercises 51

2.1 Overview

The gradient descent algorithm (including variants such as projected or stochastic gradient descent) is the most useful workhorse for minimizing loss functions in practice. The algorithm is extremely simple and surprisingly robust in the sense that it also works well for many loss functions that are not convex. While it is easy to construct (artificial) non-convex functions on which gradient descent goes completely astray, such functions do not seem to be typical in practice; however, understanding this on a theoretical level is an open problem, and only few results exist in this direction.

The vast majority of theoretical results concerning the performance of gradient descent hold for convex functions only. In this and the following chapters, we will present some of these results, but maybe more importantly, the main ideas behind them. As it turns out, the number of ideas that we need is rather small, and typically, they are shared between different results. Our approach is therefore to fully develop each idea once, in the context of a concrete result. If the idea reappears, we will typically only discuss the changes that are necessary in order to establish a new result from this idea. In order to avoid boredom from ideas that reappear too often, we omit other results and variants that one could also get along the lines of what we discuss.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function. We also assume that f has a global minimum $\mathbf{x}^\star$, and the goal is to find (an approximation of) $\mathbf{x}^\star$. This usually means that for a given $\varepsilon > 0$, we want to find $\mathbf{x} \in \mathbb{R}^d$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^\star) < \varepsilon.$$

Notice that we are not making an attempt to get near to $\mathbf{x}^\star$ itself — there can be several minima $\mathbf{x}_1^\star \neq \mathbf{x}^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star) = f(\mathbf{x}^\star)$.

Table 2.1 gives an overview of the results that we will prove. They concern several variants of gradient descent as well as several classes of functions. The significance of each algorithm and function class will briefly be discussed when it first appears.

In Chapter 6, we will also look at gradient descent on functions that are not convex. In this case, provably small approximation error can still be obtained for some particularly well-behaved functions (we will give an example). For smooth (but not necessarily convex) functions, we gener-

	Lipschitz convex functions	smooth convex functions	strongly convex functions	smooth & strongly convex functions
gradient descent	Thm. 2.1 $\mathcal{O}(1/\varepsilon^2)$	Thm. 2.7 $\mathcal{O}(1/\varepsilon)$		Thm. 2.12 $\mathcal{O}(\log(1/\varepsilon))$
accelerated gradient descent		Thm. 2.8 $\mathcal{O}(1/\sqrt{\varepsilon})$		
projected gradient descent	Thm. 3.2 $\mathcal{O}(1/\varepsilon^2)$	Thm. 3.4 $\mathcal{O}(1/\varepsilon)$		Thm. 3.5 $\mathcal{O}(\log(1/\varepsilon))$
proximal gradient descent		Thm. 3.14 $\mathcal{O}(1/\varepsilon)$		
subgradient descent	Thm. 4.7 $\mathcal{O}(1/\varepsilon^2)$		Thm. 4.11 $\mathcal{O}(1/\varepsilon)$	
stochastic gradient descent	Thm. 5.1 $\mathcal{O}(1/\varepsilon^2)$		Thm. 5.2 $\mathcal{O}(1/\varepsilon)$	

Table 2.1: Results on gradient descent. Below each theorem, the number of steps is given which the respective variant needs on the respective function class to achieve additive approximation error at most ε.

ally cannot show convergence in error, but a (much) weaker convergence property still holds.

2.2 The algorithm

Gradient descent is a very simple iterative algorithm for finding the desired approximation $\mathbf{x}$, under suitable conditions that we will get to. It computes a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of vectors such that $\mathbf{x}_0$ is arbitrary, and for each $t \geq 0$, $\mathbf{x}_{t+1}$ is obtained from $\mathbf{x}_t$ by making a step of $\mathbf{v}_t \in \mathbb{R}^d$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{v}_t.$$

How do we choose $\mathbf{v}_t$ in order to get closer to optimality, meaning that $f(\mathbf{x}_{t+1}) < f(\mathbf{x}_t)$?

From differentiability of f at $\mathbf{x}_t$ (Definition 1.7), we know that for $\|\mathbf{v}_t\|$ tending to 0,

$$f(\mathbf{x}_t + \mathbf{v}_t) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{v}_t + \underbrace{r(\mathbf{v}_t)}_{o(\|\mathbf{v}_t\|)} \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{v}_t.$$

To get any decrease in function value at all, we have to choose $\mathbf{v}_t$ such that $\nabla f(\mathbf{x}_t)^\top \mathbf{v}_t < 0$. But among all steps $\mathbf{v}_t$ of the same length, we should in fact choose the one with the most negative value of $\nabla f(\mathbf{x}_t)^\top \mathbf{v}_t$, so that we maximize our decrease in function value. This is achieved when $\mathbf{v}_t$ points into the direction of the negative gradient $-\nabla f(\mathbf{x}_t)$. But as differentiabilty guarantees decrease only for small steps, we also want to control how far we go along the direction of the negative gradient.

