NUMERICAL INTEGRATORS THAT CONTRACT VOLUME

ROBERT I. MCLACHLAN AND G.R.W. QUISPEL

Abstract. We study numerical integrators that contract phase space volume even when the ODE does so at an arbitrarily small rate. This is done by a splitting into two-dimensional contractive systems. We prove a sufficient condition for Runge-Kutta methods to have the appropriate contraction property for these two-dimensional systems; the midpoint rule is an example.

1. Introduction. What is a dissipative system? In physics, the term usually refers to possession of a scalar function (such as energy) which decreases in time, and one speaks of, e.g., the dissipative pendulum, $\ddot{x} = -\sin x - \varepsilon\dot{x}$, for which $\frac{d}{dt}(\frac{1}{2}\dot{x}^2 - \cos x) = -\varepsilon\dot{x}^2 \le 0$. (See [1, 6] for some general formulations of such systems.) In dynamical systems, it usually refers to a decrease of phase space volume in time, as in the "dissipative Hénon map" $(x, y) \mapsto (y, 1 + bx - ay^2)$, with Jacobian determinant $-b$—phase space area decreases if $|b| < 1$. Another example is the famous Lorenz system, which contracts volume at a constant rate. In the numerical analysis of ODEs, it has been used to describe systems that decrease some norm of the solution, either in the sense that $\frac{d}{dt}\|x\|^2 < a - b\|x\|^2$ for some a, $b > 0$, or $\frac{d}{dt}\|x\|^2 < 0$ for all $\|x\| > R > 0$ [11].

In the field of geometric integration, much work has been done in maintaining the preservation of a conserved quantity (first integral) [3, 6, 7], the decrease of a dissipated quantity (a Lyapunov function) [6, 7], or the preservation of phase space volume [2, 5]. Here we look at the missing case, and study how to maintain the property of contracting phase space volume.

Consider the ODE

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n \tag{1}$$

with solution $x(t)$ and Jacobian (first variation) $A(t) = \partial x(t)/\partial x(0)$ which evolves according to

$$\dot{A} = FA, \quad A(0) = I,$$

where $F(x) = df(x)$ is the derivative of the vector field f. We have

$$\begin{aligned} \frac{d}{dt}\det A &= \det A \,\mathrm{tr}\left(A^{-1}\dot{A}\right) \\ &= \det A \,\mathrm{tr}\, F \end{aligned}$$

so that phase space volume contracts, is preserved, or expands when $\mathrm{tr}\, F < 0$, $\mathrm{tr}\, F = 0$, or $\mathrm{tr}\, F > 0$ for all x, respectively. $\mathrm{tr}\, F$ is the divergence or trace of the vector field f. Strongly contractive systems are those for which there is a b such that

Research at MSRI is supported in part by NSF grant DMS-9701755.

$\operatorname{tr} F < b < 0$. In this case any consistent numerical integrator will be contractive for small enough time step h. Therefore we concentrate on weak contraction (defined below), which is a closed property and more difficult to preserve. It turns out that requiring contractivity for *all* $h > 0$ and all contractive f is prohibitively difficult, which leads to the following definition. We consider one-step methods $x_{n+1} = g(x_n)$ with Jacobian $A = dg(x_0)$.

Definition 1. *The ODE (1) is (weakly) contractive if* $\operatorname{tr} F \le 0$ *for all* x. *An integrator is (weakly) contractive if for any matrix norm* $\|\cdot\|$ *and all* $L > 0$ *there is a time step* $h^* > 0$ *such that* $|\det A| \le 1$ *for all* $0 < h < h^*$, *for all* x, *and for all* f *such that* $\|F\| < L$ *and* $\operatorname{tr} F \le 0$.

That is, there might be stiffness problems (for large L, h^* might be small), but the time step needed to preserve contractivity should not tend to zero as $\operatorname{tr} F \to 0$. Note that a contractive integrator as defined here is not necessarily volume-preserving when the ODE is, nor is the relative amount of contraction necessarily correct as $\operatorname{tr} F \to 0$. These would be true if we added the requirement $\ln(\det A)/h \operatorname{tr} F \to 1$ uniformly as $\operatorname{tr} F \to 0$ uniformly, for all fixed $h < h^*$. The midpoint rule (see Proposition 3, below) satisfies this, for example.

Since there are no known linearly covariant volume-preserving schemes in more than two dimensions [2], we expect that the same is true here, and we immediately consider systems in two dimensions.

