Análisis Numérico

ISBN: 978-956-306-072-0
Registro de Propiedad Intelectual: 200.529
Colección: Herramientas para la formación de profesores de matemáticas.
Diseño: Jessica Jure de la Cerda.
Diseño de Ilustraciones: Cristina Felmer Plominsky, Catalina Frávega Thomas.
Diagramación: Pedro Montealegre Barba, Francisco Santibáñez Palma.
Financiamiento: Proyecto Fondef D05I-10211.

Datos de contacto para la adquisición de los libros:
Para Chile:
1. En librerías para clientes directos.
2. Instituciones privadas directamente con:
   Juan Carlos Sáez C.
   Director Gerente
   Comunicaciones Noreste Ltda.
   J.C. Sáez Editor
   jcsaezc@vtr.net
   www.jcsaezeditor.blogspot.com
   Oficina: (56 2) 3260104 - (56 2) 3253148
3. Instituciones públicas o fiscales: www.chilecompra.cl

Desde el extranjero:
1. Liberalia Ediciones: www.liberalia.cl
2. Librería Antártica: www.antartica.cl
3. Argentina: Ediciones Manantial: www.emanantial.com.ar
4. Colombia: Editorial Siglo del Hombre
   Fono: (571) 3377700
5. España: Tarahumara, tarahumara@tarahumaralibros.com
   Fono: (34 91) 3656221
6. México: Alejandría Distribución Bibliográfica, alejandria@alejandrialibros.com.mx
   Fono: (52 5) 556161319 - (52 5) 6167509
7. Perú: Librería La Familia, Avenida República de Chile # 661
8. Uruguay: Dolmen Ediciones del Uruguay
   Fono: 00-598-2-7124857

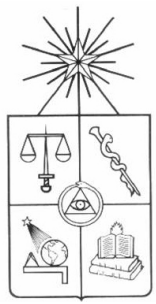
Análisis Numérico | Axel Osses A.
Departamento de Ingeniería Matemática, Universidad de Chile
axosses@dim.uchile.cl

ESTA PRIMERA EDICIÓN DE 2.000 EJEMPLARES
Se terminó de imprimir en febrero de 2011 en WORLDCOLOR CHILE S.A.

ANÁLISIS NUMÉRICO

Axel Osses A.

Universidad de Chile



Dirección del Proyecto Fondef D05I-10211
Herramientas para la Formación de Profesores de Matemática

Patricio Felmer, Director del Proyecto
Universidad de Chile.

Leonor Varas, Directora Adjunta del Proyecto
Universidad de Chile.

Salomé Martínez, Subdirectora de Monografías
Universidad de Chile.

Cristián Reyes, Subdirector de Estudio de Casos
Universidad de Chile.

Presentación de la Colección

La colección de monografías que presentamos es el resultado del generoso esfuerzo de los autores, quienes han dedicado su tiempo y conocimiento a la tarea de escribir un texto de matemática. Pero este esfuerzo y generosidad no se encuentra plenamente representado en esta labor, sino que también en la enorme capacidad de aprendizaje que debieron mostrar, para entender y comprender las motivaciones y necesidades de los lectores: Futuros profesores de matemática.

Los autores, encantados una y otra vez por la matemática, sus abstracciones y aplicaciones, enfrentaron la tarea de buscar la mejor manera de traspasar ese encanto a un futuro profesor de matemática. Éste también se encanta y vibra con la matemática, pero además se apasiona con la posibilidad de explicarla, enseñarla y entregarla a los jóvenes estudiantes secundarios. Si la tarea parecía fácil en un comienzo, esta segunda dimensión puso al autor, matemático de profesión, un tremendo desafío. Tuvo que salir de su oficina a escuchar a los estudiantes de pedagogía, a los profesores, a los formadores de profesores y a sus pares. Tuvo que recibir críticas, someterse a la opinión de otros y reescribir una y otra vez su texto. Capítulos enteros resultaban inadecuados, el orden de los contenidos y de los ejemplos era inapropiado, se hacía necesario escribir una nueva versión y otra más. Conversaron con otros autores, escucharon sus opiniones, sostuvieron reuniones con los editores. Escuchar a los estudiantes de pedagogía significó, en muchos casos, realizar eventos de acercamiento, desarrollar cursos en base a la monografía, o formar parte de cursos ya establecidos. Es así que estas monografías recogen la experiencia de los autores y del equipo del proyecto, y también de formadores de profesores y estudiantes de pedagogía. Ellas son el fruto de un esfuerzo consciente y deliberado de acercamiento, de apertura de caminos, de despliegue de puentes entre mundos, muchas veces, separados por falta de comunicación y cuya unión es vital para el progreso de nuestra educación.

La colección de monografías que presentamos comprende una porción importante de los temas que usualmente encontramos en los currículos de formación de profesores de matemática de enseñanza media, pero en ningún caso pretende ser exhaustiva. Del mismo modo, se incorporan temas que sugieren nuevas formas de abordar los contenidos, con énfasis en una matemática más pertinente para el futuro profesor, la que difiere en su enfoque de la matemática para un ingeniero o para un licenciado en matemática, por ejemplo. El formato de monografía, que aborda temas específicos

con extensión moderada, les da flexibilidad para que sean usadas de muy diversas maneras, ya sea como texto de un curso, material complementario, documento básico de un seminario, tema de memoria y también como lectura personal. Su utilidad ciertamente va más allá de las aulas universitarias, pues esta colección puede convertirse en la base de una biblioteca personal del futuro profesor o profesora, puede ser usada como material de consulta por profesores en ejercicio y como texto en cursos de especialización y post-títulos. Esta colección de monografías puede ser usada en concepciones curriculares muy distintas. Es, en suma, una herramienta nueva y valiosa, que a partir de ahora estará a disposición de estudiantes de pedagogía en matemática, formadores de profesores y profesores en ejercicio.

El momento en que esta colección de monografías fue concebida, hace cuatro años, no es casual. Nuestro interés por la creación de herramientas que contribuyan a la formación de profesores de matemática coincide con un acercamiento entre matemáticos y formadores de profesores que ha estado ocurriendo en Chile y en otros lugares del mundo. Nuestra motivación nace a partir de una creciente preocupación en todos los niveles de la sociedad, que ha ido abriendo paso a una demanda social y a un interés nacional por la calidad de la educación, expresada de muy diversas formas. Esta preocupación y nuestro interés encontró eco inmediato en un grupo de matemáticos, inicialmente de la Universidad de Chile, pero que muy rápidamente fue involucrando a matemáticos de la Pontificia Universidad Católica de Chile, de la Universidad de Concepción, de la Universidad Andrés Bello, de la Universidad Federico Santa María, de la Universidad Adolfo Ibáñez, de la Universidad de La Serena y también de la Universidad de la República de Uruguay y de la Universidad de Colorado de Estados Unidos.

La matemática ha adquirido un rol central en la sociedad actual, siendo un pilar fundamental que sustenta el desarrollo en sus diversas expresiones. Constituye el cimiento creciente de todas las disciplinas científicas, de sus aplicaciones en la tecnología y es clave en las habilidades básicas para la vida. Es así que la matemática actualmente se encuentra en el corazón del currículo escolar en el mundo y en particular en Chile. No es posible que un país que pretenda lograr un desarrollo que involucre a toda la sociedad, descuide el cultivo de la matemática o la formación de quienes tienen la misión de traspasar de generación en generación los conocimientos que la sociedad ha acumulado a lo largo de su historia.

Nuestro país vive cambios importantes en educación. Se ha llegado a la convicción que la formación de profesores es la base que nos permitirá generar los cambios cualitativos en calidad que nuestra sociedad ha impuesto. Conscientes de que la tarea formativa de los profesores de matemática y de las futuras generaciones de jóvenes es extremadamente compleja, debido a que confluyen un sinnúmero de factores y disciplinas, a través de esta colección de monografías, sus editores, autores y todos los que han participado del proyecto en cada una de sus etapas, contribuyen a esta tarea, poniendo a disposición una herramienta adicional que ahora debe tomar vida propia en los formadores, estudiantes, futuros profesores y jóvenes de nuestro país.

Patricio Felmer y Salomé Martínez
Editores

Agradecimientos

Agradecemos a todos quienes han hecho posible la realización de este proyecto Fondef: “Herramientas para la formación de Profesores de Matemáticas". A Cristián Cox, quien apoyó con decisión la idea original y contribuyó de manera crucial para obtener la participación del Ministerio de Educación como institución asociada. Agradecemos a Carlos Eugenio Beca por su apoyo durante toda la realización del proyecto. A Rafael Correa, Edgar Kausel y Juan Carlos Sáez, miembros del Comité Directivo. Agradecemos a Rafael Benguria, Servet Martínez y Fidel Oteiza, miembros del Comité Editorial de la colección, quienes realizaron valiosos aportes a los textos. A José Sánchez, entonces Decano de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción y a Guillermo Marshall, quién fuera Decano de la Facultad de Matemáticas de la Pontificia Universidad Católica de Chile. A ambos agradecemos por su decisiva contribución para lograr la integridad de la colección de 15 monografías. Agradecemos a Víctor Campos, Ejecutivo de Proyectos de Fondef, por su colaboración y ayuda en las distintas etapas del proyecto.

En este volumen manisfestamos nuestro especial agradecimiento a Jaime San Martín, director del Centro de Modelamiento Matemático de la Universidad de Chile, por su constante apoyo durante toda la realización de este proyecto. Más aun, su apoyo decidido y generoso que permitió que esta monografía sea parte de la colección. También queremos reconocer su valioso aporte a la educación manifiestado desde la dirección del Centro de Modelamiento Matemático, el cual ha permitido un fuerte impulso al involucramiento de matemáticos activos en esta importante tarea.

Agradecemos también a Bárbara Ossandón de la Universidad de Santiago, a Jorge Ávila de la Universidad Católica Silva Henríquez, a Víctor Díaz de la Universidad de Magallanes, a Patricio Canelo de la Universidad de Playa Ancha en San Felipe y a Osvaldo Venegas y Silvia Vidal de la Universidad Católica de Temuco, quienes hicieron posible las visitas que realizamos a las carreras de pedagogía en matemática. Agradecemos a todos los evaluadores, alumnos, académicos y profesores -cuyos nombres no incluimos por ser más de una centena- quienes entregaron sugerencias, críticas y comentarios a los autores, que ayudaron a enriquecer cada uno de los textos.

Agradecemos a Marcela Lizana por su impecable aporte en todas las labores administrativas del proyecto, a Aldo Muzio por su colaboración en la etapa de evaluación, y también a Anyel Alfaro por sus contribuciones en la etapa final del proyecto y en la difusión de los logros alcanzados.

Dirección del Proyecto

Índice General

A mis padres. A Marianela,
Alexandra y Maximiliano.

Prefacio

Esta monografía es un curso de Análisis Numérico orientado a docentes y estudiantes de pedagogía en matemáticas, física o ciencias. El texto puede ser usado en un curso semestral para presentar las herramientas básicas del Análisis Numérico, con una visión contemporánea centrada en el concepto de algoritmo aproximante, sin perder de vista, por un lado, su conexión con la historia del desarrollo de las matemáticas o la física y, por otro, su impacto en la sociedad actual. Después del curso, el alumno debería ser capaz de entender y formular soluciones aproximantes a problemas simples de la matemática del continuo y al mismo tiempo atesorar la importancia que tiene hoy este proceso científico-tecnológico en el desarrollo de nuestro conocimiento y de nuestra forma de vivir.

El objetivo principal del Análisis Numérico es estudiar *cómo aproximar los problemas del continuo que aparecen en fenómenos físicos y vivientes*. Esto se ve concretizado en el diseño y análisis matemático de algoritmos aproximantes y en la aplicación de dichos algoritmos para simular los fenómenos de la naturaleza utilizando máquinas computadoras. Calcular con un computador para modelar problemas reales no es una tarea fácil y hay que entender muy bien las posibilidades y las limitaciones del uso del cálculo aproximado y de los algoritmos utilizados. Y esto vale desde aproximar un simple número o calcular una raíz, hasta resolver sistemas lineales de gran tamaño y complejas ecuaciones diferenciales.

El primer capítulo comienza con una discusión sobre errores y redondeo. Esto es por una razón práctica, ya que muchos de estos conceptos se utilizan después como herramientas para interpretar el resultado de los cálculos, sin embargo, esto *no significa que el redondeo y el error sea el tema central del Análisis Numérico*. Ésta es una percepción errónea muy común, incluso entre los propios matemáticos. Muchos estudiantes se desmotivan, pues el análisis de errores tiende a ser muy técnico. Es por esto que en este texto se introducen los conceptos de propagación de errores, cifras significativas, errores absoluto y relativo, truncatura y redondeo, deliberadamente contextualizados en temas de interés como son el efecto mariposa, la corrección relativista, el interés bancario y el cálculo de promedios entre otros ejemplos. Además, en este capítulo se provee la solución de la mayoría de los ejercicios propuestos.

A partir del segundo capítulo nos adentramos pues en el verdadero objeto del Análisis Numérico que es, como mencionamos antes, al análisis de los algoritmos aproximantes del continuo. Partimos con las aproximaciones del número π ¿Por qué π y no el número e ó $\sqrt{2}$? La razón es que la definición de π es tan simple como la razón de la longitud de una circunferencia con su diámetro. Además, esta constante aparece por doquier en las matemáticas y las ciencias, dado que está profundamente relacionada con la medición y la periodicidad. Pero el desarrollo decimal de π es infinito, y aproximar su valor ha sido un verdadero desafío al intelecto humano desde la antiguedad hasta nuestros días, desde el algoritmo de duplicación de Arquímides hasta el llamado algoritmo BBP. Esta gran historia tiene una justa mezcla entre teoría y práctica que fascina. No podía estar ausente en este texto, ya que, aparte de su interés histórico, entrega al profesor de matemáticas o ciencias actual un excelente ejemplo de cómo conectar los conceptos y métodos matemáticos con su uso práctico. En efecto, el problema de aproximar π sirve como una introducción a la idea central de algoritmos aproximantes. Al mismo tiempo, aparecen de manera intuitiva y natural la necesidad de usar herramientas de manejo de errores y de representación de números en distintas bases con números bien concretos, lo que educa la familiaridad con los números. Finalmente, permite dar los primeros pasos en la implementación práctica en un computador de simples algoritmos para aproximar π.

En el tercer capítulo se revisa lo que se sería el ABC del Análisis Numérico clásico: los métodos iterativos para aproximar los ceros de una función, como las iteraciones de punto fijo o el método de Newton-Raphson; la interpolación y aproximación de funciones por polinomios, sean de Taylor, Lagrange o Newton, y, finalmente, las siempre útiles fórmulas de cuadratura para aproximar integrales, incluyendo la fórmula de los trapecios y la de Simpson. En este capítulo son especialmente importantes los ejemplos y los ejercicios que permiten ir comprendiendo y relacionando mejor los diferentes algoritmos que van surgiendo, muy particularmente, el estudio de la aproximación de la raíz cuadrada de un número y la aproximación de áreas de figuras geométricas.

En el cuarto capítulo presentamos uno de los grandes logros del Análisis Numérico que es el estudio y resolución eficiente de los sistemas lineales. Dada la extensión y propósito del presente texto, no hacemos una revisión exhausiva de los métodos para resolver sistemas lineales, sino que hacemos especial énfasis en poder referirnos a tres casos muy frecuentes en la práctica: el estudio de los sistemas lineales de gran tamaño, de los sistemas lineales mal condicionados y de los sistemas lineales sobredeterminados. Hoy en día esto tiene gran utilidad, ya que muchos problemas aplicados se reducen a sistemas lineales con alguna de estas características y todos ellos presentan dificultades para resolverse con precisión en un computador. Al saber cómo se realiza una tomografía computarizada se descubre la importancia de este tipo de sistemas lineales en el mundo contemporáneo.

Finalmente, en el último capítulo revisamos brevemente la resolución numérica de ecuaciones diferenciales. Este capítulo es optativo si se hace un curso semestral, dependiendo de la orientación del alumno. Utilizando las técnicas de integración del segundo capítulo, se presentan desde el método de Euler al de Runge-Kutta, principalmente de manera algorítmica, sin entrar demasiado en detalles para justificar los órdenes de convergencia. Se explica la diferencia entre métodos explícitos e implícitos y se discute la noción de estabilidad numérica. Los algoritmos se presentan progresivamente a través de ejemplos relacionados con la dinámica de crecimiento de poblaciones, epidemias y el estudio de recursos naturales.

A lo largo de todo el texto se han utilizado planillas de cálculo y programas computacionales para implementar e ilustrar los diferentes algoritmos, construyendo tablas y figuras. Cada vez que se hace uso de una de dichas planillas o programas, esto se indica con este símbolo al margen del texto, lo que invita al lector o al profesor  a consultar el anexo donde se resumen y describen dichas planillas o programas. El uso de planillas de cálculo se pensó desde un comienzo, principalmente en el segundo capítulo, dado que es una herramienta a la que tiene fácil acceso un alumno o profesor y cuyo uso desarrolla el pensamiento algorítmico y la familiaridad con los números. El desarrollo de programas se utiliza desde el tercer capítulo, en el que las planillas resultan ya insuficientes o engorrosas de usar, pero la programación se realiza en el lenguaje intuitivo de la computación científica. Este conjunto de planillas de cálculo y programas resumidas en el anexo, y que se pueden encontar el la página web del autor o de esta colección, constituyen un complemento fundamental del presente texto.

Quisiera finalmente agradecer a todos aquellos investigadores, profesores y alumnos quienes han revisado y evaluado los manuscritos originales de esta monografía, o me han hecho comentarios, entregando valiosos consejos y permitiendo mejorar el material inicial. Quería mencionar, en particular, aunque puedo olvidar algunos nombres, a mis colegas de la Universidad de Chile, Patricio Felmer, Salomé Martínez, María Leonor Varas y Jaime Ortega, a Raúl Navarro (UMCE), Jorge Wevar Negrier (U. de Los Lagos), Richard Lagos (U. de Magallanes), Rafael Benguria (U. Católica), Barbara Ossandón (USACH), Omar Gil (U. de la República, Uruguay), Pascal Frey (U. Paris 6) y especialmente a los alumnos de la carrera de Pedagogía en Matemáticas de la Universidad de Magallanes por sus interesantes opiniones.

Axel Osses
Santiago, 31 de enero de 2011

Capítulo 1: Propagación de errores y redondeo

> *"Cayó al suelo una cosa exquisita, una cosa pequeña que podía destruir todos los equilibrios, derribando primero la línea de un pequeño dominó, y luego de un gran dominó, y luego de un gigantesco dominó, a lo largo de los años."* Ray Bradbury *(El sonido del trueno, 1952)*

En 1952 el escritor estadounidense Ray Bradbury escribió un interesante relato de ciencia ficción. Se trata de un grupo de exploradores que viajan al pasado prehistórico de la Tierra. Por accidente, al salirse del sendero de seguridad, uno de ellos pisa una pequeña mariposa. Aparentemente esto no tiene la menor importancia, pero, al volver al presente, miles de años después, los viajeros advierten ligeros cambios en la composición del aire que se respira y no sólo eso, sino que además se dan cuenta que ¡ha cambiado el resultado de la última elección presidencial! En efecto, aquella insignificante mariposa desaparecida había cambiado el devenir. Hoy, este concepto de extrema sensibilidad del devenir de un sistema físico ante un pequeño error en las condiciones iniciales, se conoce popularmente como el *efecto mariposa*. [1] [2]

Miremos el efecto de una mariposa desaparecida como un pequeño error o diferencia, que va afectando los hechos futuros en una escala cada vez más grande. El vuelo de esa mariposa habría creado una ligera brisa que se habría amplificado en un viento considerable hasta ser un tornado y esto habría tenido a la larga consecuencias importantes en el clima futuro.[3]

La probabilidad de que esto ocurra parece en realidad muy pequeña. El sentido común nos dice que el clima futuro no puede depender de una simple mariposa. De hecho, todos estamos concientes de que habrá cambios climáticos en el futuro y calentamiento global y eso no cambiará simplemente por una mariposa más o menos. Entonces, el efecto mariposa ¿es sólo ciencia-ficción?

Antes de seguir discutiendo la validez del efecto mariposa, aceptemos al menos que la historia de Bradbury ilustra lo importante que podría llegar a ser un pequeño error si se *propaga* y *amplifica* por mucho tiempo.

---

[1] El término fue popularizado gracias a los trabajos del meteorólogo Edward Lorenz sobre la teoría del caos, cf. **[28]**, **[9]**.

[2] Isaac Asimov utilizó también la idea del efecto mariposa en su novela de ciencia-ficción "El Fin de la Eternidad" publicada en 1955. Se trata de ejecutores encargados de alterar sutilmente el curso del tiempo para proteger a la humanidad del sufrimiento.

[3] Hay un probervio chino que dice: "el aleteo de las alas de una mariposa se puede sentir al otro lado del mundo". En términos modernos, se suele traducir como "el aleteo de una mariposa en Hong Kong puede desatar una tormenta en Nueva York".

1.1 Propagación de errores

Analicemos más de cerca este fenómeno de la *propagación de errores*. Para ello, consideremos un cálculo muy simple. Supongamos que queremos encontrar el entero más cercano a la expresión:

$$100\sqrt{2}\sqrt{3}.$$

El problema es que no conocemos las raíces exactamente. Así es que decidimos utilizar valores aproximados de $\sqrt{2}$ y $\sqrt{3}$ con dos decimales: [4]

$$\sqrt{2} \approx 1,41 \qquad \sqrt{3} \approx 1,73.$$

El cálculo aproximado (indicado aquí y en lo que sigue con el signo $\approx$) nos entrega el resultado:

$$100\sqrt{2}\sqrt{3} \approx 100 \times 1,41 \times 1,73 = 243,93.$$

Es decir, el entero más cercano sería 244. Sin embargo, un cálculo más preciso con una calculadora nos entrega el resultado:

$$100\sqrt{2}\sqrt{3} \approx 244,94897\ldots$$

y el entero más cercano resulta ser en realidad 245 y no 244.

Veremos que este tipo de diferencias en el resultado esperado se deben justamente a la *propagación* de errores de redondeo y que son especialmente notorios cuando se hacen multiplicaciones o divisiones. Para ello, tratemos de expresar explícitamente los errores a través de símbolos. Llamemos Δx al error que se comete en la aproximación de $\sqrt{2}$ con dos decimales, esto es:

$$\sqrt{2} = 1,41 + \Delta x$$

(note que el símbolo que corresponde ahora es $=$). Del mismo modo, llamemos Δy al error que se comete al aproximar $\sqrt{3}$ con dos decimales:

$$\sqrt{3} = 1,73 + \Delta y.$$

Los errores Δx, Δy no los conocemos exactamente, de lo contrario, podríamos conocer también las raíces exactamente. Una buena idea entonces es *estimarlos* o, lo que es lo mismo, acotarlos para saber qué tan grandes pueden ser. Por ejemplo, al aproximar $\sqrt{2} \approx 1,41$ con *dos* decimales estamos diciendo que la aproximación con *tres* decimales puede ser una de las siguientes:

$$1,405;\ 1,406;\ 1,407;\ 1,408;\ 1,409;\ 1,410;\ 1,411;\ 1,412;\ 1,413;\ 1,414$$

o, más precisamente, algún valor en el intervalo comprendido entre $1,405$ (incluido) y $1,415$ (no incluido), lo cual se expresa comúnmente así:[5]

$$\sqrt{2} = 1,41 \pm 0,005$$

---

[4] Esto se puede hacer por *encajonamientos sucesivos*, por ejemplo. Ver Ejercicio 1.1 más adelante. Existen otros métodos más eficientes, ver la sección sobre aproximación de raíces cuadradas del Capítulo 3.

[5] Ver también la sección sobre precisión y cifras significativas más adelante.

y quiere decir que, en el peor de los casos, el error al aproximar $\sqrt{2}$ con dos decimales se puede acotar por

$$|\Delta x| \leq 0,005 = 5 \times 10^{-3}.$$

Del mismo modo, el error al aproximar $\sqrt{3}$ con dos decimales se puede acotar por

$$|\Delta y| \leq 0,005 = 5 \times 10^{-3}.$$

Utilizando esto último, realicemos ahora con cuidado la operación de multiplicar las dos raíces, considerando explícitamente como se *propagan* los errores :

$$\begin{aligned} \underbrace{\sqrt{2}\sqrt{3}}_{\text{valor exacto}} &= (1,41 + \Delta x) \times (1,73 + \Delta y) \\ &= \underbrace{1,41 \times 1,73}_{\text{valor aproximado}} + \underbrace{1,73 \times \Delta x + 1,41 \times \Delta y + \Delta x \Delta y.}_{\text{error propagado}} \end{aligned}$$

Para estimar el error, tomemos valor absoluto de la diferencia entre el valor exacto y el aproximado:

$$\begin{aligned} |\sqrt{2}\sqrt{3} - 1,41 \times 1,73| &= |1,73 \times \Delta x + 1,41 \times \Delta y + \Delta x \Delta y| \\ &\leq 1,73 \times |\Delta x| + 1,41 \times |\Delta y| + |\Delta x||\Delta y| \\ &= 1,73 \times 0,005 + 1,41 \times 0,005 + 0,005 \times 0,005 \\ &= 0,015725. \end{aligned}$$

Esto es, el error propagado está acotado por $1,5725 \times 10^{-2}$. Notar que el término de segundo orden $\Delta x \Delta y$ es muy pequeño comparado con los términos de primer orden proporcionales a Δx ó Δy. Por esta razón es que, en la propagación de errores, son los términos de primer orden los que más interesan (ver Cuadro 1.1 más adelante) y los de orden superior usualmente se desprecian.

Finalmente, multipliquemos por 100 y notemos que el error también se amplifica por 100

$$\begin{aligned} |\underbrace{100\sqrt{2}\sqrt{3}}_{\text{valor exacto}} - \underbrace{100 \times 1,41 \times 1,73}_{\text{valor aproximado}}| &= 100 \times |\sqrt{2}\sqrt{3} - 1,41 \times 1,73| \\ &\leq 100 \times 0,015725 \\ &= 1,5725. \end{aligned}$$

El resultado final indica que la diferencia entre el valor exacto y el aproximado puede llegar a ser ¡mayor que un entero!, aunque las raíces estuvieran correctamente aproximadas hasta la centésima. Y esto fue una buena estimación, pues la diferencia real entre el valor exacto y el aproximado, calculada con una calculadora de bolsillo, es de $1,01897\ldots$

✍ **Ejercicio 1.1.** Para aproximar $\sqrt{2}$ utilizamos la propiedad de que si $0 < x^2 < y^2$ entonces $0 < x < y$. En efecto, $\sqrt{2}$ debe estar entre 1 y 2 pues $1^2 < 2 < 2^2$. Luego probamos con $\frac{1+2}{2} = 1,5$ observando que $1,5^2 = 2,25 > 2$, y entonces $1 < \sqrt{2} < 1,5$.

Luego seguimos con $\frac{1+1,5}{2} = 1,25$ viendo que $1,25^2 = 1,5625 < 2$ y luego $1,25 < \sqrt{2} < 1,5$ y así sucesivamente. Utilice este algoritmo de *encajonamientos sucesivos* para aproximar $\sqrt{2}$ con tres decimales.[6]

Notemos un hecho muy importante en la propagación de errores: que al multiplicar, los errores se amplifican por los factores involucrados en el producto. Por ejemplo, volviendo al efecto mariposa, se sabe que en los fenómenos atmosféricos hay muchos efectos multiplicativos. El siguiente ejercicio ilustra este hecho.[7]

✍ **Ejercicio 1.2.** Al formarse una *celda convectiva* en la atmósfera, un rollo de aire que gira a medida que el aire caliente sube y el aire frío baja, el sentido de su giro X se puede representar por un valor comprendido entre 0 y 1, de modo que si $X > 0{,}5$ el giro se produce el sentido de los punteros del reloj y si $X < 0{,}5$ el giro se produce en sentido contrario de los punteros del reloj. Suponga que si X_n representa el valor del giro durante la hora n, entonces el valor del giro X_{n+1} en la próxima hora $n+1$ está dado por la siguiente regla multiplicativa:[8]

$$X_{n+1} = 3{,}9 \times X_n(1 - X_n).$$

Si $X_0 = 0{,}5$ es el valor observado inicialmente (sin giro), calcule el valor de X_{20} luego de 20 horas e indique a qué sentido de giro corresponde. Tome ahora un valor inicial de $X_0 = 0{,}501$ y realice de nuevo el cálculo anterior. Verifique que el giro final es exactamente ¡opuesto al anterior! Para facilitar los cálculos se puede usar una ❐✍ calculadora o una planilla de cálculo cómo se indica en la Figura 1.1.

Hay algo más que usted debe saber sobre la propagación de errores para tener un panorama más completo del asunto. En el caso de $\sqrt{2}$ y $\sqrt{3}$, las aproximaciones a tres cifras significativas y los valores más exactos son, respectivamente:

$$\sqrt{2} \approx 1,41, \qquad \sqrt{2} = 1,414421356\ldots$$
$$\sqrt{3} \approx 1,73, \qquad \sqrt{3} = 1,732050807\ldots$$

es decir, en ambos casos, la aproximación es *menor* que el valor exacto. En este caso se dice que son aproximaciones *por defecto* . En cambio, al aproximar $\sqrt{7}$ con tres cifras significativas y compararla con el valor más exacto:

$$\sqrt{7} \approx 2,65, \qquad \sqrt{7} = 2,645751311\ldots$$

vemos que la aproximación es *mayor* que el valor exacto. En este caso se dice que la aproximación es *por exceso* . Entonces, al escribir

$$\sqrt{2} = 1,41 + \Delta x \qquad \text{ó} \qquad \sqrt{3} = 1,73 + \Delta y$$

---

[6] Véase el Capítulo 3 para más métodos de aproximación de la raíz cuadrada.

[7] Y reproduce uno de los momentos científicos más importantes del siglo XX, que fue justamente una discrepancia de cálculo de este tipo observada por E. Lorenz, cf. [**28**].

[8] Véase el Capítulo 5 para más aspectos de esta regla llamada *modelo logístico discreto*.

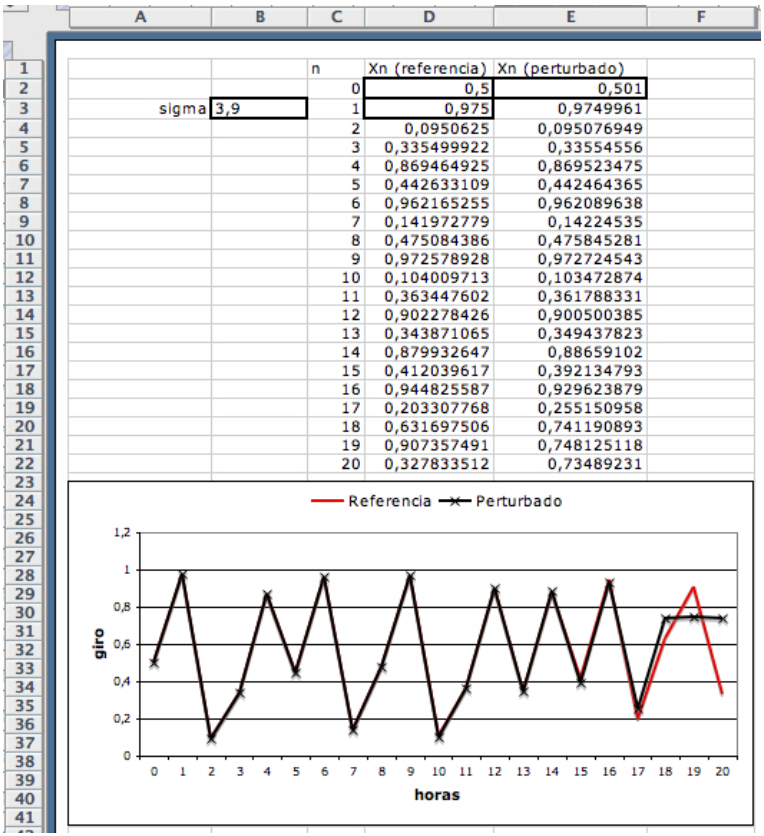


FIGURA 1.1. Planilla de cálculo para el Ejercicio 1.2 que ilustra el efecto mariposa. Primero se ingresan en las celdas `D2`, `E2` y `B3` los valores iniciales de referencia (0,5) y perturbado (0,501) de la variable X y el valor de la constante multiplicativa $\sigma = 3{,}9$. Se ingresa la fórmula en la celda `D3=$B$3*D2*(1-D2)` la que simplemente se copia en `E3` y en todas las demás celdas de las columnas `D` y `E` bajo ellas. Al principio los cálculos coinciden aproximadamente pero luego de $n = 17$ divergen claramente uno del otro.

tendremos que Δx, Δy son cantidades positivas (o sea son aproximaciones por defecto), en cambio cuando escribimos

$$\sqrt{7} = 2,65 + \Delta z$$

tendremos que Δz es negativo (aproximación por exceso).

Una vez dicho esto, veamos el efecto que tiene el signo de estos errores al multiplicar:

$$\sqrt{2}\sqrt{3} = (1,41 + \Delta x)(1,73 + \Delta y) = 1,41 \times 1,73 + 1,73 \times \Delta x + 1,41 \times \Delta y + \Delta x \Delta y$$

$$\sqrt{2}\sqrt{7} = (1,41 + \Delta x)(2,65 + \Delta z) = 1,41 \times 2,65 + 2,65 \times \Delta x + 1,41 \times \Delta z + \Delta x \Delta z.$$

En el primer caso, todos los términos de error tienen igual signo y en este caso los errores se suman o se *acumulan*. Se habla es este caso de acumulación de errores. En el segundo caso, algunos términos de error son positivos, mientras que otros son negativos (recuerde que $\Delta z < 0$), de modo que algunos errores se restan o se *cancelan*.

Por ejemplo, si modificamos un poco el ejemplo anterior y buscamos ahora el entero más cercano a:

$$100\sqrt{2}\sqrt{7}$$

aproximando, obtenemos

$$100\sqrt{2}\sqrt{7} \approx 100 \times 1,41 \times 2,65 = 373,64 \approx 374$$

que en este caso ¡sí es la respuesta correcta! lo que se verifica fácilmente con una calculadora de bolsillo:

$$100\sqrt{2}\sqrt{7} = 374,16757\ldots \approx 374.$$

En este caso hubo cancelación de errores debido a que se aproximó $\sqrt{2}$ por defecto y $\sqrt{7}$ por exceso y un error compensó al otro. Esto nos lleva a la siguiente reflexión.

1.2 La balanza del error y el dado cargado

Pensemos ahora en una balanza en que ponemos de un lado el valor exacto y del otro el valor aproximado. Un equilibrio perfecto de la balanza sería un cálculo perfecto, pero al aproximar los errores van inclinando la balanza a uno u otro lado. Un error por exceso inclinaría la balanza para el lado del valor aproximado, mientras que uno por defecto para el lado del valor exacto. A medida que calculamos a veces los errores nos alejan del resultado e inclinan más y más la balanza y otras veces nos acercan al equilibrio según cómo se va produciendo acumulación o cancelación de los errores.

Entonces, aparece un hecho que parece anular el efecto mariposa. Si se producen con igual frecuencia las acumulaciones y cancelaciones de errores, por ejemplo, sobreviven o desaparecen algunas mariposas, pero en promedio su número no cambia, entonces la balanza estará siempre cerca del equilibrio y no habrá un efecto notorio en el futuro. Por lo tanto, para hacerlo más creíble, deberíamos modificar el relato de Bradbury como sigue: los viajeros debieron hacer desaparecer sistemáticamente muchas mariposas, por ejemplo, pensemos que en cada viaje capturaban miles de ellas. Entonces, sí resulta probable que este pequeño error pudiera amplificarse en el tiempo. Es como el efecto de un dado cargado.

Como vemos, un pequeño error, sobretodo si este persiste sistemáticamente y con un cierto signo constante, podría perfectamente propagarse y, en un proceso multiplicativo, amplificarse cada vez más hasta producir cambios enormes en el proceso que

se considere. Para reafirmar que esto no es pura ciencia-ficción, piense que las mayores emisiones antropogénicas de gases con efecto invernadero, que han persistido desde la revolución industrial de principios del siglo XX, están provocando paulatinamente cambios más y más grandes en el clima del planeta. Se estima que incluso, aunque estas emisiones no aumenten o desaparezcan completamente, el cambio resulta ya irreversible.

✍ **Ejercicio 1.3.** Coloque cinco piezas de dominó en posición vertical y en fila, sobre un papel blanco que pueda rayar. Marque muy bien con un lápiz la posición inicial de cada una de las piezas. Haga caer la primera y dibuje con una lápiz la posición donde cae la última. Repita el experimento tratando de comenzar exactamente de la misma configuración inicial y de hacer caer la primera pieza de dominó siempre de la misma forma. Vaya registrando sobre el papel las distintas posiciones dónde cae la última pieza. Observe e intente asociar este experimento con las nociones de propagación de errores. Piense qué pasaría si se aumenta el número de piezas.

✍ **Ejercicio 1.4.** El siguiente problema aparece en una olimpiada de matemáticas:[9] encuentre el entero más cercano a la expresión $87 \times (15 - 4\sqrt{2}\sqrt{7})$. El alumno concursante hace un cálculo aproximando las raíces involucradas con dos decimales y obtiene $4,698$, por lo que su respuesta es 5. Para verificar, el alumno hace ahora el cálculo aproximando las raíces con tres decimales y el resultado que obtiene esta vez es $2,977488$, por lo que la respuesta sería 3. ¿Puede explicar el porqué de esta diferencia y si alguno de ambos resultados es correcto? *Solución: En este caso hay propagación de errores y estos pueden amplificarse considerablemente pues se trata de multiplicaciones. El error al aproximar con dos decimales puede estimarse tal como se vio en la sección anterior por* $87 \times (0,005 \times \sqrt{2} + 0{,}005 \times \sqrt{7}) \approx 1,77$. *En cambio, si el cálculo se hace con tres decimales el error sería* 10 *veces menor, esto es de* $0,177$. *Esto explica la diferencia en dos unidades en el resultado. Por otro lado, el error en el caso de tres decimales es menor que* $0,5$ *por lo que éste corresponde al resultado correcto.*

✍ **Ejercicio 1.5.** *Propagación de errores e interés bancario.* Se quiere calcular cuánto será, al cabo de un año, un depósito a plazo inicial de un millón de pesos con un interés mensual del $0,4\,\%$. Para ello hay que hacer el cálculo:

$$1000000 \times 1,004^{12} \approx \$1049070$$

De modo que se ganan \$49070 al cabo de un año. Lamentablemente, el banco no puede asegurar una tasa de interés constante, y ésta puede variar del $0,2\,\%$ al $0,6\,\%$. Utilizando propagación de errores, estime cuánto puede llegar a variar el cálculo anterior debido a la incertidumbre en la tasa de interés. *Solución: Tenemos que:*

$$(x + \Delta x)^n = x^n + nx^{n-1}\Delta x + \textit{términos en } (\Delta x)^2, (\Delta x)^3, \ldots, (\Delta x)^n$$

[9] Tomado del caso *¿Con cuántos decimales debo aproximar? o El Principio de la Tetera* del Proyecto FONDEF D05I-10211

de modo que, despreciando los términos de orden dos o más, el error se propaga en el cálculo del interés aproximadamente como

$$nx^{n-1}\Delta x.$$

En nuestro caso $n = 12$, $x = 1,004$ *y el error es* $\Delta x = 0{,}002$. *Aplicando esto, y sin olvidar amplificar por un millón, el error acarreado se estima así:*

$$1000000 \times 12 \times 1,004^{11} \times 0,002 \approx \$25077$$

o sea en el peor de los casos podríamos obtener \$25077 *menos o ganar* \$25077 *pesos más contando las variaciones de la tasa de interés. Esto nos da como ganancia en los casos extremos aproximadamente:*

$$\$1023993 \qquad \$1074147$$

que puede comparar con los valores exactos calculados con una calculadora con el peor $(0,2\,\%)$ *y el mejor interés* $(0,6\,\%)$*:*

$$\$1024266 \qquad \$1074424.$$

✍ **Ejercicio 1.6.** El Cuadro 1.1 resume el efecto de la propagación de errores en el resultado de varias operaciones corrientes. Obtenga los resultados usted mismo operando con las expresiones $x + \Delta x$ e $y + \Delta y$. Para ello, desprecie los términos de orden superior o igual a dos.

Operación	Error propagado aproximado
$x + y$	$\Delta x + \Delta y$
ax	$a\Delta x$
xy	$y\Delta x + x\Delta y$
$1/x$	$-\Delta x/x^2$
x^n	$nx^{n-1}\Delta x$

CUADRO 1.1. Propagación de errores.

✍ **Ejercicio 1.7.** Si quisiéramos agregar al Cuadro 1.1 la operación $\sqrt{x}$ ¿Cuál sería una estimación del error propagado? Puede serle útil la siguiente aproximación: si $|a|$ es mucho más pequeño que 1, entonces, haciendo una aproximación de Taylor[10], se obtiene:

$$\sqrt{1+a} \approx 1 + \frac{a}{2} + \text{términos de orden dos o más en } a.$$

Solución: $\frac{1}{2\sqrt{x}}\Delta x$.

[10]Vea el Capítulo 3.

✍ **Ejercicio 1.8.** ¿Hay alguna relación entre las expresiones de error propagado del Cuadro 1.1 y la derivada? Investigue.
Solución: tiene que ver con las derivadas parciales, esto es, si $f(x, y)$ es la expresión en la que queremos propagar el error, entonces el error propagado aproximado corresponde a $\frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y$. Esto se puede obtener del desarrollo de Taylor de f en dos variables, asumiendo que la función es diferenciable.

✍ **Ejercicio 1.9.** En 1792, el filósofo George Berkeley hizo una polémica crítica a Isaac Newton cuando calculaba la derivada. Por ejemplo, al calcular la derivada de x^n como el límite de $((x + \Delta x)^n - x^n)$ dividido por Δx y obtener como en el Ejercicio 1.5 la expresión nx^{n-1}, Berkeley decía que si una cantidad Δx llegaba a desaparecer en un cálculo, el cálculo dejaba automáticamente de ser válido, por lo que la derivada no tenía validez. ¿Qué opina usted?, ¿estamos usando en la propagación de errores cantidades realmente "infinitesimales"?

1.3 Cifras significativas: la corrección relativista

Las transformaciones de Galileo expresan una simple ley de la física: si vamos en la carretera en un bus a velocidad $v_1 = 20\ m/s$ y caminamos por el pasillo hacia adelante con una velocidad $v_2 = 1\ m/s$, para un observador que está parado en la carretera nuestra velocidad resultante será de

$$v_1 + v_2 = 20 + 1 = 21\ m/s$$

Nada más simple. Sin embargo, en 1950, Albert Einstein observó que esta ley violaba un postulado fundamental de la teoría de la relatividad especial: que la velocidad de la luz $c = 300\ 000\ 000\ m/s$ es siempre constante, independientemente del sistema de referencia en que se mida. Si el bus viajara a $20\ m/s$ y encendemos una linterna apuntando su foco hacia adelante del bus, para un observador parado en el camino, se sumaría la velocidad de la luz a la del bus, de modo que la luz que emerge de la linterna alcanzaría una velocidad de $300\ 000\ 020\ m/s$ para dicho observador, mayor que la velocidad de la luz.

Del mismo modo, para la luz que emerge de una estrella que se mueve, ésta se movería a una velocidad mayor a la de la luz cuando se emite en el sentido en que se mueve la estrella.[11] Así que las transformaciones de Galileo debían ser corregidas.

Al modificar la ley de suma de velocidades con la restricción de que la velocidad de la luz fuera constante, Einstein redescubrió las llamadas *transformaciones de Lorenz*, que representan una corrección relativista a la ley de suma de velocidades de Galileo. Estas transformaciones expresan que en vez de $v_1 + v_2$, debemos considerar la fórmula:

$$\frac{v_1 + v_2}{1 + \frac{v_1\, v_2}{c^2}}$$

donde c es la velocidad de la luz.

[11] La verdad es que las velocidad de la luz no varía pero sí su longitud de onda, lo que se conoce como el *efecto Doppler*.

