SIAM J. NUMER. ANAL.
Vol. 21, No. 1, February 1984

0036-1429/84/2101-0015 $01.25/0

A NEW MEASURE OF RELATIVE ERROR FOR VECTORS*

J. D. PRYCE†

Abstract. The paper defines a new measure of the relative error of a vector y as an approximation to a vector x, under an arbitrary norm. This is shown to be a metric whereas the traditional measure $\|y-x\|/\|x\|$ is only approximately so. It generalizes F. W. J. Olver's corresponding measure for scalars, has an explicit formula in inner product spaces, and is proposed as a tool for roundoff error analyses.

1. Background. Recently F. W. J. Olver has introduced a new definition of the relative error of one number as an approximation to another (Olver (1978)) and in a series of papers has shown it to be a most useful practical tool in roundoff error analysis, allowing one to simplify many classical analyses and, generally, to eliminate most of the "factors slightly in excess of 1" which are used to make approximate error bounds into rigorous ones. See especially Olver (1982) and Olver and Wilkinson (1982). In particular the technique has yielded the first fully rigorous, computable, asymptotically optimal, and reasonably computationally efficient a posteriori error bounds for the solution of a linear system $Ax=b$ by Gaussian elimination. The rigour derives from the fact that one can keep track of *all* roundoff errors concisely (including those during computation of the bounds) in a way that the analysis of (say) Stummel (1981) does not.

The traditional definition of the relative error of y as an approximation to x ($x\neq 0$) is

$$\rho_0(x,y)=|(y-x)/x|. \tag{1.1a}$$

The new definition, for real x, y of the same sign, is

$$\rho(x,y)=|\ln (y/x)|. \tag{1.1b}$$

Alternatively,

$$\begin{aligned}\rho_0(x,y)&=|t|\quad\text{where } y=(1+t)x,\\ \rho(x,y)&=|t|\quad\text{where } y=e^t x\end{aligned} \tag{1.2a, b}$$

and it is clear that ρ and ρ_0 differ only by quantities of second order as either of them tends to 0. The advantage of ρ over ρ_0 is that it is a *metric*, that is, the axioms

M1. $\rho(x,x)=0, \rho(x,y)>0$ if $x\neq y$;
M2. $\rho(x,y)=\rho(y,x)$;
M3. $\rho(x,z)\leqq\rho(x,y)+\rho(y,z)$

are exactly true of ρ, but not true of ρ_0 (though approximately so when ρ_0 is small).

For vectors also, a concept of relative error is useful in roundoff error analysis and elsewhere. The usual definition is by analogy with (1.1a):

$$\rho_0(x,y)=\|y-x\|/\|x\| \tag{1.3}$$

relative to some vector norm, and many error analyses in linear algebra give bounds that are conveniently expressed in such a form, though such bounds may be less sharp than corresponding elementwise bounds.

* Received by the editors July 9, 1982, and in revised form March 29, 1983.

† Computer Science Department, University of Bristol, Bristol, England BS8 1TW.

The main aims of the paper are as follows.

1. Of course ρ_0 in (1.3) is not a metric any more than is (1.1a). We show that any normed space admits a *relative-error metric* ρ which reduces to (1.1b) in the scalar case and satisfies

$$\rho/\rho_0 \to 1 \quad \text{as } \rho \text{ or } \rho_0 \text{ tends to zero.}$$

2. For complex x and y, formula (1.1b) still defines a metric if we take the principal value of ln (the triangle law follows from the fact that $t = \ln z$ is the solution of smallest absolute value to the equation $z = e^t$). Equivalently

$$\rho(x, y) = \{(\ln |y/x|)^2 + (\arg (y/x))^2\}^{1/2}, \qquad x, y \in \mathbb{C} \sim \{0\}. \tag{1.4}$$

We show that in any inner-product space there is a simple formula for ρ generalizing (1.4).

3. In $\mathbb{C}$ one can write

$$\rho(x, y) = \min \{|t| : t \in \mathbb{C}, y = e^t x\}. \tag{1.5}$$

We show that in any normed space this generalizes to a relation between ρ and equations $y = e^T x$ where T is a linear operator; the relation is especially close in an inner-product space. We also derive relations between $\rho(x, y)$ and $\rho(Ax, Ay)$ where A is a linear operator.

4. In § 6 we translate the results into Olver's "rp" (relative precision) notation and show that many of the results of Olver (1978) extend with minimal changes to the vector case.

The final section contains some pointers for future work, and a few open problems.

This paper was written in ignorance of the article by Ziv (1982), to which one of the referees drew my attention. Ziv studies the relative error measure $d(x, y) = \|x - y\|/\max\{\|x\|, \|y\|\}$, and I have added to the final section some comparisons of the two measures.

2. Mathematical preliminaries. In what follows we shall assume X is a real or complex normed space unless stated otherwise. A *path* in X means a subset C of X which is the range of a mapping

$$t \mapsto x(t),$$

defined on a compact interval $a \leqq t \leqq b$ in $\mathbb{R}$, and piecewise smooth in the norm topology, that is, there are points $a = t_0 < t_1 < \cdots < t_N = b$ such that on each $[t_{i-1}, t_i]$,

$$x'(t) = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}$$

exists in the norm topology and is norm-continuous. Any reparametrization $s \mapsto x(\phi(s))$, with ϕ strictly increasing, and ϕ and its inverse function piecewise smooth, is regarded as defining the same path C.

