NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

A Stochastic Approach to Error Estimates for Iterative Linear Solvers, Part 1

by

Gene H. Golub and Hallgeir Melbø

PREPRINT
NUMERICS NO. 3/2001

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/numerics/2001/N3-2001.ps
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

A STOCHASTIC APPROACH TO ERROR ESTIMATES FOR ITERATIVE LINEAR SOLVERS PART 1

G.H. GOLUB[1] AND H. MELBØ[2]

[1] *Computer Science Department, Stanford University, Stanford CA 94305, USA. email: golub@sccm.stanford.edu*

[2] *Department of Mathematics, Norwegian University of Science and Technology 7491 Trondheim, Norway. email: hallgeir@math.ntnu.no*

Abstract.

We consider the problem of estimating the magnitude of the error of an iterative linear solver after k iterations. Assuming that the initial error can be described using a probability distribution we derive l_2-estimates for the magnitude of the error in the average case. In Part 1 the ideas are presented and applied to a simple splitting method, while Part 2 extends the same ideas to the Conjugate Gradient method.

AMS subject classification: 65F99.

Key words: Stochastic error estimates, iterative solvers

1 Introduction.

Reliable estimates of the size of the error is relevant to devising a good stopping criterion for iterative linear solvers. The usual way of stopping is to calculate the residual and then stop when some function of the norm of the residual is small. Normally, this is a good stopping criterion if we are interested in a solution that approximately satisfies the equation. However, such a solution does not necessarily give a good approximation to the exact solution. Indeed, if the condition number of the matrix is large, then it is possible that the magnitude of the residual is not a good measure of the size of the error; the latter may be much larger.

One possibility for estimating the size of the error is to use classical upper bounds for the norm of the error vector, derived from the spectral radius of the matrix. Unfortunately, these upper bounds typically over-estimate the actual error.

In this article we will introduce statistical estimates that are aimed at estimating the l_2-norm of the error, in the *average case*, after k iterations with an iterative linear solver.

In Part 1 we will describe how to evaluate the expected value and the variance of the error for a simple splitting method associated with a symmetric linear system, and show how the orthogonality of the eigenvectors can be used. The initial error can be spanned, and the associated coefficients are assumed to arise from some probability distribution. Both a normal distribution and a more sophisticated distribution will be considered. In Part 2 we will discuss how these ideas can be extended to more advanced techniques such as the Conjugate Gradient method.

Our analytical observations will be accompanied by several numerical examples.

2 Estimates for a simple iterative solver

As a first approach, we will look at an iterative solver which is very simple, but which should give a good indication of how the model will work for more complicated iterative methods. The system to be solved is

$$Ax = b,$$

where $A \in \mathbf{R}^{n\times n}$ is symmetric and nonsingular, and where $x, b \in \mathbf{R}^n$. After a suitable transformation, the matrix is split into

$$A = I - B\ .$$

The system can now be written as

$$(I - B)x \quad = \quad b\ ,$$

and we consider the iterative method

$$x_k \quad = \quad Bx_{k-1} + b\ . \tag{2.1}$$

This formulation is very close to what is known as Jacobi's method, and we can easily find an expression for the error $e_k = x - x_k$ by subtracting $x = Bx + b$ from equation (2.1), yielding

$$e_k = Be_{k-1}\ , \tag{2.2}$$

and

$$\|e_k\| = \|Be_{k-1}\| = \|B^k e_0\| \leq \|B^k\|\|e_0\| \leq \|B\|^k\|e_0\|.$$

For any e_0, a sufficient condition for $\|e_k\|$ to approach zero as k goes to infinity is

$$\|B\| < 1.$$

This gives us $\|B\|^k\|e_0\|$ as an *upper bound* for the error. Let $\lambda_i(B)$ be the eigenvalues of B, and define the spectral radius of B as

$$\rho(B) = \max_{1\leq i\leq n} |\lambda_i(B)|.$$

Then $\rho(B) < 1$ is also a sufficient condition for the method to converge. In what follows, norms without subscripts are l_2-norms. Since A is assumed to be symmetric, B is also symmetric. Therefore it has orthogonal eigenvectors and it can be decomposed into

$$B = Q \Lambda Q^T,$$

where Q is an orthogonal matrix ($QQ^T = I$), with the normalised eigenvectors of B as its columns. The matrix Λ is diagonal, with the eigenvalues of B as its entries. As Q has orthonormal columns there exists a vector $\hat{\alpha}$ such that for any initial vector $e_0 = x - x_0$ we have

$$e_0 = Q\hat{\alpha} = \sum_{i=1}^{n} \alpha_i q_i. \tag{2.3}$$

The q_i's are the columns of Q. Now the following assumption is the core of the stochastic approach:

We assume that the coefficients α_i in the linear combination of eigenvectors (2.3) are stochastic and that they all have the same distribution.

