Práctica 2: Zeros

Marco Praderio y Marta Cavero

1. Considerar la equación polinómica

$$x^3 = x + 40$$

1.1. Comprobar que al evaluar en doble y simple precisión la expresión de la raiz real de la ecuación anterior (que se obtiene a partir de las fórmulas de Cardano.Vieta) dada por la fórula

$$\alpha = \left(20 + \frac{1}{9}\sqrt{32397}\right)^{\frac{1}{3}} + \left(20 - \frac{1}{9}\sqrt{32397}\right)^{\frac{1}{3}}$$

proporciona un resultado con error de cancelación. Estimar este error.

Para calcular la raiz real del polinomio

$$P(x) = x^3 - x - 40$$

utilizaremos la fórmula de Cardano Vieta dada por

$$\alpha = \left(20 + \frac{1}{9}\sqrt{32397}\right)^{\frac{1}{3}} + \left(20 - \frac{1}{9}\sqrt{32397}\right)^{\frac{1}{3}}$$

Esperamos que esta formula produzca un error de cancelación al hacer la operación

$$\left(20 - \frac{1}{9}\sqrt{32397}\right)^{\frac{1}{3}}$$

Para poder observar claramente este error hemos realizado un programa en C para obtener el resultado numérico de la fórmula de Cardano Vieta calculada con precisiónes simple y doble.
Tras aplicar la fórmula obtenemos el valor 3.517393514 calculando con precisión doble y el valor 3.517362595 calculando con precisión simple los cuales muestran discrepancias a partir del quinto decimal. Aunque a primera vista esta diferencia puede no parecer muy grande si consideramos que la diferencia debido errores de representación entre un valor normalizado guardado con precisión simple y otro guardado con precisión doble es de aproximadamente 10^{-7} entonces nos daremos cuenta de que tener discrepancias a partir del quinto decimal es significativo.
Para estudiar mejor como se propaga el error debido a la cancelación hemos hecho los cálculos de propagación del error en punto flotante obteniendo que la cancelación aumenta el error relativo en un orden de 10^6 Además evaluando el polinomio en la raiz obtenida con doble precisión no obtenemos como resultado 0 como deberiamos si no que obtenemos -3.275602011854062e-12 lo cual nos confirma que efectivamente hay errores relevantes asociado al calculo de α mediante la fórmula de Cardano Vieta dada.

1.2. Aplica el método de Newton a la función empezando con $x_0 = 2$.

$$f(x) = x^3 - x - 40$$

utilizando precisión simple y doble. Estimar el número de iteraciones necesarias para obtener una aproximación de la raiz con 8 y 16 decimales correctos respectivamente.

Hemos realizamos un programa que aplica el método de Newton para encontrar la raíz real del polinomio

$$P(x) = x^3 - x - 40$$

tomando como raíz inicial el valor $x_0 = 2$.
Por otro lado hemos realizado un cálculo teórico para predecir el número máximo de iteraciones para obtener una aproximación de la raiz con 8 o 16 decimales correctos empezando en el punto x=2 (suponiendo operaciones exactas). En el caso de 8 decimales correctos hemos calculado que son necesarias como mucho 26 iteraciones. Mientras que en el caso de querer 16 decimales correctos son necesarias un máximo de 49 iteraciones.
Este cálculo teórico lo hemos realizado utilizado la siguiente fórmula dada en clase de teoria para estimar una cota superior del error en la aproximación de la raíz.

$$\frac{K^{n-1}}{1-K}|x_2 - x_1| > \varepsilon \Longrightarrow n \leq 1 + \frac{\ln\left(\frac{\varepsilon(1-K)}{|x_1-x_2|}\right)}{\ln(K)}$$

Donde $K = 0{,}45$ es una cota superior de la derivada primera del método de punt fijo dado por el método de Newton[1] en el intervalo entre la raiz y el punto $x_1 = 5{,}09090909091$, donde se encuentran todos los elementos de la sucesión exceptuando x_0. Mientras que ε es el valor absoluto del error en el iterado número n. Insistimos nuevamente en que estos cálculos no tienen en cuenta errores en las operaciónes y representaciones de resultados que se podrán detectar a partir de la diferencia entre los resultados obtenidos realizando los cálculos con precisión simple y con precisión doble. Además es importante mencionar que estas cotas superiores en el nombre de iteraciones no tienen porque ser iguales al número de iteraciones necesarias para aproximar la raiz con la precisión deseada. de hecho, dado que la fórmula dada sirve para acotar el número de iteraciones en el caso de sucesiones que convergen linealmente y el método de Newton converge de forma cuadrática lo más probable es que se necesiten menos iteraciones para obtener el resultado deseado.

