SIAM J. SCI. COMPUT.
Vol. 21, No. 3, pp. 1045–1065
© 1999 Society for Industrial and Applied Mathematics

THE MIDPOINT SCHEME AND VARIANTS FOR HAMILTONIAN SYSTEMS: ADVANTAGES AND PITFALLS*

URI M. ASCHER† AND SEBASTIAN REICH‡

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 the 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 typically still better than the corresponding ones for explicit schemes.

Key words. midpoint scheme, Hamiltonian systems, highly oscillatory problems, decoupling, stability, differential-algebraic equations, molecular dynamics, multibody systems

AMS subject classifications. 65L10, 65L20

PII. S1064827597316059

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 nonsmall 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 [12, 22]. Much attention has been paid recently in this context to the (implicit) midpoint scheme and some of its variants; see, e.g., [10, 14, 25] (see also [5] for a different orientation). The midpoint scheme, and more generally Gauss-collocation schemes, are algebraically stable and 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

*Received by the editors January 31, 1997; accepted for publication (in revised form) June 12, 1998; published electronically December 3, 1999.

http://www.siam.org/journals/sisc/21-3/31605.html

†Institute of Applied Mathematics and Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (ascher@cs.ubc.ca). The work of this author was partially supported by 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.

exhibit instabilities in various stiff situations [1, 2, 10, 17]. 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, a more severe $k = O(\varepsilon)$ is required (see Example 3.2). Moreover, if k^2/ε is not small, then significant error growth may be experienced when fast and slow modes are coupled by a nonconstant 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 ($0 \le t \le T$)

$$\dot{q} = p, \tag{1.1a}$$

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

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

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

is an integral invariant of this ODE system. We are interested in solutions with bounded energy as $\varepsilon \to 0$.

The presence of the strong potential in (1.1) introduces, in general, highly oscillatory solution modes with amplitude $O(\varepsilon)$ in q and $O(\varepsilon^0)$ in $p = \dot{q}$. To achieve a pointwise accurate approximate solution one must then choose $k \le c_1 \varepsilon$, where c_1 is a constant such that each oscillation is resolved by the mesh. In this case, it is better to choose the explicit Verlet scheme (which is a staggered midpoint scheme; see section 3) than one of the implicit variants of midpoint. On the other hand, if the oscillations are small in q and p, say,[2] $|\dot{p}| = O(\varepsilon^0)$, then we want to be able to choose $\varepsilon \ll k \ll 1$. Even in the more general situation, when $|\dot{p}| = O(\varepsilon^{-1})$, one could hope to accurately recover slowly varying quantities, assuming that k is small enough with respect to these slow variations. For example, one could expect that the Hamiltonian (1.2) is conserved up to the order of the method independently of the value of ε.

To be more precise, let us rewrite (1.1) as an index-1 DAE,

$$\dot{q} = p, \tag{1.3a}$$

$$\dot{p} = -\operatorname{grad} V(q) - G(q)^T \lambda, \tag{1.3b}$$

$$\varepsilon^2 \lambda = g(q). \tag{1.3c}$$

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

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

1. The exact solution satisfies

$$|g(q)| = O(\varepsilon^2).$$

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

$$\dot{q} = p, \tag{1.4a}$$

$$\dot{p} = -\operatorname{grad} V(q) - G(q)^T \lambda, \tag{1.4b}$$

$$0 = g(q). \tag{1.4c}$$

Here one may hope to accurately recover the entire solution, when using an appropriate implicit scheme with $\varepsilon \ll k \ll 1$.

2. The exact solution satisfies

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

The balance in powers of ε in (1.1) is achieved with $|p| = O(\varepsilon^0)$, $|\dot{p}| = O(\varepsilon^{-1})$. Upon discretizing with $\varepsilon \ll k \ll 1$, one still hopes to accurately recover q and H, for instance. 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[3] may turn out to be (1.4) (as in Example 3.1) or it may be different, incorporating an $O(1)$ correcting potential term (as in Example 3.2).

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 is a quadratic invariant which is reproduced exactly by the midpoint scheme.) We analyze this error in section 2.2 and explain its misleadingly regular form. The same type of error appears later in the more involved nonlinear Example 3.2.

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 nondecoupled form is discretized. These errors may induce instability. We prove in section 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. [1, 2, 5]). For systems with a rotational symmetry this has been shown rigorously in [10].

