

Chapter 15

Iterative methods

15.1 Introduction

This chapter describes two important classes of iterative methods for the solution of systems of linear equations

$$\mathbf{Ax} = \mathbf{b} \tag{15.1}$$

when \mathbf{A} is square, but has properties that prevent an economical solution by the decomposition

$$\mathbf{A} = \mathbf{LU} \tag{15.2}$$

as in Gaussian elimination. For example, \mathbf{A} may be very large, as in finite-difference methods for solving partial differential equations in three or more dimensions, or \mathbf{A} may be obtainable only as the result of numerical computation, in the form of the image \mathbf{Ac} of a given vector \mathbf{c} .

The utility of an iterative method for solving Eq. (15.1) depends on how rapidly the method converges to a satisfactorily accurate solution vector \mathbf{x} . Every matrix-vector multiplication \mathbf{Ac} requires $\sim O(N^2)$ floating-point operations (using the obvious, unsophisticated algorithm). Since Gaussian elimination requires only $\sim O(N^3/3)$ operations, it follows that, to be useful, an iterative method should require significantly fewer than $\sim O(N/3)$ matrix-vector multiplications.

Iterative methods of one class make use of contraction mappings, which shrink the distance between any two vectors (or, more generally, any two points in a metric space). For example, a matrix \mathbf{B} such that $\|\mathbf{B}\| < 1$ in some consistent matrix norm shrinks the distance between vectors: $\|\mathbf{B}(\mathbf{x} - \mathbf{y})\| \leq$

⁰ © Copyright 1993, 1994, 1995, 2005, 2006 by C. D. Cantrell. All rights reserved. This document may not be reproduced or transmitted in whole or in part by any mechanical, electronic or optical means, or by any combination of such means, without the written permission of the author.

$\|\mathbf{B}\| \|\mathbf{x} - \mathbf{y}\| < \|\mathbf{x} - \mathbf{y}\|$. Jacobi and Gauss-Seidel iterative methods use a constant contraction matrix \mathbf{B} .

Contraction mappings based on the construction of Krylov subspaces can achieve significant reductions in the number of iterations required to attain acceptable accuracy in solving Eq. (15.1). A Krylov subspace is spanned by an initial vector \mathbf{c} and a set of images under \mathbf{A} , such as $\{\mathbf{c}, \mathbf{A}\mathbf{c}, \dots, \mathbf{A}^{n-1}\mathbf{c}\}$, where \mathbf{c} is not an eigenvector of \mathbf{A} . If \mathbf{A} is non-singular and $\mathbf{c} \neq \mathbf{0}$, then there exists a Krylov subspace that contains the solution vector \mathbf{x} . Unfortunately the dimension of such a Krylov subspace may be so large that there is no computational advantage over solving the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ by Gaussian elimination. As a result, Krylov subspace methods are most often employed in an iterative mode, to obtain successively improved approximations to the solution. In order to converge, a Krylov-based iterative method must satisfy the contraction mapping theorem with respect to successive approximations to the solution. Section 15.4 provide a brief introduction to Krylov subspaces. Section 15.5 discusses what is probably the most important Krylov method, the method of conjugate gradients.

15.2 Contraction mappings

The foundation of iterative methods in many areas of numerical computation is the contraction mapping theorem, which we now state and prove.

Let \mathcal{M} be a complete metric space under a metric ρ . A mapping $\tau : \mathcal{M} \rightarrow \mathcal{M}$ is called a **contraction mapping** if and only if

$$\exists c \in (0, 1) : \exists : \forall x, y \in \mathcal{M} : \rho(\tau x, \tau y) \leq c\rho(x, y). \quad (15.3)$$

In other words, a contraction mapping shrinks the distance between points. We call the constant c the **contraction factor**.

The **contraction mapping theorem** asserts that, if a contraction mapping is iterated, every sequence of points

$$\forall x \in \mathcal{M} : x_n := \tau x_{n-1} \quad \text{where} \quad x_0 := x \quad (15.4)$$

converges to a unique point x_f in \mathcal{M} ,

$$\exists! x_f \in \mathcal{M} : \exists : \forall x \in \mathcal{M} : \lim_{n \rightarrow \infty} x_n = x_f, \quad (15.5)$$

and x_f is the unique fixed point of the contraction mapping:

$$\tau x_f = x_f. \quad (15.6)$$

For example, the Newton-Raphson method for finding a root of a function f , Eq. (3.20), makes use of the mapping

$$\tau_N : x_k \mapsto x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \quad (15.7)$$

If there is an interval $\mathcal{M} = [a, b]$ on which τ_N is a contraction mapping under the absolute-value metric $\rho(x, x') = |x - x'|$, then the Newton-Raphson method is guaranteed to converge to a root of f in $[a, b]$. In other words, if there exists a positive real number $c < 1$ such that $|\tau_N(x) - \tau_N(x')| \leq c|x - x'|$ for all $x, x' \in \mathcal{M}$, then the Newton-Raphson method will converge. In Fig. 3.4, for example, the x -intersection points of the tangents to f at $x = 1.0$ and $x = 1.5$ are $\tau_N(1.0) = 1.5$ and $\tau_N(1.5) = 1.875$; evidently $|\tau_N(1.5) - \tau_N(1.0)| = 0.375 < |1.5 - 1.0| = 0.5$. On the interval $\mathcal{M} = [1, 2]$, which is a metric space under the absolute-value norm, and for $f(x) = x^{-1} - 0.5$, τ_N is a contraction mapping with $c = 0.5$. Therefore, according to the contraction mapping theorem, Newton-Raphson iteration converges to the root of f , $x = 2.0$, from any starting point in the interval $[1, 2]$.

To prove the contraction mapping theorem, it is enough to show that $\{x_n\}$ is a Cauchy sequence. Since we have assumed that the metric space \mathcal{M} is complete, every Cauchy sequence approaches a limit in \mathcal{M} .

To show that $\{x_n\}$ is a Cauchy sequence, we start with Eq. (15.3), which implies a bound on the distance between x_m and x_n that depends on m and n . Assuming (without loss of generality) that $m \geq n$ (and therefore $c^m \leq c^n$), we get

$$\begin{aligned}
\rho(x_m, x_n) &= \rho(\tau^m x, \tau^n x) \\
&= \rho(\tau^{m-n} \tau^n x, \tau^n x) \\
&\leq c \rho(\tau^{m-n-1} x, \tau^{n-1} x) \\
&\leq c^n \rho(x_{m-n}, x) \\
&\leq c^n (\rho(x_{m-n}, x_{m-n-1}) + \cdots + \rho(x_1, x)) \\
&\leq c^n (c^{m-n-1} \rho(x_1, x) + \cdots + \rho(x_1, x)) \\
&\leq c^n (c^{m-n-1} + \cdots + c + 1) \rho(x_1, x) \\
&\leq c^n \left(\frac{1 - c^{m-n}}{1 - c} \right) \rho(x_1, x) = \left(\frac{c^n - c^m}{1 - c} \right) \rho(x_1, x) \\
&\leq \left(\frac{c^n}{1 - c} \right) \rho(x_1, x),
\end{aligned} \tag{15.8}$$

which depends only on n . It follows that

$$\forall \epsilon > 0 : \exists n : \exists \forall m \geq n : \rho(x_m, x_n) \leq \left(\frac{c^n}{1 - c} \right) \rho(x_1, x) < \epsilon, \tag{15.9}$$

where n may depend on the starting point x . It follows that, for every starting point $x \in \mathcal{M}$, $\{\tau^n x\}$ is a Cauchy sequence, which converges to a point $x_f \in \mathcal{M}$ by virtue of the completeness of \mathcal{M} .

If there are two such fixed points, x_f and $y_f \neq x_f$, then

$$\rho(x_f, y_f) = \rho(\tau x_f, \tau y_f) \leq c \rho(x_f, y_f) < \rho(x_f, y_f), \tag{15.10}$$

a contradiction. Hence the fixed point x_f is unique.

Exercises for Section 15.2

15.2.1 Show that the mapping $x \mapsto g(x)$, where g is defined in Eq. (3.48), is a contraction mapping on the complete metric space $\mathcal{M} = [\sqrt{a}, \infty)$ under the absolute-value norm.

15.2.2 Prove that the mapping

$$\tau x = ax + b, \quad (15.11)$$

where a and b are real constants and x is real, is a contraction mapping on the complete metric space \mathbb{R} under the absolute-value norm if and only if $|a| < 1$.

