where $\boldsymbol{x}$ is the solution to $\boldsymbol{Ax}=\boldsymbol{f}$. This, however, is obviously impossible in practice, because the solution $\boldsymbol{x}$ is not known. Therefore, alternative strategies must be used. For example, one can enforce the orthogonality $\boldsymbol{r}^{(p)} \perp K_p(\boldsymbol{A},\boldsymbol{r}^{(0)})$, i.e., require that $\forall \boldsymbol{u} \in K_p(\boldsymbol{A},\boldsymbol{r}^{(0)}):\ (\boldsymbol{u},\boldsymbol{Ax}^{(p)}-\boldsymbol{f})=0$. This leads to the Arnoldi method, which is also called the full orthogonalization method (FOM). This method is known to reduce to conjugate gradients (Section 5.6) if $\boldsymbol{A}=\boldsymbol{A}^*>0$. Alternatively, one can minimize the Euclidean norm of the residual $\boldsymbol{r}^{(p)}=\boldsymbol{Ax}^{(p)}-\boldsymbol{f}$:

$$\boldsymbol{x}^{(p)} = \arg\min_{z\in N_p} \|\boldsymbol{Az}-\boldsymbol{f}\|_2.$$

This strategy defines the so-called method of generalized minimal residuals (GMRES) that we describe in Section 6.3.2.

It is clear though that before FOM, GMRES, or any other Krylov subspace method can actually be implemented, we need to thoroughly describe the minimization search space N_p of (6.70). This is equivalent to describing the Krylov subspace $K_p(\boldsymbol{A},\boldsymbol{r}^{(0)})$. In other words, we need to construct a basis in the space $K_p(\boldsymbol{A},\boldsymbol{r}^{(0)})$ introduced by Definition 6.1 for $p=m$ and $\boldsymbol{r}^{(0)}=\boldsymbol{u}$.

The first result which is known is that in general the dimension of the space $K_m(\boldsymbol{A},\boldsymbol{u})$ is a non-decreasing function of m. This dimension, however, may actually be lower than m as it is not guaranteed ahead of time that all the vectors: $\boldsymbol{u}, \boldsymbol{Au}, \boldsymbol{A}^2\boldsymbol{u}$, ..., $\boldsymbol{A}^{m-1}\boldsymbol{u}$ are linearly independent.

For a given m one can obtain an orthonormal basis in the space $K_m(\boldsymbol{A},\boldsymbol{u})$ with the help of the so-called Arnoldi process, which is based on the well known Gram-Schmidt orthonormalization algorithm (all norms are Euclidean):

$$\begin{aligned}
&\boldsymbol{u}_1 = \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|},\\
&\boldsymbol{v}_1 = \boldsymbol{Au}_1 - (\boldsymbol{u}_1,\boldsymbol{Au}_1)\boldsymbol{u}_1, \quad \boldsymbol{u}_2 = \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|},\\
&\boldsymbol{v}_2 = \boldsymbol{Au}_2 - (\boldsymbol{u}_1,\boldsymbol{Au}_2)\boldsymbol{u}_1 - (\boldsymbol{u}_2,\boldsymbol{Au}_2)\boldsymbol{u}_2, \quad \boldsymbol{u}_3 = \frac{\boldsymbol{v}_2}{\|\boldsymbol{v}_2\|},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\boldsymbol{v}_k = \boldsymbol{Au}_k - \sum_{j=1}^{k} (\boldsymbol{u}_j,\boldsymbol{Au}_k)\boldsymbol{u}_j, \quad \boldsymbol{u}_{k+1} = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{6.71}$$

Obviously, all the resulting vectors $\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_k,\ldots$ are orthonormal. If the Arnoldi process terminates at step m, then the vectors $\{\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_m\}$ will form an orthonormal basis in $K_m(\boldsymbol{A},\boldsymbol{u})$. The process can also terminate prematurely, i.e., yield $\boldsymbol{v}_k=\boldsymbol{0}$ at some $k<m$. This will indicate that the dimension of the corresponding Krylov subspace is lower than m. Note also that the classical Gram-Schmidt orthogonalization is prone to numerical instabilities. Therefore, in practice one often uses its stabilized version (see Remark 7.4 on page 219). The latter is not completely fail proof either, yet it is more robust and somewhat more expensive computationally.

