2.1. **Perturbation theory for linear systems of equations.** We next try to understand for what types of matrices, small changes in the entries of the matrix or small changes in the right hand side have the potential to produce large changes in the solution. This is relevant, because we will have roundoff errors in our computation and one way to view the computed solution is that it is the exact solution of a problem where A and b are perturbed by small changes.

Before obtaining our main result, we need one preliminary result.

Lemma 1. *Let E be an $n \times n$ matrix with $\|E\| < 1$ for some norm. Then $I + E$ is non-singular and $\|(I + E)^{-1}\| \leq 1/(1 - \|E\|)$.*

Proof. If $I + E$ were singular, then there would be a vector $y \neq 0$ such that $(I + E)y = 0$. Then $Ey = -y$ and so $\|Ey\| = \|y\|$. Hence,

$$\|E\| = \max_{x \neq 0} \frac{\|Ex\|}{\|x\|} \geq \frac{\|Ey\|}{\|y\|} = 1.$$

This contradicts the assumption that $\|E\| < 1$, so E must be non-singular. To establish the bound, let $G = (I + E)^{-1}$. Then $(I + E)G = I$, so $G = I - EG$. Hence,

$$\|G\| = \|I - EG\| \leq \|I\| + \|EG\| \leq 1 + \|E\|\|G\|.$$

Then $\|G\|(1 - \|E\|) \leq 1$ and so $\|(I + E)^{-1}\| \equiv \|G\| \leq 1/(1 - \|E\|)$. □

Using this result, we can now establish a bound on the relative error in the computation of the solution of a linear system of equations.

Theorem 1. *Suppose A is a non-singular $n \times n$ matrix and $b \in \mathbb{R}^n$ is a given vector. Let x_T and x_C satisfy $Ax_T = b$ and $(A + \delta A)x_C = b + \delta b$ (i.e., x_T is the true solution and x_C is the computed solution, assumed to be the exact solution of a perturbed problem). If $\|\delta A\|\|A^{-1}\| < 1$, then*

$$\frac{\|x_T - x_C\|}{\|x_T\|} \leq \frac{\mu}{1 - \mu \frac{\|\delta A\|}{\|A\|}} \left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right], \qquad \text{where } \mu = \|A\|\|A^{-1}\|.$$

Proof. Now

$$(A + \delta A)(x_T - x_C) = Ax_T + \delta A x_T - b - \delta b = \delta A x_T - \delta b.$$

Hence,

$$(I + A^{-1}\delta A)(x_T - x_C) = A^{-1}\delta A x_T - A^{-1}\delta b$$

and so

$$x_T - x_C = (I + A^{-1}\delta A)^{-1}(A^{-1}\delta A x_T - A^{-1}\delta b).$$

Then taking norms and using Lemma 1 with $E = A^{-1}\delta A$, we get

$$\begin{aligned}\|x_T - x_C\| &\le \|(I + A^{-1}\delta A)^{-1}\|[\|A^{-1}\|\|\delta A\|\|x_T\| + \|A^{-1}\|\|\delta b\|] \\ &\le \frac{1}{1 - \|A^{-1}\delta A\|}[\|A^{-1}\|\|\delta A\|\|x_T\| + \|A^{-1}\|\|\delta b\|] \\ &\le \frac{1}{1 - \|A^{-1}\|\|\delta A\|}[\|A^{-1}\|\|\delta A\|\|x_T\| + \|A^{-1}\|\|\delta b\|\frac{\|A\|\|x_T\|}{\|b\|}],\end{aligned}$$

where we used the fact that $\|b\| \le \|A\|\|x_T\|$, and the hypothesis $\|\delta A\|\|A^{-1}\| < 1$. Dividing through by $\|x_T\|$ and writing $\|A^{-1}\| = \mu/\|A\|$, with $\mu = \|A\|\|A^{-1}\|$, gives the result of the theorem. □

The quantity μ is called the condition number of the matrix A. Note if $\delta A = 0$, then

$$\frac{\|x_T - x_C\|}{\|x_T\|} \le \mu\frac{\|\delta b\|}{\|b\|}.$$

This result shows that making a small change in the right hand side will cause only a small change in the solution if μ is not too large. Hence μ is a measure of the ill-conditioning of the system. Note that $\|I\| \le \|A\|\|A^{-1}\| = \mu(A)$, so $\mu(A) \ge 1$.

