Chapter 7

Iterative methods for linear systems

In this chapter we revisit the problem of solving linear systems of equations, but now in the context of large sparse systems for which direct methods are too expensive, in memory and execution time.

We introduce instead iterative methods, for which matrix sparsity is exploited to develop fast algorithms with a low memory footprint.

7.1 Stationary iterative methods

Iterative methods

For a given nonsingular matrix $A \in \mathbb{R}^{n\times n}$ and vector $b \in \mathbb{R}^n$, we consider the problem of finding a vector $x \in \mathbb{R}^n$, such that

$$Ax = b, \tag{7.1}$$

where the size n of the system is large, and the matrix A is *sparse* with the number of nonzero elements being $O(n)$ and not $O(n^2)$.

We do not seek to construct the exact solution $x = A^{-1}b$, but instead we will develop *iterative methods* based on algorithms that generate a sequence of approximations $\{x^{(k)}\}_{k\geq 0}$ that converges towards x, with

$$e^{(k)} = x - x^{(k)}, \tag{7.2}$$

the *error* at iteration k.

The error is not directly computable since the exact solution is unknown, but the error can be expressed in terms of the *residual* $r^{(k)} = b - Ax^{(k)}$, as

$$r^{(k)} = b - Ax^{(k)} = Ax - Ax^{(k)} = Ae^{(k)}, \tag{7.3}$$

so that for $\|\cdot\| = \|\cdot\|_2$, we have that

$$\|e^{(k)}\| = \|A^{-1}r^{(k)}\| \leq \|A^{-1}\|\|r^{(k)}\|, \tag{7.4}$$

and similarly

$$\|r^{(0)}\| = \|Ae^{(0)}\| \leq \|A\|\|e^{(0)}\|. \tag{7.5}$$

The *condition number* of A relative to the norm $\|\cdot\|$ is defined as

$$\kappa(A) = \|A\|\|A^{-1}\|, \tag{7.6}$$

which together with (7.4) and (7.5) provides an estimate of the relative error in terms of the relative residual.

Theorem 10 (Error estimate)**.** *For $\{x^{(k)}\}_{k\geq 0}$ a sequence of approximate solutions to the linear system of equations $Ax = b$, the relative error can be estimated as*

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \kappa(A)\frac{\|r^{(k)}\|}{\|r^{(0)}\|}. \tag{7.7}$$

The error estimate (7.7) may be used as a stopping criterion for when to terminate an iterative algorithm,

$$\frac{\|r^{(k)}\|}{\|r^{(0)}\|} < TOL, \tag{7.8}$$

with $TOL > 0$ the chosen tolerance.

Although, to use the relative error with respect to the initial approximation can be problematic, since the choice of $x^{(0)}$ may be completely arbitrary and not of significance for the problem at hand. Instead it is more suitable to formulate a stopping criterion based on the following condition,

$$\frac{\|r^{(k)}\|}{\|b\|} < TOL, \tag{7.9}$$

corresponding to $x^{(0)} = 0$.

Stationary iterative methods

Stationary iterative methods are formulated as a linear *fixed point iteration* of the form

$$x^{(k+1)} = Mx^{(k)} + c, \tag{7.10}$$

with $M \in \mathbb{R}^{n\times n}$ the *iteration matrix*, $\{x^{(k)}\}_{k\geq 0} \subset \mathbb{R}^n$ a sequence of approximations, and $c \in \mathbb{R}^n$ a vector.

Theorem 11 (Banach fixed point theorem for matrices)**.** *If* $\|M\| < 1$*, the fixed point iteration (7.10) converges to the solution of the equation* $x = Mx + c$*.*

Proof. For any $k > 1$, we have that

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &= \|Mx^{(k)} - Mx^{(k-1)}\| = \|M(x^{(k)} - x^{(k-1)})\| \\ &\leq \|M\|\|x^{(k)} - x^{(k-1)}\| \leq \|M\|^k \|x^{(1)} - x^{(0)}\|. \end{aligned}$$