Let us now introduce the matrix $\boldsymbol{U}_k = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k]$ composed of the n-dimensional orthonormal vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$ as columns. From the last equation (6.71) is is clear that the vector $\boldsymbol{A}\boldsymbol{u}_k$ is a linear combination of the vectors $\boldsymbol{u}_{k+1}, \boldsymbol{u}_k, \ldots, \boldsymbol{u}_1$. Consequently, we can write:

$$\boldsymbol{A}\boldsymbol{U}_k = \boldsymbol{U}_{k+1}\boldsymbol{H}_k, \tag{6.72}$$

where $\boldsymbol{H}_k$ is a matrix in the upper Hessenberg form. This matrix has $k+1$ rows and k columns, and all its entries below the first sub-diagonal are equal to zero:

$$\boldsymbol{H}_k = \{\mathfrak{h}_{i,j} \,|\, i = 1,2,\ldots,k+1,\ j = 1,2,\ldots,k,\ \mathfrak{h}_{i,j} = 0 \text{ for } i > j+1\}.$$

Having introduced the procedure for building a basis in a Krylov subspace, we can now analyze a particular iteration method — the GMRES.

6.3.2 GMRES

In GMRES, we are minimizing the Euclidean norm of the residual, $\|\cdot\| \equiv \|\cdot\|_2$:

$$\begin{aligned} \|\boldsymbol{r}^{(p)}\| = \|\boldsymbol{A}\boldsymbol{x}^{(p)} - \boldsymbol{f}\| = \min_{z \in N_p} \|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{f}\|, \\ N_p = \boldsymbol{x}^{(0)} + K_p(\boldsymbol{A}, \boldsymbol{r}^{(0)}). \end{aligned} \tag{6.73}$$

Let $\boldsymbol{U}_p$ be an $n \times p$ matrix with orthonormal columns such that these columns form a basis in the space $K_p(\boldsymbol{A}, \boldsymbol{r}^{(0)})$; this matrix is obtained by means of the Arnoldi process described in Section 6.3.1. Then we can write:

$$\boldsymbol{x}^{(p)} = \boldsymbol{x}^{(0)} + \boldsymbol{U}_p \boldsymbol{w}^{(p)},$$

where $\boldsymbol{w}^{(p)}$ is a vector of dimension p that needs to be determined. Hence, for the residual $\boldsymbol{r}^{(p)}$ we have according to formula (6.72):

$$\boldsymbol{r}^{(p)} = \boldsymbol{r}^{(0)} + \boldsymbol{A}\boldsymbol{U}_p\boldsymbol{w}^{(p)} = \boldsymbol{r}^{(0)} + \boldsymbol{U}_{p+1}\boldsymbol{H}_p\boldsymbol{w}^{(p)} = \boldsymbol{U}_{p+1}(\boldsymbol{q}^{(p+1)} + \boldsymbol{H}_p\boldsymbol{w}^{(p)}), \tag{6.74}$$

where $\boldsymbol{q}^{(p+1)}$ is a $p+1$-dimensional vector defined as follows:

$$\boldsymbol{q}^{(p+1)} = \big[\|\boldsymbol{r}^{(0)}\|, \underbrace{0, 0, \ldots, 0}_{p}\big]^T. \tag{6.75}$$

Indeed, as the first column of the matrix $\boldsymbol{U}_{p+1}$ generated by the Arnoldi process (6.71) is given by $\boldsymbol{r}^{(0)}/\|\boldsymbol{r}^{(0)}\|$, we clearly have: $\boldsymbol{U}_{p+1}\boldsymbol{q}^{(p+1)} = \boldsymbol{r}^{(0)}$ in formula (6.74). Consequently, minimization in the sense of (6.73) reduces to:

$$\|\boldsymbol{r}^{(p)}\| = \min_{\boldsymbol{w}^{(p)} \in \mathbb{R}^p} \|\boldsymbol{q}^{(p+1)} + \boldsymbol{H}_p\boldsymbol{w}^{(p)}\|. \tag{6.76}$$

Equality (6.76) holds because $\boldsymbol{U}_{p+1}^T\boldsymbol{U}_{p+1} = \boldsymbol{I}_{p+1}$, which implies that for the Euclidean norm $\|\cdot\| \equiv \|\cdot\|_2$ we have: $\|\boldsymbol{U}_{p+1}(\boldsymbol{q}^{(p+1)} + \boldsymbol{H}_p\boldsymbol{w}^{(p)})\| = \|\boldsymbol{q}^{(p+1)} + \boldsymbol{H}_p\boldsymbol{w}^{(p)}\|$.

Problem (6.76) is a typical problem of solving an overdetermined system of linear algebraic equations in the sense of the least squares. Indeed, the matrix $\boldsymbol{H}_p$ has $p+1$ rows and p columns, i.e., there are $p+1$ equations and only p unknowns. Solutions of such systems can, generally speaking, only be found in the weak sense, in particular, in the sense of the least squares. The concept of weak, or generalized, solutions, as well as the methods for their computation, are discussed in Chapter 7.

