Contents

Chapter 1

Theory of Convex Functions

Contents

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

1.1 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

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

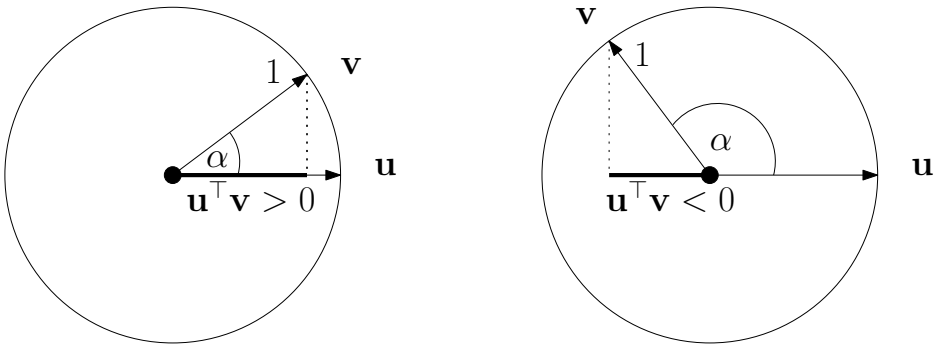
For nonzero vectors, this is equivalent to

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

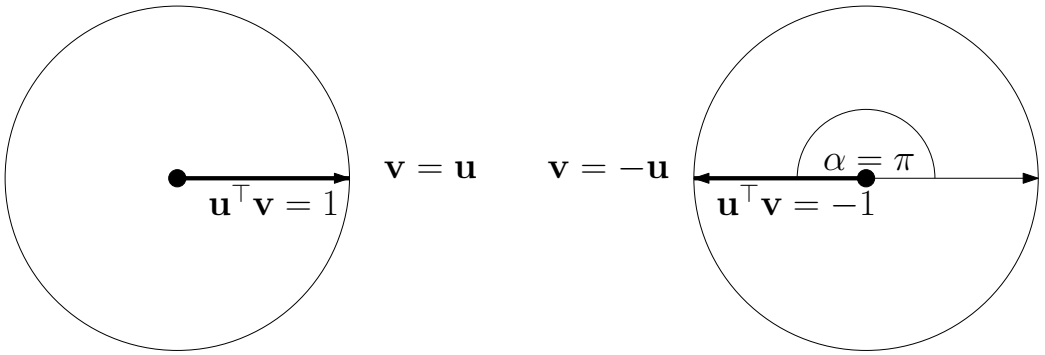
and this fraction can be used to define the angle α between $\mathbf{u}$ and $\mathbf{v}$:

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

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,

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

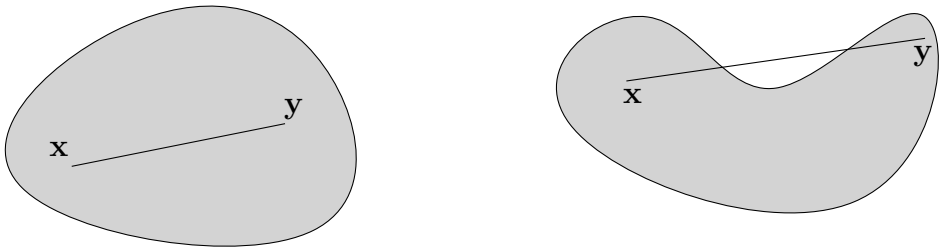


Figure 1.1: A convex set (left) and a non-convex set (right)

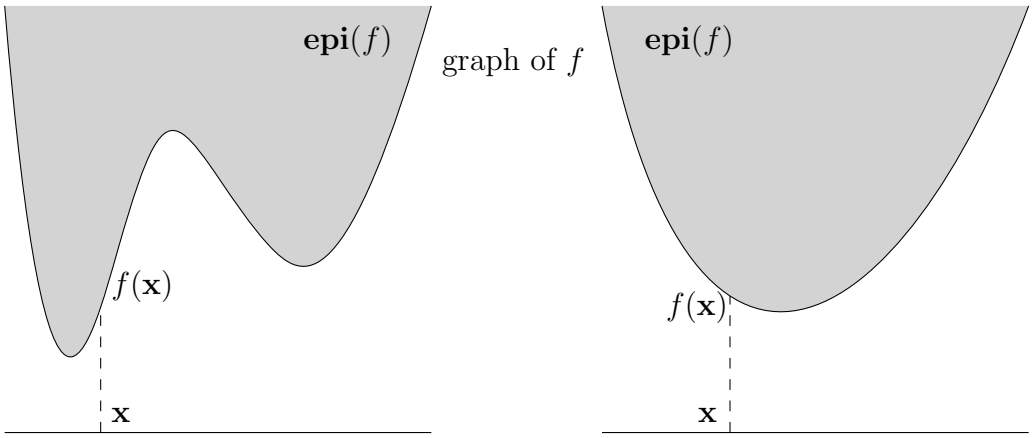


Figure 1.2: Graph and epigraph of a non-convex function (left) and a convex function (right)

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.

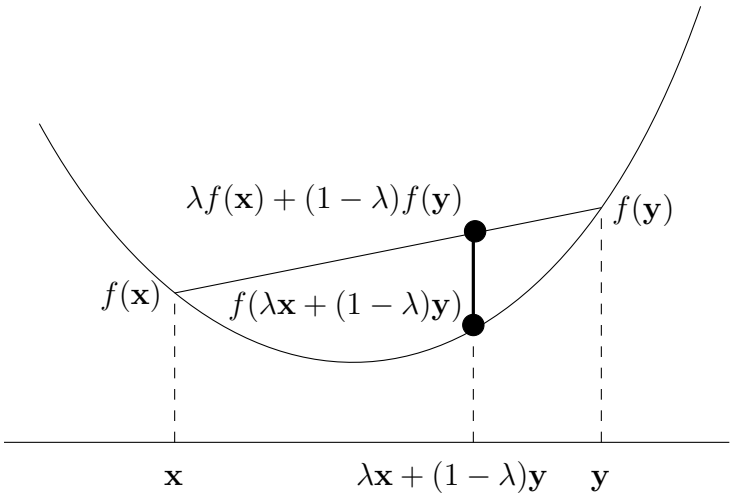


Figure 1.3: A convex function

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.

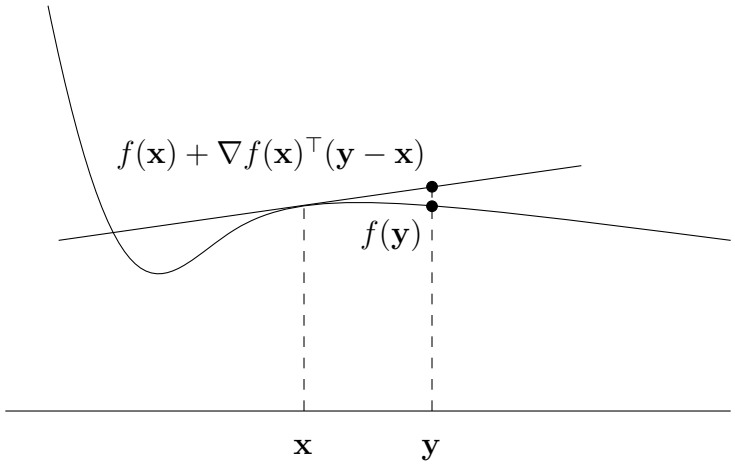


Figure 1.4: If f is differentiable at $\mathbf{x}$, the graph of f is locally (around $\mathbf{x}$) approximated by a tangent hyperplane

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 $t \in (0, 1)$,

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

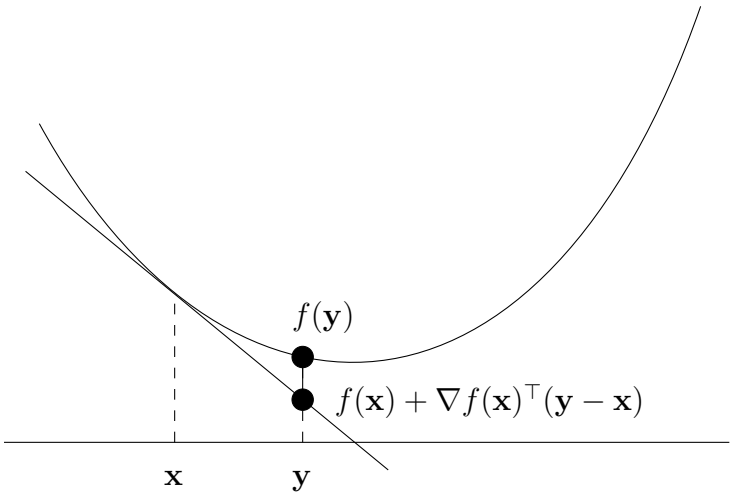


Figure 1.5: First-order characterization of convexity

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.

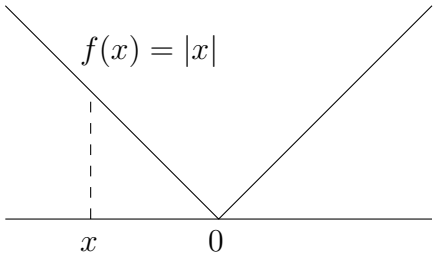


Figure 1.6: A non-differentiable convex function

1.4.3 Second-order characterization of convexity

If $f : \mathbf{dom}(f) \to \mathbb{R}$ is twice continuously differentiable (meaning that the function ∇f is differentiable and $\nabla^2 f$ is continuous), 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 continuously 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 continuously differentiable. If the Hessian $\nabla^2 f(\mathbf{x}) \succ \mathbf{0}$ for every $x \in \mathbf{dom}(f)$ (i.e., $\mathbf{z}^\top \nabla^2 f(\mathbf{x}) \mathbf{z} > 0$ for any $\mathbf{z} \neq \mathbf{0}$), then f is strictly convex.*

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

Lemma 1.20. *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 (left).

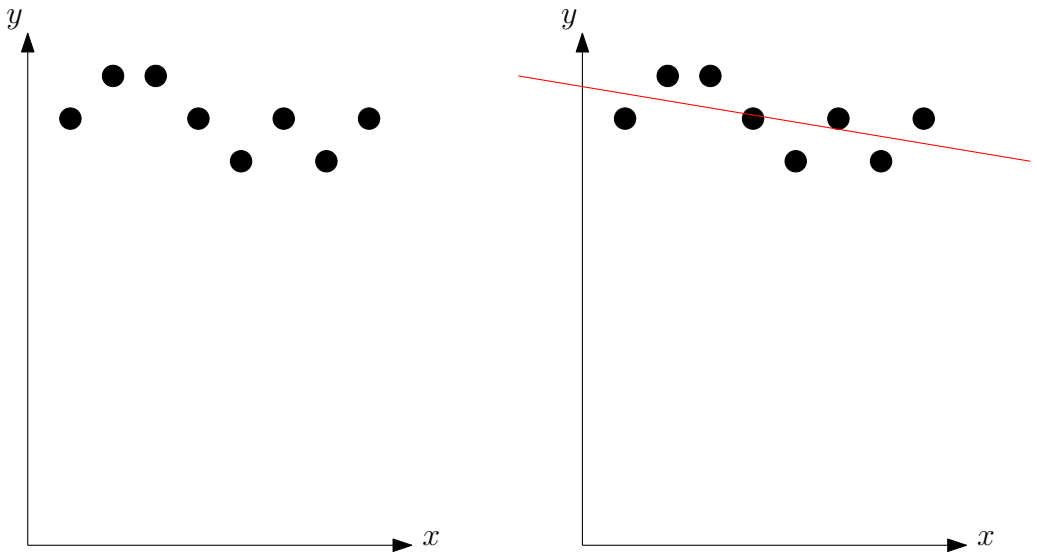


Figure 1.7: Data points in $\mathbb{R}^2$ (left) and least-squares fit (right)

Also, for simplicity (and quite appropriately in this case), let us restrict to fitting a linear model, of more formally to fit non-vertical lines of the form $y = w_0 + w_1 x$. If (x_i, y_i) is the i-th data point, the least squares fit chooses w_0, w_1 such that the *least squares objective*

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

is minimized. It easily follows from Lemma 1.13 that f is convex. In fact,

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

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 $\mathbf{x} \in X$.

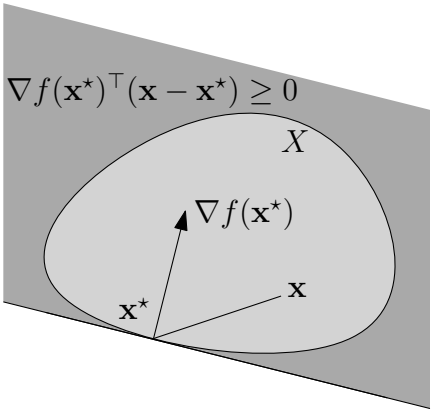


Figure 1.8: Optimality condition for constrained optimization

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

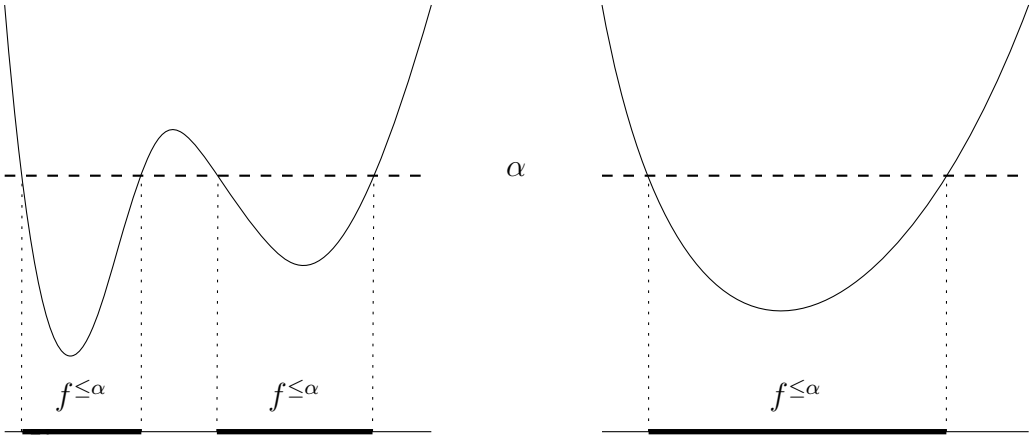


Figure 1.9: Sublevel set of a non-convex function (left) and a convex function (right)

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.

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

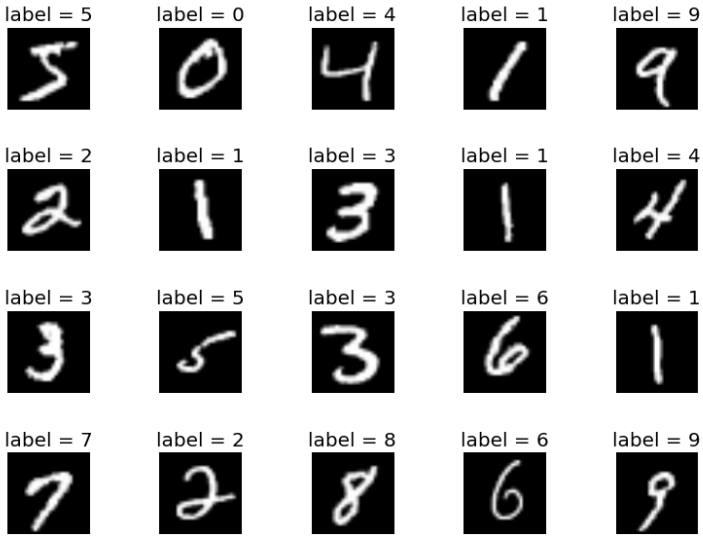


Figure 1.10: Some training images from the MNIST data set (picture from http://corochann.com/mnist-dataset-introduction-1138.html)

error induced by a given matrix W.

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

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]

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

[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.

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* $\mathbf{x}_1, \dots, \mathbf{x}_n$, where each vector $\mathbf{x}_i \in \mathbb{R}^d$ consists of d input variables; then we have n *outputs* $y_1, \dots, y_n \in \mathbb{R}$. 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* $w_0, \mathbf{w} = (w_1, \dots, w_d) \in \mathbb{R}^d$, 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).

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

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

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

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)

$\mathbf{w}^*$, 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.

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

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

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

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

tion 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 $(0.2, 0)$.

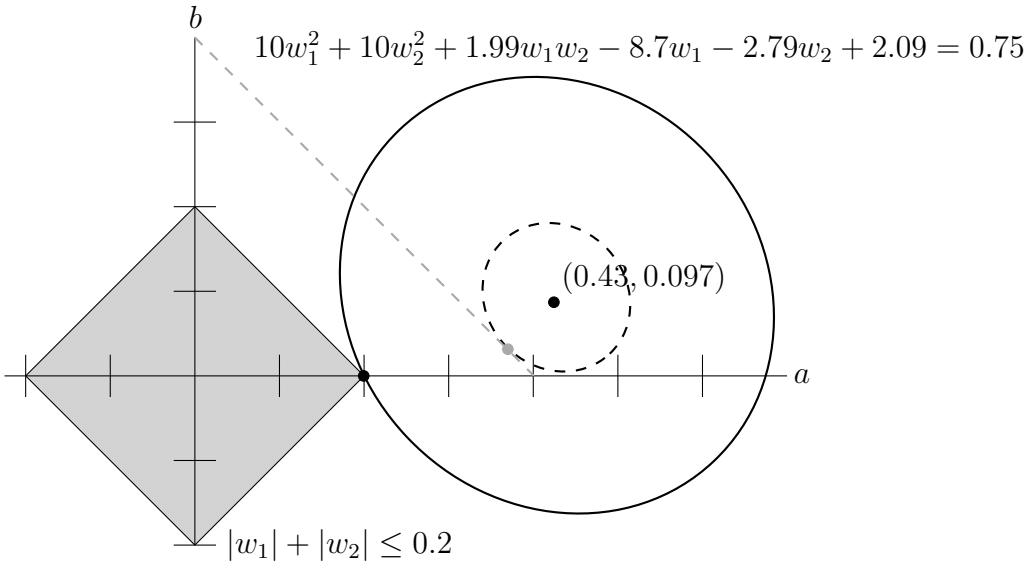


Figure 1.11: Lasso

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, 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

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

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

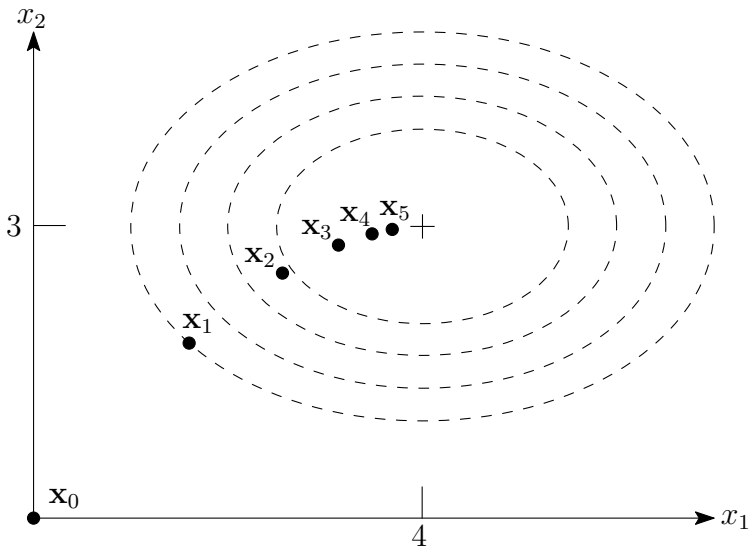


Figure 2.1: Example run of gradient descent on the quadratic function $f(x_1, x_2) = 2(x_1 - 4)^2 + 3(x_2 - 3)^2$ with global minimum $(4, 3)$; we have chosen $\mathbf{x}_0 = (0, 0), \gamma = 0.1$; dashed lines represent level sets of f (points of constant f-value)

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.

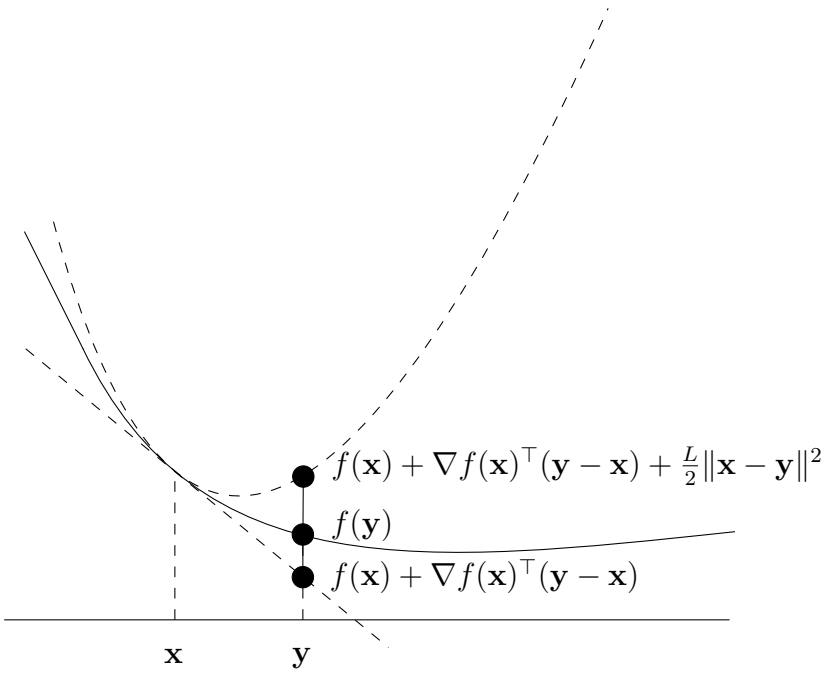


Figure 2.2: A smooth convex function

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 $\mu = 0$.

