Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

URI ASCHER AND SEBASTIAN REICH

The midpoint scheme and variants for Hamiltonian systems: advantages and pitfalls

Preprint SC 97-51 (December 1997)

The midpoint scheme and variants for Hamiltonian systems: advantages and pitfalls

Uri M. Ascher* Sebastian Reich†

14th January 1998

Abstract

The (implicit) midpoint scheme, like higher order Gauss-collocation schemes, is algebraically stable and symplectic, and it preserves quadratic integral invariants. It may appear particularly suitable for the numerical solution of highly oscillatory Hamiltonian systems, such as those arising in molecular dynamics or structural mechanics, because there is no stability restriction when it is applied to a simple harmonic oscillator. Although it is well-known that the midpoint scheme may also exhibit instabilities in various stiff situations, one might still hope for good results when resonance-type instabilities are avoided.

In this paper we investigate the suitability of the midpoint scheme for highly oscillatory, frictionless mechanical systems, where the step-size k is much larger than the system's small parameter ε, in case that the solution remains bounded as $\varepsilon \to 0$. We show that in general one must require that k^2/ε be small enough, or else, even the errors in slowly varying quantities like the energy may grow undesirably (especially when fast and slow modes are tightly coupled) or, worse, the computation may yield misleading information. In some cases this may already happen when $k = O(\varepsilon)$. The same holds for higher order collocation at Gaussian points. The encountered restrictions on k are still better than the corresponding ones for explicit schemes.

1 Introduction

Classical discretization methods for differential equations typically find pointwise-accurate approximate solutions when the number of discretization steps taken is not too large and when the size of the discretization steps is smaller than any scale of the differential problem. In this work, however, we consider situations in which there may be a large number of non-small discretization steps taken. In such cases, dictated by necessity of computational feasibility in many applications (e.g. molecular dynamics, fluid flow), a lot more attention must be paid, not only to choosing appropriate discretization schemes but also to investigating the quality and meaning of the obtained results.

It is well-known that symplectic and time-reversible discretization schemes possess particularly attractive properties when applied over a long time[1] to Hamiltonian systems [22, 12]. Much attention has been paid recently in this context to the (implicit) midpoint scheme and some of its variants; see, e.g. [25, 10, 14] (see also [5] for a different orientation). The midpoint scheme, and more generally Gauss-collocation schemes, are algebraically stable, symplectic [22], and they preserve quadratic integral invariants [8]. They may appear particularly suitable for the numerical solution of highly oscillatory systems [11], although it is known that these schemes may also exhibit instabilities in

*Institute of Applied Mathematics and Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4 (ascher@cs.ubc.ca). The work of this author was partially supported under NSERC Canada Grant OGP0004306.

†Konrad-Zuse-Zentrum, Takustr. 7, D-14195 Berlin, Germany (reich@zib.de).

[1]More specifically, one must consider $Nk \gg 1$, where k is the discretization step-size and N is the number of discretization steps needed.

various stiff situations [10, 17, 1, 2]. However, whereas for the usual, highly damping stiff initial value problem better alternative schemes exist (e.g. [13]), no clear winners are known for the highly oscillatory case.

It may be argued that when computing approximate solutions to a complicated problem it is better to get a clearly unstable computation (where the approximate solution blows up or behaves otherwise unphysically) than to compute a wrong solution which looks physical – in the latter situation the user may receive no direct warning, and may be led to wrong conclusions regarding the object of the simulation. We show in this paper that, unfortunately, slowly varying quantities may be approximated by wrong, slowly varying quantities when a system with highly oscillatory solution components is approximated by a discretization scheme with a step-size k, unless k^2/ε is small enough, where ε is the fast scale of the differential system. In some instances, $k = O(\varepsilon)$ is required. Moreover, if k^2/ε is not small then significant error growth may be experienced when fast and slow modes are coupled by a non-constant transformation. The midpoint scheme is still among the more robust choices, but its imperfections, coupled with its implicitness, make the choice of "best" scheme and the interpretation of computational results more involved.

To be concrete, consider the initial value problem for the friction-free mechanical system ($0 \le t \le T$)

$$\dot{q} = p \tag{1.1a}$$
$$\dot{p} = -\mathrm{grad}V(q) - \varepsilon^{-2} G^T g(q) \tag{1.1b}$$

where q are the generalized positions of the bodies, p the generalized momenta, V and g are known potential functions of moderate size, $G(q) = \frac{\partial g}{\partial q}$, and $0 < \varepsilon \ll 1$ is the small parameter. It is well-known that the constancy in time of the Hamiltonian

$$H(q,p) = \frac{1}{2} p^T p + V(q) + \frac{1}{2\varepsilon^2} g(q)^T g(q) \tag{1.2}$$

is an integral invariant of this ODE system.

The presence of strong potentials introduces highly oscillatory solution modes. If the amplitude of such modes is large then to achieve a pointwise accurate approximate solution one must choose $k \le c_1 \varepsilon$, where c_1 is a constant such that each oscillation is reseolved by the mesh. It is better then to choose an explicit Verlet scheme (which is a staggered midpoint scheme, see §3) than one of the implicit variants of midpoint. But if the solution sought is bounded as $\varepsilon \to 0$ then we want to be able to choose $0 < \varepsilon \ll k \ll 1$. Now there is no hope to accurately recover rapidly varying solution quantities (i.e. quantities with $O(\varepsilon^{-1})$ derivatives), but one still hopes to accurately recover slowly varying quantities, such as the Hamiltonian, assuming that k is small enough with respect to this slow variation. The slow variation of the Hamiltonian often implies that fast modes are only $O(\varepsilon)$ in magnitude, and thus it may not be necessary to have $k \le c_1 \varepsilon$ for the sake of accuracy in the slowly varying quantities.

Let us rewrite (1.1) as an index-1 DAE,

$$\dot{q} = p \tag{1.3a}$$
$$\dot{p} = -\mathrm{grad}V(q) - G(q)^T \lambda \tag{1.3b}$$
$$\varepsilon^2 \lambda = g(q) \tag{1.3c}$$

We distinguish between two cases; which of these occurs depends on the initial conditions.

1. The exact solution satisfies[2]
$$|g(q)| = O(\varepsilon^2)\,.$$

[2]We denote, for any vector x, $|x| = \sqrt{x^T x}$.

In this case $|\lambda| = O(1)$ in (1.3). (This case is considered, e.g., in [11, 19].) In the passage to the limit, $\varepsilon \to 0$, we obtain the reduced system

$$\begin{aligned} \dot{q} &= p && \text{(1.4a)} \\ \dot{p} &= -\mathrm{grad}V(q) - G^T\lambda && \text{(1.4b)} \\ 0 &= g(q)\,. && \text{(1.4c)} \end{aligned}$$

2. The exact solution satisfies

$$|g(q)| = O(\varepsilon)\,.$$

The balance in powers of ε in (1.1) is achieved with $|p| = O(1)$, $|\dot{p}| = O(\varepsilon^{-1})$. This case is more prevalent in molecular dynamics calculations (although Case 1 occurs as well). Here we obtain $|\lambda| = O(\varepsilon^{-1})$ in (1.3c) and the passage to the limit is less clear. The limit DAE may turn out to be (1.4) (as in Examples 3.1 and 3.2) or it may be different, incorporating an $O(1)$ correcting potential term (as in Example 3.3).

Whether or not (1.4) defines the correct reduced solution, it turns out that $O(k^2/\varepsilon)$ errors in slowly varying quantities are unavoidable in general when the midpoint scheme is applied. There are two sources for this. One is already apparent when considering the error in the adiabatic invariant in a linear oscillator with a slowly varying frequency. (If the frequency is constant then the energy constancy is a quadratic invariant which is reproduced exactly by the midpoint scheme.) We analyze this error in §2.2 and explain its misleadingly regular form. The same type of error appears later in §3, in a more involved example.

The other source of error is more generic and may generate very large errors. When the solution variables in (1.1) contain coupled fast and slow components such that the decoupled form is simple enough that the midpoint scheme performs well when discretizing it directly, the *decoupling transformation* from the given system to its decoupled form may still yield $O(k^2/\varepsilon)$ errors in the discretization scheme when the non-decoupled form is discretized. These errors may induce instability. We prove in §2.3 a restricted stability condition for a linear example: starting from a pair of uncoupled fast and slow linear oscillators, for which midpoint stability can be proved, a smooth and well-conditioned coupling transformation yields a formulation whose direct midpoint discretization results in large errors which may grow rapidly in t unless the stability restriction is satisfied. The explanation must then be in the misrepresentation of the continuous transformation by the discrete one (cf. [2, 1, 5]).

All of the examples in §3 exhibit this phenomenon, and it is demonstrated and discussed further in that section. We show that unstable midpoint calculations which get worse as k is increased for a fixed ε can be easily encountered (i.e., this is not an esoteric phenomenon). The methodology for creating such examples is simple and it is demonstrated in Examples 3.1 and 3.2 which are related to the linear Example 2.2.

Apparent instabilities in midpoint calculations with large k for this class of problems have been observed in practice. Some explanation has been given in terms of resonance phenomena [17, 23]. Such explanations may suggest, however, that the step size k may be judiciously increased to avoid resonance spots – an unlikely cure. The difficulties we highlight here are present also for linear problems and for discretization schemes which attempt to avoid resonance instabilities.

The midpoint scheme preserves the energy only when the energy expression is at most quadratic (see [8, 3]). Straightforward projection schemes, which correct the midpoint scheme, say, at the end of each step to preserve H, typically lose symplecticity and even time-reversibility. In [6] we propose a method which preserves the energy and retains time-reversibility. We embed the Hamiltonian system in a projected DAE and propose a specially modified midpoint scheme for DAEs of pure index 2. When this scheme is applied to the projected DAE a time-reversible scheme is obtained.