We say C goes *from* x_0 *to* y_0 where $x_0 = x(a)$, $y_0 = x(b)$. The path $-C$ denotes the same set parametrized in the reverse direction, e.g. by $x(-t)$, $-b \leqq t \leqq -a$, and goes from y_0 to x_0. If C_1 goes from x_0 to y_0, C_2 from y_0 to z_0, then $C_1 + C_2$ denotes the set $C_1 \cup C_2$ suitably reparametrized as a path from x_0 to z_0.

We make much use of (scalar) integrals of the form

$$\int_C F \|dx\| \tag{2.1a}$$

where F is some function defined on C. This is short for

$$\int_a^b F(x(t))\|x'(t)\|\,dt \tag{2.1b}$$

and since $\|x'(t)\|$ is piecewise continuous such an integral exists in the Riemann sense when F is continuous, and the value is independent of the parametrization. One easily sees also that the integral over $-C$ is equal to that over C, and the integral over C_1+C_2 is the sum of those over C_1 and C_2.

We also use integrals

$$\int_C F\,d\|x-x_0\| \underset{\text{def}}{=} \int_a^b F(x(t))\frac{d}{dt}\|x(t)-x_0\|\,dt. \tag{2.2}$$

Since the norm may not be differentiable more care is needed here. In fact since $x(t)$ is continuously differentiable on the subintervals $[t_{i-1}, t_i]$ above, it is Lipschitzian, i.e. $\|x(s)-x(t)\|\leqq L|s-t|$ holds for some constant L. The inequality

$$|\,\|x-x_0\|-\|y-y_0\|\,|\leqq\|x-y\|$$

then implies that the scalar function $\theta(t)=\|x(t)-x_0\|$ is Lipschitzian, and thus is absolutely continuous, has a derivative a.e., and is the Lebesgue integral of its derivative. We conclude that, for continuous F, integrals (2.2) always exist in the Lebesgue sense.

Of course the norms used in practice (the 1, 2 and ∞ norms on $\mathbb{R}^n$ for instance) are piecewise smooth, making this machinery unnecessary, but it is nice to cover the general case.

For integrals (2.2), the integral over C_1+C_2 is the sum of those over C_1 and C_2 but the integral over $-C$ is minus that over C.

3. Definition and basic properties.

DEFINITION 3.1. The *relative distance* $\rho(x, y)$ between nonzero x, y in the normed space X is the geodesic distance defined by the metric element

$$d\rho=\frac{\|dx\|}{\|x\|}, \tag{3.1}$$

that is,

$$\rho(x,y)=\inf_C\int_C\frac{\|dx\|}{\|x\|}=\inf_C\int_a^b\frac{\|x'(t)\|}{\|x(t)\|}dt \tag{3.2}$$

over all paths $C: t\mapsto x(t)$, $a\leqq t\leqq b$ from x to y, and not passing through 0. (The last condition is optional since Lemma 3.4 part (c) shows the integral diverges to $+\infty$ for any path through 0.)

Of the metric axioms, the first half of M1, $\rho(x,x)=0$, is clear, and M2, M3 follow easily from the equations

$$\int_{-C}\frac{\|dx\|}{\|x\|}=\int_C\frac{\|dx\|}{\|x\|},\qquad \int_{C_1+C_2}\frac{\|dx\|}{\|x\|}=\int_{C_1}\frac{\|dx\|}{\|x\|}+\int_{C_2}\frac{\|dx\|}{\|x\|}.$$

The remaining fact, that $\rho(x,y)>0$ whenever $x\neq y$, follows from the next lemma, which gives sharp bounds on the relation between ρ and the "traditional" ρ_0 in (1.3).

LEMMA 3.2. (a) *For any* $x_0, y_0 \neq 0$ *let* ρ *denote* $\rho(x_0, y_0)$ *and* ρ_0 *denote* $\rho_0(x_0, y_0)$. *Then*

$$\ln(1+\rho_0) \leqq \rho \leqq -\ln(1-\rho_0) \tag{3.3}$$

or equivalently

$$e^{\rho}-1 \geqq \rho_0 \geqq 1-e^{-\rho} \tag{3.4}$$

provided that $\rho_0 < 1$ *in the right-hand inequality of* (3.3).

(b) *Whenever* $\rho_0 \leqq \delta < 1$ *we have*

$$1-\delta \leqq \frac{-\delta}{\ln(1-\delta)} \leqq \frac{\rho_0}{\rho} \leqq \frac{\delta}{\ln(1+\delta)} \leqq 1+\delta. \tag{3.5}$$

Proof. (a) Take any path C: $x(t)$ from x_0 to y_0. By the triangle inequality

$$\frac{d}{dt}\|x-x_0\| \leqq \left\|\frac{dx}{dt}\right\| \text{ on } C \tag{3.6}$$

and $\|x_0\|+\|x-x_0\| \geqq \|x\|$, so that

$$\begin{aligned}\int_C \frac{\|dx\|}{\|x\|} &\geqq \int_C \frac{d\|x-x_0\|}{\|x_0\|+\|x-x_0\|}\\ &= [\ln(\|x_0\|+\|x-x_0\|)]_{x_0}^{y_0}\\ &= \ln(1+\rho_0) \text{ by definition of } \rho_0.\end{aligned}$$

Taking the infimum over all paths gives the left hand of (3.3).

