Math 714: Homework 0 solutions[1]

1. (a) Write the recursion as $x_{k+1} = f(x_k)$ where $f(x) = ax^2 + bx$. If the sequence converges, it must converge to a fixed point where $x = f(x)$, and hence

$$0 = ax^2 + bx - x = x(ax - (1 - b)). \tag{1}$$

Therefore either $x = 0$ or $x = \frac{1-b}{a}$.

(b) When x_k is close to a fixed point, the sequence will converge if $|f'| < 1$, and diverge if $|f'| > 1$. In this case $f'(x) = 2ax + b$. At $x = 0$,

$$|f'(0)| = |b|, \tag{2}$$

and therefore the sequence will converge if $|b| < 1$ and diverge if $|b| > 1$. The convergence/divergence of the sequence for $b = \pm 1$ may depend on the precise value and sign of x_0,[2] which is not specified in the question, so we do not address this.

Now consider the other potential fixed point at $x = \frac{b-1}{a}$. Then

$$\left| f'\left(\frac{b-1}{a}\right) \right| = |(2 - 2b) + b| = |2 - b| \tag{3}$$

and hence the sequence converges for $b \in (1, 3)$. Again, we do not address the possible convergence for $b = 1$ or $b = 3$.

2. The program `cheby_2d.py` calculates the Chebyshev polynomials $T_k(x)$ for $k = 0, 1, \ldots, 5$. Figure 1 shows a two-dimensional plot of the function $T_3(x)T_5(y)$ in the region $(x, y) \in [-1, 1]^2$.

3. (a) For the floating point addition,

$$(4 + 2)(1 - \epsilon) < 4 \oplus 2 < (4 + 2)(1 + \epsilon) \tag{4}$$

and therefore

$$6 - 6\epsilon < 4 \oplus 2 < 6 + 6\epsilon. \tag{5}$$

The floating point division will accrue another relative error up to ϵ in magnitude. In addition, the value will be minimized if $4 \oplus 2$ takes its maximum value, and it will be maximized if $4 \oplus 2$ takes its minimum value. Hence

$$\frac{3}{6 + 6\epsilon}(1 - \epsilon) < 3 \oslash (4 \oplus 2) < \frac{3}{6 - 6\epsilon}(1 + \epsilon) \tag{6}$$

so

$$\frac{1 - \epsilon}{2(1 + \epsilon)} < 3 \oslash (4 \oplus 2) < \frac{1 + \epsilon}{2(1 + \epsilon)}. \tag{7}$$

[1] Written by Chris H. Rycroft.

[2] For example, when $b = 1$ and $a < 0$, then $x_{k+1} = ax_k^2 + x_k$. When $x_k \neq 0$, $x_{k+1} < x_k$. If x_0 is small and positive, then (x_k) is a decreasing sequence bounded below by 0, so it converges. If x_0 is negative, then (x_k) is a decreasing unbounded sequence and it diverges.

