Numerical Solution of Linear, Homogeneous Differential Equation Systems via Padé Approximation

Kenneth C. Johnson

KJ Innovation

kjinnovation@earthlink.net

(Posted April 1, 2015; latest update April 22, 2016.)

Abstract

This paper reports work-in-progress on the solution of first-order, linear, homogeneous differential equation systems, with non-constant coefficients, by generalization of the Padé-approximant method for exponential matrices.

1. Introduction

A system of first-order, linear, homogeneous differential equations is of the form

$$F'[x] = D[x]\,F[x]\,, \tag{1}$$

where F and D are matrix functions of a scalar argument x, $D[x]$ is a known coefficient matrix, and $F[x]$ is to be determined from a specified initial value (e.g. $F[0]$). (Following the Mathematica convention, square braces "$[\ldots]$" are used in this paper to delimit function arguments, while round braces "$(\ldots)$" are reserved for grouping.) Typically, methods such as Runge-Kutta [1] are used to calculate numerical solutions of Eq. (1). But in the constant-coefficient case (x-independent D) solutions have an exponential-matrix representation, e.g.,

$$F'[x] = D\,F[x] \quad \rightarrow \quad F[x] = \exp[D\,x]\,F[0]\,. \tag{2}$$

The exponential matrix $\exp[D\,x]$ can be calculated using a Padé approximation for small x (using a "scale-and-square" method to build up $\exp[D\,x]$ for large x) [2].

The Padé-approximant method can also be extended for the case of non-constant coefficients. This paper briefly outlines work-in-progress on the method, which may be generalized and elaborated upon in future work. Section 2 introduces Padé approximation in the context of Eq. (1); section 3 summarizes standard exponential matrix approximation methods for the constant-coefficient case; and section 4 presents several Padé-approximant formulas for the case of non-constant coefficients. The Appendix provides Mathematica code validating the results of section 4.

2. Application of the Padé-approximant method to Eq. (1)

Eq. (1) is solved by a multi-step method in which an approximation of $F[x + \Delta x]$ is determined from a previously computed estimate of $F[x]$, for some small increment Δx. It will be convenient to denote the integration step Δx as $2h$, and to locate the x origin at the center of the integration interval. Thus, the problem is to find an approximation to $F[h]$ given a predetermined estimate of $F[-h]$. The approximation is represented as

$$F[h] \approx Q[h]^{-1} P[h] F[-h], \tag{3}$$

where $P[h]$ and $Q[h]$ are matrix-valued, polynomial functions of h determined to minimize the error in Eq. (3) under the premise of Eq. (1). Specifically, we require that

$$Q[h] F[h] - P[h] F[-h] = O h^{2n+1}, \tag{4}$$

where $2n$ is the approximation order. (The order is limited to being even, as explained below.)

Making the substitution $h \to -h$ in Eq. (4), we obtain the similar expression

$$P[-h] F[h] - Q[-h] F[-h] = O h^{2n+1}, \tag{5}$$

Assuming that P and Q are uniquely determined by some type of definition criteria, it can be inferred from the similarity of Eq's. (4) and (5) that

$$P[h] = Q[-h], \tag{6}$$

Thus, we seek to determine a polynomial function $Q[h]$ such that

$$Q[h] F[h] - Q[-h] F[-h] = O h^{2n+1}, \tag{7}$$

$Q[0]$ is set equal to the identity matrix $\mathbf{I}$,

$$Q[0] = \mathbf{I}. \tag{8}$$

Eq. (7) is an odd function of h, so a Taylor series expansion of the expression will contain only odd powers of h and the error order on the right side of Eq. (7) is also an odd power of h. The approximation order (i.e., the error order minus one) is even.

Due to the odd symmetry of Eq. (7), an order-n polynomial $Q[h]$ has sufficient degrees of freedom to achieve order-$2n$ accuracy of Eq. (7). This is a key benefit of the Padé approximation, which remains true for a non-constant coefficient matrix $D[h]$, although the advantage is diminished in this case because the calculation of $Q[-h]$ also entails evaluation of an order-n polynomial. (For the constant-D case, the calculation of $Q[-h]$ adds very little computational overhead because the even and odd parts of the polynomial $Q[h]$ can be computed separately and subtracted to get $Q[-h]$.) Nevertheless, Padé approximants such as those outlined in section 4 can have advantages of computational efficiency and accuracy relative to standard techniques such as Runge-Kutta.

