BIT Numer Math (2015) 55:531–558
DOI 10.1007/s10543-014-0503-3

Stiffness 1952–2012: Sixty years in search of a definition

Gustaf Söderlind · Laurent Jay · Manuel Calvo

Received: 5 February 2013 / Accepted: 30 May 2014 / Published online: 13 June 2014

Abstract Although stiff differential equations is a mature area of research in scientific computing, a rigorous and computationally relevant characterization of stiffness is still missing. In this paper, we present a critical review of the historical development of the notion of stiffness, before introducing a new approach. A functional, called the *stiffness indicator*, is defined terms of the logarithmic norms of the differential equation's vector field. Readily computable along a solution to the problem, the stiffness indicator is independent of numerical integration methods, as well as of operational criteria such as accuracy requirements. The stiffness indicator defines a local *reference time scale* Δt, which may vary with time and state along the solution. By comparing Δt to the range of integration T, a large *stiffness factor* $T/\Delta t$ is a *necessary condition* for stiffness. In numerical computations, Δt can be compared to the actual step size h, whose stiffness factor $h/\Delta t$ depends on the choice of integration method. Thus Δt embodies the mathematical aspects of stiffness, while h accounts for its numerical and operational aspects.To demonstrate the theory, a number of highly nonlinear

Communicated by Anne Kværnø.

In memory of our colleague and friend Jan G. Verwer (1946–2011).

G. Söderlind (✉)
Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden
e-mail: Gustaf.Soderlind@na.lu.se

L. Jay
Department of Mathematics, The University of Iowa, 14 MacLean Hall, Iowa City, IA 52242-1419, USA
e-mail: laurent-jay@uiowa.edu

M. Calvo
Departamento de Matemática Aplicada, Pza. San Francisco s/n, Universidad de Zaragoza, 50009 Zaragoza, Spain
e-mail: calvo@unizar.es

test problems are solved. We show, *inter alia*, that the stiffness indicator is able to distinguish the complex and rapidly changing behavior at (locally unstable) turning points, such as those observed in the van der Pol and Oregonator equations. The new characterization is mathematically rigorous, and in full agreement with observations in practical computations.

Keywords Initial value problems · Stability · Logarithmic norms · Stiffness · Stiffness indicator · Stiffness factor · Reference time scale · Step size

Mathematics Subject Classification (2000) 65L04 · 65L05

1 Introduction

Over the years, many different numerical methods and codes have been designed for the efficient solution of stiff initial value problems. Nevertheless, 60 years after the seminal paper of Curtiss and Hirschfelder [5] which opened this field, stiffness has yet to be properly defined. As Hairer and Wanner remark, [11, p. 1]:

> While the intuitive meaning of stiff is clear to all specialists, much controversy is going on about its correct mathematical definition [...]. The most pragmatical opinion is historically the first one (Curtiss and Hirschfelder 1952): *stiff equations are equations where certain implicit methods, in particular BDF, perform better, usually tremendously better, than explicit ones*.

Although the (originally) emphasized passage is usually attributed to Curtiss and Hirschfelder [5], it is in fact not contained in that paper. Reportedly it reflects the opinion of the authors at the time. But even though most researchers agree with this characterization, it immediately leads to difficulties. For example, the "performance" of computational methods is part of the characterization, and it is not mentioned that superior performance crucially depends on "certain implicit methods" having large enough stability regions. Further, it offers no clues as to how one can find out whether an initial value problem exhibits stiffness, without trying different solvers. Dekker and Verwer [8, p. 5] write:

> The problems called stiff are diverse and it is rather cumbersome to give a mathematically rigorous definition of stiffness.

This is confirmed by Shampine [16], who writes:

> A major difficulty is that stiffness is a complex of related phenomena, so that it is not easy to say what stiffness is.

In fact, according to Cash [4], there is (as yet?) no proper definition:

> One of the major difficulties associated with the study of stiff differential systems is that a good mathematical definition of the concept of stiffness does not exist.

In spite of this, research has not suffered, as Ekeland et al. [9] suggest:

> It is perhaps true that a precise definition of stiffness is not crucial for practical purposes.

Dekker and Verwer [8, p. 13] appear to support a similar view:

> Stiff problems from practice are well recognized.

Nevertheless, even though stiffness is phenomenologically well understood, the lack of a proper definition is unsatisfactory, not least from a pedagogical perspective. There is a need to define stiffness in a reasonably rigorous, simple and mathematically appealing way, rather than relying on descriptive approaches, in terms of operational criteria, method classes, software performance, or various notions of how "computationally demanding" a problem is or might be.

In this paper we construct a new concept of stiffness that is both mathematically rigorous and of direct computational relevance. This includes an effort of unifying many previous well-motivated and justified characterizations of stiffness. We limit ourselves to considering initial value problems in which perturbations are, at least in part, strongly damped in forward time. Thus we do not consider weakly damped highly oscillatory problems to be stiff—such problems are of a different nature and pose other requirements than those associated with strongly damped systems.

The approach we take is inspired by the following quote (original emphasis) from Dekker and Verwer, [8, p. 5]:

> *The essence of stiffness is that the solution to be computed is slowly varying but that perturbations exist which are rapidly damped.*

Shampine [17, p. 4] has expressed the same view, perhaps even more aptly:

> A way we prefer for describing the latter condition is that [the solution] is very unstable in the opposite direction [of time].

Noting that both quotes refer to the solution of the problem, we view them as *a posteriori* characterizations, as opposed to *a priori* characterizations, which would make no use of the problem's solution, but only of the vector field itself. Another observation is that stiffness is related to the stability of a given solution, and therefore requires a *norm* for its analysis. Thus we abandon the "classical" discussion of stiffness in terms of eigenvalues, in favor of characterizing damping rates in terms of *logarithmic norms*. The latter are used to construct a new functional, the *stiffness indicator*, which is readily computable a posteriori, along a solution to the differential equation.

The stiffness indicator depends exclusively on the problem, and is *independent of computational criteria* such as method order, stability regions and accuracy requirements. It may vary in magnitude along the solution, and determines a local *reference time scale* $\Delta t > 0$, such that any given relevant time scale or step size $h > 0$ (which may depend on method choice and operational criteria etc.) can be assessed with respect to stiffness through a simple, dimensionless *stiffness factor*, $h/\Delta t$. Thus a time scale h is "stiff" whenever $h/\Delta t \gg 1$. The practical use of this theory will be clearly demonstrated by solving a number of nontrivial, nonlinear test problems, analyzing how stiffness varies along the computed solution.

2 A brief review and critique of the notion of stiffness

It is impossible to give a complete review of the historical development of the notion of stiffness in a short space, but a few carefully selected quotes from the literature serve well to illustrate how various mathematical thoughts and characterizations have evolved. Hairer and Wanner [11, p. 2] begin their treatment of stiff problems with a deceptively simple sentence:

> Stiff equations are problems for which explicit methods don't work.

On closer inspection, it is quite informative. Thus explicit methods may suffer a breakdown when applied to the initial value problem $\dot{x} = f(x);\ x(0) = x_0$. The key word is *explicit*—it means that the method only uses direct evaluations of the function f to advance the solution, with no algebraic equation solving involved. The breakdown occurs in problems where $f'(x)$ is "large," causing severe time step size restrictions in explicit methods. The issue is resolved by instead using dedicated implicit methods, which invoke Newton-type methods for nonlinear equation solving.

This phenomenon is not unique to differential equations. It is also well known in *optimization*, where simple gradient methods have similar limitations. Thus, using steepest descent to minimize a convex functional $F(x)$ leads to the iteration $x^{m+1} = x^m - h\,\mathrm{grad}_x F(x^m)$, where the line search step size h is changed on every step. But this is merely the explicit Euler method applied to the ODE $\dot{x} = -\mathrm{grad}_x F(x)$. In problems with steep gradients, then, the iteration may slow dramatically, making Newton-type methods preferable in spite of their relative complexity. In ODE terms, this corresponds to replacing the explicit Euler method by an implicit method, or a Rosenbrock-type method, designed for solving stiff ODEs.

Likewise, in *nonlinear equations*, fixed point iteration $x^{m+1} = f(x^m)$ does not converge for problems where $f'(x)$ is large. Unsurprisingly, the remedy is to turn to Newton-type methods. In a similar way, in *iterative solvers* for large linear systems, convergence may become painfully slow unless appropriate "preconditioning" is used. For example, if (damped) Jacobi iteration is used to solve the (discrete) Laplace equation $\Delta u = 0$, the iteration is equivalent to using explicit Euler (in pseudo time) to solve the diffusion equation $u_t = \Delta u$ with the step size on the CFL limit. To speed up this unacceptably slow process, a preconditioner is necessary. Once again, this corresponds to replacing the explicit method by a Rosenbrock-type method better suited to solving the diffusion equation, well-known to be stiff.

In view of the manifold instances of closely related, well-understood problems in other areas of numerical analysis, it ought to be possible to define stiffness in mathematical terms. The question is: for what problems are explicit methods inadequate for solving $\dot{x} = f(x)$? As the common pattern above indicates, the step size h of explicit methods is restricted by the magnitude of $f'(x)$. It is therefore reasonable to expect that this restriction can be determined from $f'(x)$ alone, without making additional assumptions on operational criteria such as method choice and accuracy requirements. This is indeed the case, and the crucial issue is to expose how a given $f'(x)$ is associated to a characteristic time scale h.

The first mention of the word "stiff" by Curtiss and Hirschfelder in 1952 [5] discusses a scalar differential equation of the form

$$\dot{x} = \frac{1}{a(t,x)}(x - g(t)), \tag{2.1}$$

where we have changed variable names to match the notation used throughout this paper. For this equation, Curtiss and Hirschfelder characterize stiffness as follows:

> If Δt is the desired resolution of t or the interval which will be used in the numerical integration, the equation is "stiff" if
>
> $$\left|\frac{a(t,x)}{\Delta t}\right| \ll 1$$
>
> and g is well behaved.

Although this first characterization is imprecise and relies on a simple scalar model equation, its importance lies in relating a time scale Δt to the decay rate of transients. This rate is supposedly governed by the coefficient $a(t,x)$, which is assumed to be negative. As the coefficient depends on t and x, the time scale Δt will vary accordingly along the solution. A simplified case is to take $a(t,x)$ constant, say $a(t,x) = 1/\lambda < 0$. Then the equation is "stiff" with respect to Δt if $|\lambda \Delta t| \gg 1$.

To extend the characterization above, linear systems of equations of the form $\dot{x} = Ax + g(t)$ were considered. A common early approach is the one explained by Lambert [13, pp. 231ff]. It is assumed that the eigenvalues of A are located in the left half plane; some have large negative real parts while others have small. The *stiffness ratio* is defined as $\max |\mathrm{Re}\,\lambda[A]| / \min |\mathrm{Re}\,\lambda[A]|$, and a large stiffness ratio is claimed to be characteristic of stiff systems.