15.3 Jacobi and Gauss-Seidel iteration

15.3.1 Jacobi iteration

To motivate iterative approaches to solving Eq. (15.1), we separate \mathbf{A} into strictly lower-triangular, diagonal, and strictly upper-triangular matrices,

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}, \quad (15.12)$$

where

$$\mathbf{D} = \text{diag}[a_1^1, \dots, a_n^n]. \quad (15.13)$$

We assume that all of the diagonal elements of \mathbf{A} are non-zero. (This can be ensured by pivoting.) Then \mathbf{D}^{-1} exists, and it makes sense to move the non-diagonal terms of Eq. (15.1) to the right-hand side, formally solving for the solution vector \mathbf{x} in terms of itself, the off-diagonal parts of \mathbf{A} and the right-hand side \mathbf{b} :

$$\mathbf{D}\mathbf{x} = -(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}. \quad (15.14)$$

In component form, this equation reads

$$a_i^i x^i = - \sum_{j=1, j \neq i}^n a_j^i x^j + b^i. \quad (15.15)$$

The simplest implementation of Eq. (15.14) as an iterative method is

$$\mathbf{D}\mathbf{x}_{k+1} = -(\mathbf{L} + \mathbf{U})\mathbf{x}_k + \mathbf{b}, \quad (15.16)$$

which reads

$$a_i^i x_{k+1}^i = - \sum_{j=1, j \neq i}^n a_j^i x_k^j + b^i \quad (15.17)$$

in component form. Solving for \mathbf{x}_{k+1} , we obtain **Jacobi's method**:

$$\mathbf{x}_{k+1} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}_k + \mathbf{D}^{-1}\mathbf{b}. \quad (15.18)$$

Eq. (15.18) is of the static iterative form

$$\mathbf{x}_{k+1} = \mathbf{B}\mathbf{x}_k + \mathbf{c} \quad (15.19)$$

with a constant matrix

$$\mathbf{B}_J = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}) \quad (15.20)$$

and vector

$$\mathbf{c}_J = \mathbf{D}^{-1}\mathbf{b}. \quad (15.21)$$

The convergence properties of mappings of the form (15.19) are easily characterized, as we show now.

The vector space \mathbb{R}^n , to which the solution \mathbf{x} belongs, is a metric space under any vector norm, including the computationally useful norms $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$, with the definition

$$\rho(\mathbf{y}', \mathbf{y}) = \|\mathbf{y}' - \mathbf{y}\|. \quad (15.22)$$

We show that a mapping

$$\mathbf{y} \mapsto \tau_{\mathbf{B}}\mathbf{y} := \mathbf{B}\mathbf{y} + \mathbf{c} \quad (15.23)$$

is a contraction mapping if

$$\|\mathbf{B}\| < 1, \quad (15.24)$$

where $\|\mathbf{B}\|$ is the matrix norm that is consistent with the vector norm in the sense that, for all vectors \mathbf{y} ,

$$\|\mathbf{B}\mathbf{y}\| \leq \|\mathbf{B}\| \|\mathbf{y}\|. \quad (15.25)$$

The proof is straightforward:

$$\begin{aligned} \rho(\tau_{\mathbf{B}}\mathbf{y}', \tau_{\mathbf{B}}\mathbf{y}) &= \|\mathbf{B}\mathbf{y}' + \mathbf{c} - (\mathbf{B}\mathbf{y} + \mathbf{c})\| \\ &= \|\mathbf{B}(\mathbf{y}' - \mathbf{y})\| \\ &\leq \|\mathbf{B}\| \|\mathbf{y}' - \mathbf{y}\| = \|\mathbf{B}\| \rho(\mathbf{y}', \mathbf{y}), \end{aligned} \quad (15.26)$$

from which it follows that $\|\mathbf{B}\| < 1$ guarantees that $\tau_{\mathbf{B}}$ is a contraction mapping with a contraction factor equal to $\|\mathbf{B}\|$.

For example, suppose that we are using the infinity norm $\|\mathbf{x}\|_\infty$. Then the coefficient matrix \mathbf{A} is strictly row diagonally dominant, Eq. (9.335), if and only if

$$\|\mathbf{B}_J\|_\infty < 1. \quad (15.27)$$

Therefore Jacobi iteration converges if \mathbf{A} is strictly row diagonally dominant. A similar assertion for strict column diagonal dominance can be proved using the 1 norm.

15.3.2 Gauss-Seidel iteration

One might think that, if Jacobi's method converges, meaning that \mathbf{x}_{k+1} is closer to the solution \mathbf{x} than \mathbf{x}_k , then it would be advantageous to use the improved values as soon as possible. If we compute the components of \mathbf{x}_{k+1} beginning with $i = 1$, then Eq. (15.17) implies that, for a given value of $i > 1$, the components x_{k+1}^j for which $1 \leq j < i$ have already been computed, and therefore can be used to compute x_{k+1}^i . This observation motivates the modified iterative equation

$$\mathbf{D}\mathbf{x}_{k+1} = -\mathbf{L}\mathbf{x}_{k+1} - \mathbf{U}\mathbf{x}_k + \mathbf{b} , \quad (15.28)$$

which is equivalent to the component form

$$a_i^i x_{k+1}^i = - \sum_{j=1}^{i-1} a_j^i x_{k+1}^j - \sum_{j=i+1}^n a_j^i x_k^j + b^i . \quad (15.29)$$

Solving Eq. (15.28) for \mathbf{x}_{k+1} , we get the **Gauss-Seidel method**

$$\mathbf{x}_{k+1} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\mathbf{x}_k + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b} , \quad (15.30)$$

which is of the same form as Eq. (15.19) with a constant matrix equal to

$$\mathbf{B}_{GS} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U} \quad (15.31)$$

and a constant vector equal to

$$\mathbf{c}_{GS} = (\mathbf{D} + \mathbf{L})^{-1}\mathbf{b} . \quad (15.32)$$

If $\|\mathbf{B}_{GS}\|_\infty < 1$, then Eq. (15.30) defines a contraction mapping with a contraction factor equal to $\|\mathbf{B}_{GS}\|_\infty$.

The Gauss-Seidel method also converges by the contraction mapping theorem if the coefficient matrix \mathbf{A} is strictly row diagonally dominant. To show this, we begin by writing Eq. (15.31) in terms of the matrices

$$\mathbf{L}' := \mathbf{D}^{-1}\mathbf{L}, \quad \mathbf{U}' := \mathbf{D}^{-1}\mathbf{U} \Rightarrow \mathbf{U} = \mathbf{D}\mathbf{U}' . \quad (15.33)$$

Then

$$(\mathbf{D} + \mathbf{L})^{-1} = (\mathbf{D} + \mathbf{D}\mathbf{L}')^{-1} = (\mathbf{1} + \mathbf{L}')^{-1}\mathbf{D}^{-1} \quad (15.34)$$

and therefore

$$\mathbf{B}_{GS} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U} = -(\mathbf{1} + \mathbf{L}')^{-1}\mathbf{D}^{-1}\mathbf{D}\mathbf{U}' = -(\mathbf{1} + \mathbf{L}')^{-1}\mathbf{U}' . \quad (15.35)$$

If we let

$$\mathbf{z} = \mathbf{B}_{GS}\mathbf{y} = -(\mathbf{1} + \mathbf{L}')^{-1}\mathbf{U}'\mathbf{y} \quad (15.36)$$

then

$$(\mathbf{1} + \mathbf{L}')\mathbf{z} = -\mathbf{U}'\mathbf{y} \quad (15.37)$$

and therefore

$$\mathbf{z} = -\mathbf{L}'\mathbf{z} - \mathbf{U}'\mathbf{y}. \quad (15.38)$$

We now use the fact that

$$\|\mathbf{B}_{GS}\|_\infty = \max_{\mathbf{y}} \frac{\|\mathbf{z}\|_\infty}{\|\mathbf{y}\|_\infty} \quad (15.39)$$

to show that, if \mathbf{A} is strictly row diagonally dominant, then $\|\mathbf{B}_{GS}\|_\infty < 1$. From Eq. (15.38) it follows that

$$\|\mathbf{z}\|_\infty \leq \|\mathbf{L}'\mathbf{z}\|_\infty + \|\mathbf{U}'\mathbf{y}\|_\infty \leq \|\mathbf{L}'\|_\infty \|\mathbf{z}\|_\infty + \|\mathbf{U}'\|_\infty \|\mathbf{y}\|_\infty. \quad (15.40)$$

Then

$$\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{y}\|_\infty} \leq \|\mathbf{L}'\|_\infty \frac{\|\mathbf{z}\|_\infty}{\|\mathbf{y}\|_\infty} + \|\mathbf{U}'\|_\infty \quad (15.41)$$

and

$$\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{y}\|_\infty} \leq \frac{\|\mathbf{U}'\|_\infty}{1 - \|\mathbf{L}'\|_\infty}. \quad (15.42)$$

Strict row diagonal dominance implies that

$$1 > \|\mathbf{L}'\|_\infty + \|\mathbf{U}'\|_\infty \quad (15.43)$$

and therefore that

$$\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{y}\|_\infty} \leq \frac{\|\mathbf{U}'\|_\infty}{1 - \|\mathbf{L}'\|_\infty} < \frac{\|\mathbf{U}'\|_\infty}{\|\mathbf{U}'\|_\infty + \|\mathbf{L}'\|_\infty - \|\mathbf{L}'\|_\infty} = 1. \quad (15.44)$$

It follows that $\|\mathbf{B}_{GS}\|_\infty < 1$. Therefore Eq. (15.30) defines a contraction mapping, and the convergence of Gauss-Seidel iteration is assured for a strictly row diagonally dominant coefficient matrix \mathbf{A} .

