Chapter 5

Linear Systems – Large Matrices

In practice, one often encounters linear systems with large sparse matrices. It is not uncommon to meet a $100,000 \times 100,000$ matrix with only about 10^6 nonzero entries. When A is tridiagonal (or, more generally, banded) it was already discussed in the exercises that the solution of such large linear systems is still feasible through an LU factorization. However, in practice one often encounters more complicated sparsity patterns for which it is not possible to carry out the LU factorization without excessive cost. In the following, we discuss several iterative methods that only involve matrix-vector products with sparse matrices.

5.1 Jacobi and Gauss-Seidel methods

The Jacobi method is the simplest iterative method for $A\mathbf{x} = \mathbf{b}$. Its idea consists of considering the jth equation and considering all but the jth variable fixed. This computation is performed repeatedly for $j = 1, \ldots, n$. For example, for $n = 3$, this corresponds to rewriting $A\mathbf{x} = \mathbf{b}$ as

$$\begin{aligned} x_1 &= (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \\ x_2 &= (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \\ x_3 &= (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33}\,. \end{aligned}$$

Let $\mathbf{x}^{(0)}$ be given[14] then the Jacobi method is the recursion defined by

$$x_i^{(k+1)} := \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) \Big/ a_{ii}, \quad i = 1, \ldots, n. \tag{5.1}$$

One obvious possibility for improvement of (5.1) is to use the newest available

[14] The starting vector $\mathbf{x}^{(0)}$ is often set to zero or chosen randomly.

information, that is to use $x_j^{(k+1)}$ instead of $x_j^{(k)}$ for $j < i$:

$$x_i^{(k+1)} := \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) \Big/ a_{ii}\,. \tag{5.2}$$

◇

For the purpose of analysis, it is better to rewrite (5.1) and (5.2) in the form of matrix operations. We write

$$A = L + D + U \tag{5.3}$$

where

$$D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix},\ L = \begin{pmatrix} 0 & & & 0 \\ a_{21} & \ddots & & \\ & & \ddots & \\ a_{n1} & & a_{n,n-1} & 0 \end{pmatrix},\ U = \begin{pmatrix} 0 & a_{12} & & a_{1n} \\ & \ddots & & \vdots \\ & & \ddots & a_{n-1,n} \\ & 0 & & 0 \end{pmatrix}.$$

Jacobi method: From (5.1) it follows that

$$\begin{aligned} D\mathbf{x}^{(k+1)} &= \mathbf{b} - L\mathbf{x}^{(k)} - U\mathbf{x}^{(k)} \iff \\ \mathbf{x}^{(k+1)} &= D^{-1}\big(\mathbf{b} - L\mathbf{x}^{(k)} - U\mathbf{x}^{(k)}\big) \\ &= \underbrace{-D^{-1}(L+U)}_{B^J}\ \mathbf{x}^{(k)} + D^{-1}\,\mathbf{b} \\ &= \qquad B^J \qquad\ \mathbf{x}^{(k)} + \mathbf{f}\,. \end{aligned} \tag{5.4}$$

Gauss-Seidel method: From (5.2) it follows that

$$\begin{aligned} & \mathbf{x}^{(k+1)} = D^{-1}\big(\mathbf{b} - L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)}\big) && \iff \\ (D+L)\ & \mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + \mathbf{b} && \iff \\ & \mathbf{x}^{(k+1)} = -(D+L)^{-1}\,U\mathbf{x}^{(k)} + (D+L)^{-1}\mathbf{b} && \iff \\ & \mathbf{x}^{(k+1)} = B^{\mathrm{GS}}\,\mathbf{x}^{(k)} + \mathbf{f} \end{aligned} \tag{5.5}$$

with

$$B^{\mathrm{GS}} = -(D+L)^{-1}\,U, \quad \mathbf{f} = (D+L)^{-1}\,\mathbf{b}.$$

5.2 Splitting methods

Both, the Jacobi and Gauss-Seidel methods are splitting methods. Consider some splitting of A:

$$A = P - N, \qquad P, N \in \mathbb{R}^{n\times n}, \tag{5.6}$$

where P is usually called *preconditioner* and it is assumed that it is relatively easy to solve linear systems with P. Given such a splitting, we reformulate $A\mathbf{x} = \mathbf{b}$ as the fixed point equation

$$\mathbf{x} = B\mathbf{x} + \mathbf{f}, \quad B := P^{-1}N. \tag{5.7}$$

The corresponding fixed point iteration is given by

$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{f}, \tag{5.8}$$

with $\mathbf{f} = P^{-1}\mathbf{b}$. Equivalently, this corresponds to solving the linear system

$$P\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + \mathbf{b}$$

in each step.

It is easy to check that (5.8) corresponds to the Jacobi and Gauss-Seidel methods when choosing $P = D$ (diagonal preconditioner) and $P = D + L$, respectively.

For analyzing the convergence of (5.8), we let $\rho(A)$ denote the spectral radius of A, that is,

$$\rho(A) := \max\{|\lambda| \colon \lambda \in \mathbb{C} \text{ is an eigenvalue of } A\}.$$

Lemma 5.1 *Let $A \in \mathbb{R}^{n\times n}$. Then $A^k \to 0$ for $k \to \infty$ if and only if $\rho(A) < 1$.*

Proof. EFY. □

By subtracting $\mathbf{x} = B\mathbf{x} + \mathbf{f}$ from (5.8), we obtain the error recurrence

$$\mathbf{e}^{(k+1)} = B\mathbf{e}^{(k)}, \qquad k = 0, 1, 2, ..., \qquad \text{where} \quad \mathbf{e}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}. \tag{5.9}$$

We have $\mathbf{x}^{(k)} \to \mathbf{x}$ if and only if $\|\mathbf{e}^{(k)}\| = \|\mathbf{x}^{(k)} - \mathbf{x}\| \to 0$. Because of

$$\mathbf{e}^{(k)} = B\mathbf{e}^{(k-1)} = B^2\mathbf{e}^{(k-2)} = \ldots = B^k\mathbf{e}^{(0)}$$

we obtain convergence for *every* starting vector $\mathbf{x}^{(0)} \in \mathbb{R}^n$, if $B^k \to 0$ for $k \to \infty$. By Lemma 5.1, this holds if and only if the spectral radius $\rho(B)$ is smaller than 1.

Proving statements about the spectral radius is usually very difficult. Instead one uses the well-known fact that the spectral radius is bounded by any operator norm. We will illustrate this principle by showing that the Jacobi method converges for strictly diagonally dominant matrices.

Definition 5.2 *A matrix $A \in \mathbb{R}^{n\times n}$ is called* **strictly diagonally dominant by rows** *if*

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n\,,$$

and **strictly diagonally dominant by columns** *if*

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|, \quad i = 1, \ldots, n\,.$$

Theorem 5.3 *Let A be strictly diagonally dominant by rows or columns. Then the Jacobi method converges.*

Proof. We will prove the statement when A is strictly diagonally dominant by rows; the column case is handled analogously. We recall that the iteration matrix for the Jacobi method is given by $B^J = -D^{-1}(L+U)$ and hence

$$\|B^J\|_\infty = \max_{i=1,\dots,n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \,/\, |a_{ii}| < 1\,.$$

By the discussion above, this implies $\rho(B^J) \leq \|B^J\|_\infty < 1$ and hence the Jacobi method converges. □

Theorem 5.3 also holds for the Gauss-Seidel method but the proof is more difficult. Moreover, one can show that the Gauss-Seidel method converges for every symmetric positive definite matrix A.

5.3 Richardson method

The Richardson method for approximating the solution of $A\mathbf{x} = \mathbf{b}$ takes the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\, P^{-1}\, \mathbf{r}^{(k)}\,, \tag{5.10}$$

where $\alpha > 0$ is an acceleration parameter, P is a preconditioner (that is easy to solve linear systems with), and $\mathbf{r}^{(k)}$ is the residual $\mathbf{x}^{(k)}$ defined by

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}. \tag{5.11}$$

We again get an error recurrence of the form (5.9), with the iteration matrix $B = I - \alpha P^{-1}A$. Note that the Jacobi and Gauss-Seidel methods can be viewed as special cases of the Richardson method *if* $\alpha = 1$. However, in contrast to the Jacobi and Gauss-Seidel methods, the Richardson method can always be made convergent by choosing α appropriately. When A, P are symmetric positive definite then $P^{-1}A$ has positive real eigenvalues (Proof EFY) and the following result can be established.