Therefore, the step of gradient descent is defined by

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t). \tag{2.1}$$

Here, γ is a fixed *stepsize*, but it may also make sense to have γ depend on t. For now, γ is fixed. We hope that for some reasonably small integer t, in the t-th iteration we get that $f(\mathbf{x}_t) - f(\mathbf{x}^\star) < \varepsilon$; see Figure 2.1 for an example.

Now it becomes clear why we are assuming that $\mathbf{dom}(f) = \mathbb{R}^d$: The update step (2.11) may in principle take us "anywhere", so in order to get a well-defined algorithm, we want to make sure that f is defined and differentiable everywhere.

The choice of γ is critical for the performance. If γ is too small, the process might take too long, and if γ is too large, we are in danger of overshooting. It is not clear at this point whether there is a "right" stepsize.

2.3 Vanilla analysis

Let $\mathbf{x}_t$ be some iterate in the sequence (2.11). We abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$, and will relate this vector to our current direction from an optimum $\mathbf{x}_t - \mathbf{x}^\star$. By definition of gradient descent (2.11), $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$, hence

$$\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma} (\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star). \tag{2.2}$$

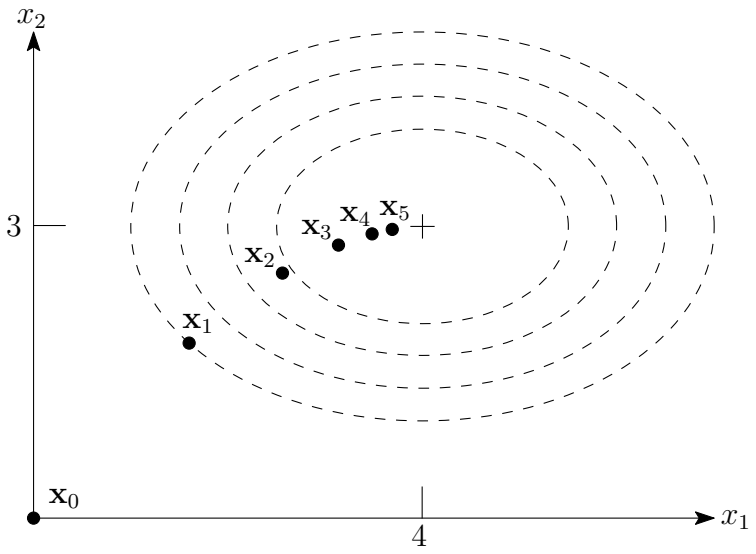


Figure 2.1: Example run of gradient descent on the quadratic function $f(x_1, x_2) = 2(x_1 - 4)^2 + 3(x_2 - 3)^2$ with global minimum $(4, 3)$; we have chosen $\mathbf{x}_0 = (0, 0), \gamma = 0.1$; dashed lines represent level sets of f (points of constant f-value)

Now we apply (somewhat out of the blue, but this will clear up in the next step) the basic vector equation $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$ (a.k.a. the cosine theorem) to rewrite the same expression as

$$\begin{aligned} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) &= \frac{1}{2\gamma} \left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \right) \\ &= \frac{1}{2\gamma} \left(\gamma^2 \|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \right) \\ &= \frac{\gamma}{2} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \right) \end{aligned} \tag{2.3}$$

Next we sum this up over the iterations t, so that the latter two terms in the bracket cancel in a telescoping sum.

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) &= \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2 \right) \\ &\leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \end{aligned} \tag{2.4}$$

So far, we have not used any properties of the function f or its gradient $\mathbf{g}_t$, except the definition of the update step $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathbf{g}_t$. Now we invoke convexity of f, or more precisely the first-order characterization of convexity (1.4) with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star). \tag{2.5}$$

Hence we further obtain

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2. \tag{2.6}$$

This gives us an upper bound for the *average* error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$, $t = 0, \ldots, T-1$, hence in particular for the error incurred by the iterate with the smallest function value. The last iterate is not necessarily the best one: gradient descent with fixed stepsize γ will in general also make steps that overshoot and actually increase the function value; see Exercise 12(i).

The question is of course: is this result any good? In general, the answer is no. A dependence on $\|\mathbf{x}_0 - \mathbf{x}^\star\|$ is to be expected (the further we start from $\mathbf{x}^\star$, the longer we will take); the dependence on the squared gradients $\|\mathbf{g}_t\|^2$ is more of an issue, and if we cannot control them, we cannot say much.

