Errors in Gaussian Elimination

Many would mark the birth of modern numerical analysis as a branch of mathematics with the 1947 paper of von Neumann and Goldstine: "Numerical Inverting of Matrices of High Order". A perspective hinted at, if not explicitly stated there, of viewing the computed solution as the exact solution to another problem, is called *backward error analysis.* If we can show that our computed solution is always the exact solution to a nearby problem, then we call the method *backward stable.*

Without pivoting, GE is not stable. Here is a backward error result that applies when no zero pivots are encountered: If $\tilde{L}$ and $\tilde{U}$ are the computed versions of L and U, respectfully, then there exists an $\delta A \in \mathbb{R}^{n\times n}$ for which

$$\tilde{L}\tilde{U} = A + \delta A, \qquad \text{where} \qquad \frac{\|\delta A\|}{\|A\|} = \frac{\|L\|\|U\|}{\|A\|}\mathrm{O}(\mu).$$

This result does not imply backward stability because $\|L\|$ or $\|U\|$ can be arbitrarily large. But with partial pivoting $\|L\| = \mathrm{O}(n)$ and the only concern is with $\|U\|$. Turning to U we define the *growth factor* for GE to be

$$\rho = \|U\|/\|A\|.$$

The analogous backward error result for GEPP is then

$$\tilde{L}\tilde{U} = \tilde{P}(A + \delta A), \qquad \text{where} \qquad \frac{\|\delta A\|}{\|A\|} = \rho n\mathrm{O}(\mu).$$

This implies GEPP is backward stable for fixed n if $\rho = \mathrm{O}(1)$.

For fixed n and nonsingular A, ρ cannot be arbitrarily large, so GEPP is technically backward stable. On the other hand, we know examples for which $\|U\| = \mathrm{O}(2^n)$ (and this really violates the spirit of $\mathrm{O}(\mu)$). We haven't (yet) run into such examples in applications, so a popular compromise is to call GEPP "backward stable in practice": in real world problems GEPP has (thus far) given the exact factorization of a matrix relatively close to A.

Backward and forward substitution, on the other hand, are clearly backward stable. The result for back substitution is that the computed $\tilde{x}$ satisfies

$$(R + \delta R)\tilde{x} = b, \qquad \text{where} \qquad \frac{\|\delta R\|}{\|R\|} = \mathrm{O}(\mu).$$

Combining the results above, we can say the that the computed solution, $\tilde{x}$ to $Ax = b$, using G.E.P.P with forward and backward substitution, satisfies

$$(A + \delta A)\tilde{x} = b, \qquad \text{where} \qquad \frac{\|\delta A\|}{\|A\|} = \rho n^3\mathrm{O}(\mu).$$

The n^3 term above (a product of 3 upper bounds that depend on norms) appears quite pessimistic, for in practice we see

$$\frac{\|\delta A\|}{\|A\|} = \rho n\mathrm{O}(\mu).$$