Theorem 5.4 *Given an invertible preconditioner P, assume that $P^{-1}A$ has positive real eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0\,.$$

Then the Richardson method (5.10) *converges if and only if* $0 < \alpha < 2/\lambda_1$.
The choice $\alpha = \alpha_{\text{opt}} := 2/(\lambda_1 + \lambda_n)$ *minimizes the spectral radius of B, that is,*

$$\rho_{\text{opt}} = \min_{\alpha>0} |\,\rho(B_\alpha)| = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\,. \tag{5.12}$$

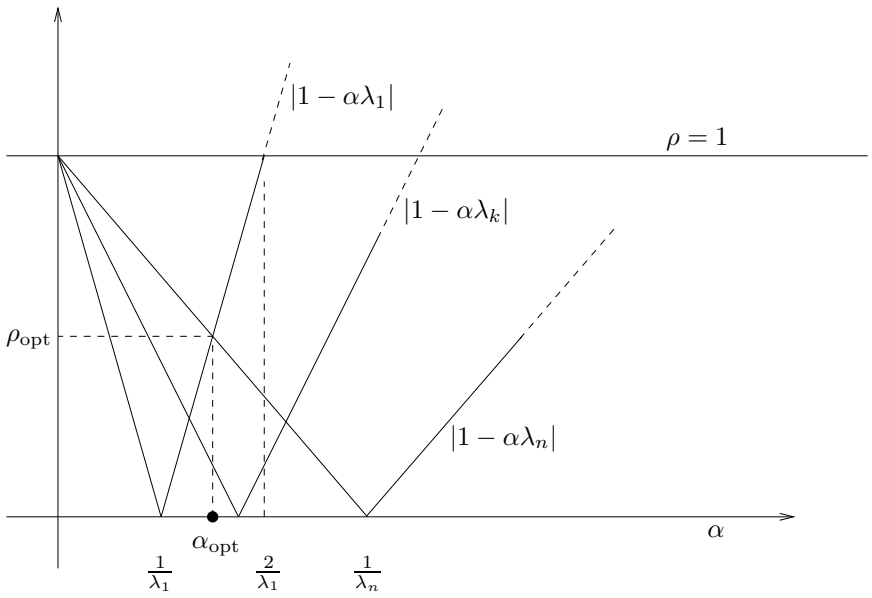

Figure 5.1: Spectral radius of B_α as a function of the eigenvalues of $P^{-1}A$.

Proof. The eigenvalues of B_α are given by $\lambda_i(B_\alpha) = 1 - \alpha\lambda_i$ for $i = 1, \ldots, n$. Hence, (5.10) converges if and only if $|\lambda_i(B_\alpha)| < 1$ for $i = 1, \ldots, n$, or, equivalently, if $0 < \alpha < 2/\lambda_1$. It follows (see Figure 5.1) that $\rho(B_\alpha)$ is minimal for $1-\alpha\lambda_n = \alpha\lambda_1 - 1$, which is satisfied when choosing $\alpha = 2/(\lambda_1 + \lambda_n)$. ∎

Remark 5.5 When A, P are symmetric positive definite then $\kappa_2(P^{-1}\,A) = \lambda_1/\lambda_n$ (proof EFY) and, hence,

$$\rho_{\text{opt}} = \frac{\lambda_1/\lambda_n - 1}{\lambda_1/\lambda_n + 1} = \frac{\kappa_2(P^{-1}A) - 1}{\kappa_2(P^{-1}A) + 1} \tag{5.13}$$

It follows that the convergence rate of the Richardson method depends only on $\kappa_2(P^{-1}A)$ ab.

