Average Condition Number for Solving Linear Equations

N. Weiss
Department of Mathematics
Queens College of the City University of New York
Flushing, New York 11367

G. W. Wasilkowski* and H. Woźniakowski*
Department of Computer Science *and* *Institute of Informatics*
Columbia University in the City of New York *University of Warsaw*
New York, New York 10027 *Warsaw, Poland*

and

M. Shub*
IBM-Yorktown Heights, New York 10598
and the Graduate Center of the
City University of New York
Flushing, New York 11367

Submitted by G. W. Stewart

ABSTRACT

We study an average condition number and an average loss of precision for the solution of linear equations and prove that the average case is strongly related to the worst case. This holds if the perturbations of the matrix are measured in Frobenius or spectral norm or componentwise. In particular, for the Frobenius norm we show that one gains about $\log_2 n + 0.9$ bits on the average as compared to the worst case, n being the dimension of the system of linear equations.

1. INTRODUCTION

In this paper we analyze an average condition number of the solution of a linear system. We consider the numerical solution of $Ax = b$ in a floating

*This research was supported in part by the National Science Foundation under Grant DCR-82-14322 (G. W. W. and H. W.) and Grant MCS-82-01267 (M. S.).

52 Vanderbilt Ave., New York, NY 10017 0024-3795/86/$3.50

point arithmetic. Here A is an $n \times n$ real nonsingular matrix and b an $n \times 1$ nonzero vector. The coefficients of A and b may be known only to some order of precision due to measurement errors or rounding errors.

Even if A and b are known exactly, at best we can count in floating point arithmetic on computing the vector $\tilde{x}$ which is the exact solution of a slightly perturbed system,

$$(A+E)\tilde{x}=b, \tag{1.1}$$

where the matrix E is "small" relative to A. In fact, commonly used algorithms such as Gaussian elimination with pivoting or the Householder or Gram-Schmidt algorithm produce $\tilde{x}$ for which the matrix E satisfies

$$\|E\| \leqslant \rho\|A\| \tag{1.2}$$

for some norm $\|\cdot\|$ and for ρ which is usually a small multiple of the relative precision of floating point arithmetic; see e.g., [9]. It was shown in [6] that (1.2) can be improved by a few steps of iterative refinement. That is, one computes $\tilde{x}$ for which the matrix E satisfies

$$|e_{ij}| \leqslant \rho|a_{ij}| + O(\rho^2), \tag{1.3}$$

where e_{ij} and a_{ij} are the entries of E and A respectively. The inequalities (1.2) and (1.3) are reasonable hypothesis to make also if E represents measurement errors.

Let $x = A^{-1}$ be the exact solution. The error of $\tilde{x} = \tilde{x}(E)$ is given by

$$x - \tilde{x}(E) = (I + A^{-1}E)^{-1}A^{-1}Ex = A^{-1}Ex + O(\rho^2), \tag{1.4}$$

assuming that ρ is sufficiently small. We are interested in the error of the kth component of $x - \tilde{x}(E), k = 1,2,\dots,n,$ for E satisfying (1.2) or (1.3). For simplicity we drop $O(\rho^2)$ terms in (1.3) and (1.4), and estimate

$$|x - \tilde{x}(E)|_k \cong |A^{-1}Ex|_k, \tag{1.5}$$

where $|z|_k$ denotes the absolute value of the kth component of the vector z. The matrix E belongs to the set $\mathbf{E}$ which is defined by either (1.2) or (1.3). We note that

$$\frac{|A^{-1}Ex|_k}{\|x\|} \leqslant \rho \operatorname{cond}(A,k), \tag{1.6}$$

where $\operatorname{cond}(A,k)=(1/\rho\|x\|)\sup\{|A^{-1}Ex|_k\colon E\in \mathbf{E}\}$ is the (worst case) condition number of the matrix A. Note that $\operatorname{cond}(A,k)$ depends also on the vector x. For simplicity we do not list x as an argument of cond.

It is a common belief that the upper bound (1.6) is realistic for most matrices E; see [7, p. 195]. We prove that this is indeed the case by considering the average condition number. Specifically, let μ be Lebesgue measure in R^{n^2} normalized so that $\mu(E)=1$. Define the average condition number in L_p, $p\geqslant 1$, as

$$\operatorname{cond}^{\mathrm{avg}}(A,k,p)=\frac{1}{\rho\|x\|}\left\{\int_{\mathbf{E}}|A^{-1}Ex|_k^p\mu(dE)\right\}^{1/p}. \tag{1.7}$$

We show in Sections 2 to 4 that $\operatorname{cond}^{\mathrm{avg}}(A,k,p)$ is comparable to $\operatorname{cond}(A,k)$. This holds for all values of p and the set $\mathbf{E}$ defined by (1.2) for the Frobenius or spectral norm, as well as for $\mathbf{E}$ defined by (1.3). For instance, for the Frobenius norm we have

$$\begin{aligned}\frac{\operatorname{cond}^{\mathrm{avg}}(A,k,1)}{\operatorname{cond}(A,k)}&\cong\sqrt{\frac{2}{\pi}}\,\frac{1}{n},\\ \frac{\operatorname{cond}^{\mathrm{avg}}(A,k,2)}{\operatorname{cond}(A,k)}&=\frac{1}{\sqrt{n^2+2}}.\end{aligned} \tag{1.8}$$

This means that for modest n, the average condition number is roughly the same as the worst case one.

We now comment on the definition of the average condition number. Elements of the matrix E are regarded in (1.7) as uniformly distributed inside the ball (1.2) or (1.3). Clearly, the assumption about uniform distribution is unrealistic if E is fully deterministic and depends on coefficients of A and b, a specific algorithm used for the solution of a linear system as well as floating point arithmetic. In such a case, our results can be interpreted as saying that even a hypothetical assumption of uniform distribution of elements of E does not lead to a substantial gain, since the average condition number is comparable to the worst case one. On the other hand, one may argue that each individual rounding error resembles a random process with uniform distribution and, quoting Wilkinson [8, p. 25], "We may expect that the rounding errors in a computation will be more or less randomly distributed." In any case, uniform distribution is a crude assumption if E represents roundoff errors.

The situation changes if E represents measurement errors, especially if their bound is significantly larger than the relative precision of floating point arithmetic. Then it seems reasonable to assume that elements of E are independent and identically distributed, and uniform distribution is one of the possible distributions to be considered.

