

PUBLICATION III

**Correction to partial fraction  
decomposition coefficients for  
Chebyshev rational approximation  
on the negative real axis**

arXiv:1206.2880v1 [math.NA] (2012).



**CORRECTION TO PARTIAL FRACTION DECOMPOSITION  
COEFFICIENTS FOR CHEBYSHEV RATIONAL  
APPROXIMATION ON THE NEGATIVE REAL AXIS**

MARIA PUSA

ABSTRACT. Chebyshev rational approximation can be a viable method to compute the exponential of matrices with eigenvalues in the vicinity of the negative real axis, and it was recently applied successfully to solving nuclear fuel burnup equations. Determining the partial fraction decomposition (PFD) coefficients of this approximation can be difficult and they have been provided (for approximation orders 10 and 14) by Gallopoulos and Saad in “Efficient solution of parabolic equations by Krylov approximation methods”, SIAM J. Sci. Stat. Comput., 13(1992). It was recently discovered that the order 14 coefficients contain errors and result in  $10^2$  times poorer accuracy than expected by theory. The purpose of this note is to provide the correct PFD coefficients for approximation orders 14 and 16 and to briefly discuss the approximation accuracy resulting from the erroneous coefficients.

1. CHEBYSHEV RATIONAL APPROXIMATION

This note concerns the computation of matrix exponential based on the Chebyshev rational approximation method (abbreviated CRAM in [7]) on the negative real axis. In this approach, the matrix exponential  $e^{\mathbf{A}t}$  is approximated by a rational matrix function  $\hat{r}(\mathbf{A}t)$ , where the rational function  $\hat{r}(z)$  is chosen as the best rational approximation of the exponential function on the negative real axis  $\mathbb{R}_-$ . Let  $\pi_{k,k}$  denote the set of rational functions  $r_{k,k}(x) = p_k(x)/q_k(x)$  where  $p_k$  and  $q_k$  are polynomials of order  $k$ . The CRAM approximation of order  $k$  is defined as the unique rational function  $\hat{r}_{k,k} = \hat{p}_k(x)/\hat{q}_k(x)$  satisfying

$$(1) \quad \sup_{x \in \mathbb{R}_-} |\hat{r}_{k,k}(x) - e^x| = \inf_{r_{k,k} \in \pi_{k,k}} \left\{ \sup_{x \in \mathbb{R}_-} |r_{k,k}(x) - e^x| \right\}.$$

The asymptotic convergence of this approximation on the negative real axis is remarkably fast with the convergence rate  $O(H^{-k})$ , where  $H = 9.28902549\dots$  is called the Halphen constant [4]. It was recently discovered by Stahl and Schmelzer [11] that this convergence extends to compact subsets on the complex plane and also to Hankel contours in  $\mathbb{C} \setminus \mathbb{R}_-$ . The application of this approximation to computing the matrix exponential was originally made famous by Cody, Meinardus, and Varga in 1969 in the context of rational approximation of  $e^{-x}$  on  $[0, \infty)$  and it was recently resurfaced by Trefethen, Weideman, and Schmelzer [12]. For self-adjoint and negative semi-definite matrices, the method is

---

*Date:* 12 June 2012.

*2000 Mathematics Subject Classification.* 41-04, 65F60.

*Key words and phrases.* best rational approximation, CRAM, partial fraction decomposition, matrix exponential.

guaranteed to yield an error bound in 2-norm that corresponds to the maximum error of the rational approximation on the negative real axis. This has also been the main context for scientific applications [2, 9, 10]. Recently, the method has also been successfully applied to non-self-adjoint matrices with eigenvalues near the negative real axis [7, 6]. These specific matrices arise from a reactor physics application, where the changes in nuclide concentrations due to radioactive decay and neutron-induced reactions are governed by a linear system  $\mathbf{x}' = \mathbf{A}\mathbf{x}$  known as the burnup equations.

