Stabilization of invariants of discretized
differential systems

Uri M. Ascher*
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
Canada V6T 1Z4
asher@cs.ubc.ca
May 16, 1996

Abstract

Many problems of practical interest can be modeled by differential systems
where the solution lies on an invariant manifold defined explicitly by algebraic
equations. In computer simulations, it is often important to take into account
the invariant’s information, either in order to improve upon the stability of the
discretization (which is especially important in cases of long time integration)
or because a more precise conservation of the invariant is needed for the given
application.

In this paper we review and discuss methods for stabilizing such an invariant.
Particular attention is paid to post-stabilization techniques, where the stabi-
lication steps are applied to the discretized differential system. We summarize
theoretical convergence results for these methods and describe the application
of this technique to multibody systems with holonomic constraints. We then
briefly consider collocation methods which automatically satisfy certain, rela-
tively simple invariants. Finally, we consider an example of a very long time
integration and the effect of the loss of symplecticity and time-reversibility by
the stabilization techniques.

1 Differential systems with invariants

The numerical simulation of a differential system which has an invariant manifold has
received a lot of attention recently. In this work we consider the relatively simple case

*The work of this author was partially supported under NSERC Canada Grant OGP0004306. It
was written while the author was visiting Prof. M. Lentini and the CESMa group in Universidad
Simon Bolivar, Venezuela.
where the manifold can be defined by a set of explicitly specified algebraic equations. Concretely, consider the nonlinear \textit{Differential system}

\[
\frac{dz}{dt} \equiv \dot{z} = f(z)
\]  

(1.1)

and assume for simplicity that for each initial value vector \(z(0) = z_0\) there is a unique \(z(t)\) satisfying (1.1). Now, suppose in addition that there is an \textit{Invariant set} \(\mathcal{M}\) defined by the algebraic equations

\[
0 = h(z)
\]  

(1.2)

such that if \(h(z_0) = 0\) then \(h(z(t)) = 0\) for all \(t\). \(^1\)

The equations (1.1), (1.2) together may be viewed as an overdetermined system which the exact solution satisfies for each initial value \(z_0\). Once the system is discretized and approximated in order to solve it numerically, however, the full overdetermined system may not have a solution anymore. The question then becomes how to design a numerical method which preserves important properties given by the invariant, or which takes advantage of the additional information given by (1.2) to improve on a method’s stability or on the solution’s quality in case of a long time integration. A standard discretization for the initial value problem for (1.1) determines the approximate solution uniquely, but better results are often obtainable upon taking advantage of knowing (1.2).

It is possible to consider approaches on the continuous or on the discrete level. It is also possible to consider projection onto the invariant or just stabilization with respect to it. There are also “automatic” methods. This gives five different approaches (in addition to the one which simply ignores the invariant’s information):

1. \textbf{Projected invariant} (e.g. [21, 8]): With \(D(z)\) such that \(\exists (HD)^{-1}\) (e.g. \(D = HT\)), consider the related differential-algebraic equations (DAE)

\[
\begin{align*}
\dot{z} &= f(z) - D(z)y \\
0 &= h(z)
\end{align*}
\]  

(1.3a) (1.3b)

This DAE has pure index 2 (“Hessenberg form”) [11, 26]. The exact solution gives \(y = 0\), but this is no longer true in general for any discretization.

\(^1\)We assume sufficient smoothness as necessary, and use the following notation and additional assumptions:

- \(z(t), f(z(t))\) are vectors of length \(n_z\) and \(h(z(t))\) has length \(n_y \leq n_z\) for each \(t\).
- Denote \(H = h_z\) and assume that \(\text{rank}(H) = n_y\) for each \(t\).
- The functions \(f\) and \(h\) can depend explicitly on time \(t\) – we assume they don’t for notational simplicity.
2. **Stabilization** of the differential system with respect to the invariant (e.g. [5]). This relates closely also to Baumgarte’s stabilization [9].

Of course, both the projected invariant and the stabilization reformulations are succeeded by discretization of the resulting DAE or DE. In the next two approaches, one discretizes (1.1) and considers using the invariant information at the end of each time step.

3. **Coordinate projection** (e.g. [17, 2, 18]) applies a projection onto the algebraic manifold at the end of each time step.

4. **Post-stabilization** (e.g. [5, 4, 14]) applies a stabilization step with respect to the manifold at the end of each time step.

5. The “**automatic**” approach (e.g. [15, 33, 13, 31]) attempts to find a discretization for (1.1) which automatically satisfies also the equations (1.2).

Of these approaches, we ultimately concentrate on post-stabilization for general-purpose use. In the following sections it will hopefully become clear why. In §2 we briefly describe DAEs which are intimately related to ODEs with invariants. We have already seen in (1.3) that DAEs are obtained from ODEs with invariants, but a popular solution method for nonstiff DAEs also generates invariants. In §3 we then consider stabilization techniques. In §4 we consider stabilization of discrete dynamical systems, and in particular post-stabilization. We derive appropriate methods for constrained mechanical systems. Most of our exposition follows and describes the results of [14] and our related ones.

In §5 we then consider some situations where stabilization or projection onto the manifold may not be the best way to proceed. The “automatic” approach, if it is applicable without complications, has the attraction of not destroying other present structure (e.g. time-reversibility in certain situations) and of possible simplicity, although it appears more suitable for particular applications than for general purposes. A survey of such methods is well beyond the scope of this work, but we show which invariants collocation methods preserve. The essence appears in Cooper [15], but this result has been rediscovered a few times, and the proof for collocation is particularly elegant and revealing (and in some ways more general, too), relating the issue to quadrature precision alone. Applications include certain mechanical systems with quadratic constraints, preserving orthogonality of a solution matrix and (therefore) isospectral flows [16, 13]. Finally, we consider an example of symplectic vs projected invariant methods.