2. Dissipative schemes in two dimensions.

Example 2. *Euler's method is not contractive in two dimensions.* We have $x_{n+1} = x_n + hf(x_n)$ so $A = I + hF$. In two dimensions,

$$\det A = \det \begin{pmatrix} 1 + hF_{11} & hF_{12} \\ hF_{21} & 1 + hF_{22} \end{pmatrix} = 1 + h \operatorname{tr} F + h^2 \det F.$$

So $\det A \le 1$ for all h if $\det F \le 0$, and $\det A \le 1$ for

$$h \le \frac{-\operatorname{tr} F}{\det F}$$

if $\det F > 0$, so small contractivity can require a small time step to be captured.

Note that since $\det A = \det(I + hF) = \prod(1 + h\lambda_i)$, where λ_i are the eigenvalues of F, Euler's method *is* contractive in n dimensions on systems with bounded negative eigenvalues. We look at this further in Section 3.

Proposition 3. *The midpoint rule,* $x_{n+1} = x_n + hf(\bar{x})$, $\bar{x} = (x_n + x_{n+1})/2$, *is contractive in two dimensions.*

Proof. We have

$$A = \Big(I - \frac{1}{2}hF(\bar{x})\Big)^{-1}\Big(I + \frac{1}{2}hF(\bar{x})\Big),$$

so

$$\det A = \frac{1 + he + h^2 d}{1 - he + h^2 d}$$

where $e = \frac{1}{2}\operatorname{tr} F(\bar{x})$, $d = \frac{1}{4}\det F(\bar{x})$. Thus $(\det A)^2 \le 1$ if

$$\left(1 + he + h^2 d\right)^2 \le \left(1 - he + h^2 d\right)^2$$

or

$$e\left(1 + h^2 d\right) \le 0$$

Since $e \le 0$, this is true for all h if $e = 0$ (the well-known result that the midpoint rule is area-preserving, or symplectic), for all h if $d \ge 0$, or for $h < 1/\sqrt{-d}$ if $d < 0$. □

Proposition 3 can be generalized as follows.

Proposition 4. *The symplectic Runge-Kutta methods with $b_i > 0$ for all i are contractive in two dimensions.*

Proof. For the terminology, see [9]. Our proof closely follows their proof of symplecticity. An s-stage Runge-Kutta method is defined by

$$X_i = x_n + h\sum_{j=1}^{s} a_{ij} f(X_j), \tag{2}$$

$$x_{n+1} = x_n + h\sum_{j=1}^{s} b_j f(X_j), \tag{3}$$

and is symplectic if $b_i b_j - b_i a_{ij} - b_j a_{ji} = 0$ for all i and j. Note that in two dimensions, $A^T J A = J \det A$, where $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, so we evaluate the left hand side. Let $D_i = df(X_i) = F(X_i) dX_i =: F_i A_i$. Differentiating (3) gives

$$\begin{aligned} A^T J A &= \Big(I + h\sum_i b_i D_i\Big)^T J \Big(I + h\sum_j b_j D_j\Big) \\ &= J + h\sum_i b_i \left(J D_i + D_i^T J\right) + h^2 \sum_{i,j} b_i b_j D_i^T J D_j. \end{aligned}$$

Differentiating (2) gives

$$A_i = I + h\sum_j a_{ij} D_j \tag{4}$$

or

$$J D_i = A_i^T J D_i - h\sum_j a_{ij} D_j^T J D_i\,.$$

Inserting,

$$A^T J A = J + h\sum_i b_i \left(A_i^T J D_i + D_i^T J A_i\right) + h^2 \sum_{i,j} \left(b_i b_j - b_i a_{ij} - b_j a_{ji}\right) D_i^T J D_j.$$

The last term is zero because of the assumption on the coefficients b_i, a_{ij}. Now $D_i = F_i A_i$, so

$$\begin{aligned} A^T J A &= J + h \sum_i b_i A_i^T (J F_i + F_i^T J) A_i \\ &= J + h \sum_i b_i A_i^T J A_i \operatorname{tr} F_i \\ &= J \Big(1 + h \sum_i b_i \det A_i \operatorname{tr} F_i \Big) \end{aligned}$$

so

$$\det A = 1 + h \sum_i b_i \det A_i \operatorname{tr} F_i.$$

From (4), $\det A_i$ is bounded and equal to $1 + \mathcal{O}(h)$. Using $b_i > 0$ and $\operatorname{tr} F_i \le 0$ gives the result. □

The assumption $b_i > 0$ is necessary. Suppose there are $s = 2$ stages with $b_1 > 0$ and $b_2 < 0$. Then the vector (b_i) lies in the fourth quadrant, and all we know of the vector $(\operatorname{tr} F_i)$ is that it lies in the third quadrant. In regions where the trace varies relatively quickly, the angle between these two vectors can be less than $\frac{\pi}{2}$, leading to $(b_i) \cdot (\operatorname{tr} F_i) > 0$ and $\det A > 1$.

