Lecture 24

Conjugate Gradient Method

Learning Outcomes

- **Understanding Conjugate Gradient Method:**
  - Acquire knowledge on the Conjugate Gradient Method, focusing on the concept of A-orthogonal or conjugate search directions, and how it minimizes the function g over the set $x^{(0)} + \text{span}\{d^{(0)}, \ldots, d^{(l-1)}\}$, providing a more efficient approach compared to the Steepest Descent Method.
- **Mastering the Implementation of the Method:**
  - Develop proficiency in implementing the Conjugate Gradient Method, understanding the computation of the scalars $\alpha^{(k-1)}$ and $\beta^{(k)}$, and how to update the search direction efficiently, ensuring that only one matrix-vector multiplication per iteration step is required.
- **Analyzing the Methods Efficiency:**
  - Gain insights into the efficiency of the Conjugate Gradient Method, learning how the method avoids the zigzag path observed in the Steepest Descent Method by choosing A-orthogonal search directions, and how it can be the global minimizer of g and provide the desired solution to $Ax = b$.

The steepest descent method considered a nonlinear iterative method for obtaining a solution of the form,

$$x^{(k)} = x^{(k-1)} + \alpha^{(k-1)} d^{(k-1)}$$

which actually says that at each step the current guess is a linear combination of the search directions $d^{(0)}, \ldots, d^{(k-1)}$,

$$x^{(k)} = x^{(0)} + \alpha_0 d^{(0)} + \cdots + \alpha^{(k-2)} d^{(k-2)} + \alpha^{(k-1)} d^{(k-1)}$$

This $x^{(k)}$ is shown to be the minimiser of the function $g(x^{(k)}) = \frac{1}{2}\|r^{(k)}\|^2_{A^{-1}}$ over the set of all points along these search directions. However, since $x^{(k)}$ is a linear combination of all previous search directions a better approach would be to compute $x^{(k)}$ such that it is the minimiser of the span of all these search directions.
Another important observation is that in the steepest descent we search in the direction of the gradient and then we minimise the cost function in that search direction which results in the

zig-zag to the solutions in a large number of steps. Thus an additional improvement could be introduced by considering a search direction that is not just the gradient but a composite of the gradient and the previous search directions,

$$d^{(k)} = r^{(k)} + \beta^{(k)} d^{(k-1)}$$

To introduce these two improvements, the following questions need to be answered,

- How to find the step size, $\alpha^{(k)}$ of a given search direction,
- How to find the amount, $\beta^{(k)}$ of the previous search direction.

To address the above two questions,

- We will assume, as in the previous lecture, that $A \in \mathbb{R}^{n\times n}$ is positive definite throughout.
- In SD, subsequent search directions were orthogonal with respect to the standard scalar product. Now, we consider A-orthogonal (or conjugate) search directions $d^{(k)}$, i.e., $\langle d^{(i)}, d^{(j)}\rangle_A = 0$ if $i \neq j$. This has the following advantage which addresses the first question above.

Lemma 24.1. *Assume that the search directions $d^{(0)}, \dots, d^{(l-1)}$ in the iteration (23.3) form an A-orthogonal set. Then $x^{(l)}$ minimises g over the set $x^{(0)} + \operatorname{span}\{d^{(0)}, \dots, d^{(l-1)}\}$.*

Proof. Consider the map

$$h : \mathbb{R}^l \to \mathbb{R}, \quad h(\gamma_0, \dots, \gamma_{l-1}) = g\Big(x^{(0)} + \sum_{i=0}^{l-1} \gamma_i d^{(i)}\Big).$$

Since h is convex and tends to infinity as $\|\gamma_0\|, \dots, \|\gamma_{l-1}\| \to \infty$, it has a unique minimum $\hat\gamma$. Recalling that $\nabla g(x) = Ax - b = -r$ and using the A-orthogonality of the search directions we obtain for all $m = 0, \dots, l-1$ that

$$\begin{aligned} 0 = \frac{\partial}{\partial \gamma_m} h(\hat\gamma) &= \Big\langle \nabla g\Big(x^{(0)} + \sum_{i=0}^{l-1} \hat\gamma_i d^{(i)}\Big), d^{(m)} \Big\rangle \\ &= \Big\langle Ax^{(0)} - b + \sum_{i=0}^{l-1} \hat\gamma_i A d^{(i)}, d^{(m)} \Big\rangle = -\langle r^{(0)}, d^{(m)}\rangle + \sum_{i=0}^{l-1} \hat\gamma_i \underbrace{\langle Ad^{(i)}, d^{(m)}\rangle}_{=1 \text{ if } i=m,\ =0 \text{ else}} \end{aligned}$$

