Mem. Fac. Eng., Kyoto Univ. Vol. 51, No. 4 (1989)

Numerical Calculation for Discretization of Continuous Quadratic Performance Index

By

Tomomichi HAGIWARA, Yumi SAITO and Mituhiko ARAKI

(Received June 30, 1989)

Abstract

A new procedure of numerical calculation to discretize a quadratic performance index defined for a linear time-invariant continuous system is proposed. The procedure is based on the Padé approximation with scaling and repeated squaring. Theoretical bounds of truncation errors involved in the resulting discretized weighting matrices are provided in terms of the maximum singular value norm for the proposed procedure. It is also shown that the paper by Van Loan which proposed another procedure of numerical calculation for the same problem contains some errors and a certain modification is required for his procedure. Numerical examples show that the new procedure is superior to the old one (of the modified version) from the viewpoint of accuracy, efficiency, and reliability.

1. Introduction

In sampled-data control of a continuous system using a zero-order hold, the optimal control which minimizes a continuous quadratic performance index posed on the continuous system is proposed[2),3)]. Such optimal control (the outline of which is reviewed in Section 2) has two advantages over the ordinary sampled-data optimal control under the standard discrete quadratic performance index. First, due to the use of the continuous performance index, it is possible to take account of the intersample behaviour of the system. Secondly, some information can be obtained as to the range of acceptable values for the sampling period by investigating the relation between the sampling period and the optimal performance index[2)].

On the other hand, the following difficulty is also present. In order to solve the above-mentioned optimal control problem[2),3)], it is necessary to convert the problem into an equivalent discrete optimal control problem[2),3)]. As we shall see

Department of Electrical Engineering II

in Section 2, this conversion consists of derivations of coefficient matrices for the sampled-data system, and weighting matrices for the discretized quadratic performance index. Since these matrices are expressed as complicated integrals involving matrix exponentials (see Section 2 for details), it is quite difficult to calculate the numerical values of these matrices by an analytic means, although they are required for the solution of the optimal control problem.

On account of this difficulty, some procedures have been proposed to numerically calculate these matrices[1),4),5)]. Among them, one of the most effective procedures available at present is the one proposed by Van Loan[1)], where the Padé approximation and the doubling formulae are utilized. This procedure is quite convenient since the bounds of truncation errors involved in the approximated matrices are provided in terms of a certain matrix norm. In addition, the precision and amount of computational effort can be regulated through a tuning parameter. As a matter of fact, however, the error bounds given by Van Loan[1)] are erroneous, as shown in Section 4.

In the present paper, we propose a new procedure for the numerical calculation of the above-mentioned coefficient and weighting matrices, where the Padé approximation and the doubling formulae are likewise utilized (Section 3). The bounds of truncation errors involved in the matrices calculated by the proposed procedure are also provided in terms of the maximum singular value norm. On the other hand, the error bounds for the Van Loan method[1)]described in terms of the Frobenius norm are shown to be erroneous on account of an intrinsic property of the Frobenius norm. In order to assure the correct theoretical error bounds, in Section 3, Van Loan's procedure is modified to use the maximum singular value norm rather than the Frobenius norm. Furthermore, theoretical error bounds are given for the modified procedure in terms of the maximum singular value norm (Section 4). Finally, the method proposed in the present paper is shown to be superior to the (modified) Van Loan method with respect to accuracy, efficiency and reliability by numerical examples (Section 5).

2. Expressions for Discretized Coefficient and Weighting Matrices

In this section, we review the sampled-data optimal control problem under a continuous quadratic performance index, and give expressions for the discretized coefficient matrices and weighting matrices.

Suppose that a linear time-invariant continuous system

$$dx(t)/dt=A_c x(t)+B_c u(t) \quad x\in \mathrm{R}^n,\ u\in \mathrm{R}^m \tag{2.1}$$

be given, and consider the sampled-data optimal control problem under the continuous quadratic performance index

$$J=\int_0^{\infty} \{x'(t)Q_c x(t)+u'(t)R_c u(t)\}dt \tag{2.2}$$

and the constraint

$$u(t)=u(kT) \quad (kT\leqq t<\overline{k+1}T). \tag{2.3}$$

The constraint (2.3) implies that a zero-order hold is employed at the input terminal of (2.1), where T denotes the sampling period. In (2.2), the symbol ' denotes the transpose, and Q_c and R_c are both symmetric. It would be usually assumed that Q_c and R_c are, respectively, positive semidefinite and positive definite, and that (A_c, B_c) and $(A_c, Q_c^{1/2})$ are, respectively, stabilizable and detectable. However, these assumptions are not always prerequisite in the following.

In view of the piecewise constant property (2.3) of the manipulating variables, the state transition equation (2.1) and the performance index (2.2) can be equivalently rewritten by discrete expressions[2),3)] as

$$x(\overline{k+1}T)=A(T)x(kT)+B(T)u(kT) \tag{2.4}$$

and

$$J=\sum_{k=0}^{\infty} \{x'(kT)Q(T)x(kT)+2x'(kT)S(T)u(kT)+u'(kT)R(T)u(kT)\}, \tag{2.5}$$

respectively, where

$$A(T)=\exp(A_cT) \tag{2.6}$$

$$B(T)=\int_0^T \exp(A_c t)B_c dt \tag{2.7}$$

$$Q(T)=\int_0^T \exp(A_c't)Q_c \exp(A_c t)dt \tag{2.8}$$

$$S(T)=\int_0^T \exp(A_c't)Q_c \int_0^t \exp(A_c s)B_c\, dsdt \tag{2.9}$$

$$R(T)=R_cT+\int_0^T\left[\int_0^t \exp(A_c s)B_c ds\right]'Q_c \int_0^t \exp(A_c s)B_c\, dsdt. \tag{2.10}$$

Thus, the original optimal control problem with the constraint (2.3) reduces to the optimal control problem without constraint. Therefore, the optimal control law can be easily derived by solving the optimal control problem of (2.4) and (2.5). In order to solve this problem, we require the numerical values of the matrices (2.6)-(2.10). Unfortunately, however, the expressions for these matrices are very much complicated, and it is quite difficult to calculate them by an analytic means.

The purpose of this paper is to provide an effective procedure for the numerical calculation of these matrices, and to study the numerical properties of the procedure.

3. A New Procedure for Numerical Calculation of the Discretized Matrices and Its Comparison with the Procedure by Van Loan

In this section, we first give our basic idea for the calculation of matrices given by (2.6)-(2.10). Secondly, we propose a new procedure for the numerical calculation of these matrices using the notions of the Padé approximation and the doubling formulae. Finally, we review the procedure proposed by Van Loan[1], and make some fundamental comparisons between these two procedures.

3.1 Basic Idea for the Procedure to Be Proposed

First, noting that

$$\exp\left(\begin{bmatrix} A_c & B_c \\ O & O \end{bmatrix} t\right) = \begin{bmatrix} \exp(A_c t) & \int_0^t \exp(A_c s) B_c ds \\ O & I_m \end{bmatrix}, \tag{3.1}$$

it can be readily obtained from (2.8)-(2.10) that

$$\begin{bmatrix} Q(T) & S(T) \\ S'(T) & W(T) \end{bmatrix} = \int_0^T \exp\left(\begin{bmatrix} A_c & B_c \\ O & O \end{bmatrix} t\right) \begin{bmatrix} Q_c & O \\ O & O \end{bmatrix} \exp\left(\begin{bmatrix} A_c & B_c \\ O & O \end{bmatrix} t\right) dt, \tag{3.2}$$

where

$$W(T) = R(T) - R_c T. \tag{3.3}$$

It is easy to see that the right hand side of (3.2) has the form of (2.8) with (A_c, Q_c) replaced by

$$\left(\begin{bmatrix} A_c & B_c \\ O & O \end{bmatrix}, \begin{bmatrix} Q_c & O \\ O & O \end{bmatrix}\right). \tag{3.4}$$

Here, it is known[1] that $Q(T)$ given by (2.8) can be calculated as

$$Q(T) = F_3'(T) G_2(T), \tag{3.5}$$

where $F_3(T)$ and $G_2(T)$ are given by

$$\exp\left(\begin{bmatrix} -A_c' & Q_c \\ O & A_c \end{bmatrix} T\right) = \begin{bmatrix} F_2(T) & G_2(T) \\ O & F_3(T) \end{bmatrix}. \tag{3.6}$$

In view of (3.2), it follows that $Q(T)$, $S(T)$ and $W(T)$ can be calculated at one time using appropriate submatrices of $\exp(C_0 T)$, where

$$C_0 = \begin{pmatrix} -A_c' & O & Q_c & O \\ -B_c' & O & O & O \\ O & O & A_c & B_c \\ O & O & O & O \end{pmatrix}. \tag{3.7}$$

Equivalently, we have only to calculate

$$\exp(CT) = \begin{pmatrix} F_1(T) & G_1(T) & H_1(T) & K_1(T) \\ O & F_2(T) & G_2(T) & H_2(T) \\ O & O & F_3(T) & G_3(T) \\ O & O & O & F_4(T) \end{pmatrix}, \tag{3.8}$$

where

$$C = \begin{pmatrix} O & I_n & O & O \\ I_m & O & O & O \\ O & O & I_n & O \\ O & O & O & I_m \end{pmatrix}^{-1} C_0 \begin{pmatrix} O & I_n & O & O \\ I_m & O & O & O \\ O & O & I_n & O \\ O & O & O & I_m \end{pmatrix} = \begin{pmatrix} O & -B_c' & O & O \\ O & -A_c' & Q_c & O \\ O & O & A_c & B_c \\ O & O & O & O \end{pmatrix}. \tag{3.9}$$

To summarize, it is easy to verify that $Q(T)$, $S(T)$ and $W(T)$ can be obtained as

$$Q(T) = F_3'(T) G_2(T) \tag{3.10}$$

$$S(T) = F_3'(T) H_2(T) \tag{3.11}$$

$$W(T) = G_3'(T) H_2(T) + K_1(T). \tag{3.12}$$

Needless to say, $R(T)$ can be obtained from (3.3) as

$$R(T) = W(T) + R_c T. \tag{3.13}$$

Furthermore, from inspection of (3.1), (3.8) and (3.9), it follows immediately that

$$A(T) = F_3(T) \tag{3.14}$$

$$B(T) = G_3(T). \tag{3.15}$$

The expressions (3.10)-(3.15) are convenient since the calculation of the right hand sides does not require any integration, but requires only an exponential of a matrix. By combining a method of numerical calculation of the matrix

exponential given by (3.8), a procedure for the numerical calculation of matrices (2.6)-(2.10) will be proposed in the next subsection. Before proceeding, note the following. Instead of (3.2), we can derive

$$\begin{pmatrix} Q(T) & S(T) \\ S'(T) & R(T) \end{pmatrix} = \int_0^T \exp\left(\begin{bmatrix} A_c & B_c \\ O & O \end{bmatrix}' t\right) \begin{bmatrix} Q_c & O \\ O & R_c \end{bmatrix} \exp\left(\begin{bmatrix} A_c & B_c \\ O & O \end{bmatrix} t\right) dt. \quad (3.16)$$

Therefore, we have another alternative which calculates $\exp(C_1 T)$ instead of (3.8), where

$$C_1 = \begin{pmatrix} O & I_n & O & O \\ I_m & O & O & O \\ O & O & I_n & O \\ O & O & O & I_m \end{pmatrix}^{-1} \begin{pmatrix} -A_c' & O & Q_c & O \\ -B_c' & O & O & R_c \\ O & O & A_c & B_c \\ O & O & O & O \end{pmatrix} \begin{pmatrix} O & I_n & O & O \\ I_m & O & O & O \\ O & O & I_n & O \\ O & O & O & I_m \end{pmatrix}$$
$$= \begin{pmatrix} O & -B_c' & O & R_c \\ O & -A_c' & Q_c & O \\ O & O & A_c & B_c \\ O & O & O & O \end{pmatrix}. \quad (3.17)$$

However, this alternative will not be adopted for two reason. First, a nonzero submatrix R_c in (3.17) is disadvantageous as compared with (3.9) from the viewpoint of the amount of computational effort. Secondly, as will be clarified by the truncation error analysis in the next section, having more nonzero submatrices seems to be undesirable from the viewpoint of precision of the calculation, too.