These methods actually preserve area when $\operatorname{tr} F = 0$. In fact, this is not necessary for contractivity in two dimensions, because we can allow a small amount of "numerical contractivity" even as $\operatorname{tr} F \to 0$; away from $\operatorname{tr} F = 0$ the inherent contractivity of the ODE contributes. It turns out that only methods of order 2, 3, 6, 7, ... , can achieve this.

Lemma 5. *Let $R(z)$ be the linear stability polynomial of a consistent Runge-Kutta method. In two dimensions, the method is contractive on linear ODEs if there is a $u^* > 0$ such that*

$$R(u)R(-u) \le 1, \quad R(iu)R(-iu) \le 1$$

for all $0 \le u < u^*$.

Proof. In n dimensions, a Runge-Kutta method on linear problems $\dot{x} = Fx$ has derivative $A = R(hF)$. Therefore $\det A = \prod_i R(h\lambda_i) = 1 + h \operatorname{tr} F + \mathcal{O}(h^2)$, so the method is contractive if $\operatorname{tr} F < 0$. If $\operatorname{tr} F = 0$ we have to examine $\det A$ in more detail. In $n = 2$ dimensions, there are only two such cases: the eigenvalues can be $(u, -u)$ or $(iu, -iu)$. This gives the result. □

We note that the result also applies to nonlinear problems with 1-stage methods, since then $F(x)$ is evaluated at only a single point.

Proposition 6. *Let the method have order p, so that $R(z) = e^z + az^{p+1} + bz^{p+2} + \mathcal{O}(z^{p+3})$. If $4|(p+1)$ and $a < 0$, or if $4|(p+2)$ and $b < a$, then the method is contractive on linear problems in two dimensions.*

Proof. We expand

$$\begin{aligned} R(z)R(-z) - 1 &= e^{-z}(az^{p+1} + bz^{p+2}) + e^z(a(-z)^{p+1} + b(-z)^{p+2}) + \dots \\ &= az^{p+1}(1 - (-1)^p) + (b - a)z^{p+2}(1 + (-1)^p) + \dots \end{aligned}$$

The leading term must be negative for $z = u$ and for $z = iu$, so it must be a fourth power. If p is even, the leading term is z^{p+2} so $4|(p+2)$ and we need $b < a$; if p is odd, the leading term is z^{p+1} so $4|(p+1)$ and we need $a < 0$. $\square$

An example is any 3-stage, 3rd order Runge-Kutta, which has $R(z) = 1+z+\frac{1}{2}z^2+\frac{1}{6}z^3$.

This result can be extended to more dimensions. For example, a longer calculation shows that Proposition 6 holds with $p = 3$ in three dimensions. We are not sure how it extends to nonlinear systems. It seems that if the eigenvalues of F are varying rapidly, contractivity could be lost.

3. More than two dimensions. For systems in more than two dimensions, we generalize the volume-preserving method of Feng and Wang ([2]; see also [5]). We write the ODE as a sum of two-dimensional contractive systems (i.e., ones for which $\dot{x}_i = 0$ except for two indices i), apply a contractive method to each term, and compose the resulting maps with positive time steps. Since contractivity is a semi-group property, we can build a contractive integrator of order 1 or 2 in this way [5]. This relies on the following proposition.

Proposition 7. *Any C^{r+1} contractive ODE is the sum of two-dimensional C^r contractive ODEs.*

Proof. Consider $\dot{x} = f(x)$, $F = df$. We shall write f in the form $f_i = \sum_j \partial_j L_{ij}$ (where $\partial_j = \partial/\partial_{x_j}$.)

Let $s_{ij}(x)$ be n^2 functions with

$$s_{ij}(x) + s_{ji}(x) \geq 0 \quad \text{and} \quad \sum_{i,j=1}^{n} s_{ij}(x) = 1$$

for all x. Let

$$S_{ij} = \iint s_{ij}(x)\,\mathrm{tr}\,F(x)\,dx_i\,dx_j$$

where any values of the indefinite integrals can be taken. Let

$$\tilde{f}_i = f_i - \sum_j \partial_j S_{ij} = f_i - \sum_j \int s_{ij}\,\mathrm{tr}\,F\,dx_i,$$

so that

$$\mathrm{tr}\,d\tilde{f} = \Big(1 - \sum_{i,j} s_{ij}\Big)\,\mathrm{tr}\,F = 0.$$

Thus, $\tilde{f}$ is traceless and can be written as

$$\tilde{f}_i = \sum_j \partial_j A_{ij}$$

where the matrix A is antisymmetric and as smooth as f [2, 5]. Therefore

$$f_i = \sum_j \partial_j (A_{ij} + S_{ij}) \tag{5}$$

or $L = A + S$.