so that $\hat\gamma_m = \langle r^{(0)}, d^{(m)}\rangle / \|d^{(m)}\|_A^2$.
On the other hand, the optimal step length is $\alpha^{(m)} = \langle d^{(m)}, r^{(m)}\rangle / \|d^{(m)}\|_A^2$. But using the iterative formula for the residual (23.6)

$$\langle d^{(m)}, r^{(m)}\rangle = \langle d^{(m)}, r^{(m-1)}\rangle - \alpha^{(m-1)} \underbrace{\langle d^{(m)}, Ad^{(m-1)}\rangle}_{=0} = \cdots = \langle d^{(m)}, r^{(0)}\rangle$$

whence $\hat\gamma_m = \alpha^{(m)}$ so that $(\alpha^{(0)}, \dots, \alpha^{(l-1)})$ is the minimiser of h.
Consequently, $x^{(l)} = x^{(0)} + \sum_{i=0}^{l-1} \alpha^{(i)} d^{(i)}$ computed by the nonlinear iteration is the minimum of g on $x^{(0)} + \operatorname{span}\{d^{(0)}, \dots, d^{(l-1)}\}$ as asserted. □

For $l = n$ and on assuming that $d^{(k)} \neq 0$ for all $k = 0, \dots, n-1$ we have that $x^{(0)} + \operatorname{span}\{d^{(0)}, \dots, d^{(n-1)}\} = \mathbb{R}^n$, so the above lemma then means that $x^{(n)}$ is the global minimiser of g and the desired solution to $Ax = b$. Going back to the case $n = 2$ and Figure 23.1, choosing $d^{(0)} = r^{(0)}$ as first search direction this means that the second search direction would ensure jumping from $x^{(1)}$ immediately to the minimum and we would avoid the zigzag path. So the big questions is: How can we obtain A-orthogonal search directions?
More precisely, given $d^{(0)}, \dots, d^{(k-1)}$ (and the $x^{(i)}$ and $r^{(i)}$), how can an appropriate $d^{(k)}$ A-orthogonal to all the previous search directions be obtained? The following ideas go back to Hestenes and Stiefel:

1. Given any $v \notin \operatorname{span}\{d^{(0)}, \dots, d^{(k-1)}\}$, such a vector can be computed via the Gram-Schmidt orthogonalisation method (applied with the A-scalar product, of course):

$$\tilde{d}^{(k)}(v) := v - \frac{\langle d^{(k-1)}, v\rangle_A}{\|d^{(k-1)}\|_A^2} d^{(k-1)} - \frac{\langle d^{(k-2)}, v\rangle_A}{\|d^{(k-2)}\|_A^2} d^{(k-2)} - \dots - \frac{\langle d^{(0)}, v\rangle_A}{\|d^{(0)}\|_A^2} d^{(0)}.$$

   Apart from stability issues, this becomes very expensive when k becomes big.

2. The choice $v = r^{(k)} = -\nabla g(x^{(k)})$ is quite reasonable: If $r^{(k)} \in \operatorname{span}\{d^{(0)}, \dots, d^{(k-1)}\}$ then $r^{(k)} = 0$ because $x^{(k)} = x$ already is the minimum by the preceding result in Lemma 24.1 and we would stop the iteration anyway. Moreover, since $r^{(k)}$ points in the direction of the steepest descent it gives a good idea into which direction roughly to proceed next. So set $d^{(k)} := \tilde{d}^{(k)}(r^{(k)})$.

3. Set $d^{(0)} = r^{(0)}$. Consequently, the first step of the iteration is the same as for **SD**.

The most amazing point with the above choices is that $\langle d^{(i)}, r^{(k)}\rangle_A = 0$ for $i = 0, \dots, k-2$ as we shall prove later on. This means that

$$d^{(k)} = r^{(k)} \underbrace{- \frac{\langle d^{(k-1)}, r^{(k)}\rangle_A}{\|d^{(k-1)}\|_A^2}}_{=:\beta^{(k)}} d^{(k-1)} - \sum_{i=0}^{k-2} \underbrace{\frac{\langle d^{(i)}, r^{(k)}\rangle_A}{\|d^{(k-2)}\|_A^2}}_{=0} d^{(i)} = r^{(k)} + \beta^{(k)} d^{(k-1)},$$

hence *the update of the search direction in fact is cheap!*
In the algorithm below the scalars $\alpha^{(k-1)}$ and $\beta^{(k)}$ are computed with slightly different formulas that will turn out to be algebraically equivalent (see Lemma 25.1) but in practice turned out to perform somewhat better.