2.4 Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Here is the cheapest "solution" to squeeze something out of the vanilla analysis (2.4): let us simply assume that all gradients of f are bounded in norm. Equivalently, such functions are Lipschitz continuous over $\mathbb{R}^d$

by Theorem 1.10. (A small subtetly here is that in the situation of real-valued functions, Theorem 1.10 is talking about the spectral norm of the $(1 \times d)$-matrix (or row vector) $\nabla f(\mathbf{x})^\top$, while below, we are talking about the Euclidean norm of the (column) vector $\nabla f(\mathbf{x})$; but these two norms are the same; see Exercise 10.)

Assuming bounded gradients rules out many interesting functions, though. For example, $f(x) = x^2$ (a supermodel in the world of convex functions) already doesn't qualify, as $\nabla f(x) = 2x$—and this is unbounded as x tends to infinity. But let's care about supermodels later.

Theorem 2.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$. Choosing the stepsize*

$$\gamma := \frac{R}{B\sqrt{T}},$$

gradient descent (2.11) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{RB}{\sqrt{T}}.$$

Proof. This is a simple calculation on top of (2.6): after plugging in the bounds $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\mathbf{g}_t\| \leq B$, we get

$$\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2}B^2T + \frac{1}{2\gamma}R^2,$$

so want to choose γ such that

$$q(\gamma) = \frac{\gamma}{2}B^2T + \frac{R^2}{2\gamma}$$

is minimized. Setting the derivative to zero yields the above value of γ, and $q(R/(B\sqrt{T})) = RB\sqrt{T}$. Dividing by T, the result follows. □

This means that in order to achieve $\min_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \varepsilon$, we need

$$T \geq \frac{R^2B^2}{\varepsilon^2}$$

many iterations. This is not particularly good when it comes to concrete numbers (think of desired error $\varepsilon = 10^{-6}$ when R, B are somewhat larger). On the other hand, the number of steps does not depend on d, the dimension of the space. This is very important since we often optimize in high-dimensional spaces. Of course, R and B may depend on d, but in many relevant cases, this dependence is mild.

What happens if we don't know R and/or B? An idea is to "guess" R and B, run gradient descent with T and γ resulting from the guess, check whether the result has absolute error at most ε, and repeat with a different guess otherwise. This fails, however, since in order to compute the absolute error, we need to know $f(\mathbf{x}^\star)$ which we typically don't. But Exercise 13 asks you to show that knowing R is sufficient.

2.5 Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Our workhorse in the vanilla analysis was the first-order characterization of convexity: for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}). \tag{2.7}$$

Next we want to look at functions for which $f(\mathbf{y})$ can be bounded *from above* by $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$, up to at most quadratic error. The following definition applies to all differentiable functions, convexity is not required.

Definition 2.2. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subseteq \mathbf{dom}(f)$ convex and $L \in \mathbb{R}_+$. Function f is called* smooth *(with parameter L) over X if*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{2.8}$$

If $X = \mathbf{dom}(f)$, f is simply called smooth.

Recall that (2.7) says that for any $\mathbf{x}$, the graph of f is above its tangential hyperplane at $(\mathbf{x}, f(\mathbf{x}))$. In contrast, (2.8) says that for any $\mathbf{x} \in X$, the graph of f is below a not-too-steep tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$; see Figure 2.2.

This notion of smoothness has become standard in convex optimization, but the naming is somewhat unfortunate, since there is an (older) definition of a smooth function in mathematical analysis where it means a function that is infinitely often differentiable.

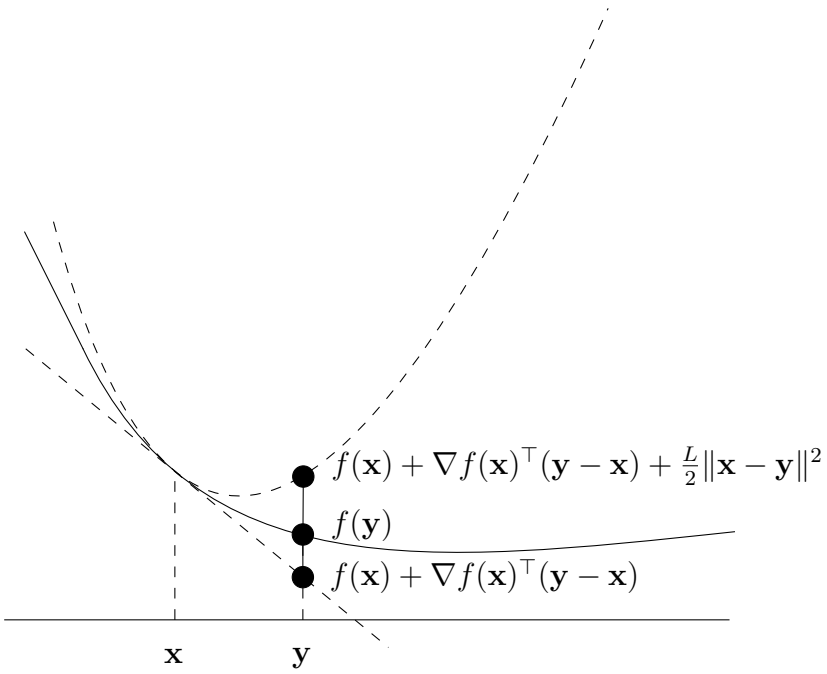


Figure 2.2: A smooth convex function

Let us discuss some cases. If $L = 0$, (2.7) and (2.8) together require that

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f),$$

meaning that f is an affine function. A simple calculation shows that our supermodel function $f(x) = x^2$ is smooth with parameter $L = 2$:

$$\begin{aligned} f(y) = y^2 &= x^2 + 2x(y-x) + (x-y)^2 \\ &= f(x) + f'(x)(y-x) + \frac{L}{2}(x-y)^2. \end{aligned}$$

More generally, we also claim that all quadratic functions of the form $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$ are smooth, where Q is a $(d \times d)$ matrix, $\mathbf{b} \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Because $\mathbf{x}^\top Q\mathbf{x} = \mathbf{x}^\top Q^\top\mathbf{x}$, we get that $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} = \frac{1}{2}\mathbf{x}^\top(Q + Q^\top)\mathbf{x}$, where $\frac{1}{2}(Q + Q^\top)$ is symmetric. Therefore, we can assume without loss of generality that Q is symmetric, i.e., it suffices to show that quadratic functions defined by symmetric functions are smooth.

Lemma 2.3 (Exercise 11)**.** *Let $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$, where Q is a symmetric $(d \times d)$ matrix, $\mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$. Then f is smooth with parameter $2\|Q\|$, where $\|Q\|$ is the spectral norm of Q (Definition 1.9).*

The (univariate) convex function $f(x) = x^4$ is not smooth (over $\mathbb{R}$): at $x = 0$, condition (2.8) reads as

$$y^4 \leq \frac{L}{2} y^2,$$

and there is obviously no L that works for all y. The function is smooth, however, over any bounded set X (Exercise 16).

In general—and this is the important message here—only functions of asymptotically at most quadratic growth can be smooth. It is tempting to believe that any such "subquadratic" function is actually smooth, but this is not true. Exercise 12(iii) provides a counterexample.

While bounded gradients are equivalent to Lipschitz continuity of f (Theorem 1.10), smoothness turns out to be equivalent to Lipschitz continuity of ∇f—if f is convex over the whole space. In general, Lipschitz continuity of ∇f implies smoothness, but not the other way around.

Lemma 2.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.*

(i) f is smooth with parameter L.

(ii) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

We will derive the direction (ii)⇒(i) as Lemma 6.1 in Chapter 6 (which neither requires convexity nor domain $\mathbb{R}^d$). The other direction is a bit more involved. A proof of the equivalence can be found in the lecture slides of L. Vandenberghe, `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf`.

The operations that we have shown to preserve convexity (Lemma 1.13) also preserve smoothness. This immediately gives us a rich collection of smooth functions.

Lemma 2.5 (Exercise 14)**.**

(i) Let $f_1, f_2, \ldots, f_m$ be smooth with parameters $L_1, L_2, \ldots, L_m$, and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then the function $f := \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$ over $\mathbf{dom}(f) := \bigcap_{i=1}^m \mathbf{dom}(f_i)$.

(ii) Let $f : \mathbf{dom}(f) \to \mathbb{R}$ with $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ be smooth with parameter L, and let $g : \mathbb{R}^m \to \mathbb{R}^d$ be 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 smooth with parameter $L\|A\|^2$ on $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$, where $\|A\|$ is the spectral norm of A (Definition 1.9).

We next show that for smooth convex functions, the vanilla analysis provides a better bound than it does under bounded gradients. In particular, we are now able to serve the supermodel $f(x) = x^2$.

We start with a preparatory lemma showing that gradient descent (with suitable stepsize γ) makes progress in function value on smooth functions in every step. We call this *sufficient decrease*, and maybe suprisingly, it does not require convexity.

Lemma 2.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L according to (2.8). With*

$$\gamma := \frac{1}{L},$$

gradient descent (2.11) satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

More specifically, this already holds if f is smooth with parameter L over the line segment connecting $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.

Proof. We apply the smoothness condition (2.8) and the definition of gradient descent that yields $\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$. We compute

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

□

Theorem 2.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L according to (2.8). Choosing stepsize*

$$\gamma := \frac{1}{L},$$

gradient descent (2.11) yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof. We apply sufficient decrease (Lemma 2.6) to bound the sum of the $\|\mathbf{g}_t\|^2 = \|\nabla f(\mathbf{x}_t)\|^2$ after step (2.6) of the vanilla analysis as follows:

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \leq \sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_T). \tag{2.9}$$

With $\gamma = 1/L$, (2.6) then yields

$$\begin{aligned}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) &\leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2,\end{aligned}$$

equivalently

$$\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2. \tag{2.10}$$

Because $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ for each $0 \leq t \leq T$ by Lemma 2.6, by taking the average we get that

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

□

This improves over the bounds of Theorem 2.1. With $R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$, we now only need

$$T \geq \frac{R^2 L}{2\varepsilon}$$

iterations instead of R^2B^2/ε^2 to achieve absolute error at most ε.

