## Boundary Value Problems and Iterative Methods for Linear Systems

### 1 Introduction

In many applications we have to solve a linear system \( Au = b \) where the matrix \( A \in \mathbb{R}^{n \times n} \) and the right hand side vector \( b \in \mathbb{R}^n \) are given. If \( n \) is not too large (say, less than \( 10^4 \)) we can use Gaussian elimination. In Matlab this means we can use \( u = A \backslash b \) which uses Gaussian elimination with partial pivoting (i.e., row permutations).

Many practical problems lead to much larger linear systems. E.g., if we consider temperatures in a unit cube \([0, 1] \times [0, 1] \times [0, 1] \subset \mathbb{R}^3\). We subdivide each edge into \( N = 100 \) subintervals of length \( h = 1/N \). This leads to a grid of \( n = N^3 = 10^6 \) grid points in the volume of the cube. If we have some physical law relating the temperatures at the grid points we obtain a linear system of about \( 10^6 \) unknowns.

If we have a "full matrix" (i.e., all entries nonzero) \( A \in \mathbb{R}^{n \times n} \) with \( n = 10^6 \) we have a problem: Just storing the entries of the matrix would take \( n^2 \times 8 \) bytes, i.e.,

\[
10^{12} \cdot 8 \text{ bytes} = 8000 \text{ Gigabytes}
\]

which is not practical. Solving a linear system with Gaussian elimination takes \( \frac{2}{3}n^3 \) floating point operations. For \( n = 10^6 \) this would be \( \frac{2}{3} \cdot 10^9 \cdot 8 \) floating point operations. Currently a very fast processor (like core i9 or Xeon) can perform up to 100 Gigaflops for one core (1 Gigaflop means \( 10^9 \) floating point operations per second). Hence \( \frac{2}{3} \cdot 10^{18} \) operations take about

\[
\frac{2}{3} \cdot 10^{18} / 10^9 \text{ seconds} = \frac{2}{3} \cdot 10^9 \text{ seconds} \approx 21 \text{ years}
\]

of computation time (if we had sufficient memory).

However, for most large linear systems we have a "sparse matrix" (i.e., most entries are zero): Each of the unknowns is only coupled with a small number of other unknowns, e.g., 6 other unknowns. This means that each row of the matrix contains at most \( 7 \) nonzeros. Hence the number of nonzeros is only \( Cn \) rather than \( n^3 \). Matlab has a sparse matrix format which uses storage proportional to the number of nonzeros. If the matrix has \( Cn \) nonzeros we need \( Cn \) storage, and we can compute a matrix-vector product \( Av \) with \( Cn \) floating point operations (rather than \( n^2 \) as in the case of a full matrix).

For a large sparse linear system \( Au = b \) there are two types of algorithms available:

**direct solvers** like Gaussian elimination give the exact solution of the linear system in finitely many operations (up to roundoff error): We can use standard Gaussian elimination with partial pivoting and find the LU decomposition. However, the factors \( L, U \) will usually have MANY MORE NONZEROS than the original matrix \( A \) (so-called “fill-in”), hence this will not be feasible for large \( n \). It turns out that using the unknowns in a different order (this corresponds to a column permutation of the matrix) can dramatically decrease the number of nonzeros in the factors \( L, U \). This is what Matlab does by default if you type \( u = A \backslash b \) with a matrix \( A \) in sparse format.

**iterative solvers** start with an initial guess. Each iteration improves the error, and we iterate until the error is below a tolerance. For each iteration we use matrix-vector products like \( Av \) or \( A^T v \). Hence typically each operation uses \( Cn \) operations.

**Example:** For Richardson iteration we start with an initial guess \( u^0 \) and compute the residual \( r^0 := b - Au^0 \). Then we define \( u^{k+1} := u^k + \alpha r^k \). Here we choose a fixed value \( \alpha > 0 \) which is “sufficiently
small”. If the matrix $A$ has $C_n$ nonzeros each iteration costs $C'n$ operations. We need to find out how many iterations we need to reach our tolerance.

We will see: Typically the condition number of the matrix increases with $n$, e.g., $\text{cond}(A) \approx cn^\alpha$.

This means that the number of iterations $k$ will also grow with $n$, i.e., $k \approx c'n^\beta$.

**Claim:** Typically, for problems in $\mathbb{R}^d$ with $d \geq 2$ iterative methods are competitive, and for $d \geq 3$ iterative methods are usually superior to direct methods.

We will consider linear systems $Au = b$ where

- $A \in \mathbb{R}^{n \times n}$ is sparse
- $v^\top Av > 0$ for all nonzero vectors $v \in \mathbb{R}^n$

This is satisfied in many practical applications. In particular “equilibrium problems” in elasticity, fluid dynamics and heat flow problems give linear systems where $A$ has these properties.

A very important special case occurs if we have additionally

- $A$ is symmetric

In this case $A$ is symmetric positive definite, and this will allow us to obtain more efficient algorithms.

As typical applications we will consider boundary value problems for elliptic partial differential equations:

- “diffusion problems” (e.g., equilibrium problems for elastic membrane or heat flow)
- “convection-diffusion problems”

Note that the iterative methods we discuss are useful in many other applications.

In the first case $A$ is symmetric positive definite. In the second case $A$ is no longer symmetric, but still satisfies $v^\top Av > 0$ for nonzero $v$.

We will discuss the following iterative methods:

- Richardson iteration with fixed $\alpha$
- minimal residual method and “General Minimal Residual Method” (GMRES)
- “Conjugate Gradient Method” (CG)
- Preconditioned GMRES and preconditioned CG methods

## 2 Richardson iteration for problems with “coercivity condition”

Assume that $V$ is a vector space with an inner product $(u, v)_V$ and norm $\|u\|_V$. Assume that $V$ with this norm is complete. Such a space is called a Hilbert space. We will consider real Hilbert spaces.

Examples:

- $V = \mathbb{R}^n$ with $(u, v) = \sum_{j=1}^{n} u_j v_j$
- $V = L^2(\Omega)$ with $(u, v) = \int_{\Omega} u(x)v(x)dx$
- $V = H^1(\Omega)$ with $(u, v) = \int_{\Omega} [u(x)v(x) + \nabla u(x) \cdot \nabla v(x)]dx$

For a function $F : V \to V$ (may be nonlinear) and $b \in V$ we want to find $u \in V$ such that

$$F(u) = b$$

We have an initial guess $u_0 \in V$. In order to check the equation we compute the residual $r_0 := b - F(u_0)$. If $r_0 = 0$ we are done, otherwise we should use some correction based on $r_0$.

This motivates the following iterative method, called **Richardson iteration**: We have an initial guess $u_0$.

Pick $\alpha \in \mathbb{R}$ and define

$$u_{k+1} := u_k + \alpha (b - F(u_k)) \quad \text{for } k = 0, 1, 2, \ldots$$

If the function $F$ satisfies a “Lipschitz condition” (1) and a “coercivity condition” (2) the problem has a unique solution, and the Richardson iteration converges to this solution for $\alpha > 0$ sufficiently small:
Lemma 2.1. Let $V$ be a Hilbert space. Assume that the function $F: V \to V$ satisfies with constants $L, \gamma > 0$ for all $u, v \in V$

\begin{align*}
\|F(v) - F(u)\| &\leq L \|v - u\| \quad \text{"Lipschitz condition"} \\
(F(v) - F(u), v - u) &\geq \gamma \|v - u\|^2 \quad \text{"coercivity condition"}
\end{align*}

Then the equation $F(u) = b$ with $b \in V$ has a unique solution $u \in V$. The inverse mapping satisfies for $b,c \in V$

$$\|F^{-1}(c) - F^{-1}(b)\| \leq \gamma^{-1} \|c - b\|.$$

Proof. Consider the Richardson iteration $u_{k+1} = G(u_k)$ with $G(u) := u + \alpha(b - F(u))$ with $\alpha > 0$. We claim that $G$ is a contraction if $\alpha$ is small: With $e := v - u$ we have

$$\|G(v) - G(u)\|^2 = \|e - \alpha(F(v) - F(u))\|^2 = \|e\|^2 - 2\alpha\langle F(v) - F(u), e \rangle + \alpha^2\|F(v) - F(u)\|^2 \leq (1 - 2\alpha\gamma + \alpha^2L^2)\|e\|^2.$$ 

For $g(\alpha) := 1 - 2\alpha\gamma + \alpha^2L^2$ we have $g(0) = 1$ and $g'(0) = -2\gamma < 0$, so $G$ is a contraction for sufficiently small $\alpha$. By completing the square we get

$$g(\alpha) = L^2\left(\alpha - \frac{\gamma}{L^2}\right)^2 + 1 - \frac{\gamma^2}{L^2}$$

and see that $g(\alpha) < 1$ for $\alpha \in (0, 2\gamma/L^2)$; we can minimize $g(\alpha)$ by choosing $\alpha = \gamma/L^2$ and obtain

$$\|G(v) - G(u)\| \leq (1 - \frac{\gamma^2}{L^2})\|v - u\|.$$

By the contraction mapping theorem the equation $G(u) = u \iff F(u) = b$ has a unique solution. We obtain (3) from $\gamma\|v - u\|^2 \leq (F(v) - F(u), v - u) \leq \|F(v) - F(u)\|\|v - u\|.$

Algorithm 1: Use a fixed $\alpha$. If we know some (possibly nonoptimal) constants $L, \gamma$ satisfying (1), (2) we use $\alpha := \gamma/L^2$. Then we have convergence

$$\|u_{k+1} - u_*\| \leq \left(1 - \frac{\gamma^2}{L^2}\right)^{1/2}\|u_k - u_*\|.$$ 

Algorithm 2: If we don’t know the constants $\gamma, L$ we can choose for each step a parameter $\alpha_k$ by “line search”.