It is possible to prove that x_C, the solution obtained using Gaussian elimination, is the exact solution of a system $(A + \delta A)x_C = b$ in which bounds can be given for δA (called backward error analysis).

Theorem 2. *Let A be an $n \times n$ nonsingular matrix and assume that full or partial pivoting is used in the elimination process. Let $\rho = \max_{1\le i,j\le k\le n} |a_{ij}^{(k)}|$ and let u denote the unit roundoff error of the machine (i.e., the smallest positive u such that $1 + u > 1$). Then we have the following results. (i) The matrices L and U computed using Gaussian elimination satisfy $LU = A + E$, with $\|E\|_\infty \le n^2\rho\|A\|_\infty u$, (ii) The approximate solution x_C satisfies $(A + \delta A)x_C = b$, with $\|\delta A\|_\infty/\|A\|_\infty \le 1.01(n^3 + 3n^2)\rho u$.*

Hence, by the perturbation theorem, we have:

Corollary 1.

$$\frac{\|x_T - x_C\|}{\|x_T\|} \le \frac{\mu}{1 - \mu\frac{\|\delta A\|}{\|A\|}}\left[1.01(n^3 + 3n^2)\rho u\right].$$

This bound is quite pessimistic, since ρ can be as much as 2^{n-1} for partial pivoting, although more typically $\rho \sim n$.

We next introduce some useful definitions.

Definition: A matrix A is *strictly* diagonally dominant if

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

and is *weakly* diagonally dominant if

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

and strict inequality holds for at least one value of i.

Definition: A matrix A of order n is *irreducible* if $n = 1$ or if $n > 1$ and for any i and j such that $1 \leq i, j \leq n$ with $i \neq j$, either $a_{ij} \neq 0$ or there exists a sequence $i_1, i_2, \ldots, i_s$ such that $a_{i,i_1} a_{i_1,i_2} \cdots a_{i_s,j} \neq 0$.

One can interpret this definition using graph theory. Consider n points labeled 1 through n. For each i and j such that $a_{ij} \neq 0$, draw an arrow from point i to point j, i.e., $1 \leftarrow 2$, if $a_{21} \neq 0$. If it is possible to get from each node to any other node by a sequence of such arrows, then A is irreducible.

Example: $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$. The only paths we have to check correspond to the off-diagonal elements which are zero. In this case, $a_{13} = 0$ and $a_{31} = 0$. But since a_{12} and $a_{23} \neq 0$ we have $a_{12}a_{23} = 0$ and since a_{32} and $a_{21} \neq 0$, we have $a_{32}a_{21} \neq 0$. Hence A is irreducible.

In terms of the graph theory approach described above, since a_{12} and $a_{23} \neq 0$ we have $1 \rightarrow 2 \rightarrow 3$. Since a_{32} and $a_{21} \neq 0$, we have $1 \leftarrow 2 \leftarrow 3$. Since we can get from each node to any other node by a sequence of such arrows, A is irreducible. Note that A is also weakly diagonally dominant.

The relevance of these concepts is seen in the following results.

Theorem 3. *If A is either strictly diagonally dominant or both weakly diagonally dominant and irreducible, then A has an LU decomposition without pivoting.*

Theorem 4. *If A is a real symmetric matrix with non-negative diagonal elements which is either strictly diagonally dominant or both weakly diagonally dominant and irreducible, then A is positive definite (i.e., $x^T Ax > 0$ for $x \neq 0$).*