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. 2.11 $\mathcal{O}(\log(1/\varepsilon))$
projected gradient descent	Thm. 3.2 $\mathcal{O}(1/\varepsilon^2)$	Thm. 3.4 $\mathcal{O}(1/\varepsilon)$		Thm. 3.5 $\mathcal{O}(\log(1/\varepsilon))$
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 (2.1) 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 (2.1). 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 (2.1), $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$, hence

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

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

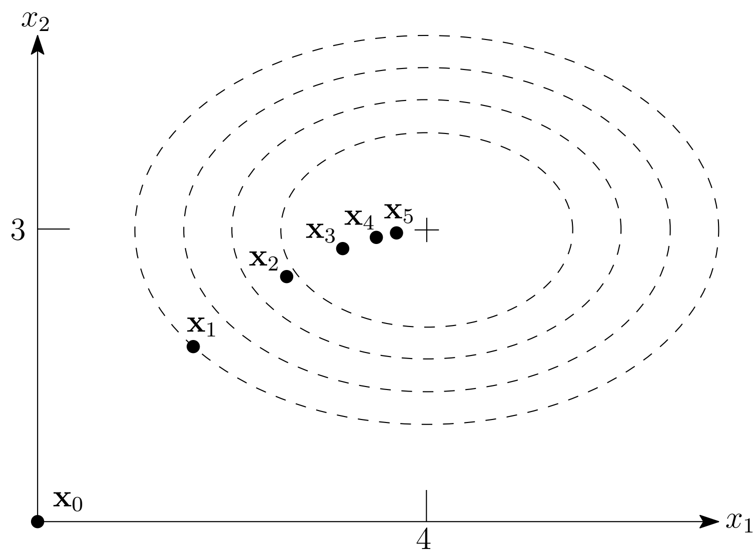

Figure 2.1: Example run of gradient descent on the quadratic function $f(x_1, x_2) = 2(x_1 - 4)^2 + 3(x_2 - 3)^2$ with global minimum $(4, 3)$; we have chosen $\mathbf{x}_0 = (0, 0), \gamma = 0.1$; dashed lines represent level sets of f (points of constant f-value)

cosine theorem) to rewrite the same expression as

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

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

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

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 (1.3) 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

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

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 13(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 11.)

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 (2.1) 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. $\square$

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 14 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.

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.

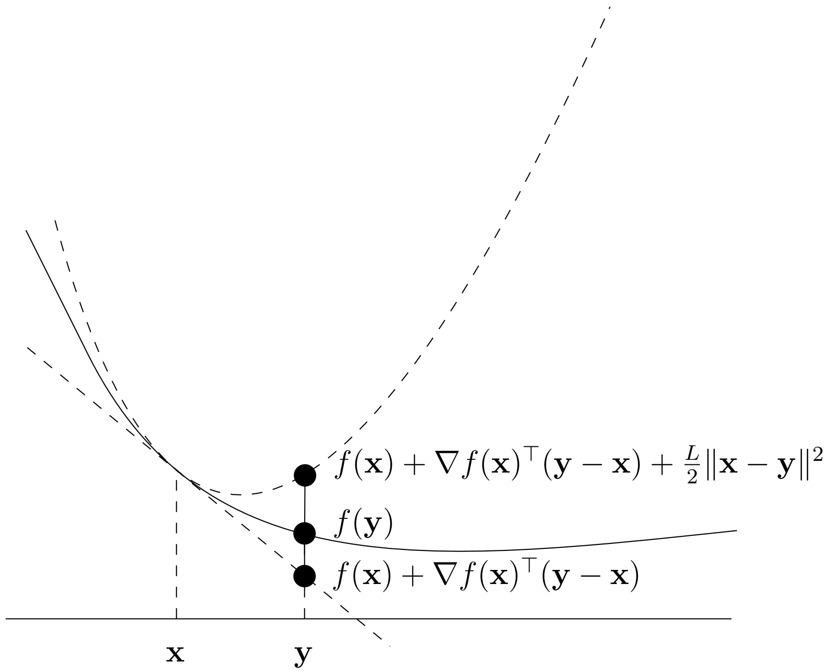

Figure 2.2: A smooth convex function

Lemma 2.3 (Exercise 12)**.** *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 17).

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 13(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 15)**.**

(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 (2.1) 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 (2.1) 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 16 shows that we do not need to know L to obtain the same asymptotic runtime.