In the meantime, let us mention that if the Arnoldi orthogonalization process (6.71) does not terminate on or before $k = p$, then the minimization problem (6.76) has a unique solution. As shown in Section 7.2, this is an implication of the matrix $\boldsymbol{H}_p$ being full rank. The latter assertion, in turn, is true because according to formula (6.72), equation number k of (6.71) can be recast as follows:

$$\boldsymbol{A}\boldsymbol{u}_k = \boldsymbol{v}_k + \sum_{j=1}^{k} \mathfrak{h}_{jk}\boldsymbol{u}_j = \boldsymbol{u}_{k+1}\|\boldsymbol{v}_k\| + \sum_{j=1}^{k} \mathfrak{h}_{jk}\boldsymbol{u}_j = \sum_{j=1}^{k+1} \mathfrak{h}_{jk}\boldsymbol{u}_j,$$

which means that $\mathfrak{h}_{k+1,k} = \|\boldsymbol{v}_k\| \neq 0$. As such, all columns of the matrix $\boldsymbol{H}_p$ are linearly independent since every column has an additional non-zero entry $\mathfrak{h}_{k+1,k}$ compared to the previous column. Consequently, the vector $\boldsymbol{w}^{(p)}$ can be obtained as a solution to the linear system (see Theorem 7.1 on page 215):

$$\boldsymbol{H}_p^T\boldsymbol{H}_p\boldsymbol{w}^{(p)} = -\boldsymbol{H}_p^T\boldsymbol{q}^{(p+1)}. \tag{6.77}$$

The solution $\boldsymbol{w}^{(p)}$ of system (6.77) is unique because the matrix $\boldsymbol{H}_p^T\boldsymbol{H}_p$ is non-singular (Exercise 2). In practice, one does not normally reduce the least squares minimization problem (6.76) to linear system (6.77) since this reduction may lead to the introduction of large additional errors (amplification of round-off). Instead, problem (6.76) is solved using the $\boldsymbol{QR}$ factorization of the matrix $\boldsymbol{H}_p$, see Section 7.2.2.

Note that in the course of the previous analysis we assumed that the dimension of the Krylov subspaces $K_p(\boldsymbol{A},\boldsymbol{r}^{(0)})$ would increase monotonically as a function of p. Let us now see what happens if the alternative situation takes place, i.e., if the Arnoldi process terminates prematurely.

THEOREM 6.5

Let p be the smallest integer number for which the Arnoldi process (6.71) terminates:

$$\boldsymbol{A}\boldsymbol{u}_p - \sum_{j=1}^{p} (\boldsymbol{u}_j, \boldsymbol{A}\boldsymbol{u}_p)\boldsymbol{u}_j = \boldsymbol{0}.$$

Then the corresponding iterate yields the exact solution:

$$\boldsymbol{x}^{(p)} = \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{f}.$$

PROOF By hypothesis of the theorem, $\boldsymbol{A}\boldsymbol{u}_p \in K_p$. Consequently, $\boldsymbol{A}K_p \subset K_p$. This implies [cf. formula (6.72)]:

$$\boldsymbol{A}\boldsymbol{U}_p = \boldsymbol{U}_p\tilde{\boldsymbol{H}}, \tag{6.78}$$

where $\tilde{\boldsymbol{H}}$ is a $p \times p$ matrix. This matrix is non-singular because otherwise it would have had linearly dependent rows. Then, according to formula (6.78), each column of $\boldsymbol{AU}_p$ could be represented as a linear combination of only a subset of the columns from $\boldsymbol{U}_p$ rather than as a linear combination of all of its columns. This, in turn, means that the Arnoldi process terminates earlier than $k = p$, which contradicts the hypothesis of the theorem.

For the norm of the residual $\boldsymbol{r}^{(p)} = \boldsymbol{Ax}^{(p)} - \boldsymbol{f}$ we can write:

$$\|\boldsymbol{r}^{(p)}\| = \|\boldsymbol{Ax}^{(p)} - \boldsymbol{f}\| = \|\boldsymbol{A}(\boldsymbol{x}^{(p)} - \boldsymbol{x}^{(0)}) + \boldsymbol{r}^{(0)}\|. \tag{6.79}$$

Next, we notice that since $\boldsymbol{x}^{(p)} \in N_p$, then $\boldsymbol{x}^{(p)} - \boldsymbol{x}^{(0)} \in K_p(\boldsymbol{A}, \boldsymbol{r}^{(0)})$, and consequently, $\exists \boldsymbol{w} \in \mathbb{R}^p : \ \boldsymbol{x}^{(p)} - \boldsymbol{x}^{(0)} = \boldsymbol{U}_p \boldsymbol{w}$, because the columns of the matrix $\boldsymbol{U}_p$ provide a basis in the space $K_p(\boldsymbol{A}, \boldsymbol{r}^{(0)})$. Let us also introduce a p-dimensional vector $\boldsymbol{q}^{(p)}$ (similar to the $p+1$-dimensional vector $\boldsymbol{q}^{(p+1)}$ given by (6.75)):