## 2 Differential-algebraic equations (DAEs)

The general form of DAEs is

\[ F(\dot{v}, v, t) = 0 \]
with $F_0$ singular (see, e.g., [11, 26]). However, here we consider further only differential equations with constraints, i.e. the semi-explicit case. Assume also, for simplicity, that the algebraic variables do not appear in the constraints and only appear linearly elsewhere:

\[
\begin{align*}
\dot{x} &= \hat{f}(x) - B(x)y \\
0 &= g(x)
\end{align*}
\]

(2.1a)  
(2.1b)

This DAE obviously has a very similar form to (1.3), except that $y \neq 0$ in general.\(^2\)

A class of higher-order, higher-index DAEs is given by:

\[
\begin{align*}
\ddot{x} &= \hat{f}(x, \dot{x}) - B(x, \dot{x})y \\
0 &= g(x)
\end{align*}
\]

(2.2a)  
(2.2b)

Here, if $\exists (GB)^{-1}$ then the DAE has index 3. The equations for multibody systems with holonomic constraints are in this form. It is well-known (e.g. [11]) that usual direct discretization schemes are ineffective here.

Even higher-order, higher-index DAEs can be easily incorporated into the discussion. In fact, everything we’ll say for (2.2) extends to the $m$th-order, $(m + 1)$-index DAE

\[
\begin{align*}
\frac{d^m x}{dt^m} &= \hat{f}(x, \dot{x}, \ldots) - B(x, \dot{x}, \ldots)y \\
0 &= g(x)
\end{align*}
\]

(2.3a)  
(2.3b)

with $m$ replacing 2.

2.1 Converting a DAE to a DE with invariant

It is well-known that carrying out the repeated constraint differentiation that is used to define a differential index (e.g. [26]) and eliminating the algebraic variables $y$ yields an ODE with invariant for $x$ and possibly its derivatives. For the index-2 problem, differentiate the constraints once and eliminate,

\[
y = (GB)^{-1} G \hat{f}(x)
\]

- $x(t), f(x(t))$ are vectors of length $n_x$ and $y(t), g(x(t))$ have length $n_y \leq n_x$ for each $t$.
- Denote $G = g_x$ and assume $\text{rank}(G) = n_y$ for each $t$.
- If $\exists (GB)^{-1}$ for each $t$ then the DAE has index 2.
- $\hat{f}, B$ and $g$ can depend explicitly on time $t$ – we assume they don’t for notational simplicity.
obtaining the DE with invariant

\[
\dot{x} = [I - B(GB)^{-1}G] \dot{f}(x)
\]

\[
0 = g(x)
\]

so in (1.1), (1.2): \(z \leftarrow x, \ f(z) \leftarrow [I - B(GB)^{-1}G] \dot{f}, \ h(z) \leftarrow g(x)\). Also, \(H \leftarrow g_x = G\).

For the index-3 problem, differentiate the constraints twice and eliminate,

\[
y = (GB)^{-1}[G \dot{f}(x) + \dot{x}^T g_{xx} \dot{x}] 
\]

obtaining the DE with invariant

\[
\dot{x} = [I - B(GB)^{-1}G] \dot{f}(x) - B(GB)^{-1} \ddot{x}^T g_{xx} \ddot{x}
\]

\[
0 = g(x)
\]

\[
0 = G(x) \ddot{x}
\]

so \(z \leftarrow (x, \dot{x}), \ h(z) \leftarrow (g(x), G(x) \dot{x}), \ H \leftarrow \begin{pmatrix} G & 0 \\ L & G \end{pmatrix} \quad L = \dot{x}^T g_{xx}\).

Note that in the index-2 case we have an integral invariant:

\[
Hf = 0 \quad \forall z
\] (2.4)

However, in the index-(\(m + 1\)) case, \(m \geq 2\), the condition (2.4) does not hold. In particular, for \(m = 2\),

\[
Hf = \begin{pmatrix} G(x) \dot{x} \\ 0 \end{pmatrix}
\]

**Application:** Euler-Lagrange equations for constrained multibody systems

The Euler-Lagrange equations for a multibody system with holonomic constraints is by far the most popular application of higher index DAEs (e.g. [26]), and we mention it quickly here for later reference. Consider

\[
\dot{q} = v 
\] (2.5a)

\[
M(q) \ddot{v} = \dot{f}(q, v) - G^T(q) \lambda
\] (2.5b)

\[
0 = g(q) 
\] (2.5c)

where \(q\) – generalized positions, \(v\) – generalized velocities, \(t\) – time, and \(\lambda\) – Lagrange multipliers. (NB \(q, v\) and \(\lambda\) are functions of \(t\).) Denoting \(\frac{\partial}{\partial q} = G\), assume that \(G\) has full rank and that \(M\) is symmetric positive definite.

The once-differentiated constraints on the velocity level are then

\[
\dot{g} = G(q)v = 0
\] (2.6)
Differentiating again, the constraints on acceleration level are
\[
\ddot{y} = G(q) \dot{v} + L(q, v)v = 0, \quad L = v^T g_{qq} \tag{2.7}
\]
Eliminating \(\lambda\) now gives
\[
\lambda = \Lambda(q, v) = (GM^{-1} G^T)^{-1}(GM^{-1} \dot{f} + L v) \tag{2.8}
\]
Substituting (2.8) into (2.5b) we get an ODE for \( q \) and \( v \), with an invariant manifold defined by (2.5c), (2.6).

2.2 Basic approaches for solving DAEs

Most of the many methods proposed in the literature for numerically solving DAEs divide into two or three basic approaches:

1. View the DAE as a limit of a singularly perturbed DE, e.g., replace the constraints \( g(x) = 0 \) by
\[
\epsilon E \ddot{y} = g(x)
\]
\[
or\]
\[
\epsilon E \dot{y} = g(x)
\]

\( E \) nonsingular and \( 0 < \epsilon \ll 1 \).

This includes applying stiff discretization methods to the original DAE (where \( \epsilon \) never really shows up but the approach is still stiff), as well as usual penalty and regularization techniques. A sample of works can be found in [20, 22, 23, 29, 26, 6].

The basic advantage of this approach is generality and robustness using stiff solvers.\(^3\) Very popular codes such as DASSL [29], RADAU5 [26] and others are based on this. The basic disadvantage is that the numerical approach introduces stiffness (although there are exceptions, see e.g. [6]), so methods must necessarily be implicit and stiffly stable — this may be inefficient if the original problem is nonstiff. Many DAEs arising in practice, especially in multibody systems simulation, are in fact nonstiff. Good invariant-based methods are often more effective then.

