10 Error Bounds for Linear Systems of Equations

In many applications that give rise to linear systems of equations and least-squares problems, the right-hand side vector represents a measured quantity and therefore is contaminated by measurement error. It is the aim of the present lecture to investigate how this error affects the accuracy of the computed solution.

Example 10.1

Two kinds of cells on a Petri dish multiply (by division!) at different known rates. It is quite easy to see the cells with a microscope, but difficult to tell which kind they are. The total number of cells has been measured twice with an hour between measurements. We would like to determine how many cells there will be one hour later.

Let a_1 be the number of cells of the first kind available at the first measurement at time t_1. At time, $t > t_1$, the number of cells of this population then is $x_1(t) = a_1 \exp(\beta_1(t - t_1))$, where β_1 is known to be 2 and the unit for t is hours.

Similarly, let a_2 be the number of cells of the second kind available at the first measurement at time t_1. At time, $t > t_1$, the number of cells of this population is $x_2(t) = a_2 \exp(\beta_2(t - t_1)$ for $\beta_2 = 2.2$. We have measured the total number of cells at time t_1 to be 124 and at time $t_2 = t_1 + 1$ to be 1038. How many cells can we expect at time $t_2 + 1$?

One way to solve this problem is to set up a linear system of equations for a_1 and a_2. Having determined these coefficients, the total number of cells at time t is given by

$$x(t) = x_1(t) + x_2(t) = a_1 \exp(\beta_1(t - t_1)) + a_2 \exp(\beta_2(t - t_1). \tag{1}$$

Using the above equation at times t_1 and t_2 gives

$$\begin{aligned} a_1 + a_2 &= 124 \\ a_1 \exp(\beta_1) + a_2 \exp(\beta_2) &= 1038 \end{aligned}$$

with $\exp(\beta_1) = 7.39$ and $\exp(\beta_2) = 9.03$. We express this linear system of equations in the form

$$A\mathbf{a} = \mathbf{b}, \tag{2}$$

where

$$A = \begin{bmatrix} 1 & 1 \\ 7.39 & 9.03 \end{bmatrix}, \qquad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 124 \\ 1038 \end{bmatrix}.$$

The solution of this system is about $a_1 = 50$ and $a_2 = 74$, which yields

$$x(2) = 50 \exp(4) + 74 \exp(4.4) = 8757. \tag{3}$$

[0] Version November 12, 2013

We may wonder how many digits in the computed value for $x(2)$ are correct. There are many sources of errors. The coefficients β_1 and β_2 are likely not to be exact, the model (1) only is an approximation of reality, and the number of cells measured at times t_1 and t_2 is likely to be contaminated by error, because the cells multiply rapidly and their number may change during the measurement process. In this section, we will discuss the sensitivity of the computed solution $\mathbf{a}$ of (2) to errors in the measured data $\mathbf{b}$. □

10.1 The geometry of sensitivity

We illustrate the sensitivity of the solution of linear systems of equations to perturbations in the right-hand side with a couple of figures. Introduce the linear system of equations

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 &= b_1 \qquad (4) \\ a_{2,1}x_1 + a_{2,2}x_2 &= b_2 \qquad (5) \end{aligned}$$

Each equation represents a line in the (x_1, x_2)-plane. The solution of this system amounts to the familiar problem of finding the point at which two lines intersect. If these lines intersect at a unique point, then the linear system of equations has a unique solution. On the other hand, if the lines are parallel, then two cases have to be distinguished: i) if the lines are distinct, then the linear system of equations has no solution, and ii) if the lines coalesce, then the system has infinitely many solutions. The angle between the lines determines how accurately the intersection can be computed in the presence of errors in the data. We will discuss two particular choices of coefficients $a_{i,j}$ in the equations (4) and (5).