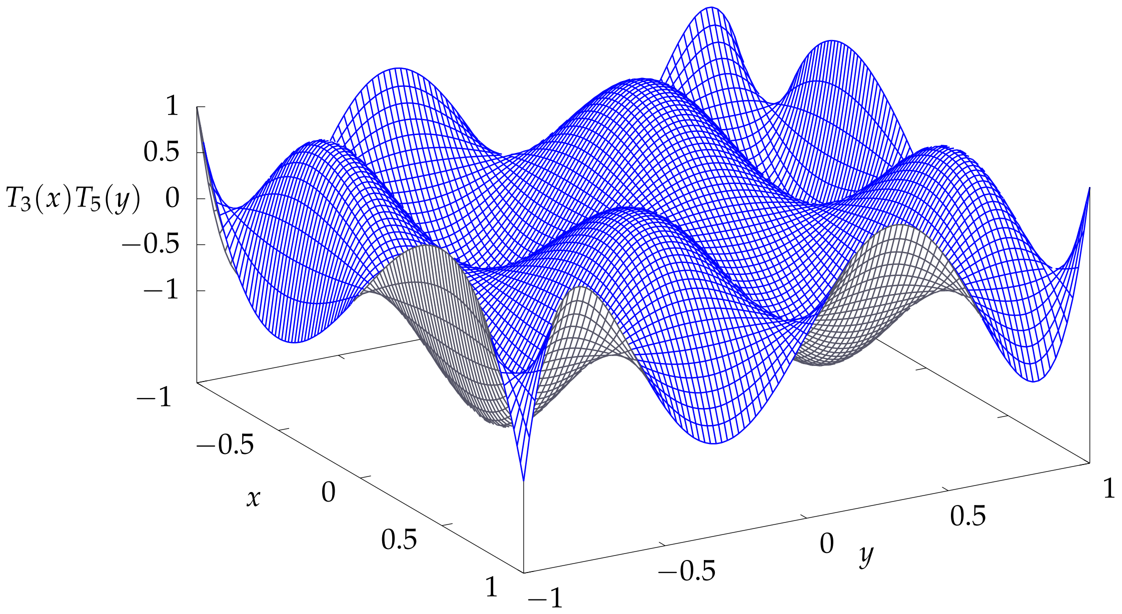

Figure 1: Plot of the product of Chebyshev polynomials $T_3(x)T_5(y)$ considered in question 2.

(b) To simplify Eq. (7), note that

$$\frac{1}{1+\epsilon} = \sum_{n=0}^{\infty}(-\epsilon)^n = 1 - \epsilon + O(\epsilon^2) \tag{8}$$

and

$$\frac{1}{1-\epsilon} = \sum_{n=0}^{\infty}\epsilon^n = 1 + \epsilon + O(\epsilon^2). \tag{9}$$

Therefore if terms of $O(\epsilon^2)$ are neglected, then Eq. (7) becomes

$$\frac{(1-\epsilon)(1-\epsilon)}{2} < 3 \oslash (4 \oplus 2) < \frac{(1+\epsilon)(1+\epsilon)}{2} \tag{10}$$

and hence writing $S = 3/(4+2) = \frac{1}{2}$ and $\tilde{S} = 3 \oslash (4 \oplus 2)$,

$$S(1-2\epsilon) < \tilde{S} < S(1+2\epsilon) \tag{11}$$

so $|S - \tilde{S}| < 2S\epsilon = \epsilon$. Hence $\lambda = 1$.

4. (a) Throughout this equation, $\|\cdot\|$ is taken to mean the Euclidean norm. The first two parts of this problem can be solved using diagonal matrices only. Consider first

$$B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \tag{12}$$

Then $\|B\| = 2$, $\|B^{-1}\| = 1$ and hence $\kappa(B) = 2$. Similarly, $\kappa(C) = 2$. Adding the two matrices together gives

$$B + C = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} = 3I \tag{13}$$

and hence $\kappa(B+C) = \|3I\| \, \|\frac{1}{3}I\| = 3 \times \frac{1}{3} = 1$. For these choices of matrices, $\kappa(B+C) < \kappa(B) + \kappa(C)$.

(b) If

$$B = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{14}$$

then $\kappa(B) = 2$ and $\kappa(C) = 1$. Adding the two matrices together gives

$$B + C = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix} \tag{15}$$

and hence $\kappa(B+C) = 5$. Therefore $\kappa(B+C) > \kappa(B) + \kappa(C)$.

(c) Let A be an invertible 2×2 symmetric matrix. First, note that

$$\|2A\| = \max_{v \neq 0} \frac{\|2Av\|}{\|v\|} = \max_{v \neq 0} \frac{2\|Av\|}{\|v\|} = 2 \max_{v \neq 0} \frac{\|Av\|}{\|v\|} = 2\|A\|. \tag{16}$$

Similarly, note that $\|(2A)^{-1}\| = \|\frac{1}{2}A^{-1}\| = \frac{1}{2}\|A^{-1}\|$. Hence

$$\kappa(2A) = \|2A\| \, \|(2A)^{-1}\| = 2\|A\| \times \frac{1}{2}\|A^{-1}\| = \|A\| \, \|A^{-1}\| = \kappa(A). \tag{17}$$