The above assumption is reasonable as long as A has orthogonal eigenvectors, and for the following theory to be applicable it also needs real eigenvalues. The matrix A being symmetric is therefore crucial throughout the article.

As all the α_i's have the same distribution we will just write α where there can be no confusion. The vector of α_i's will be denoted by $\hat{\alpha}$. Under the above assumption we will find expressions for the expected value and the variance of $\|e_k\|^2$.

Using equation (2.2) and writing $(\cdot,\cdot)$ for the usual l_2-inner product we get

$$\begin{aligned} \|e_k\|^2 &= (B^k e_0, B^k e_0) \\ &= e_0^T B^{2k} e_0 \\ &= \hat{\alpha}^T Q^T Q \Lambda^{2k} Q^T Q \hat{\alpha} \\ &= \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2k}. \end{aligned}$$

Here only the α's are stochastic (for a particular A), and we can easily find the expected value and the variance of the expression.

$$E\left(\|e_k\|^2\right) = E\left(\alpha^2\right) \sum_{i=1}^{n} \lambda_i^{2k}\ , \tag{2.4}$$

$$\begin{aligned} Var\left(\|e_k\|^2\right) &= Var\left(\alpha_1^2 \lambda_1^{2k} + \ldots + \alpha_n^2 \lambda_n^{2k}\right) \\ &= Var\left(\alpha^2 \lambda_1^{2k}\right) + \ldots + Var\left(\alpha^2 \lambda_n^{2k}\right) + 2\sum_{i<j} Cov(\alpha_i^2, \alpha_j^2) \lambda_i^{2k} \lambda_j^{2k} \\ &= Var\left(\alpha^2\right) \sum_{i=1}^{n} \lambda_i^{4k} + 2Cov_{i \neq j}(\alpha_i^2, \alpha_j^2) \sum_{i<j} \lambda_i^{2k} \lambda_j^{2k}. \end{aligned} \tag{2.5}$$

Note that the sums may also be written as

$$\begin{aligned}\sum_{i=1}^{n} \lambda_i^{2k} &= tr(B^{2k})\ ,\\ \sum_{i<j} \lambda_i^{2k}\lambda_j^{2k} &= \frac{1}{2}\left[tr(B^{2k})\right]^2 - \frac{1}{2}tr(B^{4k})\ ,\end{aligned}$$

where $tr(B)$ is the trace of B. To evaluate the expressions for the expected value and the variance, we need $E\left(\alpha^2\right)$, $Var\left(\alpha^2\right)$ and $Cov(\alpha_i^2, \alpha_j^2)$ which will be determined from the chosen distribution. This will be addressed in the next section. The second problem will be to estimate the sums. What makes this task hard is that most of the λ's are assumed to be unknown.

3 Statistical issues

In this section we will look at two different, but related, distributions and find algebraic expressions for the three necessary coefficients in equation (2.4) and (2.5). A typical choice of distribution for the α's would be to assume

$$\alpha \sim N(0, \sigma^2).$$

It is reasonable to choose a distribution that is symmetric around zero since we do not want to be biased towards either positive or negative coefficients. Also it seems reasonable to choose a distribution that has a positive probability for all values of α, which eliminates uniform distributions. (If we have a lot of knowledge about the solution, uniform distributions can still apply.) Since a normal distribution satisfies all this, in addition to being a good description of lots of natural phenomena, it seems to be a natural choice for practical cases.

If the α's are assumed to be $N(0, \sigma^2)$ the following results are well known:

$$\begin{aligned}E(\alpha^2) &= \sigma^2\\ Var(\alpha^2) &= 2\sigma^4\\ Cov(\alpha_i^2, \alpha_j^2) &= 0.\end{aligned} \tag{3.1}$$

The second distribution we will consider is the *Dirichlet distribution*. When testing the estimates we would often like to run lots of test cases with random α's drawn from a normal distribution. This construction might however lead to e_0's of different lengths, which makes it difficult to compare the $||e_k||$'s fairly. A solution is to normalise $||e_0||$ before we start the iterations. The requirement is then,

$$||e_0||^2 = \sum_{i=1}^{n} \alpha_i^2 = 1. \tag{3.2}$$

Starting with normally distributed coefficients

$$\beta_i \sim N(0, 1),$$

the normalised coefficients that satisfy equation (3.2) will be

$$\alpha_i = \frac{\beta_i}{\sqrt{\sum_{j=1}^{n} \beta_j^2}} \qquad i = 1...n. \tag{3.3}$$