Algorithm 13 CG (conjugate gradient method)

input: $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ positive definite, $b, x^{(0)} \in \mathbb{R}^n$, $\varepsilon_r > 0$.
output: $x \in \mathbb{R}^n$ with $\|Ax - b\|_2 \leq \varepsilon_r$.

$d^{(0)} = r^{(0)} = b - Ax^{(0)}$
if $\|r^{(0)}\|_2 \leq \varepsilon_r$ **then**
  return $x^{(0)}$
else
  for $k = 1, 2, \ldots$ **do**
    $h^{(k-1)} = Ad^{(k-1)}$
    $\alpha^{(k-1)} := \|r^{(k-1)}\|_2^2 / d^{(k-1)^T} h^{(k-1)}$, where we used assertion 4 from Lemma 25.1
    $x^{(k)} := x^{(k-1)} + \alpha^{(k-1)} d^{(k-1)}$
    $r^{(k)} := r^{(k-1)} - \alpha^{(k-1)} h^{(k-1)}$
    if $\|r^{(k)}\|_2 \leq \varepsilon_r$ **then**
      return $x^{(k)}$
    end if
    $\beta^{(k)} := \|r^{(k)}\|_2^2 / \|r^{(k-1)}\|_2^2$, where we used assertion 5 from Lemma 25.1
    $d^{(k)} := r^{(k)} + \beta^{(k)} d^{(k-1)}$
  end for
end if

Note that the algorithm is formulated such that only one matrix-vector multiplication per iteration step is required.

Example: Consider the data

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad x^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We then have $d^{(0)} = r^{(0)} = b$.
Step $k = 1$:

$$h^{(0)} := Ad^{(0)} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$\alpha^{(0)} := \frac{\|r^{(0)}\|_2^2}{\|d^{(0)}\|_A^2} = \frac{1}{\langle d^{(0)}, h^{(0)} \rangle} = \frac{1}{2}$$

$$x^{(1)} := x^{(0)} + \alpha^{(0)} d^{(0)} = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}$$

$$r^{(1)} := r^{(0)} - \alpha^{(0)} h^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}$$

$$\beta^{(1)} := \frac{\|r^{(1)}\|_2^2}{\|r^{(0)}\|_2^2} = \frac{1}{4}$$

$$d^{(1)} := r^{(1)} + \beta^{(1)} d^{(0)} = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1/4 \\ 1/2 \end{pmatrix}$$

Step $k = 2$:

$$h^{(1)} := Ad^{(1)} = \begin{pmatrix} 0 \\ 3/4 \end{pmatrix}$$
$$\alpha^{(1)} := \frac{\|r^{(1)}\|_2^2}{\langle d^{(1)}, h^{(1)} \rangle} = \frac{1/4}{3/8} = \frac{2}{3}$$
$$x^{(2)} := x^{(1)} + \alpha^{(1)} d^{(1)} = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix} + \frac{2}{3} \begin{pmatrix} 1/4 \\ 1/2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$
$$r^{(2)} := r^{(1)} - \alpha^{(1)} h^{(1)} = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix} - \frac{2}{3} \begin{pmatrix} 0 \\ 3/4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Lecture 25

More on CG

Learning Outcomes

- Understanding the significance of using conjugate or A-orthogonal search directions in the Conjugate Gradient (CG) method to improve convergence compared to the steepest descent method, and recognizing the efficiency of updating the search direction in each step through Gram-Schmidt orthogonalization.
- Gaining insights into the properties and assertions of vectors and iterates computed by CG, such as A-orthogonality of the vectors $d^{(0)}, \ldots, d^{(k)}$ and the relationships between the vectors $r^{(k)}$, $d^{(k)}$, and A as expressed in Lemma and Proof sections, contributing to the understanding of the update formulas and orthogonality conditions in CG.
- Acquiring knowledge about the Krylov subspaces and their role in iterative methods like GMRES and BiCGstab, and understanding the characteristic property of these methods where the increment lies in the actual Krylov subspace, leading to insights into the convergence and solution reach of the CG algorithm in solving system of linear equations (SLEs).

We have motivated to use conjugate or A-orthogonal search directions in order to improve the convergence in comparison with the straight forward steepest descent method. An open issue has been the claim that the update of the search direction in each step, which is based on a Gram-Schmidt orthogonalisation, is cheap as most of the terms drop out.