Now suppose that A is a symmetric invertible matrix. Then there exists an orthogonal matrix Q and a diagonal matrix D such that

$$A = Q^{\mathsf{T}} D Q. \tag{18}$$

Since $Q^{\mathsf{T}} Q = Q Q^{\mathsf{T}} = I$, it follows that

$$A^2 = Q^{\mathsf{T}} D Q Q^{\mathsf{T}} D Q = Q^{\mathsf{T}} D^2 Q. \tag{19}$$

The matrix norm of $\|A\|$ is

$$\|A\| = \max_{v \neq 0} \frac{\|Q^{\mathsf{T}} D Q v\|}{\|v\|}. \tag{20}$$

Since Q corresponds to a rotation or reflection, it preserves distances under the Euclidean norm and hence $\|Qw\| = \|w\| = \|Q^{\mathsf{T}} w\|$ for an arbitrary vector w. Therefore

$$\|A\| = \max_{v \neq 0} \frac{\|DQv\|}{\|Qv\|} = \max_{u \neq 0} \frac{\|Du\|}{\|u\|} = \|D\| \tag{21}$$

where $u = Qv$. Similarly $\|A^{-1}\| = \|Q^{\mathsf{T}} D^{-1} Q\|$, and since D^{-1} is also diagonal it follows that $\|A^{-1}\| = \|D^{-1}\|$, so $\kappa(A) = \kappa(D)$. With reference to the condition number notes, $\kappa(A) = |\alpha\beta^{-1}|$ where α is the diagonal entry with largest magnitude and $|\beta|$ is the diagonal entry with the smallest entry with smallest magnitude.

Since D^2 is also diagonal, it follows that $\|A^2\| = \|D^2\|$. The diagonal entry of D^2 with the largest amplitude will be α^2, and the diagonal entry of D^2 with the smallest amplitude will be β^{-2}. Hence

$$\kappa(A^2) = |\alpha^2\beta^{-2}| = (\kappa(A))^2 . \tag{22}$$

(d) The result for that $\kappa(2A) = \kappa(A)$ is true for arbitrary 2×2 invertible matrices. The derivation that was considered in part (c) did not rely on the matrix being symmetric.

The result about $\kappa(A^2)$ does not generalize to arbitrary matrices. If

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{23}$$

then

$$A^2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} . \tag{24}$$

One can numerically verify that $\kappa(A^2) = 5.828$ but $(\kappa(A))^2 = 6.854$, so the two do not agree.

5. (a) The program `gamma_vander.py` computes the polynomials p_k as described in the question, using the Vandermonde matrix method. Since this question only considers low-order polynomials, the Vandermonde matrix is well-conditioned and gives accurate answers. With Vandermonde interpolation, it is more straightforward to evaluate the coefficients b_i of the polynomial when expressed as $\sum_i b_i x^i$.

The polynomials are shown in Table 1, along with the values of $|p_k(\frac{3}{2}) - \Gamma(\frac{3}{2})|$. In this case, the most accurate value is given when $k = 2$. This is because the rapidly growing values of $\Gamma(x)$ cause large oscillations in the interpolating polynomial when k gets large.

(b) Table 2 lists the polynomials q_k that interpolate $\log(\Gamma(x))$ at $x = 1, 2, \ldots, k+1$. The absolute errors between $\exp(q_k(\frac{3}{2}))$ and $\Gamma(\frac{3}{2})$ are also shown. In this case the most accurate value is given by $k = 5$. Because the points of $\log(\Gamma(x))$ do not grow as rapidly, the interpolating polynomial can match the points better.