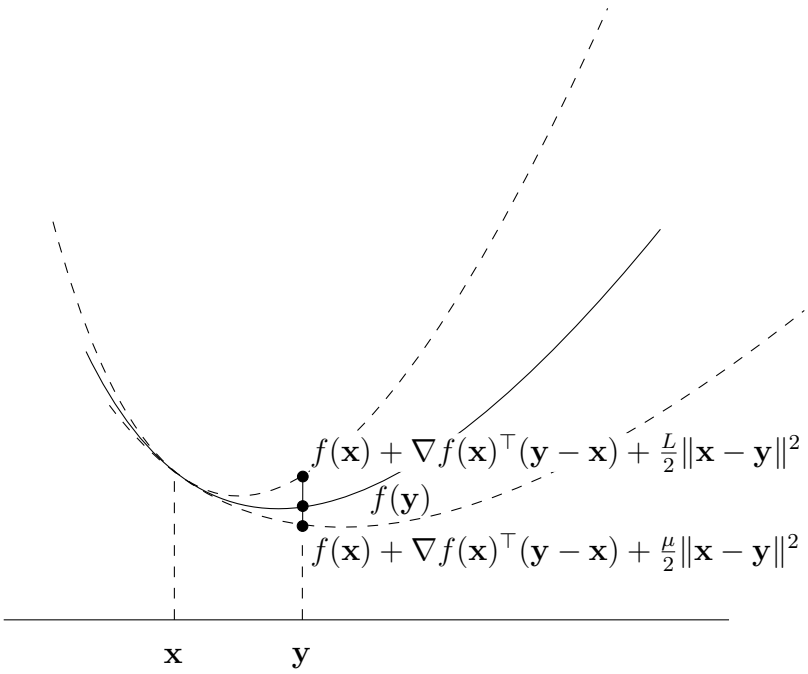


Figure 2.3: A smooth and strongly convex function

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)*

Chapter 3

Projected and Proximal Gradient Descent

Contents

3.1 The Algorithm

Another way to control gradients in (2.4) is to minimize f over a closed convex subset $X \subseteq \mathbb{R}^d$. For example, we may have a constrained optimization problem to begin with (for example the LASSO in Section 1.7.2), or we happen to know some region X containing a global minimum $\mathbf{x}^\star$, so that we can restrict our search to that region. In this case, gradient descent also works, but we need an additional *projection step*. After all, it can happen that some iteration of (2.11) takes us "into the wild" (out of X) where we have no business to do. *Projected* gradient descent is the following modification. We choose $\mathbf{x}_0 \in X$ arbitrary and for $t \geq 0$ define

$$\begin{aligned} \mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), && (3.1) \\ \mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \operatorname*{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2. && (3.2) \end{aligned}$$

This means, after each iteration, we project the obtained iterate $\mathbf{y}_{t+1}$ back to X. This may be very easy (think of X as the unit ball in which case we just have to scale $\mathbf{y}_{t+1}$ down to length 1 if it is longer). But it may also be very difficult. In general, computing $\Pi_X(\mathbf{y}_{t+1})$ means to solve an auxiliary convex constrained minimization problem in each step! Here, we are just assuming that we can do this. The projection is well-defined since $d_{\mathbf{y}} := \|\mathbf{x} - \mathbf{y}\|^2$ has bounded sublevel sets. Moreover, $d_{\mathbf{y}}(\mathbf{x})$ is strictly convex, so the minimum over X (that exists by continuity of $d_{\mathbf{y}}$ and compactness of X intersected with any nonempty sublevel set) is unique by Lemma 1.20. We note that finding an initial $\mathbf{x}_0 \in X$ also reduces to projection (of $\mathbf{0}$, for example) onto X.

3.2 Bounded gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

As in the unconstrained case, let us first assume that gradients are bounded by a constant B—this time over X. This implies that f is B-Lipschitz over X (see Theorem 1.10), but the converse may not hold.

To show that the vanilla analysis still goes through, we need the following

Fact 3.1. *Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then*

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

Part (i) says that the vectors $\mathbf{x} - \Pi_X(\mathbf{y})$ and $\mathbf{y} - \Pi_X(\mathbf{y})$ form an obtuse angle, and (ii) equivalently says that the square of the long side $\mathbf{x} - \mathbf{y}$ in the triangle formed by the three points is at least the sum of squares of the two short sides; see Figure 3.1.

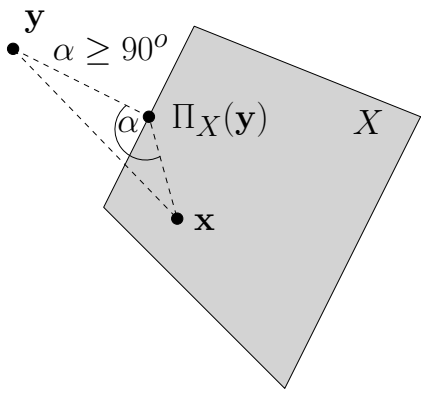


Figure 3.1: Illustration of Fact 3.1

Proof. $\Pi_X(\mathbf{y})$ is by definition a minimizer of the (differentiable) convex function $d_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|^2$ over X, and (i) is just the equivalent optimality condition of Lemma 1.22. We need X to be closed in the first place in order to ensure that we can project onto X (see Exercise 20 applied with $d_{\mathbf{y}}(\mathbf{x})$). Indeed, for example, 1 has no closest point in the set $[-\infty, 0) \in \mathbb{R}^1$.

Part (ii) follows from (i) via the (by now well-known) equation $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$. □

If we minimize f over a closed and *bounded* (= compact) convex set X, we get the existence of a minimizer and a bound R for the initial distance to it for free; assuming that f is *continuously* differentiable, we also have a bound B for the gradient norms over X. This is because then $\mathbf{x} \mapsto \|\nabla f(\mathbf{x})\|$ is a continuous function that attains a maximum over X. In this case, our vanilla analysis yields a much more useful result than the one in Theorem 2.1, with the same stepsize and the same number of steps.

Theorem 3.2. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and differentiable, $X \subseteq \mathbf{dom}(f)$ closed and convex, $\mathbf{x}^\star$ a minimizer of f over X; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, and that $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x} \in X$. Choosing the constant stepsize*

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

projected gradient descent (3.1) with $\mathbf{x}_0 \in X$ yields

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

Proof. The only required changes to the vanilla analysis are that in steps (2.2) and (2.3), $\mathbf{x}_{t+1}$ needs to be replaced by $\mathbf{y}_{t+1}$ as this is the real next (non-projected) gradient descent iterate after these steps; we therefore get

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2\right). \tag{3.3}$$

From Fact 3.1 (ii) (with $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$), we obtain $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$, hence we get

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \tag{3.4}$$

and return to the previous vanilla analysis for the remainder of the proof. □

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

We recall from Definition 2.2 that f that is smooth over X if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{3.5}$$

To minimize f over X, we use projected gradient descent again. The runtime turns out to be the same as in the unconstrained case. Again, we have sufficient decrease. This is not obvious from the following lemma, but you are asked to prove it in Exercise 19.

Lemma 3.3. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable and smooth with parameter L over a closed and convex set $X \subseteq \mathbf{dom}(f)$, according to (3.5). Choosing stepsize*

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

projected gradient descent (3.1) with arbitrary $\mathbf{x}_0 \in X$ satisfies

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

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 proceed similar to the proof of the "unconstrained" sufficient decrease Lemma 2.6, except that we now need to deal with projected gradient descent. We again start from smoothness but then use $\mathbf{y}_{t+1} = \mathbf{x}_t - \nabla f(\mathbf{x}_t)/L$, followed by the usual equation $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$:

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

□

Theorem 3.4. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbf{dom}(f)$ be a closed convex set, and assume that there is a minimizer $\mathbf{x}^\star$ of f over X; furthermore, suppose that f is smooth over X with parameter L according to (3.5). Choosing stepsize*

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

projected gradient descent (3.1) with $\mathbf{x}_0 \in X$ satisfies

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

Proof. The plan is as in the proof of Theorem 2.7 to use the inequality

$$\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \tag{3.6}$$

resulting from sufficient decrease (Lemma 3.3) to bound the squared gradient $\|\mathbf{g}_t\|^2 = \|\nabla f(\mathbf{x}_t)\|^2$ in the vanilla analysis. Unfortunately, (3.6) has an extra term compared to what we got in the unconstrained case. But we can compensate for this in the vanilla analysis itself. Let us go back to its "constrained" version (3.3), featuring $\mathbf{y}_{t+1}$ instead of $\mathbf{x}_{t+1}$:

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2\right).$$

Previously, we applied $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$ (Fact 3.1(ii)) to get back on the unconstrained vanilla track. But in doing so, we dropped a term that now becomes useful. Indeed, Fact 3.1(ii) actually yields $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$, so that we get the following upper bound for $\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$:

$$\frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right). \tag{3.7}$$

Using $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ from convexity, we have (with $\gamma = 1/L$) that

$$\begin{aligned}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) &\leq \sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \qquad (3.8)\\ &\leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.\end{aligned}$$

To bound the sum of the squared gradients, we use (3.6):

$$\begin{aligned}\frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 &\leq \sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right)\\ &= f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.\end{aligned}$$

Plugging this into (3.8), the extra terms cancel, and we arrive—as in the unconstrained case—at

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

The statement follows as in the proof of Theorem 2.7 from the fact that due to sufficient decrease (Exercise 19), the last iterate is the best one. □

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

Assuming that f is smooth *and* strongly convex over a set X, we can also prove fast convergence of projected gradient descent. This does not require any new ideas, we have seen all the ingredients before.

We recall from Definition 2.9 that f is strongly convex with parameter $\mu > 0$ over X if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{3.9}$$

Theorem 3.5. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbf{dom}(f)$ be a nonempty closed and convex set and suppose that f is smooth over X with parameter L according to (3.5) and strongly convex over X with parameter $\mu > 0$ according to (3.9). Exercise 20 asks you to prove that there is a unique minimizer $\mathbf{x}^\star$ of f over X. Choosing*

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

projected gradient descent (3.1) 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:

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

We note that this is *almost* the same result as in Theorem 2.12 for the unconstrained case; in fact, the result in part (i) is identical, but in part (ii), we get an additional term. This is due to the fact that in the constrained case, we cannot argue that $\nabla f(\mathbf{x}^\star) = \mathbf{0}$. In fact, this additional term is the dominating one, once the error becomes small. It has the effect that the required number of steps to reach error at most ε will roughly double, in comparison to the bound of Theorem 2.12.

Proof. In the strongly convex case, the "constrained" vanilla bound (3.7)

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

on $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ can be strengthened to

$$\frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \tag{3.10}$$

Now we proceed as in the proof of Theorem 2.12 and rewrite the latter bound into a bound on $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$ that is

$$2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 + (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2,$$

so we have geometric decrease in squared distance to $\mathbf{x}^\star$, up to some noise. Again, we show that by sufficient decrease, the noise in this bound disappears. From Lemma 3.3, we know that

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

and using this, the noise can be bounded. Multiplying the previous inequality by $2/L$, and rearranging the terms we get:

$$\frac{2}{L}\left(f(\mathbf{x}^\star) - f(\mathbf{x}_t)\right) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \leq 0.$$

With $\gamma = 1/L$, this exactly shows that the noise is nonpositive. This yields (i). The bound in (ii) follows from smoothness (2.8):

$$\begin{aligned}
f(\mathbf{x}_T) - f(\mathbf{x}^\star) &\leq \nabla f(\mathbf{x}^\star)^\top(\mathbf{x}_T - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_T\|^2 \\
&\leq \|\nabla f(\mathbf{x}^\star)\|\,\|\mathbf{x}_T - \mathbf{x}^\star\| + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_T\|^2 \text{ (Cauchy-Schwarz)} \\
&\leq \|\nabla f(\mathbf{x}^\star)\|\left(1 - \frac{\mu}{L}\right)^{T/2}\|\mathbf{x}_0 - \mathbf{x}^\star\| + \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.
\end{aligned}$$

□

3.5 Projecting onto ℓ_1-balls

Problems that are ℓ_1-regularized appear among the most commonly used models in machine learning and signal processing, and we have already discussed the Lasso as an important example of that class. We will now address how to perform projected gradient as an efficient optimization for ℓ_1-constrained problems. Let

$$X = B_1(R) := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i| \leq R \right\}$$

be the ℓ_1-ball of radius $R > 0$ around $\mathbf{0}$, i.e., the set of all points with 1-norm at most R. Our goal is to compute $\Pi_X(\mathbf{v})$ for a given vector $\mathbf{v}$, i.e. the projection of $\mathbf{v}$ onto X; see Figure 3.2.

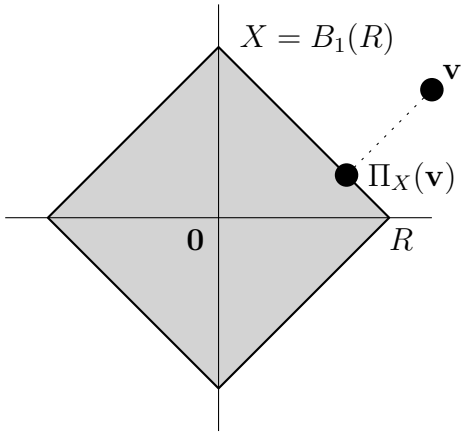


Figure 3.2: Projecting onto an ℓ_1-ball

At first sight, this may look like a rather complicated task. Geometrically, X is a *cross polytope* (square for $d = 2$, octahedron for $d = 3$), and as such it has 2^d many facets. But we can start with some basic simplifying observations.

Fact 3.6. *We may assume without loss of generality that (i) $R = 1$, (ii) $v_i \geq 0$ for all i, and (iii) $\sum_{i=1}^{d} v_i > 1$.*

Proof. If we project $\mathbf{v}/R$ onto $B_1(1)$, we obtain $\Pi_X(\mathbf{v})/R$ (just scale Figure 3.2), so we can restrict to the case $R = 1$. For (ii), we observe that simultaneously flipping the signs of a fixed subset of coordinates in both $\mathbf{v}$ and $\mathbf{x} \in X$ yields vectors $\mathbf{v}'$ and $\mathbf{x}' \in X$ such that $\|\mathbf{x} - \mathbf{v}\| = \|\mathbf{x}' - \mathbf{v}'\|$; thus, $\mathbf{x}$ minimizes the distance to $\mathbf{v}$ if and only if $\mathbf{x}'$ minimizes the distance to $\mathbf{v}'$. Hence, it suffices to compute $\Pi_X(\mathbf{v})$ for vectors with nonnegative entries. If $\sum_{i=1}^d v_i \leq 1$, we have $\Pi_X(\mathbf{v}) = \mathbf{v}$ and are done, so the interesting case is (iii). □

Fact 3.7. *Under the assumptions of Fact 3.6, $\mathbf{x} = \Pi_X(\mathbf{v})$ satisfies $x_i \geq 0$ for all i and $\sum_{i=1}^d x_i = 1$.*

Proof. If $x_i < 0$ for some i, then $(-x_i - v_i)^2 \leq (x_i - v_i)^2$ (since $v_i \geq 0$), so flipping the i-th sign in $\mathbf{x}$ would yield another vector in X at least as close to $\mathbf{v}$ as $\mathbf{x}$, but such a vector cannot exist by strict convexity of the squared distance. And if $\sum_{i=1}^d x_i < 1$, then $\mathbf{x}' = \mathbf{x} + \lambda(\mathbf{v} - \mathbf{x}) \in X$ for some small positive λ, with $\|\mathbf{x}' - \mathbf{v}\| = (1 - \lambda)\|\mathbf{x} - \mathbf{v}\|$, again contradicting the optimality of $\mathbf{x}$. □

Corollary 3.8. *Under the assumptions of Fact 3.6,*

$$\Pi_X(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{x} \in \Delta_d} \|\mathbf{x} - \mathbf{v}\|^2,$$

where

$$\Delta_d := \left\{ \mathbf{x} \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0 \ \forall i \right\}$$

is the standard simplex.

This means, we have reduced the projection onto an ℓ_1-ball to the projection onto the standard simplex; see Figure 3.3.

To address the latter task, we make another assumption that can be established by suitably permuting the entries of $\mathbf{v}$ (which just permutes the entries of its projection onto Δ_d in the same way).

Fact 3.9. *We may assume without loss of generality that $v_1 \geq v_2 \geq \cdots \geq v_d$.*

Lemma 3.10. *Let $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x} \in \Delta_d} \|\mathbf{x} - \mathbf{v}\|^2$. Under the assumption of Fact 3.9, there exists (a unique) $p \in \{1, \ldots, d\}$ such that*

$$\begin{aligned} x_i^\star &> 0, \quad i \leq p, \\ x_i^\star &= 0, \quad i > p. \end{aligned}$$

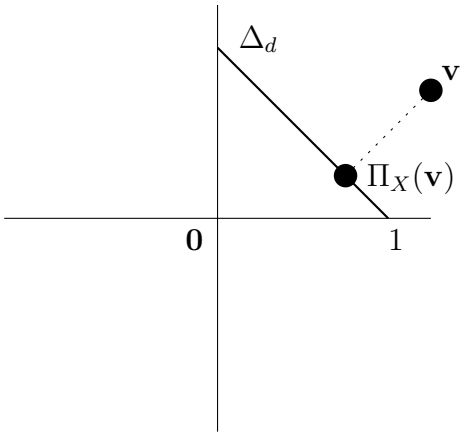


Figure 3.3: Projecting onto the standard simplex

Proof. We are using the optimality criterion of Lemma 1.22:

$$\nabla d_{\mathbf{v}}(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) = 2(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star) \geq 0, \quad \mathbf{x}\in\Delta_d, \tag{3.11}$$

where $d_{\mathbf{v}}(\mathbf{z}) := \|\mathbf{z}-\mathbf{v}\|^2$ is the squared distance to $\mathbf{v}$.

Because $\sum_{i=1}^d x_i^\star = 1$, there is at least one positive entry in $\mathbf{x}^\star$. It remains to show that we cannot have $x_i^\star = 0$ and $x_{i+1}^\star > 0$. Indeed, in this situation, we could decrease $x_{i+1}^\star$ by some small positive ε and simultaneously increase $x_i^\star$ to ε to obtain a vector $\mathbf{x}\in\Delta_d$ such that

$$(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star) = (0-v_i)\varepsilon - (x_{i+1}^\star - v_{i+1})\varepsilon = \varepsilon(\underbrace{v_{i+1}-v_i}_{\leq 0} - \underbrace{x_{i+1}^\star}_{>0}) < 0,$$

contradicting the optimality (3.11). □

But we can say even more about $\mathbf{x}^\star$.

Lemma 3.11. *Under the assumption of Fact 3.9, and with p as in Lemma 3.10,*

$$x_i^\star = v_i - \Theta_p, \quad i\leq p,$$

where

$$\Theta_p = \frac{1}{p}\Big(\sum_{i=1}^p v_i - 1\Big).$$

Proof. Suppose $x_i^\star - v_i < x_j^\star - v_j$ for some $i, j \le p$. As before, we could then decrease $x_j^\star > 0$ by some small positive ε and simultaneously increase $x_i^\star$ by ε to obtain $\mathbf{x} \in \Delta_d$ such that

$$(\mathbf{x}^\star - \mathbf{v})^\top (\mathbf{x} - \mathbf{x}^\star) = (x_i^\star - v_i)\varepsilon - (x_j^\star - v_j)\varepsilon = \varepsilon(\underbrace{(x_i^\star - v_i) - (x_j^\star - v_j)}_{<0}) < 0,$$

again contradicting (3.11). The expression for Θ_p is then obtained from

$$1 = \sum_{i=1}^{p} x_i^\star = \sum_{i=1}^{p} (v_i - \Theta_p) = \sum_{i=1}^{p} v_i - p\Theta_p.$$

□

Let us summarize the situation: we now have d candidates for $\mathbf{x}^\star$, namely the vectors

$$\mathbf{x}^\star(p) := (v_1 - \Theta_p, \dots, v_p - \Theta_p, 0, \dots, 0), \quad p \in \{1, \dots, d\}, \tag{3.12}$$

and we just need to find the right one. In order for candidate $\mathbf{x}^\star(p)$ to comply with Lemma 3.10, we must have

$$v_p - \Theta_p > 0, \tag{3.13}$$

and this actually ensures $\mathbf{x}^\star(p)_i > 0$ for all $i \le p$ by the assumption of Fact 3.9 and therefore $\mathbf{x}^\star(p) \in \Delta_d$. But there could still be several values of p satisfying (3.13). Among them, we simply pick the one for which $\mathbf{x}^\star(p)$ minimizes the distance to $\mathbf{v}$. It is not hard to see that this can be done in time $\mathcal{O}(d \log d)$, by first sorting v and then carefully updating the values Θ_p and $\|\mathbf{x}^\star(p) - \mathbf{v}\|^2$ as we vary p to check all candidates.

But actually, there is an even simpler criterion that saves us from comparing distances.

Lemma 3.12. *Under the assumption of Fact 3.9, with $\mathbf{x}^\star(p)$ as in (3.12), and with*

$$p^\star := \max\Big\{p \in \{1, \dots, d\} : v_p - \frac{1}{p}\Big(\sum_{i=1}^{p} v_i - 1\Big) > 0\Big\},$$

it holds that

$$\operatorname*{argmin}_{\mathbf{x} \in \Delta_d} \|\mathbf{x} - \mathbf{v}\|^2 = \mathbf{x}^\star(p^\star).$$

The proof is Exercise 21. Together with our previous reductions, we obtain the following result.

Theorem 3.13. *Let $\mathbf{v} \in \mathbb{R}^d$, $R \in \mathbb{R}_+$, $X = B_1(R)$ the ℓ_1-ball around $\mathbf{0}$ of radius R. The projection*

$$\Pi_X(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{v}\|^2$$

of $\mathbf{v}$ onto $B_1(R)$ can be computed in time $\mathcal{O}(d \log d)$.

This can be improved to time $\mathcal{O}(d)$, based on the observation that a given p can be compared to the value $p^\star$ in Lemma 3.12 in linear time, without the need to presort $\mathbf{v}$ [DSSSC08].

3.6 Proximal gradient descent

Many optimization problems in applications come with additional structure. An important class of objective functions is composed as

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}) \tag{3.14}$$

where g is a "nice" function, where as h is a "simple" additional term, which however doesn't satisfy the assumptions of niceness which we used in the convergence analysis so far. In particular, an important case is when h is not differentiable.

The classical gradient step for unconstrained minimization of a function g can be equivalently written as

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^d} \; g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{x}_t\|^2 \tag{3.15}$$

$$= \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^d} \; \frac{1}{2\gamma} \|\mathbf{y} - (\mathbf{x}_t - \gamma \nabla g(\mathbf{x}_t))\|^2 \,. \tag{3.16}$$

To obtain the last equality, we have just completed the quadratic $\|\mathbf{v}\|^2 + 2\mathbf{v}^\top \mathbf{w} + \|\mathbf{w}\|^2 = \|\mathbf{v} + \mathbf{w}\|^2$ for $\mathbf{v} := \gamma \nabla g(\mathbf{x}_t)$ and $\mathbf{w} := \mathbf{y} - \mathbf{x}_t$. Here it is crucial that $\mathbf{v}$ is independent of the optimization variable $\mathbf{y}$, so therefore the term can be ignored when taking the argmin. The scaling by $\frac{1}{2\gamma}$ is also irrelevant but we keep it for better illustrating the next step.

The interpretation of the above equivalent reformulation of the classic gradient step is important for us, and is what has enabled the previous convergence analysis in Section 2.5 for smooth unconstrained optimization: For the particular choice of stepsize $\gamma := \frac{1}{L}$ which we have used, the above formulation shows that the gradient descent step exactly minimizes the local quadratic model of g at our current iterate $\mathbf{x}_t$, formed by the smoothness property with parameter L as defined in (2.8).