It can now be shown (see e.g. Melbø[8]) that the α_i^2's are distributed according to a *Dirichlet distribution* with the parameters $\nu_i = 1/2$ for $i = 1...n$. This is denoted $(\alpha_1^2, ..., \alpha_n^2) \sim \mathcal{D}(1/2, ...1/2)$. The quantities of interest are then found to be (see e.g. Johnson and Kotz [7])

$$\begin{aligned} E\left(\alpha^2\right) &= \frac{1}{n} \\ Var\left(\alpha^2\right) &= \frac{2\,(n-1)}{n^2\,(n+2)} \\ Cov\left(\alpha_i^2\,, \alpha_j^2\right) &= \frac{-2}{n^2\,(n+2)}\,. \end{aligned} \tag{3.4}$$

For this distribution the last term in equation (2.5) is difficult to compute and it is tempting to leave it out. The sum is obviously positive, so dropping the whole covariance term will make the variance larger, in which case we will make an over estimate. It is easy to give an upper limit for this sum, which will be skipped when using the simple iterative solver. For simplicity let $\rho \equiv \rho(B)$, then

$$\begin{aligned} \sum_{i<j} \lambda_i^{2k} \lambda_j^{2k} &= \frac{1}{2}\left[tr(B^{2k})\right]^2 - \frac{1}{2} tr(B^{4k}) \\ &\leq \frac{1}{2}\left(n\rho^{2k}\right)^2 \\ &= \frac{n^2}{2}\rho^{4k}. \end{aligned}$$

As we require $\rho < 1$ to assure convergence, the absolute value of the term being skipped,

$$\begin{aligned} T &= \left| Cov\left(\alpha_i^2\,, \alpha_j^2\right) \sum_{i<j} \lambda_i^{2k} \lambda_j^{2k} \right| \\ &\leq \frac{\rho^{4k}}{n+2}\,, \end{aligned}$$

will approach zero as k increases, and it is small since we can assume n to be large.

Dropping the whole term may seem quite a rough approximation, but we are seeking only an order of magnitude estimate for $||e_k||^2$, and it turns out that the variance will not play a very important role even if the covariance is skipped. The covariance plays an even smaller role when CG is applied, so from now on this term will be left out.

4 Estimating the sums

What is left to estimate in equations (2.4) and (2.5) are the sums. When using the simple solver we will rewrite the sums as integrals and then use a quadrature rule. The sums can be rewritten as

$$\begin{aligned}\sum_{i=1}^{n} \lambda_i^m &= \int_{\lambda_{min}}^{\lambda_{max}} \delta(\lambda - \lambda_j)\, \lambda^m \, d\lambda \\ &= \int_{\lambda_{min}}^{\lambda_{max}} \omega\,(\lambda)\, g_m(\lambda)\, d\lambda \; . \end{aligned} \tag{4.1}$$

Here $\omega\,(\lambda)$ is a weight function, $g_m(\lambda)$ is the function to be integrated and $\delta\,(\lambda - \lambda_j)$ is Dirac's delta function which is nonzero each time the variable λ hits one of the eigenvalues of B. The delta function is a valid weight function for quadrature rules.

We will use a Gauß-Radau quadrature rule to approximate the integral, which means that one of the abscissas is fixed to one end of the interval while the other abscissas are placed in positions that give us optimal order for the approximation. We therefore assume that λ_{min} and λ_{max} are known so we can choose to fix the abscissa to either side. To decide the unknown variables in such a quadrature rule we need to compute the integral (4.1) exactly for as high powers of λ as we can afford. In this case that will be the same as calculating the trace of B^m for m as large as possible. The following moments are easy to compute:

$$\begin{aligned}\mu_0 &= \sum_{i=1}^{n} \lambda_i^0 = trace\,(B^0) = n \\ \mu_1 &= \sum_{i=1}^{n} \lambda_i = trace\,(B) \\ \mu_2 &= \sum_{i=1}^{n} \lambda_i^2 = trace\,(B^2) = \sum_{i,j} b_{ij}^2 = \|B\|_F^2 \; ,\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. More moments can of course be calculated, but not without having to perform matrix-matrix products which are relatively expensive. The Gauß-Radau quadrature rule generally looks like