Find $\alpha_k > 0$ such that for $u_{k+1} = u_k + \alpha_k r_k$ the new residual $r_{k+1} = b - F(u_{k+1})$ has minimal norm, i.e., $f_k(\alpha) := \|b - F(u_k + \alpha r_k)\|^2$ becomes minimal for $\alpha = \alpha_k$. Since

$$f_k(\alpha) = \|r_k + F(u_k) - F(u_k + \alpha r_k)\|^2 = \|r_k\|^2 - 2\alpha^{-1}\left(F(u_k + \alpha r_k) - F(u_k), \alpha r_k\right) + \|F(u_k + \alpha r_k) - F(u_k)\|^2 \geq \gamma\|r_k\|^2 \leq (L\alpha\|r_k\|)^2$$

$$f_k(\alpha) \leq (1 - 2\alpha\gamma + \alpha^2L^2)\|r_k\|^2$$

If we perform an exact line search then we must have

$$\|r_{k+1}\|^2 = f_k(\alpha_k) \leq f_k(\gamma/L^2) \leq \left(1 - \frac{\gamma}{L^2}\right)\|r_k\|^2, \quad \|r_k\| \leq \left(1 - \frac{\gamma}{L^2}\right)^{k/2}\|r_0\| \to 0 \quad \text{as } k \to \infty$$

If we pick any $\alpha_k \in (0, 2\gamma/L^2)$ we will have $(1 - 2\alpha\gamma + \alpha^2L^2) < 1$ and hence $\|r_{k+1}\|^2 = f_k(\alpha_k) < \|r_k\|^2$. I.e., for $\alpha_k$ sufficiently small we will have $\|r_{k+1}\| < \|r_k\|$.

In practice we can use an approximate line search: Pick $q < 1$ and stop the search when you achieve $\|r_{k+1}\| \leq q\|r_k\|$. This works if we pick $q \geq (1 - \frac{\gamma}{L^2})^{1/2}$. Then we obtain

$$\|r_k\| = \|b - F(u_k)\| \leq q^k\|r_0\| \to 0 \quad \text{as } k \to \infty$$

Note that (3) implies

$$\|u_k - u_*\| \leq \gamma^{-1}\|F(u_k) - b\|$$

and $\|u_k - u_*\| \leq \gamma^{-1}q^k\|r_0\| \to 0$ as $k \to \infty$. 

3
Corollary 2.2. Let $V$ be a Hilbert space. Assume that the function $F: V \to V'$ satisfies with constants $L, \gamma > 0$ for all $u, v \in V$

$$\|F(v) - F(u)\|_{V'} \leq L \|v - u\|$$

$$[F(v) - F(u)](v - u) \geq \gamma \|v - u\|^2$$

Then the equation $F(u) = f$ with $f \in V'$ has a unique solution $u \in V$.

Proof. By the Riesz representation theorem there is a linear mapping $\phi: V' \to V$ such that $\ell(v) = (\phi \ell, v)$ for all $\ell \in V'$, $v \in V$, and $\|\phi \ell\|_V = \|\ell\|_{V'}$. We define $\tilde{F} := \phi \circ F: V \to V$ and can apply the previous Lemma. \qed

Corollary 2.3. (Lax-Milgram) Let $V$ be a Hilbert space. Assume that the bilinear form $a: V \times V \to \mathbb{R}$ satisfies with constants $L, \gamma > 0$ for all $u, v \in V$

$$|a(u, v)| \leq L \|u\| \|v\|$$

$$a(u, u) \geq \gamma \|u\|^2$$

Then the equation there is a unique $u \in V$ which satisfies

$$a(u, v) = f(v) \quad \text{for all } v \in V$$

and we have $\|u\| \leq \gamma^{-1} \|f\|_{V'}$.

Proof. Define $F: V \to V'$ by $F(u) := a(u, \cdot)$. By the definition of $\|\cdot\|_{V'}$, the function $F$ satisfies the assumptions of 2.2. The estimate for $\|u\|$ follows from $\gamma \|u\|^2 \leq a(u, u) = f(u) \leq \|f\|_{V'} \|u\|$. \qed

Special case: Linear problem $Au = b$ in $V = \mathbb{R}^n$

Corollary 2.4. Assume $A \in \mathbb{R}^{n \times n}$ satisfies for all $u \in \mathbb{R}^n$

$$(Au, u) \geq \gamma \|u\|^2_2 \quad (4)$$

Then the equation $Au = b$ has for $b \in \mathbb{R}^n$ a unique solution and we have $\|A^{-1}\|_2 \leq \gamma^{-1}$. The sharp values of $L, \gamma$ from (6), (7) are given by

$$L = \|A\|_2, \quad \gamma = \lambda_{\min}(A_{\text{symm}}) \quad \text{where } A_{\text{symm}} := \frac{1}{2} (A + A^\top)$$

We can find the solution of $Au = b$ with the following iterative methods:

1. Richardson iteration with fixed $\alpha$: Let $L := \|A\|_2$, then for $\alpha \in (0, 2\gamma/L^2)$ the iteration

$$u^{k+1} = u^k + \alpha(b - Au^k) \quad (5)$$

converges. In particular, for $\alpha = \gamma/L^2$ we have $\|u^{k+1} - u\|_2 \leq (1 - \gamma^2/L^2)^{1/2} \|u^k - u\|_2$. A drawback of the method is that we need to know some (possibly nonoptimal) constants $\gamma, L$ satisfying (17), $\|A\|_2 \leq L$ in order to choose $\alpha$ (or we have to experiment with different values of $\alpha$).

2. 1-step minimum residual: We use (5) and choose $\alpha$ so that the norm of the new residual $r^{k+1} = b - Au^{k+1}$ becomes minimal:

$$\|r^{k+1}\|_2^2 = \|r^k - \alpha A r^k\|_2^2 = \|r^k\|_2^2 - 2\alpha \langle Ar^k, r^k \rangle + \alpha^2 \|Ar^k\|_2^2$$

i.e.,

$$\alpha_k := \frac{\langle Ar^k, r^k \rangle}{\|Ar^k\|_2^2}$$
then
\[ \|r^{k+1}\|_2^2 = \|r^k\|_2^2 - \frac{(A^k, r^k)^2}{\|A^k\|_2^2} \leq \left(1 - \frac{\gamma^2}{L^2}\right) \|r^k\|_2^2 \]

Therefore the residuals \( r^k = A(u - u^k) \) converge and
\[ \|r^k\|_2 \leq (1 - \gamma^2 / L^2)^{k/2} \|r^0\|_2, \quad \|u^k - u\|_2 \leq \gamma^{-1} \left(1 - \frac{\gamma^2}{L^2}\right)^{k/2} \|r^0\|_2 \]

This method corresponds to the first step of the GMRES method, or the GMRES(1) method which is restarted after every step. The full GMRES method minimizes the residuals over multiple directions, so the norm of the residual can only be lower. Hence the above estimates for \( \|r^k\|_2 \) and \( \|u^k - u\|_2 \) also hold for the GMRES method.

3 Equilibrium Problems

3.1 Abstract setting: Symmetric bilinear form \( a(u, v) \)

We want to find a “displacement” \( u \in V \). Here \( V \) is a complete vector space with a norm \( \|v\|_V \).