Although such a span in negative real parts of eigenvalues is often observed, it is neither necessary nor sufficient for stiffness, as Byrne and Hindmarsh [3, pp. 3ff] pointed out. First, as Curtiss and Hirschfelder had already suggested, there are scalar stiff problems (hence with a unit stiffness ratio); second, the stiffness ratio is not related to any time scale; and third, stiffness often varies along the solution and is not necessarily a global property. Worse still, the stiffness ratio breaks down for singular matrices, for which the characterization is simply misleading.

For nonlinear systems Lambert suggests checking the stiffness ratio of the eigenvalues of the Jacobian matrix, [13, p. 232]. This approach is highly questionable, however, and Artemiev and Averina [1, p. 6] note:

> For example, a nonlinear autonomous ODE system with $f(0) = 0$ can be called stiff if real parts [sic] of the eigenvalues of the Jacobi matrix $[f'(0)]$ satisfy the above conditions. However, the famous van der Pol equation, which is frequently used as an example of stiff ODEs, does not satisfy this definition. […] it is impossible to determine stiffness only by means of eigenvalues of the Jacobi matrix for nonlinear systems […]

Although this criticism is in part valid (eigenvalues cannot be used) it is off the mark, as the zero solution of the van der Pol equation is an unstable equilibrium. Thus stiffness does not occur near the origin, but only *along the limit cycle*. As the stability properties near the attractor are nontrivial, however, a precise stiffness characterization

must rely on local information about the solution. The a posteriori stiffness indicator constructed in this paper will address this issue and will be demonstrated to resolve the questions raised by Artemiev and Averina.

As an alternative to using eigenvalues, Higham and Trefethen [12] suggest characterizing stiffness in terms of *pseudospectra*, as non-normality can have a significant effect on the transient behavior of the system. It is correctly argued that short-term behavior is more important than the long-term, asymptotic behavior associated with the eigenvalues. Further, unlike the eigenvalues, the pseudospectrum depends on the choice of norm, which is essential since stability analysis in general requires a norm.

Earlier attempts avoiding eigenvalues (and including nonlinear systems) characterize stiffness as systems with *large Lipschitz constants*, see e.g. Dahlquist [7]. If such a system is integrated using an explicit method, the maximal step size is inversely proportional to the Lipschitz constant, $h \sim 1/L$. Hence, if the range of integration is $[0, T]$, the number of steps required to complete the integration is $N \sim LT$ and may become exceedingly large if $LT \gg 1$. The latter condition essentially defines what is meant by a "large Lipschitz constant," and combines an equation property, L, with a time scale, T.

But a large Lipschitz constant does not capture stiffness either. The main shortcoming is that it does not distinguish between solving the problem in forward and reverse time. While changing the direction of time leaves the Lipschitz constant invariant, stability changes completely. Yet, with an appropriate method, some systems with arbitrarily large Lipschitz constants—stiff problems—can be solved "efficiently," by which we mean that an accurate solution can be obtained with a "small" number of steps $N \ll LT$. In fact, N may even be independent of the magnitude of LT. Discussing error bounds, Dahlquist recognized this already in 1958, [6, p. 52]:

> The Lipschitz constant is then wrong in principle.

Dahlquist also pioneered the study of A-stability, investigating methods having unbounded stability regions, for the purpose of solving stiff problems efficiently. Interestingly, this analysis can be carried out starting from the *scalar* test equation

$$\dot{x} = \lambda x. \tag{2.2}$$

Although (2.2) cannot represent stiffness according to some of the early characterizations, the point is that a method's stability region is determined by $h\lambda$, a dimensionless product of the problem parameter λ and the method parameter (time scale) h. In particular, an A-stable method (whose stability region covers $\mathbb{C}^-$) does not impose any stability restrictions on the step size h if $\operatorname{Re}\lambda \le 0$, no matter how large $|\lambda|$ is. Noting that in (2.2) the Lipschitz constant is $L = |\lambda|$, this means that we may use "large step sizes" in the sense that $hL \gg 1$. This can be compared to the previously mentioned product LT, with the distinction that in (2.2) the sign of $\operatorname{Re}\lambda$ (or equivalently, the direction of integration) matters.

A completely different approach is proposed by Brugnano et al. [2]. They suggest defining the "stiffness ratio" by

$$\rho = \sup_{\eta} \frac{\|x(\cdot\,;\eta)\|_{L^1[0,T]}}{\|x(\cdot\,;\eta)\|_{L^\infty[0,T]}},$$

where η is the initial condition of the solution x on $[0, T]$, i.e., $x(0; \eta) = \eta$. Interestingly, their notion can also be used for boundary value problems (although computational difficulties are then different in nature). However, since ρ is simply a constant, it is a "global" notion and therefore unable to distinguish whether stiffness varies along the solution. Essentially, ρ is the ratio of the range of integration T to the shortest intrinsic time scale T^*, as defined by the fastest decay rate.

The examples above are far from exhaustive, but they illustrate that defining stiffness is nontrivial, with a large variety of ideas in the vast, existing literature, see e.g. [18] for further critique. In the end, researchers have often preferred to rely on experience: every practicing numerical analyst learns what is special about stiff problems by solving a few. Even so, we shall develop the idea, expressed by Dekker and Verwer as well as by Shampine, that in order to characterize stiffness one must consider perturbation damping rates, as well as the stability of the solution to the system.

3 A priori stiffness—an informal characterization

Separating the mathematical properties of stiffness and operational criteria can be done, arguing that certain mathematical criteria are *necessary* for stiffness, and would also be *sufficient* under additional operational conditions. We shall seek such necessary conditions, and distinguish between an a priori characterization of stiffness, in terms of the system of differential equations alone, and an a posteriori characterization, which also assumes that a solution $x(t)$ is known, defining stiffness with respect to that solution. The a posteriori analysis makes it possible to localize stiffness and quantify it at any given point in time and state space. The latter approach is treated separately in the next section.

The scalar Prothero–Robinson equation [15]

$$\dot{x} = \lambda\big(x - g(t)\big) + \dot{g}(t); \quad x(0) \neq g(0), \quad t \in [0, T], \tag{3.1}$$

where $\lambda \ll -1$, is a well-known model problem similar to the Curtiss and Hirschfelder equation (2.1), see also Gear [10, p. 211]. The solution is

$$x(t) = \mathrm{e}^{t\lambda}\big(x(0) - g(0)\big) + g(t), \tag{3.2}$$

where $g(t)$ is the particular solution and the exponential term correspond to transients. As a method of order p recovers polynomial solutions up to degree p exactly, the appropriate step size for computing the particular solution is only a matter of how well $g(t)$ can be locally approximated by a polynomial $P(t)$ of degree p. The error $|g(t) - P(t)|$ will remain sufficiently small on a grid of mesh width H, say. Although H could vary along the solution, we shall in this preliminary discussion assume that it is constant.

If an explicit method is used to solve the problem, there is a potential conflict between the time scale H of $g(t)$ and the time scale of the exponential $\mathrm{e}^{t\lambda}$. Thus numerical stability will restrict the actual step size $h > 0$ such that $|h\lambda| \approx 1$, irrespective of whether the transient has decayed or not. As $h \sim 1/|\lambda|$, *the integration effectively stalls if* $\lambda \to -\infty$, and even a moderately large negative λ may require $h \ll H$. This condition represents stiffness, which can be quantified in terms of the magnitude of the dimensionless quantity $H\lambda$, or, preferably, H/h. The problem is *non-stiff* if $|H\lambda| \lessapprox 1$ and *stiff* if $H\lambda \ll -1$. Should λ be positive, however, the homogeneous solution never decays. It dominates the solution (3.2), and stiffness will not occur.

As a scalar model problem is too simplistic, we need to extend the analysis. The Prothero–Robinson problem can be generalized to nonlinear systems of the form

$$\dot{x} = f\big(x - g(t)\big) + \dot{g}(t), \quad t \in [0, T], \tag{3.3}$$

where the vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $f(0) = 0$ and a monotonicity condition[1]

$$m[f] \cdot \|u - v\|^2 \le \langle u - v, f(u) - f(v)\rangle \le M[f] \cdot \|u - v\|^2, \qquad u, v \in D. \tag{3.4}$$

Here $m[f]$ and $M[f]$ are the lower (g.l.b.) and upper (l.u.b.) logarithmic Lipschitz constants, respectively [20]. We further assume that $x(0) - g(0) \in D \subset \mathbb{R}^n$.

Letting e^{tf} denote the flow of $\dot{z} = f(z)$, the solution of (3.3) is structurally analogous to the linear problem (3.2). Thus it can be written

$$x(t) = \mathrm{e}^{tf}\big(x(0) - g(0)\big) + g(t),$$

where we assume that the flow is positively invariant on the simply connected set D, i.e., $e^{tf}(D) \subseteq D$ for $t \in [0, T]$. This implies that D contains the vector $x(t) - g(t)$ for all $t \in [0, T]$.

Again, suppose that $g(t)$ can be approximated on a mesh of width $H < T$. As $e^{tf}(0) = 0$, transients are bounded by $\|\mathrm{e}^{tf}\big(x(0) - g(0)\big)\| \le L[\mathrm{e}^{tf}] \cdot \|x(0) - g(0)\|$ for $t \ge 0$, where $L[\mathrm{e}^{tf}]$ is the l.u.b. Lipschitz constant of the flow on D. This Lipschitz constant is in turn bounded below and above by

$$\mathrm{e}^{tm[f]} \le L[\mathrm{e}^{tf}] \le \mathrm{e}^{tM[f]}; \qquad t \ge 0,$$

where both inequalities are sharp for small $t \ge 0$, [20]. Therefore, the maximal perturbation *growth rate* is determined by $M[f]$, while the fastest perturbation *decay rate* is determined by $m[f]$. Note that the fastest decay rate in forward time corresponds the fastest growth rate in reverse time—this is reflected by the property $M[-f] = -m[f]$, which follows directly from (3.4).

[1] The analysis in this paper can be carried out with respect to any given norm with analogous results. However, for simplicity we have chosen to work with inner product norms, later further specialized to the usual Euclidean norm.

Let us first assume that the problem is *dissipative*, i.e. $M[f] < 0$. As transients decay, all solutions approach $g(t)$, with stiffness associated with the fastest decay rate. For an explicit method, the step size h would have to be restricted such that $|hm[f]| \approx 1$ in order to maintain numerical stability. Thus $h \sim 1/|m[f]|$, even after transients have decayed. In analogy with the scalar case, where $m[f] = \lambda$, stiffness is characterized by the condition $h \ll H$, or equivalently $Hm[f] \ll -1$.

In the scalar problem (3.1) we saw that stiffness does not occur when $\lambda > 0$. In the system (3.3), however, the situation is more intricate. Assuming that the solution is moderately unstable in forward time ($M[f] > 0$), it may still exhibit stiffness, depending on the relative magnitudes of $M[f]$ and $m[f] < 0$. The maximum growth rate of e^{tf} then defines the time scale $H > 0$, such that $HM[f] \approx 1$, as large steps cannot be used to resolve an unstable solution. Meanwhile, the maximum decay rate gives a time scale $h > 0$, such that $hm[f] \approx -1$ as before. Stiffness still occurs if $h \ll H$, which obviously implies $Hm[f] \ll -HM[f] \approx -1$. Hence the stiffness condition can be expressed in dimensionless form as

$$H(m[f] + M[f]) \ll -1, \tag{3.5}$$

relating the problem parameters $m[f]$ and $M[f]$ to the time scale H. Note that the two logarithmic Lipschitz constants are *added*. To our knowledge, there is no mention of a similar stiffness condition in the literature. It will be further elaborated as an a posteriori condition, and is the key notion in our characterization of stiffness.