2.6 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 (2.1) with stepsize $\gamma = 1/2$ satisfies

$$f(x_T) \leq \frac{1}{T}x_0^2. \tag{2.11}$$

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, (2.1) 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 (2.1) 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

$$f(x_T) = f\left(\frac{x_0}{2^T}\right) = \frac{1}{2^{2T}}x_0^2. \tag{2.12}$$

This is still vastly better than the bound of (2.11)! While (2.11) requires $T \approx x_0^2/\varepsilon$ to achieve $f(x_T) \leq \varepsilon$, (2.12) 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.

2.7 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 2.8. *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*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{2.13}$$

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

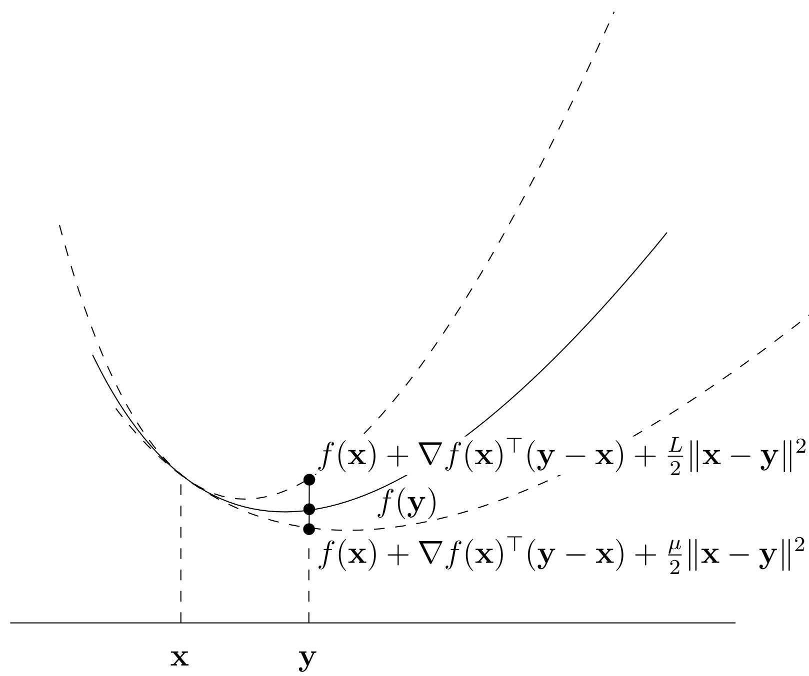

Figure 2.3: A smooth and strongly convex function

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 (2.13) 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 (2.13) with $\mu = 0$.

Lemma 2.9 (Exercise 18)**.** *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 12 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 2.10 (Exercise 19)**.** *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

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2. \tag{2.14}$$

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:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2. \tag{2.15}$$

Theorem 2.11. *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 21 asks you to prove that there is a unique global minimum $\mathbf{x}^\star$ of f. Choosing*

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

gradient descent (2.1) 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:

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

Proof. For (i), we show that the noise in (2.15) disappears. By sufficient decrease (Lemma 2.6), we know that

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

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

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

Hence, (2.15) actually yields

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

and

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

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

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

□

This implies that after

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

iterations, we reach absolute error at most ε.

2.8 Exercises

Exercise 11. *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*

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

Solution: This is clear if $\mathbf{c} = \mathbf{0}$. Otherwise, setting $\mathbf{x} = \mathbf{c}$, we have

$$\frac{|\mathbf{c}^\top \mathbf{x}|}{\|\mathbf{x}\|} = \frac{|\mathbf{c}^\top \mathbf{c}|}{\|\mathbf{c}\|} = \frac{\|\mathbf{c}\|^2}{\|\mathbf{c}\|} = \|\mathbf{c}\|,$$

so the spectral norm is at least the Euclidean norm. On the other hand, the Cauchy-Schwarz inequality yields that for all $\mathbf{x} \neq \mathbf{0}$,

$$\frac{|\mathbf{c}^\top \mathbf{x}|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{c}\| \, \|\mathbf{x}\|}{\|\mathbf{x}\|} = \|\mathbf{c}\|,$$

so the spectral norm is at most the Euclidean norm.

Exercise 12. *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 $2\|Q\|$.*

Solution: As the function $\mathbf{x} \mapsto \mathbf{b}^\top \mathbf{x} + c$ is affine and hence smooth with parameter 0, it suffices by Lemma 2.5 to restrict ourselves to the case $f(\mathbf{x}) := \mathbf{x}^\top Q\mathbf{x}$.