Una de las cosas más sorprendentes es que esta nueva ley de suma de velocidades se obtuvo primero algebraicamente, a partir justamente del supuesto de que c sea constante, y no a través de la experiencia directa. Esto se debe a que la corrección que introduce la relatividad es extremadamente pequeña e imperceptible considerando las velocidades a las que estamos habituados (o sea de 1 m/s a, digamos, 320 m/s que corresponde a la velocidad del sonido).

Para verificar esto[12], consideremos de nuevo nuestro ejemplo del viaje en bus en la carretera, donde $v_1 = 20\ m/s$ y $v_2 = 1\ m/s$, entonces el cálculo relativista para la suma de velocidades sería:

$$\frac{20+1}{1+\frac{20\times 1}{300\ 000\ 000^2}}.$$

Si usted intenta hacer este cálculo en una calculadora de bolsillo probablemente le dará como resultado 21. Pero este no puede ser el resultado correcto ya que es fácil verificar que ¡el resultado debe ser estrictamente menor que 21! pues estamos dividiendo 21 por un número mayor que 1.

Utilizaremos entonces la aproximación siguiente válida para $|a|$ mucho menor que 1:

$$\frac{1}{1+a} \approx 1-a$$

la cual puede comprobarse a partir de la identidad siguiente:[13]

$$\frac{1}{1+a} = 1-a+\frac{a^2}{1+a}$$

donde se ha despreciado el término de segundo orden $\frac{a^2}{1+a}$. Utilizando esta aproximación, se obtiene:

$$\frac{21}{1+\frac{20}{9\times 10^{16}}} = \frac{21}{1+2,22\times 10^{-16}} \approx 21(1-2,2\overline{2}\times 10^{-16}) = 20,999999999999995338$$

es decir, ¡la corrección relativista se produce solamente en el décimoquinto decimal o después de la décimosexta cifra! Para detectar esto, el observador de la carretera habría requerido una regla y cronómetro que midiera longitudes y tiempos con la precisión de mil millonésimas de millonésimas de metro y de segundo respectivamente.

✍ **Ejercicio 1.10.** Estime cuál sería la corrección relativista para la velocidad relativa que llevan dos aviones que viajan en la misma dirección a 400 m/s y 200 m/s respectivamente. *Solución: $v_1 = 400$, $v_2 = -200$, con esto la velocidad clásica es $v_1 + v_2 = 400 - 200 = 200\ m/s$ y la relativista es*

$$\frac{200}{1-\frac{200\times 400}{9\times 10^{16}}} \approx 200(1+8,8\overline{8}\times 10^{-13}) = 200,0000000002\ m/s.$$

---

[12]Este ejemplo está inspirado del libro: *Álgebra Recreativa* de Y. Perelman, cf. **[22]**.

[13]O también haciendo un desarrollo de Taylor en torno a $a = 0$ de $f(a) = 1/(1+a)$. Véase el Capítulo 3 para polinomios de Taylor.

✍ **Ejercicio 1.11.** El juego del *teléfono* puede servir para motivar la idea de propagación de errores. En un grupo de personas formando una cadena, el primero piensa una frase y se la dice al oído una sola vez al del lado, quien a su vez transmite el mensaje tal y como lo ha oído al siguiente de la cadena y así sucesivamente. El último de la cadena dice en voz alta la frase que llegó a sus oídos y se compara con la original. Haga una analogía entre el juego del teléfono y una sucesión de operaciones matemáticas, ¿por qué el juego funciona mejor si hay más participantes?
Solución: cada transmisión representa una operación en la que se comete un pequeño error. Sin embargo, éste se propaga y se amplifica cada vez más. Si hay más participantes, el mensaje se transforma más y hay mayor propagación de los errores, por lo que el mensaje final puede resultar más alterado y divertido.

1.4 Precisión y cifras significativas

Retomemos el ejemplo del párrafo anterior. Llamemos x la velocidad "exacta", es decir, la velocidad que tiene incluida la corrección relativista y $\widetilde{x}$ la velocidad "aproximada"predicha por la mecánica clásica:

$$x = 20,99999999999995338 \qquad \text{y} \qquad \widetilde{x} = 21.$$

Para apreciar mejor la diferencia entre ambas velocidades, podemos utilizar el error absoluto o el error relativo (también llamado error porcentual si se multiplica por un factor 100):

Errores absoluto y relativo

$$e = |x - \widetilde{x}| = \text{error absoluto}, \qquad \epsilon = \frac{|x - \widetilde{x}|}{|\widetilde{x}|} = \text{error relativo}.$$

Con estas definiciones, los errores al no considerar la corrección relativista son:

$$e = 4,66 \times 10^{-16} \qquad \epsilon = 2,21 \times 10^{-17}.$$

Notar que, dado que los errores expresan cantidades extremadamente pequeñas, resulta especialmente cómodo utilizar la escritura en *notación científica*, la que recordamos aquí al lector[14]:

Notación científica

$$x = m \times 10^e \qquad m = \text{mantisa}, \quad e = \text{exponente},$$
$$\text{por convención} \quad 1 \leq |m| < 10 \quad \text{excepto si} \quad x = 0.$$

[14] En algunas calculadoras o computadores, la mantisa se elige de la forma: $0,1 \leq |m| < 1$ si el número es no nulo, pero aquí adoptamos el uso más común en las calculadoras corrientes que es $1 \leq |m| < 10$.

Por ejemplo:

$$0,001356 = \underbrace{1,356}_{m\text{: mantisa}} \times \underbrace{10^{-3}}_{-3\text{: exponente}} .$$

Introduzcamos ahora el concepto de *cifras significativas* (abreviado aquí como "c.s"). Diremos que un número (no nulo) está dado con p cifras significativas si su *mantisa* está dada con p cifras cuando está expresado en notación científica.

Por ejemplo $12345000 = 1,2345 \times 10^7$ está dado con 5 cifras significativas, en cambio el mismo número escrito así: $12345000 = 1,23450 \times 10^7$ está dado ¡con 6 cifras significativas! En general, es posible deducir con cuántas cifras significativas está dado un número:

$$\begin{aligned} 1;\ 0,02;\ -1 \times 10^5 \quad &\rightarrow \quad p = 1 \text{ c.s.} \\ 1,0;\ 0,022;\ -1,2 \times 10^5 \quad &\rightarrow \quad p = 2 \text{ c.s.} \\ 1,00;\ 0,0222;\ -1,20 \times 10^5,\ 101 \quad &\rightarrow \quad p = 3 \text{ c.s.} \end{aligned}$$

y solamente resulta ambiguo en números enteros que terminan en ceros:

$$100;\ 12300;\ -200 \times 10^5 \quad \rightarrow \quad \text{no se pueden determinar las c.s.}$$

en cuyo caso es mejor especificar cuántas de las cifras dadas son significativas.

Cuando indicamos que el volumen de una caja es de 1 litro, no es lo mismo decir que es de $1,0$ litro, o de $1,00$ litros, pues cada vez estamos dando el volumen con mayor *precisión*. ¿Qué es la precisión?, simplemente la agudeza con que está representado el volumen real por el número. Por ejemplo, si decimos que el volumen es $1,00$, el volumen real podría ser:

$$0,995;\ 0,996;\ 0,997;\ 0,998;\ 0,999;\ 1,000;\ 1,001;\ 1,002;\ 1,003;\ 1,004$$

o más generalmente, cualquier número en el intervalo[15]

$$[0,995, 1,005).$$

Esto se expresa usualmente así:

$$1,00 \pm 0,005$$

de donde la precisión o error máximo que podemos cometer al entregar el valor del volumen es de:

$$0,005 = 5 \times 10^{-3}.$$

La figura 1.2 puede servir para entender mejor la situación.

De manera más general, la equivalencia entre precisión y número de cifras significativas está dada por:

[15] Aquí se adopta la convención que incluye el extremo izquierdo del intervalo. Ver más adelante la convención sobre esta elección en la sección sobre redondeo.

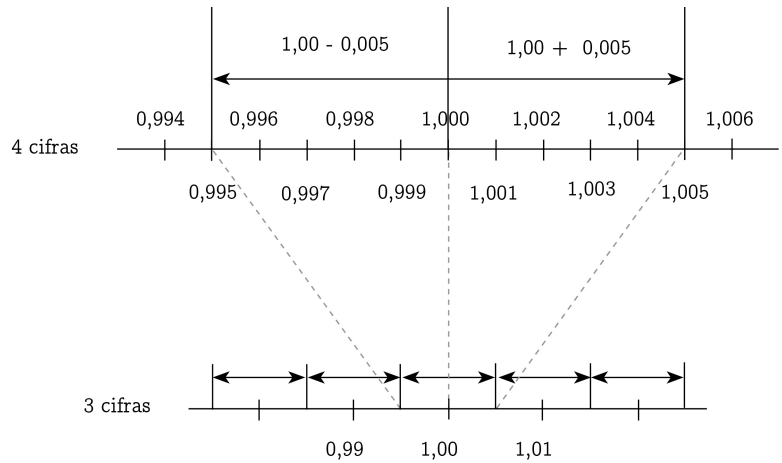


FIGURA 1.2. Precisión al utilizar tres cifras significativas. Cada número escrito con tres cifras significativas, representa en realidad un conjunto de números en un intervalo de radio $0,005 = 5 \times 10^{-3}$, lo que corresponde justamente a la precisión con que se trabaja.

PRECISIÓN Y NÚMERO DE CIFRAS SIGNIFICATIVAS

$$\begin{aligned} \eta &= \text{precisión} = 5 \times 10^{-p}, \qquad p\text{: cifras significativas,} \\ p &= \left\lfloor -\log\left(\frac{\eta}{5}\right) \right\rfloor \qquad \text{(aproximado al entero inferior).} \end{aligned}$$

La cantidad η se conoce como la *precisión relativa* con que se expresan los números y se realizan los cálculos al considerar p cifras significativas. En la mayoría de las calculadoras científicas se trabaja con 16 cifras significativas, lo que equivale a una precisión de $\eta = 5 \times 10^{-16}$. La calculadora puede representar números menores que la precisión (por ejemplo 10^{-99}), pero no distingue entre dos mantisas que están a una distancia menor que la precisión.

En nuestro ejemplo relativista, ya vimos que no es fácil hacer cálculos de corrección relativista con una calculadora normal y corriente (al menos para velocidades expresadas en m/s), y esto se debe a que serían necesarias más cifras que las que maneja internamente la calculadora para poder apreciar los efectos de dicha corrección. Más precisamente, observemos que el error relativo introducido por la corrección relativista es de

$$\epsilon = 2,21 \times 10^{-17}$$

lo que significa que para detectarlo necesitaríamos una precisión menor o igual que este error, esto es

$$\eta = 5 \times 10^{-p} \leq 2,21 \times 10^{-17}.$$

Despejando, se obtiene que el número de cifras significativas debe satisfacer:

$$p \geq -\log\left(\frac{2,21 \times 10^{-17}}{5}\right) = 17,36$$

lo que indica que se requerirían 18 cifras significativas y la mayoría de las calculadoras de bolsillo maneja solamente 16 cifras (véase el Ejercicio 1.13).

✍ **Ejercicio 1.12.** Tome una calculadora científica y sume:

$$1 + 1 \times 10^{-17}.$$

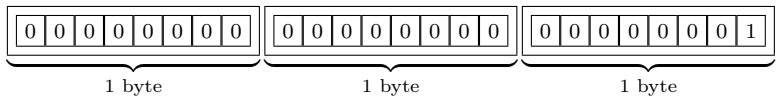
Observe el resultado que arroja la calculadora. Si el resultado es 1, ¿cómo explica esto?

Solución: El resultado correcto debería ser

$$1,00000000000000001$$

pero seguramente la calculadora trabaja con una precisión de 10^{-16}*, es decir, con 16 cifras significativas, y no es posible representar el número anterior que tiene* 18 *cifras significativas.*

✍ **Ejercicio 1.13.** Para efectos de los cálculos internos de un computador[16], la mantisa y exponente de los números se almacena en forma *binaria*. Cada *bit* de memoria puede almacenar un 0 ó un 1. Los bits se agrupan en grupos de 8, llamados *bytes*. En un ordenador con precisión de 32 bits, el signo del número se almacena en un bit (0: positivo, 1: negativo), el exponente en 8 bits y la mantisa en 23 bits, pero hay un *bit oculto* adicional que siempre es 1, así es que para efectos prácticos puede suponerse que son en realidad 24. Por ejemplo, la siguiente mantisa de 24 bits (3 bytes):



corresponde a 2^{-24} por lo que la precisión con que se puede trabajar es la mitad, o sea 2^{-25}. Esto significa que se trabaja con 7 cifras significativas ¿Por qué? En el caso de los computadores de aritmética de 64 bits, los cálculos se realizan con 16 cifras significativas pues se utilizan 53 bits para la mantisa, incluyendo el bit oculto. ¿Por qué? ¿Qué se puede decir sobre la precisión de un procesador que utiliza n bits para almacenar la mantisa, incluyendo el bit oculto?

Solución: Con n *bits, la precisión es de* $2^{-n}/2 = 2^{-(n+1)}$ *y las cifras significativas el entero inferior de*

$$-\log\left(2^{-(n+1)}\right) = (n+1)\log 2.$$

[16]Este ejercicio está basado en el estándar IEEE 754, que establece que de los 32 bits reservados para el número, se reserva uno para el signo, 8 para el exponente y 23 para la mantisa. En el caso de 64 bits la asignación es 1, 11 y 52, respectivamente.

Si $n = 24$ bits entonces el número de cifras significativas es de $25 \log 2 = 7,5 \approx 7$. Si $n = 53$ bits, el número de cifras significativas es de $54 \log 2 = 16,25 \approx 16$.

✍ **Ejercicio 1.14.** Siguiendo con el ejercicio anterior, es importante no confundir la *precisión* con que se trabaja, que depende del número de cifras de la mantisa m, con el *rango* de números con que se trabaja, o sea, el rango comprendido entre el número más pequeño representable y el número más grande, que depende del rango del exponente e. Investigue sobre el rango de números que se pueden representar en aritmética de 32 y 64 bits.

✍ **Ejercicio 1.15.** ¿Quiere saber cuál es la precisión con que trabaja su calculadora? Con ella, divida 1 por 6. Probablemente le arrojará en pantalla algo así como $0,166666666667$. ¿Cuál es el número de cifras significativas y la precisión relativa con que trabaja la calculadora? [17]

Solución: Si la cifra 7 aparece en la posición decimal 12 como en el ejemplo, entonces el error relativo es de:

$$\frac{|0,166666666667 - 0,1\overline{6}|}{1/6} = 1 \times 10^{-12}$$

dividiendo por la mitad se obtiene la precisión (2×10^{-13}) y tomando menos el logaritmo en base 10 se obtiene que la calculadora trabaja con 12 cifras significativas.

1.5 Truncatura y redondeo

Fuentes comunes de error en los cálculos son la *truncatura* y el *redondeo*. Tanto la truncatura como el redondeo son formas de aproximar un número real x por otro número $\widetilde{x}$ que tiene un cierto número p de cifras significativas dado.

Estudiar la truncatura y sobretodo el redondeo puede resultar un asunto algo árido y técnico, pero es necesario entenderlo cabalmente si se quiere hablar con propiedad cuando se hacen aproximaciones y se manejan errores en situaciones cotidianas tan importantes como son el cálculo de notas de evaluaciones, el cálculo de porcentajes, el cálculo con números decimales, etc.

Consideremos el siguiente ejemplo de motivación: en un sistema de notas del $1,0$ al $7,0$, donde la nota de aprobación es $4,0$, el promedio final de un alumno es de $3,945$. Y queremos aproximar este promedio por uno de solamente dos cifras significativas (o lo que es lo mismo, de una cifra decimal) ¿debemos aproximarlo por $3,9$ o por $4,0$? Es decir, ¿el alumno reprueba o aprueba?

Truncar un número real x a p cifras es aproximarlo, eliminando en su mantisa las últimas cifras de x hasta obtener un número $\widetilde{x}$ de p cifras significativas, pero sin cambiar el valor de las primeras cifras ya existentes. Por ejemplo, si $x = 3,945$, truncado a dos cifras nos da $\widetilde{x} = 3,9\not{4}\not{5}$.

[17] Algunas calculadoras trabajan con una mayor precisión interna y solamente se despliegan en pantalla algunos decimales, en este ejercicio suponemos que la calculadora depliega todas las cifras con las que internamente trabaja.

Redondear un número real x a p cifras es aproximarlo por el número más cercano $\widetilde{x}$ de p cifras significativas.[18] [19] Por ejemplo, si $x = 3,945$ haciendo la diferencia se puede observar que $3,945$ está más cerca de $3,9$ (diferencia $0,045$) que de $4,0$ (diferencia $0,055$) por lo que su redondeo a dos cifras es $\widetilde{x} = 3,9$.

Al redondear, en el caso de haber dos números posibles $\widetilde{x}$ a la misma distancia de x, o sea un empate, se debe especificar cuál es la convención usada: esto es, si se escogerá siempre el mayor o el menor de entre ellos. Por ejemplo, si $x = 3,95$, este se encuentra a igual distancia de los números de dos cifras $3,9$ y $4,0$. En este caso no hay un redondeo a dos cifras único y la respuesta depende de la convención. Si la convención es escoger siempre el número mayor, el redondeo será $4,0$, pero si la convención es tomar siempre el menor, el redondeo será $3,9$.

¿Hay alguna regla simple para redondear correctamente? Sí, pero hay que tener cuidado. Por ejemplo, asumiendo la convención de redondear en caso de empate al número mayor, si primero redondeamos $3,945$ a tres cifras nos da $3,95$ y si luego redondeamos $3,95$ a dos cifras nos da $4,0$. En cambio, ya vimos que si redondeamos $3,945$ directamente a dos cifras da $3,9$. Esto es, una serie de redondeos con una cifra menos cada vez no necesariamente equivale al redondeo directo reduciendo varias cifras simultáneamente. Esto hay que considerarlo si se quiere establecer una regla correcta de redondeo:

Regla de redondeo (eligiendo el mayor si hay empate)

Una vez identificada la *última* cifra significativa:

1. Si la cifra inmediatamente *posterior* es un dígito de 0 a 4, esta cifra y las siguientes se eliminan o se reemplazan por cero si están en posiciones de la unidad, la decena o más.
2. Si la cifra inmediatamente *posterior* es un dígito de 5 a 9, se aplica la misma regla anterior, pero además se aumenta en una unidad la última cifra significativa (y hay acarreo de cifras si ésta es 9).

Un ejemplo de la aplicación de la regla de redondeo se puede ver en el Cuadro 1.2. Se puede también redondear *por arriba* o *por abajo* según si al redondear, la aproximación $\widetilde{x}$ se toma *mayor o igual* o *menor o igual* que x (en valor absoluto[20]) respectivamente.

[18] No es la única forma de redondeo, ver los ejercicios sobre la "ley de redondeo" y la "regla del banquero" más adelante.

[19] A veces también se habla de redondear a un cierto número de decimales. Esto coincide con redondear al número de cifras significativas si $0,1 \leq |x| < 10$. Por otro lado, redondear a p cifras significativas es lo mismo que redondear la mantisa a p decimales.

[20] Lo del valor absoluto es necesario si queremos considerar números negativos: redondear por arriba es elegir el número más cercano en la dirección *opuesta al cero*, y esto se logra si se comparan los números en valor absoluto.

c.s.	redondear a la más cercana...	última cifra significativa	sube o no	redondeo final
7	milésima $0,001$	$2795,25\mathbf{6}2$	no sube	$2795,25\mathbf{6}$
6	centésima $0,01$	$2795,2\mathbf{5}62$	sube	$2795,2\mathbf{6}$
5	décima $0,1$	$2795,\mathbf{2}562$	sube	$2795,\mathbf{3}$
4	unidad 1	$279\mathbf{5},2562$	no sube	$279\mathbf{5}$
3	decena 10	$27\mathbf{9}5,2562$	sube (hay acarreo)	$2\mathbf{8}00$
2	centena 100	$2\mathbf{7}95,2562$	sube	$2\mathbf{8}00$
1	unidad de mil 1000	$\mathbf{2}795,2562$	sube	$\mathbf{3}000$

CUADRO 1.2. Redondeo de $2795,2562$ a un número cada vez menor de cifras significativas.

Por ejemplo, $3,945$ redondeado por abajo a dos cifras es $3,9$, pero redondeado por arriba a dos cifras es $4,0$. De hecho, redondear corresponde a elegir el menor entre los redondeos por arriba y por abajo.

Entonces, ¿cuál es la aproximación correcta de $3,945$? Vimos que según el tipo de aproximación que usemos el promedio del alumno con $3,945$ es redondeado a veces $3,9$ y a veces $4,0$. Así es que para poder dar una respuesta satisfactoria a esta pregunta, debemos seguir ahondando todavía más en el asunto.

Antes de seguir, ilustramos en la Figura 1.3 el uso de la truncatura y los distintos tipos de redondeo en una planilla de cálculo. ❐✍

Como regla general, resulta mejor redondear que truncar. Esto si nuestro objetivo es el de aproximar con la mayor precisión. En efecto, al truncar, la información de las cifras eliminadas no se utiliza; en cambio, al redondear, hay un proceso de *optimización* al buscar el número más cercano y todas las cifras del número pueden resultar importantes en el resultado del redondeo. Aunque truncar resulta más rápido que redondear, vale la pena el esfuerzo de redondear.

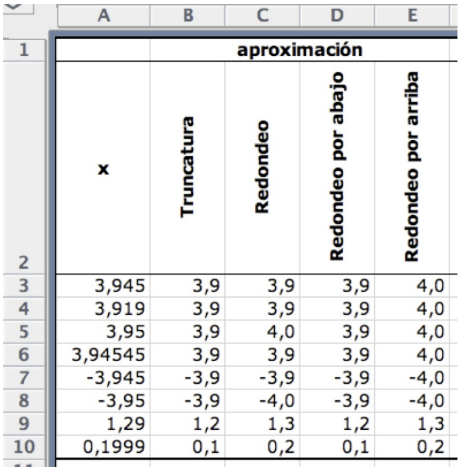
Por otro lado, el mejor redondeo es el usual, en que se elige el menor entre el redondeo por arriba y el redondeo por abajo. En este caso, se tiene que el error relativo satisface:

ERROR RELATIVO AL REDONDEAR A p CIFRAS SIGNIFICATIVAS

$$\text{error relativo} = \frac{|x-\widetilde{x}|}{|x|} \leq 5\times 10^{-p} = \text{precisión}$$

x: valor real
$\widetilde{x}$: valor redondeado
p: cifras significativas

Por ejemplo, la precisión relativa al trabajar con dos cifras significativas es de 5×10^{-2} o sea $0,05$ que es media décima. Cualquier error relativo menor que media décima resulta *no significativo*, en cambio, los errores mayores o iguales que media décima resultan *significativos*.



	A	B	C	D	E
1		aproximación			
2	x	Truncatura	Redondeo	Redondeo por abajo	Redondeo por arriba
3	3,945	3,9	3,9	3,9	4,0
4	3,919	3,9	3,9	3,9	4,0
5	3,95	3,9	4,0	3,9	4,0
6	3,94545	3,9	3,9	3,9	4,0
7	-3,945	-3,9	-3,9	-3,9	-4,0
8	-3,95	-3,9	-4,0	-3,9	-4,0
9	1,29	1,2	1,3	1,2	1,3
10	0,1999	0,1	0,2	0,1	0,2

FIGURA 1.3. Ejemplos de truncatura y redondeo a dos cifras significativas en una planilla de cálculo. En la celda `A3` se ingresa el valor a truncar o redondear. Las celdas `B3=TRUNCAR(A3;1)`, `C3=REDONDEAR(A3;1)`, `D3=REDONDEAR.MENOS(A3;1)`, `E3=REDONDEAR.MAS(A3;1)` corresponden a truncar, redondear, redondear por abajo y redondear por arriba respectivamente. La cifra después del punto y coma indica el número de decimales con los que se redondea o se trunca.

Ahora, al calcular el error relativo que se comete al aproximar $3,945$ por $4,0$ resulta $(4-3,945)/4 = 1,375\times10^{-2} < 0,05$ que no es significativo y al calcular el error relativo de aproximar $3,945$ por $3,9$ resulta $(3,945-3,9)/4 = 1,125\times10^{-2} < 0,05$ que tampoco es significativo. Entonces, resulta imposible dirimir solamente por redondeo si la nota del alumno será $3,9$ ó $4,0$ con la precisión de media décima con la que trabajamos. Es por tanto necesario *añadir otro criterio*. Por ejemplo, si pensamos que hay errores aleatorios asociados a los instrumentos de medición, pruebas o exámenes, que se utilizaron [21], entonces, ante dicha incertidumbre, y dado que la calificación es un resultado de alta consecuencia para el estudiante, ya que implica la reprobación o aprobación del alumno, si queremos inclinar la balanza a favor de éste, podríamos escoger la nota $4,0$.

✍ **Ejercicio 1.16.** *La ley del redondeo.* En Argentina existe la llamada "ley del redondeo" decretada el año 2004. La ley estipula que: "en todos aquellos casos en los que surgieran del monto total a pagar diferencias menores a cinco centavos y fuera imposible la devolución del vuelto correspondiente, la diferencia será siempre a favor

[21] Llamado "el temblor de la mano" por un veterano profesor. La *Teoría de la medición* es el área de la educación que se preocupa de este tipo de errores.

del consumidor". ¿Qué tipo de redondeo es el que hay que aplicar por ley? Si una llamada de teléfono sale \$0,78 pesos, y no se tienen monedas para dar vuelto, ¿cuánto se debe cobrar?
Solución: se debe aplicar un redondeo por abajo al múltiplo de 5 centavos más cercano. En el caso de la llamada telefónica, se debe cobrar solamente \$0,75.

✍ **Ejercicio 1.17.** *La recaudación hormiga.* Hay consumidores que reclaman que algunas empresas efectúan una recaudación hormiga cuando solicitan a sus clientes el redondeo del vuelto para donaciones. Afirman que dichas empresas se ahorran impuestos al hacer dichas donaciones a nombre de ellas mismas y no de cada donante anónimo. Estime el monto total recaudado diariamente por un gran supermercado por donaciones gracias al redondeo. Para ello, identifique los factores involucrados.
Solución: estimamos que en un día pasan por una caja 100 clientes por hora, que hay 20 cajas que trabajan 15 horas al día y que existen 200 sucursales de la empresa repartidas a lo largo del país. Suponiendo que se recaudan 5 pesos de redondeo por cada compra, la recaudación total sería de 3 millones de pesos diarios.

✍ **Ejercicio 1.18.** *Cifras significativas versus número de decimales.* Fijar un número de cifras significativas no es lo mismo que fijar un cierto número de decimales, excepto si se refieren a la mantisa. Las cifras significativas controlan el error relativo, en cambio el número de decimales controlan el error absoluto. Suponga que las medidas de una puerta son: 2 metros y 13 centímetros de alto, $89,5$ centímetros de ancho y $4,47$ centímetros de espesor. Exprese las medidas con dos cifras significativas en centímetros, redondeando si es necesario. Verifique que en las tres medidas el error relativo que se comete es menor o igual que 5×10^{-2} y que corresponden a tomar un decimal en las mantisas. Entregue ahora las medidas de la puerta con 1 decimal exacto en centímetros y verifique que el error absoluto es menor que 10^{-1} centímetros.

✍ **Ejercicio 1.19.** *La regla del banquero.* Cuando el último dígito que se eliminará es 5, el precedente, de ser impar, se aproxima al dígito par superior más cercano. Por ejemplo: $1,4145$ se aproxima a 4 cifras como $1,414$, pero $1,4155$ se aproxima a $1,414$. ¿Qué ventaja tiene esta regla respecto del redondeo usual? [22]
Solución: la idea es que la mitad de las veces ante un empate, el número será redondeado hacia arriba y la otra mitad hacia abajo, logrando que los errores de redondeo no se acumulen sistemáticamente en un único sentido y puedan cancelarse.

✍ **Ejercicio 1.20.** No siempre el número de cifras exactas es sinónimo de precisión. Encuentre una aproximación de 1 que no tenga ninguna cifra exacta, pero que tenga una precisión de 10^{-8}.
Solución: $0,99999999$.

[22] Esta regla de redondeo, llamada *insesgada*, es la que se usa por defecto en los redondeos internos de los computadores modernos para evitar propagación de errores.

Capítulo 2: Aproximando
$\pi = 3{,}14159265358979323846264338327950288\ldots$

"El cumpleaños de π promueve la celebración de la educación matemática, el gozo colectivo por las matemáticas, y el interés multicultural y atemporáneo de π. Educadores, estudiantes y padres se mezclan en una variedad de actividades públicas, expresando en forma imaginativa su pasión por la naturaleza creativa de las matemáticas"
`www.megsl.org/pi.html`

2.1 El día de π

La constante matemática π (pi) se define geométricamente como el cuociente entre el perímetro ℓ de un círculo y su diámetro (el doble del radio r):

$$\pi \equiv \frac{\ell}{2r} = 3{,}14159265358\ldots \approx 3{,}14$$

y esta razón no depende del radio del círculo en cuestión por lo que π resulta ser una constante geométrica universal [1]. El símbolo de π fue tomado de la primera letra del vocablo *perímetro* en griego ($\pi\epsilon\rho\iota\mu\epsilon\tau\rho\sigma\nu$) por el matemático galés Sir William Jones en 1706[2] y popularizado después por Leonhard Euler en el siglo XVIII, por lo que tiene más de 300 años de uso[3].

¿Por qué escoger un símbolo? Pues porque π es un número real cuya serie decimal no se repite, esto es, π es un número irracional[4] por lo que no tiene una escritura en dígitos, decimales o fracciones finita, lo que amerita un símbolo, al igual que la base de los logaritmos Neperianos e o ciertas raíces $\sqrt{2}$, $\sqrt{3}$.

Cada 14 de marzo (fecha indicada por 3-14 en los países anglosajones) se celebra el día o cumpleaños de π y cada cierto tiempo se celebran también cada uno de los nuevos dígitos de π conocidos. Hoy en día, el récord es de más de un millón de millones de decimales sucesivos calculados ($1{,}24\times10^{12}$). Para tener una idea de este gigantesco número, piense en que si se repartieran los dígitos de π ya calculados entre la población mundial actual, nos tocarían 190 a cada uno. Este cálculo le tomó más de 600 horas a

[1] Véase el Ejercicio 2.3.

[2] W. Jones, *Synopsis Palmariorum Matheseos, A New Introduction to the Mathematics*, J. Matthews, London, 1706.

[3] Para más detalles históricos véanse las referencias **[4]**, **[3]**, **[5]**.

[4] Véase el Ejercicio 2.13.

un supercomputador en Japón el año 2002. El desafío no termina ahí, ya que hoy se conocen métodos para calcular un dígito cualquiera de π sin necesariamente calcular los dígitos precedentes. Es así como el cálculo de los decimales de π ha desafiado a cada una de las generaciones de computadores (y de matemáticos e informáticos ¡por supuesto!) de nuestro tiempo.

Pero, ¿cómo se pueden calcular los dígitos de π? Existen varios tipos de métodos. La mayoría de ellos aproximan este número irracional por una *sucesión convergente*, de cuyo límite puede desprenderse el valor de π. Mientras más rápida es la convergencia de dicha sucesión, se considera que el método es más eficiente numéricamente, aunque por supuesto también es deseable la simplicidad y la elegancia del método. Esta búsqueda de más y más precisión y belleza, ha motivado el desarrollo de muchas ramas de la matemática, especialmente en la teoría de números y en el análisis real y complejo. A continuación veremos algunos de estos algoritmos, partiendo de los más simples, hasta llegar a algunos de los más sofisticados. Para implementar los algoritmos en la práctica, utilizaremos planillas de cálculo electrónicas, ya que hoy en día están al alcance de todos.

2.2 Fracciones de historia

En el papiro de Rhind , datado del 1600 a.C., que es posiblemente una transcripción de un escrito babilonio aún más antiguo del 2000 a.C., se puede leer que el área de un cuadrado de lado $9x$ es la misma que la de un círculo de diámetro $8x$. Si se utiliza que el área de un círculo es una cuarta parte de π por el cuadrado del diámetro, se obtiene para π el valor:

$$4 \times (8/9)^2 = 256/81 = 3,16049382716$$

que es una *aproximación de* π con dos cifras exactas por una fracción racional. Los egipcios no estaban interesados en calcular π como un número en sí, pero les interesaba saber cómo construir un cuadrado con la misma área que un círculo dado o viceversa. Como este valor aproximado por una fracción racional se han encontrado muchos otros en la historia, como muestra el Cuadro 2.1, sin embargo, estas fracciones no proveen un método sistemático para aproximar π con cada vez más *precisión*. Esto nos lleva más bien al estudio de *algortimos aproximantes* que veremos a continuación.

2.3 El algoritmo aproximante de Arquímides

Un *algoritmo aproximante* de π es una serie de instrucciones ordenadas o *etapas o iteraciones*, cada una de ellas con un número finito de cálculos simples a realizar, que nos permite acercarnos cada vez con más precisión a π.

Un algoritmo aproximante puede verse entonces como una sucesión convergente a π, donde el cálculo de cada término S_n de dicha sucesión es una etapa o iteración del algoritmo, cálculo que consiste en un número finito de operaciones simples (sumas, restas, multiplicaciones, divisiones, potencias y raíces). De este modo se tiene que:

fracción racional	decimal equivalente	origen histórico	cifras exactas de π
$\frac{25}{8}$	$\mathbf{3,1}25$	Babilonios 2000 a.C.	2
$\frac{256}{81}$	$\mathbf{3,1}6049382716$	Papiro de Rhind 1600 a.C.	2
$\frac{355}{113}$	$\mathbf{3,141592}920354$	Tsu Chung Chih, 450 d.C.	7
$\frac{103993}{33102}$	$\mathbf{3,14159265}3012$	Euler, siglo XVIII	10
$\frac{4913549396}{1564031349}$	$\mathbf{3,14159265358}8$	Indeterminado	12

CUADRO 2.1. Aproximaciones de π a través de fracciones racionales en las que se indican en negrita las cifras exactas que aproximan π con cada vez más precisión.

ALGORITMO APROXIMANTE PARA π

$$\lim_{n\to\infty} S_n = \pi.$$

Arquímides (287–212 a.C.) fue uno de los primeros en establecer un algoritmo aproximante para π basado en un método geométrico. Para explicarlo, recordemos que π se define como la razón entre el perímetro de un círculo y su diámetro. De esto se desprende que el perímetro de un círculo de radio r está dado por:

$$\ell = 2\pi r.$$

Si el círculo es *unitario*, o sea de radio $r = 1$, entonces su perímetro es 2π y así tenemos que:

$$\pi = \text{ semiperímetro de un círculo unitario.}$$

La idea de Arquímides es encajonar el semi-perímetro de un círculo de radio unitario por el de dos polígonos regulares, uno inscrito y el otro circunscrito, con cada vez más lados. Por ejemplo, primero por hexágonos regulares (6 lados, ver Figura 2.1), luego dodecágonos regulares (12 lados) y así sucesivamente duplicando cada vez el número de lados. De esta manera se construyen dos sucesiones convergentes a π, una (que llamaremos p_n) que converge por valores menores o *por defecto* y la otra (que llamaremos q_n) que converge por valores mayores o *por exceso*.

Este procedimiento se denomina *algoritmo de duplicación de Arquímides*. Consideremos inicialmente los dos hexágonos regulares que aparecen en la Figura 2.1,

uno inscrito y el otro circunscrito a la circunferencia, y llamemos p_0 y q_0 a sus semiperímetros. Duplicando cada vez el número de lados, se obtiene una serie de polígonos regulares inscritos y circunscritos de 6×2^n lados. Definamos como p_n y q_n los semi-perímetros de dichos polígonos. Si $2\theta_n$ es el ángulo del centro de los polígonos correspondiente a la iteración n, utilizando trigonometría (ver Figura 2.1) es fácil ver que los lados de los polígonos inscrito y circunscrito tienen longitudes $2\sin\theta_n$ y $2\tan\theta_n$ respectivamente. Multiplicando por el número de lados y dividiendo por dos (semiperímetro) se obtiene que:

$$p_n = 6 \times 2^n \sin\theta_n, \qquad p_{n+1} = 6 \times 2^{n+1} \sin(\theta_n/2)$$
$$q_n = 6 \times 2^n \tan\theta_n, \qquad q_{n+1} = 6 \times 2^{n+1} \tan(\theta_n/2).$$

Es claro de la geometría del problema, que las sucesiones están encajonadas, esto es,

$$p_n < p_{n+1} < q_{n+1} < q_n$$

de modo que ambas sucesiones de números reales (una creciente y acotada y la otra decreciente y acotada) son convergentes a reales p_∞ y q_∞ respectivamente[5].

Encontremos ahora una *relación de recurrencia* entre los términos de estas sucesiones, esto es, una regla explícita que nos permita calcular p_{n+1} y q_{n+1} una vez calculados los términos precedentes p_n y q_n. Es fácil verificar través de identidades trigonométricas que[6]:

Algoritmo de duplicación de Arquímides

$$q_{n+1} = \frac{2p_n\, q_n}{p_n + q_n}, \qquad p_{n+1} = \sqrt{p_n\, q_{n+1}} \qquad n \geq 0,$$

de modo que sabiendo que los semi-perímetros iniciales son $p_0 = 3$, $q_0 = 2\sqrt{3}$ se puede obtener primero $q_1 = 2p_0q_0/(p_0 + q_0)$ y luego $p_1 = \sqrt{p_0\, q_1}$ y así sucesivamente, se pueden calcular los términos con subíndice $n + 1$ a partir de los términos con subíndice n. Tomando límite en la segunda recurrencia se obtiene que $p_\infty^2 = p_\infty\, q_\infty$ de donde $p_\infty = q_\infty$ y los dos límites coinciden con la longitud del semicírculo unitario que es por definición π (ver también el Ejercicio 2.2).

❒✍ Construya una planilla de cálculo para este algoritmo como la que se muestra en la Figura 2.2. Para ello, defina cinco columnas con etiquetas: n (columna A), p_n (columna B), q_n (columna C), *error por defecto* $\pi - p_n$ y *error por exceso* $q_n - \pi$ (columnas C y D) usando la primera fila de la hoja de cálculo. Luego inicialice en la segunda fila de la hoja con los valores de p_0 y q_0. En la tercera fila agregue las cuatro fórmulas indicadas en la Figura 2.2 que se copiarán en las restantes filas respectivas. Debería obtener el Cuadro 2.2, donde se escogió un formato de 11 decimales para las aproximaciones y un formato de notación científica con un decimal para los errores. Se puede observar en la tabla las sucesivas aproximaciones de π por defecto y por

[5]Axioma del supremo (o ínfimo) del cuerpo de los reales. Toda sucesión creciente y acotada superiormente tiene límite real. Lo mismo vale para una sucesión decreciente y acotada inferiormente.

[6]Véase el Ejercicio 2.1

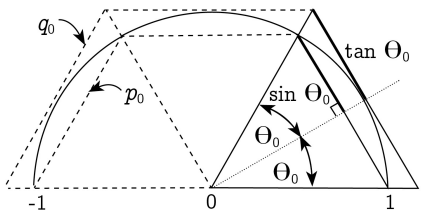


FIGURA 2.1. Configuración inicial ($n = 0$) para la aproximación del semi-perímetro de un círculo unitario por el semi-perímetro de polígonos regulares inscritos y circunscritos de 6×2^n lados.

	A	B	C	D	E
1	n	p_n	q_n	error defecto	error exceso
2	0	3,00000000000	3,46410161514	1,4E-01	3,2E-01
3	1	3,10582854123	3,21539030917	3,6E-02	7,4E-02
4	2	3,13262861328	3,15965994210	9,0E-03	1,8E-02
5	3	3,13935020305	3,14608621513	2,2E-03	4,5E-03
6	4	3,14103195089	3,14271459965	5,6E-04	1,1E-03
7	5	3,14145247229	3,14187304998	1,4E-04	2,8E-04
8	6	3,14155760791	3,14166274706	3,5E-05	7,0E-05
9	7	3,14158389215	3,14161017660	8,8E-06	1,8E-05
10	8	3,14159046323	3,14159703432	2,2E-06	4,4E-06
11	9	3,14159210600	3,14159374877	5,5E-07	1,1E-06
12	10	3,14159251669	3,14159292739	1,4E-07	2,7E-07

FIGURA 2.2. Planilla de cálculo para obtener el Cuadro 2.2. Los cuadros enmarcados corresponden a `B2=3`, `C2=2*RCUAD(3)`, `C3=2*B2*C2/(B2+C2)`, `B3=RCUAD(B2*C3)` que se copian hacia abajo en cada columna.

exceso, utilizando las sucesiones p_n y q_n. Se obtienen en cada iteración cada vez más decimales o cifras significativas[7] de π.

✍ **Ejercicio 2.1.** Obtenga las fórmulas de recurrencia del algoritmo de duplicación de Arquímides usando las identidades trigonométricas del ángulo medio:

$$\sin^2\frac{\theta}{2} = \frac{1-\cos\theta}{2}, \quad \cos^2\frac{\theta}{2} = \frac{1+\cos\theta}{2}, \quad \tan\frac{\theta}{2} = \frac{\sin\theta}{1+\cos\theta} = \frac{1-\cos\theta}{\sin\theta}.$$

Solución: pruebe primero las identidades siguientes:

$$\frac{\sin\theta\tan\theta}{\sin\theta+\tan\theta} = \tan\frac{\theta}{2}, \qquad \tan\frac{\theta}{2}\sin\theta = 2\sin^2\frac{\theta}{2}.$$

[7]Véase el Capítulo 1.

n	perímetro interno p_n	perímetro externo q_n	error por defecto $\pi - p_n$	error por exceso $q_n - \pi$	cifras exactas	c.s.
0	**3**,00000000000	**3**,46410161514	1,4E-01	3,2E-01	1	1
1	**3,1**0582854123	**3**,21539030917	3,6E-02	7,4E-02	1	2
2	**3,1**3262861328	**3,1**5965994210	9,0E-03	1,8E-02	2	2
3	**3,1**3935020305	**3,14**608621513	2,2E-03	4,5E-03	2	3
4	**3,14**103195089	**3,14**271459965	5,6E-04	1,1E-03	3	4
5	**3,14**145247229	**3,14**187304998	1,4E-04	2,8E-04	4	4
6	**3,1415**5760791	**3,14**166274706	3,5E-05	7,0E-05	4	5
7	**3,1415**8389215	**3,14**161017660	8,8E-06	1,8E-05	4	5
8	**3,14159**046323	**3,14159**703432	2,2E-06	4,4E-06	6	6
9	**3,14159**210600	**3,14159**374877	5,5E-07	1,1E-06	6	7
10	**3,14159**251669	**3,14159**292739	1,4E-07	2,7E-07	7	7

CUADRO 2.2. Aproximaciones de π a través del método de duplicación de Arquímides en las que se indica el error por defecto y exceso, el número de cifras exactas (en negrita) y de cifras significativas (c.s.) para el máximo error en cada iteración.

✍ **Ejercicio 2.2.** Usando el conocido límite:

$$\lim_{x\to 0} \frac{\sin x}{x} = 1$$

pruebe que $\lim_{n\to\infty} p_n = \pi$ eligiendo x adecuadamente, donde p_n es la sucesión de Arquímides.
Solución: tomando $x = \frac{2\pi}{6\times 2^{n+1}}$ *se obtiene*

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty} 6\times 2^n \sin\theta_n = \pi \lim_{n\to\infty} \frac{\sin\frac{2\pi}{6\times 2^{n+1}}}{\frac{2\pi}{6\times 2^{n+1}}} = \pi \lim_{x\to 0} \frac{\sin x}{x} = \pi.$$

✍ **Ejercicio 2.3.** Se dijo antes que π se puede definir como la razón entre el perímetro y el diámetro de un círculo, y que esta razón es independiente del tamaño del círculo. Para probar esto, suponga que la razón en el círculo unitario es π y demuestre que la razón sigue siendo la misma para un círculo de radio r.
Solución: basta considerar que las sucesiones p_n *y* q_n *(perímetros de los polígonos inscritos y circunscritos respectivamente) de Arquímides en el caso de un círculo unitario, se multiplican por un factor* r *en el caso de un círculo de radio* r*. De modo que el semiperímetro de un círculo de radio* r *queda aproximado por:*

$$\lim_{n\to\infty} rp_n = \lim_{n\to\infty} rq_n = \pi r$$

esto es, el perímetro es $2\pi r$*, así que la razón con el diámetro* $2r$ *es nuevamente* π*.*

2.4 Análisis de convergencia

El *análisis de convergencia* de un algoritmo aproximante S_n consiste en estudiar cuál es la precisión del algoritmo al aproximar π en la iteración n. Para ello introducimos el *error de aproximación* del algoritmo en la iteración n por

$$e_n = |\pi - S_n| \quad \text{o bien} \quad \epsilon_n = \frac{|\pi - S_n|}{\pi}$$

según si se usa el error absoluto e_n o el error relativo ϵ_n. Diremos que la precisión es mayor si el error de aproximación (absoluto o relativo) es menor y viceversa.