Comprobemos los resultados teóricos de forma numérica. Para hacerlo calcularemos la sucesión hasta que la diferencia en valor absoluto entre un iterado y el iterado sucesivo sea menor que $\frac{10^{-8}}{2}$ o que $\frac{10^{-16}}{2}$ y nos detendremos cuando se den estas dos condiciones. Estas condiciones de parada son suficientes para asegurar que los resultados obtenidos tendrán 8 y 16 decimales correctos dado que el método empleado es cuadrático y por lo tanto el número de cifras significativas tiende a doblarse cuando los iterados están lo suficientemente cerca de la convergencia (cosa que pasará cuando se cumplan las condiciones mencionadas). Tras ejecutar el programa unas cuantas veces hemos notado que dado que la raiz real del polinomio $p(x)$ és del orden de las unidades entonces representarla con 8 decimales correctos és equivalente a decir que se tienen mas de 8 cifras significativas. Dado que la mantissa de las variables tipo float està limitada a 24 bits implicitos es imposible para un float representar 8 cifras significativas. Por lo tanto no tiene sentido decir que la sucesión calculada con precisión simple llegue a tener 8 o 16 decimales correctos dado que, antes de que eso, llegaremos a un iterado con el màximo número posible de decimales correctos y a partir de allí ya no aumentará la precisión[2]

Teniendo esto en mente y tras ejecutar el programa vemos que en la iteración número 6 del calculo con precisión simple llegamos a tener el máximo número de decimales correctos (6) con el valor 3.5173936. En el caso del calculo con precisión simple obtenemos en el iterado 7 el resultado 3.517393514 con 8 decimales correctos. En el iterado sucesivo (el 8) doblamos el número de decimales correctos (16) con el valor 3.517393514052818.

1.3. Considera la ecuación polinómica

$$x^3 = x + 400$$

Obtén una fórmula de Cardano.-Vieta para el calculo de la raiz real, β. Comprueba que dicha raiz cumple que

$$2 \leq \beta \leq 8$$

Estima el error de cancelación calculando la fórmula explicita en doble precisión.

Para calcular la raíz real del polinomio $P(x) = x^3 - x - 400$ aplicaremos la fórmula de Cardano Vieta dada por

$$x = \sqrt[3]{200 + \sqrt{200^2 - \frac{1}{3}^3}} + \sqrt[3]{200 - \sqrt{200^2 - 1/3^3}}$$

[1] $g(x) = x - \frac{x^3 - x - 40}{3x^2 - 1}$
[2] El mismo argumento és válido para la variable con precisión doble pero con un número mucho más elevado de decimales correctos.

La cual producirá error de cancelación en la resta de la segunda raiz cúbica en cuanto 200 es aproximadamente igual a $\sqrt{200^2 - 1/3^3}$. Realizando el cálculo con doble precisión obtenemos como resultado el valor

$$\alpha = 7{,}413302725855254$$

el cual está comprendido entre 2 y 8 pero está claramente contaminado con error de cancelación dado que $P(\alpha) \neq 0$ al contrario de lo que se deberia cumplir si el resultado fuese exacto.

1.4. Aplicar los siguientes métodos iterativos para obtener los 16 decimales correctos de la raiz y compara el orden de convergencia numérica. Considerar la posible aceleración mediante la iteración de Aitken sobre las sucesiones de iterados obtenidas. Discutir en su caso la mejora.

Calculemos ahora la raiz del polinomio con hasta 16 decimales correctos mediante los métodos de Bisección, de la secante y el método de Newton. Sucesivamente intentaremos aplicar el Método de aceleración de Aitken para acelerar las tres sucesiones obtenidas. Es fácil comprobar que, teóricamente, el método de Aitken se puede aplicar a las tres sucesiones siempre y cuando las sucesiones tengan un mínimo de iterados[3] y la raiz exacta no forme parte de las sucesiones.

1.4.1. Método de la bisección partiendo del intervalo [2,8]

Aplicando el método de la bisección en el intervalo [2,8] y con 50 iteraciones hemos llegado al resultado 7.413302725857898 Con únicamente error de representación dado que, afortunadamente, hemos caido en el valor exacto de la raiz en uno de los iterados.

El métode de acceleración de Aitken no ha logrado reducir el número de iterados. Esto puede ser debido al hecho que la sucesión contenga inesperadamente el valor exacto[4] cosa que no cumple las hipótesis para aplicar el metodo de aceleración de Aitken. No obstante si dispusieramos de precisión infinita para así poder tener infinitos iterados del método de la bisección[5] el método acelerado por Aitken deberia, teóricamente, acabar convergiendo de manera más rápida que el método sin acelerar.

1.4.2. Método de la secante partiendo del intervalo [2,8]

Aplicando el método de la secante en el intervalo [2,8] y con 8 iteraciones hemos llegado al resultado 7.413302725857898 Con únicamente error de representación dado que, afortunadamente. al igual que en el caso del método de la bisección, hemos caído en el valor exacto de la raiz en uno de los iterados.

El métode de acceleración de Aitken no ha logrado acelerar la sucesión. Esto es probablemente debido al hecho que la sucesión dada por el método de la secante tiene únicamente 8 iterados y, por lo tanto, no tiene sentido acelerarla dado que ya es bastante rápida por si sola.

1.4.3. Método de Newton partiendo del pivote $x_0 = 2$

Aplicando el método de Newton y con 10 iteraciones hemos llegado al resultado 7.413302725857898 Con un error tan pequeño que la máquina no es capaz de representarlo y, por lo tanto, es equivalente a tener únicamente error de representación[6].

