Chapter

New Matrix Series Formulae for Matrix Exponentials and for the Solution of Linear Systems of Algebraic Equations

Ioan R. Ciric

Abstract

The solution of certain differential equations is expressed using a special type of matrix series and is directly related to the solution of general systems of algebraic equations. Efficient formulae for matrix exponentials are derived in terms of rapidly convergent series of the same type. They are essential for two new solution methods, especially beneficial for large linear systems, namely an iterative method and a method based on an exact matrix product formula. The computational complexity of these two methods is analysed, and for both of them, the number of matrix exponential-vector multiplications required for an imposed accuracy can be predetermined in terms of the system condition. The total number of arithmetic operations involved is roughly proportional to n^2, where n is the matrix dimension. The common feature of all the series in the results presented is that starting with a first term that is already well-conditioned, each subsequent term is computed by multiplication with an even better conditioned matrix, tending quickly to the identity matrix. This contributes substantially to the stability of the numerical computation. A very efficient method based on the numerical integration of a special kind of differential equations, applicable to even ill-conditioned systems, is also presented.

Keywords: matrix equations, matrix exponentials, numerical solutions

1. Introduction

New matrix series expressions were recently derived by the author [1] for the solution of simple first order differential equations associated with general systems of linear algebraic equations. These differential equations describe the orthogonal trajectories of a family of hypersurfaces that represent a quadratic functional related to the linear algebraic system. The solution of the latter can be obtained by minimizing the functional along an orthogonal trajectory instead of applying various techniques based on minimization along conjugate gradient directions or based on minimized iterations [2]. Since the solutions of the differential equations considered are simply related to the solutions of the corresponding algebraic systems through matrix exponentials, there is the possibility to develop efficient solution methods if only the matrix exponentials could be used in numerical calculations accurately and with a small computational effort. A survey of various existent algorithms for computing matrix exponentials and a useful discussion of the difficulties involved are presented in [3].

In the present work, we use new formulae for arbitrary matrix exponentials that contain highly convergent infinite series which allow accurate and stable numerical computations. Employing these formulae, two new solution methods are proposed which are particularly efficient for large-scale general linear algebraic systems.

2. Differential equations associated with linear systems of algebraic equations

We start with simple vector differential equations whose solutions are related to the solution of general systems of linear algebraic equations. Later, in Section 5 we construct differential equations that allow an efficient numerical integration in order to obtain the solution of these systems.

2.1 Matrix series solution of some vector differential equations

Consider a first order vector differential equation of the form

$$\frac{dx}{dv} = f(v)(Ax - b) \tag{1}$$

where $A \in R^{n\times n}$ is a general nonsingular matrix, $b \in R^n$ is a given n-dimensional vector, $x = (x_1, x_2, \dots, x_n)^T$ is the unknown n-dimensional vector, T indicates the transpose, and $f(v)$ is a continuous function over some interval of the real variable v. With the condition

$$x(v_o) = x_o \tag{2}$$

x_o being a given vector, (1) has a unique solution over the interval considered [4]. Integrating both sides of (1) from v_o to v yields

$$Ax(v) - b = e^{A[g(v)-g(v_o)]}(Ax_o - b) \tag{3}$$

where

$$g(v) \equiv \int f(v)\,dv \tag{4}$$

is a primitive for $f(v)$, i.e., $f(v) = dg/dv$. Thus, (1) can be written in the form

$$\frac{dx}{dv} = f(v)e^{A[g(v)-g(v_o)]}(Ax_o - b) \tag{5}$$

Integrating again both sides from v_o to v gives

$$x(v) = x_o + e^{-Ag(v_o)}\sum_{k=0}^{\infty}\frac{A^k}{(k+1)!}\left[(g(v))^{k+1} - (g(v_o))^{k+1}\right](Ax_o - b) \tag{6}$$

Two particular cases are presented.

If $f(v)$ is chosen to be $f(v) = -1$ then $g(v) = -v$. Choosing $v_o = 0$ gives $g(v_o) = 0$ and (6) becomes

$$x(v) = e^{-Av}x_o + v\sum_{k=0}^{\infty}(-1)^k \frac{(Av)^k}{(k+1)!}b \tag{7}$$

If $f(v) = 1/v$ then $g(v) = \ln v$. With $v_o = 1$ and $g(v_o) = 0$ (6) becomes now

$$x(v) = x_o + (\ln v)\sum_{k=0}^{\infty}\frac{(A\ln v)^k}{(k+1)!}(Ax_o - b) \tag{8}$$

The solution of (1) with $f(v) = 1/v$ can also be obtained by employing a Taylor series expansion about $x_o = 1$. Indeed,

$$x(v) = x(1) + \frac{(v-1)}{1!}x^{(1)}(1) + \frac{(v-1)^2}{2!}x^{(2)}(1) + \cdots + \frac{(v-1)^k}{k!}x^{(k)}(1) + \cdots \tag{9}$$

and evaluating the derivatives yields

$$x(v) = x_o - (1-v)\left[I + \sum_{k=1}^{\infty}\frac{(1-v)^k}{k+1}(I-A)\left(I - \frac{A}{2}\right)\cdots\left(I - \frac{A}{k}\right)\right](Ax_o - b) \tag{10}$$

where I is the identity matrix. The series in (10) is convergent for $v \in (0, 2)$ but its rate of convergence is, in general, very small for v very close to zero.

