Journal of Lie Theory
Volume **11** (2000) 231–254

Lie groups and error analysis

Jeremy Schiff and Steve Shnider

Communicated by K. H. Hofmann and P. Olver

Abstract. A new approach to error analysis is introduced, based on the observation that many numerical procedures can be interpreted as computations of products in a suitable Lie group. The absence of an additive error law for such procedures is intimately related to the nonexistence of bi-invariant metrics on the relevant groups. Introducing the notion of an almost $\mathrm{Inn}(G)$ invariant metric (a left invariant, almost $\mathrm{Inn}(G)$ invariant metric can be constructed on any locally compact connected group having a countable basis for its identity neighborhoods), we show how error analysis can nevertheless be done for such procedures. We illustrate for what we call "scalar calculations without writing to memory"; the Horner algorithm for evaluation of a polynomial is such a calculation, and we give explicit error bounds for a floating point implementation of the Horner algorithm, and demonstrate their usefulness numerically. A left invariant, almost $\mathrm{Inn}(G)$ invariant metric on a group induces a metric on a homogeneous space of the group with useful properties for error analysis; treating $\mathbf{R}$ as a homogeneous space of the group of affine transformations of $\mathbf{R}$ we compute a new metric that unifies absolute and relative error.
Mathematics Subject Classification Numbers. Primary: 65G99. Secondary: 22E99, 65G05.

1. Introduction

The aim of this paper is to introduce a new way of approaching error analysis. Error analysis is the unpopular subject of examining the effects on numerical algorithms of inaccuracies in input data and the fact that digital computers cannot do exact arithmetic. Our method revolves around two central observations:

1. *Many numerical algorithms can be interpreted as the computation of the product of a (typically large) number of elements in a suitable Lie or matrix group (or, in greater generality, as the computation of the action of such a product on a point in a manifold on which the relevant group acts).* The example of this on which we shall concentrate in the current paper is the class of calculations that can be carried out on a pocket calculator which performs just the 4 basic arithmetic operations, and has no memory. Starting with some number on the screen, all we are allowed to do is to add another number to it, or multiply it by another number (subtractions and divisions are regarded as just additions and multiplications by inverses); at no

ISSN 0949–5932 / \$2.50

stage are we allowed to record an intermediate result for later use. We call such calculations *scalar calculations without writing to memory.* The Horner algorithm for evaluation of the polynomial $p(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$ in nested form $a_n + x(a_{n-1} + x(a_{n-2} + x(\ldots)))$ is precisely the calculation "start with a_0, multiply by x, add a_1, multiply by x, add $a_2, \ldots$, multiply by x, add a_n", and is an example of a scalar calculation without writing to memory. Denoting the affine transformation $\xi \mapsto a\xi + b$ by $g[a,b]$, $a,b \in \mathbf{R}$, $a \neq 0$, the Horner algorithm (assuming $x \neq 0$) can evidently be written

$$p(x) = g_n \circ g_{n-1} \circ \ldots \circ g_1(a_0) \ ,$$
$$\text{where} \quad g_1 = g[x,a_1],\ g_2 = g[x,a_2], \ldots, g_n = g[x,a_n].$$

We see that *the Horner algorithm applied to x is precisely the computation of a product of elements in the affine group (depending on the parameters a_i and x) acting on a_0.* The same is true for any scalar calculation without writing to memory.

Many other numerical algorithms can be viewed in a similar light, including the Newton-Raphson method for solution of nonlinear equations, Gaussian elimination (with or without pivoting) for matrix inversion, and all the standard methods for integration of ordinary differential equations. In the current paper we only explore in detail the case of scalar calculations without writing to memory; we do however develop a theory that we expect will be applicable in some generality.

2. The effect of flawed data on such algorithms is that instead of asking the computer to find the product $g'_n g'_{n-1} \ldots g'_1$ that is needed, we in fact request the product $g_n g_{n-1} \ldots g_1$ of group elements $g_1, \ldots, g_n$ which are in some sense "close" to the correct ones $g'_1, \ldots, g'_n$. The effect of rounding errors is that the computer does not compute $g_n g_{n-1} \ldots g_1$, but rather $k_n g_n k_{n-1} g_{n-1} \ldots k_1 g_1$, where the group elements $k_1, \ldots, k_n$ are typically "close" to the identity. The latter statement about rounding errors requires verification on a case-to-case basis, and once again, we shall only explore here the case of scalar calculations without writing to memory; we have indications that it is true in some generality, but in some cases it might be necessary to enlarge the group. Note that if we take $k'_1 = \ldots k'_n = I$, where I is the identity in the group, we can write the desired product $k'_n g'_n k'_{n-1} g'_{n-1} \ldots k'_1 g'_1$; thus we see that rounding error is basically on the same footing as flawed data error. In the general theory section of this paper we therefore just treat flawed data error. In the example of scalar calculations without writing to memory which we treat in detail, we consider the effect of both kinds of error.

Given the two basic observations above, the fundamental quantitive problem of error analysis is to bound the discrepancy of the products $g'_n g'_{n-1} \ldots g'_1$ and $g_n g_{n-1} \ldots g_1$ given some measure of the discrepancy of g_1 from g'_1, g_2 from g'_2 etc. We are led naturally, therefore, to look at metrics on Lie groups. Section 2 of this paper develops the general theory of metrics on Lie groups and on the associated homogeneous spaces (manifolds on which there is a smooth, transitive group action). The most significant results are as follows:

I) If the group admits a bi-invariant metric d, then there is an additive error law $d(g'_n g'_{n-1} \ldots g'_1, g_n g_{n-1} \ldots g_1) \leq d(g'_1, g_1) + d(g'_2, g_2) + \ldots + d(g'_n, g_n)$.

II) General groups do not admit bi-invariant metrics, but we can construct left invariant metrics with a property we call "almost Inn(G) invariance"; for such a

metric d there is an error law of the form error law $d(g'_n g'_{n-1} \cdots g'_1, g_n g_{n-1} \cdots g_1) \leq d(g'_1, g_1) + \rho(g_1) d(g'_2, g_2) + \rho(g_2 g_1) d(g'_3, g_3) + \ldots + \rho(g_{n-1} \cdots g_2 g_1) d(g'_n, g_n)$, where ρ is a scalar function on the group, determined by d.
III) A left invariant, almost Inn(G) invariant metric on a Lie group induces a metric on associated homogeneous spaces, which also has suitable properties for the purpose of error analysis.

In Section 3 we compute a left invariant, almost Inn(G) invariant metric for the group of affine transmformations of the line, and other associated quantities. The induced metric on the line, treated as a homogeneous space of the group, is an interesting new metric that unifies absolute and relative error. Section 3 prepares for Section 4, where we use the metric on the group of affine transformations to derive an error bound for scalar calculations without writing to memory. In Section 5, this is specialized to give an error bound for Horner's algorithm, and some numerical experiments are performed, showing the bound to be quite reasonable. From a technical standpoint the calculations involved in obtaining the bound are hard, and the bound displays no particular advantages over those obtained by other methods [9]. We feel, however, that from a conceptual standpoint the method we propose is easier, and expect this method to be easier to apply in more complicated situations where ad hoc methods fail. Section 6 contains concluding remarks. The appendix, written by Karl H. Hofmann, presents a theorem describing a large class of topological groups which have left invariant, almost Inn(G) invariant metrics. In the body of the paper we have restricted attention to Lie groups, for which we can explicitly construct such a metric, which can then be used in calculations.

2. Metrics on Lie Groups and Homogeneous Spaces

As explained in the introduction, we wish to examine the product of approximations to a set of elements of a Lie group as an approximation to the product of those elements. Let G be the Lie group, which we assume to be finite dimensional and connected. We need a notion of closeness in G, so we assume we are given a metric on G, i.e. a function $d : G \times G \to \mathbf{R}$ such that for all $g, h, k \in G$

$$\begin{aligned} &d(g,h) \geq 0 \quad \text{and} \quad d(g,h) = 0 \;\Leftrightarrow\; g = h \\ &d(g,h) = d(h,g) \\ &d(g,h) + d(h,k) \geq d(g,k). \end{aligned}$$

We assume that the metric topology and the Lie group topology on G coincide (i.e. the identity map, as a map from G with its given topology to G with the metric topology, is a homeomorphism). We wish to consider $d(g'_n g'_{n-1} \cdots g'_1, g_n g_{n-1} \cdots g_1)$, where we are given $g_1, \ldots, g_n$, and $d(g'_i, g_i)$ for $i = 1, \ldots, n$, but we do not know $g'_1, \ldots, g'_n$.

We will take d to be a left invariant metric, i.e. we will impose that for all $g, h, k \in G$

$$d(kg, kh) = d(g,h). \tag{1}$$

The rationale for this is that if the last group element in our product is known exactly, i.e. $g_n = g'_n$, then we want it not to contribute to the error estimate. Now it would also be desirable that if the first element in our product were known exactly then it should not contribute to the error estimate. So it is tempting to impose right invariance of d as well. Unfortunately, the groups we wish to consider include groups that do not admit bi-invariant (i.e. left and right invariant) metrics. Fortunately, though, there are left invariant metrics with sufficiently good properties that some error analysis can be done.