Unfortunately, however, our scheme and other similar ones do not perform well for highly oscillatory systems, because the projected DAE is very difficult to solve with a large step size. Imposing energy conservation by the discrete scheme adversely affects its ability to handle highly oscillatory nonlinear problems when $k > \varepsilon$.

We summarize our experience as follows. Consider (1.1) again, with fast and slow modes clearly separable. In case that the correct smooth manifold is only $O(\varepsilon^2)$ – or even $O(\varepsilon)$ – away from the reduced solution manifold of (1.4), and if ε is really small, so that this difference is tolerable as an error, then solving (1.4), e.g. using SHAKE or RATTLE (see [15] and references therein) or one of the more general stabilization techniques [4], may be the most efficient way to proceed. On the other extreme, there are faster, explicit schemes such as (3.8) if we are willing to use $k \leq c_1\varepsilon$. Away from these possibilities, e.g. if the limit DAE is not known and choosing a large k is desirable for efficiency reasons, the midpoint scheme and other implicit variants may have something to offer. But caution must be exercised when they are used. In particular, as $\varepsilon/k \to 0$ *all* implicit Runge-Kutta schemes with a nonsingular coefficient matrix tend to approximate (1.4), so if the limit DAE is different from (1.4) (as in Example 3.3) then some misleading calculations may arise; moreover, energy-preserving variants [6, 20, 9] do not help in such circumstances, and caution dictates simply taking $k \leq c_2\varepsilon$ then, where typically it is still possible to choose the constant c_2 such that $c_2 > c_1$.

2 Highly oscillatory linear problems

Consider an ODE system

$$\dot{y} = f(y) \qquad 0 \leq t \leq T. \tag{2.1}$$

Using a discretization with a step-size k, the *midpoint scheme* reads

$$y_n - y_{n-1} = k\ f(\frac{y_n + y_{n-1}}{2}). \tag{2.2}$$

This is a simple, symmetric, symplectic 2nd order implicit Runge-Kutta scheme. It is an instance of collocation at Gaussian points, with one collocation point per step. Gauss-collocation schemes also preserve quadratic integral invariants and possess symplecticity and algebraic stability [13, 3]: writing such an s-stage scheme in the usual Runge-Kutta notation

$$\begin{array}{c|c} c & A \\ \hline & b \end{array}$$

we have

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \qquad i, j = 1, \ldots, s.$$

2.1 Linear oscillator

Consider the harmonic oscillator equations in the highly oscillatory case $0 < \varepsilon \ll 1$,

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -\varepsilon^{-2} q \end{aligned} \tag{2.3}$$

where q and p are scalar dependent variables. The exact solution for the initial conditions $p(0) = \beta, q(0) = 0$ is

$$q(t) = \varepsilon\beta \sin(\varepsilon^{-1}t), \quad p(t) = \beta\cos(\varepsilon^{-1}t)\,.$$

The Hamiltonian

$$H(q,p) = p^2/2 + \varepsilon^{-2}q^2/2$$

is conserved along solutions. Applying now a step of an s-stage Runge-Kutta scheme, denote by (s-vectors) $\hat{q}$ and $\hat{p}$ the corresponding approximations obtained for the solution at the stages between t_{n-1} and $t_n = t_{n-1} + k$. We easily obtain

$$\begin{pmatrix} I & -kA \\ k\varepsilon^{-1}A & \varepsilon I \end{pmatrix} \begin{pmatrix} \hat{q} \\ \hat{p} \end{pmatrix} = O(\varepsilon)\,.$$

Assuming that the coefficients matrix A is nonsingular (which holds, e.g., for all Gauss and Radau collocation as well as for all DIRK schemes [13]) we immediately obtain for the intermediate stages that

$$\hat{q} = O(\varepsilon^2/k), \quad \hat{p} = O(\varepsilon/k).$$

Comparing this with the exact solutions, we conclude that the approximation obtained for the intermediate stages may be good for $\beta = O(\varepsilon/k)$ but it is $O(1)$ away in $\hat{p}$ for $\beta = 1$, say, when $\varepsilon/k \ll 1$. The error at p_n is then $O(1)$ as well. Note that the error in the variable q can be neglected in appropriate circumstances, due to the small amplitude of the oscillations in $q(t)$, even though the relative error is $O(1)$. The mesh values p_n may oscillate or not, but in any case they will be wrong in absolute value, as $\varepsilon \to 0$. For Radau collocation the mesh points are also internal stages, so $q_n = O(\varepsilon^2/k)$, $p_n = O(\varepsilon/k)$. Then the solution corresponds to the exact solution subject to entirely different initial conditions, which may be particularly misleading.

Now, symmetric Runge-Kutta schemes are P-stable [19]. Moreover, for the midpoint scheme the Hamiltonian H is conserved exactly, because it happens to be a quadratic invariant. Hence we obtain a very stable scheme for any step-size k which may, however, produce wrong, physically-looking solutions.

Note the marked difference between the harmonic oscillator equations and the classical "test equation"

$$\dot{y} = -\varepsilon^{-1} y\,.$$

For the latter, a simple scheme like backward Euler (or more generally, Radau collocation) produces the exact solution as $\varepsilon/k \to 0$, regardless of the initial condition.

2.2 Linear oscillator with slowly varying frequency

Let us now modify the oscillator (2.3) by introducing a time-dependent, slowly-varying change in frequency; namely

$$\begin{aligned} \dot{q} &= \omega^2(t) p \\ \dot{p} &= -\varepsilon^{-2} q \end{aligned} \tag{2.4}$$

where $\omega(t) > 0$, $0 \le t \le T$. For example, take

$$\omega(t) = 1 + t\,.$$

The modified system is still Hamiltonian with energy

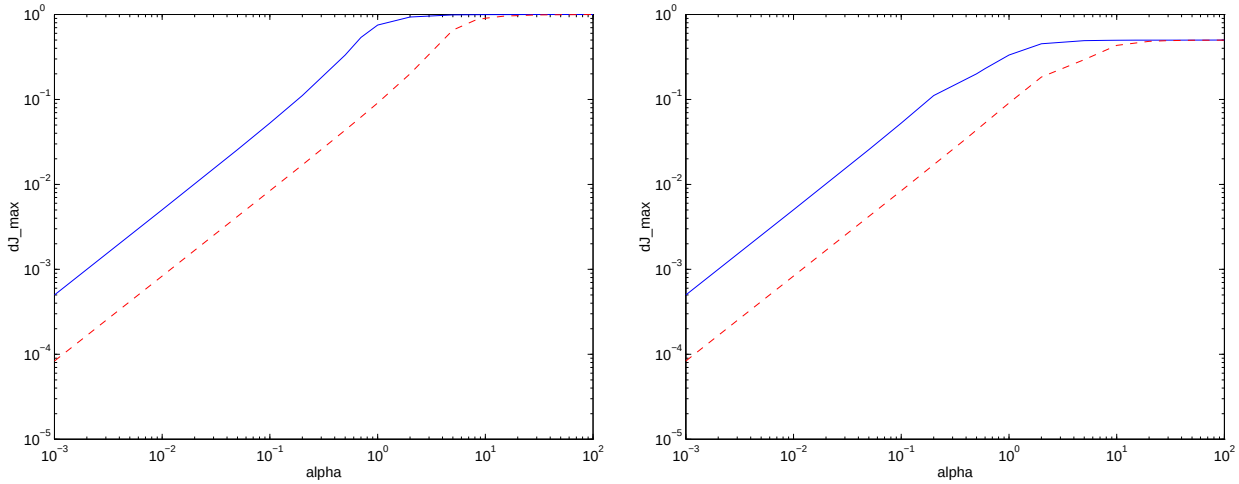
$$H(q,p,t) = \omega^2(t) p^2/2 + \varepsilon^{-2} q^2/2\,.$$

Since the energy is now time-dependent, it is no longer a first integral. However, there exists an *adiabatic invariant*

$$J(q,p,t) = H(q,p,t)/\omega(t) \;=\; \omega(t) p^2/2 + \varepsilon^{-2}\omega^{-1}(t) q^2/2$$

(see, e.g., [16]) such that, for ε small enough,

$$\left[J(t) - J(0)\right]/J(0) = O(\varepsilon)$$



(a) Initial conditions $q(0) = 0, p(0) = 1$

(b) Initial conditions $q(0) = 1, p(0) = 0$

Figure 2.1: Error in adiabatic invariant vs. $\alpha = \frac{k^2}{4\varepsilon}$: midpoint in solid line, 3-point Gauss collocation in dashed line

over a time interval of length

$$T = c_1 \, e^{c_2/\varepsilon} \, ,$$

$1 > c_1, c_2 > 0$ appropriate constants depending on $\omega(t)$ [24, 18]. In particular, starting from $p(0) = \beta$, $q(0) = 0$, J is bounded in terms of β and varies smoothly for $0 \leq t \leq 1$ and ε small enough. Note that J is proportional to the ratio of total energy and frequency of the oscillator.

Let us now numerically integrate the equations by the implicit midpoint scheme with a step-size $k \ll 1$ and possibly $\varepsilon/k \ll 1$. Since energy is no longer a first integral, nothing immediate can be said a priori about the numerically computed solutions (that they will be wrong in pointwise absolute value is clear, but we hope that less is needed for the accurate computation of slowly varying quantities). Still, one might expect that the adiabatic invariant J will be approximately conserved by the midpoint scheme.