There are simple interpretations for linear constant coefficient systems, i.e., when $f(x) = Ax$. A condition $H(M[A] - m[A]) \gg 1$ bounds the "gap" in real parts of the eigenvalues of HA, but the stiffness condition $H(M[A] + m[A]) \ll -1$ is entirely different. Thus stiffness is not caused by a large gap (which is nonexistent in the scalar case, as $M[\lambda] = m[\lambda] = \lambda$), but rather by *the necessary condition* $Hm[A] \ll -1$. Nevertheless, $M[A]$ is included in (3.5), where the sum $m[A] + M[A]$ reflects that *a positive* $M[A]$ *can partly or completely offset stiffness*, if $M[A]$ is of the same order as $m[A]$. This is readily observable in practical computations, and will be demonstrated to be an important effect e.g. in the van der Pol equation, which is not dissipative.

We finally note that the new characterization is consistent with most of the quotes mentioned in the previous sections. For example, a large value of $|m[f]|$ implies a *large Lipschitz constant*, [7], but not the other way around; further, having *eigenvalues far into the left half-plane* will imply that $m[f]$ is large and negative, but the gap in eigenvalues is immaterial; finally, the particular solution is *highly unstable in reverse time*, [17], and it features *rapidly damped transients*, [8]. Thus most previous characterizations of stiffness are covered by (3.5), although the latter is more general and will be seen to provide essential new insight.

Parabolic differential equations is a classical example of stiffness in applied mathematics. Consider the diffusion problem

$$u_t = u_{xx} \equiv \mathcal{L}u; \qquad u(t, 0) = u(t, 1) = 0,$$

with suitable initial data. Diffusion is an irreversible process, mathematically reflected by the flow $e^{t\mathcal{L}}$ being a *semigroup*. Thus, on the space $C_0^1[0,1] \cap L^2[0,1]$, we have $m[\mathcal{L}] = -\infty$. As a consequence, for any reasonable spatial discretization, stiffness is inevitable. This is directly linked to the fact that the diffusion equation in reverse time is *ill-posed*, cf. Shampine's remark on the solution of a stiff problem being "very unstable" in reverse time, [17].

Let us discretize $\partial^2/\partial x^2$ using equidistant second order finite differences. Taking $\Delta x = 1/(N+1)$, we approximate $u_t = u_{xx}$ by the method of lines ODE $\dot{U} = T_{\Delta x} U$, where the $N \times N$ matrix $T_{\Delta x} = (1/\Delta x^2) \cdot \text{tridiag}(1 \quad -2 \quad 1)$ has N eigenvalues

$$\lambda_k = -4(N+1)^2 \sin^2\left(\frac{k\pi}{2(N+1)}\right); \quad k = 1:N.$$

As $T_{\Delta x}$ is a normal matrix, choosing the Euclidean norm gives sharp logarithmic norms in (3.4), which then takes the form

$$m_2[T_{\Delta x}] \cdot \|U\|_2^2 \le U^{\mathrm{T}} T_{\Delta x} U \le M_2[T_{\Delta x}] \cdot \|U\|_2^2,$$

with $m_2[T_{\Delta x}] = \lambda_N \approx -4/\Delta x^2$ and $M_2[T_{\Delta x}] = \lambda_1 \approx -\pi^2$. Even without making use of a known solution, we can conclude that if any explicit method were used for forward time integration, numerical stability would restrict the time step size Δt such that $|\Delta t \cdot m_2[T_{\Delta x}]| \lessapprox 1$. This implies that

$$\frac{\Delta t}{\Delta x^2} \lessapprox \frac{1}{4},$$

known as a *CFL condition*. Thus Δt is restricted by reverse dynamics in terms of $m_2[T_{\Delta x}]$. Therefore, stiffness increases without bound when the spatial resolution is increased ($\Delta x \to 0$), even if the solution is smooth.

As is well known, CFL conditions can be overcome by using unconditionally stable methods for time discretization; this is exactly the same as overcoming stiffness by using an appropriate implicit time stepping method. It should however be noted that we here rule out hyperbolic PDEs, although explicit time discretizations lead to CFL conditions also for such problems. But hyperbolic problems are well-posed in forward as well as reverse time, and exhibit little or no damping. They are related to highly oscillatory problems, which are usually not regarded as stiff, [8, p. 9].

4 A posteriori stiffness: a rigorous characterization

Although the a priori analysis above may offer sufficient insight e.g. for linear systems with constant coefficients, it is often far too crude for nonlinear systems, as the stability properties of the flow can vary significantly in the domain of interest. Accordingly, as Byrne and Hindmarsh [3] point out, stiffness typically varies along the solution.

An a posteriori characterization of stiffness is *local in time as well as in space*. This is achieved by studying the *variational equation* associated with the differential equation, in particular the *short-time* growth of *small* perturbations. Computable local

information on stiffness is derived from the evolution of infinitesimal perturbations on infinitesimal time intervals. The a posteriori characterization is otherwise in line with the previous a priori characterization of stiffness.

Consider the initial value problem

$$\dot{x} = f(t, x); \quad x(0) = x_0, \tag{4.1}$$

where $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be $\mathcal{C}^1$. A perturbed solution $x(t) + \delta x(t)$, with $\delta x(0) \neq 0$, then satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta x = f\left(t, x(t) + \delta x\right) - f\left(t, x(t)\right).$$

As $f \in \mathcal{C}^1$, infinitesimal perturbations satisfy the variational equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta x = J(t, x(t))\, \delta x, \tag{4.2}$$

where $J(t, x) = D_x f(t, x)$. Taking the inner product with δx, we obtain

$$\langle \delta x,\, J(t, x(t))\, \delta x \rangle = \left\langle \delta x,\, \frac{\mathrm{d}}{\mathrm{d}t}\delta x \right\rangle = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\delta x\|^2 = \|\delta x\|^2 \frac{\mathrm{d}}{\mathrm{d}t}\log\|\delta x\|. \tag{4.3}$$

Next, the g.l.b. and l.u.b. logarithmic norms at $(t, x(t))$ are the best possible bounds satisfying

$$m[J(t, x(t))] \leq \frac{\langle \delta x,\, J(t, x(t))\, \delta x \rangle}{\|\delta x\|^2} \leq M[J(t, x(t))] \tag{4.4}$$

for all $\delta x \neq 0$. From (4.3) we then obtain the differential inequalities

$$m[J(t, x(t))] \leq \frac{\mathrm{d}}{\mathrm{d}t}\log\|\delta x\| \leq M[J(t, x(t))]. \tag{4.5}$$

Whereas the right differential inequality bounds the maximum growth of perturbations, the left limits the maximum decay rate. The two inequalities (4.5) are sharp for short-term propagation, although not for the same initial condition, as the quadratic form in (4.4) attains its extreme values at different vectors δx.

Three observations are of great importance. First, by considering perturbation propagation on infinitesimal time intervals, there is no need to integrate the differential inequalities in (4.5). Second, considering infinitesimal perturbations justifies the use of the variational equation without viewing it as an approximation. Third, and most importantly, using logarithmic norms accounts for the topology and overcomes the difficulties associated with using eigenvalues of local linearizations.

To construct a computable a posteriori characterization of stiffness, based on the variational equation (4.2) along a solution $x(t)$, we make the following definition:

Definition 4.1 For a given matrix $A \in \mathbb{R}^{n\times n}$ the *stiffness indicator* is defined by

$$\sigma[A] = \frac{m[A] + M[A]}{2}. \tag{4.6}$$

Although (4.6) deviates slightly from (3.5) and the previously discussed special cases by introducing the average instead of the sum $m[A]+M[A]$, it is preferable due to its structural properties.

Theorem 4.1 *The stiffness indicator has the following elementary properties:*

1. $\sigma[0] = 0$
2. $\sigma[I] = 1$
3. $\sigma[sI + A] = s + \sigma[A]; \quad s \in \mathbb{R}$
4. $\sigma[\alpha A] = \alpha\,\sigma[A]; \quad \alpha \in \mathbb{R}$
5. $|\sigma[A] - \sigma[B]| \le \|A - B\|$
6. $m[A] \le \sigma[A] \le M[A]$.

The first three properties are simple consequences of the properties of logarithmic norms. The fourth follows from the logarithmic norm property $M[-A] = -m[A]$, and implies that the stiffness indicator has *odd parity*, i.e., $\sigma[-A] = -\sigma[A]$. Hence, if a linear system $\dot{x} = Ax$ is considered in reverse time, we need to solve $\dot{x} = -Ax$, and the construction above guarantees that the stiffness indicator handles forward and reverse time in a consistent way. The fifth property implies that the stiffness indicator is *continuous* and is proved by noting that both the lower and upper logarithmic norms are continuous and satisfy $|m[A]-m[B]| \le \|A-B\|$ and $|M[A]-M[B]| \le \|A-B\|$, respectively. Finally, the last property follows trivially from the average. Note that the eigenvalues $\lambda[A]$ are similarly bounded, as $m[A] \le \operatorname{Re}\lambda[A] \le M[A]$.

Depending on the problem class, the stiffness indicator obviously has the following structure:

$$\sigma[J(t,x(t))] = \begin{cases} \text{constant} & \text{if } f(t,x) \equiv Ax \\ \text{time-dependent} & \text{if } f(t,x) \equiv A(t)x \\ \text{state-dependent} & \text{if } f(t,x) \equiv f(x). \end{cases} \tag{4.7}$$

This is unaffected by the addition of a forcing function $g(t)$. The structure further implies that for linear problems, there is no difference between a priori and a posteriori stiffness, as the stiffness indicator is then *independent of the solution* x and can be determined a priori. By contrast, a posteriori stiffness is a necessary tool to address stiffness in nonlinear problems. For this reason all computational experiments at the end of this paper are nonlinear.

The stiffness indicator is easily computed. Let $\lambda[A]$ denote the eigenvalues of a matrix A. Then, for the Euclidean norm, it holds that

$$M_2[A] = \max \lambda[\mathrm{He}(A)]\,; \qquad m_2[A] = \min \lambda[\mathrm{He}(A)],$$

where $\mathrm{He}(A) = (A + A^{\mathrm{T}})/2$ denotes the Hermitian part of the matrix A. (Other norms can also be used, provided that the logarithmic norms in (4.6) are computed using well-known expressions, see e.g. [20]).

Note that if A is skew-Hermitian, then $\mathrm{He}(A) = 0$ and $M_2[A] = m_2[A] = 0$. Hence $\sigma_2[A] = 0$, implying that stiffness will not occur—skew-Hermitian systems are oscillatory. As mentioned before, we only associate stiffness with systems that at least in part exhibit strongly damped perturbations, [8, p. 9].