3. The constant-coefficient case; exponential matrices.

For the constant-coefficient case, Eq's. (2) and (7) imply that

$$Q[h] \exp[D h] - Q[-h] \exp[-D h] = O h^{2n+1}, \tag{9}$$

The function Q, denoted as Q_n for a particular approximation order $2n$, is of the form

$$Q_n[h] = \sum_{j=0}^{n} \frac{(2n-j)!n!}{j!(2n)!(n-j)!}(-2hD)^j \,, \tag{10}$$

The polynomials can be calculated from the following recursion relations,

$$\begin{aligned} &Q_0[h] = \mathbf{I}, \\ &Q_1[h] = \mathbf{I} - hD, \\ &Q_{n+1}[h] = Q_n[h] + \frac{h^2 D^2}{(2n+1)(2n-1)} Q_{n-1}[h]. \end{aligned} \tag{11}$$

The first several iterations of this recursion yield

$$Q_2[h] = \mathbf{I} - hD + \tfrac{1}{3}h^2 D^2 \,, \tag{12}$$

$$Q_3[h] = \mathbf{I} - hD + \tfrac{2}{5}h^2 D^2 - \tfrac{1}{15}h^3 D^3 \,, \tag{13}$$

$$Q_4[h] = \mathbf{I} - hD + \tfrac{3}{7}h^2 D^2 - \tfrac{2}{21}h^3 D^3 + \tfrac{1}{105}h^4 D^4 \,. \tag{14}$$

The accuracy advantage of the Padé approximant method is illustrated by comparing the accuracy error of Eq. (9) to Runge-Kutta methods. For $n = 2$, the error is approximately $\tfrac{2}{45}h^5 D^5$, which is six times smaller than the error of the classic 4th-order Runge-Kutta method. For $n = 3$, the approximate error is $-\tfrac{2}{1575}h^7 D^7$, which is smaller than the error of the 6th-order Runge-Kutta method described in [1] (top of page 192) by a factor of $3/200$.

4. The non-constant-coefficient case: some illustrative formulas

For non-constant $D[x]$ the first several expressions for $Q_n[h]$ can be generalized by replacing the D factors with linear combinations of $D[x]$ evaluated at different x's,

$$Q_1[h] = \mathbf{I} - hD[0], \tag{15}$$

$$Q_2[h] = \mathbf{I} - h(-\tfrac{1}{6}D[-h] + \tfrac{2}{3}D[0] + \tfrac{1}{2}D[h]) + \tfrac{1}{3}h^2 D[h]^2 \,, \tag{16}$$

$$\begin{aligned} Q_3[h] = \mathbf{I} - h(&\tfrac{2}{45}D[-\tfrac{1}{2}h] + \tfrac{2}{15}D[0] + \tfrac{2}{3}D[\tfrac{1}{2}h] + \tfrac{7}{45}D[h]) + \\ &(\tfrac{1}{15}D[-\tfrac{1}{2}h] + \tfrac{1}{5}D[0] + \tfrac{11}{15}D[\tfrac{1}{2}h]) \\ &(\tfrac{2}{5}h^2(\tfrac{1}{9}D[-\tfrac{1}{2}h] - \tfrac{1}{2}D[0] + D[\tfrac{1}{2}h] + \tfrac{7}{18}D[h]) - \tfrac{1}{15}h^3 D[h]^2). \end{aligned} \tag{17}$$

Eq. (17) illustrates the efficiency characteristics of the Padé approximant method. The calculation of $Q_3[h]^{-1} Q_3[-h]$ (i.e., the $Q[h]^{-1} P[h]$ factor in Eq. (3)) requires four matrix multiplies and one matrix divide, but it actually only needs three multiplies per integration step because the $D[h]^2$ term can be reused for the succeeding step (as $D[-h]^2$). The method requires four $D[x]$ function evaluations per integration step (not counting $D[h]$, which is the starting point for the succeeding step). The Padé approximation samples the function at uniform intervals, which is advantageous because interleaved data points can be added to reduce h by a factor of 2 (e.g. for using Richardson extrapolation). If the sampling does not need to be uniform, then an alternative Padé approximant using only three $D[x]$ samples per step can be used,

$$
\begin{aligned}
Q_3[h] = \mathbf{I} - h\Big(& \left(\tfrac{5}{12} - \tfrac{3\sqrt{5}}{20}\right) D[-\tfrac{1}{\sqrt{5}} h] + \left(\tfrac{5}{12} + \tfrac{3\sqrt{5}}{20}\right) D[\tfrac{1}{\sqrt{5}} h] + \tfrac{1}{6} D[h]\Big) + \\
& \Big(\left(\tfrac{1}{2} - \tfrac{\sqrt{5}}{6}\right) D[-\tfrac{1}{\sqrt{5}} h] + \left(\tfrac{1}{2} + \tfrac{\sqrt{5}}{6}\right) D[\tfrac{1}{\sqrt{5}} h]\Big) \\
& \Big(\tfrac{2}{5} h^2 \left(\tfrac{1}{12} D[-h] - \tfrac{5}{24}(\sqrt{5} - 1) D[-\tfrac{1}{\sqrt{5}} h] + \tfrac{5}{24}(\sqrt{5} + 1) D[\tfrac{1}{\sqrt{5}} h] + \tfrac{1}{2} D[h]\right) - \tfrac{1}{15} h^3 D[h]^2 \Big).
\end{aligned}
\tag{18}
$$