Further, for $m > n$,

$$\begin{aligned} \|x^{(m)} - x^{(n)}\| &= \|x^{(m)} - x^{(m-1)}\| + ... + \|x^{(n+1)} - x^{(n)}\| \\ &\leq \left(\|M\|^{m-1} + ... + \|M\|^n\right) \|x^{(1)} - x^{(0)}\|, \end{aligned}$$

so that with $\|M\| < 1$. We thus have that

$$\lim_{n\to\infty} \|x^{(m)} - x^{(n)}\| = 0, \tag{7.11}$$

that is $\{x^{(n)}\}_{n=1}^{\infty}$ is a *Cauchy sequence*, and since the vector space $\mathbb{R}^n$ is *complete*, all Cauchy sequences converge, so there exists an $x \in \mathbb{R}^n$ such that

$$x = \lim_{n\to\infty} x^{(n)}. \tag{7.12}$$

By taking the limit of both sides of (7.10) we find that x satisfies the equation $x = Mx + c$. □

Further, an equivalent condition for convergence is that the *spectral radius* $\rho(M) < 1$, with

$$\rho(A) = \max_{\lambda \in \Lambda(A)} |\lambda|. \tag{7.13}$$

In particular, for a real symmetric matrix A, the spectral radius is identical to the induced 2-norm, that is $\rho(A) = \|A\|_2$.

Richardson iteration

The linear system $Ax = b$ can be formulated as a fixed point iteration through the *Richardson iteration*, with an iteration matrix $M = I - A$,

$$x^{(k+1)} = (I - A)x^{(k)} + b, \tag{7.14}$$

which will converge if $\|I - A\| < 1$, or $\rho(A) < 1$. We note that for an initial approximation $x^{(0)} = 0$, we obtain for $k = 0$,

$$x^{(1)} = (I - A)x^{(0)} + b = b$$

for $k = 1$,

$$x^{(2)} = (I - A)x^{(1)} + b = (I - A)b + b = 2b - Ab,$$

for $k = 2$,

$$x^{(3)} = (I - A)x^{(2)} + b = (I - A)(2b - Ab) + b = 3b - 3Ab + A^2b,$$

and more generally, that the iterate $x^{(k)}$ is a linear combination of powers of the matrix A acting on b, that is

$$x^{(k)} = \sum_{i=0}^{k-1} \alpha_i A^i b, \tag{7.15}$$

with $\alpha_i \in \mathbb{R}$.

Preconditioned Richardson iteration

To improve convergence of Richardson iteration we can *precondition* the system $Ax = b$ by multiplication of both sides of the equation by a matrix B, so that we get the new system

$$BAx = Bb, \tag{7.16}$$

for which Richardson iteration will converge if $\|I - BA\| < 1$, or equivalently $\rho(BA) < 1$, and we then refer to B as an *approximate inverse* of A. The preconditioned Richardson iteration takes the form

$$x^{(k+1)} = (I - BA)x^{(k)} + Bb, \tag{7.17}$$

and the preconditioned residual $Bb - BAx^{(k)}$ is used as basis for a stopping criterion.

Iterative methods based on matrix splitting

An alternative to Richardson iteration is *matrix splitting*, where stationary iterative methods are formulated based on splitting the matrix into a sum

$$A = A_1 + A_2, \tag{7.18}$$

where A_1 is chosen as a nonsingular matrix easy to invert, such as a diagonal matrix D, a (strict) lower triangular matrix L or (strict) upper triangular matrix U, where L and U have zeros on the diagonal.

Jacobi iteration

Jacobi iteration is based on the splitting

$$A_1 = D, \quad A_2 = R = A - D, \tag{7.19}$$

which gives the iteration matrix $M_J = -D^{-1}R$ and $c = D^{-1}b$, or in terms of the elements of $A = (a_{ij})$,

$$x_i^{(k+1)} = a_{ii}^{-1}(b - \sum_{j \neq i} a_{ij} x_j^{(k)}), \tag{7.20}$$

where the diagonal matrix D is trivial to invert. To use Jacobi iteration as a preconditioner, we choose $B = D^{-1}$.