$$\boldsymbol{q}^{(p)} = \big[\|\boldsymbol{r}^{(0)}\|, \underbrace{0, 0, \ldots, 0}_{p-1}\big]^T,$$

so that $\boldsymbol{U}_p \boldsymbol{q}^{(p)} = \boldsymbol{r}^{(0)}$. Then, taking into account equality (6.78), as well as the orthonormality of the columns of the matrix $\boldsymbol{U}_p$: $\boldsymbol{U}_p^T \boldsymbol{U}_p = \boldsymbol{I}_p$, we obtain from formula (6.79):

$$\|\boldsymbol{r}^{(p)}\| = \|\boldsymbol{U}_p(\boldsymbol{q}^{(p)} + \tilde{\boldsymbol{H}}\boldsymbol{w})\| = \|\boldsymbol{q}^{(p)} + \tilde{\boldsymbol{H}}\boldsymbol{w}\|. \tag{6.80}$$

Finally, we recall that on every iteration of GMRES we minimize the norm of the residual: $\|\boldsymbol{r}^{(p)}\| \longrightarrow \min$. Then, we can simply set $\boldsymbol{w} = -\tilde{\boldsymbol{H}}^{-1}\boldsymbol{q}^{(p)}$ in formula (6.80), which immediately yields $\|\boldsymbol{r}^{(p)}\| = 0$. This is obviously a minimum of the norm, and it implies $\boldsymbol{r}^{(p)} = \boldsymbol{0}$, i.e., $\boldsymbol{Ax}^{(p)} = \boldsymbol{f} \implies \boldsymbol{x}^{(p)} = \boldsymbol{A}^{-1}\boldsymbol{f} = \boldsymbol{x}$. □

We can now summarize two possible scenarios of behavior of the GMRES iteration. If the Arnoldi process terminates prematurely at some $p < n$ (n is the dimension of the space), then, according to Theorem 6.5, $\boldsymbol{x}^{(p)}$ is the exact solution to $\boldsymbol{Ax} = \boldsymbol{f}$. Otherwise, the maximum number of iterations that the GMRES can perform is equal to n. Indeed, if the Arnoldi process does not terminate prematurely, then $\boldsymbol{U}_n$ will contain n linearly independent vectors of dimension n and consequently, $K_n(\boldsymbol{A}, \boldsymbol{r}^{(0)}) = \mathbb{R}^n$. As such, the last minimization of the residual in the sense of (6.73) will be performed over the entire space $\mathbb{R}^n$, which obviously yields the exact solution $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{f}$. Therefore, technically speaking, the GMRES can be regarded as a direct method for solving $\boldsymbol{Ax} = \boldsymbol{f}$, in much the same way as we regarded the method of conjugate gradients as a direct method (see Section 5.6).

In practice, however, the GMRES is never used in the capacity of a direct solver, it its only used as an iterative scheme. The reason is that for high dimensions n it is feasible to perform only very few iterations, and one should hope that the approximate solution obtained after these iterations will be sufficiently accurate in a given context. The limitations for the number of iterations come primarily from the large storage

requirements for the Krylov subspace basis $\boldsymbol{U}_p$, as well as from the increasing computational costs associated with solving the sequence of the least squares problems (6.76) for $p = 1, 2, \ldots$. Note that the method of conjugate gradients does not entail this type of limitations because its descent directions are automatically $\boldsymbol{A}$-orthogonal. These additional constraints that characterize the GMRES are the "price to pay" for its broader applicability and ability to handle general matrices $\boldsymbol{A}$, as opposed to only symmetric positive definite matrices, for which the method of conjugate gradients works. However, another inherent limitation of the GMRES fully translates to the method of conjugate gradients (or the other way around). Indeed, the exact solution of $\boldsymbol{Ax} = \boldsymbol{f}$ can be obtained by means of the GMRES only if the computations are conducted with infinite precision. On a finite precision computer the method is prone to numerical instabilities. No universal cure is available for this problem; some partial remedies, such as restarts, are discussed, e.g., in [Saa03].

Exercises

1. Prove that the Arnoldi process (6.71) indeed yields an orthonormal system of vectors: $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$.
2. Prove that the system matrix $\boldsymbol{H}_p^T \boldsymbol{H}_p$ in (6.77) is symmetric positive definite.

6.4 Multigrid Iterations

We have seen previously that in many cases numerical methods with superior performance can be developed at the expense of narrowing down the class of problems that they are designed to solve. In the framework of direct methods, examples include the tri-diagonal elimination (Section 5.4.2), as well as the methods that exploit the finite Fourier series and the FFT (Section 5.7). In the framework of iterative methods, a remarkable example of that kind is given by multigrid.