Para estudiar la precisión de un algoritmo aproximante, parece tentador mirar cómo aumenta el número de cifras exactas, como lo hemos hecho hasta ahora en los Cuadros 2.1 y 2.2. Lamentablemente esto induce a confusión. Es cierto que mientras más cifras exactas obtengamos, menor es el error de aproximación y por ello mayor es la precisión, pero al revés no es cierto, pues en algunos casos podemos aumentar la precisión de nuestra aproximación sin que por ello aumente necesariamente el número de cifras exactas. Por ejemplo, $0,999$ aproxima con mayor precisión a 1 que $0,9$ y, sin embargo, *el número de cifras exactas no aumenta*. Del mismo modo, si aparece un segmento como $\ldots 59999 \ldots$ en la serie de decimales de π, aproximar dicho segmento por $\ldots 60000 \ldots$ es mejor que aproximarlo por $\ldots 59000 \ldots$ aunque el segundo tenga dos cifras exactas y el primero ninguna.

Estos casos suelen ser poco frecuentes (de hecho, si asumimos que todos los dígitos del 0 al 9 aparecen en la serie de π con la misma frecuencia[8], la probabilidad de tener m nueves seguidos es de 10^{-m}), pero es mejor ser rigurosos en esto desde un principio, pues son justamente estos casos los que producen confusiones.

Por todo lo anterior, es mejor trabajar con el número de *cifras significativas* (c.s.). Este concepto lo vimos con detalle el capítulo precedente. Utilizando las ideas del capítulo anterior, diremos que si el error de aproximación (relativo) cumple:

$$5 \times 10^{-(p+1)} < \epsilon_n \leq 5 \times 10^{-p}$$

para un cierto entero p, entonces la aproximación tiene p cifras significativas (en realidad p depende de n pero no lo haremos explícito para no sobrecargar la notación). Despejando p, se obtiene entonces que

$$p \leq -\log\left(\frac{\epsilon_n}{5}\right) < p + 1$$

o sea:

Cifras significativas

$$p = \left\lfloor -\log\left(\frac{\epsilon_n}{5}\right)\right\rfloor = \left\lfloor -\log\left(\frac{e_n}{5\pi}\right)\right\rfloor,$$

[8]Esta conjetura, confirmada sólo experimentalmente, es la supuesta *normalidad* de π y su prueba es un problema abierto en matemáticas. La misma conjetura existe para otros números irracionales como $\sqrt{2}$, e, $\log(2)$.

esto es, el número de cifras significativas es el entero inferior[9] de menos el logaritmo de un quinto del error relativo. La ventaja ahora es que si el error de aproximación (absoluto o relativo) disminuye, estamos seguros que el número de cifras significativas aumenta, lo que no ocurría necesariamente con el número de cifras exactas.

En el Cuadro 2.2 se muestran en las dos últimas columnas el número de cifras exactas y el número de cifras significativas (utilizando la fórmula previa) para el peor caso, es decir, considerando el mayor error entre el error por defecto (tercera columna) y el error por exceso (cuarta columna). Notar que el número de cifras significativas va en aumento.

Para estudiar numéricamente como aumentan las cifras significativas en cada iteración, se puede graficar el logaritmo del error en función del número de iteración[10]. Por ejemplo, si se grafica el error de aproximación relativo por defecto $\epsilon_n = \frac{\pi - p_n}{\pi}$ en función de n (columna 3 del Cuadro 2.2) se obtiene una recta de pendiente aproximada $-3/5$ (hágalo usted mismo en su planilla electrónica de cálculo). De modo que si asumimos

$$\log \epsilon_n = -\frac{3}{5}n + C_1$$

donde C_1 es una cierta constante. Reemplazando esta expresión en

$$p \leq -\log \epsilon_n + \log 5 < p + 1$$

se obtiene que

$$\frac{3}{5}n + C_2 - 1 < p \leq \frac{3}{5}n + C_2$$

donde $C_2 = \log 5 - C_1$ es otra constante. La interpretación de esto es que el número de cifras significativas p crece como $3n/5$ donde n es el número de iteraciones. En otras palabras, se ganan 3 cifras significativas por cada 5 iteraciones. Se dice en este caso que el número de cifras significativas *aumenta de manera lineal* con n.

La anterior es una conjetura experimental o empírica que se obtuvo a partir de la observación de una simulación numérica, pero lamentablemente no tiene una validez general hasta no ser probada matemáticamente. Para algunos algoritmos es posible probar teóricamente cómo aumenta la precisión y en otros casos no. Pero en todos los casos las simulaciones numéricas nos pueden dar un indicio de cómo se comporta el algoritmo, lo que nos permite muchas veces hacer alguna conjetura.

En el caso particular del algoritmo de duplicación de Arquímides, es posible analizar el error teóricamente. Esto nos servirá para contrastar con la conjetura que obtuvimos numéricamente. Para ello, definamos la suma de los errores (relativos) de aproximación por defecto y por exceso del método como

$$\epsilon_n = \frac{\pi - p_n}{\pi} + \frac{q_n - \pi}{\pi} = \frac{q_n - p_n}{\pi}.$$

[9]Denotamos por $\lfloor x \rfloor$ (función cajón inferior) al entero inferior y más cercano a x.

[10]Llamado gráfico *semi-log* o semi logarítmico.

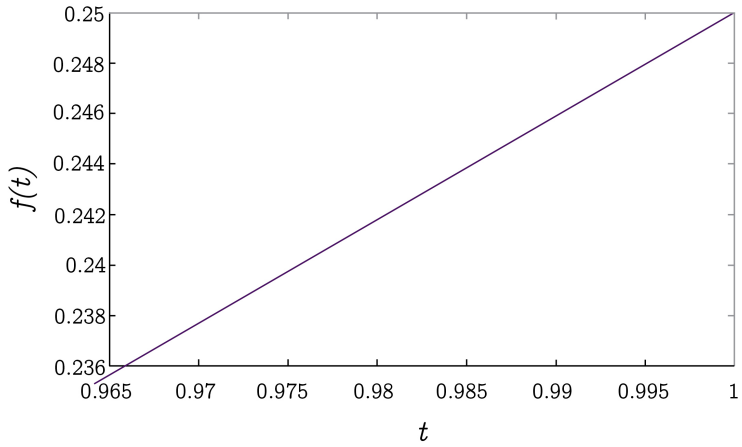


FIGURA 2.3. Gráfico de la función $f(t) = \frac{2t^2-1}{2t^2(1+t)}$ entre $t = \cos(\theta_0/2)$ y $t = 1$.

De las expresiones trigonométricas de p_n y q_n, es posible ver que:

$$\frac{\epsilon_{n+1}}{\epsilon_n} = \frac{q_{n+1} - p_{n+1}}{q_n - p_n} = \frac{2t^2 - 1}{2t^2(1+t)} \equiv f(t), \qquad t = \cos(\theta_n/2)$$

y como la función $f(t)$ es creciente en t (ver Figura 2.3), toma su valor máximo para $t = 1$ ($f(1) = 0{,}25$) que se alcanza cuando n tiende a infinito ya que $\theta_n = \frac{\pi}{6\times 2^n}$ tiende a cero y toma su valor mínimo para $\theta_0 = \frac{\pi}{6}$ (o sea $f(\cos(\theta_0/2)) = 0{,}2361$). En consecuencia:

$$0,236 \leq \frac{\epsilon_{n+1}}{\epsilon_n} \leq 0,25 \quad \text{para todo } n.$$

Esto es, la suma de los errores de aproximación se reduce al menos 4 veces en cada iteración y, por tanto, también cada uno de los errores de aproximación por defecto y por exceso de p_n y q_n. Usando que

$$\frac{\epsilon_n}{\epsilon_0} = \frac{\epsilon_n}{\epsilon_{n-1}} \cdot \frac{\epsilon_{n-1}}{\epsilon_{n-2}} \cdots \frac{\epsilon_2}{\epsilon_1} \cdot \frac{\epsilon_1}{\epsilon_0}$$

se obtiene:

$$(0,236)^n \leq \frac{\epsilon_n}{\epsilon_0} \leq (0,25)^n \quad \text{para todo } n.$$

Tomando logaritmo y definiendo $C = \log \epsilon_0$

$$-0,626n + C \leq n \log 0,236 + C \leq \log \epsilon_n \leq n \log 0,25 + C \leq -0,603n + C$$

y como $0,6 = \frac{3}{5}$ se obtiene lo mismo que ya se observó numéricamente antes, esto es, que las aproximación por defecto y por exceso del algoritmo de separación de Arquímides proveen aproximadamente 3 cifras significativas adicionales cada 5 iteraciones.

2.5 Algoritmos ineficientes y algoritmos eficientes

Un algoritmo aproximante para π muy *ineficiente* se puede escribir a partir de la célebre serie del cuadrado de los recíprocos,[11] establecida por Leonhard Euler en 1739:

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + ... = \pi.$$

Miremos la serie como una sucesión de sumas parciales donde la suma parcial S_{n+1} es S_n más $1/(n+1)^2$ esto es

> Algoritmo ineficiente para aproximar π
>
> $$S_0 = 1, \qquad S_{n+1} = S_n + \frac{1}{(n+1)^2}.$$

El error de aproximación absoluto es:

$$e_n = |S_n - \pi| = \sum_{k=1}^{\infty} \frac{1}{k^2} - \sum_{k=1}^{n} \frac{1}{k^2} = \sum_{k=n+1}^{\infty} \frac{1}{k^2}$$

y se puede estimar de la siguiente forma (ver Figura 2.4):

$$\frac{1}{n+1} = \int_{n+1}^{\infty} \frac{1}{x^2}\, dx \leq \sum_{k=n+1}^{\infty} \frac{1}{k^2} \leq \int_{n}^{\infty} \frac{1}{x^2}\, dx = \frac{1}{n}.$$

De modo que dividiendo por π para obtener el error relativo, dividiendo por cinco y tomando menos logaritmo vemos que el número de cifras significativas satisface:

$$\log n + C \leq p \leq \log(n+1) + C$$

donde C es una constante. Esto es, el número de cifras significativas que se obtienen en cada iteración de la serie de los cuadrados de los inversos de Euler *aumenta de manera logarítmica*. Para darse cuenta de lo extremadamente lenta que es esta convergencia, piense que si n es diez mil millones ($n = 10^{10}$), entonces $\log n = 10$, es decir, se necesitan sumar diez mil millones de términos de la serie para obtener con 10 cifras significativas π. Esto ilustra el hecho de que no siempre una bella fórmula teórica es útil numéricamente y que, en este caso, el algoritmo es altamente ineficiente.

Imagínese ahora un método muy *eficiente* en que se doble el número de cifras significativas en cada iteración del algoritmo. Es el caso del algoritmo desarrollado en 1976 por Brent y Salamin y basado en ideas estudiadas mucho antes por el matemático Srinivasa Ramanujan (1887-1920). Para escribirlo, tomemos los dos números reales 1 y $1/\sqrt{2}$ y el límite de las sucesiones:

$$a_1 = 1, \quad b_1 = \frac{1}{\sqrt{2}}$$

$$a_{k+1} = \frac{a_k + b_k}{2}, \quad b_{k+1} = \sqrt{a_k\, b_k}, \quad k \geq 1.$$

[11] Para una idea intuitiva de la factibilidad de esta serie, véase el Ejercicio 2.14 más adelante.

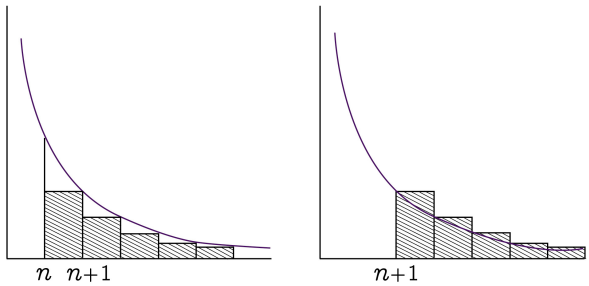


FIGURA 2.4. La serie $\displaystyle\sum_{k=n+1}^{\infty} \frac{1}{k^2}$ minora el área de $1/x^2$ desde n y la mayora desde $n+1$.

El algoritmo para aproximar π de Brent y Salamin está dado por la sucesión aproximante que converge a π siguiente:[12]

ALGORITMO DE BRENT Y SALAMIN

$$BS_n = \frac{4a_{n+1}^2}{1 - 2\sum_{k=1}^{n} 2^k (a_{k+1}^2 - b_{k+1}^2)}, \quad n \geq 0.$$

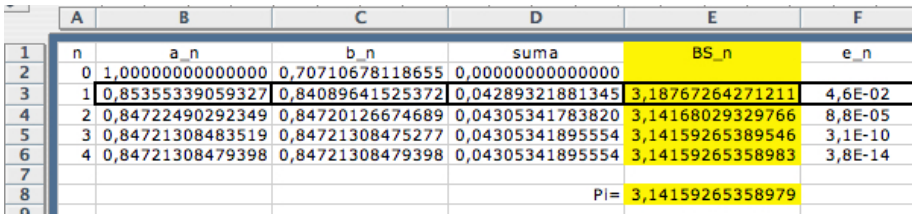
✍ **Ejercicio 2.4.** Probar que las sucesiones a_k y b_k definidas más arriba están encajonadas y convergen a un mismo límite, llamada la *media aritmético-geométrica* entre a_1 y b_1.

Para ver cómo se comporta numéricamente el algoritmo, impleméntelo en una planilla de cálculo electrónica. Siga las instrucciones de la Figura 2.5. Con este programa simple se obtienen cuatro iteraciones del algoritmo que se muestran en el Cuadro 2.3, pero no se puede seguir mucho más allá debido al límite de precisión de los cálculos. ❒✍

✍ **Ejercicio 2.5.** John Machin, profesor inglés de astronomía, fue uno de los primeros en descubrir algoritmos eficientes para aproximar π usando cálculos "a mano". Esto lo hizo en 1706, usando la identidad trigonométrica $\arctan x + \arctan y = \arctan \frac{x+y}{1-xy}$. Investigue sobre cuáles fueron dichos algoritmos y con cuántos dígitos se llegó a aproximar π usando solamente cálculos a mano.

Solución: el mismo Machin calculó 100 dígitos de π usando su algoritmo. Los algoritmos de Machin sirvieron para aproximar luego π con 707 decimales en 1873, esto lo hizo William Shanks con cálculos a mano que le tomaron varios años. En 1945, D. F. Ferguson descubrió un error y que solo eran correctos los primeros 527 dígitos, así es que Ferguson pasó un año

[12] La demostración de que este algoritmo converge a π es complicada y está fuera del alcance del presente texto. Para mayores detalles puede consultar la referencia **[10]**.



	A	B	C	D	E	F
1	n	a_n	b_n	suma	BS_n	e_n
2	0	1,00000000000000	0,70710678118655	0,00000000000000		
3	1	0,85355339059327	0,84089641525372	0,04289321881345	3,18767264271211	4,6E-02
4	2	0,84722490292349	0,84720126674689	0,04305341783820	3,14168029329766	8,8E-05
5	3	0,84721308483519	0,84721308475277	0,04305341895554	3,14159265389546	3,1E-10
6	4	0,84721308479398	0,84721308479398	0,04305341895554	3,14159265358983	3,8E-14
7						
8				Pi=	3,14159265358979	

FIGURA 2.5. Planilla de cálculo para obtener el Cuadro 2.3. Se inicializa el algoritmo con `B2=1`, `C2=1/RCUAD(2)`, `D2=0`. Los cuadros enmarcados corresponden a las fórmulas `B3=(B2+C2)/2`, `C3=RCUAD(B2*C2)`, `D3=D2+2^A3*(B3^2-C3^2)`, `E3=4*B3^2/(1-2*D3)`, que se copian hacia abajo en cada columna. Para calcular el error de aproximación se agrega la columna `E` con el valor de pi en `E8=PI()` y la fórmula `F3=ABS(E$8-E3)` que se copia hacia abajo. No olvide poner formato de celda de número de 14 decimales para las columnas de la `B` a la `E` y científica para la columna del error de aproximación `F`.

n	BS_n	error	c.e.	c.s.
1	**3,1**8767264271211	4,6E-02	2	2
2	**3,141**68029329766	8,8E-05	4	5
3	**3,1415926535**89546	3,1E-10	10	10
4	**3,14159265358983**	3,8E-14	13	14

CUADRO 2.3. Aproximación de π con el algoritmo de Brent-Salamin basado en la media aritmético-geométrica. Notar que en solamente 4 iteraciones del algoritmo se obtienen 13 cifras exactas (c.e.) de π ó 14 cifras significativas (c.s.).

calculando 808 decimales, retomando el cálculo desde el error, y finalmente, en 1947, Levi Smith y John Wrench batieron el récord llegando a 1000 dígitos. Nótese que dos años más tarde, en 1949, uno de los primeros computadores (ENIAC) calcularía 2037 dígitos de π en solamente 3 días (véase [**4**]*). De ahí en adelante la historia cambiaría, hoy en día se pueden obtener dígitos de π a una velocidad de más de 500,000 digitos de π por segundo.*

2.6 Estimaciones a priori

Hasta ahora, hemos podido acotar el error de los algoritmos aproximantes sin necesidad de calcular dichos errores explícitamente. Este tipo de estimaciones teóricas son llamadas *estimaciones a priori* y su obtención es uno de los objetivos fundamentales del análisis numérico.

En el caso del algoritmo de Brent y Salamin, es posible probar la siguiente estimación o cota superior para el error de convergencia del método [13]:

$$e_n = |BS_n - \pi| \leq C2^{n+2}\frac{1}{e^{\pi 2^{n+1}}}.$$

Tomando el opuesto del logaritmo del error vemos que el número de cifras significativas que se obtiene en cada iteración aumenta al menos de manera proporcional a 2^{n+1} para n grande. Esto es, la razón del número de cifras significativas entre dos iteraciones sucesivas es 2, o también, el número de cifras significativas se duplica en cada iteración para valores grandes de n. Este método podríamos decir que es *muy eficiente* y se habla de *super convergencia*. Mencionemos que existen incluso algoritmos en que el número de cifras exactas se triplica, cuadriplica y hasta se multiplica por 16 en cada iteración.

2.7 Acelerando la convergencia

Hay un modo de *acelerar* la convergencia de algunos algorimos, en el sentido de aumentar su velocidad de convergencia, o velocidad con que decrece el error de aproximación en función del número de iteraciones. Este método es llamado el *método de aceleración de Aitken*. Dado un algoritmo representado por una sucesión a_n convergente a π con error de aproximación $e_n = |a_n - \pi|$, definimos un nuevo algoritmo dado por la sucesión

Aceleración de Aitken

$$\overline{a}_n = a_n - \frac{(\Delta a_n)^2}{\Delta^2 a_n}, \qquad \Delta a_n = a_{n+1} - a_n, \qquad \Delta^2 a_n = \Delta a_{n+1} - \Delta a_n.$$

Por ejemplo, el siguiente diagrama indica cómo se obtiene $\overline{a}_0$ a partir de a_0, a_1 y a_2:

$$\begin{array}{ccccc} a_0 & & & & \\ & \searrow & & & \\ & & \Delta a_0 = a_1 - a_0 & & \\ & \nearrow & & \searrow & \\ a_1 & & & & \Delta^2 a_0 = \Delta a_1 - \Delta a_0 \quad \longrightarrow \quad \overline{a}_0 = a_0 - \dfrac{(\Delta a_0)^2}{\Delta^2 a_0} \\ & \searrow & & \nearrow & \\ & & \Delta a_1 = a_2 - a_1 & & \\ & \nearrow & & & \\ a_2 & & & & \end{array}$$

La sucesión $\overline{a}_n$ también converge a π, pero su convergencia puede ser más rápida. En efecto, suponiendo que nunca dividimos por cero y que en cada iteración el error disminuye según la regla $a_{n+1} - \pi = \alpha(a_n - \pi)$ para algún $0 < \alpha < 1$ (por ejemplo, vimos que este es aproximadamente el caso del algoritmo de duplicación de Arquímides

[13] La prueba de este resultado es complicada y puede encontrarse en la referencia **[10]**

con $\alpha = 0{,}25$), es fácil verificar que el nuevo error de aproximación $\overline{e}_n = |\overline{a}_n - \pi|$ satisface

$$\lim_{n\to\infty} \frac{\overline{e}_n}{e_n} = 0$$

de modo que $\overline{e}_n$ (error asociado al algoritmo con aceleración de Aitken) tiende a cero más rápidamente que e_n (error asociado al algoritmo sin aceleración).

✍ **Ejercicio 2.6.** Pruebe el límite anterior, para ello, note que $\Delta^2 a_n = a_{n+2} - 2a_{n+1} + a_n$ y pruebe que el límite pedido es de hecho el límite de $a_{n+2} - \pi - \alpha(a_{n+1} - \pi)$ que es cero.

❒✍ Veamos en la planilla de cálculo la mejora del algoritmo de duplicación de Arquímides después de aplicar la aceleración de Aitken. Por ejemplo, si la sucesión p_n se encuentra en la columna `B` de la hoja de cálculo de la Figura 2.2, basta agregar una columna `F` para Δp_n con la fórmula `F2=B3-B2` y una columna `G` para $\Delta^2 p_n$ con fórmula `G2=F3-F2` y copiarlas hacia abajo en cada columna y luego agregar una columna `H` con la fórmula `H2=B2-F2*F2/G2` para el cálculo de $\overline{p}_n$. De manera análoga se puede hacer para $\overline{q}_n$. La planilla utilizada se muestra en la Figura 2.6 y los resultados del cálculo se muestran en el Cuadro 2.4. Notar que se pierde la característica de encajonamiento de las sucesiones aceleradas $\overline{p}_n$ y $\overline{q}_n$, pero la velocidad de convergencia a π mejora en ambas.

	F	G	H	I
1	Delta p_n	Delta^2 p_n	Aitken p_n	error p_n
2	0,10582854123	-0,07902846918	3,14171703255	1,2E-04
3	0,02680007205	-0,02007848229	3,14160036162	7,7E-06
4	0,00672158977	-0,00503984192	3,14159313433	4,8E-07
5	0,00168174784	-0,00126122645	3,14159268362	3,0E-08
6	0,00042052139	-0,00031538577	3,14159265547	1,9E-09
7	0,00010513563	-0,00007885139	3,14159265371	1,2E-10
8	0,00002628424	-0,00001971316	3,14159265360	7,3E-12
9	0,00000657108	-0,00000492831	3,14159265359	4,6E-13
10	0,00000164277	-0,00000123208	3,14159265359	2,8E-14
11	0,00000041069			

FIGURA 2.6. Planilla de cálculo para obtener el Cuadro 2.4. Los cuadros enmarcados corresponden a `F2=B3-B2`, `G2=F3-F2`, `H2=B2-F2*F2/G2` que se copian hacia abajo en cada columna.

2.8 Digitalizando π

Estamos acostumbrados a trabajar con 10 dígitos, pero por muchas razones prácticas es bueno saber operar con otras bases. Por ejemplo, es bien conocido que la información digital se almacena en forma *binaria* usando ceros y unos, como es el caso en un disco digital de música o video.

n	$\overline{p}_n$	$\overline{q}_n$	error $\overline{p}_n$	error $\overline{q}_n$	c.e.	c.s.
0	**3,141**71703255	**3,14**356574216	1,2E-04	2,0E-03	3	3
1	**3,141**60036162	**3,141**71570403	7,7E-06	1,2E-04	4	5
2	**3,14159**313433	**3,141**60034099	4,8E-07	7,7E-06	4	6
3	**3,14159268**362	**3,14159**313401	3,0E-08	4,8E-07	6	7
4	**3,14159265**547	**3,14159268**362	1,9E-09	3,0E-08	8	8
5	**3,14159265**371	**3,14159265**547	1,2E-10	1,9E-09	9	9
6	**3,14159265**360	**3,14159265**371	7,3E-12	1,2E-10	9	11
7	**3,14159265359**	**3,14159265**360	4,6E-13	7,3E-12	9	12
8	**3,14159265359**	**3,14159265359**	2,8E-14	4,6E-13	10	13

CUADRO 2.4. Aproximaciones de π a través del método de separación de Arquímides con aceleración de Aitken. Notar que ahora se ganan aproximadamente 3 cifras significativas cada 3 iteraciones lo que mejora el desempeño del algoritmo original en el que se ganaban 3 cada 5 (compare con el Cuadro 2.2).

Si cada *bit* de información es 0 ó 1, entonces n bits permiten 2^n combinaciones. La unidad actual de cantidad de información es el *byte* o grupo de 8 bits, que permite representar un entero del 0 al 127. La razón de esto es que 8 bits bastan para almacenar una letra o símbolo corriente de un texto por un código del 0 al 127 llamado *código ascii*.

Digitalizar un número (o un texto, por ejemplo, si se usa el código ascii) es simplemente representarlo por bits o, en otras palabras, cambiar su representación de la base 10 a la base 2.

Para digitalizar π, primero descompongámoslo en potencias de 10: ❐✍

$$3,14159\ldots = 3\times 10^0 + 1\times 10^{-1} + 4\times 10^{-2} + 1\times 10^{-3} + 5\times 10^{-4} + 9\times 10^{-5}\ldots.$$

De esta descomposición se desprende que para obtener las cifras o dígitos de la *parte fraccionaria* de π, eso es, de la parte menor que 1, basta tomar la parte entera de multiplicar reiteradamente por 10 el resto decimal. Esto es:

$$\begin{aligned} 0,14159\ldots\times 10 = 1,4159\ldots &\rightarrow 1 \text{ (parte entera)}\\ 0,4159\ldots\times 10 = 4,159\ldots &\rightarrow 4 \text{ (parte entera)}\\ 0,159\ldots\times 10 = 1,59\ldots &\rightarrow 1 \text{ (parte entera)}\ldots \end{aligned}$$

Por otro lado, sabemos que para obtener las cifras de la *parte entera* de un número real, debemos tomar el residuo o resto de dividir reiteradamente por 10. En el caso de π, esto da simplemente el dígito 3 que lo encabeza.

Repitamos ahora el procedimiento anterior cambiando la base 10 por la base 2:

$$\begin{aligned}
&\text{Parte entera:} \\
3 : 2 = 1, \quad &\to \quad 1 \text{ (residuo)} \\
1 : 2 = 0 \quad &\to \quad 1 \text{ (residuo)} \\
&\text{Parte fraccionaria:} \\
0,141592 \times 2 = 0,283185\ldots \quad &\to \quad 0 \text{ (parte entera)} \\
0,283185 \times 2 = 0,566370\ldots \quad &\to \quad 0 \text{ (parte entera)} \\
0,566370 \times 2 = 1,132741\ldots \quad &\to \quad 1 \text{ (parte entera)} \\
0,132741 \times 2 = 0,265482\ldots \quad &\to \quad 0 \text{ (parte entera)}\ldots
\end{aligned}$$

Siguiendo así se obtiene:

$$\begin{aligned}
\pi = 3 \quad &, \quad 14159265358979323846\ldots \\
\text{residuos al dividir por 2} \longleftarrow \quad &| \quad \longrightarrow \text{partes enteras al multiplicar por 2} \\
\pi = 11 \quad &, \quad 00100100001111110110101010001000\ldots
\end{aligned}$$

que corresponde exactamente a los coeficientes de π en su descomposición en potencias de 2:

$$\pi = 1 \times 2^2 + 1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + \ldots.$$

Será de utilidad saber que n cifras decimales en base 10 equivalen aproximadamente a $3,32 \times n$ cifras en base 2. Para ver esto, simplemente preguntémonos: ¿cuántas veces debemos dividir 10 para obtener un número fraccionario? Esto corresponde a resolver la inecuación $10/2^x < 1$ cuya solución es $x > \log(2) \approx 3,32$.

2.9 Exprimiendo π gota a gota

Un algoritmo *gota a gota* es uno que genera los dígitos de π uno por uno, sin usar en el cálculo de un nuevo dígito todos los precedentes, de modo que el cálculo de cada nuevo dígito necesita siempre la misma cantidad de memoria en el computador. La ventaja es que no se requiere trabajar con números de, digamos, 100 decimales para obtener 100 dígitos de π, como es el caso de los algoritmos vistos en las secciones precedentes. La desventaja es que se debe reservar más y más memoria mientras más dígitos se deseen.

Este tipo de algoritmos fue primero descubierto para aproximar el número e en 1968 y varios años más tarde fue utilizado para aproximar π por Rabinowitz y Wagon en 1995, quienes los bautizaron como algoritmo gota a gota. El punto de partida es escribir π como

$$\pi = \mathbf{3} + \frac{1}{10}\left(\mathbf{1} + \frac{1}{10}\left(\mathbf{4} + \frac{1}{10}\left(\mathbf{1} + \frac{1}{10}\left(\mathbf{5} + \ldots \quad \right)\right)\right)\right)$$

digamos en "base" $\left[\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \ldots\right]$. Luego mirar la siguiente fórmula de Euler:

BASE PARA ALGORITMO CUENTA GOTAS

$$\pi = \mathbf{2} + \frac{1}{3}\left(\mathbf{2} + \frac{2}{5}\left(\mathbf{2} + \frac{3}{7}\left(\mathbf{2} + \frac{4}{9}\left(\mathbf{2} + \dots \quad \right)\right)\right)\right),$$

que significa que $\pi = 2,2222\ldots$ en "base" $\left[\frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{4}{9} \ldots\right]$. Usando esta idea y haciendo el correspondiente cambio de base, se puede construir un algoritmo [14] que provee los decimales de π uno a uno. La idea esta vez no es explicar detalladamente el algoritmo[15], sino que proveer una planilla que lo implementa y reconocer en ella que efectivamente se realiza un cierto cambio de base a partir de las explicaciones de la sección precedente.

Para completar la planilla de cálculo que implementa este algoritmo, siga cuidadosamente las instrucciones al pie de la Figura 2.7. Note que, como se hace un cambio de base tipo base 2 a base 10, para obtener n cifras exactas de π, se debe utilizar un número de columnas igual al entero superior a $3,32\,n$, por la razón explicada al final de la sección anterior. Por ejemplo, con las 15 columnas del ejemplo de la Figura 2.7 se pueden obtener con seguridad al menos 4 cifras de π exactas. ❒✍

2.10 Tajadas digitales de π

En 1997, Bailey, Borwein y Plouffe presentaron un algoritmo para obtener una *tajada digital* de π, esto es, cualquier dígito de π y sus vecinos a la derecha sin necesidad de calcular los dígitos anteriores. El llamado *algoritmo BBP* permite obtener tajadas de π en su expresión binaria. Por ejemplo, la fórmula:

ALGORITMO BPP EN BASE 4

$$\pi = \sum_{k=1}^{\infty} \frac{1}{4^k}\left(\frac{2}{4k+1} + \frac{2}{4k+2} + \frac{1}{4k+3}\right),$$

permite encontrar tajadas de π escrito en base 4, que se pueden traducir fácilmente al binario usando el siguiente diccionario (notar que en base 4 sólo se usan los dígitos 0, 1, 2 y 3):

$$\begin{array}{cccccc} \underbrace{00} & \underbrace{01} & \underbrace{10} & \underbrace{11} & \rightarrow & \text{base 2} \\ 0 & 1 & 2 & 3 & \rightarrow & \text{base 4} \end{array}$$

Por ejemplo, π en base 2 y base 4 se escribe así:

$$\begin{array}{cccccccccccc} \pi & = & 11 & , & 00 & 10 & 01 & 00 & 00 & 11 & 11 & 11 & 01\ldots\text{(base 2)} \\ \pi & = & 3 & , & 0 & 2 & 1 & 0 & 0 & 3 & 3 & 3 & 1\ \ldots\text{(base 4)} \end{array}$$

Imaginémonos ahora que quisiéramos calcular la cifra 1000 de π en base 4. Como en la serie que aproxima π los términos decrecen como 4^{-k}, para calcular la cifra 1000

[14] Esta planilla se adaptó de una obtenida de la página *L´univers de Pi* `www.pi314.net`. Véase la referencia **[14]**.

[15] Pueden consultarse para ello las referencias **[26]**, **[23]**.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		pi															
2	num			1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	den		1	3	5	7	9	11	13	15	17	19	21	23	25	27	29
4																	
5	Inicio		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
6																	
7	*10		20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
8	acarreo		10	12	12	12	10	12	7	8	9	0	0	0	0	0	0
9	suma	3	30	32	32	32	30	32	27	28	29	20	20	20	20	20	20
10	resto		0	2	2	4	3	10	1	13	12	1	20	20	20	20	20
11																	
12	*10		0	20	20	40	30	100	10	130	120	10	200	200	200	200	200
13	acarreo		13	20	33	40	65	48	98	88	81	170	165	156	130	84	0
14	suma	1	13	40	53	80	95	148	108	218	201	180	365	356	330	284	200
15	resto		3	1	3	3	5	5	4	8	14	9	8	11	5	14	26
16																	
17	*10		30	10	30	30	50	50	40	80	140	90	80	110	50	140	260
18	acarreo		11	24	30	40	45	54	77	96	72	70	77	72	117	112	0
19	suma	4	41	34	60	70	95	104	117	176	212	160	157	182	167	252	260
20	resto		1	1	0	0	5	5	0	11	8	8	10	21	17	9	28
21																	
22	*10		10	10	0	0	50	50	0	110	80	80	100	210	170	90	280
23	acarreo		5	6	15	36	35	36	84	80	90	120	154	120	104	126	0
24	suma	1	15	16	15	36	85	86	84	190	170	200	254	330	274	216	280
25	resto		5	1	0	1	4	9	6	10	0	10	2	8	24	0	19

FIGURA 2.7. Arriba: construcción de la planilla de cálculo para el algoritmo gota a gota. Las celdas enmarcadas a la derecha corresponden a `Q7=Q5*10`, `P8=ENTERO(Q9/Q$3)*Q$2`, `Q8=0`, `Q9=Q7+Q8`, `Q10=RESTO(Q9;Q$3)` que se copian en la misma línea hacia la izquierda. Las celdas enmarcadas a la izquierda corresponden a `C10=RESTO(C9;10)` y `B9=ENTERO(C9/10)`. Luego se copia cuatro veces todo el bloque `A7:Q10` en los bloques que se indican más abajo. Se obtienen las primeras cifras de π en `B9`, `B14`, `B19` y `B25`.

en base 4 de π se pueden despreciar los términos demasiado pequeños de esta serie. En efecto, solo un número finito de términos (del orden de 1000) van a influir sobre la cifra 1000. Se trata pues de calcular con facilidad las cifras $1000, 1001, 1002, \ldots$ de estos términos de la sumatoria para $k = 1, 2, \ldots, 1000$ y unos pocos más para poder sumarlas.

Pero al sumar dos números, el valor de la cifra 1000 del resultado depende solamente de las cifras $1000, 1001, \ldots, 1000 + m$ de los sumandos, donde m es el largo del acarreo de cifras. El entero m es en general pequeño, pues la probabilidad de un acarreo de largo m es 1 de cada 4^m.

Los términos para los que hay que calcular la cifras $1000, 1001, \ldots, 1000 + m$ son de la forma $\frac{1}{(4k+i)4^k}$ con $i = 1, 2, 3$ y hay un forma fácil de calcular sus cifras a partir de la 1000-ésima. Para ilustrar esto notemos, por ejemplo, que es fácil calcular el

decimal 1000 de $\frac{1}{13\times10^3}$ sin necesidad de calcular los decimales anteriores. En efecto, el decimal 1000 de $\frac{1}{13\times10^3}$ es el primer decimal de $1/13 = 0{,}076\ldots$ o sea 0.

Así es como el algoritmo para obtener tajadas de π funciona en base 4. Hay también una fórmula similar en base 16 (ver Ejercicio 2.11) y es posible hacerlo también en base 2. Si hubiera una serie similar para π con base 10 en el denominador, el algoritmo sería aplicable a calcular cualquier decimal de π sin conocer los anteriores, pero no se sabe si existe dicha fórmula.[16]

✍ **Ejercicio 2.7.** La probabilidad de que una aguja de largo $1/2$ intersecte al caer una de las rayas de un parquet formado por tablas paralelas de ancho 1 es $1/\pi$. La probabilidad de que dos números enteros elegidos al azar sean primos entre sí es de $6/\pi^2$. Busque una demostración de esto en la web sabiendo que se conocen por *Teorema de la aguja de Buffon* y *Teorema de Cesaro* respectivamente. Solución: *(para el Teorema de Buffon) en 1977 George Louis Leclerc (1707–1788) propuso esta relación entre la aguja que cae y π. Es fácil convencerse que la probabilidad de encontrar un entero en un intervalo cualquiera de largo $1/2$ es $1/2$. Del mismo modo, es $1/2$ la probabilidad de que una circunferencia de diámetro $1/2$, dibujada al azar en el paquet de tablas paralelas de ancho 1, intersecte una de sus rayas. Finalmente, si imaginamos que una circunferencia está formada por una infinidad de agujas giradas que comparten su punto medio, la probabilidad de intersección de una sola aguja es $1/2$ dividido por el largo de la circunferencia que es $\pi/2$, este cuociente da $1/\pi$.*[17]

✍ **Ejercicio 2.8.** Implemente numéricamente en una planilla de cálculo electrónica la aceleración de Aitken para el algoritmo basado en la serie de cuadrado de los recíprocos de Euler.

✍ **Ejercicio 2.9.** A partir de la definición que dimos para π, y sabiendo que las sucesiones aproximantes del algoritmo de Arquímides convergen a él, pruebe que el área de un círculo de radio r es πr^2.
Solución: *Siguiendo las mismas notaciones del algoritmo de duplicación de Arquímides, de la Figura 2.2, el* área *del polígonos inscritos (P_n) y circunscritos (Q_n) en un círculo de radio r se pueden expresar como:*

$$P_n = r^2 p_n \cos\theta_n, \qquad Q_n = r^2 q_n$$

como p_n y q_n convergen a π y $\cos(\theta_n)$ converge a 1, entonces P_n y Q_n convergen a πr^2.

✍ **Ejercicio 2.10.** Hay un algoritmo con una inspiración geométrica similar al algoritmo de duplicación de Arquímides y es debido a François Viète (1540-1603), solamente que se aproxima π a través del *área* de polígonos regulares inscritos en el círculo y no de su *perímetro*, como es el caso del algoritmo de Arquímides. La fórmula es:

$$\pi = 2\,\frac{2}{\sqrt{2}}\,\frac{2}{\sqrt{2+\sqrt{2}}}\,\frac{2}{\sqrt{2+\sqrt{2+\sqrt{2}}}}\cdots$$

[16] Para mayores detalles, pueden consultarse las referencias **[8]**, **[2]**.
[17] Véase `www.worsleyschool.net/science/files/buffon/buffonapplet.html`.

que se puede escribir en forma algorítmica como:

ALGORITMO DE VIÈTE

$$p_0 = 0, \qquad q_0 = 2$$
$$p_{n+1} = \sqrt{2 + p_n}, \qquad q_{n+1} = q_n \frac{2}{p_{n+1}}, \qquad n \geq 0.$$

donde q_n es la sucesión que aproxima π (notar que p_n y q_n ya no tienen el mismo significado geométrico que para el algoritmo de Arquímides). Estudie numéricamente el algoritmo de Viète del ejercicio precedente con la ayuda de una planilla de cálculo y estime el número de cifras exactas que se obtienen en cada iteración. Estudie numéricamente cómo mejora la velocidad de convergencia al acelerar el algoritmo usando el método de aceleración de Aitken.

✍ **Ejercicio 2.11.** La fórmula original para π usada por el algoritmo BPP es:

ALGORITMO BPP EN BASE 16

$$\pi = \sum_{n=1}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right).$$

que permite obtener tajadas de π en base 16 o *hexadecimal* comúnmente usado en informática. En grupos de 4 bits se pueden almacenar números del 0 al 15 y se usan los dígitos extendidos $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$ con letras adicionales del alfabeto que representan los números del 10 al 15 y se tiene el siguiente diccionario que traduce de binario a hexadecimal:

$$\begin{array}{cccccccccc} \underbrace{0000} & \underbrace{0001} & \underbrace{0010} & \underbrace{0011} & \underbrace{0100} & \underbrace{0101} & \underbrace{0110} & \underbrace{0111} & \to & \text{base 2} \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \to & \text{base 16} \end{array}$$

$$\begin{array}{cccccccccc} \underbrace{1000} & \underbrace{1001} & \underbrace{1010} & \underbrace{1011} & \underbrace{1100} & \underbrace{1101} & \underbrace{1110} & \underbrace{1111} & \to & \text{base 2} \\ 8 & 9 & \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} & \mathrm{F} & \to & \text{base 16} \end{array}$$

Usando el algoritmo de cambio de base para la base 16, reencuentre la expresión de π en binario a partir de su equivalente hexadecimal. Debería obtener algo como:

$$\begin{array}{ccccccccccccc} \pi & = & 11 & , & 0010 & 0100 & 0011 & 1111 & 0110 & 1010 & 1000 & 1000 & 1000\ldots \\ \pi & = & 3 & , & 2 & 4 & 3 & \mathrm{F} & 6 & \mathrm{A} & 8 & 8 & 8\ldots \end{array}$$

✍ **Ejercicio 2.12.** ¿Puede un decimal periódico en base 10 no serlo en otra base?

✍ **Ejercicio 2.13.** Investigue sobre una demostración de que π es irracional.
Solución: Un número irracional es el que no se puede expresar de la forma p/q con p y q enteros. Para probar que π es irracional, se puede probar que π^2 es irracional (¿por

qué?). Aunque no es fácil[18]*, es instructivo al menos conocer los pasos de una demostración. Supongamos entonces que* $\pi^2 = p/q$*. Considere ahora el polinomio de grado* $2n$ *definido por:*

$$P_n(x) = \frac{x^n(1-x)^n}{n!}$$

y pruebe que todas las derivadas $P^{(k)}$ *de* P_n *en* $x = 0$ *y* $x = 1$ *son enteras. Hágalo primero para* $x = 0$ *y deduzca por simetría que también es cierto para* $x = 1$*. Defina ahora el polinomio:*

$$Q_n(x) = q^n \left(\pi^{2n} P_n(x) - \pi^{2n-2} P_n''(x) + \pi^{2n-4} P_n^{(4)}(x) - \ldots + (-1)^n P_n^{(2n)}(x)\right)$$

y pruebe que

$$\frac{d}{dx}\left(Q_n' \sin(\pi x) - \pi Q_n(x)\cos(\pi x)\right) = \pi^2 p^n \sin(\pi x) P_n(x).$$

Integrando entre 0 *y* 1 *la expresión anterior, deduzca que la sucesión*

$$a_n = \pi^2 p^n \int_0^1 \sin(\pi x) P_n(x) dx$$

es entera. Pruebe además que $0 < a_n < \frac{\pi p^n}{n!}$ *y obtenga una contradicción para* n *suficientemente grande. (Nota:* $n!$ *crece más rápido que* p^n*).*

✍ **Ejercicio 2.14.** Investigue cómo Euler descubrió la fórmula

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + ... = \frac{\pi}{6}.$$

Solución: pruebe primero que si se tiene el polinomio

$$p(x) = 1 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

y sus raíces son $\lambda_1, \lambda_2, \ldots, \lambda_n$*, entonces*

$$a_1 = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \ldots + \frac{1}{\lambda_n}.$$

Extrapolando esta propiedad a un desarrollo de Taylor[19] *se obtiene la fórmula. En efecto, del desarrollo de Taylor de la función seno:*

$$\sin y = 1 - \frac{y^3}{3!} + \frac{y^5}{5!} - \ldots + (-1)^n \frac{y^{2n+1}}{(2n+1)!} + \ldots$$

se obtiene tomando $x = y^2$ *que*

$$\frac{\sin\sqrt{x}}{\sqrt{x}} = 1 - \frac{x}{3!} + \frac{x^2}{5!} - \ldots + (-1)^n \frac{x^n}{(2n+1)!} + \ldots$$

cuyas raíces son $\pi^2, 4\pi^2, 9\pi^2, \ldots$*. Suponiendo que la propiedad que se probó sobre la suma de los recíprocos de las raíces para polinomios se puede extender a esta serie, se obtiene el resultado de Euler.*

[18] Es mucho más fácil probar que $\sqrt{2}$ es irracional también por contradicción.
[19] Véase Capítulo 3.