Lemma 25.1. *Let $x^{(1)}, \ldots, x^{(k)}$ be the iterates computed by* **CG** *and assume that $r^{(0)}, \ldots, r^{(k)}, d^{(0)}, \ldots, d^{(k-1)} \neq 0$. Then*

(1) $\langle d^{(k-1)}, r^{(k)} \rangle = 0$,

(2) $\|d^{(k)}\|_2 \geq \|r^{(k)}\|_2 > 0$,

(3) $\alpha^{(k-1)} = \|r^{(k-1)}\|_2^2 / \|d^{(k-1)}\|_A^2 > 0$,

(4) $\langle r^{(k-1)}, r^{(k)} \rangle = 0$,

(5) $\beta^{(k)} = \|r^{(k)}\|_2^2 / \|r^{(k-1)}\|_2^2 > 0$.

Proof. Using the update formulas $r^{(k)} = r^{(k-1)} - \alpha^{(k-1)} A d^{(k-1)}$ and the one for $\alpha^{(k-1)}$ we obtain that

$$\langle d^{(k-1)}, r^{(k)} \rangle = \langle d^{(k-1)}, r^{(k-1)} \rangle - \frac{\langle d^{(k-1)}, r^{(k-1)} \rangle}{\|d^{(k-1)}\|_A^2} \langle d^{(k-1)}, A d^{(k-1)} \rangle = 0 \tag{25.1}$$

which proves (1). A consequence of this Euclidean orthogonality is that

$$\|d^{(k)}\|_2^2 = \|r^{(k)} + \beta^{(k)} d^{(k-1)}\|_2^2 = \|r^{(k)}\|_2^2 + |\beta^{(k)}|^2 \|d^{(k-1)}\|_2^2 > \|r^{(k)}\|_2^2 > 0$$

which is assertion (2). A further consequence of (25.1) is that for $k > 1$

$$\langle d^{(k-1)}, r^{(k-1)} \rangle = \langle r^{(k-1)} + \beta^{(k-1)} d^{(k-2)}, r^{(k-1)} \rangle = \|r^{(k-1)}\|_2^2,$$

and thanks to the choice $d^{(0)} = r^{(0)}$ this is also true for $k = 1$. This implies assertion (3),

$$\alpha^{(k-1)} = \frac{\langle d^{k-1)}, r^{(k-1)} \rangle}{\|d^{(k-1)}\|_A^2} = \frac{\|r^{(k-1)}\|_2^2}{\|d^{(k-1)}\|_A^2} > 0.$$

To show (4) we first observe that for $k > 1$ thanks to the A orthogonality of the $d^{(i)}$

$$\langle r^{(k-1)}, d^{(k-1)} \rangle_A = \langle d^{(k-1)}, d^{(k-1)} \rangle_A - \beta^{(k-1)} \underbrace{\langle d^{(k-2)}, d^{(k-1)} \rangle_A}_{=0} = \|d^{(k-1)}\|_A^2,$$

and by the choice $d^{(0)} = r^{(0)}$ this is also true for the case $k = 1$. Therefore, using the already proved identity (3)

$$\begin{aligned} \langle r^{(k-1)}, r^{(k)} \rangle &= \langle r^{(k-1)}, r^{(k-1)} \rangle - \alpha^{(k-1)} \langle r^{(k-1)}, A d^{(k-1)} \rangle \\ &= \langle r^{(k-1)}, r^{(k-1)} \rangle - \frac{\|r^{(k-1)}\|_2^2}{\|d^{(k-1)}\|_A^2} \|d^{(k-1)}\|_A^2 = 0. \end{aligned}$$

Finally we prove the update formula (5). Since $\alpha^{(k-1)} > 0$ we can write $-Ad^{(k-1)} = \frac{1}{\alpha^{(k-1)}}(r^{(k)} - r^{(k-1)})$, Using this and the already shown identities (4) and (3) we get

$$\beta^{(k)} = -\frac{\langle d^{(k-1)}, r^{(k)} \rangle_A}{\|d^{(k-1)}\|_A^2} = \frac{\langle r^{(k)} - r^{(k-1)}, r^{(k)} \rangle}{\alpha^{(k-1)} \|d^{(k-1)}\|_A^2} = \frac{\|r^{(k)}\|_2^2}{\|r^{(k-1)}\|_2^2} > 0.$$

□

The following lemma is central to **CG**:

Lemma 25.2. *The vectors $d^{(0)}, \ldots, d^{(k)}$ are A-orthogonal. Moreover*

$$\operatorname{span}\{r^{(0)}, \ldots, r^{(l)}\} = \operatorname{span}\{d^{(0)}, \ldots, d^{(l)}\} = \operatorname{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^l r^{(0)}\} \tag{25.2}$$

for $l = 0, \ldots, k-1$.

Proof. We start with proving the second assertion by induction.