Definition 1. Suppose $d: G \times G \to \mathbf{R}$ is a metric on G and $\alpha \in \mathrm{Aut}(G)$. Define the *norm of* α *with respect to* d by

$$\|\alpha\|_d := \inf\{C \in \mathbf{R} : \ \forall g, h \in G, \ d(\alpha(g), \alpha(h)) \leq C d(g, h)\} , \tag{2}$$

where $\|\alpha\|_d := \infty$ if the set of such C is empty. Clearly $0 < \|\alpha\|_d$ and $\|\beta \circ \alpha\|_d \leq \|\beta\|_d \|\alpha\|_d$. We call an automorphism α *bounded with respect to* d if $\|\alpha\|_d < \infty$.

Definition 2. Suppose $d: G \times G \to \mathbf{R}$ is a metric on G. We write $\mathrm{Aut}_d(G)$ for the set of all automorphisms of G that are bounded with respect to d. If Γ is a subgroup of $\mathrm{Aut}(G)$, we say d is *almost* Γ *invariant* if $\Gamma \subseteq \mathrm{Aut}_d(G)$.

Note 1. Let I_g denote the inner automorphism $I_g : h \mapsto g^{-1} h g$ and $\mathrm{Inn}(G)$ the subgroup of $\mathrm{Aut}(G)$ of inner automorphisms. A metric d is almost $\mathrm{Inn}(G)$ invariant if for all $g \in G$ we can find $\rho(g)$ such that $d(I_g(h), I_g(k)) \leq \rho(g) d(h, k)$. Given an almost $\mathrm{Inn}(G)$ invariant metric d, define $\rho_c(g) := \|I_g\|_d$, and call $\rho_c : G \to \mathbf{R}$ *the optimum* ρ*–function associated with* d. Any other function $\rho : G \to \mathbf{R}$ such that $d(I_g(h), I_g(k)) \leq \rho(g) d(h, k)$, $\forall h, k \in G$ is called an *admissible* ρ*–function.*

Note 2. It is well-known that a left invariant metric on a topological group G is equivalent to a function $\|\cdot\| : G \to \mathbf{R}$ satisfying
(1) $\forall g \in G$, $\|g\| \geq 0$, and $\|g\| = 0$ if and only if $g = I$,
(2) $\forall g \in G$, $\|g\| = \|g^{-1}\|$,
(3) $\forall g, h \in G$, $\|g\| + \|h\| \geq \|gh\|$.
The correspondence is via $d(g, h) = \|g^{-1} h\|$, $\|g\| = d(I, g)$. For a left invariant metric, the norm of an automorphism α, if it is bounded, is then the smallest number $\|\alpha\|_d$ such that $\|\alpha(g)\| \leq \|\alpha\|_d \|g\| \ \forall g \in G$.

We choose the metric on our Lie group to be left invariant and almost $\mathrm{Inn}(G)$ invariant. A bi-invariant metric is left invariant and almost $\mathrm{Inn}(G)$ invariant, with optimal ρ–function $\rho_c = 1$. For orientation, we state here the following three facts, which we discuss in more detail later:

1. Abelian Lie groups, compact topological groups such that $\{I\}$ is the intersection of a countable family of open sets [6], and direct products thereof, all admit bi-invariant metrics.
2. A necessary and sufficient condition for a connected Lie group to admit a bi-invariant metric is that the image of the adjoint representation has compact closure.
3. Every locally compact connected group having a countable basis for its identity neighborhoods admits a left invariant, almost $\mathrm{Inn}(G)$ invariant metric. See the appendix by Karl H. Hofmann.

With this preparation, we state and prove the main theorem of this section, which we consider to be of critical importance in error analysis:

Theorem 1. Let d be a left invariant, almost $\mathrm{Inn}(G)$ invariant metric on G, and let ρ be an admissible ρ-function for d. Then for all $g_1, \ldots, g_n, g'_1, \ldots, g'_n \in G$,

$$d(g'_n \cdots g'_1, g_n \cdots g_1) \leq \rho(h_n) d(g'_n, g_n) + \ldots + \rho(h_2) d(g'_2, g_2) + \rho(h_1) d(g'_1, g_1), \tag{3}$$

where

$$\begin{aligned} h_1 &= I \\ h_2 &= g_1 \\ h_3 &= g_2 g_1 \\ &\vdots \\ h_n &= g_{n-1} g_{n-2} \cdots g_2 g_1. \end{aligned}$$

The h_i that appear in this result are intermediate products that are found in computing the product $g_n g_{n-1} \ldots g_1$.

Proof. The proof by induction is straightforward. The theorem is trivially true for $n = 1$. The inductive step is proved by the following two line calculation, in which we use first the triangle inequality, and then left invariance and almost $\mathrm{Inn}(G)$ invariance:

$$\begin{aligned} d(g'_n \cdots g'_1, g_n \cdots g_1) &\le d(g'_n g'_{n-1} \cdots g'_1, g'_n g_{n-1} \cdots g_1) + d(g'_n g_{n-1} \cdots g_1, g_n g_{n-1} \cdots g_1) \\ &\le d(g'_{n-1} \cdots g'_1, g_{n-1} \cdots g_1) + \rho(h_n) d(g'_n, g_n). \qquad \blacksquare \end{aligned}$$

Note 1. Instead of using almost $\mathrm{Inn}(G)$ invariance to obtain $\rho(h_n)d(g'_n, g_n)$ in the above calculation, we could have used it to obtain $\rho(h_{n+1})d(g'_n g_n^{-1}, I)$, where $h_{n+1} = g_n g_{n-1} \ldots g_2 g_1$. By left invariance and symmetry, $d(g'_n g_n^{-1}, I) = d(g_n'^{-1}, g_n^{-1})$, and it follows that in the statement of the theorem we could replace $\rho(h_i)d(g'_i, g_i)$ by $\rho(h_{i+1})d(g_i'^{-1}, g_i^{-1})$.

Note 2. For a bi-invariant metric, $d(g'_i, g_i) = d(g_i'^{-1}, g_i^{-1})$, and we can take $\rho = 1$.

Note 3. In the application we will examine in Section 4, both g'_i and g_i belong to a subgroup H of G. We can then replace $\rho(h_i)$ in the theorem by $\rho_H(h_i)$, where ρ_H is any function such that for all $h, k \in H$, $d(g^{-1}hg, g^{-1}kg) \le \rho_H(g)d(h, k)$. This can improve the effectiveness of Theorem 1.

Two examples of Theorem 1 are very well known. In both cases the group is abelian and the metric bi-invariant:

Example 1. Let $G = \mathbf{R}$, with addition as group operation. The metric $d(g, g') = |g - g'|$ is bi-invariant, and thus we can take $\rho = 1$. The theorem is the generalized triangle inequality

$$\left| \sum_{i=1}^{n} g'_i - \sum_{i=1}^{n} g_i \right| \le \sum_{i=1}^{n} |g'_i - g_i|,$$

known in error analysis by the mnemonic "the absolute error of the sum is less than or equal to the sum of the absolute errors".

Example 2. Let $G = \mathbf{R}^+$, with multiplication as group operation. The two metrics

$$d_{\mathrm{Olver}}(g, g') = |\ln g' - \ln g| \tag{4}$$

$$d_{\mathrm{Ziv}}(g, g') = \frac{|g' - g|}{\max(g', g)} \tag{5}$$

are both bi-invariant. These metrics were introduced in [10, 11] and [19] respectively as giving alternatives to the standard notion of relative error. The rationale for having such alternatives is that the mnemonic "the relative error of the product is less than or equal to the sum of the relative errors" is only "more or less true" when the standard notion of relative error is used, but it is exact when the Olver or Ziv notions of relative error are used [10, 11, 19].

The bulk of this paper is devoted to the consequences of Theorem 1 for the case where G is the group of proper affine transformations of $\mathbf{R}$, and d is a suitable metric we will construct in section 3. The rest of this section, however, is devoted to (1) construction of a class of left invariant, almost $\mathrm{Inn}(G)$ invariant metrics on a general Lie group G, and computation of admissible ρ–functions for them, and (2) construction of an associated metric on homogeneous spaces.

2.1. Existence of left invariant, almost $\mathrm{Inn}(G)$ invariant metrics.

A fundamental result for Lie groups is the existence of a set of left invariant one forms $\omega^1, \ldots, \omega^d$, $d = \dim G$, that form a basis for the cotangent space at each point of G. By left invariant we mean as usual that if for an element $g \in G$, we define the left action of g on G, $L_g : G \to G$, by

$$L_g(k) = gk \ , \qquad k \in G,$$

then $L_g^* \omega^i = \omega^i$, for all $g \in G$ and $i = 1, \ldots, d$. We can use the left invariant one forms to construct left invariant Riemannian metrics on G of the form

$$m = \sum_{i,j} M_{ij} \omega^i \otimes \omega^j, \tag{6}$$

where M is any symmetric, positive definite $d \times d$ matrix.

A Riemannian metric on a differentiable manifold defines a line element $ds(v) = \sqrt{m(v,v)}$, and hence for any smooth path in G $\gamma : [0,1] \to G$ we can define a length functional by

$$\ell(\gamma) = \int_0^1 \gamma^*(ds) = \int_0^1 ds(\dot\gamma(t)).$$

This, in turn, allows us to define a point metric, by

$$d(g,h) = \inf\ \{\ell(\gamma) | \gamma(0) = g,\, \gamma(1) = h\},$$

where the infimum is over all smooth paths connecting g and h. That a left invariant Riemannian metric induces a left invariant point metric follows trivially from the fact that left multiplication by k establishes a length-preserving one to one correspondence between the smooth paths from g to h and the smooth paths from kg to kh.

