BIT 37:1 (1997), 24–36.

TIME-STEPPING AND PRESERVING ORTHONORMALITY [*]

DESMOND J. HIGHAM [†]

Department of Mathematics, University of Strathclyde, Livingstone Tower
Glasgow, G1 1XH, U.K. email: dhigham@na-net.ornl.gov

Abstract.

Certain applications produce initial value ODEs whose solutions, regarded as time-dependent matrices, preserve orthonormality. Such systems arise in the computation of Lyapunov exponents and the construction of smooth singular value decompositions of parametrized matrices. For some special problem classes, there exist time-stepping methods that automatically inherit the orthonormality preservation. However, a more widely applicable approach is to apply a standard integrator and regularly replace the approximate solution by an orthonormal matrix. Typically, the approximate solution is replaced by the factor Q from its QR decomposition (computed, for example, by the modified Gram-Schmidt method). However, the optimal replacement—the one that is closest in the Frobenius norm—is given by the orthonormal polar factor. Quadratically convergent iteration schemes can be used to compute this factor. In particular, there is a matrix multiplication based iteration that is ideally suited to modern computer architectures. Hence, we argue that perturbing towards the orthonormal polar factor is an attractive choice, and we consider performing a fixed number of iterations. Using the optimality property we show that the perturbations improve the departure from orthonormality without significantly degrading the finite-time global error bound for the ODE solution. Our analysis allows for adaptive time-stepping, where a local error control process is driven by a user-supplied tolerance. Finally, using a recent result of Sun, we show how the global error bound carries through to the case where the orthonormal QR factor is used instead of the orthonormal polar factor.

AMS subject classification: 65L05.

Key words: Error control, variable time-step, Lyapunov exponents, global error, polar decomposition, nearest orthonormal matrix.

1 Introduction.

This work applies some linear algebra ideas in an ordinary differential equation (ODE) context. We begin by summarising the appropriate linear algebra, and then we introduce the ODE problem.

[*] Received July 1995. Revised July 1996.

[†] This work was supported by Engineering and Physical Sciences Research Council grants GR/H94634 and GR/K80228.

Suppose that $A \in \mathbb{R}^{m\times p}$, with $m \geq p$, has full rank. Then A has a unique polar decomposition

$$A = UH, \quad U \in \mathbb{R}^{m\times p},\ U^TU = I,$$

$H \in \mathbb{R}^{p\times p}$ symmetric positive definite, and a unique QR decomposition

$$A = QR, \quad Q \in \mathbb{R}^{m\times p},\ Q^TQ = I,$$

$R \in \mathbb{R}^{p\times p}$ upper triangular with positive diagonal. We refer to U as the orthonormal polar factor and to Q as the orthonormal QR factor.

Now, consider approximating A by an orthonormal matrix, measuring perturbations in the Frobenius norm, defined by $\|A\|_F = \mathrm{trace}(A^TA)^{1/2}$. We let $\gamma(A)$ denote the corresponding departure from orthonormality, that is,

$$\gamma(A) := \min\{\,\|E\|_F : E \in \mathbb{R}^{m\times p},\ (A+E)^T(A+E) = I\,\}. \tag{1.1}$$

It is known (see, for example, [11]) that $\gamma(A) = \|U - A\|_F$; in other words, the orthonormal polar factor is a best orthonormal approximation. A useful order-of-magnitude estimate of $\gamma(A)$ can be found using the inequalities

$$\frac{\|A^TA - I\|_F}{\|A\|_2 + 1} \leq \gamma(A) \leq \|A^TA - I\|_F \tag{1.2}$$

from [11].

It is possible to construct the orthonormal polar factor U from a singular value decomposition of A, but for the application described here a more efficient approach is to use a matrix iteration. If $m = p$ and A is nonsingular, then the sequence generated by

$$\begin{aligned} Y_0 &= A, \\ Y_{i+1} &= \frac{1}{2}(Y_i + Y_i^{-T}), \quad i = 0,1,2\ldots \end{aligned}$$

converges quadratically to U [10]. We refer to this as the Newton iteration, since it can be obtained by applying Newton's Method to $Y^TY = I$. For $m > p$, if A has full rank then the Newton iteration can be applied after an initial QR decomposition: with $A = QR$, if R has the polar decomposition $R = U_RH_R$ then A has the polar decomposition $A = (QU_R)H_R$.

An alternative iteration that works for any $m \geq p$ is the Schulz iteration [13]

$$\begin{aligned} Y_0 &= A, \\ Y_{i+1} &= Y_i\left(I + \frac{1}{2}(I - Y_i^TY_i)\right), \quad i = 0,1,2\ldots. \end{aligned}$$

This iteration is locally quadratically convergent, and a sufficient condition for convergence is $\|A^TA - I\|_2 < 1$.