We also study the average loss of precision. Let $2^{\text{loss}} = |A^{-1}Ex|_k/\rho\|x\|$. Then[1]

$$\text{loss} = \text{loss}(E) = \log \frac{|A^{-1}Ex|_k}{\rho\|x\|}, \tag{1.9}$$

called the loss of precision, tells us how many bits are lost due to computational (or measurement) errors. Due to (1.6), the (worst case) loss of precision is

$$\text{loss}(A,k) = \log \text{cond}(A,k). \tag{1.10}$$

The average loss of precision is defined by

$$\text{loss}^{\text{avg}}(A,k) = \int_{\mathbf{E}} \log \frac{|A^{-1}Ex|_k}{\rho\|x\|}\mu(dE). \tag{1.11}$$

We prove that the average loss of precision is comparable to the worst case one. In particular, assuming the Frobenius norm in (1.2) we have

$$\text{loss}(A,k) - \text{loss}^{\text{avg}}(A,k) = \log n + 0.916 + O(n^{-2}). \tag{1.12}$$

As we mentioned before, the coefficients of b as well as A may not be known exactly. They are measured with some error, and instead of $Ax = b$ we have a perturbed system $(A+E)x = b - h$. The matrix E satisfies (1.2) or (1.3) with ρ depending on measurement errors. The vector h is small relative to b, i.e., $\|h\| \leqslant \eta\|b\|$ or $|h_k| \leqslant \eta|b_k|$, $k = 1,2,\dots,n$, for some small η. In Section 5 we show how the results of Sections 2 to 4 can be extended for the case when both A and b are perturbed.

Our paper is motivated by recent interesting work [1, 2, 4, 5] dealing with the average condition number $\|A^{-1}\|\cdot\|A\|$ or the average loss of precision $\log(\|A^{-1}\|\cdot\|A\|)$ of $n \times n$ random matrices. Assuming that all entries of A are independent random variables with standard normal distribution on the class of $n \times n$ real or complex matrices, it has been proved that the average

[1] Throughout, $\log \equiv \log_2$; $\ln \equiv \log_e$.

condition number is infinity for the real case and finite for the complex case. Along these lines one may analyze the average value of $\operatorname{cond}^{\mathrm{avg}}(A,k,p)$ over matrices A. Since $\operatorname{cond}^{\mathrm{avg}}(A,k,p)$ is proportional to $\operatorname{cond}(A,k)$, using the results mentioned above shows that the average value of $\operatorname{cond}^{\mathrm{avg}}(A,k,p)$ is infinity over real matrices A and finite if taken over complex matrices. In [5] it is shown that the average of $\ln(\|A^{-1}\|_2\|A\|_2)$ over $n\times n$ real matrices is between $(\frac{2}{3}-\varepsilon)\ln n$ and $(3+\varepsilon)\ln n$, where ε tends to zero if n goes to infinity. The upper bound was improved in [4] to $\frac{5}{2}\ln n+1$. Thus the gain in (1.12) is significant on going from the worst case to the average one.

2. FROBENIUS NORM

In this section we assume the Frobenius norm of matrices, $\|E\|=(\sum_{i,j=1}^{n}e_{ij}^2)^{1/2}$. The norm of a vector z is given correspondingly by $\|z\|=(\sum_{i=1}^{n}z_i^2)^{1/2}$. We assume that $\mathbf{E}$ has now the form

$$\mathbf{E}=\{E:\|E\|\leqslant\rho\|A\|\},\tag{2.1}$$

and the measure μ of a Borel subset B of $\mathbb{R}^{n^2}$ is given by

$$\mu(B)=\frac{\lambda(B\cap\mathbf{E})}{\lambda(\mathbf{E})},\tag{2.2}$$

where λ is standard Lebesgue measure on $\mathbb{R}^{n^2}$. By A_k^{-1} we denote the kth row of A^{-1}. The worst case condition number is now given by

$$\operatorname{cond}(A,k)=\|A_k^{-1}\|\cdot\|A\|.$$

THEOREM 2.1. *We have*

$$\operatorname{cond}^{\mathrm{avg}}(A,k,p)=a_{n,p}\operatorname{cond}(A,k),\tag{2.3}$$

where

$$a_{n,p}=\sqrt[p]{\frac{\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{n^2}{2}+1\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n^2+p}{2}+1\right)}},$$

and

$$\operatorname{loss}^{\text{avg}}(A,k)=\operatorname{loss}(A,k)-a_n, \tag{2.4}$$

where

$$a_n=\begin{cases}\dfrac{1}{\ln 2}\left(\ln 2+\dfrac{1}{2}+\dfrac{1}{4}+\cdots+\dfrac{1}{n^2}\right) & \text{if } n \text{ is even,}\\[2ex] \dfrac{1}{\ln 2}\left(1+\dfrac{1}{3}+\dfrac{1}{5}+\cdots+\dfrac{1}{n^2}\right) & \text{if } n \text{ is odd.}\end{cases}$$

Proof. We need to compute

$$a=\int_{\mathbf{E}} f(|A^{-1}Ex|_k)\mu(dE)$$

for $f(u)=u^p$ $(p\geqslant 1)$ and $f(u)=\log u$. Note that $(A\ {}^1Ex)_k=\sum_{i,j=1}^n m_{ki}e_{ij}x_j$, where m_{ki} are the elements of the kth row of A^{-1}. Let $t_{ij}=e_{ij}/(\rho\|A\|)$ and $y_{ij}=m_{ki}x_j$. Let t and y be the $n^2\times 1$ vectors with components t_{ij} and y_{ij}. Then

$$a=\frac{1}{\lambda(B)}\int_B f(\rho\|A\|\cdot|(y,t)|)\,dt,$$

where B is the unit ball with center zero in $\mathbb{R}^{n^2}$ and $\lambda(B)$ its Lebesgue measure.

Take an orthogonal matrix Q such that $Qy=\|y\|[1,0,\ldots,0]^T$. Note that $\|y\|=\|A_k^{-1}\|\cdot\|x\|$. Change variables by setting $u=Qt$. Since the Lebesgue measure is invariant under orthogonal transformations and $QB=B$, we have $du=dt$ and $(y,t)=\|y\|u_{11}$. Thus for $c=\rho\|A_k^{-1}\|\,\|A\|\,\|x\|$ we have

$$\begin{aligned} a&=\frac{1}{\lambda(B)}\int_B f(c|u_{11}|)\,du\\ &=\frac{1}{\lambda(B)}\int_{-1}^{1} f(c|t|)\operatorname{Vol}(n^2-1,\sqrt{1-t^2})\,dt, \end{aligned} \tag{2.5}$$

where $\operatorname{Vol}(n^2-1,\sqrt{1-t^2})$ denotes the volume of the ball in $\mathbb{R}^{n^2-1}$ with radius $\sqrt{1-t^2}$. Since

$$\lambda(B)=\int_{-1}^{+1}\mathrm{Vol}(n^2-1,\sqrt{1-t^2})\,dt$$

and $\mathrm{Vol}(n^2-1,\sqrt{1-t^2})$ is proportional to $(1-t^2)^{(n^2-1)/2}$, we have

$$a=\frac{\int_0^1 f(ct)(1-t^2)^{(n^2-1)/2}\,dt}{\int_0^1 (1-t^2)^{(n^2-1)/2}\,dt}.$$

If $f(t)=t^p$, then [3, 3.251] yields

$$a=c^p\frac{\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{n^2}{2}+1\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n^2+p}{2}+1\right)}=c^pa^p_{n,p},$$

which implies (2.3). If $f(t)=\log t$, then [3, 4.246, 4.253] yields

$$a=\log c-a_n,$$

which implies (2.4). ■

Consider now the constants $a_{n,p}$ from (2.3). It is immediate that $a_{n,2}=1/\sqrt{n^2+2}$. For arbitrary p, it follows from Stirling's formula that

$$a_{n,p}\cong\frac{\sqrt{2}}{n}\left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right)^{1/p}\qquad\text{if}\quad n^2\gg p,$$

so, in particular, $a_{n,1}\cong\sqrt{2/\pi}\,1/n$. Also, $\lim_{p\to\infty}a_{n,p}=1$ for each n.