This explains the wording "preconditioner" for P. The construction of cheap and effective preconditioner is an art by itself. First choices are D (the diagonal part of A) or incomplete LU/Cholesky decompositions of A.

5.4 Gradient method

The optimal choice of α in the Richardson method depends on the eigenvalues of A, which are usually unknown (and can be more expensive to compute than solving the linear system!). We will now take an "optimization perspective" of the Richardson method for symmetric positive definite A (with $P = I$ for simplicity), which allows choose α optimally in each step, without knowledge of the eigenvalues of A.

Theorem 5.6 *Let A be symmetric positive definite. Then $\mathbf{x}$ is the solution of $A\mathbf{x} = \mathbf{b}$ if and only if it solves the optimization problem*

$$\mathbf{x} = \arg\min_{\mathbf{y}\in\mathbb{R}^n} \Phi(\mathbf{y}), \quad \textit{with} \quad \Phi(\mathbf{y}) := \frac{1}{2}\mathbf{y}^\top A\mathbf{y} - \mathbf{y}^\top \mathbf{b}. \tag{5.14}$$

Proof. For $\triangle\mathbf{y} \in \mathbb{R}^n$ we consider

$$\begin{aligned}\Phi(\mathbf{y}+\triangle\mathbf{y}) - \Phi(\mathbf{y}) &= \frac{1}{2}\triangle\mathbf{y}^\top A\mathbf{y} + \frac{1}{2}\mathbf{y}^\top A\triangle\mathbf{y} - \triangle\mathbf{y}^\top\mathbf{b} + O(\|\triangle\mathbf{y}\|_2^2)\\ &= \triangle\mathbf{y}^\top(A\mathbf{y}-\mathbf{b}) + O(\|\triangle\mathbf{y}\|_2^2). \end{aligned}\tag{5.15}$$

By the uniqueness of the Taylor expansion, it follows that $A\mathbf{y} - \mathbf{b}$ is the gradient of Φ at $\mathbf{y}$.

If $\mathbf{x}$ solves (5.14) then the gradient of Φ at $\mathbf{x}$ is zero, that is, $A\mathbf{x} - \mathbf{b}$.[15] In the other direction, if $\mathbf{x}$ is solution of the linear system, then

$$\begin{aligned}\Phi(\mathbf{y}) &= \Phi(\mathbf{x} + (\mathbf{y}-\mathbf{x}))\\ &= \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} + (\mathbf{y}-\mathbf{x})^T(A\mathbf{x}-\mathbf{b}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T A(\mathbf{y}-\mathbf{x})\\ &= \Phi(\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T A(\mathbf{y}-\mathbf{x}),\end{aligned}$$

and since $\frac{1}{2}(\mathbf{y}-\mathbf{x})^T A(\mathbf{y}-\mathbf{x}) \geq 0$, it follows that $\Phi(\mathbf{y}) \geq \Phi(\mathbf{x})$ and hence $\mathbf{x}$ minimizes Φ. ∎

The idea behind the gradient method is to proceed in every step in the negative direction of the gradient. Let $\mathbf{x}^{(k)}$ denote the kth iterate of the method. The $(k+1)$th iterate is obtained by setting

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}.$$

The search direction $\mathbf{p}^{(k)}$ is chosen such that it minimizes the first-order term $(\mathbf{p}^{(k)})^\top(A\mathbf{y}-\mathbf{b})$ in the Taylor expansion (5.15) among all vectors of the same 2-norm. By the Cauchy-Schwarz inequality the best choice is

$$\mathbf{p}^{(k)} = -\nabla\Phi\big(\mathbf{x}^{(k)}\big) = \mathbf{b} - A\mathbf{x}^{(k)} =: \mathbf{r}^{(k)}. \tag{5.16}$$

The step size α_k is chosen such that $\Phi\big(\mathbf{x}^{(k)} + \alpha_k\mathbf{p}^{(k)}\big) = \Phi\big(\mathbf{x}^{(k+1)}\big)$ is minimal