Now, consider the matrix ODE

$$X'(t) = F(t, X(t)), \quad t > 0, \tag{1.3}$$

where the initial value $X(0) = X_0 \in \mathbb{R}^{m\times p}$ is given. We introduce the solution operator $S(t, \Delta t, Y)$ that advances the flow (1.3) from $X(t) = Y$ over time Δt; formally, for $t > 0$, $\Delta t > 0$ and $Y \in \mathbb{R}^{m\times p}$, we have

$$S(t, \Delta t, Y) = X(t + \Delta t),$$

where X is the solution of (1.3) with $X(t) = Y$.

We assume that $X_0^T X_0 = I$, and that the problem has the property of preserving orthonormality:

$$\begin{aligned} &X(t)^T X(t) = I \text{ for some } t > 0 \\ \Longrightarrow \quad & S(t, \Delta t, X(t))^T S(t, \Delta t, X(t)) = I \text{ for all } \Delta t > 0. \end{aligned} \tag{1.4}$$

The main application that we have in mind is the computation of Lyapunov exponents of a general ODE by a discrete or continuous QR algorithm [3, 2, 4, 14]. In this case, the columns of $X(t)$ represent solutions of the same ODE arising from different initial values. However, only the property (1.4) will be needed here. For certain problem subclasses that satisfy (1.4) it is possible to show that some discretisation methods share the property of preserving orthonormality. However, such methods are expensive to implement, and, furthermore, there are known to be problem classes where no standard discretization method can preserve orthonormality [4]. An attractive alternative is to use a *projected integrator*. The idea is to apply a conventional discretization method and, after each time-step, to replace the approximate solution by a matrix with orthonormal columns. The standard choice is to replace the matrix by its orthonormal QR factor, computed by the modified Gram-Schmidt method.

2 New method and error analysis.

Since the orthonormal polar factor gives an optimal perturbation in (1.1), it is natural to consider using this matrix, rather than the orthonormal QR factor, in the projected integrator. More generally, we could perform a fixed number of iterations of the Newton or Schulz iteration. In this section, we analyse the effect of the projection on the global error and the departure from orthonormality. We are concerned with finite-time error bounds, and hence throughout this section we assume that the problem (1.3) is to be solved over a finite interval $[0, T]$ with a fixed initial condition X_0.

To proceed with the analysis we must introduce some notation and state our basic assumptions. We suppose that any conventional one-step ODE method with local error control is used; for example, an (explicit or implicit) Runge-Kutta pair. We let $P^{[0]}(t, \Delta t, Y)$ denote the basic time-stepping operator; that is, $P^{[0]}(t, \Delta t, Y)$ is the result of applying the one-step method to $Y \approx X(t)$ with time-step Δt. (Note that in the case of implicit Runge-Kutta methods, it may be necessary to restrict Δt and Y in order for $P^{[0]}(t, \Delta t, Y)$ to be uniquely defined. It is clear that this could easily be accommodated in the analysis below, so, for simplicity, we do not mention this further.) Similarly, we let $P^{[k]}(t, \Delta t, Y)$ denote the result of applying the one-step method followed by k iterations of either the Newton or Schulz iteration. We also let $P^{[\infty]}(t, \Delta t, Y)$ denote the result of iterating $P^{[0]}(t, \Delta t, Y)$ to convergence with either the Newton or Schulz iteration. The discrete sequence that results from the "time-step plus iterate" process is denoted $\{X_n^{[k]}\}$. More precisely, we assume that the time-step selection and local error control procedure is based on the standard one-step method, so that a complete step has the following form. Given $X_n^{[k]} \approx X(t_n)$ and Δt_n,

- Compute $P^{[0]}(t_n, \Delta t_n, X_n^{[k]})$.
- Compute the local error estimate.
- If the local error estimate is too big, reject the step and repeat with a smaller time-step.
- Otherwise, compute the time-step Δt_{n+1} ready for use on the next step, and generate $X_{n+1}^{[k]}$ by applying k iterations of the Newton or Schulz iteration to $P^{[0]}(t_n, \Delta t_n, X_n^{[k]})$.

The local error control and time-step selection is driven by a user-supplied tolerance $\tau > 0$, a smaller τ indicating that greater accuracy is desired. We refer to [6, section II.4] for implementation details.

The **local error** committed by the basic ODE method $P^{[0]}$ over a step from $Y \approx X(t)$ to $P^{[0]}(t, \Delta t, Y) \approx X(t + \Delta t)$ is defined to be

$$P^{[0]}(t, \Delta t, Y) - S(t, \Delta t, Y).$$

We follow the approach of [9] by making simple, realistic assumptions about the effect of the local error control strategy used by the numerical method. This allows us to draw conclusions that are independent of the particular details of the method. We point out that [9] analyses errors in the continuous realm, whereas in this work we find it more convenient to work directly with the discrete approximants.