Our goal in this section is to minimize $f = g + h$, instead of only the smooth part g alone. The idea of the proximal gradient method is to modify the simple quadratic model (3.15) above, so as to make it a valid model for f, that is a model which upper bounds f at all points. The simplest way to do this is to just treat the h function separately by adding it unmodified. We obtain the update equation for *proximal gradient descent*

$$\mathbf{x}_{t+1} := \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d} \; g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}_t\|^2 + h(\mathbf{y}) \tag{3.17}$$

$$= \operatorname*{argmin}_{\mathbf{y}} \; \frac{1}{2\gamma}\|\mathbf{y} - (\mathbf{x}_t - \gamma\nabla g(\mathbf{x}_t))\|^2 + h(\mathbf{y}) \; . \tag{3.18}$$

The last formulation makes clear that the resulting update tries to combine the two goals, staying close to the classic gradient update, as well as also to minimize h.

3.6.1 The proximal gradient algorithm

We define the *proximal mapping* for a given function h, and parameter $\gamma > 0$:

$$\operatorname{prox}_{h,\gamma}(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{z}\|^2 + h(\mathbf{y}) \right\}$$

An iteration of *proximal gradient descent* is defined as

$$\mathbf{x}_{t+1} := \operatorname{prox}_{h,\gamma}(\mathbf{x}_t - \gamma\nabla g(\mathbf{x}_t)) \; . \tag{3.19}$$

This same update step can also be written in different form as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma G_\gamma(\mathbf{x}_t) \tag{3.20}$$

for $G_{h,\gamma}(\mathbf{x}) := \frac{1}{\gamma}\Big(\mathbf{x} - \operatorname{prox}_{h,\gamma}(\mathbf{x} - \gamma\nabla g(\mathbf{x}))\Big)$ being the so called generalized gradient of f.

A generalization of gradient descent. The proximal gradient descent method (3.19) is also known as generalized gradient descent. In the special case $h \equiv 0$, we of course recover classic gradient descent.

More interestingly, it is also a generalization of projected gradient descent as we have discussed in the previous sections. Given a closed convex set X, the *indicator function* of the set X is given as the convex function

$$\begin{aligned} \boldsymbol{\iota}_X : \mathbb{R}^d &\to \mathbb{R} \cup +\infty \\ \mathbf{x} &\mapsto \boldsymbol{\iota}_X(\mathbf{x}) := \begin{cases} 0 & \text{if } \mathbf{x} \in X, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{3.21}$$

When using the indicator function of our constraint set X as $h \equiv \boldsymbol{\iota}_X$, it is easy to see that the proximal mapping simply becomes

$$\begin{aligned} \operatorname{prox}_{h,\gamma}(\mathbf{z}) &:= \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{z}\|^2 + \boldsymbol{\iota}_X(\mathbf{y}) \right\} \\ &= \operatorname*{argmin}_{\mathbf{y} \in X} \ \|\mathbf{y} - \mathbf{z}\|^2 = \Pi_X(\mathbf{z}) \,, \end{aligned}$$

which is the projection of $\mathbf{z}$ onto X.

As we will see, the convergence of proximal gradient will be as fast as classic gradient descent. However, this still comes not entirely for free. In every iteration, we now have to additionally compute the proximal mapping. This can be very expensive if h is complex. Nevertheless, for some important examples of h the proximal mapping is efficient to compute, such as for the ℓ_1-norm.

3.6.2 Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Interestingly, the vanilla convergence analysis for smooth functions as in Theorem 2.7 directly applies for the more general case of proximal gradient descent. Intuitively, this means that proximal method only "sees" the nice smooth part g of the objective, and is not impacted by the additional h which it treats separately in each step.

Theorem 3.14. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L, and also h convex and $prox_{h,\gamma}(\mathbf{x}) := \operatorname{argmin}_{\mathbf{z}}\{\|\mathbf{x} - \mathbf{z}\|^2/(2\gamma) + h(\mathbf{z})\}$ can be computed. Choosing the fixed stepsize*

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

proximal gradient descent (3.19) with arbitrary $\mathbf{x}_0$ satisfies

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

Proof. The proof follows the vanilla analysis for the smooth case, applying it only to g, while always keeping h separate, as in (3.17). We leave the details as Exercise 22 for the reader. □

3.7 Exercises

Exercise 19. *Prove that in Theorem 3.4 (i),*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t).$$

Exercise 20. *Prove that under the assumptions of Theorem 3.5, f has a unique minimizer $\mathbf{x}^\star$ over any nonempty closed and convex set $X \subseteq \mathbb{R}^d$! In particular, for $X = \mathbb{R}^d$, we obtain the existence of a unique global minimum.*

Exercise 21. *Prove Lemma 3.12!*

Hint: *It is useful to prove that with $\mathbf{x}^\star(p)$ as in (3.12) and satisfying (3.13),*

$$\mathbf{x}^\star(p) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^{d} x_i = 1, x_{p+1} = \cdots = x_d = 0\}.$$

Exercise 22. *Prove Theorem 3.14!*

Chapter 4

Subgradient Descent

Contents

4.1 Subgradients

Definition 4.1. *Let* $f : \mathbf{dom}(f) \to \mathbb{R}$. *Then* $\mathbf{g} \in \mathbb{R}^d$ *is a* subgradient *of* f *at* $\mathbf{x} \in \mathbf{dom}(f)$ *if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \quad \forall \mathbf{y} \in \mathbf{dom}(f). \tag{4.1}$$

The set of subgradients of f *at* $\mathbf{x}$ *is called the* subdifferential *at* $\mathbf{x}$ *and is denoted by* $\partial f(\mathbf{x})$.

The notion of a subgradient can be seen as a generalization of the gradient, for functions which are not necessarily differentiable. A prominent example is the ℓ_1-norm, which we have discussed in Exercise 7. Figure 4.1 shows that this function has several subgradients at $x = 0$ (one-dimensional case).

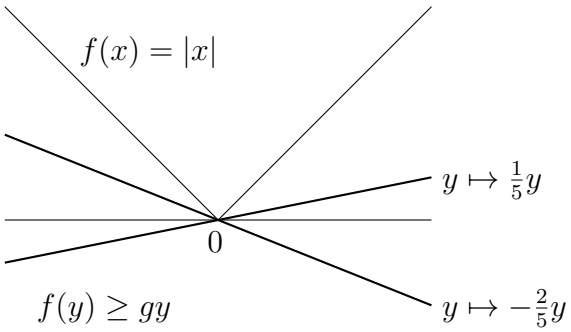


Figure 4.1: The function $f(x) = |x|$ has subgradients $g \in [-1, 1]$ at 0, since $f(y) \geq gy$ for exactly $g \in [-1, 1]$.

Lemma 4.2 (Exercise 23)**.** *If* $f : \mathbf{dom}(f) \to \mathbb{R}$ *is differentiable at* $\mathbf{x} \in \mathbf{dom}(f)$, *then* $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.

This means that in the differentiable case, there is either exactly one subgradient $\nabla f(\mathbf{x})$, or no subgradient at all (if f is *not* above its tangent hyperplane at $\mathbf{x}$; see Figure 1.4).

Definition 4.1 above looks suspiciously similar to the first-order characterization of convexity (1.4) that we discussed earlier. Indeed, the only difference is that here we have replaced $\nabla f(\mathbf{x})$ by $\mathbf{g}$. It turns out that convexity is equivalent to the existence of subgradients everywhere. So we

get a "first order characterization" of convexity that also covers the non-differentiable case.

Lemma 4.3 (Exercise 24)**.** *A function $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex if and only if $\mathbf{dom}(f)$ is convex and $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in \mathbf{dom}(f)$.*

It turns out that Lipschitz continuity can be characterized by bounded subgradients. For real-valued convex functions, this is a generalization of Lemma 1.10 to the non-differentiable case.

Lemma 4.4 (Exercise 25)**.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mathbf{dom}(f)$ open, $B \in \mathbb{R}_+$. Then the following two statements are equivalent.*

(i) $\|\mathbf{g}\| \leq B$ *for all* $\mathbf{x} \in \mathbf{dom}(f)$ *and all* $\mathbf{g} \in \partial f(\mathbf{x})$.

(ii) $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$.

Subgradient optimality condition. Subgradients also allow us to describe cases of optimality for functions which are not necessarily differentiable (and not necessarily convex), in the spirit of Lemma 1.16:

Lemma 4.5. *Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ and $\mathbf{x} \in \mathbf{dom}(f)$. If $\mathbf{0} \in \partial f(\mathbf{x})$, then $\mathbf{x}$ is a global minimum.*

Proof. By (4.1), $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$ gives

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

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

Here we see (again) that subgradients are "stronger" than gradients for differentiable functions. Indeed, if $\nabla f(\mathbf{x}) = \mathbf{0}$ for a differentiable function f and $\mathbf{x} \in \mathbf{dom}(f)$, we can only say that $\mathbf{x}$ is a *critical point*, but not necessarily a global minimum. Unlike the gradient, a subgradient yields by definition a linear lower bound to the function.

4.2 Differentiability of convex functions

Before we move on to subgradient descent, we want to get a feeling for how "wild" non-differentiable convex functions can be. The answer is: they are surprisingly tame. While there are continuous functions that are *nowhere* differentiable (the classical example is the *Weierstrass function*), convex function cannot be as pathological. In fact, a convex function f is differentiable *almost everywhere*. Formally, this means that wherever you are in $\mathbf{dom}(f)$, you find points arbitrarily close to you at which f is differentiable. In still other words, the set of points where f is not differentiable has measure 0 [Roc97, Theorem 25.5].

This does not mean that we can ignore non-differentiability in optimization. For example, as Figure 4.1 demonstrates, the global minimum $\mathbf{x}^\star$ can easily be a "kink", a point where f is not differentiable. Also, while running an iterative optimization scheme, we may always stumble upon an intermediate kink.

An important fact is the following characterization of subdifferentials;

Theorem 4.6 ([Roc97, Theorem 25.6]). *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mathbf{dom}(f)$ open, $\mathbf{x} \in \mathbf{dom}(f)$. Then $\partial f(\mathbf{x})$ is the convex hull of the set*

$$S(\mathbf{x}) = \{ \lim_{n\to\infty} \nabla f(\mathbf{x}_n) \mid \lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}\}.$$

In words, we consider sequences $(\mathbf{x}_n)_{n\in\mathbb{N}}$ that converge to $\mathbf{x}$ and for which the sequence of gradients $(\nabla f(\mathbf{x}_n))_{n\in\mathbb{N}}$ exists and also converges; the theorem says that the limit is a subgradient at $\mathbf{x}$, and that *any* subgradient can be obtained as a convex combination of such limit subgradients.

In the example of Figure 4.1, there are two types of sequences converging to 0 such the gradients converge as well. These are sequences that have almost all elements negative (gradients converge to -1), and sequences that have almost all elements positive (gradients converge to 1). Consequently, the subgradients at 0 are formed by the set $[-1, 1]$, the convex hull of -1 and 1.

4.3 The algorithm

An iteration of *subgradient descent* is defined as

$$\begin{aligned} &\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t) \\ &\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t. \end{aligned} \tag{4.2}$$

In contrast to our previous descent algorithms, we allow a time-varying stepsize here. This can of course be done for any descent algorithm but so far, we just did not need it. Later in this chapter, we will make use of a time-varying step size.

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

The following result gives the convergence for Subgradient Descent. It is identical to Theorem 2.1, up to relaxing the requirement of differentiability.

Theorem 4.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and B-Lipschitz continuous with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$. Choosing the constant stepsize*

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

subgradient descent (4.2) yields

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

Proof. The proof is identical to the one of Theorem 2.1 presented in Section 2.4. The only change is that $\mathbf{g}_t$ is a subgradient now and not a gradient, so that the inequality (2.5) now follows from the subgradient property (4.1) instead of the first-order characterization of convexity. The required bound $\|\mathbf{g}_t\|^2 \leq B^2$ follows from Lemma 4.4 ("convex and Lipschitz = bounded subgradients"). □

Projected subgradient descent. Theorem 3.2 for constrained optimization in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to the case of subgradient descent as well.

4.5 Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

(Projected) gradient descent converges in $\mathcal{O}(\log(1/\varepsilon))$ steps for functions that are both smooth and strongly convex. But if a function is non-differentiable, then it cannot be smooth under the natural definition of smoothness (Exercise 26). It can still be strongly convex, however, so it is natural to ask whether strong convexity alone allows us to obtain a convergence result. The answer is no in general, but before we discuss this, let us define strong convexity for not necessarily differentiable functions. This is straightforward; for differentiable functions, we recover Definition 2.9. Here, we restrict to the unconstrained case for simplicity.

Definition 4.8. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called* strongly convex *(with parameter μ) if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f),\ \forall \mathbf{g} \in \partial f(\mathbf{x}). \tag{4.3}$$

Actually, requiring (4.3) only for *some* $\mathbf{g} \in \partial f(\mathbf{x})$ would be another straightforward generalization of Definition 2.9, so which one is the "right" one? The answer is that it does not matter if $\mathbf{dom}(f)$ is open. We could even afford to not require *anything* for points $\mathbf{x}$ where f is not differentiable. This is a consequence of Theorem 4.6 (Exercise 27).

Strong convexity has the following useful characterization.

Lemma 4.9 (Exercise 28)**.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mathbf{dom}(f)$ open, $\mu \in \mathbb{R}_+, \mu > 0$. f is strongly convex with parameter μ if and only if $f_\mu : \mathbf{dom}(f) \to \mathbb{R}$ defined by*

$$f_\mu(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2, \quad \mathbf{x} \in \mathbf{dom}(f)$$

is convex.

Let's look at the problem with (sub)gradient descent on strongly convex functions.

Lemma 4.10 (Exercise 29)**.** *The function $f(x) = e^{|x|}$ is strongly convex with parameter $\mu = 1$.*

This function is of course far from being smooth; it grows exponentially, so there can't be any quadratic upper bounds. In fact, as strong

convexity ony requires quadratic *lower* bounds, strongly convex functions can be extremely fast-growing. In such a situation, (sub)gradient descent will overshoot already for tiny step sizes and diverge.

In case of $f(x) = e^{|x|}$, the function is differentiable at $x \neq 0$ with $f'(x) = \operatorname{sgn}(x)e^{|x|}$, so the (sub)gradient step is

$$x_{t+1} = x_t - \gamma_t \operatorname{sgn}(x_t) e^{|x_t|}.$$

For $|x|$ only mildly larger than 0, the step will overshoot the optimum $x^* = 0$ and take us (much) further away. To compensate for this, we would need extremely small stepsizes. These in turn would lead to extremely poor convergence for functions such as $f(x) = x^2/2$ (which is also strongly convex with $\mu = 1$) . Hence, there are no stepsizes that fit all strongly convex functions with a fixed strong convexity parameter μ.

To succeed with (sub)gradient descent in this situation, we therefore need to make some additional assumptions. Smoothness (quadratic upper bounds) is such an assumption, but in the non-differentiable case, this is precisely not an option. What people have done instead is to assume that the subgradients g_t that we encounter during the algorithm are bounded in norm.

To ensure bounded subgradients, we could simply assume that f is Lipschitz, but then we will only make a statement about an empty function class. The reason is that a function cannot be globally strongly convex and Lipschitz at the same time (Exercise 30). It can be strongly convex *and* have bounded gradients over a closed and bounded set X, so analyzing projected subgradient descent is an alternative.

But even when we optimize over $\mathbb{R}^d$, we may be lucky and only hit iterates with small subgradients. This will typically happen if we start sufficiently close to optimality. In this case, there are step sizes γ_t (not depending on the observed gradients) that give us useful error bounds.

Below, we prove such a bound for subgradient descent, and this result then clearly extends to gradient descent on differentiable and strongly convex (but not necessarily smooth) functions. The bound on the number of steps will be $\mathcal{O}(1/\varepsilon)$ which is of course much worse than $\mathcal{O}(\log(1/\varepsilon))$, but still better than $\mathcal{O}(1/\varepsilon^2)$ that we get in the Lipschitz case. So assuming strong convexity results in a convergence behavior as in the smooth case—*if* the gradients stay bounded, and this is what we mean by "tame".

In order to analyze subgradient descent on strongly convex functions,

we will for the first time depart from algorithm variants with a constant stepsize γ, but instead use a time-varying stepsize γ_t decreasing over time.

Theorem 4.11. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex with parameter $\mu > 0$ and let $\mathbf{x}^\star$ be the unique global minimum of f. With decreasing step size*

$$\gamma_t := \frac{2}{\mu(t+1)}, \quad t > 0,$$

subgradient descent (4.2) yields

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2B^2}{\mu(T+1)},$$

where $B = \max_{t=1}^{T} \|\mathbf{g}_t\|$.

Unlike in previous convergence results, small error is not achieved by some iterate that we have gone through, but by a convex combination of iterates.

Proof. We start from the vanilla analysis (2.3) (with $\gamma = \gamma_t$):

$$\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma_t}{2} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma_t} \left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \right).$$

Now we plug in the lower bound $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}^\star\|^2$ resulting from strong convexity to obtain (with $\|\mathbf{g}_t\|^2 \leq B^2$) that

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{B^2 \gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2. \quad (4.4)$$

Unlike in the vanilla analysis (where we had $\gamma_t = \gamma, \mu = 0$), the right-hand side does not telescope anymore when we sum over all $t \leq T$; to fix this, we precisely need the time-varying stepsize.

Let's make a small computation: to get telescoping behavior, we would need that $\gamma_t^{-1} = \gamma_{t+1}^{-1} - \mu$. For example, $\gamma_t^{-1} = \mu(1+t)$ satisfies this, but our choice $\gamma_t^{-1} = \mu(1+t)/2$ does not. Exercise 31 asks you to compute what happens when we actually choose $\gamma_t^{-1} = \mu(1+t)$; this will let you

appreciate the seemingly "wrong" choice of $\gamma_t = \frac{2}{\mu(t+1)}$ here. Plugging in this stepsize and multiplying with t on both the sides, we get

$$\begin{aligned} t \cdot \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) &\leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big) \\ &\leq \frac{B^2}{\mu} + \frac{\mu}{4}\Big(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big). \end{aligned}$$

Summing from $t = 1, \ldots, T$, we obtain a telescoping sum:

$$\sum_{t=1}^{T} t \cdot \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{TB^2}{\mu} + \frac{\mu}{4}\Big(0 - T(T+1)\left\|\mathbf{x}_{T+1} - \mathbf{x}^\star\right\|^2\Big) \leq \frac{TB^2}{\mu}.$$

Since

$$\frac{2}{T(T+1)}\sum_{t=1}^{T} t = 1,$$

Jensen's inequality (Lemma 1.5) yields

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big).$$

This in turn implies

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2B^2}{\mu(T+1)}\,.$$

□

Unlike all previous bounds, this bound seems to be independent from the initial distance $\|\mathbf{x}_o - \mathbf{x}^\star\|$ to the optimum. However, there is no free lunch here. The initial distance will typically affect the bound B (think of a quadratic function where B is proportional to $\|\mathbf{x}_o - \mathbf{x}^\star\|$).

4.6 Optimality of first-order methods

With all the convergence rates we have seen so far, a very natural question to ask is if these rates are best possible or not. Surprisingly, the rate can indeed not be improved in general.

Theorem 4.12 (Nesterov). *For any $T \leq d-1$ and starting point $\mathbf{x}_0$, there is a function f in the problem class of B-Lipschitz functions over $\mathbb{R}^d$, such that any (sub)gradient method has an objective error at least*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \geq \frac{RB}{2(1+\sqrt{T+1})} \,.$$

The above theorem applies to all first-order methods which form iterates by linearly combining past iterates and (sub)gradients, and requires the dimension d to be sufficiently large.

4.7 Exercises

Exercise 23. *Prove Lemma 4.2, meaning that a function that is differentiable at $\mathbf{x}$ has at most one subgradient there, namely $\nabla f(\mathbf{x})$.*

Exercise 24. *Prove the easy direction of Lemma 4.3, meaning that the existence of subgradients everywhere implies convexity!*

Exercise 25. *Prove Lemma 4.4 (Lipschitz continuity and bounded subgradients).*

Exercise 26. *Generalizing Definition 2.2, let us call a (not necessarily differentiable) function $f : \mathbb{R}^d \to \mathbb{R}$ smooth with parameter $L \in \mathbb{R}_+$ if for all $\mathbf{x} \in \mathbb{R}^d$, there exists a subgradient $\mathbf{g}_\mathbf{x} \in \mathbb{R}^d$ such that*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \mathbf{g}_\mathbf{x}^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

This means that for every point $\mathbf{x}$, the graph of f is below the graph of the quadratic function $f(\mathbf{x}) + \mathbf{g}_\mathbf{x}^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$.

Prove that if f is smooth according to this definition, then f is differentiable, with $\mathbf{g}_\mathbf{x} = \nabla f(\mathbf{x})$ for all $\mathbf{x}$. In particular, for differentiable functions, the notion of smoothness introduced above coincides with the one of Definition 2.2; moreover, non-differentiable functions cannot be smooth.

Does the above hold if $\mathbf{g}_\mathbf{x}$ is not a subgradient?

Exercise 27. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is convex and satisfies*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

for all $\mathbf{x}$ *such that* $\nabla f(\mathbf{x})$ *exists, and for all* $\mathbf{y}$*. Prove that this implies*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}_{\mathbf{x}}^{\top}(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

for all $\mathbf{x}$*, all* $\mathbf{g}_{\mathbf{x}} \in \partial f(\mathbf{x})$ *and all* $\mathbf{y}$*.*

Exercise 28. *Prove Lemma 4.9: f is strongly convex with parameter μ over an open domain if and only if $f_\mu : \mathbf{x} \mapsto f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex over the same domain.*

Exercise 29. *Prove Lemma 4.10: $f(x) = e^{|x|}$ is strongly convex with parameter $\mu = 1$.*

Exercise 30. *Prove that a function $f : \mathbb{R}^d \to \mathbb{R}$ cannot simultaneously be Lipschitz and strongly convex!*

Exercise 31. *Which result can you prove when you use the "telescoping stepsize"*

$$\gamma_t = \frac{1}{\mu(t+1)}$$

in Theorem 4.11 instead of $\gamma_t = \frac{2}{\mu(t+1)}$?

Chapter 5

Stochastic Gradient Descent

Contents

5.1 The algorithm

Many objective functions occurring in machine learning are formulated as *sum structured objective functions*

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}). \tag{5.1}$$

Here f_i is typically the cost function of the i-th datapoint, taken from a training set of n elements in total.