El métode de acceleración de Aitken no ha logrado acelerar la sucesión. Esto lo podemos explicar de la misma manera en la que hemos explicado este fenómeno para el método de la secante. No podemos acelerar la sucesión porque esta contiene únicamente 10 elementos y, por lo tanto, no tiene sentido acelerarla dado que ya es bastante rápida por si sola..

[3] Auque la sucesión tenga suficientes iterados como para poder aplicar el método de Aitken puede no tener suficienter iterados como para que se vean los efectos de este método.

[4] Cuando utilizo la palabra exacta me refiero a que la máquina no es capaz de distinguir entre la aproximación obtenida y el 0 de la función de manera que evaluando la función en nuestra aproximación obtenemos 0.

[5] Con precisión infinita la probabilidad de acertar por casualidad la raiz exacta del polinomio seria 0 y por lo tanto podriamos seguir aplicando el método de la bisección indeterminadamente.

[6] Sabemos por las clases de teoria que no es posible que un iterado del método de Newton coincida con el 0 de la función a analizar a menos que el primer iterado también coincida y no es el caso.

2. Sea la ecuación $f(x) = 0$ con $f(x)$ continuamente derivable, $x*$ una raiz simple, $f(x*) = 0$, con $f'(x) \neq 0$ en un entorn de $x*$. Considerar la iteración

$$x_{k+1} = x_k - b_k f(x_k)$$

donde

$$b_{k+1} = b_k(2 - f'(x_{k+1}b_k))$$

partiendo de un pivote x_0 suficientemente próximo a $x*$ con $b_0 = \frac{1}{f'(x_0)}$.

2.1. Aplicar la iteración de la ecuación polinómica del Problema 1, $x^3 = x+400$, tomando $x_0 = 6$ y $b_0 = \frac{1}{3x_0^2-1}$. Estudiar el orden de convergencia y calcular cuantas iteraciones son necesarias para tener una precisión de 13 cifras decimales correctas.

Aplicando la sucesión presentada en el problema 2 para calcular la raiz real del polinomio $P(x) = x^3 - x - 400$ calculada tres veces en el apartado anterior con una precisión de 16 decimales obtenemos que, en la iteración número 7, llegamos al resultado

7,413302725858

El cual tiene 13 decimales correctos como podemos observar comparándolo con los resultados obtenidos en el ejercicio anterior[7].

Si ahora denotamos por α el valor de la raiz encontrado en el ejercico anterior[8] y calculamos

$$\frac{x_k - \alpha}{(x_{k-1} - \alpha)^n}$$

para n=1,2,3 y k=5,6,7 los tres iterados previos a obtener la aproximación de la raiz con 13 decimales correctos obtenemos.

	k=5	k=6	k=7
n=1	0.00763839	7.00744e-05	4.42494e-07
n=2	2.0369	2.44639	220.452
n=3	543.173	85406.7	1.0983e+11

El hecho que los resultados obtenidos para n=1 tienden a 0 nos indica que el orden de convergencia de la sucesión es al menos lineal. Por otro lado el hecho que, para $n > 1$ los resultados obtenidos crezcan descontroladamente nos indica que la sucesión no llega a tener orden de convergencia cuadrático. No obstante, dado que los dos primeros cocientes en el caso de n=2 són bastante similares nos surge la duda de si el hecho de que el tercer quociente aumente es debido únicamente a la precisión finita de la doble precisión y no implique por lo tanto que el orden de convergencia no llegue a ser cuadrático. Para comprobrarlo repetiremos los calculos utilizando ahora variables del typo __float128 las cuales tienen el doble de precisión que las variables tipo double obteniendo así los resultados que podemos observar en la siguiente tabla.

	k=5	k=6	k=7
n=1	0.00763839	7.00744e-05	5.84275e-08
n=2	2.0369	2.44639	29.1088
n=3	543.173	85406.7	1.45021e+10

Como podemos observar el primer y segundo cocientes se han mantenido iguales mientras que el tercer quociente ha disminuido para los tres valores de n[9]. No obstante, en el caso de n=2, sigue siendo sensiblemente mayor que los

[7] Podemos considerar los resultados obtenidos en el ejercicio anterior como exactos dado que sabemos que tienen almenos 16 decimales correctos y estamos controlando si el resultado obtenido en este apartado tiene 13 decimales correctos. No podriamos considerar los resultados obtenidos en el apartado anterior como correctos si buscaramos comprobar si el resultado obtenido tiene 17 o más decimales correctos.

[8] $\alpha = 7{,}413302725857898$ que supondremos exacto tan y como ya hemos dicho.

[9] Esto indica tal y como cabía esperar que cuanto más se acercan los iterados a la convergencia más importancia toman los errores debidos a la finitud de las variables utilizadas.

dos primeros cocientes por lo tanto no podemos afirmar que el orden de convergencia sea cuadrático. Pero podemos sin duda alguna afirmar que el orden de convergencia, no obstante puede no llegue a ser cuadrático, es superior al lineal. Tenemos por lo tanto un orden de convergencia fraccionario entre 1 y 2 bastante mas cercano al 2 que al 1.

