Backward Error Estimates for Toeplitz and Vandermonde Systems

by

J.M. Varah

Technical Report 91-20
September 1991

Department of Computer Science
University of British Columbia
Vancouver, B.C.
CANADA V6T 1Z2

Backward Error Estimates for Toeplitz and Vandermonde Systems

J.M. Varah
Computer Science Department
University of British Columbia

Abstract

Given a computed approximate solution $\bar{x}$ to $Ax = b$, it is of interest to find nearby systems with $\bar{x}$ as exact solution, and which have the same structure as A. In this paper, we show that the distance to these nearby structured systems can be much larger than for the corresponding general perturbation for Toeplitz and Vandermonde systems. In fact, even the correctly rounded solution $\hat{x}$ may require a structured perturbation of $O(\eta \|\hat{x}\|)$, not $O(\eta)$ as might be expected.

Introduction

Given the linear system $Ax = b$ and a computed solution $\bar{x}$, it is of interest to find nearby systems for which $\bar{x}$ is the exact solution. That is, to find δA and δb such that

$$(A + \delta A)\,\bar{x} \;= b + \delta b \tag{1.1}$$

with δA and δb small.

If we define the associated residual vector

$$r = r(\bar{x}) = b - A\bar{x},$$

then (1.1) becomes

$$(\delta A)\bar{x} = r + \delta b \tag{1.2}$$

If we consider general perturbations δA and δb, then these conditions (1.1) or (1.2) do not specify them fully, and we must impose additional conditions (such as minimizing some measure of the size of δA and δb). If however the matrix A has some special form, and we are interested in maintaining this form in the allowable perturbations, then the solution of (1.1) or (1.2) becomes more complicated. In this paper, we consider the cases of A being of Toeplitz or Vandermonde form. This issue of restricted perturbations for structured systems has also been considered by Higham and Higham [6].

Notice that the scaling of the problem is important; we assume throughout that $\|A\| = O(1)$ and $\|b\| = O(1)$, so that ill-conditioning of A is reflected in $\|x\|$

being (possibly) large, but not small. In fact

$$\frac{\| b \|}{\| A \|} \leq \| x \| \leq \kappa(A)\left(\frac{\| b \|}{\| A \|}\right)$$

where $\kappa(A) = \|A\| \, \|A^{-1}\|$ is the (standard) condition number of A in any norm. We also assume throughout that A is nonsingular.

To get some sense of the size of the residual $r(\bar{x})$, it is useful to consider what happens in the best possible case, when $\bar{x} = \hat{x}$, the correctly rounded solution. Since

$$\hat{x}_i = x_i(1 + \eta_i)$$

where $|\eta_i| \leq \eta =$ machine roundoff level, we can write

$$\hat{x} = (\mathrm{I} + D_n)\, x$$

and then

$$r = r(\hat{x}) = b - A\hat{x} = -AD_n x \qquad (1.3)$$

giving

$$||r|| \,/\, ||x|| \leq \eta \, ||A||.$$

In particular, $||r|| \,/||x|| = \mathrm{O}(\eta)$ independent of the solution x. We also remark that (1.3) implies

$$|r| \leq \eta |A| \, |x|, \qquad (1.4)$$

where the inequality is meant to be taken component-wise.

Thus the most we can expect for a computed solution $\bar{x}$ is that $||r(\bar{x})|| \,/\, ||\bar{x}|| = \mathrm{O}(\eta)$. Such behaviour occurs, for example, with solutions computed by Gaussian elimination or Cholesky factorization. As a result, it is not appropriate in general to solve (1.1) or (1.2) by taking $\delta A = \mathrm{O}$. This gives $\delta b = -r$, which for $||\bar{x}||$ large means a large backward error $||\delta b|| = ||r|| = \mathrm{O}(\eta \cdot ||\hat{x}||)$ even for the correctly rounded solution. Instead, one attempts to find solutions δA, δb with

$$||\delta A|| = \mathrm{O}\left(\frac{||r||}{||x||}\right), \; ||\delta b|| = \mathrm{O}\left(\frac{||r||}{||x||}\right). \qquad (1.5)$$