Theorem 2. The metrics induced by the Riemannian metrics (6) are almost $\mathrm{Inn}(G)$ invariant.

Proof. Let $I_g : G \to G$ be conjugation by g,

$$I_g(k) = g^{-1} k g \ , \qquad k \in G,$$

and $R_g : G \to G$ be right multiplication by g,

$$R_g(k) = kg \ , \qquad k \in G.$$

We have $I_g = L_{g^{-1}} R_g$. Since right translation and left translation commute, the forms $I_g^* \omega^i = R_g^* \omega_i$ are all left invariant. The forms ω^i are a basis for the left invariant forms, so we have

$$R_g^* \omega^i = \sum_{j=1}^{d} R_j^i(g) \omega^j, \tag{7}$$

for some matrix $R_j^i(g)$. In fact the correspondence $g \mapsto R(g)^T$ gives the coadjoint representation of G. Thus

$$I_g^* m = R_g^* m = \sum_{i,j} \left(R(g)^T M R(g) \right)_{ij} \omega^i \otimes \omega^j. \tag{8}$$

The positive definite matrices M and $R(g)^T M R(g)$ represent two inner products on the space $\mathbf{R}^d$. Thus they define equivalent norms related by

$$\rho(g) = \max_{X \in \mathbf{R}^d} \sqrt{\frac{X^T R(g)^T M R(g) X}{X^T M X}} \,. \tag{9}$$

This is a uniform bound on the increase in the length of a path under the action of I_g. Almost $\mathrm{Inn}(G)$ invariance follows at once, and further, we deduce that $\rho(g)$ is an admissible ρ–function for the metric induced by the Riemannian metric m given by (6). ∎

Thus we see that a finite dimensional, connected Lie group does admit a left invariant, almost $\mathrm{Inn}(G)$ invariant metric. In the appendix, a more general result on the existence of such metrics is proved. For abelian groups, any left invariant metric is bi-invariant. For compact Lie groups, the image of the coadjoint representation on the dual space to the Lie algebra of G, $g \mapsto R(g)^T$, used above, is a subgroup of the orthogonal group of a suitable inner product. Letting M be the matrix of this inner product we have

$$R(g)^T M R(g) = M, \quad \text{therefore} \quad \rho(g) = 1,$$

so the group admits a bi-invariant metric. However, if the closure of the image of the coadjoint representation is not compact, then we cannot find a suitable orthogonal group containing $R(g)$ for all $g \in G$ and thus for any choice of M, we will have $\rho(g) \neq 1$.

We conclude the discussion of metrics on Lie groups with a simple example.

Example 3. If G is a real matrix group with typical element X, the matrix $X^{-1} dX$ is a matrix of left invariant one forms. We can use this to construct the Riemannian metrics

$$m = \sum_{(i,j)(k,l)} M_{(i,j),(k,l)} (X^{-1} dX)_{(i,j)} \otimes (X^{-1} dX)_{(k,l)}.$$

which will be nondegenerate for the appropriate choice of M. A sufficient, though not necessary, condition is that M define a positive definite inner product on the space of all $d \times d$ matrices. For example $M_{(i,j),(k,l)} = \delta_{i,k} \delta_{j,l}$, which gives the form

$$\mathrm{Tr} \left((X^{-1} dX)^T \otimes (X^{-1} dX) \right).$$

For a complex group we replace the second factor in the tensor product with the conjugate to get a positive definite form. We see that for matrix groups, where the coadjoint representation acts by conjugation, the metric is bi-invariant if the conjugation acts as an orthogonal transformation on the space of matrices, with the inner product defined by $M_{(i,j),(k,l)}$.

2.2. The Associated Metrics on Homogeneous Spaces.

Theorem 1 allows us to bound $d(g'_n \dots g'_1, g_n \dots g_1)$ given knowledge of $g_1, \dots, g_n$ and bounds on $d(g_1, g'_1), \dots, d(g_n, g'_n)$. Frequently in practice, we are given a G homogeneous space V on which G acts (transitively) from the left, and we are actually interested in comparing $g'_n \dots g'_1 v'$ with $g_n \dots g_1 v$ for some elements $v', v \in V$. For this we need a metric on V.

Theorem 3. A left invariant metric d on G induces a metric d_V on a G homogeneous space V, given by

$$d_V(v,w) = \inf_{g \in G \,:\, gv=w} d(I,g) \ . \tag{10}$$

Proof.

1. $d_V(v,w) \geq 0$ is trivial. The implication $d_V(v,w) = 0 \ \Rightarrow \ v = w$ is proved as follows. If $d_V(v,w) = 0$, then for all $\epsilon > 0$ we can find $g(\epsilon) \in G$ such that $g(\epsilon)v = w$ and $d(I, g(\epsilon)) \leq \epsilon$. Thus given a sequence $\epsilon_1, \epsilon_2, \dots$ such that $\lim_{r\to\infty} \epsilon_r = 0$, we can construct a corresponding sequence $g(\epsilon_1), g(\epsilon_2), \dots$ that converges to $I \in G$ and such that $g(\epsilon_i)v = w$ (since the metric topology is consistent with the given topology on G, convergence in the metric topology is equivalent to convergence in the given topology on G). By continuity of the G action on V it follows that $Iv = w$.

2. $d_V(v,w) = d_V(w,v)$ is straightforward:

$$\begin{aligned} d_V(w,v) &= \inf_{h\in G \,:\, hw=v} d(I,h) \\ &= \inf_{g\in G \,:\, gv=w} d(I,g^{-1}) \quad \text{(substituting } h = g^{-1}) \\ &= \inf_{g\in G \,:\, gv=w} d(I,g) \quad \text{(left invariance and symmetry of } d) \\ &= d_V(v,w) \ . \end{aligned}$$

3. To prove the triangle inequality. Take $v, w, x \in V$. For all $\epsilon > 0$ we can find g, h such that

$$\begin{aligned} gv = w \ , \qquad & d(I,g) < d_V(v,w) + \tfrac{1}{2}\epsilon \ , \\ hw = x \ , \qquad & d(I,h) < d_V(w,x) + \tfrac{1}{2}\epsilon \ . \end{aligned}$$

Since $hgv = x$, we have

$$\begin{aligned} d_V(v,x) &\leq d(I,hg) \\ &\leq d(I,h) + d(h,hg) \quad \text{(triangle inequality)} \\ &= d(I,h) + d(I,g) \quad \text{(left invariance)} \\ &< d_V(v,w) + d_V(w,x) + \epsilon \ . \end{aligned}$$

So $d_V(v,x) \leq d_V(v,w) + d_V(w,x)$. ■

Note 1. This construction differs from the standard construction of a metric on homogeneous space (with a left G action) starting from a *right* invariant metric on the group; see, for example, [2].

Note 2. The metric on a homogeneous space defined by (10) is *not* left invariant, in the sense that in general

$$d_V(gv, gw) \neq d_V(v, w).$$

If, however, d is almost $\mathrm{Inn}(G)$ invariant as well as left invariant, then we do have

$$d_V(gv, gw) \leq \rho(g^{-1}) d_V(v, w) \quad \forall g \in G\ . \tag{11}$$

This is proved in the proof of Theorem 4 below.

The following theorem shows that if the metric d is both left invariant and almost $\mathrm{Inn}(G)$ invariant then we can bound $d_V(g'v', gv)$ in terms of $d(g', g)$ and $d_V(v'v)$:

Theorem 4. Suppose the metric d on G is left invariant and almost $\mathrm{Inn}(G)$ invariant, and let ρ be an admissible ρ–function for d. Then for all $g, g' \in G$ and $v, v' \in V$

$$d_V(g'v', gv) \leq \rho(g^{-1}) \left(d(g', g) + d_V(v', v)\right)\ . \tag{12}$$

Proof. By the triangle inequality

$$d_V(g'v', gv) \leq d_V(g'v', gv') + d_V(gv', gv).$$

Using in turn the definition of d_V and the properties of d we have

$$\begin{aligned} d_V(g'v', gv') &\leq d(I, g'g^{-1}) \\ &\leq \rho(g^{-1}) d(g', g)\ . \end{aligned}$$

Similarly, we have

$$\begin{aligned} d_V(gv', gv) &= \inf_{h \in G\ :\ hgv' = gv} d(I, h) \\ &= \inf_{k \in G\ :\ kv' = v} d(I, gkg^{-1}) \\ &\leq \rho(g^{-1}) \inf_{k \in G\ :\ kv' = v} d(I, k) \\ &= \rho(g^{-1})\ d_V(v', v) \end{aligned} \qquad \blacksquare$$

Note. In practice a slightly stronger result than this is necessary. Rounding error typically means that when asked to compute gv, a computer computes gkv for some element $k \in G$ close to the identity. To estimate the error in this we use, in turn, Theorem 4, the triangle inequality, and the definition of d_V, to get

$$\begin{aligned} d_V(g'v', gkv) &\leq \rho(g^{-1}) \left(d(g', g) + d_V(v', kv)\right) \\ &\leq \rho(g^{-1}) \left(d(g', g) + d_V(v', v) + d_V(v, kv)\right) \\ &\leq \rho(g^{-1}) \left(d(g', g + d_V(v', v) + d(I, k)\right)\ . \end{aligned} \tag{13}$$

3. The Group of Proper Affine Transformations of the Line

As explained in the introduction, we wish to apply the theory of the preceeding section to provide an error bound for scalar calculations without writing to memory. For this we need a specific left invariant, almost G invariant metric on the group of (proper) affine transformations of the line, and an admissible ρ–function for this metric.