Exercise 15 shows that we do not need to know L to obtain the same asymptotic runtime.

Interestingly, the bound in Theorem 2.7 can be improved—but not by much. Fixing L and $R = \|\mathbf{x}_0 - \mathbf{x}^\star\|$, the bound is of the form $O(1/T)$. Lee and Wright have shown that a better upper bound of $o(1/T)$ holds, but that for any fixed $\delta > 0$, a lower bound of $\Omega(1/T^{1+\delta})$ also holds [LW19].

2.6 Acceleration for smooth convex functions: $\mathcal{O}(1/\sqrt{\varepsilon})$ steps

Let's take a step back, forget about gradient descent for a moment, and just think about what we actually use the algorithm for: we are minimizing a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$, where we are assuming that we have acccess to the gradient vector $\nabla f(\mathbf{x})$ at any given point $\mathbf{x}$.

But is it clear that gradient descent is the best algorithm for this task? After all, it is just *some* algorithm that is using gradients to make progress locally, but there might be other (and better) such algorithms. Let us define a *first-order method* as an algorithm that only uses gradient information to minimize f. More precisely, we allow a first-order method to access f only via an oracle that is able to return values of f and ∇f at arbitrary points. Gradient descent is then just a specific first-order method.

For any class of convex functions, one can then ask a natural question: What is the best first-order method for the function class, the one that needs the smallest number of oracle calls in the worst case, as a function of the desired error ε? In particular, is there a method that asymptotically beats gradient descent?

There is an interesting history here: in 1979, Nemirovski and Yudin have shown that *every* first-order method needs in the worst case $\Omega(1/\sqrt{\varepsilon})$ steps (gradient evaluations) in order to achieve an additive error of ε on smooth functions [NY83]. Recall that we have seen an upper bound of $O(1/\varepsilon)$ for gradient descent in the previous section; in fact, this upper bound was known to Nemirovsky and Yudin already. Reformulated in the language of the previous section, there is a first-order method (gradient descent) that attains additive error $O(1/T)$ after T steps, and all first-order methods have additive error $\Omega(1/T^2)$ in the worst case.

The obvious question resulting from this was whether there actually exists a first-order method that has additive error $O(1/T^2)$ after T steps, on every smooth function. This was answered in the affirmative by Nesterov in 1983 when he proposed an algorithm that is now known as *(Nesterov's) accelerated gradient descent* [Nes83]. Nesterov's book (Sections 2.1 and 2.2) is a comprehensive source for both lower and upper bound [Nes18].

It is not easy to understand why the accelerated gradient descent algorithm is an optimal first-order method, and how Nesterov even arrived at it. A number of alternative derivations of optimal algorithms have been given by other authors, usually claiming that they provide a more natural or easier-to-grasp approach. However, each alternative approach requires some understanding of other things, and there is no well-established "simplest approach". Here, we simply throw the algorithm at the reader, without any attempt to motivate it beyond some obvious words. Then we present a short proof that the algorithm is indeed optimal.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and smooth with parameter L. *Accelerated gradient descent* is the following algorithm: choose $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_0$ arbitrary. For $t \geq 0$, set

$$\mathbf{y}_{t+1} := \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \tag{2.11}$$

$$\mathbf{z}_{t+1} := \mathbf{z}_t - \frac{t+1}{2L}\nabla f(\mathbf{x}_t), \tag{2.12}$$

$$\mathbf{x}_{t+1} := \frac{t+1}{t+3}\mathbf{y}_{t+1} + \frac{2}{t+3}\mathbf{z}_{t+1}. \tag{2.13}$$

This means, we are performing a normal "smooth step" from $\mathbf{x}_t$ to obtain $\mathbf{y}_{t+1}$ and a more aggressive step from $\mathbf{z}_t$ to get $\mathbf{z}_{t+1}$. The next iterate $\mathbf{x}_{t+1}$ is a weighted average of $\mathbf{y}_{t+1}$ and $\mathbf{z}_{t+1}$, where we compensate for the more aggressive step by giving $\mathbf{z}_{t+1}$ a relatively low weight.