[3]The limit DAE describes the behavior of the slowly varying solution components of (1.1) as $\varepsilon \to 0$.

All of the examples in section 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 Example 3.1 which is related to the linear Example 2.2.

Instabilities in midpoint calculations with large k for this class of problems have been observed in practice before. 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 [3, 8]). Various energy conserving variants of the midpoint scheme have been considered in [6]. The most successful of these is the method investigated in [10] which can be interpreted as applying the midpoint discretization to the better-decoupled variables. All of these variants suffer the disadvantages demonstrated in Example 3.2 (see [6]).

We have organized our presentation around four examples. The two in section 2 are linear and simple enough to enable a thorough analysis. The two in section 3 are of nonlinear oscillators and serve in part to demonstrate that the pitfalls highlighted here are fundamental for many applications. In section 4 we offer conclusions.

2. Highly oscillatory linear problems. Consider an ODE system

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

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

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

This is a simple, symmetric, symplectic second-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 [3, 13]: writing such an s-stage scheme in the usual Butcher notation

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

we have

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

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

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

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 coefficient 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.

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 apply a modest modification to 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 $\omega(t) > 0$, $0 \le t \le T$. The modified system is still Hamiltonian with energy

$$H(q,p,t) = \frac{1}{2}[\omega^2(t)p^2 + \varepsilon^{-2}q^2].$$

Next, we discretize the equations (2.4) by the midpoint scheme with a step size k such that $k \ll 1$ and $\varepsilon/k \ll 1$. Clearly, the numerically computed solutions may be

wrong in pointwise absolute value; yet, one hopes that less is needed for the accurate computation of slowly varying quantities.

However, in the following example we demonstrate numerically and prove rigorously that, instead, errors in slowly varying quantities may not be small.

Example 2.1. We consider the errors in a quantity which varies particularly slowly in t. Since the energy is now time dependent, it is no longer a first integral for (2.4). However, there exists an *adiabatic invariant*

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

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

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

over a time interval of length

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

with $1 > c_1, c_2 > 0$ appropriate constants depending on $\omega(t)$ [18, 24].[4]

We have calculated the solutions for the midpoint approximation of (2.4) with

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

for various values of ε and k, measuring

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

Starting from the initial values yielding (a) $J(0) = H(0) = .5$, and (b) $J(0) = H(0) = .5\varepsilon^{-2}$, we find that the error in the adiabatic invariant depends (for ε and k small enough) only on

$$\alpha = \frac{k^2}{4\varepsilon} \tag{2.5}$$

in the manner that Figure 2.1 indicates. In particular, the error in the (almost constant) adiabatic invariant remains large if $k^2 \gg \varepsilon$, even as $k \to 0$ with α held fixed. A similar phenomenon is observed for higher-order Gauss collocation schemes.

Let us give an explanation for the observed phenomenon:

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

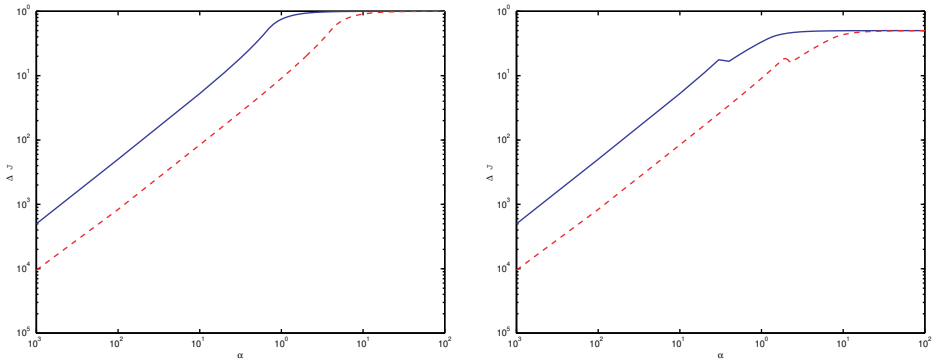
Consider the midpoint equations

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

and 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

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

[4]Note that, starting from $p(0) = \beta$, $q(0) = 0$, J is bounded in terms of β, and it varies smoothly for $0 \le t \le 1$ and ε small enough. Also, J is proportional to the ratio of total energy and frequency of the oscillator.