A proper affine transformation of $\mathbf{R}$ is a transformation $y \mapsto \alpha y + \beta$ $(\alpha > 0)$. These transformations form a connected, nonabelian, noncompact group, with a standard representation using 2×2 matrices, arising from the fact that $y \mapsto \alpha y + \beta$ can be rewritten

$$\begin{pmatrix} y \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} \alpha & \beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y \\ 1 \end{pmatrix}. \tag{14}$$

We use the method of example 3 in the previous section to construct a left invariant metric on the group. Writing

$$X = \begin{pmatrix} \alpha & \beta \\ 0 & 1 \end{pmatrix}, \tag{15}$$

we have

$$X^{-1}dX = \begin{pmatrix} \frac{d\alpha}{\alpha} & \frac{d\beta}{\alpha} \\ 0 & 0 \end{pmatrix},$$

giving us a basis of left invariant one forms on the group

$$\omega^1 = \frac{d\alpha}{\alpha}, \qquad \omega^2 = \frac{d\beta}{\alpha}.$$

Thus we have a left invariant Riemannian metric

$$\frac{d\alpha \otimes d\alpha + d\beta \otimes d\beta}{\alpha^2}. \tag{16}$$

If we write

$$g = \begin{pmatrix} \alpha' & \beta' \\ 0 & 1 \end{pmatrix},$$

it is straightforward to compute the action of right translation on the ω^i, giving, in the notation of equation (7):

$$R(g) = \begin{pmatrix} 1 & 0 \\ \frac{\beta'}{\alpha'} & \frac{1}{\alpha'} \end{pmatrix}. \tag{17}$$

As a check we observe $R(g_1)R(g_2) = R(g_2 g_1)$. Once we have identified $R(g)$, we use equation (9) to deduce that the function

$$\rho(g) = \sqrt{\frac{1}{2}\left(\left(1 + \frac{1}{\alpha'^2} + \frac{\beta'^2}{\alpha'^2}\right) + \sqrt{\left(1 + \frac{1}{\alpha'^2} + \frac{\beta'^2}{\alpha'^2}\right)^2 - \frac{4}{\alpha'^2}}\right)} \tag{18}$$

is an admissible ρ function for the metric induced by the Riemannian metric given in equation (16).

The simplest way to compute the induced point metric is using a change of coordinates. We have identified the group as the half plane $\alpha > 0$, and the metric (16) is the well known hyperbolic metric on the upper half plane. Defining x, y via the conformal map

$$i(\beta + i\alpha) = \frac{(x+iy)-1}{(x+iy)+1} , \tag{19}$$

the half plane is mapped onto the unit disk $x^2 + y^2 < 1$ with Riemannian metric

$$\frac{4(dx \otimes dx + dy \otimes dy)}{(1 - x^2 - y^2)^2} . \tag{20}$$

By the evident rotational symmetry of this, geodesics through the origin are straight lines, and the distance from the origin to a point on the circle $x^2 + y^2 = r^2$ is simply

$$\ln\left(\frac{1+r}{1-r}\right) .$$

Returning to the half plane model, this gives

$$d(I, X) = \ln\left(\frac{1+r}{1-r}\right) , \qquad r^2 = \frac{(\alpha-1)^2 + \beta^2}{(\alpha+1)^2 + \beta^2} . \tag{21}$$

This and left invariance fully determine the metric.

As mentioned in Note 3 after Theorem 1, if g_i, g_i' are both known to belong to a certain subgroup H in G, then we can replace $\rho(h_i)$ in Theorem 1 by $\rho_H(h_i)$ where ρ_H is any function satisfying $d(ghg^{-1}, gkg^{-1}) \le \rho_H(g) d(h,k)\ \forall h, k \in H$. In our application, the pairs g_i, g_i' will all be taken either from the subgroup H_1 of translations (transformations $y \mapsto y + \beta$), or from the subgroup H_2 of scalings (transformations $y \mapsto \alpha y$, $\alpha > 0$). Since both of these are one parameter subgroups it is quite straightforward to compute "optimum subgroup ρ–functions" by the formula

$$\rho_{H_r}(g) \;=\; \inf_{h,k \in H_r : h \neq k} \frac{d(g^{-1}hg, g^{-1}kg)}{d(h,k)} , \qquad\qquad r = 1, 2.$$

Writing $g = \begin{pmatrix} \alpha & \beta \\ 0 & 1 \end{pmatrix}$ we find

$$\rho_{H_1}(g) \;=\; \max_{b \ge 0} \frac{\ln(\sqrt{4\alpha^2 + b^2} + b) - \ln(\sqrt{4\alpha^2 + b^2} - b)}{\ln(\sqrt{4 + b^2} + b) - \ln(\sqrt{4 + b^2} - b)} \;=\; \max(1, \alpha^{-1}) . \tag{22}$$

Comparing with (18), we see that when β is large it is strongly preferable to use ρ_{H_1} in place of ρ. For the subgroup of scalings we find

$$\rho_{H_2}(g) = \sqrt{1 + \frac{\beta^2}{\alpha^2}} . \tag{23}$$

It is an interesting exercise to compute the metric induced by the metric d on the line $\mathbf{R}$, treated as a homogeneous space of G, following Theorem 3 in the previous section. We have

$$d_V(v, w) = \inf_{\alpha, \beta \,:\, \alpha v + \beta = w} \; d\left(I, \begin{pmatrix} \alpha & \beta \\ 0 & 1 \end{pmatrix}\right)$$

The infimum is achieved by the geodesic through the point $(\beta,\alpha) = (0,1)$ and perpendicular to the line $\alpha v + w = \beta$. Since the metric is conformally equivalent to the standard flat metric, perpendicular has the standard meaning, and the required geodesic arc is centered at the point of intersection of the line $\alpha v + w = \beta$ with the line $\alpha = 0$. The arc is parametrized by

$$(\beta,\alpha) = \left(w + \sqrt{1+w^2}\cos t, \sqrt{1+w^2}\sin t\right).$$

It is a simple business to compute the length:

$$d_V(v,w) = \ln\left(\frac{\sqrt{1+w^2}+w}{\sqrt{1+v^2}+v}\right) . \tag{24}$$

Writing $\epsilon = w - v$, the first order Taylor approximation gives

$$d_V(v,w) = \frac{|\epsilon|}{\sqrt{1+v^2}} + O(\epsilon^2) . \tag{25}$$

(24) defines an interesting, novel metric on $\mathbf{R}$. Suppose $|w - v|$ is small. From (25) we see that if $v << 1$ then $d_V(v,w)$ is close to the absolute error of w as an approximation to v, but if $v >> 1$ then $d_V(v,w)$ is close to the relative error of w as an approximation to v. The metric $d_V(v,w)$ "extrapolates" between absolute error in one limit and relative error in another limit.

We are unaware of such a metric having been written down before, but the need for such a metric has most definitely been identified. A standard method of error control when numerically approximating a quantity Y', say, is to compute two approximations Y_1 and Y_2 to Y', using a better method for Y_2 than for Y_1, and to estimate the error in Y_1 by comparing Y_1 and Y_2. The need arises to write a criterion to decide when Y_1 and Y_2 are acceptably close. An absolute error criterion, $|Y_1 - Y_2| \le \epsilon_a$ is acceptable when the magnitudes of Y_1 and Y_2 are small; but for large magnitudes a relative error criterion $|Y_1 - Y_2| \le \epsilon_r |Y_1|$ is more appropriate. The standard compromise [16] is to use a criterion of the form

$$|Y_1 - Y_2| \le \epsilon_a + \epsilon_r |Y_1|,$$

where ϵ_a and ϵ_r are independent "tolerances", often taken equal. Clearly such a criterion could be replaced by a criterion of the form $d_V(Y_1, Y_2) \le \epsilon$. We note that in our approach we can change the initial Riemannian metric on the group, and this provides a family of metrics d_V on $\mathbf{R}$ wih different "weightings" between absolute and relative error, corresponding to different possible choices of the ratio $\epsilon_a : \epsilon_r$ above.

We expect that the metric d_V on $\mathbf{R}$ might also be a good metric to use in regression analysis when there are sources of both absolute and relative error[1].