In the following, given $\delta > 0$ we define the tube

$$\mathcal{B}_\delta(t) := \{Y : Y = X(t) + E, \quad \|E\|_F \le \delta, \quad 0 \le t \le T\}.$$

ASSUMPTION 2.1. *There exists $\tau^* > 0$ such that for any $0 < \tau \le \tau^*$ a solution sequence $\{X_n^{[k]}\}_{n=0}^N$ and a time-step sequence $\{\Delta t_n\}_{n=0}^{N-1}$ are computed, with $\sum_{n=0}^{N-1} \Delta t_n = T$. (Note that the dependence of $X_n^{[k]}$, Δt_n and N upon τ is suppressed in this notation.) Further, for any $0 < \tau \le \tau^*$ the following conditions hold. First, letting $\Delta t_{\max} := \max_n\{\Delta t_n\}$,*

$$\Delta t_{\max} \to 0 \quad as \quad \tau \to 0 \quad and \quad N\Delta t_{\max} \le C, \tag{2.1}$$

for some constant C. Second, there exists a constant D such that

$$\tau \le D\Delta t_n, \quad for\ all \quad n \ge 0. \tag{2.2}$$

Third, there exist constants $K > 0$ and $\delta^* > 0$ such that the local error control method ensures that

$$\|P^{[0]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]})\|_F \leq K\Delta t_n \tau, \quad \textit{for all} \quad X_n^{[k]} \in \mathcal{B}_{\delta^*}(t_n). \tag{2.3}$$

The conditions (2.1)–(2.3) can be justified for smooth problems by asymptotic (small Δt) expansions. For (2.3), the method must use error-per-unit-step control or extrapolated error-per-step control—see, for example, [9].

We also make an assumption about F.

ASSUMPTION 2.2. *The function F in (1.3) is continuous in all variables and locally Lipschitz; that is, for any bounded set C there exists a constant L, depending upon C, such that*

$$\|F(t,X) - F(t,Y)\|_F \leq L\|X - Y\|_F, \quad \textit{for all} \quad 0 \leq t \leq T, \quad X, Y \in C.$$

We define $e_n^{[k]} := \|X_n^{[k]} - X(t_n)\|_F$ and $\gamma_n^{[k]} := \gamma(X_n^{[k]})$, which represent the global error and the departure from orthonormality of the $X_n^{[k]}$ values, respectively. The analysis in [9] shows that the basic ODE method, under Assumptions 2.1–2.2, produces a global error that can be bounded by a constant multiple of τ. It follows trivially that the corresponding departure from orthonormality satisfies the same bound. Our aim here is to investigate the global error and the departure from orthonormality of the partially projected values $\{X_n^{[k]}\}$, for fixed $k > 0$. Intuitively, we would expect the Newton or Schulz iterations to give an improvement over the departure from orthonormality that would arise with the unprojected method. Perhaps the key difficulty in the analysis is that the values $\{X_n^{[k]}\}$ are not exactly orthonormal, so the property (1.4) cannot be immediately invoked. Our approach is to set up a recurrence for the departures, $\gamma_n^{[k]}$, in addition to the usual recurrence that arises for the global error, $e_n^{[k]}$.

We begin by specifying two basic results.

RESULT 2.1. *There exist constants $\delta^*, \Delta t^* > 0$ and $K_1 > 0$ such that for $0 < \Delta t \leq \Delta t^*$ and $Y \in \mathcal{B}_{\delta^*}(t)$ we have*

$$\begin{aligned}
\gamma(P^{[k]}(t,\Delta t, Y)) &\leq K_1 \gamma(P^{[0]}(t,\Delta t,Y))^{2k}, \\
\|P^{[k]}(t,\Delta t,Y) - P^{[\infty]}(t,\Delta t,Y)\|_F &\leq K_1 \gamma(P^{[0]}(t,\Delta t,Y))^{2k}, \\
\|P^{[k]}(t,\Delta t,Y) - P^{[0]}(t,\Delta t,Y)\|_F &\leq 2\|P^{[\infty]}(t,\Delta t,Y) - P^{[0]}(t,\Delta t,Y)\|_F.
\end{aligned}$$

PROOF. The first two inequalities are direct consequences of the quadratic convergence of the iteration and the third follows from the triangle inequality. □

RESULT 2.2. *There exist constants $\delta^*, \Delta t^* > 0$ and $L > 0$ such that*

$$\|S(t,\Delta t, X) - S(t,\Delta t,Y)\|_F \leq (1 + L\Delta t)\|X - Y\|_F,$$

for all $0 < \Delta t \leq \Delta t^*$ and $X, Y \in \mathcal{B}_{\delta^*}(t)$.

PROOF. Choose any $\delta^* > 0$. Suppose $0 \leq t \leq T$ and $X, Y \in \mathcal{B}_{\delta^*}(t)$. Using the local Lipschitz property (Assumption 2.2) it follows from a standard stability bound [6, Theorem 10.2] that

$$\|S(t, \Delta t, X) - S(t, \Delta t, Y)\|_F \leq \exp(\widehat{L}\Delta t)\|X - Y\|_F, \quad \text{for all} \quad X, Y \in \mathcal{B}_{\delta^*}(t),$$

where $\widehat{L}$ is the appropriate local Lipschitz constant. (Note that we may regard (1.3) as a vector system, with the columns of $X(t)$ stacked into a vector in $\mathbb{R}^{mp}$. The result for this vector problem using $\|\cdot\|_2$ converts into a result for the matrix problem using $\|\cdot\|_F$.) Now by restricting $0 < \Delta t \leq \Delta t^*$ with $\widehat{L}\Delta t^* < .5$ we have $\exp(\widehat{L}\Delta t) < 1 + 2\widehat{L}\Delta t$. The result follows with $L = 2\widehat{L}$. □

In the following, we suppose that δ^* and τ^* are chosen so that the above assumptions and results are valid. (Note that by restricting τ^* we can ensure $\Delta t_n \leq \Delta t^*$.) Also, without comment we use the symbols K_2, K_3, K_4, K_5 to denote constants (independent of n and τ).