Because Q is symmetric, $\mathbf{x}^\top Q\mathbf{y} = \mathbf{y}^\top Q\mathbf{x}$ for any $\mathbf{x}$ and $\mathbf{y}$. Thus, a simple calculation shows that

$$\begin{aligned} f(\mathbf{y}) = \mathbf{y}^\top Q\mathbf{y} &= \mathbf{x}^\top Q\mathbf{x} + 2\mathbf{x}^\top Q(\mathbf{y}-\mathbf{x}) + (\mathbf{x}-\mathbf{y})^\top Q(\mathbf{x}-\mathbf{y}) \\ &= f(\mathbf{x}) + 2\mathbf{x}^\top Q(\mathbf{y}-\mathbf{x}) + (\mathbf{x}-\mathbf{y})^\top Q(\mathbf{x}-\mathbf{y}). \end{aligned}$$

Cauchy-Schwarz for $(\mathbf{x}-\mathbf{y})^\top Q(\mathbf{x}-\mathbf{y}) \leq \|\mathbf{x}-\mathbf{y}\| \, \|Q(\mathbf{x}-\mathbf{y})\|$, and using and the definition of spectral norm for $\|Q(\mathbf{x}-\mathbf{y})\| \leq \|Q\| \, \|\mathbf{x}-\mathbf{y}\|$ we get

$$f(\mathbf{y}) \leq f(\mathbf{x}) + 2\mathbf{x}^\top Q(\mathbf{y}-\mathbf{x}) + \|Q\| \, \|\mathbf{x}-\mathbf{y}\|^2,$$

Because $\|x-y\|^2$ vanishes as $(x-y)$ goes to 0, differentiability of f (Definition 1.7) implies that $\nabla f(\mathbf{x})^\top = 2\mathbf{x}^\top Q$, so we further get

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{2\|Q\|}{2} \|\mathbf{x}-\mathbf{y}\|^2,$$

That is, f is smooth with parameter $2\|Q\|$.

Exercise 13. *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 (2.1) 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.*

Solution:

(i) Since for all $x > 0$, $f(x) = x^{3/2}$, and for all $x < 0$, $f(x) = (-x)^{3/2}$, f is (infinitely) differentiable at every point $x \neq 0$. So we need to show that f is differentiable at the point $x = 0$. Indeed, by definition of derivative

$$f'(0) = \lim_{h \to 0} \frac{f(h) - f(0)}{h} = \lim_{h \to 0} \frac{|h|^{3/2}}{h} = \lim_{h \to 0} \text{sign}(h)|h|^{1/2} = 0.$$

To prove that f is strictly convex, we will at first show that the function $x^{3/2}$ (with domain $x > 0$) is strictly convex. Its second derivative $\frac{3}{4}x^{-1/2}$ is positive for all $x > 0$. If some function has positive second derivative at every point of its domain and the domain is open and convex, then this function is strictly convex (see the discussion after definition 1.18). Hence $x^{3/2}$ is strictly convex.

We need to show that

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \tag{2.16}$$

holds for all $x, y \in \mathbb{R}$ such that $x \neq y$ and for all $\lambda \in (0,1)$.

At first assume that both x and y are nonzero. Then $|x| > 0$, $|y| > 0$ and we get

$$\begin{aligned}
f(\lambda x + (1-\lambda)y) &= |\lambda x + (1-\lambda)y|^{3/2} \\
&\leq (|\lambda x| + |(1-\lambda)y|)^{3/2} \quad \text{(triangle inequality)} \\
&= (\lambda|x| + (1-\lambda)|y|)^{3/2} \\
&< \lambda|x|^{3/2} + (1-\lambda)|y|^{3/2} \quad \text{(strict convexity of } x^{3/2}\text{)} \\
&= \lambda f(x) + (1-\lambda)f(y).
\end{aligned}$$

It remains to show that (2.16) holds when $x = 0$ or $y = 0$. Without loss of generality, assume that $y = 0$. Then for all $x \neq 0$ and $\lambda \in (0, 1)$

$$f(\lambda x + (1-\lambda)y) = \lambda^{3/2}|x|^{3/2} < \lambda|x|^{3/2} = \lambda f(x) + (1-\lambda)f(y).$$

Since f is strictly convex and nonnegative, it has a unique global minimum $x^\star = 0$.