2.2 Relationship with linear systems of algebraic equations

Consider now a system of equations written in matrix form as

$$Ax = b \tag{11}$$

and assume that x is a continuous function of the real variable v over a certain interval. The solution of (11) can be obtained from (3) for a $v \equiv v_S$ for which

$$Ax(v_S) - b = 0 \tag{12}$$

To satisfy this condition $g(v)$ in (3) must be chosen such that

$$e^{A[g(v_s) - g(v_o)]} = 0 \tag{13}$$

For positive definite matrices A and a finite $g(v_o)$ this is achieved if $g(v) \to -\infty$ for $v \to v_S$. Thus, the solution of (11) cannot be computed directly from (6). On the other hand, in the particular case of $f(v) = 1/v, g(v) = \ln v$ the solution of (11) can be obtained from (10) with $v = v_S = 0$,

$$x_s = x_o - \left[I + \sum_{k=1}^{\infty}\frac{1}{k+1}(I-A)\left(I - \frac{A}{2}\right)\cdots\left(I - \frac{A}{k}\right)\right](Ax_o - b) \tag{14}$$

The expression in the brackets is just the inverse of A,

$$A^{-1} = I + \sum_{k=1}^{\infty} \frac{1}{k+1}(I-A)\left(I-\frac{A}{2}\right)\cdots\left(I-\frac{A}{k}\right) \tag{15}$$

The rate of convergence of the series in (14) and (15) is very small and they cannot be practically used in numerical computation for arbitrary matrices.

We note that the solution of (11) can be formally expressed from (6) as

$$x_s = x_o + e^{Ag(v_o)}\left(e^{Ag(v_o)} - e^{Ag(v)}\right)^{-1}\left(x(v) - x_o\right) \tag{16}$$

which is valid for any $v \neq v_o$ in the interval considered. To compute x_s with (16) would require the computation of $\left(e^{Ag(v_o)} - e^{Ag(v)}\right)^{-1}$. Equation (16) will be used in Subsection 4.1.

Note. The matrix product in the terms of the series in (10), (14) and (15) can be expressed as a matrix polynomial using the relationship with the Stirling numbers of the first kind $S_{k+1}^{(m)}$ [5]

$$(I-A)\left(I-\frac{A}{2}\right)\cdots\left(I-\frac{A}{k}\right) = (-1)^k \frac{1}{k!}\sum_{m=1}^{k+1} S_{k+1}^{(m)} A^{m-1} \tag{17}$$

While each new term in such series as those in (10), (14) and (15) is calculated through a multiplication with a matrix that becomes more and more well-conditioned as k increases, the computation with the expression in (17) would require successive multiplications with the same original matrix and, for each k, a new polynomial is to be constructed and new Stirling numbers have to be generated. The formulae derived in the next two sections contain the same type of series and, therefore, are simpler and more efficient to be used for numerical computations.

3. New formulae for computing matrix exponentials

In this section, we derive matrix exponential expressions which contain highly convergent infinite series that allow accurate and stable numerical computations in numerous applications. They shall also be used in the next section.

3.1 Series expressions for matrix exponentials

Consider the matrix function $v^A \equiv e^{A \ln v}$ where v is a positive real variable and A a general square matrix with real number entries. By integration,

$$\int_1^v v^{A-I} dv = \left.\frac{v^A}{A}\right|_{v=1}^{v} = \frac{1}{A}\left(e^{A\ln v} - I\right) \tag{18}$$

For $v \in (0, 2)$ the integrand can be expanded in a power series as

$$\begin{aligned} v^{A-I} = \left[1-(1-v\right]^{A-I} = I - \frac{(1-v)}{1!}(A-I) + \frac{(1-v)^2}{2!}(A-I)(A-2I) \\ - \frac{(1-v)^3}{3!}(A-I)(A-2I)(A-3I) + \cdots \end{aligned} \tag{19}$$

that can be integrated term by term. From (18) and (19)

$$e^{A\ln v} = I-(1-v)A\left[I+\sum_{k=1}^{\infty}\frac{(1-v)^k}{k+1}(I-A)\left(I-\frac{A}{2}\right)\cdots\left(I-\frac{A}{k}\right)\right] \tag{20}$$

which is valid for any A, positive definite or not, of arbitrary condition. One can see that, for a positive definite A and for $v\to 0$, the expression in the brackets of (20) gives a series expansion for the inverse A^{-1} (see (14) and (15)). Various expressions for the matrix exponential are obtained by giving particular values to v in (20). For example, for any $v=e^{-q}, q>-\ln 2$, we have

$$e^{-qA} = I-(1-e^{-q})A\left[I+\sum_{k=1}^{\infty}\frac{(1-e^{-q})^k}{k+1}(I-A)\left(I-\frac{A}{2}\right)\cdots\left(I-\frac{A}{k}\right)\right] \tag{21}$$

the series becoming less convergent as q increases above q = 0. On the other hand, with $v=1/e$ in (20) and then A replaced with qA we obtain for any real number q

$$e^{-qA} = I-(1-e^{-1})qA\left[I+\sum_{k=1}^{\infty}\frac{(1-e^{-1})^k}{k+1}(I-qA)\left(I-\frac{qA}{2}\right)\cdots\left(I-\frac{qA}{k}\right)\right] \tag{22}$$

this series being more convergent than that in (21) for greater values of q.