We have already seen an example for this: the loss function (1.11) in the handwritten digit recognition (Section 1.7.1) has one term for each of the n training images $\mathbf{x} \in P$:

$$\ell(W) = -\sum_{\mathbf{x} \in P} \ln z_{d(\mathbf{x})}(W\mathbf{x}).$$

The normalizing factor $1/n$ that we assume in the general setting (5.1) will just simplify the following a bit.

An iteration of *stochastic gradient descent* (SGD) in its basic form is defined as

$$\begin{aligned} &\text{sample } i \in [n] \text{ uniformly at random} \\ &\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t). \end{aligned} \tag{5.2}$$

This update looks almost identical to the classical gradient method, the only difference being that we have computed the gradient not of the entire f but only of one particular (randomly chosen) function f_i. As we will need varying stepsizes a bit later, we allow for the stepsize to depend on t now.

In the above setting, the update vector $\mathbf{g}_t := \nabla f_i(\mathbf{x}_t)$ is called a *stochastic gradient*. Formally, $\mathbf{g}_t$ is a vector of d random variables, but we will also simply call this a random variable.

The crucial advantage of SGD versus its classical gradient descent counterpart is the efficiency per iteration: While computing the full gradient for a sum structured problem (5.1) would require us to compute n individual gradients of the f_i functions, an iteration of SGD requires only a single one of those, and therefore is n times cheaper. SGD has therefore become the main workhorse for training machine learning models. Whether such cheaper iterations also give similar progress is another question, which we analyze next.

5.2 Unbiasedness

We would like to start with the vanilla analysis again, but now we cannot bound the random variable $\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ from below using (2.5), as the inequality

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

may hold or not hold, depending on how $\mathbf{g}_t$ turns out. But it still holds *in expectation*, as we show now.

The vector $\mathbf{g}_t$ may be far from the true gradient, and of high variance, but in expectation over the random choice of i, it does coincide with the full gradient of f. We formalize this as

$$\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t = \mathbf{x}\big] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d. \tag{5.3}$$

Here, $\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t = \mathbf{x}\big]$ is the *conditional expectation* of $\mathbf{g}_t$, given the event $\{\mathbf{x}_t = \mathbf{x}\}$. If this event is nonempty, linearity of conditional expectations yields that

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x} - \mathbf{x}^\star) \big| \mathbf{x}_t = \mathbf{x}\big] = \mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t = \mathbf{x}\big]^\top(\mathbf{x} - \mathbf{x}^\star) = \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star).$$

Using the fact that $\{\mathbf{x}_t = \mathbf{x}\}$ can occur only for $\mathbf{x}$ in some finite set X (one element for every choice of indices throughout all iterations), the partition theorem further gives us

$$\begin{aligned}
\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] &= \sum_{\mathbf{x}\in X} \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x} - \mathbf{x}^\star) \big| \mathbf{x}_t = \mathbf{x}\big]\ \mathrm{prob}(\mathbf{x}_t = \mathbf{x}) \\
&= \sum_{\mathbf{x}\in X} \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star)\ \mathrm{prob}(\mathbf{x}_t = \mathbf{x}) \\
&= \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big].
\end{aligned}$$

Hence, we have

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \geq \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big]. \tag{5.4}$$

The last inequality is by convexity, and this is means that the lower bound (2.5) holds in expectation.

Exercise 32 lets you recall some basics around conditional expectations. Under (5.3) we say that the stochastic gradient $\mathbf{g}_t$ is an *unbiased* estimator of the gradient, for any time-step t.

5.3 Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

To get a first result out of the vanilla analysis, we assumed in Section 2.4 that $\|\nabla f(\mathbf{x})\|^2 \leq B^2$ for all $\mathbf{x} \in \mathbb{R}^d$, where B was a constant. Here, we are assuming the same for the *expected* squared norms of our stochastic gradients. And we are getting the same result, except that it now holds for the *expected* function values.

Theorem 5.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, and let $\mathbf{x}^\star$ be a global minimum of f; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, and that $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$ for all t. Choosing the constant stepsize*

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

stochastic gradient descent (5.2) yields

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

Proof. Taking expectations on both sides of the vanilla analyis (2.4) and using linearity of expectations, we get

$$\sum_{t=0}^{T-1} \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\mathbf{g}_t\|^2\big] + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2. \tag{5.5}$$

By (5.4),

$$\mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] \leq \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big].$$

Plugging this into (5.5), using $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$ and $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, we get

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

from which the statement follows from the choice of γ as in Theorem 2.1. □

Constrained optimization. For constrained optimization, Theorem 5.1 for the convergence in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to constrained problems as well. After every step of SGD, projection back to X is applied as usual. The resulting algorithm is called *projected SGD*.

5.4 Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

It is possible to strengthen our above SGD analysis. One way to do so is under the additional assumption of strong convexity of the objective function f (as in Definition 2.9). Again, the proof works by "taking expectations" over a previous analysis, in this case the one for subgradient descent in the tame strongly convex case (Theorem 4.11).

Theorem 5.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$; let $\mathbf{x}^\star$ be the unique global minimum of f. With decreasing step size*

$$\gamma_t := \frac{2}{\mu(t+1)}$$

stochastic gradient descent (5.2) yields

$$\mathbb{E}\Big[f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\Big) - f(\mathbf{x}^\star)\Big] \leq \frac{2B^2}{\mu(T+1)},$$

where $B = \max_{t=1}^{T} \mathbb{E}\big[\,\|\mathbf{g}_t\|\,\big]$.

Proof. We start from the vanilla analysis (2.3) (with $\gamma = \gamma_t$) and take expectations on both sides:

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] = \frac{\gamma_t}{2}\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] + \frac{1}{2\gamma_t}\left(\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}^\star\|^2\big] - \mathbb{E}\big[\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\big]\right).$$

Now we use (5.4) along with strong convexity to get a lower bound

$$\begin{aligned}\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] &= \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \\ &\geq \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] + \frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}^\star\|^2\big]\end{aligned}$$

for the left-hand side. Combining the previous two equations and using $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$, we get the "expected version" of (4.4):

$$\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)] \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\mathbb{E}\big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\big] - \frac{\gamma_t^{-1}}{2}\mathbb{E}\big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\big].$$

The proof continues as in Theorem 4.11, with every step being the "expected version" of the corresponding step in the earlier proof. □

5.5 Stochastic Subgradient Descent

For problems which are not necessarily differentiable, we modify SGD to use a subgradient of f_i in each iteration. The update of stochastic subgradient descent is given by

$$\begin{aligned} &\text{sample } i \in [n] \text{ uniformly at random} \\ &\text{let } \mathbf{g}_t \in \partial f_i(\mathbf{x}_t) \\ &\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t. \end{aligned} \tag{5.6}$$

Let $\mathbf{g}^i : \mathbb{R}^d \to \mathbb{R}^d$ denote the function that selects the subgradient of f_i at the current point. Then we have $\mathbf{g}_t = \mathbf{g}^i(\mathbf{x}_t)$ for random i. Unbiasedness now becomes

$$\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t = \mathbf{x}\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}^i(\mathbf{x}) =: \mathbf{g}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

It is immediate from the subgradient property that $\mathbf{g}(\mathbf{x}) \in \partial f(\mathbf{x})$ if $\mathbf{g}^i(\mathbf{x}) \in \partial f_i(\mathbf{x})$ for all i. As in Section 5.2 for SGD, we then get

$$\mathbb{E}\big[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)\big] = \mathbb{E}\big[\mathbf{g}(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star)\big].$$

This in turn can be lower bounded by

$$\mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] + \frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}^\star\|^2\big],$$

with $\mu = 0$ in the convex case and $\mu > 0$ in the strongly convex case, now using $\mathbf{g}(\mathbf{x}_t)$'s subgradient property (4.1) in the convex and (4.3) in the strongly convex case instead of the first-order condition for $\nabla f(\mathbf{x}_t)$. As this lower bound is the crucial ingredient in the previous two analyses of convergence in $\mathcal{O}(1/\varepsilon^2)$ and $\mathcal{O}(1/\varepsilon)$ steps, the results directly extend to the case of subgradient descent as well.

5.6 Mini-batch variants

Instead of using a single element f_i of our sum objective (5.1) to form a stochastic gradient $\mathbf{g}_t = \nabla f_i(\mathbf{x}_t)$, another variant is to use an average of several of them:

$$\tilde{\mathbf{g}}_t := \frac{1}{m}\sum_{j=1}^{m} \mathbf{g}_t^j. \tag{5.7}$$

where $\mathbf{g}_t^j = \nabla f_{i_j}(\mathbf{x}_t)$ for an index i_j. The set of the (distinct) i_j indices is called a mini-batch, and m is the mini batch size.

Using the step direction $\tilde{\mathbf{g}}_t$ defines mini-batch SGD. For $m = 1$, we recover SGD as originally defined, while for $m = n$ we recover full gradient descent.

Mini-batch SGD can be advantageous in several applications. For example, parallelization over up to m processors will easily give a speed-up for the gradient computation, which is typically the main cost of running SGD. Here, parallelization exploits the fact that all $\mathbf{g}_t^j$ are defined at the same iterate $\mathbf{x}_t$ and can therefore be computed independently.

Taking an average of many independent random variables reduces the variance. In the context of mini-batch SGD, we obtain that for larger size of the mini-batch m our estimate $\tilde{\mathbf{g}}_t$ will be closer to the true gradient, in expectation:

$$\begin{aligned}\mathbb{E}\Big[\Big\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\Big\|^2\Big] =& \mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_t^j - \nabla f(\mathbf{x}_t)\Big\|^2\Big] \\ =& \frac{1}{m}\mathbb{E}\big[\|\mathbf{g}_t^1 - \nabla f(\mathbf{x}_t)\|^2\big] \\ =& \frac{1}{m}\mathbb{E}\big[\|\mathbf{g}_t^1\|^2\big] - \frac{1}{m}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{B^2}{m}\,.\end{aligned}$$

Using a modification of the above analysis, it is possible to use this property to relate the above convergence rate of SGD to the rate of full gradient descent.

5.7 Exercises

Exercise 32. *Let Y be a random variable over a finite probability space (Ω, prob) where $\mathrm{prob} : 2^\Omega \to [0,1]$; this avoids subtleties in defining conditional probabilities and expectations; and it covers the random variables occurring in SGD, since in each step, we are randomly choosing among a finite set of n indices. Furthermore, let $B \subseteq \Omega$ be an event.*

For nonemepty B, the conditional expectation of Y given B *is the number*

$$\mathbb{E}\big[Y\big|B\big] := \sum_{y \in Y(\Omega)} y \cdot \mathrm{prob}\big(Y = y\big|B\big).$$

where $Y = y$ is shorthand for the event $\{\omega \in \Omega : Y(\omega) = y\}$.

Finally, for two events A and $B \neq \emptyset$, the conditional probability $\mathrm{prob}\big[A\big|B\big]$ is defined as

$$\mathrm{prob}\big(A\big|B\big) := \frac{\mathrm{prob}\big(A \cap B\big)}{\mathrm{prob}\big(B\big)}.$$

If $B = \emptyset$, $\mathbb{E}\big[Y\big|B\big]$ can be defined arbitrarily.

Prove the following statements.

(i) Alternative definition of conditional expectation:

$$\mathrm{prob}\big(B\big) \cdot \mathbb{E}\big[Y\big|B\big] = \sum_{\omega \in B} Y(\omega)\mathrm{prob}(\omega).$$

(ii) Partition Theorem: Let $B_1, \ldots, B_m$ be a partition of Ω. Then

$$\mathbb{E}\big[Y\big] = \sum_{i=1}^{m} \mathbb{E}\big[Y\big|B_i\big]\mathrm{prob}\big(B_i\big).$$

(iii) Linearity of conditional expectation: For random variables $Y_1, \ldots, Y_m$ over (Ω, prob) and real numbers $\lambda_1, \ldots, \lambda_m$, and if $B \neq \emptyset$,

$$\sum_{i=1}^{m} \lambda_i \mathbb{E}\big[Y_i\big|B\big] = \mathbb{E}\big[\sum_{i=1}^{m} \lambda_i Y_i\big|B\big].$$

Chapter 6

Nonconvex functions

Contents

So far, all convergence results that we have given for variants of gradient descent have been for convex functions. And there is a good reason for this: on nonconvex functions, gradient descent can in general not be expected to come close (in distance or function value) to the global minimum $\mathbf{x}^\star$, even if there is one.

As an example, consider the nonconvex function from Figure 1.2 (left). Figure 6.1 shows what happens if we start gradient descent somewhere "to the right", with a not too large stepsize so that we do not overshoot. For any sufficiently large T, the iterate $\mathbf{x}_T$ will be close to the local minimum $\mathbf{y}^\star$, but not to the global minimum $\mathbf{x}^\star$.

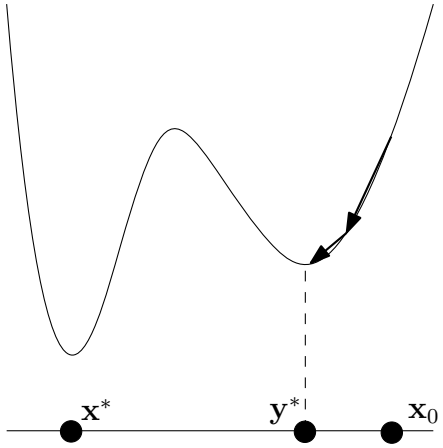


Figure 6.1: Gradient descent may get stuck in a local minimum $\mathbf{y}^\star \neq \mathbf{x}^\star$

Even if the global minimum is the unique local minimum, gradient descent is not guaranteed to get there, as it may also get stuck in a saddle point, or even fail to reach anything at all; see Figure 6.2.

In practice, variants of gradient descent are often observed to perform well even on nonconvex functions, but theoretical explanations for this are mostly missing.

In this chapter, we show that under favorable conditions, we can still say something useful about the behavior of gradient descent, even on nonconvex functions.

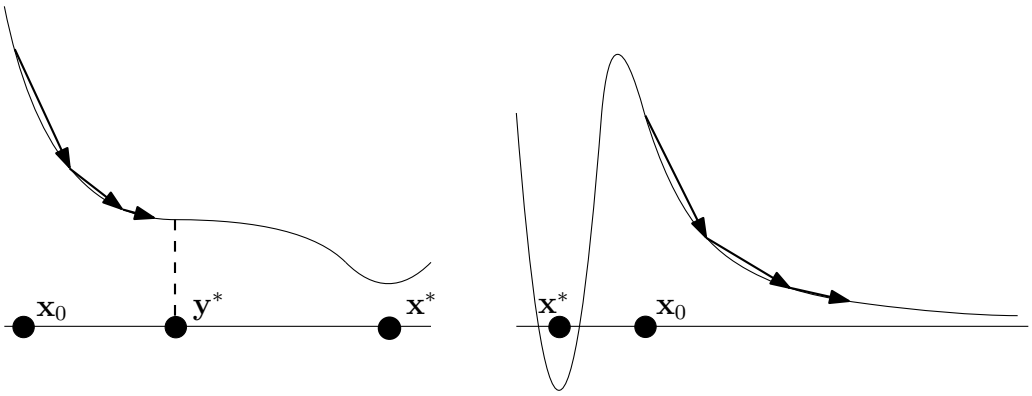


Figure 6.2: Gradient descent may get stuck in a flat region (saddle point) $\mathbf{y}^\star$ (left), or reach neither a local minimum nor a saddle point (right).

6.1 Smooth functions

A particularly low hanging fruit is the analysis of gradient descent on smooth (but not necessarily convex) functions. We recall from Definition 2.2) that a differentiable function $f : \mathbf{dom}(f) \to \mathbb{R}$ is smooth with parameter $L \in \mathbb{R}_+$ over a convex set $X \subseteq \mathbf{dom}(f)$ if

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

This means that at every point $\mathbf{x} \in X$, the graph of f is below a not-too-steep tangential paraboloid, and this may happen even if the function is not convex; see Figure 6.3.

There is a class of arbitrarily smooth nonconvex functions, namely the differentiable *concave* functions. A function f is called concave if $-f$ is convex. Hence, for all $\mathbf{x}$, the graph of a differentiable concave function is *below* the tangent hyperplane at $\mathbf{x}$, hence f is smooth with parameter $L = 0$; see Figure 6.4.

However, from our optimization point of view, concave functions are boring, since they have no global minimum (at least in the unconstrained setting that we are treating here). Gradient descent will then simply "run off to infinity".

We will therefore consider smooth functions that have a global minimum $\mathbf{x}^\star$. Are there even such functions that are not convex? Actually,

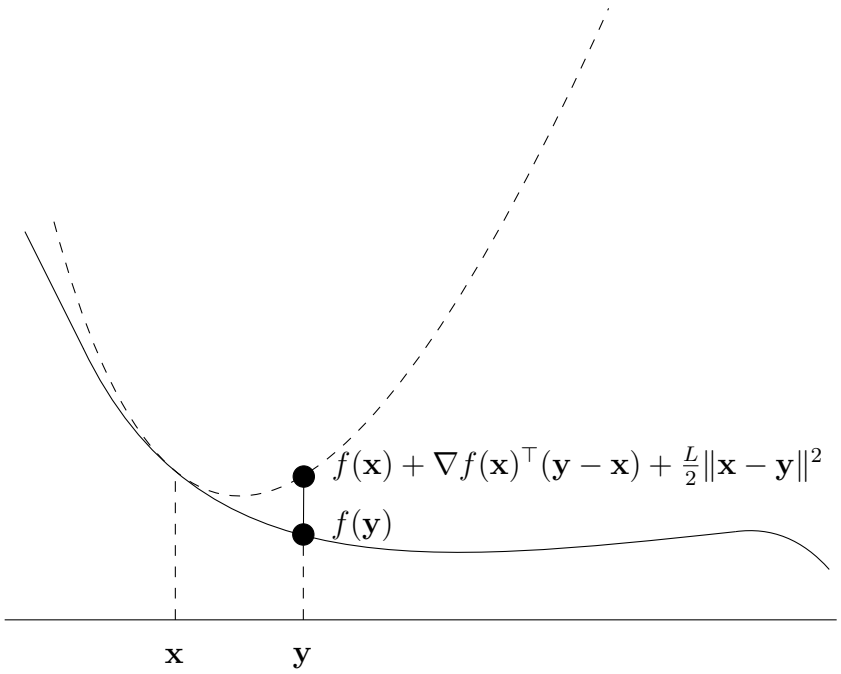


Figure 6.3: A smooth and nonconvex function

many. As we show next, any twice differentiable function with bounded Hessians over some convex set X is smooth over X. A concrete example of a smooth function that is not convex but has a global minimum (actually, many), is $f(x) = \sin(x)$.

Lemma 6.1. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable, with $X \subseteq \mathbf{dom}(f)$ a convex set, and $\|\nabla^2 f(\mathbf{x})\| \leq L$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is again spectral norm. Then f is smooth with parameter L over X.*

Proof. By Theorem 1.10 (applied to the gradient function ∇f), bounded Hessians imply Lipschitz continuity of the gradient,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x}, \mathbf{y} \in X. \tag{6.1}$$

We show that this in turn implies smoothness. This is in fact the easy direction of Lemma 2.4 (in the twice differentiable case), and we proceed as in the proof of Theorem 1.10 by employing the fundamental theorem of

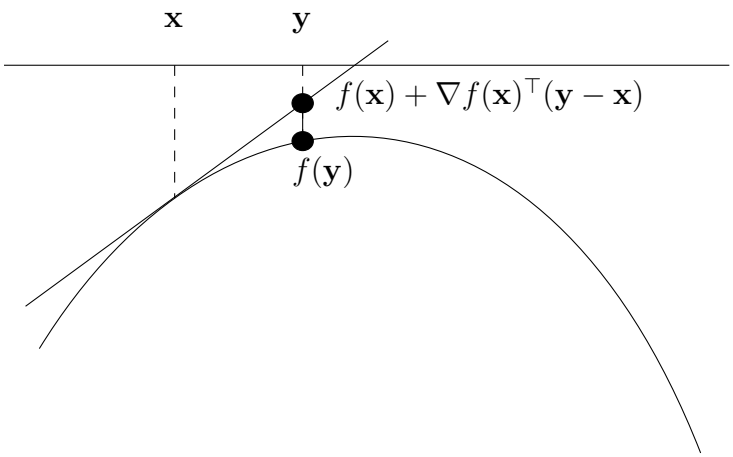


Figure 6.4: A concave function and the first-order characterization of concavity: $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

calculus with $h : \mathbf{dom}(h) \to \mathbb{R}^d$, $[0, 1] \subseteq \mathbf{dom}(h)$ given by

$$h(t) = f\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big), \quad t \in \mathbb{R},$$

in which case the chain rule (see (1.2) yields

$$h'(t) = \nabla f\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big)^\top(\mathbf{y} - \mathbf{x}).$$

For any $\mathbf{x}, \mathbf{y} \in X$, we now compute

$$
\begin{aligned}
& f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
= \; & h(1) - h(0) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
= \; & \int_0^1 h'(t) dt - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
= \; & \int_0^1 \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) dt - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
= \; & \int_0^1 \big(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \big) dt \\
= \; & \int_0^1 \big(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}) \big)^\top (\mathbf{y} - \mathbf{x}) dt \\
\leq \; & \int_0^1 \big| \big(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}) \big)^\top (\mathbf{y} - \mathbf{x}) \big| dt \\
\leq \; & \int_0^1 \big\| \big(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}) \big) \big\| \, \|(\mathbf{y} - \mathbf{x})\| \, dt \quad \text{(Cauchy-Schwarz)} \\
\leq \; & \int_0^1 L \, \|t(\mathbf{y} - \mathbf{x})\| \, \|(\mathbf{y} - \mathbf{x})\| \, dt \quad \text{(Lipschitz continuous gradients)} \\
= \; & \int_0^1 Lt \, \|\mathbf{x} - \mathbf{y}\|^2 \\
= \; & \frac{L}{2} \, \|\mathbf{x} - \mathbf{y}\|^2 .
\end{aligned}
$$

This is smoothness over X according to Definition 2.2. □

For twice differentiable functions, the converse is also (almost) true. If f is smooth over an *open* convex subset $X \subseteq \mathbf{dom}(f)$, the maximum eigenvalue of the Hessian is bounded over X (Exercise 33). We can only bound the eigenvalues from above since e.g. concave functions are smooth with parameter $L = 0$ but generally have unbounded Hessians. It is also not hard to understand why openness is necessary in general. Indeed, for a point $\mathbf{x}$ on the boundary of X, the smoothness condition does not give us any information about nearby points not in X. As a consequence, even at points with large Hessians, f might look smooth inside X. As a simple example, consider $f(x_1, x_2) = x_1^2 + Mx_2^2$ with $M \in \mathbb{R}_+$ large. The function f is smooth with $L = 2$ over $X = \{(x_1, x_2) : x_2 = 0\}$: indeed, over this set, f looks just like the supermodel. But for all $\mathbf{x}$, we have $\|\nabla^2 f(\mathbf{x})\| = 2M$.