3. Definimos la iteración para $k = 1, 2, 3, \cdots, p_x = \frac{2a_k^2}{s_k}$

$$a_k = \frac{a_{k-1}}{2} \quad b_k = \sqrt{a_{k-1}b_{k-1}}$$

$$c_k = a_k^2 - b_k^2 \quad s_k = s_{k-1} - 2^k c_k$$

tomando $a_0 = 1$, $b_0 = \frac{1}{\sqrt{2}}$ y $s_0 = \frac{1}{2}$. Teoricamente p_k converge al número π. Verificar que la convergencia es cuadrática. Determinar cuantas iteraciones debemos realizar para que el error absoluto comience a crecer y la convergencia numérica degenere (debido a la precisión finita). Investigar si la convergencia puede acelerarse.

Calculando numéricamente con precisión doble la sucesión descrita en el problema 3 obtenemos los siguientes resultados notables.

- En el iterado número 5 el error ha empezado a crecer pasando de 3.819167204710538e-14 a 8.704148513061227e-14
- El error va en aumento a cada iterado hasta llegar al iterado 51
- En el iterado número 51 el error ha empezado a decrecer pasando de 36.48482199328234 a 5.785052711079649
- En el iterado número 1000 todavia no se ha podido mejorar el resultado obtenido en el iterado 4 y dejamos de calcular elementos de la sucesión.

El hecho de que hasta el iterado 1000 no se haya podido mejorar el resultado obtenido en el iterado 4 nos indica que, a partir de este iterado, la convergencia degenera a causa de la precisión finita.

No hemos conseguido obtener 16 decimales correctos de π aplicando el método descrito en el apartado 3, no obstante si que hemos logrado,en el iterado 4, obtener una aproximación de π con 13 decimales correctos dada por el valor 3.141592653589831 el cual lleva asociado un error de 3.81917e-14

Si ahora calculamos el error asociado a las últimas 4 iteraciones previas a la mejor aproximación (incluida) de π obtenida encontramos:

iterado número 1	0,04608
iterado número 2	$8{,}76397e-05$
iterado número 3	$3{,}05667e-10$
iterado número 4	$3{,}81917e-14$

El error debido a la finitud de la representación de tipo double no nos permite hacer un analisis preciso del orden de convergencia dado que, a partir de los resultados obtenidos, podria resultar que el orden del convergecia fuese tanto lineal (con una constante assimptotica del orden de 10^{-4}) como cuadrático (el numero de cifras significativa se duplica en los iterados 2 y 3 y es razonable pensar que no se duplica en la última iteración debido a que la precisión limitada de la variable tipo double no se lo permite al generar errores de magnitud superior a 10^{-20}). Será por lo tanto necesario repetir los calculos con una precisión mayor para poder dar una respuesta concluyente.

Efectivamente repitiendo los calculos per esta vez utilizando variables del tipo __float128 obtenemos que el mejor resultado se sigue oteniendo en el iterado 4^{10} pero esta vez con una precisión mucho mayor ($\pi = 3{,}141592653589793 \pm 4{,}79389e-18$ con 17 decimales correctos). Si ahora calculamos de nuevo el error asociado a las últimas 4 iteraciones previas a la mejor aproximación (incluida) de π obtenida encontramos:

iterado número 1	0,04608
iterado número 2	$8{,}76397e-05$
iterado número 3	$3{,}05653e-10$
iterado número 4	$4{,}79389e-18$

[10] Despues de este iterado el error no para de crecer hasta el iterado 1000 en el que dejamos de iterar. Este resultado nos sigue indicando que la convergencia degenera a partir del iterado 4 debido a la precisión finita.

Como podemos observar los resultados obtenidos se han mantenido casi iguales a los que habiamos visto utilizando precisión doble excepto por el último que ha disminuído considerablemente de manera que ahora se puede ver claramente que el orden de convergencia es cuadrático dado que el número de cifras significativas para iterados cerca de la convergéncia se duplica a cada iterado.

Por último y teniendo en cuenta que las sucesiones obtenidas tienen únicamente 4 iterados no tiene sentido aplicar el método de Aitken para intentar acelerar la sucesión.

4. Cálculo aproximado de raices cuadradas.

El objetivo es obtener una aproximación de la raiz cuadrada de un número utilizando la expresión

$$\sqrt{1+x} = f(x)\sqrt{1+g(x)}$$

donde g es un infinitésimo de orden más pequeño que x para x tendiendo a 0. Si elegimos $f(x)$ como una aproximación de $\sqrt{1+x}$ entonces se puede calcular $g(x)$ como

$$g(x) = \frac{1+x}{f(x)^2} - 1$$

4.1. La función $f(x)$ puede elegirse como una función racional $p(x)/q(x)$, tal que p y q tienen el mismo grado y su desarrollo de MacLaurin coincide con el de $\sqrt{1+x}$ hasta cierto grado. Hallar una función racional, $f(x) := p(x)/q(x)$, cociente de dos linealesm tal que el desarrollo de MacLaurin de $p(x) - \sqrt{1+x}q(x)$ tenga los tres primeros términos nulos. Esta función f se conoce como el aproximante de Padé de la función $\sqrt{1+x}$.