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

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

For the new variables we get

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

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

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

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"

$$-\omega^2(t)\alpha\dot{v} = u,$$
$$\alpha\dot{u} = v,$$

or

$$\ddot{u} = -(\omega(t)\alpha)^{-2}u. \tag{2.6}$$

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 (2.6) describes again a highly oscillatory system with slowly varying frequency. Its frequency is

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

and its Hamiltonian is

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

Thus the corresponding adiabatic invariant is

$$\hat{J}(u, v, t) = \frac{1}{2}[\omega^{-1}(t)u^2 + \omega(t)v^2].$$

So, similar to the original problem formulation, we obtain

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

over a time interval

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

with $1 > \hat{c}_1, \hat{c}_2 > 0$ 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

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

Thus, we have rigorously shown that, for α small enough or large enough, the error in the adiabatic invariant depends only on $\frac{k^2}{\varepsilon}$ as $k \to 0$ and $\varepsilon \to 0$, in the manner recorded in Figure 2.1.

2.3. Coupling effects. Here we generate a simple linear example where a midpoint discretization with a large step size results in an unstable numerical process. We prove this claim rigorously and demonstrate it numerically, giving a precise condition for this instability to occur. This condition is not based on the eigenvalues of the system.

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

$$\dot{u} = v, \tag{2.7a}$$

$$\dot{v} = -\begin{pmatrix} \varepsilon^{-2} & 0 \\ 0 & 1 \end{pmatrix} (u - \bar{u}), \tag{2.7b}$$

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 time-dependent linear transformation:

$$u = Qx,\ v = Qy, Q(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix},\ K = \dot{Q}^T Q = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{2.8}$$

This yields the coupled system

$$\dot{x} = y + Kx, \tag{2.9a}$$

$$\dot{y} = -Q^T \begin{pmatrix} \varepsilon^{-2} & 0 \\ 0 & 1 \end{pmatrix} Q(x - \bar{x}) + Ky, \tag{2.9b}$$

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

Now consider the midpoint scheme applied to the coupled linear system

$$k^{-1}(x_n - x_{n-1}) = y_{n-1/2} + K x_{n-1/2}, \tag{2.10a}$$

$$k^{-1}(y_n - y_{n-1}) = -Q_{n-1/2}^T \begin{pmatrix} \varepsilon^{-2} & 0 \\ 0 & 1 \end{pmatrix} Q_{n-1/2}(x_{n-1/2} - \bar{x}_{n-1/2}) + K y_{n-1/2}, \tag{2.10b}$$

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 $\alpha > 1$, where α is defined in (2.5). 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 (2.10). 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.,

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

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

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

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 (2.7) plus additional terms. Because the midpoint discretization of (2.7) 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

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

Similarly, for u_2 we have

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

and substituting this into the previous expression we get

$$\begin{aligned} &\varepsilon k^{-1}(v_{1,n} - v_{1,n-1}) \\ &\quad = (1 - \alpha^2)^{-1}[-\varepsilon^{-1}(u_{1,n-1/2} - \bar{u}_1) + \alpha v_{2,n-1/2} + \alpha u_{1,n-1/2}/2 + \cdots]. \end{aligned}$$

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 (2.7) 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 (2.8) to

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

for any constant $0 \le \omega \ll k^{-1}$, 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 (2.9) 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 two simple nonlinear examples. These examples are deemed more representative than those in the previous section of large nonlinear systems such as those arising in molecular dynamics. However, unlike for the linear examples our explanations here stop short of complete proofs. The examples involve difficulties which can be associated with the coupling of rapidly varying and slowly varying oscillators. Example 3.2 involves, in addition, also difficulties that can be associated with those analyzed in Example 2.1.

Example 3.1. This example demonstrates the dependence of the stability of midpoint discretization on the coupling of the unknowns in a simple mechanical system. We consider a stiff spring pendulum, first in polar coordinates. The equations of motion are

$$\dot{r} = p_r, \tag{3.1a}$$

$$\dot{p_r} = -\varepsilon^{-2}(r - r_0) + p_\phi^2 r^{-3}, \tag{3.1b}$$

$$\dot{\phi} = r^{-2} p_\phi, \tag{3.1c}$$

$$\dot{p_\phi} = -(\phi - \phi_0), \tag{3.1d}$$

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

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

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 section 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 $k^2/\varepsilon \gg 1$.