As for a_n in (2.4), let

$$\gamma=\lim_{i\to\infty}\left(1+\frac{1}{2}+\cdots+\frac{1}{i}-\ln i\right)=0.577\ldots$$

be Euler's constant. It is easy to check that

$$a_n = \log n + \frac{\gamma + \ln 2}{2 \ln 2} + O(n^{-2})$$

$$= \log n + 0.916\ldots + O(n^{-2}). \tag{2.6}$$

Theorem 2.1 states that the average and worst case condition numbers and losses of precision are related. For instance, from (2.4) and (2.6) it follows that one gains roughly $\log n + 0.9$ bits, on the average, over the worst case. Specifically, if $n = 32$ and $\|A_k^{-1}\|\,\|A\| = 2^{10}$, then one loses 10 bits in the worst case and about 4.1 on the average.

3. COMPONENTWISE PERTURBATIONS

In this section we assume that the matrix $E = (e_{ij})$ satisfies (1.3). That is, let

$$\mathbf{E} = \left\{ E = (e_{ij}) : |e_{ij}| \leqslant \rho |a_{ij}| \right\}. \tag{3.1}$$

Without loss of generality assume that $a_{ij} \neq 0$ for all i, j. The measure μ is now given by

$$\mu(B) = \left(\prod_{i,j=1}^{n} 2\rho |a_{ij}| \right)^{-1} \lambda(B \cap \mathbf{E}), \tag{3.2}$$

where B is a Borel set in $\mathbb{R}^{n^2}$ and λ is Lebesgue measure.

It is easy to check that the worst case condition number cond(A, k) for $\mathbf{E}$ given by (3.1) is

$$\operatorname{cond}(A, k) = \frac{(|A^{-1}|\,|A|\,|x|)_k}{\|x\|}, \tag{3.3}$$

where $|x|$ denotes the vector with components $|x_i|$, while $|A|, |A^{-1}|$ denote matrices with elements $|a_{ij}|$ and $|m_{ij}|$, and $A^{-1} = (m_{ij})$.

THEOREM 3.1. *We have*

$$\operatorname{cond}^{\mathrm{avg}}(A,k,p)=a_{n,p}\operatorname{cond}(A,k), \tag{3.4}$$

where

$$\left[(1+\varepsilon_{n,p})\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right]^{1/p}\sqrt{\frac{2}{3}}\,\frac{1}{n}\leqslant a_{n,p}\leqslant\frac{1}{(p+1)^{1/p}}$$

with $\varepsilon_{n,2}=0$ *and* $\lim_{n\to\infty}\varepsilon_{n,p}=0$ *for all* p; *and*

$$\operatorname{loss}^{\mathrm{avg}}(A,k)=\operatorname{loss}(A,k)-a_n, \tag{3.5}$$

where

$$\frac{1}{\ln 2}\leqslant a_n\leqslant\log n+\frac{1+\varepsilon_n}{2}\left(\log 6+\frac{\gamma}{\ln 2}\right)$$

with $\lim_{n\to\infty}\varepsilon_n=0$. *Here* $\gamma=0.577\ldots$ *is Euler's constant and* $\frac{1}{2}(\log 6+\gamma/\ln 2)=1.708\ldots$.

Proof. We need to estimate

$$a=\int_{\mathbf{E}}f(|A^{-1}Ex|_k)\mu(dE)$$

for $f(u)=u^p$ and $f(u)=\log u$. Let $t_{ij}=e_{ij}/(\rho|a_{ij}|)$ and $y_{ij}=\rho m_{ki}|a_{ij}|x_j$. Let t and y be the $n^2\times 1$ vectors with components t_{ij} and y_{ij}. Then $|A^{-1}Ex|_k=|(t,y)|$. Since the Lebesgue measure is symmetric, we have

$$a=2^{-n^2}\int_{[-1,1]^{n^2}}f(|\;(t,|y|)\;|)\,dt. \tag{3.6}$$

To estimate a, we need the following

LEMMA 3.1. *Let* λ *be Lebesgue measure on* $\mathbb{R}^N$. *Let* $y=[y_1,y_2,\ldots,y_N]\in\mathbb{R}^N_+$ *and* $Y=\sum_{i=1}^N y_i>0$. *For* $u\in\mathbb{R}$ *let*

$$F(u;y)=\lambda\left(\left\{t\in[-1,1]^N:(y,t)>u\right\}\right).$$

Then for all $u \in \mathbb{R}_+$,

$$F(u; y^*) \leqslant F(u; y) \leqslant F(u; y^{**}),$$

where $y^* = (Y/N)[1, 1, \ldots, 1]$ *and* $y^{**} = [Y, 0, \ldots, 0]$.

Proof. The case $N = 2$ can be verified directly. Suppose inductively that $N > 2$ and Lemma 3.1 holds for $N - 1$. We first prove that

$$F(u; [\bar{y}, \bar{y}, \ldots, \bar{y}, y_N]) \leqslant F(u; y) \leqslant F(u; [(N-1)\bar{y}, 0, \ldots, 0, y_N]), \quad (3.7)$$

where $\bar{y} = [1/(N-1)]\Sigma_{i=1}^{N-1} y_i$. Note that

$$F(u; y) = \int_{-1}^{+1} F(u - y_N t_N; y')\, dt_N, \quad (3.8)$$

where $y' = (y_1, \ldots, y_{N-1})$ and $y = (y', y_N)$. Thus if $u \geqslant y_N$, then (3.7) follows immediately from the inductive hypothesis. Suppose therefore that $u < y_N$. Then (3.8) can be rewritten as

$$F(u; y) = \int_{-1}^{u/y_N} F(u - y_N t_N; y')\, dt_N + \int_{u/y_N}^{1} F(u - y_N t_N; y')\, dt_N.$$

