Chapter 2

Gradient Descent

Contents

2.1 The algorithm

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

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

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

Gradient descent is a very simple iterative algorithm for finding the desired approximation $\mathbf{x}$, under suitable conditions that we will get to. Gradient descent computes a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of vectors such that $\mathbf{x}_0$ is arbitrary, and

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), \quad t \geq 0. \tag{2.1}$$

Here, γ is a fixed *stepsize*, but it may also make sense to have γ depend on t. For now, γ is fixed. As the vector $-\nabla f(\mathbf{x}_t)$ points into a direction of descent of f at $\mathbf{x}_t$, the idea is to move a little bit into this direction and then iterate. We hope that after not too many iterations t, $f(\mathbf{x}_t) - f(\mathbf{x}^\star) < \varepsilon$; see Figure 2.1 for an example.

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.

2.2 Vanilla analysis

Let $\mathbf{x}_t$ be some iterate in the sequence (2.1). We do have an inequality that bounds $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$, namely the one saying that the graph of f lies above all its tangent hyperplanes; indeed, applying (1.2) with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$ and reshuffling terms, we obtain

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star). \tag{2.2}$$

By definition of gradient descent (2.1), $\nabla f(\mathbf{x}_t) = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$, hence

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star). \tag{2.3}$$

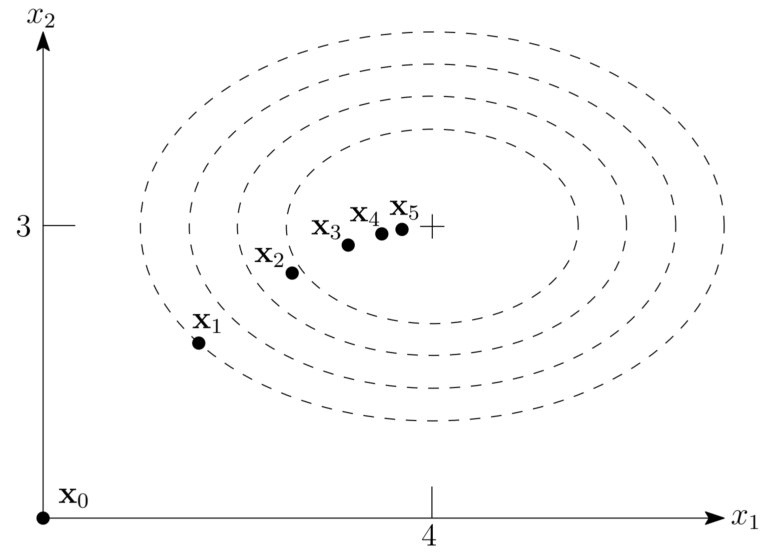


Figure 2.1: Example run of gradient descent on the quadratic function $f(x_1, x_2) = 2(x_1 - 4)^2 + 3(x_2 - 3)^2$ with global minimum $(4, 3)$; we have chosen $\mathbf{x}_0 = (0, 0), \gamma = 0.1$; dashed lines represent level sets of f (points of constant f-value)

Now we apply (somewhat out of the blue, but this will clear up in the next step) the basic vector equation $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$ to obtain

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^\star) &\leq \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \\ &= \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \quad (2.4) \end{aligned}$$

again by using the definition (2.1) of gradient descent. Next we sum this up over some initial values of t, so that the latter two terms in the bracket cancel in a telescoping sum.

$$\begin{aligned} \sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) &\leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2\right) \\ &\leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \quad (2.5) \end{aligned}$$

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 11(i).

The question is of course: is this bound 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 is more of an issue, and if we cannot control them, we cannot say much.

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

Here is the cheapest "solution" to squeeze something out of the vanilla analysis: let us simply assume that all gradients of f are bounded in norm. This rules out many interesting functions, though, since functions with bounded gradients only have at most linear growth. Equivalently, such functions are Lipschitz continuous over $\mathbb{R}^d$. But 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\nabla f(\mathbf{x})\| \leq L$ for all $\mathbf{x}$. Choosing the stepsize*

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

gradient descent (2.1) yields

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

Proof. This is a simple calculation on top of (2.5): after plugging in the bounds R and L, we want to choose γ such that

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

is minimized. Setting the derivative to zero yields the above value of γ, and $q(R/(L\sqrt{T})) = RL\sqrt{T}$. 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 $\mathcal{O}(1/\varepsilon^2)$ many iterations, considering R and L as constants. This is not particularly good when it comes to concrete numbers (think of desired error $\varepsilon = 10^{-6}$ when R, L 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 L may depend on d, but in many relevant cases, this dependence is mild.