For an explicit splitting, we take the diagonal elements $s_{ii}(x) = 0$. Then $\dot{x} = f$ is the sum of the following $n(n-1)/2$ two-dimensional ODEs:

$$\begin{aligned}\dot{x}_i &= \partial_j L_{ij}\\ \dot{x}_j &= \partial_i L_{ji}\\ \dot{x}_k &= 0 \text{ for } k \neq i, j\end{aligned}$$

for each pair (i, j) of indices from 1 to n. Each is contractive because each A piece is traceless and each S piece has trace $(s_{ij} + s_{ji}) \operatorname{tr} F \leq 0$.

One degree of smoothness is lost in this splitting, because each S piece depends on $\operatorname{tr} F$. □

An interesting solution is obtained by taking $s_{ii} = 1/n$, $s_{ij} = 0$ for $i \neq j$, and

$$nL_{ij} = \int f_i \, dx_j - \int f_j \, dx_i + \delta_{ij} \iint \operatorname{tr} F \, dx_i \, dx_i.$$

However, a more practical decomposition is to take the same S but $A_{ij} = 0$ for $|i - j| > 1$; this gives the minimum of $n - 1$ two-dimensional ODEs.

Although the above proof is constructive, it may be possible to find a more convenient splitting by ad hoc methods, in some cases leading to an explicit contractive integrator.

Firstly, if f is the sum of *integrable* contractive vector fields, then their flows can be composed to give a contractive integrator for f. For example, the Lorenz system,

$$\dot{x} = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho & -1 & 0 \\ 0 & 0 & -\beta \end{pmatrix} x + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -x_1 \\ 0 & x_1 & 0 \end{pmatrix} \nabla \left(\frac{x_2^2 + x_3^2}{2} \right),$$

is the sum of a linear, contractive part and a Poisson, non-contractive part, each of which may be solved exactly, giving an integrator with *exactly* correct contractivity.

Secondly, it may be possible to use a simpler method, such as Euler, on some of the pieces. Here are some criteria which allow this.

Proposition 8. *Euler's method is contractive in n dimensions if there is a b such that $\operatorname{tr}(F^2) > b > 0$. This condition is equivalent to $\|S\|^2 > \|A\|^2 + b$, where $F = A + S$, $A = -A^T$, $S = S^T$, and $\|\cdot\|$ is the Frobenius (sum of squares) norm. This condition is satisfied if all the eigenvalues of F are bounded away from the sectors $\frac{\pi}{4} < |\theta| < \frac{3\pi}{4}$; in particular, if they are all real and bounded away from zero.*

Proof. Let λ_i be the eigenvalues of F. For Euler's method we have

$$\begin{aligned}\ln\det A &= \ln\det(I+hF)\\ &= \ln\prod_i(1+h\lambda_i)\\ &= \sum\ln(1+h\lambda_i)\\ &= h\sum_i\lambda_i - \frac{1}{2}h^2\sum_i\lambda_i^2+\mathcal{O}(h^3)\\ &= h\operatorname{tr}F-\frac{1}{2}h^2\operatorname{tr}(F^2)+\mathcal{O}(h^3).\end{aligned}$$

If there is a b such that $\operatorname{tr}(F^2)>b>0$, this is less than 0 for all small enough h, i.e., the method is contractive. Splitting F into its symmetric and antisymmetric parts,

$$\operatorname{tr}(F^2)=\sum_{i,j}F_{ij}F_{ji}=\sum_{i,j}(S_{ij}+A_{ij})(S_{ij}-A_{ij})=\|S\|^2-\|A\|^2,$$

giving the second part of the proposition. Now $\operatorname{tr}(F^2)=\sum_i\lambda_i^2$, and if each λ_i is outside the specified sectors, then each real eigenvalue or complex conjugate pair of eigenvalues gives a positive contribution to this sum, giving the last part of the proposition. □

Note that the eigenvalues of elliptic or nearly elliptic fixed points lie near the imaginary axis—right in the middle of the "bad" sector. Perhaps this was only to be expected.

