Lecture 4

Iterative algorithms for solving a problem of the form

$$\text{minimize} \quad f(\boldsymbol{x}) \tag{4.1}$$

on $\mathbb{R}^d$ generate a sequence of vectors $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ in the hope that this sequence converges to a (local or global) minimizer $\boldsymbol{x}^*$ of (4.1). In this lecture we study what it means for a sequence to converge, and how to quantify the speed of convergence. The concepts are applied to gradient descent.

4.1 Convergence of iterative methods

A sequence of vectors $\{\boldsymbol{x}_k\}$ in $\mathbb{R}^d$, $k \geq 0$, *converges* to a vector $\boldsymbol{x}^*$ with respect to a norm $\|\cdot\|$ as $k \to \infty$, written $\boldsymbol{x}_k \to \boldsymbol{x}$, if the sequence of numbers $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ converges to zero. More formally, if for every $\varepsilon > 0$ there exists an index N such that for all $n \geq N$,

$$\|\boldsymbol{x}_n - \boldsymbol{x}^*\| < \varepsilon.$$

Iterative algorithms will rarely find the exact solution to a problem like (4.1), so we will usually be happy to find a solution that differs from the true one by at most some specified accuracy.

Definition 4.1. A sequence of vectors $\{\boldsymbol{x}_k\}$, $k \geq 0$, is said to converge to $\boldsymbol{x}^*$

(a) linearly (or Q-linear, Q for quotient), if there exist an $r \in (0, 1)$ such that for sufficiently large k,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq r\|\boldsymbol{x}_k - \boldsymbol{x}^*\|.$$

(b) superlinearly, if

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} = 0,$$

(c) with order p, if there exists a constant $M > 0$, such that for sufficiently large k,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq M\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p.$$

The case $p = 2$ is called *quadratic convergence*.

Of course, as mentioned earlier, these definitions depend on the choice of a norm. It can be shown that quadratic convergence implies superlinear convergence, and superlinear convergence implies linear convergence.

Example 4.2. Consider the sequence of numbers $x_k = 1/2^{r^k}$ for some $r > 1$. Clearly, $x_k \to x^* = 0$ as $k \to \infty$. Moreover,

$$x_{k+1} = \frac{1}{2^{r^{k+1}}} = \frac{1}{2^{r^k r}} = \left(\frac{1}{2^{r^k}}\right)^r = x_k^r,$$

which shows that the sequence has rate of convergence r.

4.2 Convergence of gradient descent

In this section, a norm $\|\cdot\|$ will refere to the 2-norm, unless otherwise stated. We now study the convergence of gradient descent for the least squares problem

$$\text{minimize} \quad f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2, \tag{4.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ of full rank. As we have seen in Lecture 3, the gradient descent method is the procedure

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k,$$

where the step length and the *residual* are given by

$$\alpha_k = \frac{\|\boldsymbol{r}_k\|^2}{\|\boldsymbol{A}\boldsymbol{r}_k\|^2}, \quad \boldsymbol{r}_k = \boldsymbol{A}^\mathsf{T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k) = -\nabla f(\boldsymbol{x}_k).$$

At the minimizer, the residual is

$$\boldsymbol{r} = -\nabla f(\boldsymbol{x}^*) = \boldsymbol{A}^\mathsf{T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*) = 0, \tag{4.2}$$

and as the sequence $\boldsymbol{x}_k$ converges to $\boldsymbol{x}^*$, the norms of the residuals converge to 0. Conversely, the residual is related to the difference $\boldsymbol{x}_k - \boldsymbol{x}$ by

$$\boldsymbol{r}_k = \boldsymbol{A}^\mathsf{T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k) = \boldsymbol{A}^\mathsf{T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^*)) = \boldsymbol{A}^\mathsf{T}\boldsymbol{A}(\boldsymbol{x_k} - \boldsymbol{x}^*), \tag{4.3}$$

where we used the "intelligent zero" (4.2). Therefore

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| = \|(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{r}\| \leq \|(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\|\|\boldsymbol{r}_k\|,$$

where $\|\boldsymbol{B}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{B}\boldsymbol{x}\|/\|\boldsymbol{x}\|$ is the operator norm of a matrix $\boldsymbol{B}$ with respect to the 2-norm. Consequently, if the sequence $\|\boldsymbol{r}_k\|$ converges to zero, so does the sequence $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. A reasonable criterium is to stop the algorithm is therefore when the residual norm $\|\boldsymbol{r}_k\|$ is below a predefined tolerance ε.