What happens if we don't know R and/or L? An idea is to "guess" R and L, 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 12 asks you to show that knowing R is sufficient.

We conclude this section by remarking that bounded gradients are actually equivalent to *Lipschitz continuity* of f.

Lemma 2.2 (Exercise 13). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, $L \in \mathbb{R}_+$. Then the following two statements are equivalent.*

(i) $\|\nabla f(\mathbf{x})\| \leq L$ *for all* $\mathbf{x} \in \mathbb{R}^d$.

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

2.4 Smoothness: $\mathcal{O}(1/\varepsilon)$ steps

Our workhorse in the vanilla analysis was the first-order characterization of convexity: for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) \tag{2.6}$$

Next we want to require that f is not "too convex", intuitively meaning that the curvature of the bowl is bounded.

Definition 2.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $L \in \mathbb{R}_+$. f is called* smooth *(with parameter L) if*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{2.7}$$

Recall that (2.6) says that for any $\mathbf{x}$, the graph of f is above its tangential hyperplane at $(\mathbf{x}, f(\mathbf{x}))$. In contrast, (2.7) says that for any $\mathbf{x}$, the graph of f is below a not-too-steep tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$; see Figure 2.2.

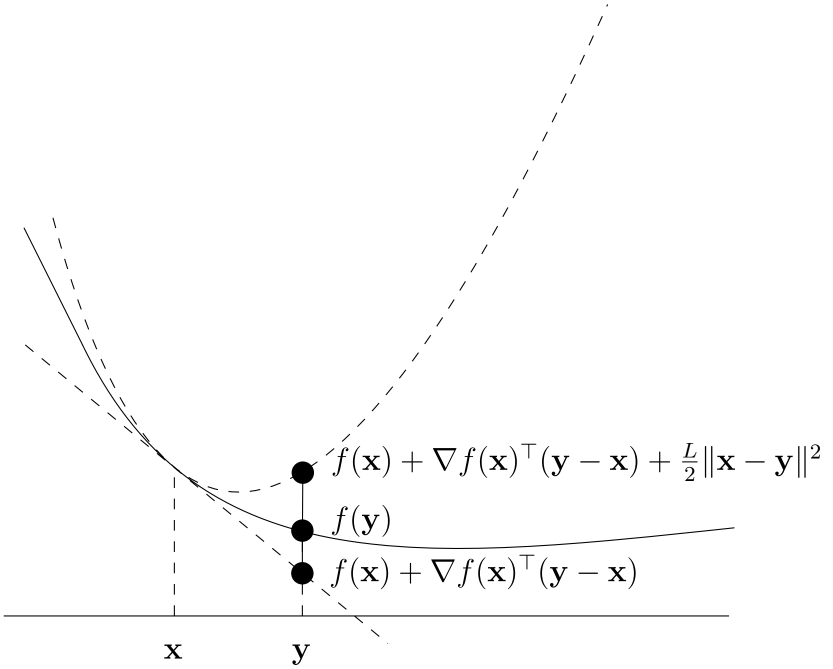

$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$
$f(\mathbf{y})$
$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$
$\mathbf{x}$
$\mathbf{y}$

Figure 2.2: A smooth convex function

Let us discuss some cases. If $L = 0$, (2.6) and (2.7) together require that

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

meaning that f is an affine function. A simple calculation shows that our supermodel function $f(x) = x^2$ is smooth with parameter $L = 2$, and the same holds for its d-dimensional generalization $f(\mathbf{x}) = \|\mathbf{x}\|^2$ (Exercise 10). The (univariate) convex function $f(x) = x^4$ is not smooth: at $x = 0$, condition (2.7) reads as

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

and there is obviously no L that works for all y. 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 11(iii) provides a counterexample.

The operations that we have shown to preserve convexity in Lemma 1.9 also preserve smoothness. This immediately gives us a rich collection of smooth functions.

Lemma 2.4 (Exercise 14)**.**

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

(ii) Let f be a convex function with $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ that is 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 convex function $f \circ g$ (that maps $\mathbf{x}$ to $f(A\mathbf{x} + \mathbf{b})$) is smooth with parameter $L\|A\|^2$, where

$$\|A\| = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

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

Corollary 2.5. *Let $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2$ be a least squares objective. Then f is smooth with parameter $L = 2\|A\|^2$.*

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

Theorem 2.6. *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 (2.7). Choosing*

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

gradient descent (2.1) with arbitrary $\mathbf{x}_0$ satisfies the following two properties.