✍ **Ejercicio 2.15.** Un *algoritmo recursivo* es aquel que en una iteración utiliza el mismo algoritmo. Por ejemplo, para calcular $n!$ podríamos decir que se calcula como n por $(n-1)!$ y a su vez $(n-1)!$ se calcula como $(n-1)$ por $(n-2)!$, etcétera, de modo que el algoritmo sería:

> ALGORITMO RECURSIVO PARA CALCULAR $n!$
> - Si $n = 1$ entonces $1! = 1$
> - Si $n \geq 2$ entonces $n! = n(n-1)!$

Escriba un algoritmo recursivo para aproximar π a partir de la *fracción continua* siguiente:

$$\frac{\pi}{4} = \cfrac{1}{1+\cfrac{1^2}{2+\cfrac{3^2}{2+\cfrac{5^2}{2+\cfrac{7^2}{2+\cfrac{9^2}{\ddots}}}}}}$$

Solución: se tiene que

$$\pi = \frac{4}{1+\alpha}$$

donde α es límite de la sucesión α_n definida por recurrencia de la siguiente manera:

$$\alpha_n = \frac{(2n+1)^2}{2+\alpha_{n+1}} \quad n = 0, 1, \ldots.$$

Para detener las iteraciones y encontrar un valor aproximado, hay que definir un valor final, por ejemplo, $\alpha_{10} = 0$ luego de $n = 9$ iteraciones.

✍ **Ejercicio 2.16.** La constante π está conectada con otras constantes famosas. A partir del límite:

$$\sqrt{2\pi} = \lim_{n\to\infty} \frac{e^n n!}{n^n \sqrt{n}}$$

aproxime π con una calculadora usando un valor aproximado de $e \approx 2,718$.

✍ **Ejercicio 2.17.** Investigue sobre el significado geométrico de la fórmula de Euler:

$$e^{i\pi} = -1$$

en el plano complejo, donde $i = \sqrt{-1}$ es la unidad imaginaria.

Solución: $e^{i\theta}$ representa una rotación en un ángulo de θ en el sentido anti-horario. Si denotamos un número complejo $a + bi$ por el par (a, b) en el plano complejo, la fórmula expresa entonces que una rotación del número real 1 considerado como el número complejo $(1, 0)$ en un ángulo de π radianes (180 grados), nos lleva al número complejo $(-1, 0)$ que no es más que el real -1.

✍ **Ejercicio 2.18.** ¿Cómo explicaría a un estudiante que el área de un círculo se expresa usualmente en centímetros cuadrados como cualquier otra área, y que no es obligatorio aunque podría ser interesante introducir centímetros "circulares" o algo así? *Solución: como parte de la discusión, se puede acotar que los centímetros cuadrados corresponden a la unidad de área de las coordenadas cartesianas que son usualmente utilizadas. Pero existen los centimetros "circulares" y corresponden a la unidad de área de las llamadas coordenadas polares.*

Capítulo 3: Ceros, Interpolación e Integración Numérica

> *"El análisis numérico es el estudio de algoritmos para los problemas de la matemática del continuo."* L. N. Trefethen, *(The definition of Numerical Analysis, 1992).*

Ya se analizó en el capítulo anterior la importancia de los algoritmos aproximantes para aproximar un número real. En este capítulo veremos tres temas que tienen que ver también con algoritmos aproximantes, pero esta vez se aplican a funciones de variable real. Los tres problemas son: a) aproximar los ceros o raíces de una función, b) aproximar la gráfica de una función y c) aproximar el área bajo la curva de una función, considerados el *abc* del análisis numérico.

3.1 Aproximando los ceros de una función

Aproximar los ceros de una función es una cuestión difícil. Por ejemplo, en el caso de las raíces de polinomios, es fácil calcular las raíces de un polinomio de grado dos, y existen fórmulas analíticas finitas para hallar las raíces de polinomios de grado 3 y 4, pero éstas fórmulas involucran radicales que hay que aproximar. Además, para polinomios de grado 5 y superior ya no existen fórmulas analíticas finitas y las raíces sólo pueden aproximarse [1] [2].

3.1.1 Iteraciones de punto fijo

Imagine que tiene usted un mapamundi y una copia exacta más pequeña de él. Si superpone el mapa más pequeño sobre el mapa más grande y se ven los dos mapas al trasluz, puede verificar que hay un (y sólo un) punto geográfico sobre el mapa pequeño que coincide exactamente con el mismo punto geográfico del mapa grande. Este ejemplo ilustra una propiedad que se conoce como la existencia y unicidad de un *punto fijo* para una cierta contracción.

---

[1] El *teorema fundamental del álgebra* establece que un polinomio de grado n a coeficientes reales o complejos tiene n raíces reales o complejas, resultado demostrado por primera vez por el matemático suizo Jean Robert Argand (1768-1822) en 1806. Aquí nos interesa aproximar las raíces reales de polinomios a coeficientes reales.

[2] El teorema de Abel-Ruffini publicado en 1824 establece que no pueden obtenerse las raíces de un polinomio de grado superior o igual a cinco aplicando únicamente un número finito de sumas, restas, multiplicaciones, divisiones y extracción de raíces a los coeficientes de la ecuación. Esto hace parte de la llamada *teoría de Galois*. Véase la Monografía *Álgebra abstracta* de esta misma colección, cf. [**18**].

Consideremos una *función contractante* de $\mathbb{R}$ en $\mathbb{R}$, es decir, una que transforma intervalos de largo a en intervalos de largo menor: $\alpha\ a$, donde $\alpha < 1$ es una *constante de contracción*. Esto es[3]:

$$|f(x) - f(y)| \leq \alpha|x - y|, \qquad \text{con } 0 < \alpha < 1.$$

Entonces, el *Teorema del punto fijo de Banach* en honor al matemático polaco Stefan Banach (1892-1945), dice que existe[4] un punto fijo x^* para f, esto es:

$$\text{Existe } x^* \in \mathbb{R} \text{ tal que } f(x^*) = x^*.$$

Del punto de vista del análisis numérico, es muy interesante la siguiente prueba de este Teorema, que utiliza un algoritmo aproximante para encontrar el punto fijo x^*, y que constituye, de hecho, un algorimo para encontrar los ceros de la función:

$$f(x) - x = 0.$$

En efecto, se demuestra que la siguiente sucesión llamada *iteración de Picard*:

Iteraciones de Picard		
Etapa 0**:**	x_0 dado,	
Etapa n**:**	$x_{n+1} = f(x_n),$	$n \geq 0,$

converge al punto fijo. Esto es:

$$x_n \to x^*.$$

Por ejemplo, supongamos que queremos resolver:

$$2\cos(x) - 3x = 0$$

lo que es equivalente a resolver:

$$x^* = \frac{2}{3}\cos(x^*).$$

Como $\frac{2}{3}\cos(x)$ es contractante de constante $\alpha = 2/3$, haciendo la iteración de punto fijo obtenemos la sucesión del Cuadro 3.1. En una calculadora científica esto es muy fácil de hacer. Simplemente comience con el valor $0,4$ y luego presione repetidamente la combinación de las cuatro teclas

$$\boxed{\cos}, \quad \boxed{\div 3}, \quad \boxed{\times 2}, \quad \boxed{=}.$$

Debería obtener los valores de la segunda columna del Cuadro 3.1 hasta ir convergiendo al valor

$$x^* = 0{,}56356920422552\ldots$$

que a partir de un momento parece invariante ante la combinación de cuatro teclas debido al límite de precisión de su calculadora. Pero, aunque calcularamos con más y más precisión, nunca llegaríamos al valor exacto de x^*, ya que resulta ser un número

[3]Notemos que toda función contractante es también continua, pues de la definición de función contractante se ve que si x tiende a y, $f(x)$ tiende a $f(y)$.

[4]Único, pero esto no nos interesará ahora.

Término	aproximación	error $\lvert x_n - x^*\rvert$
x_0	0,40000000	$1,64 \times 10^{-1}$
x_1	0,61404066	$5,05 \times 10^{-2}$
x_2	0,54488438	$1,87 \times 10^{-2}$
x_3	0,57012482	$6,56 \times 10^{-3}$
x_4	0,56122241	$2,35 \times 10^{-3}$

CUADRO 3.1. Iteraciones de punto fijo para resolver $2\cos(x)-3x = 0$ con $x_0 = 0{,}4$ indicando el error de aproximación.

irracional. Sin embargo, en términos prácticos, con ayuda de un computador *podemos aproximar numéricamente la solución y con precisión arbitraria* y eso es magnífico. Es lo mismo que pasaba al tratar de aproximar π en el Capítulo 2.

Para ver cómo se comporta la sucesión x_n antes definida, se recurre a un truco gráfico. Dibuje en un gráfico una función contractante f y la recta diagonal $y = x$. Partiendo de x_0, se encuentra verticalmente $f(x_0)$ y luego horizontalmente sobre la diagonal $y = x_1 = f(x_0)$ lo que se lleva de nuevo verticalmente a la función $f(x_1)$ que horizontalmente corresponde en la diagonal a $y = x_2 = f(x_1)$, etcétera (ver Figura 3.1, izquierda).

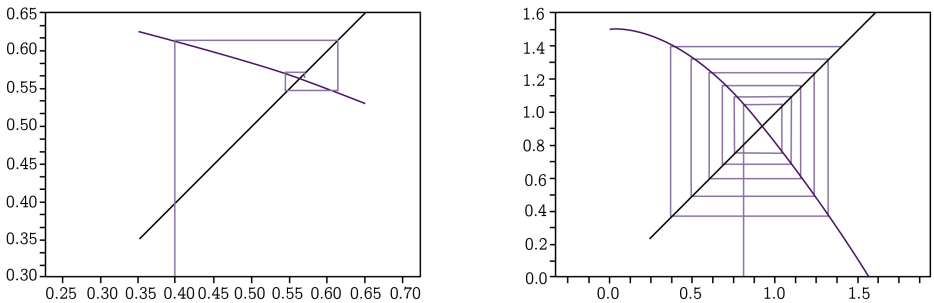


FIGURA 3.1. Izquierda: iteraciones de punto fijo para resolver la ecuación $2\cos(x) - 3x = 0$ con $x_0 = 0,4$. Derecha: iteraciones de punto fijo del Ejercicio 3.2.

Para demostrar que x_n converge a x^* se prueba que x_n es una sucesión de Cauchy, por lo tanto convergente a un cierto límite real ℓ. Si se toma límite en la recurrencia $x_{n+1} = f(x_n)$, como f es continua se obtiene $\ell = f(\ell)$, esto es, ¡el límite resulta ser exactamente un punto fijo!

Para ver que x_n es de Cauchy, primero se observa que

$$|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \leq \alpha |x_n - x_{n-1}|$$

y repitiendo el argumento n veces se obtiene

$$|x_{n+1} - x_n| \leq \alpha^n |x_1 - x_0|.$$

Luego se acota $|x_{n+k} - x_n|$ sumando y restando los términos intermedios, para llegar a:

$$\begin{aligned} |x_{n+k} - x_n| &\leq |x_{n+k} - x_{n+k-1}| + |x_{n+k-1} - x_{n+k-2}| + \ldots + |x_{n+1} - x_n| \\ &= |x_1 - x_0| \left(\sum_{i=0}^{n+k-1} \alpha^i - \sum_{i=0}^{n-1} \alpha^i \right) \\ &= |x_1 - x_0| \left(\frac{1-\alpha^{n+k}}{1-\alpha} - \frac{1-\alpha^n}{1-\alpha} \right) \end{aligned}$$

que tiende a cero cuando n y $n+k$ tienden a infinito. Hemos usado la conocida suma geométrica $\sum_{i=0}^{n} r^i = (1 - r^{n+1})/(1-r)$ para $r \neq 1$ cuya prueba por inducción se deja de ejercicio al lector.

✍ **Ejercicio 3.1.** Tomando $k \to \infty$ en la expresión anterior, pruebe que el error que se comete en las iteraciones de punto fijo está acotado como:

$$|x_n - x^*| \leq C\alpha^n.$$

¿Cuántas cifras significativas se obtienen en cada iteración?
Solución:

$$-\log \frac{C\alpha^n}{5x^*} = n \log \frac{1}{\alpha} + \log \frac{C}{5x^*}$$

esto es las cifras significativas crecen linealmente como $n \log \frac{1}{\alpha}$.

✍ **Ejercicio 3.2.** Observe qué pasa con la iteración de punto fijo si se intenta resolver: $3\cos(x) - 2x = 0$ con $x_0 = 0{,}8$ (ver Figura 3.1, derecha). Pero, ¿existe alguna solución?, ¿por qué las iteraciónes divergen?
Solución: las iteraciones divergen pues la función $f(x) = \frac{3}{2}\cos x$ *no es contactante (la constante es* $\alpha = \frac{3}{2} > 1$*).*

✍ **Ejercicio 3.3.** Encuentre los ceros del polinomio $p(x) = x^2 - 2x - 3$ usando el método de punto fijo. Pruebe usando los dos despejes posibles $x = \frac{x^2-3}{2}$ ó $x = \sqrt{2x+3}$.

✍ **Ejercicio 3.4.** Haga la prueba por indución de que $\sum_{i=0}^{n} r^i = (1 - r^{n+1})/(1-r)$ para $r \neq 1$.

✍ **Ejercicio 3.5.** Pruebe que una función $f : \mathbb{R} \to \mathbb{R}$ contractante de constante $\alpha < 1$ que además es derivable, cumple que $|f'(x)| < 1$, $\forall x \in \mathbb{R}$.

✍ **Ejercicio 3.6.** Encuentre un función derivable tal que $|f'(x)| < 1$, $\forall x \in \mathbb{R}$ pero que no sea contractante y que de hecho no tenga puntos fijos. ¿Se contradice esto con el resultado del ejercicio anterior?

3.1.2 Bisección

Busque la palabra *recurrencia* en un diccionario de la lengua española. Si abre el diccionario más o menos en la mitad encontrará la letra J. Sabe que la palabra que busca empieza con R, así es que descarta la primera mitad del diccionario y centra su búsqueda en la segunda mitad, y vuelve a separar las hojas de la J a la Z. Si abre esta vez en la letra S, sabe que la palabra que busca está entre la J y la S y selecciona esta vez esa parte para buscar, y así sucesivamente.

El algoritmo de búsqueda que está utilizando recibe el nombre de *búsqueda por bisección* y lo mismo puede hacer para buscar un nombre en una guía telefónica o un examen de un alumno en una pila ordenada alfabéticamente.

Este método de búsqueda por bisección es también usado para encontrar los ceros de una función continua. La única condición es que la función cambie de signo en el intervalo en que buscamos. Planteemos esto ahora más precisamente en términos matemáticos.

Sea f una función continua de $[a,b]$ en $\mathbb{R}$ y supongamos que f cambia de signo en dicho intervalo, esto es, por continuidad, ella se anula[5] al menos una vez en $[a,b]$. Para fijar ideas, puede pensar que f cambia de signo una sola vez, de modo que hay un único cero $z \in [a,b]$, pero en realidad el método funciona aunque haya más de un cero en $[a,b]$. Encajonemos z de la siguiente manera:

Algortimo de bisección o encajonamientos sucesivos

Etapa 0**:** $a_0 = a, \quad b_0 = b,$

Etapa n**:** $c_n = \dfrac{a_n + b_n}{2}$

Si $f(a_n) = 0$ ó $f(c_n) = 0$ ya hemos encontrado el cero.

Si $f(a_n)f(c_n) < 0$ entonces $a_{n+1} = a_n$, $b_{n+1} = c_n$,

en caso contrario $a_{n+1} = c_n$, $b_{n+1} = b_n$.

Esto es, se busca el cero z de la función en los subintervalos donde f cambia su signo. En cada iteración el largo del intervalo de búsqueda se divide por la mitad, de modo que el error de aproximación también, esto es:

$$|z - a_n| \leq |b_n - a_n| \leq \frac{(b-a)}{2^n}.$$

✍ **Ejercicio 3.7.** Entregue una intrepretación de la cota anterior en términos de cifras significativas.

Solución: en la iteración n se obtienen del orden de $n \log 2$ *cifras significativas. Esto es, el número de cifras significativas crece de manera lineal.*

En la Figura 3.2 se muestra el algoritmo de bisección aplicado para encontrar una ❒✍

[5] Teorema del valor intermedio.

solución de la ecuación $\ln(z) - \sin(z) = 0$ en el intervalo $[1, 3]$. El error de aproximación se registra en el Cuadro 3.2. El resultado obtenido después de varias iteraciones es:

$$z = 2{,}21910714891375\ldots$$

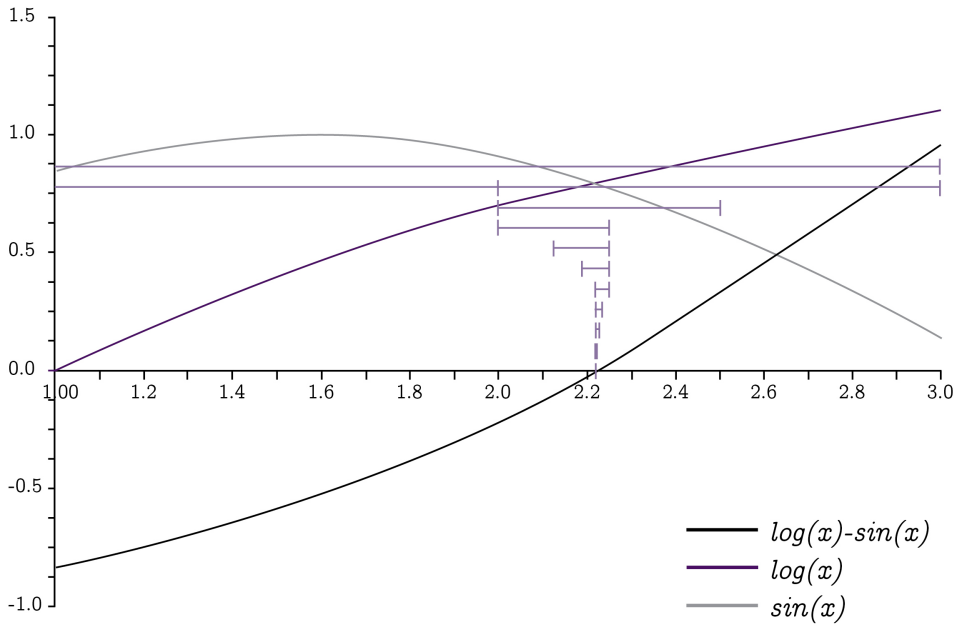


FIGURA 3.2. Algoritmo de bisección para encontrar una solución de $\ln(x) = \sin(x)$ en el intervalo $[1, 3]$. Se indican de arriba hacia abajo los sucesivos intervalos donde se busca la solución.

✍ **Ejercicio 3.8.** Si algún día tiene la dicha de visitar París, puede encontrar en la sección de impresionistas del tercer piso del Museo de Orsay el original de la famosa pintura "Bal du Moulin de la Galette" de Renoir. Hay un gentío bailando feliz bajo la luz tamizada por el follaje, pero se dice que hay un único personaje que ¡está mirándo fijo al espectador! Si el cuadro real tiene dimensiones de $131\,cm$ de alto y $175\,cm$ de largo y cada rostro distinguible ocupa a lo más $10\,cm^2$, ¿en cuántas iteraciones es seguro que se aislaría el rostro del enigmático personaje si se buscara por bisección, dividiendo sucesivamente el cuadro en mitades iguales? Indicación: alterne cortes verticales y horizontales del cuadro y piense que en cada iteración se descarta la mitad del área de búsqueda.

n	$[a_n, b_n]$	error $\lvert z - a_n \rvert$
0	$[1, 3]$	$1{,}22 \times 10^{0}$
1	$[2, 3]$	$2,19 \times 10^{-1}$
2	$[2, 2{,}5]$	$2,19 \times 10^{-1}$
3	$[2, 2{,}25]$	$2,19 \times 10^{-1}$
4	$[2{,}125, 2{,}25]$	$9,41 \times 10^{-2}$
5	$[2{,}1875, 2{,}25]$	$3,16 \times 10^{-2}$

CUADRO 3.2. Iteraciones del algoritmo de bisección para resolver $\ln(x) = \sin(x)$ indicando el error de aproximación.

3.1.3 Método de Newton-Raphson

Este método es uno de los más populares para encontrar los ceros de f, pero requiere del conocimiento de la derivada de f. La regla principal del algoritmo consiste dado x_n en encontrar la recta tangente a f en x_n y luego definir x_{n+1} como el punto de intersección entre esa recta tangente y el eje x (ver Figura 3.3). Esto es, x_{n+1} se despeja de:

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n)$$

y se obtiene:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

siempre que la derivada en x_n no se anule. Esto es, el *método de Newton-Raphson* queda así:

MÉTODO DE NEWTON-RAPHSON

Etapa 0: x_0 dado,

Etapa n: $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}, \quad n \geq 0.$

Es posible demostrar que de converger, el método de Newton tiene una convergencia cuadrática, esto es, para n grande, en cada iteración el error $e_n = |x_n - z|$ (suponemos $e_n < 1$) decrece como:

$$e_{n+1} \leq C e_n^2.$$

Como $-\log(e_{n+1}) \geq -2\log(e_n) - \log C$, el número de cifras exactas se duplica en cada iteración (ver Capítulo 2) para n grande. El método lleva también el nombre de Joseph Raphson (1648 - 1715) matemático británico contemporáneo a Isaac Newton (1643-1727).

A modo de ejemplo, apliquemos el algoritmo de Newton-Raphson para encontrar una raíz de $x^3 - x - 3$ partiendo de $x_0 = 1$. Se encuentra la sucesión:

$$x_0 = 1, \quad x_{n+1} = x_n - \frac{x_n^3 - x_n - 3}{3x_n^2 - 1}, \quad n \geq 0.$$

❒✍ La convergencia se ilustra en la Figura 3.3 y los valores que se obtienen para x_n en cada iteración se muestran en el Cuadro 3.3 así como el error de aproximación $|x_n - z|$ donde

$$z = 1{,}67169988165716\ldots$$

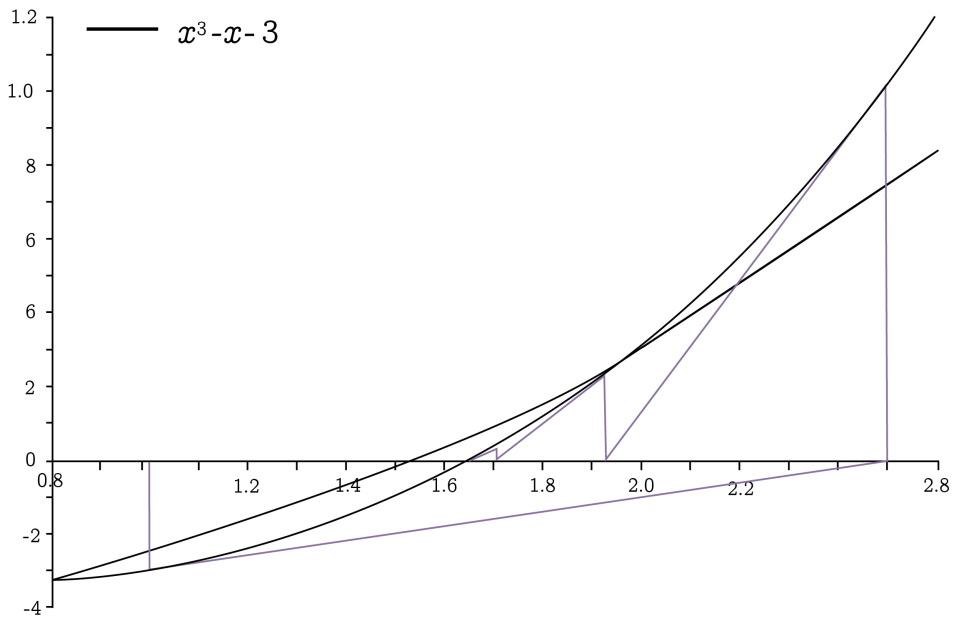


FIGURA 3.3. Algoritmo de Newton-Raphson para encontrar una raíz de $x^3 - x - 3$ partiendo de $x_0 = 1$. Se indican las rectas tangentes que se usan en cada iteración.

✍ **Ejercicio 3.9.** El siguiente es el llamado *método de la secante* para encontrar los ceros de una función:

n	x_n	error $\lvert x_n - z\rvert$
0	1,00000000	$0{,}67 \times 10^{0}$
1	2,50000000	$8,28 \times 10^{-1}$
2	1,92957747	$2,58 \times 10^{-1}$
3	1,70786640	$3,62 \times 10^{-2}$
4	1,67255847	$8,59 \times 10^{-4}$
5	1,67170038	$5,00 \times 10^{-7}$

CUADRO 3.3. Iteraciones del algoritmo de Newton-Raphson para encontrar una raíz de $x^3 - x - 3$ partiendo de $x_0 = 1$. Se indica además el error de aproximación en cada iteración.

MÉTODO DE LA SECANTE

Etapa 0**:** x_0, x_1 dados,

Etapa n**:**
$$x_{n+1} = x_n - \frac{f(x_n)}{s_n},$$
$$s_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}, \quad n \geq 1.$$

Compare con el método de Newton-Raphson. ¿Cuál es la ventaja de este método? ¿Por qué se llama "método de la secante"?

✍ **Ejercicio 3.10.** Considere el método de Newton Raphson para aproximar los ceros de la función

$$f(x) = \begin{cases} \sqrt{x} & x > 0 \\ -\sqrt{-x} & x < 0. \end{cases}$$

¿Qué ocurre con el algoritmo? Haga una representación gráfica de las iteraciones trazando las pendientes para entender mejor qué ocurre. Si ahora se utiliza el método de la secante, ¿ocurre lo mismo?

✍ **Ejercicio 3.11.** El *número aúreo* ϕ satisface

$$\phi = 1 + \frac{1}{\phi}$$

de donde se puede obtener que:

$$\phi = \frac{1 + \sqrt{5}}{2}.$$

Escriba en forma de algoritmo recursivo para obtener $\sqrt{5}$ a partir de la *fracción continua* siguiente:

$$\phi = \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{\ddots}}}}}}$$

¿Se le ocurre ahora cómo aproximar ϕ usando el método de Newton-Raphson? ¿Qué método es mejor?

✍ **Ejercicio 3.12.** Utilice el método de Newton-Raphson para aproximar las dos raíces del polinomio de cuarto grado $p(x) = x^4 - 5x^3 + 4x^2 - 3x + 2$. Nota: una vez encontrada una raíz r puede dividir el polinomio por $(x - r)$ para continuar buscando la otra raíz.[6]

Solución: las raíces son 0,8023068018257805 *y* 4,1888470295364675.

✍ **Ejercicio 3.13.** ¿Existen métodos para aproximar las raíces complejas de un polinomio? Investigue sobre el *método de Bairstow*[7] y el *método de Laguerre*.

3.1.4 Ejemplo de aplicación: cálculo de la raíz cuadrada

Utilicemos los métodos para encontrar ceros vistos anteriormente para encontar algoritmos que aproximen el valor de $\sqrt{M}$. En efecto, se trata de encontrar un cero (positivo) de la ecuación:

$$f(x) = x^2 - M.$$

- Método de bisección o de encajonamientos sucesivos para aproximar $\sqrt{M}$. Es el método más usado, pero veremos que no es el más eficiente. El método queda así una vez que lo aplicamos a este caso particular:

Algortimo de bisección para calcular $\sqrt{M}$	
Etapa 0**:**	$a_0 = 1, \quad b_0 = M,$
Etapa n**:**	$c_n = \dfrac{a_n + b_n}{2}$
	Si $a_n^2 = M$ ó $c_n^2 = M$ ya hemos encontrado la raíz.
	Si $c_n^2 > M$ entonces $a_{n+1} = a_n$, $b_{n+1} = c_n$
	en caso contrario $a_{n+1} = c_n$, $b_{n+1} = b_n$.

  Como vimos antes, las sucesiones a_n y b_n encajonan la raíz de M, esto es $a_n \leq \sqrt{M} \leq b_n$ y convergen a $\sqrt{M}$ de modo tal que el error se divide a la mitad en cada iteración.

[6] Véase el diagrama interactivo que ilustra el método de Newton-Raphson en `www.math.umn.edu/~garrett/qy/Newton.html`.

[7] Véase, por ejemplo, la página para calcular raíces de forma interactiva usando el método de Newton-Raphson y Bairstow en `jj.gbtopia.com/g_mathapplets.html`.

- Método de Newton-Raphson para aproximar $\sqrt{M}$. Observemos que $f'(x) = 2x$ de modo que

$$x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - M}{2x_n} = \frac{1}{2}\left(x_n + \frac{M}{x_n}\right).$$

Con esto, el método queda:

> MÉTODO DE NEWTON-RAPHSON PARA APROXIMAR $\sqrt{M}$
>
> **Etapa** 0**:** $x_0 = 1,$
>
> **Etapa** n**:** $x_{n+1} = \frac{1}{2}\left(x_n + \frac{M}{x_n}\right), \quad n \geq 0.$

Como vimos antes, la sucesión x_n converge a $\sqrt{M}$ de manera cuadrática, esto es, el número de cifras exactas con que se aproxima la raíz se duplica en cada iteración. Este método se utiliza desde la antiguedad y se denomina también el *método babilónico* o *método de Herón*, matemático e ingeniero griego que vivió del 70 al 10 a.C. en Alejandría.

- Método de la secante para aproximar $\sqrt{M}$. Observemos que en este caso la secante se calcula como:

$$s_n = \frac{x_n^2 - M - (x_{n-1}^2 - M)}{x_n - x_{n-1}} = x_n + x_{n-1}$$

de modo que

$$x_n - \frac{f(x_n)}{s_n} = x_n - \frac{x_n^2 - M}{x_n + x_{n-1}} = \frac{x_n x_{n-1} + M}{x_n + x_{n-1}}.$$

Con esto, el método de la secante queda:

> MÉTODO DE LA SECANTE PARA APROXIMAR $\sqrt{M}$
>
> **Etapa** 0**:** $x_0 = 1, x_1 = M$
>
> **Etapa** n**:** $x_{n+1} = \frac{x_n x_{n-1} + M}{x_n + x_{n-1}}, \quad n \geq 1.$

En este caso también se puede demostrar que la sucesión x_n converge a $\sqrt{M}$ de manera cuadrática, como en el caso del método de Newton-Raphson.

Las iteraciones y convergencia de los tres métodos se comparan en el Cuadro 3.4 donde se aproxima el valor de $\sqrt{2}$ ($M = 2$). Para obtener el valor correcto redondeado a seis decimales, el método de bisección o de encajonamiento toma 20 iteraciones, el método de Newton-Raphson, 4 iteraciones y el método de la secante, 5 iteraciones. ❒✍

n	bisección	Newton-Raphson	secante
0	1	1	1 y 2
1	1,500000	1,500000	1,333333
2	1,250000	1,416667	1,428571
3	1,375000	$\rightarrow$ 1,414214	1,413793
4	1,437500	1,414214	$\rightarrow$ 1,414214
5	1,406250	1,414214	1,414214
10	1,415039	1,414214	1,414214
15	$\rightarrow$ 1,414214	1,414214	1,414214

CUADRO 3.4. Aproximación de $\sqrt{2}$ usando los algoritmos de bisección, Newton-Raphson y secante en función del número de iteraciones de cada algoritmo. Se indica con una flecha cuando el algoritmo a logrado exactitud al redondear a seis decimales.

✍ **Ejercicio 3.14.** El siguiente es el llamado *algoritmo de Bakhshali* para aproximar $\sqrt{M}$, encontrado en 1881 en Pakistán en un antiguo manuscrito matemático:

ALGORITMO DE BAKHSHALI PARA APROXIMAR $\sqrt{M}$

- Escoja N
- Calcule $d = N^2 - M$
- Calcule $P = \frac{d}{2N}$
- Aproxime

$$\sqrt{M} \approx (N - P) - \frac{P^2}{2(N - P)}.$$

El número N podría ser cualquier número, pero se recomienda en el manuscrito comenzar con N algún natural tal que N^2 sea cercano a M de manera que la distancia d sea pequeña. Pruebe que este algoritmo no es más que aplicar dos veces seguidas el algoritmo de Newton-Raphson (o algoritmo babilonio o de Herón) comenzando con $x_0 = N$.

Solución: si $x_0 = N$, entonces iterando dos veces el algoritmo de Newton-Raphson:

$$x_1 = x_0 - \frac{x_0^2 - M}{2x_0} = N - \frac{N^2 - M}{2N} = N - \frac{d}{2N} = N - P$$

$$x_2 = x_1 - \frac{x_1^2 - M}{2x_1} = (N - P) - \frac{(N - P)^2 - M}{2(N - P)}$$

pero $(N - P)^2 - M = N^2 - 2NP + P^2 - M = d - d + P^2 = P^2$, de donde se obtiene que $\sqrt{M} \approx x_2$.

✍ **Ejercicio 3.15.** Investigue otros métodos para aproximar la raíz cuadrada.

✍ **Ejercicio 3.16.** Usando las mismas ideas del párrafo anterior, encuentre tres algoritmos distintos para aproximar $\sqrt[3]{M}$ y compárelos numéricamente al aproximar $\sqrt[3]{2}$.

✍ **Ejercicio 3.17.** ¿Se le ocurre algún algoritmo del tipo punto fijo para aproximar $\sqrt{M}$?, ¿obtiene un nuevo algoritmo?

3.2 Aproximando una función por un polinomio

Hay diversas maneras de aproximar una función por un polinomio: podemos intentar aproximar la función de modo que coincida ella y sus derivadas con las del polinomio en un punto; o de modo que la función y el polinomio coincidan en varios puntos; podemos buscar aproximarla de modo que la función y el polinomio tengan áreas bajo la curva cercanas, etc. En toda esta sección, supondremos que la función f que aproximamos es $n+1$ veces continuamente diferenciable, para un cierto n dado.

3.2.1 Polinomios de Taylor

Sin duda, los *polinomios de Taylor* son muy utilizados cuando queremos aproximar una función por un polinomio[8]. Dada una función f, imaginemos un polinomio que coincide con f en el origen y cuyas derivadas también coinciden con las derivadas de f en el origen. Es el *polinomio de Taylor de f en torno al origen.*[9]

Las *series de Taylor* se obtienen al considerar un número infinito de términos en los polinomios y fueron introducidas en toda su generalidad en 1715 por Brook Taylor (1685-1731) matemático británico, aunque muchas series, como las de las funciones trigonométricas, se conocían antes.

Si f es una función n veces derivable en el intervalo $[-a,a]$, $a>0$ (con derivadas laterales en los bordes del intervalo), el polinomio de Taylor de orden n de f en torno a 0 está dado por:

Polinomio de Taylor

$$T_n(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \ldots + \frac{x^n}{n!}f^{(n)}(0).$$

En efecto, es fácil verificar que las derivadas de T_n en cero coinciden con las derivadas de f en cero.

Para estimar el error que se comete al aproximar f por su polinomio de Taylor T_n hay al menos dos opciones extremas: una es considerar el error que se comete en un punto en particular y la otra es considerar el máximo error que se comete al recorrer todos los puntos del intervalo:

$$|f(x)-T_n(x)| \quad \text{error puntual}, \qquad \max_{x\in[-a,a]} |f(x)-T_n(x)| \quad \text{error uniforme}.$$

[8] Véase la Monografía *Cálculo Integral y Series de Potencias* en esta misma colección cf. [**1**].

[9] Las series de Taylor en torno al origen se conocen también como *series de Maclaurin*.

Es ideal que el error uniforme sea pequeño, lo que asegura también que el error puntual sea pequeño en todo el intervalo con la misma cota de error, por eso se dice *uniforme*. Al *error o distancia uniforme* lo anotaremos con dos barras así:

$$||f - T_n|| = \max_{x\in[-a,a]} |f(x) - T_n(x)|.$$

Esta misma idea nos sirve para cuantificar qué tan grande es una función, en efecto:

$$||f|| = \max_{x\in[-a,a]} |f(x)|$$

mide la distancia de f a 0 y se llama *norma uniforme o norma infinito*[10] de f.

Es posible mostrar que cuando f es $(n+1)$ veces continuamente derivable:

$$f(x) = T_n(x) + R(x), \quad R(x) \equiv \frac{x^{n+1}}{(n+1)!} f^{(n+1)}(y),$$

donde y es algún punto en el intervalo $[-a, a]$ (más precisamente entre x_0 y x por lo que y depende de x). La expresión $R(x)$ se denomina *resto de Taylor*[11]. Tomando el máximo con $x \in [-a, a]$ se obtiene que:

$$||f - T_n|| \leq \frac{a^{(n+1)}}{(n+1)!} ||f^{(n+1)}||$$

esto es, el máximo error de aproximación en el intervalo depende uniformemente del máximo valor que toma la derivada $(n+1)$ de f en dicho intervalo.

Hay funciones como $\text{sen}(x)$ o $\cos(x)$ que son tantas veces derivables como se quiera, y cuyas derivadas siempre están acotadas entre -1 y 1, entonces como:

$$\lim_{n\to\infty} \frac{a^{(n+1)}}{(n+1)!} = 0,$$

es claro que la cota del error converge a cero y, por lo tanto, sus polinomios de Taylor convergen uniformemente a ellas. Véase, por ejemplo, la Figura 3.4, donde se aproxima la función $\cos(x)$ en torno a cero.

Pero hay que tener cuidado, ya que hay otras funciones que se resisten a ser aproximadas. Para éstas, el lado derecho de la cota del error anterior no tiende a cero. Una muy simple es $f(x) = \exp\left(-\frac{1}{x^2}\right)$ que tiene todas sus derivadas nulas en $x = 0$, por lo que $T_n = 0$ por grande que sea n. Esto quiere decir que para esta función $||f^{(n+1)}||$ en cualquier intervalo que contenga al origen tiende a infinito al menos como la sucesión $\frac{(n+1)!}{a^{n+1}}$.

✍ **Ejercicio 3.18.** Para la función $f(x) = \exp\left(-\frac{1}{x^2}\right)$, calcule sus primeras derivadas y evalúelas en $x = 1$. Observe que en la derivada n hay términos que crecen en valor absoluto como $(n+1)!$.

---

[10] Se llama norma infinito porque resulta como límite cuando $p \to \infty$ de las integrales $||f||_p = \frac{1}{p}\int_{-a}^{a} |f(x)|^p\, dx$, llamadas normas L^p.

[11] Véase la Monografía *Cálculo Integral y Series* en esta misma colección cf. **[1]**.

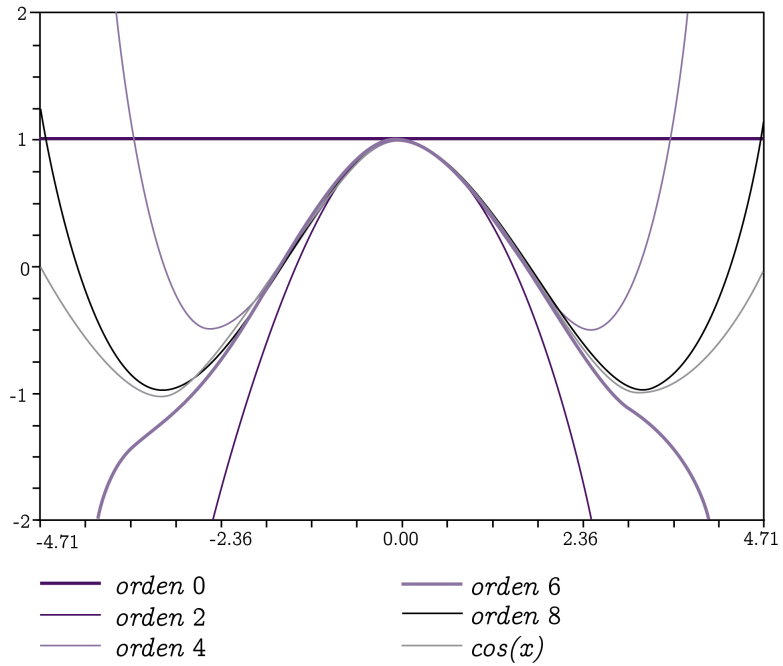


FIGURA 3.4. Aproximaciones por polinomios de Taylor de la función $\cos(x)$ en torno a cero. $T_0 = 1$, $T_2 = 1 - x^2/2!$, $T_4 = 1 - x^2/2! + x^4/4!$, etcétera.

✍ **Ejercicio 3.19.** Investigue sobre tres expresiones diferentes para el resto de Taylor $R(x)$, ¿cuál se usó aquí en el texto? *Solución: Resto de Lagrange:*

$$R(x) = \frac{f^{(n+1)}(y)}{(n+1)!}(x - x_0)^{n+1} \quad \textit{para cierto } y \textit{ entre } x \textit{ y } x_0.$$

Resto de Cauchy:

$$R(x) = \frac{f^{(n+1)}(y)}{n!}(x - y)^n(x - x_0) \quad \textit{para cierto } y \textit{ entre } x \textit{ y } x_0.$$

Resto integral:

$$R(x) = \int_{x_0}^{x} \frac{f^{(n+1)}}{n!}(x - t)^n \, dt.$$

Se utilizó el resto de Lagrange (con $x_0 = 0$).