Example 2.1 *We have calculated the solutions for the midpoint approximation of (2.4) for various values of ε and k, measuring*

$$\Delta J = \max_{0 \leq t \leq 1} |J(t) - J(0)|/J(0)$$

starting from the initial values yielding (a) $J(0) = H(0) = 0.5$, and (b) $J(0) = H(0) = 0.5\varepsilon^{-2}$. We have found that the error in the adiabatic invariant depends (for ε and k small enough) only on the ratio of k to ε/k, in the manner that Figure 2.1 indicates. In particular, the error in the (slowly varying) adiabatic invariant remains large if $k^2 \gg \varepsilon$, even as $k \to 0$. A similar phenomenon is observed for higher order Gauss collocation schemes. To illustrate the different behavior of the energy $H(t)$ and the adiabatic invariant $J(t)$, we plot both for $k = 0.01$ and $k^2/\varepsilon = 0.1$ in Figure 2.2. □

Let us now give an explanation for the phenomenon observed in Example 2.1. Define

$$\alpha = \frac{k^2}{4\varepsilon}$$

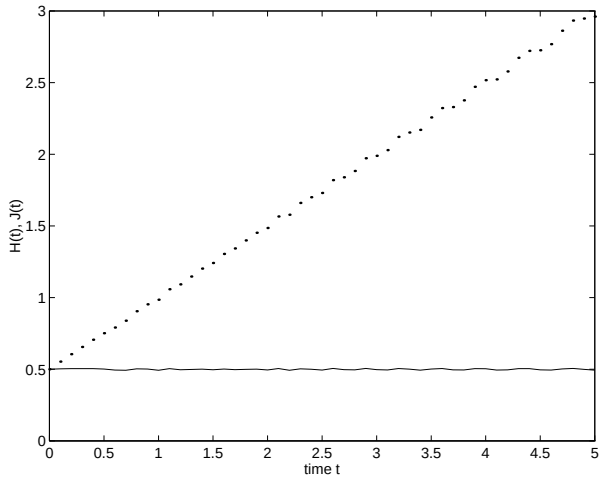


Figure 2.2: Adiabatic invariant $J(t)$ (solid line) and energy $H(t)$ (dotted line) vs. time t .

and consider the midpoint equations

$$\begin{aligned}(q_n - q_{n-1})/k &= \omega(t_{n-1/2})^2(p_n + p_{n-1})/2 \\ (p_n - p_{n-1})/k &= -\varepsilon^{-2}(q_n + q_{n-1})/2\,.\end{aligned}$$

Let $u_n := (-1)^n\varepsilon^{-1}q_n$, $v_n := (-1)^{n+1}p_n$ (cf. [2] and references therein). Note that the Hamiltonian H, and therefore also the adiabatic invariant J, satisfy

$$H(q_n, p_n, t_n) = H(\varepsilon u_n, v_n, t_n), \quad J(q_n, p_n, t_n) = J(\varepsilon u_n, v_n, t_n)\,.$$

For the new variables we get

$$\begin{aligned}(u_n + u_{n-1})/k &= -\varepsilon^{-1}\omega(t_{n-1/2})^2(v_n - v_{n-1})/2 \\ (v_n + v_{n-1})/k &= \varepsilon^{-1}(u_n - u_{n-1})/2\,.\end{aligned}$$

Multiplying both equations by $k/2$ and rearranging we get

$$\begin{aligned}(u_n + u_{n-1})/2 &= -\omega(t_{n-1/2})^2\alpha(v_n - v_{n-1})/k \\ (v_n + v_{n-1})/2 &= \alpha(u_n - u_{n-1})/k\,.\end{aligned}$$

Now observe what approximation we get as $k \to 0$ for a fixed α. In fact, this is again the midpoint scheme(!) applied to the "ghost ODE"

$$\begin{aligned}-\omega^2(t)\alpha\dot{v} &= u \\ \alpha\dot{u} &= v\end{aligned}$$

or,

$$\ddot{u} = -(\omega(t)\alpha)^{-2}u\,. \tag{2.5}$$

The features of u and v therefore depend in the limit only on α (and of course ω). Whenever $\alpha \ll 1$, the modified system (2.5) describes again a highly oscillatory system with slowly varying frequency. Its frequency is

$$\hat{\omega}(t) = (\omega(t)\alpha)^{-1}$$

and its Hamiltonian is

$$\hat{H}(u,v,t) = (\omega^2(t)\alpha)^{-1}u^2/2 + \alpha^{-1}v^2/2\,.$$

Thus the corresponding adiabatic invariant is

$$\hat{J}(u,v,t) = \omega^{-1}(t)u^2/2 + \omega(t)v^2/2$$

and so, similarly to the original problem formulation, we obtain

$$\left[\hat{J}(t) - \hat{J}(0)\right]/\hat{J}(0) = O(\alpha)$$

over a time-interval

$$\hat{T} = \hat{c}_1 e^{\hat{c}_2/\alpha}\,,$$

$1 > \hat{c}_1, \hat{c}_2 > 0$ appropriate constants, provided that α is not too large. When $\alpha \gg 1$ the ghost ODE clearly indicates that $|\dot{u}| = O(\alpha^{-2})$ and $|\dot{v}| = O(\alpha^{-1})$, hence $u(t) \approx u(0)$, $v(t) \approx v(0)$. So, for $\alpha \gg 1$ (and $\varepsilon \ll 1$), $\Delta\hat{J}$ reaches an asymptote depending only on the initial values (as well as ω and T, but no longer on α).

Finally note that

$$\hat{J}(u_n, v_n, t_n) = \hat{J}(\varepsilon^{-1}q_n, p_n, t_n) = J(q_n, p_n, t_n)\,.$$

This completely explains the results recorded in Fig. 2.1.

2.3 Coupling effects

Let us generate a simple linear example where a midpoint discretization with a large step size results in an unstable numerical process. We give a precise condition for this instability to occur.

Example 2.2 *We start with an uncoupled pair of fast and slow linear oscillators,*

$$\dot{u} = v \tag{2.6a}$$

$$\dot{v} = -\begin{pmatrix}\varepsilon^{-2} & 0\\ 0 & 1\end{pmatrix}(u - \bar{u}) \tag{2.6b}$$

where $u(t), v(t)$ and the given constant vector $\bar{u}$ each have two components. The fast linear oscillator is for $u_1,\ v_1$, and it is separated from the slow oscillator (for $u_2,\ v_2$), even though we write them as a system. The midpoint scheme is therefore stable for any step size $k > 0$ and it reproduces the (quadratic) energies for the fast and the slow oscillators.

Next, we apply the following standard, time-dependent linear transformation:

$$u = Qx,\ v = Qy, Q(t) = \begin{pmatrix}\cos t & \sin t\\ -\sin t & \cos t\end{pmatrix},\ K = \dot{Q}^T Q = \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}. \tag{2.7}$$

This yields the coupled system

$$\dot{x} = y + Kx \tag{2.8a}$$

$$\dot{y} = -Q^T\begin{pmatrix}\varepsilon^{-2} & 0\\ 0 & 1\end{pmatrix}Q(x - \bar{x}) + Ky\,, \tag{2.8b}$$

where $\bar{x} = Q^T\bar{u}$.

Now consider the midpoint scheme applied to the coupled linear system,

$$k^{-1}(x_n - x_{n-1}) = y_{n-1/2} + Kx_{n-1/2} \tag{2.9a}$$

$$k^{-1}(y_n - y_{n-1}) = -Q_{n-1/2}^T\begin{pmatrix}\varepsilon^{-2} & 0\\ 0 & 1\end{pmatrix}Q_{n-1/2}(x_{n-1/2} - \bar{x}_{n-1/2}) + Ky_{n-1/2}, \tag{2.9b}$$

where we denote, e.g., $x_{n-1/2} = \frac{x_n + x_{n-1}}{2}$*, and similarly for any other mesh quantity, and* $Q_{n-1/2} = Q(t_{n-1/2})$.

Numerical experiments clearly indicate that the approximate solution remains bounded for $\alpha < 1$*, but increases rapidly in* t *for* $\alpha > 1$ *(to recall,* $\alpha = \frac{k^2}{4\varepsilon}$*). Choosing values for* $\bar{u}$ *and the initial data as in Example 3.1 the results are similar to those in Table 3.5 and Figure 3.2 and are therefore omitted here.* □

We now prove that the midpoint scheme indeed becomes unstable for $\alpha > 1$, if α is bounded away from 1 (say $\alpha \geq 1 + \delta$ for some small positive δ), for the discretization (2.9). Denote $u_n = Q_n x_n$, $v_n = Q_n y_n$, $u_{n-1/2} = \frac{u_n + u_{n-1}}{2}$, etc. The matrix function Q is orthogonal at each t, all of its even derivatives are scalar multiples of one another, and all of its odd derivatives are scalar multiples of one another. Taylor expansions of Q_n about $Q_{n-1/2}$ or of $Q_{n-1/2}$ about Q_n thus become particularly tractable. We have, e.g.,

$$\begin{aligned} Q_{n-1/2}(x_n - x_{n-1})/k &= k^{-1}(u_n - u_{n-1})(1 - k^2/8 + O(k^4)) + K u_{n-1/2}(1 + O(k^2)) \\ Q_{n-1/2} x_{n-1/2} &= u_{n-1/2}(1 - k^2/8 + O(k^4)) + \frac{k}{4} K (u_n - u_{n-1})(1 + O(k^2)) \,. \end{aligned}$$

Note also that $Q_{n-1/2}$ commutes with K. Multiplying the discretized equations for x and for y each by $Q_{n-1/2}$ we obtain

$$\begin{aligned} k^{-1}(u_n - u_{n-1}) &= v_{n-1/2} + \frac{k}{4} K (v_n - v_{n-1}) + \dots \\ k^{-1}(v_n - v_{n-1}) &= -\begin{pmatrix} \varepsilon^{-2} & 0 \\ 0 & 1 \end{pmatrix} (u_{n-1/2} - \bar{u} + \frac{k}{4} K (u_n - u_{n-1}) - \frac{k^2}{8} u_{n-1/2}) + \dots \end{aligned}$$

where the terms in $(\dots)$ are majorized by the ones listed.

