U.C.L.A.
COMPUTATIONAL AND APPLIED MATHEMATICS

Effective Condition Numbers for Linear Systems

Tony F. Chan
David Foulser

August, 1986
(Revised February, 1987)
CAM Report 87-03

Department of Mathematics
University of California, Los Angeles
Los Angeles, CA. 90024

Effective Condition Numbers for Linear Systems

Tony F. Chan

Dept. of Mathematics
UCLA
Los Angeles, CA 90024

David Foulser

Saxpy Computer Corporation
255 San Geronimo Way
Sunnyvale, CA 94086

Abstract

When solving the linear system $A\mathbf{x} = \mathbf{b}$, the condition number $K(A) \equiv \|A\| \, \|A^{-1}\|$ is a useful measure of the sensitivity of the solution $\mathbf{x}$ under perturbations ΔA and Δb to A and $\mathbf{b}$, respectively. It is widely appreciated, however, that $K(A)$ is often an overly conservative estimate. We introduce the notion of an "effective" condition number $K_e \equiv K_e(A, \mathbf{b})$, which gives a more accurate measure of the sensitivity of $\mathbf{x}$, by taking into account the actual values of $\mathbf{b}$. The effective condition number provides insight into why observed numerical errors may fall below theoretical error bounds. We consider the effects on $\mathbf{x}$ of perturbations ΔA and $\Delta \mathbf{b}$, which may determine the influence of K_e in a particular solution algorithm. We present applications to a fast Poisson solver and a Vandermonde system solver that demonstrate the usefulness of K_e and the related perturbation analysis of linear systems and solvers.

1. Introduction

Consider the solution of the linear system

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

where A is a nonsingular $n \times n$ matrix and $\mathbf{x}$ and $\mathbf{b}$ are n-vectors. A fundamental question is the sensitivity of the solution $\mathbf{x}$ under perturbations to A and $\mathbf{b}$. Such perturbations are unavoidable, for example, when one represents (1) in finite precision numbers.

A sensitivity analysis often gives insight into the stability of numerical algorithms for solving (1). It is well-known that the condition number

$$K(A) \equiv \|A\| \, \|A^{-1}\| \tag{2}$$

Abbreviated title: Effective Condition Numbers
Keywords: Condition Number, Effective Condition Numbers, Stable Algorithm, Bjorck-Pereyra Algorithm, Fast Poisson Solver, Physical Stability, Error Analysis.

plays a central role in such an analysis. In fact, the following two perturbation bounds can be found in most textbooks on the subject [3]:

(a) If $A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$, then

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq K(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{3}$$

(b) If $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$, then

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x} + \Delta\mathbf{x}\|} \leq K(A)\frac{\|\Delta A\|}{\|A\|}. \tag{4}$$

In practice, however, the bounds in (3) and (4) are often overly conservative estimates of the actual relative errors in $\mathbf{x}$.

It should be noted that the bound in (4) is sharp in the sense that given A and $\mathbf{b}$, there exists a ΔA for which the equality sign in (4) holds for an induced matrix norm [10]. The bound in (3), however, is not as sharp. One can only say that given A, there exist $\mathbf{b}$ and $\Delta\mathbf{b}$ for which the equality in (3) holds. If both A and $\mathbf{b}$ are given, then the bound in (3) may be unachievable for any $\Delta\mathbf{b}$, and in these cases, the condition number $K(A)$ is not a good measure of the sensitivity of $\mathbf{x}$ under perturbations to $\mathbf{b}$. Various authors [4,7,9] have discussed this situation in the literature. Viewing ΔA as the perturbation to A obtained in a backward error analysis of a given method of solving (1) with finite precision numbers, the particular ΔA that produces equality in (4) may not be achieved in practice, rendering that bound overly pessimistic as well.

In general, the sensitivity of $\mathbf{x}$ in (1) depends not only on A but also on $\mathbf{b}$. For this reason, in this paper we introduce the notion of an "effective" condition number $K_e \equiv K(A, \mathbf{b})$ for the linear system (1) that explicitly accounts for the values of A and $\mathbf{b}$. This condition number is based on the singular value decomposition of A. One of our goals is to identify classes of problems for which K_e will provide a more accurate estimate of the sensitivity of $\mathbf{x}$.

For example, in the case of $\Delta \mathbf{b} \neq 0$ and $\Delta A = 0$ we shall derive a formula for K_e which not only depends on the relative sizes of the singular values of A, but also on the projection of $\mathbf{b}$ onto the left singular vectors of A.

Different algorithms for solving (1) may or may not achieve the smaller perturbations to $\mathbf{x}$ indicated by K_e. Considering the totality of rounding errors introduced in a finite precision calculation as a perturbation ΔA to the matrix A, it is clear that the size of the computed $\Delta \mathbf{x}$ depends on ΔA and $\Delta \mathbf{b}$, as well as on K_e. Thus it is also relevant to apply the effective conditioning analysis to particular solution algorithms; Theorem 2 may be interpreted in this manner.