LEMMA 2.1. *If $X_n^{[k]} \in \mathcal{B}_{\delta^*}(t_n)$ for all $n \geq 0$ and all $0 < \tau \leq \tau^*$, then*

$$\gamma(P^{[0]}(t_n, \Delta t_n, X_n^{[k]})) \leq K\Delta t_n \tau + (1 + L\Delta t_n)\gamma_n^{[k]}, \tag{2.4}$$

and

$$\gamma_{n+1}^{[k]} \leq K_1\{K\Delta t_n \tau + (1 + L\Delta t_n)\gamma_n^{[k]}\}^{2k}. \tag{2.5}$$

PROOF. Let $X_n^{[k]*}$ be a nearest orthonormal matrix to $X_n^{[k]}$. Then, from the property (1.4) of the ODE, $S(t_n, \Delta t_n, X_n^{[k]*})$ is also orthonormal, so that

$$\begin{aligned}\gamma(P^{[0]}(t_n, \Delta t_n, X_n^{[k]})) &\leq \|P^{[0]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]*})\|_F \\ &\leq \|P^{[0]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]})\|_F \\ &+ \|S(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]*})\|_F.\end{aligned}$$

Now the first term on the right hand side can be bounded using the error control assumption (2.3) and the second term can be bounded using Result 2.2, to give (2.4). Finally, using Result 2.1

$$\gamma_{n+1}^{[k]} := \gamma(P^{[k]}(t_n, \Delta t_n, X_n^{[k]})) \leq K_1\gamma(P^{[0]}(t_n, \Delta t_n, X_n^{[k]}))^{2k}.$$

Inserting (2.4) completes the proof. □

The next lemma establishes the convergence of $\gamma_n^{[k]}$. The bound is then refined in Lemma 2.3.

LEMMA 2.2. *If $X_n^{[k]} \in \mathcal{B}_{\delta^*}(t_n)$ for all $n \geq 0$ and all $0 < \tau \leq \tau^*$, then, by further reduction of τ^* if necessary, we have*

$$\gamma_n^{[k]} \leq K_2\tau, \quad \textit{for all} \quad n \geq 0, \quad 0 < \tau \leq \tau^*.$$

PROOF. By reducing τ^* in Lemma 2.1, if necessary, we can ensure that $\tau^* < 1$. Now assume that $\gamma_n^{[k]}$ is sufficiently small for

$$\gamma_n^{[k]} \leq 1 \quad \text{and} \quad (1 + L\Delta t_n)^{2k-1}\left(\gamma_n^{[k]}\right)^{2k-1} \leq \frac{1}{K_1}. \tag{2.6}$$

It then follows from (2.5) in Lemma 2.1 that

$$\gamma_{n+1}^{[k]} \leq K_1\{K\Delta t_n \tau + (1 + L\Delta t_n)\gamma_n^{[k]}\}^{2k} \leq K_3\Delta t_n \tau + (1 + L\Delta t_n)\gamma_n^{[k]}. \tag{2.7}$$

The proof can be completed with an application of the discrete Gronwall Lemma; see, for example, [7, pages 18–19]. □

LEMMA 2.3. *Under the assumptions of Lemma* 2.2,

$$\gamma_n^{[k]} \leq K_4 \tau^{2k}, \quad \textit{for all} \quad n \geq 0, \quad 0 < \tau \leq \tau^*.$$

PROOF. Applying Lemma 2.2 in (2.5) of Lemma 2.1, we have

$$\gamma_{n+1}^{[k]} \leq K_1\{K\Delta t_n \tau + (1 + L\Delta t_n)K_2\tau\}^{2k} \leq K_4\tau^{2k}.$$

□

LEMMA 2.4. *Under the assumptions of Lemma* 2.2,

$$\begin{aligned} \|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[k]})\|_F \quad &\leq \quad K_5 \Delta t_n \tau, \\ \textit{for all} \qquad & n \geq 0, \quad 0 < \tau \leq \tau^*. \end{aligned}$$

PROOF. Using Result 2.1

$$\begin{aligned} &\|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[k]})\|_F \\ &\qquad \leq 2\|P^{[\infty]}(t_n, \Delta t_n, X_n^{[k]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[k]})\|_F, \end{aligned}$$

which can be written

$$\|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[k]})\|_F \leq 2\gamma(P^{[0]}(t_n, \Delta t_n, X_n^{[k]})).$$