- the base case $l = 0$ is clear thanks to the choice $d^{(0)} = r^{(0)}$.
- Inductive hypothesis. Let $l > 0$ and assume that (25.2) is true for $l - 1$.
  Show that the result holds for l.
  - Noting that $r^{(l)} = r^{(l-1)} - \alpha^{(l-1)} A d^{(l-1)}$. By the induction hypothesis we have,

$$r^{(l-1)} \in \operatorname{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^{l-1} r^{(0)}\}$$

    and

$$d^{(l-1)} \in \operatorname{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^{l-1} r^{(0)}\}$$

multiplying this equation by A, we get

$$Ad^{(l-1)} \in \text{span}\{Ar^{(0)}, A^2r^{(0)}, \ldots, A^lr^{(0)}\}$$

now since,

$$r^{(l)} = r^{(l-1)} - \alpha^{(l-1)}Ad^{(l-1)}$$

then,

$$r^{(l)} \in \text{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^lr^{(0)}\}$$

– Noting that $d^{(l)} = r^{(l)} + \beta^{(l)}d^{(l-1)} = r^{(l-1)} - \alpha^{(l-1)}Ad^{(l-1)} + \beta^{(l)}d^{(l-1)}$. By the induction hypothesis we have,

$$d^{(l)} \in \text{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^lr^{(0)}\}$$

since,

$$r^{(l)} \in \text{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^lr^{(0)}\}$$

and

$$Ad^{(l-1)} \in \text{span}\{Ar^{(0)}, A^2r^{(0)}, \ldots, A^lr^{(0)}\}$$

- We complete the inductive step by noting that all three subspaces have the same dimension and hence must be the same subspace.

Let us now come to the assertion on the A-orthogonality. We show this by induction, too, where the case $k = 1$ trivially is fulfilled. So let $k > 1$ and assume that $d^{(0)}, \ldots, d^{(k-1)}$ are A-orthogonal.
Consider an index $i < k - 1$. Then for all $j = i + 1, \ldots, k - 1$

$$\langle d^{(i)}, r^{(j+1)}\rangle = \langle d^{(i)}, r^{(j)}\rangle - \alpha^{(j)} \underbrace{\langle d^{(i)}, Ad^{(j)}\rangle}_{=0} = \langle d^{(i)}, r^{(j)}\rangle$$

where the second term vanishes thanks to the induction hypothesis. But since $\langle d^{(i)}, r^{(i+1)}\rangle = 0$ by lemma 25.1 we conclude that

$$\langle d^{(i)}, r^{(j)}\rangle = 0, \quad j = i + 1, \ldots, k \tag{25.3}$$

where $i = 0, \ldots, k - 1$.
Let $l < k - 1$. Then thanks to (25.2), $Ad^{(l)} \in \text{span}\{r^{(l)}, r^{(l+1)}\} \subset \text{span}\{d^{(0)}, \ldots, d^{(l+1)}\}$, so that, using (25.3), $\langle d^{(l)}, r^{(k)}\rangle_A = \langle Ad^{(l)}, r^{(k)}\rangle = 0$. Consequently,

$$\langle d^{(l)}, d^{(k)}\rangle_A = \underbrace{\langle d^{(l)}, r^{(k)}\rangle_A}_{=0} + \beta^{(k)} \underbrace{\langle d^{(l)}, d^{(k-1)}\rangle_A}_{=0} = 0,$$

where we used the induction hypothesis for the second term.
In the only remaining case $l = k - 1$ it is the choice of $\beta^{(k)}$ which ensures A-orthogonality:

$$\begin{aligned}\langle d^{(k-1)}, d^{(k)}\rangle_A &= \langle d^{(k-1)}, r^{(k)}\rangle_A + \beta^{(k)}\langle d^{(k-1)}, d^{(k-1)}\rangle_A \\ &= \langle d^{(k-1)}, r^{(k)}\rangle_A - \frac{\langle d^{(k-1)}, r^{(k)}\rangle_A}{\|d^{(k-1)}\|_A^2}\|d^{(k-1)}\|_A^2 = 0.\end{aligned}$$

□

The spaces in (25.2), denoted by

$$\mathcal{K}_k(r^{(0}, A) := \text{span}\{r^{(0)}, \ldots, A^{k-1}r^{(0)}\}$$

are called Krylov subspaces and play a prominent role in other iterative methods such as GMRES and BiCGstab which, indeed, are even termed Krylov (sub)space methods. A characteristic property of these methods is that the increment lies in the actual Krylov subspace:

$$x^{(k)} - x^{(k-1)} \in \mathcal{K}_k(r^{(0)}, A).$$

A consequence of the previous results is

Theorem 25.1. *The* **CG** *algorithm reaches the exact solution to SLEs in at most n steps for any $x^{(0)}$.*

So in fact **CG** is a direct method. But in practice it is considered as an iterative method because ε_r usually is much bigger than the machine precision ε_m so that the iteration terminates with an approximation $x^{(k)}$ where k is much smaller than n.