15.3.3 Successive over-relaxation

In practical computations one often observes that Gauss-Seidel iteration converges slowly because the (vector) step taken from \mathbf{x}_k to \mathbf{x}_{k+1} is in the right direction, but is not big enough. To accelerate convergence by taking a bigger step at each iteration, it is natural to let

$$\mathbf{x}_{k+1} - \mathbf{x}_k = \omega (\mathbf{x}_{k+1}^{GS} - \mathbf{x}_k) \quad (15.45)$$

where ω is an acceleration constant and the next Gauss-Seidel approximation to the solution vector is, by definition, given by the equation

$$\mathbf{D}\mathbf{x}_{k+1}^{GS} = -\mathbf{L}\mathbf{x}_{k+1} - \mathbf{U}\mathbf{x}_k + \mathbf{b}, \quad (15.46)$$

which is the same as Eq. (15.28). Substituting into Eq. (15.45), one obtains

$$\begin{aligned}\mathbf{D}\mathbf{x}_{k+1} &= \omega\mathbf{D}\mathbf{x}_{k+1}^{GS} + (1-\omega)\mathbf{D}\mathbf{x}_k \\ &= -\omega\mathbf{L}\mathbf{x}_{k+1} + (1-\omega)\mathbf{D}\mathbf{x}_k - \omega\mathbf{U}\mathbf{x}_k + \omega\mathbf{b}.\end{aligned}\tag{15.47}$$

After collecting terms, one has

$$(\mathbf{D} + \omega\mathbf{L})\mathbf{x}_{k+1} = [(1-\omega)\mathbf{D} - \omega\mathbf{U}]\mathbf{x}_k + \omega\mathbf{b}\tag{15.48}$$

which implies that

$$\mathbf{x}_{k+1} = \mathbf{B}_{SOR}\mathbf{x}_k + \mathbf{c}_{SOR}\tag{15.49}$$

where the constant iteration matrix is

$$\mathbf{B}_{SOR} = (\mathbf{D} + \omega\mathbf{L})^{-1}[(1-\omega)\mathbf{D} - \omega\mathbf{U}]\tag{15.50}$$

and the constant iteration vector is

$$\mathbf{c}_{SOR} = \omega(\mathbf{D} + \omega\mathbf{L})^{-1}\mathbf{b}.\tag{15.51}$$

A “detail” that this derivation leaves unresolved is how to choose the acceleration constant ω , first, to ensure convergence, and second, to ensure the fastest possible rate of convergence. A theorem due to Kahan¹ states that $\omega \in (0, 2)$ is a necessary condition for convergence. Since $\omega = 1$ corresponds to the Gauss-Seidel method, it is clear that, to improve the rate of convergence beyond what is available with Gauss-Seidel iteration, one must choose $\omega \in (1, 2)$. Unfortunately it is not possible, in general, to estimate the optimal value of ω in advance for a specific linear system. Software packages designed for solving large linear systems may employ adaptive parameter estimation to adjust the value of ω as the computation progresses.

Exercises for Section 15.3

- 15.3.1** Show that, for Jacobi iteration, the solution vector \mathbf{x} of the linear system in Eq. (15.1) is a fixed point of the mapping in Eq. (15.23).
- 15.3.2** Show that, for Gauss-Seidel iteration, the solution vector \mathbf{x} of the linear system in Eq. (15.1) is a fixed point of the mapping in Eq. (15.23).
- 15.3.3** Show that, for successive over-relaxation, the solution vector \mathbf{x} of the linear system in Eq. (15.1) is a fixed point of the mapping in Eq. (15.23).

¹W. Kahan, “Gauss-Seidel methods of solving large systems of linear equations”, Ph.D. thesis, University of Toronto, 1958.

15.3.4 Use Jacobi iteration to solve the linear system in Eq. (15.1), where

$$\mathbf{A} = \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 90 \\ 70 \\ 50 \\ 30 \\ 10 \\ -10 \\ -30 \\ -50 \\ -70 \\ -90 \end{pmatrix}. \quad (15.52)$$

Also obtain the solution using the Thomas algorithm (Volume 1, p. 529) and compute the ∞ -norm of the difference between the Jacobi solution and the Thomas solution after 3, 6 and 10 Jacobi iterations. Using the UNIX `time` command, or a higher-level-language timing routine, obtain the execution time required for 3, 6 and 10 iterations, and compare with the time required for Gaussian elimination.

15.3.5 Use Jacobi iteration to solve the linear system in Eq. (15.1), where

$$\mathbf{A} = \begin{pmatrix} -30 & 16 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 16 & -30 & 16 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 16 & -30 & 16 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 16 & -30 & 16 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 16 & -30 & 16 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 16 & -30 & 16 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 16 & -30 & 16 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 16 & -30 & 16 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 16 & -30 & 16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 16 & -30 \end{pmatrix} \quad (15.53)$$

and the right-hand side \mathbf{b} is the same as in Eq. (15.52). Also obtain the solution using Gaussian elimination and compute the ∞ -norm of the difference between the Jacobi solution and the Gaussian solution after 3, 6 and 10 Jacobi iterations.

15.3.6 Use Gauss-Seidel iteration to solve the linear system in Eq. (15.1), using \mathbf{A} and \mathbf{b} from Eq. (15.52). Compute the ∞ -norm of the difference between the Gauss-Seidel solution and the Thomas solution after 3, 6 and 10 Gauss-Seidel iterations, and compare with the result of Exercise 15.3.4.

15.3.7 Use successive over-relaxation with a selection of values of $\omega \in (1, 2)$ to solve the linear system in Eq. (15.1), using \mathbf{A} and \mathbf{b} from Eq. (15.52). Compute the ∞ -norms of the differences between the successive over-relaxation solutions and the Thomas solution after 10 successive over-relaxation iterations for each value of ω , and compare with the result of Exercises 15.3.4 and 15.3.6.

15.4 Krylov subspaces

The n^{th} **Krylov subspace** generated by an initial vector \mathbf{c} and a matrix \mathbf{A} is

$$\mathcal{K}_n(\mathbf{c}, \mathbf{A}) := \text{span} \{\mathbf{c}, \mathbf{A}\mathbf{c}, \dots, \mathbf{A}^{n-1}\mathbf{c}\}. \quad (15.54)$$

The span of a set of vectors is defined in Volume 1, Eq. (5.110). It follows directly from the definition of the span that $\mathcal{K}_n(\mathbf{c}, \mathbf{A})$ is a vector subspace of the vector space to which \mathbf{c} , $\mathbf{A}\mathbf{c}$, etc., belong.

Krylov methods attempt to construct approximate solutions \mathbf{x}_n of Eq. (15.1), starting from an initial guess, \mathbf{x}_0 . The initial residual vector (Volume 1, pp. 311 and 519) is

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0. \quad (15.55)$$

The goal of a Krylov method is to find an n^{th} residual,

$$\mathbf{r}_n = \mathbf{b} - \mathbf{A}\mathbf{x}_n, \quad (15.56)$$

with an acceptably small norm for an acceptably small value of n .

In a Krylov method, the approximate solution vector \mathbf{x}_n lies in the affine subspace (Volume 1, Eq. (5.57)) that passes through \mathbf{x}_0 and is parallel to $\mathcal{K}_n(\mathbf{r}_0, \mathbf{A})$:

$$\mathbf{x}_n \in \mathbf{x}_0 + \mathcal{K}_n(\mathbf{r}_0, \mathbf{A}). \quad (15.57)$$

For example, if one's best initial guess is $\mathbf{x}_0 = \mathbf{0}$, then $\mathbf{r}_0 = \mathbf{b}$ and the approximate solution vector \mathbf{x}_n lies in the Krylov subspace $\mathcal{K}_n(\mathbf{b}, \mathbf{A})$, which is spanned by the vectors $\{\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{n-1}\mathbf{b}\}$.

One is certainly entitled to ask why it is reasonable to suppose that the solution vector lies in $\mathbf{x}_0 + \mathcal{K}_n(\mathbf{r}_0, \mathbf{A})$. To see why this makes sense, consider the example $\mathbf{x}_0 = \mathbf{0}$ for a nonsingular coefficient matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$.

According to the Cayley-Hamilton theorem, which is proved in Appendix I, every square matrix satisfies its own characteristic equation (Volume 1, section 9.3.1). Let the characteristic polynomial of \mathbf{A} be

$$\chi(\lambda) = \det[\mathbf{A} - \lambda\mathbf{1}]. \quad (15.58)$$

Explicitly,

$$\chi(\lambda) = (-1)^n \lambda^n + (-1)^{n-1} \text{trace}[\mathbf{A}] \lambda^{n-1} + \dots + \det[\mathbf{A}]. \quad (15.59)$$

The roots of the polynomial equation

$$\chi(\lambda) = 0 \quad (15.60)$$

are the eigenvalues of \mathbf{A} (Volume 1, section 9.3.1). The Cayley-Hamilton theorem asserts that

$$\chi(\mathbf{A}) = (-1)^n \mathbf{A}^n + (-1)^{n-1} \text{trace}[\mathbf{A}] \mathbf{A}^{n-1} + \dots + \det[\mathbf{A}] \mathbf{1} = \mathbf{0}, \quad (15.61)$$

where $\mathbf{0}$ is the zero matrix.