3.2.2 Polinomios de Lagrange

Otro punto de vista es intentar aproximar f por un polinomio que coincida con f en ciertos puntos $x_1, x_2, \ldots, x_n \in [-a, a]$: estos son los polinomios de Lagrange [12]. El polinomio de Lagrange de orden n asociado a f tiene la forma siguiente:

POLINOMIO DE LAGRANGE

$$L_n(x) = f(x_1)\ell_{x_1}(x) + f(x_2)\ell_{x_2}(x) + \ldots + f(x_n)\ell_{x_n}(x),$$

donde ℓ_{x_j} son polinomios de grado n que valen 1 en $x = x_j$ y 0 en $x = x_i$ con $i \neq j$ y son llamados la base de Lagrange. Así definido, el polinomio de Lagrange L_n es claramente de orden n y coincide con f en los puntos $x_1, x_2, \ldots, x_n$. Pero, ¿son fáciles de encontrar los ℓ_{x_i}? Ésta es justamente la ventaja, es muy fácil verificar que los ℓ_{x_i} se obtienen de la expresión:

BASE DE POLINOMIOS DE LAGRANGE

$$\ell_{x_i} = \frac{\prod_{j=1, i\neq j}^{n} (x - x_j)}{\prod_{j=1, i\neq j}^{n} (x_i - x_j)}.$$

ya que son polinomios de grado n y evaluando se verifica que $\ell_{x_i}(x_j)$ es cero si $i \neq j$ y uno si $i = j$. Tomemos ahora el caso particular en que $[-a, a] = [-1, 1]$, $n = 2$ y $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, entonces, aplicando la fórmula anterior se obtiene:

$$\ell_{-1} = \frac{(x-0)(x-1)}{(-1-0)(-1-1)} = \frac{1}{2}x(x-1),$$

$$\ell_0 = \frac{(x+1)(x-1)}{(0+1)(0-1)} = (x+1)(1-x), \quad \ell_1 = \frac{(x+1)(x-0)}{(1+1)(1-0)} = \frac{1}{2}x(x+1).$$

En efecto, cada uno de ellos es un polinomio de grado 2 que vale 1 en uno de los puntos $x_j \in \{-1, 0, 1\}$ y cero en los demás. Entonces, el polinomio de Lagrange de orden 2 es:

$$L_2(x) = f(-1)\ell_{-1}(x) + f(0)\ell_0(x) + f(1)\ell_1(x).$$

Para ver el error que se comete al aproximar f por L_2, escribamos la aproximación de Taylor de f de orden 2 en torno al origen (suponemos f tres veces continuamente diferenciable):

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2}f''(0) + \frac{x^3}{3!}f'''(y),$$

[12] En honor al matemático y físico italiano Joseph Louis Lagrange (1736-1813).

con y algún punto en $[-1,1]$ que depende de x. Evaluando en los tres puntos $\{-1,0,1\}$ se tiene:

$$\begin{aligned} f(-1) &= f(0) - f'(0) + \frac{1}{2}f''(0) - \frac{1}{3!}f'''(y_1) \\ f(0) &= f(0) \\ f(1) &= f(0) + f'(0) + \frac{1}{2}f''(0) + \frac{1}{3!}f'''(y_2) \end{aligned}$$

y multiplicando la primera por ℓ_{-1}, la segunda por ℓ_0, la tercera por ℓ_1 y sumando se obtiene:

$$L_2 = (\ell_{-1}+\ell_0+\ell_1)f(0)+(\ell_1-\ell_{-1})f'(0)+\frac{1}{2}(\ell_1+\ell_{-1})f''(0)+\frac{1}{3!}(f'''(y_2)\ell_1-f'''(y_1)\ell_{-1})$$

y notando que:

$$\ell_{-1}+\ell_0+\ell_1 = 1, \quad \ell_1-\ell_{-1} = x, \quad \ell_1+\ell_{-1} = x^2, \quad ||\ell_1|| \leq 1/2, \quad ||\ell_{-1}|| \leq 1/2$$

se obtiene

$$||f - L_2|| \leq \left(\frac{1}{2\cdot 3!} + \frac{1}{3!}\right) = \frac{1}{6}||f'''||.$$

Más generalmente, se puede probar[13] que si se hacen coincidir f y L_n en $n+1$ puntos equiespaciados en $[-1,1]$, y f es $n+1$ veces continuamente diferenciable, se tiene que:

$$||f - L_n|| \leq \frac{1}{2(n+1)}\left(\frac{2}{n}\right)^{(n+1)}||f^{(n+1)}||.$$

La misma observación hecha para el error de aproximación por polinomios de Taylor vale aquí, esto es, que la cota del lado derecho converge a cero si $||f^{(n+1)}||$ está acotada, pero podría también no converger a cero o incluso diverger.

✍ **Ejercicio 3.20.** Compare los errores cometidos al aproximar por polinomios de Taylor y Lagrange: $||f - T_n||$ y $||f - L_n||$. Discuta cuál sería más preciso en términos de cifras significativas si suponemos que $||f^{(n+1)}||$ está acotado. Notar que esta comparación es algo forzada, ya que el polinomio de Taylor aproxima en torno a un punto usando n derivadas en dicho punto, en cambio el polinonio de Lagrange aproxima en torno a n puntos usando solamente la función.

En el Cuadro 3.5 se calculan como ejemplo los errores de aproximación de la función coseno en $[-1,1]$ por polinomios Lagrange y de Taylor para órdenes pares crecientes El polinomio de Taylor es mejor aproximación de las derivadas cerca del origen, mientras que el de Lagrange intenta interpolar en todo el intervalo. En este sentido, las dos aproximaciones son de distinta naturaleza, el polinomio de Taylor es una aproximación local mientras que el polinomio de Lagrange es una aproximación global. ❒✍

✍ **Ejercicio 3.21.** Obtenga el Cuadro 3.5 y la Figura 3.5 (derecha) al aproximar la función e^x entre -1 y 1. Haga lo mismo para $e^{-\frac{1}{x^2}}$. Discuta los resultados.

[13] Véase la referencia [**24**].

orden n	error uniforme Lagrange $\max_{x\in[-1,1]} \lvert\cos(x) - L_n(x)\rvert$	error uniforme Taylor $\max_{x\in[-1,1]} \lvert\cos(x) - T_n(x)\rvert$
2	$9,90 \times 10^{-3}$	$4,03 \times 10^{-2}$
4	$1,27 \times 10^{-4}$	$1,36 \times 10^{-3}$
6	$9,50 \times 10^{-7}$	$2,45 \times 10^{-5}$
8	$4,70 \times 10^{-9}$	$2,73 \times 10^{-7}$

CUADRO 3.5. Error de aproximación uniforme de la función coseno en $[-1, 1]$ por polinomios Lagrange y de Taylor para órdenes pares crecientes.

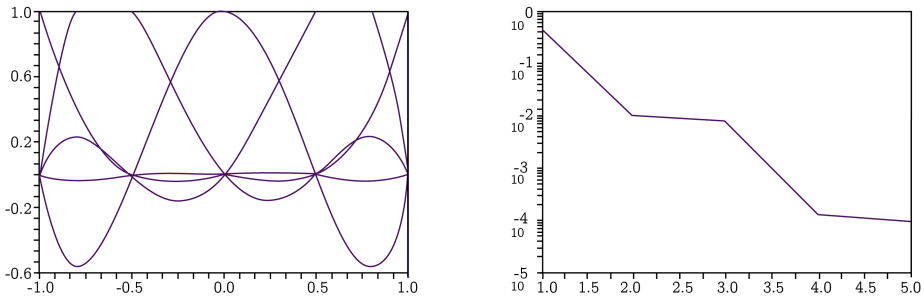


FIGURA 3.5. Izquierda: base de Lagrange para $n = 4$ en $[-1, 1]$. Observe que cada polinomio de grado 4 vale 1 en uno de los puntos del conjunto $\{-1, -1/2, 0, 1/2, 1\}$ y vale 0 en los demás. Derecha: logaritmo del error uniforme al aproximar la función coseno por su polinomio de Lagrange en $[-1, 1]$ en función del orden del polinomio.

3.2.3 Derivadas numéricas y polinomios de Newton

Una desventaja de los polinomios de Lagrange es que si agregamos un punto más de interpolación x_{n+1} y buscamos un nuevo polinomio de Lagrange, esta vez de grado $n+1$, que interpole f en dichos puntos, debemos volver a calcular todos los elementos de la base de Lagrange para construirlo. Cuando esto ocurre, resulta más útil la llamada *interpolación de Newton*. Consideremos para $h > 0$ los puntos

$$x_0, x_0 + h, x_0 + 2h, x_0 + 3h, \ldots.$$

Definamos ahora[14] las cantidades siguientes, llamadas derivadas numéricas progresivas[15]:

$$\begin{aligned}\Delta_h f(x_0) &= \frac{f(x_0+h)-f(x_0)}{h}\\ \Delta_h^2 f(x_0) &= \Delta_h(\Delta_h f(x_0))\\ \Delta_h^3 f(x_0) &= \Delta_h(\Delta_h^2 f(x_0))\\ &\vdots\end{aligned}$$

esto es, por ejemplo:

$$\Delta_h^2 f(x_0) = \frac{\frac{f(x_0+2h)-f(x_0+h)}{h} - \frac{f(x_0+h)-f(x_0)}{h}}{h} = \frac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{h^2}.$$

Consideremos ahora polinomios de la forma:

Base de Polinomios de Newton

$$\begin{aligned}N_1(x) &= (x-x_0),\\ N_2(x) &= (x-x_0)(x-x_0-h),\\ N_3(x) &= (x-x_0)(x-x_0-h)(x-x_0-2h),\\ &\vdots\end{aligned}$$

entonces, el *polinomio de Newton* de orden n de f en torno a 0 está dado por la siguiente expresión[16]:

Polinomio de Newton

$$p(x) = f(x_0) + N_1(x)\Delta_h f(x_0) + \frac{N_2(x)}{2!}\Delta_h^2 f(x_0) + \ldots + \frac{N_n(x)}{n!}\Delta_h^n f(x_0).$$

Se tiene que

$$\begin{aligned}p(x_0) &= f(x_0)\\ p(x_0+h) &= f(x_0)+f(x_0+h)-f(x_0) = f(x_0+h)\end{aligned}$$

y más generalmente se puede verificar que

$$p(x_0+kh) = f(x_0+kh), \quad \forall k = 0,\ldots,n-1,$$

[14] Análogamente como se hizo para el Delta de Aitken en el Capítulo 2.

[15] Existen las derivadas numéricas centradas y retrógradas que son respectivamente $(f(x_0+h/2)-f(x_0-h/2))/h$ y $(f(x_0)-f(x_0-h))/h$.

[16] Notar la analogía con el polinomio de Taylor.

esto es, la función y el polinomio coinciden en los n puntos $x_0, x_0+h, \ldots, x_0+(n-1)h$, pero si ahora agregamos un punto extra $x_{n+1} = x_0 + nh$ basta con sumar el término

$$+\frac{N_{n+1}(x)}{(n+1)!}\Delta_h^{n+1} f(x_0)$$

para actualizar la interpolación, sin necesidad de recalcular los términos anteriores.

✍ **Ejercicio 3.22.** Encuentre la expresión para los polinomios de orden 4 de la base de Lagrange que corresponden a la Figura 3.5 izquierda.

✍ **Ejercicio 3.23.** Usando la *aproximación de Stirling* que dice que $\ln n!$ se comporta como $n\ln n - n$ para n grande, pruebe que la cota del error para los polinomios de Lagrange es menor que la cota del error de Taylor cuando n es grande. Para ello tome $a = 1$ y estudie el límite cuando n tiende a infinito de $\frac{n!2^n}{n^{(n+1)}}$ tomando logaritmo.

✍ **Ejercicio 3.24.** Pruebe que el polinomio de Newton cumple que

$$\frac{d^n p}{dx^n}(x_0) = \Delta_h^n f(x_0).$$

Puede serle útil recordar que al derivar $k+1$ veces un polinomio de grado k, el resultado es nulo.

✍ **Ejercicio 3.25.** Investigue sobre los llamados *polinomios de Chevishev*.

✍ **Ejercicio 3.26.** Investigue sobre cuáles son las llamadas *funciones Spline*.

3.3 Aproximando el área bajo la curva de una función

Todos conocen la máxima *dividir para reinar*. Para aproximar[17] la integral de una función continua $f : [a,b] \to \mathbb{R}$ la separamos como

División base de una Cuadratura

$$\int_a^b f(x)\,dx = \int_a^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{N-1}}^{b} f(x)dx, \tag{3.1}$$

descomposición que corresponde a una subdivisión

$$[a,b] = [x_0,x_1] \cup [x_1,x_2] \cup \ldots \cup [x_{N-1},x_N]$$

en N subintervalos $[x_i, x_{i+1}]$ que para simplificar supondremos de una longitud o *paso* (constante) h:

$$h = \frac{b-a}{N}.$$

De esta forma se tiene que

$$a = x_0, \quad b = x_N, \quad x_{i+1} = x_i + h,\ i = 0, \ldots, N-1.$$

[17] Esto puede ser particularmente útil para aproximar integrales que no tienen una expresión analítica simple. Las *integrales elípticas* o la integral $\int_0^1 \frac{\sin x}{x}$ son algunos ejemplos.

Luego de esto, hacemos el siguiente cambio de variables

$$z = \frac{2}{h}(x - x_i) - 1$$

que no es más que una *transformación lineal afín* entre los intervalos

$$x \in [x_i, x_{i+1}] \quad \Leftrightarrow z \in [-1, 1]$$

y con esto se pueden reescribir cada una de las partes de la integral como

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{h}{2}\int_{-1}^{1} g_i(z)\,dz \tag{3.2}$$

donde la nueva función a integrar es ahora

$$g_i(z) = f\left(x_i + \frac{h}{2}(z+1)\right). \tag{3.3}$$

De modo que se puede reducir el problema de aproximar la integral de una función continua sobre $[a, b]$ por el de aproximar N integrales de la forma:

$$\int_{-1}^{1} g(z)dz$$

todas en el mismo *intervalo de referencia* $[-1, 1]$.

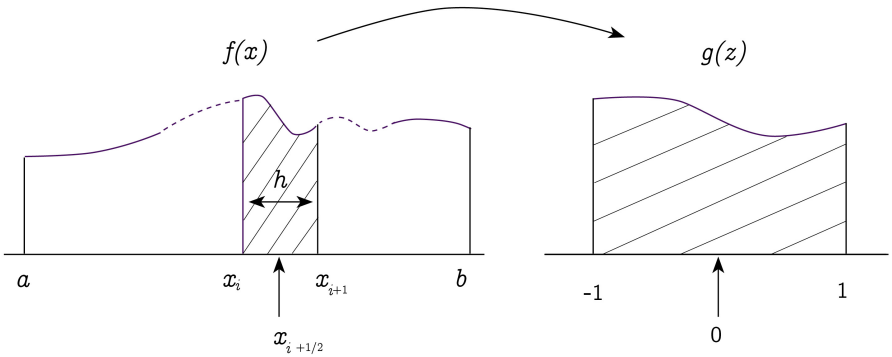


FIGURA 3.6. Reducción de la integral $\int_a^b f(x)\,dx$. Primero por subdivisión del intervalo $[a, b]$ en N subintervalos de largo $h > 0$ y luego por un cambio de variables de $[x_i, x_{i+1}]$ a $[-1, 1]$. El resultado es una suma de integrales de referencia de la forma $\int_{-1}^{1} g(z)\,dz$.

La idea más utilizada es aproximar esta última integral como una combinación lineal finita de los valores de g en ciertos puntos del intervalo $[-1, 1]$. Por ejemplo, si

se eligen los puntos $\{-1, 0, 1\}$ se tendría:

$$\int_{-1}^{1} g(z)dz \approx \alpha g(-1) + \beta g(0) + \gamma g(1),$$

donde α, β y γ son constantes a determinar. Este tipo de aproximaciones lleva a las llamadas *fórmulas de cuadratura* [18].

3.3.1 Método de rectángulos con punto medio

Aproximamos la integral de g como el área del rectángulo (Figura 3.7 (izquierda)):

$$\int_{-1}^{1} g(z)dz \approx 2g(0).$$

En este caso, reemplazando en (3.2) y luego en (3.1) se obtiene la fórmula de cuadratura:

Método de cuadratura por rectángulos (punto medio)

$$\int_a^b f(x)\,dx \approx h \sum_{i=1}^{N-1} f(x_{i+\frac{1}{2}}), \tag{3.4}$$

donde usaremos la siguiente notación para el punto medio entre x_i y x_{i+1}

$$x_{i+\frac{1}{2}} = \frac{x_i + x_{i+1}}{2}.$$

Esta forma de aproximar la integral es una *fórmula de cuadratura por rectángulos* usando el punto medio. Se puede establecer también una fórmula de rectángulos usando el extremo izquierdo x_i o el extremo derecho x_{i+1} o algún otro punto. Pero veremos más adelante que es más conveniente tomar el punto medio.

✍ **Ejercicio 3.27.** Aproxime $\int_0^1 x^2\,dx$ por el método de los rectángulos usando el extremo izquierdo y verifique que esta aproximación converge a $1/3$ cuando el número de subintervalos tiende a infinito.
Solución:

$$\int_0^1 x^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i}{n}\right)^2 = \frac{1}{n^3}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6n^3}$$

y es fácil ver que el límite de la aproximación cuando $n \to \infty$ es $\frac{2}{6} = \frac{1}{3}$. Para la suma de los cuadrados de los enteros véase el Ejercicio 4.4.

[18] Esta idea de cuadratura también es la base para la definición de Riemann de la integral, como un límite cuando $h \to 0$. Véase la Monografía *Cálculo Integral y Series de Potencias* en esta misma colección cf. [**1**].

3.3.2 Método de trapecios

Aproximamos la integral de g como el área del trapecio (Figura 3.7 (centro)):

$$\int_{-1}^{1} g(z)dz \approx g(-1) + g(1)$$

y reemplazando en (3.2) y (3.1) se obtiene la fórmula de cuadratura:

Método de cuadratura por trapecios

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\sum_{i=1}^{N-1}(f(x_i) + f(x_{i+1})), \tag{3.5}$$

conocida como la *fórmula de cuadratura por trapecios.*

✍ **Ejercicio 3.28.** Utilizando el método de los rectángulos y el método de los trapecios para aproximar el área bajo la curva de la circunferencia $f(x) = \sqrt{1-x^2}$, encuentre sumatorias que sirven de aproximaciones para π.

3.3.3 Método de Simpson

Aproximamos la integral de g como el área bajo una parábola. Para ello interpolamos la función g en los puntos $\{-1, 0, 1\}$ por un polinomio de Lagrange de grado 2 como se indica en la Figura 3.7 (derecha) recordemos que se tiene (ver sección precedente sobre interpolación):

$$g(z) = \ell_{-1}(z)g(-1) + \ell_0(z)g(0) + \ell_1(z)g(1)$$

donde ℓ_{-1}, ℓ_0, ℓ_1 es la base de Lagrange de grado 2. Recordando las expresiones explícitas que obtuvimos antes para esta base, integrando en el intervalo $[-1, 1]$ se obtiene

$$\begin{aligned}
\int_{-1}^{1} \ell_{-1}(z)\,dz &= \int_{-1}^{1} \frac{1}{2}z(z-1)\,dz = \frac{1}{3} \\
\int_{-1}^{1} \ell_0(z)\,dz &= \int_{-1}^{1} (z+1)(1-z)\,dz = \frac{4}{3} \\
\int_{-1}^{1} \ell_1(z)\,dz &= \int_{-1}^{1} \frac{1}{2}z(z+1)\,dz = \frac{1}{3}
\end{aligned}$$

es fácil ver que la integral de g en $[-1, 1]$ puede aproximarse por

$$\int_{-1}^{1} g(z)dz \approx \frac{1}{3}(g(-1) + 4g(0) + g(1)).$$

Reemplazando en (3.2) y (3.1) se obtiene la fórmula:

MÉTODO DE CUADRATURA DE SIMPSON

$$\int_a^b f(x)\,dx \approx \frac{h}{6}\sum_{i=1}^{N-1}(f(x_i)+4f(x_{i+\frac{1}{2}})+f(x_{i+1})), \tag{3.6}$$

conocida como la *fórmula de cuadratura Simpson* .

✍ **Ejercicio 3.29.** Pruebe que la fórmula de Simpson es exacta para aproximar $\int_{-1}^{1} x^3 dx$ y en general para aproximar $\int_{-1}^{1} p(x)\,dx$ para cualquier polinomio p de grado menor o igual a 3.

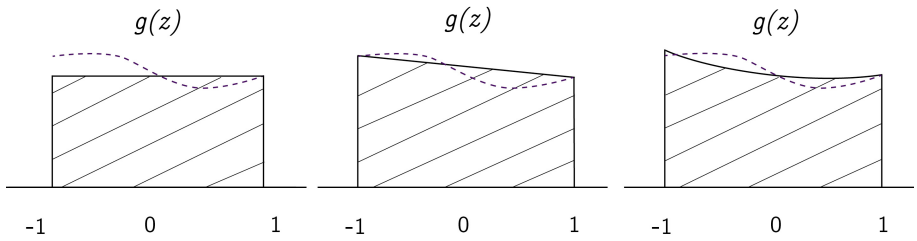


FIGURA 3.7. Ilustración de los métodos de cuadratura de rectángulos (3.4) utilizando el punto medio (izquierda), trapecios (3.5) (centro) y Simpson (3.6) (derecha). Se basan respectivamente en una interpolación constante, lineal y cuadrática de la función en los puntos $\{-1, 0, 1\}$.

3.3.4 Estimación del error de cuadratura

Supongamos que la integral $\int_a^b f(x)dx$ la hemos aproximado por la cuadratura Q_h, ya sea por rectángulos, trapecios o Simpson. Probaremos que si:

- la función $f(x)$ que integramos es $k+1$ veces continuamente diferenciable,
- la fórmula de cuadratura integra exactamente los polinomios z^i, $i = 0, \ldots, k$ en $[-1, 1]$,

entonces una estimación del *error de cuadratura* es:

$$\left| \int_a^b f(x)dx - Q_h \right| \leq C_k h^{k+1} \tag{3.7}$$

donde C_k es una constante que depende de k pero no de h.

Para ello, recordemos que de las identidades (3.2) y (3.1) se tiene que:

$$\int_a^b f(x)dx = \frac{h}{2}\sum_{i=0}^{N-1}\int_{-1}^{1} g_i(z)dz$$

donde las funciones g_i fueron definidas en (3.3). Estas funciones resultan ser también $k+1$ veces continuamente diferenciables en z por la misma definición, de modo que podemos desarrollar cada g_i en serie de Taylor en torno a cero con un error e_i:

$$g_i(z) = g_i(0) + zg_i'(0) + \frac{z^2}{2!}g_i''(0) + \ldots + \frac{z^k}{k!}g_i^{(k)}(0) + e_i(z) = T_k(z) + e_i(z).$$

Como la cuadratura es exacta para z^i, $i = 0, \ldots, k$, es también exacta al integrar $T_k(z)$ entre -1 y 1, de modo que el error de cuadratura está dado por

$$\int_a^b f(x)dx - Q_h = \frac{h}{2}\sum_{i=0}^{N-1}\int_{-1}^{1} e_i(z)dz \tag{3.8}$$

donde e_i son los restos de Taylor. Pero sabemos que el error $e_i(z)$ se puede expresar como:

$$e_i(z) = \frac{z^{k+1}}{(k+1)!}g_i^{(k+1)}(\xi)$$

donde ξ es algún punto en el intervalo $[-1, 1]$, así es que usando la regla de la cadena tenemos que

$$g_i^{(k+1)}(\xi) = \left(\frac{h}{2}\right)^{k+1} f^{(k+1)}\left(x_i + \frac{h}{2}(\xi+1)\right)$$

de donde

$$\left|\int_{-1}^{1} e(z)dz\right| \leq C_k\, h^{k+1} f^{(k+1)}(\tau_i)$$

donde C_k es una constante que sólo depende de k pero no de h, y τ_i es algún punto en el intervalo $[x_i, x_{i+1}]$. Reemplazando esto en (3.8) se obtiene finalemente (3.7).

3.3.5 Aplicación del error de cuadratura

Apliquemos ahora la estimación del error de cuadratura (3.7):

- La cuadratura con rectángulos considerando el extremo izquierdo o derecho solamente es exacta para la función constante $z^0 = 1$ en $[-1, 1]$, de modo que el error de cuadratura está acotado por h.
- La cuadratura con rectángulos considerando el punto medio y la cuadratura por trapecios son exactas para 1 y z en $[-1, 1]$, de modo que el error de aproximación en ambos casos está acotado por h^2.
- Es fácil verificar que la cuadratura de Simpson es exacta para 1, z, z^2 e incluso z^3 en $[-1, 1]$ de modo que el error de cuadratura en este caso se comporta como h^4. Es por esto que la cuadratura de Simpson es la más comúnmente usada en la práctica.

Estos comportamientos del error de cuadratura se resumen en el Cuadro 3.6.

Veamos las fórmulas de cuadratura sobre un ejemplo numérico: tomemos como caso de prueba:

$$\int_0^{\pi} \sin(x)\, dx = -\cos(x)|_0^{\pi} = -(-1) + 1 = 2.$$

Fórmula de cuadratura	aproximación de la integral en $[x_i, x_{i+1}]$	exacta para polinomios de grado k	error de cuadratura h^{k+1}
Rectángulos (pto. izq. o der.)	$hf(x_i)$ o $hf(x_{i+1})$	0	h
Rectángulos (pto. medio)	$hf(x_{i+\frac{1}{2}})$	1	h^2
Trapecios	$\frac{h}{2}(f(x_i)+f(x_{i+1}))$	1	h^2
Simpson	$\frac{h}{6}(f(x_i)+4f(x_{i+\frac{1}{2}})+f(x_{i+1}))$	3	h^4

CUADRO 3.6. Resumen de las fórmulas de cuadratura más utilizadas.

N	h	Rectángulos (izquierda)	Rectángulos (pto. medio)	Trapecios	Simpson
2	$\frac{\pi}{2}$	$1,5707963$	$2,2214415$	$1,5707963$	$2,0045598$
3	$\frac{\pi}{3}$	$1,8137994$	$2,0943951$	$1,8137994$	$2,0008632$
5	$\frac{\pi}{5}$	$1,9337656$	$2,0332815$	$1,9337656$	$2,0001095$
10	$\frac{\pi}{10}$	$1,9835235$	$2,0082484$	$1,9835235$	$2,0000068$
20	$\frac{\pi}{20}$	$1,9958860$	$2,0020576$	$1,9958860$	$2,0000004$
40	$\frac{\pi}{40}$	$1,9989718$	$2,0005141$	$1,9989718$	$2,0000000$

CUADRO 3.7. Desempeño de las fórmulas de cuadratura al aproximar $\int_0^\pi \sin(x)\,dx = 2$.

En el Cuadro 3.7 se registran los valores obtenidos con tres fórmulas de cuadratura con un número creciente de puntos N equiespaciados en el intervalo $[0, \pi]$.

✍ **Ejercicio 3.30.** Nótese que, en el Cuadro 3.7, el método de trapecios aproxima por defecto la integral $\int_0^\pi \sin(x)\,dx$. Además, el método de rectángulos con un punto a la izquierda produce los mismos resultados que el método de trapecios. Entregue una explicación gráfica a estos hechos.

✍ **Ejercicio 3.31.** Construya una tabla similar al Cuadro 3.7 para aproximar la integral $\int_0^\pi \sin^2(x)dx$. Compare los errores de cuadratura.

✍ **Ejercicio 3.32.** Investigue sobre otras fórmulas de cuadratura: cuadratura de Gauss y cuadratura de Newton-Côtes.

Capítulo 4: ¿Cómo y por qué resolver sistemas lineales?

> *"Adquirir un conocimiento es siempre útil al intelecto, aunque sólo sea para abandonar lo inútil y reservar lo que es bueno. Pues no se puede amar u odiar nada sin conocer, y el deseo del conocimiento actúa sobre el hombre como un instinto superior"* Leonardo da Vinci *(La Última Lección, 1499).*

Si Leonardo da Vinci hubiera vivido en nuestra época, de seguro se habría interesado por el análisis numérico, al menos para estudiar el vuelo de las aves o para indagar el interior de la anatomía humana. Comenzaremos este capítulo con un ejemplo que muestra lo útiles que han llegado a ser los sistemas lineales para ayudar al hombre en su empresa de volar, pero también lo complejos que pueden llegar a ser, debido en gran parte al enorme número de incógnitas que involucran.

4.1 Tiempo de cálculo

Hoy en día, la forma precisa del nuevo fuselaje de un avión de pasajeros es diseñada con ayuda de un ordenador. La forma debe satisfacer una serie de requerimientos: permitir velocidad, pero estabilidad en el vuelo; economizar el máximo de combustible y albergar el máximo de espacio al mismo tiempo. El diseño es luego probado en experimentos de cámaras de viento con una maqueta a escala antes de ser propuesta a la industria. Un diseño exitoso puede influir enormemente en la disminución de las tarifas de los vuelos y permitir a más y más gente viajar en avión.

Por ejemplo, en el cálculo de la forma perfecta de un ala de avión se deben resolver fenómenos que ocurren a la escala del centímetro sobre la superficie y cerca del ala. Como un ala real tiene una superficie de varios metros cuadrados, y en un metro cuadrado hay $10000 = 10^4$ centímetros cuadrados, es plausible pensar que se deba calcular la presión del aire sobre y en torno al ala en vuelo en más de $100000 = 10^5$ puntos. El cálculo de estas presiones es fundamental para el diseño del ala, pues una presión muy elevada o muy variable podría romperla o desestabilizla. Al mismo tiempo, se intenta lograr el máximo de presión en la parte inferior de ésta para que el aire sostenga al avión.

Las presiones en cada punto son más de 10^5 incógnitas, que resultan de resolver un sistema lineal de más de 10^5 ecuaciones[1]. La Figura 4.1 muestra una forma típica de perfil de ala de un avión, conocido como *perfil NACA*. El ala está rodeada de una malla de triángulos que se utiliza para realizar los cálculos.

[1]Sistema que resulta de discretizar el llamado sistema de Navier-Stokes.

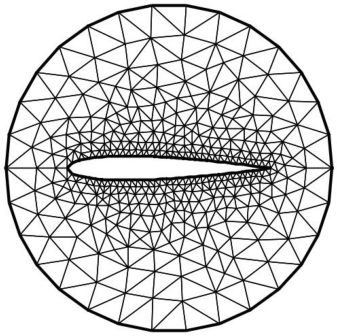


FIGURA 4.1. Perfil NACA de un ala de avión. En cada vértice de triángulo, se debe calcular la presión del aire circundante al ala.

También hemos elegido este ejemplo para ilustrar la importancia del concepto de *número de operaciones* y *tiempo de cálculo*. Pensemos en la factibilidad del cálculo para resolver un sistema lineal tan grande como el del diseño de un ala. Un procesador puede llegar hoy a una velocidad de más de 2Ghz (dos Giga-hertz), lo que significa que puede realizar hasta 2×10^9 operaciones aritméticas en un segundo. Es una cantidad impresionante. Sin embargo, ¡aún nos limita!.

En efecto, calculemos el tiempo que tomaría resolver un sistema de 10^5 por 10^5 en un tal procesador. Para ello, debemos saber primero cuántas operaciones aritméticas[2] requiere resolver un sistema de $n \times n$ cuando n es muy grande. Llamemos $Op(n)$ a esta función. Conocida ésta, el tiempo en segundos que nos toma resolver el sistema está dado por:

$$T = \frac{Op(n)}{v}$$

donde v es la velocidad del procesador en *hertz*. Si se divide este tiempo por 3600, 3600×24 ó $3600 \times 24 \times 365$ se obtiene el tiempo en horas, días o años respectivamente.

Veamos en el Cuadro 4.1 los tiempos de cálculo que se obtienen para distintas funciones $Op(n)$ al resolver un sistema de 10^5 ecuaciones y 10^5 incógnitas. Notemos que solamente $Op(n) = n^2$ o $Op(n) = n^3$ son realistas para poder realizar el cálculo pues el tiempo de cálculo va de una milésima de segundo a 6 días. Afortunadamente, veremos que el número de operaciones al resolver un sistema lineal se encuentra entre estos dos casos y explica la creciente utilización de ordenadores en el diseño de aviones hoy en día.[3] Es por esta razón de tiempo y economía de recursos que resolver

[2]Las operaciones que más cuentan son las multiplicaciones o divisiones, siendo las sumas y restas muchísimo más rápidas en un procesador.

[3]Hoy en día, se utiliza también el *cálculo paralelo*, si se tienen p procesadores en paralelo, el tiempo se reduce en el mejor de los casos en el mismo factor.

$Op(n)$	$n = 10^5$	tiempo de cálculo
n^2	10^{10}	0,001 segundo
n^3	10^{15}	6 días
n^4	10^{20}	1586 años
$n!$	¡uf!	más que el tiempo del universo

CUADRO 4.1. Tiempo de cálculo en un procesador moderno en función del número de operaciones aritméticas que toma resolver un problema de tamaño n.

eficientemente un sistema lineal ha sido y sigue siendo un problema fundamental del análisis numérico.

No desarrollaremos este ejemplo en más detalle a lo largo del capítulo, pues su objetivo era solamente motivar la importancia de la resolución de sistemas lineales y la complejidad que puede llegar a tener su resolución. Para ilustrar las diversas técnicas y métodos que iremos introduciendo a lo largo del capítulo, utilizaremos más bien ejemplos sencillos y al final abordaremos una aplicación de mayor complejidad: la tomografía computarizada.

✍ **Ejercicio 4.1.** Otro ámbito en el que se deben resolver sistemas lineales de gran tamaño es en la confección de listas de prioridad en las búsquedas de internet[4]. Investigue sobre esta aplicación. *Solución: La búsqueda se hace utilizando aquellas páginas que tienen mayor prioridad en la web, esto es, aquellas páginas que tienen una mayor probabilidad de ser visitadas y, de hecho, los resultados de una búsqueda se listan en orden de prioridad. Una forma de tener una idea del número de páginas que maneja Google, por ejemplo, es buscando el omnipresente vocablo inglés "the" y observando el número de páginas encontradas. Actualmente, ese número es cercano a* $n = 14 \times 10^9$. *Para calcular la prioridad* x_i *de la página* i, *se calcula primero una enorme matriz* A *de* $n \times n$ *donde* a_{ij} *representa la probabilidad de que un internauta salte de una página* j *a otra página* i *de la web. Luego se calcula el límite de la siguiente sucesión de sistemas lineales:*

$$Ax^1 = x^0,\ Ax^2 = x^1,\ Ax^3 = x^2,\ \ldots$$

donde x^0 *es un vector de* n *componentes, todas iguales a* $1/n$, *representando el hecho de que inicialmente todas las páginas tienen la misma prioridad para un internauta novato. Luego las iteraciones representan cómo van cambiando las prioridades a medida que el internauta navega en internet. Si el internauta navegara un tiempo infinito, finalmente correspondería resolver:*

$$Ax = x$$

que es equivalente al sistema lineal:

$$(I - A)x = 0$$

[4]Puede consultar la referencia **[21]** por ejemplo.

donde I es la matriz identidad de $n \times n$. Como la solución de este problema está definida salvo constante, se elige la constante de modo que $\sum_{i=1}^{n} x_i = 1$ y así se obtienen las prioridades x_i.

4.2 Resolución numérica de sistemas lineales

Cuando se desea resolver el sistema lineal

$$Ax = b, \qquad A \in \mathbb{R}^{m \times n}$$

éste representa generalmente un sistema de m ecuaciones (número de filas de la *matriz del sistema* A) con n incógnitas (número de columnas de la matriz A) donde el vector $x \in \mathbb{R}^n$ contiene las n incógnitas $x_1, x_2, \ldots, x_n$ en una columna y el vector $b \in \mathbb{R}^m$ o *lado derecho* contiene los términos constantes $b_1, b_2, \ldots, b_m$:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \vdots \quad \vdots \quad \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

Veremos que un sistema lineal aparece por doquier, cuando queremos, por ejemplo, encontrar los n coeficientes de un polinomio de grado $n-1$ que toma m valores dados. O para hacer una regresión lineal de modo de ajustar una nube de puntos en el plano a una recta. O en una tomografía médica. Todos esos ejemplos los estudiaremos en este capítulo.

Para comenzar, recordemos un hecho básico: que un sistema lineal puede tener ninguna, una única, o infinitas soluciones. En el caso en que A sea cuadrada y además su determinante sea no nulo, esto es, *invertible o regular*[5], sabemos que existe una única solución, que podemos escribir como:

$$x = A^{-1}b$$

donde A^{-1} es la matriz inversa de A. Sin embargo, del punto de vista del análisis numérico, para resolver el sistema lineal $Ax = b$ *no es necesario ni conveniente* calcular la inversa de A. Para explicar esto, notemos que calcular la inversa de una matriz A de $n \times n$ es equivalente a resolver n sistemas lineales. En efecto, si consideramos los n sistemas:

$$Ax_1 = e_1, \ Ax_2 = e_2, \ \ldots, \ Ax_n = e_n$$

donde e_i es el $n-$ésimo vector de la base canónica, resulta que los vectores x_i son exactamente las n columnas de la inversa. Así es pues:

$$A \begin{pmatrix} \vdots & \vdots & \ldots & \vdots \\ x_1 & x_2 & \ldots & x_n \\ \vdots & \vdots & \ldots & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & \ldots & \vdots \\ e_1 & e_2 & \ldots & e_n \\ \vdots & \vdots & \ldots & \vdots \end{pmatrix} = I.$$

[5]En oposición a no invertible o singular si $\det A = 0$.

De modo que en principio, resolver $Ax = b$ es tan o más simple que calcular la inversa de A. No solamente eso, veremos que la resolución de $Ax = b$ no solo provee una forma muy eficiente de encontrar A^{-1} sino que también podemos hallar el determinante de A. Es tan crucial esto, que el desarrollo tecnológico actual no existiría si nos hubiéramos visto obligados a resolver $Ax = b$ calculando A^{-1} con los métodos tradicionales de determinantes usando la matriz de cofactores.

Veremos, en lo que sigue de este capítulo, métodos directos para resolver el problema $Ax = b$ y otros tópicos importantes como los sistemas mal condicionados y los sistemas sobredeterminados.

4.3 Eliminación de Gauss

Consideremos el sistema lineal:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13} \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23} \ldots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33} \ldots + a_{2n}x_n &= b_2 \\ \vdots \quad \vdots \quad \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3} \ldots + a_{nn}x_n &= b_n \end{aligned}$$

que para simplificar supondremos con el mismo número n de ecuaciones que de incógnitas.

El método de *eliminación de Gauss* consiste simplemente en eliminar la primera incógnita x_1 de todas las ecuaciones salvo la primera y luego la segunda incógnita x_2 de todas las ecuaciones salvo la primera y la segunda y así sucesivamente, llegar a lo que se conoce como un sistema triangular superior:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13} \ldots + a_{1n}x_n &= b_1 \\ \tilde{a}_{22}x_2 + \tilde{a}_{23} \ldots + \tilde{a}_{2n}x_n &= \tilde{b}_2 \\ \tilde{a}_{33} \ldots + \tilde{a}_{2n}x_n &= \tilde{b}_2 \\ \vdots \quad \vdots \quad \vdots & \\ \tilde{a}_{n,n-1}x_{n-1} + \tilde{a}_{nn}x_n &= \tilde{b}_{n-1} \\ \tilde{a}_{nn}x_n &= \tilde{b}_n \end{aligned}$$

que es muy fácil de resolver por *sustitución hacia atrás*. En efecto, de la última ecuación se calcula x_n y luego se reemplaza su valor en la penúltima de donde se obtiene x_{n-1}. Luego los valores de x_n y x_{n-1} se reemplazan ambos en la antepenúltima ecuación para encontrar x_{n-2} y así sucesivamente hasta llegar a calcular todas las incógnitas x_3, x_2 y finalmente x_1.

Sistema triangular superior: sustitución hacia atrás

- **Etapa** 1**:** Se calcula x_n de la ecuación n.
- **Etapa** i**:** Se reemplazan los ya calculados $\{x_i, x_{i+1}, \ldots, x_n\}$ en la ecuación $i-1$ de donde se obtiene x_{i-1}. Esto para i de n a 2.

Este trabajo de sustitución hacia atrás toma $n(n-1)/2$ multiplicaciones al reemplazar las incógnitas ya calculadas y n divisiones por los $\tilde{a}_{ii}$ para despejar el valor de x_i. Esto es, el número de operaciones aritméticas (solamente contando multiplicaciones y divisiones y sin contar sumas o restas) es:

Operaciones aritméticas de la sustitución hacia atrás

$$Op_1 = n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}.$$

Para llegar a la forma triangular, la idea es ir eliminando los coeficientes sucesivamente por columnas como se indica esquemáticamente aquí:

$$\begin{pmatrix} \boxed{a_{11}} & a_{12} & a_{13} & \ldots & a_{nn} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{pmatrix} \rightarrow \begin{pmatrix} \boxed{a_{11}} & a_{12} & a_{13} & \ldots & a_{nn} \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \ldots & \tilde{a}_{2n} \\ 0 & \tilde{a}_{32} & \tilde{a}_{33} & \ldots & \tilde{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{n2} & \tilde{a}_{n3} & \ldots & \tilde{a}_{nn} \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{nn} \\ 0 & \boxed{\tilde{a}_{22}} & \tilde{a}_{23} & \ldots & \tilde{a}_{2n} \\ 0 & \tilde{a}_{32} & \tilde{a}_{33} & \ldots & \tilde{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{n2} & \tilde{a}_{n3} & \ldots & \tilde{a}_{nn} \end{pmatrix} \rightarrow \begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{nn} \\ 0 & \boxed{\tilde{a}_{22}} & \tilde{a}_{23} & \ldots & \tilde{a}_{2n} \\ 0 & 0 & \tilde{\tilde{a}}_{33} & \ldots & \tilde{\tilde{a}}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \tilde{\tilde{a}}_{n3} & \ldots & \tilde{\tilde{a}}_{nn} \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{nn} \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \ldots & \tilde{a}_{2n} \\ 0 & 0 & \boxed{\tilde{\tilde{a}}_{33}} & \ldots & \tilde{\tilde{a}}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \tilde{\tilde{a}}_{n3} & \ldots & \tilde{\tilde{a}}_{nn} \end{pmatrix} \rightarrow \begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{nn} \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \ldots & \tilde{a}_{2n} \\ 0 & 0 & \boxed{\tilde{\tilde{a}}_{33}} & \ldots & \tilde{\tilde{a}}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \tilde{\tilde{\tilde{a}}}_{nn} \end{pmatrix}$$

Los elementos enmarcados son llamados *pivotes* y supondremos para simplificar que ellos resultan siempre no nulos en el proceso de eliminación. Si no es así, una forma de que no aparezcan pivotes nulos es permutar la fila del pivote nulo por la fila del pivote mayor en la misma columna (esto es equivalente a permutar el orden de las ecuaciones). Esto siempre es posible si la matriz del sistema es invertible y se conoce como *método de eliminación de Gauss con pivote parcial* . La otra forma es permutar la fila y la columna del pivote nulo, por la fila y la columna del elemento mayor en

la submatriz abajo y a la derecha del pivote nulo (esto es equivalente a permutar el orden de las ecuaciones y de las variables). Esto también siempre es posible si la matriz del sistema es invertible y se conoce como *método de eliminación de Gauss con pivote total*.

Sigamos nuestro análisis suponiendo que no se anula ningún pivote. Los elementos de las submatrices abajo y a la derecha de los pivotes que son modificados en cada paso los denotamos con sucesivos tildes sobre los coeficientes.

Para anular un coeficiente q que se encuentra bajo un pivote p en la misma columna, sumamos a la fila del coeficiente q la fila del pivote p multiplicada por $-q/p$ lo que deja un 0 donde estaba q y al mismo tiempo modifica toda la fila donde estaba q solamente hacia la derecha, pero sin alterar los ceros que ya se habían obtenido con las eliminaciones anteriores. Esto corresponde esquemáticamente a la siguiente operación elemental sobre las filas de la matriz del sistema:

$$\underbrace{\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & \boxed{-\frac{q}{p}} & \cdots & \boxed{1} & \\ & & & & 1 \end{pmatrix}}_{\text{operación elemental}} \underbrace{\begin{pmatrix} \times & & & & \\ & \boxed{p} & & & \\ & & \ddots & & \\ & \boxed{q} & \cdots & \times & \\ & & & & \times \end{pmatrix}}_{\text{matriz antes de pivotear}} = \underbrace{\begin{pmatrix} \times & & & & \\ & \boxed{p} & & & \\ & & \ddots & & \\ & \boxed{0} & \cdots & \times & \\ & & & & \times \end{pmatrix}}_{\text{matriz luego de pivotear}}$$

Entonces estas operaciones de eliminación se pueden representar como una serie de matrices que pre-multiplican a la izquierda la matriz del sistema, todas ellas con determinante uno, de modo que si no hay permutación de filas, el determinante de la matriz triangular superior resultante al final del proceso de eliminación (esto es, el producto de los elementos de su diagonal) es exactamente el mismo que el determinante de la matriz original del sistema. Hay que tener en cuenta, eso sí, que un intercambio impar de filas cambia el signo del determinante (ver ejercicio más adelante).

Las operaciones elementales de eliminación también se hacen sobre el lado derecho, y esto lleva a considerar la siguiente matriz aumentada:

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{nn} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n \end{array}\right)$$

al momento de hacer la eliminación. Si se agregan más columnas se puede resolver un *sistema lineal simultáneo*. Esto sirve, por ejemplo, para calcular la inversa de la

matriz del sistema, en que se comienza con:

$$\left(\begin{array}{ccccc|cccc} a_{11} & a_{12} & a_{13} & \dots & a_{nn} & 1 & 0 & 0 & \dots\ 0 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & 0 & 1 & 0 & \dots\ 0 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & 0 & 0 & 1 & \dots\ 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots\ \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & 0 & 0 & 0 & \dots\ 1 \end{array}\right)$$

El siguiente algoritmo resume la eliminación de Gauss:

Algoritmo de eliminación de Gauss

- En la matriz aumentada, se elige el pivote p (eventualmente cambiando filas) y se le suma a cada fila bajo el pivote la fila de p por $-q/p$ donde q es elemento respectivo de la fila bajo el pivote.
- Luego se avanza al siguiente pivote hasta obtener una forma triangular superior en la matriz del sistema.