In the absence of external forces the equilibrium solution minimizes the “internal energy” \( Q(u) \).

If there are external forces: Now changing the displacement from \( u \) to \( u + v \) corresponds to a work \( \ell(v) \) (“work of external forces”). Here \( \ell: V \to \mathbb{R} \) is linear and bounded:
\[ \forall v \in V : \quad |\ell(v)| \leq C \|v\| \]

Principle of virtual displacements: The equilibrium solution \( u \in V \) minimizes
\[ F(u) := Q(u) - \ell(u) \]

For a linear problem the energy \( Q(u) \) is quadratic and has the form
\[ Q(u) = \frac{1}{2} a(u, u) \]

where \( a: V \times V \to \mathbb{R} \) is bilinear and bounded:
\[ \forall u, v \in V : \quad |a(u, v)| \leq L \|u\|_V \|v\|_V \]  \hspace{1cm} (6)

The bilinear form \( a \) is symmetric: \( a(u, v) = a(v, u) \) for all \( u, v \in V \).

We want that the energy \( Q(u) = \frac{1}{2} a(u, u) \) is positive for nonzero \( u \in V \), moreover we need a lower bound. We require that the bilinear form \( a(u, v) \) is coercive: there exists \( \gamma > 0 \) such that
\[ \forall u \in V : \quad a(u, u) \geq \gamma \|u\|_V^2 \]  \hspace{1cm} (7)

This property is also called “\( V \)-ellipticity”.

Minimization Problem (MIN): Find \( u \in V \) such that \( F(u) \) is minimal.

We claim that (MIN) is equivalent to the following linear problem:
Variational Problem (VAR): Find $u \in V$ such that

$$\forall v \in V: \quad a(u,v) = \ell(v)$$

**Theorem 3.1.** The following statements are equivalent:

1. $u$ minimizes $F(u)$ over all $u \in V$
2. $u$ satisfies (VAR)

**Proof.** (1) Assume $u \in V$ solves (MIN). Then for any $v \in V$ and any $\varepsilon \in \mathbb{R}$ we have

$$F(u) \leq F(u + \varepsilon v)$$

We have by using the bilinearity and symmetry of $a(\cdot, \cdot)$, and the linearity of $\ell(\cdot)$

$$F(u + \varepsilon v) = \frac{1}{2}a(u + \varepsilon v, u + \varepsilon v) - \ell(u + \varepsilon v) = \frac{1}{2}a(u,u) + \varepsilon a(u,v) + \frac{1}{2} \varepsilon^2 a(v,v) - \ell(u) - \varepsilon \ell(v)$$

$$F(u) = Q(u) + \varepsilon [a(u,v) - \ell(v)] + \frac{1}{2} \varepsilon^2 a(v,v)$$

Consider this as a function $f(\varepsilon)$ for $\varepsilon \in \mathbb{R}$. If this is minimal for $\varepsilon = 0$ then we must have $a(u,v) - \ell(v) = 0$ (otherwise we could achieve a smaller value with $\varepsilon \neq 0$).

(2) Assume $u$ satisfies (VAR). Then we have for any nonzero $v \in V$

$$F(u + v) = F(u) + [a(u,v) - \ell(v)] + \frac{1}{2} a(v,v)$$

0 by (VAR) $> 0$ since by coercivity $a(v,v) \geq \gamma \|v\|_V^2$ and $\|v\|_V > 0$ for nonzero $v$.

Note that $a(\cdot, \cdot)$ is an inner product on $V$ with the corresponding norm

$$\|u\|_a := a(u,u)^{1/2}$$

and by (13), (7) we have upper and lower bounds

$$\forall u \in V: \quad \gamma^{1/2} \|u\|_V \leq \|u\|_a \leq L^{1/2} \|u\|_V$$

**Theorem 3.2.** The variational problem (VAR) has a unique solution $u \in V$.

**Proof.** Note that $V$ with the inner product $a(\cdot, \cdot)$ forms a Hilbert space. The functional $\ell: V \to \mathbb{R}$ is linear and bounded:

$$|\ell(v)| \leq C \|v\|_a$$

with $C := \gamma^{-1/2} C_\ell$. By the Riesz representation theorem there exists a unique $u \in V$ such that

$$\forall v \in V: \quad a(u,v) = \ell(v)$$
3.2 Example Problem: Elastic string

The internal energy of the elastic string is given by

\[
\frac{1}{2} a(u, u), \quad a(u, v) := \int_{\Omega} c(x) u'(x) v'(x) dx
\]

where the "stiffness" \(c(x)\) satisfies

\[0 < c_{\text{min}} \leq c(x) \leq c_{\text{max}} \quad \text{for } x \in \Omega\]

For a force \(f(x)\) the work of the external force is given by

\[\ell(v) = \int_{\Omega} f(x)v(x) dx\]

Let \(H^1(\Omega)\) denote the space

\[H^1(\Omega) := \{ u \mid u \in L^2(\Omega) \text{ and } u' \in L^2(\Omega) \}\]

with the norm

\[\|u\|^2_{H^1(\Omega)} = \|u\|^2_{L^2(\Omega)} + \|u'\|^2_{L^2(\Omega)}\]

We define the space \(V\) as

\[V = H^1_0(\Omega) := \{ u \mid u \in H^1(\Omega), \ u = 0 \text{ on } \partial \Omega \}\]

with the norm \(\|u\|_{H^1}\).

We have the Poincare inequality: For \(u \in H^1_0(\Omega)\)

\[\|u\|_{L^2} \leq C_{\Omega} \|u'\|_{L^2}\]

Hence we can define \(\|u\|_V := \|u'\|_{L^2}\) as norm on the space \(V:\)

\[\|u'\|^2_{L^2} \leq \|u\|^2_{H^1} \leq (1 + C_{\Omega}) \|u'\|^2_{L^2}\]

Now we can show that \(a(\cdot, \cdot)\) satisfies the boundedness with \(C_a = c_{\text{max}}\): By Cauchy-Schwarz

\[a(u, v) \leq \int_{\Omega} c_{\text{max}} |u'| |v'| dx \leq c_{\text{max}} \|u'\|_{L^2} \|v'\|_{L^2}\]

We can also show the coercivity of \(a(\cdot, \cdot)\) with \(\gamma = c_{\text{min}}\):

\[a(u, u) = \int_{\Omega} c(x) u'(x)^2 dx \geq \int_{\Omega} c_{\text{min}} u'(x)^2 dx = c_{\text{min}} \|u'\|^2_{L^2}\]

Assume that the given force \(f(x)\) satisfies \(f \in L^2(\Omega)\). Then we can show that functional \(\ell: V \to \mathbb{R}\) is bounded with \(C_\ell = \|f\|_{L^2}\): By Cauchy-Schwarz

\[|\ell(v)| = \left| \int_{\Omega} f(x)v(x) dx \right| \leq \|f\|_{L^2} \|v\|_{L^2}\]

Hence we obtain from Theorems 3.1, 3.2 that for any given load \(f \in L^2(\Omega)\) the equilibrium problem has a unique solution \(u \in V\) satisfying (VAR).

3.3 Approximation of equilibrium solution

In our example the space \(V = H^1_0(\Omega)\) is infinite dimensional. On a computer we can only compute finitely many numbers.

Choose a finite dimensional subspace \(V_N \subset V\).
Minimization Problem (MIN\(_N\)): Find \( u_N \in V_N \) such that \( F(u_N) \) is minimal.

We claim that (MIN) is equivalent to the following linear problem:

Variational Problem (VAR\(_N\)): Find \( u_N \in V_N \) such that

\[
\forall v_N \in V_N: \quad a(u_N, v_N) = \ell(v_N)
\]

Since \( V_N \subset V \) all properties of \( a(u,v) \) and \( \ell(v) \) also hold for \( u, v \in V_N \). Therefore we obtain:

- \( u_N \) satisfies (MIN\(_N\)) \( \iff \) \( u_N \) satisfies (VAR\(_N\))
- problem (VAR\(_N\)) has a unique solution \( u_N \in V_N \)

Note that \( u_N \) is the best possible approximation of \( u \) in the space \( V_N \) with respect to \( \|\cdot\|_a \):

\[
\|u - u_N\|_a = \min_{v \in V_N} \|u - v_N\|_a
\]

To prove this we have to show that \( u_N \) satisfies the normal equations, i.e., \( a(u - u_N, v_N) = 0 \) for all \( v \in V_N \): This follows from

\[
a(u_N, v_N) \overset{(VAR)}{=} \ell(v_N) \overset{(VAR)}{=} a(u, v_N)
\]

In order to compute \( u_N \in V_N \) we choose a basis of \( V_N \). Assume \( \dim V_N = n \) we can pick a basis \( \phi_1, \ldots, \phi_n \).

