row. In particular, when the matrix A has a sparse storage organization (see Sections 5.6 and 5.8), permutation by rows is performed only when a null (or exceedingly small) pivot element is encountered.

See Exercises 5.6-5.8.

## 5.5 How accurate is the solution of a linear system?

We have already noticed in Example 5.8 that, due to roundoff errors, the product LU does not reproduce A exactly. Even though the pivoting strategy damps these errors, yet the result could sometimes be rather unsatisfactory.

**Example 5.9** Consider the linear system $\mathrm{A}_n\mathbf{x}_n = \mathbf{b}_n$, where $\mathrm{A}_n \in \mathbb{R}^{n\times n}$ is the so-called *Hilbert matrix* whose elements are

$$a_{ij} = 1/(i+j-1), \qquad i,j = 1,\dots,n,$$

while $\mathbf{b}_n$ is chosen in such a way that the exact solution is $\mathbf{x}_n = (1,1,\dots,1)^T$. The matrix $\mathrm{A}_n$ is clearly symmetric and one can prove that it is also positive definite. For different values of $n$ we use the MATLAB function `lu` to get the Gauss factorization of $\mathrm{A}_n$ with pivoting by row. Then we solve the associated linear systems (5.20) and denote by $\widehat{\mathbf{x}}_n$ the computed solution. In Figure 5.8 we report (in logarithmic scale) the relative errors

$$E_n = \|\mathbf{x}_n - \widehat{\mathbf{x}}_n\|/\|\mathbf{x}_n\|, \tag{5.22}$$

having denoted by $\|\cdot\|$ the Euclidean norm introduced in the Section 1.4.1. We have $E_n \geq 10$ if $n \geq 13$ (that is a relative error on the solution higher than 1000%!), whereas $\mathrm{R}_n = \mathrm{L}_n\mathrm{U}_n - \mathrm{P}_n\mathrm{A}_n$ is the null matrix (up to machine accuracy) for any given value of $n$. ■

On the ground of the previous remark, we could speculate by saying that, when a linear system $\mathrm{A}\mathbf{x} = \mathbf{b}$ is solved numerically, one is indeed looking for the *exact* solution $\widehat{\mathbf{x}}$ of a *perturbed* system

$$(\mathrm{A} + \delta\mathrm{A})\widehat{\mathbf{x}} = \mathbf{b} + \boldsymbol{\delta}\mathbf{b}, \tag{5.23}$$

where $\delta\mathrm{A}$ and $\boldsymbol{\delta}\mathbf{b}$ are respectively a matrix and a vector which depend on the specific numerical method which is being used. We start by considering the case where $\delta\mathrm{A} = 0$ and $\boldsymbol{\delta}\mathbf{b} \neq \mathbf{0}$ which is simpler than the most general case. Moreover, for simplicity we will also assume that A$\in \mathbb{R}^{n\times n}$ is symmetric and positive definite.

By comparing (5.1) and (5.23) we find $\mathbf{x} - \widehat{\mathbf{x}} = -\mathrm{A}^{-1}\boldsymbol{\delta}\mathbf{b}$, and thus

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| = \|\mathrm{A}^{-1}\boldsymbol{\delta}\mathbf{b}\|. \tag{5.24}$$

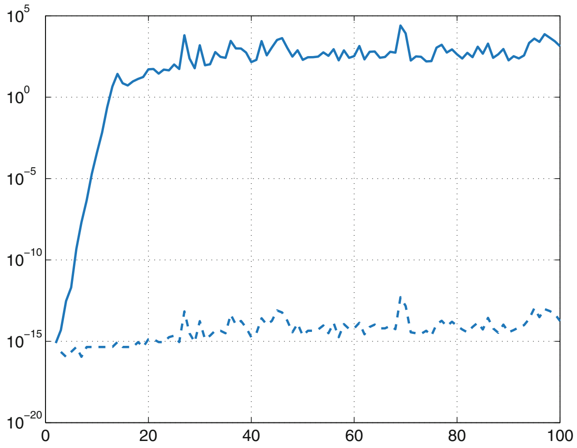

**Fig. 5.8.** Behavior versus $n$ of $E_n$ (*solid line*) and of $\max_{i,j=1,\ldots,n} |r_{ij}|$ (*dashed line*) in logarithmic scale, for the Hilbert system of Example 5.9. The $r_{ij}$ are the coefficients of the matrix $\mathrm{R}_n$