3.2 Padé Approximation, Doubling Formulae, and Proposal of a Discretization Procedure

In order to raise the idea in the previous subsection to a complete procedure, we must be provided with a numerical procedure for the calculation of the matrix exponential (3.8). It is well known that one of the most effective methods to calculate a matrix exponential is the Padé approximation with scaling and repeated squaring[6]. Hence, we apply this method to calculate (3.8). Then, we obtain an approximate value of (3.8) by

$$\exp(CT) \cong [R_{qq}(Ct_0)]^{2^j} \qquad (t_0 = T/2^j) \quad (3.18)$$

where

$$R_{qq}(A) = D_q(A)^{-1} N_q(A) \quad (\cong \exp(A)) \quad (3.19)$$

$$D_q(A)=\sum_{k=0}^{q}\beta_k(-A)^k,\quad N_q(A)=\sum_{k=0}^{q}\beta_k A^k \tag{3.20}$$

$$\beta_k=(2q-k)!q!/[(2q)!k!(q-k)!]. \tag{3.21}$$

Equivalently, β_k is given by

$$\beta_0=1 \tag{3.22}$$

and

$$\beta_k=[\prod_{i=0}^{k-1}(q-i)/(2q-i)]/k! \quad (k>0). \tag{3.23}$$

In (3.18), j is the minimum non-negative integer satisfying

$$\| Ct_0 \| = \| CT \| /2^j \leqq 1/2, \tag{3.24}$$

where $\| \cdot \|$ denotes the maximum singular value norm

$$\begin{aligned}\| A \| &=\lambda_{\max}^{1/2}(A'A) \quad (\lambda_{\max} \text{ denotes the maximum eigenvalue})\\ &=\lambda_{\max}^{1/2}(AA')\\ &=\max_{\| x \|_E=1} \| Ax \|_E \quad (\| \cdot \|_E \text{ denotes the Euclid vector norm}),\end{aligned} \tag{3.25}$$

Unless otherwise stated, $\| \cdot \|$ denotes the maximum singular value norm in the following. Furthermore, q is an arbitrary positive integer, on which the precision and amount of computational effort are dependent. In Section 5, a guideline on determination of the value of q will be given, using the results of the truncation error analysis given in the next section. But, regardless of the value of q, $D_q(Ct_0)$ is guaranteed to be non-singular from (3.24). The proof of this fact and an efficient method for calculation of $R_{qq}(Ct_0)$ are given in Appendix 1.

So far, $R_{qq}(Ct_0)$ has been obtained. For the sake of efficiency, however, we shall not repeatedly square this matrix directly as suggested by (3.18). Instead, we apply (3.10)-(3.12), (3.14), and (3.15) with T replaced by t_0 to estimate $A(t_0)$, $B(t_0)$, $Q(t_0)$, $S(t_0)$, and $W(t_0)$, and then apply

$$A(2t)=A(t)^2 \tag{3.26}$$

$$B(2t)=B(t)+A(t)B(t) \tag{3.27}$$

$$Q(2t)=Q(t)+A'(t)Q(t)A(t) \tag{3.28}$$

$$S(2t)=S(t)+A'(t)[Q(t)B(t)+S(t)] \tag{3.29}$$

$$W(2t)=2W(t)+B'(t)[Q(t)B(t)+S(t)]+S'(t)B(t) \tag{3.30}$$

repeatedly. Identities (3.26)-(3.30) can be easily derived from the definitions (2.6) -(2.10), and are referred to as the doubling formulae[1]. Since $T=2^j t_0$, repeated applications of the doubling formulae by j times yield $A(T)$, $B(T)$, $Q(T)$, $S(T)$, and $W(T)$. By adopting this alternative, the amount of computational effort can be significantly reduced[1].

To summarize, the procedure we propose in the present paper is as follows:

Procedure 1 :

(Step 1) Find the minimum non-negative integer j satisfying (3.24), and put

$$t_0=T/2^j,\ t_{k+1}=2t_k \quad (k=0,\ \cdots,\ j-1). \tag{3.31}$$

(Step 2) Calculate the approximate value of $\exp(Ct_0)$ using the Padé approximation (3.19)-(3.21), together with the efficient procedure given in Appendix 1.

$$\exp(Ct_0)\cong R_{qq}(Ct_0)=\begin{pmatrix} \hat{F}_1(t_0) & \hat{G}_1(t_0) & \hat{H}_1(t_0) & \hat{K}_1(t_0) \\ O & \hat{F}_2(t_0) & \hat{G}_2(t_0) & \hat{H}_2(t_0) \\ O & O & \hat{F}_3(t_0) & \hat{G}_3(t_0) \\ O & O & O & \hat{F}_4(t_0) \end{pmatrix} \tag{3.32}$$

(Stem 3) Calculate the approximate values of $A(t_0)$, $B(t_0)$, $Q(t_0)$, $S(t_0)$, and $W(t_0)$ by

$$A_0=\hat{F}_3(t_0) \tag{3.33}$$
$$B_0=\hat{G}_3(t_0) \tag{3.34}$$
$$Q_0=\hat{F}_3'(t_0)\hat{G}_2(t_0) \tag{3.35}$$
$$S_0=\hat{F}_3'(t_0)\hat{H}_2(t_0) \tag{3.36}$$
$$W_0=\hat{G}_3'(t_0)\hat{H}_2(t_0)+\hat{K}_1(t_0), \tag{3.37}$$

respectively.

(Step 4) Calculate the approximate values of $A(t_{k+1})$, $B(t_{k+1})$, $Q(t_{k+1})$, $S(t_{k+1})$, and $W(t_{k+1})$ by

$$A_{k+1}=A_k^2 \tag{3.38}$$
$$B_{k+1}=B_k+A_kB_k \tag{3.39}$$
$$Q_{k+1}=Q_k+A_k'Q_kA_k \tag{3.40}$$
$$S_{k+1}=S_k+A_k'[Q_kB_k+S_k] \tag{3.41}$$
$$W_{k+1}=2W_k+B_k'[Q_kB_k+S_k]+S_k'B_k, \tag{3.42}$$

respectively (for $k=0,\ \cdots,\ j-1$).

(Step 5) Let

$$A=A_j,\ B=B_j,\ Q=Q_j,\ S=S_j,\ W=W_j \tag{3.43}$$

and

$$R=W+R_cT. \tag{3.44}$$

Then A, B, Q, S, W, and R are the approximate values of $A(T)$, $B(T)$, $Q(T)$, $S(T)$, $W(T)$ and $R(T)$, respectively.

In the above procedure, q is an arbitrary positive integer. A criterion for the determination of q will be given in subsection 5.1 below. In the following, also (3.38)-(3.42) are referred to as the doubling formulae.

3.3 Procedure Proposed by Van Loan

In this subsection, we review the procedure for the calculation of matrices (2.6)-(2.10) proposed by Van Loan[1]. Details on the underlying idea for the procedure are omitted since it is explained in 1), but it would be fair to say that the idea is more technical and complicated than our idea presented in subsection 3.1.

Let $\tilde{C}$ be given by

$$\tilde{C}=\begin{pmatrix} -A_c' & I_n & O & O \\ O & -A_c' & Q_c & O \\ O & O & A_c & B_c \\ O & O & O & O \end{pmatrix}, \tag{3.45}$$

and let q be an arbitrary positive integer. Then, the procedure is as follows:

Procedure 2 :

(Step 1) Find the minimum non-negative integer j satisfying

$$\| \tilde{C}T \| /2^j \leqq 1/2, \tag{3.46}$$

and put

$$t_0=T/2^j,\ t_{k+1}=2t_k \quad (k=0, \cdots, j-1). \tag{3.47}$$

(Step 2) Caluculate the approximate value of $\exp(\tilde{C}t_0)$ using the Padé approximation (3.19)-(3.21).

$$\exp(\tilde{C}t_0)\cong R_{qq}(\tilde{C}t_0)=\begin{pmatrix}\overline{F}_1(t_0) & \overline{G}_1(t_0) & \overline{H}_1(t_0) & \overline{K}_1(t_0)\\ O & \overline{F}_2(t_0) & \overline{G}_2(t_0) & \overline{H}_2(t_0)\\ O & O & \overline{F}_3(t_0) & \overline{G}_3(t_0)\\ O & O & O & \overline{F}_4(t_0)\end{pmatrix} \tag{3.48}$$

(Step 3) Calculate the approximate values of $A(t_0)$, $B(t_0)$, $Q(t_0)$, $S(t_0)$, and $W(t_0)$ by

$$\tilde{A}_0=\overline{F}_3(t_0) \tag{3.49}$$
$$\tilde{B}_0=\overline{G}_3(t_0) \tag{3.50}$$
$$\tilde{Q}_0=\overline{F}_3'(t_0)\overline{G}_2(t_0) \tag{3.51}$$
$$\tilde{S}_0=\overline{F}_3'(t_0)\overline{H}_2(t_0) \tag{3.52}$$
$$\tilde{W}_0=[B_c\overline{F}_3'(t_0)\overline{K}_1(t_0)]+[B_c'\overline{F}_3'(t_0)\overline{K}_1(t_0)]', \tag{3.53}$$

respectively.

(Step 4) Calculate the approximate values of $A(t_{k+1})$, $B(t_{k+1})$, $Q(t_{k+1})$, $S(t_{k+1})$, and $W(t_{k+1})$ by the doubling formulae (3.38)–(3.42) (with A_i replaced by $\tilde{A}_i$, etc.), respectively (for $k=0, \cdots, j-1$).

(Step 5) Let

$$\tilde{A}=\tilde{A}_j,\ \tilde{B}=\tilde{B}_j,\ \tilde{Q}=\tilde{Q}_j,\ \tilde{S}=\tilde{S}_j,\ \tilde{W}=\tilde{W}_j \tag{3.54}$$

and

$$\tilde{R}=\tilde{W}+R_cT. \tag{3.55}$$

Then, $\tilde{A}$, $\tilde{B}$, $\tilde{Q}$, $\tilde{S}$, $\tilde{W}$, and $\tilde{R}$ are the approximate values of $A(T)$, $B(T)$, $Q(T)$, $S(T)$, $W(T)$, and $R(T)$, respectively.