Now we can write the solution \( u_N \in V_N \) as

\[
u_N = c_1 \phi_1 + \cdots + c_n \phi_n
\]

with a coefficient vector \( \vec{c} \in \mathbb{R}^n \). We need that (VAR\(_N\)) holds for \( v = \phi_1, \ldots, \phi_n \). This gives an \( n \times n \) linear system:

\[
A \vec{c} = \vec{b}
\]

where the so-called “stiffness matrix” \( A \in \mathbb{R}^{n \times n} \) is the Gram matrix with entries

\[
A_{jk} = a(\phi_k, \phi_j)
\]

and the right-hand side vector \( \vec{b} \in \mathbb{R}^n \) has entries

\[
b_j = \ell(\phi_j).
\]

This leads to the following algorithm:

1. Compute the entries of the stiffness matrix \( A \in \mathbb{R}^{n \times n} \)
2. Compute the entries of the right-hand-side vector \( b \in \mathbb{R}^n \)
3. Solve the \( n \times n \) linear system \( A \vec{c} = \vec{b} \)

We can use Gaussian elimination to solve the linear system. For a general \( n \times n \) matrix \( A \) this requires \( O(n^3) \) operations.

3.4 Variational problem with nonsymmetric \( a(u,v) \)

We also consider the case where the bilinear form \( a(u,v) \) satisfies (6) and (7), but \( a(u,v) \) is not symmetric. Note that in this case we don’t have a minimization problem (MIN), and we don’t have a norm \( \|\cdot\|_a \). So we need different proofs.

The variational formulation (VAR) has a unique solution by the Lax-Milgram theorem (2.3).

Again, we can choose a finite dimensional subspace \( V_N \subset V \) and consider the problem (VAR\(_N\)). By Lax-Milgram this problem has a unique solution \( u_N \).

Let \( v_N \in V_N \). By (VAR) we have \( a(u, v_N) = \ell(v_N) \), by (VAR\(_N\)) we have \( a(u_N, v_N) = \ell(v_N) \). Hence

\[
\forall v_N \in V_N: \quad a(u - u_N, v_N)
\]
Therefore we have to find \( \gamma \| u - u_N \|_V^2 \leq a(u - u_N, u - u_N) = a(u - u_N, u - v_N) \leq L \| u - u_N \|_V \| u - v_N \|_V \)

This gives the so-called "quasioptimality":

\[
\| u - u_N \|_V \leq \frac{L}{\gamma} \min_{v_N \in V_N} \| u - v_N \|_V
\]

I.e., the error is up to a constant \( \frac{L}{\gamma} \) optimal. Often we know something about the smoothness of \( u \), and we can prove the existence of approximations \( v_N \) with \( \| u - v_N \|_V \leq CN^{-\alpha} \). Then we obtain for the solutions \( u_N \) of \( (\text{VAR}_N) \) the same rate: \( \| u - u_N \|_V \leq \frac{L}{\gamma} CN^{-\alpha} \). We will use this idea to obtain convergence rates for the "p-version" and "h-version" methods.

### 3.5 p-version

Let \( \Omega = (-1, 1) \). We want to use finite dimensional spaces \( V_N \) so that we can obtain good approximations of our solution \( u(x) \).

One good choice are polynomials: For \( N \geq 2 \) we pick the space

\[
V_N = \{ u \in \mathcal{P}_N \mid u(-1) = u(1) = 0 \}
\]

and have \( n = \dim V_N = N - 1 \).

Since \( u \in V \) we have \( u(-1) = u(1) = 0 \) and therefore \( \int_{\Omega} u'(x)dx = u(1) - u(-1) = 0 \).

We now want to construct \( v_N \in V_N \) such that \( \| u - v_N \|_a \) is small. Note that

\[
\| u - v_N \|_a^2 \leq c_{\max} \| u' - v_N' \|_{L^2}^2
\]

Therefore we have to find \( q_{N-1} := u'_N \in \mathcal{P}_{N-1} \) such that \( \| u' - q_{N-1} \|_{L^2} \) is small. We define \( q_{N-1} \) as the \( L^2 \) projection of \( u' \) onto the space \( \mathcal{P}_{N-1} \). Then we must have \( \int_{\Omega}(u' - q_{N-1})r \, dx = 0 \) for all \( r \in \mathcal{P}_{N-1} \). In particular for \( r(x) = 1 \) we obtain \( \int_{\Omega}(u' - q_{N-1})dx = 0 \), hence \( \int_{\Omega}q_{N-1}(x)dx = 0 \). We now define \( v_N \) as the antiderivative

\[
v_N(x) := \int_{-1}^{x} q_{N-1}(t)dt,
\]

then clearly \( v_N \in \mathcal{P}_N \) and \( v_N(-1) = 0, v_N(1) = 0 \), i.e., \( v_N \in V_N \).

Now we can give an upper bound for \( \| u' - q_{N-1} \|_{L^2} \). Since \( q_{N-1} \) is the best \( L^2 \) approximation we have for any \( \tilde{q}_{N-1} \in \mathcal{P}_{N-1} \)

\[
\| u' - q_{N-1} \|_{L^2} \leq \| u' - \tilde{q}_{N-1} \|_{L^2} \leq \| u' - \tilde{q}_{N-1} \|_{L^2}
\]

where we use the weight function \( w(x) = (1 - x^2)^{-1/2} \geq 1 \) on \( \Omega = (-1, 1) \).

The function \( u' \) has a Chebyshev expansion

\[
u' = \sum_{k=0}^{\infty} a_k T_k
\]

and we choose \( \tilde{q}_{N-1} \in \mathcal{P}_{N-1} \) as the truncated Chebyshev expansion (this minimizes \( \| u' - \tilde{q}_{N-1} \|_{L^2} \))

\[
\tilde{q}_{N-1} := \sum_{k=0}^{N-1} a_k T_k
\]

Then we have

\[
\| u' - \tilde{q}_{N-1} \|_{L^2}^2 = \sum_{k=n}^{\infty} |a_k|^2 \frac{T_k^2}{\pi/2}
\]
Therefore we obtain the upper bound:

\[ \|u - u_N\|_a^2 \leq c_{\max} \frac{\pi}{2} \sum_{k=N}^{\infty} |a_k|^2 \]  \hspace{1cm} (9)

where \( a_k \) are the Chebyshev coefficients of the function \( u'(x) \).

We can use this to get results for an arbitrary interval \( \Omega = [x_{\text{left}}, x_{\text{right}}] \).

By using the results for the decay of \( |a_k| \) we obtain the following convergence rates:

- **if** \( u \) **has derivatives** \( u', \ldots, u^{(m)} \) **and** \( u^{(m)} \) **has bounded variation, with** \( m \geq 1 \):
  
  \[ \|u - u_N\|_{H^1} \leq C N^{-m} \]

  **Example:** for \( -u'' = \delta_0 \) on \( \Omega = (-1, 1) \) we obtain \( u(x) = (1 - |x|)/2 \). Here \( m = 1 \) since \( u'' \in BV \).

  Hence \( \|u - u_N\|_{H^1} \leq C n^{-1} \)

- **if** \( u \) **is analytic on** \( \Omega \):
  
  \[ \|u - u_N\|_{H^1} \leq C e^{-\beta N} \]

  with \( \beta > 0 \).

- **if** \( u \) **has the form** \( u(x) = c |x - x_0|^{\alpha} + \text{smoother terms with} \alpha > \frac{1}{2} \) **and** \( x_0 \in \Omega \):
  
  \[ \|u - u_N\|_{H^1} \leq C N^{-(\alpha - \frac{1}{2})} \]

Note that we used the estimate \( \|u' - \tilde{u}_{N-1}\|_{L^2} \leq \|u' - \tilde{u}_{N-1}\|_{L^2} \) in (8). Therefore (9) does not always give sharp results. E.g., for a singularity \( |x - x_{\text{left}}|^{\alpha} \) at the endpoint of the interval one can obtain with a different proof that \( \|u - u_N\|_{H^1} \leq C N^{-2(\alpha - \frac{1}{2})} \).

### 3.6 h-version

We have a domain \( \Omega \subset \mathbb{R}^d \) and we will consider the cases \( d = 1, d = 2 \) and \( d = 3 \).