The transformed discretized equations above consist of a midpoint discretization of the uncoupled system (2.6) plus additional terms. Because the midpoint discretization of (2.6) is stable and well-behaved, the crucial question is the effect of these additional terms, in particular in the equation involving ε^{-2}. There we have

$$\varepsilon k^{-1}(v_{1,n} - v_{1,n-1}) = -\varepsilon^{-1}(u_{1,n-1/2} - \bar{u}_1) + \alpha k^{-1}(u_{2,n} - u_{2,n-1}) + \alpha u_{1,n-1/2}/2 + \dots$$

Similarly, for u_2 we have

$$k^{-1}(u_{2,n} - u_{2,n-1}) = v_{2,n-1/2} + \alpha\varepsilon k^{-1}(v_{1,n} - v_{1,n-1}) + \dots$$

and substituting this into the previous expression we get

$$\varepsilon k^{-1}(v_{1,n} - v_{1,n-1}) = (1 - \alpha^2)^{-1}[-\varepsilon^{-1}(u_{1,n-1/2} - \bar{u}_1) + \alpha v_{2,n-1/2} + \alpha u_{1,n-1/2}/2 + \dots]$$

Consider values of α which are bounded away from $\alpha = 1$. (If $\alpha = 1$, additional terms must be considered.) We see that the method becomes unstable when $\alpha > 1$, because then the leading-order eigenvalue pair corresponding to the fast oscillator in (2.6) switches from imaginary to real, with both signs present. Experiments for ε small confirm that indeed the midpoint scheme is found unstable for all values of $\alpha > 1$ tried for this simple problem.

Note that if we generalize the coupling transformation (2.7) to

$$Q(t) = \begin{pmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{pmatrix}$$

for any constant $0 \leq \omega \ll k^{-1}$, the stability condition can be easily proved to be

$$\alpha < \omega^{-1}, \text{ or } k < 2\sqrt{\frac{\varepsilon}{\omega}}.$$

So, as ω is decreased from 1 to 0 the stability restriction gradually eases off. On the other hand, as ω is increased smaller step sizes are required for stability. Note also that the eigenvalues of the linear problem (2.8) are purely imaginary for all ω (!) Attempts to use these eigenvalues as a telltale sign for instability in the discretized problem (see, e.g., [26]) totally fail here.

3 Examples of coupled harmonic oscillators

In this section we examine in detail the performance of the midpoint scheme (and occasionally also of higher order Gaussian collocation schemes) on three simple nonlinear examples. All three examples involve difficulties which can be associated with the coupling of rapidly varying and slowly varying oscillators. Example 3.3 involves, in addition, also difficulties associated with those analyzed in Example 2.1.

Example 3.1 *We consider a simple stiff spring pendulum, first in polar coordinates. The equations of motion are*

$$\begin{aligned} \dot{r} &= p_r && \text{(3.1a)} \\ \dot{p_r} &= -\varepsilon^{-2}(r - r_0) + p_\phi^2 r^{-3} && \text{(3.1b)} \\ \dot{\phi} &= r^{-2} p_\phi && \text{(3.1c)} \\ \dot{p_\phi} &= -(\phi - \phi_0) && \text{(3.1d)} \end{aligned}$$

where r is the radius and ϕ is the angle of the pendulum. This is a Hamiltonian system with the Hamiltonian

$$H = \frac{1}{2}[p_r^2 + r^{-2}p_\phi^2 + (\phi - \phi_0)^2 + \varepsilon^{-2}(r - r_0)^2]\,. \tag{3.2}$$

For $\varepsilon \ll 1$, the system is clearly highly oscillatory in the variables (r, p_r). In our numerical experiments we used $\phi_0 = \pi/4$, $r_0 = 1$, and initial data $r(0) = r_0$, $\phi(0) = \phi_0$, $p_r(0) = \frac{1}{\sqrt{2}}$, $p_\phi(0) = -\frac{1}{\sqrt{2}}$, which yield $|r - r_0| = O(\varepsilon)$ for the exact solution.

In the limit $\varepsilon \to 0$, the reduced equations of motion for the slowly varying variables (ϕ, p_ϕ) are simply given by

$$\begin{aligned} \dot{\phi} &= r_0^{-2} p_\phi \\ \dot{p_\phi} &= -(\phi - \phi_0)\,. \end{aligned}$$

The system (3.1) can therefore be viewed as consisting of two loosely coupled linear oscillators (the coupling is nonlinear, though), one fast and one slow. It is not difficult to extend the results stated in §2.1 to this case. Thus we expect no stability difficulties and an accurate recovery of H by the midpoint scheme, provided that the step size k is small enough for the slow variables, even if $k^2/\varepsilon \gg 1$.

We now consider discretization of (3.1) by the implicit midpoint scheme. The results for various values of the step-size k, the parameter ε, and $\alpha := \frac{k^2}{4\varepsilon}$, can be found in Table 3.1. Here ΔE_F and ΔE_S are defined by

$$\Delta E_F = \max_{t\in[0,5]} |E_F(0) - E_F(t)|$$

and

$$\Delta E_S = \max_{t\in[0,5]} |E_S(0) - E_S(t)|$$

where

$$E_F(t) = \frac{1}{2}[p_r^2 + \varepsilon^{-2}(r - r_0)^2]$$

scheme s	k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	ΔE
1	0.1e-1	0.1e-2	0.25e-1	0.35e-3	0.35e-3	0.19e-6
3	0.1e-1	0.1e-2	0.25e-1	0.35e-3	0.35e-3	0.14e-6
4	0.1e-1	0.1e-2	0.25e-1	0.35e-3	0.35e-3	0.10e-6
1	0.1	0.1e-2	0.25e+1	0.34e-3	0.34e-3	0.10e-5
3	0.1	0.1e-2	0.25e+1	0.35e-3	0.35e-3	0.21e-6
4	0.1	0.1e-2	0.25e+1	0.31e-3	0.31e-3	0.20e-6
1	0.1	0.1e-5	0.25e+4	0.46e-9	0.46e-9	0.29e-11
3	0.1	0.1e-5	0.25e+4	0.27e-8	0.27e-8	0.30e-10
4	0.1	0.1e-5	0.25e+4	0.46e-8	0.46e-8	0.55e-10

Table 3.1: Example 3.1 – maximum error in the energy using collocation at s Gaussian points for the (almost) decoupled system (3.1): fast components, slow components and total energy.

is the energy in the fast variables (r, p_r) *and*

$$E_S(t) = \frac{1}{2}[r^{-2}p_\phi^2 + (\phi - \phi_0)^2]$$

is the energy in the slow variables (ϕ, p_ϕ)*. Note that* $\Delta E_S = O(\varepsilon)$ *and* $\Delta E_F = O(\varepsilon)$ *for the analytical solution. We also calculate*

$$\Delta E = \max_{t\in[0,5]} |H(0) - H(t)|$$

(which equals 0 *for the analytical solution) for the Hamiltonian* H *defined in (3.2). We list results for collocation schemes at* s *Gaussian points. The midpoint scheme corresponds to the case* $s = 1$*. The results in Table 3.1 are excellent, and no dependence on* α *is noticed. Note that the accuracy in* ΔE *improves here as* ε *is decreased with* k *held fixed and larger. This is because* H *in (3.2) is "closer" to being quadratic.*

Next, we transform the problem to Cartesian coordinates q *and corresponding momenta* $p = \dot{q}$*, where* $r(q) = |q| = \sqrt{q_1^2 + q_2^2}$ *and* $\phi(q) = arccos\,(q_1/|q|)$*. The transformed Hamiltonian is*

$$H(q,p) = \frac{1}{2}[p^T p + (\phi(q) - \phi_0)^2 + \varepsilon^{-2}(r(q) - r_0)^2]$$

and the corresponding equations of motion are

$$\dot{q} = p \tag{3.3a}$$
$$\dot{p} = -(\phi(q) - \phi_0)\nabla\phi(q) - \varepsilon^{-2}(r(q) - r_0)\nabla r(q)\,. \tag{3.3b}$$

Note that the conjugate momenta satisfy $p_r = \nabla r(q)^T p$ *and* $p_\phi = r^2\nabla\phi(q)^T p$*. It can be directly verified that*

$$\nabla r(q) = \frac{1}{r}q, \qquad \nabla\phi(q) = \frac{1}{r^2}(-q_2, q_1)^T = \frac{1}{r^2}Kq.$$

Hence

$$p = G^T p_r + B^T p_\phi = p_r\nabla r + p_\phi\nabla\phi$$

(Note that $BG^T = 0$*.) The corresponding initial data used in our numerical experiments are* $q(0) = \frac{1}{\sqrt{2}}(1,1)^T$*,* $p(0) = (1,0)^T$*.*

In the limit $\varepsilon \to 0$*, the reduced equations of motion in cartesian coordinates are given by the constrained system*

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -(\phi(q) - \phi_0)\nabla\phi(q) - \nabla r(q)\lambda \\ 0 &= r(q) - r_0\,. \end{aligned}$$