Now, applying (2.4) of Lemma 2.1 and Lemma 2.3,

$$\begin{aligned} \gamma(P^{[0]}(t_n, \Delta t_n, X_n^{[k]})) \quad &\leq \quad K\Delta t_n \tau + (1 + L\Delta t_n)\gamma_n^{[k]} \\ &\leq \quad K\Delta t_n \tau + (1 + L\Delta t_n)K_4\tau^{2k}. \end{aligned}$$

Now (2.2) completes the proof. □

LEMMA 2.5. *Under the assumptions of Lemma* 2.2,

$$e_{n+1}^{[k]} \leq K_5 \Delta t_n \tau + (1 + L\Delta t_n)e_n^{[k]}, \quad \textit{for all} \quad n \geq 0, \quad 0 < \tau \leq \tau^*.$$

PROOF. From $e_{n+1}^{[k]} := \|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X(t_n))\|_F$ we have

$$\begin{aligned} e_{n+1}^{[k]} \leq\ & \|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[k]})\|_F \\ & + \|P^{[0]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]})\|_F \\ & + \|S(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X(t_n))\|_F. \end{aligned}$$

We can bound the first, second and third terms on the right hand side by using Lemma 2.4, the error control assumption (2.3) and Result 2.2, respectively, to give the required result. □

Lemma 2.5 is the key relation that allows us to bound the global error. The final result is given the following theorem.

THEOREM 2.6. *There exist constants $\widehat{\tau} > 0$ and $C_1, C_2 > 0$ such that*

$$\left. \begin{aligned} e_n^{[k]} &\leq C_1 \tau \\ \gamma_n^{[k]} &\leq C_2 \tau^{2k} \end{aligned} \right\} \quad \textit{for all} \ \ n \geq 0, \quad 0 < \tau \leq \widehat{\tau}.$$

PROOF. Recall that Lemma 2.5 requires $X_n^{[k]} \in \mathcal{B}_{\delta^*}(t_n)$ for all $n \geq 0$ and all $0 < \tau \leq \tau^*$. Assume for the moment that this condition holds. Using "Gronwall" analysis, as in the proof of Lemma 2.2, we obtain

$$e_n^{[k]} \leq C_1 \tau.$$

Now, by further reduction of τ^*, if necessary, we can ensure that $X_n^{[k]} \in \mathcal{B}_{\delta^*}(t_n)$ for all $n \geq 0$ and all $0 < \tau \leq \tau^*$, as required. We may also invoke Lemma 2.3 to obtain the bound for $\gamma_n^{[k]}$. □

3 Discussion and further analysis.

How do the results in Theorem 2.6 compare with those for the unprojected method? Under Assumptions 2.1 and 2.2, the discrete Gronwall approach from the previous section (or the continuous analogue in [9]) can be used to show that the global error for the unprojected method satisfies a bound of the same form:

$$e_n^{[0]} \leq C_3 \tau, \ \ \text{for all} \ \ n \geq 0, \quad 0 < \tau \leq \bar{\tau}. \tag{3.1}$$

Since $X(t_n)$ is a candidate for the nearest orthonormal matrix, we can deduce from (3.1) that $\gamma_n^{[0]} \leq C_3 \tau$.

It is natural to ask how the constant C_3 in (3.1) compares with that for the projected method in Theorem 2.6. To this end, we remark that it is particularly simple to analyse the case where the Newton or Schulz iteration is continued to convergence ($k = \infty$). Here, since $X_n^{[\infty]}$ is orthonormal, we have

$$\begin{aligned} & \|P^{[\infty]}(t_n, \Delta t_n, X_n^{[\infty]}) - S(t_n, \Delta t_n, X_n^{[\infty]})\|_F \\ & \qquad \leq \|P^{[\infty]}(t_n, \Delta t_n, X_n^{[\infty]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[\infty]})\|_F \\ & \qquad\quad + \|P^{[0]}(t_n, \Delta t_n, X_n^{[\infty]}) - S(t_n, \Delta t_n, X_n^{[\infty]})\|_F \\ & \qquad \leq 2\|P^{[0]}(t_n, \Delta t_n, X_n^{[\infty]}) - S(t_n, \Delta t_n, X_n^{[\infty]})\|_F, \end{aligned}$$

using the optimality of the orthonormal polar factor $P^{[\infty]}(t_n, \Delta t_n, X_n^{[\infty]})$. Hence, with (2.3) of Assumption 2.1, we have

$$\|P^{[\infty]}(t_n, \Delta t_n, X_n^{[\infty]}) - S(t_n, \Delta t_n, X_n^{[\infty]})\|_F \leq 2K\Delta t_n \tau. \tag{3.2}$$

So the local-error-per-unit-step for the $k = \infty$ process satisfies a bound that is twice the bound for the unprojected version. It follows that the global error satisfies $e_n^{[\infty]} \leq 2C_3\tau$, where C_3 is the constant in (3.1). Using a similar argument, we can show that for any $k > 0$ the global error bound has essentially the same constant. This follows from

$$\begin{aligned}
&\|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]})\|_F \\
&\qquad \leq \|P^{[k]}(t_n, \Delta t_n, X_n^{[k]}) - P^{[\infty]}(t_n, \Delta t_n, X_n^{[k]})\|_F \\
&\qquad\qquad + \|P^{[\infty]}(t_n, \Delta t_n, X_n^{[k]}) - S(t_n, \Delta t_n, X_n^{[k]})\|_F \\
&\qquad \leq \gamma_{n+1}^{[k]} + 2K\Delta t_n \tau.
\end{aligned}$$

Since $\gamma_{n+1}^{[k]} = O(\tau^{2k})$, this local-error-per-unit-step bound, and hence the resulting global error bound, is essentially the same for general $k > 0$ as for the $k = \infty$ case in (3.2).