Now we get back to gradient descent on smooth functions with a global minimum. The punchline is so unspectacular that there is no harm in spoiling it already now: What we can prove is that $\|\nabla f(\mathbf{x}_t)\|^2$ converges to 0 at the same rate as $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ converges to 0 in the convex case. Naturally, $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ itself is not guaranteed to converge in the nonconvex case, for example if $\mathbf{x}_t$ converges to a local minimum that is not global, as in Figure 6.1.

It is tempting to interpret convergence of $\|\nabla f(\mathbf{x}_t)\|^2$ to 0 as convergence to a *critical point* of f (a point where the gradient vanishes). But this interpretation is not fully accurate in general, as Figure 6.2 (right) shows: The algorithm may enter a region where f asymptotically approaches some value, without reaching it (think of the rightmost piece of the function in the figure as $f(x) = e^{-x}$). In this case, the gradient converges to 0, but the iterates are nowhere near a critical point.

Theorem 6.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L according to Definition 2.2. Choosing stepsize*

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

gradient descent (2.11) yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big), \quad T > 0.$$

In particular, $\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big)$ for some $t \in \{0, \dots, T-1\}$. And also, $\lim_{t\to\infty} \|\nabla f(\mathbf{x}_t)\|^2 = 0$ (Exercise 34).

Proof. We recall that sufficient decrease (Lemma 2.6) does not require convexity, and this gives

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

Rewriting this into a bound on the gradient yields

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

Hence, we get a telescoping sum

$$\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq 2L\big(f(\mathbf{x}_0) - f(\mathbf{x}_T)\big) \leq 2L\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big).$$

The statement follows. □

In the smooth setting, gradient descent has another interesting property: with stepsize $1/L$, it cannot overshoot. By this, we mean that it cannot pass a critical point (in particular, not the global minimum) when moving from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$. Equivalently, with a smaller stepsize, no critical point can be reached. With stepsize $1/L$, it is possible to reach a critical point, as we have demonstrated for the supermodel function $f(x) = x^2$ in Section 2.7.

Lemma 6.3 (Exercise 35)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable; let $\mathbf{x} \in \mathbb{R}^d$ such that $\nabla f(\mathbf{x}) \neq \mathbf{0}$, i.e. $\mathbf{x}$ is not a critical point. Suppose that f is smooth with parameter L over the line segment connecting $\mathbf{x}$ and $\mathbf{x}' = \mathbf{x} - \gamma \nabla f(\mathbf{x})$, where $\gamma = 1/L' < 1/L$. Then $\mathbf{x}'$ is also not a critical point.*

Figure 6.5 illustrates the situation.

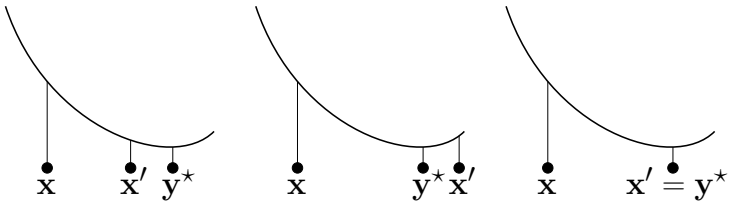


Figure 6.5: Gradient descent on smooth functions: When moving from $\mathbf{x}$ to $\mathbf{x}' = \mathbf{x} - \gamma \nabla f(\mathbf{x})$ with $\gamma < 1/L$, $\mathbf{x}'$ will not be a critical point (left); equivalently, with $\gamma = 1/L$, we cannot overshoot, i.e. pass a critical point (middle); with $\gamma = 1/L$, we may exactly reach a critical point (right).

6.2 Trajectory analysis

Even if the "landscape" (graph) of a nonconvex function has local minima, saddle points, and flat parts, it is sometimes possible to prove that gradient descent avoids these bad spots and still converges to a global minimum. For this, one needs a good starting point and some theoretical understanding of what happens when we start there—this is trajectory analysis.

In 2018, results along these lines have appeared that prove convergence of gradient descent to a global minimum in training deep *linear* linear networks, under suitable conditions. In this section, we will study a vastly simplified setting that allows us to show the main ideas (and limitations) behind one particular trajectory analysis [ACGH18].

In our simplified setting, we will look at the task of minimizing a concrete and very simple nonconvex function. This function turns out be smooth *along the trajectories* that we analyze, and this is one important ingredient. However, smoothness alone does not suffice to prove convergence to the global minimum, let alone fast convergence: As we have seen in the last section, we can in general only guarantee that the gradient norms converge to 0, and at a rather slow rate. To get beyond this, we will need to exploit additional properties of the function under consideration.

6.2.1 Deep linear neural networks

Let us go back to the problem of learning linear models as discussed in Section 1.7.2, using the example of Master's admission. We had n inputs $\mathbf{x}_1, \dots, \mathbf{x}_n$, where each input $\mathbf{x}_i \in \mathbb{R}^d$ consisted of d input variables; and we had n outputs $y_1, \dots, y_n \in \mathbb{R}$. Then we made the hypothesis that (after centering), output values depend (approximately) linearly on the input,

$$y_i \approx \mathbf{w}^\top x_i,$$

for a weight vector $\mathbf{w} = (w_1, \dots, w_d) \in \mathbb{R}^d$ to be learned.

Now we consider the more general case where there is not just one output $y_i \in \mathbb{R}$ as response to the i-th input, but m outputs $\mathbf{y}_i \in \mathbb{R}^m$. In this case, the linear hypothesis becomes

$$\mathbf{y}_i \approx W\mathbf{x}_i,$$

for a weight matrix $W \in \mathbb{R}^{m \times d}$ to be learned. The matrix that best fits this hypothesis on the given observations is the least-squares matrix

$$W^\star = \operatorname*{argmin}_{W \in \mathbb{R}^{m \times d}} \sum_{i=1}^{n} \|W\mathbf{x}_i - \mathbf{y}_i\|^2 .$$

If we let $X \in \mathbb{R}^{d \times n}$ be the matrix whose columns are the $\mathbf{x}_i$ and $Y \in \mathbb{R}^{m \times n}$ the matrix whose columns are the $\mathbf{y}_i$, we can equivalently write this as

$$W^\star = \operatorname*{argmin}_{W \in \mathbb{R}^{m \times d}} \|WX - Y\|_F^2, \tag{6.2}$$

where $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ is the *Frobenius norm* of a matrix A.

Finding W^* (the global minimum of a convex quadratic function) is a simple task that boils down to solving a system of linear equations; see also Section 1.5.2. A fancy way of saying this is that we are training a linear neural network with one layer, see Figure 6.6 (left).

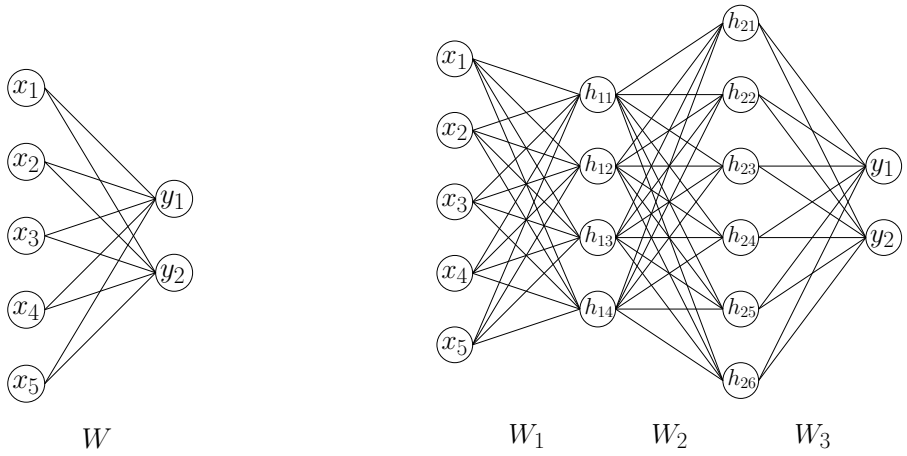


Figure 6.6: Left: A linear neural network over d input variables $\mathbf{x} = (x_1, \ldots, x_d)$ and m output variables $\mathbf{y} = (y_1, \ldots, y_m)$. The edge connecting input variable x_j with output variable y_i has a weight w_{ij} (to be learned), and all weights together form a weight matrix $W \in \mathbb{R}^{m \times d}$. Given the weights, the network computes the linear transformation $\mathbf{y} = W\mathbf{x}$ between inputs and outputs. Right: a deep linear neural network of depth 3 with weight matrices W_1, W_2, W_3. Given the weights, the network computes the linear transformation $\mathbf{y} = W_3 W_2 W_1 \mathbf{x}$.

But what if we have ℓ layers (Figure 6.6 (right)? Training such a network corresponds to minimizing

$$\|W_\ell W_{\ell-1} \cdots W_1 X - Y\|_F^2,$$

over ℓ weight matrices $W_1, \ldots, W_\ell$ to be learned. In case of linear neural networks, there is no benefit in adding layers, as any linear transformation $\mathbf{x} \mapsto W_\ell W_{\ell-1} \cdots W_1 X$ can of course be represented as $\mathbf{x} \mapsto WX$ with $W := W_{\ell-1} \cdots W_1$. But from a theoretical point of view, a deep linear neural network gives us a simple playground in which we can try to understand why training deep neural networks with gradient descent works,

despite the fact that the objective function is no longer convex. The hope is that such an understanding can ultimately lead to an analyis of gradient descent (or other suitable methods) for "real" (meaning non-linear) deep neural networks.

In the next section, we will discuss the case where all matrices are 1×1, so they are just numbers. This is arguably a toy example in our already simple playground. Still, it gives rise to a nontrivial nonconvex function, and the analysis of gradient descent on it will require similar ingredients as the one on general deep linear neural networks [ACGH18].

6.2.2 A simple nonconvex function

The function (that we consider fixed throughout the section) is $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\mathbf{x}) := \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2, \tag{6.3}$$

As d is fixed, we will abbreviate $\prod_{k=1}^{d} x_k$ by $\prod_k x_k$ throughout. Minimizing this function corresponds to training a deep linear neural network with d layers, one neuron per layer, with just one training input $x = 1$ and a corresponding output $y = 1$. Figure 6.7 visualizes the function f for $d = 2$.

First of all, the function f does have global minima, as it is nonnegative, and value 0 can be achieved (in many ways). Hence, we immediately know how to minimize this (for example, set $x_k = 1$ for all k). The question is whether gradient descent also knows, and if so, how we prove this.

Let us start by computing the gradient. We have

$$\nabla f(\mathbf{x}) = \left(\prod_k x_k - 1\right)\left(\prod_{k\neq 1} x_k\ ,\ \ldots\ ,\ \prod_{k\neq d} x_k\right)^\top. \tag{6.4}$$

What are the critical points, the ones where $\nabla f(\mathbf{x})$ vanishes? This happens when $\prod_k x_k = 1$ in which case we have a global minimum (level 0 in Figure 6.7). But there are other critical points. Whenever at least *two* of the x_k are zero, the gradient also vanishes, and the value of f is $1/2$ at such a point (point $\mathbf{0}$ in Figure 6.7). This already shows that the function cannot be convex, as for convex functions, every critical point is a global minimum (Lemma 1.16). It is easy to see that every non-optimal critical point must have two or more zeros.

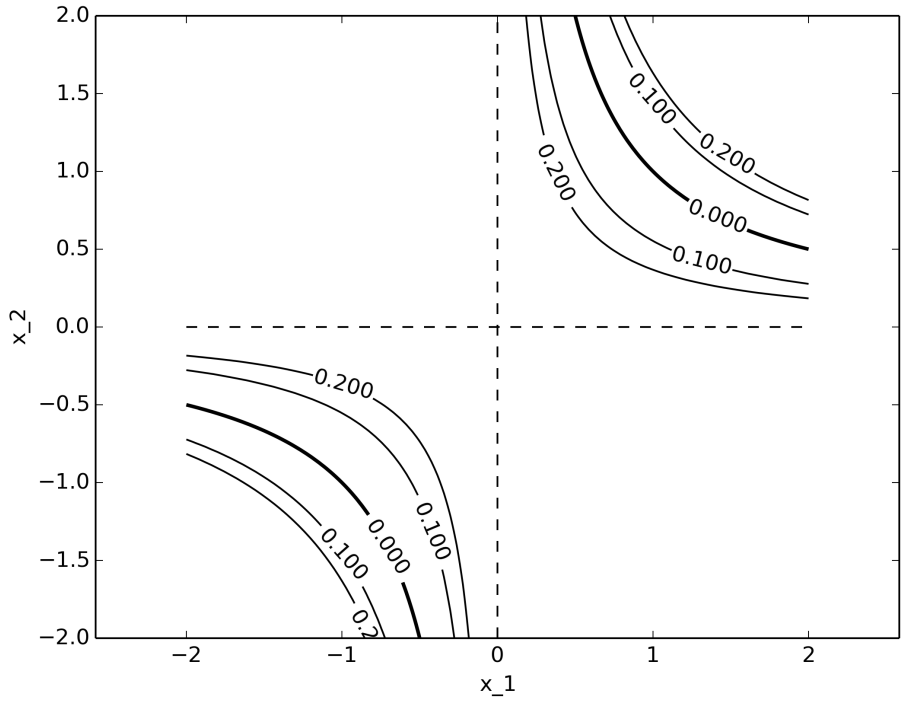


Figure 6.7: Levels sets of $f(x_1, x_2) = \frac{1}{2}(x_1 x_2 - 1)^2$

In fact, all critical points except the global minima are saddle points. This is because at any such point $\mathbf{x}$, we can slightly perturb the (two or more) zero entries in such a way that the product of all entries becomes either positive or negative, so that the function value either decreases or increases.

Figure 6.8 visualizes (scaled) negative gradients of f for $d = 2$; these are the directions in which gradient descent would move from the tails of the respective arrows. The figure already indicates that it is difficult to avoid convergence to a global minimum, but it is possible (see Exercise 37).

We now want to show that for any dimension d, and from *anywhere* in $X = \{\mathbf{x} : \mathbf{x} > \mathbf{0}, \prod_k \mathbf{x}_k \leq 1\}$, gradient descent will converge to a global minimum. Unfortunately, our function f is not smooth over X. For the analysis, we will therefore show that f is smooth along the trajectory of

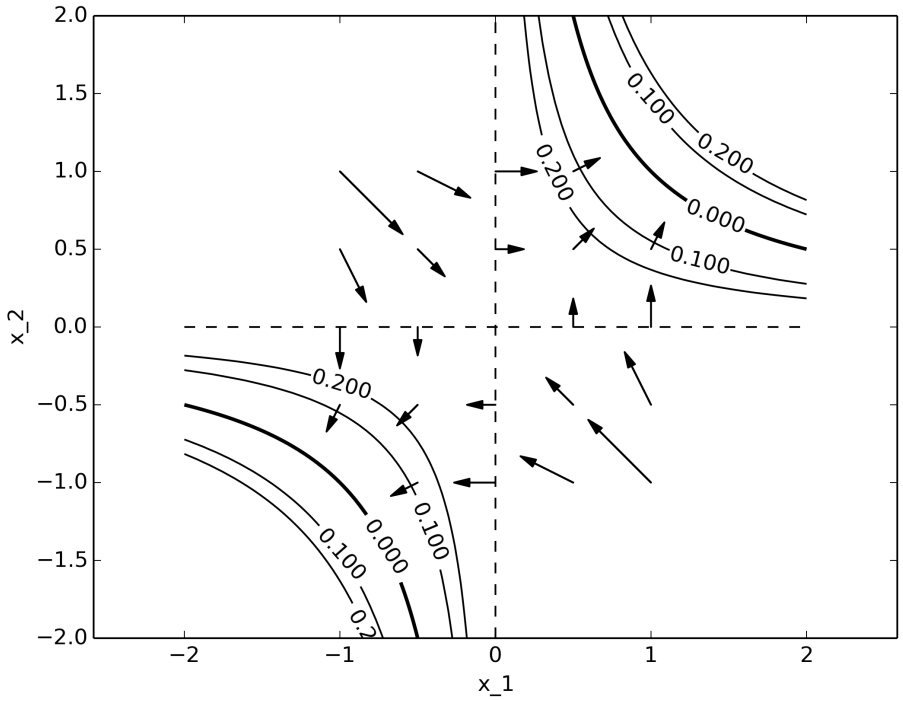


Figure 6.8: Scaled negative gradients of $f(x_1, x_2) = \frac{1}{2}(x_1 x_2 - 1)^2$

gradient descent for suitable L, so that we get sufficient decrease

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

by Lemma 2.6.

This already shows that gradient descent cannot converge to a saddle point: all these have (at least two) zero entries and therefore function value $1/2$. But for starting point $\mathbf{x}_0 \in X$, we have $f(\mathbf{x}_0) < 1/2$, so we can never reach a saddle while decreasing f.

But doesn't this mean that we necessarily *have* to converge to a global minimum? No, because the sublevel sets of f are unbounded, so it could in principle happen that gradient descent runs off to infinity while constantly improving $f(\mathbf{x}_t)$ (an example is gradient descent on $f(x) = e^{-x}$). Or some

other bad behavior occurs (we haven't characterized what can go wrong). So there is still something to prove.

How about convergence from other starting points? For $\mathbf{x} > \mathbf{0}, \prod_k \mathbf{x}_k \geq 1$, we also get convergence (Exercise 36). But there are also starting points from which gradient descent will not converge to a global minimum (Exercise 37).

The following simple lemma is the key to showing that gradient descent behaves nicely in our case.

Definition 6.4. *Let $\mathbf{x} > \mathbf{0}$ (componentwise) , and let $c \geq 1$ be a real number.* $\mathbf{x}$ *is called* c-balanced *if $x_i \leq c x_j$ for all $1 \leq i, j \leq d$.*

In fact, any initial iterate $\mathbf{x}_0 > \mathbf{0}$ is c-balanced for some (possibly large) c.

Lemma 6.5. *Let $\mathbf{x} > \mathbf{0}$ be c-balanced with $\prod_k x_k \leq 1$. Then for any stepsize $\gamma > 0$, $\mathbf{x}' := \mathbf{x} - \gamma \nabla f(\mathbf{x})$ satisfies $\mathbf{x}' \geq \mathbf{x}$ (componentwise) and is also c-balanced.*

If $c = 1$ (all entries of $\mathbf{x}$ are equal), this is easy to see since then also all entries of $\nabla f(\mathbf{x})$ in (6.4) are equal. Later we will show that for suitable step size, we also maintain that $\prod_k x'_k \leq 1$, so that gradient descent only goes through balanced iterates.

Proof. Set $\Delta := -\gamma(\prod_k x_k - 1)(\prod_k x_k) \geq 0$. Then the gradient descent update assumes the form

$$x'_k = x_k + \frac{\Delta}{x_k} \geq x_k, \quad k = 1, \ldots, d.$$

For i, j, we have $x_i \leq c x_j$ and $x_j \leq c x_i$ ($\Leftrightarrow 1/x_i \leq c/x_j$). We therefore get

$$x'_i = x_i + \frac{\Delta}{x_i} \leq c x_j + \frac{\Delta c}{x_j} = c x'_j.$$

□

6.2.3 Smoothness along the trajectory

It will turn out that our function f—despite not being globally smooth—is smooth over the trajectory of gradient descent, assuming that we start with $\mathbf{x}_0 > \mathbf{0}, \prod_k (\mathbf{x}_0) k < 1$. We will derive this from bounded Hessians. Let us therefore start by computing the Hessian matrix $\nabla^2 f(\mathbf{x})$, where by

definition, $\nabla^2 f(\mathbf{x})_{ij}$ is the j-th partial derivative of the i-th entry of $\nabla f(\mathbf{x})$. This i-th entry is

$$(\nabla f)_i = \left(\prod_k x_k - 1\right) \prod_{k \neq i} x_k$$

and its j-th partial derivative is therefore

$$\nabla^2 f(\mathbf{x})_{ij} = \begin{cases} \left(\prod_{k \neq i} x_i\right)^2, & j = i \\ 2 \prod_{k \neq i} x_k \prod_{k \neq j} x_k - \prod_{k \neq i,j} x_k, & j \neq i \end{cases}$$

This looks promising: if $\prod_k x_k \leq 1$, then we would also expect that the products $\prod_{k \neq i} x_k$ and $\prod_{k \neq i,j} x_k$ are small, in which case all entries of the Hessian are small, giving us a bound on $\|\nabla f^2 \mathbf{x}\|$ that we need to establish smoothness of f. However, for general $\mathbf{x}$, this fails. If $\mathbf{x}$ contains entries close to 0, it may happen that some terms $\prod_{k \neq i} x_k$ and $\prod_{k \neq i,j} x_k$ are actually very large.

What comes to our rescue is again c-balancedness.

Lemma 6.6. *Suppose that $\mathbf{x} > \mathbf{0}$ is c-balanced (Definition 6.4). Then for any $I \subseteq \{1, \ldots, d\}$, we have*

$$\left(\frac{1}{c}\right)^{|I|} \left(\prod_k x_k\right)^{1-|I|/d} \leq \prod_{k \notin I} x_k \leq c^{|I|} \left(\prod_k x_k\right)^{1-|I|/d}.$$

Proof. For any i, we have $x_i^d \geq (1/c)^d \prod_k x_k$ by balancedness, hence $x_i \geq (1/c)(\prod_k x_k)^{1/d}$. It follows that

$$\prod_{k \notin I} x_k = \frac{\prod_k x_k}{\prod_{i \in I} x_i} \leq \frac{\prod_k x_k}{(1/c)^{|I|} (\prod_k x_k)^{|I|/d}} = c^{|I|} \left(\prod_k x_k\right)^{1-|I|/d}.$$

The lower bound follows in the same way from $x_i^d \leq c^d \prod_k x_k$. □

This lets us bound the Hessians of c-balanced points.

Lemma 6.7. *Let $\mathbf{x} > \mathbf{0}$ be c-balanced with $\prod_k x_k \leq 1$. Then*

$$\left\|\nabla^2 f(\mathbf{x})\right\| \leq \left\|\nabla^2 f(\mathbf{x})\right\|_F \leq 3dc^2.$$

where $\|A\|_F$ is the Frobenius norm and $\|A\|$ the spectral norm.