From Eq. (15.61) and from the assumption that \mathbf{A} is nonsingular it follows that

$$\mathbf{A}^{-1} = \frac{1}{\det[\mathbf{A}]} \left((-1)^{n-1} \mathbf{A}^{n-1} + (-1)^{n-2} \text{trace}[\mathbf{A}] \mathbf{A}^{n-2} + \dots \right). \quad (15.62)$$

In this case, the solution vector

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b} \quad (15.63)$$

is a linear combination of the vectors $\mathbf{A}^{n-1} \mathbf{b}, \dots, \mathbf{b}$.

Of course, if one needs all of the vectors $\mathbf{A}^{n-1} \mathbf{b}, \dots, \mathbf{b}$ to make a satisfactorily accurate approximation to the actual solution vector \mathbf{x} , then there is no computational advantage with respect to Gaussian elimination. For this reason a Krylov method is terminated as soon as the norm of the k^{th} residual is acceptably small.

The vectors that span $\mathcal{K}_k(\mathbf{b}, \mathbf{A})$ are generally not orthogonal. As a result, computational problems may result unless an orthogonal basis is obtained. Orthogonalization of the basis of $\mathcal{K}_k(\mathbf{b}, \mathbf{A})$ is an important step in conjugate-gradient algorithms.

A necessary condition for any iterative method to define a contraction mapping is that $\{\mathbf{x}_n\}$ must be a Cauchy sequence in the vector space to which the solution vector \mathbf{x} belongs (Volume 1, p. 556). Equivalently, $\{\|\mathbf{x} - \mathbf{x}_n\|\}$ must be a Cauchy sequence in \mathbb{R} .

Although it is tempting to use convergence to zero of the sequence of residual norms, $\{\mathbf{r}_n\}$, as a criterion for convergence of the approximate solutions \mathbf{x}_n to the solution vector \mathbf{x} , it is possible for $\|\mathbf{r}_n\|$ to be small while $\|\mathbf{x} - \mathbf{x}_n\|$ is large. Although a small residual norm may be used as a practical criterion for terminating an iterative method, proofs of convergence must be based on the convergence properties of the sequence $\{\mathbf{x}_n\}$.

15.5 The conjugate-gradient method

The conjugate-gradient method is a Krylov method for solving large systems of linear equations $\mathbf{Ax} = \mathbf{b}$ when the coefficient matrix \mathbf{A} is symmetric and positive-definite:

$$\mathbf{A} \in \mathbb{R}^{n \times n}, \quad \mathbf{A}^T = \mathbf{A}, \quad (15.64)$$

$$\forall \mathbf{x} \in \mathbb{R}^n : \exists: \mathbf{x} \neq \mathbf{0} : \quad \mathbf{x}^T \mathbf{Ax} > 0. \quad (15.65)$$

In principle, the method of conjugate gradients directly constructs a solution of $\mathbf{Ax} = \mathbf{b}$ in a finite number of steps. In practice, the number of steps necessary for a full solution may be so large that, in practice, the process of construction is terminated as soon as the residual norm $\|\mathbf{b} - \mathbf{Ax}\|$ is judged to be sufficiently small.

15.5.1 Choice of a functional for minimization

One way to approach the problem of solving $\mathbf{Ax} = \mathbf{b}$ is to find a functional of \mathbf{y} that is minimized when \mathbf{y} is equal to the solution vector, \mathbf{x} . Since \mathbf{A} is symmetric and positive definite, one can define the new inner product

$$\langle\langle \mathbf{y}, \mathbf{x} \rangle\rangle := \langle \mathbf{y}, \mathbf{Ax} \rangle := \mathbf{y}^T \mathbf{Ax}. \quad (15.66)$$

Because \mathbf{A} is positive definite,

$$\langle\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle\rangle \geq 0, \quad (15.67)$$

equality occurring if and only if $\mathbf{y} = \mathbf{x}$. Now

$$\begin{aligned} \frac{1}{2} \langle\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle\rangle &= \frac{1}{2} \langle\langle \mathbf{y}, \mathbf{y} \rangle\rangle - \langle\langle \mathbf{y}, \mathbf{x} \rangle\rangle + \frac{1}{2} \langle\langle \mathbf{x}, \mathbf{x} \rangle\rangle \\ &= \frac{1}{2} \langle \mathbf{y}, \mathbf{Ay} \rangle - \langle \mathbf{y}, \mathbf{Ax} \rangle + \frac{1}{2} \langle \mathbf{x}, \mathbf{Ax} \rangle. \end{aligned} \quad (15.68)$$

So far the functional depends explicitly upon the unknown exact solution, \mathbf{x} . Therefore we use the linear equation to replace \mathbf{Ax} with \mathbf{b} in $\langle \mathbf{y}, \mathbf{Ax} \rangle$, obtaining

$$\frac{1}{2} \langle\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle\rangle = \frac{1}{2} \langle \mathbf{y}, \mathbf{Ay} \rangle - \langle \mathbf{y}, \mathbf{b} \rangle + \frac{1}{2} \langle \mathbf{x}, \mathbf{Ax} \rangle. \quad (15.69)$$

If \mathbf{A} has full rank, then $\mathbf{b} \in \text{range}[\mathbf{A}]$, and the minimum value of $\frac{1}{2} \langle\langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle\rangle$ is 0, achieved for $\mathbf{y} = \mathbf{x}$. It follows that, if \mathbf{A} has full rank, then the vector \mathbf{y} that minimizes the functional

$$\phi[\mathbf{y}] := \frac{1}{2} \langle \mathbf{y}, \mathbf{Ay} \rangle - \langle \mathbf{y}, \mathbf{b} \rangle \quad (15.70)$$

solves the system of linear equations $\mathbf{Ax} = \mathbf{b}$. The minimum value of $\phi[\mathbf{y}]$ is $-\frac{1}{2} \langle \mathbf{x}, \mathbf{Ax} \rangle$.

Since minimizing $\phi[\mathbf{y}]$ is equivalent to minimizing $\langle\langle \mathbf{z}, \mathbf{z} \rangle\rangle = \langle \mathbf{z}, \mathbf{Az} \rangle$, where $\mathbf{z} = \mathbf{y} - \mathbf{x}$, it is helpful to have a geometrical interpretation of $\langle\langle \mathbf{z}, \mathbf{z} \rangle\rangle$. The vectors \mathbf{z} that satisfy the equation

$$\langle\langle \mathbf{z}, \mathbf{z} \rangle\rangle = \langle \mathbf{z}, \mathbf{Az} \rangle = 1 \quad (15.71)$$

lie on the hyperellipsoid described by the equation

$$z^i a_{ij} z^j = 1. \quad (15.72)$$

Since \mathbf{A} is real, symmetric and positive-definite, we can find an orthogonal transformation \mathbf{V} such that

$$\mathbf{V}^T \mathbf{AV} = \mathbf{\Lambda} \quad (15.73)$$

where $\mathbf{\Lambda} = \text{diag}[\lambda_1, \dots, \lambda_n]$ and each $\lambda_i > 0$. Clearly the columns of \mathbf{V} are the vectors that define the principal axes of the hyperellipsoid (Volume 1, pp. 499–502).

Then

$$\begin{aligned}\langle \mathbf{z}, \mathbf{Az} \rangle &= \mathbf{z}^T \mathbf{Az} = \mathbf{z}^T \mathbf{V} \mathbf{\Lambda} \mathbf{Vz} \\ &= \mathbf{w}^T \mathbf{\Lambda} \mathbf{w}\end{aligned}\tag{15.74}$$

where $\mathbf{w} = \mathbf{Vz}$ is the same abstract vector as \mathbf{z} , referred to new axes. Then the equation of the hyperellipsoid is

$$\sum_{i=1}^n (w^i)^2 \lambda_i = 1.\tag{15.75}$$

Let $\lambda_i = \sigma_i = \frac{1}{c_i^2}$ where $c_i > 0$, and assume that

$$\sigma_1 = \frac{1}{c_1^2} \geq \dots \geq \frac{1}{c_n^2} = \sigma_n.\tag{15.76}$$

Then the equation of the hyperellipsoid is

$$\sum_{i=1}^n \left(\frac{w^i}{c_i} \right)^2 = 1,\tag{15.77}$$

The semi-major axis is c_n , and the semi-minor axis is c_1 . The condition number of \mathbf{A} is

$$\text{cond}_2[\mathbf{A}] = \frac{\sigma_1}{\sigma_n} = \frac{c_n^2}{c_1^2}.\tag{15.78}$$

A well-conditioned matrix corresponds to a hyperellipsoid that is nearly spherical, while a poorly conditioned matrix corresponds to a hyperellipsoid that is elongated and needle-like, or flattened and pancake-like, or both, as in an ellipsoid that is shaped like a leaf with rounded ends.