(c) The program `gamma_lagr.py` computes the polynomials p_k and q_k using Lagrange interpolation. This method is better suited to large k, because it does not suffer from the numerical conditioning problems of the Vandermonde matrix.

k	$p_k(x)$	$\lvert p_k(\frac{3}{2}) - \Gamma(\frac{3}{2})\rvert$
1	1	0.11377
2	$\frac{1}{2}x^2 - \frac{3}{2}x + 2$	0.011227
3	$\frac{1}{3}x^3 - \frac{3}{2}x^2 + \frac{13}{6}x$	0.11377
4	$\frac{3}{8}x^4 - \frac{41}{12}x^3 + \frac{93}{8}x^2 - \frac{199}{12}x + 9$	0.23779
5	$\frac{11}{30}x^5 - \frac{41}{8}x^4 + \frac{111}{4}x^3 - \frac{567}{8}x^2 + \frac{5033}{60}x - 35$	0.96534

Table 1: Interpolating polynomials $p_k(x)$ for the gamma function for $k = 1, 2, 3, 4, 5$, using control points at $x = 1, 2, \ldots, k+1$. The absolute error of the polynomial to the gamma function at $x = \frac{3}{2}$ is also listed.

k	$q_k(x)$	$\lvert \exp(q_k(\frac{3}{2})) - \Gamma(\frac{3}{2})\rvert$
1	0	0.11377
2	$0.34657x^2 - 1.0397x + 0.69315$	0.03777
3	$-0.047947x^3 + 0.63426x^2 - 1.5671x^1 + 0.98083$	0.014437
4	$0.0070791x^4 - 0.11874x^3 + 0.88203x^2 + -1.9211x^1 + 1.1507$	0.0084790
5	$-0.00097212x^5 + 0.021661x^4 - 0.20137x^3 + 1.1008x^2 - 2.1875x^1 + 1.2674$	0.0056296

Table 2: Interpolating polynomials $q_k(x)$ for $\log(\Gamma(x))$ for $k = 1, 2, 3, 4, 5$, using control points at $x = 1, 2, \ldots, k+1$. The absolute error of $\exp(q_k(x))$ to the gamma function at $x = \frac{3}{2}$ is also listed.

Figure 2 shows the absolute errors $|p_k(\frac{3}{2}) - \Gamma(\frac{3}{2})|$ and $|\exp(q_k(\frac{3}{2})) - \Gamma(\frac{3}{2})|$ for $k = 0, 1, 2, \ldots, 30$. The errors for p_k diverge approximately exponentially. The errors for $\exp(q_k)$ appear to converge as k gets large, but the convergence rate is very slow. It is unclear whether the polynomials will converge to the correct value as $k \to \infty$, or if there will be a non-zero error in the limit.

6. (a) The Taylor series of f at $x - h$ and $x + h$ are

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) + O(h^3), \tag{25}$$

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + O(h^3), \tag{26}$$

respectively. Adding these two equations together shows that

$$f(x-h) + f(x+h) = 2f(x) + h^2 f''(x) + O(h^3) \tag{27}$$

and therefore

$$h^2 f''(x) = f(x-h) - 2f(x) + f(x+h) + O(h^3). \tag{28}$$

Dividing through by h^2 shows that

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} + O(h). \tag{29}$$

Therefore $\alpha = \gamma = 1$ and $\beta = -2$.