Experts will recognize the last part of Proposition 8 as the appearance of an *order star* of a Runge-Kutta method [4] (the set $\{z : |R(z)| < |e^z|$ where $R(z)$ is the method's linear stability polynomial). For linear problems, or nonlinear problems with 1-stage methods, a method is more contractive than the flow of the ODE if $h\lambda_i$ lies in the order star of the method for each eigenvalue λ_i. However, this seems rather restrictive so we do not explore further.

Proposition 9. *There are explicit contractive integrators.*

Proof. Let f be any contractive vector field with $\|F\| < L$. Because eigenvalues vary continuously and can only become imaginary when two eigenvalues meet, and because symmetric matrices have real eigenvalues, there is a symmetric, traceless matrix M with distinct eigenvalues such that the derivative of $f_1 := f - Mx$ has real eigenvalues. Let $f_2 := Mx$ and split $f = f_1 + f_2$. f_1 is contractive and admits an explicit contractive integrator (e.g. Euler's method, see Proposition 8); f_2 is traceless and can be solved explicitly. Composing these maps gives the result. □

We close with some open questions we hope to report on in the future.

1. Are there explicit contractive integrators of any order? (Proposition 9 constructs a first order method.) There are if one only demands *linear* contractivity. The order cannot be increased by composition, because the adjoint of Euler's method—backward Euler—is not contractive for f_1.

2. The present method reduces to the volume-preserving method of Feng and Wang [2] when the vector field f is traceless. There is another approach to volume-preserving integration due to Quispel [8] and to Shang [10], which does not rely on a splitting at all; moreover, it has a generalization to systems preserving non-Euclidean measures, which we have not even considered here. Can this approach be carried over to the contractive case?
3. The splitting used in the proof of Proposition 5 writes $f = a+b$ where $\mathrm{tr}(da) = 0$ and $b \equiv 0$ when $\mathrm{tr}(df) \equiv 0$. Are there splittings with the property that $b(x) = 0$ when $\mathrm{tr}(df(x)) = 0$? If so, they could be used for systems in which $\mathrm{tr}(df(x))$ changes sign on a compact hypersurface; the interior would then be invariant and one could construct an integrator which preserved it and was contractive there. This was done for the case of dissipation of scalar functions in [7].

Acknowledgements. The authors thank John Butcher and all the organizers of ANODE for financial support and for providing the atmosphere in which this paper was conceived, and the MSRI where it was concluded. We also thank the Marsden Fund of the Royal Society of New Zealand for their financial support.

References

[1] A. Bloch, P.S. Krishnaprasad, J.E. Marsden, and T.S. Ratiu, The Euler-Poincaré equations and double bracket dissipation, *Comm. Math. Phys.* **175** (1996), 1–42.
[2] K. Feng and D.-L. Wang, Dynamical systems and geometric construction of algorithms, in *Contemporary Mathematics* **163**, eds. Z. Shi and C. Yang, 1994.
[3] O. Gonzalez, Time integration and discrete Hamiltonian systems, *J. Nonlinear Sci.* **6**(5), 449–467 (1996).
[4] A. Iserles and S.P. Nørsett, *Order Stars*, Chapman & Hall, London, 1991.
[5] R.I. McLachlan and G.R.W. Quispel, Generating functions for dynamical systems with symmetries, integrals, and differential invariants, *Physica D* **112** (1998) 298–309.
[6] R.I. McLachlan, G.R.W. Quispel, and N. Robidoux, Geometric integration using discrete gradients, *Phil. Trans. Roy. Soc. A*, to appear.
[7] R.I. McLachlan, G.R.W. Quispel, and N. Robidoux, A unified approach to Hamiltonian systems, Poisson systems, gradient systems, and systems with Lyapunov functions and/or first integrals, *Phys. Rev. Lett.*, to appear.
[8] G.R.W. Quispel, Volume-preserving integrators, *Phys. Lett.* **206A** (1995) 26-30.
[9] J.-M. Sanz-Serna and M.P. Calvo (1994), *Numerical Hamiltonian Problems*, Chapman & Hall.
[10] Shang Zai-jiu, *J. Comp. Math.* **12** (1994), 265.
[11] A.M. Stuart and A.R. Humphries, *Dynamical systems and numerical analysis*, CUP, 1996.

Robert I. McLachlan, Mathematics, Massey University, Private Bag 11–222, Palmerston North, New Zealand
E-mail address: R.McLachlan@massey.ac.nz

G.R.W. Quispel, Mathematics Department, LaTrobe University, Bundoora, Melbourne 3083, Australia
E-mail address: R.Quispel@latrobe.edu.au