Lecture 26

Computational Complexity and Error Analysis of SD and CG

Learning Outcomes

- Understanding of error analysis in Steepest Descent (SD) and Conjugate Gradient (CG) methods, focusing on convergence analysis and not considering rounding errors. This includes the ability to relate energy and Euclidean norm, and to apply lemmas and theorems to analyze the convergence rate of SD, demonstrating the impact of the condition number $\kappa_2(A)$ on the convergence rate.
- Insight into the properties and application of Chebyshev polynomials in optimizing with respect to the norm $\|\cdot\|_\infty$, and the ability to relate these polynomials to the eigenvalues of A, $\lambda_{\max}$ and $\lambda_{\min}$, and to the condition number $\kappa_2(A)$. This includes understanding the role of rescaled Chebyshev polynomial in error analysis of CG method.
- Comprehension of the computational complexity of SD and CG methods under certain assumptions, including the ability to estimate the required number of steps in terms of system size n and tolerance ε_r, and to relate the cost to achieve a certain error level in SD and CG to the system size and tolerance, considering the impact of computing Ax and the condition number $\kappa_2(A)$ on the cost.

Error analysis of SD and CG

Error analysis for iterative methods is convergence analysis, rounding errors are not taken into account. The concepts of analysing SD and CG are similar which is why they are presented together. We start with a helpful lemma relating energy and Euclidean norm.

Lemma 26.1. *Assume that $\|e^{(k)}\|_A \leq cq^k\|e^{(0)}\|_A$ for all $k \in \mathbb{N}$ and some constants $c, q > 0$. Then*

$$\|e^{(k)}\|_2 \leq \sqrt{\kappa_2(A)}cq^k\|e^{(0)}\|_2 \quad \forall k \in \mathbb{N}.$$

Proof. Denoting the minimal and maximal eigenvalue of A by $\lambda_{\min}$ and $\lambda_{\max}$, respectively, and recalling that $\kappa_2(A) = \lambda_{\max}/\lambda_{\min}$,

$$\|e^{(k)}\|_2^2 \leq \frac{1}{\lambda_{\min}}\|e^{(k)}\|_A^2 \leq \frac{1}{\lambda_{\min}}c^2q^{2k}\|e^{(0)}\|_A^2 \leq \frac{\lambda_{\max}}{\lambda_{\min}}(cq^k)^2\|e^{(0)}\|_2^2.$$

□

LECTURE 26. COMPUTATIONAL COMPLEXITY AND ERROR ANALYSIS OF SD AND CG

The first results concerns **SD**.

Theorem 26.1. *The convergence rate of* **SD** *is*

$$\|e^{(k)}\|_A \leq \left(\sqrt{1-\frac{1}{\kappa_2(A)}}\right)^k \|e^{(0)}\|_A.$$

Proof. Recalling (23.4) which with $d^{(k)} = r^{(k)}$ reads

$$g(x^{(k-1)} + \alpha r^{(k-1)}) = \frac{1}{2}\alpha^2 \|r^{(k-1)}\|_A^2 - \alpha \langle r^{(k-1)}, r^{(k-1)} \rangle + \frac{1}{2}\|r^{(k-1)}\|_{A^{-1}}^2,$$

we first observe that

$$\begin{aligned} g(x^{(k)}) &= g(x^{(k-1)} + \alpha^{(k-1)} d^{(k-1)}) \\ &= \frac{\|r^{(k-1)}\|_2^4}{2\|r^{(k-1)}\|_A^4} \|r^{(k-1)}\|_A^2 - \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_A^2} \|r^{(k-1)}\|_2^2 + \frac{1}{2}\|r^{(k-1)}\|_{A^{-1}}^2 \\ &= \frac{1}{2}\|r^{(k-1)}\|_{A^{-1}}^2 - \frac{\|r^{(k-1)}\|_2^4}{2\|r^{(k-1)}\|_A^2} \\ &= \left(1 - \frac{\|r^{(k-1)}\|_2^4}{\|r^{(k-1)}\|_A^2 \|r^{(k-1)}\|_{A^{-1}}^2}\right) \underbrace{\frac{1}{2}\|r^{(k-1)}\|_{A^{-1}}^2}_{=g(x^{(k-1)})}, \end{aligned}$$