**Error** \( \|f' - p'\|_{L^2([0,h])} \) **for linear interpolation on an interval**

Let \( f \in C([0,h]) \). Let \( p(x) = f(0) + \frac{f(h) - f(0)}{h} x \) denote the polynomial \( p \in \mathcal{P}_1 \) interpolating \( f \) at the nodes \( 0, h \). If \( f'' \) is continuous we know that \( |f(x) - p(x)| \leq \frac{h^2}{2} \|f''\|_{\infty} \) for \( x \in [0, h] \) from the classical formula for the interpolation error, i.e., the error is of order \( O(h^2) \). The error \( |f' - p'| \) for the derivatives is of order \( O(h^1) \):

1. Consider a function \( f(t) \) on the interval \( [0, 2\pi] \) with Fourier series \( f = \sum_{k=-\infty}^{\infty} c_k e^{ikt} \). The mean value of \( f \) on \( [0, 2\pi] \) is \( c_0 \), hence the Parseval identity for \( f(t) - c_0 = \sum_{k\neq 0} c_k e^{ikt} \) gives

   \[ \|f - c_0\|_{L^2([0,2\pi])}^2 = 2\pi \sum_{k\neq 0} |c_k|^2 \leq 2\pi \sum_{k=-\infty}^{\infty} k^2 |c_k|^2 = \|f'\|_{L^2([0,2\pi])}^2 \]  \hspace{1cm} (10)

   Here the last equality is the Parseval identity for \( f'(t) = \sum_{k=-\infty}^{\infty} c_k ike^{ikt} \).

2. Consider a function \( g(x) \) on an interval \( [0, h] \). Let \( c_0 = h^{-1} \int_0^h g(x)dx \) denote the mean value on \( [0, h] \). Let \( f(t) := g(t\frac{h}{2\pi}) \), then \( f'(t) = g'(t\frac{h}{2\pi}) \cdot \frac{h}{2\pi} \). Then (10) gives with the change of variables \( x = t\frac{h}{2\pi} \)

   \[ \|g - c_0\|_{L^2([0,h])}^2 \leq \left( \frac{h}{2\pi} \right)^2 \|g'\|_{L^2([0,h])}^2 \]  \hspace{1cm} (11)

3. Consider a function \( u(x) \) on an interval \( [0, h] \). Let \( p \in \mathcal{P}_1 \) denote the interpolating polynomial for the nodes \( 0, h \). Then

   \[ p'(x) = \frac{u(h) - u(0)}{h} = \frac{h}{2\pi} \int_0^h u'(t)dt, \]

i.e., \( p' \) is the mean value \( c_0 \) of the function \( g(x) := u'(x) \). Hence (11) gives

   \[ \|u' - p'\|_{L^2([0,h])} \leq \frac{h}{2\pi} \|u''\|_{L^2([0,h])} \]  \hspace{1cm} (12)
1-dimensional case

For $d = 1$ the domain is an interval $\Omega = (x_{\text{left}}, x_{\text{right}})$. We use a partition

$$x_{\text{left}} = x^0 < x^1 < \cdots < x^{N-1} < x^N = x_{\text{right}}$$

which divides the interval $(a, b)$ into $N$ subintervals and define

$$V_N := \left\{ u \in C(\Omega) \mid u|_{[x^{j-1}, x^j]} \text{ is linear for } j = 1, \ldots, N; u|_{\partial \Omega} = 0 \right\}$$

Let $h_j := x^j - x^{j-1}$ and $h_{\text{max}} := \max_{j=1,\ldots,N} h_j$. The dimension of $V_N$ is $n := N - 1$. A function $v \in V_N$ is specified by its nodal values $v_j := v(x^j)$, $j = 1, \ldots, n$. We use the nodal basis functions $\phi_1, \ldots, \phi_n$ for $V_N$: The function $\phi_j$ has nodal values $\phi_j(x^k) = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$.

The entries of the stiffness matrix $A$ and the right hand side vector $b$ are given by

$$a_{jk} = a(\phi_k, \phi_j), \quad b_j = \ell(\phi_j)$$

Note that $a(\phi_k, \phi_j) = 0$ if $|j - k| \geq 2$. Hence the matrix $A$ is tridiagonal, and the linear system $Ac = b$ can be solved with $O(n)$ operations.

For a function $u$ on $\Omega$ we denote by $u_I \in V_N$ the piecewise linear interpolation: $u_I := u(x^1)\phi_1 + \cdots + u(x^n)\phi_n$. Then we apply (12) on each subinterval $[x^{j-1}, x^j]$ and obtain

$$\|u' - u'_I\|_{L^2(\Omega)} \leq \frac{h_{\text{max}}}{2\pi} \|u''\|_{L^2(\Omega)}$$

This implies $\|u' - u'_I\|_2 \leq C h_{\text{max}}$. Therefore we obtain for the solutions $u_N$ of (VAR$_N$) from the quasioptimality

$$\|u - u_N\|_V \leq C' h_{\text{max}}$$

For equidistant nodes with $h_j = (x_{\text{right}} - x_{\text{left}})/N$ this gives $\|u - u_N\|_V \leq C'' N^{-1}$.

2-dimensional case

In two dimensions we assume that $\Omega$ is a polygon. We divide $\Omega$ into triangles $T_j$, $j = 1, \ldots, M$. The triangles should have size $h = N^{-1}$: we want that there are constants $c, C$ such

- each triangle fits into a circle of radius $Ch$
- we can fit a circle of radius $ch$ into the interior of each triangle

We define $V_N$ as

$$V_N := \left\{ u \in C(\Omega) \mid u|_{T_j} \text{ is linear for } j = 1, \ldots, M; u|_{\partial \Omega} = 0 \right\}$$

Let $x^1, \ldots, x^n$ denote the vertices of the mesh in the interior of $\Omega$. A function $v \in V_N$ is specified by the nodal values $v_j := v(x^j)$, $j = 1, \ldots, n$. We use the nodal basis functions $\phi_1, \ldots, \phi_n$ for $V_N$: The function $\phi_j$ has nodal values $\phi_j(x^k) = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$.

For a function $u$ on $\Omega$ we denote by $u_I \in V_N$ the piecewise linear interpolation: $u_I = u(x^1)\phi_1 + \cdots + u(x^n)\phi_n$. One can prove that

$$\|\nabla u - \nabla u_I\|_{L^2(\Omega)} \leq Ch \|u\|_{H^2(\Omega)}$$

where $\|u\|_{H^2(\Omega)} := \sum_{j,k=1}^2 \left\| \frac{\partial^2 u}{\partial x_j \partial x_k} \right\|_{L^2(\Omega)}^2$.

Therefore we obtain for the solutions $u_N$ of (VAR$_N$) from the quasioptimality

$$\|u - u_N\|_V \leq C' h$$
In two dimensions we assume that \( \Omega \) is a polyhedron. We divide \( \Omega \) into tetrahedra \( T_j, j = 1, \ldots, M \). The tetrahedra should have size \( h = N^{-1} \); they should fit inside a sphere of radius \( Ch \), and they should contain a sphere of radius \( ch \). We can now proceed as in the 2-dimensional case and obtain from the quasioptimality that \( \|u - u_N\|_V \leq C'h \).

4 Solution of the linear system: Direct solvers versus iterative solvers

In many applications we have to solve a linear system \( Ax = b \) with \( A \in \mathbb{R}^{n \times n} \) and \( b \in \mathbb{R}^n \) given. If \( n \) is large the solution of the linear system takes a lot of operations, and standard Gaussian elimination may take too long.

But in many cases most entries of the matrix \( A \) are zero and \( A \) is a so-called **sparse matrix**. This means each equation only couples very few of the \( n \) unknowns \( x_1, \ldots, x_n \). A typical example are discretizations of partial differential equations, see next section for an example.

**Direct solvers** will give the exact solution after finitely many operations (if we ignore roundoff errors).

**Gaussian elimination with partial pivoting:** This gives a decomposition \( LU = \begin{bmatrix} \text{row } p_1 \text{ of } A \\ \vdots \\ \text{row } p_n \text{ of } A \end{bmatrix} \) where \( L \) is lower triangular, \( U \) is upper triangular.

**Cholesky decomposition:** We need that \( A \) is symmetric positive definite. This gives a decomposition \( A = LL^\top \) where \( L \) is lower triangular.

Cholesky decomposition takes about half the number of operations of Gaussian elimination.

**Cost for Gaussian elimination and Cholesky algorithm:**

- for **full matrices** finding the decomposition takes \( Cn^3 \) operations. Once have the decomposition solving the linear system for a given vector \( b \) takes \( n^2 \) operations.
- for **band matrices with bandwidth** \( m \), i.e., \( A_{ij} = 0 \) for \( |i - j| > m \): Finding the decomposition takes \( Cm^2n \) operations, solving a linear system then takes \( Cmn \) operations.