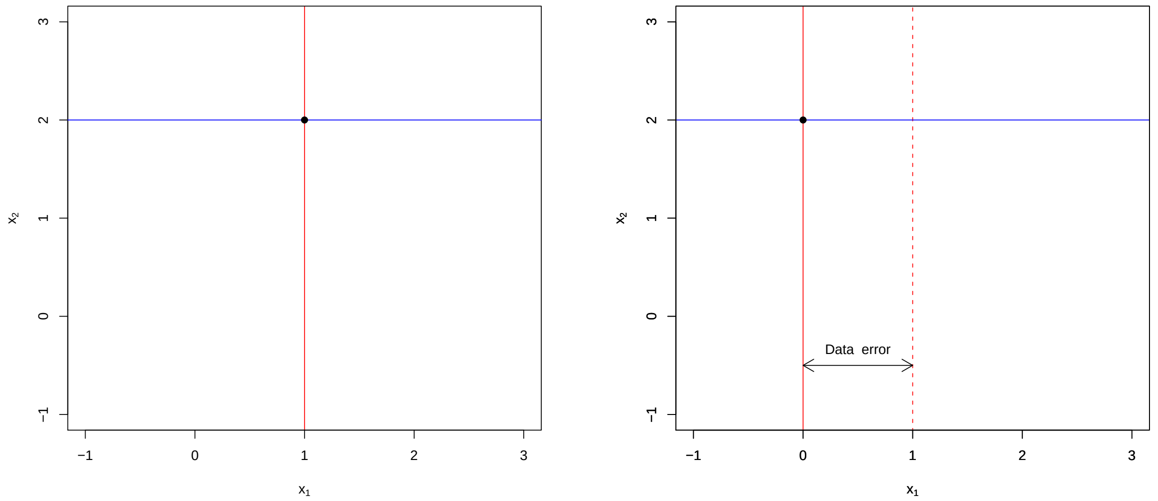
Example 10.2

Consider the linear system of equations

$$\begin{aligned} 1 \cdot x_1 + 0 \cdot x_2 &= 1 \\ 0 \cdot x_1 + 1 \cdot x_2 &= 2 \end{aligned}$$

The first equation represents the vertical line $\{(x_1, x_2) : x_1 = 1, x_2 \in \mathbb{R}\}$ and the second equation represents the horizontal line $\{(x_1, x_2) : x_1 \in \mathbb{R}, x_2 = 2\}$. These lines are perpendicular and intersect at the point $(1, 2)$; the lines are shown in Figure 1(a). This is an example of an orthogonal matrix.

Figure 1(b) depicts a plot in which the first right-hand side component is perturbed by -1 (labeled as "data error"). This perturbation may, for instance, stem from a measurement error, and we are interested in how it affects the computed solution. The equation with the perturbed right-hand side component represents the horizontal line $\{(x_1, x_2) : x_1 = 0, x_2 \in \mathbb{R}\}$. The second equation is not changed. The lines represented by these equations now intersect at the point $(0, 2)$. Figure 1(b) shows the new solution with a black disc. □



(a) (b)

Figure 1: (a) Error-free perpendicular lines; (b) Perpendicular lines with error in our knowledge of the position of the vertical line.

In the above example the error 1 in the data, as measured by the norm $\|\cdot\|$, gives rise to an error 1 in the computed solution. This depends on that the lines represented by the linear system of equations (4)-(5) are perpendicular. When the angle between the lines is acute, a larger error in the computed solution may result. This is illustrated in the following example.

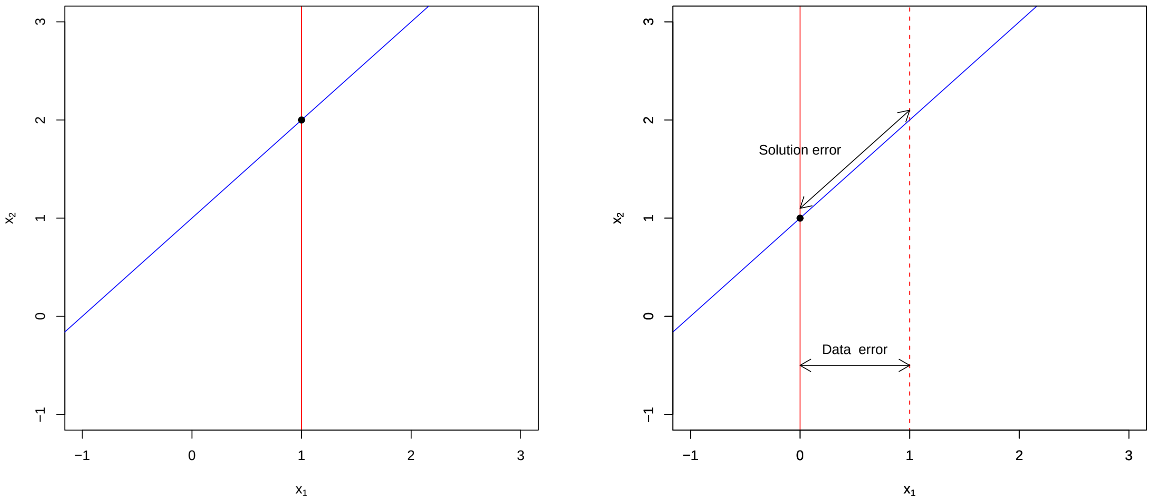
Example 10.3

Regard the linear system of equations

$$\begin{aligned} x_1 + 0 \cdot x_2 &= 1 \\ -x_1 + \quad x_2 &= 1 \end{aligned}$$