For any $v\in(1,2)$, the terms in the series expressions derived from (20) have coefficients that are alternately positive and negative. With $v=e^{1/2}$ = 1.64872127, e.g., and then replacing A with $2A$ we have

$$e^{A} = I+2(e^{1/2}-1)A\left[I+\sum_{k=1}^{\infty}\frac{(-1)^k(e^{1/2}-1)^k}{k+1}(I-2A)\left(I-\frac{2A}{2}\right)\cdots\left(I-\frac{2A}{k}\right)\right] \tag{23}$$

As well, by replacing A by $-qA$ in (23) one obtains instead of (22) an expression with alternating in sign series coefficients.

3.2 Rapidly convergent series formulae

From the basic Eq. (20) we derive now formulae which contain series that have a higher rate of convergence than those presented in the previous subsection.

Firstly, it is obvious that for values of v close to 1 the series in (20) has a high rate of convergence. For instance, with $v=1+10^{-q}, 10^{-q}\ll 1$, and replacing $A\ln(1+10^{-q})$ by A we obtain

$$e^{A} = I+10^{-q}c_qA\left[I+\sum_{k=1}^{\infty}\frac{(-1)^k 10^{-qk}}{k+1}(I-c_qA)\left(I-\frac{c_qA}{2}\right)\cdots\left(I-\frac{c_qA}{k}\right)\right] \tag{24}$$

where $c_q\equiv 1/\ln(1+10^{-q})$. This series is rapidly convergent.

Secondly, the convergence of the series in the expressions derived from (20) can further be improved by successive integrations. Indeed, integrating both sides of (20) from 1 to v, we have

$$\begin{aligned} ve^{A\ln v} = I+(1-v)(I+A)\Bigg[&-I+\frac{1-v}{2}A \\ &+A\sum_{k=1}^{\infty}\frac{(1-v)^{k+1}}{(k+1)(k+2)}(I-A)\left(I-\frac{A}{2}\right)\cdots\left(I-\frac{A}{k}\right)\Bigg] \end{aligned} \tag{25}$$

Integrating repeatedly we obtain the identity

$$\begin{aligned} v^p e^{A\ln v} &= I + p!\sum_{k=1}^{p}\frac{(-1)^k(1-v)^k}{k!(p-k)!}\left(I+\frac{A}{p}\right)\left(I+\frac{A}{p-1}\right)\cdots\left(I+\frac{A}{p-k+1}\right) \\ &\quad +(-1)^{p+1}(1-v)^{p+1}A(I+A)\left(I+\frac{A}{2}\right)\cdots\left(I+\frac{A}{p}\right)\left[\frac{1}{p+1}I\right. \\ &\quad \left.+\,p!\sum_{k=1}^{\infty}\frac{(1-v)^k}{(k+1)(k+2)\cdots(k+p+1)}(I-A)\left(I-\frac{A}{2}\right)\cdots\left(I-\frac{A}{k}\right)\right], \\ &\qquad\qquad p=1,2,\ldots;\quad v\in(0,2) \end{aligned} \tag{26}$$

Same result is obtained by replacing A with $pI + A$ in (20). This identity contains an infinite series whose coefficients decrease rapidly as p increases. Obviously, for a given A and v, (26) generates more efficient computational formulae than those in the previous subsection. As before, for $v \in (1, 2)$ the infinite series have coefficients that alternate in sign. For example, with $v = e^{1/2}$ in (26) and then replacing A by $2A$ and p by $2p$ we have instead of (23)

$$\begin{aligned} e^A &= e^{-p}\left\{I+(2p)!\sum_{k=1}^{2p}\frac{(e^{1/2}-1)^k}{k!(2p-k)!}\left(I+\frac{2A}{2p}\right)\left(I+\frac{2A}{2p-1}\right)\cdots\left(I+\frac{2A}{2p-k+1}\right)\right. \\ &\quad +2(e^{1/2}-1)^{2p+1}A(I+2A)\left(I+\frac{2A}{2}\right)\cdots\left(I+\frac{2A}{2p}\right)\left[\frac{1}{2p+1}I\right. \\ &\quad \left.\left.+(2p)!\sum_{k=1}^{\infty}\frac{(-1)^k(e^{1/2}-1)^k}{(k+1)(k+2)\cdots(k+2p+1)}(I-2A)\left(I-\frac{2A}{2}\right)\cdots\left(I-\frac{2A}{k}\right)\right]\right\}, \\ &\qquad\qquad p=1,2,\ldots \end{aligned} \tag{27}$$