In Matlab we should initialize the matrix \( A \) as a sparse matrix structure. Then Matlab will only use storage and operations to compute the nonzero elements of \( L, U \).

For a matrix \( A \) with bandwidth \( m \) the factors \( L, U \) will also have bandwidth \( m \). For a general sparse matrix \( A \), the factors \( L, U \) will usually have additional nonzero elements at locations where \( A \) had zero elements. This is called **fill-in**, and this increases the number of operations.

**Reordering:** If you use the Matlab command \( x=A\backslash b \) (where the matrix has sparse array type) then Matlab will try to renumber the unknowns in such a way that the amount of fill-in will minimized. This can substantially reduce the number of operations.

The command \texttt{spparms(’spumoni’,2)} makes Matlab print out details about the algorithms used for each following \( \backslash \) command. Try the following example for solving \(-\Delta u = 1\) on the square \( \Omega = (0,1)^2 \):

\[
N = 50; \\
A = delsq(numgrid(’S’,N+1)); \\
figure(1); spy(A) \\
n = size(A,1); \\
b = ones(n,1)/N^2; \\
spparms(’spumoni’,2) \quad \% \text{print out details about what } \backslash \quad \text{does} \\
u = A\backslash b; \\
figure(2); surf(reshape(u,N-1,N-1))
\]
There are also versions of the \texttt{lu} and \texttt{chol} commands that use reordering to minimize fill-in.

\section{Convection diffusion problem in $\mathbb{R}^d$}

We consider a typical application problem which leads to a large sparse linear system. Equilibrium problems for elastic deformations or heat transfer lead to elliptic differential equations.

\textbf{Boundary value problem}

In the convection diffusion problem we have a domain $\Omega \subset \mathbb{R}^d$. For $d = 1$ we consider an interval, for $d = 2$ we consider a polygon, for $d = 3$ we consider a polyhedron.

We want to find a function $u(x)$ for $x \in \Omega$ such that

$$-\Delta u + b \cdot \nabla u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on the boundary } \partial \Omega$$

where $b \in \mathbb{R}^n$ is a constant vector, and $f$ is a given function on $\Omega$. This is called a \textbf{boundary value problem}.

\textbf{Variational formulation}

We first want to find the “variational formulation”: If we multiply the PDE by a “test function” $v$ which is zero on the boundary and integrate over $\Omega$ we obtain after using the first Green formula

\begin{equation}
\int_{\Omega} \left[ \nabla u \cdot \nabla v + (b \cdot \nabla u)v \right] \, dx = \int_{\partial \Omega} (\partial_n u) v \, ds
\end{equation}

We use the Hilbert space $V = H^1_0(\Omega) = \{ u | \int_{\Omega} \left( \| \nabla u \|^2_2 + |u|^2 \right) \, dx < \infty, \ u|_{\partial \Omega} = 0 \}$ with the norm

$$\| u \|^2_V = \int_{\Omega} \left( \| \nabla u \|^2_2 + |u|^2 \right) \, dx = \| \nabla u \|^2_{L^2(\Omega)} + \| u \|^2_{L^2(\Omega)}.$$

Note that $\int_{\Omega} (b \cdot \nabla u) u \, dx = \int_{\Omega} (b_1 u_{x_1} + b_2 u_{x_2}) u \, dx = 0$: For $\int_{x_2=a}^{x_2=b} \int_{x_1=a(x_2)}^{x_1=b(x_2)} u_{x_1} u \, dx_1 \, dx_2 = -\int_{x_1=a}^{x_1=b} u_{x_1} u \, dx_1$ since $u$ is zero on the boundary.

We then obtain that $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is a bilinear form satisfying for all $u, v \in V$

\begin{align}
|a(u, v)| &\leq L_a \| u \|_V \| v \|_V \\
\| a(u, u) \|_V &\geq \gamma_u \| u \|^2_V
\end{align}

The first inequality follows from the Cauchy-Schwartz inequality. For the second inequality we use $a(u, v) = \int \nabla u \cdot \nabla u \, dx$ and the Poincare inequality $\| u \|^2_{L^2(\Omega)} \leq C_{\Omega} \| \nabla u \|^2_{L^2(\Omega)}$.

We obtain that $\ell : V \to \mathbb{R}$ is a linear functional such that

$$|\ell(v)| \leq C_{\ell} \| v \|_V.$$

The \textbf{variational formulation} is: Find $u \in V$ such that

$$\forall v \in V : \quad a(u, v) = \ell(v)$$

By the Lax-Milgram theorem (see Appendix A below) this variational problem has a unique solution $u \in V$. 

13
Finite element discretization

We choose a finite dimensional subspace $V_h \subset V$.

For $d = 2$ the domain $\Omega$ is a polygon, and we divide it into a mesh of triangles. (For $d = 1$ we divide the interval into subintervals, for $d = 3$ we divide the polyhedron into a mesh of tetrahedra).

Then we define $V_h$ as the space of piecewise linear functions on the mesh which are continuous in $\Omega$ and are zero on the boundary $\partial \Omega$.

The discrete problem is: Find $u_h \in V_h$ such that

$$\forall v_h \in V_h : \quad a(u_h, v_h) = \ell(v_h)$$

(15)

Since $V_h \subset V$ the inequalities (13), (14) are satisfied for $u, v \in V_h$. Hence by the Lax-Milgram theorem the discrete problem has a unique solution $u_h$.

We can specify a function $v_h \in V_h$ by specifying the values $v_1, \ldots, v_n$ at the interior nodes $x_1, \ldots, x_n$ of the mesh. The basis function $\phi_j$ is the function in $V_h$ with $\phi_j(x_j) = 1$ and $\phi(x_k) = 0$ for $k \neq j$. We can then write $u_h$ as

$$u_h = u_1 \phi_1 + \cdots + u_n \phi_n$$

where $u =$ \begin{equation*} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \end{equation*} \in \mathbb{R}^n$ is the coefficient vector. Now (15) for $v_h = \phi_1, \ldots, \phi_n$ gives the linear system

$$Au = b, \quad A_{jk} = a(\phi_k, \phi_j), \quad b_j = \ell(\phi_j).$$

Therefore the finite element method involves the following steps:

- pick a mesh on $\Omega$
- assemble the stiffness matrix $A$ and the right hand side vector $b$
- solve the linear system $Au = b$

Work for direct solvers

For $d = 1$ we obtain a tridiagonal matrix $A$. Hence the work is proportional to $N = 1/h$.

As a simple example for $d = 3$ consider the cube $\Omega = (0, 1)^3$. Let $h = 1/N$ with positive integer $N$. By using uniform grids $x_1 = j_1/N, x_2 = j_2/N, x_3 = j_3/N$ with $j_1, j_2, j_3 \in \{0, \ldots, N\}$ for each coordinate we can subdivide $\Omega$ into $N^3$ smaller cubes. We can then subdivide each of the smaller cubes into tetrahedra. We have $n = (N - 1)^3$ interior nodes with $j_1, j_2, j_3 \in \{1, \ldots, N - 1\}$, and we can order them lexicographically by $(j_1, j_2, j_3)$: $(1, 1, 1), \ldots, (1, N - 1), (1, 2, 1), \ldots, (N, N, N)$. Then the resulting stiffness matrix $A$ has size $n \times n$ with $n = (N - 1)^3$, and bandwidth $(N - 1)^2$.

This will also hold for a more general domain $\Omega \subset \mathbb{R}^d$, assuming that all triangles/tetrahedra are of “size $h$, up to a constant”: We will have $n = \dim V_h \approx ch^{-d}$ and a bandwidth $m \approx c'h^{1-d}$. Therefore the work for Gaussian elimination is with $h \approx cN^{-1}$

$$m^2n \approx C\left(N^{d-1}\right)^2 N^d = CN^{3d-2}$$

For $d = 2$ we have therefore $O(N^4)$ operations. For $d = 3$ we have $O(N^7)$ operations. Using Matlab’s reordering algorithms reduces the work to $N^3$ for $d = 2$, but for $d = 3$ it does not improve the rate $O(N^7)$.