In general, the case $\Delta A \neq 0$ and $\Delta \mathbf{b} = 0$ is more complicated and we show that $K_e = K(A)$ is the best that one can do for a general perturbation ΔA. However, for special classes of ΔA, it is possible to derive an expression that is a much sharper estimate of the sensitivity of $\mathbf{x}$ than that provided by $K(A)$. Van der Sluis considered such a family of perturbations [9] to columns of A. Here we shall show that, for a rather general class of ΔA, the previous formula for K_e is also a good measure of the sensitivity of $\mathbf{x}$ to perturbations ΔA. This result, embodied in Theorem 2, also helps to identify a class of algorithms which can take advantage of effectively well-conditioned problems.

We shall present two examples and a summary of numerical comparisons which demonstrate the usefulness of this new concept of an effective condition number. The first example is the solution of Vandermonde systems, for which we show that oscillatory right-hand sides can give much smaller relative errors in $\mathbf{x}$ than would be predicted by $K(A)$ alone. The second example is the solution of a Poisson equation, for which we show that smooth right-hand sides are the ones that are effectively well-conditioned. We also indicate why fast Poisson solvers are able to exploit the low effective condition number of this problem.

The idea behind the effective condition number is not completely new — it has been

widely appreciated that the usual condition number is not always a good estimator of the actual error in a computation [2]. For example, in the numerical simulation of semiconductor devices, the matrices that arise are often so badly conditioned that the usual condition number estimates would predict no correct digits even in 64-bit precision. However, such simulations are very successful and do produce useful results.

We present the effective condition number as a tool for investigating specific classes of problems and certain solution algorithms. It is likely that other effective condition numbers (not based on the SVD) will provide additional insight into the error behavior of linear systems and algorithms not considered here. Simpler condition numbers do exist (for instance, $K(A, \mathbf{b}) = \|A^{-1}\|\|\mathbf{b}\|/\|A^{-1}\mathbf{b}\|$), but they do not always provide the insight obtained with our effective condition number.

The structure of the paper is as follows. In the next section, we shall motivate the notion of an "effective" condition number from a physical point of view. Then in section 3, we shall develop the effective condition number $K_e(A, \mathbf{b})$. In section 4, we apply a similar analysis to the case $\Delta\mathbf{b} = 0$ and $\Delta A \neq 0$. The numerical examples will be discussed in section 5 with concluding remarks in section 6. All norms used in this paper are Euclidean norms.

2. Physically Stable Systems

The notion of an effective condition number can be applied to physical systems as well as to mathematical systems. In this section, we shall look at a simple physical system whose sensitivity depends not only on its intrinsic structural properties, but also on the nature of the external forces on it.

Consider an elastic beam supported at its two ends under load:

INSERT FIGURE 1

A simple mathematical model for this physical system is

$$u_{xx} = f(x), \quad u(0) = u(1) = 0 \tag{5}$$

where $u(x)$ is the displacement of the beam and $f(x)$ is the load. After a standard centered second-order finite difference discretization of (5). one obtains the linear system

$$L\,\mathbf{u} = \mathbf{f} \tag{6}$$

where L is a tridiagonal matrix and the components of $\mathbf{u}$ and $\mathbf{f}$ are the displacement and load at the mesh points. Thus the sensitivity of (6) under perturbations in $\mathbf{f}$ is related to the sensitivity of the physical displacement under perturbations in the load $f(x)$.

Consider a load $f(x) = \sin \pi x$

INSERT FIGURE 2

and another load $f(x) = \sin p\pi x$ where p is a large integer.

INSERT FIGURE 3

It should be physically intuitive that the situation in Fig. 3 is much more unstable than that in Fig. 2. In Fig. 3, the displacement $u(x)$ is much smaller than in Fig. 2 because the oscillating load tends to cancel itself out. Therefore, a perturbation to $f(x)$ of the form $\delta \sin \pi x$ would produce a large *relative* change in $u(x)$ for Fig. 3.

Referring to the linear system (6), it is clear that the "effective" sensitivity depends not only on the matrix L, but also on the special form of the right-hand side $\mathbf{f}$. This analogy has also been made by Hammarling and Wilkinson [4], who also discuss the case of no load with nonzero boundary conditions.

3. Effective Condition Number $K_e(A, \mathbf{b})$

We shall now make precise the notion of effective conditioning by deriving an expression for $K_e \equiv K_e(A, \mathbf{b})$ that incorporates the special form of $\mathbf{b}$. A bound on K_e involving $\Delta \mathbf{b}$ as well is developed following Theorem 1.

Let A have the following singular value decomposition:

$$A = U \Sigma V^T \tag{7}$$

where

$$U \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_n]$$

are the left singular vectors.

$$V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$$

are the right singular vectors, and

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$$

are the singular values, ordered so that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$. Thus $K(A) = \sigma_1/\sigma_n$.