El siguiente es un ejemplo para resolver un sistema de 3 por 3, encontrar el determinante de la matriz del sistema y calcular su inversa utilizando el método de eliminación de Gauss. Consideremos sistema $Ax = b$ escrito en forma de una matriz aumentada de la manera siguiente:

$$\left(\begin{array}{ccc|c|ccc} 2 & 3 & 1 & 4 & 1 & 0 & 0 \\ 4 & 1 & 1 & 4 & 0 & 1 & 0 \\ -6 & 1 & -2 & -3 & 0 & 0 & 1 \end{array}\right) \quad \begin{array}{ccc} \underbrace{} & \underbrace{} & \underbrace{} \\ A & b & I \end{array}$$

Escalonando como se explicó antes, y simultáneamente dobre todas las columnas, se obtiene:

$$\left(\begin{array}{ccc|c|ccc} \boxed{2} & 3 & 1 & 4 & 1 & 0 & 0 \\ \mathbf{4} & 1 & 1 & 4 & 0 & 1 & 0 \\ \mathbf{-6} & 1 & -2 & -3 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|c|ccc} \boxed{2} & 3 & 1 & 4 & 1 & 0 & 0 \\ \mathbf{0} & -5 & -1 & -4 & -2 & 1 & 0 \\ \mathbf{0} & 10 & 1 & 9 & 3 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|c|ccc} \boxed{2} & 3 & 1 & 4 & 1 & 0 & 0 \\ 0 & \boxed{-5} & -1 & -4 & -2 & 1 & 0 \\ 0 & \mathbf{10} & 1 & 9 & 3 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|c|ccc} \boxed{2} & 3 & 1 & 4 & 1 & 0 & 0 \\ 0 & \boxed{-5} & -1 & -4 & -2 & 1 & 0 \\ 0 & \mathbf{0} & -1 & 1 & -1 & 2 & 1 \end{array}\right)$$

con lo que queda una matriz triangular superior en el lugar de la matriz del sistema:

$$\left(\begin{array}{ccc|c|ccc} \boxed{2} & 3 & 1 & 4 & 1 & 0 & 0 \\ 0 & \boxed{-5} & -1 & -4 & -2 & 1 & 0 \\ 0 & 0 & \boxed{-1} & 1 & -1 & 2 & 1 \end{array}\right)$$

Como no hubo permutación de filas, el producto de los pivotes nos entrega el determinante:

$$\det A = \boxed{2} \times \boxed{-5} \times \boxed{-1} = 10.$$

Si seguimos escalonando, esta vez de abajo hacia arriba (lo que de hecho equivale a la sustitución hacia atrás), dividiendo antes cada fila por el pivote respectivo, queda:

$$\left(\begin{array}{ccc|c|ccc} 2 & 3 & \mathbf{1} & 4 & 1 & 0 & 0 \\ 0 & -5 & \mathbf{-1} & -4 & -2 & 1 & 0 \\ 0 & 0 & \boxed{1} & -1 & 1 & -2 & -1 \end{array}\right) \to \left(\begin{array}{ccc|c|ccc} 2 & 3 & \mathbf{0} & 5 & 0 & 2 & 1 \\ 0 & -5 & \mathbf{0} & -5 & -1 & -1 & -1 \\ 0 & 0 & \boxed{1} & -1 & 1 & -2 & -1 \end{array}\right)$$

$$\left(\begin{array}{ccc|c|ccc} 2 & \mathbf{3} & 0 & 5 & 0 & 2 & 1 \\ 0 & \boxed{1} & 0 & 1 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & \boxed{1} & -1 & 1 & -2 & -1 \end{array}\right) \to \left(\begin{array}{ccc|c|ccc} 2 & \mathbf{0} & 0 & 2 & -\frac{3}{5} & \frac{7}{5} & \frac{2}{5} \\ 0 & \boxed{1} & 0 & 1 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & \boxed{1} & -1 & 1 & -2 & -1 \end{array}\right)$$

y finalmente, dividiendo por 2 la primera fila se obtiene:

$$\left(\begin{array}{ccc|c|ccc} 1 & 0 & 0 & 1 & -\frac{3}{10} & \frac{7}{10} & \frac{1}{5} \\ 0 & 1 & 0 & 1 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & 1 & -1 & 1 & -2 & -1 \end{array}\right)$$

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{I} \quad \underbrace{\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}}_{x} \quad \underbrace{\begin{pmatrix} -\frac{3}{10} & \frac{7}{10} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 1 & -2 & -1 \end{pmatrix}}_{A^{-1}}$$

✍ **Ejercicio 4.2.** Resuelva, como se hizo anteriormente, el siguiente sistema, calculando la solución, el determinante y la inversa de la matriz del sistema.

$$\left(\begin{array}{ccc|c|ccc} 2 & 6 & 8 & 4 & 1 & 0 & 0 \\ 4 & 3 & 4 & 8 & 0 & 1 & 0 \\ 1 & 1 & 1 & 3 & 0 & 0 & 1 \end{array}\right)$$

$$\underbrace{\begin{pmatrix} 2 & 6 & 8 \\ 4 & 3 & 4 \\ 1 & 1 & 1 \end{pmatrix}}_{A} \quad \underbrace{\begin{pmatrix} 4 \\ 8 \\ 3 \end{pmatrix}}_{b} \quad \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{I}$$

Solución: $\det A = \boxed{2} \times \boxed{-9} \times \boxed{-\frac{1}{3}} = 6$

$$\left(\begin{array}{ccc|c|ccc} 1 & 0 & 0 & 2 & -\frac{1}{6} & \frac{1}{3} & 0 \\ 0 & 1 & 0 & 4 & 0 & -1 & 4 \\ 0 & 0 & 1 & 3 & \frac{1}{6} & \frac{2}{3} & -3 \end{array}\right)$$

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{I} \quad \underbrace{\begin{pmatrix} 2 \\ 4 \\ 3 \end{pmatrix}}_{x} \quad \underbrace{\begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & 0 \\ 0 & -1 & 4 \\ \frac{1}{6} & \frac{2}{3} & -3 \end{pmatrix}}_{A^{-1}}$$

4.4 Conteo del número de operaciones aritméticas

El conteo del número de operaciones aritméticas es importante, pues nos da una idea de cuánto se demorará el computador en resolver un problema. Cada operación le toma una fracción de segundo al procesador. Por ejemplo, en un computador con un reloj de 2 gigahercios (2GHz), se realizan 2×10^9 operaciones por segundo.

El número de operaciones aritméticas del método de Gauss puede ser calculado así: para cada uno de los $n-1$ pivotes p se realizan los cálculos del factor $-q/p$ para

las filas bajo él, pero además se recalculan los coeficientes en la submatriz abajo y a la derecha de él, lo que hace que haya un número de operaciones igual a la dimensión de la submatriz abajo y a la derecha del pivote, que es de $(n-i)\times(n-i+1)$ para el i-ésimo pivote. Sumando sobre todos los pivotes se obtiene:

$$\begin{aligned} Op_2 &= \sum_{i=1}^{n}(n-i)(n-i+1) \\ &= \sum_{i=1}^{n}(i-1)i \\ &= \sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i \\ &= \frac{n^3}{3}+\frac{n^2}{2}+\frac{n}{6}-\frac{n(n+1)}{2} \\ &= \frac{n^3}{3}-\frac{n}{3}. \end{aligned}$$

A esto hay que sumarle todavía las operaciones correspondientes sobre el o los lados derechos. Se realiza una operación en cada lado derecho por cada cero que se produce bajo los pivotes. Como hay $n(n-1)/2$ ceros al finalizar el procedimiento, se obtiene:

$$Op_2 = \frac{n^3}{3}-\frac{n}{3}+\frac{n(n-1)}{2}\ell, \quad \ell = \text{número de lados derechos}$$

Finalmente, sumando las operaciones de la sustitución hacia atrás, que al considerar ℓ lados derechos son

$$Op_1 = n\ell + \frac{n(n-1)}{2}$$

se obtiene para n grande, esto es, considerando sólo los términos de mayor orden:

$$Op = Op_1 + Op_2 \approx \frac{n^3}{3}+\frac{n^2\ell}{2}$$

lo que da para $\ell = 1$ (sistema simple) y $\ell = n$ (cálculo de inversa) los valores:

> Operaciones para resolver $Ax=b$ y calcular A^{-1} por Gauss
>
> $$Op(n) \approx \frac{n^3}{3}, \qquad Op(n) \approx \frac{5}{6}n^3.$$

El análisis anterior confirma lo que decíamos antes: el cálculo de la solución de un sistema o de la inversa de la matriz del sistema se pueden realizar en un tiempo similar, habiendo solamente un factor 2,5 entre ellos.

✍ **Ejercicio 4.3.** En la eliminación de Gauss con pivote parcial, si hay intercambio de filas, averigue por qué si el número de intercambios de filas es impar, el determinante de la matriz original del sistema y el determinante de la matriz triangular superior

luego de la eliminación difieren en el signo, en cambio si el número de intercambio de filas es par, son iguales.
Solución: Cada intercambio de filas multiplica el determinante por -1. *Si se realizan* n *intercambios, el factor es* $(-1)^n$, *de modo que si* n *es impar, el factor es* -1.

✍ **Ejercicio 4.4.** Pruebe por inducción que

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} = \frac{n(n+1)(2n+1)}{6}.$$

✍ **Ejercicio 4.5.** ¿Cuál es el número de operaciones para calcular el determinante de la matriz del sistema?, ¿cómo se compara este con el método de los menores principales o cofactores?[6]

✍ **Ejercicio 4.6.** Considere la matriz y el vector:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 4 & 1 & 2 \\ 2 & 3 & 4 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Resuelva por eliminación de Gauss y sustitución hacia atrás el sistema $Ax = b$, calcule el determinante de A y A^{-1}.
Solución:

$$\det(A) = -160, \quad A^{-1} = \frac{1}{160} \begin{pmatrix} 36 & -44 & -4 & -4 \\ -4 & 36 & -44 & -4 \\ -4 & -4 & 36 & -44 \\ -44 & -4 & -4 & 36 \end{pmatrix}, \qquad x = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ -1/2 \end{pmatrix}.$$

4.5 Métodos iterativos por descomposición

Existen también *métodos iterativos* para resolver sistemas lineales. Esta vez no se trata de encontrar una solución exacta en un número finito de pasos, como es el caso del método de eliminación Gauss, sino de aproximar la solución por una sucesión de soluciones que converjan a la solución exacta.

La idea de los *métodos iterativos de descomposición* es la siguiente. Si queremos resolver el sistema

$$Ax = b$$

descomponemos la matriz de la forma

$$A = A_1 + A_2$$

de modo que queda

$$A_1 x = b - A_2 x.$$

Entonces la idea es plantear el siguiente algoritmo iterativo:

[6] Véase la Monografía *Excursiones por el Álgebra Lineal y sus Aplicaciones* en esta misma colección cf. **[13]**.

Método iterativo por descomposición

Etapa 0**:** x_0 dado,

Etapa n**:** $A_1 x_{n+1} = b - A_2 x_n, \qquad n \geq 0.$

Dado que en cada etapa se debe resolver un sistema del tipo $A_1 x = \widetilde{b}$, la idea es escoger la descomposición de A de modo tal que dicho sistema sea fácil de resolver, en particular, la matriz A_1 ha de ser invertible.

Una idea es comenzar por separar A de la forma:

$$A = D + L + U$$

donde D es la diagonal de A, L es la parte triangular inferior de A y U es la parte triangular superior de A. Supondremos que la diagonal D de A es invertible, esto es, que $a_{ii} \neq 0$ para todo $i = 1, \ldots, n$.

El *método de Jacobi* consiste en aplicar el método iterativo escogiendo la descomposición:

$$A_1 = D, \quad A_2 = L + U$$

por lo que el sistema a resolver en cada etapa es de la forma $Dx = \widetilde{b}$, que es muy fácil de resolver dividiendo por los a_{ii}.

Otra opción es la del *método de Gauss-Seidel*. que consiste en aplicar el método iterativo escogiendo la descomposición:

$$A_1 = D + L, \quad A_2 = U.$$

En este caso el sistema a resolver en cada iteración es de la forma $(D + L)x = \widetilde{b}$, que es posible resolver usando sustitución hacia adelante.

Es fácil ver que el algoritmo de Jacobi queda de la forma:

Método de Jacobi

$$x_i^{n+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^n \right).$$

y también es posible verificar que el algoritmo de Gauss-Seidel queda:

Método de Gauss-Seidel

$$x_i^{n+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^n - \sum_{j>i} a_{ij} x_j^{n+1} \right).$$

Es posible demostrar que el método de Gauss-Seidel converge a la solución del sistema lineal si la matriz A es simétrica y *definida positiva*[7]. También se puede demostrar

[7]Véase la Monografía *Excursiones por el Álgebra Lineal y sus Aplicaciones* en esta misma colección cf. [**13**].

que el método de Jacobi converge si además $2D - A$ es definida positiva. Las demostraciones son complicadas y no las trataremos en este texto.[8]

✍ **Ejercicio 4.7.** Convénzase de las formas de los dos métodos dados anteriormente.

✍ **Ejercicio 4.8.** ¿Por qué el método de Gauss-Seidel sería mejor, en principio, que el método de Jacobi?
Solución: pues en el método de Gauss-Seidel se tienen en cuenta para calcular x_i^{n+1} los valores ya calculados x_j^{n+1} con $j < i$.

✍ **Ejercicio 4.9.** Aplique los métodos de Jacobi y Gauss-Seidel para resolver el sistema lineal

$$A = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix}, \qquad b = \begin{pmatrix} 13 \\ 16 \\ 19 \\ 22 \end{pmatrix}.$$

¿Qué ocurre si aplica estos métodos al ejemplo del Ejercicio 4.6? *Solución: la solución al sistema es $x = (1, 2, 3, 4)^t$. Si se aplican estos métodos al ejemplo del Ejercicio 4.6 no hay convergencia. Esto se debe a que la matriz en ese caso no es simétrica ni definida positiva. Estos métodos no sirven para resolver sistemas con cualquier matriz.*

✍ **Ejercicio 4.10.** Investigue sobre los *métodos de relajación* consistentes en elegir la descomposición $A_1 = \frac{1}{\omega}D + L$, $A_2 = U - \frac{1-\omega}{\omega}D$ donde $\omega > 0$ es un cierto parámetro a escoger. Verifique que $A_1 + A_2 = A$. ¿Para qué valor de ω se recupera el método de Gauss-Seidel? Pruebe numéricamente en el ejemplo del ejercicio anterior para valores de $\omega < 1$ (sub-relajación), $\omega > 1$ (sobre-relajación) y $\omega = 1$. Compare la convergencia en los tres casos.

4.6 Sistemas mal puestos y condicionamiento

Consideremos el sistema lineal con lado derecho $b \in \mathbb{R}^n$:

$$Ax = b$$

y supongamos que A es invertible, por lo que hay una única solución x. En un computador no siempre conocemos b con toda la precisión necesaria, así es que consideremos una perturbación $b + \delta b$ del lado derecho y estudiemos la correspondiente perturbación $x + \delta x$ en la solución del sistema:

$$A(x + \delta x) = b + \delta b.$$

Por linealidad se tiene que

$$A\delta x = \delta b,$$

de donde el error relativo que se comete en la nueva solución con un lado derecho perturbado queda dado por

$$\epsilon = \frac{|\delta x|}{|x|},$$

[8] Véase la referencia [**24**].

donde hemos utilizado la norma euclidiana de un vector $x \in \mathbb{R}^n$:

$$|x| = \sum_{i=1}^{N} |x_i|^2.$$

Lo que nos interesa es acotar el error relativo ϵ en función de alguna propiedad de la matriz A. Esta propiedad es llamada condicionamiento y la definiremos a continuación.

Para ello, introduzcamos la siguiente norma de una matriz $A \in \mathbb{R}^{n\times n}$:

NORMA ESPECTRAL

$$||A|| = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|Ax|}{|x|},$$

llamada *norma espectral*. De la definición de supremo[9] es fácil ver que

$$|Ax| \leq ||A||\,|x|, \quad \forall x \in R^n$$

y que

$$|Ax| \leq C|x|,\ \forall x \in \mathbb{R}^n \qquad \Rightarrow \qquad ||A|| \leq C,$$

esto es, la norma espectral es la mejor constante posible al estimar $|Ax|$ a partir de $|x|$. Llamaremos *número de condicionamiento* de la matriz A al número:

NÚMERO DE CONDICIONAMIENTO

$$\chi(A) = ||A||\,||A^{-1}||, \qquad \chi(A) \in [1, \infty).$$

✍ **Ejercicio 4.11.** Demuestre que $\chi(A) \geq 1$.
Solución: notemos que

$$\begin{aligned}
|x| &\leq ||A^{-1}||\,|Ax| \\
\Rightarrow \quad \frac{|Ax|}{|x|} &\geq \frac{1}{||A^{-1}||} \\
\Rightarrow \quad \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|Ax|}{|x|} &\geq \frac{1}{||A^{-1}||} \\
\Rightarrow \quad ||A|| &\geq \frac{1}{||A^{-1}||} \\
||A||\,||A^{-1}|| &\geq 1.
\end{aligned}$$

[9] Dado un subconjunto E de $\mathbb{R}$ acotado superiormente, su supremo $\sup E$ es la mínima cota superior de E.

Sigamos ahora con el estudio del sistema perturbado $A(x+\delta x) = b + \delta b$. Usando que $|Ax| \leq ||A||\,|x|$ y que $|A^{-1}\delta b| \leq ||A^{-1}||\,|\delta b|$ se tiene que

$$\begin{aligned}
\frac{|\delta x|}{|x|} &= \frac{|A^{-1}\delta b|}{|x|} \\
&\leq ||A^{-1}||\,\frac{|\delta b|}{|x|} \\
&= ||A^{-1}||\,\frac{|\delta b|}{|x|}\frac{|Ax|}{|b|} \\
&\leq ||A^{-1}||\,\frac{|\delta b|}{|x|}\frac{||A||\,|x|}{|b|} \\
&= ||A||\,||A^{-1}||\,\frac{|\delta b|}{|b|},
\end{aligned}$$

de donde se obtiene finalmente que el error relativo en la solución se amplifica a partir del error relativo en el lado derecho mediante el factor $\chi(A)$.

Amplificación del error relativo

$$\frac{|\delta x|}{|x|} \leq \chi(A)\frac{|\delta b|}{|b|}.$$

Usemos la estimación precedente para explicar lo que se conoce como el condicionamiento de los sistemas lineales: si el número de condicionamiento es grande, el error relativo en la solución se amplifica ante un error relativo en el lado derecho y decimos que el sistema está *mal condicionado*, en cambio, si el número de condicionamiento es cercano a la unidad, el error relativo en la solución se mantiene o no empeora demasiado y decimos que el sistema está *bien condicionado*.

Notemos primero que para que esta estimación sea útil, debería haber alguna forma de conocer el número de condicionamiento $\chi(A)$ con facilidad. Si lo logramos, incluso antes[10] de resolver el sistema lineal, sabremos cuál es la amplificación máxima del error que podemos esperar en la solución, que en este caso viene dada por la constante $\chi(A)$.

Hay que retener entonces que el mal condicionamiento significa una amplificación extrema de los errores[11].

4.7 Un ejemplo de mal condicionamiento

Veamos un ejemplo concreto de un sistema mal condicionado:

$$\begin{aligned}
2,0000000\,x + 3,0000000\,y &= 5,0000000 \\
2,0000001\,x + 3,0000000\,y &= 5,0000000
\end{aligned}$$

[10] *a priori* en latín.

[11] Véase en el Capítulo 5 la inestabilidad numérica al resolver ecuaciones diferenciales o la propagación de errores en el Capítulo 1 que son conceptos relacionados.

cuya solución exacta es evidentemente:

$$x = 0, \qquad y = 5/3.$$

Si ahora perturbamos ligeramente el lado derecho del sistema:

$$\begin{aligned} 2,0000000\,x + 3,0000000\,y &= 5,0000000 \\ 2,0000001\,x + 3,0000000\,y &= 5,000000\mathbf{1} \leftarrow \end{aligned}$$

la nueva solución exacta es también evidente pero ¡completamente diferente!

$$x = 1, \qquad y = 1.$$

Hay en este caso una explicación geométrica simple del mal condicionamiento: se trata de dos rectas que se intersectan que son prácticamente paralelas, de modo que una ligera perturbación en su coeficiente de posición (lado derecho del sistema) cambia radicalmente el punto de intersección. Esto es, la solución del sistema es extremadamente sensible a un cambio del lado derecho.

✍ **Ejercicio 4.12.** Respecto al ejemplo de las rectas casi paralelas, es claro que si en vez de perturbarse los coeficientes de posición se perturban las pendientes de las rectas, el punto de intersección también cambia radicalmente. ¿A qué tipo de perturbación en el sistema lineal corresponde este caso? *Solución: a una perturbación en la matriz A.*

Lo del condicionamiento grande o pequeño es algo relativo, pero en la práctica funciona así: típicamente el error relativo asociado al lado derecho es como 10^{-p}, donde p es el número de cifras significativas que se utilizan en los cálculos internos del ordenador o calculadora de bolsillo y 10^{-p} es justamente la precisión máxima de los cálculos.[12] Así es que si el condicionamiento $\chi(A)$ es cercano a 10^p, el error relativo de la solución del sistema puede llegar a $10^p \times 10^{-p} = 1$, que corresponde a un 100 % de error.

En el ejemplo, el condicionamiento de la matriz del sistema es:

$$A = \begin{pmatrix} 2,0000000 & 3,0000000 \\ 2,0000001 & 3,0000000 \end{pmatrix} \qquad \chi(A) \approx 10^8,$$

y el error relativo del lado derecho es:

$$\frac{5 - 5,0000001}{5} \approx 10^{-8}$$

que corresponde a cálculos con 8 cifras significativas, uno diría que está a salvo de todo error al resolver un sistema lineal. Pero no es así, porque un condicionamiento de la matriz del sistema cercano a 10^8 indica que un error en el lado derecho en la diez-millonésima podría arrojar un error relativo de la solución cercano a

$$10^8 \times 10^{-8} = 1$$

[12] Véase el Capítulo 1.

lo que significa que el error puede llegar a ser del 100 % en la solución. Esto parece muy pesimista pero es exactamente lo que pasa en el ejemplo anterior: si nos equivocamos en la octava cifra significativa del lado derecho del sistema, provocamos un cambio en la primera cifra de la solución, esto es, un cambio completo del resultado. Afortunadamente, eso podía preverse de la estimación a priori debido al mal condicionamiento de A.

Ahora, si el condicionamiento es $\chi(A) \approx 10^q$ y el error relativo del lado derecho es 10^{-p}, entonces el error relativo de la solución es a lo más de $10^{-(p-q)}$. Esto es, el número de cifras significativas que no cambian entre la solución exacta y la solución perturbada está dado por:

$$p - q = p - \log \chi(A).$$

Esto quiere decir que, como el computador no es capaz de almacenar con precisión la cifra significativa p del lado derecho con buena precisión, la solución del sistema pierde $\log \chi(A)$ cifras exactas. En el ejemplo anterior, cualquiera sea la precisión con la que trabajemos, siempre podremos perder hasta 8 cifras significativas al resolver el sistema lineal con la matriz A.

4.8 Cálculo del condicionamiento

A partir de la discusión de la sección anterior, podemos dar ahora una definición más precisa del condicionamiento:

Condicionamiento de un sistema lineal

- Si $q = \log \chi(A)$ es cercano o mayor a la precisión p de la máquina, decimos que el sistema está mal condicionado y la solución se podría conocer con una o incluso ninguna cifra significativa.
- Si q es menor que la precisión p de la máquina, (en el mejor de los casos cercano a cero) decimos que el sistema está bien condicionado y el número de cifras significativas calculadas exactamente en la solución es aproximadamente $p - q$.

Hay que entender que la capacidad de predicción de una estimación *a priori* tiene un límite: el error podría ser menor que lo predicho en muchos casos, pero la estimación es *óptima* en el sentido que habrá algunos casos particulares donde el error se amplificará exactamente por el número de condicionamiento[13].

A veces se prefiere considerar el recíproco del número de condicionamiento

$$r(A) = \frac{1}{\chi(A)}, \qquad r(A) \in [0, 1].$$

[13]En análisis numérico, esto se conoce como una estimación óptima del error, en el sentido que el error dado por la cota calculada se alcanza en algún caso, o como el límite de una sucesión de casos.

condicionamiento del sistema	$\chi(A) \in [1, +\infty)$	$r(A) \in [0, 1]$	cifras significativas exactas en la solución (aprox.)
bueno	cercano a 1	cercano a 1	$p - \log \chi(A)$
malo	cercano a 10^p	cercano a 10^{-p}	una o ninguna

CUADRO 4.2. Interpretando el número de condicionamiento en un ordenador o calculadora que trabaja con p cifras significativas.

Si $r(A)$ se encuentra cercano a cero, o más precisamente, a la precisión de la máquina, se habla de un sistema mal condicionado y si $r(A)$ es cercano a uno, se habla un sistema bien condicionado. Todo lo anterior se resume en el Cuadro 4.2.

Pero ¿cómo calcular $\chi(A)$? Una forma es utilizar los llamados *valores y vectores propios de* A. Esto es, consideremos los n números complejos λ (no nulos pues A es invertible) llamados valores propios de A y n vectores x no nulos llamados vectores propios de A tales que:

$$Ax = \lambda x.$$

El conjunto de los valores propios de A conforman lo que se conoce como el *espectro* de A que es, en general, un subconjunto del plano complejo, pero en el caso en que A es simétrica, es un subconjunto de los reales.

Probaremos que, para matrices *simétricas e invertibles*, se tiene que:

ESTIMACIÓN DEL NÚMERO DE CONDICIONAMIENTO

$$\chi(A) = \frac{\max_{1 \le k \le n} |\lambda_k|}{\min_{1 \le k \le n} |\lambda_k|}.$$

Esto es, el condicionamiento está dado por la razón entre el máximo y el mínimo de los valores propios de A en valor absoluto. Esta fórmula es útil ya que permite estimar el condicionamiento de A sin necesidad de invertir la matriz A. Esto se debe a que se conocen métodos numéricos eficientes para calcular los valores propios de A y que no requieren invertir A. El Ejercicio 4.13 ilustra este hecho.

✍ **Ejercicio 4.13.** Investigue sobre el *método de la potencia* para encontrar los valores propios de una matriz simétrica A y aplíquelo a la matriz A del ejemplo de sistema mal puesto de la sección precedente. A partir de esto estime $\chi(A)$.
Solución: el método consiste en iterar, a partir de un vector x_0 no nulo dado:

$$x_{n+1} = \frac{Ax_n}{|Ax_n|}.$$

Entonces $x_n^t Ax_n / |x_n|^2$ converge al valor propio λ_1 mayor en módulo y x_n converge al vector propio asociado v_1. Para encontrar el siguiente valor y vector propio se aplica el mismo

método a $A_1 = A - \lambda_1 v_1 v_1^t/|v_1|^2$. [14] *Si se aplica a la matriz* A *del ejemplo de la sección precedente se obtiene* $\lambda_1 = 5 + 6 \times 10^{-8}$. *Como la suma de los valores propios debe ser* 5 *(traza de* A*) entonces* $\lambda_2 = -6 \times 10^{-8}$. *Entonces* $\chi(A) \leq \frac{|\lambda_1|}{|\lambda_2|} \approx 10^8$.

Para demostrar la estimación del número de condicionamiento, es más simple trabajar primero con los valores propios de $A^tA = A^2$, que por ser una matriz simétrica y definida positiva, tiene valores propios reales y positivos[15] que denotamos por:

$$0 < \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_n.$$

Estos valores σ_k son también llamados los *valores singulares* de A.

La idea es probar primero (ver Ejercicio 4.15) que la norma espectral de A está dada por

$$||A||^2 = \sigma_n$$

donde σ_n es el máximo de los valores singulares de A. Y de modo similar que

$$||A^{-1}||^2 = \frac{1}{\sigma_1}$$

donde σ_1 es el mínimo de los valores singulares de A. Si aceptamos esto, la fórmula para $\chi(A)$ viene ahora del hecho de que los valores propios de $A^t = A$ son los mismos que los de A, de donde si (λ, x) es par propio de A:

$$A^tAx = A^t(\lambda x) = A(\lambda x) = \lambda^2 x$$

esto es, los λ^2 son los valores singulares de A. Entonces, salvo reordenamiento, se tiene que

$$\sigma_k = |\lambda_k|^2$$

y con esto

$$\begin{aligned} \chi(A)^2 &= ||A||^2 \, ||A^{-1}||^2 \\ &= \frac{\sigma_n}{\sigma_1} \\ &= \frac{\max_{1\leq k\leq n} |\lambda_k|^2}{\min_{1\leq k\leq n} |\lambda_k|^2} \end{aligned}$$

de donde se obtiene la caracterización buscada.

Del razonamiento anterior se deduce también que el condicionamiento de A^tA es el cuadrado del condicionamiento de A (al menos para matrices simétricas e invertibles). Esto es importante al momento de resolver las llamadas ecuaciones normales (ver más adelante).

---

[14] La justificación de este método está fuera de los contenidos de este texto. Para más detalles véanse las referencias **[7]**, **[24]**.

[15] Véase la Monografía *Excursiones por el Álgebra Lineal y sus Aplicaciones* en esta misma colección cf. **[13]**.

✍ **Ejercicio 4.14.** Si el número de condicionamento de la matriz

$$A = \begin{pmatrix} 1,0000 & 1,0000 \\ 1,0001 & 1,0000 \end{pmatrix}$$

está entre 10^4 y 10^5, y resolvemos los sistemas

$$Ax_1 = b_1 \qquad Ax_2 = b_2$$

con

$$b_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad b_2 = \begin{pmatrix} 1,000001 \\ 1 \end{pmatrix}.$$

¿Puede predecir aproximadamente cuántas cifras significativas tendrán iguales las respectivas soluciones x_1 y x_2? *Solución: el error relativo del lado derecho es de* 10^{-6} *de modo que está entre* $6-4=2$ *y* $6-5=1$ *cifras, eso es, son* 1 *ó* 2 *cifras iguales. En efecto, resolviendo los sistemas se obtiene* $x_1 = (0,1)$, $x_2 = (0{,}01, 1{,}01)$ *y solamente las dos primeras cifras son iguales en las dos soluciones.*

✍ **Ejercicio 4.15.** La idea de este ejercicio es probar la siguiente igualdad para la norma espectral de una matriz A invertible:

$$||A|| = \sqrt{\sigma_n}, \qquad ||A^{-1}|| = \frac{1}{\sqrt{\sigma_1}},$$

donde σ_n y σ_1 son respectivamente el máximo y mínimo de los valores singulares de A. Para ello siga los siguientes pasos:

- Primero pruebe el resultado de $||A||$ para una matriz $A = D$ diagonal.
- Ahora considere la descomposición de A^tA como

  $$A^tA = PDP^t$$

  donde D es una matriz diagonal con los valores singulares en la diagonal principal y P es una matriz unitaria, esto es, tal que $PP^t = I$[16]. Reemplazando esto en la expresión

  $$||A||^2 = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|Ax|^2}{|x|^2} = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^tA^tAx}{x^tx}$$

  y haciendo el cambio de variables $y = Px$ reduzca todo al caso diagonal y concluya.
- Usando que los valores propios de A^{-1} son los recíprocos de los valores propios de A deduzca la fórmula para $||A^{-1}||$.

[16] Véase la Monografía *Excursiones por el Álgebra Lineal y sus Aplicaciones* en esta misma colección cf. **[13]**.

4.9 Sistemas sobredeterminados y mínimos cuadrados

Consideremos un sistema lineal en el que tenemos más ecuaciones que incógnitas. La matriz de un tal sistema tiene más filas que columnas:

$$Ax = b, \qquad A \in \mathbb{R}^{m\times n}, \qquad m > n$$

el que denominamos como sistema lineal *sobredeterminado*.

En vez de resolver $Ax = b$ (que en general no tiene solución) la idea es resolver el problema de *aproximación por mínimos cuadrados*:

Problema de mínimos cuadrados

$$\text{Encontrar } x \text{ que minimice:} \qquad \min_{x\in\mathbb{R}^n} \frac{1}{2}|Ax - b|^2.$$

Probaremos que si A satisface la condición:

$$Ax = 0 \quad \Rightarrow \qquad x = 0,$$

esto es, si A es de rango completo n (por ejemplo esto se tiene si hay n ecuaciones linealmente independientes en el sistema[17]) entonces la solución a este problema existe, es única y está dada por

Solución de mínimos cuadrados

$$A^t Ax = A^t b.$$

Como A es de rango n, además se tiene que A^tA es una matriz de $n \times n$ invertible, lo que nos provee la solución[18]:

$$x = (A^tA)^{-1}A^t b$$

donde

Pseudoinversa de Penrose

$$A^\dagger = (A^tA)^{-1}A^t.$$

se denomina *pseudoinversa* de A o *pseudoinversa de Penrose* de A, ya que es inversa por la derecha de A (que no es cuadrada!): $AA^\dagger = I$. Con esto la solución de mínimos cuadrados viene dada por

$$x = A^\dagger b.$$

Para demostrar el resultado, primero introduzcamos la notación de producto interno:

$$x^t y = (x, y), \qquad \forall x, y \in \mathbb{R}^m.$$

[17] ídem.

[18] Recordemos que la inversión es sólo formal, porque de un punto de vista numérico es más eficiente resolver el sistema $A^tAx = A^t b$ por un método eficiente como el de Gauss por ejemplo, sin necesidad de calcular la inversa de A^tA.

Con esta notación, tenemos que

$$|x|^2 = (x, x), \quad (x, y) = (y, x)$$

y que

$$(Ax, y) = (x, A^t y), \qquad \forall x \in \mathbb{R}^n,\ y \in \mathbb{R}^m,\ A \in \mathbb{R}^{m\times n}.$$

Con esto, sea x la solución de $A^t Ax = A^t\, b$, que existe pues a causa de la condición de rango completo, la matriz cuadrada $A^t A$ también es de rango completo, por lo que resulta invertible. Sea ahora $y \neq x$ otro vector cualquiera en $\mathbb{R}^n$. Calculemos entonces

$$\begin{aligned} |Ax - y|^2 - |Ax - b|^2 &= (Ay - b, Ay - b) - (Ax - b, Ax - b) \\ &= (Ay, Ay) - 2(Ay, b) + (b, b) - (Ax, Ax) + 2(Ax, b) - (b, b) \\ &= (Ay, Ay) - 2(Ay, b) - (Ax, Ax) + 2(Ax, b) \end{aligned}$$

pero usando que

$$\begin{aligned} (Ay, b) &= (y, A^t\, b) = (y, A^t A\, x) = (Ay, Ax) \\ (Ax, b) &= (x, A^t\, b) = (x, A^t A\, x) = (Ax, Ax) \end{aligned}$$

se obtiene que

$$\begin{aligned} |Ax - y|^2 - |Ax - b|^2 &= (Ay, Ay) - 2(Ay, Ax) - (Ax, Ax) + 2(Ax, Ax) \\ &= (Ax, Ax) + (Ay, Ay) - 2(Ax, Ay) \\ &= |Ax - Ay|^2 \end{aligned}$$

esto es

$$|Ax - b|^2 = |Ay - b|^2 - |A(x - y)|^2.$$

De la condición de rango, como $x - y \neq 0$ implica que $A(x - y) \neq 0$ se tiene que

$$\frac{1}{2}|Ax - b|^2 < \frac{1}{2}|Ay - b|^2, \qquad \forall y \in \mathbb{R}^n$$

y esto es decir exactamente que en x se alcanza el mínimo.

✍ **Ejercicio 4.16.** Pruebe que bajo la hipótesis de rango completo la solución de mínimos cuadrados es única. Para ello considere otra solución distinta que también minimice $|Ax - b|^2$ y use la relación que encontramos anteriormente para llegar a una contradicción.

✍ **Ejercicio 4.17.** La idea de este problema es recuperar las fórmulas clásicas de la *regresión lineal* [19] Consideremos una cierta relación lineal de la forma

$$y(x) = ax + b.$$

Determine los parámetros a y b a partir de una serie de datos de la forma:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

de manera de minimizar el error cuadrático $\sum_{i=1}^n |y_i - y(x_i)|^2$.
Solución:

[19] Véase la Monografía *Estadística Multivariada* en esta misma colección cf. **[17]**.

- *Lleve el problema a un sistema lineal sobredeterminado con incógnitas (a, b) cuya matriz A tiene la forma:*

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}.$$

- *Plantee la solución de mínimos cuadrados del sistema anterior y encuentre que el sistema final a resolver es:*

Sistema de Regresión Lineal

$$\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum x_i\, y_i \\ \sum y_i \end{pmatrix}.$$

- *Resolviendo el sistema anterior, encuentre las fórmulas de regresión lineal que vienen incluídas en la mayoría de las calculadoras científicas:*

$$a = \frac{n\sum x_i\, y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}, \qquad b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i\, y_i)(\sum x_i)}{n\sum x_i^2 - (\sum x_i)^2}.$$

✍ **Ejercicio 4.18.** Considere la ley de desaparición natural de un cierto contaminante en la atmósfera que dejó de ser producido por la actividad antropogénica y cuya concentración decrece siguiendo la ley exponencial:

$$c(t) = ae^{-\sigma t}, \qquad t \geq 0.$$

Determine los parámetros a y $\sigma > 0$ a partir de una serie de mediciones de la concentración (mucho más que dos de ellas) en tiempos $t_i \geq 0$ de la forma:

$$(t_1, c_1),\ (t_2, c_2),\ \ldots,\ (t_n, c_n)$$

de modo tal de minimizar $\sum_{i=1}^{n} |c_i - c(t_i)|^2$.
Solución: Tomando logaritmo lleve el problema a la forma de un sistema lineal del tipo

$$\begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_n & 1 \end{pmatrix} \begin{pmatrix} -\sigma \\ \ln a \end{pmatrix} = \begin{pmatrix} \ln c_1 \\ \ln c_2 \\ \vdots \\ \ln c_n \end{pmatrix}$$

y plantee la solución por mínimos cuadrados de este sistema.

4.10 Ejemplo numérico: la matriz mágica de Durero

En el siglo XVI, Alberto Durero inmortalizó en uno de sus grabados una matriz mágica de 4×4 con 16 números enteros. Se le llama *matriz mágica* pues las sumas por filas, columnas y diagonales principales es siempre igual a 34, llamado *número mágico* de la matriz:

$$D = \begin{array}{|c|c|c|c|}\hline 16 & 3 & 2 & 13 \\ \hline 5 & 10 & 11 & 8 \\ \hline 9 & 6 & 7 & 12 \\ \hline 4 & \mathbf{15} & \mathbf{14} & 1 \\ \hline \end{array}$$

De manera anecdótica, notemos que, en la parte inferior de esta matriz que llamaremos D, las cifras coinciden con la fecha del grabado de Durero en el año 1514.

Imagínese ahora que todos los números de esta matriz D fueran desconocidos, pero que sabemos que las sumas por filas, columnas y diagonales principales es 34:

$$\begin{array}{|c|c|c|c|}\hline ? & ? & ? & ? \\ \hline ? & ? & ? & ? \\ \hline ? & ? & ? & ? \\ \hline ? & ? & ? & ? \\ \hline \end{array}\quad \begin{array}{l}\rightarrow 34\\ \rightarrow 34\\ \rightarrow 34\\ \rightarrow 34\end{array}$$

$$\begin{array}{ccccccc} \swarrow & \downarrow & \downarrow & \downarrow & \downarrow & \searrow & \\ 34 & 34 & 34 & 34 & 34 & & 34 \end{array}$$

Cada entrada de la matriz es una incógnita, digamos x_1, x_2 hasta x_{16} si numeramos de izquierda a derecha por filas de arriba a abajo, de modo que hay 16 incógnitas. Cada suma representa una ecuación lineal, por ejemplo, la suma de la primera fila y de la primera columna corresponde a las ecuaciones:

$$x_1 + x_2 + x_3 + x_4 = 34, \qquad x_1 + x_5 + x_9 + x_{13} = 34.$$

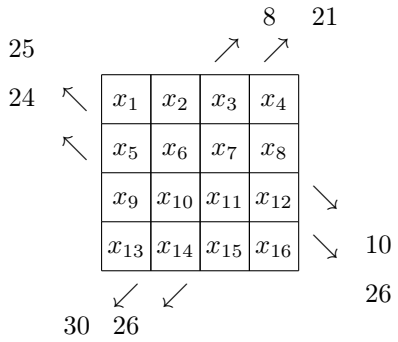
Como hay 10 sumas, esto corresponde a 10 ecuaciones para 16 incógnitas. Se puede entonces escribir un sistema lineal *subdeterminado*, pero podría haber muchas soluciones, esto es, muchas matrices mágicas D que tienen las misma suma 34 por filas, por columnas y por diagonales principales.[20]

$$\begin{array}{|c|c|c|c|}\hline x_1 & x_2 & x_3 & x_4 \\ \hline x_5 & x_6 & x_7 & x_8 \\ \hline x_9 & x_{10} & x_{11} & x_{12} \\ \hline x_{13} & x_{14} & x_{15} & x_{16} \\ \hline \end{array}\quad \begin{array}{l}\rightarrow 34\\ \rightarrow 34\\ \rightarrow 34\\ \rightarrow 34\end{array}$$

$$\begin{array}{ccccccc} \swarrow & \downarrow & \downarrow & \downarrow & \downarrow & \searrow & \\ 34 & 34 & 34 & 34 & 34 & & 34 \end{array}$$

Podemos agregar entonces más ecuaciones suponiendo, por ejemplo, que las sumas por otras 8 diagonales son también conocidas:

[20] Aunque no todas necesariamente con entradas enteros positivos.



Ahora tenemos 18 ecuaciones y 16 incógnitas, sistema lineal que se puede escribir de la forma (véase el Ejercicio 4.20)

$$Ax = b$$

donde A es una matriz de 18 filas y 16 columnas. Este sistema lineal estaría en principio *sobredeterminado*. Sin embargo, si intentamos calcular la solución al sistema por mínimos cuadrados:

$$A^t Ax = A^t b$$

resulta imposible, ya que de hecho $A^t A$ no es invertible. Esto ocurre pues la matriz A no tiene rango completo (se puede verificar que tiene rango 15 y el número de columnas es 16). En este caso la solución al sistema no es única, a pesar de que hay más ecuaciones que incógnitas. Por ejemplo, es fácil verificar que la siguiente matriz D' tiene exactamente las mismas sumas por las 5 filas, 5 columnas y 8 diagonales que la matriz de Durero D:

$$D' = \begin{array}{|c|c|c|c|}\hline 16 & 15 & -10 & 13 \\ \hline -7 & 10 & 11 & 20 \\ \hline 21 & 6 & 7 & 0 \\ \hline 4 & 3 & 26 & 1 \\ \hline \end{array}$$

Hay una forma, sin embargo, de forzar la solución a ser única. Se trata de resolver el sistema aproximado:

$$(A^t A + \varepsilon I)x^\varepsilon = A^t b$$

donde ε es un número pequeño e I es la matriz identidad. La solución es única pues en este caso la matriz $(A^t A + \varepsilon I)$ sí es invertible. Este problema corresponde a minimizar

Problema de mínimos cuadrados regularizado

Encontrar x^ε que minimice: $\min_{x\in\mathbb{R}^n} \frac{1}{2}|Ax-b|^2 + \frac{\varepsilon}{2}|x|^2$.

Si se hace el cálculo con $\varepsilon = 10^{-6}$, el resultado x^ε da aproximadamente las entradas de ¡la matriz D de Durero! con errores del orden de 10^{-4}. Este método se conoce como

método de regularización de Tikhonov. Esto quiere decir que la matriz de Durero, entre todas aquellas matrices que tienen iguales sumas como las requeridas, tiene entradas cuya suma de cuadrados es mínima. Por ejemplo, la matriz D tiene suma de cuadrados $16^2 + 3^2 + 2^2 + 13^2 + \ldots = 1496$ y la matriz D' suma de cuadrados $16^2 + 15^2 + (-10)^2 + 13^2 + \ldots = 2648$.