Tras reescribir los polinomios $p(x)$ y $q(x)$ como $p(x) = a' + b'x$ y $q(x) = a + bx$ imponemos como condicion que las tres primeras derivadas de $p(x) - \sqrt{1+x}q(x)$ evaluadas en el origen sean nulas.[11] De esta manera obtenemos las siguientes ecuaciones

$$\begin{cases} a' = a \\ 2b' = a + 2b \\ a = 4b \end{cases}$$

Resolviendo el sistema en función del parámetro b obtenemos.

$$\begin{array}{ll} a = 4b & a' = 4b \\ b = b & b' = 3b \end{array}$$

Por tanto la función f(x) es:

$$f(x) = \frac{b}{b}\frac{3x+4}{x+4} = \frac{3x+4}{x+4}$$

4.2. Siendo $a_0 = x$, $a_{n+1} = g(a_n)$ y $b_n = f(a_n)$. Comprobar que

$$\sqrt{1+x} = \left(\prod_{j=0}^{k} b_j\right)\sqrt{1+a_{k+1}}$$

Realizaremos esta demostración haciendo Uso del método inductivo.

Para $k = 0$ tenemos la identidad

$$\sqrt{1+x} = b_0 \cdot \sqrt{1+a_1} = f(a_0)\cdot\sqrt{1+f(a_0)} = f(x)\cdot\sqrt{1+g(x)}$$

la cual es cierta por definición de g y f.

Suponiendo que la identidad es cierta para $k-1$ podemos escribir el iterado número k como

$$\begin{aligned} \left(\prod_{j=0}^{k} b_j\right)\sqrt{1+a_{k+1}} &= \left(\prod_{j=0}^{k} f(a_j)\right)\sqrt{1+g(a_k)} = \\ &= \left(\prod_{j=0}^{k-1} f(a_j)\right) f(a_k)\sqrt{1+g(a_k)}\frac{\sqrt{1+a_k}}{\sqrt{1+a_k}} = \\ &= \left(\prod_{j=0}^{k-1} f(a_j)\right)\sqrt{1+g(a_{k-1})}\frac{f(a_k)\sqrt{1+g(a_k)}}{\sqrt{1+a_k}} \end{aligned} \tag{1}$$

[11] Decir esto es equivalente a decir que los tres primeros terminos del desarrollo de MacLaurin de $p(x) - \sqrt{1+x}q(x)$ són nulos.

por hipótesis de inducción tenemos que tenemos que

$$\left(\prod_{j=0}^{n-1} f(a_j)\right)\sqrt{1+g(a_{n-1})} = \sqrt{1+x}$$

y por definición de g y f tenemos que

$$f(a_k)\sqrt{1+g(a_k)} = \sqrt{1+a_k} \quad \text{para todo } a_k > 0$$

[12] Por lo tanto podemos reescribir la identidad (1) como

$$\left(\prod_{j=0}^{n} b_j\right)\sqrt{1+a_{n+1}} = \sqrt{1+x}\frac{\sqrt{1+a_n}}{\sqrt{1+a_n}} = \sqrt{1+x}$$

Queda por lo tanto demostrado que, para todo $k \in \mathbb{N} \cup \{0\}$, se cumple la identidad

$$\sqrt{1+x} = \left(\prod_{j=0}^{k} b_j\right)\sqrt{1+a_{k+1}} \tag{2}$$

4.3. Realiza programas en C para experimentar con el algoritmo anterior para una elección de f commo el cociente de dos polinommios de tercer grado.

Antes de realizar el programa en C será necesario escribir $f(x) = \frac{p(x)}{q(x)}$ como cociente de polinomios de tercer grado de manera que se anulen el máximo número de términos iniciales del desarrollo de MacLaurin de $h(x) = p(x) - \sqrt{1+x}q(x)$ tal y como hemos hecho el apartado (4.1). Reescribiendo $p(x)$ y $q(x)$ como $p(x) = a' + b'x + c'x^2 + d'x^3$ y $q(x) = a + bx + cx^2 + dx^3$ y imponiendo que se anulen los primeros siete terminos del desarrollo de MacLaurin obtenemos el sistema

$$\begin{cases} a' - a = 0 \\ -\frac{a}{2} + b' - b = 0 \\ \frac{a}{8} - \frac{b}{2} + c' - c = 0 \\ -\frac{a}{16} + \frac{b}{8} - \frac{c}{2} + d' - d = 0 \\ -\frac{5a}{64} + \frac{b}{8} - \frac{c}{4} + d = 0 \\ \frac{7a}{32} - \frac{5a}{16} + \frac{c}{2} - d = 0 \\ -\frac{21a}{64} + \frac{7b}{16} - \frac{5c}{8} + d = 0 \end{cases}$$

Resolviendo en función del parámetro d obtenemos

$$\begin{aligned} a &= 64d & a' &= 64d \\ b &= 80d & b' &= 112d \\ c &= 24d & c' &= 56d \\ d &= d & d' &= 7d \end{aligned}$$

por lo tanto la función f escrita como fracción de dos polinomios de tercer grado és

$$f(x) = \frac{d}{d}\frac{7x^3 + 56x^2 + 112x + 64}{x^3 + 24x^2 + 80x + 64} = \frac{7x^3 + 56x^2 + 112x + 64}{x^3 + 24x^2 + 80x + 64}$$

Ahora podemos realizar el programa que se encuentra adjunto el la carpeta CODIGOS bajo el nombre CalculoRaiz o en Programas práctica 2 bajo el mismo nombre. es interesante notar que la convergencia resulta extremadamente rápida[13]. En los siguientes apartados estudiaremos analíticamente con más profundidad esta velocidad de convergencia.