Next define two projection operators P_k and P_k^- by

$$P_k \equiv U_k U_k^T, \qquad 1 \leq k \leq n, \tag{8}$$

where

$$U_k = [\mathbf{u}_{n-k+1}, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n \times k},$$

and

$$P_k^- \equiv I - U_k U_k^T. \tag{9}$$

Our first principal result is

Theorem 1. Let A have the singular value decomposition (7). If $A\mathbf{x} = \mathbf{b}$ and $A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$, then

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq K_e(A, \mathbf{b}; k) \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}, \qquad \text{for} \quad 1 \leq k \leq n, \tag{10}$$

where

$$K_e(A, \mathbf{b}: k) \equiv \frac{\sigma_{n-k+1}}{\sigma_n} \left(\frac{\|P_k \mathbf{b}\|}{\|\mathbf{b}\|} \right)^{-1} \tag{11}$$

and P_k is the projection operation defined in (8). Moreover, if one defines the effective condition number K_e by:

$$K_e(A, \mathbf{b}) \equiv \min_k K_e(A, \mathbf{b}: k),$$

then

$$K_e(A, \mathbf{b}) \leq K(A). \tag{12}$$

Proof. Let

$$\mathbf{b} = \sum_{i=1}^{n} \beta_i \, \mathbf{u}_i \tag{13}$$

and

$$\Delta\mathbf{b} = \sum_{i=1}^{n} \delta_i \, \mathbf{u}_i. \tag{14}$$

Then the solution $\mathbf{x}$ can be written as

$$\mathbf{x} = \sum_{i=1}^{n} \frac{\beta_i}{\sigma_i} \mathbf{v}_i. \tag{15}$$

Taking norms, we get

$$\begin{aligned} \|\mathbf{x}\|^2 &= \sum_{i=1}^{n} \frac{\beta_i^2}{\sigma_i^2} \\ &\geq \sum_{i=n-k+1}^{n} \frac{\beta_i^2}{\sigma_i^2} \\ &\geq \frac{1}{\sigma_{n-k+1}^2} \left(\sum_{i=n-k+1}^{n} \beta_i^2 \right) \\ &= \frac{\|P_k \mathbf{b}\|^2}{\sigma_{n-k+1}^2}. \end{aligned} \tag{16}$$

Next, from the expression for $\Delta\mathbf{x}$

$$\Delta\mathbf{x} = A^{-1}\Delta\mathbf{b} = \sum_{i=1}^{n} \frac{\delta_i}{\sigma_i} \mathbf{v}_i$$

we get

$$\|\Delta\mathbf{x}\|^2 = \sum_{i=1}^{n} \left(\frac{\delta_i}{\sigma_i} \right)^2 \leq \frac{1}{\sigma_n^2} \sum_{i=1}^{n} \delta_i^2 = \frac{\|\Delta\mathbf{b}\|^2}{\sigma_n^2}. \tag{17}$$

Combining (16) and (17), we get (11). (Note that a tighter bound than (17) could be derived from additional knowledge of $\Delta\mathbf{b}$.)

Since $P_n\mathbf{b} = \mathbf{b}$, we have

$$K_e(A, \mathbf{b}; n) = K(A)$$

and so (12) follows. □

The implication of the theorem is that if, for a particular value of k, the right-hand side $\mathbf{b}$ projects largely onto the first k left singular vectors of A, *i.e.*, that $\|P_k\mathbf{b}\| \approx \|\mathbf{b}\|$, then the effective condition number K_e is approximately bounded above by σ_{n-k-1}/σ_n, which could be much smaller than $K(A) \equiv \sigma_1/\sigma_n$, *e.g.*, for $k \ll n$. A major advantage of this effective condition number is that it provides insight into why $K(A) = \|A\| \cdot \|A^{-1}\|$ may over-estimate errors in $\mathbf{x}$. From this point of view, the singular value decomposition does not play an essential role in the theorem; one may usefully consider any orthogonal factorization of A in analyzing problem-dependent conditioning.

Example 1. If $\mathbf{b} = \mathbf{u}_n$, then we can take $k = 1$, and $K_e(A, \mathbf{b}) = 1$. In other words, if the right-hand side is a multiple of the left singular vector corresponding to the smallest singular value, then the linear system $A\mathbf{x} = \mathbf{b}$ is perfectly conditioned with respect to arbitrary perturbations in $\mathbf{b}$.

Example 2. If $\mathbf{b} = \mathbf{u}_1$, then $K_e(A, \mathbf{b}) = K_e(A, \mathbf{b}: n) = K(A)$. In this case, the effective sensitivity is no smaller than the sensitivity given by $K(A)$.

Note that Example 1 does *not* necessarily imply that one can solve for $\mathbf{x}$ with a precision that is independent of $K(A)$, because numerical solution procedures typically introduce errors that can be construed as perturbations to A as well. Take, for example, Gaussian elimination with complete pivoting. We will consider a typical backward error bound [8] for the computed solution $\hat{\mathbf{x}}$ in

$$(A + E)\hat{\mathbf{x}} = \mathbf{b} \tag{18}$$

in which the error matrix E can be bounded by

$$\|E\| \leq p(n)\|A\|\varepsilon_m. \tag{19}$$

with ε_m the machine precision and $p(n)$ a slowly growing polynomial in n. From (18), one can view $\hat{\mathbf{x}}$ as the exact solution to a perturbed problem with

$$E\,\hat{\mathbf{x}} = \Delta\mathbf{b}. \tag{20}$$

Now if $\mathbf{b} = \mathbf{u}_n$, then one can expect

$$\|\hat{\mathbf{x}}\| \approx \frac{1}{\sigma_n} = \|A^{-1}\|. \tag{21}$$

Therefore, using (10) with $k = 1$ in this case gives

$$\begin{aligned}\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} &\le K_e(A, \mathbf{b})\frac{\|E\,\hat{\mathbf{x}}\|}{\|\mathbf{b}\|} \\ &\approx p(n)\|A\|\,\|A^{-1}\|\varepsilon_m \\ &= p(n)\,K(A)\varepsilon_m.\end{aligned} \tag{22}$$

Thus the computed solution could still grow with $K(A)$, and so Gaussian elimination seems unable to take advantage of an especially nice right-hand side. However, as we shall see in sections 4 and 5, it is possible for other algorithms to produce extraordinarily accurate results for right-hand sides with small effective condition numbers K_e.