Allowing general perturbartions δA, δb, one can indeed find solutions satisfying (1.5), as has been known for some time, and we review this material in Section 2. See also the excellent survey paper by Higham [5]. Then in Section 3 we consider Toeplitz perturbations of Toeplitz matrices. We are motivated to do this from interest in the stability properties of special methods available for solving Toeplitz systems, such as the Levinson method (see Golub and van Loan [4, page 183] for example). One might hope that the computed solution obtained from such a method is the exact solution of a nearby Toeplitz system. Indeed, Bunch [3] refers to this behaviour as "strong stability". However, we find that under these restrictions, the perturbations δA and δb satisfy not (1.5) but

$$||\delta A|| = O(||r||)\,,\ ||\delta b|| = O(||r||). \tag{1.6}$$

This means that for ill-conditioned Toeplitz systems, computed solutions (even correctly rounded solutions) satisfy Toeplitz systems which are as much as $O(\kappa(A)\cdot\eta)$ away from the original system. We illustrate this behaviour with some numerical examples in Section 4, and finally in Section 5 we discuss the same problem for Vandermonde systems, where the conclusion is the same.

2. General Backward Error

Consider the basic equation (1.2) with $(\delta A)_{ij} = \varepsilon_{ij}$ and $(\delta b)_i = \delta_i$. The equations decouple, and we consider the first one in detail:

$$(\varepsilon_{11} ... \varepsilon_{1n})\bar{x} = r_1 + \delta_1.$$

If $r_1 = 0$, we can take $\varepsilon_{11} = \ldots = \varepsilon_{1n} = \delta_1 = 0$. So assume $r_1 = 0$ and let

$$\varepsilon_{ij} = r_1 e_i z_i, \quad \delta_i = r_1 f y,$$

where $\{e_i\}_1^n$ and f are fixed scaling factors and$\{z_i\}_1^n$ and y are to be determined. Then the defining equation can be written

$$(e_1\bar{x}_1 \ldots e_n\bar{x}_n\ -f)\begin{pmatrix} z_1 \\ \vdots \\ z_n \\ y \end{pmatrix} = 1 \qquad (2.1)$$

or $u^T v \ = \ 1$.

Normally, besides satisfying the equation (2.1), we want to make the perturbations as small as possible in some sense, which amounts to minimizing $||v||$ for some norm. In particular, if we use a Hölder norm $||v||_q$, with dual = p,

$$1 = \left|u^T v\right| \leq ||u||_p ||v||_q$$

for any u and v satisfying (2.1), and

$$\underset{v}{min}\ ||v||_q = \frac{1}{||u||_p}.$$

We could use $p = q = 2$, but it is more natural to use $p = 1$, $q = \infty$, giving

$$\underset{v}{min}\ ||v||_\infty = \frac{1}{||u||_1} = \frac{1}{f + \Sigma e_i |\bar{x}_i|} \qquad (2.2)$$

which is attained by using $v \ni v_i = sgn(u_i)/||u||_1$.

This max norm solution translates into

$$\varepsilon_{1i} = \frac{\pm r_1 e_i}{f + \Sigma e_i |\bar{x}_i|},\ \delta_1 = \frac{\pm r_1 f}{f + \Sigma e_i |\bar{x}_i|}$$

which replicates the Oettli/Prager result [7] for general scaling factors E and f. One particular case deserves special mention: $e_i = ||A||, f = ||b||$. Then

$$e_{1j} = \frac{\pm r_1 ||A||}{||b|| + ||A||\ ||\bar{x}||},\ \delta_1 = \frac{\pm r_1 ||b||}{||b|| + ||A||\ ||\bar{x}||}$$

and similarly for the other rows. Notice that in this case, we do obtain

$$\frac{||\delta A||}{||A||} = O\left(\frac{||r||}{||\bar{x}||}\right),\ \frac{||\delta b||}{||b||} = O\left(\frac{||r||}{||\bar{x}||}\right),$$

as predicted in Section 1.