4. Scalar Calculations Without Writing to Memory

In this section we apply the theory of the previous sections to obtain an error bound for *scalar calculations without writing to memory.* We start with 0, and

[1]J.S. wishes to acknowledge J.D.Klein and D.A.Kessler for this suggestion

successively "apply" (i.e. add, subtract, multiply by or divide by) a sequence of inputs, until the output is obtained. More formally, we define a sequence y_i', $1 \le i \le n+1$ by

$$\begin{aligned} y_1' &= 0 & (26)\\ y_{i+1}' &= y_i' \circ_i x_i' \qquad 1 \le i \le n, & (27) \end{aligned}$$

where the x_i' are the inputs and $\circ_i$ is one of the binary operations $+, -, \times, \div$; y_{n+1} is the output. Note it is in fact sufficient to take $\circ_i$ to be one of the operations $+, \times$, since subtraction and division are just addition and multiplication by suitable inverses. We will also assume that we only ever multiply by positive numbers; signs can always be looked after by hand. With these restrictions, the calculation is just the application of n successive proper affine transformations of the line, each being either a translation or a scaling; more explicitly, if we define

$$g_i' = \begin{cases} \begin{pmatrix} x_i' & 0 \\ 0 & 1 \end{pmatrix} & \circ_i = \times \\ \begin{pmatrix} 1 & x_i' \\ 0 & 1 \end{pmatrix} & \circ_i = +, \end{cases} \tag{28}$$

then we have

$$\begin{pmatrix} y_{n+1}' \\ 1 \end{pmatrix} = g_n' g_{n-1}' \dots g_1' \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{29}$$

As explained in the introduction, in addition to showing that the computation we wish to perform can be considered as evaluation of a product in a suitable group, we want to show that the effects of errors in computer arithmetic can be expressed using group multiplication. To this end, let

$$g_1 = \begin{pmatrix} \alpha_1 & \beta_1 \\ 0 & 1 \end{pmatrix}, \qquad g_2 = \begin{pmatrix} \alpha_2 & \beta_2 \\ 0 & 1 \end{pmatrix}$$

be two group elements. The computer calculated product of these has the form

$$(g_2 g_1)_{\text{computer}} = \begin{pmatrix} \alpha_1\alpha_2(1+\epsilon_1) & (\alpha_2\beta_1(1+\epsilon_2)+\beta_2)(1+\epsilon_3) \\ 0 & 1 \end{pmatrix},$$

where here, for each multiplication or addition we have inserted into the standard product a "rounding error factor" of the form $1+\epsilon$, where ϵ is small (we assume multiplication by 1 can be performed without error). We observe that we can write

$$(g_2 g_1)_{\text{computer}} = \begin{pmatrix} 1+\epsilon_1 & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1\alpha_2 & \alpha_2\beta_1+\beta_2 \\ 0 & 1 \end{pmatrix}, \tag{30}$$

where

$$T = \alpha_2\beta_1(\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3-\epsilon_1) + \beta_2(\epsilon_3-\epsilon_1).$$

This substantiates our claim. We note that while the 1,1-entry in the error matrix in (30) is close to 1, the 1,2-entry need not be close to 0. This reflects the fact that rounding errors in floating point arithmetic give small relative errors, but may give large absolute ones.

We are now ready to do error analysis. We assume that instead of being given the exact inputs x_i', $1 \le i \le n$, we are given approximations x_i, $1 \le i \le n$, and we perform all operations on a digital computer, encountering rounding errors. Our aim is to find a bound on the error in the computed output. We define matrices g_i, $1 \le i \le n$ using the data x_i analogously to the definition of the g'_i from the data x'_i in equation (28). We introduce error matrices k_i, $1 \le i \le n$ which, in the manner explained above, track the effects of rounding error in the procedure of group multiplication by g_i — so the computer is actually computing $k_n g_n \dots k_1 g_1$, not $g_n \dots g_1$ (actually $k_1 = I$, but we retain it for a more symmetric looking expression). Finally, it will be useful to have expressions for the intermediate computed results, so we define

$$\begin{aligned} h_1 &= I \\ h_2 &= k_1 g_1 \\ h_3 &= k_2 g_2 k_1 g_1 \\ &\vdots \\ h_n &= k_{n-1} g_{n-1} k_{n-2} g_{n-2} \dots k_2 g_2 k_1 g_1 \\ h_{n+1} &= k_n g_n k_{n-1} g_{n-1} \dots k_2 g_2 k_1 g_1 . \end{aligned}$$

We are departing here from the notation introduced in Theorem 1 of Section 2; the above definition of the h_i, incorporating the computer error factors k_i, will be used from here on. We wish to bound $d(g'_n \dots g'_1, k_n g_n \dots k_1 g_1)$, and use this bound to extract an error bound for the computed output.

4.1. $d(g'_n \dots g'_1, k_n g_n \dots k_1 g_1)$**.**

Using Theorem 1 of Section 2, we have

$$d(g'_n \dots g'_1, k_n g_n \dots k_1 g_1) \le \sum_{i=1}^{n} \rho(h_i) d(g'_i, k_i g_i) \tag{31}$$

To apply this estimate we need to know the factors $d(g'_i, k_i g_i)$. We distinguish the cases $\circ_i = +$ and $\circ_i = \times$.

(1) $d(g'_i, k_i g_i)$, $\circ_i = +$. We have

$$g'_i = \begin{pmatrix} 1 & x'_i \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad g_i = \begin{pmatrix} 1 & x_i \\ 0 & 1 \end{pmatrix}.$$

Now

$$g_i h_i = \begin{pmatrix} (h_i)_{11} & (h_i)_{12} + x_i \\ 0 & 1 \end{pmatrix},$$

and $k_i g_i h_i$ is the computer version of this, viz.

$$k_i g_i h_i = \begin{pmatrix} (h_i)_{11} & [(h_i)_{12} + x_i](1 + \epsilon_1) \\ 0 & 1 \end{pmatrix},$$

where ϵ_1 is small, bounded in magnitude by the "machine epsilon" ϵ. (For a machine that does floating point arithmetic with radix R and N digits in the mantissa, $\epsilon = R^{1-N}$.) From this it follows that

$$k_i = \begin{pmatrix} 1 & \epsilon_1 [(h_i)_{12} + x_i] \\ 0 & 1 \end{pmatrix},$$

and

$$k_i g_i = \begin{pmatrix} 1 & x_i + \epsilon_1[(h_i)_{12} + x_i] \\ 0 & 1 \end{pmatrix}.$$

Using this we find

$$d(g_i', k_i g_i) = \ln\left(\frac{1+r}{1-r}\right) , \qquad r = \frac{|x_i' - x_i - \epsilon_1[(h_i)_{12} + x_i]|}{\sqrt{4 + |x_i' - x_i - \epsilon_1[(h_i)_{12} + x_i]|^2}}.$$

It is straightforward to check that the function $\ln((\sqrt{4+p^2}+p)/(\sqrt{4+p^2}-p))$ is an increasing function of p for $p > 0$, and thus we obtain the final result

$$d(g_i', k_i g_i) \leq \ln\left(\frac{\sqrt{4+p^2}+p}{\sqrt{4+p^2}-p}\right) , \quad p = |x_i' - x_i| + \epsilon|(h_i)_{12} + x_i| \; . \tag{32}$$

It is also straightforward to prove that $d(g_i', k_i g_i) \leq p$.

In equation (31), $d(g_i', k_i g_i)$ comes multiplied by $\rho(h_i)$. Both g_i' and $k_i g_i$ belong to the translation subgroup H_1, and we can therefore replace $\rho(h_i)$ by $\rho_{H_1}(h_i) = \max(1, 1/(h_i)_{11})$ (see (22)). Replacing ρ by ρ_{H_1} here improves the error bound we ultimately obtain, and in addition gives expressions that are easier to manipulate.

(2) $d(g_i', k_i g_i)$, $\circ_i = \times$. We have

$$g_i' = \begin{pmatrix} x_i' & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad g_i = \begin{pmatrix} x_i & 0 \\ 0 & 1 \end{pmatrix}.$$

Arguments similar to those used above give

$$k_i = \begin{pmatrix} 1+\epsilon_2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad |\epsilon_2| \leq \epsilon,$$

and we obtain the final result:

$$d(g_i', k_i g_i) \leq |\ln x_i' - \ln x_i| + \ln(1+\epsilon_2) \leq |\ln x_i' - \ln x_i| + \epsilon. \tag{33}$$

Recall that we only allow multiplications by positive numbers, so here $x_i', x_i > 0$. Ignoring the effect of computer error we see here that Olver's metric (4) reappears. We are of course now looking at the abelian subgroup of G isomorphic to $\mathbf{R}^+$ with multiplication as group operation. We note that the (bi-invariant) metric on $\mathbf{R}$ with addition as group operation given by (32) (ignoring the effects of computer arithmetic) is nonstandard.

Once again g_i' and $k_i g_i$ are in a subgroup of G, this time the group of scalings H_2. Thus we can replace $\rho(h_i)$ in equation (31) by $\rho_{H_2}(h_i) = \sqrt{1 + (h_i)_{12}^2/(h_i)_{11}^2}$.