For approximation order 8, the $Q_4[h]$ definition in Eq. (14) can be generalized for non-constant D by replacing each power D^m by a linear combination of product terms, each with m factors of the general form

$$
L[h] = c_{-3}\, D[-h] + c_{-2}\, D[-\tfrac{2}{3} h] + c_{-1}\, D[-\tfrac{1}{3} h] + c_0\, D[0] + c_1\, D[\tfrac{1}{3} h] + c_2\, D[\tfrac{2}{3} h] + c_3\, D[h]. \tag{19}
$$

The seven coefficients c_{-3}, …, c_3 in each factor are initially undetermined, except that they are constrained so that the $Q_4[h]$ representation reduces to Eq. (14) when D is constant. Eq. (7) is expanded in an order-$2n$ Taylor series, using Eq. (1) to eliminate derivatives of F. The monomial coefficients in the series must vanish; this condition leads to a set of equations from which the coefficients can be determined. (The equations may be underdetermined, or they may be overdetermined if the $Q_4[h]$ definition does not have sufficiently many summation terms.)

The above process leads to an enormously complex system of equations, but the equations can be greatly simplified by representing $L[h]$ alternatively in terms of its undetermined derivatives at $h = 0$,

$$
\begin{aligned}
L[h] = & \tfrac{1}{4}(4 d_0 - 49 d_2 + 126 d_4 - 81 d_6)\, D[0] \\
& + \tfrac{9}{16}(4 d_1 + 12 d_2 - 13 d_3 - 39 d_4 + 9 d_5 + 27 d_6)\, D[\tfrac{1}{3} h] \\
& + \tfrac{9}{16}(-4 d_1 + 12 d_2 + 13 d_3 - 39 d_4 - 9 d_5 + 27 d_6)\, D[-\tfrac{1}{3} h] \\
& + \tfrac{9}{40}(-2 d_1 - 3 d_2 + 20 d_3 + 30 d_4 - 18 d_5 - 27 d_6)\, D[\tfrac{2}{3} h] \\
& + \tfrac{9}{40}(2 d_1 - 3 d_2 - 20 d_3 + 30 d_4 + 18 d_5 - 27 d_6)\, D[-\tfrac{2}{3} h] \\
& + \tfrac{1}{80}(4 d_1 + 4 d_2 - 45 d_3 - 45 d_4 + 81 d_5 + 81 d_6)\, D[h] \\
& + \tfrac{1}{80}(-4 d_1 + 4 d_2 + 45 d_3 - 45 d_4 - 81 d_5 + 81 d_6)\, D[-h].
\end{aligned}
\tag{20}
$$

The seven undetermined constants d_0, …, d_6 are coefficients in the Taylor series expansion of $L[h]$,

$$
\begin{aligned}
L[h] = & d_0\, D[0] + d_1\, h\, D'[0] + \tfrac{1}{2} d_2\, h^2\, D''[0] + \tfrac{1}{6} d_3\, h^3\, D^{[3]}[0] \\
& + \tfrac{1}{24} d_4\, h^4\, D^{[4]}[0] + \tfrac{1}{120} d_5\, h^5\, D^{[5]}[0] + \tfrac{1}{720} d_6\, h^6\, D^{[6]}[0] + O\, h^7.
\end{aligned}
\tag{21}
$$