3. The Toeplitz Case

Now assume A is a symmetric positive definite Toeplitz matrix, and that we want δA to be symmetric and Toeplitz as well. That is,

$$\delta A = \begin{bmatrix} \varepsilon_0 & \varepsilon_1 & \cdots & \varepsilon_{n-1} \\ \varepsilon_1 & \varepsilon_0 & \cdots & \varepsilon_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n-1} & \cdots & \cdots & \varepsilon_0 \end{bmatrix}$$

In the defining equation (1.2), the key observation is to rewrite $(\delta A)\bar{x}$ as $X\underline{\varepsilon}$, where $\underline{\varepsilon}$ is the vector $(\varepsilon_0, \ldots, \varepsilon_{n-1})^T$ and (for n odd):

$$X = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \cdots & \cdots & \cdots & \bar{x}_{n-1} & \bar{x}_n \\ \bar{x}_2 & (\bar{x}_1 + \bar{x}_3) & \bar{x}_4 & \cdots & \cdots & \cdots & \bar{x}_n & 0 \\ \bar{x}_3 & (\bar{x}_2 + \bar{x}_4) & (\bar{x}_1 + \bar{x}_5) & \cdots & \cdots & & & \vdots \\ \vdots & & & \ddots & & & & \vdots \\ \bar{x}_{(n+1)/2} & \cdots & \cdots & \cdots & (\bar{x}_1 + \bar{x}_n) & 0 & \cdots & 0 \\ \vdots & & & & & & & \vdots \\ \vdots & & & & & \ddots & & \vdots \\ \bar{x}_{n-1} & (\bar{x}_{n-2} + \bar{x}_n) & \bar{x}_{n-3} & \cdots & \cdots & \cdots & \bar{x}_1 & 0 \\ \bar{x}_n & \bar{x}_{n-1} & \bar{x}_{n-2} & \cdots & \cdots & \cdots & \bar{x}_2 & \bar{x}_1 \end{bmatrix} \qquad (3.1)$$

For n even, the middle row and column are not present. Notice that each $\bar{x}_i$ appears once in each row, and that $||X||_\infty = ||\bar{x}||_1$. Also, notice that X can be singular: if $\Sigma \bar{x}_i = 0$ for example, then $Xe = 0$ for $e = (1, 1, \ldots, 1)^T$.

Using this matrix, the defining equation (1.2) reads

$$X\underline{\varepsilon} - \underline{\delta} = \underline{r}$$

where $\underline{\delta} = \delta b$, which is n equations in the $2n$ unknowns $\underline{\varepsilon}$ and $\underline{\delta}$. This can also be expressed as

$$(X \; -I)\binom{\varepsilon}{\delta} = r \text{ or } G\underline{v} = \underline{r}. \qquad (3.2)$$

Again, we would like the perturbation to be as small as possible, and thus we are led to the constrained optimization problem

$$\min_{v} \|v\| \ni Gv = r.$$

Notice that we can include component-wise scaling factors in $\underline{\varepsilon}$, $\underline{\delta}$ by using diagonal scaling factors D_ε, D_δ giving $G = (XD_\varepsilon \; -D_\delta)$.

Again the most natural norm is $\|v\|_\infty$ giving a constrained Chebyshev optimization problem, which can be rewritten using $\delta = X\varepsilon - r$ as the overdetermined discrete Chebyshev problem

$$\min_{\varepsilon} \left\| \begin{pmatrix} X \\ I \end{pmatrix} \underline{\varepsilon} - \begin{pmatrix} r \\ 0 \end{pmatrix} \right\|_\infty .$$

This problem is difficult to solve explicity, although algorithms have been developed to solve individual cases (see [1] or [2]). The basic question here is whether the solution $\|v\|_\infty = \mathrm{O}(\|r\|/\|\bar{x}\|)$.

The following example shows that this is not always true, and that the consequence is that even for a rounded solution $\hat{x}$, the closest perturbed symmetric Toeplitz system with $\hat{x}$ as exact solution, is $\mathrm{O}(\eta\|\hat{x}\|)$ away.