2. View the DAE as a DE on a manifold — integrate the DE “on the manifold”.

\(^3\)This is particularly important for more general forms than (2.1),(2.2), e.g. when (2.1b) is replaced by
\[
0 = g(x, y)
\]
for an index-2 DAE.
(a) In a state space formulation one transforms locally to minimal coordinates [32, 30, 35]. The basic advantage is that the obtained representation is compact, being in a minimal set of coordinates. Codes utilizing this approach have been written by Potra-Rheinboldt [30] and by Haug’s group (the commercial code DADD). The basic disadvantage is that transforming can be tricky or expensive. It appears that most practitioners prefer the next approach, which is we feel easier to implement and understand.

(b) Here one differentiates the constraints until $y$ can be eliminated, and obtains a DE with invariant, as was demonstrated above in §2.1. Then one integrates the DE using the invariant for stabilization. Among the many papers on this we mention [9, 18, 19, 5, 4, 14, 2, 17].

The basic advantage is that the method is simple to understand and implement, and that it does not introduce artificial stiffness (explicit integrators can be used). we mention the codes by von Schwerin et al [2, 10], Chin and Ascher [14], and the commercial code SD/FAST. There are many others. The basic disadvantage is that one must differentiate the constraints; also, we work with more than the minimal set of unknowns.

We now proceed with the last mentioned approach for DAEs, which brings us back to DEs with invariant manifolds.

## 3 Stabilization of invariants for the differential system

The old and popular Baumgarte’s method replaces the constraint $g(q) = 0$ in (2.5) by a linear combination of the constraint and its time derivatives:

\[
\begin{align*}
\dot{q} &= v \\
M(q)\ddot{v} &= \dot{f}(q, v) - G^T(q)\lambda \\
0 &= \ddot{g}(q) + \gamma_1 \dot{g}(q) + \gamma_2 g(q)
\end{align*}
\]

with the coefficients $\gamma_1, \gamma_2$ chosen such that $\rho(\tau) := \tau^2 + \gamma_1 \tau + \gamma_2$ has negative roots. For instance, $\gamma_1 = 2\gamma, \gamma_2 = \gamma^2, \gamma > 0$, yields a double root $-\gamma$. The equations to be solved involve a standard ODE solver and a Kuhn-Tucker matrix,

\[
\begin{pmatrix}
M & G^T \\
G & 0
\end{pmatrix}
\begin{pmatrix}
\dot{v} \\
\lambda
\end{pmatrix}
= 
\begin{pmatrix}
-Lv - \gamma_1 Gv - \gamma_2 g \\
\dot{f}
\end{pmatrix}
\]

This approach can obviously be extended to general DAEs of the form (2.3). However, it is well-known that there are practical difficulties in choosing the coefficients $\gamma$. This has been analyzed, explained and demonstrated in [5, 4, 14]. We quickly
mention here only that the extreme values of $\gamma = 0$ and $\gamma = \infty$ are both unsuitable, and that the best value for $\gamma$ usually depends on the discretization method and mesh.

Next, consider Stabilization for the index-2 DAE (2.1). Differentiating the constraint and eliminating $y$ gives

$$\dot{x} = f(x) \quad [= (I - B(GB)^{-1}G)\dot{f}]$$

It is easy to check that Baumgarte’s approach, $\dot{y} + \gamma y = 0$, is equivalent to replacing the ODE for $x$ by

$$\dot{x} = f(x) - \gamma B(GB)^{-1}g(x)$$

So, this is a stabilization of the ODE $\dot{x} = f(x)$ with respect to the invariant $g(x) = 0$.

No such relation exists for the higher-index case [5, 14]: In fact, the method of starting from the index-3 DAE, differentiating the constraints twice and stabilizing the resulting DE with respect to the invariant is better than Baumgarte’s method applied directly to the index-3 problem.

In [5] we proved the following results regarding Stabilization of invariants.

**Theorem 3.1** Consider again the DE (1.1) and the invariant set $\mathcal{M}$ defined by (1.2), and apply the stabilization

$$\dot{z} = f(z) - \gamma F(z)h(z) \quad (3.1)$$

where $F = D(HD)^{-1}$ (e.g. $D = H^T$).

1. Assume that there is a constant $\gamma_0$ such that

$$\|H(z)f(z)\|_2 \leq \gamma_0\|h(z)\|_2$$

for all $z$ in $\mathcal{M}$’s neighborhood. Then $\mathcal{M}$ is an asymptotically stable invariant manifold of (3.1) for $\gamma > \gamma_0$.

2. In particular, if $h$ is an integral invariant of ODE (1.1), i.e.

$$H(z)f(z) = 0 \quad \forall z,$$

then $\mathcal{M}$ is an asymptotically stable invariant manifold of (3.1) for any $\gamma > 0$.

The stabilized system (3.1) must be discretized in order to obtain a numerical solution. Often, a standard nonstiff integrator can be applied (see, e.g., [24]), yielding an effective algorithm. The direct discretization of the projected invariant equations, by contrast, is more complicated, even though the projected invariant approach is very robust [8]. Thus, (3.1) is preferable in many cases over both Baumgarte’s and the projected invariant approaches.
4 Stabilization of the discrete dynamical system

Despite the attraction of the stabilization (3.1), stabilization of the discretized system can be preferable. To see this, we use a simple example.

Example 4.1 For the scalar ODE with invariant

\[ \dot{z} = \psi(t) \]
\[ 0 = z - \psi(t) \]

with \( z(0) = \psi(0) \), where \( \psi \) is a given, sufficiently differentiable function, the exact solution is \( z = \psi(t) \). Baumgarte’s method, or (3.1), gives the stabilized DE

\[ \dot{z} = \dot{\psi}(t) - \gamma(z - \psi(t)) \]

Note that for \( \gamma = 0 \) the invariant is stable but not asymptotically stable, while for \( \gamma > 0 \) \( M \) is asymptotically stable, according to Theorem 3.1.