[15] One could complete the proof with the observation that f is strictly convex and, hence, a zero gradient characterizes the (global) minimum. We include a direct proof of the other direction for illustration.

among all choices of α, that is,

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\Phi\Big(\mathbf{x}^{(k)} + \alpha\mathbf{p}^{(k)}\Big)\bigg|_{\alpha=\alpha_k} \\ &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\frac{1}{2}\big\langle \mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)}, A(\mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)})\big\rangle - \big\langle \mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)}, \mathbf{b}\big\rangle\right]\bigg|_{\alpha=\alpha_k} \\ &= -\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)}\big\rangle + \alpha_k\big\langle \mathbf{r}^{(k)}, A\mathbf{r}^{(k)}\big\rangle, \end{aligned} \tag{5.17}$$

or, equivalently,

$$\alpha_k = \frac{\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)}\big\rangle}{\big\langle A\mathbf{r}^{(k)}, \mathbf{r}^{(k)}\big\rangle}.$$

In summary, the gradient method reads as follows: Given $\mathbf{x}^{(0)} \in \mathbb{R}^n$, let $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$. Then for all $k \geq 0$,

$$\begin{cases} \alpha_k = \frac{\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)}\big\rangle}{\big\langle A\mathbf{r}^{(k)}, \mathbf{r}^{(k)}\big\rangle}, \\ \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{r}^{(k)}, \\ \mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{r}^{(k)}. \end{cases}$$

The gradient method is thus a variant of the Richardson method (with $P = I$) for which the acceleration parameter is chosen adaptively in every iteration.

Figure 5.2 gives a visual representation of the gradient method. In particular, it seems that each direction $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)}$ for $k \geq 0$ is orthogonal to the descent direction at the previous iteration, $\mathbf{p}^{(k)} = \mathbf{r}^{(k)}$. Indeed, from equation (5.17),

$$0 = -\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)}\big\rangle + \alpha_k\big\langle \mathbf{r}^{(k)}, A\mathbf{r}^{(k)}\big\rangle = \big\langle \mathbf{r}^{(k)}, -\mathbf{r}^{(k)} + \alpha_k A\mathbf{r}^{(k)}\big\rangle = -\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k+1)}\big\rangle. \tag{5.18}$$

Therefore, $\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k+1)}\big\rangle = 0$. However, in general, $\big\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k+2)}\big\rangle \neq 0$. Indeed, in Figure 5.2, any two directions $\mathbf{r}^{(k)}$ and $\mathbf{r}^{(k+2)}$ are parallel.

Using the Kantorovich inequality, it can be shown that the gradient method converges at the rate (5.13), the convergence rate of the Richardson iteration with optimally chosen α.

5.5 The method of conjugate gradients (CG)

For ill-conditioned matrices, the gradient method makes little progress because the search directions $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots$ are too similar across several iterations. For $n = 2$ this can be nicely illustrated by the "zigzag" behavior of the gradient method.

We continue to assume that $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite. The idea of the CG method is to chose search directions that are orthogonal to each other in the inner product induced by A: $\langle \mathbf{y}, \mathbf{z}\rangle_A := \mathbf{y}^\top A\mathbf{z} = \langle \mathbf{y}, A\mathbf{z}\rangle = \langle A\mathbf{y}, \mathbf{z}\rangle$. In the first

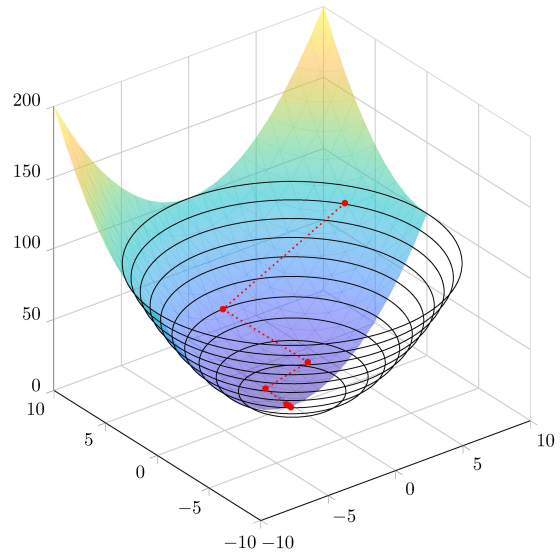

Figure 5.2: Illustration of the gradient method. In red are the descent directions $\mathbf{p}^{(k)}$. The red dots represent the intermediate solutions $\mathbf{x}^{(k)}$, $k \geq 0$. The ellipsoids represent some level curves of Φ; note that they are centered on the exact solution $\mathbf{x}$, and that the descent directions arrive tangentially towards them.