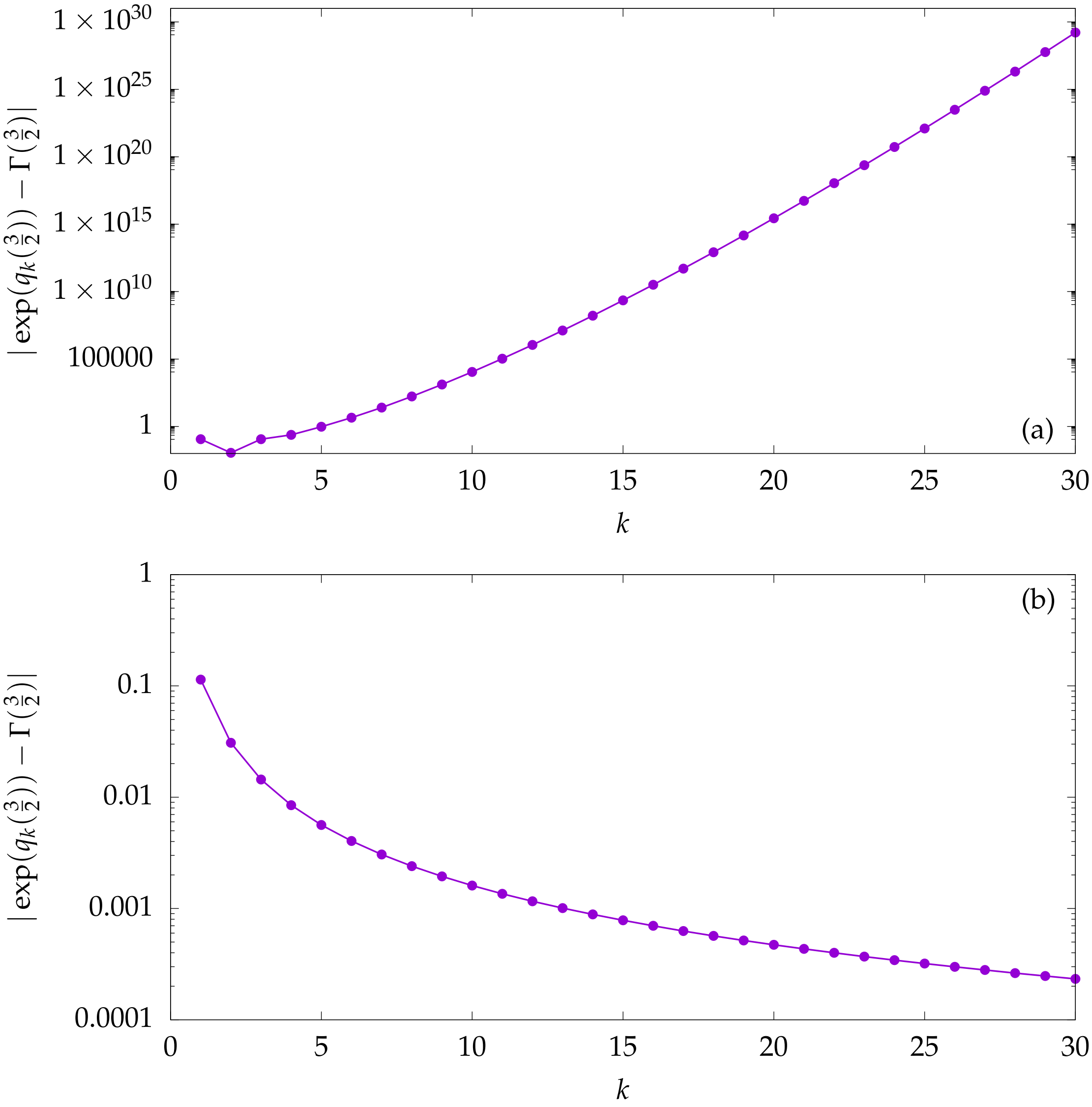

Figure 2: Errors in the interpolating polynomials, $|p_k(\frac{3}{2}) - \Gamma(\frac{3}{2})|$ and $|\exp(q_k(\frac{3}{2})) - \Gamma(\frac{3}{2})|$, for $k = 0, 1, 2, \ldots, 30$.

(b) The Taylor series of f at $x+h$ is

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(x) + \frac{h^5}{120}f^{(5)}(x) + O(h^6) \tag{30}$$

and the Taylor series of f' at $x+h$ is

$$f'(x+h) = f'(x) + hf''(x) + \frac{h^2}{2}f'''(x) + \frac{h^3}{6}f^{(4)}(x) + \frac{h^4}{24}f^{(5)}(x) + O(h^5). \tag{31}$$

Analogous expressions exist for $f(x-h)$ and $f'(x-h)$ by switching the signs of the odd powers of h in Eqs. (30) & (31). Consider the expression

$$f''(x) = \frac{af(x-h) + bf(x) + cf(x+h)}{h^2} + \frac{rf'(x-h) + sf'(x) + tf'(x+h)}{h} + O(h^4). \tag{32}$$