Following is a $Q_4[h]$ definition, which was has been formulated to minimize the number of matrix multiplies:

$$Q_4[h] = \mathbf{I} - h\,L_1[h] + L_2[h](\tfrac{121}{315}h^2\,L_3[h] - \tfrac{2}{315}h^3\,L_4[h]\,L_5[h]) \\ + \left(\tfrac{2}{45}h^2\,L_6[h] + L_2[h](-\tfrac{4}{45}h^3\,L_6[h] + \tfrac{1}{105}h^4\,D[h]^2)\right)D[h], \tag{22}$$

where

$$\begin{aligned}
L_1[h] &= \tfrac{403}{16800}D[-h] - \tfrac{279}{2800}D[-\tfrac{2}{3}h] + \tfrac{99}{800}D[-\tfrac{1}{3}h] + \tfrac{34}{105}D[0] - \tfrac{333}{5600}D[\tfrac{1}{3}h] + \tfrac{1719}{2800}D[\tfrac{2}{3}h] + \tfrac{1237}{16800}D[h] \\
L_2[h] &= \tfrac{57}{1120}D[-h] - \tfrac{243}{560}D[-\tfrac{2}{3}h] + \tfrac{1269}{1120}D[-\tfrac{1}{3}h] - \tfrac{3}{4}D[0] + \tfrac{891}{1120}D[\tfrac{1}{3}h] + \tfrac{27}{112}D[\tfrac{2}{3}h] - \tfrac{41}{1120}D[h] \\
L_3[h] &= -\tfrac{2067}{9680}D[-h] + \tfrac{6021}{4840}D[-\tfrac{2}{3}h] - \tfrac{5805}{1936}D[-\tfrac{1}{3}h] + \tfrac{1863}{484}D[0] - \tfrac{5697}{1936}D[\tfrac{1}{3}h] + \tfrac{10341}{4840}D[\tfrac{2}{3}h] - \tfrac{727}{9680}D[h] \\
L_4[h] &= \tfrac{63}{16}D[-h] - \tfrac{1809}{40}D[-\tfrac{2}{3}h] + \tfrac{2295}{16}D[-\tfrac{1}{3}h] - \tfrac{801}{4}D[0] + \tfrac{2133}{16}D[\tfrac{1}{3}h] - \tfrac{297}{8}D[\tfrac{2}{3}h] + \tfrac{233}{80}D[h] \\
L_5[h] &= \tfrac{123}{160}D[-h] - \tfrac{135}{8}D[-\tfrac{2}{3}h] + \tfrac{2295}{32}D[-\tfrac{1}{3}h] - 132\,D[0] + \tfrac{3861}{32}D[\tfrac{1}{3}h] - \tfrac{1917}{40}D[\tfrac{2}{3}h] + \tfrac{149}{32}D[h] \\
L_6[h] &= -\tfrac{6}{35}D[-h] + \tfrac{27}{10}D[-\tfrac{2}{3}h] - \tfrac{1053}{112}D[-\tfrac{1}{3}h] + \tfrac{57}{4}D[0] - \tfrac{621}{56}D[\tfrac{1}{3}h] + \tfrac{729}{140}D[\tfrac{2}{3}h] - \tfrac{277}{560}D[h]
\end{aligned} \tag{23}$$

References

[1] Butcher, John C. "On Runge-Kutta processes of high order." *Journal of the Australian Mathematical Society* 4.02 (1964): 179-194.

[2] Higham, Nicholas J. "The scaling and squaring method for the matrix exponential revisited." *SIAM review* 51.4 (2009): 747-764.

Appendix: Approximation orders of Eq's. (15)-(18), (22)