For the following, we use the norm

$$\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 := \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle = \|\boldsymbol{A}\boldsymbol{x}\|^2.$$

We have the bounds

$$\|\boldsymbol{A}^\dagger\|^{-1}\|\boldsymbol{x}\| \leq \|\boldsymbol{x}\|_{\boldsymbol{A}} \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|,$$

where $\boldsymbol{A}^\dagger = (\boldsymbol{A}^\intercal\boldsymbol{A})^{-1}\boldsymbol{A}^\intercal$ is the *Moore-Penrose pseudoinverse* of $\boldsymbol{A}$. The important consequence of these inequalities is that convergence with respect to the $\|\cdot\|_{\boldsymbol{A}}$ norm is equivalent to convergence with respect to the 2-norm.

A simple but nevertheless very useful observation is the following.

Lemma 4.1. *The difference between the function value at some $\boldsymbol{x} \in \mathbb{R}^d$ and the optimal value is*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\boldsymbol{A}}^2.$$

Proof. This proceeds by straight-forward calculation:

$$\begin{aligned}
f(\boldsymbol{x}^*) - f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\boldsymbol{A}}^2 &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}^* - b\|^2 - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*)\|^2 \\
&= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|^2 - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})\|^2 \\
&= \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|^2 - \langle \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}, (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\rangle \\
&= \langle \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\rangle \\
&= \langle \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}, \boldsymbol{A}(\boldsymbol{x}^* - \boldsymbol{x})\rangle \\
&= \langle \boldsymbol{A}^\intercal(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}), \boldsymbol{x}^* - \boldsymbol{x}\rangle = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^* - \boldsymbol{x}\rangle = 0,
\end{aligned}$$

where the last equality follows from (4.2). □

We next give a bound on the number of iterations that are guaranteed to bring the residual below a certain value. We then state (without proof) a more precise convergence bound in terms of the condition number, that shows that gradient descent for linear least squares has linear convergence.

Theorem 4.2. *Let $\varepsilon > 0$ be given and assume*

$$N > \frac{2\|\boldsymbol{A}\|^2}{\varepsilon^2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\boldsymbol{A}}^2$$

for a starting point $\boldsymbol{x}_0$, where $\boldsymbol{x}^*$ is the minimizer of $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$. If $\{\boldsymbol{x}_k\}_{k\geq 0}$ is the sequence of vectors generated by gradient descent for this function, then

$$\min_{0\leq k\leq N} \|\boldsymbol{r}_k\| < \varepsilon.$$

We first derive a bound on the decrease of the function value at each step.

Lemma 4.3. *For the function f as in* (4.1)*,*

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) = \frac{1}{2}\frac{\|\boldsymbol{r}_k\|^4}{\|\boldsymbol{A}\boldsymbol{r}_k\|^2} \geq \frac{1}{2}\frac{\|\boldsymbol{r}_k\|^2}{\|\boldsymbol{A}\|^2},$$

where $\|\boldsymbol{A}\| := \max_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|$ is the operator norm *of $\boldsymbol{A}$ with respect to the* 2*-norm.*

Proof. Using the identity $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$ and the bilinearity