[12] Dado que para todo $x > 0$ se cumlpe que $g(x) > 0$ entonces esta identidad es válida para todo elemento de la sucesión $\{a_n\}$ siempre y cuando $a_0 = 0 > x$

[13] calcula la raiz de 3 con un error de $1.e-16$ haciendo uso de únicamente tres iterados

4.4. Hallar $n > 0$ tal que la función, g, calculada mediante $g(x)$ a partir de la racional f obtenidas en los apartados (4.1) y (4.3), cumpla que $g(x) = O(x^n)$.

Antes de empezar notemos que si $g(x) = O(x^n)$ y la sucesión $\{a_n\}$ definida por $a_0 = x$ y $a_{n+1} = g(a_n)$ converge hacia 0. Entonces la sucesión convergirá con orden de convergencia n. Si ahora tenemos en cuenta la siguiente identidad ya demostrada

$$\sqrt{1+x} = \left(\prod_{j=0}^{n} b_j\right)\sqrt{1+a_{n+1}}$$

podremos intuir[14] que el producto $\prod_{j=0}^{n} b_j$ tiende a $\sqrt{1+x}$ con el mismo orden de convergencia. Por lo tanto la pregunta expuesta en este apartado nos ayudará a estudiar la convergencia de la sucesión $\{\prod_{j=0}^{n} b_j\}$

apartado (4.1) lineal La expresión de f como cociente de dos lineales es $f(x) = \frac{3x+4}{x+4}$ y podemos calcular la función g siguiendo la definición:

$$g(x) = \frac{1+x}{(f(x))^2} - 1 = \frac{x^3}{9x^2+24x+16}$$

Vamos a ver que $g(x)$ es un infinitésimo equivalente a x^3 o, lo que es lo mismo $g(x) = o(x^3)$. Veremos que esto es cierto calculando el límite de $x \to 0$ de $\frac{x^3}{g(x)}$ y viendo que este es finito y distinto de 0.

$$\lim_{x\to 0} \frac{x^3}{g(x)} = \lim_{x\to 0} x^3 \frac{9x^2+24x+16}{x^3} = \lim_{x\to 0} 9x^2+24x+16 = 16$$

apartado (4.3) cubica En este apartado tenemos que $f(x) = \frac{7x^3+56x^2+112x+64}{x^3+24x^2+80x+64}$ y, por lo tanto obtendremos

$$g(x) = \frac{1+x}{(f(x))^2} - 1 = \frac{x^7}{(x^3+24x^2+80x+64)^2}$$

Veamos que $g(x)$ es un infinitésimo equivalente a x^7 o, dicho en otras palabras, que $g(x) = O(x^7)$. Para demostrar esto es suficiente con ver que existe el límite de $x \to 0$ de $\frac{x^7}{g(x)}$ y que es finito y diferente de 0. Efectivamente, calculando el límite obtenemos

$$\lim_{x\to 0} \frac{x^7}{g(x)} = \lim_{x\to 0} x^7 \frac{(x^3+24x^2+80x+64)^2}{x^7} = \lim_{x\to 0} (x^3+24x^2+80x+64)^2 = 64^2 = 2^1 2$$

Por lo tanto $g(x) = O(x^7)$.

4.5. Comprobar que las $g(x)$ definidas a partir de las $f(x)$ encontradas en los apartados (4.1) y (4.3) són contractivas para $x > 0$.

Antes de empezar notemos que si ahora demostramos que las funciones g que se derivan de los apartados (4.1) y (4.3) són contractivas para todo $x > 0$ y tiene como punto fijo el origen. Entonces obtendremos que las sucesións $\{a_n\}$ definidas por recurrencia como $a_0 = x$ y $a_{n+1} = g(a_n)$ tendirán a 0 cuando n tienda a infinito. Por lo tanto, cuando $n \to \infty$, se cumplirá que $\sqrt{1+a_n}$ tendirá a 1. Dado que además hemos demostrado que por definición de g se cumple la identidad

$$\sqrt{1+x} = \left(\prod_{j=0}^{n} b_j\right)\sqrt{1+a_{n+1}}$$

entonces obtendremos que, cuando $n \to \infty$, el producto $\prod_{j=0}^{n} b_j$ tendirá a $\sqrt{1+x}$. Podremos por lo tanto calcular una aproximación de $\sqrt{1+x}$ (con $x > 0$) sencillamente realizando el producto $\prod_{j=0}^{n} b_j$

[14] Lo demostraremos más adelante para las funciones g encontradas en los apartados (4.1) y (4.3)

apartado (4.1) lineal Esta claro que 0 es un punto fijo de la función $g(x)$ ya que $g(0) = 0$. Ahora vamos a ver que la función g es contractiva para $x > 0$. Calculamos primero $g'(x)$ y obtenemos

$$g'(x) = \frac{3x^3 + 12x^2}{27x^3 + 108x^2 + 144x + 64}$$

Estudiamos el comportamiento de esta función.