The lines corresponding to these equations are depicted in Figure 2(a). Similarly as in Example 10.2, Figure 2(b) shows the result of perturbing the first component of the data by -1. However, differently from Example 10.2, we see that the norm of the error in the computed solution is now larger than the error in the data. (Elementary trigonometry shows that the computed solution error is $\sqrt{2}$.) □

Examples 10.2 and 10.3 illustrate that errors in the data may be magnified in the computed solution when the lines are not perpendicular. As the angle between the lines grows more acute,



(a) (b)

Figure 2: (a) Error-free non-perpendicular lines; (b) Non-perpendicular lines with error in our knowledge of the position of the red line.

the error in the computed solution may be more amplified. Our two-dimensional intuition about this phenomenon carries over to higher dimensions.

The sensitivity to error in the data is not always revealed by standard numerical methods for the solution of linear systems of equations. It is important to keep in mind that a computed solution displayed with 16 decimal digits may not be very accurate.

One approach to gain insight into the sensitivity of the solution of a linear system of equations to perturbations of the right-hand side is to solve a large number of systems of equations with perturbed right-hand sides and investigate how much the solution changes. This kind of experimental sensitivity analysis can be carried out for small to medium-sized problems only, because the computational effort typically is quite large; see Exercises 10.1 and 10.2. Moreover, while experimental sensitivity analysis provides insight into the "typical sensitivity" of the solution under perturbations of the right-hand side, one cannot be certain that the sensitivity of the solution has been fully exposed without a careful design of the experiments. If the solution is very sensitive to a small class of perturbations, but not to others, then we might not notice this fact without carefully conducted experiments. This section discusses the sensitivity of the solution of a linear system of equations in terms of properties of the matrix.

10.2 Error propagation and the condition number

Let $\tilde{\mathbf{b}}$ be the available approximation of the unknown error-free right-hand side $\mathbf{b}$. Then the norm $\|\tilde{\mathbf{b}}-\mathbf{b}\|$ provides a convenient measure of the error in $\tilde{\mathbf{b}}$. We refer to $\|\tilde{\mathbf{b}}-\mathbf{b}\|$ as the *absolute error* in $\tilde{\mathbf{b}}$, and, to the quotient $\|\tilde{\mathbf{b}}-\mathbf{b}\|/\|\mathbf{b}\|$ as the *relative error* in $\tilde{\mathbf{b}}$.

Let an estimate of the absolute or relative errors in $\tilde{\mathbf{b}}$ be available. We would like to bound the absolute or relative errors in the solution of a linear system of equations

$$A\mathbf{x}=\tilde{\mathbf{b}}$$

caused by the error in $\tilde{\mathbf{b}}$. Examples 10.2 and 10.3, and in particular Figures 1 and 2, show that the error in the solution does not only depend on the error in $\tilde{\mathbf{b}}$, but also on the matrix A. The *condition number* of a matrix, introduced below sheds light on this dependence.

Let $A\in\mathbb{R}^{n\times n}$ be a nonsingular matrix, and let $\mathbf{x}$ and $\tilde{\mathbf{x}}$ solve

$$A\mathbf{x}=\mathbf{b},\qquad A\tilde{\mathbf{x}}=\tilde{\mathbf{b}}. \tag{6}$$

It is convenient to express $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{b}}$ in the form

$$\tilde{\mathbf{x}}=\mathbf{x}+\delta\mathbf{x},\qquad \tilde{\mathbf{b}}=\mathbf{b}+\delta\mathbf{b}.$$

We would like to bound the absolute error $\|\delta\mathbf{x}\|$. Subtracting one of the equations (6) from the other one gives

$$A(\delta\mathbf{x})=\delta\mathbf{b},$$

and therefore $\delta\mathbf{x}=A^{-1}(\delta\mathbf{b})$. Application of the compatibility property of the matrix norm (see (24) of Lecture 1) to this equation yields the bound

$$\|\delta\mathbf{x}\|\leq\|A^{-1}\|\,\|\delta\mathbf{b}\|. \tag{7}$$

We see that if $\|A^{-1}\|$ is large, then the error $\delta\mathbf{x}$ in the computed solution $\tilde{\mathbf{x}}$ may be much larger than the error $\delta\mathbf{b}$ in the available right-hand side $\tilde{\mathbf{b}}$.