✍ **Ejercicio 4.19.** Pruebe, como en la sección anterior, pero sin necesidad de suponer que A es de rango completo, que resolver el sistema

$$(A^t A + \varepsilon I)x^\varepsilon = A^t b$$

es equivalente a minimizar

$$\frac{1}{2}|Ax - b|^2 + \frac{\varepsilon}{2}|x|^2$$

concluya que ambos problemas tienen una solución única.

✍ **Ejercicio 4.20.** Encuentre explícitamente el sistema lineal al que se hace referencia en esta sección. *Solución: El sistema de ecuaciones se puede escribir de la siguiente forma:*

$$\begin{pmatrix} 1111000000000000 \\ 0000111100000000 \\ 0000000011110000 \\ 0000000000001111 \\ 1000100010001000 \\ 0100010001000100 \\ 0010001000100010 \\ 0001000100010001 \\ 1000010000100001 \\ 0001001001001000 \\ 0100100000000000 \\ 0010010010000000 \\ 0010000100000000 \\ 0100001000010000 \\ 0000000000010010 \\ 0000000100100100 \\ 0000000010000100 \\ 0000100001000010 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \\ x_{16} \end{pmatrix} = \begin{pmatrix} 34 \\ 34 \\ 34 \\ 34 \\ 34 \\ 34 \\ 34 \\ 34 \\ 34 \\ 34 \\ 8 \\ 21 \\ 10 \\ 26 \\ 26 \\ 30 \\ 24 \\ 25 \end{pmatrix}$$

✍ **Ejercicio 4.21.** Si toma en promedio de los elementos entre la matriz de Durero D y la que tiene iguales sumas y se presentó en el texto D', ¿se obtiene una nueva matriz que también tiene las mismas sumas?, ¿se extiende esto a todas las combinaciones lineales de ambas matrices? *Solución: solamente a las combinaciones lineales convexas, esto es, aquellas de la forma* $\theta D + (1-\theta)D'$ *con* $0 \leq \theta \leq 1$.

✍ **Ejercicio 4.22.** ¿Cree usted que si se restringe a las incógnitas $x_1, \ldots, x_{16}$ a ser una permutación de los 16 primeros enteros, la matriz D de Durero es la única matriz mágica con número mágico 34? *Solución: No, basta considerar las matrices que se obtienen de intercambiar las columnas 2 y 3 de* D *o las filas 2 y 3 de* D.

✍ **Ejercicio 4.23.** ¿Se le ocurre una manera de expresar algunas de las reglas del popular juego *sudoku* en forma de un sistema lineal?

4.11 Ejemplo numérico: tomografía computarizada

Utilicemos lo que hemos aprendido hasta ahora para desarrollar la aplicación siguiente. Se trata de la detección de defectos por rayos X, comúnmente llamada *tomografía computarizada*. No es lo mismo que las radiografías tradicionales que usted conoce y que son algo así como sombras obtenidas al irradiar con rayos X un cuerpo. La tomografía computarizada es una técnica más sofisticada y se utiliza en medicina para obtener imágenes aún más detalladas que pueden servir para detectar tumores, aneurismas, hemorragias cerebrales, cálculos renales, entre otros. También se utiliza la misma técnica en ciertos microscopios de rayos X.

El principio fundamental para obtener una tomografía es en realidad muy simple y es el mismo presentado en la sección anterior, cuando buscábamos las entradas de una matriz a partir de las sumas por filas, columnas y diagonales.

Imagínese una sección bidimensional C del cuerpo humano, que puede ser por ejemplo una sección transversal del tórax a nivel del corazón. Esta sería como la matriz de números. Ella está conformada por distintos tejidos y fluidos, cada uno de los cuales tiene una atenuación α de rayos X distinta que lo caracteriza. Estos valores de α representarían los números desconocidos de la matriz. Por ejemplo, una porción de tejido pulmonar con un tumor atenuará los rayos X más de lo normal. La idea es encontrar la atenuación $\alpha(x)$ en cada punto x de C, y de este modo obtener un mapa coloreado del interior, donde cada color represente un tejido o fluido distinto y se puedan identificar las anomalías.

Para ello se coloca al paciente acostado boca arriba y se le irradia con rayos X de leve intensidad haciendo girar a su alrededor un emisor de rayos. Piense que uno de estos rayos X atraviesa el cuerpo C siguiendo una trayectoria recta L. El rayo parte del emisor con una intensidad inicial I_0 emergiendo luego de C con una intensidad final I que es medida por un receptor. Se sabe que el rayo pierde intensidad a medida que atraviesa C proporcionalmente a la atenuación del medio, esto es, si $\alpha = \alpha(x)$, se sabe que:

$$I = I_0 \exp(-\int_{L\cap C} \alpha(x)\, dx),$$

donde se integra la opacidad sobre la parte de la recta L que intersecta el cuerpo C. En definitiva, el aparato de rayos infrarrojos que tiene un emisor que mide I_0 y un receptor que mide I (ambos giran sincronizados en un tambor), y entonces se puede calcular menos el logaritmo de la razón de pérdida de intensidad:

$$-\ln\left(\frac{I}{I_0}\right) = \int_{L\cap C} \alpha(x)\, dx,$$

que coincide con el valor numérico de la integral de la atenuación α en la porción de recta que atraviesa el cuerpo: $L \cap C$.

Como no es uno, sino que son muchos los rayos que atraviesan C en distintas direcciones y desde distintas posiciones a medida que el aparato va girando, entonces, se pueden conocer las integrales de $\alpha(x)$ en muchas porciones de recta $L \cap C$

(ver Figura 4.2 izquierda). Estas serían las sumas por diagonales que se conocen. El problema es pues recuperar la función $\alpha(x)$ a partir de sus integrales conocidas.[21]

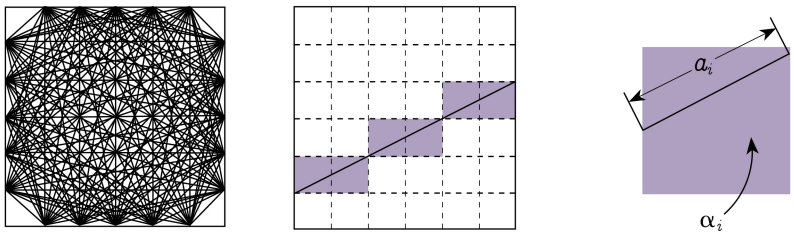


FIGURA 4.2. Izquierda: haz de rayos X en el caso de 5 emisores-receptores por lado ($n = 6$). Centro: cada rayo X atraviesa un conjunto de cuadritos $c_1, \ldots, c_m$. Derecha: la integral en cada cuadrito se aproxima por $a_i\alpha_i$, donde α_i son las incógnitas.

Para simplificar, supongamos que C es un cuadrado de lado 1 el que cuadriculamos en $n \times n$ cuadrados iguales, con $n \geq 2$. En cada lado de C, salvo en las esquinas, suponemos que hay $n-1$ emisores-receptores uniformemente repartidos como en la Figura 4.2, izquierda. Si suponemos que cada emisor-receptor se conecta con los emisores-receptores de los otros lados, hay

$$3(n-1)(n-1) + 2(n-1)(n-1) + (n-1)(n-1) = 6(n-1)^2 \text{ rayos X.}$$

En efecto, los primeros $n-1$ se conectan con los $3(n-1)$ de lo otros 3 lados y luego de esto los $n-1$ del segundo lado se conectan con los $2(n-1)$ de los dos lados restantes y finalmente los $n-1$ del tercer lado se conectan con los últimos $n-1$ del cuarto lado.

Suponemos ahora que en cada cuadrito de C la atenuación α es constante, entonces tenemos:

$$n^2 \text{ valores de } \alpha \text{ en } C.$$

Pero cuando un rayo L atraviesa C, va intersectando una serie de cuadritos pequeños $c_1, c_2, \ldots, c_m$ (ver Figura 4.2, centro). Podemos entonces aproximar la integral de α en L como la suma de las contribuciones sobre cada cuadrito:

$$\int_L \alpha(x)\, dx = a_1\alpha_1 + a_2\alpha_2 + \ldots + a_m\alpha_m,$$

donde a_i es el largo del segmento de recta $L \cap c_i$ y α_i es el valor (desconocido) de la atenuación en c_i (ver Figura 4.2, derecha).

[21] Problema conocido en matemáticas como el problema de la *transformada de Radón*.

Como las integrales $\int_L \alpha(x)dx$ son conocidas, tenemos una ecuación lineal por cada rayo X. Por otro lado, las incógnitas son los valores α_i en cada cuadrito c_i. Esto es hay:

$$n^2 \text{ incógnitas y } 6(n-1)^2 \text{ ecuaciones.}$$

Para $n \geq 2$ se tiene que $6(n-1)^2 > n^2$ de modo que hay siempre más ecuaciones que incógnitas, esto es, el sistema lineal está sobredeterminado. Si A es la matriz formada por los coeficientes a_i, el vector x es el de las incógnitas α_i y b es el vector de las integrales conocidas, entonces el sistema sobredeterminado es el siguiente:

$$Ax = b, \qquad \text{con } A \text{ de } 6(n-1)^2 \times n^2.$$

Buscamos entonces la solución de mínimos cuadrados (la que minimiza $\frac{1}{2}|Ax-b|^2$) que se obtiene finalmente resolviendo:

$$A^t Ax = A^t b.$$

La Figura 4.3 muestra un ejemplo de tomografía computarizada utilizando la técnica explicada anteriormente. Se utilizaron 9 emisores-receptores por lado y la solución de mínimos cuadrados del sistema sobredeterminado con 100 incógnitas y $6 \times 81 = 486$ ecuaciones. ❒✍

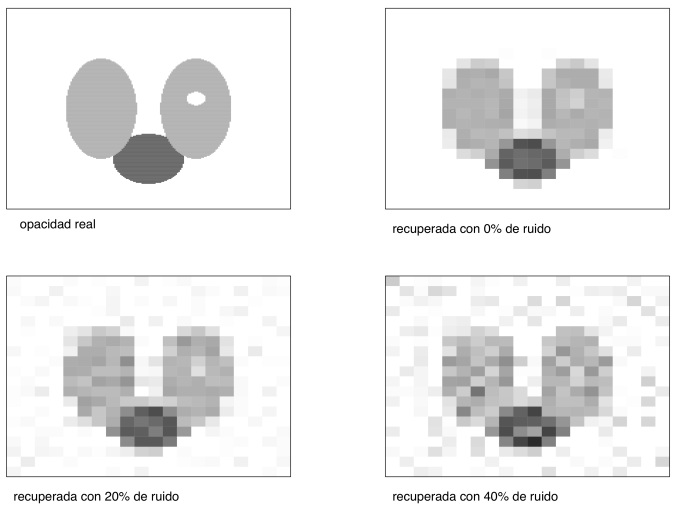


FIGURA 4.3. Recuperación de la atenuación en medio que simula dos pulmones, un corazón y un pequeño tumor. La solución se deteriora al aumentar el ruido del lado derecho b que impone un límite al tamaño del mínimo tumor detectable.

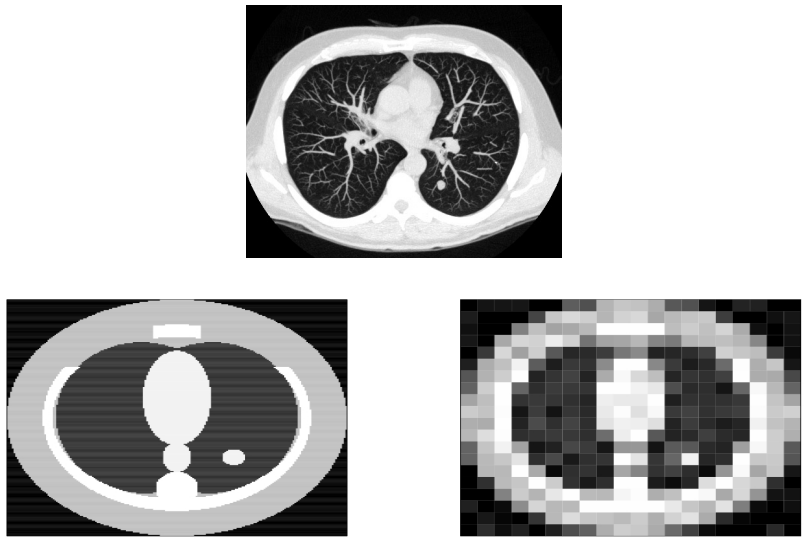


FIGURA 4.4. Izquierda: corte transversal de un paciente acostado boca arriba y mostrando un tumor en el pulmón y obtenido gracias a una tomografía computarizada de alta resolución profesional (fuente: `www.isi.uu.nl/Education/Projects/nodulesize`). Centro y derecha: dominio simulado y reconstrucción con el método visto con $n = 20$ y $m = 20$ con un 5 % de ruido en las mediciones.

✍ **Ejercicio 4.24.** Investigue sobre Johann Radón. Investigue sobre otras técnicas de inspección no invasiva y más inofensivas que los rayos X como son la tomografía por impedancia eléctrica (EIT) y la ecotomografía. Investigue sobre los trabajos del matemático argentino Alberto Calderón y relacionados con la EIT.

✍ **Ejercicio 4.25.** Averigue sobre la investigación sobre invisibilidad y metamateriales y sobre los trabajos del matemático chileno Gunther Uhlmann. Se trata, al contrario de la tomografía, esconder un medio en vez de descubrirlo. Por ejemplo, encuentre una matriz de 4×4 en que un cambio de los elementos internos de la matriz no altere la suma por filas y por columnas.
Solución: por ejemplo, el siguiente cambio es invisible a la suma por filas o columnas:

1	*1*	*1*	*1*
1	*1*	*1*	*1*
1	*1*	*1*	*1*
1	*1*	*1*	*1*

$\rightarrow$

1	*1*	*1*	*1*
1	*2*	*0*	*1*
1	*0*	*2*	*1*
1	*1*	*1*	*1*

Capítulo 5: ¿Cómo y por qué resolver ecuaciones diferenciales?

> *"En efecto, toda la dificultad de la filosofía parece consistir en buscar las fuerzas de la naturaleza a partir de los fenómenos de movimiento que ellas producen y a demostrar en seguida otros fenómenos a partir de estas fuerzas."* I. Newton, *(De Philosophiæ Naturalis Principia Mathematica).*

5.1 ¿Por qué plantear ecuaciones diferenciales?

En 1687, Isaac Newton publicaría los principios matemáticos de lo que él llamaría la *filosofía natural* en su obra *Philosophiæ Naturalis Principia Mathematica* y que sentaron las bases de la física y de la astronomía. En gran parte de su teoría, Newton se basó en que muchos fenómenos de la naturaleza pueden entenderse a través de relaciones entre el movimiento, sus variaciones y las fuerzas que lo producen.

En términos matemáticos, Newton estaba utilizando identidades que relacionan una función y sus derivadas[1], esto es, trabajaba por primera vez con *ecuaciones diferenciales*, para lo cual requirió desarrollar los elementos del cálculo diferencial e integral. [2]

Aunque Newton ocultó bajo un lenguaje geométrico sus cálculos con ecuaciones diferenciales al momento de publicar *Principia*[3], debido posiblemente a que el cálculo diferencial e integral era un conocimiento nuevo y suceptible de ser cuestionado, la facilidad con que las ecuaciones diferenciales podían expresar las leyes de la física se impuso rápidamente en el estudio de los fenómenos naturales.

Un ejemplo emblemático es el de la segunda ley de Newton. Luego de que Kepler encontrara leyes empíricas a partir de la tablas de Tycho Brahe y estableciera que las órbitas de los planetas eran elípticas, todo ello sería deducible de la segunda ley de Newton combinada con la ley de gravitación universal. La aceleración de un planeta (es decir, la segunda derivada de su posición) es proporcional a la fuerza que sobre él ejerce el sol, cuya magnitud es inversa al cuadrado de la distancia que los separa. Esta ecuación del siglo XVII estipulada en el *Principia*, tiene como solución elipses en el caso de un planeta que gira alrededor de un Sol masivo. Pero en realidad,

[1] *fluxiones* en la terminología original de Newton, que refleja además la idea de continuo.

[2] Cálculo revolucionario para la ciencia introducida en el siglo XVII no solamente por Newton, sino que también por Fermat y Leibniz

[3] Efectivamente, al hojear el libro no se encuentra ninguna ecuación diferencial.

también ejercen fuerza sobre él los demás planetas, lo que lleva a órbitas muchísimo más complicadas y al estudio posterior de órbitas caóticas en el siglo XX por Poincaré entre otros. Hoy en día, la única manera de calcular dichas órbitas con precisión es a través de cálculos numéricos que resuelven las ecuaciones diferenciales involucradas. Por ejemplo, si hay 8 planetas, se trata de un sistema de $(3+3)\times 8 = 48$ ecuaciones diferenciales, pues para cada planeta hay que encontrar las tres coordenadas de su posición y las tres coordenadas de su velocidad en el espacio.

Hoy en día, el análisis numérico de ecuaciones diferenciales y su resolución por computador, son una herramienta fundamental y omnipresente para la comprensión y relación con el mundo que nos rodea. Éste se usa, para citar otros dos ejemplos importantes, para el estudio de poblaciones (crecimiento, recursos naturales, epidemias) y para realizar simulaciones atmosféricas (pronóstico meteorológico, cambio climático, calidad del aire).

En este capítulo, primero estudiaremos algunos métodos numéricos para resolver ecuaciones diferenciales y revisaremos algunos conceptos básicos: el orden de un método, la noción de estabilidad y la diferencia entre métodos explícitos e implícitos. Luego, aplicaremos los algoritmos aprendidos a la resolución de ecuaciones diferenciales ordinarias para estudiar diversos aspectos de la dinámica de poblaciones: crecimiento, epidemias y competencia.

La mayoría de las veces, las ecuaciones diferenciales planteadas parten de consideraciones simples e intuitivas de la realidad y, sin embargo, nos llevan a analizar y cuantificar situaciones complejas que están lejos de la comprensión inmediata, lo que nos ayuda a reinterpretar con profundidad la realidad que las originó. El análisis numérico provee de algoritmos apropiados que, una vez implementados en un computador, nos permite simular dichos modelos en un verdadero laboratorio virtual.

5.2 Discretizando el problema de Cauchy

Para comenzar, consideremos una ecuación diferencial ordinaria con condición inicial, lo que se conoce como el *problema de Cauchy*. Es éste el problema que nos interesa resolver numéricamente [4]:

Problema de Cauchy o de valor inicial

$$\begin{aligned} x'(t) &= f(x(t),t), \qquad x\in\mathbb{R}, \quad t\in[0,T]. \\ x(0) &= x_0 \text{ dado.} \end{aligned}$$

donde x' denota la derivada temporal de la función x, y la función f la suponemos continua en sus dos variables. El tiempo total de evolución es $T>0$ y $x(0)$ es la *condición inicial* de partida en $t=0$.

Para fijar ideas pensemos en x como una función del tiempo a valores escalares, y f una función de $\mathbb{R}\times\mathbb{R}$ en $\mathbb{R}$, pero todos los algoritmos de resolución que veremos a

[4] Llamado así en honor al matemático francés Auguste Louis Cauchy (1760-1848) quien estudiara existencia y unicidad de solución para dicho problema.

continuación se aplican también al caso de sistemas de ecuaciones diferenciales donde $x \in \mathbb{R}^n$, $n \geq 1$ y $f : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^n$, por lo que la ecuación resulta vectorial. Este es el caso de una ecuación con dos derivadas $y'' = g(y,t)$, por ejemplo, que se puede reducir a un sistema vectorial de dos ecuaciones ($n = 2$) con una sola derivada:

$$x \equiv (y, y'), \quad x' = (y', y'') = (y', g(y,t)) \equiv f(x,t)$$

También es el caso del modelo epidemiológico SIR, o del modelo de Lotka-Volterra que se verán más adelante, en que $n = 3$.

También para fijar ideas, supongamos que bajo ciertas hipótesis adicionales[5]sobre f, se tiene que para cada condición inicial $x(0)$, el problema de Cauchy anterior tiene una única solución continua definida en el intervalo $[0,T]$. Supondremos pues que dicha solución existe, es única y es continua:

$$t \in [0,T] \quad \rightarrow \quad x(t) \in \mathbb{R}.$$

Nuestro objetivo es aproximar dicha solución, utilizando métodos numéricos. La primera idea fundamental es considerar la *forma integral de la ecuación* que se obtiene justamente integrando el problema de Cauchy entre 0 y t:

Forma integral del Problema de Cauchy

$$x(t) = x(0) + \int_0^t f(x(s), s)\, ds, \qquad t \in [0,T].$$

Luego, para aproximar la solución $x(t)$, subdividimos el intervalo temporal $[0,T]$ en N subintervalos $[t_n, t_{n+1}]$ de largo h:

$$h = \frac{T}{N} > 0$$

y si unimos todos los subintervalos[6] nos queda:

$$[0,T] = [t_0, t_1] \cup [t_1, t_2] \cup \ldots \cup [t_{N-1}, t_N]$$

donde

$$\begin{aligned} t_0 &= 0 \\ t_{n+1} &= t_n + h, \quad n = 0, \ldots, N-1. \end{aligned}$$

Ahora escribimos la forma integral de la ecuación diferencial en el intervalo $I_n \equiv [t_n, t_{n+1}]$ como:

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(x(s), s)\, ds \tag{5.1}$$

[5]Véase la Monografía *Modelando Fenómenos de Evolución* en esta misma colección cf. **[12]**.
[6]Por simplicidad, se considera aquí un paso fijo h, pero podría ser un paso variable h_n.

En seguida, todos los métodos que veremos consisten esencialmente en aproximar el valor de la función incógnita x en los puntos t_n por una sucesión x_n:

$$x_n \approx x(t_n)$$

y reemplazar el término integral de la derecha en (5.1) por una fórmula de cuadratura o de integración numérica de las que estudiamos en el Capítulo 3, haciendo los ajustes que sean necesarios para encontrar un algoritmo realizable en la práctica.

5.3 Orden del algoritmo. Error local y global.

Cualquiera sea el algoritmo, nos interesa que los *errores de discretización globales*

$$e_n = x_n - x(t_n)$$

sean pequeños, esto es, que la solución discretizada se mantenga cerca de la solución exacta. Sin embargo, los errores se van acumulando[7] al pasar de un intervalo I_n al siguiente I_{n+1}. Para saber cuál es el error de discretización local en cada intervalo I_n, hay que considerar el problema de Cauchy local

$$\begin{aligned} z' &= f(z,t),\ t \in [t_n, t_{n+1}] \\ z(t_n) &= x_n \end{aligned}$$

es decir, considerar la solución exacta si ésta coincidiera con la numérica en $t = t_n$. En general $z(t_{n+1})$ difiere de x_{n+1} lo que lleva a definir el *error de discretización local*:

$$d_n = x_{n+1} - z(t_{n+1}).$$

El máximo p tal que

$$|d_n| \leq Ch^{p+1}$$

para una constante C independiente de h es llamado el *orden* del método.

Esto se explica pues el error de discretización global e_N luego de $N = h/T$ iteraciones será comparable con la suma de los N errores locales d_n, $n = 0, \ldots, N-1$:

$$|e_N| \leq \sum_{n=0}^{N-1} d_n \leq (C/T)h^p$$

por lo que el error acumulado al final del cálculo medido en el último intervalo es como h^p. Es como si tuviéramos una cuenta de ahorro donde depositamos cada mes d_n y acumulamos e_n con un interés que dependerá de $\partial f/\partial x$.

No analizaremos el orden de cada uno de los métodos que veremos en detalle, aunque sí lo indicaremos. Sin embargo, es fácil convencerse de que el orden del método depende principalmente del error de la cuadratura que se escoge para aproximar el término integral de la derecha en (5.1). Entonces, cuadraturas por rectángulos, trapecios y Simpson generarán métodos de cada vez mayor orden.[8]

---

[7] Incluyendo también los errores de redondeo.

[8] Véanse los métodos de cuadratura en el Capítulo 3.

5.4 Métodos de tipo Euler de orden 1

Hacemos una cuadratura del tipo rectángulo[9] para

$$\int_{t_n}^{t_{n+1}} f(x(s), s)\, ds \approx (t_{n+1} - t_n) f(x(t_n), t_n) = h\, f(x(t_n), t_n).$$

Esto motiva el siguiente algoritmo llamado *método de Euler progresivo* para aproximar la solución de una ecuación diferencial ordinaria:

Método de Euler progresivo (orden 1)

$$\begin{aligned} x_0 &= x(0), \\ x_{n+1} &= x_n + h\, f(x_n, t_n). \end{aligned}$$

Este método calcula el nuevo valor de la función x_{n+1} en el tiempo t_{n+1} directamente a partir del valor anterior x_n en el tiempo precedente t_n. El método anterior también puede verse como una *discretización*[10] de la ecuación diferencial por una ecuación de diferencias:

$$\frac{x_{n+1} - x_n}{h} = f(x_n, t_n),$$

donde a la izquierda se ha reemplazado la derivada continua por una diferencia [11].

Si en vez de la aproximación anterior colocamos:

$$\frac{x_{n+1} - x_n}{h} = f(x_{n+1}, t_{n+1})$$

obtenemos el algoritmo llamado *método de Euler retrógrado*

Método de Euler retrógrado (orden 1)

$$\begin{aligned} x_0 &= x(0), \\ x_{n+1} &= x_n + h\, f(x_{n+1}, t_{n+1}). \end{aligned}$$

Conocer el nuevo valor de x_{n+1} a partir del valor de x_n requiere ahora la resolución de una ecuación algebraica no-lineal[12] que involucra f. En este caso se dice que x_{n+1} está implícitamente dado a partir de x_n. Este algoritmo hace parte de los llamados *métodos implícitos* en contraposición a los *métodos explícitos* como es el caso del esquema de Euler progresivo.

---

[9] Véase el Capítulo 3.

[10] Discretización del problema continuo, pues se reemplaza la variable continua por una que toma un número discreto de valores (finito o numerable).

[11] Lo que ya se había visto en el Capítulo 3 cuando se estudiaron las derivadas numéricas denotadas por Δ_h.

[12] Para ello son útiles los métodos como el método de Newton-Raphson vistos en el Capítulo 3.

5.5 Un primer ejemplo numérico

Para ver cómo funcionan los métodos anteriores, consideremos el siguiente problema de Cauchy como ejemplo:

$$\begin{aligned} x'(t) &= 2x(t) + t, \qquad t \in [0,1] \\ x(0) &= x_0 \text{ dado} \end{aligned}$$

cuya solución exacta se obtiene fácilmente multiplicando la ecuación por el factor integrante e^{-2t}:

$$\begin{aligned} x'\, e^{-2t} - 2x\, e^{-2t} &= te^{-2t} \\ (x\, e^{-2t})' &= te^{-2t} \\ x\, e^{-2t} &= C + \int t\, e^{-2t}\, dt \\ x &= Ce^{2t} + e^{2t} \int te^{-2t}\, dt \\ x &= Ce^{2t} - \frac{t}{2} - \frac{1}{4} \end{aligned}$$

y evaluando la constante con la condición inicial se obtiene $C = x(0) + \frac{1}{4}$ de donde tenemos la expresión para la solución exacta de la EDO

$$x(t) = \left(x(0) + \frac{1}{4}\right) e^{2t} - \frac{t}{2} - \frac{1}{4}, \qquad t \in [0,1].$$

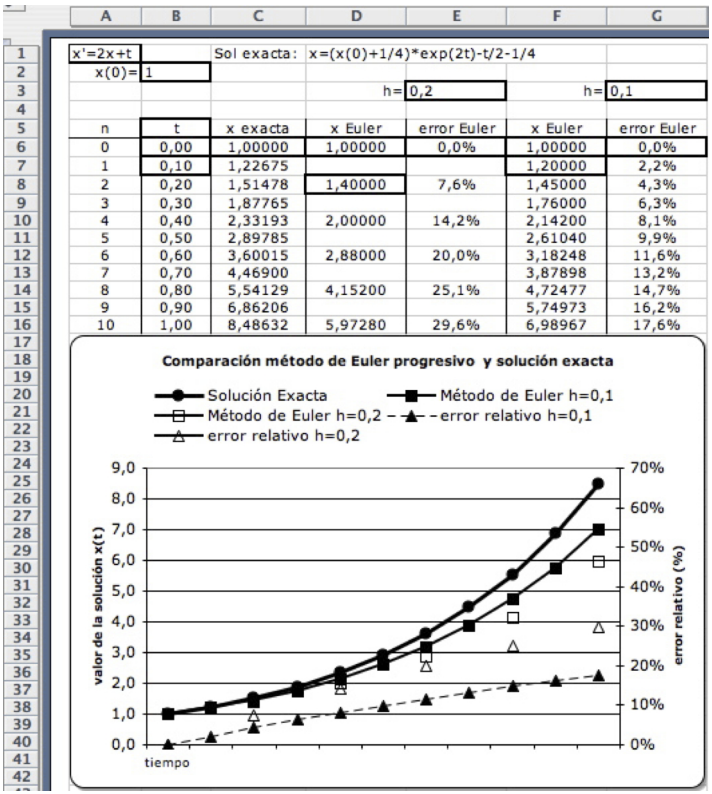
❒✍ En la Figura 5.1 se muestra cómo implementar el método de Euler progresivo en una planilla de cálculo para el ejemplo indicado, comparando los resultados para dos pasos distintos de discretización $h = 0,2$ y $h = 0,1$.

En el cuadro de la Figura 5.1, también se calcula el error relativo cometido en la iteración n del algoritmo:

$$\epsilon_n = \frac{|x(t_n) - x_n|}{|x(t_n)|}$$

que se puede expresar en porcentaje. El error del método progresivo comienza con un error 0 pues se parte de la condición inicial exacta $x_0 = x(0)$ y luego va creciendo hasta alcanzar un $29,6\,\%$ después de 5 iteraciones si $h = 0,2$. Si se divide el paso a la mitad, esto es $h = 0,1$, el error máximo alcanza esta vez para el método progresivo se reduce a $17,6\,\%$ después de 10 iteraciones. Notemos que reducir el paso a la mitad, equivale a duplicar el número de iteraciones. En cualquier caso el error del método es considerable al final del intervalo de tiempo.

✍ **Ejercicio 5.1.** Se observa de lo anterior que si el paso h se reduce a la mitad, el error también se reduce a la mitad, ¿a qué se puede deber esto?



	A	B	C	D	E	F	G
1	x'=2x+t		Sol exacta:	x=(x(0)+1/4)*exp(2t)-t/2-1/4			
2	x(0)=	1					
3				h=	0,2	h=	0,1
4							
5	n	t	x exacta	x Euler	error Euler	x Euler	error Euler
6	0	0,00	1,00000	1,00000	0,0%	1,00000	0,0%
7	1	0,10	1,22675			1,20000	2,2%
8	2	0,20	1,51478	1,40000	7,6%	1,45000	4,3%
9	3	0,30	1,87765			1,76000	6,3%
10	4	0,40	2,33193	2,00000	14,2%	2,14200	8,1%
11	5	0,50	2,89785			2,61040	9,9%
12	6	0,60	3,60015	2,88000	20,0%	3,18248	11,6%
13	7	0,70	4,46900			3,87898	13,2%
14	8	0,80	5,54129	4,15200	25,1%	4,72477	14,7%
15	9	0,90	6,86206			5,74973	16,2%
16	10	1,00	8,48632	5,97280	29,6%	6,98967	17,6%

FIGURA 5.1. Planilla de cálculo que implementa el método de Euler progresivo. Se inicializa el método con el valor inicial $x(0)$ y de los pasos h en `B2=1`, `E3=0,2` y `G3=0,1`. Luego se rellenan las celdas `B6=0` (tiempo inicial), `D6=F6=$B$2` (condiciones iniciales). Los restantes cuadros enmarcados corresponden a las fórmulas `B7=B6+`$G3$ (avance del tiempo), `C6=($B$2+1/4)*EXP(2*B6)-B6/2-1/4` (solución exacta), `D8=D6+$E$3*(2*D6+B6)`, `F7=F6+G$3*(2*F6+B6)` (estas dos últimas la iteración del método de Euler progresivo) que se copian hacia abajo en cada columna. Para calcular el error relativo de aproximación se agregan las columnas `E` y `G` con las fórmulas `E6=ABS($C6-D6)/ABS($C6)`, `G6=ABS($C6-F6)/ABS($C6)` que se copian hacia abajo. No olvide poner formato de celda de porcentaje para las columnas del error `E` y `G`. Abajo se muestra un gráfico asociado a las columnas `C`, `F` y `G`.

5.6 Estabilidad e inestabilidad numérica

Ahora que ya conocemos dos algoritmos de discretización de ecuaciones diferenciales, revisemos sucintamente los conceptos de *estabilidad e inestabilidad numérica*.

En el Capítulo 1 mencionamos el relato de Ray Bradbury y el efecto mariposa, que ilustra muy bien la extrema sensibilidad que puede tener un proceso de evolución ante pequeñas variaciones de las condiciones iniciales, lo que lleva al concepto de inestabilidad. La misma noción se puede aplicar a un método numérico de resolución de una ecuación diferencial.

Intuitivamente, un algoritmo de resolución numérica de un problema de Cauchy es *estable numéricamente*, si pequeñas alteraciones de la condición inicial x_0 o pequeños errores de redondeo en los cálculos, conllevan a errores $e_n = x_n - x(t_n)$ que no se amplifican demasiado en el tiempo. Por el contrario, un método es *inestable numéricamente* si pequeños errores en la condición inicial x_0 o en los cálculos llevan a una amplificación no acotada del error e_n.

Para ilustrar mejor los conceptos de estabilidad e inestabilidad que ocurren al discretizar una ecuación diferencial, tomemos como ejemplo el siguiente problema de Cauchy:

$$\begin{aligned} x' &= -20x \\ x(0) &= 1 \end{aligned}$$

cuya solución exacta es

$$x = e^{-20t}$$

y aproximemos $x(t_n)$ por x_n usando los métodos de Euler progresivo y retrógrado respectivamente.

Notemos que como la solución exacta tiende a cero cuando t tiende infinito, se tiene que $x(t_n) \to 0$, entonces el comportamiento asintótico del error

$$e_n = x_n - x(t_n)$$

es el mismo que tiene la sucesión x_n cuando n tiende a infinito.

1. Para el ejemplo considerado, el método de Euler progresivo queda dado por la recurrencia:

$$\begin{aligned} x_0 &= 1 \\ x_{n+1} &= x_n - 20\,h\,x_n = (1 - 20\,h)x_n \end{aligned}$$

de donde se obtiene fácilmente que

$$x_n = r^n, \qquad r \equiv 1 - 20\,h.$$

Analicemos el comportamiento asintótico de x_n (y consecuentemente del error e_n) cuando n tiende a infinito para distintos valores del paso h:

a) Si $0 < h \leq \frac{1}{20}$ entonces $0 \leq r < 1$ y x_n tiende a cero.

b) Si $\frac{1}{20} \leq h < \frac{2}{20}$ entonces $-1 < r \leq 0$ y x_n también tiende a cero, pero alternando de signo.

c) Si $h = \frac{2}{20}$ entonces $r = -1$ y $x_n = (-1)^n$ no converge a cero pero se mantiene acotada.

d) Si $h > \frac{2}{20}$ entonces $r < -1$ y x_n diverge en valor absoluto a infinito y sin valor absoluto toma valores cada vez más grandes con signos alternados.

En el caso d) hay inestabilidad numérica pues el error no es acotado y en los otros casos a), b) y c) se dice que hay estabilidad numérica pues el error es acotado [13] . De modo que el método de Euler progresivo se dice que es *condicionalmente estable* bajo la condición de estabilidad:

$$h \leq \frac{20}{2} = 0{,}1.$$

La *estabilidad condicional* resulta ser una característica usual de los métodos explícitos.

2. Analicemos ahora la estabilidad del método de Euler retrógrado que está dado por:

$$\begin{aligned} x_0 &= 1 \\ x_{n+1} &= x_n - 20\,h\,x_{n+1} \end{aligned}$$

entonces

$$x_{n+1} = (1 + 20\,h)^{-1} x_n$$

de donde, como antes

$$x_n = r^n, \qquad r \equiv (1 + 20\,h)^{-1}.$$

El comportamiento asintótico de x_n (recordemos que es también el del error e_n) es ahora mucho más simple de analizar, ya que cualquiera sea el paso $h > 0$ se tiene que x_n converge a cero. Por esta razón el método de Euler retrógrado se dice que es *incondicionalmente estable* . La *estabilidad incondicional* es una característica usual de los métodos implícitos.

Como conclusión del análisis precedente, digamos que el error de los métodos de Euler progresivo y retrógrado es similar cuando h es pequeño y ambos métodos son estables, pero si h crece, el método de Euler progresivo se vuelve inestable.

Este fenómeno se aprecia en la Figura 5.2 donde se comparan los dos métodos para valores crecientes del paso: $h = 0,05$, $h = 0,09$, $h = 0,1$ y $h = 0,11$ respectivamente. Para obtener estos gráficos se utilizó una planilla de cálculo que se explica en la Figura 5.3. ❒✍

✍ **Ejercicio 5.2.** Pruebe que la condición de estabilidad del método de Euler progresivo, cuando se aproxima el problema de Cauchy con $\sigma > 0$

$$\begin{aligned} x' &= -\sigma x \\ x(0) &= x_0 \text{ dado} \end{aligned}$$

esta dada en este caso por $h \leq \frac{2}{\sigma}$. Pruebe varios valores distintos de σ y h en forma numérica y para ello construya la planilla que se explica en la Figura 5.3, con la que se obtuvo de hecho los gráficos de la Figura 5.2. Experimente numéricamente qué ocurre para ambos métodos si $h > \frac{2}{\sigma}$.

[13] Notemos, sin embargo, que en los casos b) y c) el error es acotado, pero oscilante. Se habla en este caso de oscilaciones numéricas, que pueden ser un preludio a la inestabilidad.

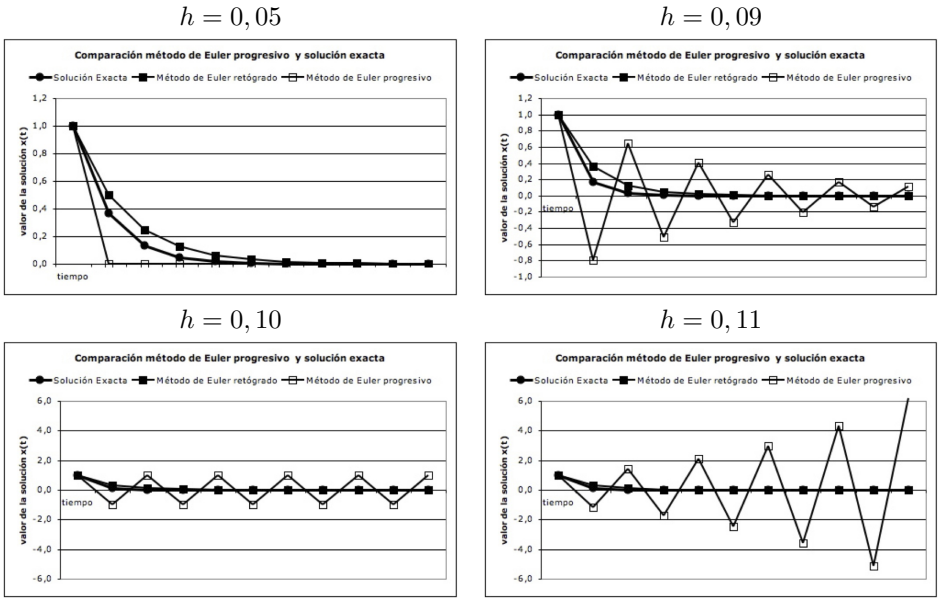


FIGURA 5.2. Comparación de los dos métodos de Euler progresivo (explícito) y de Euler retrógrado (implícito) para valores crecientes del paso h. El método de Euler se vuelve inestable si el paso h supera el valor $0,1$ mientras que el método retrógrado se mantiene siempre estable.

5.7 Método de Euler en una ecuación escalar: la población mundial

La Figura 5.4 muestra el poblamiento de la Tierra en los últimos 500 años. Se espera que para el año 2050, la población mundial crezca a más de 9×10^9 habitantes.

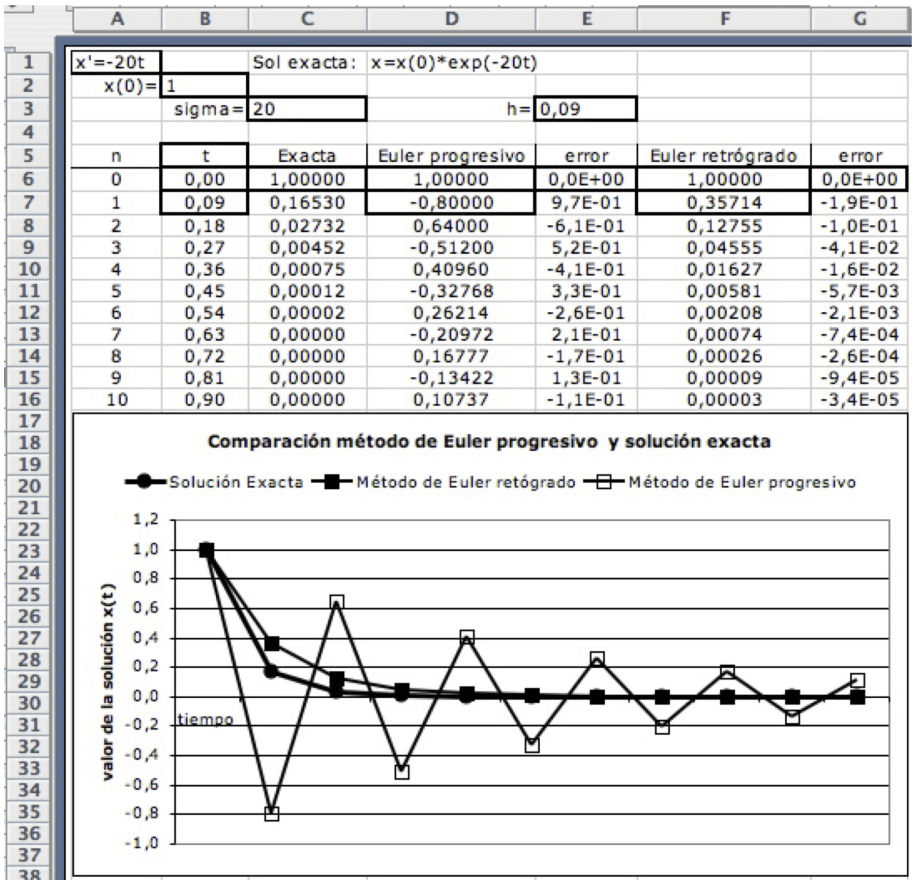
El modelo más sencillo para la evolución P de una población es el de una EDO lineal de primer orden homogénea, llamado *modelo malthusiano*:

$$P' = \sigma(t)P = f(P,t)$$

donde σ representa la evolución de la tasa de crecimiento anual neta de la población a través del tiempo (tasa de nacimiento menos tasa de mortalidad) [14].

Apliquemos el algoritmo de Euler progresivo para simular el crecimiento de la población mundial a partir del año 2000. Tomemos un paso $h = 1$ (un año) y llamemos P_n, σ_n a la población y tasa de crecimiento el año n:

[14]Cuya solución exacta es para una población inicial P_0 en $t = t_0$ $P(t) = P(t_0) \exp\left(\int_{t_0}^{t} \sigma(s)ds\right)$.