TABLE 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.*

Scheme s	k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	ΔE
1	.1e-1	.1e-2	.25e-1	.35e-3	.35e-3	.19e-6
3	.1e-1	.1e-2	.25e-1	.35e-3	.35e-3	.14e-6
4	.1e-1	.1e-2	.25e-1	.35e-3	.35e-3	.10e-6
1	.1	.1e-2	.25e+1	.34e-3	.34e-3	.10e-5
3	.1	.1e-2	.25e+1	.35e-3	.35e-3	.21e-6
4	.1	.1e-2	.25e+1	.31e-3	.31e-3	.20e-6
1	.1	.1e-5	.25e+4	.46e-9	.46e-9	.29e-11
3	.1	.1e-5	.25e+4	.27e-8	.27e-8	.30e-10
4	.1	.1e-5	.25e+4	.46e-8	.46e-8	.55e-10

We now discretize (3.1) by the midpoint scheme. The results for various values of the step size k, the problem parameter ε, and α of (2.5) 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

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

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

$$\dot{q} = p, \tag{3.3a}$$

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

TABLE 3.2

In Cartesian coordinates—maximum error in the energy using midpoint: 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	$\alpha^{-1}\Delta E$
.1e-1	.1e-2	.25e-1	.48e-2	.35e-3	.35	.45e-2	.18
.1e-1	.2e-3	.125	.22e-1	.71e-4	.36	.22e-1	.18
.1e-1	.1e-3	.25	.42e-1	.34e-4	.34	.42e-1	.17
$\sqrt{10}$e-2	.1e-2	.25	.42e-1	.34e-3	.34	.42e-1	.17
$\sqrt{5}$e-2	.1e-2	.125	.22e-1	.35e-3	.35	.22e-1	.18
.1e-1	.1e-2	.25e-1	.48e-2	.35e-3	.35	.45e-2	.18
.5e-2	.25e-3	.25e-1	.45e-2	.89e-4	.35	.44e-2	.18
.1e-1	.1e-2	.25e-1	.48e-2	.35e-3	.35	.45e-2	.18
.2e-1	.4e-2	.25e-1	.59e-2	.14e-2	.36	.45e-2	.18
.1e-1	.1e+1	.25e-4	.15	.15	.15	.62e-5	.25
.1e-1	.1e-1	.25e-2	.40e-2	.35e-2	.35	.45e-3	.18
.1e-1	.1e-5	.25+2	.13e+3	.19e-4	.19e+2	.13e+3	.50e+1

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

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

(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

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

We now consider the midpoint discretization of (3.3). The results for various values of k, ε, and α 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 three and four 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.

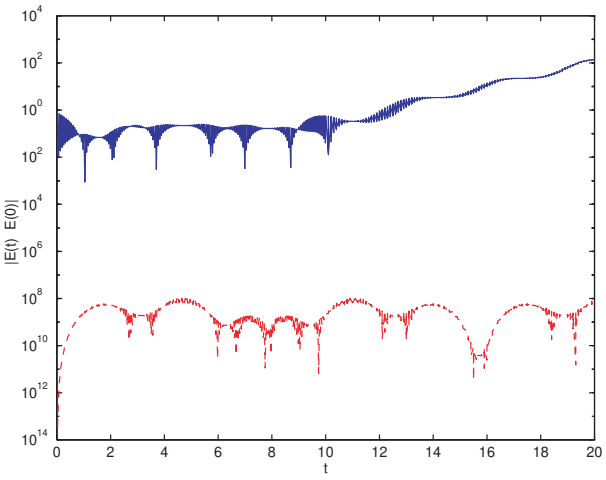


FIG. 3.1. *Errors in total energy for* $\varepsilon = .1e-3$, $\alpha = 2.5$, *and* $T = 20$, *using midpoint for Example* 3.1 *in Cartesian and polar coordinates. The error in Cartesian coordinates eventually grows unacceptably.*

To understand the observations above, 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 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 as 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. section 2.3 and [2, 5]). 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 unfavorably on the solution, then the scheme for (3.3) may become unstable, as we proved in section 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 (2.8) of Example 2.2. This gives for x and y the system

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

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

TABLE 3.3

In Cartesian coordinates—maximum error in the energy using collocation at three Gaussian points: fast components, slow components, and total energy.