Before proceeding, we want to make one remark. We note that K_e, as defined in Theorem 1, is independent of $\Delta\mathbf{b}$. A tighter bound in (10) can be obtained by taking into account additional knowledge about $\Delta\mathbf{b}$. It can be verified that instead of the bound in (17), we could use

$$\|\Delta\mathbf{x}\|^2 \le \frac{1}{\sigma_{n-\ell}^2}\|P_\ell^\perp\Delta\mathbf{b}\|^2 + \frac{1}{\sigma_n^2}\|P_\ell\Delta\mathbf{b}\|^2. \tag{23}$$

to derive

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \sqrt{\left(\frac{\sigma_{n-k-1}}{\sigma_{n-\ell}}\left(\frac{\|P_k\mathbf{b}\|}{\|P_\ell^\perp\Delta\mathbf{b}\|}\right)^{-1}\right)^2 + \left(\frac{\sigma_{n-k-1}}{\sigma_n}\left(\frac{\|P_k\mathbf{b}\|}{\|P_\ell\Delta\mathbf{b}\|}\right)^{-1}\right)^2}. \tag{24}$$

for $1 \le k, \ell \le n$, and therefore in this case the bound depends on $\Delta\mathbf{b}$. The bound in (10) corresponds to choosing $\ell = n$ in (24). However, if $\Delta\mathbf{b}$ lies primarily within the span of

a few left singular vectors corresponding to large singular values, then a much improved bound can be derived. For example, if there is an ℓ^* such that $P_\ell \Delta \mathbf{b} = 0$ and $P_\ell^+ \Delta \mathbf{b} = \Delta \mathbf{b}$ for $\ell \leq \ell^*$, then by choosing $\ell = \ell^*$ in (24) we can refine (10) to

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\sigma_{n-k+1}}{\sigma_{n-\ell^*}} \left(\frac{\|P_k \mathbf{b}\|}{\|\mathbf{b}\|} \right)^{-1} \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|},$$

which could be much tighter than (10) if $\sigma_{n-\ell^*} \gg \sigma_n$. To give a more specific example, let $\Delta \mathbf{b} = \delta \mathbf{u}_1$; then by taking $\ell^* = n - 1$ we have improved (10) by a factor of σ_1/σ_n.

4. Effective Conditioning Under Perturbations to A

We now consider the solution of (1) under perturbations to A, thus solving the nearby $(A + \Delta A)\hat{\mathbf{x}} = \mathbf{b}$, and apply the techniques of section 3 to develop bounds on the sensitivity of $\hat{\mathbf{x}}$. We represent the error bound under perturbations ΔA as $K_e^* \equiv K_e^*(A, \mathbf{b}, \Delta A)$. It is useful to note that K_e^* applies to backward error perturbations as well as true perturbations to A. Thus K_e^* can be used to investigate the error properties of numerical algorithms. The following two examples show that, for the perturbed problem $(A + \Delta A)\hat{\mathbf{x}} = \mathbf{b}$, an error bound K_e^* that is to be sharper than the upper bound $K(A)$ must employ information about the nature of the perturbation ΔA.

Example 3. If $\Delta A = \varepsilon A$, then it is easy to verify that

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x} + \Delta \mathbf{x}\|} = \varepsilon = \frac{\|\Delta A\|}{\|A\|}.$$

Then $K_e^* = 1$ in this case, independent of A and $\mathbf{b}$, but depending on the structure of ΔA.

Example 4. If $A = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$, the right-hand side is $\mathbf{b} = (0, \ldots, 0, b_n)^T$, and $\Delta A = \delta I$, then one can verify that

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} = \frac{\delta}{\lambda_n} = \frac{\lambda_1}{\lambda_n} \frac{\|\Delta A\|}{\|A\|} = K(A) \frac{\|\Delta A\|}{\|A\|}.$$

Thus $K_e^* = K(A)$ in this case, due to the form of ΔA.

The above examples show that the special form of ΔA has a crucial effect on a problem's effective conditioning. If one is willing to make assumptions about ΔA, then sharper bounds can be derived. For instance, Van der Sluis [9] has considered special classes of ΔA based on perturbations to columns of A. In this paper we consider the effect of a different class of ΔA, which depends on a spectral decomposition of A.