Theorem 2.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L according to (2.8). Accelerated gradient descent (2.11), (2.12), and (2.13), yields*

$$f(\mathbf{y}_T) - f(\mathbf{x}^\star) \leq \frac{2L\left\|\mathbf{z}_0 - \mathbf{x}^\star\right\|^2}{T(T+1)}, \quad T > 0.$$

Comparing this bound with the one from Theorem 2.7, we see that the error is now indeed $O(1/T^2)$ instead of $O(1/T)$; to reach error at most ε,

accelerated gradient descent therefore only needs $O(1/\sqrt{\varepsilon})$ steps instead of $O(1/\varepsilon)$.

Proof. The analysis uses a *potential function argument* [BG17]. We assign a potential $\Phi(t)$ to each time t and show that $\Phi(t+1) \leq \Phi(t)$. The potential is

$$\Phi(t) := t(t+1)\left(f(\mathbf{y}_t) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|^2.$$

If we can show that the potential always decreases, we get

$$\underbrace{T(T+1)\left(f(\mathbf{y}_T) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_T - \mathbf{x}^\star\right\|^2}_{\Phi(T)} \leq \underbrace{2L\left\|\mathbf{z}_0 - \mathbf{x}^\star\right\|^2}_{\Phi(0)},$$

from which the statement immediately follows. For the argument, we need three well-known ingredients: (i) sufficient decrease (Lemma 2.6) for step (2.11) with $\gamma = 1/L$:

$$f(\mathbf{y}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2; \tag{2.14}$$

(ii) the vanilla analysis (Section 2.3) for step (2.12) with $\gamma = \frac{t+1}{2L}$, $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$:

$$\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star) = \frac{t+1}{4L}\|\mathbf{g}_t\|^2 + \frac{L}{t+1}\left(\|\mathbf{z}_t - \mathbf{x}^\star\|^2 - \|\mathbf{z}_{t+1} - \mathbf{x}^\star\|^2\right); \tag{2.15}$$

(iii) convexity:

$$f(\mathbf{x}_t) - f(\mathbf{w}) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{w}), \quad \mathbf{w} \in \mathbb{R}^d. \tag{2.16}$$

On top of this, we perform some simple calculations next. By definition, the potentials are

$$\begin{aligned}\Phi(t+1) &= t(t+1)\left(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)\right) + 2(t+1)\left(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_{t+1} - \mathbf{x}^\star\right\|^2 \\ \Phi(t) &= t(t+1)\left(f(\mathbf{y}_t) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|^2\end{aligned}$$

Now,

$$\Delta := \frac{\Phi(t+1) - \Phi(t)}{t+1}$$

can be bounded as follows.

$$\begin{aligned}
\Delta &= t\left(f(\mathbf{y}_{t+1})-f(\mathbf{y}_t)\right)+2\left(f(\mathbf{y}_{t+1})-f(\mathbf{x}^\star)\right)+\frac{2L}{t+1}\left(\|\mathbf{z}_{t+1}-\mathbf{x}^\star\|^2-\|\mathbf{z}_t-\mathbf{x}^\star\|^2\right)\\
&\overset{(2.15)}{=} t\left(f(\mathbf{y}_{t+1})-f(\mathbf{y}_t)\right)+2\left(f(\mathbf{y}_{t+1})-f(\mathbf{x}^\star)\right)+\frac{t+1}{2L}\|\mathbf{g}_t\|^2-2\mathbf{g}_t^\top(\mathbf{z}_t-\mathbf{x}^\star)\\
&\overset{(2.14)}{\leq} t\left(f(\mathbf{x}_t)-f(\mathbf{y}_t)\right)+2\left(f(\mathbf{x}_t)-f(\mathbf{x}^\star)\right)-\frac{1}{2L}\|\mathbf{g}_t\|^2-2\mathbf{g}_t^\top(\mathbf{z}_t-\mathbf{x}^\star)\\
&\leq t\left(f(\mathbf{x}_t)-f(\mathbf{y}_t)\right)+2\left(f(\mathbf{x}_t)-f(\mathbf{x}^\star)\right)-2\mathbf{g}_t^\top(\mathbf{z}_t-\mathbf{x}^\star)\\
&\overset{(2.16)}{\leq} t\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{y}_t)+2\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star)-2\mathbf{g}_t^\top(\mathbf{z}_t-\mathbf{x}^\star)\\
&= \mathbf{g}_t^\top((t+2)\mathbf{x}_t-t\mathbf{y}_t-2\mathbf{z}_t)\\
&\overset{(2.13)}{=} \mathbf{g}_t^\top\mathbf{0}=0.
\end{aligned}$$