(i) Function values are monotone decreasing:

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

(ii)

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

Proof. For (i), we directly apply the smoothness condition (2.7) and the definition of gradient descent that yields $\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$. We compute

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

In particular, this lets us now bound the sum of squared gradients after step (2.5) of the vanilla analysis:

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \leq \sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_T). \tag{2.8}$$

With $\gamma = 1/L$, (2.5) 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

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

Hence, by (i),

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

□

This improves over the bounds of Theorem 2.1. Again assuming that L and $\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$ are constant, we now only need $\mathcal{O}(1/\varepsilon)$ iterations instead of $\mathcal{O}(1/\varepsilon^2)$ to achieve $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \varepsilon$. Exercise 15 shows that we do not need to know L to obtain the same asymptotic runtime.

While bounded gradients are equivalent to Lipschitz continuity of f, smoothness turns out to be equivalent to Lipschitz continuity of ∇f.

Lemma 2.7. *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* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

A proof can for example be found in the lecture slides of L. Vandenberghe, `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf`.

2.5 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 2.6, gradient descent (2.1) with stepsize $\gamma = 1/2$ satisfies

$$f(x_t) \leq \frac{1}{t} x_0^2. \tag{2.10}$$

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 haven't 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

$$f(x_t) = f\left(\frac{x_0}{2^t}\right) = \frac{1}{2^{2t}} x_0^2. \tag{2.11}$$

This is still vastly better than the bound of (2.10)! While (2.10) requires $t \approx x_0^2/\varepsilon$ to achieve $f(x_t) \leq \varepsilon$, (2.11) only requires

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

which is an exponential improvement in the number of steps.

2.6 Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ 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 $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mu \in \mathbb{R}_+, \mu > 0$. f is called* strongly convex *(with parameter μ) if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{2.12}$$

While smoothness according to (2.7) says that for any $\mathbf{x}$, 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.

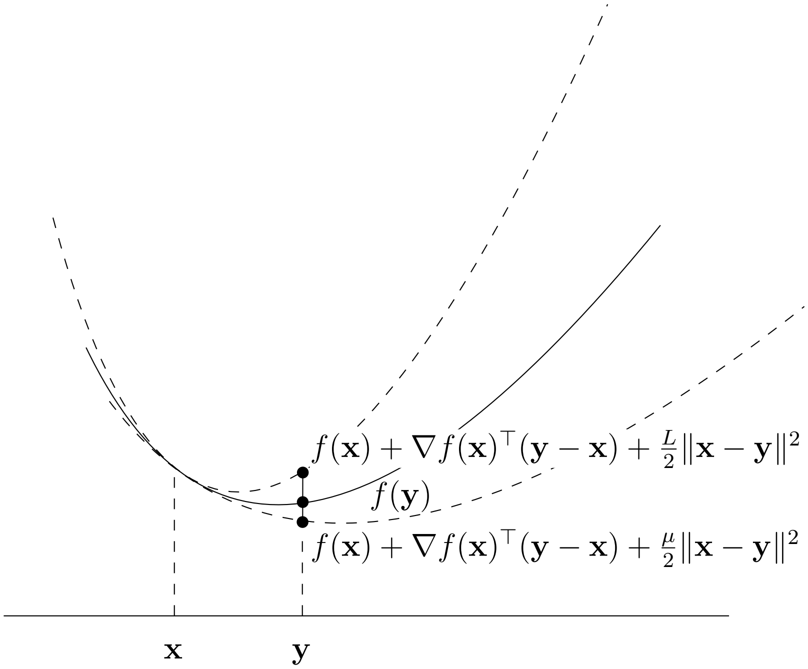
We can also interpret (2.12) as a strengthening of the first-order characterization of convexity. In the form of (2.6) this reads as

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

and therefore says that every convex function satisfies (2.12) with $\mu = 0$.

Lemma 2.9 (Exercise 17)**.** *If f 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 10 again 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.



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

Figure 2.3: A smooth and strongly convex function

Lemma 2.10 (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. From the vanilla analysis, we already have an inequality that potentially allows us to get started on this, namely (2.4) that we derived from the first-order characterization:

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

If f is strongly convex, we can start from the strengthening (2.12) instead to obtain

$$f(\mathbf{x}_t)-f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2. \tag{2.13}$$

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:

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

Theorem 2.11. *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 (2.7) and strongly convex with parameter $\mu > 0$ according to (2.12). 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)

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

Proof. For (i), we show that the noise in (2.14) disappears. From Theorem 2.6 (i), 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 $\gamma = 1/L$:

$$\begin{aligned} 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 &= \frac{2}{L}(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 \\ &\leq -\frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 = 0. \end{aligned}$$

Hence, (2.14) actually yields

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

The bound in (ii) follows from smoothness (2.7), using $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ (Lemma 1.13):

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

□

This implies that in order to reach absolute error at most ε, we only need $\mathcal{O}(\log \frac{1}{\varepsilon})$ iterations, where the constant behind the big-$\mathcal{O}$ is roughly L/μ.