A necessary condition for stiffness is that $\sigma[J(t, x(t))]$ is "large" and negative, (3.5). In order to quantify this statement, we need to specify what is meant by "large," by relating the stiffness indicator to a time scale.

Definition 4.2 Let $J(t, x(t))$ denote the Jacobian along a solution $x(t)$ of (4.1) on $[0, T]$. For forward time integration, we define the *reference time scale* $\Delta t > 0$ at t by

$$\Delta t = \begin{cases} T & \text{if} \quad \sigma[J(t, x(t))] \geq 0 \\ \min(T, -1/\sigma[J(t, x(t))]) & \text{if} \quad \sigma[J(t, x(t))] < 0. \end{cases} \tag{4.8}$$

The different clauses in (4.8) simply guarantee that $\Delta t \in (0, T]$, as neither a negative time scale nor one larger than T are meaningful. Note that just like the stiffness indicator, the reference time scale is exclusively a problem property, *independent of the choice of integration method*. It does not represent the step size in use. As the stiffness indicator is typically negative along a stable solution, the reference time scale at t is often simply given by $\Delta t = -1/\sigma[J(t, x(t))]$.

In practical computations, the step size h will be determined by method properties, problem properties, solution properties and other operational criteria, such as accuracy requirements. However, if an explicit integrator is used, the method's local step size h cannot essentially exceed the problem's local reference time scale Δt. Assuming that $\sigma[J(t, x(t))] < 0$ on $[0, T]$, its *computational work per integrated unit of time* is proportional to $-\sigma[J(t, x(t))]$. In order to complete the integration, an explicit method using adaptive step size control therefore requires a total number of steps

$$N = -K_{\mathcal{M}} \int_0^T \sigma[J(t, x(t))] \, \mathrm{d}t, \tag{4.9}$$

where $K_{\mathcal{M}}$ is a method dependent constant of moderate size. It follows that N is large if $\sigma[J(t, x(t))]$ is large and negative over a significant portion of $[0, T]$. This estimate will also be verified in the numerical experiments.

Definition 4.3 Let Δt be the local reference time scale of the solution $x(t)$ of (4.1). The solution is called *locally stiff* at t with respect to a given time scale $h > 0$ if

$$S(t, h) := \frac{h}{\Delta t} \gg 1. \tag{4.10}$$

The quantity $S(t, h)$ is called the *stiffness factor at t with respect to the time scale h*.

The time scale h can be chosen in different ways. For example, taking $h = T$, the definition says that the solution $x(t)$ is locally stiff at t *with respect to the range of*

integration T, if $S(t,T) = -\sigma[TJ(t,x(t))] \gg 1$. Since only time scales $0 < h \leq T$ are of interest, we have $S(t,h) \leq S(t,T)$. The method independent characterization $S(t,T) = T/\Delta t \gg 1$ is therefore a *necessary condition* for stiffness.

In practical computations h is naturally interpreted as the local *step size* in use. The step size stiffness factor $S(t,h) = h/\Delta t$ then measures how much larger the actual step size is compared to the local reference time scale. Note that even though the stiffness indicator and reference time scale are constant in a linear, constant coefficient system, the step size stiffness factor $S(t,h)$ may still vary along the solution due to adaptive step size selection.

The local condition (4.10) can also be extended to a global characterization.

Definition 4.4 Let Δt be the local reference time scale at t. The solution is called *stiff* with respect to a time interval $[t_0, t_1] \subset [0,T]$ if

$$G(t_0,t_1) := \int_{t_0}^{t_1} \frac{\mathrm{d}\tau}{\Delta t(\tau)} \gg 1. \tag{4.11}$$

Here $G(0,T) \gg 1$ is a method-independent necessary condition for stiffness, quantifying the *stiffness of the problem*. If the stiffness indicator remains large and negative, $G(0,T)$ corresponds to the integral in (4.9). Further, as $\sigma[J(t,x(t))]$ is a continuous function of the state, it typically changes at the same rate as the Jacobian. Therefore, as computational experiments confirm, if $t_1 = t_0 + h$, where h is "small," we have $G(t_0, t_0+h) \approx S(t_0,h)$, which may still be huge. Thus the stiffness indicator and $G(0,T)$ provide method-independent information on stiffness, while the step size stiffness factors $S(t,h)$ provide information on how a particular method is doing when engaging a stiff problem.

For any quantification of stiffness, it is important to establish *scaling and invariance properties*. As stiffness depends in part on the range of integration, it is desirable to normalize the integration interval. Introducing a new "dimensionless" independent variable $\theta = t/T$ in (4.1), we have $\mathrm{d}t = T\mathrm{d}\theta$, and (4.1) is transformed into

$$\frac{\mathrm{d}x}{\mathrm{d}\theta} = T \cdot f(\theta T, x)\,; \quad \theta \in [0,1].$$

The corresponding normalized variational equation is

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\delta x = T \cdot J(\theta T, x(\theta))\,\delta x.$$

Consequently, the *normalized stiffness indicator* is $\sigma[TJ(\theta T,x)] = T\sigma[J(t,x)]$. Assuming that the stiffness indicator is negative, we further have

$$-1 = \Delta t \cdot \sigma[J(t,x)] = \Delta\theta \cdot T \cdot \sigma[J(t,x)] = \Delta\theta \cdot \sigma[TJ(\theta T,x)],$$

implying that the *normalized reference time scale* $\Delta\theta$ is the same fraction of $[0, 1]$ as Δt is of $[0, T]$. This invariance supports viewing $\sigma[TJ(t, x)]$ as an "absolute" quantification of stiffness.

However, some caution is required. For example, in the problem

$$\dot{x} = \sin t - x\,; \quad t \in [0, T], \tag{4.12}$$

the reference time scale is $\Delta t = 1$, so (4.12) is not a stiff problem on the unit interval, $T = 1$. As the normalized problem is

$$\frac{\mathrm{d}x}{\mathrm{d}\theta} = T \cdot (\sin(T\theta) - x)\,; \quad \theta \in [0, 1].$$

the reference time scale is $\Delta\theta = 1/T$, suggesting that the problem would exhibit stiffness for large T. However, T also affects the angular frequency in the forcing function, and to resolve the particular solution, actual step sizes h_θ will be of the order $Th_\theta \ll \pi$. This implies that h_θ will not exceed the normalized reference time scale $\Delta\theta = 1/T$. Due to the sampling theorem, stiffness will never be encountered in (4.12), no matter how large T is.

This is not a contradiction, as $\Delta\theta \ll 1$ is only a *necessary condition* for stiffness. In real computations, the reference time scale Δt is to be compared to the *proper step size* h, and stiffness is only encountered when the actual stiffness factor is $h/\Delta t \gg 1$.

The invariance properties also apply to nonlinear time transformations. Consider a time transformation

$$t = \Psi(\theta)$$

such that $\Psi \in \mathcal{C}^1$ is *monotonically increasing*, with $\Psi(0) = 0$ and $\Psi(1) = T$. This implies that there is a one-to-one correspondence between t and θ. Then $\mathrm{d}t = \Psi'(\theta)\mathrm{d}\theta$, with $\Psi'(\theta) > 0$. The transformed variational equation is

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\delta x = \Psi'(\theta) \cdot J(\Psi(\theta), x(\Psi(\theta)))\,\delta x$$

with transformed stiffness indicator

$$\sigma[\Psi'(\theta) \cdot J(\Psi(\theta), x(\Psi(\theta)))] = \Psi'(\theta) \cdot \sigma[J(t, x(t))].$$

However, corresponding to the differential relation $\mathrm{d}t = \Psi'(\theta)\mathrm{d}\theta$, the local time scale transformation at this point is $\Delta t = \Psi'(\theta)\Delta\theta$. Hence *the reference time scale is locally invariant* as a fraction of the integration interval.

For the simple case $\Psi(\theta) = \theta T$, we naturally obtain the previously discussed linear scaling. A similar analysis can also be carried out for terminal value problems and reverse time integration, by using the time transformation $\theta = (T - t)/T$, provided that proper attention is given to the resulting sign changes and the odd parity of the stiffness indicator.

5 Computational experiments

We shall demonstrate the new theory, analyzing some highly nonlinear equations that have previously eluded analysis in terms of "classical" stiffness concepts. The a posteriori stiffness indicator and stiffness factor offer new and detailed insight into the character and behavior of these nonlinear systems, such as how stiffness varies along the solution. High precision solutions and the a posteriori stiffness indicator were computed with MATLAB's `ode15s` solver. As a nonstiff reference, `ode45` was chosen to compare stiff and nonstiff solvers. Finally, for moderate precision stiff computations, a special version of `ode23s` was developed, where the standard step size selection scheme was replaced by a controller based on the digital filter H211PI, [19], in order to generate smooth step size sequences instead of the non-smooth, piecewise constant sequences from the standard version. The new solver, designated `ode23sdc` for "digital control," was used to demonstrate how adaptive step sizes and their corresponding stiffness factors vary along the solution of the problems. The computational experiments verify the theoretical claims, and show that a highly detailed quantitative analysis of stiffness is possible.

The first problem is the van der Pol equation,

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= \mu \cdot (1 - x_1^2)x_2 - x_1,\end{aligned}$$

with initial conditions $x(0) = (2 \quad 0)^{\mathrm{T}}$ and $\mu = 200$. The Jacobian depends only on the state, implying that stiffness indicator and reference time scale also depend on the state and vary along the solution.