In summary, for any $k > 0$ the projected method has (a) a global error bound that is no more than a factor two bigger than the unprojected method and (b) a departure from orthonormality that is $O(\tau^{2k})$. It is worth emphasising that we have compared global error *bounds* only, and the analysis was based on a local Lipschitz condition (Assumption 2.2). For problems with special structure, it may be possible to exploit the smallness of $\gamma_n^{[k]}$ to refine the error analysis for the projected methods.

Next we show that it is possible to analyse the projected integrator that uses the orthonormal QR factor of $P^{[0]}$ rather than the orthonormal polar factor. We make use of the following result of Sun [18, Lemma 2.4]. This lemma compares the approximating power of the orthonormal QR factor with that of the orthonormal polar factor.

LEMMA 3.1 (SUN, 1995). *Let $A \in \mathbb{R}^{m\times p}$ have full rank. Let U and Q denote the orthonormal polar and QR factors of A, respectively, as described in Section* 1. *Then if $\|A^TA - I\|_2 < 1$, we have*

$$\|A - Q\|_F \leq \frac{(1 + \|A\|_2)}{\sqrt{2}\,(1 - \|A^TA - I\|_2)}\|A - U\|_F.$$

We let $X_n^{[\mathrm{QR}]}$ denote the result of projecting onto the orthonormal QR factor (rather than iterating towards the polar factor). Thus, $X_{n+1}^{[\mathrm{QR}]}$ is the orthonormal QR factor of $P^{[0]}(t_n, \Delta t_n, X_n^{[\mathrm{QR}]})$. We use $P^{[\mathrm{QR}]}$ to denote the corresponding operator, so that $X_{n+1}^{[\mathrm{QR}]} = P^{[\mathrm{QR}]}(t_n, \Delta t_n, X_n^{[\mathrm{QR}]})$.

Now, since $X_n^{[\mathrm{QR}]}$ is orthonormal, the error control assumption (2.3) gives

$$\begin{aligned}
\gamma\left(P^{[0]}(t_n, \Delta t_n, X_n^{[\mathrm{QR}]})\right) &\leq \|P^{[0]}(t_n, \Delta t_n, X_n^{[\mathrm{QR}]}) - S(t_n, \Delta t_n, X_n^{[\mathrm{QR}]})\|_F \\
&\leq K\Delta t_n \tau.
\end{aligned}$$

So, from (1.2),

$$P^{[0]}(t_n, \Delta t_n, X_n^{[QR]})^T P^{[0]}(t_n, \Delta t_n, X_n^{[QR]}) = I + O(\tau \Delta t_n).$$

Hence, using $\|A\|_2^2 = \|A^T A\|_2$, we have

$$\|P^{[0]}(t_n, \Delta t_n, X_n^{[QR]})\|_2 = 1 + O(\tau \Delta t_n).$$

Thus, applying Lemma 3.1,

$$\begin{aligned} &\|P^{[0]}(t_n, \Delta t_n, X_n^{[QR]}) - P^{[QR]}(t_n, \Delta t_n, X_n^{[QR]})\|_F \\ &\quad \leq \frac{2 + O(\tau \Delta t_n)}{\sqrt{2}(1 - O(\tau \Delta t_n))} \\ &\qquad\qquad \times \|P^{[0]}(t_n, \Delta t_n, X_n^{[QR]}) - P^{[\infty]}(t_n, \Delta t_n, X_n^{[QR]})\|_F, \end{aligned}$$

which gives

$$\begin{aligned} (3.3) \quad &\|P^{[0]}(t_n, \Delta t_n, X_n^{[QR]}) - P^{[QR]}(t_n, \Delta t_n, X_n^{[QR]})\|_F \\ &\quad \leq (\sqrt{2} + O(\tau \Delta t_n)) \\ &\qquad\qquad \times \|P^{[0]}(t_n, \Delta t_n, X_n^{[QR]}) - P^{[\infty]}(t_n, \Delta t_n, X_n^{[QR]})\|_F. \end{aligned}$$