Multigrid methods have been originally developed for solving elliptic boundary value problems discretized by finite differences (Chapter 12). A key distinctive characteristic of these methods is that the number of iterations required for reducing the initial error by a prescribed factor does not depend on the dimension of the grid at all. Accordingly, the required number of arithmetic operations is directly proportional to the grid dimension N^n, where N is the number of grid nodes along one coordinate direction and n is the dimension of the space $\mathbb{R}^n$. This is clearly an asymptotically unimprovable behavior, because the overall number of quantities to be computed (solution values on the grid) is also directly proportional to the grid dimension. As the grid dimension determines the condition number of the corresponding matrix,[2]

[2]The latter is typically inversely proportional to the square of the grid size: $\mu = \mathcal{O}(h^{-2})$, see formula (5.115), i.e., $\mu = \mathcal{O}(N^2)$.

we conclude that the number of multigrid iterations needed for achieving a given accuracy does not depend on the condition number μ. In contradistinction to that, for the best iterative methods we have analyzed before, the Chebyshev method (Section 6.2.1) and the method of conjugate gradients (Section 6.2.2), the number of iterations is proportional to the square root of the condition number $\sqrt{\mu}$, see formula (6.65). Accordingly, the required number of arithmetic operations is $\mathcal{O}(N^{n+1})$.

Multigrid iterations will apply to basically the same range of elliptic finite-difference problems to which the Richardson iterations apply. An additional constraint is that of the "smoothness," or "regularity," of the first eigenfunctions of the corresponding operator (matrix). For elliptic problems, it normally holds.

A rigorous analysis of multigrid is quite involved. Therefore, we will restrict ourselves to a qualitative description of its key idea (Section 6.4.1) and of the simplest version of the actual numerical algorithm (Section 6.4.2). Further detail, general constructions, and proofs can be found in the literature quoted in Section 6.4.3.

6.4.1 Idea of the Method

Introduce a uniform Cartesian grid on the square $D = \{(x,y)|0 \le x \le 1, 0 \le y \le 1\}$:

$$(x_{m_1}, y_{m_2}) = (m_1 h, m_2 h), \quad m_1, m_2 = 0, 1, \ldots, M, \quad h = M^{-1},$$

define the grid boundary Γ_h as the set of nodes that belong to $\Gamma = \partial D$:

$$\Gamma_h = \{(x_{m_1}, y_{m_2}) | m_1 = 0, M \text{ or } m_2 = 0, M\},$$

and consider the same homogeneous finite-difference Dirichlet problem for the Poisson equation as we analyzed in Section 5.1.3:

$$\begin{aligned} -\Delta_h u_{m_1,m_2} \equiv -\Big(&\frac{u_{m_1+1,m_2} - 2u_{m_1,m_2} + u_{m_1-1,m_2}}{h^2} \\ &+ \frac{u_{m_1,m_2+1} - 2u_{m_1,m_2} + u_{m_1,m_2-1}}{h^2}\Big) = f_{m_1,m_2}, \\ &m_1, m_2 = 1, 2, \ldots, M-1, \\ u\big|_{\Gamma_h} &= 0. \end{aligned} \tag{6.81}$$

For solving problem (6.81), we will use the standard stationary Richardson iteration (Section 6.1) as our starting point:

$$\begin{aligned} u^{(p+1)}_{m_1,m_2} &= u^{(p)}_{m_1,m_2} + \tau \Delta_h u^{(p)}_{m_1,m_2} + \tau f_{m_1,m_2}, \\ m_1, m_2 &= 1, 2, \ldots, M-1, \quad p = 0, 1, 2, \ldots \\ u^{(p+1)}\big|_{\Gamma_h} &= 0, \quad u^{(0)}_{m_1,m_2} \text{ is given.} \end{aligned} \tag{6.82}$$

Iterations (6.82) are generally known to converge slowly. This slowness, however, is not uniform across the spectrum of the problem. To see that, let us introduce the

error $\varepsilon^{(p)} = u - u^{(p)}$ of the iterate $u^{(p)}$ and represent it in the form of a finite Fourier series according to the methodology of Section 5.7:

$$\varepsilon^{(p)} = \sum_{r,s=1}^{M-1} [1 - \tau\lambda_{rs}]^p c_{rs}^{(0)} \psi^{(r,s)}. \tag{6.83}$$

In formula (6.83), $\psi^{(r,s)}$ are eigenfunctions of the discrete Laplacian $-\Delta_h$ given by (5.105):

$$\psi^{(r,s)} = \left\{ 2 \sin \frac{r\pi m_1}{M} \sin \frac{s\pi m_2}{M} \right\}, \quad r,s = 1,2,\ldots,M-1,$$

and λ_{rs} are the corresponding eigenvalues given by (5.109):

$$\lambda_{rs} = \frac{4}{h^2} \left(\sin^2 \frac{r\pi}{2M} + \sin^2 \frac{s\pi}{2M} \right), \quad r,s = 1,2,\ldots,M-1.$$