iterate, we still choose $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$, as in the gradient method, and set $\mathbf{x}^{(1)} := \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)}$. To choose the next search direction, we set

$$\mathbf{p}^{(1)} = \mathbf{r}^{(1)} - \beta_0 \mathbf{p}^{(0)},$$

where the parameter β_0 is chosen such that $\mathbf{p}^{(1)}$ is orthogonal (conjugate) to $\mathbf{p}^{(0)}$. This corresponds to one step of the Gram-Schmidt process in the A-inner product:

$$\beta_0 = \frac{\langle \mathbf{r}^{(1)}, \mathbf{p}^{(0)} \rangle_A}{\langle \mathbf{p}^{(0)}, \mathbf{p}^{(0)} \rangle_A}.$$

More generally, we choose

$$\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \sum_{i=0}^{k} \frac{\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(i)} \rangle_A}{\langle \mathbf{p}^{(i)}, \mathbf{p}^{(i)} \rangle_A} \mathbf{p}^{(i)}, \tag{5.19}$$

As for the gradient method, for all $k \geq 0$, $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$, and α_k is chosen so that $\Phi\left(\mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}\right) = \Phi\left(\mathbf{x}^{(k+1)}\right)$ is minimal. By following the exact same steps as in (5.17), we obtain:

$$\alpha_k = \frac{\langle \mathbf{p}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)} \rangle}.$$

The crucial observation, which makes CG efficient for larger k, is that most terms in (5.19) vanish.

Lemma 5.7 *With the notation introduced above, assume that $\alpha_i \neq 0$ for $i = 0, \ldots, k$. Then*

$$\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(0)} \rangle_A = \cdots = \langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(k-1)} \rangle_A = 0.$$

Proof. We start by noting that $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ implies the residual recursion

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A \mathbf{p}^{(k)}. \tag{5.20}$$

Step 1: Alternative formula for α_k. Using the definition of α_k, it follows from (5.20) that

$$\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(k)} \rangle = \langle \mathbf{r}^{(k)}, \mathbf{p}^{(k)} \rangle - \alpha_k \langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A = 0.$$

Because $\{\mathbf{p}^{(0)}, \ldots, \mathbf{p}^{(k)}\}$ is an A-orthogonal basis, we have for $i < k$ that

$$\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(i)} \rangle = \langle \mathbf{r}^{(k)}, \mathbf{p}^{(i)} \rangle + \alpha_k \langle \mathbf{p}^{(k)}, \mathbf{p}^{(i)} \rangle_A = \langle \mathbf{r}^{(k)}, \mathbf{p}^{(i)} \rangle,$$

and we can conclude inductively that

$$\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(i)} \rangle = 0, \quad i = 0, \ldots, k. \tag{5.21}$$

EFY: Show that this relation implies

$$\alpha_k = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)} \rangle}. \tag{5.22}$$

Step 2: Orthogonality of residuals. Note that

$$\langle \mathbf{r}^{(k)}, \mathbf{p}^{(k)} \rangle_A = \langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A.$$

For $k = 0$, this follows from $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$. For $k > 0$, this follows from applying the A-inner product with $p^{(k+1)}$ to both sides of (5.19), exploiting A-orthogonality, and shift $k + 1$ to k. Together with (5.22) it follows from (5.20) that

$$\langle \mathbf{r}^{(k+1)}, \mathbf{r}^{(k)} \rangle = \langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle - \alpha_k \langle \mathbf{r}^{(k)}, A\mathbf{p}^{(k)} \rangle = \langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle - \alpha_k \langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)} \rangle = 0.$$

For $i < k$, it follows from (5.19) that $\langle \mathbf{p}_k, \mathbf{r}_i \rangle_A = 0$. We conclude from (5.20) that

$$\langle \mathbf{r}^{(k+1)}, \mathbf{r}^{(i)} \rangle = 0, \quad i < k.$$

In other words, $\{\mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(k+1)}\}$ is an orthogonal basis.