Summing up the results of this section we have

$$\begin{aligned} d(g_n' \dots g_1', k_n g_n \dots k_1 g_1) \;\; \leq & \sum_{i\,:\,\circ_i=+} \max\left(1, (h_i)_{11}^{-1}\right) \ln\left(\frac{\sqrt{4+p_i^2}+p_i}{\sqrt{4+p_i^2}-p_i}\right) \\ & + \sum_{i\,:\,\circ_i=\times} \sqrt{1 + \frac{(h_i)_{12}^2}{(h_i)_{11}^2}} \left(|\ln x_i' - \ln x_i| + \epsilon\right) \; , \end{aligned} \tag{34}$$

$$p_i = |x'_i - x_i| + \epsilon|(h_i)_{12} + x_i|.$$

4.2. The error in the computed output.

Having computed a bound for $\eta = d(g'_n \dots g'_1, k_n g_n \dots k_1 g_1)$, we now need to use this to obtain an error bound for the output. We could invoke the general results on metrics on homogeneous spaces derived in Section 2, but the situation is so simple we prefer to work directly. We have (equation (29)):

$$\begin{aligned} \begin{pmatrix} y'_{n+1} \\ 1 \end{pmatrix} &= g'_n g'_{n-1} \dots g'_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= h_{n+1} \mathcal{M} \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \end{aligned} \tag{35}$$

where $\mathcal{M} = h_{n+1}^{-1} g'_n g'_{n-1} \dots g'_1$. Using left invariance we see that $d(I, \mathcal{M}) \leq \eta$. We thus have

$$y'_{n+1} = (h_{n+1})_{12} + (h_{n+1})_{11} \mathcal{M}_{12} \ . \tag{36}$$

This is the result we need. $(h_{n+1})_{12}$ is the computed output. $(h_{n+1})_{11}$ is the computed value of the product of all inputs used for multiplications, i.e. the computed value of

$$\prod_{i \,:\, \circ_i = \times} x'_i \ . \tag{37}$$

And, finally, $\mathcal{M}_{12}$ is constrained by the requirement $d(I, \mathcal{M}) \leq \eta$. Since $\ln((1+r)/(1-r))$ is an increasing function of r (for $0 < r < 1$), $d(I, \mathcal{M}) \leq \eta \Rightarrow r \leq \tanh(\eta/2)$, where $r^2 = ((\mathcal{M}_{11} - 1)^2 + \mathcal{M}_{12}^2)/((\mathcal{M}_{11} + 1)^2 + \mathcal{M}_{12}^2)$ (using (21), the formula for d). Writing $\mathcal{M}_{12}^2$ in terms of r^2 and $\mathcal{M}_{11}^2$ it is easy to show that $\mathcal{M}_{12}^2 \leq 4r^2/(1-r^2)^2$, and combining this with the inequality $r \leq \tanh(\eta/2)$ gives

$$|\mathcal{M}_{12}| \leq \sinh \eta \ . \tag{38}$$

The form of our final result merits further comment. Equation (36) takes the form

$$\text{exact output} \;=\; \text{computed output} \;+\; \zeta \;.\; \text{computed product} \ , \tag{39}$$

where by "computed product" we mean the computed value of the product (37), and ζ, which measures the error, is bounded. ζ is neither an absolute error nor a relative error, but has one very pleasing property which neither of the standard error measures have, and which make it a very sensible measure to use. Suppose we were to continue our calculation by either adding or multiplying by a further input, which is exactly known (we also ignore the errors of computer arithmetic for this step). If the new input is added, it is added to both the exact output and the computed output, while the "computed product" is unchanged. Therefore, that such an operation does not change ζ. If the new input is multiplied, then not only do the exact output and computed output get multiplied, but so does the "computed product", and once again ζ is unchanged. This property of invariance of the error under further exact operations, is a consequence of left invariance of the metric.

5. The Horner Algorithm

The Horner algorithm is the most efficient algorithm for evaluation of polynomials. The polynomial

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

is evaluated in nested form

$$(\ldots((a_0x + a_1)x + a_2)\,x + \ldots + a_{n-1})\,x + a_n.$$

This is a scalar calculation without writing to memory, consisting of the $2n+1$ operations

$$\text{add } a_0,\quad \text{multiply by } x,\quad \text{add } a_1,\quad \text{multiply by } x,\quad \ldots\quad \text{add } a_n.$$

Let us denote by μ_i the computed intermediate result after addition of a_i ($i = 0,\ldots,n$), i.e.

$$\begin{aligned}
\mu_0 &= \text{computed value of } a_0\\
\mu_1 &= \text{computed value of } a_0x + a_1\\
\mu_2 &= \text{computed value of } (a_0x + a_1)x + a_2\\
&\vdots
\end{aligned}$$

We assume $x > 0$. In equation (34), h_i is the computed results prior to the "application of" (i.e. addition of or multiplication by) x_i. In our calculation, the addition of a_i ($i = 0,\ldots,n$) is the $(2i{+}1)$th operation, and we have $(h_{2i+1})_{11} = x^i$ and $(h_{2i+1})_{12} = \mu_i - a_i$. The $(i+1)$th multiplication by x ($i = 0,\ldots,n-1$) is the $2(i+1)$th operation, and we have $(h_{2(i+1)})_{11} = x^i$ and $(h_{2i+1})_{12} = \mu_i$. The considerations of the previous section thus give

$$|\text{exact result} - \text{computed result}| \le (\sinh\eta)\ (\text{computed value of } x^n)$$

$$\eta = \sum_{i=0}^{n} \max\left(1, \frac{1}{x^i}\right) \ln\left(\frac{\sqrt{4+p_i^2}+p_i}{\sqrt{4+p_i^2}-p_i}\right) + \sum_{i=0}^{n-1} \sqrt{1+\left(\frac{\mu_i}{x^i}\right)^2}\,(\epsilon + |\ln x' - \ln x|)\ , \tag{40}$$

$$p_i = |a_i' - a_i| + \epsilon|\mu_i|.$$

Here $|\ln x' - \ln x|$ is just the Olver relative error in the input x, which must be specified. If there is no intrinsic error in x, only error from representation as a "computer number", we can take $|\ln x' - \ln x|$ to be ϵ. Similarly, $|a_i' - a_i|$ is the absolute error in the coefficient a_i, which, if there is no intrinsic error can be taken to be $|a_i|\epsilon$. Finally we note that in the formula above, if p_i is small, a computer might well miscalculate the factor $\ln\ (\sqrt{4+p_i^2}+p_i)/(\sqrt{4+p_i^2}-p_i)$, and hence we should program the computer to just use the approximation p_i for this expression when p_i is small. Apart from this, we can be confident that a computer will compute η accurately enough for practical purposes.

We present the results of two explicit computations showing the usefulness of these bounds when applied suitably.

x	computed e^x	computed error bound	real e^x
0.30	1.3498588	4.67e-07	1.3498588
0.40	1.4918246	5.69e-07	1.4918247
0.50	1.6487212	6.89e-07	1.6487213
0.60	1.8221188	8.35e-07	1.8221188
0.70	2.0137529	1.03e-06	2.0137527
0.80	2.2255411	1.36e-06	2.2255409
0.90	2.4596033	2.04e-06	2.4596031
1.00	2.7182822	3.65e-06	2.7182818

Table 1: Results for e^x using single precision

Example 1. Computation of e^x, $0 < x < 1$. We first consider using single precision floating point arithmetic, with a machine epsilon of $2^{1-24} \approx 1.2 \times 10^{-7}$. For this level of accuracy we can approximate e^x by

$$e^x \approx \sum_{n=0}^{10} \frac{x^n}{n!} . \tag{41}$$

By the remainder theorem for Taylor series the error in this approximation for $0 < x < 1$ does not exceed $e/11! \approx 7 \times 10^{-8}$. In table 1 we list the results of computations using this approximation, with error bounds computed by our methods, and real values to 8 significant figures of e^x.

The results are pleasing; the error bounds are evidently of the right order of magnitude. The table however only starts with $x = 0.3$. When run for small x the error bounds start to grow very rapidly — see table 2. This has an explanation: for small x, the last terms in the series (41) are superfluous. For such x, at the start of the implementation of Horner's algorithm we are doing operations that do not give any contribution to the final result, but that do (since the group metric used is not right invariant) give contributions to the error. Similarly, if we were to take a series for e^x including terms of degree up to 20, we would not get any reasonable results for $x < 1$. *Superfluous initial operations must be avoided for our bounds to be useful.*

Repeating this exercise in double precision, with a machine epsilon of $2^{1-53} \approx 2.2 \times 10^{-16}$, we use

$$e^x \approx \sum_{n=0}^{18} \frac{x^n}{n!} . \tag{42}$$

for which the error does not exceed $e/19! \approx 2 \times 10^{-17}$. Results are given in table 3; here real values of e^x are given to 17 significant figures. Again the error bounds obtained for small x are very poor, for the reasons explained, and are not given. But in the range of values of x where we are not doing large numbers of superfluous operations, the bounds are very acceptable (here we must remember we are looking at a 37-step calculation).

x	computed e^x	computed error bound	real e^x
0.13	1.1388284	1.13e+93	1.1388284
0.14	1.1502738	9.62e+40	1.1502738
0.15	1.1618342	1.04e+17	1.1618342
0.16	1.1735109	2.78e+05	1.1735109
0.17	1.1853049	4.40e-01	1.1853049
0.18	1.1972173	4.58e-04	1.1972174
0.19	1.2092496	1.29e-05	1.2092496
0.20	1.2214028	2.06e-06	1.2214028
0.21	1.2336781	8.39e-07	1.2336781
0.22	1.2460768	5.58e-07	1.2460767
0.23	1.2586001	4.71e-07	1.2586000
0.24	1.2712492	4.42e-07	1.2712492
0.25	1.2840254	4.35e-07	1.2840254
0.26	1.2969302	4.36e-07	1.2969301
0.27	1.3099644	4.42e-07	1.3099645
0.28	1.3231299	4.50e-07	1.3231298
0.29	1.3364275	4.58e-07	1.3364275