Next, consider the straight-line segment

$$C\colon x(t)=x_0+t(y_0-x_0), \quad 0 \leqq t \leqq 1$$

from x_0 to y_0. On C we have $\|dx\| = \|y_0-x_0\|\,dt$ and $\|x\| \geqq \|x_0\| - t\|y_0-x_0\|$, so that

$$\begin{aligned}\rho \leqq \int_C \frac{\|dx\|}{\|x\|} &\leqq \int_0^1 \frac{\|y_0-x_0\|}{\|x_0\|-t\|y_0-x_0\|}\,dt\\ &= [-\ln(\|x_0\|-t\|y_0-x_0\|)]_0^1\\ &= -\ln(1-\rho_0)\end{aligned}$$

provided that $\|y_0-x_0\| < \|x_0\|$, i.e. that $\rho_0 < 1$. This gives the right-handed side of (3.3). Inequalities (3.4) follow from (3.3), and the proof is complete.

(b) The inner inequalities follow from part (a) and the fact that $t/\ln(1+t)$ is an increasing function for $t > -1$; the outer inequalities are elementary.

THEOREM 3.3. ρ *is a metric, and*

$$\frac{\rho_0(x, y)}{\rho(x, y)} \to 1, \quad \textit{uniformly on } X,$$

as either ρ *or* $\rho_0 \to 0$.

Proof. Immediate from the lemma and the preceding remarks.

We now derive some elementary properties of the ρ metric. Part (a) is an obvious scale-invariance; (c), (d) are useful and not so obvious; and (e), (f) show that (3.3), (3.4) are sharp.

LEMMA 3.4. *For any nonzero* x_0, y_0

(a) $\rho(\lambda x_0, \lambda y_0) = \rho(x_0, y_0)$ *for any scalar* $\lambda \neq 0$.

(b) *Let X be the real line. Then for* x_0, y_0 *of the same sign*, $\rho(x_0, y_0) = |\ln(y_0/x_0)|$.

(c) $\rho(x_0, y_0) \geqq |\ln(\|y_0\|/\|x_0\|)|$.

(d) *Hence*, $\rho(x_0, y_0) \geqq \rho(\|x_0\|, \|y_0\|)$.

(e) *The right-hand bounds of* (3.3), (3.4) *are attained when* $y_0 = \alpha x_0, 0 < \alpha \leqq 1$.

(f) *The left-hand bounds of* (3.3), (3.4) *are attained when* $y_0 = \alpha x_0, \alpha \geqq 1$.

Proof. (a) Clear since the defining integral (3.2) is unaltered if the path $x(t)$ from x_0 to y_0 is replaced by $\lambda x(t)$ from λx_0 to λy_0.

(b) Clearly $\rho(x_0, y_0) = \rho(-x_0, -y_0)$, so we may assume $x_0 > 0$, $y_0 > 0$. By part (c), $\rho(x_0, y_0) \geqq |\ln(|y_0|/|x_0|)| = |\ln(y_0/x_0)|$. Suppose $x_0 \leqq y_0$. Then taking C as the path $x = t$, for $x_0 \leqq t \leqq y_0$ gives

$$\rho(x_0, y_0) \leqq \int_C \frac{|dx|}{|x|} = \ln(y_0/x_0).$$

Similarly if $y_0 \leqq x_0$, we get $\rho(x_0, y_0) \leqq -\ln(y_0/x_0)$. Together these prove the required equality.

(c) On any path C: $x(t)$ from x_0 to y_0 we have (cf. (3.6))

$$\left\|\frac{dx}{dt}\right\| \geqq \left|\frac{d}{dt}\|x\|\right|,$$

so

$$\int_C \frac{\|dx\|}{\|x\|} \geqq \int_C \frac{|d\|x\||}{\|x\|} \geqq \left|\int_C \frac{d\|x\|}{\|x\|}\right|$$
$$= |[\ln\|x\|]_{x=x_0}^{y_0}| = |\ln(\|y_0\|/\|x_0\|)|.$$

Taking the infimum over paths gives the result.

(d) Immediate from (b), (c).

(e) When $y_0 = \alpha x_0$ where $0 < \alpha \leqq 1$, the right hand of (3.3) gives $\rho(x_0, y_0) \leqq -\ln(1-(1-\alpha)) = |\ln \alpha|$, whereas part (c) of the lemma gives the reverse inequality.

(f) When $y = \alpha x_0$ where $\alpha \geqq 1$, the left-hand side of (3.3) gives $\rho(x_0, y_0) \geqq \ln(1+\alpha-1) = \ln \alpha$, whereas if C is the path $x = tx_0, 1 \leqq t \leqq \alpha$ then $\rho(x_0, y_0) \leqq \int_C \|dx\|/\|x\| = \ln \alpha$, the reverse inequality. QED

In the interests of conceptual simplicity we now give a characterization of the ρ metric purely in terms of finite sums instead of integrals.

Let us define a *mesh M* from x to y in X to be any sequence of nonzero x_i in X $(i = 0, \cdots, N)$ with $x_0 = x, x_N = y$. We define the *relative meshsize* $r(M)$ of M by

$$r(M) = \max_{1 \leqq i \leqq N} \rho_0(x_{i-1}, x_i) \tag{3.7}$$

and the ρ_0*-length* of M, which we denote by $\rho_0(M)$, by

$$\rho_0(M) = \sum_{i=1}^{N} \rho_0(x_{i-1}, x_i). \tag{3.8}$$

THEOREM 3.5. (a) *Let* $0 < \delta < 1$. *Then for any nonzero x, y in X*,

$$\frac{-\delta}{\ln(1-\delta)} \rho(x, y) \leqq \inf_{r(M) \leqq \delta} \rho_0(M) \leqq \frac{\delta}{\ln(1+\delta)} \rho(x, y).$$

(b) *For any nonzero x, y in X*,

$$\rho(x, y) = \lim_{r(M) \to 0} \inf \rho_0(M)$$

where in each case M ranges over all meshes from x to y.