Using the estimates $\|v\|_A^2 \leq \lambda_{\max}\|v\|_2^2$ and $\|v\|_{A^{-1}}^2 \leq \frac{1}{\lambda_{\min}}\|v\|_2^2$ we obtain that this is

$$\begin{aligned} &\leq \left(1 - \frac{\|r^{(k-1)}\|_2^4}{\lambda_{\max}\|r^{(k-1)}\|_2^2 \frac{1}{\lambda_{\min}}\|r^{(k-1)}\|_2^2}\right) g(x^{(k-1)}) \\ &= \left(1 - \frac{1}{\kappa_2(A)}\right) g(x^{(k-1)}). \end{aligned}$$

Therefore

$$g(x^{(k)}) \leq \left(1 - \frac{1}{\kappa_2(A)}\right)^k g(x^{(0)})$$

Using that $g(x^{(l)}) = \frac{1}{2}\|e^{(l)}\|_A^2$ for $l = 0, k$ (see (23.2)) yields the assertion. □

For **CG** we have the following result where $\mathcal{P}^k$ denotes the set of polynomials p of degree $\leq k$ with $p(0) = 1$ and $\Lambda(A)$ is the set of eigenvalues of A:

Theorem 26.2. *If* **CG** *has not yet converged after step k then*

$$\|e^{(k)}\|_A = \inf_{p \in \mathcal{P}^k} \|p(A)e^{(0)}\|_A \leq \inf_{p \in \mathcal{P}^k} \max_{\lambda \in \Lambda(A)} |p(\lambda)| \|e^{(0)}\|_A.$$

We only prove the first equality. The second one is an exercise.

Proof. The proof utilizes polynomial properties and CG method's construction to express the error $e^{(k)}$ in terms of a polynomial applied to A and the initial error $e^{(0)}$.

1. Error is represented as $e^{(k)} = x - x^{(k)}$.

2. Since $x^{(k)}$ lies in $x^{(0)} + \text{span}\{r^{(0)}, \ldots, A^{k-1}r^{(0)}\}$, we can find coefficients η_j such that:

$$e^{(k)} = e^{(0)} + \sum_{j=0}^{k-1} \eta_j A^{j+1} e^{(0)}$$

3. Constructing the polynomial $q(\lambda) = 1 + \sum_{j=1}^{k} \eta_{j-1}\lambda^j$, we have $e^{(k)} = q(A)e^{(0)}$.

4. It follows that:

$$\|e^{(k)}\|_A = \|q(A)e^{(0)}\|_A.$$

5. The proof shows that this q provides the smallest A-norm for $e^{(k)}$ over all polynomials of degree k.

The theorem provides insights into the CG method's effectiveness, showing that at each step k, CG minimizes the error over a set of polynomials of A, thus offering a strong theoretical foundation for its convergence properties. □

Chebyshev polynomials

These polynomials are defined by

$$T_n(x) := \frac{1}{2}\big((x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n\big), \quad x \in [-1, 1]$$

and fulfil the recursive formula

$$T_0(x) = 1, \quad T_1(x) = x, \qquad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ n = 1, 2, \ldots$$

A further definition is

$$T_n(x) = \cos(n \arccos(x)), \quad x \in (-1, 1).$$

We see that $\max_{|x|\le 1} |T_n(x)| \le 1$, and indeed the Chebyshev polynomials play an important role when optimising with respect to the norm $\|\cdot\|_\infty$.
Let $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximal and minimal eigenvalue of A and consider the rescaled Chebyshev polynomial

$$p(x) := T_n\Big(\gamma - \frac{2x}{\lambda_{\max} - \lambda_{\min}}\Big)\Big/T_n(\gamma)$$

where

$$\gamma = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{\lambda_{\max}/\lambda_{\min} + 1}{\lambda_{\max}/\lambda_{\min} - 1} = \frac{\kappa_2(A) + 1}{\kappa_2(A) - 1}, \tag{26.1}$$

where we recall that $\|A\|_2 = \lambda_{\max}$ and $\|A^{-1}\|_2 = \lambda_{\min}^{-1}$ so that $\kappa_2(A) = \lambda_{\max}/\lambda_{\min}$. For $x \in [\lambda_{\min}, \lambda_{\max}]$ we have that $\gamma - 2x/(\lambda_{\max} - \lambda_{\min}) \in [-1, 1]$, and since then $|T_n(x)| \le 1$ we arrive at

$$|p(x)| \le 1/T_n(\gamma), \quad x \in [\lambda_{\min}, \lambda_{\max}]. \tag{26.2}$$

Writing $\kappa := \kappa_2(A)$ we first observe that

$$\frac{\kappa+1}{\kappa-1} \pm \sqrt{\frac{(\kappa+1)^2}{(\kappa-1)^2} - 1} = \frac{\kappa+1}{\kappa-1} \pm \sqrt{\frac{(\kappa+1)^2 - (\kappa-1)^2}{(\kappa-1)^2}}$$
$$= \frac{\kappa + 1 \pm \sqrt{4\kappa}}{\kappa - 1} = \frac{(\sqrt{\kappa} \pm 1)^2}{(\sqrt{\kappa}+1)(\sqrt{\kappa}-1)} = \Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^{\pm 1}.$$