Taking $v = 1 + 10^{-q}$, with $q>0$ and $c_q \equiv 1/\ln(1 + 10^{-q})$, (26) gives (compare with (24))

$$\begin{aligned} e^A &= (1+10^{-q})^{-p}\left\{I+p!\sum_{k=1}^{p}\frac{10^{-qk}}{k!(p-k)!}\left(I+\frac{c_qA}{p}\right)\left(I+\frac{c_qA}{p-1}\right)\cdots\left(I+\frac{c_qA}{p-k+1}\right)\right. \\ &\quad +10^{-(p+1)q}c_qA(I+c_qA)\left(I+\frac{c_qA}{2}\right)\cdots\left(I+\frac{c_qA}{p}\right)\left[\frac{1}{p+1}I\right. \\ &\quad \left.\left.+p!\sum_{k=1}^{\infty}\frac{(-1)^k10^{-qk}}{(k+1)(k+2)\cdots(k+p+1)}(I-c_qA)\left(I-\frac{c_qA}{2}\right)\cdots\left(I-\frac{c_qA}{k}\right)\right]\right\}, \\ &\qquad\qquad p=1,2,\ldots \end{aligned} \tag{28}$$

Notice that, in the new formulae derived from (26) the infinite series are very rapidly convergent, with their rate of convergence increasing when the parameter p increases. Highly accurate numerical results can be generated with only a small number of terms retained in the infinite series of these formulae (see Section 4).

Note. All the formulae presented in this section remain valid if A is changed in $-A$. Obviously, in all these expressions A can be replaced by a real number and the identity matrix I by 1, yielding a few novel identities and summation formulae for series of real numbers. Also, the expression in the brackets of (26) for $v \to 0$ is just $(I + A/p)^{-1}/p$ if $I + A/p$ is positive definite and, thus, we obtain another identity, i.e.,

$$\begin{aligned} p!\sum_{k=1}^{p}\frac{(-1)^{k+1}}{k!(p-k)!}\left(I+\frac{A}{p}\right)\left(I+\frac{A}{p-1}\right)\cdots\left(I+\frac{A}{p-k+1}\right) \\ = I+\frac{(-1)^{p+1}}{p}A(I+A)\left(I+\frac{A}{2}\right)\cdots\left(I+\frac{A}{p-1}\right),\quad p=2,3,\dots \end{aligned} \tag{29}$$

which reduces for $A = 0$ to an elementary binomial sum.

4. Solution of general linear systems

In what follows, we apply the matrix exponential formulae from the previous section and present a new iteration procedure and a matrix product formula for the solution of large systems of linear algebraic equations.

4.1 An iterative method

Equation (16) can be written in the form

$$x(v)=x_o+\left(I-e^{A\left[g(v)-g(v_o)\right]}\right)\left(x_S-x_o\right) \tag{30}$$

where $x_S = x(v_S)$ is the solution of (11), $x_o = x(v_o)$, A is a positive definite matrix and $g(v)$ is a function of the real variable v such that $g(v) \to -\infty$ for $v \to v_s$ (see Subsection 2.2). To get $x(v)$ for values of v very close to v_S we choose an adequate $g(v)$ and a formula for the matrix exponential from Section 3. When $g(v_o) = 0$, applying (22) for instance gives

$$\begin{aligned} x(v)=x_o+(1-e^{-1})g(v)\Bigg[I+\sum_{k=1}^{\infty}\frac{(1-e^{-1})^k}{k+1}(I+Ag(v)) \\ \times\left(I+\frac{A}{2}g(v)\right)\cdots\left(I+\frac{A}{k}g(v)\right)\Bigg](Ax_o-b) \end{aligned} \tag{31}$$

If $g(v) \equiv \ln v$, with $v_o = 1$ and $v_s = 0$, we compute $x\left(e^{-N}\right)$ for $N \gg 1$ which is closer to the solution x_S,

$$\begin{aligned} x(e^{-N})\equiv x_S^{(1)}=x_S^{(0)}-N(1-e^{-1})\Bigg[I+\sum_{k=1}^{\infty}\frac{(1-e^{-1})^k}{k+1}(I-NA) \\ \times\left(I-\frac{N}{2}A\right)\cdots\left(I-\frac{N}{k}A\right)\Bigg](Ax_S^{(0)}-b) \end{aligned} \tag{32}$$

where $x_S^{(0)} \equiv x_o$. This equation is applied iteratively by replacing $x_S^{(1)}$ and $x_S^{(0)}$, respectively, with $x_S^{(i)}$ and $x_S^{(i-1)}$, $i = 2, 3, \dots$, until $x_S^{(i)}$ satisfies (11) with a desired accuracy.

To evaluate the amount of computation necessary to obtain the solution of (11) with a certain accuracy, let us take N such that $N\|A\| = 10$ when one needs about 30 terms in the infinite series, i.e., 30 matrix-vector multiplications. The number of iterations increases with the condition number of A. To see this and to determine the corresponding number of iterations, consider (11) with $b = 0$ and A replaced with a diagonal matrix whose entries are positive numbers, the greatest of these being 1, and whose condition is the same as that of A. The solution of this system is

$x_S = 0$ and the components of the solution of (1), with $f(v) = 1/v$, are $x_{o_k} v^{\lambda_k}$ where $\lambda_k, k = 1, 2, \ldots, n$, are the entries in the diagonal matrix. In order to make the magnitudes of all these components at least 1%, e.g., of the corresponding magnitudes of the initial components, one needs no iteration if the condition number is less than 2, but 5 and, respectively, 46 iterations are needed if the condition number is 10 and 100.