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	ΔE
.1e-1	.1e-2	.25e-1	.36e-3	.35e-3	.10e-4
.1e-1	.2e-3	.125	.14e-3	.70e-4	.67e-4
.1e-1	.1e-3	.25	.29e-3	.35e-4	.26e-3
.1e-1	.25e-4	1.0	.41e-2	.83e-5	.41e-2
$2\sqrt{10}$e-2	.1e-2	1.0	.45e-2	.33e-3	.42e-2
$\sqrt{10}$e-2	.1e-2	.25	.64e-3	.35e-3	.29e-3
$\sqrt{5}$e-2	.1e-2	.125	.40e-3	.33e-3	.79e-4
.1e-1	.1e-1	.25e-1	.36e-3	.35e-3	.10e-5
.5e-2	.25e-3	.25e-1	.91e-4	.88e-4	.34e-5
.1e-1	.1e-2	.25e-1	.36e-3	.35e-3	.10e-4
.2e-1	.4e-2	.25e-1	.14e-2	.14e-2	.81e-5
.1e-1	.1e+1	.25e-4	.15	.15	.20e-14
.1e-1	.1e-1	.25e-2	.35e-2	.35e-2	.22e-8
.1e-1	.1e-4	.25e+1	.29e-1	.32e-5	.29e-1

TABLE 3.4

In Cartesian coordinates—maximum error in the energy using collocation at four Gaussian points: fast components, slow components, and total energy.

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔE_F	ΔE_S	ΔE
.1e-1	.1e-2	.25e-1	.35e-3	.35e-3	.48e-5
.1e-1	.2e-3	.125	.10e-3	.71e-4	.31e-4
.1e-1	.1e-3	.25	.15e-3	.35e-4	.11e-3
.1e-2	.25e-4	1.0	.18e-2	.86e-5	.18e-2
$2\sqrt{10}$e-2	.1e-2	1.0	.22e-2	.34e-3	.19e-2
$\sqrt{10}$e-2	.1e-2	.25	.49e-3	.35e-3	.15e-3
$\sqrt{5}$e-2	.1e-2	.125	.40e-3	.35e-3	.51e-4
.1e-1	.1e-2	.25e-1	.35e-3	.35e-3	.48e-5
.5e-2	.25e-3	.25e-1	.91e-4	.88e-4	.25e-5
.1e-1	.1e-2	.25e-1	.35e-3	.35e-3	.48e-5
.2e-1	.4e-2	.25e-1	.14e-2	.14e-2	.15e-5
.1e-1	.1e+1	.25e-4	.15	.15	.30e-14
.1e-1	.1e-1	.25e-2	.35e-2	.35e-2	.11e-10
.1e-1	.1e-4	.25e+1	.11e-1	.30e-5	.11e-1

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 $\varepsilon = .1e - 3$.

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). The resulting example was considered in [11, 14, 19]. The Jacobian of the linearized problem in Cartesian coordinates was observed to have large $O(k^2/\varepsilon)$ eigenvalues [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. Experiments with initial data corresponding to cases 1 and 2 of section 1 reveal a similar qualitative behavior, which is somewhat different from that suggested in [11, 14]. A fuller exposition of these experiments can be obtained by request via e-mail from one of the authors.

In the next and final example we discuss a case where (1.4) is not the correct limit

TABLE 3.5

Transformed by a time-dependent orthogonal transformation from polar coordinates—maximum error in the energy using midpoint: 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	$\alpha^{-1}\Delta E$
.1e-1	.1e-2	.25e-1	.22e-1	.36e-3	.36	.21e-1	.85
.1e-1	.2e-3	.125	.11	.72e-4	.36	.11	.90
.1e-1	.1e-3	.25	.23	.36e-4	.36	.23	.93
$\sqrt{10}$e-2	.1e-2	.25	.23	.37e-3	.37	.23	.92
$\sqrt{5}$e-2	.1e-2	.125	.11	.36e-3	.36	.11	.90
.1e-1	.1e-2	.25e-1	.22e-1	.36e-3	.36	.21e-1	.85
.5e-2	.25e-3	.25e-1	.22e-1	.89e-4	.36	.22e-1	.86
.1e-1	.1e-2	.25e-1	.22e-1	.36e-3	.36	.21e-1	.85
.2e-1	.4e-2	.25e-1	.23e-1	.14e-2	.36	.22e-1	.87
.1e-1	.1e+1	.25e-4	.15	.15	.15	.40e-4	.16e+1
.1e-1	.1e-1	.25e-2	.57e-2	.35e-2	.35	.21e-2	.85