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{x}\rangle + 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \boldsymbol{y}\rangle = \|\boldsymbol{x}\|^2 + 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \|\boldsymbol{y}\|^2,$$

we compute

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k) &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}_k + \alpha_k \boldsymbol{A}\boldsymbol{r}_k - \boldsymbol{b}\|^2 \\ &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}\|^2 + \langle \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}, \alpha_k \boldsymbol{A}\boldsymbol{r}_k\rangle + \frac{1}{2}\alpha_k^2\|\boldsymbol{A}\boldsymbol{r}_k\|^2 \\ &= \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}\|^2 + \alpha_k\langle \boldsymbol{A}^\intercal(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}), \boldsymbol{r}_k\rangle + \frac{1}{2}\alpha_k^2\|\boldsymbol{A}\boldsymbol{r}_k\|^2 \\ &= f(\boldsymbol{x}_k) - \alpha_k\langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle + \frac{1}{2}\alpha_k^2\|\boldsymbol{A}\boldsymbol{r}_k\|^2. \end{aligned} \tag{4.4}$$

Substituting $\alpha_k = \|\boldsymbol{r}_k\|^2/\|\boldsymbol{A}\boldsymbol{r}_k\|^2$, we get

$$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k) - \frac{\|\boldsymbol{r}_k\|^4}{\|\boldsymbol{A}\boldsymbol{r}_k\|^2} + \frac{1}{2}\frac{\|\boldsymbol{r}_k\|^4}{\|\boldsymbol{A}\boldsymbol{r}_k\|^2} = f(\boldsymbol{x}_k) - \frac{1}{2}\frac{\|\boldsymbol{r}_k\|^4}{\|\boldsymbol{A}\boldsymbol{r}_k\|^2}.$$

Since

$$\frac{\|\boldsymbol{A}\boldsymbol{r}_k\|}{\|\boldsymbol{r}_k\|} \leq \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{A}\|,$$

we get $\|\boldsymbol{A}\boldsymbol{r}_k\| \leq \|\boldsymbol{A}\|\|\boldsymbol{r}_k\|$, and applying this to (4.4), the inequality

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2}\frac{\|\boldsymbol{r}_k\|^2}{\|\boldsymbol{A}\|^2}.$$

Rearranging the gives the claim. □

Proof of Theorem 4.2. Let

$$g_N := \min_{0\leq k\leq N} \|\boldsymbol{r}_k\|.$$

Then since $(N+1)g_N^2 \leq \sum_{k=0}^N \|\boldsymbol{r}_k\|^2$, and using Lemma 4.3,

$$g_N \leq \sqrt{\frac{1}{N+1}\sum_{k=0}^{N}\|\boldsymbol{r}_k\|^2} \leq \frac{\sqrt{2}\|\boldsymbol{A}\|}{\sqrt{N+1}}\sqrt{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)} = \frac{\sqrt{2}\|\boldsymbol{A}\|}{\sqrt{N+1}}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\boldsymbol{A}}.$$

We know that $g_N < \varepsilon$ if $\sqrt{2\|\boldsymbol{A}\|^2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\boldsymbol{A}}/\sqrt{N+1} < \varepsilon$. Solving the latter for N, we get that

$$N > \frac{2\|\boldsymbol{A}\|^2}{\varepsilon^2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\boldsymbol{A}}^2 - 1.$$

will guarantee a residual smaller than ε. □

The following theorem shows that the gradient descent method for linear least squares converges linearly with respect to the $\boldsymbol{A}$ norm. The statement involves the *condition number* of $\boldsymbol{A}$[1]. This quantity is defined as

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^\dagger\|.$$

Theorem 4.4. *The error in the $k+1$-th iterate is bounded by*

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq \left(\frac{\kappa^2(\boldsymbol{A}) - 1}{\kappa^2(\boldsymbol{A}) + 1}\right)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|.$$

In particular, the gradient descent algorithm converges linearly. This Theorem follows from Theorem 4.2 using an inequality known as Kantorivich's inequality.

[1]The concept of condition number, introduced by Alan Turing while in Manchester, is one of the most important ideas in numerical analysis, as it is indispensable in studying the performance of numerical algorithms.