In (3.46), $\|\cdot\|$ denoted the Frobenius norm

$$\|A\|=[\sum_i\sum_j a_{ij}^2]^{1/2}\quad (A=(a_{ij})) \tag{3.56}$$

in the original procedure proposed by Van Loan[1], but we should modify the procedure and consider (3.46) in terms of the maximum singular value norm in the following. The reason for this will be clarified in the next section.

Thus, Procedure 1 proposed in the present paper turns out to be quite similar to the previously proposed Procedure 2. In particular, since the right-lower $\overline{2n+m}\times\overline{2n+m}$ submatrices of C and $\tilde{C}$ coincide, it follows from inspection of these two Procedures that the approximate values of $A(T)$, $B(T)$, $Q(T)$, and $S(T)$ obtained by Procedure 1 are exactly the same as those obtained by Procedure 2 provided that q and j in the two Procedures coincide, even if we

take account of the effect of rounding errors. Moreover, the amount of computational effort required for each of these four discretized matrices is also exactly the same for these two Procedures.

On the other hand, as for $W(T)$ and $R(T)$ these two Procedures result in different approximate values, and the newly proposed Procedure 1 generally yields approximate values which are not less accurate than those by Procedure 2, as will be shown in the following two sections. Furthermore, since $C\in R^{\overline{2n+2m}\times\overline{2n+2m}}$ whereas $\tilde{C}\in R^{\overline{3n+m}\times\overline{3n+m}}$, and since $n\geqq m$ in general, it follows that Procedure 1 requires a less amount of calculation than Procedure 2 in (Step 1), (Step 2) and (Step 3). (Step 4) of the doubling formulae requires the same amount of computational effort in each Procedure.

4. Truncation Error Analysis

4.1 Bounds of Truncation Errors

The purpose of this section is to show that the truncation errors involved in the matrices calculated by Procedure 1 which we proposed in subsection 3.2 are bounded by the inequalities

$$\| A-A(T) \| \leqq \tau_A\theta(T) \tag{4.1}$$

$$\| B-B(T) \| \leqq \tau_B\theta(T) \tag{4.2}$$

$$\| Q-Q(T) \| \leqq \tau_Q\theta^2(T) \tag{4.3}$$

$$\| S-S(T) \| \leqq \tau_S\theta^2(T) \tag{4.4}$$

$$\| R-R(T) \| = \| W-W(T) \| \leqq \begin{cases} \tau_R\theta^4(T/2) & (j>0) \quad (4.5) \\ \tau_R\theta^2(T) \quad (\leqq\tau_R\theta^4(T/2)) & (j=0) \quad (4.6) \end{cases}$$

where

$$\tau_A=\varepsilon T\exp(\varepsilon T) \tag{4.7}$$

$$\tau_B=\varepsilon T\exp(\varepsilon T)[1+\alpha T/2] \tag{4.8}$$

$$\tau_Q=\varepsilon T\exp(2\varepsilon T)[1+\alpha T] \tag{4.9}$$

$$\tau_S=\varepsilon T\exp(2\varepsilon T)[1+(\alpha+\varepsilon)T]^2 \tag{4.10}$$

$$\tau_R=4\varepsilon T\exp(2\varepsilon T)[\{1+(\alpha+\varepsilon)T/2\}^3+1] \tag{4.11}$$

$$\varepsilon=2^{3-2q}\| C \|(q!)^2/[(2q)!(2q+1)!] \quad (\geqq 0) \tag{4.12}$$

$$\alpha=\max\{\| B_c \|, \| Q_c \|\} \quad (\geqq 0) \tag{4.13}$$

$$\theta(t)=\max_{0\leqq s\leqq t}\| \exp(A_c s) \| . \tag{4.14}$$

The proofs of (4.1)-(4.6) are almost parallel to those of Theorems 2-6 of 1), which gave the bounds of truncation errors for the original form of

Procedure 2 in terms of the Frobenius norm. Those proofs were based on the inequalities

$$\| A+B \| \leqq \| A \| + \| B \| \tag{4.15}$$
$$\| AB \| \leqq \| A \| \cdot \| B \| \tag{4.16}$$
$$\| A' \| = \| A \| \tag{4.17}$$
$$\| \mathrm{sub}(A) \| \leqq \| A \| \tag{4.18}$$
$$\| \exp(At) \| \leqq \exp(at) \qquad (\| A \| \leqq a,\ t \geqq 0), \tag{4.19}$$

where $\| \cdot \|$ denotes the Frobenius norm and sub(A) denotes an arbitrary matrix obtained by deleting some rows and/or columns of the matrix A. But, the inequality (4.19) does not hold for the Frobenius norm (for example, consider the case where $t=0$) on account of its intrinsic property

$$\| I_n \| = n^{1/2}. \tag{4.20}$$

Therefore, the truncation error bounds given in 1) for the original Van Loan method turn out to be erroneous. This is why we have modified his original procedure and have given the modified procedure using the maximum singular value norm (see subsection 3.3).

The truncation errors for the modified procedure, Procedure 2, can be shown to be bounded by the inequalities

$$\| \tilde{A} - A(T) \| \leqq \tilde{\tau}_A \theta(T) \tag{4.21}$$
$$\| \tilde{B} - B(T) \| \leqq \tilde{\tau}_B \theta(T) \tag{4.22}$$
$$\| \tilde{Q} - Q(T) \| \leqq \tilde{\tau}_Q \theta^2(T) \tag{4.23}$$
$$\| \tilde{S} - S(T) \| \leqq \tilde{\tau}_S \theta^2(T) \tag{4.24}$$
$$\| \tilde{R} - R(T) \| = \| \tilde{W} - W(T) \| \leqq \tilde{\tau}_R \theta^4(T/2) \qquad (j>0) \tag{4.25}$$
$$\leqq \tilde{\tau}_R \theta^2(T) \quad (\leqq \tilde{\tau}_R \theta^4(T/2)) \qquad (j=0) \tag{4.26}$$

where

$$\tilde{\tau}_A = \tilde{\varepsilon} T \exp(\tilde{\varepsilon} T) \tag{4.27}$$
$$\tilde{\tau}_B = \tilde{\varepsilon} T \exp(\tilde{\varepsilon} T)[1+\alpha T/2] \tag{4.28}$$
$$\tilde{\tau}_Q = \tilde{\varepsilon} T \exp(2\tilde{\varepsilon} T)[1+\alpha T] \tag{4.29}$$
$$\tilde{\tau}_S = \tilde{\varepsilon} \exp(2\tilde{\varepsilon} T)[1+(\alpha+\tilde{\varepsilon})T]^2 \tag{4.30}$$
$$\tilde{\tau}_R = 4\tilde{\varepsilon} T \exp(2\tilde{\varepsilon} T)[\{1+(\alpha+\tilde{\varepsilon})T/2\}^3+\alpha] \tag{4.31}$$
$$\tilde{\varepsilon} = 2^{3-2q} \| \tilde{C} \| (q!)^2/[(2q)!(2q+1)!] \tag{4.32}$$

and α and $\theta(t)$ are defined by (4.13) and (4.14), respectively. Since (4.15)-(4.19) hold for the maximum singular value norm, most of the arguments in the proofs of Theorems 2-6 of 1) are justified as the derivation of truncation error

bounds for Procedure 2 in terms of the maximum singular value norm. However, some of the proofs of 1) are still erroneous in several other respects as will be pointed out in subsection 4.3. Therefore, the error bounds (4.24)-(4.26) are different from those given in 1), even if we neglect the difference of the underlying definitions of the matrix norm.

In the next subsection, we give some preliminary results, which are quite similar to those results given in 1). Then, in subsection 4.3, we give complete proofs for (4.1)-(4.6). The proofs of (4.21)-(4.26) will be omitted because they are parallel to the proofs of (4.1)-(4.6).

4.2 Preliminary Results for Error Analysis

Lemma 1 (Theorem 1, Van Loan 1978) :

$$\exp\left(\begin{bmatrix} \Gamma_1 & \Delta_1 & \Theta_1 & \Lambda_1 \\ O & \Gamma_2 & \Delta_2 & \Theta_2 \\ O & O & \Gamma_3 & \Delta_3 \\ O & O & O & \Gamma_4 \end{bmatrix} t\right) = \begin{bmatrix} \Xi_1(t) & \Phi_1(t) & \Psi_1(t) & \Omega_1(t) \\ O & \Xi_2(t) & \Phi_2(t) & \Psi_2(t) \\ O & O & \Xi_3(t) & \Phi_3(t) \\ O & O & O & \Xi_4(t) \end{bmatrix} \tag{4.33}$$

holds true where

$$\Xi_i(t) = \exp(\Gamma_i t) \quad (i=1, 2, 3, 4) \tag{4.34}$$

$$\Phi_i(t) = \int_0^t \exp\{\Gamma_i(t-s)\}\,\Delta_i \exp(\Gamma_{i+1}s)\,ds \quad (i=1, 2, 3) \tag{4.35}$$

$$\begin{aligned} \Psi_i(t) = &\int_0^t \exp\{\Gamma_i(t-s)\}\,\Theta_i \exp(\Gamma_{i+2}s)\,ds \\ &+ \int_0^t \int_0^s \exp\{\Gamma_i(t-s)\}\,\Delta_i \exp\{\Gamma_{i+1}(s-r)\}\,\Delta_{i+1} \exp(\Gamma_{i+2}r)\,dr\,ds \\ &\qquad (i=1, 2) \end{aligned} \tag{4.36}$$

$$\begin{aligned} \Omega_1(t) = &\int_0^t \exp\{\Gamma_1(t-s)\}\,\Lambda_1 \exp(\Gamma_4 s)\,ds \\ &+ \int_0^t \int_0^s \exp\{\Gamma_1(t-s)\}\,[\Theta_1 \exp\{\Gamma_3(s-r)\}\,\Delta_3 + \Delta_1 \exp\{\Gamma_2(s-r)\}\,\Theta_2] \\ &\quad \times \exp(\Gamma_4 r)\,dr\,ds \\ &+ \int_0^t \int_0^s \int_0^r \exp\{\Gamma_1(t-s)\}\,\Delta_1 \exp\{\Gamma_2(s-r)\}\,\Delta_2 \exp\{\Gamma_3(r-w)\}\,\Delta_3 \\ &\quad \times \exp(\Gamma_4 w)\,dw\,dr\,ds. \end{aligned} \tag{4.37}$$

Lemma 2 : If $R_{qq}(Ct_0)$ is computed according to (Step 2) of Procedure 1, then

$$R_{qq}(Ct_0) = \exp(\hat{C}t_0) \quad (\hat{C}=C+E) \tag{4.38}$$

holds true for some E, where

$$E=\begin{pmatrix} O & E_2 & E_5 & E_7 \\ O & -E_1' & E_3 & E_6 \\ O & O & E_1 & E_4 \\ O & O & O & O \end{pmatrix} \quad \text{(partitioned as } C\text{)}. \tag{4.39}$$

Furthermore,

$$A_cE_1=E_1A_c \tag{4.40}$$

$$\| E_i \| \leq \varepsilon \quad (i=1, \cdots, 7), \tag{4.41}$$

where ε is given by (4.12).