In order to find an upper bound for the right-hand side of (5.24), we proceed as follows. Since A is symmetric and positive definite, the set of its eigenvectors $\{\mathbf{v}_i\}_{i=1}^n$ provides an orthonormal basis of $\mathbb{R}^n$ (see [QSS07, Chapter 5]). This means that

$$\mathrm{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ i = 1,\ldots,n, \qquad \mathbf{v}_i^T \mathbf{v}_j = \delta_{ij},\ i,j = 1,\ldots,n,$$

where $\lambda_i$ is the eigenvalue of A associated with $\mathbf{v}_i$ and $\delta_{ij}$ is the Kronecker symbol. Consequently, a generic vector $\mathbf{w} \in \mathbb{R}^n$ can be written as

$$\mathbf{w} = \sum_{i=1}^n w_i \mathbf{v}_i,$$

for a suitable (and unique) set of coefficients $w_i \in \mathbb{R}$. We have

$$\begin{aligned}\|\mathrm{A}\mathbf{w}\|^2 &= (\mathrm{A}\mathbf{w})^T(\mathrm{A}\mathbf{w})\\ &= [w_1(\mathrm{A}\mathbf{v}_1)^T + \ldots + w_n(\mathrm{A}\mathbf{v}_n)^T][w_1\mathrm{A}\mathbf{v}_1 + \ldots + w_n\mathrm{A}\mathbf{v}_n]\\ &= (\lambda_1 w_1 \mathbf{v}_1^T + \ldots + \lambda_n w_n \mathbf{v}_n^T)(\lambda_1 w_1 \mathbf{v}_1 + \ldots + \lambda_n w_n \mathbf{v}_n)\\ &= \sum_{i=1}^n \lambda_i^2 w_i^2.\end{aligned}$$

Denote by $\lambda_{max}$ the largest eigenvalue of A. Since $\|\mathbf{w}\|^2 = \sum_{i=1}^n w_i^2$, we conclude that

$$\|\mathrm{A}\mathbf{w}\| \le \lambda_{max} \|\mathbf{w}\| \quad \forall \mathbf{w} \in \mathbb{R}^n. \tag{5.25}$$

In a similar manner, we obtain

$$\|\mathrm{A}^{-1}\mathbf{w}\| \le \frac{1}{\lambda_{min}} \|\mathbf{w}\|,$$

upon recalling that the eigenvalues of $A^{-1}$ are the reciprocals of those of A. This inequality enables us to draw from (5.24) that

$$\frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{1}{\lambda_{min}} \frac{\|\boldsymbol{\delta}\mathbf{b}\|}{\|\mathbf{x}\|}. \tag{5.26}$$

Using (5.25) once more and recalling that $A\mathbf{x} = \mathbf{b}$, we finally obtain

$$\boxed{\frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\lambda_{max}}{\lambda_{min}} \frac{\|\boldsymbol{\delta}\mathbf{b}\|}{\|\mathbf{b}\|}} \tag{5.27}$$

We can conclude that the relative error in the solution depends on the relative error in the data through the following constant ($\geq 1$)

$$\boxed{K(A) = \frac{\lambda_{max}}{\lambda_{min}}} \tag{5.28}$$

which is called *spectral condition number of the matrix* A. $K(A)$ can be computed in MATLAB using the command `cond`. `cond`