Proof. (a) For any mesh M from x to y with $r(M) \leqq \delta$ we have by (3.5)

$$\frac{\rho_0(x_{i-1}, x_i)}{\rho(x_{i-1}, x_i)} \geqq \frac{-\delta}{\ln(1-\delta)}$$

so

$$\frac{-\delta}{\ln(1-\delta)}\rho(x, y) \leqq \frac{-\delta}{\ln(1-\delta)} \sum_1^N \rho(x_{i-1}, x_i)$$
$$\leqq \sum_1^N \rho_0(x_{i-1}, x_i) = \rho_0(M)$$

proving the left-hand inequality of (a).

Now, given $\varepsilon > 0$, choose a path C: $x(t)$, $a \leqq t \leqq b$ from x to y, such that

$$\int_C \frac{\|dx\|}{\|x\|} \leqq \rho(x, y) + \varepsilon. \tag{3.9}$$

By a compactness argument we can choose points t_i in $[a, b]$,

$$a = t_0 < t_1 < \cdots < t_N = b$$

such that the corresponding $x_i = x(t_i)$ form a mesh M with $r(M) \leqq \delta$. Let C_i be the portion of C between t_{i-1} and t_i; then for each i

$$\frac{\rho_0(x_{i-1}, x_i)}{\rho(x_{i-1}, x_i)} \leqq \frac{\delta}{\ln(1+\delta)}, \tag{3.10}$$

by (3.5). Thus by (3.9)

$$\rho(x, y) + \varepsilon \geqq \int_C \frac{\|dx\|}{\|x\|} = \sum_i \int_{C_i} \frac{\|dx\|}{\|x\|} \geqq \sum_i \rho(x_{i-1}, x_i)$$
$$\geqq \sum_i \frac{\ln(1+\delta)}{\delta} \rho_0(x_{i-1}, x_i) \quad \text{by (3.10)}$$
$$= \frac{\ln(1+\delta)}{\delta} \rho_0(M).$$

Since ε is arbitrary the right-hand inequality of (a) follows.

(b) This follows at once, since by definition

$$\liminf_{r(M)\to 0} \rho_0(M) = \lim_{\delta \to 0} \inf_{r(M) \leqq \delta} \rho_0(M). \qquad \text{QED}$$

4. $\mathbb{C}$, $\mathbb{R}$ and inner-product spaces. In $\mathbb{C}$, with the usual norm $|z|$, the extra algebraic structure gives

$$\frac{\|dz\|}{\|z\|} = \left|\frac{dz}{z}\right| = |dw| \quad \text{where } w = \ln z,$$

so that on any region Z of the z-plane admitting a branch of $\ln z$, the mapping $w = \ln z$ is an isometry of Z with the ρ-metric onto a region W of the w-plane with the usual metric. Thus geodesics in $\mathbb{C}$ exist, and are the images under $z = \exp w$ of straight lines in the w-plane, i.e. equiangular spirals centered on the origin. It follows easily that ρ reduces to the formula (1.1b). Since each z is the image of many w's differing by multiples of $2\pi i$, there are many geodesics from x to y, the shortest one corresponding to the principal value of ln.

In $\mathbb{R}$ there are no paths from positive to negative x values, but it is natural to think of $\mathbb{R}$ as embedded in $\mathbb{C}$. For instance -2 then approximates to $+3$ with relative error $((\ln 3/2)^2+\pi^2)^{1/2}$.

In Euclidean space the symmetry of the norm makes it intuitively clear that the ρ-shortest path from x to y lies in the plane Oxy, and since a plane is isometric to $\mathbb{C}$ this will yield an explicit formula for $\rho(x, y)$. We now make this argument precise.

THEOREM 4.1. *Let X be a real inner-product space of dimension $\geqq 2$ and let nonzero $x, y \in X$ lie in a plane (2-dimensional subspace) P. Then*

(a) *The ρ-distance of x, y as elements of P equals their ρ-distance as elements of X.*

$$(4.1)\quad \text{(b)} \qquad \rho(x, y)=[\{\ln(\|y\|/\|x\|)\}^2+\{\text{angle }(x, y)\}^2]^{1/2}$$

where angle $(x, y)=\cos^{-1}(x\cdot y/\|x\|\|y\|)$, *in the range* 0 *to* π, $x\cdot y$ *denoting the inner-product.*

Proof. (a) Let the two meanings of ρ be distinguished as $\rho_P(x, y)$ and $\rho_X(x, y)$. The key is to construct a (not necessarily linear) map $T: X\to P$ which nowhere increases ρ_0, i.e.

$$\rho_0(Tu, Tu')\leqq \rho_0(u, u') \quad \text{for any } u, u',$$

and which leaves x, y fixed. For then T maps any mesh M from x to y, of meshsize $\leqq\delta$, to a mesh TM lying in P, also of meshsize $\leqq\delta$ and satisfying

$$\rho_0(TM)\leqq\rho_0(M).$$

Theorem 3.5 then yields

$$\rho_P(x, y)\leqq\rho_X(x, y),$$

while the reverse inequality is always true in any normed space (because X has more paths to take the infimum over), so that

$$\rho_P(x, y)=\rho_X(x, y),$$

which is what we wish to prove.