Table 2: Further results for e^x using single precision, showing the divergence of the error bound for small x

x	computed e^x	computed error bound	real e^x
0.30	1.34985 88075 76003 2	8.6921e-16	1.34985 88075 76003 1
0.40	1.49182 46976 41270 3	1.0600e-15	1.49182 46976 41270 3
0.50	1.64872 12707 00128 2	1.2813e-15	1.64872 12707 00128 1
0.60	1.82211 88003 90509 1	1.5380e-15	1.82211 88003 90509 0
0.70	2.01375 27074 70476 2	1.8430e-15	2.01375 27074 70476 5
0.80	2.22554 09284 92467 4	2.2888e-15	2.22554 09284 92467 6
0.90	2.45960 31111 56949 4	3.5899e-15	2.45960 31111 56949 7
1.00	2.71828 18284 59044 6	1.0357e-14	2.71828 18284 59045 2

Table 3: Results for e^x using double precision

Example 2. Computation of $P_{20}(y), P_{30}(y)$ (the Legendre polynomials of orders 20 and 30) for $0 < y < 1$. P_{20} and P_{30} are given by

$$\begin{aligned} P_{20}(y) \;=\; & \frac{34461632205}{262144} y^{20} - \frac{83945001525}{131072} y^{18} + \frac{347123925225}{262144} y^{16} \\ & - \frac{49589132175}{32768} y^{14} + \frac{136745788725}{131072} y^{12} - \frac{29113619535}{65536} y^{10} \\ & + \frac{15058768725}{131072} y^{8} - \frac{557732175}{32768} y^{6} + \frac{334639305}{262144} y^{4} - \frac{4849845}{131072} y^{2} \\ & + \frac{46189}{262144} \end{aligned}$$

$$\begin{aligned} P_{30}(y) \;=\; & \frac{1}{67108864}\Big(7391536347803839\, y^{30} - 54496920530418135\, y^{28} \\ & +180700315442965395\, y^{26} - 355924863751295475\, y^{24} \\ & +463373879223384675\, y^{22} - 419762220002360235\, y^{20} \\ & +271274904083157975\, y^{18} - 126155198555389575\, y^{16} \\ & +42051732851796525\, y^{14} - 9888133564634325\, y^{12} \\ & +1591748329916745\, y^{10} - 166966608033225\, y^{8} + 10529425731825\, y^{6} \\ & -347123925225\, y^{4} + 4508102925\, y^{2} - 9694845\Big)\,. \end{aligned}$$

The large coefficients in these polynomials cause significant errors; intuitively we expect the error to increase with y, as increasing y will increase the magnitude of typical intermediate results, exacerbating the effect of both data and rounding errors. (Of course, direct computations of Legendre polynomials are known to be poor in comparision to use of recursion formulae.) Computations of P_{20} and P_{30} were only carried out in double precision; single precision is not accurate enough.

y^2	$P_{20}(y)$ (exact, 18 s.f.)	$P_{20}(y)$ (computed, 16 s.f.)	observed error	error bound
0.1	+0.17201 11111 54253 006	+0.17201 11111 54252 9	1×10^{-16}	2.818×10^{-14}
0.2	−0.18565 68320 00000 000	−0.18565 68320 00013 6	1.4×10^{-14}	3.391×10^{-13}
0.3	+0.14980 20876 17509 842	+0.14980 20876 17508 3	1.5×10^{-15}	2.103×10^{-12}
0.4	+0.01907 91261 89453 1250	+0.01907 91261 89242 87	2.1×10^{-13}	9.215×10^{-12}
0.5	−0.19306 51776 49259 567	−0.19306 51776 50343 1	1.1×10^{-12}	3.231×10^{-11}
0.6	+0.17258 13820 00000 000	+0.17258 13820 02626 1	2.6×10^{-12}	9.663×10^{-11}
0.7	+0.02195 16386 42518 9972	+0.02195 16386 49841 23	7.3×10^{-12}	2.565×10^{-10}
0.8	−0.19839 71436 69921 875	−0.19839 71436 54722 3	1.5×10^{-11}	6.191×10^{-10}
0.9	+0.27598 84681 48118 973	+0.27598 84681 71748 0	2.4×10^{-11}	1.384×10^{-09}
1.0	+1.00000 00000 00000 00	+1.00000 00000 00000	$< 10^{-15}$	2.903×10^{-09}

Table 4: Results for P_{20}

Results are presented in tables 4 and 5. For different values of y^2 we give both correct (to the number of significant figures given) and computed values, and observed errors and our theoretical error bounds. Bearing in mind that in the computation of P_{20} (P_{30}) we perform 21 (31) operations, and for each one the error bound allows for the greatest possible data and rounding errors, the ratios we observed between the error bounds and the observed errors do not seem

y^2	$P_{30}(y)$ (exact, 16 s.f.))	$P_{30}(y)$ (computed, 16 s.f.)	observed error	error bound
0.1	+0.13718 28671 79238 7	+0.13718 28671 79244 3	5.6×10^{-15}	∞
0.2	+0.00093 67781 70368 0000	+0.00093 67781 71234 2204	8.7×10^{-13}	2.27×10^{-11}
0.3	−0.06192 85876 58071 01	−0.06192 85876 48536 39	1.0×10^{-11}	3.23×10^{-10}
0.4	+0.07388 48092 54138 81	+0.07388 48093 99027 06	1.4×10^{-10}	2.94×10^{-9}
0.5	−0.06638 90524 49721 93	−0.06638 90514 08116 03	1.0×10^{-9}	1.92×10^{-8}
0.6	+0.05831 24130 36576 00	+0.05831 24191 42637 85	6.1×10^{-9}	9.86×10^{-8}
0.7	−0.07445 29637 68226 14	−0.07445 29364 75398 69	2.7×10^{-8}	4.24×10^{-7}
0.8	+0.15333 49912 37899 6	+0.15333 50892 66750 0	9.8×10^{-8}	1.58×10^{-6}
0.9	−0.23556 76702 40547 5	−0.23556 71536 06672 3	5.2×10^{-7}	5.28×10^{-6}
1.0	+1.00000 00000 00000	+1.00000 13974 21315	1.4×10^{-6}	1.60×10^{-5}

Table 5: Results for P_{30}

unreasonable. A firmer indication of sharpness by order of the bounds can be obtained doing calculations in which a controlled data error is introduced that is still small, but dominates all the other errors. A few experiments of this sort that we performed (introducing errors of 10^{-3} into specific coefficients a_i of the polynomials) gave pleasing results.

The undesirable feature of the results is quite evident — for $P_{30}(y)$ with $y^2 = 0.1$ we found an enormous error bound, which we have indicated as ∞. This has already been explained in example 1: for small y we are doing unnecessary operations which contribute heavilly to the error but not the result. We can overcome this by prefacing our evaluation routines with a routine that decides how many terms in the polynomial we need to evaluate, for given y.

6. Concluding Remarks

In this paper we have outlined a new approach to error analysis, and implemented it to obtain error estimates for a floating point implementation of Horner's algorithm. The main open question we leave, and which will require much work to resolve, is how useful our approach will be for computations involving other, higher-dimensional groups. As we have already claimed in the introduction, many numerical procedures can be viewed as the computation of a group product — for example, Gaussian elimination for an $N \times N$ matrix simply consists of applying a sequence of $GL(N)$ transformations to the matrix. The question is whether we can compute sufficient information about left invariant, almost $\mathrm{Inn}(G)$ invariant metrics on the relevant groups, with associated ρ–functions, that will make our approach practically viable. Amongst that groups which merit investigation, the groups $GL(N)$ play a crucial role, as other groups of interest arise as subgroups of $GL(N)$ (for example the group of affine transformations of an $m \times n$ matrix Y, i.e. the transformations $Y \to AY + YB + C$, where A is $m \times m$, B is $n \times n$ and C is $m \times n$ is a subgroup of $GL(m+n)$). A computation of a suitable metric, and ρ–functions, just for $GL(2)$ would already give some very useful results in error analysis of the implementation of two-term recurrence relations, and the computation of rational functions.

We conclude with some comments on the literature. After Olver introduced his notion of relative error (4) [10, 11], it was generalized to vectors both by Ziv

[19] and by Pryce [17, 18]. The work of Pryce has something in common with our work on homogeneous space metrics: if X is a Banach space, Pryce considers the possibility of defining a metric on X by $d(x,y) = \min\{\|T\| : y = e^T x\}$, where T is a bounded linear operator on X. The metrics of Pryce can be computed [3, 4] in certain simple cases.

Olver, with collaborators, has written a series of papers [1, 14, 12, 13] exploiting his notion of relative error metric in different calculations. Ziv has done likewise in the paper [20], which also relates to the question of polynomial evaluation. Ziv considers a different scheme of error analysis in [21]. Another paper we found interesting on the subject of polynomial evaluation is that of Oliver [8]. Finally, we mention the two papers [15] and [5], which we found to give useful perspectives on the subject of error analysis.

Acknowledgments.

J.S. wishes to thank D.A.Kessler and A.Ziv for discussions. S.S. wishes to thank the Department of Mathematics at the University of New Hampshire, Durham NH for hospitality in the 1996-7 academic year when most of this work was done. Both of us wish to thank Karl H. Hofmann for his extremely thorough and useful editing work on this paper, and for the appendix.