Example:

$$A = \begin{bmatrix} 1 & 1-\mu & 1-\alpha \\ 1-\mu & 1 & 1-\mu \\ 1-\alpha & 1-\mu & 1 \end{bmatrix}$$

with μ and α small and positive. (In fact, μ is not crucial in what follows.) One eigenvalue $\lambda_1 = \alpha$ with corresponding eigenvector $(1,\, 0,\, -1)^T$.

Hence when $b = (1, 0, -1)^T$, the solution $x = \frac{1}{\alpha}(1, 0, -1)^T$. Now take the rounded solution $\hat{x} = \frac{1}{\alpha}(1 + \eta_1,\, 0,\, -(1 + \eta_3))^T$. A short calculation gives, using $\hat{\eta} = \eta_1 - \eta_3$,

$$r(\hat{x}) = -\frac{\hat{\eta}}{\alpha}(1,\, 1,\, 1)^T + 0(\eta)$$

and

$$X = \frac{1}{\alpha}\begin{bmatrix} 1+\eta_1 & 0 & -(1+\eta_3) \\ 0 & \hat{\eta} & 0 \\ -(1+\eta_3) & 0 & 1+\eta_1 \end{bmatrix}.$$

Thus the equations $Gv = r$ are:

$$\begin{gathered} (1+\eta_1)v_1 - (1+\eta_3)v_3 - \alpha v_4 = -\hat{\eta} \\ \hat{\eta}v_2 - \alpha v_5 = -\hat{\eta} \\ -(1+\eta_3)v_1 + (1+\eta_1)v_3 - \alpha v_6 = -\hat{\eta} \end{gathered}$$

It is easy to see that the minimax solution of these equations has all $|v_i| = \hat{\eta}/(\alpha+\hat{\eta})$. In fact,

$$v_1 = v_2 = v_3 = -\hat{\eta}/(\alpha+\hat{\eta}),\ v_4 = v_5 = v_6 = \hat{\eta}/(\alpha+\hat{\eta}).$$

Hence for this example, we have explicitly

$$\begin{gathered} ||\hat{x}-x||/||x|| = O(\eta),\ ||r(\hat{x})|| = O(\eta/\alpha), \\ ||\delta A|| = ||\delta b|| = O(\eta/\alpha). \end{gathered}$$

Also, one can find a much closer general perturbation δA with $(A+\delta A)\hat{x} = b$; in fact

$$\delta A = \frac{-\hat{\eta}}{2}\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} + O(\eta^2)$$

Later on, we provide numerical evidence of the same behaviour, using rounded solutions, and solutions computed by the Cholesky and Levinson algorithms. All exhibit the same behaviour as above.

Now return to the basic problem (3.2). Although an explicit form of the solution in the max norm is not available, we can find an approximate solution by solving the corresponding constrained least squares problem

$$min\|v\|_2 \ni Gv = r.$$

If $M_2 = \overset{min}{v}\|v\|_2$ and $M_\infty = \overset{min}{v}\|v\|_\infty$, then

$$\frac{1}{\sqrt{n}} M_2 \leq M_\infty \leq M_2$$

and hence the solutions differ in value by at most a factor $\sqrt{n}.$

Moreover, the least squares solution has a simple form:

$$v = G^T z,\ Gv = r$$

that is,

$$\left(I + XX^T\right)z = r,\ \begin{pmatrix}\varepsilon \\ \delta\end{pmatrix} = \begin{pmatrix}X^T z \\ -z\end{pmatrix}. \tag{3.3}$$

Alternatively, one can solve the overdetermined problem

$$\underset{\underline{\varepsilon}}{min} \left\| \begin{pmatrix}X \\ I\end{pmatrix}\underline{\varepsilon} - \begin{pmatrix}r \\ 0\end{pmatrix} \right\|_2$$

with solution $(I + X^T X)\underline{\varepsilon} = X^T r$ and $\delta = X\underline{\varepsilon} - r$.