What is remarkable in the iterative method based on (32) is that, for matrices with same condition number and same norm, the number of iterations required is the same, independently of the size of the matrices. Considering approximately $2n^2$ arithmetic operations for one matrix-vector multiplication, where n is the number of equations and unknowns in (11), the total number of arithmetic operations required is, thus, proportional to only n^2. In the examples given above one has to perform, respectively, $60n^2$, $300n^2$ and $2760n^2$ arithmetic operations. Assuming only $2n^3/3$ arithmetic operations for the Gaussian elimination procedure, the method presented in this subsection requires less computation for the same examples if, respectively, $n>90$, $n>450$ and $n>4140$. One can also notice that the application of Eq. (32) leads to the actual solution of (11) independently of the small error introduced in the computation at each iteration.

4.2 A matrix product formula

The original general system (11) is replaced with an equivalent system such that its solution is obtained in terms of matrix exponentials for which highly convergent and accurate series formulae have been derived in Section 3.

Namely, instead of (11) we use the system

$$\left(I - e^{-\alpha A}\right)x = b_\alpha \tag{33}$$

where α is a real scalar to be chosen, $\alpha \neq 0$, and

$$b_\alpha = \alpha\left[1 - \frac{(\alpha A)}{2!} + \frac{(\alpha A)^2}{3!} - \frac{(\alpha A)^3}{4!} + \cdots\right]b \tag{34}$$

Assuming A to be positive definite, α is taken positive. Then, since $\left\|e^{-\alpha A}\right\| < 1$ for a normal matrix, the solution can be expressed as

$$x_s = \left(I - e^{-\alpha A}\right)^{-1} b_\alpha = \left[\sum_{k=0}^{\infty} e^{-k\alpha A}\right] b_\alpha \tag{35}$$

with the norm of the matrix exponentials decreasing when k increases [6]

$$\left\|e^{-k\alpha A}\right\| < e^{-k\alpha\lambda} \tag{36}$$

where λ is the smallest eigenvalue of A. b_α can be accurately computed by using instead of (34) an equivalent expression, for instance (see (22))

$$b_\alpha = (1 - e^{-1})\alpha\left[I + \sum_{k=1}^{\infty} \frac{(1 - e^{-1})^k}{k+1}(I - \alpha A)\left(I - \frac{\alpha A}{2}\right)\cdots\left(I - \frac{\alpha A}{k}\right)\right]b \tag{37}$$

If the infinite series in (35) is truncated to $k = N_S$ the rest of the series has a norm

$$\left\|R_{N_s}\right\| < \frac{e^{-(N_s+1)\alpha\lambda}}{1-e^{-\alpha\lambda}} \tag{38}$$

Much less numerical computation (see below) is needed if the infinite series in (35) is transformed into an infinite product using the identity [5]

$$(1-u)^{-1} = \prod_{k=0}^{\infty}(1+u^{2^k}), \quad 0<u<1 \tag{39}$$

which is also valid for matrices whose norm is less than 1. Thus, (35) becomes

$$x_s = \left[\prod_{k=0}^{\infty}(1+e^{-2^k\alpha A})\right]b_\alpha \tag{40}$$

with the norm of the exponentials $e^{-2^k\alpha A}$ decreasing very rapidly when k increases. Truncating the infinite product to $k = N_P$, i.e., $N_P + 1$ factors, leaves a remaining factor

$$F_{N_P} = \prod_{k=N_P+1}^{\infty}(I+e^{-2^k\alpha A}) \tag{41}$$

whose norm is

$$\left\|F_{N_P}\right\| < \prod_{k=N_P+1}^{\infty}(1+e^{-2^k\alpha\lambda}) = \frac{1}{1-e^{-\alpha\lambda}}\frac{1}{\prod_{k=0}^{N_P}(1+e^{-2^k\alpha\lambda})} \tag{42}$$

Let us compare the maximum value of the norm of the truncated matrix in the brackets of (35) and (40) with that of the corresponding untruncated matrix in order to get a rough estimate of the numbers N_S and N_P of matrix exponentials involved in the numerical computation to achieve a certain accuracy. This will also allow to estimate the total number of matrix-vector multiplications necessary to obtain the solution. The ratio of the maximum norm of the truncated matrix to the maximum norm of the untruncated matrix in the brackets of (35) and (40) is, respectively, (see (38) and (42))

$$\rho = 1-e^{-(N_S+1)\alpha\lambda} \tag{43}$$

and

$$\rho = (1-e^{-\alpha\lambda})\prod_{k=0}^{N_P}(1+e^{-2^k\alpha\lambda}) \tag{44}$$