In words, projecting onto the orthonormal QR factor corresponds to a perturbation that is no more than a factor $\sqrt{2}$ (plus higher order terms) bigger than the optimal perturbation given by the orthonormal polar factor. Hence, the local-error-per-unit-step can be bounded; using (2.3), (3.3) and the optimality of the polar factor we have

$$\begin{aligned} &\|P^{[QR]}(t_n, \Delta t_n, X_n^{[QR]}) - S(t_n, \Delta t_n, X_n^{[QR]})\|_F \\ &\quad \leq \|P^{[QR]}(t_n, \Delta t_n, X_n^{[QR]}) - P^{[0]}(t_n, \Delta t_n, X_n^{[QR]})\|_F \\ &\qquad + \|P^{[0]}(t_n, \Delta t_n, X_n^{[QR]}) - S(t_n, \Delta t_n, X_n^{[QR]})\|_F \\ &\quad \leq (\sqrt{2} + 1 + O(\tau \Delta t_n)) K \Delta t_n \tau. \end{aligned}$$

So we have a bound similar to (3.2) on the local-error-per-unit-step. The constant $2K$ arising for the polar factor case ($k = \infty$) is increased to $(\sqrt{2}+1)K$ (plus higher order terms) for the QR case. Thus, there is a corresponding increase to $(\sqrt{2} + 1)C_3$ in the resulting global error bound.

We mention that Dieci et al. [2] and Lord [14] also consider the convergence behaviour of the QR-based projection method. Lord establishes convergence with fixed time-steps on linear problems, and Dieci et al. [2, Lemma 4.2] show that the size of the projection is of the same asymptotic order as the local error. However, neither reference determines a constant (such as $(\sqrt{2} + 1)C_3$) for the leading term in a global error bound. Bunse-Gerstner et al. [1, page 26 and Theorem 14] consider projecting onto a nearest orthonormal matrix as part of a time-stepping algorithm for computing an analytic singular value decomposition of a path of matrices. Using the optimality of the projection, they prove convergence of a constant time-step, first-order, Euler-based scheme. In the same context Mehrmann and Rath [16, page 83] use the QR-based projection.

We have based our approach on the idea of projecting towards orthonormality at the end of each time-step, and then advancing from this new value. It is possible, of course, to perform all time-steps with the standard method, $P^{[0]}$, and then to project the values $\{X_n^{[0]}\}$ towards orthonormality. If the solution is required only at certain time values, this approach offers computational advantages, since only the relevant $X_n^{[0]}$ matrices need to be projected. Further, if we project to the optimal orthonormal polar factor then, since $X(t_n)$ is orthonormal, the perturbations are no bigger than $e_n^{[0]}$. So, from (3.1), we obtain the global error bound $2C_3\tau$. Similarly, projecting to the orthonormal QR factor leads to the global error bound $(\sqrt{2}+1)C_3\tau$. In other words, projecting *after* the integration has been completed leads to the same bounds as projecting *during* the integration. However, it seems intuitively preferable to project during the integration, so that the numerical solution is continually forced towards orthonormality.

With regard to cost, each Schulz iteration requires $2m^2p + 2mp^2$ flops, using the terminology of [5, section 1.2.4], where a flop is any floating point operation. However, as mentioned in [13], the iteration is rich in matrix multiplication, making it extremely attractive for modern computer architectures. Hence, for the Schulz iteration, the raw flop count may not be an appropriate measure of cost. (In fact, an alternative iteration specifically designed for parallel computers is derived in [12].) A QR decomposition computed by the modified Gram-Schmidt method costs $2mp^2$ flops [5, section 5.2.9]. When $m = p$, one Newton iteration costs $8m^3/3$ flops. Overall, performing one iteration of Schulz for $m < p$ or Newton for $m = p$ is likely to require roughly the same computational effort as the QR decomposition, and gives slightly better global error bounds. Whether the cost of projection forms a significant part of the complete integration process is heavily dependent upon the nature of the ODE system.

We illustrate some of these ideas with numerical tests on Example 5.2 of [2]. We implemented the algorithms in Matlab [15] with the underlying time-stepping and error control based on the built-in `ode45.m` function (which uses an explicit Runge-Kutta pair of orders four and five). The upper plot in Figure 3.1 shows the global error scaled by the tolerance, that is $e_n^{[k]}/\tau$, for

- the unprojected scheme, $k = 0$: thin solid line,
- the scheme with one Newton iteration, $k = 1$: dashed line,
- the scheme with two Newton iterations, $k = 2$: dotted line.

A tolerance of $\tau = 10^{-5}$ was used. In the upper plot, the two projected schemes produced visually indistinguishable global errors which are roughly two orders of magnitude smaller than those for the unprojected scheme. (The number of time-steps was identical for the three schemes.) The improvement in the departure from orthogonality caused by projection is clearly visible in the lower plot. In the $k = 2$ case the departure from orthogonality is consistent with rounding errors.