$$\lim_{x\to\infty} g'(x) = \frac{1}{9}$$

Como la segunda derivada es estrictamente positiva para $x > 0$, podemos afirmar que $g'(x) < \lim_{x\to\infty} g'(x)$ para todo $x > 0$.Por otro lado, $g'(x) > g'(0) = 0$ para todo $x > 0$. Podemos así concluir que $0 < g'(x) < 1$ para todo $x > 0$.

apartado (4.3) cubica

Ver que 0 es punto fijo para g es trivial dado que, por definición de $g(x)$ tenemos que $g(0) = \frac{0}{64^2} = 0$

Veamos ahora que la función g es contractiva para todo $x > 0$.
Calculando $g'(x)$ y $g''(x)$ obtenemos que, para $x > 0$, ambas són estrictamente positivas.
Explicitamente tenemos que

$$g'(x) = x^6 \frac{x^3 + 72x^2 + 400x + 448}{(x^3 + 24x^2 + 80x + 64)^3}$$

Por lo tanto podemos afirmar que $\lim_{x\to\infty} g'(x) = 1$. Dado que la derivada segunda de x es Estrictamente positiva para $x > 0$ entonces podemos afirmar que la función $g'(x)$ es estrictamente creciente para $x > 0$. Podemos por lo tanto concluir que $g'(x) < \lim_{x\to\infty} g'(x) = 1$ para todo $x > 0$. Además, análogamente, tenemos que $g'(x) > g'(0) = 0$ para todo $x > 0$.
En resumen tenemos que, para todo $x > 0$, se cumple $0 < g'(x) < 1$.

Cojamos ahora $x_1 > x_2 > 0$ culesquiera. Aplicando el teorema del valor medio tendremos que

$$g(x_1) - g(x_2) = g'(c)(x_1 - x_2)$$

donde c es un valor desconocido entre x_2 y x_1 y, por lo tanto, mayor que 0. Poniendo módulos en la identidad y aplicando que $|g'(c)| < 1$ obtenemos

$$|g(x_1) - g(x_2)| = |g'(c)(x1 - x_2)| = |g'(c)||x1 - x_2| < |x1 - x_2|$$

Queda así demostrado que $g(x)$ es contractiva para todo $x > 0$ con el origen como punto fijo. Por lo tanto, podemos aplicar el razonamiento ya hecho para demostrar que el productorio $\prod_{j=0}^{n} b_j$ tiende a $\sqrt{1+x}$ cuando n tiende a infinito.

4.6. Comprobar tanto con las f y g encontradas en el apartado (4.1) como con las f y g encontradas en el apartado (4.3) se cumple la desigualdad

$$\left| \sqrt{1+x} - \prod_{j=0}^{k} b_j \right| \leq \frac{a_{k+1}}{2} \sqrt{1+x}$$

Antes de demostrar la desigualdad es aconsejable mostrar cuales són las cosas que se podrian deducir en el caso de que fuese cierta.
Si la desigualdad fuese cierta, dado que $\lim_{n\to\infty} a_n = 0$, entonces tendriamos que $\lim_{n\to\infty} \prod_{j=0}^{n} b_j = \sqrt{1+x}$ tal y como ya habiamos visto en el apartado (4.5).
Pero además podriamos confirmar lo que sospechabamos en el apartado (4.4) o sea que la velocidad de convergencia del producto es igual a la velocidad de convergencia de la sucesión $\{a_n\}$[15] la cual ya hemos visto en (4.4) que tiene un orden de convergencia 3 en el caso de que f sea quociente de dos lineales (la f encontrada en (4.1)) y 7 en el caso de que f sea quociente de dos cúbicas (la f encontrada en (4.3)).
Por último podriamos utilizar la inequación para poder imprimir por pantalla una cota del error que el programa CalculoRaiz comete al aproximar $\sqrt{1+x}$ por el producto $\prod_{j=0}^{n} b_j$

[15]de hecho dado que a_n multiplicada por una constante representa una cota superior del error obtendriamos que el producto converge almenos igual de rápido que esta sucesión pero no hemos demostrado que no pueda convergir aún más rápido.

Para demostrar la desigualdad tendremos que expandir por Taylor de orden 2 alrededor del orígen la función $\sqrt{1+a_{n+1}}$ que aparece en la identidad (2) de esta manera obtendremos

$$\sqrt{1+x} = \left(\prod_{j=0}^{n} b_j\right)\sqrt{1+a_{n+1}} = \left(\prod_{j=0}^{n} b_j\right)\left(1+\frac{a_{n+1}}{2}-\frac{c^2}{8}\right) \tag{3}$$

Donde c es un valor desconocido entre 0 y a_{n+1}.
Dado que $c^2 > 0$ y el producto es positivo en cuanto todos los b_j són positivos[16] podemos convertir la identidad (3) en la inequación

$$\sqrt{1+x} \leq \left(\prod_{j=0}^{n} b_j\right)\left(1+\frac{a_{n+1}}{2}\right) \tag{4}$$

Reescribiendo (4) obtenemos

$$\sqrt{1+x} - \prod_{j=0}^{n} b_j \leq \frac{a_{n+1}}{2}\prod_{j=0}^{n} b_j \tag{5}$$