Example 10.1 cont'd

Assume that the cell population at time t_1 was miscounted by one unit. Then $\|\delta\mathbf{b}\|=1$. The inverse of the matrix in (2) is of norm 7.18. We therefore obtain from (7) the bound

$$\|\delta\mathbf{a}\|\leq 7.18\cdot 1.$$

It follows that the coefficient a_2 of the solution $\mathbf{a}$ is bounded by 82 (if we assume that the error in $\mathbf{b}$ does not change the coefficient a_1). Replacing 74 by 82 in (3) yields that $x(2)=9409$. This is 652 larger than the value reported in (3). We conclude that the determined number of cells at time t_2+1 can be quite sensitive to errors in the cell count. □

In some applications a bound for the relative error $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ is more relevant than a bound for the absolute error $\|\delta\mathbf{x}\|$. We therefore derive a bound for the former. The compatibility property ((24) of Lecture 1) applied to the left-hand side equation in (6) can be expressed as

$$\frac{1}{\|\mathbf{x}\|} \leq \|A\| \frac{1}{\|\mathbf{b}\|}. \tag{8}$$

Combining the bounds (7) and (8) shows that

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{9}$$

The quantity

$$\kappa(A) = \|A\|\,\|A^{-1}\| \tag{10}$$

is referred to as the condition number of A. It bounds the relative error in $\mathbf{x}$ in terms of the relative error in $\mathbf{b}$. We conclude that the relative error in the computed solution $\tilde{\mathbf{x}}$ may be much larger than the relative error in the available right-hand side $\tilde{\mathbf{b}}$ if the the condition number $\kappa(A)$ is large.

It follows from equations (13) and (15) of Lecture 7 that

$$\kappa(A) = \frac{\sigma_1}{\sigma_n}, \tag{11}$$

where σ_1 and σ_n are the largest and smallest singular values of A. While the derivation of (10) required the matrix A to be square, formula (11) does not; it is also valid for $m \times n$ matrices with $m > n$. This is commented on further below. It follows from (11) that

$$\kappa(A) \geq 1.$$

A matrix with a condition number fairly close to unity is said to be *well conditioned*. Matrices with a large condition number are said to be *ill conditioned*.

Example 10.4

Let A be the matrix of Example 10.3. Then A has norm 1.62 and its inverse

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

has the same norm. Therefore, $\kappa(A) = 2.62$.

The error-free right-hand of Example 10.3 is $\mathbf{b} = [1, 2]^T$ and the error satisfies $\|\delta\mathbf{b}\| = 1$. Therefore, $\|\delta\mathbf{b}\|/\|\mathbf{b}\| = 0.45$. We obtain from (7) and (9) that

$$\|\delta\mathbf{x}\| \leq 1.62 \cdot 1, \qquad \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq 2.62 \cdot 0.45 = 1.18.$$

□

Example 10.5

Let $\tilde{\mathbf{x}}$ denote an approximate solution of the linear system of equations

$$A\mathbf{x} = \mathbf{b},$$

and define the associated residual error

$$\tilde{\mathbf{r}} = \mathbf{b} - A\tilde{\mathbf{x}}.$$

We may consider $\tilde{\mathbf{r}}$ an error in $\mathbf{b}$ and apply the formulas (7) and (9) to bound the absolute and relative errors in $\tilde{\mathbf{x}}$; see Exercises 10.3 and 10.4. $\Box$

Let $A \in \mathbb{R}^{m\times n}$ with $m \geq n$. Proceeding analogously as in Section 7.2, we obtain

$$\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \min_{\|\mathbf{x}\|=1} \|U\Sigma V^T\mathbf{x}\| = \min_{\|\mathbf{x}\|=1} \|\Sigma V^T\mathbf{x}\| = \min_{\|\mathbf{y}\|=1} \|\Sigma\mathbf{y}\| = \sigma_n. \tag{12}$$

Combining (11), (12) with (15) of Lecture 7 shows that

$$\kappa(A) = \frac{\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}{\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|}. \tag{13}$$

This expression also can be used as the definition of the condition number. It agrees with (10) for square matrices, and has the advantage of applying to $m \times n$ matrices with $m \geq n$. Moreover, the quotient (13) suggests a geometric interpretation of the the condition number: the condition number is the quotient of how much a matrix can stretch and shrink the unit sphere.