To illustrate the computation complexity when using (35) or (40), assume that $\|A\| = 1$ and $\alpha = 20$. If ρ is imposed to be ≈ 0.99, for example, one needs $N_S = 2$, i.e., three terms in (35) and $N_P = 1$, i.e., two factors in (40) if $\lambda = 10^{-1}$. If $\lambda = 10^{-2}$ these numbers increase to $N_S = 23$ and $N_P = 4$, and when $\lambda = 10^{-3}$ one gets $N_S = 230$, but N_P only increases to $N_P = 7$. It is clear that applying the formula (40) the number of exponentials needed in the numerical computation is much smaller than that if formula (35) would be applied. For all the matrix exponentials involved in the numerical computation we use the formula (28) containing a highly convergent series, such that

$$
\begin{aligned}
e^{-\alpha A} = (1+10^{-q})^{-p} \Bigg\{ & I + p!\sum_{k=1}^{p} \frac{10^{-qk}}{k!(p-k)!}\left(I-\frac{c_q\alpha A}{p}\right)\left(I-\frac{c_q\alpha A}{p-1}\right)\cdots\left(I-\frac{c_q\alpha A}{p-k+1}\right) \\
& -10^{-(p+1)q} c_q\alpha A(I-c_q\alpha A)\left(I-\frac{c_q\alpha A}{2}\right)\cdots\left(I-\frac{c_q\alpha A}{p}\right)\Bigg[\frac{I}{p+1} \\
& + p!\sum_{k=1}^{\infty}\frac{(-1)^k 10^{-qk}k!}{(k+p+1)!}(I+c_q\alpha A)\left(I+\frac{c_q\alpha A}{2}\right)\cdots\left(I+\frac{c_q\alpha A}{k}\right)\Bigg]\Bigg\}, \\
& \qquad\qquad p = 1,2,\ldots
\end{aligned}
\tag{45}
$$

where $q>o$ and $c_q \equiv 1/\ln(1+10^{-q})$. With $\alpha\|A\| = 20$ and choosing $q = 1$ and $p = 10$, for instance, $e^{-\alpha A}$ is determined accurately by retaining 50 terms in the infinite series and, thus, to multiply $e^{-\alpha A}$ with a vector one needs 71 matrix-vector multiplications. To compute x_s from (40) one has to use repeatedly the multiplication of e^{-20A} with a vector. For a matrix A with $\lambda = 10^{-1}$ one has to retain two factors in (40) and, thus, to multiply e^{-20A} and $e^{-2\times 20A}$ with a vector. This means to use repeatedly 3 times the multiplication of e^{-20A} with a vector which requires, therefore, $3 \times 71 = 213$ matrix-vector multiplications. When $\lambda = 10^{-2}$ the infinite product in (40) is truncated at $k = N_P = 4$ and this requires the multiplication of $e^{-2^k\times 20A}$, $k = 0,1,2,3,4$, with a vector, i.e., to use repeatedly 31 times the multiplication of e^{-20A} with a vector, for a total of $31 \times 71 = 2201$ matrix-vector multiplications. We also have to add the matrix-vector multiplications required to compute b_α in (37). A very accurate result for b_α when $\alpha = 20$ can be achieved by applying four times the series in the brackets of (37) for $\alpha = 5$, each time retaining 30 terms. This requires a total of about 120 multiplications of a matrix $I - 5A/k$ with a vector. In all the matrix-vector multiplications involved when applying (45), the matrices are in the form $I \pm c_q\alpha A/k$ and become better and better conditioned as k increases.

Adding up the number of arithmetic operations involved shows that, with respect to the classical Gaussian elimination method, the procedure presented in this subsection is advantageous for very large systems (11). Namely, assuming same accuracy and only $2n^3/3$ arithmetic operations for the Gaussian elimination, with the data given above, one has to have $n>3 \times (213 + 120) = 999$ equations and unknowns if $\lambda = 10^{-1}$ and $n>3 \times (2201 + 120) = 6963$ equations and unknowns if $\lambda = 10^{-2}$ for the proposed method to be more advantageous. For a given $\alpha\|A\|$, one application of $e^{-\alpha A}$ requires a determined finite number of matrix-vector multiplications, independently of the size of A. It is remarkable, as for the iterative method in the previous subsection, that for a given condition of A, one has to apply $e^{-\alpha A}$ a well-determined number of times and, thus, the total number of arithmetic operations necessary to compute the solution with an imposed accuracy is proportional to only n^2.

It should be noted that, since the infinite series in the expression (45) is truncated and thus determined with a finite accuracy, the accuracy of the solution x_S becomes compromised after a too big a number of matrix exponential-vector multiplications. This is why, the worse conditioned systems (11) should be appropriately preconditioned. Practically, the computation with (40) is continued factor by factor and the accuracy of x is checked after each step.

5. Solution of general linear systems by numerical integration of differential equations

In this section, we introduce first order differential equations whose numerical integration allows to efficiently find the solution of linear systems of algebraic equations. Differential equations of the type of those in (1), with $f(v) = -1$ or $f(v) = 1/v$, cannot be used for this purpose due to the fact that the first and higher order derivatives of $x(v)$ tend to infinite values as x tends to the solution x_S of (11) (see Section 2).