For $t \in [u/y_N, 1]$,

$$F(u - y_N t_N; y') = \lambda_{N-1}(\{t' \in [-1, 1]^{N-1} : (t', y') > u - y_N t_N\})$$

$$= \lambda_{N-1}(-\{t' \in [-1, 1]^{N-1} : (t', y') < y_N t_N - u\}),$$

where λ_{N-1} is Lebesgue measure on $\mathbb{R}^{N-1}$. Since λ_{N-1} is symmetric and $\lambda_{N-1}([-1, 1]^{N-1}) = 2^{N-1}$, we get $F(u - y_N t_N; y') = 2^{N-1} - F(y_N t_N - u; y')$. Thus

$$F(u; y) = \int_{-1}^{u/y_N} F(u - y_N t_N; y')\, dt_N$$

$$- \int_{u/y_N}^{1} F(y_N t_N - u; y')\, dt_N + 2^{N-1}\left(1 - \frac{u}{y_N}\right).$$

Change the variable t_N in the second integral by setting $x = -t_N + 2u/y_N$. Then

$$\begin{aligned} F(u;y) &= \int_{-1}^{u/y_N} F(u - y_N t_N; y')\,dt_N \\ &\quad - \int_{-1+2u/y_N}^{u/y_N} F(u - y_N x; y')\,dx + 2^{N-1}\left(1 - \frac{u}{y_N}\right) \\ &= \int_{-1}^{-1+2u/y_N} F(u - y_N t_N; y')\,dt_N + 2^{N-1}\left(1 - \frac{u}{y_N}\right). \qquad (3.9) \end{aligned}$$

Note that for every $t_N \in [-1, -1+2u/y_N]$, $u - y_N t_N \geqslant 0$. Thus we can apply the inductive hypothesis to (3.9) to get

$$\begin{aligned} F(u;y) &\geqslant \int_{-1}^{-1+2u/y_N} F(u - y_N t_N; [\bar{y},\ldots,\bar{y}])\,dt_N + 2^{N-1}\left(1 - \frac{u}{y_N}\right) \\ &= F(u; [\bar{y}, \bar{y},\ldots,\bar{y}, y_N]), \end{aligned}$$

and

$$\begin{aligned} F(u;y) &\leqslant \int_{-1}^{-1+2u/y_N} F(u - y_N t_N;[(N-1)\bar{y},0,\ldots,0])\,dt_N + 2^{N-1}\left(1 - \frac{u}{y_N}\right) \\ &= F(u;[(N-1)\bar{y},0,\ldots,0,y_N]). \end{aligned}$$

This completes the proof of (3.7). Thus the point y^* at which $F(u;\cdot)$ takes its minimum has its first $N-1$ components equal to each other. Since $F(u;y)$ does not depend on the permutations of the components of y and since $N-1 \geqslant 2$, all the components of y^* are equal.

To prove that $F(u;\cdot)$ takes its maximum at y^{**}, permute components in (3.7) to get

$$F(u;y) \leqslant F(u;[(N-1)\bar{y}, y_N,0,\ldots,0]). \qquad (3.10)$$

Using (3.7) on the right-hand side of (3.10), we have

$$F(u;y) \leqslant F(u;[Y,0,\ldots,0]) = F(u;y^{**})$$

as claimed. ■

COROLLARY 3.1. *Let y, y^*, and y^{**} be as in Lemma* 3.1. *Suppose that f is continuous and increasing on* $(0,+\infty)$. *Set* $\Phi(y) = \int_{[-1,1]^N} f(|(y,t)|)\,dt$. *Then*

$$\Phi(y^*) \leqslant \Phi(y) \leqslant \Phi(y^{**}).$$

Proof. For $j = 1,2,\ldots,$ and $i = 0,1,\ldots,\lceil Y2^j \rceil$, $Y = \sum_{i=1}^N y_i$, set

$$f_{j,i} = \begin{cases} 1 & \text{if } \ i2^{-j} \leqslant |x|, \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $f = f_{j,i}$ then $\Phi(y) = 2F(i2^{-j}; y)$. For the given function f define

$$f_j = f(2^{-j})f_{j,0} + \sum_{i=1}^{\lceil Y2^j \rceil} \left[f((i+1)2^{-j}) - f(i2^{-j})\right] f_{j,i}.$$

Since f is increasing, $f_j(x)$ decreases to $f(x)$ a.e. as $j \to +\infty$, $|x| \leqslant Y$. Corollary 3.1 is thus a consequence of Lemma 3.1 and the monotone convergence theorem. ■

We obtain bounds for the a in (3.6) by applying Corollary 3.1 with $N = n^2$. Let $c = \sum_{i,j=1}^n |y_{i,j}|$, so that $c = \rho\|x\|\operatorname{cond}(A,k)$. For $f(u) = u^p$ we have

$$a = 2^{-n^2}\int_{[-1,1]^{n^2}} \left(\sum_{i,j=1}^{n} |y_{ij}|\,|t_{ij}| \right)^p dt$$

$$\leqslant c^p \frac{1}{2}\int_{-1}^{1} |t_{11}|^p\, dt_{11} = \frac{1}{p+1} c^p,$$

which yields $a_{n,p} \leqslant 1/(p+1)^{1/p}$ in (3.4). For $f(u) = \log u$, by the same reasoning,

$$a \leqslant \log c + \frac{1}{2}\int_{-1}^{1} \log|t_{11}|\, dt_{11} = \log c - \frac{1}{\ln 2}$$

which yields $a_n \geqslant 1/\ln 2$ in (3.5).

Turning back to the case $f(x)=x^p$, we again apply Corollary 3.1 to get

$$a \geqslant c^p 2^{-n^2} \int_{[-1,1]^{n^2}} \left| \frac{1}{n^2} \sum_{i,j=1}^{n} t_{ij} \right|^p dt.$$

The variables t_{ij} can be treated as independent, identically distributed random variables with mean zero and variance $\frac{1}{3}$. The central limit theorem implies that $z_n(t)=(1/n^2)\sum_{i,j=1}^n t_{ij}$ has a distribution on $[-1,1]^{n^2}$ with respect to $2^{-n^2}dt$ which approaches the normal distribution with mean zero and variance $\sigma = 1/(3n^2)$. Therefore

$$\begin{aligned} 2^{-n^2}\int_{[-1,1]^{n^2}} |z_n(t)|^p\, dt &= (1+\varepsilon_{n,p}) \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{+\infty} |z|^p e^{-z^2/(2\sigma)}\, dz \\ &= (1+\varepsilon_{n,p}) \frac{(2\sigma)^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \\ &= (1+\varepsilon_{n,p}) \frac{1}{\sqrt{\pi}} \left(\frac{2}{3}\right)^{p/2} \frac{1}{n^p} \Gamma\left(\frac{p+1}{2}\right), \end{aligned}$$

where $\lim_{n\to\infty} \varepsilon_{n,p} = 0$ and, of course, $\varepsilon_{n,2}=0$. This proves the first inequality in (3.4). If we now set $f(u)=\log u$, then Corollary 3.1 yields

$$a \geqslant \log c + 2^{-n^2} \int_{[-1,1]^{n^2}} \log\left(\frac{1}{n^2} \left| \sum_{i,j=1}^{n} t_{ij} \right| \right) dt.$$