15.5.2 Construction of approximate solutions

In an iterative approach to minimizing $\phi[\mathbf{y}]$, one constructs a sequence $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k, \dots\}$ of approximate solutions and attempts to reduce the norm of the residual

$$\mathbf{r}_k := \mathbf{b} - \mathbf{Ax}_k\tag{15.79}$$

at each step. Let

$$\mathbf{x}_k := \mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k,\tag{15.80}$$

where we write $\alpha_k \mathbf{p}_k$ instead of \mathbf{p}_k in order to facilitate the normalization of \mathbf{p}_k .

Having found approximate solutions for $k = 1, \dots, n-1$, and supposing for a moment that we have found a suitable \mathbf{p}_k , we minimize $\phi[\mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k]$ with respect to α in order to obtain α_k . Now

$$\begin{aligned}
\phi[\mathbf{x}_{k-1} + \alpha \mathbf{p}_k] &= \frac{1}{2} \langle \mathbf{x}_{k-1} + \alpha \mathbf{p}_k, \mathbf{A}(\mathbf{x}_{k-1} + \alpha \mathbf{p}_k) \rangle - \langle \mathbf{x}_{k-1} + \alpha \mathbf{p}_k, \mathbf{b} \rangle \\
&= \frac{1}{2} \langle \mathbf{x}_{k-1}, \mathbf{A} \mathbf{x}_{k-1} \rangle + \frac{1}{2} \alpha \langle \mathbf{p}_k, \mathbf{A} \mathbf{x}_{k-1} \rangle + \frac{1}{2} \alpha \langle \mathbf{x}_{k-1}, \mathbf{A} \mathbf{p}_k \rangle \\
&\quad + \frac{1}{2} \alpha^2 \langle \mathbf{p}_k, \mathbf{A} \mathbf{p}_k \rangle - \langle \mathbf{x}_{k-1}, \mathbf{b} \rangle - \alpha \langle \mathbf{p}_k, \mathbf{b} \rangle \\
&= \frac{1}{2} \langle \mathbf{x}_{k-1}, \mathbf{A} \mathbf{x}_{k-1} \rangle + \alpha \langle \mathbf{p}_k, \mathbf{A} \mathbf{x}_{k-1} \rangle \\
&\quad + \frac{1}{2} \alpha^2 \langle \mathbf{p}_k, \mathbf{A} \mathbf{p}_k \rangle - \langle \mathbf{x}_{k-1}, \mathbf{b} \rangle - \alpha \langle \mathbf{p}_k, \mathbf{b} \rangle \\
&= \left[\frac{1}{2} \langle \mathbf{x}_{k-1}, \mathbf{A} \mathbf{x}_{k-1} \rangle - \langle \mathbf{x}_{k-1}, \mathbf{b} \rangle \right] - \alpha \langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle + \\
&\quad \frac{1}{2} \alpha^2 \langle \mathbf{p}_k, \mathbf{A} \mathbf{p}_k \rangle. \tag{15.81}
\end{aligned}$$

The minimum of the right-hand side of Eq. (15.81) occurs when

$$\alpha = \alpha_k := \frac{\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A} \mathbf{p}_k \rangle}. \tag{15.82}$$

The computational usefulness of our approach depends on choosing \mathbf{p}_k well.

An obvious but poor choice of the vectors \mathbf{p}_k defines the **method of steepest descent**,

$$\mathbf{p}_k = \mathbf{r}_{k-1}. \tag{15.83}$$

When \mathbf{A} is ill-conditioned this method is exceedingly slow to converge. To understand why, we note that

$$\frac{\partial \phi}{\partial x_{k-1}^j} = (\mathbf{A} \mathbf{x}_{k-1} - \mathbf{b})^j = -\mathbf{r}_{k-1}^j. \tag{15.84}$$

Then the residual \mathbf{r}_k is equal to minus the gradient of ϕ at \mathbf{x}_{k-1} . When we change \mathbf{x}_{k-1} by $\alpha_k \mathbf{r}_k$, we move in the direction of maximum negative change of ϕ (hence the name, “steepest descent”). Unfortunately, when \mathbf{A} is ill-conditioned and when \mathbf{x}_{k-1} does not lie on the major axis, then the direction of steepest descent is very nearly transverse to the major axis. Thus the sequence $\{\mathbf{x}_k\}$ wanders back and forth across a narrow valley with steep walls, descending very slowly towards the global minimum.

A better approach is to ensure that \mathbf{x}_k is one of the partial sums in an expansion of the exact solution, \mathbf{x} , in terms of n vectors \mathbf{p}_k which are mutually \mathbf{A} -orthogonal, *i.e.*, which are mutually orthogonal under the inner product $\langle \langle \cdot, \cdot \rangle \rangle$:

$$k \neq l \Rightarrow \langle \langle \mathbf{p}_k, \mathbf{p}_l \rangle \rangle = 0. \tag{15.85}$$

Mutual \mathbf{A} -orthogonality guarantees that the vectors $\{\mathbf{p}_k\}$ are linearly independent, hence a basis. Let

$$\mathbf{x}_k = \sum_{i=1}^k \frac{\langle \langle \mathbf{p}_i, \mathbf{x} \rangle \rangle}{\langle \langle \mathbf{p}_i, \mathbf{p}_i \rangle \rangle} \mathbf{p}_i \Rightarrow \mathbf{x}_k = \mathbf{x}_{k-1} + \frac{\langle \langle \mathbf{p}_k, \mathbf{x} \rangle \rangle}{\langle \langle \mathbf{p}_k, \mathbf{p}_k \rangle \rangle} \mathbf{p}_k, \quad \mathbf{x}_0 = \mathbf{0}. \tag{15.86}$$

Then there can be at most n \mathbf{x}'_k 's, and we are guaranteed that convergence will occur in n steps. A more useful point of view is that exact orthogonality is hard to achieve numerically, and n may be very large; therefore we would be well advised to look upon this as another iterative approach.

To turn this idea into an iterative method, we must get rid of the unknown \mathbf{x} . That's easy:

$$\langle\langle \mathbf{p}_k, \mathbf{x} \rangle\rangle = \langle \mathbf{p}_k, \mathbf{Ax} \rangle = \langle \mathbf{p}_k, \mathbf{b} \rangle. \quad (15.87)$$

Since $\mathbf{b} = \mathbf{Ax}_{k-1} + \mathbf{r}_{k-1}$, we have

$$\langle \mathbf{p}_k, \mathbf{b} \rangle = \langle \mathbf{p}_k, \mathbf{Ax}_{k-1} \rangle + \langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle = \langle\langle \mathbf{p}_k, \mathbf{x}_{k-1} \rangle\rangle + \langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle. \quad (15.88)$$

But \mathbf{x}_{k-1} is a linear combination of $\mathbf{p}_1, \dots, \mathbf{p}_{k-1}$, to which \mathbf{p}_k is \mathbf{A} -orthogonal. Then

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k \quad (15.89)$$

where, as before,

$$\alpha_k = \frac{\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{Ap}_k \rangle}. \quad (15.90)$$

Clearly this is a variant of Gram-Schmidt orthogonalization, adapted to a linear-equation problem with a symmetric, positive-definite coefficient matrix.

The remaining "detail" is to find an algorithm for constructing an \mathbf{A} -orthogonal set $\{\mathbf{p}_i\}$. Since we know already that \mathbf{r}_{k-1} is in the direction of steepest descent, we choose \mathbf{p}_k as the component of \mathbf{r}_{k-1} which is \mathbf{A} -orthogonal to the subspace

$$\mathcal{W}_{k-1} := \text{span}[\mathbf{p}_1, \dots, \mathbf{p}_{k-1}]. \quad (15.91)$$

We now construct the \mathbf{A} -orthogonal projector on \mathcal{W}_{k-1} . Since $\langle \mathbf{p}_i, \mathbf{Ay} \rangle = \langle \mathbf{Ap}_i, \mathbf{y} \rangle$, \mathbf{A} -orthogonality to $\{\mathbf{p}_1, \dots, \mathbf{p}_{k-1}\}$ is the same as ordinary orthogonality to $\{\mathbf{Ap}_1, \dots, \mathbf{Ap}_{k-1}\}$. Then the \mathbf{A} -orthogonal projector (call it \mathbf{P}_{k-1}) on \mathcal{W}_{k-1} is the same as the ordinary orthogonal projector on $\mathbf{A}\mathcal{W}_{k-1} = \text{span}[\mathbf{Ap}_1, \dots, \mathbf{Ap}_{k-1}]$.