Proof: It can be easily verified that the proof of this lemma is parallel to that of Lemma 1 of 1) (save that $\| \cdot \|$ denotes the maximum singular value norm), if we scrutinize the underlying result (Moler and Van Loan[6], 1978, Appendix 1, Lemma 4) and the algorithm of the Padé approximation. Details are omitted for the sake of brevity.

Lemma 3: If A_k, B_k, Q_k, and S_k are defined by Procedure 1, then

$$A_k=\exp\{(A_c+E_1)t_k\} \tag{4.42}$$

$$B_k=\int_0^{t_k} \exp\{(A_c+E_1)s\}(B_c+E_4)ds \tag{4.43}$$

$$Q_k=\int_0^{t_k} \exp\{(A_c+E_1)'s\}(Q_c+E_3)\exp\{(A_c+E_1)s\}ds \tag{4.44}$$

$$S_k=\int_0^{t_k} \exp\{(A_c+E_1)'s\}E_6ds + \int_0^{t_k} \exp\{(A_c+E_1)'s\}(Q_c+E_3)\int_0^s \exp\{(A_c+E_1)(s-r)\}(B_c+E_4)drds. \tag{4.45}$$

Furthermore, W_0 is given by

$$W_0=(\varepsilon_1+\varepsilon_2)+(\varepsilon_3+\varepsilon_4+\varepsilon_5+\varepsilon_6) \tag{4.46}$$

where

$$\varepsilon_1=\int_0^{t_0} (B_c+E_4)'\exp\{(A_c+E_1)'(t_0-s)\}ds \int_0^{t_0} \exp\{-(A_c+E_1)'(t_0-s)\}E_6ds \tag{4.47}$$

$$\varepsilon_2=\int_0^{t_0}(B_c+E_4)'\exp\{(A_c+E_1)'(t_0-s)\}ds\int_0^{t_0}\exp\{-(A_c+E_1)'(t_0-s)\}$$
$$\times(Q_c+E_3)\int_0^s\exp\{(A_c+E_1)(s-r)\}(B_c+E_4)drds \tag{4.48}$$

$$\varepsilon_3=\int_0^{t_0}E_7ds \tag{4.49}$$

$$\varepsilon_4=\int_0^{t_0}\int_0^s E_5\exp\{(A_c+E_1)(s-r)\}(B_c+E_4)drds \tag{4.50}$$

$$\varepsilon_5=\int_0^{t_0}\int_0^s(-B_c'+E_2)\exp\{-(A_c+E_1)'(s-r)\}E_6drds \tag{4.51}$$

$$\varepsilon_6=\int_0^{t_0}\int_0^s(-B_c'+E_2)\exp\{-(A_c+E_1)'(s-r)\}(Q_c+E_3)$$
$$\times\int_0^r\exp\{(A_c+E_1)(r-w)\}(B_c+E_4)dwdrds. \tag{4.52}$$

Proof: Equations (4.42)-(4.45) with $k=0$ and (4.46) immediately follow from Lemma 1 and Lemma 2, where

$$\varepsilon_1+\varepsilon_2=\hat{G}_3'(t_0)\hat{H}_2(t_0) \tag{4.53}$$
$$\varepsilon_3+\varepsilon_4+\varepsilon_5+\varepsilon_6=\hat{K}_1(t_0). \tag{4.54}$$

Equations (4.42)-(4.45) with $k>0$ can be derived by induction with respect to k. See the proof of Lemma 3 of 1) for detail.

Remark: Equations (4.42)-(4.45) imply that the calculations based on the doubling formulae and the direct squaring (3.18) yield the same results for matrices A, B, Q, and S (if we neglect the effect of the rounding errors). However, as far as the matrix R is concerned, in general, these two calculations yield slightly different results.

4.3 Proofs of the Truncation Error Bounds

In this subsection, we prove the inequalities (4.1)-(4.6) using the preliminary results given in the previous subsection. Inequalities $1\leqq\exp(\varepsilon t^1)\leqq\exp(\varepsilon t^2)$ $(0\leqq t^1\leqq t^2)$ and $1\leqq\theta(t^1)\leqq\theta(t^2)$ $(0\leqq t^1\leqq t^2)$ will be used repeatedly with no particular comments wherever they are necessary.

Proof of (4.1)

It follows from Lemma 3 that

$$A=\exp\{(A_c+E_1)T\}. \tag{4.55}$$

Therefore, we obtain from (4.14), (4.16), and (4.40) that

$$\begin{aligned}\| A-A(T) \| &= \| \exp\{(A_c+E_1)T\}-\exp(A_cT) \| \\ &\leqq \| \exp(A_cT) \| \times \| \exp(E_1T)-I_n \| \\ &\leqq \theta(T) \| \exp(E_1T)-I_n \| .\end{aligned} \tag{4.56}$$

Since

$$\begin{aligned}\| \exp(E_1s)-I_n \| &= \| \int_0^s E_1\exp(E_1r)dr \| \\ &\leqq \int_0^s \| E_1 \| \times \| \exp(E_1r) \| dr \\ &\leqq \int_0^s \varepsilon \cdot \exp(\varepsilon s)dr \\ &= \varepsilon s \cdot \exp(\varepsilon s)\end{aligned} \tag{4.57}$$

hold true from (4.16), (4.19), and (4.41), the inequality (4.1) follows readily from (4.56) and (4.57).

Proof of (4.2)

Since

$$B=\int_0^T \exp\{(A_c+E_1)s\}(B_c+E_4)ds \tag{4.58}$$

holds true from Lemma 3, we get

$$\begin{aligned}B-B(T)=&\int_0^T [\exp\{(A_c+E_1)s\}-\exp(A_cs)]B_cds \\ &+\int_0^T \exp\{(A_c+E_1)s\}E_4ds.\end{aligned} \tag{4.59}$$

Therefore, we obtain from (4.15) and (4.16) that

$$\begin{aligned}\| B-B(T) \| \leqq &\int_0^T \| \exp(A_cs) \| \times \| \exp(E_1s)-I \| \times \| B_c \| ds \\ &+\int_0^T \| \exp(A_cs) \| \times \| \exp(E_1s) \| \times \| E_4 \| ds.\end{aligned} \tag{4.60}$$

By using (4.13), (4.14), (4.19), (4.41), and (4.57), we obtain

$$\begin{aligned}\| B-B(T) \| &\leqq \int_0^T \theta(T) \cdot \varepsilon s\exp(\varepsilon T) \cdot \alpha ds+\int_0^T \theta(T) \cdot \exp(\varepsilon T) \cdot \varepsilon ds \\ &=\tau_B\theta(T).\end{aligned} \tag{4.61}$$

Proof of (4.3)

It follows from Lemma 3 and (2.8) that

$$Q-Q(T)=\int_0^T [\exp\{(A_c+E_1)'s\}(Q_c+E_3)\exp\{(A_c+E_1)s\}-$$
$$\exp(A_c's)Q_c\exp(A_cs)]ds$$
$$=\int_0^T[\exp\{(A_c+E_1)s\}-\exp(A_cs)]'Q_c\exp\{(A_c+E_1)s\}ds$$
$$+\int_0^T\exp(A_c's)Q_c[\exp\{(A_c+E_1)s\}-\exp(A_cs)]ds$$
$$+\int_0^T\exp\{(A_c+E_1)s\}'E_3\exp\{(A_c+E_1)s\}ds. \tag{4.62}$$

Therefore, taking the norms of both sides and using (4.17), we obtain

$$\|Q-Q(T)\|\leqq\int_0^T\theta(T)\varepsilon s\cdot\exp(\varepsilon T)\cdot\alpha\cdot\theta(T)\exp(\varepsilon T)ds$$
$$+\int_0^T\theta(T)\cdot\alpha\cdot\theta(T)\varepsilon s\cdot\exp(\varepsilon T)ds$$
$$+\int_0^T\theta(T)\exp(\varepsilon T)\cdot\varepsilon\cdot\theta(T)\exp(\varepsilon T)ds$$
$$\leqq\varepsilon\exp(2\varepsilon T)[2\alpha\int_0^T sds+T]\theta^2(T)$$
$$=\tau_Q\theta^2(T). \tag{4.63}$$

Proof of (4.4)

It follows from Lemma 3 and (2.9) that

$$S-S(T)=\int_0^T\exp\{(A_c+E_1)'s\}(Q_c+E_3)\int_0^s\exp\{(A_c+E_1)(s-r)\}(B_c+E_4)drds$$
$$+\int_0^T\exp\{(A_c+E_1)'s\}E_6ds$$
$$-\int_0^T\exp(A_c's)Q_c\int_0^s\exp\{A_c(s-r)\}B_cdrds$$
$$=\int_0^T[\exp\{(A_c+E_1)s\}-\exp(A_cs)]'Q_c\int_0^s\exp\{(A_c+E_1)(s-r)\}B_cdrds$$
$$+\int_0^T\exp(A_c's)Q_c\int_0^s[\exp\{(A_c+E_1)(s-r)\}-\exp\{A_c(s-r)\}]B_cdrds$$
$$+\int_0^T\exp\{(A_c+E_1)'s\}(Q_c+E_3)\int_0^s\exp\{(A_c+E_1)(s-r)\}E_4drds$$
$$+\int_0^T\exp\{(A_c+E_1)'s\}E_3\int_0^s\exp\{(A_c+E_1)(s-r)\}B_cdrds$$
$$+\int_0^T\exp\{(A_c+E_1)'s\}E_6ds. \tag{4.64}$$

Taking the norms of both sides, we obtain

$$\|S-S(T)\|\leqq\int_0^T\theta(T)\varepsilon s\cdot\exp(\varepsilon T)\cdot\alpha\int_0^s\theta(T)\exp(\varepsilon T)adrds$$
$$+\int_0^T\theta(T)\cdot\alpha\int_0^s\theta(T)\varepsilon r\cdot\exp(\varepsilon T)\cdot adrds$$
$$+\int_0^T\theta(T)\exp(\varepsilon T)\cdot(\alpha+\varepsilon)\int_0^s\theta(T)\exp(\varepsilon T)\cdot\varepsilon drds$$

$$+\int_0^T \theta(T)\exp(\varepsilon T)\cdot\varepsilon\int_0^s \theta(T)\exp(\varepsilon T)\cdot\alpha dr ds$$
$$+\int_0^T \theta(T)\exp(\varepsilon T)\cdot\varepsilon ds$$
$$\leqq \varepsilon\exp(2\varepsilon T)[\alpha^2T^3/3+\alpha^2T^3/6+(\alpha+\varepsilon)T^2/2+\alpha T^2/2+T]\theta^2(T)$$
$$\leqq \tau_S\theta^2(T). \tag{4.65}$$

Proof of (4.5) *and* (4.6)

It follows from Lemma 3 and (3.12) that

$$W_0-W(t_0)=\varepsilon_1+[\varepsilon_2-G_3'(t_0)H_2(t_0)]+\varepsilon_3+\varepsilon_4+\varepsilon_5+[\varepsilon_6-K_1(t_0)]. \tag{4.66}$$