But this asymptotic stability does not necessarily guarantee a vanishing drift: Consider forward Euler with step size \( k \):

\[ z_{n+1} = z_n + k[\dot{\psi}(t_n) - \gamma(z_n - \psi(t_n))] \]

The best choice for \( \gamma \) is the one which yields no error accumulation. This is obtained by \( \gamma = 1/k \), giving

\[ z_{n+1} = \psi(t_n) + k\dot{\psi}(t_n) \]

(Note that \( \gamma \) does indeed depend on the discretization step size.) So, the drift

\[ z_{n+1} - \psi(t_{n+1}) = -\frac{k^2}{2} \ddot{\psi}(t_n) + O(k^3) \]

may not decrease and may even grow arbitrarily with \( k \) fixed, if \( \ddot{\psi} \) grows. Such is the case, for instance, for \( \psi(t) = \sin t^2 \) as \( t \) grows.

On the other hand, the post-stabilization procedure described next produces the exact solution for this simple example.

4.1 Post-stabilization

Suppose we use a one-step method of order \( p \) with a step size \( k \) for \( \dot{z} = f(z) \). Thus, if at time \( t_n \) the approximate solution is \( z_n \), application of the method gives

\[ \ddot{z}_{n+1} = \phi_k^f(z_n) \]

as the approximate solution at \( t_{n+1} = t_n + k \). (e.g. forward Euler: \( \phi_k^f(z_n) = z_n + kf(z_n) \).)
The post-stabilization approach modifies $\tilde{z}_{n+1}$ at the end of the time step to produce $z_{n+1}$, which better approximates the invariant’s equations:

$$
\tilde{z}_{n+1} = \phi^T_k(z_n)
$$

(4.1a)

$$
z_{n+1} = \tilde{z}_{n+1} - F(\tilde{z}_{n+1})h(\tilde{z}_{n+1})
$$

(4.1b)

An analysis of post-stabilization methods was carried out by H. Chin [14]. Here we only summarize his results and slightly generalize them. The 2-norm is used throughout, and smoothness and boundedness as necessary are assumed. An invariant manifold $\mathcal{M}$ defined by (1.2) is said to be decreasing in the vector field (1.1) if for any solution $z(t)$ of (1.1) and $\alpha \geq 0$,

$$
\|h(z(t + \alpha))\| \leq \|h(z(t))\|
$$

(4.2)

It is easy to see, by checking that $\frac{d}{dt} h^T h \leq 0$, that if

$$
\|H(z)f(z)\| \leq \gamma_0 \|h(z)\|
$$

(4.3)

then the decreasing property holds for the stabilized DE

$$
\dot{z} = f(z) - \gamma H^T (HHT)^{-1}(z)h(z)
$$

for any $\gamma \geq \gamma_0$. In particular, this property holds for the original system (1.1) in case of an integral invariant (for which $\gamma_0 = 0$ in (4.3)).

Consider the post-stabilization (4.1) and assume that

$$
\|I - HF\| \leq \rho < 1
$$

(4.4)

From (4.1b), by Taylor’s expansion, we have

$$
h(\tilde{z}_{n+1}) = (I - HF(\tilde{z}_{n+1}))h(\tilde{z}_{n+1}) + \Delta(\tilde{z}_{n+1})
$$

where

$$
\|H(\tilde{z}_{n+1})\| \leq M_1, \quad \|\Delta(\tilde{z}_{n+1})\| \leq M_1 \|h(\tilde{z}_{n+1})\|^2
$$

Then

$$
\|h(z_{n+1})\| \leq (\rho + M_1 \|h(\tilde{z}_{n+1})\|) \|h(\tilde{z}_{n+1})\|
$$

(4.5)

Now, let $\tau_n$ be the local truncation error,

$$
\tau_n \leq M_2 k^{\rho+1}
$$

and denote by $z(t, t_n, z_n)$ the exact solution of (1.1) satisfying

$$
z(t_n, t_n, z_n) = z_n
$$
Then
\[ \| h(\tilde{z}_{n+1}) \| \leq \| h(z(t_{n+1}, t_n, z_n)) \| + M_1 \tau_n \] (4.6)

If the decreasing property holds for (1.1) then
\[ \| h(z(t_{n+1}, t_n, z_n)) \| \leq \| h(z_n) \| \]
Substituting this into (4.6) and then into (4.5), and skipping details, we arrive at
\[ \| h(z_{n+1}) \| \leq \hat{\rho} \| h(z_n) \| + M_3 \tau_n \] (4.7)
where \( \hat{\rho} = \rho(1 + O(k)) < 1 \) for \( k \) small enough. From this we get the estimate
\[ \| h(z_n) \| \leq M k^{p+1} \quad \forall n \] (4.8)
with a moderate constant \( M \). Next we quickly obtain also a bound for the global error,
\[ z_n - z(t_n) = O(k^p) \] (4.9)

If the decreasing property does not hold for (1.1), Chin considers the auxiliary problem
\[ \dot{y} = f(y) - \gamma_0 H^T(HH^T)^{-1} h(y) \]
\[ y(t_n) = z_n \]
for which \( \mathcal{M} \) is decreasing, so \( \| h(y(t_{n+1})) \| \leq \| h(z_n) \| \). Now, (4.6) is replaced by
\[ \| h(\tilde{z}_{n+1}) \| \leq \| h(y(t_{n+1})) \| + M_4(\tau_n + \| h(z_n) \|) \leq M_5(\| h(z_n) \| + \tau_n) \]
where we have skipped details again. In place of (4.7) we get
\[ \| h(z_{n+1}) \| \leq M_5 \hat{\rho} \| h(z_n) \| + M_5 \tau_n \]
but now we are not assured that \( M_5 \hat{\rho} < 1 \), unless \( \rho = O(k) \). If \( \rho = O(k) \) then for \( k \) small enough we obtain the bounds (4.8),(4.9) as before. Moreover, \( \| h(\tilde{z}_{n+1}) \| = O(k^{p+1}) \), so inserting into (4.5) we further obtain
\[ \| h(z_n) \| \leq K(\rho k^{p+1} + k^{2(p+1)}) \] (4.10)
where the constant \( K \) depends only on the local properties of \( z(t, t_n, z_n) \) (so usual stepsize selection strategies keep it of moderate size). To summarize:

**Theorem 4.1** Assume sufficient smoothness near the compact manifold \( \mathcal{M} \) defined by (1.2). Assume that (4.4) holds as well, and that either \( \rho = O(k) \) or \( \mathcal{M} \) is decreasing in (1.1). Then the bounds (4.10) and (4.9) hold. If \( HF = I \) then
\[ \| h(z_n) \| = O(k^{2(p+1)}) \]
The post-stabilization method (4.1) with \( F = H^T(HH^T)^{-1} \) (for which obviously \( HF = I \)), coincides with one Newton step for the coordinate projection method [34, 17]. With the latter method, given as before
\[
\hat{z}_{n+1} = \phi_k^f(z_n)
\]
one determines \( z_{n+1} \) as the minimizer of \( \|z_{n+1} - \hat{z}_{n+1}\| \) such that
\[
0 = h(z_{n+1})
\]
(an energy norm can replace the 2-norm in this minimization with an obvious modification in \( F \)). In particular, the two methods almost coincide when \( k \) is very small.

For this reason, there has been a tendency in the trade to view these two methods as minor variants of each other. They are different approaches, though. Note that Theorem 4.1 implies that the first Newton iteration of the coordinate projection method is already accurate to \( O(k^{d+1}) \) and that no additional iteration is needed for maintaining this accuracy level of the invariant in later time steps. Below, we will also find it useful for efficiency reasons to consider stabilization matrices of a different form for post-stabilization.

**Example 4.2** The Modified Kepler problem [25, 12, 33] is a standard test problem.

With \( r = \sqrt{q_1^2 + q_2^3} \) a radius and the Hamiltonian
\[
e(q, v) = \frac{v_1^2 + v_2^3}{2} - \frac{1}{r} - \frac{\varepsilon}{2r^3}
\]
the differential system is
\[
\dot{q} = \frac{\partial e}{\partial v} = v
\]
\[
\dot{v} = -\frac{\partial e}{\partial q} = -\left(\frac{q}{r^3} + \frac{3\varepsilon q}{2r^5}\right) = \phi(q)
\]
The initial conditions are
\[
q_1(0) = 1 - c, \; q_2(0) = 0, \; v_1(0) = 0, \; v_2(0) = \sqrt{\frac{1 + c}{1 - c}}
\]
There are two parameters at our disposal, \( c \) and \( \varepsilon \). The invariant here is the constancy of the Hamiltonian:
\[
h(z)(t) \equiv e(q, v)(t) - e(q, v)(0) = 0 \quad \forall t
\]
In [5] we considered the unmodified problem \( \varepsilon = 0 \), for which the solution has a period \( 2\pi \). Thus, the error in the solution can be measured at integer multiples of \( 2\pi \). Results using (3.1) and \( F = H^T(HH^T)^{-1} \) are recorded in Table 3.1 of [5]. They demonstrate
a drastic improvement when using the forward Euler scheme and a rather significant improvement when using the (symplectic) midpoint scheme as well.

To these results we add here (see Table 4.1) results using post-stabilization with the same $F$ ('post-stab') and also using an explicit 2nd order Runge-Kutta scheme with and without post-stabilization (denoted 'post-stab-eRK' and 'eRK', resp.), and the projected midpoint scheme [7] ('proj-midpt') for the projected invariant formulation. All runs are with uniform time steps $k$ and $c = 0.6$. To recall, the midpoint scheme reads

$$q_{n+1} - q_n = kv_{n+1/2}$$
$$v_{n+1} - v_n = k\phi(q_{n+1/2})$$

whereas the projected midpoint scheme for the DAE

$$\dot{q} = v - \phi(q)y$$
$$\dot{v} = \phi(q) + vy$$
$$0 = h(q,v)$$

reads: First calculate $q_{n+1}, v_{n+1}$ satisfying

$$q_{n+1} - q_n = k[v_{n+1/2} - \phi(q_{n+1/2})y_{n+1/2}]$$
$$v_{n+1} - v_n = k[\phi(q_{n+1/2}) + v_{n+1/2}y_{n+1/2}]$$
$$0 = h(q_{n+1/2}, v_{n+1/2})$$

Then project:

$$q_{n+1} \leftarrow q_{n+1} - \phi(q_{n+1})\lambda, \quad v_{n+1} \leftarrow v_{n+1} + v_{n+1}\lambda$$

with $\lambda$ chosen such that

$$h(q_{n+1}, v_{n+1}) = 0$$

Note that the projected midpoint scheme has better stability properties and preserves the invariant, but the symmetry of the original ODE is lost.

We observe the second order accuracy of all methods considered and the invariant’s accuracy order $2(p + 1) = 6$ of the post-stabilization methods. The stabilization methods improve on the constant of the global error, compared to their unstabilized counterparts, but not on the order. The cheapest method here, for the given range of time integration, is the post-stabilized explicit scheme. The projected midpoint scheme is more expensive than the rest and is not worth its price, despite being most accurate for a given time step.