We construct T as follows. Choose a coordinate system $(s, t)\to si+tj$ in P, relative to orthonormal vectors i, j in P chosen so that x and y lie in the upper half-plane $t\geqq 0$. Now, any $u\in X$ can uniquely be written as

$$u=si+v \quad \text{where } v\perp i.$$

Define

$$Tu=si+\|v\|j.$$

It is easy to verify the properties $Tx=x$, $Ty=y$ and

$$\|Tu\|=\|u\| \qquad (u\in X),$$

$$\|Tu-Tu'\|\leqq\|u-u'\| \qquad (u, u'\in X).$$

The last two relations imply $\rho_0(Tu, Tu')\leqq\rho_0(u, u')$, so T has the required properties and the proof is complete.

(b) Equation (4.1) holds for the case when $X=\mathbb{R}^2$, being simply the translation of (1.4) in terms of the $\mathbb{R}^2$ norm and inner product. Since any plane in a real inner-product space is norm and inner-product isomorphic to $\mathbb{R}^2$, it holds in P also. Since $\rho_P(x, y)=\rho_X(x, y)$ by part (a), it holds in X. QED

COROLLARY 4.2. *Equation* (4.1) *holds also in a complex inner-product space X provided we define*

$$\text{angle}\,(x, y)=\cos^{-1}\left(\frac{\operatorname{Re}\,(x\cdot y)}{\|x\|\|y\|}\right).$$

Proof. X becomes a real inner-product space with the same norm, if we restrict scalar multiplication to real scalars and define the "real inner product"

$$x\cdot_R y=\operatorname{Re}\,(x\cdot y)$$

5. Connections with linear mappings. We now need to assume the normed space X is complete, i.e. a Banach space. Of course the case $X=\mathbb{R}^n$ or $\mathbb{C}^n$ continues to be the one of most practical interest.

It was pointed out in § 1 that in the scalar case

$$\rho(x, y)=\min\{|t|\colon y=e^t x\};\tag{5.1a}$$

equivalently

$$\rho(x, y)\leqq\delta\quad\text{iff}\quad y=e^t x\quad\text{where } |t|\leqq\delta.\tag{5.1b}$$

Is there something analogus to (5.1a, b) for vectors? We give a partial answer in the next theorem. Recall that for any member T of the set $\mathbb{B}(X)$ of bounded linear operators on the Banach space X there exists the *exponential* of T

$$e^T=I+T+T^2/2!+T^3/3!+\cdots\tag{5.2}$$

with the property that $e^S e^T=e^{S+T}$ if $\c{S}$, T commute (but not in general); in particular e^T has inverse e^{-T}, and the set of operators e^{sT} for all scalar s forms a group.

THEOREM 5.1. *Assume* $x_0, y_0\neq 0$. (*a*) *If* $y_0=e^T x_0$ *where* $\|T\|\leqq\delta$, *then* $\rho(x_0, y_0)\leqq\delta$.

(b) *In a real inner-product space of dimension* $\geqq 2$, $y_0=e^T x_0$ *where* $\|T\|\leqq\delta$, *if and only if* $\rho(x_0, y_0)\leqq\delta$.

Proof. (a) Consider the path

$$C\colon x=e^{tT}x_0,\qquad 0\leqq t\leqq 1$$

from x_0 to y_0. It is easy to see that, along C,

$$dx=Te^{tT}x_0\,dt=Tx\,dt$$

and by the definition of ρ,

$$\rho(x_0, y_0)\leqq\int_C\frac{\|dx\|}{\|x\|}=\int_0^1\frac{\|Tx\|}{\|x\|}dt\leqq\int_0^1\|T\|\,dt=\|T\|\leqq\delta.\tag{5.3}$$

(b) We have to prove the reverse implication when X is an inner-product space. First, it is true when $X=\mathbb{R}^2$. This is because multiplication by a complex $c=a+ib$ can be regarded as the linear mapping T_c on $\mathbb{R}^2$ with matrix

$$T_c=\begin{pmatrix}a & -b\\ b & a\end{pmatrix}$$

and the correspondence $c\to T_c$ is an isometric algebra-isomorphism of $\mathbb{C}$ into $\mathbb{B}(\mathbb{R}^2)$. Thus, for x_0, y_0 in $\mathbb{R}^2=\mathbb{C}$,

$$y_0=e^c x_0,\quad |c|\leqq\delta\quad\text{iff}\quad y_0=e^{T_c}x_0,\quad \|T_c\|\leqq\delta.$$

Now take the general case of x_0, y_0 in an inner product space X. There is a plane P containing x_0, y_0. Let $\rho(x_0, y_0) \leqq \delta$; then since P is isometrically isomorphic to $\mathbb{R}^2$ there is $T \in \mathbb{B}(P)$ with $\|T\| \leqq \delta$, $y_0 = e^T x_0$ by the previous paragraph.

Extend T from P to X by defining it to be zero on the orthogonal complement of P. The extended T still has $\|T\| \leqq \delta$, $y_0 = e^T x_0$, and the proof is complete. QED

In a general (not inner-product) space X it seems unlikely that such a strong result as Theorem 5.1(b) holds, but one can get something similar involving finite products of exponentials as we now show.