Hence, we indeed have $\Phi(t+1)\leq\Phi(t)$. □

2.7 Interlude

Let us get back to the supermodel $f(x)=x^2$ (that is smooth with parameter $L=2$, as we observed before). According to Theorem 2.7, gradient descent (2.11) with stepsize $\gamma=1/2$ satisfies

$$f(x_T)\leq\frac{1}{T}x_0^2. \tag{2.17}$$

Here we used that the minimizer is $x^\star=0$. Let us check how good this bound really is. For our concrete function and concrete stepsize, (2.11) reads as

$$x_{t+1}=x_t-\frac{1}{2}\nabla f(x_t)=x_t-x_t=0,$$

so we are always done after one step! But we will see in the next section that this is only because the function is particularly beautiful, and on top of that, we have picked the best possible smoothness parameter. To simulate a more realistic situation here, let us assume that we have not looked at the supermodel too closely and found it to be smooth with parameter $L=4$ only (which is a suboptimal but still valid parameter). In this case, $\gamma=1/4$ and (2.11) becomes

$$x_{t+1}=x_t-\frac{1}{4}\nabla f(x_t)=x_t-\frac{x_t}{2}=\frac{x_t}{2}.$$

So, we in fact have

$$f(x_T) = f\left(\frac{x_0}{2^T}\right) = \frac{1}{2^{2T}} x_0^2. \tag{2.18}$$

This is still vastly better than the bound of (2.17)! While (2.17) requires $T \approx x_0^2/\varepsilon$ to achieve $f(x_T) \leq \varepsilon$, (2.18) requires only

$$T \approx \frac{1}{2} \log\left(\frac{x_0^2}{\varepsilon}\right),$$

which is an exponential improvement in the number of steps.

2.8 Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

The supermodel function $f(x) = x^2$ is not only smooth ("not too curved") but also *strongly convex* ("not too flat"). It will turn out that this is the crucial ingredient that makes gradient descent fast.

Definition 2.9. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a convex and differentiable function, $X \subseteq \mathbf{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called* strongly convex *(with parameter μ) over X if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{2.19}$$

If $X = \mathbf{dom}(f)$, f is simply called strongly convex.

While smoothness according to (2.8) says that for any $\mathbf{x} \in X$, the graph of f is *below* a *not-too-steep* tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$, strong convexity means that the graph of f is *above* a *not-too-flat* tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$. The graph of a smooth *and* strongly convex function is therefore at every point wedged between two paraboloids; see Figure 2.3.

We can also interpret (2.19) as a strengthening of convexity. In the form of (2.7), convexity reads as

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f),$$

and therefore says that every convex function satisfies (2.19) with $\mu = 0$.

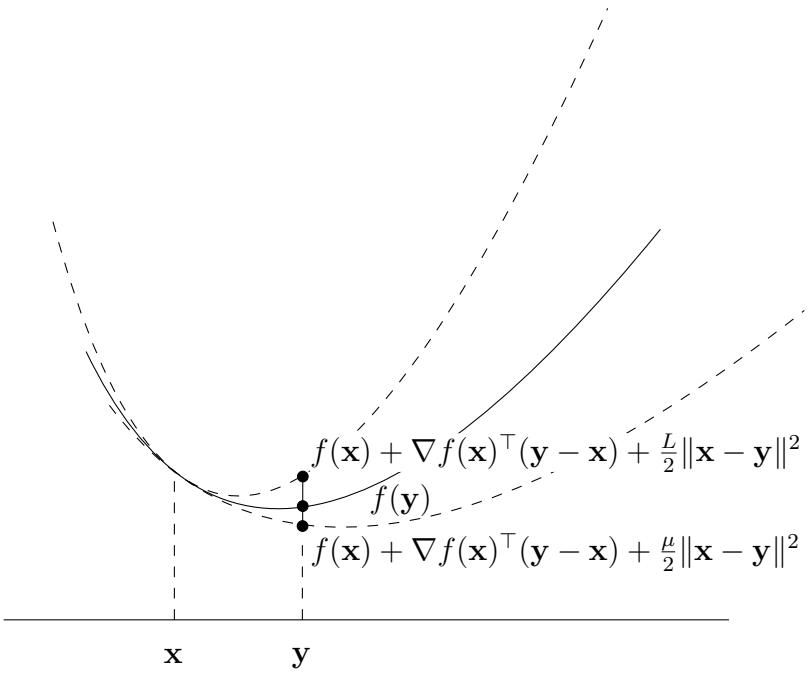


Figure 2.3: A smooth and strongly convex function

Lemma 2.10 (Exercise 17)**.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex with parameter $\mu > 0$, then f is strictly convex and has a unique global minimum.*

The supermodel $f(x) = x^2$ is particularly beautiful since it is both smooth and strongly convex with the same parameter $L = \mu = 2$ (going through the calculations in Exercise 11 will reveal this). We can easily characterize the class of particularly beautiful functions. These are exactly the ones whose sublevel sets are ℓ_2-balls.

Lemma 2.11 (Exercise 18)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex with parameter $\mu > 0$ and smooth with parameter μ. Prove that f is of the form*

$$f(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{x} - \mathbf{b}\|^2 + c,$$

where $\mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$.