Note that the (symplectic) midpoint scheme loses all significant digits for $k = 0.01\pi$ before reaching $t = 50\pi$. The pointwise error does not explode, however, but remains $O(1)$. Also, the error in the Hamiltonian remains the same, depending only on $k$, not
<table>
<thead>
<tr>
<th>method</th>
<th>$k$</th>
<th>$q_2(2\pi)$</th>
<th>$q_2(4\pi)$</th>
<th>$q_2(20\pi)$</th>
<th>$q_2(50\pi)$</th>
<th>$|h|_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>midpt</td>
<td>.01π</td>
<td>.16</td>
<td>.30</td>
<td>.72</td>
<td>.10</td>
<td>.42e-2</td>
</tr>
<tr>
<td>eRK</td>
<td>.01π</td>
<td>.12</td>
<td>.18</td>
<td>.67</td>
<td>.52</td>
<td>.36e-1</td>
</tr>
<tr>
<td>post-stab-midpt</td>
<td>.01π</td>
<td>.54e-2</td>
<td>.11e-1</td>
<td>.54e-1</td>
<td>.13</td>
<td>.81e-7</td>
</tr>
<tr>
<td>post-stab-eRK</td>
<td>.01π</td>
<td>.40e-2</td>
<td>.81e-2</td>
<td>.40e-1</td>
<td>.10</td>
<td>.15e-6</td>
</tr>
<tr>
<td>proj-midpt</td>
<td>.001π</td>
<td>.16e-2</td>
<td>.32e-2</td>
<td>.16e-1</td>
<td>.40e-1</td>
<td>.42e-4</td>
</tr>
<tr>
<td>eRK</td>
<td>.001π</td>
<td>.15e-2</td>
<td>.29e-2</td>
<td>.12e-1</td>
<td>.20e-1</td>
<td>.41e-4</td>
</tr>
<tr>
<td>post-stab-midpt</td>
<td>.001π</td>
<td>.54e-4</td>
<td>.11e-3</td>
<td>.54e-3</td>
<td>.14e-2</td>
<td>.83e-13</td>
</tr>
<tr>
<td>post-stab-eRK</td>
<td>.001π</td>
<td>.40e-4</td>
<td>.81e-4</td>
<td>.40e-3</td>
<td>.10e-2</td>
<td>.86e-13</td>
</tr>
<tr>
<td>proj-midpt</td>
<td>.001π</td>
<td>.14e-4</td>
<td>.29e-4</td>
<td>.14e-3</td>
<td>.36e-3</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.1: Errors for Kepler’s problem using various 2nd order schemes

on the interval length. Calculations with the post-stabilized midpoint scheme up to $t = 2000\pi$ yield similar conclusions regarding the invariant’s error for it as well (but not for the post-stabilized explicit Runge-Kutta scheme, where a smaller step size is found necessary).

4.2 Choosing the stabilization matrix $F$

The question of choosing $F$ for the post-stabilization step (4.1b) was investigated in [4, 14] in the context of the Euler-Lagrange equations. We now extend the discussion to the $m$-th order system (2.3) in a straightforward manner.

From Theorem 4.1 it is clear that the smaller $\|I - HF\|$ is, the more effective the post-stabilization step. The choice $F = HT(HHT)^{-1}$ which was used in Example 4.2 above, or more generally the choice corresponding to one Newton step of coordinate projection $F = D(HD)^{-1}$, achieves the minimum $HF = I$. However, this choice may be expensive. In particular, for the Euler-Lagrange equations (2.5), it is desirable to avoid the complicated and expensive matrix $L$ of (2.7).

If we choose for the index-3 problem (2.2)

$$F = \begin{pmatrix}
B(GB)^{-1} & 0 \\
0 & B(GB)^{-1}
\end{pmatrix}$$

(4.11)

(or the sometimes better choice

$$F = \begin{pmatrix}
G^T(GG^T)^{-1} & 0 \\
0 & G^T(GG^T)^{-1}
\end{pmatrix}$$

which, however, requires an additional cost) then

$$HF = \begin{pmatrix}
I & 0 \\
L & I
\end{pmatrix}, \quad \bar{L} = LB(GB)^{-1}$$

14
so \( HF \neq I \). Note, however, that

\[
(I - HF)^2 = 0
\]

The effect of \( HF = I \) can therefore be achieved by applying post-stabilization with the cheap \( F \) of (4.11) twice. 

For the index-\((m + 1)\) problem (2.3) we have

\[
(I - HF)^m = 0
\]

for

\[
F = \text{diag}\{ B(GB)^{-1} \}_m
\]

so \( m \) applications per step are used. The decomposition (or “inversion”) needed for evaluating \( F \) is performed once and this is frozen for further application at the same time step (possibly a few time steps even).

The superiority of the scheme using (4.11) is demonstrated numerically in [4, 14].

The error order \( h(z_{n+1}) = O(k^{2[p+1]}) \) is clearly observed.

The application for multibody systems with holonomic constraints then reads:

1. Starting with \((q_n, v_n)\) at \( t = t_n \), use a favourite ODE integration scheme \( \phi_k \) (e.g. Runge-Kutta or multistep) to advance the system

\[
\begin{align*}
\dot{q} &= v \\
M(q)\dot{v} &= \tilde{f}(p, v) - G^T(q)\lambda \\
0 &= G(q)\dot{v} + L(q, v)v
\end{align*}
\]

by one step. Denote the resulting values at \( t_{n+1} = t_n + k \) by \((\tilde{q}_{n+1}, \tilde{v}_{n+1})\).

2. Post-stabilize:

\[
\begin{pmatrix}
q_{n+1} \\
v_{n+1}
\end{pmatrix} = \begin{pmatrix}
\tilde{q}_{n+1} \\
\tilde{v}_{n+1}
\end{pmatrix} - F(\tilde{q}_{n+1}, \tilde{v}_{n+1})h(\tilde{q}_{n+1}, \tilde{v}_{n+1})
\]

3. Set \( \tilde{q}_{n+1} \leftarrow q_{n+1}, \tilde{v}_{n+1} \leftarrow v_{n+1} \) and repeat step 2 with the same \( F \).

In case of nonholonomic constraints the DAE has index-2 and only one application of \( F \) per step is needed.

Chin and Ascher wrote a code called RKSTAB for constrained multibody systems incorporating the above post-stabilization procedure. The code modifies DOPR5 [24], which is based on the Dormand-Prince Runge-Kutta pair with local error control. The post-stabilization procedure costs a small fraction of the cost of the Dormand-Prince step. We refer the reader to [14] for further details, justification and demonstration of the effectiveness of this code.
5 Where stabilizing methods get competition