2.7 Exercises

Exercise 10. *Prove that $f(\mathbf{x}) = \|\mathbf{x}\|^2$ is smooth with parameter $L = 2$.*

Solution: Since $\nabla f(\mathbf{x}) = 2\mathbf{x}$, for all $\mathbf{y}, \mathbf{x} \in \mathbb{R}^d$ we get

$$\begin{aligned} f(\mathbf{y}) = \|\mathbf{y}\|^2 &= \|\mathbf{x} - \mathbf{y}\|^2 + 2\mathbf{x}^\top \mathbf{y} - \|\mathbf{x}\|^2 \\ &= \|\mathbf{x}\|^2 - 2\mathbf{x}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{y} + \|\mathbf{x} - \mathbf{y}\|^2 \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + 2 \cdot \frac{1}{2} \cdot \|\mathbf{x} - \mathbf{y}\|^2 \end{aligned}$$

By definition of smoothness, f is smooth with parameter $L = 2$.

Exercise 11. *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

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

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 1.14). Hence $x^{3/2}$ is strictly convex.

We need to show that

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \tag{2.15}$$

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 (2.15) 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 2.6 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 12. *In order to obtain average error at most ε in Theorem 2.1, we need to choose iteration number and step size as*

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

If R or L are 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

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

many gradient descent steps!

Solution: The idea is to guess L. The first guess is $L = \varepsilon/R$; if this guess is correct, we can choose $T = 1$. Otherwise, we keep doubling L (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 $\|\nabla f(\mathbf{x}_t)\| \leq L$ for $t = 0, \ldots, T-1$, 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

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

and the total number of iterations at most

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

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

Exercise 13. *Prove Lemma 2.2! (Lipschitz continuity and bounded gradients)*

Solution: Suppose that (i) holds. By Taylor's theorem,

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

for some $\mathbf{z}$ on the line segment between $\mathbf{x}$ and $\mathbf{y}$. By the Cauchy-Schwarz inequality,

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \|\nabla f(\mathbf{z})\|\|(\mathbf{x} - \mathbf{y})\| \leq L\|(\mathbf{x} - \mathbf{y})\|,$$

and since the same argument can be made for $f(\mathbf{y}) - f(\mathbf{x})$, (ii) follows. Suppose now that (ii) holds. It's time to recall that f is differentiable if for every $\mathbf{x}$,

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

where $r_{\mathbf{x}}$ is an error function satisfying $r(\mathbf{v})/\|\mathbf{v}\| \to 0$ when $\|\mathbf{v}\| \to 0$. In particular, for $h \in \mathbb{R}_+$,

$$f(\mathbf{x} + h\nabla f(\mathbf{x})) - f(\mathbf{x}) = h\|\nabla f(\mathbf{x})\|^2 + r_{\mathbf{x}}(h\nabla f(\mathbf{x})) \leq Lh\|\nabla f(\mathbf{x})\|,$$

where the inequality uses (ii). This yields

$$\|\nabla f(\mathbf{x})\| + \frac{r_{\mathbf{x}}(h\nabla f(\mathbf{x}))}{\|h\nabla f(\mathbf{x})\|} \leq L,$$

and as the error term tends to zero as $h \to 0$, (i) follows.

Exercise 14. *Prove Lemma 2.4! (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 of multivariate calculus), 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 15. *In order to obtain average error at most ε in Theorem 2.6, 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

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

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

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

Solution: The required inequality reads as

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

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

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

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

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

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 1.8. For $y \in (-a, a)$, $x = y$ is therefore indeed a minimum of r_y over $(-a, a)$ by Lemma 1.12. As we have

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

smoothness follows with $L = 60a^2$ (the required inequality at the boundaries $y = a, -a$ holds by continuity).

Exercise 17. *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 (2.12) yields

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

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 1.14. To prove that there is a global minimum, we show that sublevel sets are bounded and then apply the Weierstrass Theorem 1.19. Suppose $f(\mathbf{y}) \leq \alpha$. 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 18. *Prove Lemma 2.10! (Strongly convex and smooth functions)*

Solution: If the parameters of smoothness and strong convexity are both μ, the two inequalities (2.7) and (2.12) 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 1.13. 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 $c = f(\mathbf{x}^\star)$.

Bibliography

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

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

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

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 07 2008.

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

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

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

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

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

[Vis14] Nisheeth Vishnoi. Lecture notes on fundamentals of convex optimization, 2014. https://tcs.epfl.ch/files/content/sites/tcs/files/Lec3-Fall14-Web.pdf.

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