Notice that from this formulation, it easily follows that if $r(\bar{x})$ is such that $r^T X = 0$, then $\underline{\varepsilon} = 0$ and $\underline{\delta} = -r$, which gives a solution (as we mentioned earlier) that is unacceptably large when $\|\bar{x}\|$ is large. However this is by no means the only troublesome case, as we see below.

Of course, the more acceptable computational method for finding ν, at least in cases where X is ill-conditioned, is to use the QR decomposition (Golub and van Loan [4]):

$$(i)\quad set\ G^T = \begin{pmatrix}X^T \\ -I\end{pmatrix} = Q\begin{pmatrix}R \\ 0\end{pmatrix} = (Q_1 \mid Q_2)\begin{pmatrix}R \\ 0\end{pmatrix}$$

$$(ii)\quad find\ w = Q^T v \quad via \left\{\begin{matrix}R^T w^{(1)} = r \\ w^{(2)} = 0\end{matrix}\right\}$$

$$(iii)\quad find\ v = Qw = Q_1 w^{(1)}.$$

An even more explicit formulation of the minimum least squares solution can be derived from (3.3) using the SVD.

Theorem 1: Let $X = UDV^T$ be the singular value decomposition of the matrix X in (3.1), and let $r = \Sigma\beta_i u^{(i)}$ be the expansion of $r = r(\bar{x})$ in the singular vectors $\left\{u^{(i)}\right\}$. Then the solution z to (3.3) has the expansion

$$z = \Sigma\left(\frac{\beta_i}{1+\sigma_i^2}\right)u^{(i)}$$

and the minimal least squares solution v has

$$||v||_2 = z^T r = \Sigma\frac{\beta_i^2}{1+\sigma_i^2}.$$

Proof: by substitution.

This theorem gives a complete description of the minimum restricted perturbation in the ℓ_2 sense which makes $\bar{x}$ exact. This basic result can be applied in various ways, as we now explore.

Theorem 2: Suppose the computed solution $\bar{x}$ is such that the SVD coefficients $\{\beta_i\}$ of $r(\bar{x})$ satisfy

$$|\beta_i| \leq c\sqrt{1+\sigma_i^2}\,\eta.$$

Then $||v||_2 \leq c\sqrt{n}\,\eta$ and the minimum perturbation $||\delta A||$, $||\delta b|| = \mathrm{O}(\eta)$.
Proof: again, direct substitution.

A result like this clearly can hold for solutions $||x|| = \mathrm{O}(1)$, whatever $\kappa(A)$ is, as long as the algorithm used to compute $\bar{x}$ produces $r(\bar{x}) = \mathrm{O}(\eta)$. Since $||X||_\infty = ||\bar{x}||_1$, $\mathrm{O}(1) = \sigma_1(X) \geq \sigma_2 \geq \ldots \geq \sigma_n$, and hence all $\beta_i = \mathrm{O}(\eta)$. Notice that near-singularity of X is immaterial.

Now consider large $||x||$, and recall that $||x||$ can be as large as $\kappa(A)$. Assume that $\bar{x}$ produces a residual with $||r(\bar{x})|| = O(\eta \cdot ||\bar{x}||)$, as occurs with $\bar{x} = \hat{x}$, the correctly rounded solution. Then some $\{\beta_i\}$ are also this large, and some $\{\sigma_i\}$ are $O(||\bar{x}||)$. If the large $\{\beta_i\}$ occur only for large $\{\sigma_i\}$, then Theorem 2 can still hold. However, experimental evidence with ill-conditioned symmetric positive definite Toeplitz systems indicates that in most cases all $\beta_i = O(\eta||\bar{x}||)$, whether $\bar{x}$ was the correctly rounded solution, the Cholesky solution, or the solution computed using the Levinson algorithm. We present some examples in the next section. In such cases, as long as not all $\sigma_i = O||\bar{x}||$ (i.e. some $\sigma_i = O(1)$ or smaller), the minimum ℓ_2 perturbation is $O(\eta||\bar{x}||)$, which can be $O(\kappa(A) \cdot \eta)$ if x fully reflects the ill-condition of A. Notice that this does not require X to have very small singular values, only that there is an appreciable spread in their range.