LEMMA 5.2. *Given nonzero* x_0, y_0 *with* $\rho_0(x_0, y_0) \leqq \delta < 1$, *we have* $y_0 = e^T x_0$ *for some* T *such that* $\|T\| \leqq -\ln(1-\delta)$. *In fact* T *can be chosen to have rank one.*

Proof. By the Hahn–Banach theorem there is a bounded linear functional f on X such that $f(x_0) = \|f\| \cdot \|x_0\| = 1$. Let S be the rank-one operator defined by

$$Sx = f(x)(y_0 - x_0), \qquad (x \in X)$$

so that

$$\|S\| \leqq \|f\| \, \|y_0 - x_0\| = \frac{\|y_0 - x_0\|}{\|x_0\|} = \rho_0(x_0, y_0) \leqq \delta < 1,$$

and $S^2 = \alpha S$, where $\alpha = f(y_0 - x_0)$ satisfies $|\alpha| \leqq \delta < 1$. Define

$$T = \frac{\ln(1+\alpha)}{\alpha} S.$$

Then it is easily verified that T has the required properties.

THEOREM 5.3. *Let* x_0, y_0 *lie in a Banach space* X *and let* $\theta > \rho(x_0, y_0)$. *Then there exists a finite sequence of (rank-one) operators* $T_1, \cdots, T_n$ *with* $\sum \|T_i\| < \theta$ *and*

$$y_0 = e^{T_n} e^{T_{n-1}} \cdots e^{T_1} x_0.$$

Proof. We can choose δ so that $0 < \delta < 1$ and

$$\frac{-\ln(1-\delta)}{\ln(1+\delta)} \rho(x_0, y_0) < \theta \tag{5.4}$$

(since the left side tends to $\rho(x_0, y_0)$ as $\delta \to 0$). Then let $\varepsilon > 0$ be the difference between the right and left sides of this inequality. By Theorem 3.5, and using its notation, there is a mesh M of points $u_0, \cdots, u_n$ from x_0 to y_0 such that if we write

$$\delta_i = \rho_0(u_{i-1}, u_i) \qquad (i = 1, \cdots, n)$$

then $\delta_i \leqq \delta$ and

$$\rho_0(M) \equiv \sum \delta_i < \frac{\delta}{\ln(1+\delta)} \rho(x_0, y_0) + \frac{-\delta}{\ln(1-\delta)} \varepsilon. \tag{5.5}$$

By the last lemma we can write

$$u_i = e^{T_i} u_{i-1}$$

where T_i has rank one and $\|T_i\| \leqq -\ln(1-\delta_i)$. Then

$$y_0 = u_n = e^{T_n} e^{T_{n-1}} \cdots e^{T_1} x_0$$

and

$$\begin{aligned}\sum \|T_i\| &\leqq \sum \left(\frac{-\ln(1-\delta_i)}{\delta_i}\right)\delta_i \\ &\leqq -\frac{\ln(1-\delta)}{\delta}\sum \delta_i \quad \text{since } \delta_i \leqq \delta < 1 \\ &< -\frac{\ln(1-\delta)}{\delta}\left(\frac{\delta}{\ln(1+\delta)}\rho(x_0, y_0) + \frac{-\delta}{\ln(1-\delta)}\varepsilon\right) \\ &= \theta\end{aligned}$$

by (5.4), (5.5) and the definition of ε. QED

THEOREM 5.4. *For any invertible $A \in \mathbb{B}(X)$ and any nonzero x, y*

$$\rho_0(Ax, Ay) \leqq \kappa\rho_0(x, y) \tag{5.6}$$

and

$$\rho(Ax, Ay) \leqq \kappa\rho(x, y) \tag{5.7}$$

where $\kappa = \kappa(A)$ is the usual condition number $\|A\|\,\|A^{-1}\|$.

Proof. The first comes from the inequalities

$$\begin{aligned}\|Ay - Ax\| &\leqq \|A\|\,\|y - x\| \\ \|AX\| &\geqq \|A^{-1}\|^{-1}\|x\|,\end{aligned} \tag{5.8}$$

and the second then follows using Theorem 3.5, much as in the proof of Theorem 4.1.

Inequality (5.6) is sharp (in finite dimensions) because we can always choose x and then y so that relations (5.8) are equalities. But (5.7) will not be sharp in general.

With the more general definition

$$\kappa(A) = \max_{\|x\|=1} \|Ax\| / \min_{\|x\|=1} \|Ax\|,$$

the theorem remains true for linear maps from one normed space to another.

6. Towards practicalities.

In this section we translate the results of the preceding sections into the ap and rp notation of Olver (1978), and list the similarities and differences between Olver's scalar case and the vector case. For brevity we label formulae with Roman numerals following Olver's labelling where appropriate, with the codes

O(x.x) for a formula in Olver § x.x,

IP for a formula valid only in an inner product space.

6.1. Basic properties.

We extend Olver's notation in the obvious way for vectors $a, \tilde{a}$:

$a \simeq \tilde{a}$; ap(α) means $\|a - \tilde{a}\| \leqq \alpha$;

$a \simeq \tilde{a}$; rp(α) means $\rho(a, \tilde{a}) \leqq \alpha$.