Theorem 2. Let A and $\hat{A}$ be matrices having nearby singular value decompositions (with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$ and $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$, $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \ldots \geq \hat{\sigma}_n > 0$)

$$A = U \Sigma V^T \quad \text{and} \quad \hat{A} = \hat{U} \hat{\Sigma} \hat{V}^T \tag{25}$$

such that

$$\begin{aligned} U^T \hat{U} &= I + E_u, \quad \|E_u\| \leq \varepsilon_u \\ V^T \hat{V} &= I + E_v, \quad \|E_v\| \leq \varepsilon_v \end{aligned} \tag{26}$$

and

$$\Sigma \hat{\Sigma}^{-1} = I + E_\sigma, \quad \|E_\sigma\| \leq \varepsilon_\sigma.$$

for small positive values ε_u, ε_v, and ε_σ such that $\varepsilon_u + \varepsilon_v + \varepsilon_\sigma < 1$. (In this case $\|A - \hat{A}\|/\|A\|$ is necessarily small, although the converse is not true.) If $A\mathbf{x} = \mathbf{b}$ and $\hat{A}\hat{\mathbf{x}} = \mathbf{b}$, then

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\sigma_{n-k-1}}{\sigma_n} \left(\frac{\|P_k \mathbf{b}\|}{\|\mathbf{b}\|} \right)^{-1} \left(\frac{\varepsilon_\sigma + \varepsilon_u + \varepsilon_v}{1 - (\varepsilon_\sigma + \varepsilon_u + \varepsilon_v)} \right) \quad \text{for} \quad 1 \leq k \leq n \,. \tag{27}$$

Proof:

Writing $\hat{\mathbf{x}} - \mathbf{x}$ as $(\hat{A}^{-1} - A^{-1})\mathbf{b}$, it follows that

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}\| &= \|(\hat{V}\hat{\Sigma}^{-1}\hat{U}^T - V\Sigma^{-1}U^T)\mathbf{b}\| \\ &= \|(V^T\hat{V}\hat{\Sigma}^{-1}\hat{U}^T U - \Sigma^{-1})U^T\mathbf{b}\| \\ &= \| \left((I + E_v)\Sigma^{-1}(I + E_\sigma)(I + E_u) - \Sigma^{-1} \right) U^T\mathbf{b}\|. \end{aligned} \tag{28}$$

Repeated use of the triangle inequality gives

$$\|\hat{\mathbf{x}} - \mathbf{x}\| \leq \frac{\|\mathbf{b}\|}{\sigma_n}(\varepsilon_u + \varepsilon_v + \varepsilon_\sigma + \varepsilon_u\varepsilon_\sigma + \varepsilon_u\varepsilon_v + \varepsilon_v\varepsilon_\sigma + \varepsilon_u\varepsilon_v\varepsilon_\sigma),$$

which is conveniently bounded above by

$$\|\hat{\mathbf{x}} - \mathbf{x}\| \leq \frac{\|\mathbf{b}\|}{\sigma_n} \frac{(\varepsilon_u + \varepsilon_v + \varepsilon_\sigma)}{(1 - (\varepsilon_u + \varepsilon_v + \varepsilon_\sigma))}. \tag{29}$$

The denominator is split into its projections on P_k and $P_k^{\perp}$ as

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} \frac{\beta_i^2}{\sigma_i^2} \geq \frac{\|P_k\mathbf{b}\|^2}{\sigma_{n-k+1}^2}, \quad \text{for} \quad 1 \leq k \leq n. \tag{30}$$

Combining terms yields (27). □

Due to the form of perturbations $A - \hat{A}$,

$$K_e^*(A, \mathbf{b}, \Delta A) = K_e(A, \mathbf{b}) = \min_{1 \leq k \leq n} \frac{\sigma_{n-k+1}}{\sigma_n} \left(\frac{\|P_k\mathbf{b}\|}{\|\mathbf{b}\|} \right)^{-1}.$$

In general, if $\varepsilon_A = \|\hat{A}^{-1} - A^{-1}\|/\|A^{-1}\|$ then

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq K_e(A, \mathbf{b})\varepsilon_A,$$

although ε_A may be much larger than $\|\Delta A\|/\|A\|$. Theorem 2 determines $\varepsilon_A = (\varepsilon_\sigma + \varepsilon_u + \varepsilon_v)(1 - (\varepsilon_\sigma + \varepsilon_u + \varepsilon_v))^{-1}$ as a normalized bound measuring the difference of certain pairs of matrices A^{-1} and $\hat{A}^{-1}$. The conditions of Theorem 2 can be relaxed to consider other cases in which $\|(\hat{A}^{-1} - A^{-1})\mathbf{b}\|$ is small.

The import of Theorem 2 is that the solutions of certain pairs of related linear systems can differ by approximately $K_e(A, \mathbf{b})\frac{\|\Delta A\|}{\|A\|}$, which may be less than $K(A)\frac{\|\Delta A\|}{\|A\|}$. In terms of numerical algorithms, if a method of solving $A\mathbf{x} = \mathbf{b}$ in finite precision actually solves the nearby problem $\hat{A}\hat{\mathbf{x}} = \mathbf{b}$ with $\hat{A}$ satisfying the assumptions (26), then the actual solution error is determined by the effective condition number K_e and the small perturbation

ε_A. For a stable right-hand side $\mathbf{b}$ (*i.e.*, $\|P_k\mathbf{b}\|_2 \approx \|\mathbf{b}\|_2$ while $\sigma_{n-k-1} \ll \sigma_1$) and an effectively well-conditioned solution algorithm (*i.e.*, $\varepsilon_A \approx \frac{\|\Delta A\|}{\|A\|}$) the error $\|\hat{\mathbf{x}} - \mathbf{x}\|$ could be much smaller than suggested by $K(A) = \sigma_1/\sigma_n$. In this case, the analysis of errors in computation incorporates information about the effective condition number and the errors introduced by the solution algorithm. The fast Poisson solver example of section 5.2 will make clear the usefulness of this combined approach.