From (4.47), (4.49), (4.50), and (4.51), we obtain

$$\|\varepsilon_1\| \leqq \int_0^{t_0}(\alpha+\varepsilon)\theta(t_0)\exp(\varepsilon t_0)ds\int_0^{t_0}\bar{\theta}(t_0)\exp(\varepsilon t_0)\varepsilon ds$$
$$=\varepsilon(\alpha+\varepsilon)t_0^2\exp(2\varepsilon t_0)\theta(t_0)\bar{\theta}(t_0) \tag{4.67}$$

$$\|\varepsilon_3\| \leqq \int_0^{t_0}\varepsilon ds$$
$$=\varepsilon t_0 \tag{4.68}$$

$$\|\varepsilon_4\| \leqq \int_0^{t_0}\int_0^s \varepsilon\theta(t_0)\exp(\varepsilon t_0)(\alpha+\varepsilon)drds$$
$$=\varepsilon(\alpha+\varepsilon)t_0^2\exp(\varepsilon t_0)\theta(t_0)/2 \tag{4.69}$$

$$\|\varepsilon_5\| \leqq \int_0^{t_0}\int_0^s(\alpha+\varepsilon)\bar{\theta}(t_0)\exp(\varepsilon t_0)\varepsilon drds$$
$$=\varepsilon(\alpha+\varepsilon)t_0^2\exp(\varepsilon t_0)\bar{\theta}(t_0)/2 \tag{4.70}$$

where

$$\bar{\theta}(t)=\max_{0\leqq s\leqq t}\|\exp(-A_c s)\| \quad (\geqq 1). \tag{4.71}$$

Next, from Lemma 1 and from (4.48), we obtain

$$\varepsilon_2-G_3'(t_0)H_2(t_0)=\int_0^{t_0}(B_c+E_4)'\exp\{-(A_c+E_1)'s\}ds$$
$$\times\int_0^{t_0}\exp\{(A_c+E_1)'s\}(Q_c+E_3)$$

$$\times \int_0^s \exp\{(A_c+E_1)(s-r)\}(B_c+E_4)drds$$

$$-\int_0^{t_0} B_c'\exp(-A_c's)ds$$

$$\times \int_0^{t_0} \exp(A_c's)Q_c \int_0^s \exp\{A_c(s-r)\}B_c drds$$

$$=\int_0^{t_0} B_c'[\exp\{-(A_c+E_1)s\}-\exp(-A_cs)]'ds$$

$$\times \int_0^{t_0} \exp\{(A_c+E_1)'s\}Q_c \int_0^s \exp\{(A_c+E_1)(s-r)\}B_c drds$$

$$+\int_0^{t_0} B_c'\exp(-A_c's)ds \int_0^{t_0} [\exp\{(A_c+E_1)s\}-\exp(A_cs)]'Q_c$$

$$\times \int_0^s \exp\{(A_c+E_1)(s-r)\}B_c drds$$

$$+\int_0^{t_0} B_c'\exp(-A_c's)ds \int_0^{t_0} \exp(A_c's)Q_c$$

$$\times \int_0^s [\exp\{(A_c+E_1)(s-r)\}-\exp\{A_c(s-r)\}]B_c drds$$

$$+\int_0^{t_0} (B_c+E_4)'\exp\{-(A_c+E_1)'s\}ds$$

$$\times \int_0^{t_0} \exp\{(A_c+E_1)'s\}E_3 \int_0^s \exp\{(A_c+E_1)(s-r)\}$$

$$\times (B_c+E_4)drds$$

$$+\int_0^{t_0} (B_c+E_4)'\exp\{-(A_c+E_1)'s\}ds$$

$$\times \int_0^{t_0} \exp\{(A_c+E_1)'s\}Q_c \int_0^s \exp\{(A_c+E_1)(s-r)\}E_4 drds$$

$$+\int_0^{t_0} E_4'\exp\{-(A_c+E_1)'s\}ds$$

$$\times \int_0^{t_0} \exp\{(A_c+E_1)'s\}Q_c \int_0^s \exp\{(A_c+E_1)(s-r)\}B_c drds. \tag{4.72}$$

Taking the norms of both sides, we obtain

$$
\begin{aligned}
&\| \varepsilon_2 - G_3'(t_0)H_2(t_0) \| \\
&\leqq \int_0^{t_0} \alpha\bar{\theta}(t_0)\varepsilon s \cdot \exp(\varepsilon t_0)ds \\
&\qquad \times \int_0^{t_0} \theta(t_0)\exp(\varepsilon t_0)\alpha \int_0^s \theta(t_0)\exp(\varepsilon t_0)\alpha dr ds \\
&\quad + \int_0^{t_0} \alpha\bar{\theta}(t_0)ds \int_0^{t_0} \theta(t_0)\varepsilon s \cdot \exp(\varepsilon t_0)\alpha \int_0^s \theta(t_0)\exp(\varepsilon t_0)\alpha dr ds \\
&\quad + \int_0^{t_0} \alpha\bar{\theta}(t_0)ds \int_0^{t_0} \theta(t_0)\alpha \int_0^s \theta(t_0)\varepsilon r \cdot \exp(\varepsilon t_0)\alpha dr ds \\
&\quad + \int_0^{t_0} (\alpha+\varepsilon)\bar{\theta}(t_0)\exp(\varepsilon t_0)ds \\
&\qquad \times \int_0^{t_0} \theta(t_0)\exp(\varepsilon t_0)\varepsilon \int_0^s \theta(t_0)\exp(\varepsilon t_0)(\alpha+\varepsilon)dr ds \\
&\quad + \int_0^{t_0} (\alpha+\varepsilon)\bar{\theta}(t_0)\exp(\varepsilon t_0)ds \int_0^{t_0} \theta(t_0)\exp(\varepsilon t_0)\alpha \int_0^s \theta(t_0)\exp(\varepsilon t_0)\varepsilon dr ds \\
&\quad + \int_0^{t_0} \varepsilon\bar{\theta}(t_0)\exp(\varepsilon t_0)ds \int_0^{t_0} \theta(t_0)\exp(\varepsilon t_0)\alpha \int_0^s \theta(t_0)\exp(\varepsilon t_0)\alpha dr ds \\
&\leqq \varepsilon[\alpha^3 t_0^4/4 + \alpha^3 t_0^4/3 + \alpha^3 t_0^4/6 + (\alpha+\varepsilon)^2 t_0^3/2 \\
&\qquad + \alpha(\alpha+\varepsilon)t_0^3/2 + \alpha^2 t_0^3/2]\exp(3\varepsilon t_0)\theta^2(t_0)\bar{\theta}(t_0) \\
&\leqq \varepsilon[3\alpha^3 t_0^4/4 + 3(\alpha+\varepsilon)^2 t_0^3/2]\exp(3\varepsilon t_0)\theta^2(t_0)\bar{\theta}(t_0). \qquad (4.73)
\end{aligned}
$$

On the other hand, we obtain from Lemma 1 and (4.52) that

$$
\| \varepsilon_6 - K_1(t_0) \| \leqq \varepsilon[(\alpha+\varepsilon)^2 t_0^3/2 + \alpha^3 t_0^4/8]\exp(2\varepsilon t_0)\theta(t_0)\bar{\theta}(t_0) \qquad (4.74)
$$

in like manner (the detail is omitted for the sake of brevity).

Finally, it follows from (4.66)-(4.70), (4.73), and (4.74) that

$$
\begin{aligned}
&\| W_0 - W(t_0) \| \\
&\quad \leqq \varepsilon t_0[(\alpha+\varepsilon)t_0 + \{3\alpha^3 t_0^3/4 + 3(\alpha+\varepsilon)^2 t_0^2/2\} + (\alpha+\varepsilon)t_0/2 + (\alpha+\varepsilon)t_0/2 \\
&\qquad + \{(\alpha+\varepsilon)^2 t_0^2/2 + \alpha^3 t_0^3/8\}]\exp(3\varepsilon t_0)\theta^2(t_0)\bar{\theta}(t_0) + \varepsilon t_0 \\
&\quad = \varepsilon t_0[2(\alpha+\varepsilon)t_0 + 2(\alpha+\varepsilon)^2 t_0^2 + 7\alpha^3 t_0^3/8]\exp(3\varepsilon t_0)\theta^2(t_0)\bar{\theta}(t_0) + \varepsilon t_0. \qquad (4.75)
\end{aligned}
$$

Here, note that

$$\varepsilon t_0 \leqq 1/6 \tag{4.76}$$

$$\alpha t_0 \leqq 1/2 \tag{4.77}$$

and

$$\| -A_c \| t_0 \leqq 1/2 \tag{4.78}$$

hold true from (3.24), (4.12), and (4.18). Furthermore, (4.78), together with (4.17) and (4.19), yields

$$\bar{\theta}(t_0) \leqq \exp(1/2) \leqq 7/4. \tag{4.79}$$

Applying (4.76), (4.77), and (4.79) to (4.75), we obtain

$$\| W_0 - W(t_0) \| \leqq \varepsilon t_0 [7\theta^2(t_0) + 1], \tag{4.80}$$

from which (4.6) follows readily (note that $t_0 = T$ if $j = 0$).

So far, we have obtained a bound of truncation error involved in W_0. To complete the proof of (4.5), we must investigate how the truncation error propagates by applying the doubling formulae. Subtracting (3.30) with t replaced by t_k from (3.42), and taking the norms of both sides, we obtain

$$\begin{aligned} \| W_{k+1} - W(t_{k+1}) \| \leqq 2 \| W_k - W(t_k) \| + 2 \| B_k' S_k - B'(t_k) S(t_k) \| \\ + \| B_k' Q_k B_k - B'(t_k) Q(t_k) B(t_k) \| \\ (k = 0, \cdots, j-1). \end{aligned} \tag{4.81}$$

Concerning the second and the third terms of the ffiight hand side, we have

$$\begin{aligned} \| B_k' S_k - B'(t_k) S(t_k) \| \leqq \varepsilon \theta^3(t_k) \exp(3\varepsilon t_k) t_k^2 (\alpha + \varepsilon) \\ \times [1 + 3(\alpha + \varepsilon) t_k/2 + 3(\alpha + \varepsilon)^2 t_k^2/4] \end{aligned} \tag{4.82}$$

and

$$\| B_k' Q_k B_k - B'(t_k) Q(t_k) B(t_k) \| \leqq \varepsilon \theta^4(t_k) \exp(4\varepsilon t_k) t_k^3 (\alpha + \varepsilon)^2 [3 + 2(\alpha + \varepsilon) t_k]. \tag{4.83}$$

The proof of (4.82) is given in Appendix 2. The proof of (4.83) is similar. From (4.81)-(4.83), we get

$$\| W_{k+1} - W(t_{k+1}) \| \leqq 2 \| W_k - W(t_k) \| + \delta_k \quad (k = 0, \cdots, j-1), \tag{4.84}$$

where

$$\delta_k = \varepsilon \theta^4(t_k) \exp(4\varepsilon t_k) t_k^2 (\alpha + \varepsilon) [2(\alpha + \varepsilon) t_k + 2]^2. \tag{4.85}$$

Then, the repeated application of (4.84) yields

$$\begin{aligned}\| W-W(T) \| &= \| W_j-W(t_j) \| \\ &\leqq 2 \| W_{j-1}-W(t_{j-1}) \| +\delta_{j-1} \\ &\cdots\cdots \\ &\leqq 2^j \| W_0-W(t_0) \| +\sum_{k=0}^{j-1} 2^{j-k-1}\delta_k \\ &\leqq 2^j \| W_0-W(t_0) \| \\ &\quad +\varepsilon\theta^4(t_{j-1})\exp(4\varepsilon t_{j-1})(\alpha+\varepsilon)[2(\alpha+\varepsilon)t_{j-1}+2]^2 \sum_{k=0}^{j-1} 2^{k-j-1}t_k^2. \quad (4.86)\end{aligned}$$

Noting that $2^j t_0=T$, $t_{j-1}=T/2$, $\theta(t)\leqq\theta^2(t/2)$ $(t\geqq 0)$, and

$$\begin{aligned}\sum_{k=0}^{j-1} 2^{j-k-1}t_k^2 &= \sum_{k=0}^{j-1} 2^{k-j-1}T^2 \\ &\leqq T^2/2, \quad (4.87)\end{aligned}$$

we obtain from (4.80) and (4.86) that

$$\begin{aligned}\| W-W(T) \| &\leqq \varepsilon T[7\theta^4(T/2)+1] \\ &\quad +\varepsilon T^2\theta^4(T/2)\exp(2\varepsilon T)(\alpha+\varepsilon)[(\alpha+\varepsilon)T+2]^2/2 \\ &\leqq \varepsilon T\exp(2\varepsilon T)[8+(\alpha+\varepsilon)T\{(\alpha+\varepsilon)T+2\}^2/2]\theta^4(T/2) \\ &\leqq \tau_R\theta^4(T/2). \quad (4.88)\end{aligned}$$

This completes the proof of (4.5).