Work of direct solvers for the convection-diffusion problem:

<table>
<thead>
<tr>
<th></th>
<th>$d = 1$</th>
<th>$d = 2$</th>
<th>$d = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian elimination (using band structure)</td>
<td>$N^1$</td>
<td>$N^4$</td>
<td>$N^7$</td>
</tr>
<tr>
<td>Gaussian elimination with reordering</td>
<td>$N^1$</td>
<td>$N^3$</td>
<td>$N^7$</td>
</tr>
</tbody>
</table>
Estimates \((A\vec{u}, \vec{v}) \leq L \|\vec{u}\|_2 \|\vec{v}\|_2\) and \((A\vec{u}, \vec{u}) \geq \gamma \|\vec{u}\|_2^2\) for the stiffness matrix \(A\)

A function \(v_h \in V_h\) is given by a coefficient vector \(\vec{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}\). For the function \(v_h\) we have the norms \(\|v_h\|_{L^2}\) and \(\|\nabla v_h\|_{L^2}\). How are these norms related to the norm \(\|\vec{v}\|_2\) the coefficient vector \(\vec{v}\)?

We assume that all triangles are of “size \(h\), up to a constant”. More precisely: We assume that a circle of radius \(c_0 h\) fits inside each triangle, and each triangle fits inside a circle of radius \(C_0 h\). Then one can show that there exist constants \(c_1, c_2, c_3\) depending on \(c_0\) and \(C_0\) such that

\[
c_1 h^{d/2} \|\vec{v}\|_2 \leq \|v_h\|_{L^2} \leq c_2 h^{d/2} \|\vec{v}\|_2
\]

This implies

\[
\|v_h\|_{H^1}^2 = \|\nabla v_h\|_{L^2}^2 + \|v_h\|_{L^2}^2 \leq c' h^{d-2} \|\vec{v}\|_2^2
\]

Hence we obtain for functions \(u_h, v_h \in V_h\) with coefficient vectors \(\vec{u}, \vec{v}\)

\[
(A\vec{u}, \vec{v}) = |a(u_h, v_h)| \leq L_a \|u_h\|_{H^1} \|v_h\|_{H^1} \leq L_a c' h^{d-2} \|\vec{u}\|_2 \|\vec{v}\|_2
\]

\[
(A\vec{u}, \vec{u}) = a(u_h, u_h) \geq \gamma_a \|u_h\|_{L^2}^2 \geq \gamma_a c_1^2 h^d \|\vec{u}\|_2^2
\]

Hence we obtain

\[
L = \|A\|_2 \leq C h^{d-2}, \quad \gamma = \lambda_{\text{min}} \left( \frac{1}{2}(A + A^T) \right) \geq C' h^d
\]

We can split the bilinear form \(a(u, v)\) into a diffusion part and a convection part

\[
a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx + \int_{\Omega} (b \cdot \nabla u) v \, dx,
\]

hence we have \(A = A_{\text{diff}} + A_{\text{conv}}\) with

\[
|A_{\text{diff}} \vec{u}, \vec{v}| = \left| \int_{\Omega} \nabla u_h \cdot \nabla v_h \, dx \right| \leq \|\nabla u_h\|_{L^2} \|\nabla v_h\|_{L^2} \leq c_3 h^{d-2} \|\vec{u}\|_2 \|\vec{v}\|_2
\]

\[
|A_{\text{conv}} \vec{u}, \vec{v}| = \left| \int_{\Omega} \nabla u_h \cdot (b v_h) \, dx \right| \leq \|\nabla u_h\|_{L^2} \|b\|_{L^2} \|v_h\|_{L^2} \leq c_3 c_2 \|b\|_{L^2} h^{d-1} \|\vec{u}\|_2 \|\vec{v}\|_2
\]

6 1-step minimum residual method aka GMRES(1)

We want to solve the linear system \(Au = b\) where \(A \in \mathbb{R}^{n \times n}\) and \(b \in \mathbb{R}^n\). We assume that \(A\) is positive definite:

\[
(Au, u) \geq \gamma \|u\|_2^2 \quad \text{for all } u \in \mathbb{R}^n
\]

The current guess is \(u^{(k)}\). We compute the residual \(r^{(k)} := b - Au^{(k)}\) and define the new guess as

\[
u^{(k+1)} := u^{(k)} + \alpha_k r^{(k)}
\]

where we choose \(\alpha_k\) such that the new residual has \(r^{(k+1)} := b - Au^{(k+1)}\) has minimum norm \(\|r^{(k+1)}\|_2\), yielding

\[
\alpha_k := \frac{(A r^{(k)}, r^{(k)})}{\|A r^{(k)}\|_2^2}
\]

Note that each step requires one matrix-vector product \(A r^{(k)}\) (and a few inner products of vectors).
If the matrix $A$ satisfies (16), (17) we obtained earlier (see Appendix A below)

$$\|r^{(k+1)}\|_2 \leq (1 - K^{-1})^{1/2} \|r^{(k)}\|_2$$

with $K = \left( \frac{L}{\gamma} \right)^2$ (18)

implying

$$\|r^{(k)}\|_2 \leq (1 - K^{-1})^{k/2} \|r^{(0)}\|_2, \quad \|u^{(k)} - u^{(0)}\|_2 \leq \tilde{\gamma}^{-1} (1 - K^{-1})^{k/2} \|r^{(0)}\|_2$$

(19)

Let $\kappa := L/\gamma$. Since $\|A^{-1}\|_2 \leq \gamma^{-1}$ we have $\text{cond}_2(A) \leq \kappa:

$$\text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 \leq L\gamma^{-1} = \kappa.$$}

If the matrix $A$ is symmetric we have $\text{cond}_2(A) = \kappa$:

$$L = \|A\|_2 = \lambda_{\text{max}}(A), \quad \gamma = \lambda_{\text{min}}(A), \quad \text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = L\gamma^{-1} = \kappa.$$}

Assume that we have $\|r^{(k+1)}\| \leq q \|r^{(k)}\|$ with $q = 1 - \varepsilon$. In order to achieve $\|r^{(k)}\|_2 \leq \delta$ we need to pick $k$ such that

$$\|r^{(k)}\|_2 \leq q^k \|r^{(0)}\|_2 \leq \delta, \quad \text{hence} \quad k \geq \frac{\log(\delta/\|r^{(0)}\|_2)}{\log q}$$

For $q = 1 - \varepsilon$ the first order Taylor approximation gives $\log(1 - \varepsilon) \approx -\varepsilon$, hence we need approximatively

$$k \geq \varepsilon^{-1} \log \left( \frac{\|r^{(0)}\|_2}{\delta} \right) = \varepsilon^{-1} C_{\varepsilon}$$

steps for the iterative method.

Here we have $q = (1 - K^{-1})^{1/2} \approx 1 - \frac{1}{2} K^{-1}$ using Taylor. Hence we need

$$k \geq 2C_{\varepsilon} K = 2C_{\varepsilon} \kappa^2$$

steps for the iterative method.

Note that for the convection diffusion problem we have $\kappa = Ch^{-2}$ and $q \leq 1 - ch^4$. Therefore it would seem that we need $Ch^{-4}$ steps of our iterative method. But it turns out that this estimate is too pessimistic. Actually we have $q \leq 1 - ch^2$ and we need only $Ch^{-2}$ steps of our iterative method as we will see in the next section.

7 Sharper estimates for the convergence factor

Symmetric case

Recall that

$$\|r^{(k+1)}\|_2 = \|(I - \alpha A)r^{(k)}\|_2 \leq \|I - \alpha A\|_2 \|r^{(k)}\|_2.$$}

If $A$ is symmetric, then also $I - \alpha A$ is symmetric and we have with the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $A$

$$\|I - \alpha A\|_2 = \max_{j=1,\ldots,n} |1 - \alpha \lambda_j|$$

(20)

If $A$ is positive definite, the eigenvalues are positive, and we can minimize (20) by choosing $\alpha$ such that

$$-1 + \alpha \lambda_{\text{max}} = -(1 - \alpha \lambda_{\text{min}})$$

yielding

$$\alpha = \frac{2}{\lambda_{\text{max}} + \lambda_{\text{min}}}, \quad \|I - \alpha A\|_2 = \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}} + \lambda_{\text{min}}} = 1 - \frac{2}{\kappa + 1} \quad \text{with} \quad \kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$$

So the convergence factor is $q = 1 - \frac{2}{\kappa + 1}$ and the number of iterations is proportional to $\kappa = \text{cond}_2(A)$ (and not $\kappa^2$ as the earlier estimate (18) would suggest).
Nonsymmetric case

We can write \( A \) as a sum of a symmetric part \( H \) and antisymmetric part \( S \):

\[
A = H + S, \quad H := \frac{1}{2}(A + A^\top), \quad S := \frac{1}{2}(A - A^\top)
\]

We assume that \( A \) is positive definite, i.e., \((Av, v) = (Hv, v) > 0\) for \( v \neq 0 \). Let \( u \) denote the current guess, and \( r := b - Au \) the residual.