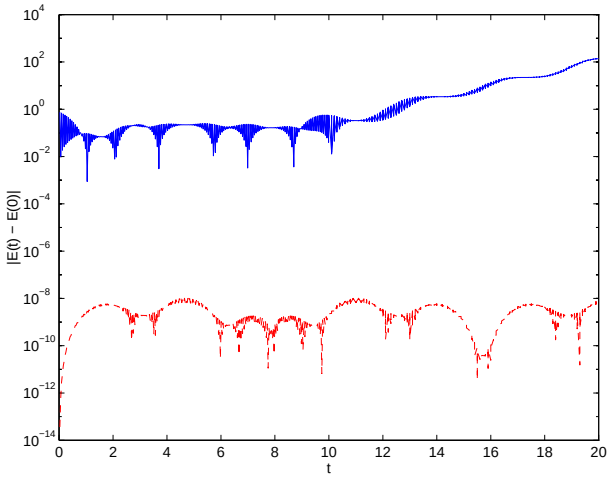


Figure 3.1: Errors in total energy for $\varepsilon = 0.1e-3$, $\alpha = 2.5$ and $T = 20$, using midpoint for Example 3.1 in Cartesian and polar coordinates. The error in Cartesian coordinates eventually grows unacceptably. .

step-size k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	$\varepsilon^{-1}\Delta E_S$	ΔE	$\alpha^{-1}\Delta E$
0.1e-1	0.1e-2	0.25e-1	0.48e-2	0.35e-3	0.35	0.45e-2	0.18
0.1e-1	0.2e-3	0.125	0.22e-1	0.71e-4	0.36	0.22e-1	0.18
0.1e-1	0.1e-3	0.25	0.42e-1	0.34e-4	0.34	0.42e-1	0.17
$\sqrt{10}$e-2	0.1e-2	0.25	0.42e-1	0.34e-3	0.34	0.42e-1	0.17
$\sqrt{5}$e-2	0.1e-2	0.125	0.22e-1	0.35e-3	0.35	0.22e-1	0.18
0.1e-1	0.1e-2	0.25e-1	0.48e-2	0.35e-3	0.35	0.45e-2	0.18
0.5e-2	0.25e-3	0.25e-1	0.45e-2	0.89e-4	0.35	0.44e-2	0.18
0.1e-1	0.1e-2	0.25e-1	0.48e-2	0.35e-3	0.35	0.45e-2	0.18
0.2e-1	0.4e-2	0.25e-1	0.59e-2	0.14e-2	0.36	0.45e-2	0.18
0.1e-1	1	0.25e-4	0.15	0.15	0.15	0.62e-5	0.25
0.1e-1	0.1e-1	0.25e-2	0.40e-2	0.35e-2	0.35	0.45e-3	0.18
0.1e-1	0.1e-5	0.25+2	0.13e+3	0.19e-4	0.19e+2	0.13e+3	0.50e+1

Table 3.2: Example 3.1 in Cartesian coordinates – maximum error in the energy using midpoint: fast components, slow components and total energy.

We now consider the discretization of (3.3) by the implicit midpoint rule. The results for various values of the step-size k, the parameter ε, and $\alpha := \frac{k^2}{4\varepsilon}$, can be found in Table 3.2. These results suggest that, provided that $\varepsilon \ll \alpha \ll 1$, $\Delta E = O(\alpha)$ while $\Delta E_S = O(\varepsilon)$. This implies also $\Delta E_F = \Delta E - \Delta E_S = O(\alpha)$.

For $\alpha = 25$ (and interval length $T = 5$) an instability is suggested. The error grows rapidly as T is increased. The instability does not occur only for isolated values of α, unlike a resonance-type phenomenon. The error in the total energy for $\alpha = 2.5$ and $T = 20$ is plotted in Figure 3.1.

In Tables 3.3 and 3.4 we list results for similar experiments where instead of using the midpoint scheme we have used collocation at 3 and 4 Gaussian points, respectively. These results show a significant improvement in accuracy for both schemes, provided that $\varepsilon \ll 1$ and $\alpha \ll 1$. Note that $\Delta E_S = O(\varepsilon)$. When α is large, these schemes too do not perform well and instabilities start to develop.

Let us consider the two possibilities for arriving at an approximate solution for the problem in Cartesian coordinates, starting from the polar coordinate formulation (3.1). One possibility is to discretize (3.1) and then transform the obtained values pointwise from the polar variables to

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	ΔE
0.1e-1	0.1e-2	0.25e-1	0.36e-3	0.35e-3	0.10e-4
0.1e-1	0.2e-3	0.125	0.14e-3	0.70e-4	0.67e-4
0.1e-1	0.1e-3	0.25	0.29e-3	0.35e-4	0.26e-3
0.1e-1	0.25e-4	1.0	0.41e-2	0.83e-5	0.41e-2
$2\sqrt{10}$e-2	0.1e-2	1.0	0.45e-2	0.33e-3	0.42e-2
$\sqrt{10}$e-2	0.1e-2	0.25	0.64e-3	0.35e-3	0.29e-3
$\sqrt{5}$e-2	0.1e-2	0.125	0.40e-3	0.33e-3	0.79e-4
0.1e-1	0.1e-1	0.25e-1	0.36e-3	0.35e-3	0.10e-5
0.5e-2	0.25e-3	0.25e-1	0.91e-4	0.88e-4	0.34e-5
0.1e-1	0.1e-2	0.25e-1	0.36e-3	0.35e-3	0.10e-4
0.2e-1	0.4e-2	0.25e-1	0.14e-2	0.14e-2	0.81e-5
0.1e-1	1	0.25e-4	0.15	0.15	0.2e-14
0.1e-1	0.1e-1	0.25e-2	0.35e-2	0.35e-2	0.22e-8
0.1e-1	0.1e-4	0.25e+1	0.29e-1	0.32e-5	0.29e-1

Table 3.3: Example 3.1 in Cartesian coordinates – maximum error in the energy using collocation at 3 Gaussian points: fast components, slow components and total energy.

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	ΔE
0.1e-1	0.1e-2	0.25e-1	0.35e-3	0.35e-3	0.48e-5
0.1e-1	0.2e-3	0.125	0.10e-3	0.71e-4	0.31e-4
0.1e-1	0.1e-3	0.25	0.15e-3	0.35e-4	0.11e-3
0.1e-2	0.25e-4	1.0	0.18e-2	0.86e-5	0.18e-2
$2\sqrt{10}$e-2	0.1e-2	1.0	0.22e-2	0.34e-3	0.19e-2
$\sqrt{10}$e-2	0.1e-2	0.25	0.49e-3	0.35e-3	0.15e-3
$\sqrt{5}$e-2	0.1e-2	0.125	0.40e-3	0.35e-3	0.51e-4
0.1e-1	0.1e-2	0.25e-1	0.35e-3	0.35e-3	0.48e-5
0.5e-2	0.25e-3	0.25e-1	0.91e-4	0.88e-4	0.25e-5
0.1e-1	0.1e-2	0.25e-1	0.35e-3	0.35e-3	0.48e-5
0.2e-1	0.4e-2	0.25e-1	0.14e-2	0.14e-2	0.15e-5
0.1e-1	1	0.25e-4	0.15	0.15	0.3e-14
0.1e-1	0.1e-1	0.25e-2	0.35e-2	0.35e-2	0.11e-10
0.1e-1	0.1e-4	0.25e+1	0.11e-1	0.30e-5	0.11e-1

Table 3.4: Example 3.1 in Cartesian coordinates – maximum error in the energy using collocation at 4 Gaussian points: fast components, slow components and total energy.

step-size k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	$\varepsilon^{-1}\Delta E_S$	ΔE	$\alpha^{-1}\Delta E$
0.1e-1	0.1e-2	0.25e-1	0.22e-1	0.36e-3	0.36	0.21e-1	0.85
0.1e-1	0.2e-3	0.125	0.11	0.72e-4	0.36	0.11	0.90
0.1e-1	0.1e-3	0.25	0.23	0.36e-4	0.36	0.23	0.93
$\sqrt{10}$e-2	0.1e-2	0.25	0.23	0.37e-3	0.37	0.23	0.92
$\sqrt{5}$e-2	0.1e-2	0.125	0.11	0.36e-3	0.36	0.11	0.90
0.1e-1	0.1e-2	0.25e-1	0.22e-1	0.36e-3	0.36	0.21e-1	0.85
0.5e-2	0.25e-3	0.25e-1	0.22e-1	0.89e-4	0.36	0.22e-1	0.86
0.1e-1	0.1e-2	0.25e-1	0.22e-1	0.36e-3	0.36	0.21e-1	0.85
0.2e-1	0.4e-2	0.25e-1	0.23e-1	0.14e-2	0.36	0.22e-1	0.87
0.1e-1	1	0.25e-4	0.15	0.15	0.15	0.40e-4	0.16e+1
0.1e-1	0.1e-1	0.25e-2	0.57e-2	0.35e-2	0.35	0.21e-2	0.85

Table 3.5: Example 3.1 transformed by a time-dependent orthogonal transformation from polar coordinates – maximum error in the energy using midpoint: fast components, slow components and total energy.

discrete values for q and p. Since the transformation is well-conditioned and the results in Table 3.1 are accurate, we obtain an accurate approximation for the problem in Cartesian coordinates even for very small ε and large α. But the other possibility, of transforming the problem first *to (3.3) and* then *discretizing by the same method, does not work so well. Thus, the main source of error in Tables 3.2–3.4 can be interpreted as arising from the inaccurate reproduction by the numerical discretization of the coupling transformation from (3.1) to (3.3). Indeed, this transformation depends on the solution and as such is time-varying, giving rise in general to $O(k^2)$ discrepancies (cf. §2.3 and [5, 2]). For (3.1b) we have then the approximation*

$$-\varepsilon^{-2}((r_n + r_{n-1})/2 + O(k^2) - r_0) = -\varepsilon^{-2}((r_n + r_{n-1})/2 + \varepsilon O(\varepsilon^{-1}k^2) - r_0)$$

(recall that $r(q) = r_0 + O(\varepsilon)$ in this example). Hence we see that perturbations of size $O(\varepsilon^{-1}k^2)$ can be introduced into the midpoint discretization of (3.1). If these perturbations are large and depend unfavourably on the solution then the scheme for (3.3) may become unstable, as we proved in §2.3 for a simpler, linear example. This also typically causes difficulties in the Newton iterations for the nonlinear problem.