Applying the central limit theorem a second time, we see that

$$a \geqslant \log c + (1+\varepsilon_n) \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{+\infty} \log|z|\, e^{-z^2/(2\sigma)}\, dz,$$

where $\lim_{n\to\infty}\varepsilon_n = 0$. The last integral can be rewritten as

$$\begin{aligned} \frac{2}{\sqrt{\pi}} \int_0^\infty \log(\sqrt{2\sigma}\ t)\, e^{-t^2}\, dt &= \tfrac{1}{2}\log\tfrac{2}{3} - \log n \\ &\quad + \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-t^2} \log t\, dt \\ &= -\log n + \tfrac{1}{2}\log\tfrac{2}{3} - \frac{\gamma}{2\ln 2} - 1 \end{aligned}$$

due to [3, 4.333]. This proves the lower bound on $-a_n$ and completes the proof. ■

REMARK 3.1. Note that recourse to the central limit theorem can be avoided by an explicit formula for

$$\int_{[-1,1]^{n^2}} f(|z_n(t)|)\,dt.$$

Specifically, if f_n is determined by requiring that $f_n^{(n^2)} = f$ and $f_n^{(i)}(0) = 0$, $i = 0, 1, \ldots, n^2 - 1$, then repeated integration yields

$$\int_{[-1,1]^{n^2}} f(|z_n(t)|)\,dt = 2(n^2)^{n^2} \sum_{j=0}^{\lfloor n^2/2 \rfloor} \binom{n^2}{j}(-1)^j f_n\left(1 - \frac{2j}{n^2}\right). \quad (3.11)$$

For example, if $f(x) = \log x$ then

$$f_n(x) = \frac{x^{n^2}}{(n^2)!}\left[\log x - \frac{1}{\ln 2}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n^2}\right)\right].$$

By using (3.11) with $f(x) = x$ and $f(x) = \log x$, for instance, we determined that in (3.4), $\varepsilon_{n,1} > 0$ for small n and already $\varepsilon_{2,1} < 0.014$, and that in (3.5), $\varepsilon_n < 0$ for small n and $|\varepsilon_2| < 0.014$.

For $f(x) = x^2$ we have the exact formula for a in (3.6),

$$a = \sum_{i,j=1}^{n} |y_{ij}|^2 2^{-n^2} \int_{[-1,1]^{n^2}} t_{ij}^2\,dt_{ij} = \frac{1}{3} \sum_{i,j=1}^{n} |y_{ij}|^2.$$

The best one can do with this is to obtain the bounds in (3.4),

$$\frac{1}{\sqrt{3}\,n} \leqslant a_{n,2} \leqslant \frac{1}{\sqrt{3}}. \qquad ■$$

We conclude this section by showing that the bounds in Theorem 3.1 are sharp. From the proof of the theorem, it is clear that the average condition number and average loss will be close to the given upper bounds whenever there exists (i_0, j_0) such that $|y_{i_0 j_0}| \gg |y_{ij}|$ if $(i, j) \neq (i_0, j_0)$, where $y_{ij} = m_{ki} a_{ij} x_j$ as usual. This is easy enough to arrange. One can take $x = [1, 0, \ldots, 0]$, $a_{11} = 1$, and $|a_{ij}| \ll 1$ if $i \neq 1$.

Similarly, the average condition number and loss will be close to the given lower bounds if all but a few $|y_{ij}|$ are equal to each other, and the rest are smaller. Letting $x=[1,\dots,1]$ and noting that $(m_{ki})=A_k^{-1}=r^{-1}[1,\dots,1]$ whenever

$$\sum_{i=1}^{n} a_{ij}=\begin{cases}0, & j\neq k,\\ r, & j=k,\end{cases} \tag{3.12}$$

we see that it will suffice to find $A=(a_{ij})$ which is non-singular, satisfies (3.12), and is such that $|a_{ij}|=1$ for most (i,j) and $a_{ij}=0$ otherwise.

We now find such a matrix. For any m, let $B_m=(b_{ij})$, $C_m=(c_{ij})$ be the $m\times m$ matrices given by

$$b_{ij}=\begin{cases}1, & i\geqslant j,\\ -1, & i<j,\end{cases}\qquad c_{ij}=\begin{cases}2, & j=i,\\ -2, & j=i+1,\\ 0 & \text{otherwise.}\end{cases}$$

Note that B_m and C_m are nonsingular. If $n=2m$, set

$$A_n=\left[\begin{array}{c|c}-B_m+C_m & B_m\\ \hline B_m & -B_m\end{array}\right].$$

Then $\det A_n=\det C_m \det(-B_m)\neq 0$; A_n satisfies (3.12) with $k=1$, $r=2$; and $|a_{ij}|=1$ for all (i,j). If n is odd, modify A_{n-1} to $\tilde{A}_{n-1}$ by changing $a_{11}=1$ to $\tilde{a}_{11}=0$, and set

$$A_n=\left[\begin{array}{c|c}\tilde{A}_{n-1} & c\\ \hline -1,0,\dots,0 & 1\end{array}\right],$$

where c is the last column of $\tilde{A}_{n-1}$. Then $\det A_n=\det\tilde{A}_{n-1}\neq 0$; A_n satisfies (3.12) with $k=n$, $r=1$; and $|a_{ij}|=1$ except that $a_{11}=0$, $a_{nj}=0$ for $j=2,3,\dots,n-1$.

4. SPECTRAL NORM

In this section we assume the spectral norm on matrices, $\|E\|=\sup\{\|Ex\|/\|x\|: \|x\|\neq 0\}$. **E**, μ, and A_k^{-1} are as in Section 2, and it is easy to see that again

$$\operatorname{cond}(A,k)=\|A_k^{-1}\|\,\|A\|.$$

If now $B=\{E:\|E\|\leqslant 1\}$, then the first equality in (2.5) continues to hold, but we have been unable to calculate for $t\in(0,1)$ the n^2-1 dimensional volume of the set of matrices in B with the fixed element $a_{11} = t$. Consequently our estimates are not as sharp as in Section 2.