Proof. The fact that $\|A\| \leq \|A\|_F$ is Exercise 38. To bound the Frobenius norm, we use the previous lemma to compute

$$\left|\nabla^2 f(\mathbf{x})_{ii}\right| = \left|\left(\prod_{k\neq i} x_i\right)^2\right| \leq c^2$$

and for $i \neq j$,

$$\left|\nabla^2 f(\mathbf{x})_{ij}\right| \leq \left|2\prod_{k\neq i} x_k \prod_{k\neq j} x_k\right| + \left|\prod_{k\neq i,j} x_k\right| \leq 3c^2.$$

Hence, $\|\nabla^2 f(\mathbf{x})\|_F^2 \leq 9d^2c^4$. Taking square roots, the statement follows. □

This now implies smoothness of f along the whole trajectory of gradient descent, under the usual "smooth stepsize" $\gamma = 1/L = 1/3dc^2$.

Lemma 6.8. *Let $\mathbf{x} > \mathbf{0}$ be c-balanced with $\prod_k x_k < 1$, $L = 3dc^2$. Let $\gamma := 1/L$. Then for all $0 \leq \nu \leq \gamma$,*

$$\mathbf{x}' := \mathbf{x} - \nu\nabla f(\mathbf{x}) \geq \mathbf{x}$$

is c-balanced with $\prod_k x'_k \leq 1$, and f is smooth with parameter L over the line segment connecting $\mathbf{x}$ and $\mathbf{x} - \gamma\nabla f(\mathbf{x})\}$.

Proof. We get that $\mathbf{x}' \geq \mathbf{x} > \mathbf{0}$ is c-balanced by Lemma 6.5. Furthermore, $\nabla f(\mathbf{x}) \neq \mathbf{0}$ (due to $\mathbf{x}' > \mathbf{0}, \prod_k x_k < 1$, we can't be at a critical point). By Lemma 6.3 (no overshooting), we can't reach $\prod_k x'_k = 1$ (global minimum) for $\nu < \gamma$, as f is smooth with parameter L over $X = \mathrm{conv}\{\mathbf{x}, \mathbf{x}'\}$ for the smallest such ν, using Lemma 6.1 with the bound on the Hessians from the previous lemma. By continutity, we therefore get $\prod_k x'_k \leq 1$ for all $\nu \leq \gamma$, and f is smooth with parameter L over the line segment connecting $\mathbf{x}$ and $\mathbf{x}'$ for $\nu = \gamma$. □

6.2.4 Convergence

Theorem 6.9. *Let $c \geq 1$ and $\delta > 0$ such that $\mathbf{x}_0 > \mathbf{0}$ is c-balanced with $\delta \leq \prod_k (\mathbf{x}_0)_k < 1$. Choosing stepsize*

$$\gamma = \frac{1}{3dc^2},$$

gradient descent satisfies

$$f(\mathbf{x}_T) \leq \left(1 - \frac{\delta^2}{3c^4}\right)^T f(\mathbf{x}_0), \quad T \geq 0.$$

This means that the loss indeed converges to its optimal value 0, and does so with a fast exponential error decrease. Exercise 39 asks you to prove that also the iterates themselves converge (to an optimal solution), so gradient descent will not run off to infinity.

Proof. For each $t \geq 0$, f is smooth over $\operatorname{conv}(\{\mathbf{x}_t, \mathbf{x}_{t+1}\})$ with parameter $L = 3dc^2$, hence Lemma 2.6 yields sufficient decrease:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{6dc^2} \left\|\nabla f(\mathbf{x}_t)\right\|^2. \tag{6.5}$$

For every c-balanced $\mathbf{x}$ with $\delta \leq \prod_k x_k \leq 1$, we have

$$\begin{aligned}
\left\|\nabla f(\mathbf{x})\right\|^2 &= 2f(\mathbf{x}) \sum_{i=1}^{d} \left(\prod_{k \neq i} x_k\right)^2 \\
&\geq 2f(\mathbf{x}) \frac{d}{c^2} \left(\prod_k x_k\right)^{2-2/d} \qquad \text{(Lemma 6.6)} \\
&\geq 2f(\mathbf{x}) \frac{d}{c^2} \left(\prod_k x_k\right)^2 \\
&\geq 2f(\mathbf{x}) \frac{d}{c^2} \delta^2.
\end{aligned}$$

Then, (6.5) further yields

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{6dc^2} 2f(\mathbf{x}_t) \frac{d}{c^2} \delta^2 = f(\mathbf{x}_t) \left(1 - \frac{\delta^2}{3c^4}\right),$$

proving the theorem. □

This looks great: just as for strongly convex functions, we seem to have fast convergence since the function value goes down by a constant factor in each step. There is a catch, though. To see this, consider the starting

solution $\mathbf{x}_0 = (1/2, \ldots, 1/2)$. This is c-balanced with $c = 1$, but the δ that we get is $1/2^d$. Hence, the "constant factor" is

$$\left(1 - \frac{1}{3 \cdot 4^d}\right),$$

and we need $T \approx 4^d$ to reduce the initial error by a constant factor not depending on d.

Indeed, for this starting value $\mathbf{x}_0$, the gradient is exponentially small, so we are crawling towards the optimum at exponentially small speed. In order to get polynomial-time convergence, we need to start with a δ that decays at most polynomially with d. For large d, this requires us to start very close to optimality. As a concrete example, let us try to achieve a constant δ (not depending on d) with a 1-balanced solution of the form $x_i = (1 - b/d)$ for all i. For this, we need that

$$\left(1 - \frac{b}{d}\right)^d \approx e^{-b} = \Omega(1),$$

and this requires $b = O(1)$. Hence, we need to start at distance $O(1/\sqrt{d})$ from the optimal solution $(1, \ldots, 1)$.

The problem is due to constant stepsize. Indeed, f is locally much smoother at small $\mathbf{x}_0$ than Lemma 6.8 predicts, so e could afford much larger steps in the beginning. The lemma covers the "worst case" when we are close to optimality already.

So could we improve using a time-varying stepsize? The question is moot: if we know the function f under consideration, we do not need to run any optimization in the first place. The question we were trying to address is whether and how a *standard* gradient descent algorithm is able to optimize nonconvex functions as well. Above, we have given a (partially satisfactory) answer for a concrete function: yes, it can, but at a very slow rate, if d is large and the starting point not close to optimality yet.

6.3 Exercises

Exercise 33. *Let $f : \mathbb{R}^n \to \mathbb{R}$ twice differentiable, with $X \subseteq \mathbf{dom}(f)$ an open convex set, and suppose that f is smooth with parameter L over X. Prove that*

under these conditions, the largest eigenvalue of the Hessian $\lambda_{max}(\nabla^2 f(\mathbf{x})) \leq L$ *for all* $\mathbf{x} \in X$.

Exercise 34. *Prove that the statement of Theorem 6.2 implies that*

$$\lim_{t\to\infty} \|\nabla f(\mathbf{x}_t)\|^2 = 0.$$

Exercise 35. *Prove Lemma 6.3 (gradient descent does not overshoot on smooth functions).*

Exercise 36. *Consider the function* $f(\mathbf{x}) = \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2$. *Prove that for any starting point* $\mathbf{x}_0 \in X = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} > \mathbf{0}, \prod_k \mathbf{x}_k \geq 1\}$ *and any* $\varepsilon > 0$, *gradient descent attains* $f(\mathbf{x}_T) \leq \varepsilon$ *for some iteration* T.

Exercise 37. *Consider the function* $f(\mathbf{x}) = \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2$. *Prove that for even dimension* $d \geq 2$, *there is a point* $\mathbf{x}_0$ *(not a critical point) such that gradient descent does not converge to a global minimum when started at* $\mathbf{x}_0$, *regardless of step size(s).*

Exercise 38. *Prove that for any matrix* A, $\|A\| \leq \|A\|_F$, *where* $\|\cdot\|$ *is the spectral norm and* $\|\cdot\|_F$ *the Frobenius norm.*

Exercise 39. *Prove that the sequence* $(\mathbf{x}_T)_{T\geq 0}$ *of iterates in Theorem 6.9 converges to a an optimal solution* $\mathbf{x}^\star$.

Chapter 7

Newton's Method

Contents

7.1 1-dimensional case

The Newton method (or Newton-Raphson method, invented by Sir Isaac Newton and formalized by Joseph Raphson) is an iterative method for finding a zero of a differentiable univariate function $f : \mathbb{R} \to \mathbb{R}$. Starting from some number x_0, it computes

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \geq 0. \tag{7.1}$$

Figure 7.1 shows what happens. x_{t+1} is the point where the tangent line to the graph of f at $(x_t, f(x_t))$ intersects the x-axis. In formulas, x_{t+1} is the solution of the linear equation

$$f(x_t) + f'(x_t)(x - x_t) = 0,$$

and this yields the update formula (7.1).

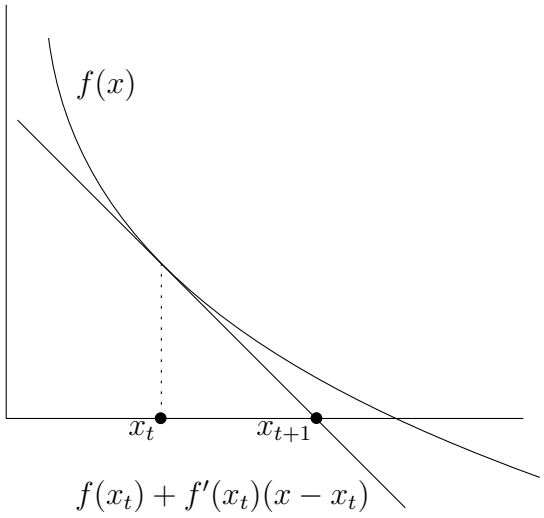


Figure 7.1: One step of Newton's method

The Newton step (7.1) obviously fails if $f'(x_t) = 0$ and may get out of control if $|f'(x_t)|$ is very small. Any theoretical analysis will have to make suitable assumptions to avoid this. But before going into this, we look at Newton's method in a benign case.

Let $f(x) = x^2 - R$, where $R \in \mathbb{R}_+$. f has two zeros, $\sqrt{R}$ and $-\sqrt{R}$. Starting for example at $x_0 = R$, we hope to converge to $\sqrt{R}$ quickly. In this case, (7.1) becomes

$$x_{t+1} = x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right). \tag{7.2}$$

This is in fact the *Babylonian method* to compute square roots, and here we see that it is just a special case of Newton's method.

Can we prove that we indeed quickly converge to $\sqrt{R}$? What we immediately see from (7.2) is that all iterates will be positive and hence

$$x_{t+1} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right) \geq \frac{x_t}{2}.$$

So we cannot be too fast. Suppose $R \geq 1$. In order to even get $x_t < 2\sqrt{R}$, we need at least $T \geq \log(R)/2$ steps. It turns out that the Babylonian method starts taking off only when $x_t - \sqrt{R} < 1/2$, say (Exercise 40 asks you to prove that it takes $\mathcal{O}(\log R)$ steps to get there).

To watch takeoff, let us now suppose that $x_0 - \sqrt{R} < 1/2$, so we are starting close to $\sqrt{R}$ already. We rewrite (7.2) as

$$x_{t+1} - \sqrt{R} = \frac{x_t}{2} + \frac{R}{2x_t} - \sqrt{R} = \frac{1}{2x_t}\left(x_t - \sqrt{R}\right)^2. \tag{7.3}$$

Assuming for now that $R \geq 1/4$, all iterates have value at least $\sqrt{R} \geq 1/2$, hence we get

$$x_{t+1} - \sqrt{R} \leq \left(x_t - \sqrt{R}\right)^2.$$

This means that the error goes to 0 *quadratically*, and

$$x_T - \sqrt{R} \leq \left(x_0 - \sqrt{R}\right)^{2^T} < \left(\frac{1}{2}\right)^{2^T}, \quad T \geq 0. \tag{7.4}$$

What does this tell us? In order to get $x_T - \sqrt{R} < \varepsilon$, we only need $T = \log\log(\frac{1}{\varepsilon})$ steps! Hence, it takes a while to get to roughly $\sqrt{R}$, but from there, we achieve high accuracy very fast.

Let us do a concrete example (with IEEE 754 double arithmetic). If $R = 1000$, we need 7 steps to get $x_7 - \sqrt{1000} < 1/2$, and then just 3 more steps to get x_{10} equal to $\sqrt{1000}$ up to the machine precision (53 binary digits). In this last phase, we essentially double the number of correct digits in each iteration!

7.2 Newton's method for optimization

Suppose we want to find a global minimum $x^\star$ of a differentiable convex function $f : \mathbb{R} \to \mathbb{R}$ (assuming that a global minimum exists). Lemmata 1.16 and 1.17 guarantee that we can equivalently search for a zero of the derivative f'. To do this, we can apply Newton's method if f is *twice* differentiable; the update step then becomes

$$x_{t+1} := x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - f''(x_t)^{-1} f'(x_t), \quad t \geq 0. \tag{7.5}$$

There is no reason to restrict to $d = 1$. Here is Newton's method for minimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$. We choose $\mathbf{x}_0$ arbitrarily and then iterate:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t), \quad t \geq 0. \tag{7.6}$$

The update vector $\nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$ is the result of a matrix-vector multiplication: we invert the Hessian at $\mathbf{x}_t$ and multiply the result with the gradient at $\mathbf{x}_t$. As before, this fails if the Hessian is not invertible, and may get out of control if the Hessian has small norm.

We have introduced iteration (7.6) simply as a (more or less natural) generalization of (7.5), but there's more to it. If we consider (7.6) as a special case of a general update scheme

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H(\mathbf{x}_t) \nabla f(\mathbf{x}_t),$$

where $H(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is some matrix, then we see that also gradient descent (2.11) is of this form, with $H(\mathbf{x}_t) = \gamma I$. Hence, Newton's method can also be thought of as "adaptive gradient descent" where the adaptation is w.r.t. the local geometry of the function at $\mathbf{x}_t$. Indeed, as we show next, this allows Newton's method to converge on *all* nondegenerate quadratic functions in one step, while gradient descent only does so with the right stepsize on "beautiful" quadratic functions whose sublevel sets are Euclidean balls (Exercise 18).

Lemma 7.1. *A* nondegenerate *quadratic function is a function of the form*

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top M \mathbf{x} - \mathbf{q}^\top \mathbf{x} + c,$$

where $M \in \mathbb{R}^{d \times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d, c \in R$. Let $\mathbf{x}^\star = M^{-1}\mathbf{q}$ be the unique solution of $\nabla f(\mathbf{x}) = \mathbf{0}$ (the unique global minimum if f is convex). With any starting point $\mathbf{x}_0 \in \mathbb{R}^d$, Newton's method (7.6) yields $\mathbf{x}_1 = \mathbf{x}^\star$.

Proof. We have $\nabla f(\mathbf{x}) = M\mathbf{x} - \mathbf{q}$ (this implies $\mathbf{x}^\star = M^{-1}\mathbf{q}$) and $\nabla^2 f(\mathbf{x}) = M$. Hence,

$$\mathbf{x}_0 - \nabla^2 f(\mathbf{x}_0)^{-1}\nabla f(\mathbf{x}_0) = \mathbf{x}_0 - M^{-1}(M\mathbf{x}_0 - \mathbf{q}) = M^{-1}\mathbf{q} = \mathbf{x}^\star.$$

□

In particular, Newton's method can solve an invertible system $M\mathbf{x} = \mathbf{q}$ of linear equations in one step. But no miracle is happening here, as this step involves the inversion of the matrix $\nabla^2 f(\mathbf{x}_0) = M$.

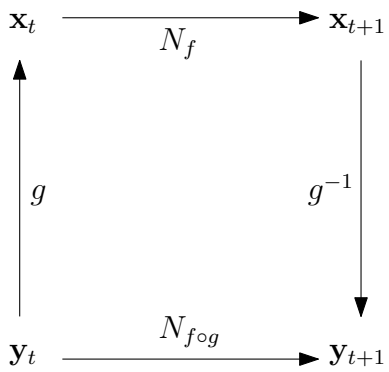
More generally, the behavior of Newton's method is affine invariant. By this, we mean that it is invariant under any invertible affine transformation, as follows:

Lemma 7.2 (Exercise 41)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable, $A \in \mathbb{R}^{d\times d}$ an invertible matrix, $\mathbf{b} \in \mathbb{R}^d$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be the (bijective) affine function $g(\mathbf{y}) = A\mathbf{y} + \mathbf{b}, \mathbf{y} \in \mathbb{R}^d$. Finally, for a twice differentiable function $h : \mathbb{R}^d \to \mathbb{R}$, let $N_h : \mathbb{R}^d \to \mathbb{R}^d$ denote the Newton step for h, i.e.*

$$N_h(\mathbf{x}) := \mathbf{x} - \nabla^2 h(\mathbf{x})^{-1}\nabla h(\mathbf{x}),$$

whenever this is defined. Then we have $N_{f\circ g} = g^{-1} \circ N_f \circ g$.

This says that in order to perform a Newton step for $f \circ g$ on $\mathbf{y}_t$, we can transform $\mathbf{y}_t$ to $\mathbf{x}_t = g(\mathbf{y}_t)$, perform the Newton step for f on $\mathbf{x}$ and transform the result $\mathbf{x}_{t+1}$ back to $\mathbf{y}_{t+1} = g^{-1}(\mathbf{x}_{t+1})$. Another way of saying this is that the following diagram commutes:



Hence, while gradient descent suffers if the coordinates are at very different scales, Newton's method doesn't.

We conclude the general exposition with another interpretation of Newton's method: each step minimizes the local second-order Taylor approximation.

Lemma 7.3 (Exercise 44)**.** *Let f be convex and twice differentiable at $\mathbf{x}_t \in \mathbf{dom}(f)$, with $\nabla^2 f(\mathbf{x}_t) \succ 0$ being invertible. The vector $\mathbf{x}_{t+1}$ resulting from the Netwon step (7.6) satisfies*

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t).$$

7.3 Once you're close, you're there...

We will prove a result about Newton's method that may seem rather weak: under suitable conditions, and starting close to the global minimum, we will reach distance at most ε to the minimum within $\log\log(1/\varepsilon)$ steps. The weak part here is of course not the number of steps $\log\log(1/\varepsilon)$—this is much faster than anything we have seen so far—but the assumption that we are starting close to the minimum already. Under such an assumption, we say that we have a *local convergence* result.

To compensate for the above weakness to some extent, we will be able to handle non-convex functions as well. More precisely, we show that under the aforementioned suitable conditions, and starting close to a critical point, we will reach distance at most ε to the critical point within $\log\log(1/\varepsilon)$ steps. This can of course only work if the conditions ensure that we are close to only one critical point; so we have a unique critical point nearby, and Newton's method will have no choice other than to converge to it.

For convex functions, we can ask about *global convergence* results that hold for every starting point. In general, such results were unknown for Newton's method as in (7.6) until recently. Under a stability assumption on the Hessian, global convergence was shown to hold by [KSJ18]. There are some variants of Newton's method for which such results can be proved, most notably the cubic regularization variant of Nesterov and Polyak [NP06]. Weak global convergence results can be obtained by adding

a step size to (7.6) and always making only steps that decrease the function value (which may not happen under the full Newton step).

An alternative is to use gradient descent to get us sufficiently close to the global minimum, and then switch to Newton's method for the rest. In Chapter 2, we have seen that under favorable conditions, we may know when gradient descent has taken us close enough.

In practice, Newton's method is often (but not always) much faster than gradient descent in terms of the number of iterations. The price to pay is a higher iteration cost, since we need to compute (and invert) Hessians.

After this disclaimer, let us state the main result right away. We follow Vishnoi [Vis15], except that we do not require convexity.

Theorem 7.4. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable with a critical point $\mathbf{x}^\star$. Suppose that there is a ball $X \subseteq \mathbf{dom}(f)$ with center $\mathbf{x}^\star$ such that the following two properties hold.*

(i) Bounded inverse Hessians: *There exists a real number $\mu > 0$ such that*

$$\|\nabla^2 f(\mathbf{x})^{-1}\| \leq \frac{1}{\mu}, \quad \forall \mathbf{x} \in X.$$

(ii) Lipschitz continuous Hessians*: There exists a real number $B \geq 0$ such that*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq B\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

In both cases, the matrix norm is the spectral norm defined in Lemma 2.5. Property (i) in particular stipulates that Hessians are invertible at all points in X.

Then, for $\mathbf{x}_t \in X$ and $\mathbf{x}_{t+1}$ resulting from the Newton step (7.6), we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{B}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

As an example, let us consider a nondegenerate quadratic function f (constant Hessian $M = \nabla^2 f(\mathbf{x})$ for all $\mathbf{x}$; see Lemma 7.1). Then f has exactly one critical point $\mathbf{x}^\star$. Property (i) is satisfied with $\mu = 1/\|M^{-1}\|$ over $X = \mathbb{R}^d$; property (ii) is satisfied for $B = 0$. According to the statement of the theorem, Newton's method will thus reach $\mathbf{x}^\star$ after one step—which we already know from Lemma 7.1.

In general, there could be several critical points for which properties (i) and (ii) hold, and it may seem surprising that the theorem makes a

statement about all of them. But in fact, if $\mathbf{x}_t$ is far away from such a critical point, the statement allows $\mathbf{x}_{t+1}$ to be even further away from it; we cannot expect to make progress towards all critical points simultaneously. The theorem becomes interesting only if we are *very close* to some critical point. In this case, we will actually converge to it. In particular, this critical point is then isolated and the only one nearby, so that Newton's method cannot avoid getting there.

Corollary 7.5 (Exercise 42)**.** *With the assumptions and terminology of Theorem 7.4, and if $\mathbf{x}_0 \in X$ satisfies*

$$\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq \frac{\mu}{B},$$

then Newton's method (7.6) yields

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

Hence, we have a bound as (7.4) for the last phase of the Babylonian method: in order to get $\|\mathbf{x}_T - \mathbf{x}^\star\| < \varepsilon$, we only need $T = \log\log(\frac{1}{\varepsilon})$ steps. But before this fast behavior kicks in, we need to be μ/B-close to $\mathbf{x}^\star$ already. The fact that $\mathbf{x}_0$ is this close to only *one* critical point necessarily follows.

An intuitive reason for a unique critical point near $\mathbf{x}_0$ (and for fast convergence to it) is that under our assumptions, the Hessians we encounter are almost constant when we are close to $\mathbf{x}^\star$. This means that locally, our function behaves almost like a nondegenerate quadratic function which has truly constant Hessians and allows Newton's method to convergence to its unique critical point in one step (Lemma 7.1).