The construction of the example where the midpoint method does not perform adequately for large α is simple, but general. Other simple coupling transformations may be harmful as well. For instance, consider the linear time-dependent transformation (2.7) of Example 2.2. This gives for x and y the system

$$\dot{x} = Q^T \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix} Qy + Kx \tag{3.4a}$$

$$\dot{y} = Q^T \begin{pmatrix} -\varepsilon^{-2}(r(x) - r_0) + p_\phi(y)^2 r(x)^{-3} \\ -(\phi(x) - \phi_0) \end{pmatrix} + Ky\,. \tag{3.4b}$$

The results listed in Table 3.5 exhibit similar phenomena for this transformed system as for the problem in Cartesian coordinates. A plot of the error in total energy H is given in Figure 3.2 for $\alpha = 1.25$, $T = 20$ and $\varepsilon = 0.1e-3$.

□

Example 3.2 *Qualitatively similar results are obtained when we replace the potential $V = \frac{1}{2}(\phi - \phi_0)^2$ by the potential $V = -q_2$ in (3.2). This example was considered in [11, 14, 19]. The expression for the fast energy $E_F(t)$ remains the same as in Example 3.1, while the slow energy becomes*

$$E_S(t) = \frac{1}{2} r^{-2} p_\phi^2 - q_2\,.$$

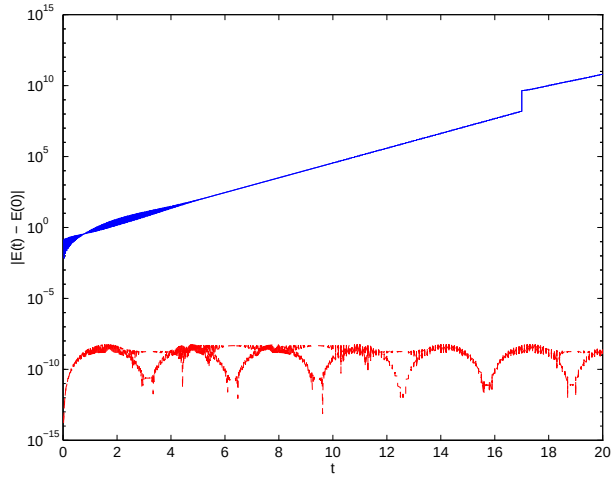


Figure 3.2: Errors in total energy for $\varepsilon = 0.1e-3$, $\alpha = 1.25$ and $T = 20$, using midpoint for Example 3.1 in linearly-transformed and polar coordinates. The error in the transformed problem eventually grows unacceptably.

The results, and their explanations, are sufficiently similar to those in Tables 3.1–3.4 so that we omit reproducing them in this writeup. Note that the fast and slow variables here are r and ϕ respectively, as in the previous example, with $q_2 = r\sin(\phi)$ in the expression for E_S. The corresponding equations of motion in (r, ϕ)-coordinates are given by

$$\begin{aligned} \dot{r} &= p_r \\ \dot{p}_r &= -\varepsilon^{-2}(r-1) + \sin(\phi) + p_\phi^2 r^{-3} \\ \dot{\phi} &= r^{-2} p_\phi \\ \dot{p}_\phi &= r\cos(\phi) \end{aligned}$$

Hence in the fast variables we essentially have again a linear harmonic oscillator.

Note that the Jacobian of the linearized problem in Cartesian coordinates has been observed to have large $O(k^2/\varepsilon)$ eigenvalues for this example [19, 26]. This is usually a good indication of potential trouble which, however, does not form a proof of instability, does not always raise an alarm when there is trouble (e.g. Example 2.2 with $\alpha > 1$), and does not offer a methodology for generating such examples.

Next, we vary the initial conditions so as to correspond to Cases 1 and 2 in the Introduction. So we choose as initial values (keeping also $r_0 = 1$)

- *(i)* $q = (1, 0)$, $p = (0, 0)$
- *(ii)* $q = (1 - \varepsilon, 0)$, $p = (0, 0)$.

The solution for case (i) is smooth to our working accuracy, and the initial energies are $E_F(0) = E_S(0) = 0$. Also, for the exact solution, $\Delta E_S = O(\varepsilon^2)$ and $r(t) - r_0 = O(\varepsilon^2)$. In contrast, for case (ii) $E_F(0) = 0.5$ (indeed, clearly $\frac{\partial E_F}{\partial r} = \frac{\partial H}{\partial r} = O(\varepsilon^{-1})$ when $|r - r_0| \geq \varepsilon$) and, as in the previous examples, there is a rapid oscillation of magnitude and frequency $O(\varepsilon)$ on top of a smooth solution curve. In particular, $\Delta E_S = O(\varepsilon)$ and $r(t) - r_0 = O(\varepsilon)$. Finally, we increase the integration interval to $T = 20$, to allow instabilities, if there are any, to develop even for the smooth case (i).

The results are listed in Tables 3.6 and 3.7. For the midpoint scheme in case (ii) we again observe $\Delta E = O(\alpha)$, for k, ε and α small enough. For case (i) this also seems to hold, but less

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	$(i)\Delta E$	$(i)\Delta E_S$	(ii) ΔE	(ii) ΔE_S
0.1	0.1e-2	0.25e+1	0.30e+5	0.27e+1	0.46e+5	0.23e+2
0.5e-1	0.1e-2	0.63	0.20	0.94e-3	0.50	0.32e-2
0.25e-1	0.1e-2	0.16	0.12e-1	0.24e-3	0.15	0.32e-2
0.125e-1	0.1e-2	0.39e-1	0.82e-3	0.63e-4	0.39e-1	0.31e-2
0.1e-2	0.1e-2	0.25e-3	0.41e-6	0.49e-5	0.25e-3	0.30e-2
0.1	0.1e-3	0.25e+2	0.50e+6	0.41e+2	0.36e+7	0.40e+2
0.25e-1	0.1e-3	0.16e+1	0.37e+3	0.32e-1	0.20e+7	0.11e+2
0.125e-1	0.1e-3	0.39	0.76e-1	0.59e-4	0.36	0.35e-3
0.1e-2	0.1e-3	0.25e-2	0.35e-5	0.42e-6	0.25e-2	0.30e-3

Table 3.6: Example 3.2 – maximum error in the energy for cases (i) and (ii) using midpoint: total energy and slow energy.

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	$(i)\Delta E$	$(i)\Delta E_S$	(ii) ΔE	(ii) ΔE_S
0.2	0.1e-2	0.10e+2	0.26e+7	0.23e+5	0.11e+6	0.15e+3
0.1	0.1e-2	0.25e+1	0.11e-5	0.19e-5	0.21	0.28e-2
0.5e-1	0.1e-2	0.63	0.49e-8	0.43e-5	0.69e-2	0.30e-2
0.25e-1	0.1e-2	0.16	0.34e-10	0.45e-5	0.50e-3	0.30e-2
0.125e-1	0.1e-2	0.39e-1	0.15e-10	0.45e-5	0.58e-4	0.30e-2
0.1e-2	0.1e-2	0.25e-3	0.28e-11	0.45e-5	0.18e-9	0.30e-2
0.1	0.1e-3	0.25e+2	0.36e+8	0.24e+4	0.40e+9	0.56e+7
0.25e-1	0.1e-3	0.16e+1	0.20e-8	0.35e-7	0.49e-1	0.30e-3
0.125e-1	0.1e-3	0.39	0.15e-8	0.45e-7	0.26e-2	0.30e-3
0.1e-2	0.1e-3	0.25e-2	0.54e-9	0.45e-7	0.41e-6	0.30e-3

Table 3.7: Example 3.2 – maximum error in the energy for cases (i) and (ii) using collocation at 3 Gaussian points: total energy and slow energy.

clearly so. When α is too large the solution quality is poor, or the nonlinear (Newton) iterations do not converge in most time steps. For the higher order Gauss collocation schemes the errors are smaller for α small enough, but when α is not small the solution quality is poor here too. We have not detected for this example the qualitative difference between the schemes suggested in [11, 14], viz., we have not found that higher order Gauss collocation converge while the midpoint scheme diverges as $k \to 0$ with α held large enough and fixed, as t is increased.