Equating terms proportional to $h^{-2}f(x), h^{-1}f'(x), f''(x), \ldots, h^3 f^{(5)}(x)$ yields the linear system

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & -1 & 0 & 1 \\ -\frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{24} & 0 & \frac{1}{24} & -\frac{1}{6} & 0 & \frac{1}{6} \\ -\frac{1}{120} & 0 & \frac{1}{120} & \frac{1}{24} & 0 & \frac{1}{24} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ r \\ s \\ t \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{33}$$

Solving this system using the program `deriv.py` shows that $(a, b, c, r, s, t) = (2, -4, 2, \frac{1}{2}, 0, -\frac{1}{2})$, and therefore

$$f''(x) = \frac{2f(x-h) - 4f(x) + 2f(x+h)}{h^2} + \frac{f'(x-h) - f'(x+h)}{2h} + O(h^4). \tag{34}$$

(c) For the function $f(x) = e^{4\sin x}$, the first and second derivatives are

$$f'(x) = 4\cos x e^{4\sin x}, \qquad f''(x) = 4e^{4\sin x}(4\cos^2 x - \sin x). \tag{35}$$

The program `deriv2.py` calculates second derivative of f using the formulae in Eqs. (29) & (34), using grid spacings of $h = 2^{-k}$ for $k = 0, 1, 2, \ldots, 23$. The absolute error magnitudes E are shown in Fig. 3. For Eq. (29), the errors follow

$$E \approx 125.9h^{1.995}. \tag{36}$$

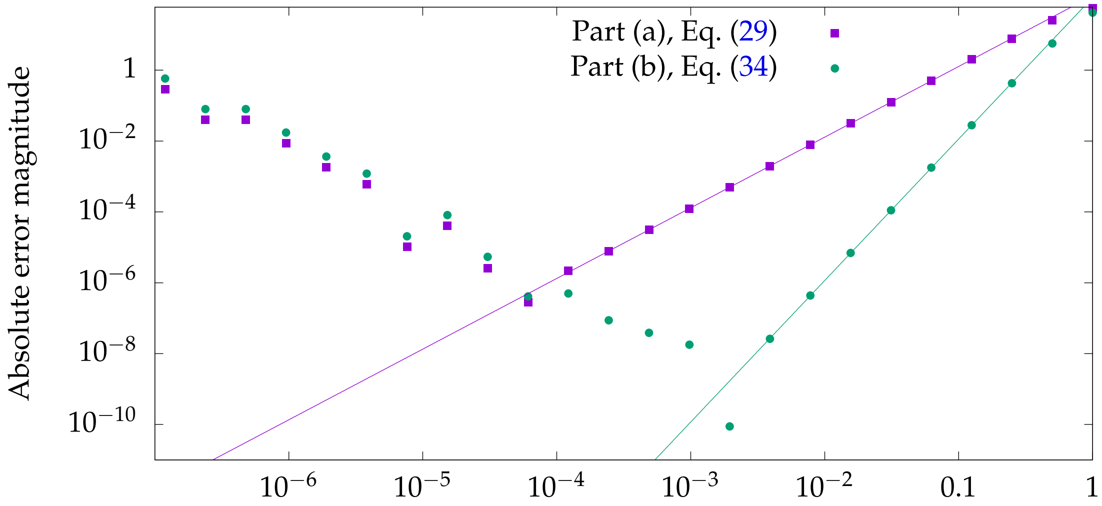

Figure 3: Absolute error magnitudes for the two finite-difference formulae for the second derivative of $f(x) = e^{4\sin x}$ at $x = 1$.

Hence the error appears to scale like h^2 one order higher than $O(h)$ as given in Eq. (29). This is still consistent with the formula, and is due to cancellation of the leading order error term due to symmetry. For Eq. (34), the errors follow

$$E \approx 112.7h^{3.994} \tag{37}$$

which is consistent the expected $O(h^4)$ error.