Lemma 7.6 (Exercise 43)**.** *With the assumptions and terminology of Theorem 7.4, and if $\mathbf{x}_0 \in X$ satisfies*

$$\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq \frac{\mu}{B},$$

then the Hessians in Newton's method satisfy the relative error bound

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

We now still owe the reader the proof of the main convergence result, Theorem 7.4:

Proof of Theorem 7.4. To simplify notation, let us abbreviate $H := \nabla^2 f$, $\mathbf{x} = \mathbf{x}_t, \mathbf{x}' = \mathbf{x}_{t+1}$. Subtracting $\mathbf{x}^\star$ from both sides of (7.6), we get

$$\begin{aligned}\mathbf{x}' - \mathbf{x}^\star &= \mathbf{x} - \mathbf{x}^\star - H(\mathbf{x})^{-1}\nabla f(\mathbf{x}) \\ &= \mathbf{x} - \mathbf{x}^\star + H(\mathbf{x})^{-1}(\nabla f(\mathbf{x}^\star) - \nabla f(\mathbf{x})) \\ &= \mathbf{x} - \mathbf{x}^\star + H(\mathbf{x})^{-1}\int_0^1 H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x}))(\mathbf{x}^\star - \mathbf{x})dt,\end{aligned}$$

using the fundamental theorem of calculus and the chain rule as in (1.2) with $h(t) = \nabla f(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x}))$. With

$$\mathbf{x} - \mathbf{x}^\star = H(\mathbf{x})^{-1}H(\mathbf{x})(\mathbf{x} - \mathbf{x}^\star) = H(\mathbf{x})^{-1}\int_0^1 -H(\mathbf{x})(\mathbf{x}^\star - \mathbf{x})dt,$$

we further get

$$\mathbf{x}' - \mathbf{x}^\star = H(\mathbf{x})^{-1}\int_0^1 \left(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right)(\mathbf{x}^\star - \mathbf{x})dt.$$

Taking norms, we have

$$\|\mathbf{x}' - \mathbf{x}^\star\| \leq \|H(\mathbf{x})^{-1}\| \cdot \left\|\int_0^1 \left(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right)(\mathbf{x}^\star - \mathbf{x})dt\right\|,$$

where we have used that $\|A\mathbf{y}\| \leq \|A\| \cdot \|\mathbf{y}\|$ for any matrix $A \in \mathbb{R}^{d\times d}$ and any vector $\mathbf{y} \in \mathbb{R}^d$ which follows directly from the definition of the spectral norm. As we also have

$$\left\|\int_0^1 \mathbf{g}(t)dt\right\| \leq \int_0^1 \|\mathbf{g}(t)\|dt$$

for any vector-valued function $\mathbf{g}$ (Exercise 46), we can further bound

$$\begin{aligned}\|\mathbf{x}' - \mathbf{x}^\star\| &\leq \|H(\mathbf{x})^{-1}\|\int_0^1 \left\|\left(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right)(\mathbf{x}^\star - \mathbf{x})\right\|dt \\ &\leq \|H(\mathbf{x})^{-1}\|\int_0^1 \left\|H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right\| \cdot \|\mathbf{x}^\star - \mathbf{x}\|dt \\ &= \|H(\mathbf{x})^{-1}\| \cdot \|\mathbf{x}^\star - \mathbf{x}\|\int_0^1 \left\|H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right\|dt.\end{aligned}$$

We can now use the properties (i) and (ii) (bounded inverse Hessians, Lipschitz continuous Hessians) to conclude that

$$\|\mathbf{x}' - \mathbf{x}^\star\| \leq \frac{1}{\mu}\|\mathbf{x}^\star - \mathbf{x}\| \int_0^1 B\|t(\mathbf{x}^\star - \mathbf{x})\|dt = \frac{B}{\mu}\|\mathbf{x}^\star - \mathbf{x}\|^2 \underbrace{\int_0^1 t dt}_{1/2}.$$

□

How realistic are properties (i) and (ii)? If f is twice *continuously* differentiable (meaning that the second derivative $\nabla^2 f$ is continuous), then we will always find suitable values of μ and L over a ball X with center $\mathbf{x}^\star$—provided that $\nabla^2 f(\mathbf{x}^\star) \neq 0$.

Indeed, already in the one-dimensional case, we see that under $f''(x^\star) = 0$ (vanishing second derivative at the global minimum), Newton's method will in the worst reduce the distance to $x^\star$ at most by a constant factor in each step, no matter how close to $x^\star$ we start. Exercise 45 asks you to find such an example. In such a case, we have linear convergence, but the fast quadratic convergence ($\mathcal{O}(\log\log(1/\varepsilon))$ steps cannot be proven.

One way to ensure bounded inverse Hessians is to require strong convexity over X.

Lemma 7.7 (Exercise 47)**.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable and strongly convex with parameter μ over an open convex subset $X \subseteq \mathbf{dom}(f)$ according to Definition 2.9, meaning that*

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

Then $\nabla^2 f(\mathbf{x})$ is invertible and $\|\nabla^2 f(\mathbf{x})^{-1}\| \leq 1/\mu$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is the spectral norm defined in Lemma 2.5.

7.4 Exercises

Exercise 40. *Consider the Babylonian method (7.2). Prove that we get $x_T - \sqrt{R} < 1/2$ for $T = \mathcal{O}(\log R)$.*

Exercise 41. *Prove Lemma 7.2!*

Exercise 42. *Prove Corollary 7.5!*

Exercise 43. *Prove Lemma 7.6!*

Exercise 44. *Prove Lemma 7.3!*

Exercise 45. *Let $\delta > 0$ be any real number. Find an example of a convex function $f : \mathbb{R} \to \mathbb{R}$ such that (i) the unique global minimum $x^\star$ has a vanishing second derivative $f''(x^\star) = 0$, and (ii) Newton's method satisfies*

$$|x_{t+1} - x^\star| \geq (1 - \delta)|x_t - x^\star|,$$

for all $x_t \neq x^\star$.

Exercise 46. *This exercise is just meant to recall some basics around integrals. Show that for a vector-valued function $\mathbf{g} : \mathbb{R} \to \mathbb{R}^d$, the inequality*

$$\left\| \int_0^1 \mathbf{g}(t)dt \right\| \leq \int_0^1 \|\mathbf{g}(t)\|dt$$

holds, where $\|\cdot\|$ is the 2-norm (always assuming that the funtions under consideration are integrable)! You may assume (i) that integrals are linear:

$$\int_0^1 (\lambda_1 g_1(t) + \lambda_2 g_2(t))dt = \lambda_1 \int_0^1 g_1(t)dt + \lambda_2 \int_0^1 g_2(t)dt,$$

And (ii), if $g(t) \geq 0$ for all $t \in [0, 1]$, then $\int_0^1 g(t)dt \geq 0$.

Exercise 47. *Prove Lemma 7.7! You may want to proceed in the following steps.*

(i) Prove that the function $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex over X (see also Exercise 28).

(ii) Prove that $\nabla^2 f(\mathbf{x})$ is invertible for all $\mathbf{x} \in X$.

(iii) Prove that all eigenvalues of $\nabla^2 f(\mathbf{x})^{-1}$ are positive and at most $1/\mu$.

(iv) Prove that for a symmetric matrix M, the spectral norm $\|M\|$ is the largest absolute eigenvalue.

Chapter 8

Quasi-Newton Methods

Contents

The main computational bottleneck in Newton's method (7.6) is the computation and inversion of the Hessian matrix in each step. This matrix has size $d \times d$, so it will take up to $\mathcal{O}(d^3)$ time to invert it (or to solve the system $\nabla^2 f(\mathbf{x}_t)\Delta\mathbf{x} = -\nabla f(\mathbf{x}_t)$ that gives us the next Newton step $\Delta\mathbf{x}$). Already in the 1950s, attempts were made to circumvent this costly step, the first one going back to Davidon [Dav59].

In this chapter, we will (for a change) not prove convergence results; rather, we focus on the development of Quasi-Newton methods, and how state-of-the-art methods arise from first principles. To motivate the approach, let us go back to the 1-dimensional case.

8.1 The secant method

Like Newton's method (7.1), the secant method is an iterative method for finding a zero of a univariate function. Unlike Newton's method, it does not use derivatives and hence does not require the function under consideration to be differentiable. In fact, it is (therefore) much older than Newton's method. Reversing history and starting from the Newton step

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \geq 0,$$

we can derive the secant method by replacing the derivative $f'(x_t)$ with its finite difference approximation

$$\frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}.$$

As we (in the differentiable case) have

$$f'(x_t) = \lim_{x \to x_t} \frac{f(x_t) - f(x)}{x_t - x},$$

we get

$$\frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}} \approx f'(x_t)$$

for $|x_t - x_{t-1}|$ small. As the method proceeds, we expect consecutive iterates x_{t-1}, x_t to become closer and closer, so that the *secant step*

$$x_{t+1} := x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}, \quad t \geq 1 \tag{8.1}$$

approximates the Newton step (*two* starting values x_0, x_1 need to be chosen here). Figure 8.1 shows what the method does: it constructs the line through the two points $(x_{t-1}, f(x_{t-1}))$ and $(x_t, f(x_t))$ on the graph of f; the next iterate x_{t+1} is where this line intersects the x-axis. Exercise 48 asks you to formally prove this.

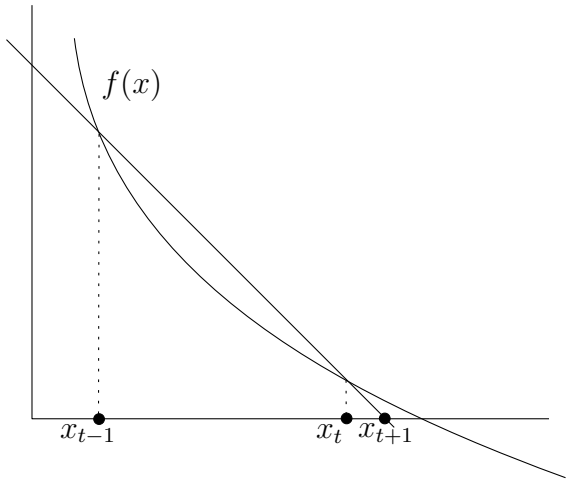


Figure 8.1: One step of the secant method

Convergence of the secant method can be analyzed, but we don't do this here. The main point for us is that we now have a *derivative-free* version of Newton's method.

When the task is to optimize a differentiable univariate function, we can apply the secant method to its derivative to obtain the secant method for optimization:

$$x_{t+1} := x_t - f'(x_t)\frac{x_t - x_{t-1}}{f'(x_t) - f'(x_{t-1})}, \quad t \geq 1. \tag{8.2}$$

This is a *second-derivative-free* version of Newton's method (7.5) for optimization. The plan is now to generalize this to higher dimensions to obtain a *Hessian-free* version of Newton's method (7.6) for optimization over $\mathbb{R}^d$.

8.2 The secant condition

Applying finite difference approximation to the second derivative of f (we're still in the 1-dimensional case), we get

$$H_t := \frac{f'(x_t) - f'(x_{t-1})}{x_t - x_{t-1}} \approx f''(x_t),$$

which we can write as

$$f'(x_t) - f'(x_{t-1}) = H_t(x_t - x_{t-1}) \approx f''(x_t)(x_t - x_{t-1}). \tag{8.3}$$

Now, while Newton's method for optimization uses the update step

$$x_{t+1} = x_t - f''(x_t)^{-1} f'(x_t), \quad t \geq 0,$$

the secant method works with the approximation $H_t \approx f''(x_t)$:

$$x_{t+1} = x_t - H_t^{-1} f'(x_t), \quad t \geq 1. \tag{8.4}$$

The fact that H_t approximates $f''(x_t)$ in the twice differentiable case was our motivation for the secant method, but in the method itself, there is no reference to f'' (which is exactly the point). All that is needed is the *secant condition* from (8.3) that defines H_t:

$$f'(x_t) - f'(x_{t-1}) = H_t(x_t - x_{t-1}). \tag{8.5}$$

This view can be generalized to higher dimensions. If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, (8.4) becomes

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H_t^{-1} \nabla f(\mathbf{x}_t), \quad t \geq 1, \tag{8.6}$$

where $H_t \in \mathbb{R}^{d \times d}$ is now supposed to be a symmetric matrix satisfying the d-dimensional secant condition

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1}). \tag{8.7}$$

8.3 Quasi-Newton methods

If f is twice differentiable, the secant condition (8.7) along with the first-order Taylor approximation of $\nabla f(\mathbf{x})$ yields the d-dimensional analog of (8.3):

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \approx \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}_{t-1}),$$

We might therefore hope that $H_t \approx \nabla^2 f(\mathbf{x}_t)$, and this would mean that (8.6) approximates Newton's method. Therefore, whenever we use (8.6) with a *symmetric* matrix satisfying the secant condition (8.7), we say that we have a *Quasi-Newton method*.

In the 1-dimensional case, there is only one Quasi-Newton method—the secant method (8.1). Indeed, equation (8.5) uniquely defines the number H_t in each step.

But in the d-dimensional case, the matrix H_t in the secant condition is underdetermined, starting from $d = 2$: Taking the symmetry requirement into account, (8.7) is a system of d equations in $d(d+1)/2$ unknowns, so if it is satisfiable at all, there are many solutions H_t. This raises the question of which one to choose, and how to do so efficiently; after all, we want to get some savings over Newton's method.

Newton's method is a Quasi-Newton method if and only if f is a non-degenerate quadratic function (Exercise 49). Hence, Quasi-Newton methods do not generalize Newton's method but form a family of related algorithms.

The first Quasi-Newton method was developed by William C. Davidon in 1956; he desperately needed iterations that were faster than those of Newton's method in order obtain results in the short time spans between expected failures of the room-sized computer that he used to run his computations on.

But the paper he wrote about his new method got rejected for lacking a convergence analysis, and for allegedly dubious notation. It became a very influential Technical Report in 1959 [Dav59] and was finally officially published in 1991, with a foreword giving the historical context [Dav91]. Ironically, Quasi-Newton methods are today the methods of choice in a number of relevant machine learning applications.

8.4 Greenstadt's approach *(Optional Material)*

For efficieny reasons (we want to avoid matrix inversions), Quasi-Newton methods typically directly deal with the inverse matrices H_t^{-1}. Suppose that we have the iterates $\mathbf{x}_{t-1}, \mathbf{x}_t$ as well as the matrix H_{t-1}^{-1}; now we want to compute a matrix H_t^{-1} to perform the next Quasi-Newton step (8.6). How should we choose H_t^{-1}?

We draw some intuition from (the analysis of) Newton's method. Recall that we have shown $\nabla^2 f(\mathbf{x}_t)$ to fluctuate only very little in the region of extremely fast convergence (Lemma 7.6); in fact, Newton's method is optimal (one step!) when $\nabla^2 f(\mathbf{x}_t)$ is actually constant— this is the case of a quadratic function, see Lemma 7.1. Hence, in a Quasi-Newton method, it also makes sense to have that $H_t \approx H_{t-1}$, or $H_t^{-1} \approx H_{t-1}^{-1}$.

Greenstadt's approach from 1970 [Gre70] is to update H_{t-1}^{-1} by an "error matrix" E_t to obtain

$$H_t^{-1} = H_{t-1}^{-1} + E_t.$$

Moreover, the errors should be as small as possible, subject to the constraints that H_t^{-1} is symmetric and satisfies the secant condition (8.7). A simple measure of error introduced by an update matrix E is its squared *Frobenius norm*

$$\|E\|_F^2 := \sum_{i=1}^{d} \sum_{j=1}^{d} e_{ij}^2.$$

Since Greenstadt considered the resulting Quasi-Newton method as "too specialized", he searched for a compromise between variability in the method and simplicity of the resulting formulas. This led him to minimize the error term

$$\|AEA^\top\|_F^2,$$

where $A \in \mathbb{R}^{d \times d}$ is some fixed invertible transformation matrix. If $A = I$, we recover the squared Frobenius norm of E.

Let us now fix t and simplify notation by setting

$$\begin{aligned} H &:= H_{t-1}^{-1}, \\ H' &:= H_t^{-1}, \\ E &:= E_t, \\ \boldsymbol{\sigma} &:= \mathbf{x}_t - \mathbf{x}_{t-1}, \\ \mathbf{y} &= \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \\ \mathbf{r} &= \boldsymbol{\sigma} - H\mathbf{y}. \end{aligned}$$

The update formula then is

$$H' = H + E, \tag{8.8}$$

and the secant condition $\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1})$ becomes

$$H'\mathbf{y} = \boldsymbol{\sigma} \quad (\Leftrightarrow E\mathbf{y} = \mathbf{r}). \tag{8.9}$$

Greenstadt's approach can now be distilled into the following convex constrained minimization problem in the d^2 variables E_{ij}:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|AEA^\top\|_F^2 \\ \text{subject to} & E\mathbf{y} = \mathbf{r} \\ & E^\top - E = 0 \end{array} \tag{8.10}$$

8.4.1 The method of Lagrange multipliers

Minimization subject to equality constraints can be done via the method of *Lagrange multipliers*. Here we need it only for the case of *linear* equality constraints in which case the method assumes a very simple form.

Theorem 8.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $C \in \mathbb{R}^{m\times d}$ for some $m \in \mathbb{N}$, $\mathbf{e} \in \mathbb{R}^m$, $\mathbf{x}^\star \in \mathbb{R}^d$ such that $C\mathbf{x}^\star = \mathbf{e}$. Then the following two statements are equivalent.*

(i) $\mathbf{x}^\star = \operatorname{argmin}\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d, C\mathbf{x} = \mathbf{e}\}$

(ii) There exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ such that

$$\nabla f(\mathbf{x}^\star)^\top = \boldsymbol{\lambda}^\top C.$$

The entries of $\boldsymbol{\lambda}$ are known as the Lagrange multipliers.

Proof. The easy direction is (ii)⇒(i): if $\boldsymbol{\lambda}$ as specified exists and $\mathbf{x} \in \mathbb{R}^d$ satisfies $C\mathbf{x} = \mathbf{e}$, we get

$$\nabla f(\mathbf{x}^\star)^\top(\mathbf{x} - \mathbf{x}^\star) = \boldsymbol{\lambda}^\top C(\mathbf{x} - \mathbf{x}^\star) = \boldsymbol{\lambda}^\top(\mathbf{e} - \mathbf{e}) = 0.$$

Hence, $\mathbf{x}^\star$ is a minimizer of f over $\{\mathbf{x} \in \mathbb{R}^d : C\mathbf{x} = \mathbf{e}\}$ by the optimality condition of Lemma 1.22.

The other direction is Exercise 50. □

8.4.2 Application to Greenstadt's Update

In order to apply this method to (8.10), we need to compute the gradient of $f(E) = \frac{1}{2}\|AEA^\top\|_F^2$. Formally, this is a d^2-dimensional vector, but it is customary and more practical to write it as a matrix again,

$$\nabla f(E) = \left(\frac{\partial f(E)}{\partial E_{ij}}\right)_{1\le i,j\le d}.$$

Fact 8.2 (Exercise 51)**.** *Let $A, B \in \mathbb{R}^{d\times d}$ two matrices. With $f : \mathbb{R}^{d\times d} \to \mathbb{R}$, $f(E) := \frac{1}{2}\|AEB\|_F^2$, we have*

$$\nabla f(E) = A^\top AEBB^\top.$$

The second step is to write the system of equations $E\mathbf{y} = \mathbf{r}, E^\top - E = 0$ in Greenstadt's convex program (8.10) in matrix form $C\mathbf{x} = \mathbf{e}$ so that we can apply the method of Lagrange multipliers according to Theorem 8.1.

As there are $d + d^2$ equations in d^2 variables, it is best to think of the rows of C as being indexed with elements $i \in [d] := \{1, \dots, d\}$ for the first d equations $E\mathbf{y} = \mathbf{r}$, and pairs $(i, j) \in [d] \times [d]$ for the last d^2 symmetry constraints (more than half of which are redundant but we don't care). Columns of C are indexed with pairs (i, j) as well.

Let us denote by $\boldsymbol{\lambda} \in \mathbb{R}^d$ the Lagrange multipliers for the first d equations and $\Gamma \in \mathbb{R}^{d\times d}$ the ones for the last d^2 ones.

In column (i, j) of C corresponding to variable E_{ij}, we have entry y_j in row i as well as entries 1 (row (j, i)) and -1 (row (i, j)). Taking the inner product with the Lagrange multipliers, this column therefore yields

$$\lambda_i y_j + \Gamma_{ji} - \Gamma_{ij}.$$

After aggregating these entries into a $d \times d$ matrix, Theorem 8.1 tells us that we should aim for equality with $\nabla f(E)$ as derived in Fact 8.2. We have proved the following intermediate result.

Lemma 8.3. *An update matrix $E^\star$ satisfying the constraints $E\mathbf{y} = \mathbf{r}$ (secant condition in the next step) and $E^\top - E = 0$ (symmetry) is a minimizer of the error function $f(E) := \frac{1}{2}\|AEA^\top\|_F^2$ subject to the aforementioned constraints if and only if there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^d$ and a matrix $\Gamma \in \mathbb{R}^{d\times d}$ such that*

$$WE^\star W = \boldsymbol{\lambda}\mathbf{y}^\top + \Gamma^\top - \Gamma, \tag{8.11}$$

where $W := A^\top A$ (a symmetric and positive definite matrix).

Note that $\boldsymbol{\lambda}\mathbf{y}^\top$ is the *outer product* of a column and a row vector and hence a matrix. As we assume A to be invertible, the quadratic function $f(E)$ is easily seen to be strongly convex and as a consequence has a unique minimizer $E^\star$ subject to the set of linear equations in (8.10) (see Lemma 2.10 which also applies if we minimize over a closed set). Hence, we know that the minimizer $E^\star$ and corresponding Lagrange multipiers $\boldsymbol{\lambda}, \Gamma$ exist.

8.4.3 The Greenstadt family

We need to solve the system of equations

$$\begin{aligned} E\mathbf{y} &= \mathbf{r}, && (8.12)\\ E^\top - E &= 0, && (8.13)\\ WEW &= \boldsymbol{\lambda}\mathbf{y}^\top + \Gamma^\top - \Gamma. && (8.14) \end{aligned}$$

This system is linear in $E, \boldsymbol{\lambda}, \Gamma$, hence easy to solve computationally. However, we want a formula for the unique solution $E^\star$ in terms of the parameters $W, \mathbf{y}, \boldsymbol{\sigma} = \mathbf{r} + H\mathbf{y}$. In the following derivation, we closely follow Greenstadt [Gre70, pages 4–5].