The amplification factors $\nu_{rs}(\tau) \stackrel{\text{def}}{=} [1 - \tau\lambda_{rs}]$ in formula (6.83) belong to the interval: $\nu_{\min} \le \nu_{rs} \le \nu_{\max}$, where:

$$\nu_{\min} = 1 - \tau\lambda_{M-1,M-1} \approx 1 - 8\tau M^2 \quad \text{and} \quad \nu_{\max} = 1 - \tau\lambda_{1,1} \approx 1 - 2\tau\pi^2.$$

Let us specify the iteration parameter τ as follows:

$$\tau = \frac{1}{5M^2}. \tag{6.84}$$

This choice of τ guarantees that if at least one of the numbers r or s is greater than $M/2$, then

$$|\nu_{rs}| < \frac{3}{5}.$$

Therefore, the contribution of the high frequency harmonics $\psi^{(r,s)}$ (with $r \ge M/2$ or $s \ge M/2$) to the error (6.83) reduces by almost a factor of two on every iteration. As such, this contribution soon becomes small and after several iterations (6.82) the error $\varepsilon^{(p)}$ will be composed primarily of the smooth components on a given grid, i.e., of the low frequencies $\psi^{(r,s)}$ that correspond to $r < M/2$ and $s < M/2$. Indeed, the amplification factors $\nu_{rs}(\tau) = 1 - \tau\lambda_{rs}$ for low frequency harmonics $\psi^{(r,s)}$ are closer to 1. The slowest decaying harmonic is $\psi^{(1,1)}$, because for the parameter τ chosen by formula (6.84) the resulting amplification factor is

$$\nu_{1,1} = 1 - \tau\lambda_{1,1} \approx 1 - \frac{2\pi^2}{5M^2}, \tag{6.85}$$

which is clearly very close to 1 for large M. Hence we see that the high frequency error content on a given grid decays fast, whereas the low frequencies decay slowly. It will therefore be natural to consider a given problem on a sequence of grids with different fineness. In doing so, the key idea is to have a special grid for every part of the spectrum, such that the corresponding harmonics on this grid can be regarded as

high frequencies. These high frequencies (i.e., short waves on the scale of the grid size) will decay fast in the course of the Richardson iteration (6.82).

Let $u^{(p)}$ be the approximate solution obtained by the iteration process (6.82). For simplicity, we will use an alternative notation $u^{(p)} = \mathfrak{U}$ hereafter. Let us also re-denote the error of the iterate $u^{(p)}$: $\varepsilon^{(p)} = u - u^{(p)} = u - \mathfrak{U} = \phi$. If we knew the error ϕ, then we would have immediately found the solution: $u = \mathfrak{U} + \phi$. We, however, do not know ϕ per se, we only know that it solves the boundary value problem:

$$-\Delta_h \phi = -\eta, \quad \phi\big|_{\Gamma_h} = 0, \tag{6.86}$$

where η is the residual of the iterate $u^{(p)}$ that it generates if substituted into (6.81):

$$\eta = -\Delta_h u^{(p)} - f \equiv -\Delta_h \mathfrak{U} - f.$$

Problem (6.86) with the correction ϕ as the unknown is only simpler than the original problem (6.81) in the sense that ϕ is known to be a smooth grid function ahead of time (with little or no high frequency content). Therefore, to approximately compute ϕ we can consider the same problem as (6.86) but on a twice as coarse grid. If M is even, then this new grid with size $2h$ instead of h will have $M/2+1$ nodes in each direction and will merely be a sub-grid of the original grid with every other node in every coordinate direction dropped out. The new problem can be written as

$$-\Delta_{2h} \tilde{\phi} = -\tilde{\eta}, \quad \tilde{\phi}\big|_{\Gamma_{2h}} = 0, \tag{6.87}$$

where the tildes denote the quantities on the coarser grid. The grid boundary is now given by:

$$\Gamma_{2h} = \{(2hm_1, 2hm_2) \,|\, m_1 = 0, M/2 \ \text{ or } \ m_2 = 0, M/2\}.$$

Note that the transition from the fine grid to a coarser grid is often called restriction. Problem (6.87) is to be solved by the Richardson iteration similar to (6.82):

$$\begin{gathered} \tilde{\phi}^{(p+1)}_{m_1,m_2} = \tilde{\phi}^{(p)}_{m_1,m_2} + \tilde{\tau}\Delta_{2h}\tilde{\phi}^{(p)}_{m_1,m_2} - \tilde{\tau}\tilde{\eta}_{m_1,m_2}, \\ m_1, m_2 = 1, 2, \ldots, \tilde{M} - 1, \quad p = 0, 1, 2, \ldots \\ \tilde{\phi}^{(p+1)}\big|_{\Gamma_{2h}} = 0, \quad \tilde{\phi}^{(0)}_{m_1,m_2} = 0, \end{gathered} \tag{6.88}$$

where $\tilde{M} = M/2$ and $\tilde{\tau} = 4\tau$ [see formula (6.84)].