5. Applications

In this section we present two examples that demonstrate the relevance of effective condition numbers. In the first example, we provide insight into the well-conditioning of the Bjorck-Pereyra algorithm [1] for Vandermonde systems by reference to Theorem 1. The second example indicates that a fast Poisson solver can achieve substantially better error performance than Gaussian elimination, as suggested by Theorem 2. Numerical simulations verifying this hypothesis are summarized in Fig. 4.

5.1 Vandermonde Systems

A Vandermonde matrix is defined in terms of a set of scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ by

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & & \alpha_n \\ \vdots & \vdots & & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{bmatrix} \in \mathbb{R}^{n \times n} . \tag{31}$$

The associated linear systems

$$\text{Primal} \quad A\mathbf{x} = \mathbf{b} \tag{32}$$

$$\text{Dual} \quad A^T\mathbf{z} = \mathbf{f} \tag{33}$$

arise in a variety of applications, such as polynomial interpolation and function approximation. It is well known that Vandermonde systems can be extremely ill-conditioned. Yet

there exist efficient algorithms [1] which have been observed empirically to give extraordinarily accurate solutions.

Higham in a recent paper [5] proved that, under the ordering assumption $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_n$, the Bjorck-Pereyra algorithm [1] produces a relative error in the computed nonzero components of $\mathbf{x}$ that is *independent* of the condition number $K(A)$, *provided* that the right-hand side $\mathbf{b}$ in (32) has a sign oscillation property:

$$(-1)^i b_i \geq 0, \quad i = 1, 2, \ldots, n. \tag{34}$$

As an application of our Theorem 1, we shall now show that, under the ordering assumption, the singular vectors $\mathbf{u}_n$ and $\mathbf{v}_n$ of A have a sign oscillation property and thus right-hand side vectors $\mathbf{b}$ with property (34) can induce a small effective condition number K_e. The backward error analysis perturbations to A in solving (32)–(33) then determine the size of the resulting errors. The present argument explains why an extremely accurate solution method might exist, due to properties of the Vandermonde system. Higham's paper [5] indicates why the Bjorck-Pereyra algorithm computes such accurate results, given an effectively well-conditioned problem instance.

For $0 < \alpha_1 < \ldots < \alpha_n$, the matrix A is totally positive [6]. Thus its singular vectors have the sign oscillation property that $\mathbf{u}_k$ and $\mathbf{v}_k$ have $k-1$ changes in sign. In particular, the left singular vector $\mathbf{u}_n$ corresponding to the smallest singular value has the sign oscillation property (34). For most vectors $\mathbf{b}$ obeying (34) the projection $\|P_1\mathbf{b}\|/\|\mathbf{b}\|$ will be non-negligible, and hence K_e is small for this case. Thus there can exist Vandermonde solvers able to exploit the small effective condition number K_e of right-hand sides $\mathbf{b}$ with the sign property (34). Apparently the Bjorck-Pereyra algorithm is one such method.

5.2 The Fast Poisson Solver Example

Because the special perturbations in $\dot{A}$ assumed in (26) are observed in practice, The-

orem 2 applies to certain commonly used solution methods for linear systems. However, Theorem 2 does not apply universally. For instance, it is widely appreciated that Gaussian elimination does not necessarily solve a related system $\hat{A}\hat{\mathbf{x}} = \mathbf{b}$ with $\hat{A}$ satisfying the assumptions in (26), if we interpret ε_u, ε_v, ε_σ to be bounded in size by some polynomial function in n times the machine precision.

The class of Fourier transform methods applied to Poisson's equation on $n+2$ grid points with zero boundary conditions, namely Eq. (6) in section 2, has the special structure necessary to employ Theorem 2. It is well known that L has the eigen-decomposition

$$L = U^{-1}\,\Lambda\,U \in \mathbb{R}^{n\times n} \tag{35}$$

where

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \quad \lambda_j = -4(n+1)^2 \sin^2\left(\frac{\pi j}{n+1}\right), \quad 1 \le j \le n \tag{36}$$

and

$$U = (u_{ij})_{i,j=1}^{n}: \quad u_{ij} = \sin\left(\frac{ij\pi}{n+1}\right). \tag{37}$$

The fast transform methods use the *a priori* knowledge of the decomposition (35-37) to compute $\mathbf{u}$ as $\hat{U}^{-1}\hat{\Lambda}^{-1}\hat{U}\,\mathbf{f}$, where $\hat{U}^{-1}$, $\hat{\Lambda}^{-1}$, and $\hat{U}$ obey (26) with ε_u, ε_v, and ε_σ no more than approximately n times the unit roundoff error. The errors incurred lie in the floating representation of the factors and in the matrix vector multiplications. The matrix multiplications are accomplished by fast Fourier transforms, which are known to be quite stable, and the divisors $\{\hat{\sigma}_i\}_{i=1}^{n}$ are related to the $\{\sigma_i\}_{i=1}^{n}$ by $\|\sigma_i - \hat{\sigma}_i\| \le \varepsilon\|\sigma_i\|$ for $1 \le i \le n$, similar to the perturbations of Example 3.