Define the matrices

$$\mathbf{Q}_{k-1} := (\mathbf{p}_1, \dots, \mathbf{p}_{k-1}), \quad (15.92)$$

which clearly have n rows and $k-1$ columns, and the matrices

$$\mathbf{C}_{k-1} := \mathbf{AQ}_{k-1} = (\mathbf{Ap}_1, \dots, \mathbf{Ap}_{k-1}). \quad (15.93)$$

Then the \mathbf{A} -orthogonal projector on \mathcal{W}_{k-1} is the ordinary orthogonal projector on $\text{range}[\mathbf{C}_{k-1}]$.

According to Eq. (9.327),

$$\mathbf{P}_{k-1} = \mathbf{P}_{\text{range}[\mathbf{C}_{k-1}]} = \mathbf{C}_{k-1} \mathbf{C}_{k-1}^+, \quad (15.94)$$

where \mathbf{C}_{k-1}^+ is the generalized inverse (Moore-Penrose inverse) of \mathbf{C}_{k-1} (Volume 1, p. 512). The component of \mathbf{r}_{k-1} which is \mathbf{A} -orthogonal to \mathcal{W}_{k-1} is therefore

$$\boxed{\mathbf{p}_k := (\mathbf{1} - \mathbf{P}_{k-1})\mathbf{r}_{k-1} = (\mathbf{1} - \mathbf{C}_{k-1}\mathbf{C}_{k-1}^+)\mathbf{r}_{k-1}.} \quad (15.95)$$

Another way to make the same statement is to say that

$$\mathbf{w}_{k-1} := \mathbf{C}_{k-1}\mathbf{C}_{k-1}^+\mathbf{r}_{k-1} \in \mathcal{W}_{k-1} \quad (15.96)$$

solves the least-squares problem of minimizing

$$\|\mathbf{p}\|_2^2 = \|\mathbf{r}_{k-1} - \mathbf{w}\|_2^2 \quad (15.97)$$

with respect to $\mathbf{w} \in \mathbf{A}\mathcal{W}_{k-1}$.

Rather than compute the residual \mathbf{r}_k as $\mathbf{b} - \mathbf{A}\mathbf{x}_k$, we can get the same result with less computational effort as follows:

$$\mathbf{r}_k - \mathbf{r}_{k-1} = \mathbf{b} - \mathbf{A}\mathbf{x}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_{k-1}) = \mathbf{A}(\mathbf{x}_{k-1} - \mathbf{x}_k). \quad (15.98)$$

It follows from this equation and Eq. (15.89) that

$$\boxed{\mathbf{r}_k - \mathbf{r}_{k-1} = -\alpha_k \mathbf{A}\mathbf{p}_k.} \quad (15.99)$$

(Note that we have to compute $\mathbf{A}\mathbf{p}_k$ in any case, in order to compute α_k .)

15.5.3 Conjugate gradient algorithm (CGA)

We now have a workable algorithm for minimizing $\phi[\mathbf{y}]$:

Choose $\mathbf{x}_0 = \mathbf{0}$. Then $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 = \mathbf{b}$. Set $\mathbf{P}_1 = \mathbf{r}_0 = \mathbf{b}$. [This choice of \mathbf{x}_0 does not work if \mathbf{b} is an eigenvector of \mathbf{A} , for, if $\mathbf{A}\mathbf{b} = \lambda\mathbf{b}$, then the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = \frac{1}{\lambda}\mathbf{b}$.]

For $k = 1, \dots, n-1$:

$$\begin{aligned} \text{Compute } \alpha_k &= \frac{\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle} \\ \mathbf{r}_k &= \mathbf{r}_{k-1} - \alpha_k \mathbf{A}\mathbf{p}_k \\ \mathbf{x}_k &= \mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k \\ \mathbf{p}_{k+1} &= (\mathbf{1} - \mathbf{C}_{k-1}\mathbf{C}_{k-1}^+)\mathbf{r}_k. \end{aligned} \quad (15.100)$$

For example,

$$\begin{aligned} \alpha_k &= \frac{\langle \mathbf{p}_1, \mathbf{r}_0 \rangle}{\langle \mathbf{p}_1, \mathbf{A}\mathbf{p}_1 \rangle} = \frac{\langle \mathbf{b}, \mathbf{b} \rangle}{\langle \mathbf{b}, \mathbf{A}\mathbf{b} \rangle} \\ \mathbf{x}_1 &= \underbrace{\mathbf{x}_0}_{\mathbf{0}} + \frac{\langle \mathbf{b}, \mathbf{b} \rangle}{\langle \mathbf{b}, \mathbf{A}\mathbf{b} \rangle} \mathbf{b} \\ \mathbf{r}_1 &= \mathbf{b} - \mathbf{A}\mathbf{x}_1 = \mathbf{b} - \frac{\langle \mathbf{b}, \mathbf{b} \rangle}{\langle \mathbf{b}, \mathbf{A}\mathbf{b} \rangle} \mathbf{A}\mathbf{b}. \end{aligned} \quad (15.101)$$

Note that $\langle \mathbf{r}_0, \mathbf{r}_1 \rangle = \langle \mathbf{b}, \mathbf{r}_1 \rangle = 0$ and that $\langle \mathbf{p}_1, \mathbf{r}_1 \rangle = \langle \mathbf{b}, \mathbf{r}_1 \rangle = 0$ while $\langle \mathbf{p}_1, \mathbf{r}_0 \rangle \neq 0$. If \mathbf{b} is an eigenvector of \mathbf{A} , then $\mathbf{r}_1 = \mathbf{0}$.

The above method requires the computation of $\mathbf{C}_{k-1} = \mathbf{A}\mathbf{Q}_{k-1}$ and a singular-value decomposition to find \mathbf{C}_{k-1}^+ . If \mathbf{A} is a very large matrix, then we need to find a better way.

We begin with the observation that $\mathbf{p}_1, \dots, \mathbf{p}_{k-1}$ are orthogonal to \mathbf{r}_k . Certainly $\mathbf{p}_k \perp \mathbf{r}_k$:

$$\langle \mathbf{p}_k, \mathbf{r}_k \rangle = \langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle - \alpha_k \langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle = 0 \quad (15.102)$$

since

$$\alpha_k = \frac{\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle}. \quad (15.103)$$

For any \mathbf{p}_j (where $j < k$), we can write \mathbf{r}_k as the telescoping sum

$$\begin{aligned} \mathbf{r}_k &= (\mathbf{r}_k - \mathbf{r}_{k-1}) + (\mathbf{r}_{k-1} - \mathbf{r}_{k-2}) + \cdots + (\mathbf{r}_{j+1} - \mathbf{r}_j) + \mathbf{r}_j \\ &= -\alpha_k \mathbf{A}\mathbf{p}_k - \alpha_{k-1} \mathbf{A}\mathbf{p}_{k-1} - \cdots - \alpha_{j+1} \mathbf{A}\mathbf{p}_{j+1} + \mathbf{r}_j \\ &\Rightarrow \langle \mathbf{p}_j, \mathbf{r}_k \rangle = \langle \mathbf{p}_j, \mathbf{r}_j \rangle = 0. \end{aligned} \quad (15.104)$$

Since

$$\langle \mathbf{p}_j, \mathbf{A}\mathbf{p}_{j+1} \rangle = \cdots = \langle \mathbf{p}_j, \mathbf{A}\mathbf{p}_k \rangle = 0, \quad (15.105)$$

it follows that $\boxed{\mathbf{r}_k \perp \mathcal{W}_k}$.

In other words, as we construct $\mathcal{W}_k = \text{span}[\mathbf{p}_1, \dots, \mathbf{p}_k]$, for increasing k , we restrict \mathbf{r}_k to a dimensionally smaller and smaller part of the range of \mathbf{A} .

We show that our conjugate gradient algorithm implies that if \mathbf{b} is not an eigenvector of \mathbf{A} , then

$$\boxed{\mathcal{W}_k = \text{span}[\mathbf{r}_0, \dots, \mathbf{r}_{k-1}] = \text{span}[\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}] = \mathcal{K}_k(\mathbf{b}, \mathbf{A})}, \quad (15.106)$$

which establishes that \mathcal{W}_k is a Krylov subspace, and therefore that the solution \mathbf{x} lies in a Krylov subspace.