Then ap has all the properties I to VI of Olver § 2.2, while rp has the following properties from Olver § 3.2: Let

$$\alpha \simeq \tilde{a}; \quad \text{rp}(\alpha).$$

Then

O(3.2)I. Symmetry: $\bar{a} \simeq a$; rp(α).

O(3.2)II. Inclusion: $a \simeq \bar{a}$; rp(δ) for any $\delta \geqq \alpha$.

O(3.2)III. For any nonzero scalar k,

$$ka \simeq k\bar{a}; \quad \text{rp}(\alpha).$$

O(3.2)VI. If, also,

$$\bar{a} \simeq \bar{\bar{a}}; \quad \text{rp}(\delta)$$

then

$$a \simeq \bar{\bar{a}}; \quad \text{rp}(\alpha + \delta).$$

These follow at once from Lemma 3.4(a) and the fact that ρ is a metric. There is no analogue of Olver's IV, V relating to powers, products and quotients but we have some new properties:

(IP)VII. $a \simeq \bar{a}$; rp(α) iff $\bar{a} = e^T a$ where $T \in \mathbb{B}(X)$, $\|T\| \leqq \alpha$,

which comes from Theorem 5.1(b).

VIII. For any invertible linear operator A,

$$Aa \simeq A\bar{a}; \quad \text{rp}(\kappa\alpha), \qquad \kappa = \|A\| \cdot \|A^{-1}\|,$$
$$A^{-1}a \simeq A^{-1}\bar{a}; \quad \text{rp}(\kappa\alpha)$$

which comes from Theorem 5.4. In particular

(IP)IX. If A is a unitary operator, or scalar multiple thereof,

$$Aa \simeq A\bar{a}; \quad \text{rp}(\alpha).$$

Lemma 3.4(d) implies

X. $$\|a\| \simeq \|\bar{a}\|; \quad \text{rp}(\alpha).$$

6.2. Conversion between ap and rp. The rules for conversion extend Olver § 3.4. They follow at once from the definition of ρ_0 and Lemma 3.2:

O(3.4)I. $a \simeq \bar{a}$ ap(α) implies $a \simeq \bar{a}$; rp($-\ln(1 - \alpha/\|\bar{a}\|)$);

O(3.4)II. $a \simeq \bar{a}$; rp(α) implies $a \simeq$; ap($\|\bar{a}\|(e^\alpha - 1)$).

6.3. Relation between rp of vector and rp of its components. Let X be $\mathbb{R}^n$ or $\mathbb{C}^n$ with the l_p norm ($1 \leqq p \leqq \infty$)—$p = 1$, 2 or ∞ being the cases of most interest. The following generalizes the rules for conversion from real to complex rp, (Olver, p. 391), though the proof seems to be quite different, and simpler than Olver's.

XI. If $a = (a_1, \cdots, a_n)$ and $\bar{a} = (\bar{a}_1, \cdots, \bar{a}_n)$ and

$$a_i \simeq \bar{a}_i; \quad \text{rp}(\alpha_i), \qquad i = 1, \cdots, n$$

then

$$a \simeq \bar{a}; \quad \text{rp}(\max_i \alpha_i).$$

Proof. For each i we have $\bar{a}_i = e^{t_i} a_i$ with $|t_i| \leqq \alpha_i$. Let T be the $n \times n$ matrix diag $(t_1, t_2, \cdots, t_n)$. Then

$$\bar{a} = e^T a$$

and, for any l_p norm,

$$\|T\| = \max_i |t_i| \leqq \max_i \alpha_i$$

so the result follows from Theorem 5.1.

There is no converse relation: $\rho(a, \bar{a})$ can be as small as we like but $\rho(a_i, \bar{a}_i)$ for a given i can be arbitrarily large, e.g. if $a_i = 0$, $\bar{a}_i \neq 0$.

Olver shows that with the usual definition of *roundoff unit* u for floating point arithmetic on a computing facility one has

$$x \bullet y \simeq x \mathbin{\tilde{\bullet}} y; \mathrm{rp}(u)$$

where $\bullet$ is one of the four basic arithmetic operations, $x \bullet y$ is the true result and $x \mathbin{\tilde{\bullet}} y$ is how it is actually computed by the machine. From this and XI follows:

XII. Let x, y be vectors in $\mathbb{R}^n$, held exactly in the machine. Let z be their componentwise sum, difference, product or quotient, and let $\bar{z}$ be the corresponding vector as actually computed. Then

$$z \simeq \bar{z}; \quad \mathrm{rp}(u).$$

The same will hold for complex vectors if "roundoff unit" is defined suitably (see Olver, p. 390).

6.4. Sum of a sequence of terms. Olver's Theorem 6.1 on the addition of complex numbers also generalizes (the proof below is essentially Olver's):

XIII. Let $a_1, \cdots, a_n$ and $\bar{a}_1, \cdots, \bar{a}_n$ be vectors in X with

$$a_j \simeq \bar{a}_j; \quad \mathrm{rp}(\alpha_j), \qquad j = 1, \cdots, n,$$

and assume that $\bar{a}_1 + \cdots + \bar{a}_n \neq 0$ and

$$\theta = \left\{\sum_j \|\bar{a}_j\|(e^{\alpha_j} - 1)\right\} \Big/ \left\|\sum \bar{a}_j\right\|$$

satisfies $0 \leqq \theta < 1$. Then

$$\sum_j a_j \simeq \sum_j \bar{a}_j; \quad \mathrm{rp}(-\ln(1-\theta)).$$

Proof. By Lemma 3.2 (the left-hand side of (3.4)),

$$\|a_j - \bar{a}_j\| \leqq \|\bar{a}_j\|(e^{\alpha_j} - 1)$$

so that by the definition of ρ_0 and the triangle inequality

$$\rho_0\left(\sum_j \bar{a}_j, \sum_j a_j\right) \leqq \sum_j \|a_j - \bar{a}_j\| \Big/ \left\|\sum_j \bar{a}_j\right\| \leqq \sum_j \|\bar{a}_j\|(e^{\alpha_j} - 1) / \left\|\sum \bar{a}_j\right\| = \theta < 1$$

and the result follows from the right-hand inequality of (3.3).