(ii) We need to find x_0 such that

$$|x_1|^{3/2} = |x_0 - \gamma f'(x_0)|^{3/2} > |x_0|^{3/2}.$$

We may assume that $x_0 > 0$. Then $f'(x_0) = \frac{3}{2}x_0^{1/2}$. We get

$$|x_1|^{3/2} = |x_0 - \frac{3}{2}\gamma x_0^{1/2}|^{3/2} = |x_0|^{3/2}|\frac{3}{2}\gamma x_0^{-1/2} - 1|^{3/2}.$$

If $0 < x_0 < \frac{9}{4}\gamma^2$, then $|x_1|^{3/2} > |x_0|^{3/2}$.

(iii) Suppose that f is smooth. Then by Theorem 2.7 there exists a stepsize in gradient descent (2.1) applied to f such that for all points x_0, $f(x_1) \leq f(x_0)$, which is a contradiction to point (ii).

(iv) Suppose that f is smooth with some parameter L. Since $X \neq \{0\}$, it contains some point $a \neq 0$. Then by convexity of X, the closed interval with endpoints a and 0 is a subset of X. Take $y \neq 0$ from this interval such that $|y| < \frac{4}{L^2}$ and $x = 0$. By definition of smoothness

$$f(y) = |y|^{3/2} \leq f(x) + f'(x)(y - x) + \frac{L}{2}|x - y|^2 = \frac{L}{2}|y|^2.$$

We get $|y|^{1/2} \geq \frac{2}{L}$, a contradiction to $|y| < \frac{4}{L^2}$.

Exercise 14. *In order to obtain average error at most ε in Theorem 2.1, we need to choose iteration number and step size 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!

Solution: The idea is to guess B. The first guess is $B = \varepsilon/R$; if this guess is correct, we can choose $T = 1$. Otherwise, we keep doubling B (which keeps quadrupling T), until the guess is correct (which must eventually happen if some global bound on the $\|\nabla f(\mathbf{x})\|$ exists). How can we check that a guess is correct? We can't, but the calculations show that in order to obtain error at most ε, we only need that $\|\nabla f(\mathbf{x}_t)\| \leq B$ for $t = 0, \ldots, T-1$, and this *can* be checked. It follows that the successful guess will not exceed the true B by more than a factor of two, so the number of iterations for the successful guess is at most

$$4\left(\frac{RB}{\varepsilon}\right)^2,$$

and the total number of iterations at most

$$\frac{16}{3}\left(\frac{RB}{\varepsilon}\right)^2,$$

using that $\sum_{i=0}^{k} 4^i = (4^{k+1} - 1)/3$.

Exercise 15. *Prove Lemma 2.5! (Operations which preserve smoothness)*

Solution: For (i), we sum up the weighted smoothness conditions for all the f_i to obtain

$$\sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) \leq \sum_{i=1}^{m} \lambda_i f_i(\mathbf{y}) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \sum_{i=1}^{m} \lambda_i \frac{L_i}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

As the gradient is a linear operator, this equivalently reads as

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\sum_{i=1}^{m} \lambda_i L_i}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$

and the statement follows. For (ii), we apply smoothness of f at $\mathbf{x}' = A\mathbf{x} + \mathbf{b}$ and $\mathbf{y}' = A\mathbf{y} + \mathbf{b}$ to obtain

$$f(A\mathbf{x} + \mathbf{b}) \leq f(A\mathbf{y} + \mathbf{b}) + \nabla f(A\mathbf{x} + \mathbf{b})^\top (A(\mathbf{y} - \mathbf{x})) + \frac{L}{2}\|A(\mathbf{x} - \mathbf{y})\|^2.$$

As $\nabla(f \circ g)(\mathbf{x})^\top = \nabla f(A\mathbf{x} + \mathbf{b})^\top A$ (chain rule (Lemma 1.8), using that $Dg(\mathbf{x}) = A$, an easy consequence of Definition 1.7). This equivalently reads as

$$(f \circ g)(\mathbf{x}) \leq (f \circ g)(\mathbf{y}) + \nabla(f \circ g)(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|A(\mathbf{x} - \mathbf{y})\|^2.$$

The statement now follows from $\|A(\mathbf{x} - \mathbf{y})\| \leq \|A\|\|\mathbf{x} - \mathbf{y}\|$.

Exercise 16. *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!

Solution: The idea is to guess L. The first guess is $L = 2\varepsilon/R^2$; if this guess is correct, we can choose $T = 1$. Otherwise, we keep doubling L (which keeps doubling T), until the guess is correct (which must eventually happen if some global smoothness parameter exists). How can we check that a guess is correct? We can't, but the calculations show that in order to obtain error at most ε, we only need that

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

and this *can* be checked. It follows that the successful guess will not exceed the true L by more than a factor of two, so the number of iterations for the successful guess is at most

$$2\frac{R^2 L}{2\varepsilon},$$

and the total number of iterations at most

$$4\frac{R^2L}{2\varepsilon},$$

using that $\sum_{i=0}^{k} 2^i = 2^{k+1} - 1$.