**Remark 5.3** The MATLAB command `cond(A)` allows the computation of the condition number of any type of matrix `A`, even those which are not symmetric and positive definite. It is worth mentioning that there exist various definitions of condition number of a matrix. For a generic matrix A, the command `cond(A)` computes the value $K_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2$, where we define $\|A\|_2 = \sqrt{\lambda_{max}(A^T A)}$. We note that if A is not symmetric and positive definite, $K_2(A)$ can be very far from the spectral condition number $K(A)$. For a sparse matrix A, the command `condest(A)` computes an approximation (at low computational cost) of the condition number $K_1(A) = \|A\|_1 \cdot \|A^{-1}\|_1$, being $\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$ the so-called *1-norm* of A. Other definitions for the condition number are available for nonsymmetric matrices, see [QSS07, Chapter 3]. ■ `condest`

A more involved proof would lead to the following more general result in the case where A is symmetric and positive definite and $\delta A$ is an arbitrary symmetric and positive definite matrix, "small enough" to satisfy $\lambda_{max}(\delta A) < \lambda_{min}(A)$:

$$\boxed{\frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{K(A)}{1-\lambda_{max}(\delta A)/\lambda_{min}(A)} \left(\frac{\lambda_{max}(\delta A)}{\lambda_{max}(A)} + \frac{\|\boldsymbol{\delta}\mathbf{b}\|}{\|\mathbf{b}\|}\right)} \tag{5.29}$$

Finally, if A and $\delta A$ are not symmetric positive definite matrices, and $\delta A$ is such that $\|\delta A\|_2 \, \|A^{-1}\|_2 < 1$, the following estimate holds:

$$\boxed{\frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{K_2(A)}{1-K_2(A)\|\delta A\|_2/\|A\|_2} \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\boldsymbol{\delta}\mathbf{b}\|}{\|\mathbf{b}\|}\right)} \tag{5.30}$$

If $K(\mathrm{A})$ is “small”, that is of the order of unity, A is said to be *well conditioned*. In that case, small errors in the data will lead to errors of the same order of magnitude in the solution. This would not occur in the case of *ill conditioned* matrices.

**Example 5.10** For the Hilbert matrix introduced in Example 5.9, $K(\mathrm{A}_n)$ is a rapidly increasing function of $n$. One has $K(\mathrm{A}_4) > 15000$, while if $n > 13$ the condition number is so high that MATLAB warns that the matrix is “close to singular”. Actually, $K(\mathrm{A}_n)$ grows at an exponential rate, $K(\mathrm{A}_n) \simeq e^{3.5n}$ (see, [Hig02]). This provides an indirect explanation of the bad results obtained in Example 5.9. ■

Inequality (5.27) can be reformulated by the help of the *residual* $\mathbf{r}$

$$\mathbf{r} = \mathbf{b} - \mathrm{A}\widehat{\mathbf{x}}. \tag{5.31}$$

Should $\widehat{\mathbf{x}}$ be the exact solution, the residual would be the null vector. Thus, in general, $\mathbf{r}$ can be regarded as an *estimator* of the error $\mathbf{x} - \widehat{\mathbf{x}}$. The extent to which the residual is a good error estimator depends on the size of the condition number of A. Indeed, observing that $\boldsymbol{\delta}\mathbf{b} = \mathrm{A}(\widehat{\mathbf{x}} - \mathbf{x}) = \mathrm{A}\widehat{\mathbf{x}} - \mathbf{b} = -\mathbf{r}$, we deduce from (5.27) that

$$\boxed{\frac{\|\mathbf{x} - \widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq K(\mathrm{A}) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}} \tag{5.32}$$

Thus if $K(\mathrm{A})$ is “small”, we can be sure that the error is small provided that the residual is small, whereas this might not be true when $K(\mathrm{A})$ is “large”.

**Example 5.11** The residuals associated with the computed solution of the linear systems of Example 5.9 are very small (their norms vary between $10^{-16}$ and $10^{-11}$); however the computed solutions differ remarkably from the exact solution. ■

See Exercises 5.9-5.10.

## 5.6 How to solve a tridiagonal system

In many applications (see for instance Chapter 8), we have to solve a system whose matrix has the form

$$\mathrm{A} = \begin{bmatrix} a_1 & c_1 & & 0 \\ e_2 & a_2 & \ddots & \\ & \ddots & & c_{n-1} \\ 0 & & e_n & a_n \end{bmatrix}.$$