## 2. PARTIAL FRACTION DECOMPOSITION FORM

The main difficulty in using CRAM for computing the matrix exponential is determining the coefficients of the rational function for a given  $k$ . In principle, the polynomial coefficients of  $\hat{p}_k$  and  $\hat{q}_k$  can be computed with Remez-type methods but this requires delicate algorithms combined with high-precision arithmetics. Fortunately, these coefficients have been computed to a high accuracy by Carpenter et al. for approximation orders  $k = 0, 1, \dots, 30$  and they are provided in [1]. In practical applications, however, it is usually advantageous to employ CRAM in the partial fraction decomposition (PFD) form. For simple poles, this composition takes the form

$$(2) \quad \hat{r}_{k,k}(z) = \alpha_0 + \sum_{j=1}^k \frac{\alpha_j}{z - \theta_j},$$

where  $\alpha_0$  is the limit of the function  $\hat{r}_{k,k}$  at infinity, and  $\alpha_j$  are the residues at the poles  $\theta_j$ :

$$(3) \quad \alpha_j = \frac{\hat{p}_k(\theta_j)}{\hat{q}'_k(\theta_j)}.$$

Since the coefficients of  $\hat{r}_{k,k}$  are real, its poles form conjugate pairs, so the computational cost can be reduced to half for a real variable  $x$

$$(4) \quad \hat{r}_{k,k}(x) = \alpha_0 + 2 \operatorname{Re} \left( \sum_{j=1}^{k/2} \frac{\alpha_j}{x - \theta_j} \right)$$

and the matrix exponential solution may be approximated as

$$(5) \quad \mathbf{x} = \alpha_0 \mathbf{x}_0 + 2 \operatorname{Re} \left( \sum_{j=1}^{k/2} \alpha_j (\mathbf{A}t - \theta_j \mathbf{I})^{-1} \mathbf{x}_0 \right)$$

for a real matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ .

Although the PFD coefficients can in principle be computed from the polynomial coefficients, the computation of the polynomial roots may be ill-conditioned and requires great care. The PFD coefficients for approximation orders 10 and 14 have been provided in [3], and the given coefficients for  $k = 14$  have been used in several applications including the matrix exponential computing package EXPKIT [10] and the reactor physics code Serpent [5]. However, in the latter context, it was recently observed that these reported coefficients contain errors and do not correspond to the true best approximation [6]. To illustrate this, Figure 1 shows the error of order 14 approximation on the negative real axis computed using two different sets of coefficients: the partial fraction coefficients from [3], with the corresponding

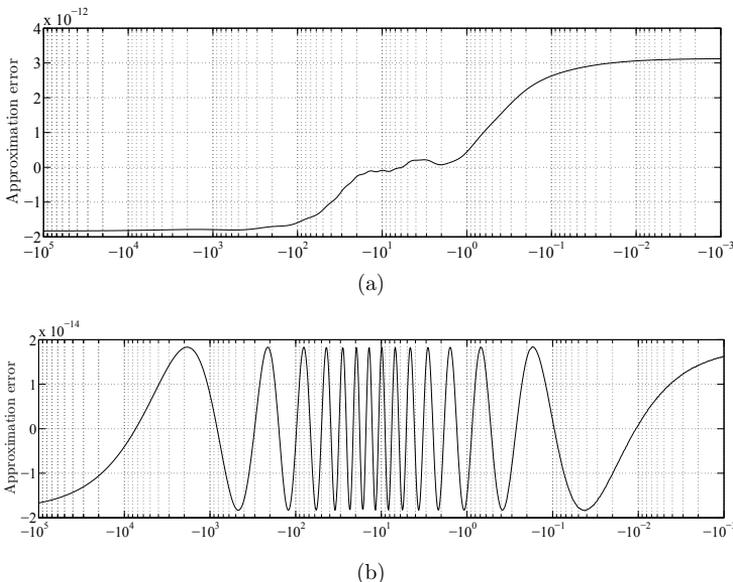


FIGURE 1. (a) Plot of  $e^x - \tilde{r}_{14,14}(x)$  on the negative real axis with  $\tilde{r}_{14,14}$  computed based on the partial fraction coefficients from [3], (b) Plot of  $e^x - \hat{r}_{14,14}(x)$  based on the polynomial coefficients from [1]. The plots were computed using high-precision arithmetics with 32 digits.

approximation denoted by  $\tilde{r}_{14,14}$ , and the polynomial coefficients from [1], with the corresponding approximation denoted by  $\hat{r}_{14,14}$ . According to theory, a necessary and sufficient condition for the best approximation is that the corresponding error function equioscillates, i.e. there exists a set of points where it attains its maximum absolute value with alternating signs. Notice that the approximation computed with the coefficients from [3] does not exhibit this behavior and in addition results in a  $10^2$  times poorer accuracy than expected by theory.