5.1 Methods which satisfy the constraints automatically

Some instances of invariants which arise in practice are such that certain Runge-Kutta discretizations retain them without any further change. The most popular of these has been the case of quadratic, integral invariants [15, 1, 16, 13].\(^4\) We now consider the case where an \(s\)-stage Runge-Kutta scheme is actually a collocation scheme, i.e. in advancing from values \(z_n\) at \(t_n\) to values \(z_{n+1}\) at \(t_{n+1}\) we collocate using a polynomial \(z_\pi(t)\) of order \(s+1\) (i.e. degree at most \(s\))^5 satisfying

\[
\begin{align*}
z_\pi(t_n) &= z_n \\
z_\pi'(t_i) &= f(z_\pi(t_i), t_i), \ i = 1, \ldots, s
\end{align*}
\]

and set \(z_{n+1} = z_\pi(t_{n+1})\). Here

\[
t_i = t_n + (t_{n+1} - t_n)c_i
\]

where \(c_i\) are the nonconfluent Runge-Kutta abscissae, assumed to have quadrature precision \(p\), \(s < p \leq 2s\). We consider a nonautonomous version of the basic problem:

\[
\begin{align*}
\dot{z} &= f(z, t) \\
0 &= h(z, t)
\end{align*}
\]

The requirement for integral invariant then reads

\[
0 = Hf + h_t \quad \forall (z, t)
\]

Now,

\[
h(z_{n+1}, t_{n+1}) - h(z_n, t_n) = \int_{t_n}^{t_{n+1}} h(\dot{z}_\pi(\tau), \tau) d\tau = \int_{t_n}^{t_{n+1}} (H \dot{z}_\pi + h_t)(\tau) d\tau
\]

Note that \(\dot{z}_\pi \in P_s\). If \(H(\dot{z}_\pi(\tau), \tau) \in P_{p-s+1}\) and \(h_t(\dot{z}_\pi(\tau), \tau) \in P_p\) then \(H \dot{z}_\pi + h_t \in P_p\) and the quadrature is exact:

\[
\int_{t_n}^{t_{n+1}} (H \dot{z}_\pi + h_t)(\tau) d\tau = k \sum_{i=1}^{a} b_i(H \dot{z}_\pi + h_t)(z_\pi(t_i), t_i)
\]

\(^4\)There are a number of recent efforts, which are not reviewed here, to extend this in various ways. For instance, see [13, 31]. Also, any Runge-Kutta method will preserve a linear, autonomous, integral invariant, such as arise for instance in linear programming methods; see e.g., [27].

\(^5\)We denote this \(z_\pi \in P_{s+1}\), where \(P_l\) stands for the class of all polynomials of degree \(< l\) over a given interval \([t_n, t_{n+1}]\).

16
At the collocation points, (5.1) is applied. So, if (5.4) holds then the integral vanishes and
\[ h(z_{n+1}, t_{n+1}) = h(z_n, t_n) \]
i.e., the invariant is preserved.

In the case of a quadratic invariant, \( h_t = 0, H \dot{z} \in P_{2q} \), so only a method with \( p = 2s \) will do. This singles out collocation at Gaussian points. We have proved:

**Theorem 5.1** Consider the DE (5.2) with invariant (5.3) satisfying (5.4), to which a collocation method of precision \( p \) is applied.

1. If \( H(z, \tau) \in P_{p-2s+1} \) and \( h_t(z, \tau) \in P_{p} \) then the invariant (5.3) is preserved at mesh points.

2. Collocation at Gaussian points, and only at Gaussian points, automatically preserves quadratic invariants.

Other, non-collocation \( s \)-stage Runge-Kutta schemes in the usual Butcher notation (e.g. [24]) must satisfy
\[ b_i a_{ij} + b_j a_{ji} = b_i b_j, \quad i, j = 1, \ldots s \]
to preserve quadratic invariants [15]. We are not aware of any such method for general problems which offers advantage over collocation at Gaussian points [3, 15, 13].

**Example 5.1** For Example 4.1 with \( \psi(t) = t^3 \), we have \( H = 1, H_t = -3t^2 \in P_3 \), so the midpoint scheme does not preserve the invariant (in fact the error grows linearly [5]), but higher order collocation methods do.

**Example 5.2** The non-dimensionalized equations in Cartesian coordinates for the simple pendulum can be written as (e.g. [26, 19])
\[ \dot{q}_1 = v_1, \quad \dot{v}_1 = -q_1 \lambda, \quad \dot{q}_2 = v_2, \quad \dot{v}_2 = -q_2 \lambda - 1 \]
\[ q_1^2 + q_2^2 = 1 \]
This is perhaps the simplest instance of (2.5). Note that \( g(q) \) appears in quadratic form.

Carrying out the elimination (2.8) yields the ODE
\[ \dot{q}_1 = v_1, \quad \dot{v}_1 = \frac{-q_1}{q_1^2 + q_2^2}(v_1^2 + v_2^2 - q_2) \]
\[ \dot{q}_2 = v_2, \quad \dot{v}_2 = \frac{-q_2}{q_1^2 + q_2^2}(v_1^2 + v_2^2 - q_2) - 1 \]
with the invariant
\[ q_1^2 + q_2^2 = 1 \]
\[ q_1 v_1 + q_2 v_2 = 0 \]

Note that the second of these equations alone defines an invariant manifold as well.

Now, when we apply the midpoint scheme to this ODE we observe that the second invariant equation, \( \dot{g}(q) = 0 \), holds at mesh points, whereas the original constraint \( g(q) = 0 \) does not, even though both constraints are quadratic. This is because the integral invariant condition (5.4) holds for \( q_1 v_1 + q_2 v_2 = 0 \) but not for \( q_1^2 + q_2^2 = 1 \). (See the comment following (2.4).)

**Example 5.3** Consider the matrix differential system

\[ \dot{U} = A(U, t)U \] (5.5)

where \( A \) and \( U \) are \( n_u \times n_u \), \( n_z = n_u^2 \) and \( A \) is skew-symmetric for all \( U, t \):

\[ A^T = -A \]

If \( \dot{V} = -VA \), where \( V \) has the same dimensions as \( U \), then \( V(t)U(t) \) is clearly independent of \( t \), because \( \dot{V}U + V\dot{U} = 0 \). Now, skew-symmetry allows one to identify \( V \) with \( UT \), so

\[ UTU = \text{constant} \]

Hence, starting from

\[ UTU(0) = I \]

(e.g. \( U(0) = I \)) yields an orthogonal matrix function \( U(t) \) for all \( t \). The next question is regarding the preservation of this orthogonality in the numerical integration. This is treated in [16].