Here below, we construct ordinary differential equations for $x(v)$ which satisfy the condition that the first few derivatives are finite when $x(v)$ tends to x_S and, therefore, are particularly useful for an accurate computation of x_S. Let us consider the system (11) with a symmetric positive definite matrix. A quadratic functional

$$F(x) = \frac{1}{2}x^T Ax - x^T b + \frac{1}{2}b^T A^{-1}b \tag{46}$$

is associated with (11) [6] whose minimum value is $F(x_S) = 0$. Define now a real variable $v, v \geq 0$, such that

$$F(x) = v^r \tag{47}$$

where r is a real number to be chosen, $r>0$, with $v = 0$ corresponding to the solution $x = x_S$ and $v = v_o$ to an initial point $x_o, F(x_o) = v_o^r$. Then,

$$\frac{dF}{dv} = rv^{r-1} = (Ax - b)^T \frac{dx}{dv} \tag{48}$$

and, thus,

$$\frac{dx}{dv} = rv^{r-1}\frac{Ax - b}{\left\|Ax - b\right\|^2} \tag{49}$$

This is the differential equation to be integrated from $v = v_o$ to $v = v_S = 0$. The second derivative of x is obtained in the form

$$\begin{aligned}\frac{d^2x}{dv^2} = r(r-1)v^{r-2}\frac{Ax - b}{\left\|Ax - b\right\|^2} + (rv^{r-1})^2\frac{1}{\left\|Ax - b\right\|^4}\Bigg\{ A \\ -2\frac{\left[(Ax - b)^T A(Ax - b)\right]}{\left\|Ax - b\right\|^2}\Bigg\}(Ax - b)\end{aligned} \tag{50}$$

Higher order derivatives can be worked out if needed.

In order to see the behaviour of the derivatives close to the solution x_S, Eqs. (49) and (50) are rewritten as

$$\frac{dx}{dv}=r\frac{(F(x))^{1-\frac{1}{r}}(Ax-b)}{\left\|Ax-b\right\|^2} \tag{51}$$

and

$$\begin{aligned}\frac{d^2x}{dv^2}=r(r-1)\frac{(F(x))^{1-\frac{2}{r}}(Ax-b)}{\left\|Ax-b\right\|^2}+r^2\frac{(F(x))^{2-\frac{2}{r}}}{\left\|Ax-b\right\|^4}\Bigg\{&A\\ -2\frac{\left[(Ax-b)^T A(Ax-b)\right]}{\left\|Ax-b\right\|^2}\Bigg\}&(Ax-b)\end{aligned} \tag{52}$$

with $F(x)$ in (46) put in the form

$$F(x)=\frac{1}{2}(Ax-b)^T(x-x_S) \tag{53}$$

Notice that as x tends to x_s, when $\|Ax-b\|\equiv\varepsilon$ tends to zero,

$$\left\|\frac{dx}{dv}\right\|\sim K_1(v)\varepsilon^{1-\frac{2}{r}},\quad \left\|\frac{d^2x}{dv^2}\right\|\sim K_2(v)\varepsilon^{1-\frac{4}{r}} \tag{54}$$

where $K_1(v)$ and $K_2(v)$ are finite when $v\to 0$. Therefore, as $x\to x_S$, $\|dx/dv\|\to 0$ if $r>2$ and $\|d^2x/dv^2\|\to 0$ if $r>4$.

Another differential equation we present here is

$$\frac{dx}{ds}=\frac{Ax-b}{\left\|Ax-b\right\|} \tag{55}$$

with the second derivative

$$\frac{d^2x}{ds^2}=\frac{1}{\left\|Ax-b\right\|^2}\left\{A-\frac{[(Ax-b)^T A(Ax-b)]}{\left\|Ax-b\right\|^2}\right\}(Ax-b) \tag{56}$$

In this case, always

$$\left\|\frac{dx}{ds}\right\|=1 \tag{57}$$

even for $x\to x_S$, but the second derivative tends to an infinite value when $x\to x_S$

$$\left\|\frac{d^2x}{ds^2}\right\|\sim\ K(s)\varepsilon^{-1} \tag{58}$$

where $K(s)$ is finite and $\varepsilon\equiv\|Ax-b\|$. The relationship between the differentials of the variables v and s in (49) and (55) is

$$ds = \frac{r(F(x))^{1-\frac{1}{r}}}{\left\|Ax-b\right\|} dv \tag{59}$$

and for $x \to x_S$ we have (see (54))

$$\frac{ds}{dv} \sim K_1(v)\,\varepsilon^{1-\frac{2}{r}} \tag{60}$$

The differential Eqs. (49) in v and (55) in s require practically the same amount of computation for their right-hand sides, i.e., one matrix-vector multiplication. The first derivatives dx/dv for $r = 2$ and dx/ds remain finite when x tends to x_S, while the second derivatives increase to infinite values as in (54) and (58). For $r = 4$ the second derivative in (52) remains finite when x tends to x_S (see (54)), while the first derivative and the ratio ds/dv tend to zero as in (54) and (60), respectively. If $r>4$ even $\left\|d^2x/dv^2\right\|$ tends to zero as in (54).