Gauss-Seidel iteration

Gauss-Seidel iteration is based on the splitting

$$A_1 = L, \quad A_2 = R = A - L, \tag{7.21}$$

which gives the iteration matrix $M_{GS} = -L^{-1}R$ and $c = L^{-1}b$, or

$$x_i^{(k+1)} = a_{ii}^{-1}(b - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)}), \tag{7.22}$$

where the lower triangular matrix L is inverted by forward substitution.

7.2 Krylov methods

Krylov subspace

A *Krylov method* is an iterative method for the solution of the system $Ax = b$ based on, for each iteration, finding an approximation $x^{(k)} \approx x = A^{-1}b$ in a *Krylov subspace* $\mathcal{K}_k$, spanned by the vectors $b, Ab, ..., A^{k-1}b$, that is

$$\mathcal{K}_k = \langle b, Ab, ..., A^{k-1}b \rangle. \tag{7.23}$$

The basis for Krylov methods is that, by the *Cayley-Hamilton theorem*, the inverse of a matrix A^{-1} is a linear combination of its powers A^k, which is also expressed in (7.15).

GMRES

The idea of *GMRES* (generalized minimal residuals) is that, at each step k of the iteration, find the vector $x^{(k)} \in \mathcal{K}_k$ that minimizes the norm of the residual $r^{(k)} = b - Ax^{(k)}$, which corresponds to the least squares problem

$$\min_{x^{(k)} \in \mathcal{K}_k} \|b - Ax^{(k)}\|. \tag{7.24}$$

But instead of expressing the approximation $x^{(k)}$ as a linear combination of the Krylov vectors $b, Ab, ..., A^{k-1}b$, which leads to an unstable algorithm, we construct an orthonormal basis $\{q_j\}_{j=1}^{k}$ for $\mathcal{K}_k$, such that

$$\mathcal{K}_k = \langle q_1, q_2, ..., q_k \rangle, \tag{7.25}$$

with Q_k the $n \times k$ matrix with the basis vectors q_j as columns.

Thus we can express the approximation as $x^{(k)} = Q_k y$, with $y \in \mathbb{R}^k$ a vector with the coordinates of $x^{(k)}$, so that the least squares problem take the form

$$\min_{y \in \mathbb{R}^k} \|b - AQ_k y\|. \tag{7.26}$$

The *Arnoldi iteration* constructs a partial similarity transformation of A into an *Hessenberg matrix* $\tilde{H}_k \in \mathbb{R}^{k+1 \times k}$,

$$AQ_k = Q_{k+1}\tilde{H}_k, \tag{7.27}$$

that is

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \left[\begin{array}{c|c|c} & & \\ q_1 & \cdots & q_k \\ & & \end{array}\right] = \left[\begin{array}{c|c|c} & & \\ q_1 & \cdots & q_{k+1} \\ & & \end{array}\right] \begin{bmatrix} h_{11} & \cdots & h_{1k} \\ h_{21} & \cdots & \\ & \ddots & \vdots \\ & & h_{k+1k} \end{bmatrix},$$

and multiplication of (7.26) by Q_{k+1}^T does not change the norm, so that the least squares problem takes the form,

$$\min_{y \in \mathbb{R}^k} \|Q_{k+1}^T b - \tilde{H}_k y\|, \tag{7.28}$$

where we note that since $q_1 = b/\|b\|$, we have that $Q_{k+1}^T b = \|b\| e_1$ with $e_1 = (1, 0, ..., 0)^T$ the first vector in the standard basis in $\mathbb{R}^{k+1}$, so that

$$\min_{y \in \mathbb{R}^k} \|\|b\| e_1 - \tilde{H}_k y\|, \tag{7.29}$$

which is the least squares problem we solve for y at each iteration k, to get $x^{(k)} = Q_k y$.