We have computed solutions to the equation $L\mathbf{u} = \mathbf{f}$ for several choices of $\mathbf{f}$ and several numbers of mesh points m. Figure 4 shows the errors recorded in solving with a smooth right-hand side $f_k = \sin\left(\frac{2\pi k}{m+1}\right)$ by the fast sine method and by Gaussian elimination with

partial pivoting, and in solving a random right-hand side high in oscillatory components by banded Gaussian elimination (in both one and two dimensions). Our theory in section 4 predicts that the effective condition number for the fast sine solver operating on smooth right-hand sides should be much smaller than $K(L) = O(m^2)$.

The order growth of the observed error has two components, a polynomial in m (which seems well approximated by the average number of arithmetic operations per point) and the condition of the linear system being solved. In Fig. 4, the curves show error growth proportional to $m^{0.73}$, $m^{1.9}$, $m^{1.9}$, and $m^{3.9}$ for the four cases. Because the average numbers of arithmetic operations per point are, respectively, $O(\log m)$, $O(1)$, $O(1)$, and $O(m^2)$ as $m \to \infty$, the observed effective condition numbers of the solved systems are of orders $m^{0.73}$, $m^{1.9}$, $m^{1.9}$, and $m^{1.9}$, respectively. While the fast sine solver does not achieve $O(1)$ effective conditioning, it does perform substantially better than the tridiagonal Gaussian elimination solver on the smooth (sinusoidal) input. Furthermore, it is clear that effective well-conditioning of the problem instance is not employed by the Gaussian elimination solver, as both smooth and noisy right-hand sides yield high error growths. This example illustrates the benefits of using a solution method that can take advantage of a known matrix decomposition to exploit an effectively well-conditioned problem instance, in particular a fast Poisson solver composed entirely of fast transforms.

INSERT FIGURE 4

6. Concluding Remarks

In this paper, we have introduced the concept of an "effective" condition number for linear systems, which provides much sharper estimates of the solution's sensitivity to perturbations in $\mathbf{b}$. We derived an expression for K_e depending not only on the ratios of certain singular values of A but also on the projection of $\mathbf{b}$ onto the right singular vectors of A. We showed that for special $\mathbf{b}$s, the effective condition number could be much smaller than the usual condition number. A similar analysis was employed to demonstrate that certain pairs of nearby linear systems maintain the effective condition number. Moreover, algorithms exist which can take advantage of effective well-conditioning to produce unexpectedly accurate results. The usefulness of the concept was demonstrated in two applications.

The concept of effective condition numbers and the related analysis of nearby linear systems rigorously explain why one may obtain better numerical results than expected from $K(A)$ alone (*e.g.*, the fast Poisson solver example in sec. 5), provide an intuitive framework for applying this concept to other situations (*e.g.*, semiconductor simulation) and may lead one to investigate properties of other algorithms that take advantage of intrinsic stability in the problem instance.

References

1. A. Bjorck and V. Pereyra. *Solution of Vandermonde systems of equations*. Math. Comp. **24**. 893-903 (1970).

2. C. de Boor and H.-O. Kreiss. *On the condition number of linear systems associated with BVPs of ODEs*. SIAM J. Num. Anal. **23**(5). 936 939 (1986).

3. G. H. Golub and C. F. Van Loan. Matrix Computations, Johns Hopkins University Press. Baltimore (1985).

4. S. Hammarling and J. H. Wilkinson. *On linear systems arising from finite difference approximations to elliptic differential equations*. NPL report DNACS 34/80. National Physics Laboratory. England (1980).

5. N. J. Higham. *Error analysis of the Bjorck-Pereyra algorithms for solving Vandermonde systems*. Numerical Analysis Report No. 108. Department of Mathematics. Univ. of Manchester. Manchester, England (1985).

6. S. Karlin. Total Positivity. vol. I, Stanford University Press, Stanford (1968).

7. R. D. Skeel. *Scaling for numerical stability in Gaussian Elimination*. J. Assoc. Comput. Mach. **26**. 494-526 (1979).

8. G. W. Stewart. Introduction to Matrix Computations, Academic Press, New York (1973).

9. A. Van der Sluis. *Stability of solutions of linear algebraic systems*. Numer. Math. **14**. 246-251 (1970).

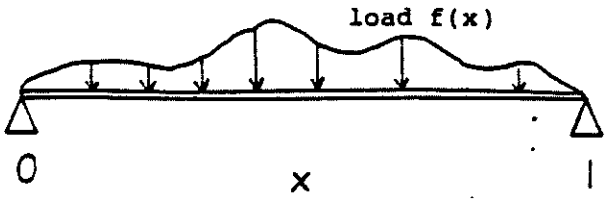
10. A. Van der Sluis. *Condition numbers and equilibration of matrices*. Numer. Math. **14**. 14-23 (1969).

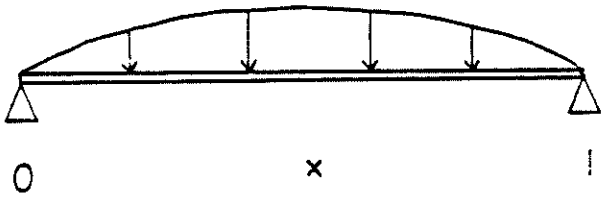
Fig. 1. Elastic beam under load.

Fig. 2. Beam under sinusoidal load $\sin \pi x$.

Fig. 3. Beam under sinusoidal load $\sin p\pi x$.