Once we have a unique global minimum $\mathbf{x}^\star$, we can attempt to prove that $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^\star$ in gradient descent. We start from the vanilla analysis

(2.3) and plug in the lower bound $\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$ resulting from strong convexity. We get

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2. \tag{2.20}$$

Rewriting this yields a bound on $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$ in terms of $\|\mathbf{x}_t - \mathbf{x}^\star\|^2$, along with some "noise" that we still need to take care of:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2. \tag{2.21}$$

Theorem 2.12. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Suppose that f is smooth with parameter L according to (3.5) and strongly convex with parameter $\mu > 0$ according to (3.9). Exercise 20 asks you to prove that there is a unique global minimum $\mathbf{x}^\star$ of f. Choosing*

$$\gamma := \frac{1}{L},$$

gradient descent (2.11) with arbitrary $\mathbf{x}_0$ satisfies the following two properties.

(i) Squared distances to $\mathbf{x}^\star$ are geometrically decreasing:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2, \quad t \geq 0.$$

(ii) The absolute error after T iterations is exponentially small in T:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof. For (i), we show that the noise in (2.21) disappears. By sufficient decrease (Lemma 2.6), we know that

$$f(\mathbf{x}^\star) - f(\mathbf{x}_t) \leq f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2,$$

and hence the noise can be bounded as follows, using $\gamma = 1/L$, multiplying by 2γ and rearranging the terms, we get:

$$2\gamma\left(f(\mathbf{x}^\star) - f(\mathbf{x}_t)\right) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 \leq 0,$$

Hence, (2.21) actually yields

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 = \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

and

$$\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

The bound in (ii) follows from smoothness (2.8), using $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ (Lemma 1.17):

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \nabla f(\mathbf{x}^\star)^\top(\mathbf{x}_T - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 = \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2.$$

□

This implies that after

$$T \geq \frac{L}{\mu}\ln\left(\frac{R^2 L}{2\varepsilon}\right),$$

iterations, we reach absolute error at most ε.

2.9 Exercises

Exercise 10. *Let $\mathbf{c} \in \mathbb{R}^d$. Prove that the spectral norm of $\mathbf{c}^\top$ equals the Euclidean norm of $\mathbf{c}$, meaning that*

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{|\mathbf{c}^\top\mathbf{x}|}{\|\mathbf{x}\|} = \|\mathbf{c}\|.$$

Exercise 11. *Prove Lemma 2.3: The quadratic function $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$ is smooth with parameter $2\|Q\|$.*

Exercise 12. *Consider the function $f(x) = |x|^{3/2}$ for $x \in \mathbb{R}$.*

(i) Prove that f is strictly convex and differentiable, with a unique global minimum $x^\star = 0$.

(ii) Prove that for every fixed stepsize γ in gradient descent (2.11) applied to f, there exists x_0 for which $f(x_1) > f(x_0)$.

(iii) Prove that f is not smooth.

(iv) Let $X \subseteq \mathbb{R}$ be a closed convex set such that $0 \in X$ and $X \neq \{0\}$. Prove that f is not smooth over X.

Exercise 13. *In order to obtain average error at most ε in Theorem 2.1, we need to choose iteration number and stepsize as*

$$T \geq \left(\frac{RB}{\varepsilon}\right)^2, \quad \gamma := \frac{R}{B\sqrt{T}}.$$

If R or B are unknown, we cannot do this.

But suppose that we know R. Develop an algorithm that—not knowing B—finds a vector $\mathbf{x}$ such that $f(\mathbf{x}) - f(\mathbf{x}^\star) < \varepsilon$, using at most

$$\mathcal{O}\left(\left(\frac{RB}{\varepsilon}\right)^2\right)$$

many gradient descent steps!

Exercise 14. *Prove Lemma 2.5! (Operations which preserve smoothness)*

Exercise 15. *In order to obtain average error at most ε in Theorem 2.7, we need to choose*

$$\gamma := \frac{1}{L}, \quad T \geq \frac{R^2 L}{2\varepsilon},$$

if $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$. If L is unknown, we cannot do this.

But suppose that we know R. Develop an algorithm that—not knowing L—finds a vector $\mathbf{x}$ such that $f(\mathbf{x}) - f(\mathbf{x}^\star) < \varepsilon$, using at most

$$\mathcal{O}\left(\frac{R^2 L}{2\varepsilon}\right)$$

many gradient descent steps!

Exercise 16. *Let $a \in \mathbb{R}$. Prove that $f(x) = x^4$ is smooth over $X = (-a, a)$ and determine a concrete smoothness parameter L.*

Exercise 17. *Prove Lemma 2.10! (Strongly convex functions have unique global minimum)*

Exercise 18. *Prove Lemma 2.11! (Strongly convex and smooth functions)*