This assertion is true for $k = 1$. Proceeding by induction, we assume that it holds for k and that $\mathbf{r}_1, \dots, \mathbf{r}_{k-1}$ are nonzero. Then

$$\begin{aligned} \mathbf{r}_k &= \mathbf{r}_{k-1} - \alpha_k \mathbf{A}\mathbf{p}_k \\ \Rightarrow \mathbf{r}_k &\in \text{span}[\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}, \mathbf{A}^k\mathbf{b}] \\ \Rightarrow \text{span}[\mathbf{r}_0, \dots, \mathbf{r}_k] &\subseteq \text{span}[\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^k\mathbf{b}]. \end{aligned} \quad (15.107)$$

Now

$$\mathbf{p}_{k+1} = \mathbf{r}_k - \mathbf{C}_k \mathbf{C}_k^+ \mathbf{r}_k = \mathbf{r}_k - \mathbf{A}\mathbf{Q}_k \mathbf{C}_k^+ \mathbf{r}_k = \mathbf{r}_k - \mathbf{A}\mathbf{Q}_k \mathbf{z}_k, \quad (15.108)$$

where $\mathbf{z}_k := \mathbf{C}_k^+ \mathbf{r}_k$. Then $\mathbf{Q}_k \mathbf{z}_k$ is a linear combination of the columns $\mathbf{p}_1, \dots, \mathbf{p}_k$, hence a linear combination of $\mathbf{b}, \dots, \mathbf{A}^{k-1} \mathbf{b}$. Clearly, then, $\mathbf{A} \mathbf{Q}_k \mathbf{z}_k$ is a linear combination of $\mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}$. It follows that

$$\mathbf{p}_{k+1} \in \text{span}[\mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}]. \quad (15.109)$$

$$\Rightarrow \mathcal{W}_{k+1} \subseteq \text{span}[\mathbf{b}, \mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}]. \quad (15.110)$$

$$\begin{aligned} \text{Now } \dim[\mathcal{W}_{k+1}] &= k+1 \leq \dim[\text{span}[\mathbf{b}, \mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}]] \\ &\leq k+1 \end{aligned} \quad (15.111)$$

(since there are $k+1$ vectors $\mathbf{b}, \mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}$). It follows that

$$\mathcal{W}_{k+1} = \text{span}[\mathbf{b}, \mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}]. \quad (15.112)$$

It is also the case that $\mathbf{A} \mathcal{W}_k = \mathbf{A} \text{span}[\mathbf{b}, \dots, \mathbf{A}^{k-1} \mathbf{b}] \subset \mathcal{W}_{k+1}$ and that the set $\{\mathbf{b}, \mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}\}$ is linearly independent (otherwise a $k+1$ -dimensional space would be spanned by fewer than $k+1$ linearly independent vectors).

To show that $\text{span}[\mathbf{r}_0, \dots, \mathbf{r}_k] = \mathcal{W}_{k+1}$, we must show that the \mathbf{r}_j 's are linearly independent. By the inductive hypothesis, $\mathcal{W}_j = \text{span}[\mathbf{r}_0, \dots, \mathbf{r}_{j-1}]$ for $j = 1, \dots, k$. But for every such j , $\mathbf{r}_j \perp \mathcal{W}_j \Rightarrow \mathbf{r}_j \perp \mathbf{r}_0, \dots, \mathbf{r}_j \perp \mathbf{r}_{j-1}$ for every $j = 1, \dots, k$. Then: \mathbf{r}_j is linearly independent of $\{\mathbf{r}_0, \dots, \mathbf{r}_{j-1}\}$

$$\Rightarrow \{\mathbf{r}_0, \dots, \mathbf{r}_k\} \text{ is linearly independent}$$

$$\Rightarrow \dim[\text{span}[\mathbf{r}_0, \dots, \mathbf{r}_k]] = k+1$$

$$\Rightarrow \boxed{\text{span}[\mathbf{r}_0, \dots, \mathbf{r}_k] = \text{span}[\mathbf{b}, \mathbf{A} \mathbf{b}, \dots, \mathbf{A}^k \mathbf{b}] = \mathcal{W}_{k+1} = \mathcal{K}_{k+1}(\mathbf{b}, \mathbf{A})}. \quad (15.113)$$

We now simplify the expression for \mathbf{p}_{k+1} ,

$$\mathbf{p}_{k+1} = \mathbf{r}_k - \mathbf{P}_k \mathbf{r}_k, \quad (15.114)$$

in order to eliminate matrix multiplication by the projector \mathbf{P}_k . It is possible to ensure that \mathbf{p}_{k+1} is \mathbf{A} -orthogonal to \mathcal{W}_k much more simply than with a projection operator, by setting

$$\mathbf{p}_{k+1} = \mathbf{r}_k + \beta_k \mathbf{p}_k \quad (15.115)$$

and choosing the constant β_k appropriately.

For every $j = 1, \dots, k-1$,

$$\langle \mathbf{A} \mathbf{p}_j, \mathbf{p}_{k+1} \rangle = \langle \mathbf{A} \mathbf{p}_j, \mathbf{r}_k \rangle + \beta_k \langle \mathbf{A} \mathbf{p}_j, \mathbf{p}_k \rangle. \quad (15.116)$$

If $\{\mathbf{p}_1, \dots, \mathbf{p}_k\}$ is a mutually \mathbf{A} -orthogonal set, then the second term on the right vanishes. Certainly $\mathbf{A} \mathbf{p}_j \in \mathcal{W}_k$ (since $\mathbf{A} \mathcal{W}_j \subset \mathcal{W}_{j+1} \subseteq \mathcal{W}_k$). Since

$\mathbf{r}_k \perp \mathcal{W}_k$, the first term on the right vanishes, establishing that \mathbf{p}_{k+1} is \mathbf{A} -orthogonal to $\mathbf{p}_1, \dots, \mathbf{p}_{k-1}$. If $j = k$, then $\langle \mathbf{A}\mathbf{p}_k, \mathbf{p}_{k+1} \rangle = 0$ if and only if

$$\boxed{\beta_k = -\frac{\langle \mathbf{p}_k, \mathbf{A}\mathbf{r}_k \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle}} \quad (15.117)$$

We can use Eq. (15.117) to simplify the numerator of our expansion for α_k ,

$$\alpha_k = \frac{\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle}. \quad (15.118)$$

Since $\mathbf{p}_k = \mathbf{r}_{k-1} + \beta_{k-1}\mathbf{p}_{k-1}$, it follows that

$$\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle = \langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle \quad (15.119)$$

because $\langle \mathbf{p}_k, \mathbf{r}_{k-1} \rangle = 0$. Then

$$\boxed{\alpha_k = \frac{\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle}}. \quad (15.120)$$

The recurrence relation for the residual vector,

$$\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_k \mathbf{A}\mathbf{p}_k, \quad (15.121)$$

can be used to simplify our expression for β_k . We have

$$\langle \mathbf{r}_k, \mathbf{r}_k \rangle = -\alpha_k \langle \mathbf{r}_k, \mathbf{A}\mathbf{p}_k \rangle \quad (15.122)$$

since $\mathbf{r}_k \perp \mathbf{r}_{k-1}$. Then

$$\langle \mathbf{r}_k, \mathbf{A}\mathbf{p}_k \rangle = -\frac{1}{\alpha_k} \langle \mathbf{r}_k, \mathbf{r}_k \rangle. \quad (15.123)$$

It follows that

$$\beta_k = -\frac{\langle \mathbf{p}_k, \mathbf{A}\mathbf{r}_k \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle} = \frac{1}{\alpha_k} \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle} = \frac{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle}{\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle} \cdot \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle} \quad (15.124)$$

and therefore that

$$\boxed{\beta_k = \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle}}. \quad (15.125)$$

The resulting **conjugate-gradient algorithm (CGA)** is

$$\mathbf{x}_0 = \mathbf{0}, \quad \mathbf{r}_0 = \mathbf{b}, \quad \mathbf{p}_1 = \mathbf{r}_0 = \mathbf{b};$$

For $k = 1, \dots, n-1$:

$$\begin{aligned}\alpha_k &= \frac{\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle} \\ \mathbf{r}_k &= \mathbf{r}_{k-1} - \alpha_k \mathbf{A}\mathbf{p}_k \\ \mathbf{x}_k &= \mathbf{x}_{k-1} - \alpha_k \mathbf{p}_k \\ \beta_k &= \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle} \\ \mathbf{p}_{k+1} &= \mathbf{r}_k + \beta_k \mathbf{p}_k\end{aligned}$$

15.5.4 Convergence of the CGA

If $\mathbf{A} = \mathbf{1} + \mathbf{B}$ and $\text{rank}[\mathbf{B}] = r$, then the CGA converges in at most $r + 1$ steps, for

$$\begin{aligned}\dim[\mathcal{W}_k] &= \dim[\text{span}[\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}]] \\ &= \dim[\text{span}[\mathbf{b}, \mathbf{B}\mathbf{b}, \dots, \mathbf{B}^{k-1}\mathbf{b}]] \\ &\leq r + 1,\end{aligned}\tag{15.126}$$

since $r = \dim[\text{range}[\mathbf{B}]]$ by definition.

For example, if $\mathbf{B} = \mathbf{0}$ (i.e., \mathbf{A} is the identity matrix), then $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$, $\mathbf{p}_1 = \mathbf{b}$,

$$\alpha_1 = \frac{\langle \mathbf{r}_0, \mathbf{r}_0 \rangle}{\langle \mathbf{p}_1, \mathbf{A}\mathbf{p}_1 \rangle} = \frac{\langle \mathbf{b}, \mathbf{b} \rangle}{\langle \mathbf{b}, \mathbf{b} \rangle} = 1, \quad \text{and} \quad \mathbf{x}_1 = \mathbf{x}_0 + \alpha_1 \mathbf{A}\mathbf{p}_1 = \mathbf{b}.$$

In this trivial example, the CGA converges in $r + 1$ steps, where $r = 0$. In general, the convergence of the CGA is more rapid, the more nearly diagonal \mathbf{A} is.