After discovering the erroneous behavior induced by the coefficients from [3], partial fraction coefficients for approximation orders  $k = 14$  and  $k = 16$  were computed from the polynomial coefficients provided in [1] and subsequently reported in [6]. The computed PFD coefficients are repeated here in Tables 1 and 2. The computations were performed with MATLAB's Symbolic Toolbox using high precision arithmetics with 200 digits to ensure a sufficient accuracy. In Tables 1 and 2 the coefficients have been rounded off to 20 digits. The coefficients in [1] have been also given with 20 digits' accuracy, and based on our experience, the approximation order  $k = 16$  is the highest for which this accuracy is sufficient for computing the PFD coefficients. For lower approximation orders,  $1 \leq k \leq 13$ , the PFD coefficients can be accurately computed with the approximative Carathéodory–Fejér method and a MATLAB script is provided for this purpose in [8].

TABLE 1. Recomputed values for the partial fraction decomposition coefficients for CRAM approximation of order 14.

Coefficient	Real part	Imaginary part
$\theta_1$	$-8.897\,773\,186\,468\,888\,819\,9 \times 10^0$	$+1.663\,098\,261\,990\,208\,530\,4 \times 10^1$
$\theta_2$	$-3.703\,275\,049\,423\,448\,060\,3 \times 10^0$	$+1.365\,637\,187\,148\,326\,817\,1 \times 10^1$
$\theta_3$	$-0.208\,758\,638\,250\,130\,125\,1 \times 10^0$	$+1.099\,126\,056\,190\,126\,091\,3 \times 10^1$
$\theta_4$	$+3.993\,369\,710\,578\,568\,519\,4 \times 10^0$	$+6.004\,831\,642\,235\,037\,317\,8 \times 10^0$
$\theta_5$	$+5.089\,345\,060\,580\,624\,506\,6 \times 10^0$	$+3.588\,824\,029\,027\,006\,510\,2 \times 10^0$
$\theta_6$	$+5.623\,142\,572\,745\,977\,124\,8 \times 10^0$	$+1.194\,069\,046\,343\,966\,976\,6 \times 10^0$
$\theta_7$	$+2.269\,783\,829\,231\,112\,709\,7 \times 10^0$	$+8.461\,737\,973\,040\,221\,401\,9 \times 10^0$
$\alpha_1$	$-7.154\,288\,063\,589\,067\,285\,3 \times 10^{-5}$	$+1.436\,104\,334\,954\,130\,011\,1 \times 10^{-4}$
$\alpha_2$	$+9.439\,025\,310\,736\,168\,877\,9 \times 10^{-3}$	$-1.718\,479\,195\,848\,301\,751\,1 \times 10^{-2}$
$\alpha_3$	$-3.763\,600\,387\,822\,696\,871\,7 \times 10^{-1}$	$+3.351\,834\,702\,945\,010\,421\,4 \times 10^{-1}$
$\alpha_4$	$-2.349\,823\,209\,108\,270\,119\,1 \times 10^1$	$-5.808\,359\,129\,714\,207\,400\,4 \times 10^0$
$\alpha_5$	$+4.693\,327\,448\,883\,129\,304\,7 \times 10^1$	$+4.564\,364\,976\,882\,776\,079\,1 \times 10^1$
$\alpha_6$	$-2.787\,516\,194\,014\,564\,646\,8 \times 10^1$	$-1.021\,473\,399\,905\,645\,143\,4 \times 10^2$
$\alpha_7$	$+4.807\,112\,098\,832\,508\,890\,7 \times 10^0$	$-1.320\,979\,383\,742\,872\,388\,1 \times 10^0$
$\alpha_0$	$+1.832\,174\,378\,254\,041\,275\,1 \times 10^{-14}$	$+0.000\,000\,000\,000\,000\,000\,0 \times 10^0$

TABLE 2. Computed values for the partial fraction decomposition coefficients for CRAM approximation of order 16.