THEOREM 4.1. *We have*

$$\operatorname{cond}^{\text{avg}}(A, k, p) = a_{n,p} \operatorname{cond}(A, k), \tag{4.1}$$

where

$$\left(\frac{n^2}{n^2 + p} b_{n,p}^2 \right)^{1/p} \leqslant a_{n,p} \leqslant \left(\frac{n^2}{n^2 + p} b_{n,p} \right)^{1/p}$$

with

$$b_{n,p} = \frac{\Gamma\left(\dfrac{p+1}{2} \right) \Gamma\left(\dfrac{n}{2} \right)}{\sqrt{\pi}\, \Gamma\left(\dfrac{n+p}{2} \right)},$$

and

$$\operatorname{loss}^{\text{avg}}(A, k) = \operatorname{loss}(A, k) - a_n, \tag{4.2}$$

where

$$b_n + \frac{1}{n^2 \ln 2} \leqslant a_n \leqslant 2b_n + \frac{1}{n^2 \ln 2}$$

with

$$b_n = \begin{cases} \dfrac{1}{\ln 2} \left(\ln 2 + \dfrac{1}{2} + \dfrac{1}{4} + \cdots + \dfrac{1}{n-2} \right) & \text{if } n \text{ is even,} \\ \dfrac{1}{\ln 2} \left(1 + \dfrac{1}{3} + \cdots + \dfrac{1}{n-2} \right) & \text{if } n \text{ is odd.} \end{cases} \qquad \blacksquare$$

As in Section 2, we have an exact value for $b_{n,2}$ and asymptotic results otherwise: $b_{n,2} = 1/\sqrt{n}$, $b_{n,1} \cong \sqrt{2/(n\pi)}$,

$$b_{np} \cong \sqrt{\frac{2}{n}} \left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}} \right)^{1/p} \qquad \text{if} \quad n^2 \gg p,$$

and

$$\begin{aligned} b_n &= \tfrac{1}{2}\log n + \frac{\gamma + \ln 2}{2\ln 2} + O(n^{-1}) \\ &= \tfrac{1}{2}\log n + 0.916\ldots + O(n^{-1}). \end{aligned}$$

Proof of Theorem 4.1. As before, we can write

$$a = \int_{\mathrm{E}} f(|A^{-1}Ex|_k)\mu(dE) = \frac{1}{\lambda(B)}\int_B f(\rho\|A\|\,|A^{-1}Yx|_k)\,dY.$$

Note now that $(A^{-1}Yx)_k = (A_k^{-1}, Yx)$ and that for fixed U in O_n, the group of orthogonal matrices, the map $Y \to UY$ preserves both Lebesgue measure dY and the ball B. Thus, we can write

$$a = \frac{1}{\lambda(B)}\int_B f\big(\rho\|A\|\big|(A_k^{-1}, UYx)\big|\big)\,dY. \tag{4.3}$$

Now let ν be Haar measure on O_n, normalized so that $\nu(\mathrm{O}_n) = 1$. Integrating over O_n both sides of (4.3) and then reversing the order of integration on the right, we see that

$$a = \frac{1}{\lambda(B)}\int_B\left[\int_{\mathrm{O}_n} f\big(\rho\|A\|\cdot\big|(A_k^{-1}, UYx)\big|\big)\nu(dU)\right]dY.$$

On the other hand, for any $r > 0$ and $\xi, \eta \in \mathbb{R}^n$,

$$\int_{\mathrm{O}_n} f(r|(\xi, U\eta)|)\nu(dU) = \frac{1}{\sigma(\Sigma)}\int_\Sigma f(r\|\xi\|\cdot\|\eta\|\cdot|\zeta_1|)\sigma(d\zeta),$$

where $\Sigma = \{\zeta = [\zeta_1, \ldots, \zeta_n] : \|\zeta\| = 1\}$ and σ is the usual measure on Σ. And so

$$a = \frac{1}{\lambda(B)}\int_B\left[\frac{1}{\sigma(\Sigma)}\int_\Sigma f(b\|Yx\|\,|\zeta_1|)\sigma(d\zeta)\right]dY, \tag{4.4}$$

with $b = \rho\|A\|\cdot\|A_k^{-1}\|$.

To obtain a lower bound for a, we notice that for each matrix Y, there is a unit vector ξ_Y with the property that $\|Y\eta\| \geqslant \|Y\| \cdot |(\xi_Y, \eta)|$ for each $\eta \in \mathbb{R}^n$. (In fact, if $Y = PW$, with P positive definite and W orthogonal, we can take ξ_Y to be $W^{-1}\xi_P$, where ξ_P is a unit eigenvector for the largest eigenvalue of P.) From this, since f is in any case an increasing function, it follows that

$$\begin{aligned}\int_B f(b\|Yx\|)\,dY &= \int_{\mathrm{O}_n}\left[\int_B f(b\|YVx\|)\,dY\right]\nu(dV)\\ &\geqslant \int_{\mathrm{O}_n}\left[\int_B f(b\|Y\|\cdot|(\xi_Y, Vx)|)\,dY\right]\nu(dV)\\ &= \int_B\left[\frac{1}{\sigma(\Sigma)}\int_\Sigma f(b\|Y\|\cdot\|x\|\cdot|\zeta_1|)\,d\zeta\right]dY, \qquad (4.5)\end{aligned}$$

where the last equality comes from reversing the order of integration in the previous integral and arguing as we did to establish (4.4).

A simple observation will lead to an upper bound for a:

$$\int_B f(b\|Yx\|)\,dY \leqslant \int_B f(b\|Y\|\,\|x\|)\,dY. \qquad (4.6)$$

Suppose now that $f(t) = t^p$. Combining (4.4), (4.5), and (4.6), we see that

$$b^2_{n,p} I_{n,p} c^p \leqslant a \leqslant b_{n,p} I_{n,p} c^p, \qquad (4.7)$$

where

$$b_{n,p} = \frac{1}{\sigma(\Sigma)}\int_\Sigma |\zeta_1|^p \sigma(d\zeta) = \frac{\int_0^1 t^p(1-t^2)^{(n-3)/2}\,dt}{\int_0^1 (1-t^2)^{(n-3)/2}\,dt},$$

$$I_{n,p} = \frac{1}{\lambda(B)}\int_B \|Y\|^p\,dY \quad \text{and} \quad c = b\|x\| = \rho\|A\|\,\|A_k^{-1}\|\,\|x\|.$$

The explicit formula for $b_{n,p}$ follows as in Section 2. As for $I_{n,p}$, if $Y = rY'$, $r > 0$, $\|Y'\| = 1$, then $dY = r^{n^2-1}\,dr\,\alpha(dY')$, where α is an appropriate measure on $\{Y' : \|Y'\| = 1\}$, so $[1/\lambda(B)]\int_B \|Y\|^p\,dY = \int_0^1 r^{p+n^2-1}dr/\int_0^1 r^{n^2-1}dr = n^2/(n^2+p)$. Taking pth roots in (4.7), we have (4.1).