4. Some Numerical Results

The first example is the 3x3 matrix from Section 3:

$$A = \begin{bmatrix} 1 & 1-\mu & 1-\alpha \\ 1-\mu & 1 & 1-\mu \\ 1-\alpha & 1-\mu & 1 \end{bmatrix}.$$

We took $\alpha = 10^{-6}, \mu = \alpha/3$. A has then two eigenvalues near 10^{-6}. For various data vectors b, we computed solutions to $Ax = b$ as follows:

$$\begin{aligned} &(i)\ \bar{x} = \mathit{Cholesky\ solution} \\ &(ii)\ \bar{\bar{x}} = \mathit{Levinson\ solution} \\ &(iii)\ \hat{x} = \mathit{Correctly\ rounded\ solution,\ obtained\ from\ \bar{x}\ using} \\ &\qquad\qquad \mathit{double\ precision\ iterative\ refinement} \end{aligned}$$

Working precision was long precision on an IBM mainframe, with special routines for "double long" calculations, so $\eta \simeq 10^{-16}$. For each approximate solution, we computed the residual r, the minimal least squares solution $v = \binom{\epsilon}{\delta}$ from (3.3), the singular values $\sigma_i(X)$, and the coefficients $\beta_i(r)$.

In Case 1, b reflects the ill-condition of A, and $||x||$ is large. For each approximate solution, the closest Toeplitz system with that exact solution is

$O(\eta||x||)$ away. In Case 2 however, $||x|| = O(1)$ and even though X is singular, the closest perturbation is now only $O(\eta)$.

$$\textbf{Case 1}: \ b = (-0.72, 0.55, 0.22)^T \ \ ||x||_\infty = 4.8\text{x}10^6$$
$$\sigma_i(X): \ 1.0\text{x}10^7, 2.9\,\text{x}10^6, 0.32$$

	$\|\|r\|\|_\infty$	$\|\|v_{min}\|\|_2$	β
$\bar{\text{x}}$	$1.0\text{x}10^{-10}$	$4.7\text{x}10^{-11}$	$-.25\text{x}10^{-10}, -.14\text{x}10^{-9}, .50\text{x}10^{-10}$
$\bar{\bar{\text{x}}}$	$3.2\text{x}10^{-10}$	$3.9\text{x}10^{-11}$	$-.15\text{x}10^{-9}, -.33\text{x}10^{-9}, -.41\text{x}10^{-10}$
$\hat{\text{x}}$	$1.8\text{x}10^{-10}$	$8.7\text{x}10^{-11}$	$-.86\text{x}10^{-10}, .28\text{x}10^{-9}, -.92\text{x}10^{-10}$

$$\textbf{Case 2}: \ b = (-0.58, \ -.58, \ 0.58)^T \ \ ||x||_\infty = 0.19$$
$$\sigma_i(X): \ .60, \ .21, \ .84\text{x}10^{-9}$$

	$\|\|r\|\|_\infty$	$\|\|v_{min}\|\|_2$	β
$\bar{x}$	$1.0\text{x}10^{-17}$	$1.4\text{x}10^{-17}$	$-.16\text{x}10^{-16}, .47\text{x}10^{-18}, .93\text{x}10^{-18}$
$\bar{\bar{x}}$	$2.4\text{x}10^{-17}$	$3.2\text{x}10^{-17}$	$-.37\text{x}10^{-16}, .37\text{x}10^{-17}, .31\text{x}10^{-17}$
$\hat{x}$	$2.1\text{x}10^{-17}$	$3.1\text{x}10^{-17}$	$.36\text{x}10^{-16}, -.40\text{x}10^{-17}, -.88\text{x}10^{-26}$