The period of the limit cycle is $T \approx 1.6\mu$ when μ is large. We have therefore chosen to solve the problem over the interval $[0, 2\mu]$ in order to account for a full period. Introducing a new independent variable $\theta = t/(2\mu)$ we solve the *normalized equation*,

$$\begin{aligned}\frac{\mathrm{d}x_1}{\mathrm{d}\theta} &= 2\mu \cdot x_2\\ \frac{\mathrm{d}x_2}{\mathrm{d}\theta} &= 2\mu^2 \cdot (1 - x_1^2)x_2 - 2\mu \cdot x_1,\end{aligned}$$

for $\theta \in [0, 1]$, with normalized Jacobian

$$J(x) = \begin{pmatrix} 0 & 2\mu \\ -4\mu^2 x_1 x_2 - 2\mu & 2\mu^2(1 - x_1^2) \end{pmatrix}.$$

Its Hermitian part is

$$\mathrm{He}(J(x)) = \begin{pmatrix} 0 & -2\mu^2 x_1 x_2 \\ -2\mu^2 x_1 x_2 & 2\mu^2(1 - x_1^2) \end{pmatrix},$$

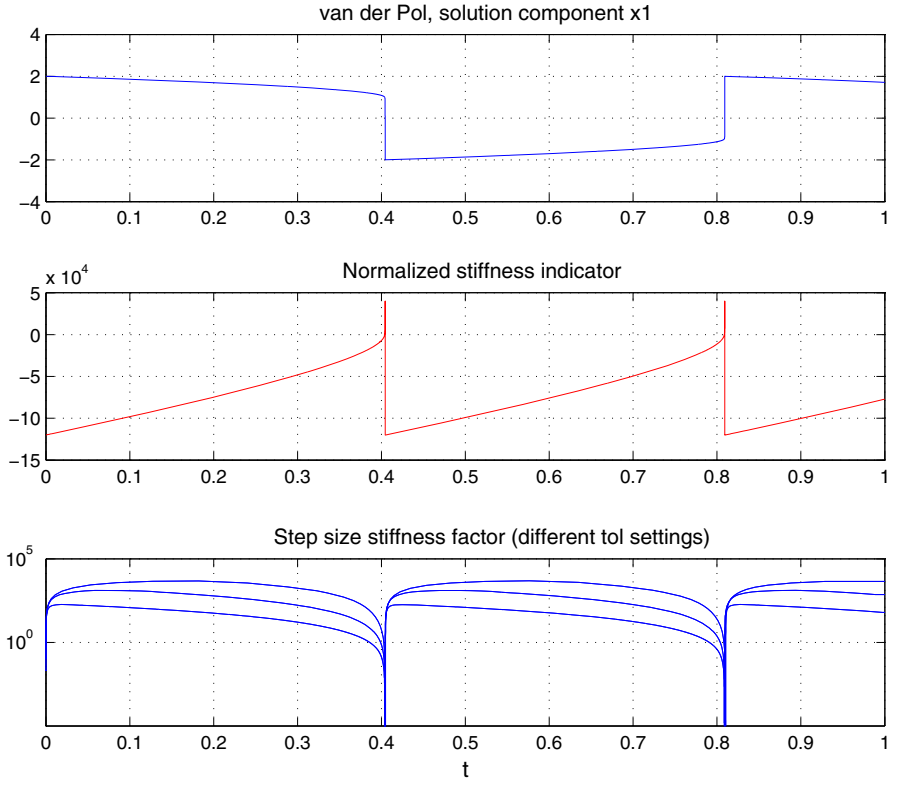

Fig. 1 *The van der Pol equation at* $\mu = 200$. Solution component x_1 (*top*) and normalized stiffness indicator $\sigma_2[J(x)] = (m_2[J(x)] + M_2[J(x)])/2$ (*center*) are plotted vs normalized time on [0, 1]. Large negative values of $\sigma_2[J(x)]$ correspond to stiffness. At turning points, however, where $|x_1| < 1$ and $\sigma_2[J(x)]$ becomes positive, there are extremely short nonstiff intervals. As the minimum value of $\sigma_2[J(x)]$ is $-3\mu^2$, effective stiffness is proportional to μ^2. The stepsizes used in `ode23sdc` for TOL $= 10^{-4}, 10^{-6}, 10^{-8}$ were used to compute actual step size stiffness factors (*bottom*), with the lowest graph corresponding to TOL $= 10^{-8}$. As TOL increases by two orders of magnitude, step sizes increase by one order of magnitude, causing a corresponding increase in the stiffness factor, which is large everywhere, except at the turning points

and $M_2[J]$ and $m_2[J]$ are, respectively, the largest and smallest eigenvalue of $\mathrm{He}(J(x))$. The two eigenvalues have opposite signs as $M_2[J]\cdot m_2[J] = -4\mu^4 x_1^2 x_2^2 \leq 0$. Thus the problem is not dissipative, and $M_2[J] \geq 0$ will affect stiffness. The normalized stiffness indicator is half the trace of $\mathrm{He}(J)$. Hence

$$\sigma_2[J(x)] = \frac{m_2[J(x)] + M_2[J(x)]}{2} = \mu^2(1 - x_1^2),$$

showing that stiffness depends exclusively on μ and x_1. More precisely, it scales like $\mathrm{O}(\mu^2)$, and *only occurs for* $|x_1| > 1$. Computational results are shown in Fig. 1.

As normalized stiffness grows like $\mathrm{O}(\mu^2)$, we also verify the work estimate (4.9) along the stiff branch of the solution from $\theta = 0$ to $\theta = 0.4$. In Fig. 1, e.g., we can graphically estimate

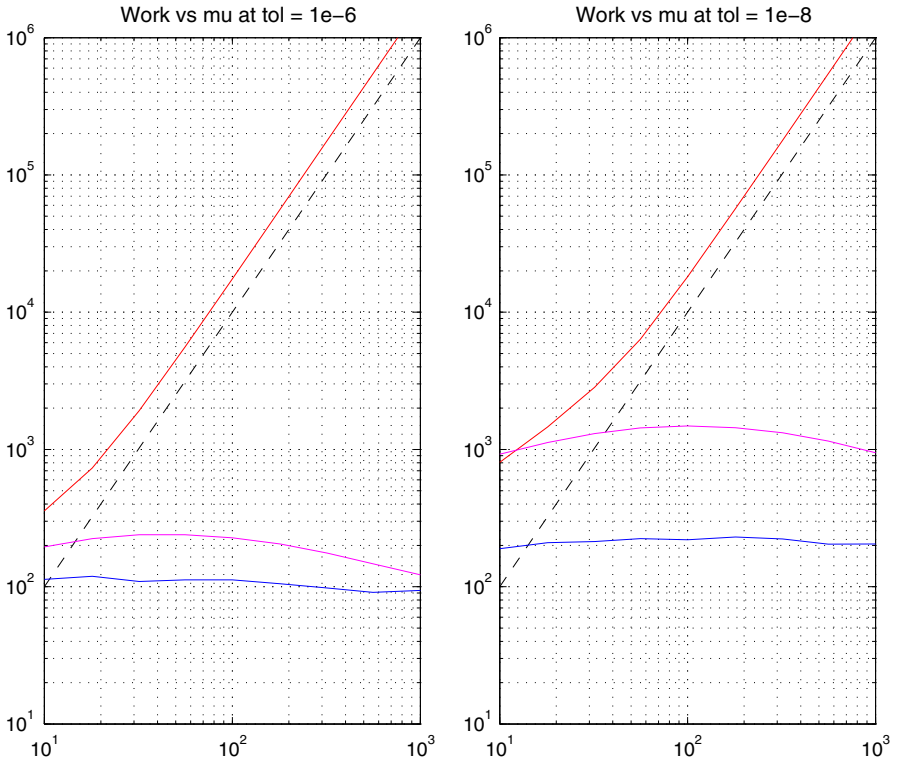

Fig. 2 *Computational effort.* Number of steps required to integrate the van der Pol equation from $\theta = 0$ to $\theta = 0.4$ is plotted as a function of μ. Left panel shows results with absolute and relative error tolerances of TOL $= 10^{-6}$; right panel with TOL $= 10^{-8}$. Upper solid curve was obtained using MATLAB's explicit Runge–Kutta solver `ode45`. A dashed reference line of slope 2 demonstrates that work is $O(\mu^2)$ as predicted by theory, independent of TOL. The two lower graphs in each panel were obtained using MATLAB's stiff solvers `ode15s` (*lower curve*) and `ode23sdc` (*middle curve*). Work is then effectively independent of μ, but instead depends on TOL. The integration interval covers the stiff branch only

$$N = -K_{\mathcal{M}} \int_0^{0.4} \sigma[J(x)]\,\mathrm{d}\theta \approx 15 \times 10^4 \cdot \frac{0.4-0}{2} \cdot K_{\mathcal{M}} = 30{,}000\, K_{\mathcal{M}},$$

where $K_{\mathcal{M}}$ is a method-dependent constant. To verify the estimate, we checked the number of steps used by the explicit adaptive Runge–Kutta method `ode45`, see Fig. 2. For $\mu = 200$, approximately 70, 000 steps were needed, *independent of* TOL, corresponding to $K_{\texttt{ode45}} \approx 7/3$. The value of $K_{\texttt{ode45}}$ accounts for method properties, notably the size of the stability region, which limits the step size. The method independent stiffness indicator, on the other hand, offers an accurate representation of stiffness derived from problem properties alone.

By contrast (but still in agreement with theory), in the stiff solvers `ode15s` and `ode23sdc` the computational effort depends on TOL, but is practically *independent of* μ, see Fig. 2. In `ode15s` at TOL $= 10^{-6}$, the average step size is $h_\theta = 0.4/100$,

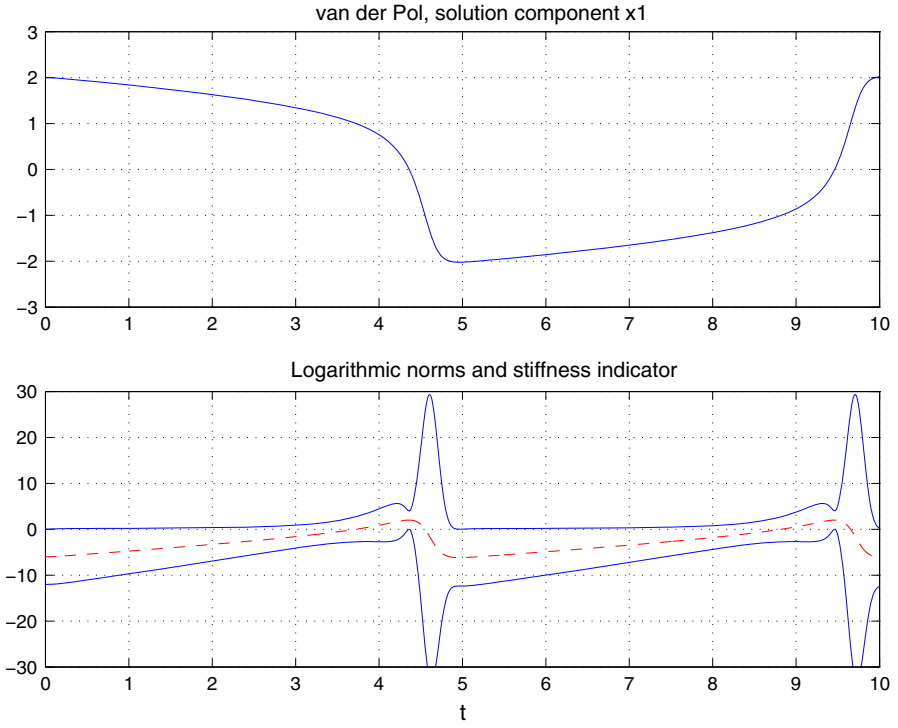

Fig. 3 *Logarithmic norms in the van der Pol equation.* Solution component x_1 (*top*) is plotted vs time over a full period for $\mu = 4$. Lower panel shows upper logarithmic norm $M_2[J(x)]$ and lower logarithmic norm $m_2[J(x)]$ (*both solid*), with stiffness indicator $\sigma_2[J(x)]$ (*dashed*) between the two. Note that $m_2[J(x)] \ll -1$ attains its *largest negative values near the turning point*, and that we simultaneously have $M_2[J(x)] \gg 1$. The stiffness indicator $\sigma_2[J(x)] = (m_2[J(x)] + M_2[J(x)])/2$ is *positive* at the turning point, implying that the reference time scale is $\Delta t = T$ there. The behavior is similar for larger values of μ

implying that $S(\theta, h_\theta) \sim \mu^2/100$. The stiffness factor roughly estimates the efficiency gain of the stiff solver; at $O(\mu^2)$ the gain is potentially stupendous.