Similarly, combining (4.4), (4.5), and (4.6) when $f(t) = \log t$ leads to

$$-2b_n - I_n + \log c \leqslant a \leqslant -b_n - I_n + \log c, \tag{4.8}$$

where

$$b_n = -\frac{1}{\sigma(\Sigma)}\int_\Sigma \log(|\zeta_1|)\,\sigma(d\zeta_1), \qquad I_n = -\frac{1}{\lambda(B)}\int_B \log\|Y\|\,dY,$$

and (4.8) yields (4.2) as (4.7) did (4.1). ■

5. PERTURBATIONS OF A AND b

We now indicate how the results of Sections 2 to 4 can be extended for perturbations of A and b. We begin with the case when the matrix A is unperturbed. That is, consider the perturbed system

$$A\tilde{x} = b - h,$$

where the vector h satisfies

$$\|h\| \leqslant \eta\|b\| \tag{5.1}$$

or

$$|h_i| \leqslant \eta|b_i|, \qquad i = 1,2,\dots,n, \tag{5.2}$$

for some small η. Since Frobenius and spectral norms of vectors are the same, we need to consider now only these two cases. We have $\tilde{x} = \tilde{x}(h)$ and

$$\frac{|x - \tilde{x}(h)|_k}{\|x\|} = \frac{|A^{-1}h|_k}{\|x\|} \leqslant \eta\,\mathrm{cond}(b,k),$$

where the worst case condition number is now equal to

$$\mathrm{cond}(b,k) = \begin{cases} \dfrac{\|A_k^{-1}\|\,\|Ax\|}{\|x\|} & \text{if } h \text{ satisfies (5.1)},\\[2ex] \dfrac{(|A^{-1}||Ax|)_k}{\|x\|} & \text{if } h \text{ satisfies (5.2)}. \end{cases}$$

The average condition number $\operatorname{cond}^{\mathrm{avg}}(b,k,p)$ and the average loss $\operatorname{loss}^{\mathrm{avg}}(b,k)$ of precision are defined as in (1.7) and (1.11). Since $|A^{-1}h|_k$ is of the same form as $|A^{-1}Ex|_k$, the results of Sections 2 and 3 go through with n replaced by $\sqrt{n}$.

For the norm perturbations we have

$$\operatorname{cond}^{\mathrm{avg}}(b,k,p)=a_{n,p}\operatorname{cond}(b,k)$$

with

$$a_{n,p}=\sqrt[p]{\frac{\Gamma\left(\dfrac{p+1}{2}\right)\Gamma\left(\dfrac{n}{2}+1\right)}{\sqrt{\pi}\,\Gamma\left(\dfrac{n+p}{2}+1\right)}},$$

$$\operatorname{loss}^{\mathrm{avg}}(b,k)=\operatorname{loss}(b,k)-a_n$$

with $a_n=(\ln 2+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{n})/\ln 2$ for n even and $a_n=(1+\frac{1}{3}+\cdots+\frac{1}{n})/\ln 2$ for n odd.

For the componentwise perturbations we have

$$\operatorname{cond}^{\mathrm{avg}}(b,k,p)=a_{n,p}\operatorname{cond}(b,k)$$

with

$$(1+\varepsilon_{n,p})\left(\frac{\Gamma\left(\dfrac{p+1}{2}\right)}{\sqrt{\pi}}\right)^{1/p}\sqrt{\frac{2}{3}}\,\frac{1}{\sqrt{n}}\leqslant a_{n,p}\leqslant\frac{1}{(p+1)^{1/p}},$$

where $\lim_{n\to\infty}\varepsilon_{n,p}=0$ for all p; and

$$\operatorname{loss}^{\mathrm{avg}}(b,k)=\operatorname{loss}(b,k)-a_n$$

with

$$\frac{1}{\ln 2}\leqslant a_n\leqslant\tfrac{1}{2}\log n+\frac{1+\varepsilon_n}{2}\left(\log 6+\frac{\gamma}{\ln 2}\right),$$

where $\lim_{n\to\infty}\varepsilon_n=0$.

We now consider the general case where both A and b are perturbed, i.e.,

$$(A+E)\tilde{x}=b-h.$$

Then $\tilde{x}=\tilde{x}(E,h)$, and dropping the second order term, we have

$$x-\tilde{x}(E,h)\cong A^{-1}Ex+A^{-1}h. \tag{5.3}$$

Consider first componentwise perturbations, $|e_{ij}|\leqslant\rho|a_{ij}|$, $|h_i|\leqslant\eta|b_i|$. Then the worst case error is given by

$$\frac{|A^{-1}Ex+A^{-1}h|_k}{\|x\|}\leqslant e^{\mathrm{wor}}=\rho\frac{\left(|A^{-1}||A||x|\right)_k}{\|x\|}+\eta\frac{\left(|A^{-1}|\cdot|Ax|\right)_k}{\|x\|}.$$

It is easy to observe that the results of Section 3 go through with n replaced by $\sqrt{n^2+n}$. That is, let

$$\mathbf{B}=\left\{(E,h):|e_{ij}|\leqslant\rho|a_{ij}|,|h_i|\leqslant\eta|b_i|\right\}\subset\mathbb{R}^{n^2+n},$$

and let μ be Lebesgue measure normalized so that $\mu(\mathbf{B})=1$. Then

$$e^{\mathrm{avg}}=\left(\int_{\mathbf{B}}\left(\frac{|A^{-1}Ex+A^{-1}h|_k}{\|x\|}\right)^p\mu(d(E,h))\right)^{1/p}=a_{n,p}e^{\mathrm{wor}},$$

where

$$(1+\varepsilon_{n,p})\left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right)^{1/p}\sqrt{\frac{2}{3}}\frac{1}{\sqrt{n^2+n}}\leqslant a_{n,p}\leqslant\frac{1}{(p+1)^{1/p}}$$

with $\lim_{n\to\infty}\varepsilon_{n,p}=0$ for all p. For the average loss of precision we have

$$\mathrm{loss}^{\mathrm{avg}}=\int_{\mathbf{B}}\log\frac{|A^{-1}Ex+A^{-1}h|_k}{\|x\|}\mu(d(E,h))=\log e^{\mathrm{wor}}-a_n,$$

where

$$\frac{1}{\ln 2} \leqslant a_n \leqslant \tfrac{1}{2}\log(n^2+n) + \frac{1+\varepsilon_n}{2}\left(\log 6 + \frac{\gamma}{\ln 2}\right)$$

with $\lim_{n\to\infty}\varepsilon_n = 0$.