$$\int_{\lambda_{min}}^{\lambda_{max}} \omega\,(\lambda)\, g_m(\lambda)\, d\lambda \; = \; w_1 g_m(t_1) \; + \; \sum_{i=2}^{p} w_i g_m(t_i) \; + \; Err \; ,$$

where w_i are the weights, t_i are the abscissas and in our case $g_m(\lambda) = \lambda^m$ and $p = 2$ since we can compute only three moments. Err is the error term. Our approximation will then be

$$\sum_{i=1}^{n} \lambda_i^m \approx \bar{\mu}_m = w_1 t_1^m \; + \; w_2 t_2^m \; , \tag{4.2}$$

in which $\bar{\mu}_m$ is our approximation to μ_m. In cases where more moments are known, Davis and Rabinowitz [3] present a very straight forward way of finding the weights and abscissas for this type of quadrature rules. However, since we here have only three known moments it is just as easy to follow Bai and Golub [1], where the two following sets of equations are derived,

$$\begin{aligned} c\mu_2 + d\mu_1 - \mu_0 &= 0 \\ ct_1^2 + dt_1 - 1 &= 0 \end{aligned}$$

and

$$\begin{aligned} \mu_1 &= w_1 t_1 + w_2 t_2 \\ \mu_2 &= w_1 t_1^2 + w_2 t_2^2 \; , \end{aligned}$$

in addition to the relation $t_2 = -\frac{1}{ct_1}$. This means we can solve the first 2×2 system for c and d since t_1 is prescribed, then find t_2 and solve the second system for w_1 and w_2. This completes the Gauß-Radau quadrature rule, which is of order 2 since we will have one fixed and one free abscissa.

If we look at the error term for this rule ($g_m(\lambda) = \lambda^m, \; p = 2$), it can be shown that

$$Err = \frac{1}{6} m\,(m-1)(m-2)\eta^{m-3} \int_{\lambda_{min}}^{\lambda_{max}} \omega\,(\lambda)(\lambda - t_1)(\lambda - t_2)^2 \, d\lambda \; ,$$

for some η satisfying $\lambda_{min} < \eta < \lambda_{max}$. The term that decides the sign of Err is $(\lambda - t_1)$, so by prescribing t_1 to be λ_{min} we would get $Err > 0$ which gives a lower bound for the integral. Our choice will be $t_1 = \lambda_{max}$, which makes $Err < 0$ and gives an upper bound.

The cost of estimating the sums in equation (2.4) and (2.5) will be solving the 2×2 systems and evaluating λ^m for two values of λ, which is very cheap. Also it is usually very easy to obtain a good estimate for the largest eigenvalue, see e.g. Parlett [9].

5 A numerical experiment for the simple method

We are now able to compute the error estimates for the simple method described by equation (2.1). To test our estimates, we have run 40 tests, always solving the same problem, but using a different initial vector each time. The problem which we have considered is

$$Ax = 0 \; ,$$

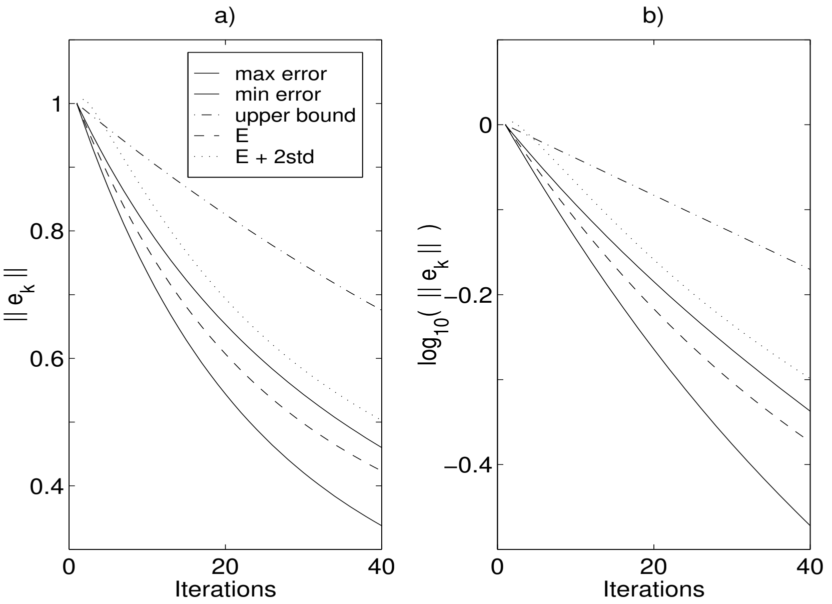

Figure 5.1: Plot a) shows the estimate compared to the upper bound and the largest and the smallest error over the 40 runs. In plot b) the same is shown with a log_{10} scale. The estimate is marked with E, and the labels on the lines are the same in both plots.