Finally, the error in the total energy for $\alpha = 1.5625$ and $T = 20$ with the initial values of Example 3.1 is plotted in Figure 3.3. While the error for the polar formulation remains small (even though the decoupling between fast and slow variables is not as "clean" as in Example 3.1 [19]) the error for the Cartesian formulation grows large. The error does not explode, though – the pattern is more complex. More typically in our calculations, Newton's method ceases to converge when the errors become large (e.g. if we try $\varepsilon = 1.d-5$ with the same α). No convergence or accuracy problems are detected for the polar formulation using also more stringent values of ε and α. □

Example 3.3 *A more complex situation arises for the stiff "reversed" pendulum [7], where the radius $r(t)$ is the slow variable and the angle $\phi(t)$ is the fast variable. We use the same notation (hence also the same transformation between polar and Cartesian coordinates), constants and starting values as for Example 3.1.*

The Hamiltonian in polar coordinates is now

$$H(p,q) = E(r,\phi) = \frac{1}{2}[p_r^2 + r^{-2}p_\phi^2 + \varepsilon^{-2}(\phi-\phi_0)^2 + (r-r_0)^2]\,. \tag{3.5}$$

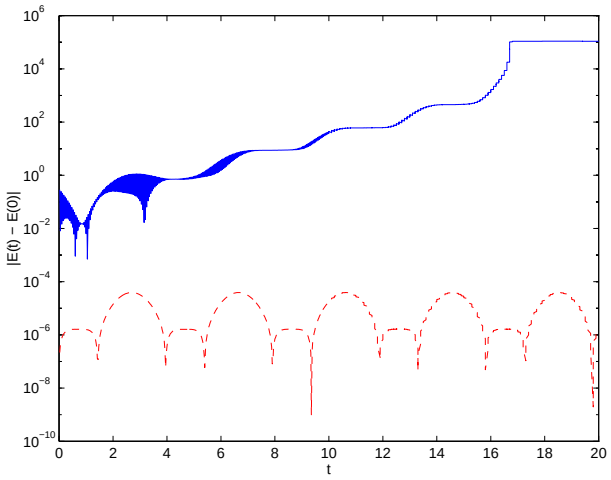


Figure 3.3: Errors in total energy for $\varepsilon = 0.1e-3$, $\alpha = 1.5625$ and $T = 20$, using midpoint for Example 3.2 in Cartesian and polar coordinates. The error in Cartesian coordinates becomes large and grows like a staircase.

The equations of motion are

$$\begin{aligned} \dot{r} &= p_r && (3.6a)\\ \dot{p_r} &= -(r-r_0) + p_\phi^2 r^{-3} && (3.6b)\\ \dot{\phi} &= r^{-2} p_\phi && (3.6c)\\ \dot{p_\phi} &= -\varepsilon^{-2}(\phi - \phi_0)\,. && (3.6d) \end{aligned}$$

For $\varepsilon \ll 1$, the system is clearly highly oscillatory in the variables (ϕ, p_ϕ). The essential analytical difference between this example and the previous two is that here

$$GG^T = \nabla\phi^T \nabla\phi = r^{-2}$$

which varies slowly in t (because r is a slow variable), whereas before we had $GG^T = \nabla r^T \nabla r \equiv 1$. It transpires that in the limit $\varepsilon \to 0$, the reduced equations of motion for the slowly varying variables (r, p_r) are given by [21, 7]

$$\begin{aligned} \dot{r} &= p_r \\ \dot{p_r} &= -(r-r_0) + c r^{-2}\,. \end{aligned}$$

Note the correcting force term $F = cr^{-2}$, where c is an appropriate constant determined below.

The fast system can be considered as a harmonic oscillator (2.4) with slowly varying frequency $\omega \sim r^{-1}$. The relevant energy is $E_F(t) = \frac{1}{2}[r^{-2}p_\phi^2 + \varepsilon^{-2}(\phi - \phi_0)^2]$. The corresponding adiabatic invariant is

$$J(t) := \frac{E_F(t)}{r^{-1}(t)} = r(t) E_F(t)$$

which is preserved up to $O(\varepsilon)$ terms. Thus, as we let $\varepsilon \to 0$,

$$E_F(t) = r^{-1}(t)\, J(0)$$

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔJ	$\alpha^{-1}\Delta J$	ΔE_S	$\alpha\varepsilon^{-1}\Delta E_S$	ΔE
0.1e-1	0.1e-2	0.25e-1	0.69e-2	0.28	0.69e-2	0.17	0.57
0.1e-1	0.2e-3	0.125	0.41e-1	0.33	0.29e-3	0.18	0.51
0.1e-1	0.1e-3	0.25	0.77e-1	0.31	.78e-4	0.19	0.40
0.1e-1	0.1e-3	0.25	0.77e-1	0.31	.78e-4	0.19	0.40
$\sqrt{50}$e-3	0.1e-3	0.125	0.43e-1	0.34	0.15e-3	0.18	0.51
$\sqrt{10}$e-3	0.1e-3	0.25e-1	0.73e-2	0.29	0.71e-3	0.18	0.60
0.1e-2	0.1e-4	0.25e-1	0.74e-2	0.30	0.71e-4	0.18	0.60
0.2e-2	0.4e-4	0.25e-1	0.73e-2	0.29	0.28e-3	0.18	0.60
0.4e-2	0.16e-3	0.25e-1	0.73e-2	0.29	0.11e-2	0.18	0.60
0.2e-2	0.1e-4	0.1	0.33e-1	0.33	0.18e-4	0.18	0.64
0.2e-2	1	0.1e-5	0.11	0.11e+6	0.15	0.15e-6	0.38e-6
0.1e-3	0.1e-2	0.25e-5	0.92e-4	0.37e+2	0.17	0.42e-3	0.77e-3
0.1e-4	0.1e-2	0.25e-7	0.92e-4	0.37e+4	0.17	0.42e-5	0.77e-5
0.1e-1	0.1e-5	0.25e+2	0.57	0.23e-1	0.37e-9	0.92e-2	0.26e+1
0.1	0.1e-5	0.25e+4	0.60	0.24e-3	0.44e-15	0.11e-5	0.27e+1

Table 3.8: Example 3.3 – maximum errors for the reversed pendulum using midpoint: polar coordinates.

and, since the time average of p_ϕ^2 satisfies

$$r^{-2}(t)\,\langle p_\phi^2\rangle(t) \quad = \quad E_F(t)$$

(the time average is taken over an interval long compared to ε and short compared to 1), the p_ϕ^2 term in the slow part can be eliminated by means of

$$\langle p_\phi^2\rangle(t) \quad = \quad r(t)\,J(0)\,.$$

This yields the desired correcting force with $c = J(0)$. For the chosen initial conditions, we have $J(0) = 0.5$.

In Table 3.8 we list results using the midpoint scheme to discretize (3.6), for various values of the step-size k, the parameter ε, and $\alpha := \frac{k^2}{4\varepsilon}$. Here ΔE_S, ΔJ, and ΔE are defined by

$$\Delta J = \max_{t\in[0,5]} |J(0) - J(t)|\,,$$

$$\Delta E_S = \max_{t\in[0,5]} |E_S(0) - E_S(t)|\,,$$

and

$$\Delta E = \max_{t\in[0,5]} |E(0) - E(t)|$$

where $E_S(t) = p_r^2/2 + 1/2(r - r_0)^2 = E - E_F$ is the energy in the slow variables (r, p_r) (without the correcting potential energy term!) and $J(t)$ is the above defined adiabatic invariant corresponding to the fast variables. The exact solution satisfies $\Delta J = O(\varepsilon)$ (cf. §2.2).

The midpoint results recorded in Table 3.8 are significantly worse than those in Table 3.1. They clearly suggest that when $k > \varepsilon$ and α is not large, $\Delta J = O(\alpha)$, and $\Delta E_S = O(\varepsilon/\alpha) = O(\varepsilon^2/k^2)$. The total energy error is much larger than in Tables 3.1 and 3.2, and advises what sort of errors can be expected without the special effects of almost quadratic invariants. For these calculations, $\Delta E_F = \Delta E - \Delta E_S \approx \Delta E$. The basic reason for this additional complication is that the fast linear oscillator (3.6c)-(3.6d) has a slowly varying frequency. This is explained in §2.2 (cf. Example 2.1).

Perhaps the most troubling aspect of the results recorded in Table 3.8 is the behaviour of ΔE_S when $\varepsilon \ll k$. The indicated behaviour may well suggest to an unaware scientist that $\Delta E_S \to 0$ as

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔJ	$\alpha^{-1}\Delta J$	ΔE_S	$\alpha\varepsilon^{-1}\Delta E_S$	ΔE
0.1e-1	0.1e-2	0.25e-1	0.72e-2	0.29	0.69e-2	0.17	0.57
0.1e-1	0.2e-3	0.125	0.53e-1	0.42	0.29e-3	0.18	0.70
0.1e-1	0.1e-3	0.25	0.13	0.54	.78e-4	0.19	0.96
0.1e-1	0.1e-3	0.25	0.13	0.54	.78e-4	0.19	0.96
$\sqrt{50}$e-3	0.1e-3	0.125	0.54e-1	0.43	0.14e-3	0.18	0.70
$\sqrt{10}$e-3	0.1e-3	0.25e-1	0.75e-2	0.30	0.71e-3	0.18	0.60
0.1e-2	0.1e-4	0.25e-1	0.75e-2	0.30	0.71e-4	0.18	0.61
0.2e-2	0.4e-4	0.25e-1	0.75e-2	0.30	0.28e-3	0.18	0.61
0.4e-2	0.16e-3	0.25e-1	0.75e-2	0.30	0.11e-2	0.18	0.60
0.2e-2	0.1e-4	0.1	0.34e-1	0.34	0.18e-4	0.18	0.58
0.2e-2	1	0.1e-5	0.11	0.11e+6	0.15	0.15e-6	0.25e-6
0.1e-3	0.1e-2	0.25e-5	0.92e-4	0.37e+2	0.17	0.42e-3	0.77e-3
0.1e-4	0.1e-2	0.25e-7	0.92e-4	0.37e+4	0.17	0.42e-5	0.77e-5

Table 3.9: Example 3.3 – maximum errors for the reversed pendulum using midpoint: Cartesian coordinates.