With $M := W^{-1}$ (which exists since $W = A^\top A$ is positive definite), (8.14) can be rewritten as

$$E = M\left(\boldsymbol{\lambda}\mathbf{y}^\top + \Gamma^\top - \Gamma\right)M. \tag{8.15}$$

Transposing this system (using that M is symmetric) yields

$$E^\top = M\left(\mathbf{y}\boldsymbol{\lambda}^\top + \Gamma - \Gamma^\top\right)M.$$

By symmetry (8.13), we can subtract the latter two equations to obtain

$$M\left(\boldsymbol{\lambda}\mathbf{y}^\top - \mathbf{y}\boldsymbol{\lambda}^\top + 2\Gamma^\top - 2\Gamma\right)M = 0.$$

As M is invertible, this is equivalent to

$$\Gamma^\top - \Gamma = \frac{1}{2}\left(\mathbf{y}\boldsymbol{\lambda}^\top - \boldsymbol{\lambda}\mathbf{y}^\top\right),$$

so we can eliminate Γ by substituting back into (8.15):

$$E = M\left(\boldsymbol{\lambda}\mathbf{y}^\top + \frac{1}{2}\left(\mathbf{y}\boldsymbol{\lambda}^\top - \boldsymbol{\lambda}\mathbf{y}^\top\right)\right)M = \frac{1}{2}M\left(\boldsymbol{\lambda}\mathbf{y}^\top + \mathbf{y}\boldsymbol{\lambda}^\top\right)M. \tag{8.16}$$

To also eliminate $\boldsymbol{\lambda}$, we now use (8.12)—the secant condition in the next step—to get

$$E\mathbf{y} = \frac{1}{2}M\left(\boldsymbol{\lambda}\mathbf{y}^\top + \mathbf{y}\boldsymbol{\lambda}^\top\right)M\mathbf{y} = \mathbf{r}.$$

Premultiplying with $2M^{-1}$ gives

$$2M^{-1}\mathbf{r} = \left(\boldsymbol{\lambda}\mathbf{y}^\top + \mathbf{y}\boldsymbol{\lambda}^\top\right)M\mathbf{y} = \boldsymbol{\lambda}\mathbf{y}^\top M\mathbf{y} + \mathbf{y}\boldsymbol{\lambda}^\top M\mathbf{y}.$$

Hence,

$$\boldsymbol{\lambda} = \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2M^{-1}\mathbf{r} - \mathbf{y}\boldsymbol{\lambda}^\top M\mathbf{y}\right). \tag{8.17}$$

To get rid of $\boldsymbol{\lambda}$ on the right hand side, we premultiply this with $\mathbf{y}^\top M$ to obtain

$$\underbrace{\mathbf{y}^\top M\boldsymbol{\lambda}}_{z} = \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2\mathbf{y}^\top\mathbf{r} - (\mathbf{y}^\top M\mathbf{y})(\underbrace{\boldsymbol{\lambda}^\top M\mathbf{y}}_{z})\right) = \frac{2\mathbf{y}^\top\mathbf{r}}{\mathbf{y}^\top M\mathbf{y}} - \underbrace{\boldsymbol{\lambda}^\top M\mathbf{y}}_{z}$$

It follows that

$$z = \boldsymbol{\lambda}^\top M\mathbf{y} = \frac{\mathbf{y}^\top\mathbf{r}}{\mathbf{y}^\top M\mathbf{y}}.$$

This in turn can be substituted into the right-hand side of (8.17) to remove $\boldsymbol{\lambda}$ there, and we get

$$\boldsymbol{\lambda} = \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2M^{-1}\mathbf{r} - \frac{(\mathbf{y}^\top\mathbf{r})}{\mathbf{y}^\top M\mathbf{y}}\mathbf{y}\right).$$

Consequently,

$$\begin{aligned}
\boldsymbol{\lambda}\mathbf{y}^\top &= \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2M^{-1}\mathbf{r}\mathbf{y}^\top - \frac{(\mathbf{y}^\top\mathbf{r})}{\mathbf{y}^\top M\mathbf{y}}\mathbf{y}\mathbf{y}^\top\right), \\
\mathbf{y}\boldsymbol{\lambda}^\top &= \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2\mathbf{y}\mathbf{r}^\top M^{-1} - \frac{(\mathbf{y}^\top\mathbf{r})}{\mathbf{y}^\top M\mathbf{y}}\mathbf{y}\mathbf{y}^\top\right).
\end{aligned}$$

This gives us an explicit formula for E, by substituting the previous expressions back into (8.16). For this, we compute

$$\begin{aligned}
M\boldsymbol{\lambda}\mathbf{y}^\top M &= \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2\mathbf{r}\mathbf{y}^\top M - \frac{(\mathbf{y}^\top\mathbf{r})}{\mathbf{y}^\top M\mathbf{y}}M\mathbf{y}\mathbf{y}^\top M\right), \\
M\mathbf{y}\boldsymbol{\lambda}^\top M &= \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(2M\mathbf{y}\mathbf{r}^\top - \frac{(\mathbf{y}^\top\mathbf{r})}{\mathbf{y}^\top M\mathbf{y}}M\mathbf{y}\mathbf{y}^\top M\right),
\end{aligned}$$

and consequently,

$$E = \frac{1}{2}M\left(\boldsymbol{\lambda}\mathbf{y}^\top + \mathbf{y}\boldsymbol{\lambda}^\top\right)M = \frac{1}{\mathbf{y}^\top M\mathbf{y}} \left(\mathbf{r}\mathbf{y}^\top M + M\mathbf{y}\mathbf{r}^\top - \frac{(\mathbf{y}^\top\mathbf{r})}{\mathbf{y}^\top M\mathbf{y}}M\mathbf{y}\mathbf{y}^\top M\right). \tag{8.18}$$

Finally, we use $\mathbf{r} = \boldsymbol{\sigma} - H\mathbf{y}$ to obtain the update matrix $E^\star$ in terms of the original parameters $H = H_{t-1}^{-1}$ (previous approximation of the inverse Hessian that we now want to update to $H_t^{-1} = H' = H + E^\star$), $\boldsymbol{\sigma} = \mathbf{x}_t - \mathbf{x}_{t-1}$ (previous Quasi-Newton step) and $\mathbf{y} = \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})$ (previous change in gradients). This gives us the Greenstadt family of Quasi-Newton methods.

Definition 8.4. *Let $M \in \mathbb{R}^{d\times d}$ be a symmetric and invertible matrix. Consider the Quasi-Newton method*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H_t^{-1}\nabla f(\mathbf{x}_t), \quad t \geq 1,$$

where $H_0 = I$ (or some other positive definite matrix), and $H_t^{-1} = H_{t-1}^{-1} + E_t$ is chosen for all $t \geq 1$ in such a way that H_t^{-1} is symmetric and satisfies the secant condition

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1}).$$

For any fixed t, set

$$\begin{aligned} H &:= H_{t-1}^{-1}, \\ H' &:= H_t^{-1}, \\ \boldsymbol{\sigma} &:= \mathbf{x}_t - \mathbf{x}_{t-1}, \\ \mathbf{y} &:= \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \end{aligned}$$

and define

$$\begin{aligned} E^\star = \frac{1}{\mathbf{y}^\top M\mathbf{y}}\Big(\boldsymbol{\sigma}\mathbf{y}^\top M &+ M\mathbf{y}\boldsymbol{\sigma}^\top - H\mathbf{y}\mathbf{y}^\top M - M\mathbf{y}\mathbf{y}^\top H \\ &- \frac{1}{\mathbf{y}^\top M\mathbf{y}}(\mathbf{y}^\top\boldsymbol{\sigma} - \mathbf{y}^\top H\mathbf{y})M\mathbf{y}\mathbf{y}^\top M\Big). \end{aligned} \tag{8.19}$$

If the update matrix $E_t = E^\star$ is used, the method is called the Greenstadt method *with parameter M.*

8.4.4 The BFGS method

In his paper, Greenstadt suggested two obvious choices for the matrix M In Definition 8.4, namely $M = H$ (the previous approximation of the inverse Hessian) and $M = I$. In the next paper of the same issue of the same

journal, Goldfarb suggested to use the matrix $M = H'$, the *next* approximation of the inverse Hessian. Even though we don't yet have it, we can use it in the formula (8.19) since we know that H' will by design satisfy the secant condition $H'\mathbf{y} = \boldsymbol{\sigma}$. And as M always appears next to $\mathbf{y}$ in (8.19), $M\mathbf{y} = H'\mathbf{y} = \boldsymbol{\sigma}$, so H' disappears from the formula!

Definition 8.5. *The* BFGS method *is the Greenstadt method with parameter $M = H' = H_t^{-1}$ in step t, in which case the update matrix $E^\star$ assumes the form*

$$\begin{aligned} E^\star &= \frac{1}{\mathbf{y}^\top\boldsymbol{\sigma}}\Big(2\boldsymbol{\sigma}\boldsymbol{\sigma}^\top - H\mathbf{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\mathbf{y}^\top H - \frac{1}{\boldsymbol{\sigma}^\top\mathbf{y}}(\mathbf{y}^\top\boldsymbol{\sigma} - \mathbf{y}^\top H\mathbf{y})\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\Big) \\ &= \frac{1}{\mathbf{y}^\top\boldsymbol{\sigma}}\Big(-H\mathbf{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\mathbf{y}^\top H + \Big(1 + \frac{\mathbf{y}^\top H\mathbf{y}}{\mathbf{y}^\top\boldsymbol{\sigma}}\Big)\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\Big)., \end{aligned} \tag{8.20}$$

where $H = H_{t-1}^{-1}, \boldsymbol{\sigma} = \mathbf{x}_t - \mathbf{x}_{t-1}, \mathbf{y} = \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})$.

We leave it as Exercise 52 (i) to prove that the denominator $\mathbf{y}^\top\boldsymbol{\sigma}$ appearing twice in the formula is positive, unless the function f is flat between the iterates $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$. And under $\mathbf{y}^\top\boldsymbol{\sigma} > 0$, the BFGS method has another nice property: if the previous matrix H is positive definite, then also the next matrix H' is positive definite; see Exercise 52 (ii). In this sense, the matrices H_t^{-1} behave like proper inverse Hessians.

The method is named after Broyden, Fletcher, Goldfarb and Shanno who all came up with it independently around 1970. Greenstadt's name is mostly forgotten.

Let's take a step back and see what we have achieved. Recall that our starting point was that Newton's method needs to compute and invert Hessian matrices in each iteration and therefore has in practice a cost of $O(d^3)$ per iteration. Did we improve over this?

First of all, any method in Greenstadt's family avoids the computation of Hessian matrices altogether. Only gradients are needed. In the BFGS method in particular, the cost per iteration drops to $O(d^2)$. Indeed, the computation of the update matrix $E^\star$ in Definition 8.5 reduces to matrix-vector multiplications and outer-product computations, all of which can be done in $O(d^2)$ time.

Newton and Quasi-Newton methods are often performed with *scaled steps*. This means that the iteration becomes

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t H_t^{-1}\nabla f(\mathbf{x}_t), \quad t \geq 1, \tag{8.21}$$

for some $\alpha_t \in \mathbb{R}_+$. This parameter can for example be chosen such that $f(\mathbf{x}_{t+1})$ is minimized (line search). Another approach is *backtracking line search* where we start with $\alpha_t = 1$, and as long as this does not lead to sufficient progress, we halve α_t. Line search ensures that the matrices H_t^{-1} in the BFGS method remain positive definite [Gol70].

As the Greenstadt update method just depends on the step $\boldsymbol{\sigma} = \mathbf{x}_t - \mathbf{x}_{t-1}$ but not on how it was obtained, the update works in exactly the same way as before even if scaled steps are being used.

8.4.5 The L-BFGS method

In high dimensions d, even an iteration cost of $O(d^2)$ as in the BFGS method may be prohibitive. In fact, already at the end of the 1970s, the first *limited memory* (and limited time) variants of the method have been proposed. Here we essentially follow Nocedal [Noc80]. The idea is to use only information from the previous m iterations, for some small value of m, and "forget" anything older. In order to describe the resulting L-BFGS method, we first rewrite the BFGS update formula in product form.

Observation 8.6. *With $E^\star$ as in Definition 8.5 and $H' = H + E^\star$, we have*

$$H' = \left(I - \frac{\boldsymbol{\sigma}\mathbf{y}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) H \left(I - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) + \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}. \tag{8.22}$$

To verify this, simply expand the product in the right-hand side and compare with (8.20).

We further observe that we do not need the actual matrix $H' = H_t^{-1}$ to perform the next Quasi-Newton step (8.6), but only the vector $H'\nabla f(\mathbf{x}_t)$. Here is the crucial insight.

Lemma 8.7. *Let H, H' as in Observation 8.6, i.e.*

$$H' = \left(I - \frac{\boldsymbol{\sigma}\mathbf{y}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) H \left(I - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) + \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}.$$

Let $\mathbf{g}' \in \mathbb{R}^d$. Suppose that we have an oracle to compute $\mathbf{s} = H\mathbf{g}$ for any vector $\mathbf{g}$. Then $\mathbf{s}' = H'\mathbf{g}'$ can be computed with one oracle call and $O(d)$ additional arithmetic operations, assuming that $\boldsymbol{\sigma}$ and $\mathbf{y}$ are known.

Proof. From (8.22), we conclude that

$$H'\mathbf{g}' = \underbrace{\underbrace{\left(I - \frac{\boldsymbol{\sigma}\mathbf{y}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) \underbrace{H \underbrace{\left(I - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\mathbf{g}'}_{\mathbf{g}}}_{\mathbf{s}}}_{\mathbf{w}} + \underbrace{\frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\mathbf{g}'}_{\mathbf{h}}}_{\mathbf{z}}.$$

We compute the vectors $\mathbf{h}, \mathbf{g}, \mathbf{s}, \mathbf{w}, \mathbf{z}$ in turn. We have

$$\mathbf{h} = \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\mathbf{g}' = \boldsymbol{\sigma}\frac{\boldsymbol{\sigma}^\top\mathbf{g}'}{\mathbf{y}^\top\boldsymbol{\sigma}},$$

so $\mathbf{h}$ can be computed with two inner products, a real division, and a multiplication of $\boldsymbol{\sigma}$ with a scalar. For $\mathbf{g}$, we obtain

$$\mathbf{g} = \left(I - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\mathbf{g}' = \mathbf{g}' - \mathbf{y}\frac{\boldsymbol{\sigma}^\top\mathbf{g}'}{\mathbf{y}^\top\boldsymbol{\sigma}}.$$

which is a multiplication of $\mathbf{y}$ with a scalar that we already know, followed by a vector addition. To get $\mathbf{s} = H\mathbf{g}$, we call the oracle. For $\mathbf{w}$, we similarly have

$$\mathbf{w} = \left(I - \frac{\boldsymbol{\sigma}\mathbf{y}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\mathbf{s} = \mathbf{s} - \boldsymbol{\sigma}\frac{\mathbf{y}^\top\mathbf{s}}{\mathbf{y}^\top\boldsymbol{\sigma}},$$

which is one inner product (the other one we already know), a real divison, a multiplication of $\boldsymbol{\sigma}$ with a scalar, and a vector addition. Finally,

$$H'\mathbf{g}' = \mathbf{z} = \mathbf{w} + \mathbf{h}$$

is a vector addition. In total, we needed three inner product computations, three scalar multiplications, three vector additions, two real divisions, and one oracle call. □

How do we implement the oracle? We simply apply the previous Lemma recursively. Let

$$\begin{aligned}\boldsymbol{\sigma}_k &= \mathbf{x}_k - \mathbf{x}_{k-1},\\ \mathbf{y}_k &= \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})\end{aligned}$$

be the values of $\boldsymbol{\sigma}$ and $\mathbf{y}$ in iteration $k \leq t$. When we perform the Quasi-Newton step $\mathbf{x}_{t+1} = \mathbf{x}_t - H_t^{-1}\nabla f(\mathbf{x}_t)$ in iteration $t \geq 1$, we have already computed these vectors for $k = 1, \ldots, t$. Using Lemma 8.7, we could therefore call the recursive procedure in Figure 8.2 with $k = t, \mathbf{g}' = \nabla f(\mathbf{x}_t)$ to compute the required vector $H_t^{-1}\nabla f(\mathbf{x}_t)$ in iteration t. To maintain the immediate connection to Lemma 8.7, we refrain from introducing extra variables for values that occur several times; but in an actual implementation, this would be done, of course.

function BFGS-STEP($k, \mathbf{g}'$) ▷ returns $H_k^{-1}\mathbf{g}'$
  if $k = 0$ **then**
    return $H_0^{-1}\mathbf{g}'$
  else ▷ apply Lemma 8.7
    $\mathbf{h} = \boldsymbol{\sigma}\dfrac{\boldsymbol{\sigma}_k^\top \mathbf{g}'}{\mathbf{y}_k^\top \boldsymbol{\sigma}_k}$
    $\mathbf{g} = \mathbf{g}' - \mathbf{y}\dfrac{\boldsymbol{\sigma}_k^\top \mathbf{g}'}{\mathbf{y}_k^\top \boldsymbol{\sigma}_k}$
    $\mathbf{s} =$ BFGS-STEP($k-1, \mathbf{g}$)
    $\mathbf{w} = \mathbf{s} - \boldsymbol{\sigma}_k\dfrac{\mathbf{y}_k^\top \mathbf{s}}{\mathbf{y}_k^\top \boldsymbol{\sigma}_k}$
    $\mathbf{z} = \mathbf{w} + \mathbf{h}$
    return $\mathbf{z}$
  end if
end function

Figure 8.2: Recursive view of the BFGS method. To compute $H_t^{-1}\nabla f(\mathbf{x}_t)$, call the function with arguments $(t, \nabla f(\mathbf{x}_t))$; values $\boldsymbol{\sigma}_k, \mathbf{y}_k$ from iterations $1, \ldots, t$ are assumed to be available.

By Lemma 8.7, the runtime of BFGS-STEP$(t, \nabla f(\mathbf{x}_t))$ is $O(td)$. For $t > d$, this is slower (and needs more memory) than the standard BFGS step according to Definition 8.5 which always takes $O(d^2)$ time.

The benefit of the recursive variant is that it can easily be adapted to a step that is *faster* (and needs *less* memory) than the standard BFGS step. The idea is to let the recursion bottom out after a fixed number m of recursive calls (in practice, values of $m \leq 10$ are not uncommon). The step then has runtime $O(md)$ which is a substantial saving over the standard step if m is much smaller than d.

The only remaining question is what we return when the recursion now bottoms out prematurely at $k = t - m$. As we don't know the matrix H_{t-m}^{-1}, we cannot return $H_{t-m}^{-1}\mathbf{g}'$ (which would be the correct output in this case). Instead, we pretend that we have started the whole method just now and use our initial matrix H_0 instead of H_{t-m}.[1] The resulting algorithm is depicted in Figure 8.3.

function L-BFGS-STEP($k, \ell, \mathbf{g}'$) ▷ $\ell \leq k$; returns $\mathbf{s}' \approx H_k^{-1}\mathbf{g}'$
  if $\ell = 0$ **then**
    return $H_0^{-1}\mathbf{g}'$
  else ▷ apply Lemma 8.7
    $\mathbf{h} = \boldsymbol{\sigma}\dfrac{\boldsymbol{\sigma}_k^\top \mathbf{g}'}{\mathbf{y}_k^\top \boldsymbol{\sigma}_k}$
    $\mathbf{g} = \mathbf{g}' - \mathbf{y}\dfrac{\boldsymbol{\sigma}_k^\top \mathbf{g}'}{\mathbf{y}_k^\top \boldsymbol{\sigma}_k}$
    $\mathbf{s} =$ L-BFGS-STEP($k-1, \ell-1, \mathbf{g}$)
    $\mathbf{w} = \mathbf{s} - \boldsymbol{\sigma}_k\dfrac{\mathbf{y}_k^\top \mathbf{s}}{\mathbf{y}_k^\top \boldsymbol{\sigma}_k}$
    $\mathbf{z} = \mathbf{w} + \mathbf{h}$
    return $\mathbf{z}$
  end if
end function

Figure 8.3: The L-BFGS method. To compute $H_t^{-1}\nabla f(\mathbf{x}_t)$ based on the previous m iterations, call the function with arguments $(t, m, \nabla f(\mathbf{x}_t))$; values $\boldsymbol{\sigma}_k, \mathbf{y}_k$ from iterations $t - m + 1, \ldots, t$ are assumed to be available.

Note that the L-BFGS method is still a Quasi-Newton method as long as $m \geq 1$: if we go through at least one update step of the form $H' = H + E$, the matrix H' will satisfy the secant condition by design, irrespective of H.

[1] In practice, we can do better: as we already have some information from previous steps, we can use this information to construct a more tuned H_0. We don't go into this here.

8.5 Exercises

Exercise 48. *Consider a step of the secant method:*

$$x_{t+1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}, \quad t \geq 1.$$

Assuming that $x_t \neq x_{t-1}$ and $f(x_t) \neq f(x_{t-1})$, prove that the line through the two points $(x_{t-1}, f(x_{t-1}))$ and $(x_t, f(x_t))$ intersects the x-axis at the point $x = x_{t+1}$.

Exercise 49. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function with nonzero Hessians everywhere. Prove that the following two statements are equivalent.*

(i) f is a nondegenerate quadratic function, meaning that

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top M\mathbf{x} - \mathbf{q}^\top\mathbf{x} + c,$$

where $M \in \mathbb{R}^{d\times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d, c \in R$ (see also Lemma 7.1).

(ii) Applied to f, Newton's update step

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t), \quad t \geq 1$$

defines a Quasi-Newton method for all $\mathbf{x}_0, \mathbf{x}_1 \in \mathbb{R}^d$.

Exercise 50. *Prove the direction (i)⇒(ii) of Theorem 8.1! You may want to do proceed in the following steps.*

1. *Prove the* Poor Man's Farkas Lemma*: a system of linear equations $A\mathbf{x} = \mathbf{b}$ in d variables has a solution if and only for all $\boldsymbol{\lambda} \in \mathbb{R}^d$, $\boldsymbol{\lambda}^\top A = \mathbf{0}^\top$ implies $\boldsymbol{\lambda}^\top\mathbf{b} = 0$. (You may use the fact that the row rank of a matrix equals its column rank.)*
2. *Argue that $\mathbf{x}^\star = \operatorname{argmin}\{\nabla f(\mathbf{x}^\star)^\top\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, C\mathbf{x} = \mathbf{e}\}$.*
3. *Apply the Poor Man's Farkas Lemma.*

Exercise 51. *Prove Fact 8.2!*

Exercise 52. *Consider the BFGS method (Definition 8.5).*

(i) *Prove that* $\mathbf{y}^\top \sigma > 0$*, unless* $\mathbf{x}_t = \mathbf{x}_{t-1}$*, or* $f(\lambda \mathbf{x}_t + (1-\lambda)\mathbf{x}_{t-1}) = \lambda f(\mathbf{x}_t) + (1-\lambda) f(\mathbf{x}_{t-1})$ *for all* $\lambda \in (0,1)$*.*

(ii) *Prove that if* H *is positive definite and* $\mathbf{y}^\top \sigma > 0$*, then also* H' *is positive definite. You may want to use the product form of the BFGS update as developed in Observation 8.6.*