Each iteration (6.88) is four times less expensive than one iteration (6.82), because there are four times fewer computational nodes. Moreover, as $\tilde{\tau} = 4\tau$ the slowest decaying component of the error is still decreasing faster on the coarser grid than on the original grid. Indeed, according to (6.85) we have:

$$\tilde{\nu}_{1,1} = 1 - \tilde{\tau}\tilde{\lambda}_{1,1} \approx 1 - \frac{2\pi^2}{5\tilde{M}^2} = 1 - 4\frac{2\pi^2}{5M^2} < \nu_{1,1}. \tag{6.89}$$

Let us require that for a given $\sigma \in (0,1)$:

$$(\tilde{\nu}_{1,1})^p \approx \left(1 - \frac{2\pi^2}{5\tilde{M}^2}\right)^p \leq \sigma.$$

Then, assuming that the subtrahend in formula (6.89) is still small, $8\pi^2/3M^2 \ll 1$, i.e., that M is large, we can use the Taylor formula for $\ln(\,\cdot\,)$ and write:

$$p \ln\left(1 - \frac{2\pi^2}{5\tilde{M}^2}\right) \le \ln\sigma \quad \Longrightarrow \quad -p\frac{2\pi^2}{5\tilde{M}^2} \le \ln\sigma$$
$$\Longrightarrow \quad p \ge -\frac{5\tilde{M}^2}{2\pi^2}\ln\sigma = -\frac{1}{4}\frac{5M^2}{2\pi^2}\ln\sigma.$$

As such, for reducing the contribution of $\tilde{\psi}^{(1,1)}$ into the error by a prescribed factor σ we will need approximately four times fewer iterations on the coarse grid with size $2h$ than for reducing the contribution of $\psi^{(1,1)}$ on the fine grid with size h.

Let us denote by $\tilde{\Phi}$ the grid function obtained as a result of the iteration process (6.88). This function is defined on the coarse grid with size $2h$. We will interpolate it (linearly) from this coarse grid onto the original fine grid with size h and obtain the function Φ. Note that the transition from the coarse grid to a finer grid is referred to as prolongation in the multigrid framework. In doing so, the smooth components will be obtained almost correctly on the fine grid. The corresponding relative error of interpolation will be small for a smooth interpolated function. However, the Fourier expansion of the interpolation error will contain all harmonics, because the interpolation error itself has kinks at the interpolation nodes and cannot be regarded as a smooth function. Moreover, as the grid function $\tilde{\Phi}$ is obtained by iteration (6.88), it has a non-smooth component of its own. The latter has basically nothing to do with the correction ϕ that we are looking for. It will, however, yield an additional (random) contribution to the non-smooth part of the resulting interpolant Φ. Altogether, we conclude that the smooth component of the sum $\mathfrak{U}+\Phi$ on the fine grid will be close to the smooth component of the unknown solution $u = \mathfrak{U} + \phi$, whereas the non-smooth component may not necessarily be very small and will basically have a random nature. Therefore, after the prolongation it is necessary to perform a few more fine grid iterations (6.82) while choosing $\mathfrak{U}+\Phi$ as their initial guess. This will facilitate a rapid suppression of the non-smooth component of the error introduced by interpolation, because every iteration (6.82) damps the high frequency part of the spectrum by almost a factor of two.

6.4.2 Description of the Algorithm

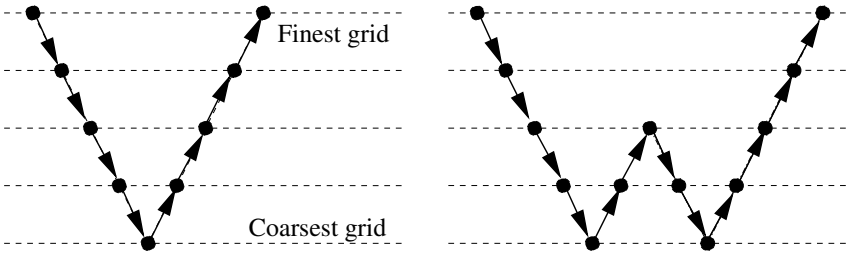
The speedup of convergence achieved using a coarser grid with size $2h$ and the iteration process (6.88) may still be insufficient. If M is large, the complexity of the coarse grid problem (6.87) will nonetheless remain fairly high. Therefore, when solving this problem it may be advisable to coarsen the grid one more time and obtain yet another problem similar to (6.87), but on the grid of size $4h$. For simplicity, let us assume that the initial grid dimension is given by a power of two, $M = 2^k$. Then, a number of coarsening steps can be performed, and a sequence of embedded grids and respective problems of type (6.87) can be introduced and exploited.