The need to use the sum $(m_2[J] + M_2[J])/2$ in the definition of the stiffness indicator (4.6) is demonstrated in Fig. 3. For clarity the computations were carried out in non-stiff conditions at $\mu = 4$ (without normalization), as the graphs otherwise present too tall and sharp peaks to reveal essential features. Although we have seen that stiffness does not occur at a given time scale h unless $m_2[hJ] \ll -1$, *this condition is offset at the turning point*, where $|x_1| < 1$. Thus, in spite of $m_2[J]$ reaching its extreme negative value there, we simultaneously have $M_2[J] \gg 1$. Consequently, stiffness cannot be characterized by the condition $m_2[J] \ll -1$ alone.

A close-up of actual transition dynamics for $\mu = 200$ shows that there is no stiffness during the transition, see Fig. 4. Unfortunately, one cannot superimpose plots for different tolerances, since at this resolution, the global error causes a noticeable phase shift of the turning point, making a direct comparison difficult.

Finally, the van der Pol equation has an *unstable equilibrium* at $x = 0$. Since $\sigma_2[J(x)] = \mu^2(1 - x_1^2)$, we have $\sigma_2[J(0)] = \mu^2 > 0$ along the unstable equilibrium

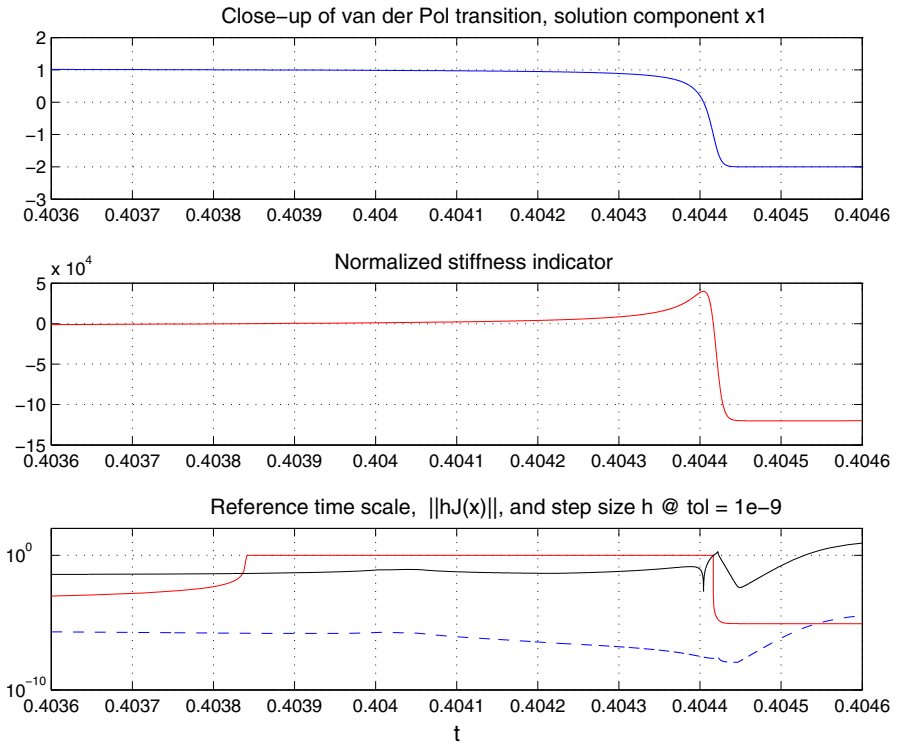

Fig. 4 *Close-up of transition point.* Solution component x_1 (*top*) and normalized stiffness indicator (*center*) are plotted to reveal transition dynamics for $\theta \in [0.4036, 0.4046]$ and $\mu = 200$. The three graphs in the bottom panel are, from top to bottom at $\theta = 0.404$, the reference time scale $\Delta\theta$; the norm $\|hJ(x)\|_2$; and the actual step size h *(dashed)* generated by `ode23sdc` running at TOL $= 10^{-9}$. As $h \approx 10^{-6}$, the interval comprises approximately 1,000 steps. The flat part of $\Delta\theta$ includes the interval where $|x_1| \leq 1$ and $\sigma_2[(J(x)]$ is positive. As $h \ll \Delta\theta$ except at the end of the interval, there is no stiffness during the transition. $\|hJ(x)\|_2$ measures the local Lipschitz constant scaled by the actual step size h. It is of order 10^{-2} except after the transition point, where the step size quickly ramps up, causing $\|hJ(x)\|_2$ and the stiffness factor $S(\theta, h)$ to grow larger than 1 at $\theta \approx 0.40453$, marking the onset of stiffness

solution, with a reference time scale $\Delta\theta = 1$. Hence *stiffness does not occur near* $x = 0$, but only, as already demonstrated, along parts of the limit cycle, when $|x_1| > 1$. This resolves the issue raised by Artemiev and Averina, [1, p. 6].

In order to demonstrate the stiffness indicator for a nonstiff equation, we solved the separable Lotka–Volterra equation

$$\begin{aligned} \dot{x}_1 &= x_1(a - bx_2) \\ \dot{x}_2 &= x_2(cx_1 - d), \end{aligned}$$

choosing the parameters $a = 3$, $b = 9$ and $c = d = 15$, taking $x(0) = (1 \quad 1)^{\mathrm{T}}$. For these data, the solution is periodic with $T \approx 1$, eliminating the need to normalize the interval. Thus we solved

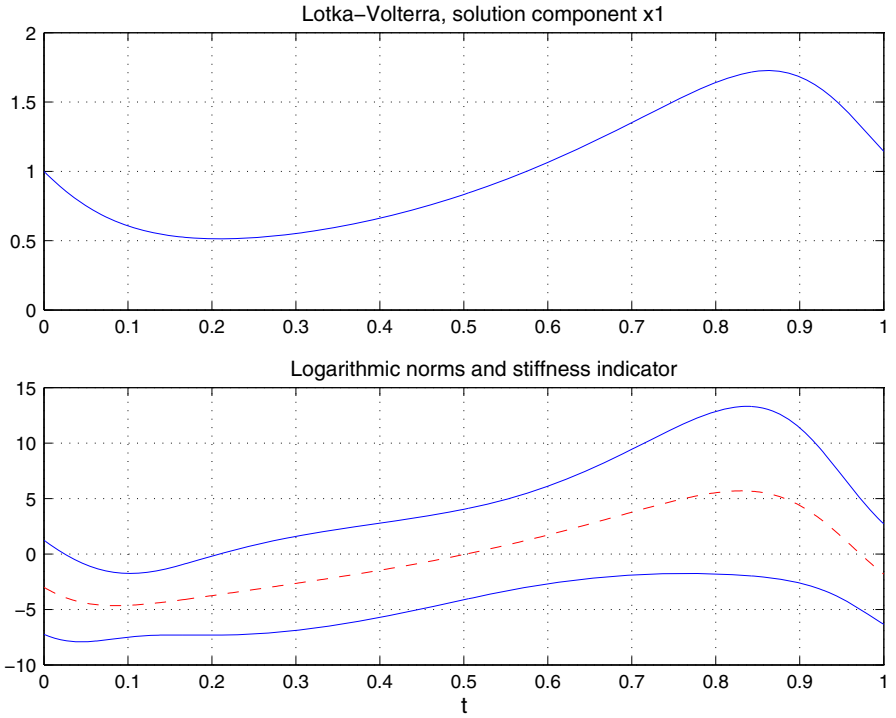

Fig. 5 *Lotka–Volterra equation.* Solution component x_1 (*top*) plotted vs time. The *bottom panel* shows the stiffness indicator $\sigma_2[J(x)]$ (*dashed*) as a function of time along the solution, as well as the upper and lower logarithmic norms (*both solid*). As expected, there is no indication of stiffness. Note that the system is *dissipative* ($M_2[J(x)] < 0$) approximately in $[0.02, 0.2]$ and *indefinite* otherwise

$$\dot{x}_1 = 3x_1 - 9x_1x_2$$
$$\dot{x}_2 = 15x_1x_2 - 15x_2$$

for $t \in [0, 1]$. Just like in the van der Pol equation, the stiffness indicator can be computed analytically, $\sigma_2[J(x)] = (15x_1 - 9x_2 - 12)/2$. It is plotted in Fig. 5. At $\min \sigma_2[J(x)] \approx -5$ stiffness does not occur, and the step sizes used by an adaptive explicit Runge–Kutta method are well within the reference time scale.

Perhaps less obviously, Fig. 5 also shows that the time integral of $\sigma_2[J(x)]$ over a full period in the Lotka–Volterra equation is zero. In fact, for every Lotka–Volterra equation with positive coefficients and positive initial values, it can be shown that $\log\sqrt{x_1x_2}$ is a primitive function of $\sigma_2[J(x)]$ along solutions. Hence, over a full period,

$$\log\sqrt{x_1(T)x_2(T)} - \log\sqrt{x_1(0)x_2(0)} = \int_0^T \sigma_2[J(x(\tau))]\,\mathrm{d}\tau = 0.$$

This corresponds to well known properties of the Poincaré map and the integral of the trace. In fact, in 2×2 systems $\dot{x} = f(x)$, the evolution of a phase volume

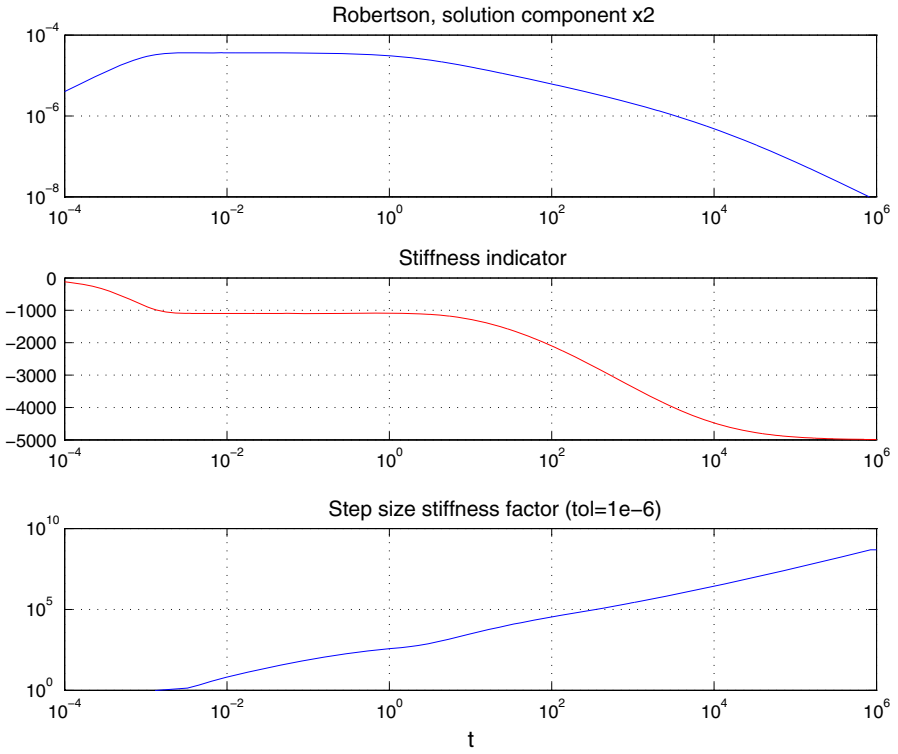

Fig. 6 *The Robertson equation.* Solution component x_2 (*top*; log-log scale) and non-normalized stiffness indicator $\sigma_2[J(x)]$ *(center)* for $t \in [0, 10^6]$. The problem is extremely stiff. With reference time scale $\Delta t \sim 2 \times 10^{-4}$ and integration range $T = 10^6$, one can estimate $G(0, T) \approx 5 \times 10^9$. When `ode23sdc` solves the problem at TOL $= 10^{-6}$, its step size stiffness factor (*bottom*) almost reaches 10^9 towards the end of the integration interval

element $V = \mathrm{d}x_1 \wedge \mathrm{d}x_2$ is given by $\dot{V} = \mathrm{div}_x(f) \cdot V = \mathrm{trace}[\mathrm{grad}_x(f)] \cdot V = (m_2[J(x)] + M_2[J(x)]) \cdot V$. Thus, although phase volume is not constant in the Lotka–Volterra equation, it is nevertheless conserved over a full period, in forward as well as reverse time.