Coefficient	Real part	Imaginary part
$\theta_1$	$-1.084\,391\,707\,869\,698\,802\,6 \times 10^1$	$+1.927\,744\,616\,718\,165\,228\,4 \times 10^1$
$\theta_2$	$-5.264\,971\,343\,442\,646\,889\,5 \times 10^0$	$+1.622\,022\,147\,316\,792\,730\,5 \times 10^1$
$\theta_3$	$+5.948\,152\,268\,951\,177\,480\,8 \times 10^0$	$+3.587\,457\,362\,018\,322\,282\,9 \times 10^0$
$\theta_4$	$+3.509\,103\,608\,414\,918\,097\,4 \times 10^0$	$+8.436\,198\,985\,884\,375\,082\,6 \times 10^0$
$\theta_5$	$+6.416\,177\,699\,099\,434\,192\,3 \times 10^0$	$+1.194\,122\,393\,370\,138\,687\,4 \times 10^0$
$\theta_6$	$+1.419\,375\,897\,185\,665\,978\,6 \times 10^0$	$+1.092\,536\,348\,449\,672\,258\,5 \times 10^1$
$\theta_7$	$+4.993\,174\,737\,717\,996\,399\,1 \times 10^0$	$+5.996\,881\,713\,603\,942\,226\,0 \times 10^0$
$\theta_8$	$-1.413\,928\,462\,488\,886\,211\,4 \times 10^0$	$+1.349\,772\,569\,889\,274\,538\,9 \times 10^1$
$\alpha_1$	$-5.090\,152\,186\,522\,491\,565\,0 \times 10^{-7}$	$-2.422\,001\,765\,285\,228\,797\,0 \times 10^{-5}$
$\alpha_2$	$+2.115\,174\,218\,246\,603\,090\,7 \times 10^{-4}$	$+4.389\,296\,964\,738\,067\,391\,8 \times 10^{-3}$
$\alpha_3$	$+1.133\,977\,517\,848\,393\,052\,7 \times 10^2$	$+1.019\,472\,170\,421\,585\,645\,0 \times 10^2$
$\alpha_4$	$+1.505\,958\,527\,002\,346\,752\,8 \times 10^1$	$-5.751\,405\,277\,642\,181\,997\,9 \times 10^0$
$\alpha_5$	$-6.450\,087\,802\,553\,964\,659\,5 \times 10^1$	$-2.245\,944\,076\,265\,209\,605\,6 \times 10^2$
$\alpha_6$	$-1.479\,300\,711\,355\,799\,971\,8 \times 10^0$	$+1.768\,658\,832\,378\,293\,790\,6 \times 10^0$
$\alpha_7$	$-6.251\,839\,246\,320\,791\,889\,2 \times 10^1$	$-1.119\,039\,109\,428\,322\,848\,0 \times 10^1$
$\alpha_8$	$+4.102\,313\,683\,541\,002\,127\,3 \times 10^{-2}$	$-1.574\,346\,617\,345\,546\,819\,1 \times 10^{-1}$
$\alpha_0$	$+2.124\,853\,710\,495\,223\,748\,8 \times 10^{-16}$	$+0.000\,000\,000\,000\,000\,000\,0 \times 10^0$