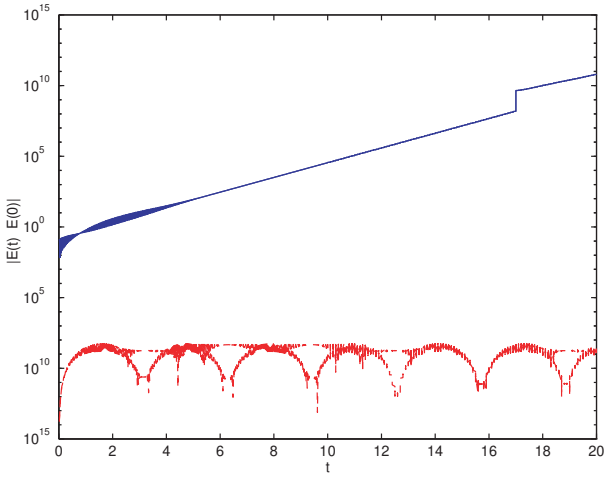


FIG. 3.2. *Errors in total energy for* $\varepsilon = .1e-3$, $\alpha = 1.25$, *and* $T = 20$, *using midpoint for Example* 3.1 *in linearly transformed and polar coordinates. The error in the transformed problem eventually grows unacceptably.*

DAE for (1.1). Here the midpoint discretization may wrongly approximate slowly varying solution quantities (cf. [6, 10] and Example 2.2), even in polar coordinates.

Example 3.2. 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

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

The equations of motion are

$$\dot{r} = p_r, \tag{3.6a}$$

$$\dot{p_r} = -(r - r_0) + p_\phi^2 r^{-3}, \tag{3.6b}$$

$$\dot{\phi} = r^{-2} p_\phi, \tag{3.6c}$$

$$\dot{p_\phi} = -\varepsilon^{-2}(\phi - \phi_0)\,. \tag{3.6d}$$

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 [7, 21]

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

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 a 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 $\varepsilon \to 0$,

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

and, since the time average of p_ϕ^2 satisfies

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

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

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

This yields the desired correcting force with $c = J(0)$. For the chosen initial conditions, we have $J(0) = .5$.

In Table 3.6 we list results using the midpoint scheme to discretize (3.6), for various values of the step size k, the problem parameter ε, and α of (2.5). Here ΔE_S, ΔJ, and ΔE are defined by

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

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

TABLE 3.6
Maximum errors for the reversed pendulum using midpoint: polar coordinates.

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔJ	$\alpha^{-1}\Delta J$	ΔE_S	$\alpha\varepsilon^{-1}\Delta E_S$	ΔE
.1e-1	.1e-2	.25e-1	.69e-2	.28	.69e-2	.17	.57
.1e-1	.2e-3	.125	.41e-1	.33	.29e-3	.18	.51
.1e-1	.1e-3	.25	.77e-1	.31	.78e-4	.19	.40
.1e-1	.1e-3	.25	.77e-1	.31	.78e-4	.19	.40
$\sqrt{50}$e-3	.1e-3	.125	.43e-1	.34	.15e-3	.18	.51
$\sqrt{10}$e-3	.1e-3	.25e-1	.73e-2	.29	.71e-3	.18	.60
.1e-2	.1e-4	.25e-1	.74e-2	.30	.71e-4	.18	.60
.2e-2	.4e-4	.25e-1	.73e-2	.29	.28e-3	.18	.60
.4e-2	.16e-3	.25e-1	.73e-2	.29	.11e-2	.18	.60
.2e-2	.1e-4	.1	.33e-1	.33	.18e-4	.18	.64
.2e-2	.1e+1	.1e-5	.11	.11e+6	.15	.15e-6	.38e-6
.1e-3	.1e-2	.25e-5	.92e-4	.37e+2	.17	.42e-3	.77e-3
.1e-4	.1e-2	.25e-7	.92e-4	.37e+4	.17	.42e-5	.77e-5
.1e-1	.1e-5	.25e+2	.57	.23e-1	.37e-9	.92e-2	.26e+1
.1	.1e-5	.25e+4	.60	.24e-3	.44e-15	.11e-5	.27e+1

and

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

where $E_S(t) = \frac{1}{2}[p_r^2 + (r - r_0)^2] = E - E_F$ 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. section 2.2).