$\varepsilon \to 0$. *In actuality,* $\Delta E_S \approx 0.17$ *for the exact solution when* ε *is small, as the numerical results using* $k < \varepsilon$ *clearly indicate;* ΔE_S *does not shrink because* E_S *does not include the correcting potential. When* $\varepsilon \ll k$*, however, we get the approximation* $p_\phi = O(\varepsilon/k)$ *at internal stages (recall §2.1), so* $p_\phi^2 r^{-3} = O(\varepsilon^2/k^2)$ *in (3.6b). In (3.6a)-(3.6b) we get (erroneously) an* $O(\varepsilon^2/k^2)$ *perturbation of a linear oscillator whose energy is* E_S*, explaining the computed results for* ΔE_S*. In short, the effect observed here is that the correcting force is approximated to* 0 *as* $\varepsilon/k \to 0$*, yielding a plausible-looking yet* wrong *result!* [3]

Next we consider the problem in Cartesian coordinates. The Hamiltonian here is

$$H(q,p) = \frac{1}{2}[p^T p + \varepsilon^{-2}(\phi(q) - \phi_0)^2 + (r(q) - r_0)^2]$$

and the equations of motion are

$$\dot{q} = p \tag{3.7a}$$
$$\dot{p} = -\varepsilon^{-2}(\phi(q) - \phi_0)\nabla\phi(q) - (r(q) - r_0)\nabla r(q) \tag{3.7b}$$

In the limit $\varepsilon \to 0$ *we get the (nonobvious) DAE*

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -(r(q) - r_0)\nabla r(q) - J(0)r^{-2}\nabla r(q) - \nabla\phi(q)\lambda \\ 0 &= \phi(q) - \phi_0 \end{aligned}$$

which contains the correcting force term $F = J(0)r^{-2}$*.*

We now consider discretization of (3.7) by the implicit midpoint scheme. The results are listed in Table 3.9. Not surprisingly, the error behaviour is more complicated here than in Table 3.8 or in Table 3.2, although there is certainly no additional accuracy. The results for E_S *are still misleading, as in Table 3.8. The errors recorded in Table 3.9 are generally close to the corresponding ones in Table 3.8. This closeness breaks down, though, when* α *is increased further and coupling effects take over. For* $\alpha > 1$ *we did not obtain convergence at all steps for the nonlinear Newton iteration when discretizing (3.7); in contrast, no such problem was observed for (3.6).*

For this problem, then, reliability requires using $k = O(\varepsilon)$*. Next, we set* $k \leq \varepsilon$ *and calculate solutions using the midpoint scheme, as well as the Verlet scheme. The latter can be viewed as*

[3] It may be argued that the error in total energy issues a warning that smooth quantities may not be approximated well here. This is true, but note that J is still approximated much better, and that ΔE_S appears to have a regular shape when $\varepsilon \ll k$, despite the large error in energy.

method	k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔJ	ΔE_S	ΔE
midpoint	0.1e-1	0.1e-1	0.25e-2	0.11e-2	0.14	0.72e-1
Verlet	0.1e-1	0.1e-1	0.25e-2	0.19	0.19	0.22
midpoint	0.1e-2	0.1e-2	0.25e-3	0.11e-3	0.14	0.72e-1
Verlet	0.1e-2	0.1e-2	0.25e-3	0.19	0.19	0.22
midpoint	0.1e-2	0.1e-1	0.25e-4	0.93e-3	0.17	0.77e-3
Verlet	0.1e-2	0.1e-1	0.25e-4	0.13e-2	0.17	0.15e-2
midpoint	0.1e-3	0.1e-2	0.25e-5	0.92e-4	0.17	0.77e-3
Verlet	0.1e-3	0.1e-2	0.25e-5	0.12e-2	0.17	0.15e-2

Table 3.10: Example 3.3 – maximum errors for the reversed pendulum using midpoint and Verlet.

a staggered midpoint *scheme: for the partitioned system (1.1) one discretizes (1.1a) centred at* $t_{n-1/2} = (t_n + t_{n-1})/2$ *and (1.1b) centred at* t_n*,*

$$\begin{aligned} \frac{q_n - q_{n-1}}{k} &= p_{n-1/2} \\ \frac{p_{n+1/2} - p_{n-1/2}}{k} &= \text{grad } V(q_n) + \varepsilon^{-2} G^T g(q_n) \end{aligned}$$

yielding the usual 3-point scheme, [4]

$$q_{n+1} - 2q_n + 2q_{n-1} = -k^2[\text{grad } V(q_n) + \varepsilon^{-2} G^T g(q_n)]\,. \tag{3.8}$$

This scheme is explicit (because the Hamiltonian is separated) and has a stability restriction when applied to a linear oscillator (2.3), requiring that

$$k \le 2\varepsilon\,. \tag{3.9}$$

There is no such requirement for the midpoint scheme, despite the closeness between these two discretizations.

The results are listed in Table 3.10. The midpoint results for ΔE_S *are no longer misleading. For* $k = 0.1\varepsilon$ *the listed total error using the midpoint scheme is only slightly better than the (cheaper, explicit) staggered midpoint scheme, but the error in* J *is smaller. For* $k = \varepsilon$ *the errors in both total energy and in* J *are much smaller using the midpoint scheme, indicating its potential attraction despite the possible setbacks for* α *not small.* □

Acknowledgement: We thank Bob Skeel for many fruitful discussions that have led to significant improvements of our original manuscript.

References

[1] U. Ascher. Two families of symmetric difference schemes for singular perturbation problems. In U. Ascher and R. Russell, editors, *Numerical Boundary Value ODEs*. Birkhauser, 1985.

[2] U. Ascher. On symmetric schemes and differential-algebraic equations. *SIAM J. Scient. Comput.*, 10:937–949, 1989.

[3] U. Ascher. Stabilization of invariants of discretized differential systems. *Numerical Algorithms*, 14:1–23, 1997.

[4] The variant in [25] shifts the mesh by $k/2$.

[4] U. Ascher, H. Chin, L. Petzold, and S. Reich. Stabilization of constrained mechanical systems with daes and invariant manifolds. *J. Mech. Struct. Machines*, 23:135–158, 1995.

[5] U. Ascher, R. Mattheij, and R. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. SIAM, 1995.

[6] U.M. Ascher and S. Reich. On difficulties in integrating highly oscillatory hamiltonian systems. 1997. Manuscript.

[7] F.A. Bornemann and Ch. Schutte. A mathematical approach to smoothed molecular dynamics: Correcting potentials for freezing bond angles. Technical report, Konrad-Zuse-Zentrum Berlin, 1995.

[8] G.J. Cooper. Stability of runge–kutta methods for trajectory problems. *IMA J. Numer. Anal.*, 7:1–13, 1987.

[9] O. Gonzalez. Mechanical systems subject to holonomic constraints: differential-algebraic formulations and conservative integration. *Physica D*, 1997. to appear.

[10] O. Gonzalez and J. Simo. On the stability of symplectic and energy-momentum algorithms for nonlinear hamiltonian systems with symmetry. *Comp. Meth. in Appl. Mech. Eng.*, 134:197–222, 1996.

[11] E. Hairer and L. Jay. Implicit Runge-Kutta methods for higher-index differential-algebraic systems. *WSSIAA, Contributions in Numerical Mathematics*, 2:213–224, 1993.

[12] E. Hairer and D. Stoffer. Reversible long term integration with variable step sizes. *SIAM J. Scient. Comput.*, 18:257–269, 1997.

[13] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, 1991.

[14] L. Jay and L. Petzold. Highly oscillatory systems and periodic-stability. manuscript, 1995.

[15] B. Leimkuhler and R. Skeel. Symplectic integrators in constrained hamiltonian systems. *J. Comp. Phys.*, 112:117–125, 1994.

[16] A.J. Lichtenberg and M.A. Lieberman. *Regular and Stochastic Motions*. Springer Verlag, 1983.

[17] M. Mandziuk and T. Schlick. Resonance in the dynamices of chemical systems simulated by the implicit midpoint scheme. *Chem. Phys. Lett.*, 237:525–535, 1995.

[18] A.I. Neishtadt. The separation of motions in systems with rapidily rotating phase. *J. Appl. Math. Mech.*, 48:133–139, 1984.

[19] L.R. Petzold, L.O. Jay, and J. Yen. Numerical solution of highly oscillatory ordinary differential equations. *Acta Numerica*, pages 437–484, 1997.

[20] S. Reich. Enhanced energy conserving methods. *BIT*, 36:122–134, 1996.

[21] H. Rubin and P. Ungar. Motion under a strong constraining force. *Comm. Pure Appl. Math.*, 10:65–87, 1957.

[22] J.M. Sanz-Serna and M.P. Calvo. *Numerical Hamiltonian Problems*. Chapman and Hall, 1994.

[23] T. Schlick, M. Mandziuk, R.D. Skeel, and K. Srinivas. Nonlinear resonance artifacts in molecular dynamics simulations. 1997. manuscript.

[24] M. Shimada and H. Yoshida. Long term conservation of adiabatic invariants. *Publ. Astron. Soc. Japan*, 48:147–155, 1996.

[25] R.D. Skeel, G. Zhang, and T. Schlick. A family of symplectic integrators: stability, accuracy, and molecular dynamics applications. *SIAM J. Sci. Comput.*, 18:203–222, 1997.

[26] R.D. Skeel and M. Zhang. Cheap implicit symplectic integrators. *Appl. Num. Math.*, 1997. To appear.