Consider now norm perturbations, i.e., suppose that E satisfies (1.2) for the Frobenius or spectral norm, and h satisfies (5.1). Let

$$\mathbf{B} = \{(E,h): \|E\| \leqslant \rho\|A\|,\ \|h\| \leqslant \eta\|b\|\} \subset \mathbb{R}^{n^2+n},$$

and again let μ be Lebesgue measure normalized to make $\mu(\mathbf{B}) = 1$. In this case we can obtain

$$\max(\rho\,\mathrm{cond}^{\mathrm{avg}}(A,k,p), \eta\,\mathrm{cond}^{\mathrm{avg}}(b,k,p))$$
$$\leqslant e^{\mathrm{avg}} \leqslant \rho\,\mathrm{cond}^{\mathrm{avg}}(A,k,p) + \eta\,\mathrm{cond}^{\mathrm{avg}}(b,k,p), \qquad (5.4\mathrm{a})$$
$$\max(\log\rho + \mathrm{loss}^{\mathrm{avg}}(A,k), \log\eta + \mathrm{loss}^{\mathrm{avg}}(b,k))$$
$$\leqslant \mathrm{loss}^{\mathrm{avg}} \leqslant \log[\rho\,\mathrm{cond}^{\mathrm{avg}}(A,k,1) + \eta\,\mathrm{cond}^{\mathrm{avg}}(b,k,1)]. \qquad (5.4\mathrm{b})$$

The upper bound in (5.4a) follows from the triangle inequality for L^p. In (5.4b), the upper bound follows from the observation that the log function is concave on $(0,+\infty)$.

As for the lower bounds in (5.4), write $\mathbf{B} = \mathbf{B}_1 \times \mathbf{B}_2$, $\mu = \mu_1 \times \mu_2$ where $\mathbf{B}_1$ and $\mathbf{B}_2$ are balls in $\mathbb{R}^{n^2}$ (with respect to either spectral or Frobenius norm) and $\mathbb{R}^n$, respectively, and the μ_j are the appropriately normalized Lebesgue measures. The desired inequalities amount to the assertion that if $f(u) = \log u$ or $f(u) = u^p$, and $a = \int_{\mathbf{B}} f(|A^{-1}Ex + A^{-1}h|_k)\mu(d(E,h))$, then

$$a \geqslant a_1 = \int_{\mathbf{B}_1} f(|A^{-1}Ex|_k)\mu_1(dE), \qquad (5.5\mathrm{a})$$

$$a \geqslant a_2 = \int_{\mathbf{B}_2} f(|A^{-1}h|_k)\mu_2(dh). \qquad (5.5\mathrm{b})$$

At this stage, notice that the reasoning leading to (4.4) is valid when B is the unit ball in the Frobenius norm as well as in the spectral norm, and that with respect to either norm one can also prove the following slight variant, in which B_1 is the unit ball in $\mathbb{R}^{n^2}$ in either norm, Σ is again the unit sphere in $\mathbb{R}^n$, and $c \in \mathbb{R}$:

$$\int_{B_1} f(|(A^{-1}Ex)_k + c|)\mu_1(dE)$$

$$= \frac{1}{\lambda(B_1)}\int_{B_1} f(|\rho\|A\|(A^{-1}Yx)_k + c|)\,dY$$

$$= \frac{1}{\lambda(B_1)}\int_{B_1}\left[\frac{1}{\sigma(\Sigma)}\int_{\Sigma} f(|\rho\|A\|\,\|A_k^{-1}\|\,\|Yx\|\zeta_1 + c|)\sigma(d\zeta)\right]dY.$$

Recalling that $\mu_2(\mathbf{B}_2) = 1$, we see that (5.5a) follows if we set $c = (A^{-1}h)_k$ in (5.6) and then apply the general fact that if f is any nondecreasing function on $(0, \infty)$ and $r, c \in \mathbb{R}$, then

$$\int_{\Sigma} f(|r\zeta_1 + c|)\sigma(d\zeta) \geqslant \int_{\Sigma} f(|r\zeta_1|)\sigma(d\zeta). \tag{5.7}$$

As in the proof of Corollary 3.1, it suffices to show that (5.7) holds whenever f is the characteristic function of an interval $[t, +\infty)$ and $t \geqslant 0$, i.e., that

$$\sigma(\{\zeta \in \Sigma : |r\zeta_1 + c| \geqslant t\}) \geqslant \sigma(\{\zeta \in \Sigma : |r\zeta| \geqslant t\}). \tag{5.8}$$

Since (5.8) follows easily from the nature of σ [see (4.7)], our proof of (5.5a) is finished. Moreover, there is an obvious analog of (5.6) for $\mathbf{B}_2$, which together with (5.7) leads to (5.6b).

We note that it is substantially easier to establish the lower bounds in (5.4) if either $f(u) = u^p$ or the norm in $\mathbb{R}^{n^2}$ is the Frobenius norm.

As an example, if we combine (5.4a) with our formulas for $\text{cond}^{\text{avg}}(b, k, 2)$ and $\text{cond}^{\text{avg}}(A, k, 2)$ for the Frobenius norm, we have when $p = 2$

$$\|A_k^{-1}\|\max\left(\rho\frac{\|A\|}{\sqrt{n^2+2}}, \eta\frac{\|Ax\|}{\|x\|\sqrt{n+2}}\right)$$

$$\leqslant e^{\text{avg}} \leqslant \|A_k^{-1}\|\left(\rho\frac{\|A\|}{\sqrt{n^2+n}} + \eta\frac{\|Ax\|}{\|x\|\sqrt{n+2}}\right).$$

Similarly, combining (5.4b) with our previous estimates for $\text{loss}^{\text{avg}}(A, k)$ and $\text{cond}^{\text{avg}}(A, k, 1)$ and observing also that if $c_1, c_2 > 0$ then $\log(c_1 + c_2) \leqslant 1$

$+\max(\log c_1, \log c_2)$, we have in the Frobenius norm case

$$\text{loss}^{\text{avg}} = \log\|A_k^{-1}\| + \max\left(\log\rho + \log\|A\| - \log n,\ \log\eta + \log\frac{\|Ax\|}{\|x\|} - \tfrac{1}{2}\log n\right) + b_n,$$

where $-0.916\ldots \leqslant b_n \leqslant \frac{1}{2}\log(2/\pi) + 1 = 0.674\ldots.$

REFERENCES

1 L. Blum and M. Shub, Evaluating rational functions: Infinite precision is finite cost and tractable on average, *SIAM J. Comput.*, to appear.

2 J. Demmel, A numerical analyst's Jordan canonical form, Ph.D. Thesis, Univ. of California at Berkeley, 1983.

3 I. S. Gradshteyn and I. W. Ryzhik, *Table of Integrals Series and Products*, 4th ed., Academic, 1965.

4 E. Kostlan, Complexity theory of numerical linear algebra, to appear.

5 A. Ocneanu, On the loss of precision in solving large linear systems, to appear.

6 R. D. Skeel, Iterative refinement implies numerical stability for Gaussian elimination, *Math. Comp.* 35 (151):817–832 (1980).

7 G. W. Stewart. *Introduction to Matrix Computations*, Academic, 1973.

8 J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, 1963.

9 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford U.P., 1965.

Received 28 February 1985; revised 28 October 1985