The midpoint results recorded in Table 3.6 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 section 2.2 (cf. Example 2.1).

Perhaps the most troubling aspect of the results recorded in Table 3.6 is the behavior of ΔE_S when $\varepsilon \ll k$. The indicated behavior may well suggest to an unaware scientist that $\Delta E_S \to 0$ as $\varepsilon \to 0$. In actuality, $\Delta E_S \approx .17$ 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 section 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![5]

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

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

[5]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.

TABLE 3.7
Maximum errors for the reversed pendulum using midpoint: Cartesian coordinates.

k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔJ	$\alpha^{-1}\Delta J$	ΔE_S	$\alpha\varepsilon^{-1}\Delta E_S$	ΔE
.1e-1	.1e-2	.25e-1	.72e-2	.29	.69e-2	.17	.57
.1e-1	.2e-3	.125	.53e-1	.42	.29e-3	.18	.70
.1e-1	.1e-3	.25	.13	.54	.78e-4	.19	.96
.1e-1	.1e-3	.25	.13	.54	.78e-4	.19	.96
$\sqrt{50}$e-3	.1e-3	.125	.54e-1	.43	.14e-3	.18	.70
$\sqrt{10}$e-3	.1e-3	.25e-1	.75e-2	.30	.71e-3	.18	.60
.1e-2	.1e-4	.25e-1	.75e-2	.30	.71e-4	.18	.61
.2e-2	.4e-4	.25e-1	.75e-2	.30	.28e-3	.18	.61
.4e-2	.16e-3	.25e-1	.75e-2	.30	.11e-2	.18	.60
.2e-2	.1e-4	.1	.34e-1	.34	.18e-4	.18	.58
.2e-2	.1e+1	.1e-5	.11	.11e+6	.15	.15e-6	.25e-6
.1e-3	.1e-2	.25e-5	.92e-4	.37e+2	.17	.42e-3	.77e-3
.1e-4	.1e-2	.25e-7	.92e-4	.37e+4	.17	.42e-5	.77e-5

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 midpoint scheme. The results are listed in Table 3.7. Not surprisingly, the error behavior is more complicated here than in Table 3.6 or in Table 3.2, although there is certainly no additional accuracy. The results for E_S are still misleading, as in Table 3.6. The errors recorded in Table 3.7 are generally close to the corresponding ones in Table 3.6. 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 $k \le \varepsilon$ and calculate solutions using the midpoint scheme, as well as the Verlet scheme. The latter can be viewed as a *staggered midpoint* scheme: for the partitioned system (1.1) one discretizes (1.1a) centered at $t_{n-1/2} = (t_n + t_{n-1})/2$ and (1.1b) centered 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,[6]

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

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

TABLE 3.8
Maximum errors for the reversed pendulum using midpoint and Verlet.

Method	k	ε	$\alpha = \frac{k^2}{4\varepsilon}$	ΔJ	ΔE_S	ΔE
midpoint	.1e-1	.1e-1	.25e-2	.11e-2	.14	.72e-1
Verlet	.1e-1	.1e-1	.25e-2	.19	.19	.22
midpoint	.1e-2	.1e-2	.25e-3	.11e-3	.14	.72e-1
Verlet	.1e-2	.1e-2	.25e-3	.19	.19	.22
midpoint	.1e-2	.1e-1	.25e-4	.93e-3	.17	.77e-3
Verlet	.1e-2	.1e-1	.25e-4	.13e-2	.17	.15e-2
midpoint	.1e-3	.1e-2	.25e-5	.92e-4	.17	.77e-3
Verlet	.1e-3	.1e-2	.25e-5	.12e-2	.17	.15e-2

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 3.8. The midpoint results for ΔE_S are no longer misleading. For $k = .1\varepsilon$ 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 being small.

4. Conclusions. In this paper we have investigated 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 showed that in general one must require that k^2/ε be small enough, or else, even the errors in the energy and other slowly varying quantities 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 $k = O(\varepsilon)$ and the high frequency solution components are not accurately resolved. The same holds for higher-order collocation at Gaussian points. The encountered restrictions on k are typically still better than the corresponding ones for explicit schemes.