Proofs of (4.21)-(4.26) are omitted since they are parallel to those of Theorems 2-6 of 1), save that $\| \cdot \|$ denotes the maximum singular value norm. For completeness, however, we shall point out the following errors in the proofs in 1).

a) The omitted part, consisting of simple calculations, in the proof of Theorem 5 is wrong.

b) Some of the preliminary results used in the proof of Van Loan's Theorem 5 are also wrong on account of errors in calculation:

b1) Lemma 4 of 1) should read as follows:

$$\| \tilde{W}_0-W(t_0) \| \leqq \tilde{\varepsilon}t_0\theta^2(t_0)[\alpha^3(t_0^3+t_0^2)/2+\alpha^2(t_0^2+t_0/6+1.44t_0)+\alpha(1.36t_0+3)]. \quad (4.89)$$

b2) The bounds for $\| \tilde{B}_k'\tilde{S}_k-B'(t_k)S(t_k) \|$ and $\| \tilde{B}_k'\tilde{Q}_k\tilde{B}_k-B'(t_k)Q(t_k)B(t_k) \|$ used in Van Loan's proof are wrong, and the corrected bounds are the right hand sides of (4.82) and (4.83) with ε replaced by $\tilde{\varepsilon}$,

respectively.

Correction of a) leads to the bounds of (4.24), and correction of b) leads to the bounds of (4.25) and (4.26). Details are omitted for the sake of brevity.

5. Numerical Study

The purpose of this section is to study some numerical examples, and to compare Procedures 1 and 2 from the numerical aspects. In order to make the comparison fair, in subsection 5.1 we describe the method of determining the value of q in the Padé approximation (3.19)-(3.21). Then, in subsection 5.2, we give the results of numerical examples studied by the authors. Those results will indicate that Procedure 1, which we propose in this paper, is superior to Procedure 2, which is the modified version of the Van Loan method[1], from the viewpoint of accuracy, efficiency, and reliability.

5.1 Guideline on Determination of the Value of q

Suppose that A_c, B_c, Q_c, R_c, and T be given. Then, $\theta(T)$ and $\theta(T/2)$ are determined by (4.14), and α is determined by (4.13). Therefore, in order to reduce the right hand sides of (4.1)-(4.6) (resp. (4.21)-(4.26)), we have to use a large value of q to reduce ε (resp. $\tilde{\varepsilon}$). Accordingly, we adopt the following procedure[1] to determine the value of q.

Procedure for Determination of the Value of q

(Step 1) Determine a "degree of tolerance" τ_0 for the truncation error of the approximated matrices considering the conflicting desire to reduce the amount of calculation.

(Step 2) Let q_0 be the minimum value of q satisfying

$$\tau_A \leqq \tau_0,\ \tau_B \leqq \tau_0,\ \tau_Q \leqq \tau_0,\ \tau_S \leqq \tau_0,\ \tau_R \leqq \tau_0 \tag{5.1}$$
$$(\text{resp. } \tilde{\tau}_A \leqq \tau_0,\ \tilde{\tau}_B \leqq \tau_0,\ \tilde{\tau}_Q \leqq \tau_0,\ \tilde{\tau}_S \leqq \tau_0,\ \tilde{\tau}_R \leqq \tau_0).$$

(Step 3) Determine q by

$$q = \min\{q_0,\ q_{\max}\}, \tag{5.2}$$

where $q_{\max}$ is an upper bound placed on the value of q according to the precision of the machine employed.

The rationale for (Step 3) is that a larger q should be employed only if the truncation errors, which we are trying to regulate through q, dominate the

Table. 1 Computation Results of Example 1
(Sampling Period T=1.0, Tolerance τ_0=1.0D−04)

(a) Computation Results and True Values

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	6	7	—
j	7	7	—
q	4	4	—
$A(T)$			
$A(1, 1)$	0.4775281427D+00	0.4775281427D+00	0.4775281430D+00
$A(1, 2)$	−0.5221553628D+00	−0.5221553628D+00	−0.5221553630D+00
$A(1, 3)$	−0.3510589330D+00	−0.3510589330D+00	−0.3510589330D+00
$A(2, 1)$	0.8554821487D+00	0.8554821487D+00	0.8554821480D+00
$A(2, 2)$	−0.9945236572D+00	−0.9945236572D+00	−0.9945236570D+00
$A(2, 3)$	−0.7021178661D+00	−0.7021178661D+00	−0.7021178660D+00
$A(3, 1)$	−0.8554821487D+00	−0.8554821487D+00	−0.8554821480D+00
$A(3, 2)$	0.1012839296D+01	0.1012839296D+01	0.1012839296D+01
$A(3, 3)$	0.7204335050D+00	0.7204335050D+00	0.7204335050D+00
$B(T)$			
$B(1, 1)$	0.1999431436D+01	0.1999431436D+01	0.1999431436D+01
$B(1, 2)$	−0.3394449326D+01	−0.3394449326D+01	−0.3394449325D+01
$B(2, 1)$	0.1148224077D+01	0.1148224077D+01	0.1148224072D+01
$B(2, 2)$	−0.6155423363D+01	−0.6155423363D+01	−0.6155423359D+01
$B(3, 1)$	−0.1665397155D+00	−0.1665397155D+00	−0.1665397110D+00
$B(3, 2)$	0.7627949905D+01	0.7627949905D+01	0.7627949901D+01
$Q(T)$			
$Q(1, 1)$	0.9934877780D+01	0.9934877780D+01	0.9934877720D+01
$Q(1, 2)$	−0.1108568965D+02	−0.1108568965D+02	−0.1108568953D+02
$Q(1, 3)$	−0.9123023947D+01	−0.9123023947D+01	−0.9123023900D+01
$Q(2, 2)$	0.1366870754D+02	0.1366870754D+02	0.1366870748D+02
$Q(2, 3)$	0.1150451516D+02	0.1150451516D+02	0.1150451512D+02
$Q(3, 3)$	0.1029179557D+02	0.1029179557D+02	0.1029179555D+02
$S(T)$			
$S(1, 1)$	0.3515982356D+01	0.3515982356D+01	0.3515982340D+01
$S(1, 2)$	−0.2487596341D+02	−0.2487596341D+02	−0.2487596341D+02
$S(2, 1)$	−0.2516164484D+01	−0.2516164484D+01	−0.2516164470D+01
$S(2, 2)$	0.3094693521D+02	0.3094693521D+02	0.3094693518D+02
$S(3, 1)$	−0.1194242586D+01	−0.1194242586D+01	−0.1194242580D+01
$S(3, 2)$	0.2429316620D+02	0.2429316620D+02	0.2429316617D+02
$R(T)$			
$R(1, 1)$	0.1529648648D+02	0.1529652522D+02	0.1529648659D+02
$R(1, 2)$	−0.4373425687D+01	−0.4373404397D+01	−0.4373425530D+01
$R(2, 2)$	0.1099996702D+03	0.1099997055D+03	0.1099996704D+03

rounding errors (see 1) for detail).

5.2 Numerical Examples

In this subsection, we give the computation results of several numerical examples, where we used the FACOM M-780 computer of the Data Processing Center of Kyoto University.

Example 1 : Let us consider the same example as Van Loan studied in 1), which is given by

$$A_c=\begin{pmatrix} 2 & -8 & -6 \\ 10 & -19 & -12 \\ -10 & 15 & 8 \end{pmatrix},\quad B_c=\begin{pmatrix} 5 & 1 \\ 1 & 4 \\ 3 & 2 \end{pmatrix},\quad Q_c=\begin{pmatrix} 4 & 1 & 2 \\ 1 & 3 & 1 \\ 2 & 1 & 5 \end{pmatrix},\quad R_c=\begin{pmatrix} 3 & 1 \\ 1 & 4 \end{pmatrix}$$

and

$$T=1.$$

Table. 1 Computation Results of Example 1
(Sampling Period T=1.0, Tolerance τ_0=1.0D−04)

(b) Residual Errors and Error Bounds

	Proposed Method		Modified Van Loan Method	
	Residual Error	Error Bound	Residual Error	Error Bound
$A(T)$	0.1045473D−08	0.1774545D−06	0.1045473D−08	0.1773322D−06
$B(T)$	0.8630867D−08	0.8028854D−06	0.8630867D−08	0.8023322D−06
$Q(T)$	0.1949802D−06	0.5962737D−05	0.1949802D−06	0.5958628D−05
$S(T)$	0.4067861D−07	0.4799358D−04	0.4067861D−07	0.4796051D−04
$R(T)$	0.3474921D−06	0.4834754D−02	0.5805190D−04	0.5143591D−02

Letting $\tau_0=10^{-4}$ in (5.1), we obtain the results given in Table 1. In the table, "Proposed Method" implies Procedure 1, and "Modified Van Loan Method" implies Procedure 2. "CPU Time" denotes the total computation time required to compute $A(T)$, $B(T)$, $Q(T)$, $S(T)$, and $R(T)$. "j" and "q" respectively denote the values of j and q in each procedure. "Residual Error" and "Error Bound" respectively denote the left hand side and the right hand side of (4.1)-(4.5) or (4.21)-(4.25).

As we mentioned in subsection 3.2, the computation results by Procedure 1 and Procedure 2 exactly coincide as far as $A(T)$, $B(T)$, $Q(T)$, and $S(T)$ are concerned. We can verify this fact from Table 1. However, Table 1 shows that the computation results of $R(T)$ by Procedure 1 and Procedure 2 are different, and Procedure 1 yields a more accurate result than Procedure 2.