In fact, the orthogonality condition

\[ UTU(t) - UTU(0) = 0 \] (5.6)

is a quadratic integral invariant of (5.5). Hence, by theorem 5.1 collocation at Gaussian points automatically preserves the condition (5.6).

Collocation at Gaussian points is a relatively expensive, fully implicit Runge-Kutta method, though. In [16] the authors consider, in addition to it, a post-stabilization technique where after each step of a non-special (e.g. explicit) order-\( p \) discretization of (5.5), producing say \( \dot{U}_{n+1} \), a modified Gram-Schmidt algorithm is applied to recover orthogonality, say

\[ \dot{U}_{n+1} = \dot{U}_{n+1}R_{n+1} \]

where \( R_{n+1} \) is upper triangular and \( U_{n+1} \) is orthogonal (see also [28]). Writing the orthogonality conditions (5.6) in vector form,

\[ h(z) : UT_j = \delta_{ij}, \quad i = 1, \ldots, n_u, \quad j = 1, \ldots, i \]
where $U_j$ is the $j$th column of $U$ (note that the Jacobian matrix $H$ is $\frac{1}{2}n_u(n_u + 1) \times n_u^2$), it can be verified that this modified Gram-Schmidt procedure is equivalent to a particular choice of $F$ in (4.1) which satisfies

$$HF = I + O(k^{p+1})$$

Using Theorem 4.1 we therefore obtain global convergence:

$$U_n - U(t_n) = O(k^p)$$

A convincing numerical demonstration of this method is given in [16].

**Example 5.4** A number of interesting applications lead to problems of isospectral flow [13], where one seeks a matrix function satisfying

$$\begin{align*}
\dot{L} &= AL - LA \\
L(0) &= L_0
\end{align*}$$

(5.7)

for a given initial value matrix $L_0$, where $A(L)$ is again skew-symmetric.

It can be easily verified that

$$L = UL_0U^T$$

where $U(t)$ is the orthogonal matrix function satisfying

$$\dot{U} = AU, \quad U(0) = I$$

Hence this application is a special case of the one considered in the previous example.

### 5.2 On symplectic methods

A good post-stabilization method seems preferable in many cases to collocation at Gaussian points if the initial value problem to be integrated is nonstiff. However, Gaussian collocation is symplectic [24, 12, 25], and stabilized methods generally are not (even if the stabilization is applied to a symplectic method). Of course, it is not reasonable to expect of any discretization to reproduce all properties of the exact solution of a given DE. What one needs to do is identify the important properties and the cost for maintaining them at the discrete level. Here, we next address a case where $nk$ is so large that no pointwise accuracy is left in the approximate solution (i.e. the pointwise error norm is of the order of the solution norm). The question is, are other important features of the differential system still decently reproduced? We only consider one example below.

**Example 5.5** We revisit the modified Kepler problem, Example 4.2 [25, 12, 33]. We take $c = 0.6, \quad \varepsilon = 0.01$. In Figure 5.1 we plot $q_2$ vs $q_1$ for $0 \leq t \leq t_f = 500$. The corresponding figures using the midpoint scheme and the projected midpoint scheme
for the projected invariant formulation with 5000 uniform steps are plotted in Figure 5.2 (cf. [25]).

Repetition of this experiment with 10000 uniform steps confirms that the midpoint scheme retains a much less noisy structure than the projected midpoint scheme, even though the latter is pointwise more accurate in the early goings. The results using the projected scheme look noisy. The symplectic midpoint scheme clearly reproduces the feature that the solution of this dynamical system lies on a torus. The radius of this torus is off by about 10%, though (Table 5.1). Worse, the “physically looking” midpoint solution pattern is inaccurate, and this requires a user to be particularly astute in order not to be misled.

Still, it can be claimed that the unstabilized but symplectic scheme has an advantage. It is interesting to note that the advantage of the symplectic scheme does not come through by merely looking at the data in Table 5.1. Also, as is well-known, the symplectic structure disappears when a sequence of random steps is used in the numerical integration, see Figure 5.3. The projected scheme is better then. We refer to [25] for a method with a particular nonuniform stepsize selection which retains time-reversibility for a symmetric scheme.

Acknowledgment I am grateful to E. Hairer, A. Iserles, P. Lin and S. Reich for comments, advices and stimulating discussions.
Figure 5.2: Modified Kepler solutions, uniform steps
Midpoint, 10000 random steps

Projected Midpoint, 10000 random steps

Figure 5.3: Modified Kepler solutions, nonuniform steps
\begin{center}
\begin{tabular}{|l|l|c|c|}
\hline
method & steps & uniform? & $eh$ & ro \\
\hline
midpl & 25000 & uniform & .049 & 1.16 \\
proj-midpl & 25000 & uniform & $1.0^{-7}$ & 1.3015 \\
midpl & 50000 & random & .154 & 1.26 \\
midpl & 50000 & random & .160 & 1.82 \\
proj-midpl & 50000 & random & $1.0^{-7}$ & 1.306 \\
proj-midpl & 50000 & random & $1.0^{-7}$ & 1.309 \\
midpl & 10000 & uniform & DIVERGED & \\
proj-midpl & 10000 & uniform & $1.0^{-7}$ & 1.25 \\
\hline
\end{tabular}
\end{center}

Table 5.1: errors in invariant (eh) and max$\|q\|$ (ro) (exact $\approx 1.313$) for $t_f = 2500$

References


strained multibody systems. Technical report, IWR Universitat Heidelberg, June
1993.


[14] H. Chin. Stabilization methods for simulations of constrained multibody dy-
namics. PhD thesis, Institute of Applied Mathematics, University of British


to continuous orthonormalization techniques. SIAM J. Num. Anal., 31:261–281,
1994.

mechanical systems with algebraic constraints. SIAM J. Numer. Anal., 30:1467,
1993.

of Technology, 1990.


[20] C.W. Gear. The simultaneous numerical solution of differential-algebraic equa-


[22] C.W. Gear. Differential-algebraic equations, indices and integral algebraic equa-

In Proc. ODE Meeting, Oberwolfach, West Germany, 1981.

24