We summarize our experience as follows. Consider (1.1) again, with fast and slow modes clearly separable. In case that the solution of (1.1) oscillates with an amplitude of size $O(\varepsilon)$ about the reduced solution manifold

$$\mathcal{M} = \{(q,p) : g(q) = 0,\ G(q)p = 0\}$$

of (1.4), and if ε is so small 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. More generally, accurate approximations for slowly varying solution quantities are obtained for ε small by solving the *correct* limit DAE, if it is known. On the other extreme, there are fast, explicit schemes such as (3.8) if we are willing to accurately resolve the high frequency solution components, i.e., to use $k \le c_1\varepsilon$. 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 3.2) then some misleading calculations may arise. Moreover, energy-preserving variants [6, 9, 10, 20] do not help in such circumstances, and caution dictates simply taking $k \le c_2\varepsilon$ then, where typically it is still possible to choose the constant c_2 such that $c_2 > c_1$.

Acknowledgments. 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 Numerical Boundary Value ODEs, U. Ascher and R. Russell, eds., Birkhauser, 1985.
[2] U. Ascher, *On symmetric schemes and differential-algebraic equations*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 937–949.
[3] U. Ascher, *Stabilization of invariants of discretized differential systems*, Numer. Algorithms, 14 (1997), pp. 1–23.
[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 (1995), pp. 135–158.
[5] U. Ascher, R. Mattheij, and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, SIAM, Philadelphia, 1995.
[6] U. Ascher and S. Reich, *On difficulties in integrating highly oscillatory Hamiltonian systems*, in Algorithms for Macromolecular Modelling, P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R. D. Skeel, eds., Springer-Verlag, 1998, pp. 281–296.
[7] F. A. Bornemann and Ch. Schutte, *A Mathematical Approach to Smoothed Molecular Dynamics: Correcting Potentials for Freezing Bond Angles*, Konrad-Zuse-Zentrum, Berlin, 1995.
[8] G. J. Cooper, *Stability of Runge–Kutta methods for trajectory problems*, IMA J. Numer. Anal., 7 (1987), pp. 1–13.
[9] O. Gonzalez, *Mechanical systems subject to holonomic constraints: Differential-algebraic formulations and conservative integration*, Phys. D, submitted.
[10] O. Gonzalez and J. Simo, *On the stability of symplectic and energy-momentum algorithms for nonlinear Hamiltonian systems with symmetry*, Comput. Methods Appl. Mech. Engrg., 134 (1996), pp. 197–222.
[11] E. Hairer and L. Jay, *Implicit Runge–Kutta methods for higher-index differential-algebraic systems*, WSSIAA, Contrib. Numer. Math., 2 (1993), pp. 213–224.
[12] E. Hairer and D. Stoffer, *Reversible long-term integration with variable stepsizes*, SIAM J. Sci. Comput., 18 (1997), pp. 257–269.
[13] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II*: Stiff and Differential-Algebraic Problems*, Springer-Verlag, New York, Berlin, 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. Comput. Phys., 112 (1994), pp. 117–125.
[16] A. J. Lichtenberg and M. A. Lieberman, *Regular and Stochastic Motions*, Springer-Verlag, Berlin, New York, 1983.
[17] M. Mandziuk and T. Schlick, *Resonance in the dynamics of chemical systems simulated by the implicit midpoint scheme*, Chem. Phys. Lett., 237 (1995), pp. 525–535.
[18] A. I. Neishtadt, *The separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech., 48 (1984), pp. 133–139.
[19] L. R. Petzold, L. O. Jay, and J. Yen, *Numerical solution of highly oscillatory ordinary differential equations*, Acta Numerica, VI (1997), pp. 437–484.
[20] S. Reich, *Enhanced energy conserving methods*, BIT, 36 (1996), pp. 122–134.
[21] H. Rubin and P. Ungar, *Motion under a strong constraining force*, Comm. Pure Appl. Math., 10 (1957), pp. 65–87.

[22] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.
[23] T. Schlick, M. Mandziuk, R. D. Skeel, and K. Srinivas, *Nonlinear resonance artifacts in molecular dynamics simulations*, J. Comput. Phys., 139 (1998), pp. 1–19.
[24] M. Shimada and H. Yoshida, *Long term conservation of adiabatic invariants*, Publ. Astron. Soc. Japan, 48 (1996), pp. 147–155.
[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 (1997), pp. 203–222.
[26] R. D. Skeel and M. Zhang, *Cheap implicit symplectic integrators*, Appl. Numer. Math., 25 (1997), pp. 297–302.