Table 2 Computation Results of Example 2
(Sampling Period $T=0.5$)

(a) Tolerance $\tau_0=1.0D-03$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	5	5	—
j	3	3	—
q	3	3	—
$R(T)$			
$R(1, 1)$	0.5830816355D+01	0.5830343095D+01	0.5830816355D+01
$R(1, 2)$	0.3906887864D+01	0.3906940565D+01	0.3906887864D+01
$R(2, 2)$	0.4462709805D+01	0.4463499478D+01	0.4462709800D+01
Residual Error	0.4683042D−08	0.7918734D−03	—
Error Bound	0.1679959D−01	0.2121498D−01	—

(b) Tolerance $\tau_0=1.0D-06$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	4	5	—
j	3	3	—
q	4	4	—
$R(T)$			
$R(1, 1)$	0.5830816355D+01	0.5830305458D+01	0.5830816355D+01
$R(1, 2)$	0.3906887864D+01	0.3906939648D+01	0.3906887864D+01
$R(2, 2)$	0.4462709800D+01	0.4463547361D+01	0.4462709800D+01
Residual Error	0.7348489D−12	0.8395461D−03	—
Error Bound	0.1666605D−04	0.2104636D−04	—

(c) Tolerance $\tau_0=1.0D-08$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	5	5	—
j	3	3	—
q	5	5	—
$R(T)$			
$R(1, 1)$	0.5830816355D+01	0.5830284549D+01	0.5830816355D+01
$R(1, 2)$	0.3906887864D+01	0.3906939153D+01	0.3906887864D+01
$R(2, 2)$	0.4462709800D+01	0.4463573988D+01	0.4462709800D+01
Residual Error	0.2344582D−13	0.8660696D−03	—
Error Bound	0.1052150D−07	0.1328684D−07	—

Table 3 Computation Results of Example 2
(Sampling Period $T=1.0$)

(a) Tolerance $\tau_0=1.0D-02$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	4	6	—
j	4	4	—
q	3	3	—
$R(T)$			
$R(1, 1)$	0.4383702173D+02	0.4383607521D+02	0.4383702173D+02
$R(1, 2)$	0.3165953692D+02	0.3165964232D+02	0.3165953692D+02
$R(2, 2)$	0.1118674436D+03	0.1118690230D+03	0.1118674433D+03
Residual Error	0.2704600D−06	0.1584007D−02	—
Error Bound	0.3892434D+01	0.4078258D+01	—

(b) Tolerance $\tau_0=1.0D-04$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	5	6	—
j	4	4	—
q	4	4	—
$R(T)$			
$R(1, 1)$	0.4383702173D+02	0.4383599993D+02	0.4383702173D+02
$R(1, 2)$	0.3165953692D+02	0.3165964049D+02	0.3165953692D+02
$R(2, 2)$	0.1118674433D+03	0.1118691185D+03	0.1118674433D+03
Residual Error	0.3842951D−10	0.1679092D−02	—
Error Bound	0.3861453D−02	0.4045800D−02	—

(c) Tolerance $\tau_0=1.0D-08$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	5	6	—
j	4	4	—
q	5	5	—
$R(T)$			
$R(1, 1)$	0.4383702173D+02	0.4383595812D+02	0.4383702173D+02
$R(1, 2)$	0.3165953692D+02	0.3165963950D+02	0.3165953692D+02
$R(2, 2)$	0.1118674433D+03	0.1118691717D+03	0.1118674433D+03
Residual Error	0.6463794D−12	0.1732139D−02	—
Error Bound	0.2437786D−05	0.2554167D−05	—

In the following examples, we show only the computation results of $R(T)$, and compare the numerical properties of the two procedures by the results, since the computation results of the other matrices by the two procedures coincide, and they are quite accurate, as in Table 1.

Example 2 : In this example, let us consider the case where

$$A_c=\begin{pmatrix}1 & 0 & 0\\ 0 & 3 & -5\\ 0 & 0 & -2\end{pmatrix},\quad B_c=\begin{pmatrix}4 & 3\\ 1 & 1\\ 1 & 4\end{pmatrix},\quad Q_c=\begin{pmatrix}3 & 0 & 1\\ 0 & 2 & 0\\ 1 & 0 & 2\end{pmatrix},\quad R_c=\begin{pmatrix}5 & 2\\ 2 & 1\end{pmatrix}.$$

The computation results for $T=0.5$ are shown in Table 2, where τ_0 in (5.1) is set to 10^{-3}, 10^{-6}, and 10^{-8} to change the value of q. From this table, we observe that the residual error for Procedure 1 becomes very small as the value of q becomes large, which is a natural consequence of (4.5). However, the residual error for Procedure 2 becomes large as q becomes large, and sometimes goes beyond the error bound given by (4.25). This is the case also for $T=1$, as shown in Table 3. Therefore, this example indicates that Procedure 2 is sensitive to rounding errors, and is not reliable from the viewpoint of numerical stability.

In the above two examples, $\alpha>1$ holds true. Therefore, in view of (4.5) and (4.25), the comparison might not be fair. In the following example, we study the case of $\alpha<1$.

Example 3 : Let us consider the case where

$$A_c=\begin{pmatrix}-3 & 0 & 0\\ 0 & -5 & 0\\ 0 & 0 & -1\end{pmatrix},\quad B_c=\begin{pmatrix}0.4\\ 0.4\\ 0.4\end{pmatrix},\quad Q_c=\begin{pmatrix}0.2 & 0 & 0\\ 0 & 0.1 & 0\\ 0 & 0 & 0.5\end{pmatrix},\quad R_c=0.3$$

and

$$T=0.2$$

Letting $\tau_0=10^{-3}$, we obtain the results of Table 4. Although $\alpha<1$ holds true, the computation result by Procedure 1 is more accurate.

Finally, we study an example in which the matrix A_c has complex eigenvalues.

Table 4 Computation Results of Example 3
(Sampling Period T=0.2, Tolerance τ_0=1.0D−03)

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	3	4	—
j	2	2	—
q	3	3	—
$R(T)$ $R(1, 1)$	0.6026136905D−01	0.6026195045D−01	0.6026136905D−01
Residual Error	0.2530138D−12	0.5814068D−06	—
Error Bound	0.1117063D−04	0.1050289D−04	—

Example 4 : Let us consider the case where

$$A_c=\begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & -2 & 1 \end{pmatrix}, \quad B_c=\begin{pmatrix} 0 & 1 & 0 \\ -2 & 0 & 0 \\ 0 & 0 & 3 \\ -4 & 0 & 0 \\ 0 & 0 & 5 \end{pmatrix},$$

$$Q_c=\begin{pmatrix} 25 & 0 & 0 & -10 & 0 \\ 0 & 16 & -12 & 0 & -4 \\ 0 & -12 & 9 & 0 & 3 \\ -10 & 0 & 0 & 4 & 0 \\ 0 & -4 & 3 & 0 & 1 \end{pmatrix}, \quad R_c=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

and

$$T=0.1.$$

Then, we obtain the results shown in Table 5. The results also indicate the unreliability of Procedure 2 as in Example 2.

So far, we have compared Procedure 1 and Procedure 2 from the viewpoint of accuracy and reliability. Concerning efficiency, we can see that Procedure 1 is more efficient than Procedure 2, in most cases, from Table 1-5 (see CPU Time). As a consequence, we obtain that Procedure 1, which we propose in this paper, is superior to Procedure 2, which is the modified procedure of the Van

Table 5 Computation Results of Example 4
(Sampling Period $T=0.1$)

(a) Tolerance $\tau_0=1.0D-01$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	8	8	—
j	3	3	—
q	3	3	—
$R(T)$			
$R(1, 1)$	0.1428258817D+00	0.1428258564D+00	0.1428258817D+00
$R(1, 2)$	0.1328560008D−01	0.1328560660D−01	0.1328560008D−01
$R(1, 3)$	0.3199660087D−01	0.3200093969D−01	0.3199660087D−01
$R(2, 2)$	0.2077364883D+00	0.2077369726D+00	0.2077364883D+00
$R(2, 3)$	−0.1264520762D−02	−0.1263544248D−02	−0.1264520762D−02
$R(3, 3)$	0.3586229941D+00	0.3586288642D+00	0.3586229941D+00
Residual Error	0.5730437D−13	0.8264456D−05	—
Error Bound	0.2764715D−03	0.7693399D−03	—

(a) Tolerance $\tau_0=1.0D-04$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	9	8	—
j	4	4	—
q	3	3	—
$R(T)$			
$R(1, 1)$	0.1428258817D+00	0.1428258519D+00	0.1428258817D+00
$R(1, 2)$	0.1328560008D−01	0.1328560766D−01	0.1328560008D−01
$R(1, 3)$	0.3199660087D−01	0.3200124335D−01	0.3199660087D−01
$R(2, 2)$	0.2077364883D+00	0.2077370068D+00	0.2077364883D+00
$R(2, 3)$	−0.1264520762D−02	−0.1263473050D−02	−0.1264520762D−02
$R(3, 3)$	0.3586229941D+00	0.3586292704D+00	0.3586229941D+00
Residual Error	0.1185061D−14	0.8839285D−05	—
Error Bound	0.2742748D−06	0.7632279D−06	—

(a) Tolerance $\tau_0=1.0D-07$

	Proposed Method	Modified Van Loan Method	True Value
CPU Time (ms)	9	9	—
j	5	5	—
q	3	3	—
$R(T)$			
$R(1, 1)$	0.1428258817D+00	0.1428258495D+00	0.1428258817D+00
$R(1, 2)$	0.1328560008D−01	0.1328560825D−01	0.1328560008D−01
$R(1, 3)$	0.3199660087D−01	0.3200141206D−01	0.3199660087D−01
$R(2, 2)$	0.2077364883D+00	0.2077370258D+00	0.2077364883D+00
$R(2, 3)$	−0.1264520762D−02	−0.1263433496D−02	−0.1264520762D−02
$R(3, 3)$	0.3586229941D+00	0.3586294961D+00	0.3586229941D+00
Residual Error	0.1176503D−14	0.9158657D−05	—
Error Bound	0.1731533D−09	0.4818358D−09	—

Loan method, from the viewpoint of accuracy, efficiency, and reliability.

6. Conclusion

In this paper, we proposed a new procedure of numerical calculation for the discretization of continuous quadratic performance index. The procedure is based on the Padé approximation with scaling and repeated squaring[6]. Furthermore, we gave the bounds of the truncation error of the calculation in terms of the maximum singular value norm.

The proposed procedure is similar to the existing procedure given by Van Loan[1]. Concerning the latter procedure, we clarified that the truncation error bounds given in 1) are erroneous due to some misled sub-estimation of norm. In the course of the correction of the sub-estimation, we also showed that the Van Loan procedure itself should be modified. Lastly, we compared our procedure with the modified Van Loan procedure by numerical examples. As a result, the procedure we proposed in this paper turned out to be superior from the viewpoint of accuracy, efficiency and reliability.