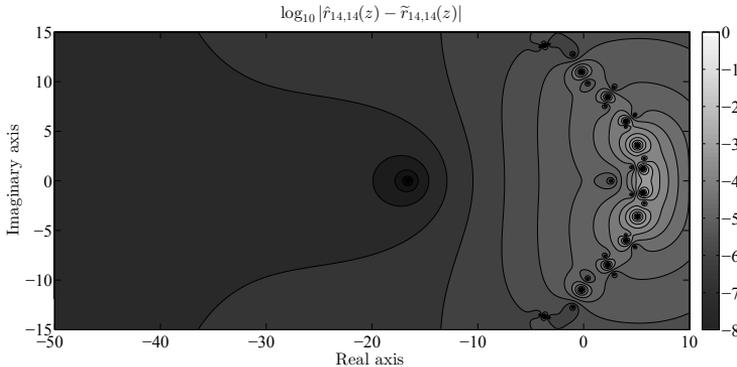


FIGURE 2. Plot of  $\log_{10} |\hat{r}_{14,14}(z) - \tilde{r}_{14,14}(z)|$  in the complex plane.  $\hat{r}_{14,14}$  was computed based on the partial fraction coefficients from Table 1 and  $\tilde{r}_{14,14}$  was formed by truncating these coefficients to 6 significant digits. The poles of  $\hat{r}_{14,14}$  have been marked to the plot with asterisks.

### 3. ANALYSIS OF INACCURATE PFD COEFFICIENTS FOR $k = 14$

To analyze the effect of inaccurate PFD coefficients denoted by  $\{\tilde{\theta}_j\}$  and  $\{\tilde{\alpha}_j\}$ , let  $\tilde{r}$  denote the corresponding rational approximation. The error caused by the inaccuracies in the PFD coefficients may be estimated

$$(6) \quad |\hat{r}_{k,k}(z) - \tilde{r}_{k,k}(z)| \lesssim |\alpha_0 - \tilde{\alpha}_0| + \sum_{j=1}^k \frac{|\alpha_j|}{|z - \theta_j|^2} |\theta_j - \tilde{\theta}_j| + \frac{1}{|z - \theta_j|} |\alpha_j - \tilde{\alpha}_j|,$$

indicating that the error is the greatest in the vicinity of the poles. It can also be seen from Eq. (6) that the inaccuracy related to the poles has a greater impact near the poles, whereas the error related to the residues should begin to dominate the total error farther away from the poles. By comparing the old and the recomputed PFD coefficients for  $k = 14$ , it can be seen that the poles all agree to about 6 digits whereas the residues agree to about 5 digits.<sup>1</sup> The most dramatic discrepancy occurs for the coefficient  $\alpha_0$  for which the significands agree to 5 digits but the exponent value given in [3] is  $-12$ , although the correct value is  $-14$ .

On the grounds of Eq. (6), it can be estimated that coefficients with 6 correct digits should produce a rational function whose deviation from  $\hat{r}_{14,14}(z)$  is at most of the order of  $10^{-3}$  on the negative real axis. Figure 2 shows the difference between  $\hat{r}_{14,14}(z)$  and a rational function  $\tilde{r}_{14,14}(z)$  that was formed by truncating the PFD coefficients of  $\hat{r}_{14,14}$  to 6 significant digits. Interestingly, as can be seen from Fig. 1a, the approximation  $\tilde{r}_{14,14}(x)$ , corresponding to the PFD coefficients from [3], yields a significantly better accuracy of order  $10^{-12}$  than is expected based on the accuracy of the coefficients alone.

<sup>1</sup>Notice that the PFD coefficients in [3] are given for the rational approximation of  $e^{-x}$  on  $[0, \infty)$  and that they have been multiplied by a factor of two making Eq. (37) in [3] equivalent to Eq. (5).

To investigate the matter further, let us now take the poles  $\{\tilde{\theta}_j\}$  reported in [3] as a starting point for constructing a rational approximation of order 14. The poles  $\{\tilde{\theta}_j\}$  define a polynomial

$$(7) \quad \tilde{q}_{14}(x) = \prod_{j=1}^{14} (x - \tilde{\theta}_j)$$

whose values agree to about 6 digits with the values of the correct polynomial  $\hat{q}_{14}(x)$  on the negative real axis. The residues at the poles  $\{\tilde{\theta}_j\}$  cannot be computed in a fully consistent manner, since the poles do not correspond to the true zeros of  $\hat{q}_{14}$ . However, two alternative approaches for computing the residues can be considered. One possibility is to use the correct rational function  $\hat{r}_{14,14}$  and Eq. (3) to compute the residues, but this is inconsistent as Eq. (3) only holds at the true poles. Another option is to define a new rational function using  $\tilde{q}_{14}$  as the denominator and the correct polynomial  $\hat{p}_{14}$  as the numerator, after which the residues can be computed exactly using symbolic arithmetics. With both of these approaches we obtain a rational approximation, whose accuracy is of the order of  $10^{-6}$  on the negative real axis. It is also worth mentioning that forming the rational function based on the poles  $\{\tilde{\theta}_j\}$  and the correct residues  $\{\alpha_j\}$  from Table 1 yields an approximation whose accuracy is of the order of  $10^{-7}$  on the negative real axis.