Appendix: On the Existence of Left Invariant, Almost Inn(G) Invariant Metrics, by Karl H. Hofmann

In the body of this paper, an explicit construction was given for a left invariant, almost Inn(G) invariant metric on a finite dimensional, connected Lie group. In this appendix the following more general result is proved:

Theorem. Every locally compact connected group G having a countable basis for its identity neighborhoods admits a left invariant, almost $\mathrm{Inn}(G)$–invariant metric.

Lemma A. On a connected Lie group G, for each norm $\|\cdot\|$ on the Lie algebra $\mathfrak{g}$ the associated metric is almost $\mathrm{Aut}(G)$-invariant.

Proof. When G is a Lie group and $\mathfrak{g}$ its Lie algebra we use the left translations $L_g\colon G \to G$, $L_g x = gx$ and the induced isomorphism $TL_g\colon T_0(G) \to T_g(G)$ of tangent spaces to identify $T_g(G)$ with $T_0(G) = \mathfrak{g}$. Any norm $\|\cdot\|$ on $\mathfrak{g}$ (for instance a euclidean one) induces a norm on the tangent spaces; accordingly, if $\gamma\colon [0,1] \to G$ is a differentiable curve on G, its arc length $\ell(\gamma) = \int_0^1 \|\dot\gamma(t)\| dt$ (where $\dot\gamma(t) \in \mathfrak{g}$) is well defined, and on G we obtain an associated left invariant metric defined by

$$d(g,h) = \inf\{\ell(\gamma) : \gamma \text{ is a differentiable curve with } \gamma(0) = g,\, \gamma(1) = h\}.$$

Suppose $\alpha \in \mathrm{Aut}(G)$ is an automorphism of G. Its differential $\mathfrak{L}(\alpha)$ (at the origin) is an automorphism of the Lie algebra $\mathfrak{g}$. Let $\|\mathfrak{L}(\alpha)\|$ denote the operator norm with respect to the given norm $\|\cdot\|$ on $\mathfrak{g}$. Then $\|\mathfrak{L}(\alpha)(X)\| \le \|\mathfrak{L}(\alpha)\| \cdot \|X\|$ for each $X \in \mathfrak{g}$. Let γ be a differentiable curve on G and let $\alpha\gamma\colon [0,1] \to G$ denote the differentiable curve defined by $\alpha\gamma(t) = \alpha(\gamma(t))$. Then $\ell(\alpha\gamma) = \int_0^1 \|(\alpha\gamma)'(t)\| dt = \int_0^1 \|\mathfrak{L}(\alpha)\dot\gamma(t)\| dt \le \int_0^1 \|\mathfrak{L}(\alpha)\| \cdot \|\dot\gamma(t)\| dt = \|\mathfrak{L}(\alpha)\| \cdot \int_0^1 \|\dot\gamma(t)\| dt = \|\mathfrak{L}(\alpha)\| \cdot \ell(\gamma)$. It follows that $d(\alpha(g), \alpha(h)) \le \|\mathfrak{L}(\alpha)\| \cdot d(g,h)$. ■

In particular, a connected Lie group G admits a left invariant, almost $\mathrm{Inn}(G)$ invariant metric.

If d is a left invariant metric on a topological group (compatible with the topology) we say that (G,d) is a *metric group*.

Lemma B. Let (G,d) be a metric group and N a closed normal subgroup. Set $D(Ng,Nh)=\inf_{n\in N} d(nh,g)$. Then $(G/N,D)$ is a metric group. If d is almost $\mathrm{Inn}(G)$–invariant with an admissible ρ-function ρ_G, then D is almost $\mathrm{Inn}(G/N)$–invariant with an admissible ρ-function $\rho_{G/N}$, where $\rho_{G/N}(Ng)=\inf_{n\in N}\rho_G(ng)$.

Proof. The first assertion belongs to the body of metric group theory (and is very familiar for normed vector spaces). As to the second, observe that $nxg=xn'g$ with $n'=x^{-1}nx\in N$ and that $\inf_{n\in N} d(nxg,yg)=\inf_{m,n\in N} d(mxg,nyg)=\inf_{m,n\in N} d(xng,ymg)=\inf_{m,n\in N} d(nxmg,ymg)$; accordingly, $D(Nxg,Nyg)=\inf_{n\in N} d(nxg,yg)=\inf_{m,n\in N} d(nxmg,ymg)\le\inf_{n\in N}\inf_{m\in N}\rho_G(mg)d(nx,y)=\rho_{G/N}(Ng)D(Nx,Ny)$. ■

Now we use a Theorem due to Iwasawa [7], p. 547, Theorem 11, which is readily converted into the global statement below, since every local Lie group is isomorphic as a local group to the identity neighborhood of a simply connected Lie group, and since a local group morphism of an identity neighborhood of a simply connected group into any group extends to a morphism.

Lemma C. If G is a locally compact connected group, then there is a compact normal subgroup N of G, a simply connected Lie group L, and a discrete central subgroup H of $N\times L$ such that $G\cong(N\times L)/H$.

Proof of Theorem. By Lemmas B and C, the assertion of the Theorem holds if it holds for groups of the form $N\times L$ for a compact first countable group N and a connected Lie group L. The group N has a bi-invariant metric d_N (see item 1 preceding Theorem 1 in Section 2 of the paper). By Lemma A, the group L has a left invariant, almost $\mathrm{Aut}(G)$–invariant metric d_L. Then the max-metric $\mathcal{D}$ on $N\times L$ given by $\mathcal{D}((n_1,x_1),(n_2,x_2))=\max\{d_N(n_1,n_2),d_L(x_1,x_2)\}$ is a left invariant, almost $\mathrm{Inn}(N\times L)$–invariant metric. ■

References

[1] Clenshaw, C. W., and F. W. J. Olver, *An Unrestricted Algorithm for the Exponential Function*, SIAM J. Numer. Anal. **17** (1980), 310–331.

[2] Dieudonné, J., "Treatise on Analysis, Vol II", Academic Press, New York, 1970, 53–54.

[3] Govaerts, W., *The Geodesics of Pryce's Relative Distance in* $\mathbf{R}^3_\infty$, SIAM J. Numer. Anal. **23** (1986), 1295–1302.

[4] —, *The Relative Distance of J. D. Pryce in* $\mathbf{R}^3_\infty$, J. Approx. Th. **56** (1989), 13–29.

[5] Henrici, P., *A Model for the Propagation of Rounding Error in Floating Arithmetic*, in: K. L. E. Nickel, Ed., " Interval Mathematics," Academic Press, New York, 1980, 49–73.

[6] Hewitt, E., and K. A. Ross, "Abstract Harmonic Analysis I (Second Edition)", Grundlehren der Mathematischen Wissenschaften 115, Springer, 1979, Chapter 2, Theorem 8.6.

[7] Iwasawa, K., *On some types of topological groups,*, Annals of Math. **50** (1949), 507–558.

[8] Oliver, J., *The Accurate Evaluation of Polynomial Approximations to Library Functions*, IMA J. Numer. Anal. **2** (1982), 63–72.

[9] Olver, F. W. J., *Error Bounds for Polynomial Evaluation and Complex Arithmetic*, IMA J. Numer. Anal. **6** (1986), 373–379.

[10] —, *A New Approach to Error Arithmetic*, SIAM J. Numer. Anal. **15** (1978), 368–393.

[11] —, *Further Developments of Rp and Ap Error Analysis*, IMA J. Numer. Anal. **2** (1982), 249–274.

[12] —, *Error Bounds for Arithmetic Operations on Computers Without Guard Digits*, IMA J. Numer. Anal. **3** (1983), 153–160.

[13] —, *Error Bounds for Linear Recursion Relations*, Math.Comp. **50** (1988), 481–499.

[14] Olver, F. W. J., and J. H. Wilkinson, *A Posteriori Error Bounds for Gaussian Elimination*, IMA J. Numer. Anal. **2** (1982), 377–406.

[15] Porta, H., and K. B. Stolarsky, *Means that Minimize Relative Error, and an Associated Integral Equation*, J. Math. Anal. Appl. **122** (1987), 95–113.

[16] Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, "Numerical Recipes in C, Second Edition",Cambridge Univ. Press, 1992, Section 16.2.

[17] Pryce, J. D., *A New Measure of Relative Error for Vectors*, SIAM J. Numer. Anal. **21** (1984), 202–215.

[18] —, *Multiplicative Error Analysis of Matrix Transformation Algorithms*, IMA J. Numer. Anal. **5** (1985), 437–445.

[19] Ziv, A., *Relative Distance—An Error Measure in Round-Off Error Analysis*, Math. Comp. **39** (1982), 563–569.

[20] —, *A Stable Method for Evaluation of a Polynomial and a Rational Function of One Variable*, Numer. Math. **41** (1983), 309–319.

[21] —, *Converting Approximate Error Bounds into Exact Ones*, Math. Comp. **64** (1995), 265–277.

Jeremy Schiff and
Steve Shnider
Department of Mathematics and
Computer Science
Bar-Ilan University
Ramar Gan 5200, Israel
schiff@math.biu.ac.il
shnider@math.biu.ac.il

Received November 6, 1997
and in final form October 6, 2000