The next approximation is \( u^{\text{new}} = u + \alpha r \), with the residual \( r^{\text{new}} = b - Au^{\text{new}} = (I - \alpha A)r \). Hence

\[
\|r^{\text{new}}\|^2 = ((I - \alpha A)r, (I - \alpha A)r) = \|r\|^2 - 2\alpha (Ar, r) + \alpha^2 \|Ar\|^2
\]

Note that both \( \|Ar\| \) and \( (Ar, r)^{1/2} = (Hr, r)^{1/2} \) define norms on \( \mathbb{R}^n \). Therefore there exists \( C > 0 \) such that

\[
\|Ar\|^2 \leq C (Ar, r) \quad \text{for all } r \in \mathbb{R}^d
\]  

(21)

Then

\[
\|r^{\text{new}}\|^2 \leq \|r\|^2 + [ -2\alpha + C\alpha^2 ] (Ar, r)
\]

The bound is minimal for \( \alpha = C^{-1} \), and with this we get

\[
\|r^{\text{new}}\|^2 \leq \|r\|^2 - C^{-1} (Ar, r) \leq \left[ 1 - \frac{\gamma}{C} \right] \|r\|^2
\]

It remains to find \( C \) such that (21) holds: with \( v = Ar \) we get

\[
(Ar, r) = (Hr, r) = (HA^{-1}v, A^{-1}v)
\]

Then we obtain \( (v, B^{-1}w) \leq C(w, w) \) with \( C = \lambda_{\text{min}}(B^{-1}) = \lambda_{\text{max}}(B^{-1}) \) since \( B \) is symmetric. Hence we need an estimate \( (w, B^{-1}w) \leq C(w, w) \) with \( B^{-1} = AH^{-1}A^\top \). Using \( A^\top = H - S \) we get

\[
(w, B^{-1}w) = ((H - S)w, H^{-1}(H - S)w) = (Hw, w) - (Sw, w) - (Hw, H^{-1}Sw) + (Sw, H^{-1}Sw) \\
\leq \lambda_{\text{max}}(H) \|w\|^2 + \lambda_{\text{min}}(H) \|Sw\|^2
\]

Since \( (Sw, Sw) = (-S^2w, w) \leq \rho(S)^2 \|w\|^2 \) we obtain

\[
C = \lambda_{\text{max}}(H) + \frac{\rho(S)^2}{\lambda_{\text{min}}(H)}
\]

Note:

\begin{itemize}
  \item the eigenvalues of \( H \) are real and positive.
  \item the eigenvalues of \( S \) are of the form \( \pm \alpha_j, \gamma \) with \( \alpha_j \geq 0 \).
\end{itemize}

Proof: Let \( \mu_j \) denote the eigenvalues of \( \tilde{S} \). The matrix \( S^2 \) is symmetric and has real eigenvalues \( \mu_j^2 \leq 0 \) because \( (S^2w, w) = -(Sw, Sw) \leq 0 \). Since the matrix \( S \) is real, taking the complex conjugate of \( Sv = \mu_j v \) gives \( S\bar{v} = \bar{\mu}_j \bar{v} \). Hence \( \bar{\mu}_j \) is also an eigenvalue of \( S \).

Theorem 7.1. Let \( A \in \mathbb{R}^{n \times n} \), let \( H := \frac{1}{2}(A + A^\top), \quad S := \frac{1}{2}(A - A^\top) \). If \( A \) is positive definite, i.e., \( \lambda_{\text{min}}(H) > 0 \) the 1-step minimum residual iteration satisfies

\[
\left\|r^{(k+1)}\right\|_2 \leq \left(1 - K^{-1}\right)^{1/2} \left\|r^{(k)}\right\|_2 \quad K := \text{cond}_2(H) + \left(\frac{\rho(S)}{\lambda_{\text{min}}(H)}\right)^2
\]  

(22)

Note: The number of iterations is proportional to \( K \). In our earlier estimate (19) we had \( K = \left(\frac{\|A\|_2}{\gamma}\right)^2 \) whereas we now obtain

\[
K = \frac{\|H\|_2}{\gamma} + \left(\frac{\|S\|_2}{\gamma}\right)^2
\]

This shows that for a symmetric matrix \( A = H \) the number of steps is proportional to the condition number. If we have nonsymmetric \( A = H + S \) then \( K \) increases by \( \left(\frac{\|S\|_2}{\gamma}\right)^2 \). So we see that the quadratic term \( \left(\frac{\|A\|_2}{\gamma}\right)^2 \) in our earlier estimate is actually only caused by the antisymmetric part \( S \).
Application to convection diffusion problem

Recall that \( A = A_{\text{diff}} + A_{\text{conv}} \) with the symmetric matrix \( H = A_{\text{diff}} \) and the antisymmetric matrix \( S = A_{\text{conv}} \) and

\[
C_1 h^d \|u\|_2^2 \leq (A_{\text{diff}}u, u) \leq C_2 h^{d-2} \|u\|_2^2
\]
\[
|(A_{\text{conv}}u, v)| \leq C_3 h^{d-1} \|u\|_2 \|v\|_2
\]

Therefore we have

\[
\lambda_{\text{min}}(H) \geq C_1 h^d, \quad \lambda_{\text{max}}(H) \leq C_2 h^{d-2}, \quad \rho(S) = \|S\|_2 \leq C_3 h^{d-1}
\]

yielding with (22)

\[
K := \text{cond}_2(H) + \left( \frac{\rho(S)}{\lambda_{\text{min}}(H)} \right)^2 \leq \frac{C_2}{C_1} h^{-2} + \left( \frac{C_3}{C_1} h^{-1} \right)^2 = C h^{-2}, \quad q = \left( 1 - C^{-1} h^2 \right)^{1/2}
\]

This means that we need \( C' h^{-2} \) steps of the iterative method to reduce the norm of the residual by a fixed factor.

**Note:** \( C_3 \) is proportional to \( \|b\|_2 \), so we obtain \( K = \left( C + C' \|b\|_2^2 \right) h^{-2} \). So for a problem with strong convection the number of iterations can be very large.

Recall that the stiffness matrix \( A \) is of size \( n \times n \) with \( cn \) nonzero elements where \( n \approx ch^{-d} \approx cN^d \) (for the meshsize \( h \approx 1/N \)).

Therefore the work of a matrix-vector product is given by the number of nonzero matrix elements \( cN^d \).

The work of one step of the 1-step min. res. method is one matrix-vector product, and some inner products, so the work per step is \( c'N^d \).

The number of steps is proportional to \( h^{-2} \approx N^2 \) if we want to achieve a residual with \( \|r\|_2 \leq \delta \). Hence the total work for our iterative method with \( q = 1 - ch^2 \) is

\[
CN^2 N^d
\]

If we had an iterative method with \( q = 1 - ch \) we would obtain a total work of \( CNN^d \) instead.

**Summary:** Work for solving convection-diffusion problem in \( \Omega \subset \mathbb{R}^d \) with meshsize \( h = 1/N \)

The number of unknowns is \( n = O(N^d) \). An ideal algorithm would use \( O(n) = O(N^d) \) operations.

<table>
<thead>
<tr>
<th></th>
<th>( d = 1 )</th>
<th>( d = 2 )</th>
<th>( d = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian elimination (using band structure)</td>
<td>( N^1 )</td>
<td>( N^4 )</td>
<td>( N^7 )</td>
</tr>
<tr>
<td>Gaussian elimination with reordering</td>
<td>( N^1 )</td>
<td>( N^3 )</td>
<td>( N^7 )</td>
</tr>
<tr>
<td>GMRES(1) method: ( q = 1 - ch^2 )</td>
<td>( N^3 )</td>
<td>( N^4 )</td>
<td>( N^5 )</td>
</tr>
<tr>
<td>CG method for symmetric ( A ): ( q = 1 - ch )</td>
<td>( N^2 )</td>
<td>( N^3 )</td>
<td>( N^4 )</td>
</tr>
</tbody>
</table>

Note that for \( d = 1 \) using an iterative method is pointless. For \( d = 2 \) the direct solver with reordering is better than the 1-step min. res. method. For \( d = 3 \) the iterative method is clearly better than the direct method.

In the case of a symmetric matrix we can construct an iterative method with \( q = 1 - ch \). This is the **conjugate gradient method** which we will discuss next.