The article [3] by Gallopoulos and Saad does not indicate, how the reported PFD coefficients were computed, but based on the observations regarding the accuracy of the resulting approximation, it is evident that the values given for the residues somehow compensate for the inaccuracies in the poles  $\{\tilde{\theta}_j\}$  and it seems likely that they have been optimized to minimize the deviation from  $\hat{r}_{14,14}$  on the negative real axis. In fact, using the poles  $\{\tilde{\theta}_j\}$  and standard least squares optimization in MATLAB with  $10^7$  points chosen from the interval  $[-10^3, -10^{-10}]$ , we were able to produce residues yielding only a slightly worse accuracy of order  $10^{-11}$ . In any case, it should be noted that optimizing the residues properly in the Chebyshev sense would essentially form a problem of comparable difficulty as the original problem of determining  $\hat{r}_{14,14}$ .

#### REFERENCES

- [1] A. J. CARPENTER, A. RUTTAN, AND R. S. VARGA, *Extended numerical computations on the 1/9 conjecture in rational approximation theory*, in Rational Approximation and Interpolation, P. R. Graves-Morris, E. B. Saff, and R. S. Varga, eds., vol. 1105 of Lecture Notes in Mathematics, Springer-Verlag, 1984, pp. 383–411.
- [2] W. J. CODY, G. MEINARDUS, AND R. S. VARGA, *Chebyshev rational approximations to  $e^{-x}$  in  $[0, \infty)$  and applications to heat-conduction problems*, J. Approx. Theory, 2 (1969), pp. 50–65.
- [3] E. GALLOPOULOS AND Y. SAAD, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 1236–1264.
- [4] A. A. GONCHAR AND E. A. RAKHMANOV, *Equilibrium distributions and degree of rational approximation of analytic functions*, Math. USSR Sb., 62 (1989).
- [5] J. LEPPÄNEN, *Serpent*, a Continuous-energy Monte Carlo Reactor Physics Burnup Calculation Code, URL: <http://montecarlo.vtt.fi>, VTT Technical Research Centre of Finland, 2012.
- [6] M. PUSA, *Rational approximations to the matrix exponential in burnup calculations*, Nucl. Sci. Eng., 169 (2011), pp. 155–167.
- [7] M. PUSA AND J. LEPPÄNEN, *Computing the matrix exponential in burnup calculations*, Nucl. Sci. Eng., 164 (2010), pp. 140–150.
- [8] T. SCHMELZER, *Carathéodory–Fejér approximation*, Matlab Central, 2008.

- [9] T. SCHMELZER AND L. N. TREFETHEN, *Evaluating matrix functions for exponential integrators via Carathéodory–Fejér approximation and contour integrals*, Electron. Trans. Numer. Anal., 28 (2007), pp. 1–18.
- [10] R. B. SIDJE, *Expokit: a software package for computing matrix exponentials*, ACM Trans. Math. Softw., 24 (1998), pp. 130–156.
- [11] H. STAHL AND T. SCHMELZER, *An extension of the ‘1/9’-problem*, Journal of Computational and Applied Mathematics, 233 (2009), pp. 821–834.
- [12] L. N. TREFETHEN, J. A. C. WEIDEMAN, AND T. SCHMELZER, *Talbot quadratures and rational approximations*, BIT., 46 (2006), pp. 653–670.

VTT TECHNICAL RESEARCH CENTRE OF FINLAND, NUCLEAR ENERGY, P.O. BOX 1000, FI-02044 VTT, FINLAND  
*E-mail address:* maria.pusa@vtt.fi