Algorithm 7: GMRES

$q_1 = b/\|b\|$
while $\|r^{(k)}\|/\|r^{(0)}\| \geq TOL$ **do**
  Arnoldi iteration $\rightarrow Q_k, \tilde{H}_k$ ▷ `partial similarity transform`
  $\min_{y \in \mathbb{R}^k} \|\|b\|e_1 - \tilde{H}_k y\|$ ▷ `least squares problem`
  $x^{(k)} = Q_k y$ ▷ `construct solution`
end

Conjugate Gradient method

For a symmetric positive definite matrix A, we can define the *A-norm* of a vector $x \in \mathbb{R}^n$, as

$$\|x\|_A = (x, Ax)^{1/2}, \tag{7.30}$$

with $(\cdot, \cdot)$ the l_2-norm. The *Conjugate Gradient method* (CG) is based on minimization of the error $e^{(k)} = x - x^{(k)}$ in the A-norm, or equivalently, by (7.3), minimization of the residual $r^{(k)} = b - Ax^{(k)}$ in the A^{-1}-norm,

$$\|e^{(k)}\|_A = (e^{(k)}, Ae^{(k)})^{1/2} = (e^{(k)}, r^{(k)})^{1/2} = (A^{-1}r^{(k)}, r^{(k)})^{1/2} = \|r^{(k)}\|_{A^{-1}},$$

to compare to GMRES where the residual is minimized in the l_2-norm. Further, to solve the minimization problem in CG we do not solve a least squares problem over the Krylov subspace $\mathcal{K}_k$, but instead we iteratively construct a *search direction* $p^{(k)}$ and a *step length* $\alpha^{(k)}$ to find the new approximate solution $x^{(k)}$ from the previous iterate $x^{(k-1)}$. In particular, this means that we do not have to store the Krylov basis.

Algorithm 8: Conjugate Gradient method

$x^{(0)} = 0$, $r^{(0)} = b$, $p^{(k)} = r^{(0)}$
while $\|r^{(k)}\|/\|r^{(0)}\| \geq TOL$ **do**
  $\alpha^{(k)} = \|r^{(k-1)}\|/\|p^{(k-1)}\|_A$ ▷ `step length`
  $x^{(k)} = x^{(k-1)} + \alpha^{(k)} p^{(k-1)}$ ▷ `approximate solution`
  $r^{(k)} = r^{(k-1)} - \alpha^{(k)} A p^{(k-1)}$ ▷ `residual`
  $\beta^{(k)} = \|r^{(k)}\|/\|r^{(k-1)}\|$ ▷ `improvement`
  $p^{(k)} = r^{(k)} + \beta^{(k)} p^{(k-1)}$ ▷ `search direction`
end

The key to the success of the CG method is that the residuals are mutually orthogonal,

$$(r^{(k)}, r^{(j)}) = 0, \quad \forall j < k, \tag{7.31}$$

and that the search directions are *A-conjugate*,

$$(p^{(k)}, p^{(j)})_A = 0, \quad \forall j < k, \tag{7.32}$$

where $(\cdot,\cdot)_A$ is the *weighted inner product*, defined for symmetric positive definite matrices as

$$(x, y)_A = x^T Ay = (Ay)^T x = y^T A^T x = y^T Ax = (y, x)_A, \tag{7.33}$$

where we note that $(\cdot,\cdot)_A$ induces the *A-norm*,

$$\|x\|_A = (x, x)_A^{1/2}, \tag{7.34}$$

which is also referred to as the *energy norm* for the equation $Ax = b$.

Theorem 12 (CG characteristics)**.** *For the CG method applied to the equation $Ax = b$, with A an $n \times n$ symmetric positive definite matrix, the orthogonality relations (7.31) and (7.32) are true, and*

$$\begin{aligned} \mathcal{K}_k &= \langle b, Ab, ..., A^{k-1}b\rangle = \langle x^{(1)}, x^{(2)}, ..., x^{(k)}\rangle \\ &= \langle p^{(0)}, p^{(1)}, ..., p^{(k-1)}\rangle = \langle r^{(0)}, r^{(1)}, ..., r^{(k-1)}\rangle, \end{aligned}$$

with the approximate solutions $x^{(k)}$, search directions $p^{(k)}$ and residuals $r^{(k)}$ constructed from Algorithm 8. Further, $x^{(k)}$ is the unique point in $\mathcal{K}_k$ that minimizes $\|e^{(k)}\|_A$, and the convergence is monotonic, that is

$$\|e^{(k)}\|_A \leq \|e^{(k-1)}\|_A, \tag{7.35}$$

with $e^{(k)} = 0$ for some $k \leq n$.