Next we solved the non-periodic Robertson problem,

$$\begin{aligned}
\dot{x}_1 &= -k_1 x_1 + k_3 x_2 x_3 \\
\dot{x}_2 &= k_1 x_1 - k_2 x_2^2 - k_3 x_2 x_3 \\
\dot{x}_3 &= k_2 x_2^2
\end{aligned}$$

with $x(0) = (1 \quad 0 \quad 0)^{\mathrm{T}}$ and parameters $k_1 = 0.04$, $k_2 = 3 \times 10^7$ and $k_3 = 10^4$. The problem is solved over an extremely long time, $t \in [0, 10^6]$, and solutions are typically plotted vs a logarithmic time scale, see Fig. 6. For this reason, no time rescaling was used, and the non-normalized stiffness indicator $\sigma_2[J(x)]$ is plotted for $t \in [0, 10^6]$.

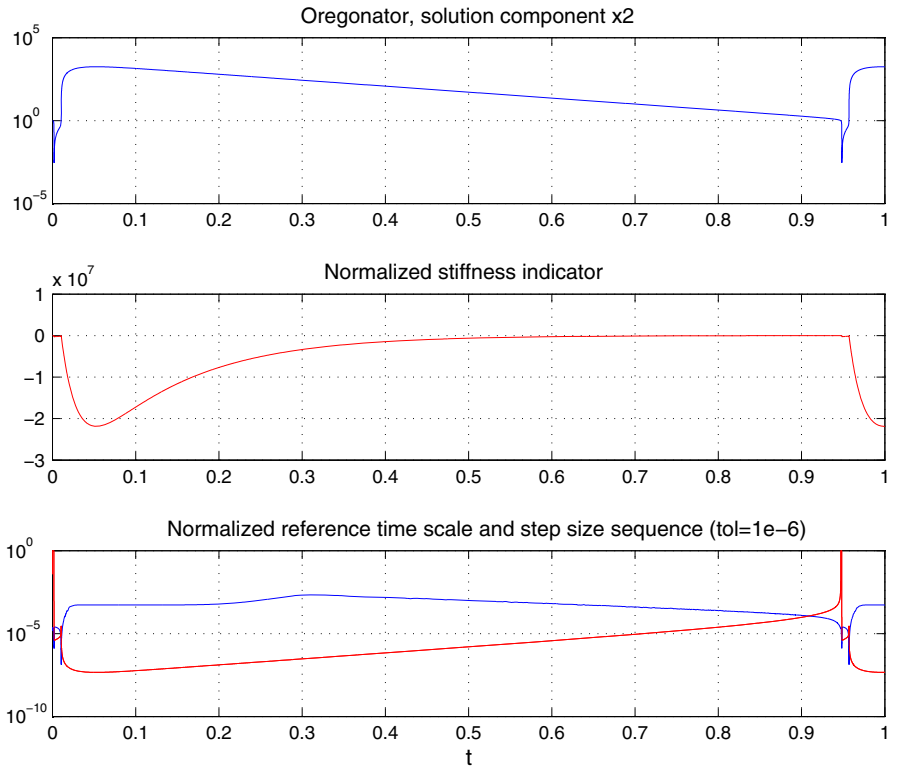

Fig. 7 *Oregonator equation.* Solution component x_2 (*top*; log scale) and normalized stiffness indicator $\sigma_2[J(x)]$ (*center*; linear scale). *Bottom panel* (log scale) shows normalized reference time scale $\Delta\theta$ (*lower curve*) as well as actual step size used in `ode23sdc` at TOL $= 10^{-6}$ (*upper curve*). For most of the integration, the step size exceeds $\Delta\theta$ by some four orders of magnitude, corresponding to a step size stiffness factor of 10^4. Close examination shows that the step size becomes smaller than $\Delta\theta$ at $\theta \approx 0.91$. Then $\sigma_2[J(x)]$ attains a maximum of 3.4×10^3 at $\theta \approx 0.9482$, when a complex double transition is triggered. Between the transitions, at $\theta \approx 0.95$, the stiffness indicator has dropped back to $\sigma_2[J(x)] \approx -2 \times 10^5$ and the step size again exceeds $\Delta\theta$. After the second transition, the indicator soon returns to its minimum, $\sigma_2[J(x)] \approx -2 \times 10^7$. Peace and quiet are restored, with a step size stiffness factor exceeding 10^4

We then solved the Oregonator equation

$$\begin{aligned}
\dot{x}_1 &= s(x_1 - x_1 x_2 + x_2 - q x_1^2) \\
\dot{x}_2 &= (x_3 - x_2 - x_1 x_2)/s \\
\dot{x}_3 &= w(x_1 - x_3)
\end{aligned}$$

with initial values $x(0) = (1 \quad 1 \quad 2)^{\mathrm{T}}$ and parameters $s = 77.27$, $q = 8.375 \times 10^{-6}$ and $w = 0.161$. The solution is periodic with $T \approx 300$; the problem is normalized using scaled time $\theta = t/320$. The initial values and the rescaling factor differ slightly from the standard choice in the Bari test set, see [14]. The normalized stiffness indicator is plotted in Fig. 7. Thus, at $\min \sigma_2[J(x)] \approx -2 \times 10^7$ the problem is very stiff, except at the transitions. The problem was solved with `ode15s` as well as `ode23sdc`.

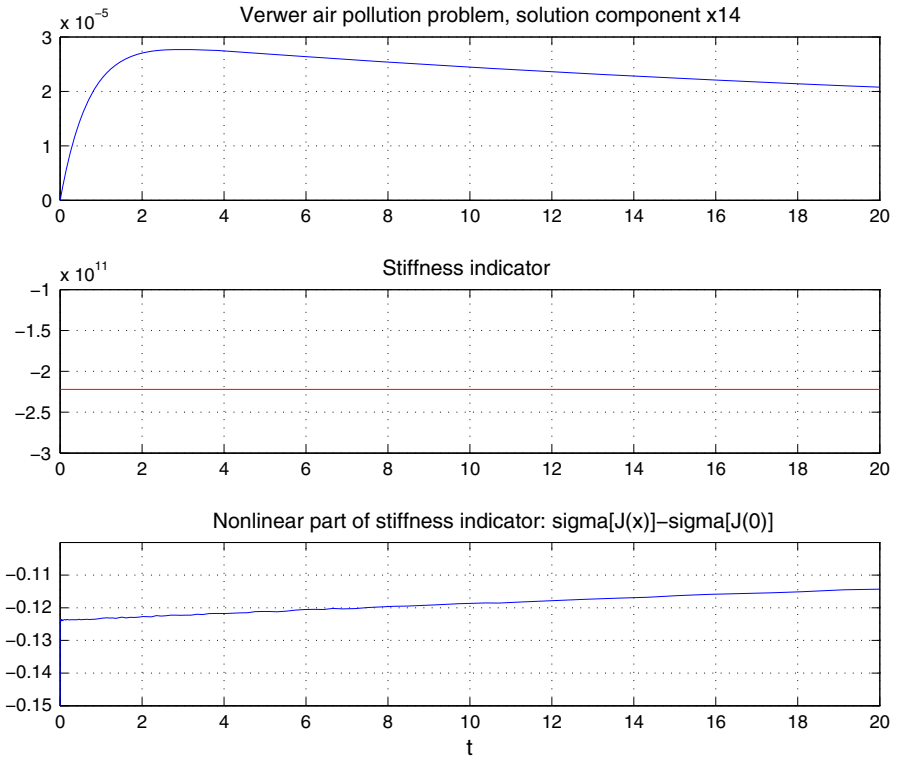

Fig. 8 *Verwer's air pollution model.* Solution component x_{14} (*top*) and non-normalized stiffness indicator $\sigma_2[J(x)]$ (*center*) for $t \in [0, 20]$. The problem is extremely stiff with a reference time scale of $\Delta t \approx 5 \times 10^{-12}$ and $G(0, T) \approx 4.4 \times 10^{12}$. The stiffness indicator is nearly constant, suggesting a linear constant coefficient behavior. As the graph of $\sigma_2[J(x)] - \sigma_2[J(0)]$ demonstrates (*bottom*), nonlinearity makes a negligible contribution to stiffness

Finally, we solved Jan Verwer's air pollution model from the Bari Test Set [14]. This is a large chemical reaction kinetics problem consisting of 20 equations with 25 widely varying reaction rate constants. The singular Jacobian can be written

$$J(x) = A \cdot K \cdot (R_0 + R_1 x),$$

where $A \in \mathbb{R}^{20\times 25}$ is an "incidence matrix" of the reactions and $K \in \mathbb{R}^{25\times 25}$ is a diagonal matrix of reaction rate constants. As the elemental reactions are of the form $r_k(x) = x_i$ or $r_k = x_i x_j$ (with $i \neq j$), it holds that

$$\text{grad}_x r(x) = R_0 + R_1 x,$$

where $R_0 \in \mathbb{R}^{25\times 20}$ is a constant matrix, and $R_1 \in \mathbb{R}^{25\times 20\times 20}$ is a 3-tensor, implying that $R_1 x \in \mathbb{R}^{25\times 20}$. Thus $J(x) = J(0) + (J(x) - J(0))$, where only $J(x) - J(0) = AKR_1 x$ contributes to nonlinearity. The Jacobian is dominated by $J(0) = AKR_0$, however, as $\|J(x) - J(0)\|_2 / \|J(x)\|_2 \approx 10^{-3}$ along the solution.

