

## Problem Set 2 — Solutions (Gradient Descent)

### Gradient Descent

**Exercise 14.** Prove Lemma 2.4: The quadratic function  $f(\mathbf{x}) = \mathbf{x}^\top Q \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$ ,  $Q$  symmetric, 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 2.6 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  $\|\mathbf{x} - \mathbf{y}\|^2$  vanishes as  $(\mathbf{x} - \mathbf{y})$  goes to 0, differentiability of  $f$  (Definition 1.5) 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 17.** Prove Lemma 2.6! (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 1.7), using that  $Dg(\mathbf{x}) = A$ , an easy consequence of Definition 1.5). 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 18.** In order to obtain average error at most  $\varepsilon$  in Theorem 2.8, we need to choose

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

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

Now suppose that we know  $R$  but not  $L$ . This means, we know a concrete number  $R$  such that  $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$ ; we also know that there exists a number  $L$  such that  $f$  is smooth with parameter  $L$ , but we don't know a concrete such number.

Develop an algorithm that—not knowing  $L$ —finds a vector  $\mathbf{x}$  such that  $f(\mathbf{x}) - f(\mathbf{x}^*) < \varepsilon$ , 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  $\varepsilon$ , 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 19.** 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

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

Now, if we choose a value of  $L$  for which  $r_y(x)$  is convex on  $(-a, a)$ , the minimum is given by  $r'_y(x) = 0$ . There are multiple choices for  $L$  for which this works out, but here we try  $L = 60a^2$ : For  $L = 60a^2$ , we get

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

on  $(-a, a)$ , so the function is convex on this interval as a consequence of Lemma 1.18. Because  $r'_y(y) = 0$ ,  $x = y$  is therefore a minimum of  $r_y$  over  $(-a, a)$  by Lemma 1.22. As we have

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

smoothness follows with  $L = 60a^2$ . (Note: this constant is not necessarily tight.)