It can be shown² that

$$\|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}} \leq \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}} \left[\frac{\sqrt{\text{cond}_2[\mathbf{A}]} - 1}{\sqrt{\text{cond}_2[\mathbf{A}]} + 1} \right]^k\tag{15.127}$$

where $\|\mathbf{y}\|_{\mathbf{A}} := \langle \mathbf{y}, \mathbf{A}\mathbf{y} \rangle^{1/2}$. This establishes that $\mathbf{x}_k \mapsto \mathbf{x}_{k+1}$ is a contraction mapping.

Although this error estimate is often unrealistically conservative, it implies that one way to guarantee that the CGA converges rapidly is to ensure that \mathbf{A} (or some symmetric, positive-definite transform of \mathbf{A}) has a condition number approximately equal to 1.

²D.G Luenberger, *Introduction to Linear & Nonlinear Programming* (Addison-Wesley, 1993), p. 187.

15.5.5 Preconditioning a matrix for the CGA

The goal of preconditioning is to make \mathbf{A} as diagonal as possible, and as well-conditioned as possible, given that we must compute cheaply. Since \mathbf{A} is real, symmetric and positive-definite, there exists a real orthogonal matrix \mathbf{V} such that

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \quad (15.128)$$

$$\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{1} \quad (15.129)$$

$$\mathbf{\Lambda} = \text{diag}[\lambda_1, \dots, \lambda_n] \quad (15.130)$$

$$\lambda_1 \geq \dots \geq \lambda_n > 0. \quad (15.131)$$

Let

$$\left. \begin{aligned} \mathbf{C} &:= \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}^T \\ \mathbf{\Lambda}^{1/2} &:= \text{diag}[\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}] \end{aligned} \right\} \Rightarrow \mathbf{C}^T = \mathbf{C}. \quad (15.132)$$

Then

$$\begin{aligned} \mathbf{C}^{-1} &= \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}^T \\ \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1} &= \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2}\mathbf{V}^T = \mathbf{1}. \end{aligned} \quad (15.133)$$

It follows that there always exists a symmetric, positive-definite matrix \mathbf{C} such that $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}$ has condition number 1. In practice, one tries to lower the condition number, but not necessarily to 1.

The system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is equivalent to the system

$$\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}\mathbf{C}\mathbf{x} = \mathbf{C}^{-1}\mathbf{b}. \quad (15.134)$$

Let

$$\left. \begin{aligned} \mathbf{A}' &:= \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}, \\ \mathbf{x}' &:= \mathbf{C}\mathbf{x}, \\ \mathbf{b}' &:= \mathbf{C}^{-1}\mathbf{b} \end{aligned} \right\} \Rightarrow \mathbf{A}'\mathbf{x}' = \mathbf{b}'. \quad (15.135)$$

If we apply the CGA to the new system, obtaining

$$\alpha'_k = \frac{\langle \mathbf{r}'_{k-1}, \mathbf{r}'_{k-1} \rangle}{\langle \mathbf{p}'_k, \mathbf{A}'\mathbf{p}'_k \rangle}, \quad \mathbf{r}'_k = \mathbf{r}'_{k-1} - \alpha'_k \mathbf{A}'\mathbf{p}'_k, \quad (15.136)$$

we have $\mathbf{x}'_k = \mathbf{x}'_{k-1} + \alpha'_k \mathbf{p}'_k$, $\beta'_k = \frac{\langle \mathbf{r}'_k, \mathbf{r}'_k \rangle}{\langle \mathbf{r}'_{k-1}, \mathbf{r}'_{k-1} \rangle}$, and

$$\mathbf{p}'_{k+1} = \mathbf{r}'_k + \beta'_k \mathbf{p}'_k. \quad (15.137)$$

In order to avoid the computational labor needed to compute $\mathbf{A}' = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}$, we define $\mathbf{C}^{-1}\mathbf{b}'' := \mathbf{b}'$, $\mathbf{C}\mathbf{p}''_k := \mathbf{p}'_k$, $\mathbf{C}\mathbf{x}''_k := \mathbf{x}'_k$, and $\mathbf{C}^{-1}\mathbf{r}''_k := \mathbf{r}'_k$. Since $\mathbf{C}^{-1}\mathbf{r}_k = \mathbf{C}^{-1}\mathbf{b} - \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}\mathbf{C}\mathbf{x}_k$, it follows that $\mathbf{b}'' = \mathbf{b}$,

$$\alpha'_k = \frac{\langle \mathbf{r}''_{k-1}, \mathbf{C}^{-2}\mathbf{r}''_{k-1} \rangle}{\langle \mathbf{p}''_k, \mathbf{A}\mathbf{p}''_k \rangle}, \quad \beta'_k = \frac{\langle \mathbf{r}''_k, \mathbf{C}^{-2}\mathbf{r}''_k \rangle}{\langle \mathbf{r}''_{k-1}, \mathbf{C}^{-2}\mathbf{r}''_{k-1} \rangle}, \quad (15.138)$$

and $\mathbf{x}_k'' = \mathbf{x}_{k-1}'' + \alpha_k' \mathbf{p}_k''$, $\mathbf{r}_k'' = \mathbf{r}_{k-1}'' - \alpha_k' \mathbf{A} \mathbf{p}_k''$.

The change of variables from primed to double-primed makes \mathbf{A} appear instead of $\mathbf{C}^{-1} \mathbf{A} \mathbf{C}^{-1}$. Indeed, \mathbf{C} enters only in the inner products $\langle \mathbf{r}_j'', \mathbf{C}^{-2} \mathbf{r}_j'' \rangle$.

Since it is preferable to obtain $\mathbf{C}^{-2} \mathbf{r}_j''$ as the solution of a system of linear equations, we let

$$\begin{aligned} \mathbf{M} &:= \mathbf{C}^2 \\ \mathbf{M} \mathbf{z}_j &:= \mathbf{r}_j''. \end{aligned} \quad (15.139)$$

Then

$$\mathbf{p}_{k+1}'' = \mathbf{z}_k + \beta_k' \mathbf{p}_k''. \quad (15.140)$$

The **preconditioned conjugate-gradient algorithm** is

$$\mathbf{x}_0 = \mathbf{0}, \quad \mathbf{r}_0 = \mathbf{b}; \text{ solve } \mathbf{M} \mathbf{z}_0 = \mathbf{b}; \quad \mathbf{p}_1 = \mathbf{z}_0$$

For $k = 1, \dots, n-1$:

$$\begin{aligned} \alpha_k' &= \frac{\langle \mathbf{z}_{k-1}, \mathbf{r}_{k-1} \rangle}{\langle \mathbf{p}_k, \mathbf{A} \mathbf{p}_k \rangle} \\ \mathbf{r}_k &= \mathbf{r}_{k-1} - \alpha_k \mathbf{A} \mathbf{p}_k \\ \mathbf{x}_k &= \mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k \\ \text{solve } \mathbf{M} \mathbf{z}_k &= \mathbf{r}_k \\ \beta_k' &= \frac{\langle \mathbf{z}_k, \mathbf{r}_k \rangle}{\langle \mathbf{z}_{k-1}, \mathbf{r}_{k-1} \rangle} \\ \mathbf{p}_{k+1} &= \mathbf{z}_k + \beta_k' \mathbf{p}_k, \end{aligned} \quad (15.141)$$

Since

$$\left\{ \begin{array}{l} \langle \mathbf{r}_j', \mathbf{r}_l' \rangle = 0 \\ \langle \mathbf{p}_j', \mathbf{A} \mathbf{p}_l' \rangle = 0 \end{array} \right\} \quad \text{if } j \neq l \quad (15.142)$$

we have

$$\left\{ \begin{array}{l} \langle \mathbf{r}_j, \mathbf{M}^{-1} \mathbf{r}_l \rangle = 0 \\ \langle \mathbf{p}_j, \mathbf{A} \mathbf{p}_l \rangle = 0 \end{array} \right\} \quad \text{if } j \neq l. \quad (15.143)$$

In order for this approach to be useful, \mathbf{M} must be chosen such that the computational effort required to solve $\mathbf{M} \mathbf{z}_j = \mathbf{r}_j$ is much smaller than $\frac{1}{N}$ times the effort normally required to solve $\mathbf{A} \mathbf{x} = \mathbf{b}$ in approximately N steps.

Exercises for Section 15.5

15.5.1 Prove that the minimum of the right-hand side of Eq. (15.81) occurs when Eq. (15.82) holds. [Note that you need to establish that Eq. (15.82) defines an extremum, and then that the extremum is, in fact, a minimum.]

15.5.2 Use the conjugate-gradient algorithm (CGA) to solve the system in Eq. (15.52). Compute the ∞ -norm of the difference between the CGA solution and the Thomas solution after 3, 6 and 10 CGA iterations. Using the UNIX `time` command, or a higher-level-language timing routine, obtain the execution time required for 3, 6 and 10 iterations, and compare with the time required for Gaussian elimination.