Fig. 4. Relative error in computing solutions to the Poisson equation $L\mathbf{x} = \mathbf{b}$ in single precision.


load f(x)
0
x
1


0
x
1

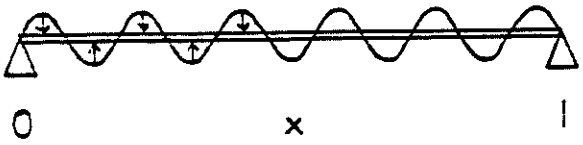
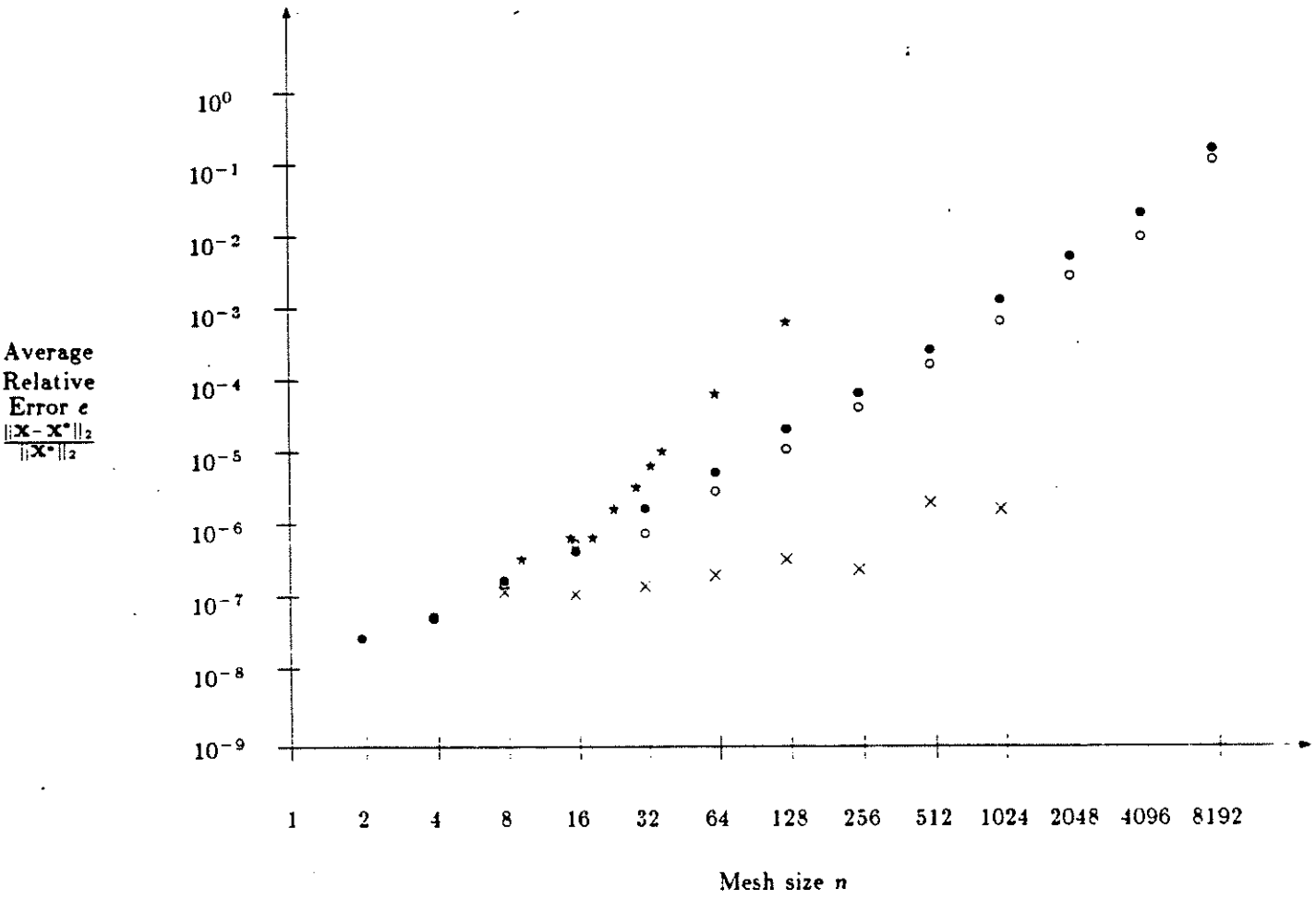

0
x
1

Fig. 4. Relative Error in Computing Solutions to the Poisson Equation $L\mathbf{x} = \mathbf{b}$ in Single Precision



- Smooth right-hand side b. Fast Sine Transform [1D], 1 repetition
- Smooth right-hand side b. SGTSL tridiagonal solver [1D], 1 repetition
- Random $\mathbf{x}^*$. SGTSL tridiagonal solver [1D], 20 repetitions
- Random $\mathbf{x}^*$. SGBSL banded solver [2D], 20 repetitions

A linear regression fit of the observed errors to the equation $\text{err} = \gamma n^c 2^{-24}$ yields the following coefficients.

Test case	γ	c	err($n = 1$)
·	$2.50 \cdot 10^{-1}$	0.73	$10^{-7.627}$
○	$3.39 \cdot 10^{-2}$	1.86	$10^{-8.606}$
•	$3.37 \cdot 10^{-2}$	1.92	$10^{-8.698}$
*	$1.14 \cdot 10^{-4}$	3.89	$10^{-11.167}$