Equations (49) and (55) can be integrated by classical numerical methods. Since we are not looking for an accurate solution of these equations all along from x_o to x_S but for finding accurately the final value $x = x_S$, we can use a lower order method, for instance, even Euler's method [7]. This yields an approximate value of x_S which is to be used as initial point for repeating the numerical integration procedure. As we get closer to the solution x_S, we decrease the step size in order to reduce the error. In the case of Euler's method the error is determined in terms of the norm of the second derivative. Higher order numerical integration methods can also be used in order to increase the computation efficiency.

To find a starting point for the integration procedure which is reasonably close to the solution point, one can minimize $F(x)$ in (46) along the normal direction, followed by a minimization of the distance to the solution point x_S along the direction of the normal to F [8]. These two preliminary steps are repeated a few times as needed.

Numerical experiments have been performed applying Euler's method to (49) for $r = 2,\ r = 4$ and $r = 8$, and to (55). Systems (11) of various sizes have been automatically generated and the differential Eqs. (49) for $r = 2$ and $r = 4$, and (55) have produced results with the least amount of computation when imposing an accuracy of 1%.

For matrices which are not symmetric positive definite, (46) is replaced with $F(x) = 1/2\ (Ax-b)^T(Ax-b)$.

6. Conclusions

A special type of matrix series are used in Section 2 to express the relationship between some first order ordinary differential equations and systems of linear algebraic equations and, also, in Section 3 to derive efficient formulae for matrix exponentials that allow accurate and stable numerical computations in various applications. The main feature of these series consists in the fact that, starting with their first term which is already a matrix substantially better conditioned than the original problem matrix, each of the subsequent terms is obtained through a multiplication with a better and better conditioned matrix that tends to the identity matrix. The new matrix exponential formulae contain very rapidly convergent series and can be applied to general, arbitrarily conditioned, positive definite or not matrices. They are used in Section 4 for two new methods of solution for general linear algebraic systems. One is an iterative method which corresponds to the solution of the differential Eq. (1) with $f(v) = 1/v$. It is based on the exact analytical expressions (30)–(32) that always yield results converging finally to the exact solution of the system (11). In a second method, the original algebraic system (11) is replaced with an equivalent system containing a matrix exponential $e^{-\alpha A}$ such that instead of inverting the system matrix A we have now to invert $I - e^{-\alpha A}$. The exact analytical solution is obtained in the form of a series of matrix exponentials which is transformed into an infinite matrix product in order to reduce substantially the necessary amount of computation. It should be remarked that, since the number of matrix-vector multiplications required for the application of one matrix exponential-vector multiplication only depends on the norm of the matrix while the number of matrix exponential-vector multiplications depends on the condition of the system matrix, the total number of arithmetic operations needed to achieve an imposed accuracy when applying each of the two methods is practically proportional to n^2, where n is the dimension of the matrix. The two methods require a comparable total amount of computation. It is also remarkable that for both methods the necessary amount of computation can be roughly predicted beforehand in terms of the system size, the system condition and the desired accuracy.

In Section 5, a powerful method is presented based on the numerical integration of specially constructed ordinary differential equations.

Author details

Ioan R. Ciric
The University of Manitoba, Winnipeg, Canada

*Address all correspondence to: ioan.ciric@umanitoba.ca

References

[1] Ciric IR. Rapidly convergent matrix series for solving linear systems of equations. In: Proceedings of the 17th International Symposium on Antenna Technology and Applied Electromagnetics (ANTEM); 10-16 July 2016; Montréal, Canada: IEEE; 2016. DOI: 10.1109/ANTEM.2016.755022/ 978-1-4673-8478-0

[2] Lanczos C. Applied Analysis. New York: Dover Publications, Inc.; 1988. ISBN: 0-486-65656-X (Paperback). CIP: 88-3961

[3] Moler CB, Van Loan CF. Nineteen dubious ways to compute the exponential of a matrix. SIAM Review. 1978;**20**:801-836. DOI: 10.1137/ S00361445024180

[4] Hurewicz W. Lectures on Ordinary Differential Equations. New York: Dover; 1990. ISBN: 0-486-66420-I (Paperback). QA372.H93

[5] Gradshteyn IS, Ryzhik IM. Table of Integrals, Series and Products. 5th ed. New York: Academic Press; 1994. ISBN: 0-12-294755-X. QA55.G6613

[6] Golub GH, Van Loan CF. Matrix Computations. 2nd ed. Baltimore: Johns Hopkins; 1989. ISBN: 0-8018-3772-3, 0-8018-3739-1 (Paperback). 88-4504

[7] Burden RL, Faires JD. Numerical Analysis. 5th ed. Boston: PWS-KENT; 1993. ISBN: 0-534-93219-3. CIP: 92-32192

[8] Ciric IR. A geometric property of the functional associated with general systems of algebraic equations. In: Proceedings of the International Symposium on Antenna Technology and Applied Electromagnetics and Canadian Radio Sciences Conference (ANTEM/URSI); 17-19 July 2006; Montréal, Canada: IEEE; 2006. p. 311-315. ISBN: 978-0-9738425-1-7