7. Summary, conclusions, questions. We have shown that each norm on a vector space X defines a relative-error metric ρ on X which is asymptotically equal to the "traditional" measure $\|y - x\|/\|x\|$ for small errors. For inner-product spaces ρ has a simple explicit formula, and for real inner-product spaces we have shown that the ρ-ball of radius δ round x is just the set of $e^T x$ where T is a linear operator of norm $\leqq \delta$.

We are able to extend many of the rp and ap results of Olver (1978) to the vector case, adding several new facts that are not relevant for the scalar case. In particular

VIII of § 6.1 relates the relative perturbations of x and b in a linear system $Ax = b$, with the condition number appearing in the expected way. The next stage of development will be to extend the theory to operators and, especially, matrices. In particular one needs a way to handle perturbations to A. (It is not clear that the obvious extension of ρ to matrices, via integrals

$$\int \frac{\|dA\|}{\|A\|},$$

interacts in a useful way with vectors or with the exponential of an operator.) If this can be done we shall have a practicable theory comparable to that of Olver for the scalar case. We emphasize that it is not intended that the ρ-distance of two vectors should often (if ever) be computed numerically: indeed except for the Euclidean norm we do not know how to do so. Olver and Wilkinson (1982) base their analysis on the ρ-metric (for scalars), but go to some lengths to replace it at the numerical level by cheaply computed but rigorous and asymptotically correct bounds on ρ. This seems the correct approach. The merit of our measure is that it exists, for any norm; and that it is related to cheaper measures by results such as Lemma 3.2.

Ziv (1982) studies what he terms "relative distance", defined by

$$d(x, y) = \frac{\|x - y\|}{\max\{\|x\|, \|y\|\}}.$$

In our notation, this equals $\min\{\rho_0(x, y), \rho_0(y, x)\}$; thus by Lemma 3.2 it is related to ρ by the inequalities

$$e^{\rho} - 1 \geqq d \geqq 1 - e^{-\rho}.$$

The measure d is not a metric in a general normed space (e.g. $d(x, z) \leqq d(x, y) + d(y, z)$ fails in $\mathbb{R}^2$ with the 1-norm if $x = (1, 0)$, $y = (1, 1)$, and $z = (0, 1)$), but Ziv proves it is a metric in an inner-product space. Since it is so simple to compute, relative distance should prove a useful numerical tool though it lacks some of the elegant properties of the metric discussed in this paper.

Finally, here are some problems, which seem to range from the amusing to the deep.

Q1. The author in Pryce (1981) introduced the notation $1(\delta)$ (for real $\delta > 0$) to mean "some number e^t where $|t| \leqq \delta$," or in the more precise language suitable for this paper, the *set*

$$1(\delta) = \{e^t : |t| \leqq \delta\}$$

so that $y \simeq x$; rp(δ) is equivalent to $y \in 1(\delta)x$. In the vector case $1(\delta)$ should be a set of operators. Should we define

$$1(\delta) = \{e^T : T \in \mathbb{B}(x), \|T\| < \delta\} \text{ ?}$$

Or in view of Theorem 5.3 should the definition involve products of e^{T_i} with $\sum \|T_i\| \leqq \delta$?

Q2. Find an explicit formula for ρ when X is $\mathbb{R}^2$ with the ∞-norm. Extend to $\mathbb{R}^n$.

Q3. In an inner product space, given $x_1, \cdots, x_k$ and $y_1, \cdots, y_k$ with $\rho(x_i, y_i) = \delta_i$, what can you say about the smallest norm of a T (if any exists) such that $e^T x_i = y_i$ for each i? If this is too hard, what about the cases

(a) where $k = 2$;

(b) where the x_i, and the y_i, are orthogonal sets;

(c) in the limit as the $\delta_i \to 0$;

(d) where instead of e^T we consider products of e^{T_i} as in Q1.

Answers to Q3 would help to show how the ρ-metric interacts with perturbations to matrices.

REFERENCES

F. W. J. OLVER (1978), *A new approach to error arithmetic*, this Journal, 15, pp. 368–393.

——— (1982), *Further developments of rp and ap error analysis*, IMA J. Numer. Anal., 2, pp. 249–274.

F. W. J. OLVER AND J. H. WILKINSON (1982), *A posteriori error bounds for Gaussian elimination.* IMA J. Numer. Anal., 2, pp. 377–406.

J. D. PRYCE (1981), *Roundoff error analysis with fewer tears*, Bull. Inst. Maths. Applics., 17, pp. 40–47.

F. STUMMELL (1981), *Forward error analysis of Gaussian elimination*, Maths. Comp., 37, pp. 435–473.

A. ZIV (1982), *Relative distance—an error measure in roundoff error analysis*, Maths. Comp., 39, pp. 563–569.