Además el producto es creciente porque

$$\prod_{j=0}^{n+1} b_j = f(a_n)\prod_{j=0}^{n} b_j$$

y tenemos que

Apartado (4.1) lineal

$$f(x) = \frac{3x+4}{x+4} = 1+\frac{2x}{x+4} > 1 \text{ para todo } x > 0$$

Apartado (4.3) cúbica

$$f(x) = \frac{7x^3+56x^2+112x+64}{x^3+24x^2+80x+64} = 1+\frac{6x^3+32x^2+32x}{x^3+24x^2+80x+64} > 1 \text{ para todo } x > 0$$

Por lo tanto el producto es creciente en los dos casos y, dado que en los dos casos tiende a $\sqrt{1+x}$, podemos deducir que, para todo $n \in \mathbb{N}$ se cumple que

$$\prod_{j=0}^{n} b_j < \sqrt{1+x}$$

Combinando esto con la equación (5) podemos escribir

$$\sqrt{1+x} - \prod_{j=0}^{n} b_j \leq \frac{a_{n+1}}{2}\prod_{j=0}^{n} b_j < \frac{a_{n+1}}{2}\sqrt{1+x} \tag{6}$$

Dado que además, tal y como hemos dicho, $\sqrt{1+x} > \prod_{j=0}^{n} b_j$ entonces tendremos que

$$\left|\sqrt{1+x} - \prod_{j=0}^{n} b_j\right| = \sqrt{1+x} - \prod_{j=0}^{n} b_j$$

Combinando esto con (6) obtenemos finalmente que

$$\left|\sqrt{1+x} - \prod_{j=0}^{n} b_j\right| \leq \frac{a_{n+1}}{2}\sqrt{1+x} \tag{7}$$

tal y como queriamos demostrar.

[16] Esto se puede ver fácilmente a partir de la hipótesis $a_0 = x > 0$, el hecho que para las dos f y las dos g se cumple que $f(x)$,$g(x) > 0$ si $x > 0$ y la definición de $b_j = f(g^j(a_0))$

4.7. Para una elección de f como el cociente de dos polinomios de tercer grado ver que

$$|\sqrt{2} - b_0 b_1 b_2| < 5 \times 10^{-255}$$

Cogiendo $x = a_0 = 1$ y $n = 2$ y aplicando la desigualdad (7) tenemos que

$$\left| \sqrt{1+x} - \prod_{j=0}^{n} b_j \right| = |\sqrt{2} - b_0 b_1 b_2| \leq \frac{a_{n+1}}{2}\sqrt{1+x} = \frac{a_3}{2}\sqrt{2} = \frac{\sqrt{2}}{2} g(g(g(a_0))) = \frac{\sqrt{2}}{2} g(g(g(1))) \tag{8}$$

Si ahora recordamos que en el apartado (4.5) demostramos que la función g a la que nos referimos en esta fórmula es creciente y, por lo tanto, si $x_2 > x_1$ tendremos que $g(x_2) > g(x_1)$. Sabiendo esto, recordando la definición de $g(x) = \frac{x^7}{(x^3+24x^2+80x+64)^2}$ y notando que, para $x > 0$ se cumple que $g(x) < \frac{x^7}{64^2} = 2^{-12}x^7$ entonces podemos continuar modificando la desigualdad (8) como

$$|\sqrt{2} - b_0 b_1 b_2| < \frac{\sqrt{2}}{2} g(g(g(1))) = \frac{\sqrt{2}}{2} g\left(g\left(\frac{1}{169^2}\right)\right) = \frac{\sqrt{2}}{2} g(g(13^{-4})) < \frac{\sqrt{2}}{2} g(2^{-12}13^{-28}) < \frac{\sqrt{2}}{2} 2^{-13} 2^{-84} 13^{-196} =$$
$$= 2^{-97} 13^{-196} \sqrt{2} < 10^{-247} 10^{-0,52} \sqrt{2} < 0{,}5 \cdot 10^{-247} = 5 \cdot 10^{-248}$$

[17] No hemos logrado llegar a demostrar la desigualdad que se pedia pero nos hemos acercado bastante. Hemos realizado un programa en C que calculase $\frac{\sqrt{2}}{2} g(g(g(1)))$ con una precisión de __float128[18] para ver si conseguiamos obtener una mejor cota del error. No obstante el programa no mejora la cota que ya teniamos dado que nos devuelve un resultado comprendido entre 5×10^{-248} y 5×10^{-249}. Aun así podemos considerar que 5×10^{-248} es una cota del error sorprendente considerando que solo se necesitan 3 iterados para conseguirla.[19]

[17] Aplicamos que

$$13^{-196} = 10^{\log_{10}(13^{-196})} = 10^{-196 \log_{10}(13)} < 10^{-218} 10^{-0,33}$$

y que

$$2^{-97} = 10^{\log(2^{-97})} = 10^{-97 \log(2)} < 10^{-29} 10^{-0,19}$$

.

[18] Se puede encontrar el programa en la carpeta CODIGOS bajo el nombre de Problema4.7

[19] El resultado obtenido haciendo la misma operación con los programas CalculoRaiz y Problema4 no da el mismo resultado debito a errores de calculo en punto flotante.