The calculations underlying Eq's. (15)-(18) and (22) require non-commutative symbolic algebra. The following results are obtained using the NCAlgebra package for Mathematica, from the University of California, San Diego (http://math.ucsd.edu/~ncalg/). The Mathematica code loads the NCAlgebra package, adds some additional functionality, and verifies Eq. (9) with $Q[x]$ defined by any of Eq's. (15)-(18), (22).

```
(* Load NCAlgebra package (http://math.ucsd.edu/~ncalg/) *)
<< NC`
<< NCAlgebra`

(* Make all variables commutative by default.
  (Override the default noncommutativity of single-letter lowercase variables.) *)
Remove[a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z]

(* Dfn, F, and Q represent matrices. ("1" represents the identity matrix.) *)
SetNonCommutative[Dfn, F, Q];

(* Series and O (e.g. O[h]^n) do not work with NC types
 (e.g.: try Dfn[h]**F[h]+O[h]^2 or Series[Dfn[h]**F[h],{h,0,1}]). Define a variant that does. *)
NCSeries[f_, {x_, x0_, n_}] := NCExpand[Sum[(D[f, {x, j}] / j! /. x → x0) (x - x0)^j, {j, 0, n}]] + O[x - x0]^(n + 1);

(* substD is a substitution rule for reducing derivatives of F using the relation F'[h]==Dfn[h]**F[h].
   Use "//. substD" to eliminate all F derivatives.
   (Use ":>" here, not "->"; otherwise the substitutions will not work when x or n has a preassigned value.) *)
substD = Derivative[n_][F][x_] :> Derivative[n - 1][Dfn[#] ** F[#] &][x];


(* Eq 15 *)
Q[h_] := 1 - h Dfn[0];
NCExpand[Normal[NCSeries[Q[h] ** F[h] - Q[-h] ** F[-h], {h, 0, 2}]] //. substD]

0


(* Eq 16 *)
Q[h_] := 1 - h (-1/6 Dfn[-h] + 2/3 Dfn[0] + 1/2 Dfn[h]) + 1/3 h^2 Dfn[h] ** Dfn[h];
NCExpand[Normal[NCSeries[Q[h] ** F[h] - Q[-h] ** F[-h], {h, 0, 4}]] //. substD]

0


(* Eq 17 *)
Q[h_] := 1 - h (2/45 Dfn[-h/2] + 2/15 Dfn[0] + 2/3 Dfn[h/2] + 7/45 Dfn[h]) +
   (1/15 Dfn[-h/2] + 1/5 Dfn[0] + 11/15 Dfn[h/2]) ** (2/5 h^2 (1/9 Dfn[-h/2] - 1/2 Dfn[0] + Dfn[h/2] + 7/18 Dfn[h]) - 1/15 h^3 Dfn[h] ** Dfn[h]);
NCExpand[Normal[NCSeries[Q[h] ** F[h] - Q[-h] ** F[-h], {h, 0, 6}]] //. substD]

0


(* Eq 18 *)
Q[h_] :=
  1 - h ((5/12 - 3 Sqrt[5]/20) Dfn[-h/Sqrt[5]] + (5/12 + 3 Sqrt[5]/20) Dfn[h/Sqrt[5]] + 1/6 Dfn[h]) + ((1/2 - Sqrt[5]/6) Dfn[-h/Sqrt[5]] + (1/2 + Sqrt[5]/6) Dfn[h/Sqrt[5]]) **
   (2/5 h^2 (1/12 Dfn[-h] - 5/24 (Sqrt[5] - 1) Dfn[-h/Sqrt[5]] + 5/24 (Sqrt[5] + 1) Dfn[h/Sqrt[5]] + 1/2 Dfn[h]) - 1/15 h^3 Dfn[h] ** Dfn[h]);
NCExpand[Normal[NCSeries[Q[h] ** F[h] - Q[-h] ** F[-h], {h, 0, 6}]] //. substD]

0
```

```
(* Eq 22 *)
L1[h_] := 403/16800 Dfn[-h] - 279/2800 Dfn[-2 h/3] + 99/800 Dfn[-h/3] + 34/105 Dfn[0] - 333/5600 Dfn[h/3] + 1719/2800 Dfn[2 h/3] + 1237/16800 Dfn[h];
L2[h_] := 57/1120 Dfn[-h] - 243/560 Dfn[-2 h/3] + 1269/1120 Dfn[-h/3] - 3/4 Dfn[0] + 891/1120 Dfn[h/3] + 27/112 Dfn[2 h/3] - 41/1120 Dfn[h];
L3[h_] := -2067/9680 Dfn[-h] + 6021/4840 Dfn[-2 h/3] - 5805/1936 Dfn[-h/3] + 1863/484 Dfn[0] - 5697/1936 Dfn[h/3] + 10341/4840 Dfn[2 h/3] - 727/9680 Dfn[h];
L4[h_] := 63/16 Dfn[-h] - 1809/40 Dfn[-2 h/3] + 2295/16 Dfn[-h/3] - 801/4 Dfn[0] + 2133/16 Dfn[h/3] - 297/8 Dfn[2 h/3] + 233/80 Dfn[h];
L5[h_] := 123/160 Dfn[-h] - 135/8 Dfn[-2 h/3] + 2295/32 Dfn[-h/3] - 132 Dfn[0] + 3861/32 Dfn[h/3] - 1917/40 Dfn[2 h/3] + 149/32 Dfn[h];
L6[h_] := -6/35 Dfn[-h] + 27/10 Dfn[-2 h/3] - 1053/112 Dfn[-h/3] + 57/4 Dfn[0] - 621/56 Dfn[h/3] + 729/140 Dfn[2 h/3] - 277/560 Dfn[h];
Q[h_] := 1 - h L1[h] + L2[h] ** (121/315 h^2 L3[h] - 2/315 h^3 L4[h] ** L5[h]) +
    (2/45 h^2 L6[h] + L2[h] ** (-4/45 h^3 L6[h] + 1/105 h^4 Dfn[h] ** Dfn[h])) ** Dfn[h];
NCExpand[Normal[NCSeries[Q[h] ** F[h] - Q[-h] ** F[-h], {h, 0, 8}]] //. substD]
```

```
0
```