Step 3: Conclusion. Using once more (5.20) we obtain

$$\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(i)} \rangle_A = \langle \mathbf{r}^{(k+1)}, A\mathbf{p}^{(i)} \rangle = \frac{1}{\alpha_k} \langle \mathbf{r}^{(k+1)}, \mathbf{r}^{(i)} - \mathbf{r}^{(i+1)} \rangle = 0.$$

□

The result of Lemma 5.7 allows us to rewrite (5.19) as

$$\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \beta_k \mathbf{p}^{(k)}, \quad \beta_k = \frac{\langle \mathbf{r}^{(k+1)}, \mathbf{p}^{(k)} \rangle_A}{\langle \mathbf{p}^{(k)}, \mathbf{p}^{(k)} \rangle_A}.$$

In summary, the CG method reads as follows. Given $\mathbf{x}^{(0)} \in \mathbb{R}^n$, let $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ and $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$. Then for all $k \geq 0$,

$$\begin{cases} \alpha_k = \frac{\langle \mathbf{p}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)} \rangle}; \\ \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}; \\ \mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{p}^{(k)}; \\ \beta_k = \frac{\langle \mathbf{r}^{(k+1)}, A\mathbf{p}^{(k)} \rangle}{\langle \mathbf{p}^{(k)}, A\mathbf{p}^{(k)} \rangle}; \\ \mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \beta_k \mathbf{p}^{(k)}. \end{cases}$$

Remark 5.8

- For CG to be well-defined, $\mathbf{p}^{(k)}$ has to be different from 0 at each iteration $k \geq 0$. But if for some $k \geq 0$, $\mathbf{p}^{(k)} = 0$, then $\mathbf{x}^{(k)} = \mathbf{x}$.
- At each iteration, the CG method considers a descent direction that is linearly independent from (since A-orthogonal to) the previous descent directions, and minimizes the quadratic form Φ in this new descent direction.

Theorem 5.9 *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Then the CG method yields after at most n iterations the exact solution (assuming exact arithmetic).*

Proof. As discussed above, in the unlikely case that $\mathbf{p}^{(k)} = 0$ for $k \leq n-1$ then CG has found already the exact solution. Otherwise, $\left\{\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n-1)}\right\}$ forms an A-orthogonal basis of $\mathbb{R}^n$. Because of (5.21), the vector $\mathbf{r}^{(n)}$ is orthogonal to the space $\operatorname{span}\left\{\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n-1)}\right\} = \mathbb{R}^n$. Consequently, $\mathbf{r}^{(n)} = \mathbf{0}$, which implies $\mathbf{x}^{(n)} = \mathbf{x}$. ∎

Theorem 5.9 is misleading because in practice one never wants to run n iterations of CG. Instead, one hopes to stop the method much earlier, as soon as the error is below a certain tolerance. The following theorem shows that the error of CG decreases quickly (and thus CG can be stopped after a few iterations) if the condition number of A is not too high.

Theorem 5.10 *Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive definite and consider the linear system $A\mathbf{x} = \mathbf{b}$. For $k \geq 0$, let $\mathbf{e}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x} \in \mathbb{R}^n$, where $\mathbf{x}^{(k)}$ is the k-th iterate of CG. Then,*

$$\left\|\mathbf{e}^{(k)}\right\|_A \leq 2\frac{C^k}{1 + C^{2k}}\left\|\mathbf{e}^{(0)}\right\|_A, \quad \textit{with} \quad C := \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}.$$

Proof. See Theorem 3.1.1 in [A. Greenbaum. Iterative methods for solving linear systems. SIAM, 1987]. ∎