References

1) C. F. Van Loan; IEEE Trans. Automat. Contr.; **AC-23**, 395 (1978)
2) A. H. Levis, R. A. Schlueter and M. Athans; Int. J. Contr.; **13**, 343 (1971)
3) P. Dorato and A. H. Levis; IEEE Trans. Automat. Contr.; **AC-16**, 613 (1971)
4) J. C. Johnson and C. L. Phillips; IEEE Trans. Automat. Contr.; **AC-16**, 204 (1971)
5) E. S. Armstrong; IEEE Trans. Automat. Contr.; **AC-23**, 478 (1978)
6) C. Moler and C. Van Loan; SIAM Rev.; **20**, 801 (1978)
7) R. A. Horn and C. A. Johnson; "Matrix Analysis," Cambridge University Press (1985)

Appendix 1

In this appendix, we provide an efficient method to calculate the Padé approximation $R_{qq}(Ct_0)$.

We first prove that $D_q(A)$ is non-singular if $\| A \| \leqq 1/2$. Since $D_q(O)=I$ is non-singular, it suffices to consider the case of $\| A \| \neq 0$. Noting that $\beta_k < 1$ ($k \geqq 1$) holds true from (3.23), we obtain

$$\rho(D_q(A)-I) \leqq \| D_q(A)-I \|$$

$$= \| \sum_{k=1}^{q} \beta_k (-A)^k \|$$

$$< \sum_{k=1}^{q} \| A \|^k$$

$$< \| A \| / \{1 - \| A \| \}$$

$$\leqq 1, \tag{A1.1}$$

where $\rho(\cdot)$ denotes the spectral radius. Inequality (A 1.1) implies that $D_q(A)$ is non-singular. This, together with (3.24), shows that $D_q(Ct_0)$ is actually non-singular.

We next give an efficient procedure to calculate the powers of the matrix C given by (3.9): For $k \geqq 1$, C^k is given by

$$C^k = \begin{pmatrix} O & (-1)^k U_k' & Z_k & Y_k \\ O & (-1)^k X_k' & P_k & V_k \\ O & O & X_k & U_k \\ O & O & O & O \end{pmatrix}, \tag{A1.2}$$

where

$$Y_1 = O \tag{A1.3}$$
$$Z_1 = O \tag{A1.4}$$
$$V_1 = O \tag{A1.5}$$
$$P_1 = Q_c \tag{A1.6}$$
$$U_1 = B_c \tag{A1.7}$$
$$X_1 = A_c \tag{A1.8}$$

and

$$Y_{k+1} = -B_c' V_k \qquad (k = 1, \cdots, q-1) \tag{A1.9}$$
$$Z_{k+1} = -B_c' P_k \qquad (k = 1, \cdots, q-1) \tag{A1.10}$$
$$V_{k+1} = -A_c' V_k + Q_c U_k \qquad (k = 1, \cdots, q-1) \tag{A1.11}$$
$$P_{k+1} = -A_c' P_k + Q_c X_k \qquad (k = 1, \cdots, q-1) \tag{A1.12}$$
$$U_{k+1} = X_k B_c \qquad (k = 1, \cdots, q-1) \tag{A1.13}$$
$$X_{k+1} = A_c X_k \qquad (k = 1, \cdots, q-1). \tag{A1.14}$$

Equations (A 1.2)-(A 1.14) can be easily proved by induction if we note that

$$A_c X_k = X_k A_c \qquad (k = 1, \cdots, q-1) \tag{A1.15}$$
$$X_k B_c = A_c U_k \qquad (k = 1, \cdots, q-1) \tag{A1.16}$$

Thus, $D_q(Ct_0)$ and $N_q(Ct_0)$ can be calculated efficiently.

Denoting

$$D_q(Ct_0)=\begin{pmatrix} I & D_{12} & D_{13} & D_{14} \\ O & D_{22} & D_{23} & D_{24} \\ O & O & D_{33} & D_{34} \\ O & O & O & I \end{pmatrix}, \quad N_q(Ct_0)=\begin{pmatrix} I & N_{12} & N_{13} & N_{14} \\ O & N_{22} & N_{23} & N_{24} \\ O & O & N_{33} & N_{34} \\ O & O & O & I \end{pmatrix}, \tag{A1.17}$$

and expanding both sides of $D_q(Ct_0)R_{qq}(Ct_0)=N_q(Ct_0)$ using the expression of (3.32), we obtain

$$D_{33}\hat{F}_3(t_0)=D_{22}' \tag{A1.18}$$
$$D_{33}\hat{G}_3(t_0)=N_{34}-D_{34} \tag{A1.19}$$
$$D_{22}\hat{G}_2(t_0)=N_{23}-D_{23}\hat{F}_3(t_0) \tag{A1.20}$$
$$D_{22}\hat{H}_2(t_0)=N_{24}-D_{23}\hat{G}_3(t_0)-D_{24} \tag{A1.21}$$
$$\hat{K}_1(t_0)=N_{14}-D_{12}\hat{H}_2(t_0)-D_{13}\hat{G}_3(t_0)-D_{14}, \tag{A1.22}$$

where

$$N_{33}=\sum_{k=0}^{q}\beta_k A_c^k=[\sum_{k=0}^{q}\beta_k\{-(-A_c')\}^k]'=D_{22}' \tag{A1.23}$$

is used in (A 1.18). Solving linear equations (A 1.18)-(A 1.21), we can obtain $\hat{F}_3(t_0)$, $\hat{G}_3(t_0)$, $\hat{G}_2(t_0)$, $\hat{H}_2(t_0)$, and $\hat{K}_1(t_0)$. The other submatrices of $R_{qq}(Ct_0)$ need not be calculated since they are not required in Procedure 1.

It is possible to solve the equations (A 1.18)-(A 1.21) by an iterative algorithm (Horn and Johnson[7] 1985, Problem 1, p. 350) since $\rho(D_{22}-I)<1$ and $\rho(D_{33}-I)<1$ hold true from (A1.1). For example, $\hat{F}_3(t_0)$ can be obtained from (A 1.18) by an iterative procedure

$$\hat{F}_3^{(k+1)}=D_{22}'-(D_{33}-I)\hat{F}_3^{(k)}. \tag{A1.24}$$

Then, from $\rho(D_{33}-I)<1$,

$$\hat{F}_3^{(k)}\to\hat{F}_3(t_0) \quad (k\to\infty) \tag{A1.25}$$

holds true regardless of the choice of the initial $F_3^{(0)}$. Similar procedures can be applied to obtain $\hat{G}_3(t_0)$, $\hat{G}_2(t_0)$, and $\hat{H}_2(t_0)$. These iterative procedures would be effective in the case where the order n and the number of inputs m are relatively large.

Appendix 2

In this appendix, the proof of the inequality (4.82) is given. A similar proof of (4.83) is omitted for the sake of brevity.

It follows from (2.7), (2.9), and Lemma 3 that

$$
\begin{aligned}
&B_k'S_k - B'(t_k)S(t_k) \\
&= \int_0^{t_k} (B_c+E_4)'\exp\{(A_c+E_1)'s\}ds \int_0^{t_k} \exp\{(A_c+E_1)'s\}E_6 ds \\
&\quad + \int_0^{t_k} (B_c+E_4)'\exp\{(A_c+E_1)'s\}ds \\
&\qquad \times \int_0^{t_k} \exp\{(A_c+E_1)'s\}E_3 \int_0^s \exp\{(A_c+E_1)(s-r)\}(B_c+E_4)drds \\
&\quad + \int_0^{t_k} (B_c+E_4)'\exp\{(A_c+E_1)'s\}ds \\
&\qquad \times \int_0^{t_k} \exp\{(A_c+E_1)'s\}Q_c \int_0^s \exp\{(A_c+E_1)(s-r)\}E_4 drds \\
&\quad + \int_0^{t_k} E_4'\exp\{(A_c+E_1)'s\}ds \\
&\qquad \times \int_0^{t_k} \exp\{(A_c+E_1)'s\}Q_c \int_0^s \exp\{(A_c+E_1)(s-r)\}B_c drds \\
&\quad + \int_0^{t_k} B_c'[\exp\{(A_c+E_1)s\} - \exp(A_c s)]'ds \\
&\qquad \times \int_0^{t_k} \exp\{(A_c+E_1)'s\}Q_c \int_0^s \exp\{(A_c+E_1)(s-r)\}B_c drds \\
&\quad + \int_0^{t_k} B_c'\exp(A_c's)ds \\
&\qquad \times \int_0^{t_k} [\exp\{(A_c+E_1)s\} - \exp(A_c s)]'Q_c \\
&\qquad \times \int_0^s \exp\{(A_c+E_1)(s-r)\}B_c drds \\
&\quad + \int_0^{t_k} B_c'\exp(A_c's)ds \\
&\qquad \times \int_0^{t_k} \exp(A_c's)Q_c \\
&\qquad \times \int_0^s [\exp\{(A_c+E_1)(s-r)\} - \exp\{A_c(s-r)\}]B_c drds. \qquad \text{(A2.1)}
\end{aligned}
$$

Taking the norms of both sides, we obtain

$$
\begin{aligned}
&\| B_k'S_k - B'(t_k)S(t_k) \| \\
&\quad \leqq (\alpha+\varepsilon)\theta(t_k)\exp(\varepsilon t_k)t_k \cdot \theta(t_k)\exp(\varepsilon t_k)\varepsilon t_k \\
&\qquad + (\alpha+\varepsilon)\theta(t_k)\exp(\varepsilon t_k)t_k \cdot \theta(t_k)\exp(\varepsilon t_k)\varepsilon\theta(t_k)\exp(\varepsilon t_k)(\alpha+\varepsilon)t_k^2/2 \\
&\qquad + (\alpha+\varepsilon)\theta(t_k)\exp(\varepsilon t_k)t_k \cdot \theta(t_k)\exp(\varepsilon t_k)\alpha\theta(t_k)\exp(\varepsilon t_k)\varepsilon t_k^2/2 \\
&\qquad + \varepsilon\theta(t_k)\exp(\varepsilon t_k)t_k \cdot \theta(t_k)\exp(\varepsilon t_k)\alpha\theta(t_k)\exp(\varepsilon t_k)\alpha t_k^2/2 \\
&\qquad + \alpha\theta(t_k)\varepsilon\exp(\varepsilon t_k)t_k^2/2 \cdot \theta(t_k)\exp(\varepsilon t_k)\alpha\theta(t_k)\exp(\varepsilon t_k)\alpha t_k^2/2 \\
&\qquad + \alpha\theta(t_k)t_k \cdot \theta(t_k)\varepsilon\exp(\varepsilon t_k)\alpha\theta(t_k)\exp(\varepsilon t_k)\alpha t_k^3/3 \\
&\qquad + \alpha\theta(t_k)t_k \cdot \theta(t_k)\alpha\theta(t_k)\varepsilon\exp(\varepsilon t_k)\alpha t_k^3/6 \\
&\quad \leqq \varepsilon\theta^3(t_k)\exp(3\varepsilon t_k)t_k^2(\alpha+\varepsilon)[1+(\alpha+\varepsilon)t_k/2+\alpha t_k/2 \\
&\qquad\qquad + \alpha t_k/2+\alpha^2 t_k^2/4+\alpha^2 t_k^2/3+\alpha^2 t_k^2/6] \\
&\quad \leqq \varepsilon\theta^3(t_k)\exp(3\varepsilon t_k)t_k^2(\alpha+\varepsilon)[1+3(\alpha+\varepsilon)t_k/2+3(\alpha+\varepsilon)^2 t_k^2/4]. \qquad \text{(A2.2)}
\end{aligned}
$$

This completes the proof of (4.82).