Thanks to (27.1) and (27.2)

$$\begin{aligned} T_n(\gamma) &= \frac{1}{2}\left(\left(\frac{\kappa+1}{\kappa-1}+\sqrt{\frac{(\kappa+1)^2}{(\kappa-1)^2}-1}\right)^n+\left(\frac{\kappa+1}{\kappa-1}-\sqrt{\frac{(\kappa+1)^2}{(\kappa-1)^2}-1}\right)^n\right) \\ &= \frac{1}{2}\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n+\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n\right). \end{aligned}$$

Using (27.2) we see that for all $x \in [\lambda_{\min}, \lambda_{\max}]$

$$|p(x)| \leq 2\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n+\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n\right)^{-1} \tag{26.3}$$

Since $p \in \mathcal{P}^k$ for $k = n$ and since estimate (27.3) holds true for all $x \in \Lambda(A) \subset [\lambda_{\min}, \lambda_{\max}]$ we deduce from Theorem 27.2 the following result:

Theorem 26.3. [6.20] *If* **CG** *has not yet converged after step k then*

$$\|e^{(k)}\|_A \leq 2\left(\left(\frac{\sqrt{\kappa_2(A)}+1}{\sqrt{\kappa_2(A)}-1}\right)^k+\left(\frac{\sqrt{\kappa_2(A)}+1}{\sqrt{\kappa_2(A)}-1}\right)^{-k}\right)^{-1}\|e^{(0)}\|_A.$$

Computational Complexity

For the computational complexity we proceed analogously as in the context of the linear iterative methods: Get an estimate for the required number of steps in terms of the system size n and the tolerance ε_r, and then multiply with the cost per step.

Assumption 26.1.

1. *Computing Ax involves a cost of $\Theta(n^\alpha)$ as $n \to \infty$ with some $\alpha \in [1, 2]$.*
2. *$\kappa_2(A) = \Theta(n^\beta)$ as $n \to \infty$ with some $\beta \geq 0$.*
3. *$\|e^{(0)}\|_A$ is uniformly bounded in n.*

Theorem 26.4. *Under Assumption 27.1, the cost to achieve $\|e^{(k)}\|_A \leq \varepsilon_r$ with* **SD** *is bounded by a function $C(n, \varepsilon_r)$ satisfying*

$$C(n, \varepsilon_r) = \Theta(n^{\alpha+\beta}\log(\varepsilon_r^{-1})) \quad \text{as } (n, \varepsilon_r) \to (\infty, 0).$$

Proof. By Theorem 27.1, $\|e^{(k)}\|_A \leq \left(\sqrt{1-\frac{1}{\kappa_2(A)}}\right)^k \|e^{(0)}\|_A$, hence it is sufficient to achieve that

$$\frac{\|e^{(0)}\|_A}{\varepsilon_r} \leq \left(\frac{1}{\sqrt{1-\frac{1}{\kappa_2(A)}}}\right)^k \quad \Leftrightarrow \quad k \geq \frac{\log(\|e^{(0)}\|_A)+\log(\varepsilon_r^{-1})}{\log\left(1/\sqrt{1-\frac{1}{\kappa_2(A)}}\right)} =: k^\sharp(n, \varepsilon_r).$$

Using the Taylor expansion we see that

$$\log\left(\frac{1}{\sqrt{1-x}}\right) = \frac{1}{2}x + O\left(x^2\right) \text{ as } x \to 0,$$

and with $x = 1/\kappa_2(A) = \Theta(n^{-\beta})$ one can proceed as in the proof of Theorem 22.1 to show the assertion. □

Theorem 26.5. *Under Assumption 27.1, the cost to achieve* $\|e^{(k)}\|_A \leq \varepsilon_r$ *with* **CG** *is bounded by a function* $C(n, \varepsilon_r)$ *satisfying*

$$C(n, \varepsilon_r) = \Theta(n^{\alpha+\frac{1}{2}\beta} \log(\varepsilon_r^{-1})) \quad \text{as } (n, \varepsilon_r) \to (\infty, 0).$$

(Spot the difference to the previous theorem!)

Proof. As for the previous theorem but based on the estimate in Theorem 27.3. The fact that there only $\sqrt{\kappa_2(A)}$ appears rather than $\kappa_2(A)$ as in the estimate for **SD** (see Theorem 27.1) leads to the prefactor $\frac{1}{2}$ in front of β. □