As a second example, consider the prolate matrix (see Slepian [8]) of order 11 with $a_{ij} = \gamma_{|j-i|}$,

$$\gamma_k = \frac{sin(\pi k/2)}{\pi k}, \ \gamma_o = 1/2.$$

This matrix is positive definite symmetric and Toeplitz, and its smallest eigenvalue λ_1 is $O(10^{-7})$. The behaviour generally with various data vectors b is similar to that of the first example, and we mention only one case, where b is the eigenvector corresponding to λ_1. For this case (and for b= other eigenvectors as well), the matrix X is near-singular, but this does not affect the perturbation behaviour. The β—coefficients are all $O(10^{-10})$.

$$b = \textit{first eigenvector}, \; ||x||_\infty = 5.4\text{x}10^6$$
$$\sigma_i(X): \; 3.3\text{x}10^7, \ldots, 7.3\text{x}10^{-8}$$

	$\|\|r\|\|_\infty$	$\|\|v_{min}\|\|_2$
$\bar{x}$	$3.1\text{x}10^{-10}$	$2.5\text{x}10^{-10}$
$\bar{\bar{x}}$	$1.5\text{x}10^{-10}$	$1.1\text{x}10^{-10}$
$\hat{x}$	$3.7\text{x}10^{-11}$	$2.1\text{x}10^{-11}$

5. The Vandermonde Case

Now consider the (primal) Vandermonde system $Ax = b$,

$$A = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \vdots & & & \\ \alpha_1^{n-1} & \ldots & \ldots & \alpha_n^{n-1} \end{bmatrix},$$

and again image a computed solution $\bar{x}$ as the exact solution of a nearby system of the same form $\overline{A}\bar{x} = b + \delta b$. Notice that the perturbation from A to $\overline{A}$ is no longer linear, as in the Toeplitz case.

One could again derive the equations satisfied by the perturbations, but one can see the extent of the perturbation required much more simply. Consider the first equation of the perturbed systems: since $\overline{a}_{ij} = a_{ij} = 1$, the equation gives

$$\delta_1 = -r_1,$$

where δ_1 is the first component of δb. As we have already seen, the residual norm $||r(\bar{x})||$ is at best $O(\eta||\bar{x}||)$ even for the correctly rounded solution, so unless the first component r_1 is unusually small (for large $||\bar{x}||$), we already have $|\delta_1| = O(\eta||\bar{x}||)$ and thus the full perturbation required is at least this large. However there is no reason to believe that $r_1 = b_1 - \Sigma\bar{x}_i$ will be unusually small, and this can be easily verified numerically. Indeed, typically

$|r_1| = O(\eta||\bar{x}||)$ for $\bar{x}$ the Gaussian elimination solution, the solution using the special Björck/Pereyra methods ([4], page 178), or the correctly rounded solution.

So again for Vandermonde systems, as in the Toeplitz case, the size of the restricted perturbations required can be much larger than for general perturbations, by a factor $||\bar{x}||$, which can be as large as $\kappa(A)$.

References

1. I. Barrodale and C. Phillips, Algorithm 495: Solution of an overdetermined system of linear equations in the Chebychev norm. ACM Trans. Math. Soft. 1 (1975), 264–270.

2. R. Bartels, A. Conn, and C. Charalambous, On Cline's direct method for solving overdetermined linear systems in the ℓ_∞ sense. SIAM J. Num. Anal. 15 (1978), 255–270.

3. J. Bunch, The weak and strong stability of algorithms in numerical linear algebra. Lin. Alg. Appl. 88/89(1987), 49–66.

4. G. Golub and C. van Loan, Matrix Computations (2nd ed.), Johns Hopkins Press, 1989.

5. N. Higham, How accurate is Gaussian elimination? In Numerical Analysis 1989, Proceedings of the 13th Dundee Conference, eds. D. Griffiths and G. Watson, Longman Scientific and Technical, (1990), pgs. 137–154.

6. D. Higham and N. Higham, Backward error and condition of structured linear system. Univ. of Manchester Math. Dept. Numerical Analysis Report #192, Sept. 1990.

7. W. Oettli and W. Prager, Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. Num. Math. 6 (1964), 405–409.

8. D. Slepian, Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: the discrete case. Bell System Tech. J. 57 (1978), 1371–1430.