Exercise 17. *Let $a \in \mathbb{R}$. Prove that $f(x) = x^4$ is smooth over $X = (-a, a)$ and determine a concrete smoothness parameter L.*

Solution: The required inequality reads as

$$y^4 \leq x^4 + 4x^3(y-x) + \frac{L}{2}(x-y)^2 = -3x^4 + 4x^3y + \frac{L}{2}(x^2 - 2xy + y^2) =: r_y(x).$$

We therefore want to ensure that $r_y(x) \geq y^4$ for all $x, y \in (-a, a)$. This is the case if and only if

$$\min\{r_y(x) : x \in [-a, a]\} \geq y^4, \quad \forall y \in [-a, a].$$

To minimize $r_y(x)$, we compute derivatives and get

$$\begin{aligned} r_y'(x) &= -12x^3 + 12x^2y + Lx - Ly, \\ r_y''(x) &= -36x^2 + 24xy + L. \end{aligned}$$

We have $r_y'(y) = 0$. Moreover, for $L = 60a^2$, we get

$$r_y''(x) \geq -36a^2 - 24a^2 + L \geq 0,$$

so r_y is convex over $(-a, a)$ as a consequence of the second-order characterization Lemma 1.12. For $y \in (-a, a)$, $x = y$ is therefore indeed a minimum of r_y over $(-a, a)$ by Lemma 1.16. As we have

$$r_y(y) = y^4,$$

smoothness follows with $L = 60a^2$.

Exercise 18. *Prove Lemma 2.9! (Strongly convex functions have unique global minimum)*

Solution: Let $\mathbf{x} \neq \mathbf{y}, \lambda \in (0, 1)$ and $z = \lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$. As $\mathbf{z} \neq \mathbf{x}, \mathbf{y}$, strong convexity (2.13) yields

$$\begin{array}{rcl} f(\mathbf{x}) & > & f(\mathbf{z}) + \nabla f(\mathbf{z})^\top(\mathbf{x} - \mathbf{z}) = f(\mathbf{z}) + (1 - \lambda)\nabla f(\mathbf{z})^\top(\mathbf{x} - \mathbf{y}), \\ f(\mathbf{y}) & > & f(\mathbf{z}) + \nabla f(\mathbf{z})^\top(\mathbf{y} - \mathbf{z}) = f(\mathbf{z}) + \lambda\nabla f(\mathbf{z})^\top(\mathbf{y} - \mathbf{x}). \end{array}$$

Adding up these two inequalities with multiples λ and $1 - \lambda$, respectively, the gradient terms cancel, and we get

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

This is strict convexity according to Definition 1.18. To prove that there is a global minimum, we show that sublevel sets are bounded and then apply the Weierstrass Theorem 1.24. Let $\mathbf{y} \in f^{\leq\alpha}$ for some $\alpha \in \mathbb{R}$. By strong convexity, we then have

$$\alpha \geq f(\mathbf{y}) \geq f(\mathbf{0}) - \nabla f(\mathbf{0})^\top(-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{y}\|^2 \geq f(\mathbf{0}) - \|\nabla f(\mathbf{0})\|\|\mathbf{y}\| + \frac{\mu}{2}\|\mathbf{y}\|^2,$$

using the Cauchy-Schwarz inequality ($\mathbf{v}^T\mathbf{w} \leq \|v\|\|w\|$). Hence,

$$\|\mathbf{y}\|\left(\frac{\mu}{2}\|\mathbf{y}\| - \|\nabla f(\mathbf{0})\|\right) \leq \alpha - f(\mathbf{0}),$$

which implies that $\|\mathbf{y}\|$ is bounded.

Exercise 19. *Prove Lemma 2.10! (Strongly convex and smooth functions)*

Solution: If the parameters of smoothness and strong convexity are both μ, the two inequalities (2.8) and (2.13) enforce the equality

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

We also know from Lemma 2.9 that there is a (unique) global minimum $\mathbf{x}^\star$ which satisfies $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ by Lemma 1.17. Hence,

$$f(\mathbf{y}) = f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{y}\|^2, \quad \forall \mathbf{y} \in \mathbb{R}^d,$$

and the statement follows with $\mathbf{b} = \mathbf{x}^\star$ and $c = f(\mathbf{x}^\star)$.