Similarly as the condition number was helpful for establishing the bound (9) for the relative error in the solution due to a perturbation of the right-hand side $\mathbf{b}$, the condition number also enters naturally when bounding the error in the computed solution caused by round-off errors during the computations. For instance, assume that the matrix $A \in \mathbb{R}^{n\times n}$ is nonsingular, and let $\hat{\mathbf{x}}$ denote the exact solution of the linear system of equations $A\mathbf{x} = \mathbf{b}$. Let $\tilde{\mathbf{x}}$ be the approximate solution computed in finite precision arithmetic with the aid of the QR factorization of A described in Lecture 6. Then using the property (16) of Lecture 6, one can show that

$$\frac{\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}}\|} = \mathcal{O}(\kappa(A)\mathsf{eps}). \tag{14}$$

Thus, if A is not very ill-conditioned, then the method delivers accurate answers in the presence of round-off errors.

Finally, let $A \in \mathbb{R}^{m\times n}$, $m > n$, and consider the least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|. \tag{15}$$

The bound (14) also holds for consistent least-squares problems, i.e., when the value of the expression (15) is zero. If the least-squares problem is inconsistent, then the error in the computed solution may be larger; in addition to (14) the error bound contains a term proportional to $(\kappa(A)^2 \min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|\mathsf{eps})$.

Exercise 10.1

Generate the Toeplitz matrix A, solution $\mathbf{x}$, and right-hand side $\mathbf{b}$ with the MATLAB/Octave commands

$$A = \mathit{toeplitz}(\mathbf{v}); \qquad \mathbf{x} = \mathit{ones}(10,1); \qquad \mathbf{b} = A * \mathbf{x};$$

where $\mathbf{v} = [1, 1/2, 1/3, \ldots, 1/10]^T$. Determine experimentally the sensitivity to errors in the solution of

$$A\mathbf{x} = \tilde{\mathbf{b}},$$

when $\tilde{\mathbf{b}}$ is a perturbation of the error-free right-hand side $\mathbf{b}$. The perturbation $\delta\mathbf{b} = \tilde{\mathbf{b}} - \mathbf{b}$ should be of norm 0.1 and have normally distributed components with zero mean. Vectors $\delta\mathbf{b}$ with this property can be generated with the MATLAB/Octave function *randn*. This function determines vectors with normally distributed components, which have to be scaled appropriately. Determine the smallest constant c, such that the inequality

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq c\,\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

holds for your experiments. In view of the bound (9), do your experiments suggest that A is well conditioned or ill conditioned? Compare the computed value of c with $\kappa(A)$. The latter can be evaluated with the MATLAB/Octave command $\mathsf{cond}(A)$. Which one is larger, c or $\kappa(A)$? Is this to be expected? □

Exercise 10.2

Carry out the above experiments with the Toeplitz matrix A replaced by a Hilbert matrix of order 10. The latter can be generated with the MATLAB/Octave command

$$A = \mathit{hilb}(10);$$

Hilbert matrices are very ill-conditioned. Compare the computed coefficient c with the condition number of A. □

Exercise 10.3

Let A be the Toeplitz matrix of Exercise 10.1 and let $\mathbf{b} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{10}$. Solve the linear system of equations $A\mathbf{x} = \mathbf{b}$ by using the backslash operator in MATLAB/Octave. Evaluate $\mathbf{r} = \mathbf{b} - A\mathbf{x}$. Determine a bound for the relative error in $\mathbf{x}$. □

Exercise 10.4

Repeat Exercise 10.3 with A the Hilbert matrix of Exercise 10.2. □

Exercise 10.5

Let $A = QR$ be the QR factorization of $A \in \mathbb{R}^{m \times n}$, $m \geq n$. Show that $\kappa(A) = \kappa(R)$, cf. (14). □