where A is very close to the usual 3-node finite difference discretisation of the one-dimensional Laplacian. However, a small adjustment is made to the Laplacian to make $\|B\| < 1$. The test matrix can be written as

$$A = (U_{xx} + I)/100 \ , \qquad A \in \mathbf{R}^{100\times 100} \ ,$$

where I is the identity and U_{xx} is the discrete Laplacian. This choice makes $\|B\| = 0.9$. The initial vectors are different linear combinations of the eigenvectors of B where the coefficients are drawn from a standard normal distribution and normalised according to equation (3.3). This means that

$$(\alpha_1^2, \dots, \alpha_n^2) \sim \mathcal{D}\,(1/2, .., 1/2) \ ,$$

where α_i are the normalised coefficients. Thus we *know* that the assumptions are satisfied. Note that for this problem the initial vector is the same as the initial error, e_0. We can now compute estimates for $E(\|e_k\|^2)$ and $Var(\|e_k\|^2)$ using (2.4), (2.5), (3.4) (with covariance equal to zero) and (4.2).

The results of the experiment can be seen in Figure 5.1. In plot a) we are using a linear scale on the y-axis, while in plot b) we have a $\log_{10}$ scale. The "max error" and "min error" lines indicate the maximum and minimum error over the 40 runs. The “upper bound” line shows an upper bound computed as $||B||^k$. Finally, the “E” line is our estimate of the expected value using the upper bound for the sum, and “std” is the square root of the estimated variance. All computations have been carried out for $||e_k||^2$, and at the end we have taken the square root, which means "E + 2std" actually is

$$\begin{aligned} \text{“}E + 2std\text{”} &= \sqrt{E(||e_k||^2) + 2\sqrt{Var(||e_k||^2)}} \\ &\neq E(||e_k||) + 2\sqrt{Var(||e_k||)}\,. \end{aligned}$$

The quantity $E(x) + 2\sigma(x)$ is of interest because for a normally distributed variable x it makes the upper limit for a 95% confidence interval. In our case it makes no sense with a lower estimate as it may become negative, while in practice zero is always the lower bound. The value for λ_{max} was found using the function $eig()$ in Matlab.

From the plot we can see that our estimate follows the actual error quite closely, and compared to the upper bound the estimate works very well. For the iterations beyond what is shown in Figure 5.1, the lines straighten out and become parallel to the bound in the plot with the logarithmic y-axis.

6 Conclusion

We have derived a statistically based error estimate for a simple iterative solver, and in a test example we have seen that the estimator shows the right behaviour. It is worth noting that the estimate is much cheaper to compute than the upper bound for this method, as the upper bound requires computation of the norm of a matrix, preferably the l_2-norm.

This algorithm, however, is not an effective method for solving equations, but it clearly shows a way of deriving stochastic error estimates. In Part 2 we will extend our results to the Conjugate Gradient method for which more extensive testing will be performed.

REFERENCES

1. Z.Bai and G.H.Golub. *Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices.* SCCM-96-06, Stanford University. http://www-sccm.stanford.edu/reports.html#start-1996
2. P.Concus, G.H.Golub and D.P.O'Leary. *A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations.* In "Sparse Matrix Computations", ed. J.R.Bunch and D.J.Rose, Academic Press, NY, 1976.
3. P.J.Davis and P.Rabinowitz. *Methods of numerical integration.* pp. 14-19, 109-112 Academic Press, 1984.
4. G.H.Golub. *Bounds for the round-off errors in the Richardson second order method*, BIT, 2(4) (1962), pp. 212–223.
5. G.H.Golub and G.Meurant. *Matrices, moments and quadrature.* SCCM-93-07, Stanford University. http://www-sccm.stanford.edu/reports.html#start-1993.
6. G.H.Golub and C.F.Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 3rd edition, 1996.
7. N.L.Johnson and S.Kotz. *Encyclopedia of statistical sciences, Volume 2.* pp. 386-387, John Wiley & Sons, 1982.
8. H.Melbø. *Preconditioning and error estimates for iterative linear solvers.* Masters thesis, Dept. of Math., NTNU.
http://www.math.ntnu.no/~hallgeir/tekster
9. B.N.Parlett. *The Symmetric Eigenvalue Problem.* SIAM, 1998.