On the initial fine grid, we first make several iterations (6.82) to smooth out the error, i.e., reduce its high frequency content. As the error itself is not known, we

can monitor the residual $-\Delta_h u^{(p)} - f$ instead, because it also becomes smoother in the course of iteration (6.82). The result of these iterations $u^{(p)} = \mathfrak{U}$ is to be stored in the computer memory. Then we consider a coarser grid problem (6.87) for the correction ϕ, make several iterations (6.88) in order to smooth out the correction to the correction, and again store the result $\tilde{\Phi}$ in the computer memory (it requires four times less space than $\mathfrak{U}$). To actually compute the correction to $\tilde{\Phi}$, we consider yet another coarser grid problem, this time with size $4h$, perform several iterations with step $\tilde{\tau}^{(2)} = 4\tilde{\tau} = 16\tau$ and store the result $\tilde{\Phi}^{(2)}$. This nested process of computing corrections to corrections on twice as coarse embedded grids is run k times until the coarsest grid is reached and the corresponding correction $\tilde{\Phi}^{(k)}$ is obtained.

Then, we start the process of returning to the fine grid. First we interpolate the coarsest grid correction $\tilde{\Phi}^{(k)}$ to the second to last grid, which is twice as fine. On this grid, we add the correction $\Phi^{(k-1)}$ to the previously stored solution and also make several iterations to damp the interpolation error. The result of these iterations is interpolated to the next finer grid and then used on this grid for correcting the stored function $\tilde{\Phi}^{(k-2)}$. Subsequently, several iterations are conducted and yet another interpolation is made. On the second to last step, once the correction $\Phi^{(2)}$ is introduced and iterations performed on the grid $2h$, we obtain the last correction $\tilde{\Phi}$, interpolate it to the finest grid h, make several iterations (6.82) starting with the initial guess $\mathfrak{U} + \Phi$, and obtain the final result.

In the modern theory of multigrid methods, the algorithm we have just described is referred to as a V-cycle; it is schematically shown in Figure 6.3(a). In practice, several consecutive V-cycles may be required for obtaining a sufficiently accurate approximation to the solution u of problem (6.81). Alternatively, one can use the so-called W-cycles that are shown schematically in Figure 6.3(b). Each individual W-cycle may be more expensive computationally than the corresponding V-cycle. However, fewer W-cycles are normally required for reducing the initial error by a prescribed factor.



(a) V-cycle. (b) W-cycle.

FIGURE 6.3: Multigrid cycles.

6.4.3 Bibliography Comments

The concept of what has later become known as multigrid methods was first introduced in a 1961 paper by Fedorenko [Fed61]. The author called his iteration scheme a relaxation method. In a subsequent 1964 paper [Fed64], Fedorenko has also provided first estimates of the convergence rate of his relaxation method when it was applied to solving a finite-difference Dirichlet problem. Similar estimates for other problems were later obtained by Bakhvalov [Bak66] and by Astrakhantsev [Ast71]. A detailed summary of this early development of multigrid methods can be found in the review paper by Fedorenko [Fed73].

Subsequent years, starting from the mid-seventies, witnessed a rapid growth of attention to multigrid methods, and an "explosion" of work on their theoretical analysis, algorithmic implementation, and applications to a wide variety of problems far beyond simple elliptic discretizations. These methods have proven extremely successful and superior to other techniques even when their actual performance for difficult problems was not as good as predicted theoretically, say, for the Poisson equation. For example, multigrid methods have enabled a historical breakthrough in the performance of numerical solvers used in computational fluid dynamics for the quantitative analysis of aerodynamic configurations. This dramatic progress in the development of multigrid is associated with the names of many researchers; fundamental contributions were made by Brandt, Hackbusch, Jameson, and others. Advances in the area of multigrid methods are summarized in a number of papers and books, see, e.g., [Bra77], [Bra84, Hac85, Wes92, Bra93, BHM00, TOS01].

A separate research direction in this area is the so-called algebraic multigrid methods, when similar multilevel ideas are applied directly to a given matrix, without any regard to where this matrix originates from. This approach, in particular, led to the development of the recursive ordering preconditioners.

Exercises

1. Redefine the notion of the high frequencies on the grid as those harmonics $\psi^{(r,s)}$ for which both $r > M/2$ and $s > M/2$. What value of the iteration parameter τ shall one choose instead of (6.84) so that to guarantee the best possible damping of the high frequencies by the iteration scheme (6.82)? What is the corresponding maximum value of the amplification factor: $\max_{r>\frac{M}{2}, s>\frac{M}{2}} |\nu_{rs}|$?

   Hint. Use the condition $\nu_{\min}(\tau) \equiv \nu_{M-1,M-1}(\tau) = -\nu_{\frac{M}{2},\frac{M}{2}}(\tau)$. Explain why this choice of τ will guarantee the best damping.