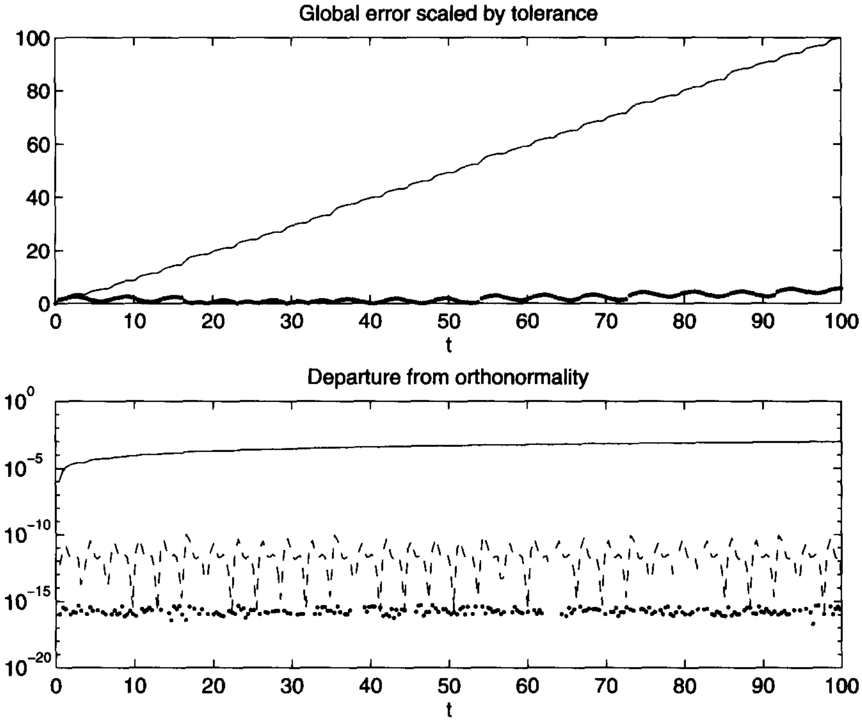

Figure 3.1: Effect of projection on global error and departure from orthogonality: see text for details.

Similar results arose with the Schulz iteration. We also implemented schemes that, on each time-step, replaced the approximate answer with (a) the orthonormal polar factor, and (b) the orthonormal QR factor, computed by the modified Gram-Schmidt method. We found very little difference between these two schemes and the $k = 2$ Newton or Schulz versions in our tests.

We also looked at the approach of applying the standard scheme and simply projecting to the orthonormal polar factor at the end of the integration. On some test problems, this gave an endpoint global error as small as that arising from projection at each time-step. However, this was not always the case—for example, on Problem 2 from [8, section 3] projection at each time-step gave an endpoint global error that was smaller by more than an order of magnitude than endpoint projection, with the same number of steps.

As a final point, we remark that the analysis in section 2 could be applied in more general circumstances where the solution operator has an invariant manifold and there is a convenient way to compute an optimal projection. Shampine [17] discusses this issue and develops some theory, although his assumptions are very different from those made in this work.

REFERENCES

1. Angelika Bunse-Gerstner, Ralph Byers, Volker Mehrmann, and Nancy K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–39.
2. Luca Dieci, Robert D. Russell, and Erik S. van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal., 31 (1994), pp. 261–281.
3. Luca Dieci, Robert D. Russell, and Erik S. Van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal., to appear.
4. Luca Dieci and Erik S. Van Vleck, *Computation of a few Lyapunov exponents for continuous and discrete dynamical systems*, Appl. Numer. Math., to appear.
5. Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, second edition, 1989. ISBN 0-8018-3739-1.
6. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer Verlag, second edition, 1993.
7. P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, New York, 1962.
8. Desmond J. Higham. Runge–Kutta type methods for orthogonal integration. Technical Report NA/168, University of Dundee, 1996, Appl. Numer. Math., to appear.
9. Desmond J. Higham and Andrew M. Stuart, *Analysis of the dynamics of local error control via a piecewise continuous residual*, Technical Report SCCM-95-03, Stanford University, 1995.
10. Nicholas J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Stat. Comput., 7:4 (1986), pp. 1160–1174.
11. Nicholas J. Higham, *Matrix nearness problems and applications*, in M. J. C. Gover and S. Barnett, editors, Applications of Matrix Theory, pp. 1–27. Oxford University Press, 1989.
12. Nicholas J. Higham and Pythagoras Papadimitriou, *A parallel algorithm for computing the polar decomposition*, Parallel Computing, 20 (1994), pp. 1161–1173.
13. Nicholas J. Higham and Robert S. Schreiber, *Fast polar decomposition of an arbitrary matrix*, SIAM J. Sci. Stat. Comput., 11:4 (1990), pp. 648–655.
14. G. J. Lord, *Analysis of numerical methods suitable for computing Lyapunov exponents*, Technical Report 95/02, University of Bath, 1995.
15. The MathWorks, Inc. *MATLAB User's Guide.* Natick, Massachusetts, 1992.
16. Volker Mehrmann and Werner Rath, *Numerical methods for the computation of analytic singular value decompositions*, Electronic Transactions on Numerical Analysis, 1 (1993), pp. 72–88.
17. L. F. Shampine, *Conservation laws and the numerical solution of ODEs*, Comp. Maths. Applics., 12B (1986), pp. 1287–1296.
18. Ji-guang Sun, *A note on backward perturbations for the Hermitian eigenvalue problem*, BIT, 35 (1995), pp. 385–393.