Although the problem is extremely stiff, it presents no real difficulties to a stiff solver. Thus, depending on the tolerance setting, MATLAB's `ode15s` solves the problem over the interval [0, 20] in a few hundred steps. In line with (4.7), the stiffness indicator is nearly constant, see Fig. 8, as the linear part $J(0)$ dominates the Jacobian $J(x)$. In fact, whereas

$$\sigma_2[J(x)] \approx -2.2 \times 10^{11}$$

and

$$\sigma_2[J(x)] - \sigma_2[J(0)] \approx -0.12,$$

the nonlinear part $J(x) - J(0)$ *makes a negligible contribution to stiffness*. Thus, although the stiffness indicator does vary, the problem largely behaves like a linear constant coefficient system, which explains the moderate computational effort needed by a stiff solver.

6 Conclusions

Although stiff differential equations is a mature area of research, the notion of stiffness has resisted rigorous characterization for 60 years. The object of this paper has been to present a critical review of the classical ideas, as well as to introduce a new characterization, splitting stiffness into two parts—a mathematical part which is exclusively problem dependent, and a numerical part that depends on the numerical method and operational criteria.

Based on sharp short-term bounds on perturbation growth and decay rates in the variational equation, the a posteriori stiffness indicator,

$$\sigma[J(t,x)] = \frac{m[J(t,x)] + M[J(t,x)]}{2}, \tag{6.1}$$

depends exclusively on the mathematical problem. It is readily computable along the solution $x(t)$, and relates the solution to a local *reference time scale* $\Delta t > 0$, essentially defined by $\Delta t = -1/\sigma[J(t,x)]$ whenever the stiffness indicator is negative. Irrespective of method order or accuracy requirements, no explicit time stepping method can use step sizes much larger than Δt without loss of stability. The reference time scale Δt therefore represents a barrier beyond which special methods may be necessary.

By relating Δt to the range of integration T, stiffness is mathematically characterized by a large *stiffness factor* $T/\Delta t$. As our test problems demonstrate, this factor may take extremely large values. The stiffness indicator further picks up qualitative changes, such as unstable transitions at turning points in nonlinear systems, encountered e.g. in the van der Pol and Oregonator equations.

Whether stiffness will be encountered in actual computations depends on operational criteria, such as the combination of method, algorithm, implementation and accuracy requirement, and whether they will suggest using a step size $h > \Delta t$ in a given problem. As computational experiments demonstrate, the step size stiffness

factor $h/\Delta t$ may become extremely large, but as $h/\Delta t < T/\Delta t$ this can only occur if the method independent characterization $T/\Delta t \gg 1$ holds. The latter is therefore a *necessary condition* for encountering stiffness.

The stiffness indicator (6.1) is expressed in terms of the lower and upper logarithmic norms of the Jacobian $J(t, x)$, where—at least in principle—any suitable norm could be used. For simplicity we have chosen to work with inner product norms in this paper. Specifically, all computations have been carried out in the Euclidean norm. Although simple to work with, this choice may not always be ideal, even though it has proven remarkably robust in the strongly nonlinear test problems studied here.

With the tools presented in this paper, it is no longer difficult to say what stiffness is. Even in nonlinear problems with complex behavior, stiffness can be adequately characterized and quantified. For example, by all objective standards, with $G(0, T) \approx 4.4 \times 10^{12}$ (cf. (4.11) and Fig. 8), Verwer's air pollution problem is stiff, plain and simple, although it is not particularly hard to solve. In the van der Pol equation, what is observed in practice is fully explained in theory: stiffness scales like $\mathrm{O}(\mu^2)$ and only occurs for $|x_1| > 1$, away from the transition points. In the Lotka–Volterra equation, stiffness is not present at all. Likewise, no separable Hamiltonian problem is ever stiff, for if $H(p, q) = T(p) + U(q)$, with

$$\begin{aligned} \dot{p} &= -H_q \\ \dot{q} &= H_p, \end{aligned}$$

the Hermitian part of the Jacobian becomes

$$\mathrm{He}(J(p, q)) = \frac{1}{2} \begin{pmatrix} 0 & T_{pp} - U_{qq} \\ T_{pp} - U_{qq} & 0 \end{pmatrix}.$$

The logarithmic norms of $J(p, q)$ therefore satisfy the quadratic eigenvalue problem

$$\frac{1}{4} (T_{pp} - U_{qq})^{\mathrm{T}} (T_{pp} - U_{qq}) q = \mu^2 q.$$

Hence the upper and lower logarithmic norms are $M_2[J(p, q)] = +\|T_{pp} - U_{qq}\|_2/2$ and $m_2[J(p, q)] = -\|T_{pp} - U_{qq}\|_2/2$, respectively. As their sum is zero, so is the stiffness indicator. Therefore, according to this theory, *every separable Hamiltonian problem is nonstiff*. This is an unconditional, structural property.

In line with the theoretical understanding of Hamiltonian systems, this result is not surprising, but nevertheless unexpected. Thus the comprehensive theoretical treatment of stiffness presented in this paper offers a qualitative as well as quantitative representation of stiffness, and is able to distinguish a variety of phenomena that have previously eluded a proper, common analysis.

There is one further substantive argument supporting the construction of the stiffness indicator. Thus, in the Lotka–Volterra equation, we noted that a phase volume element $V = \mathrm{d}x_1 \wedge \mathrm{d}x_2$ evolves according to $\dot{V} = \mathrm{div}_x(f) \cdot V = \mathrm{trace}[\mathrm{grad}_x(f)] \cdot V$. Now, in the 2×2 case, we have $\mathrm{trace}[\mathrm{grad}_x(f)] = m_2[J(x)] + M_2[J(x)] = 2\sigma_2[J(x)]$.

Consequently, a zero Euclidean stiffness indicator implies that *phase volume is conserved* (i.e., phase volume is "incompressible"), while a negative stiffness indicator shows that phase volume is compressed, even though (like in the van der Pol equation) the system might not be dissipative itself. More importantly, and independent of the dimension of the system, every separable Hamiltonian system conserves phase volume, and also has a zero stiffness indicator, as seen above. This implies that the suggested construction of the stiffness indicator is qualitatively compatible with the Liouville theorem, as well as the theory of separable Hamiltonian systems.

One could then ask whether it is possible to base a stiffness indicator on the evolution of the wedge product $V = \wedge_k \mathrm{d}x_k$. This would be in line with the well known result, for linear constant coefficient systems $\dot{x} = Ax$, that $\det[\mathrm{e}^{tA}] = \mathrm{e}^{t\,\mathrm{trace}[A]}$. However, because $\mathrm{div}_x(f) = \mathrm{trace}[J(x)] = \sum \lambda_k[J(x)]$, such an approach would make the indicator depend on the intermediate eigenvalues of $J(x)$. That would technically be wrong, as stiffness depends on the *extreme eigenvalues*. This is the reason for using the lower and upper logarithmic norms of $J(x)$ rather than its trace. Further, there are scaling reasons. For example, in a method of lines discretization, the trace will not capture stiffness, due to its dependence on the dimension of the spatial discretization.

Just like the trace and divergence, the stiffness indicator $\sigma_2[J(x)]$ is based on *a sum*, $(m_2[J(x)] + M_2[J(x)])/2$. This is the key feature that makes the proposed theory compatible with the theory of Hamiltonian systems, however unimportant or counterintuitive such a connection might at first seem to be.

As a last remark, we note that Curtiss and Hirschfelder's original discussion [5] was close to our theory. It gives a result identical to ours if $a(t, x) = a(t)$ in their model problem (2.1). Then $\sigma[J(t)] = 1/a(t)$, with reference time scale $\Delta t = -a(t)$, implying that Curtiss and Hirschfelder defined their problem as stiff if the stiffness factor $h/\Delta t \gg 1$, where h is "the desired resolution [of time] or the interval which will be used in the numerical integration".

Thus, after 60 years, we have come full circle.

Acknowledgments The substantial input from the reviewers is gratefully acknowledged.

References

1. Artemiev, S., Averina, T.: Numerical Analysis of Systems of Ordinary and Stochastic Differential Equations. VSP, Utrecht (1997)
2. Brugnano, L., Mazzia, F., Trigiante, D.: Fifty years of stiffness. In: Simos, T.E. (ed.) Recent Advances in Computational and Applied Mathematics, pp. 1–21. Springer, Berlin (2011)
3. Byrne, G.D., Hindmarsh, A.C.: Stiff ODE solvers: a review of current and coming attractions. J. Comp. Phys. **70**, 1–62 (1987)
4. Cash, J.R.: Efficient numerical methods for the solution of stiff initial-value problems and differential-algebraic equations. Proc. R. Soc. Lond. A **459**, 797–815 (2003)
5. Curtiss, C.F., Hirschfelder, J.O.: Integration of stiff equations. Proc. Natl. Acad. Sci. **38**, 235–243 (1952)
6. Dahlquist, G.: Stability and error bounds in the numerical integration of ordinary differential equations. Almqvist & Wiksells, Uppsala (1959)
7. Dahlquist, G.: A numerical method for some ordinary differential equations with large Lipschitz constants. In: Morrell, A.J.H. (ed.) Proceedings of IFIP Congress. Information Processing 68, Edinburgh, UK, vol. 1, Mathematics, Software, pp. 183–186 (1968)

8. Dekker, K., Verwer, J.G.: Stability of Runge-Kutta methods for stiff nonlinear differential equations. CWI Monographs, vol. 2. North-Holland, Amsterdam (1984)
9. Ekeland, K., Owren, B., Øines, E.: Stiffness detection and estimation of dominant spectrum with explicit Runge-Kutta methods. ACM Trans. Math. Softw. **24**, 368–382 (1998)
10. Gear, C.W.: Numerical initial value problems in ordinary differential equations. Prentice Hall, Englewood Cliffs (1971)
11. Hairer, E.; Wanner, G.: Solving ordinary differential equations II. Stiff and differential-algebraic problems, second revised edition. Comput. Math., vol. 14. Springer, Berlin (1996)
12. Higham, D.J., Trefethen, L.N.: Stiffness of ODEs. BIT **33**, 285–303 (1993)
13. Lambert, J.D.: Computational Methods in Ordinary Differential Equations. Wiley, London (1973)
14. Mazzia, F., Magherini, C.: Test Set for Initial Value Problem Solvers, Release 2.4 (2008). http://www.dm.uniba.it/testset/report/testset
15. Prothero, A., Robinson, A.: On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. Math. Comp. **28**, 145–162 (1974)
16. Shampine, L.: Evaluation of a test set for stiff ODE solvers. ACM Trans. Math. Softw. **7**, 409–420 (1981)
17. Shampine, L.: What is stiffness? In: Aiken, R.C. (ed.) Stiff Computation. Oxford University Press, New York (1985)
18. Spijker, M.N.: Stiffness in numerical initial-value problems. J. Comp. Appl. Math. **72**, 393–406 (1996)
19. Söderlind, G.: Digital filters in adaptive time-stepping. ACM-TOMS **29**, 1–26 (2003)
20. Söderlind, G.: The logarithmic norm. History and modern theory. BIT **46**, 631–652 (2006)