	A	B	C	D	E	F	G
1	x'=-20t		Sol exacta:	x=x(0)*exp(-20t)			
2	x(0)=	1					
3		sigma=	20		h=	0,09	
4							
5	n	t	Exacta	Euler progresivo	error	Euler retrógrado	error
6	0	0,00	1,00000	1,00000	0,0E+00	1,00000	0,0E+00
7	1	0,09	0,16530	-0,80000	9,7E-01	0,35714	-1,9E-01
8	2	0,18	0,02732	0,64000	-6,1E-01	0,12755	-1,0E-01
9	3	0,27	0,00452	-0,51200	5,2E-01	0,04555	-4,1E-02
10	4	0,36	0,00075	0,40960	-4,1E-01	0,01627	-1,6E-02
11	5	0,45	0,00012	-0,32768	3,3E-01	0,00581	-5,7E-03
12	6	0,54	0,00002	0,26214	-2,6E-01	0,00208	-2,1E-03
13	7	0,63	0,00000	-0,20972	2,1E-01	0,00074	-7,4E-04
14	8	0,72	0,00000	0,16777	-1,7E-01	0,00026	-2,6E-04
15	9	0,81	0,00000	-0,13422	1,3E-01	0,00009	-9,4E-05
16	10	0,90	0,00000	0,10737	-1,1E-01	0,00003	-3,4E-05

FIGURA 5.3. Planilla de cálculo que ilustra la estabilidad condicional del método de Euler progresivo y la estabilidad incondicional del método de Euler retrógrado y que sirve para realizar el Ejercicio 5.2. Se inicializa similarmente a la planilla de la Figura 5.1 más las fórmulas `C6=$B$2*EXP(-$C$3*B6)` (solución exacta), `D7=(1-$C$3*$E$3)*D6` y `F7=1/(1+$C$3*$E$3)*F6` que corresponden a $x_{n+1} = (1 - \sigma h)x_n$ y $x_{n+1} = (1 + \sigma h)^{-1}x_n$ respectivamente (ver texto) y las diferencias (no relativas) con la solución `E6=$C6-D6`, `G6=$C6-F6` que se copian hacia abajo en cada columna respectiva.

ALGORITMO DE EULER PROGRESIVO PARA EL MODELO DE MATHUS

$$P_{n+1} = P_n + \sigma_n P_n, \quad n \geq 2000, \quad P(2000) = 6000 \text{ millones de hbts.}$$

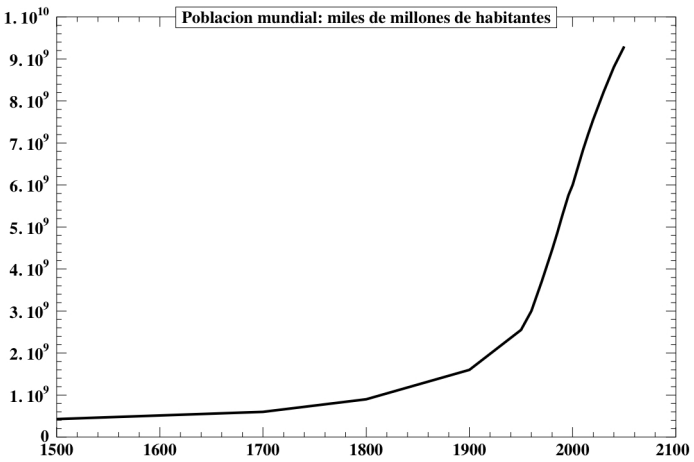


FIGURA 5.4. Población mundial desde el año 1500 proyectada al 2050 (fuente: `www.census.gov/ipc/www/idb`).

El crecimiento de la población mundial en los últimos 50 años se ha extrapolado hasta el 2050 como se muestra en la Figura 5.5 superior. Observe una ligera inflexión en la década de los 90'debido a una disminución sostenida de la tasa de crecimiento. En la Figura 5.5 inferior, se muestra la tasa de crecimiento σ que corresponde al porcentaje de crecimiento de la población cada año. A partir de dicha figura, es razonable suponer que σ decrecerá linealmente en el futuro disminuyendo desde su valor actual de 1,25 % un 0,5 % cada 30 años, esto es:

$$\sigma(t) = -0{,}5\,\%/30(t - 2000) + 1{,}25\,\%.$$

Con este dato, intentemos estimar la población mundial desde el 2000 hasta los años 2075 y 2100.

❒✍ Es posible hacer una simple planilla de cálculo implementando el método de Euler progresivo para esta ecuación. Los resultados obtenidos los resumimos en el Cuadro 5.1. Los cálculos nos permiten predecir que habrá un máximo de población en el año 2075 con un valor de $P_{2075} \approx 9629$ millones de hbts.

✍ **Ejercicio 5.3.** Realice una simulación similar a la que se hizo para la población mundial en el caso particular de un país como Chile. Eso sí, debe considerar además la evolución de la inmigración neta (inmigración menos emigración) representada por

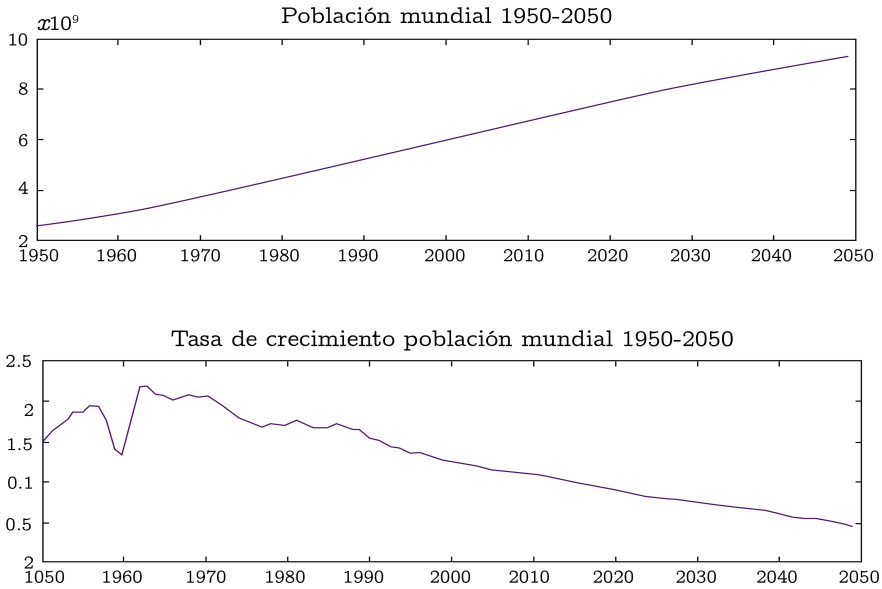


FIGURA 5.5. Población mundial y tasa de crecimiento entre 1950 y 2050 (fuente: `www.census.gov/ipc/www/idb`).

una función Q. El modelo corregido es una EDO lineal de primer orden no homogénea:[15]

$$P' = \sigma(t)P + Q(t).$$

Para ello, escriba el algoritmo de Euler progresivo para el modelo de este ejercicio y simule en una planilla de cálculo el impacto en el periodo $2000 - 2040$ que tendría en la población chilena una inmigración que crece año a año en Q_0 y que comenzó el año 2000, esto es, $Q(t) = Q_0(t - 2000)$. Utilice para ello una población inicial de $P_{2000} = 15,2$ millones de hbts. y utilice una tasa de crecimiento σ que puede modelar en forma lineal a partir de los datos de la Figura 5.6 como se hizo antes.

5.8 Pérdida de estabilidad numérica: crecimiento logístico

Un modelo más realista de crecimiento de poblaciones debe tener en cuenta las restricciones de recursos: alimento, agua, energía, etc. El modelo logístico tiene en cuenta este efecto y se escribe:

$$P' = \sigma P(M - P), \quad P(0) = P_0$$

[15] En este caso, la solución es $P(t) = P(t_0)\exp\left(\int_{t_0}^{t}\sigma(s)ds\right) + \int_{t_0}^{t}\exp\left(\int_{s}^{t}\sigma(s)ds\right)Q(s)ds$.

n año	σ_n	P_n hbts (miles de mill.)
2000	0,0125	6000
2001	0,0123	6075
⋮	⋮	⋮
2074	0,0002	9627
2075	**0,0000**	**9629**
2076	-0,0002	9629
⋮	⋮	⋮
2099	-0,0040	9196
2100	-0,0042	9159

CUADRO 5.1. Método de Euler progresivo aplicado al modelo de crecimiento de población mundial. Se estima un máximo de población el año 2075, exactamente cuando la tasa de crecimiento σ comienza a ser negativa.

donde M es una población máxima alcanzable que depende de los recursos disponibles y supondremos para simplificar que $\sigma > 0$ es constante. Notemos que si P es menor que M entonces el lado derecho de la ecuación logística es positivo, por lo que P' es positivo y la población crece, pero a medida que P se acerca a M, el lado derecho es más y más cercano a cero haciendo que la población crezca cada vez menos.

El método de Euler progresivo se escribe en este caso como:

MÉTODO DE EULER PROGRESIVO: MODELO LOGÍSTICO DISCRETO

$$P_{n+1} = P_n + h\sigma P_n(M - P_n).$$

Este modelo tiene un comportamiento asombrosamente complejo. En efecto, si calculamos los puntos de acumulación de la sucesión aproximante P_n para una condición inicial cualquiera, pero para valores de σ crecientes entre $1{,}8 \leq \sigma \leq 3{,}0$, con $M = 100$ y $h = 1$, se obtiene la impresionante Figura 5.7 [16]. Este fenómeno se debe a que el algoritmo presenta oscilaciones numéricas que lo hacen cada vez menos estable a medida que σ crece y h se mantiene constante ($h = 1$). Se observa que aparecen sucesivamente 1,2,4,8,16,... puntos de equilibrio hasta un régimen aparentemente alternado entre orden y desorden, o más precisamente, entre periódico y caótico [17].

[16] Véase `brain.cc.kogakuin.ac.jp/~kanamaru/Chaos/e/BifArea`.

[17] En realidad, se puede demostrar que la aparición de periodos sigue un orden predeterminado, cf. [**28**].

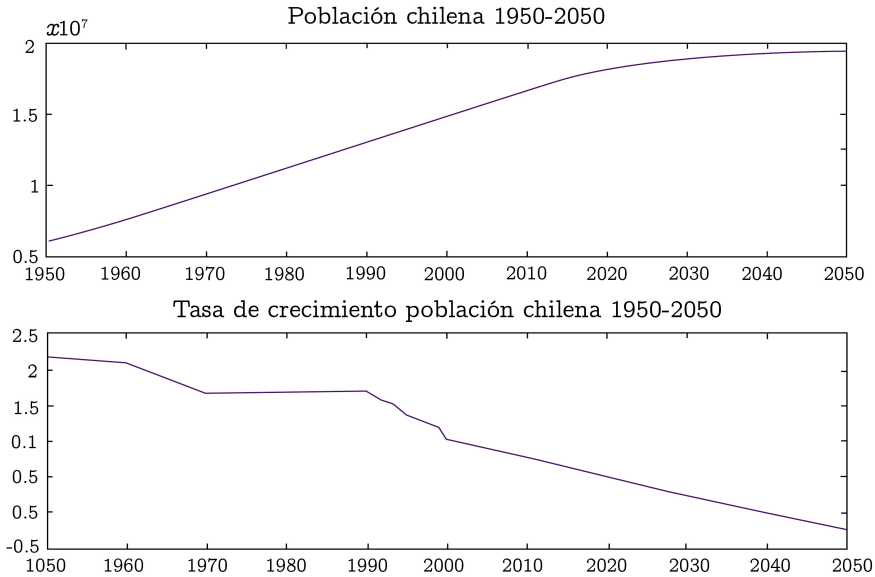


FIGURA 5.6. Población y tasa de crecimiento de Chile entre 1950 y 2050 (fuente: `www.census.gov/ipc/www/idb`).

✍ **Ejercicio 5.4.** Tomando el modelo logístico simplificado $P_{n+1} = \sigma P_n(1-P_n)$, estudie numéricamente el comportamiento de este sistema discreto para $\sigma = 2$, $\sigma = 3{,}2$, $\sigma = 3{,}5$, $\sigma = 3{,}56$, etc. El gráfico de P_{n+1}en función de P_n resulta una parábola. Grafíquela. Ahora, dibuje segmentos entre los sucesivos puntos $(P_n, P_{n+1}) - (P_{n+1}, P_{n+1}) - (P_{n+1}, P_{n+2})$ para un valor de σ dado. El diagrama resultante se conoce como el *mapeo logístico*[18]. Es similar al diagrama de punto fijo de la Figura 3.1.

✍ **Ejercicio 5.5.** Para hallar la solución exacta de la ecuación logística haga el cambio de variables $z = \dfrac{1}{P}$ con $z' = \dfrac{-P'}{P^2}$ y obtenga $z' = -\sigma(Mz - 1)$, esto es, $z' = -M\sigma z + \sigma$ (EDO lineal) de donde puede deducir que:

$$\frac{1}{P} = \exp\left(-\int_0^t M\sigma(s)ds\right)\left(\frac{1}{P_0} + \int_0^t \exp\left(\int_0^s M\sigma(s)ds\right)\sigma(s)ds\right).$$

Reordenando obtenga que: $P = \dfrac{P_0 M}{P_0 + (M - P_0)\exp(-M\sigma t)}$. A qué valor converge P si $t \to \infty$, ¿es esto coherente con los resultados numéricos que se obtienen?

[18] Véase un diagrama interactivo en `www.lboro.ac.uk/departments/ma/gallery/doubling`.

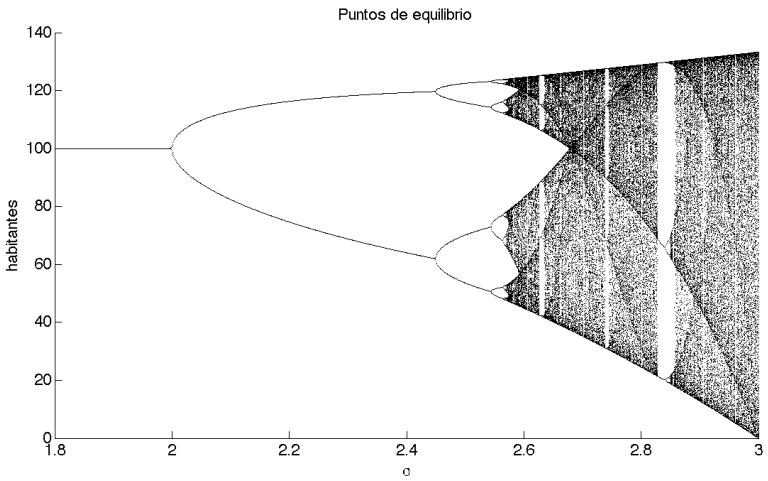


FIGURA 5.7. Árbol de bifurcaciones de Feigenbaum.

5.9 Método Euler en un caso vectorial: el crecimiento de una epidemia

El ser humano a sufrido diferentes plagas o pandemias durante su historia. Hubo, por ejemplo, tres pandemias de gripe durante el siglo XX, la gripe española (1918-1919) con cerca de 40 millones de muertos, la gripe asiática (1957-1958) con cerca de 4 millones de decesos y la gripe de Honk Kong (1968-1969) con cerca de 2 millones de muertos. La más reciente pandemia gripal corresponde a la influenza humana (2009) con cerca de 20.000 muertes.

En epidemiología, existen los llamados modelos SIR, que sirven para modelar la evolución de una epidemia o pandemia. La idea es separar una población de N individuos en tres clases: suceptibles (S), infectados (I) y recuperados (R), cumpliéndose que $S + I + R = N$. El número de contagios es proporcional al número de encuentros que se producen entre individuos infectados e individuos suceptibles, lo que resulta proporcional al producto SI, con una constante de proporcionalidad $\beta > 0$ llamada tasa de infección. La población de infectados crece de acuerdo a este número de infectados, mientras que la población de suceptibles decrece de acuerdo a este número. Por otro lado, hay un número de infectados que se recupera (o muere) de acuerdo a una tasa de recuperación $\gamma > 0$. Con esto, el sistema de ecuaciones es el siguiente:

$$\left\{ \begin{array}{rcl} S' & = & -\beta SI \\ I' & = & \beta SI - \gamma I \\ R' & = & \gamma I \end{array} \right. \tag{5.2}$$

con condiciones iniciales el número de suceptibles, infectados y recuperados inicial $S(0)$, $I(0)$, $R(0)$ tales que $S(0) + I(0) + R(0) = N$.

Este simple modelo fue introducido por Anderson Gray McKendrick (1876 - 1943) epidemiólogo británico y William Ogilvy Kermack (1898 - 1970) químico británico y es también llamado modelo de Kermack-McKendrick en epidemiología.

Para resolver numéricamente este sistema, podemos aplicar un método de tipo Euler, para lo cual resulta conveniente considerar el sistema en forma vectorial, esto es de la forma:

$$(S', I', R') = f(S, I)$$

donde f es la función de $\mathbb{R}^2$ en $\mathbb{R}^3$ dada por

$$f(S, I) = (-\beta SI, \beta SI - \gamma I, \gamma I).$$

Notar que f no depende explícitamente de R. Con esto, el método de Euler se escribe muy simplemente así:

Método de Euler progresivo para modelo SIR

$$\begin{aligned}(S_0, I_0, R_0) &= (S(0), I(0), R(0)) \\ (S_{n+1}, I_{n+1}, R_{n+1}) &= (S_n, I_n, R_n) + hf(S_n, I_n).\end{aligned}$$

Utilizando una sencilla planilla de cálculo de la Figura 5.8, podemos calcular la evolución de una epidemia en que inicialmente hay 2 millones de suceptibles ($S(0) = 2\,mdh$ en millones de habitantes), un millón de infectados ($I(0) = 1\,mdh$) y ningún recuperado ($R(0) = 0\,mdh$). La población total es $N = 3\,mdh$ y estudiamos su evolución diariamente. Suponiendo constantes $\beta = 1$ y $\gamma = 1$ la evolución de las tres curvas de suceptibles $S(t)$, infectados $I(t)$ y recuperados $R(t)$ se grafica en la Figura 5.8. ❐✍

Notemos que el número de infectados crece inicialmente hasta alcanzar un máximo y luego disminuye.

A los epidemiólogos les interesa que el número de infectados no crezca inicialmente, esto es que

$$I'(0) < 0$$

de la primera ecuación de (5.2), es fácil ver que esto se tiene siempre que

$$\frac{S(0)\beta}{\gamma} < 1$$

este número se interpreta como el número de infecciones secundarias producidas por una infección inicial, o umbral epidemiológico, si es menor que uno, cada persona contagiada infecta a menos de una persona en promedio y la epidemia no crece inicialmente, si es mayor que uno, cada persona infectada infecta a más de una persona en promedio y la epidemia crece inicialmente. En nuestro caso se tiene que:

$$\frac{S(0)\beta}{\gamma} = 2 > 1$$

lo que explica por qué la epidemia crece inicialmente.

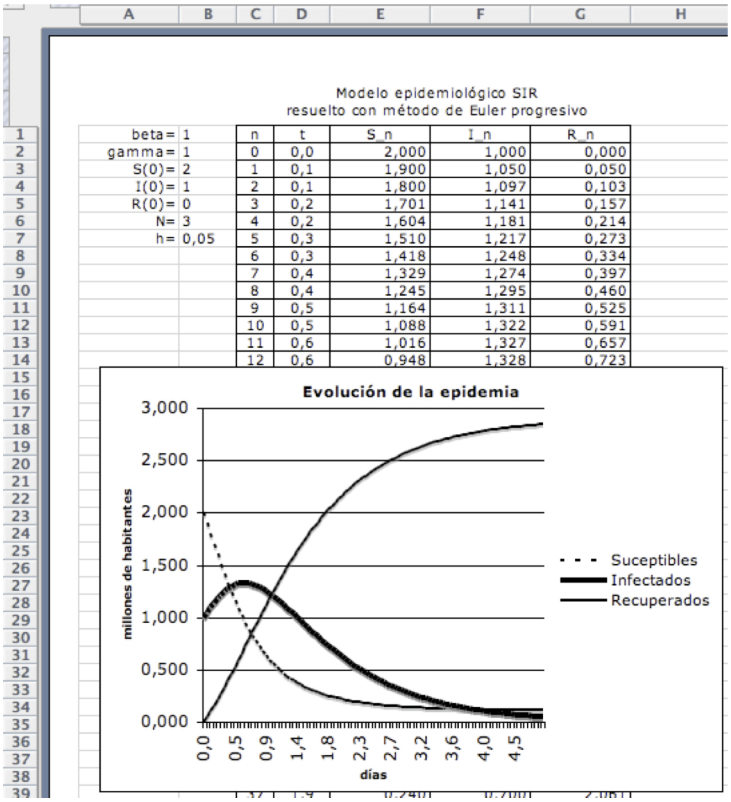


FIGURA 5.8. Evolución de una epidemia siguiendo el modelo SIR de Hardy-Weinberg. Para construir la planilla se inicializa el método con los valores iniciales indicados para β, γ, $S(0)$, $I(0)$ y $R(0)$. Se estipula además el valor de N y del paso h. Los valores de $S(0)$, $I(0)$ y $R(0)$ se copian en las celdas `E2`, `F2` y `G2` respectivamente. Luego se inicializan las celdas `E3=E2-$B$7*$B$1*E2*F2`, `F3=F2+$B$7*($B$1*E2*F2-$B$2*F2)` y `G3=G2+$B$7*$B$2*F2` y se copian hacia abajo tantas veces como se desee. Después de 98 iteraciones se obtiene $S_{98} = 0,109$, $I_{98} = 0,048$, $R_{98} = 2,843$.

Si tomamos $\beta = 0{,}5$ tendríamos

$$\frac{S(0)\beta}{\gamma} = 1$$

luego este es el valor crítico para la tasa de infección. Para valores de $\beta > 0{,}5$ la epidemia crece inicialmente, para valores de $\beta < 0{,}5$ la epidemia decrece inicialmente.

Notemos además que las tres curvas parecen converger ciertos valores límite S_∞, I_∞, R_∞ constantes cuando ha transcurrido suficiente tiempo. No es fácil calcular estos valores, llamados valores de equilibrio, del sistema de ecuaciones. En efecto, si suponemos que las tres curvas tienen derivada nula, esto es $S' = 0$, $I' = 0$, $R' = 0$, del sistema (5.2) obtenemos el siguiente sistema de ecuaciones no lineales:

$$\begin{aligned} -\beta S_\infty I_\infty &= 0 \\ \beta S_\infty I_\infty - \gamma I_\infty &= 0 \\ \gamma I_\infty &= 0 \end{aligned}$$

De la tercera de estas ecuaciones podemos deducir fácilmente que $I_\infty = 0$, esto es, que el número de infectados tiende a cero. Pero, de las otras dos ecuaciones no podemos encontrar el valor límite S_∞ de individuos suceptibles que no fueron nunca infectados. Si conociéramos este valor podríamos calcular también el número de recuperados $R_\infty = N - S_\infty$ una vez terminada la epidemia. En este caso resulta útil el cáculo numérico que nos provee de valores aproximados de S_∞ y R_∞.

✍ **Ejercicio 5.6.** Hay otra forma de aproximar S_∞. Divida la primera y la tercera ecuación en (5.2) para obtener que

$$(\ln S)' = -\frac{\beta}{\gamma} R'$$

integrando esto y tomando límite deduzca que

$$S_\infty = S(0) e^{-\frac{\beta}{\gamma} R_\infty}.$$

Luego reemplazando $R_\infty = N - S_\infty$, encuentre que S_∞ debe ser un cero de la función no lineal:

$$f(x) = x - S_0 \exp\left(-\frac{\beta}{\gamma}(N - x)\right) = 0.$$

Esta ecuación no tiene una solución analítica exacta, pero su solución se puede aproximar usando un método numérico para encontrar los ceros de la función $f(x)$. Considerando $\beta = 1$, $\gamma = 1$, $N = 3$ y $S(0) = 2$, encuentre una solución aproximada de $f(x) = x - e^{x-3} = 0$ utilizando el método de Newton-Raphson visto en el Capítulo 3. Compare esta solución aproximada con el valor asintótico obtenido en la Figura 5.8. Solución: $S_\infty = x \approx 0,111296644$ y en el cuadro se obtiene $S_\infty \approx 0,109$ para $h = 0,05$ y 98 iteraciones. Para obtener un valor más cercano se debe disminuir el paso y aumentar el número de iteraciones.

✍ **Ejercicio 5.7.** Estudie numéricamente el siguiente sistema de ecuaciones diferenciales utilizando el método de Euler progresivo, tal cual como se hizo para el modelo

	a (p)	b (q)
a (p)	población x: aa (p^2)	población y: ab (pq)
b (q)	población y: ab (pq)	población z: bb (q^2)

CUADRO 5.2. Tabla de combinación de alelos indicando sus frecuencias $(p+q=1)$.

SIR (5.2). Considere el sistema lineal

$$\begin{aligned} x' &= -qx + \frac{p}{2}y \\ y' &= qx - \frac{1}{2}y + pz \\ z' &= \frac{q}{2}y - pz, \end{aligned}$$

donde p, q son constantes no negativas con $p+q=1$. El sistema anterior modela las poblaciones x de "aa", y de "ab" y z de "bb", donde "a" y "b" son dos alelos de un mismo gen que aparecen con frecuencias p y q en una población T. Resolviendo el sistema numéricamente, encuentre que cuando $t \to \infty$ se llega aproximadamente a la proporción de equilibrio:

$$x_\infty : y_\infty : z_\infty = p^2 : 2pq : q^2$$

que se ilustra en el Cuadro 5.2, llamado *equilibrio de Hardy-Weinberg*. De hecho, estas ecuaciones fueron encontradas independientemente por el matemático británico Godfrey Harol Hardy (1877-1947) y el médico alemán Wilhelm Weinberg (1862-1937) [19].

5.10 Métodos de tipo Runge-Kutta de orden 2 y 4

Pasemos ahora a estudiar algunos métodos de orden 2. Para ello, hagamos una cuadratura del tipo trapecios[20] para la integral en la identidad (5.1):

$$\int_{t_n}^{t_{n+1}} f(x(s), s)\, ds \approx \frac{h}{2}\left(f(x(t_n), t_n) + f(x(t_{n+1}), t_{n+1})\right.$$

lo que inspira el siguiente algoritmo llamado *método de Heun* y que hace parte de los métodos de Runge-Kutta de orden 2:

[19] Hardy, G. H, Mendelian proportions in a mixed population. *Science* 28 (1908), 49–50. Weinberg, W. Über den Nachweis der Vererbung beim Menchen. *Jahresh. Verein f. vaterl. Naturk.* in Wruttemberg 64 (1908), 368–382. Hardy y Weinberg encontraron este modelo independientemente. Se dice que el matemático Hardy expresó discretamente al publicar su artículo: "esto es algo que podría interesar a los biólogos".

[20] Véase el Capítulo 3.

MÉTODO DE HEUN (RUNGE-KUTTA ORDEN 2)

$$\begin{aligned} x_0 &= x(0) \\ g_1 &= f(x_n, t_n) \\ g_2 &= f\left(x_n + h\, g_1, t_{n+1}\right) \\ x_{n+1} &= x_n + \frac{h}{2}\left(g_1 + g_2\right). \end{aligned}$$

✍ **Ejercicio 5.8.** ¿En qué tipo de cuadratura cree que está inspirado el siguiente algoritmo llamado método Euler modificado?

MÉTODO DE EULER MODIFICADO (RUNGE-KUTTA ORDEN 2)

$$\begin{aligned} x_0 &= x(0) \\ g_1 &= f(x_n, t_n) \\ g_2 &= f\left(x_n + \frac{h}{2} g_1, t_n + \frac{h}{2}\right) \\ x_{n+1} &= x_n + h\, g_2. \end{aligned}$$

¿Es este nuevo método implícito o explícito? Sepa que este método también hace parte de los métodos de Runge-Kutta de orden 2.

✍ **Ejercicio 5.9.** Compare numéricamente el error del método de Heun y de Euler modificado (que son ambos métodos de orden 2) para la ecuación $x' = 2x + t$. Para ello, modifique la planilla de cálculo ya construída para los métodos de Euler de orden 1 en la sección anterior. ❒✍

✍ **Ejercicio 5.10.** Al hacer una cuadratura del tipo Simpson[21] para la integral en la identidad (5.1):

$$\int_{t_n}^{t_{n+1}} f(x(s), s)\, ds \approx \frac{h}{6}\left(f(x(t_n), t_n) + 2f(x(t_{n+\frac{1}{2}}), t_{n+\frac{1}{2}}) + f(x(t_{n+1}), t_{n+1})\right)$$

se inspira el siguiente algoritmo que hace parte de los métodos de Runge-Kutta de orden 4 explícitos. Este algoritmo es uno de los más utilizados para resolver ecuaciones diferenciales ordinarias:

[21] Véase el Capítulo 3.

Método de Runge-Kutta de orden 4 explícito

$$\begin{aligned} x_0 &= x(0), \\ g_1 &= f(x_n, t_n), \\ g_2 &= f\left(x_n + \frac{h}{2}\, g_1, t_{n+\frac{1}{2}}\right), \\ g_3 &= f(x_n + \frac{h}{2} g_2, t_{n+\frac{1}{2}}), \\ g_4 &= f\left(x_n + h\, g_3, t_{n+1}\right), \\ x_{n+1} &= x_n + \frac{h}{6}\left(g_1 + 2g_2 + 2g_3 + g_4\right). \end{aligned}$$

Verifique que si f no depende de x se tiene que $g_2 = g_3$ y se recupera en este caso la idea de la cuadratura de Simpson vista en el Capítulo 3.

5.11 Ejemplo de Runge-Kutta: la pesca en el Mar Adriático

Después de la primera guerra, los pescadores del Mar Adriático estaban sorprendidos, pues la cantidad de presa para pescar había disminuido en vez de aumentar, a pesar de que durante los años de la guerra se había dejado de pescar. Se propuso el siguiente modelo para explicar la situación:

$$\begin{aligned} \frac{x'}{x} &= a - by \\ \frac{y'}{y} &= -c + dx \end{aligned}$$

Si x e y representan las poblaciones de presa y predador respectivamente, la presa crece relativamente ($a > 0$) si no hay predador y el predador decrece relativamente ($-c < 0$) si no hay presa. Por otro lado, los encuentros entre presa y predador favorecen a los predadores ($d > 0$) y merman las presas ($-b < 0$). Es fácil ver que si $x' = y' = 0$ entonces hay un punto de equilibrio dado por

$$x = \frac{c}{d}, \qquad y = \frac{a}{b}.$$

Estas ecuaciones son un clásico modelo del tipo predador-presa, y son llamadas ecuaciones de Lotka-Volterra en honor a Vito Volterra (1860-1940) físico y matemático italiano y su contemporáneo Alfred J. Lotka (1880-1949) quien era un químico americano.

Resulta cómodo representar la solución $x(t)$, $y(t)$ del sistema de Lotka-Volterra como pares ordenados $(x(t), y(t))$, los que se dibujan como puntos en un plano, llamado *plano de fases*. La curva descrita por la solución, partiendo del punto inicial (x_0, y_0), al variar t se donomina *trayectoria*. Asimismo, el punto de equilibrio del sistema también puede representarse por un punto del plano:

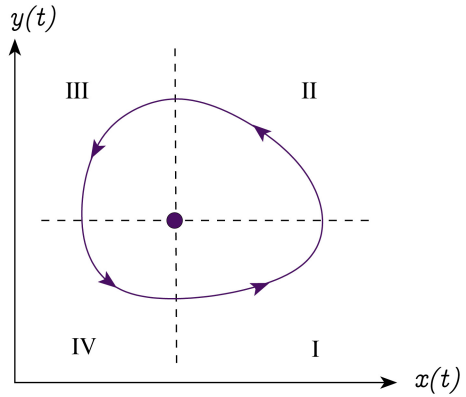


FIGURA 5.9. Órbita del sistema de Lotka-Volterra en el plano de fases en torno al punto de equilibrio. En el cuadrante I, las presas $x(t)$ y predadores $y(t)$ aumentan. En el cuadrante II, las presas comienzan a disminuir, pero los predadores siguen en aumento. En III, tanto presas como predadores disminuyen. En el cuadrante IV, las presas comienzan de nuevo a aumentar, mientras los predadores siguen disminuyendo. Esta dinámica poblacional se repite cada ciclo pasando nuevamente por I, II, III y IV y así sucesivamente.

$$(x, y) = \left(\frac{c}{d}, \frac{a}{b}\right).$$

Es posible mostrar[22] que las poblaciones de predador y presa, si no están en equilibrio, describen curvas cerradas que al cabo de un cierto tiempo vuelven a pasar por el punto inicial y giran (o se dice que oscilan) en torno al punto de equilibrio. A estas trayectorias cerradas se les llama *órbitas periódicas* (ver Figura 5.9).

Si ahora se provoca un cambio en los parámetros de la forma

$$\begin{aligned} a &\rightarrow a + \Delta a \\ c &\rightarrow c - \Delta a \end{aligned}$$

esto equivale a pasar de una situación con pesca a una nueva situación sin pesca. Este cambio hace que el punto de equilibrio se mueva de la posición en el plano de la siguiente forma:

$$\left(\frac{c}{d}, \frac{a}{b}\right) \quad \rightarrow \quad \left(\frac{c - \Delta c}{d}, \frac{a + \Delta a}{b}\right)$$

[22]Véase la referencia [**16**].

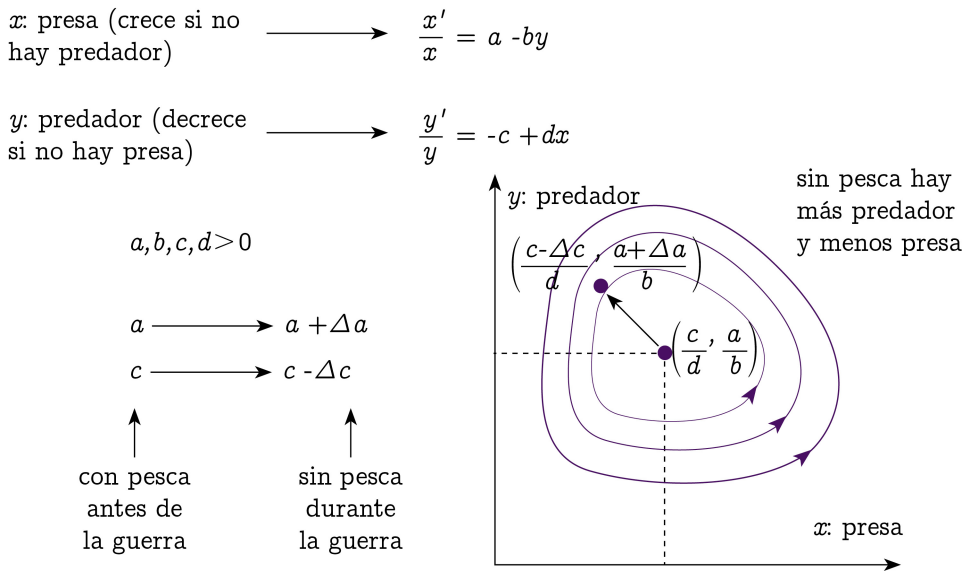


FIGURA 5.10. Modelo de explicación de la disminución de presas después de la Primera Guerra Mundial en el Mar Adriático.

esto es, en una disminución de las presas y un aumento de los predadores en una situación sin pesca, lo que se traduce en un desplazamiento de las órbitas del plano hacia arriba y hacia la izquierda, La Figura 5.10 muestra una síntesis del análisis anterior.

El resultado, que explicaba este aparentemente extraño fenómeno, fue muy celebrado en la época y hasta hoy este modelo es utilizado en problemas más complejos de planificación de pesca y de otros recursos renovables. Para la resolución numérica de este modelo utilizaremos un método de Runge-Kutta de orden 4. Notar que en este caso la función f es una función de $\mathbb{R}^2$ en $\mathbb{R}^2$:

$$f(x,y) = \begin{pmatrix} x(a-by) \\ y(-c+dx) \end{pmatrix}$$

que no depende explícitamente del tiempo (pues supusimos a, b, c y d constantes) y el par (x,y) se maneja como vector. Con todo, el método numérico queda:

RUNGE-KUTTA ORDEN 4 PARA LOTKA-VOLTERRA

$$
\begin{aligned}
(x_0, y_0) &= (x(0), y(0)) \\
g_1 &= f(x_n, y_n) \\
g_2 &= f\left((x_n, y_n) + \frac{h}{2} g_1\right) \\
g_3 &= f\left((x_n, y_n) + \frac{h}{2} g_2\right) \\
g_4 &= f\left((x_n, y_n) + h\, g_3\right) \\
(x_{n+1}, y_{n+1}) &= (x_n, y_n) + \frac{h}{6}\left(g_1 + 2g_2 + 2g_3 + g_4\right)
\end{aligned}
$$

De hecho, es posible verificar que un método de tipo Euler de orden 1 o de Runge-Kutta de orden 2, no provee suficiente precisión para que las órbitas sean cerradas, y es por esto que es necesario utilizar un método de orden 4.

✍ **Ejercicio 5.11.** Construya una planilla de cálculo para obtener las órbitas de Lotka-Volterra utilizando el método de Runge-Kutta de orden 4. La planilla debería ser como la de la Figura 5.11. ❒✍

a= 1 h= 0,41
b= 1
c= 1
d= 1

n	x	y	g1x	g1y	g2x-	g2y-	g2x	g2y	g3x-	g3y-	g3x	g3y	g4x-	g4y-	g4x	g4y
0	**1,5**	**1**	0	0,5	1,5	1,103	-0,15	0,551	1,468	1,113	-0,17	0,521	1,432	1,214	-0,31	0,524
1	**1,44**	**1,22**	-0,31	0,53	1,372	1,325	-0,45	0,493	1,344	1,318	-0,43	0,453	1,26	1,402	-0,51	0,365
2	**1,26**	**1,41**	-0,51	0,366	1,155	1,482	-0,56	0,23	1,146	1,454	-0,52	0,212	1,047	1,494	-0,52	0,07
3	**1,04**	**1,5**	-0,52	0,064	0,936	1,51	-0,48	-0,1	0,945	1,478	-0,45	-0,08	0,858	1,464	-0,4	-0,21
4	**0,85**	**1,46**	-0,39	-0,21	0,772	1,419	-0,32	-0,32	0,787	1,397	-0,31	-0,3	0,725	1,341	-0,25	-0,37
5	**0,72**	**1,34**	-0,24	-0,37	0,672	1,262	-0,18	-0,41	0,686	1,253	-0,17	-0,39	0,651	1,177	-0,12	-0,41
6	**0,65**	**1,17**	-0,11	-0,41	0,627	1,09	-0,06	-0,41	0,638	1,091	-0,06	-0,39	0,626	1,013	-0,01	-0,38
7	**0,63**	**1,01**	-0,01	-0,38	0,625	0,933	0,042	-0,35	0,634	0,939	0,039	-0,34	0,642	0,87	0,083	-0,31
8	**0,64**	**0,87**	0,084	-0,31	0,659	0,805	0,128	-0,27	0,668	0,813	0,125	-0,27	0,693	0,759	0,167	-0,23
9	**0,69**	**0,76**	0,168	-0,23	0,728	0,71	0,211	-0,19	0,737	0,718	0,208	-0,19	0,779	0,68	0,249	-0,15
10	**0,78**	**0,68**	0,25	-0,15	0,831	0,649	0,292	-0,11	0,84	0,657	0,288	-0,11	0,898	0,636	0,327	-0,06
11	**0,9**	**0,64**	0,328	-0,06	0,966	0,622	0,365	-0,02	0,973	0,631	0,359	-0,02	1,046	0,628	0,389	0,029
12	**1,05**	**0,63**	0,39	0,029	1,126	0,634	0,413	0,08	1,131	0,644	0,403	0,084	1,211	0,662	0,409	0,14
13	**1,21**	**0,66**	0,41	0,141	1,297	0,69	0,401	0,205	1,295	0,704	0,384	0,207	1,37	0,747	0,347	0,276
14	**1,37**	**0,75**	0,348	0,277	1,443	0,803	0,284	0,356	1,43	0,819	0,258	0,352	1,477	0,891	0,161	0,425
15	**1,48**	**0,89**	0,161	0,428	1,513	0,979	0,032	0,503	1,487	0,994	0,009	0,484	1,484	1,09	-0,13	0,527

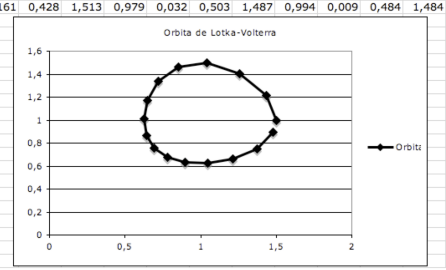


FIGURA 5.11. Planilla de cálculo para el Ejercicio 5.11.

Los resultados numéricos obtenidos con el método de Runge-Kutta de orden 4 descrito anteriormente se muestran en la Figura 5.12, donde se hizo la simulación del paso de una situación con pesca (líneas oscuras) a una situación sin pesca (líneas claras). Observe que el punto de equilibrio que está en el centro de las órbitas se desplaza arriba y hacia la izquierda como predice la teoría.

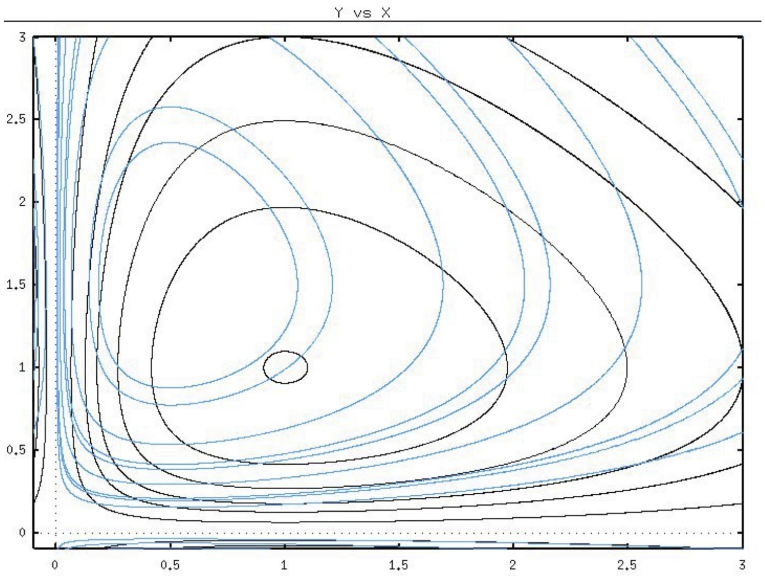


FIGURA 5.12. Resultado de la aproximación utilizando el método de Runge-Kutta de orden 4 para estudiar los equilibrios con (orbitas centradas a la derecha y abajo) y sin pesca (órbitas centradas a la izquierda y arriba). Se observa una disminución de las presas (eje x) de 1 a 0.5 a pesar de la prohibición de pesca. Esto se explica por el aumento de predadores (eje y) de 1 a 1.5. Hay un desplazamiento del punto de equilibrio de las coordenadas (1,1) a (0.5, 1.5). Para el gráfico se utilizó *xppaut* (`www.math.pitt.edu/~bard/xpp/xpp.html`).

Apéndice A: Programas computacionales

A continuación se entrega una lista de los programas computacionales que acompañan esta monografía. Estos pueden encontrarse en la página web del autor o de esta colección. Los algoritmos de planilla de cálculo utilizados a lo largo del libro fueron realizados en *excel*[1] para ilustrar los Capítulos 1, 2 y 5. Estos archivos tienen extensión *xls*. La elaboración de dichas planillas ha sido explicada en detalle durante el desarrollo de cada ejemplo en esta monografía y en principio se requiere un mínimo conocimiento de las reglas de su uso[2]. Los restantes algoritmos fueron escritos en la plataforma *scilab*[3] y fueron desarrollados, sobre todo, para ilustrar los ejemplos de los Capítulos 3, 4 y 5. Estos archivos tienen extensión *sci*.

A.1 Listado de los programas utilizados en este texto

Los programas son listados en el orden en que aparecen en la monografía. Cada vez que se utiliza o hace referencia a uno de ellos en el texto, aparece un símbolo ❒✍ en el margen de la página correspondiente. Los programas *scilab* marcados con ∗∗ se detallan en la siguiente sección. Por razones prácticas, los nombres de programas no llevan tildes.

Programas Capítulo 1. Propagación de errores y redondeo

- ❒✍ `Cap1_efecto_mariposa.xls` (pág. 26)
  Planilla en que se ilustra el efecto de propagación de errores y sensibilidad con respecto a las condiciones iniciales o efecto mariposa, y se genera la Figura 1.1.
- ❒✍ `Cap1_redondeo.xls` (pág. 39)
  Planilla en que se ilustran los distintos usos del redondeo y la truncatura explicados en la Sección 1.5. De este archivo se obtuvo la Figura 1.3.

Programas Capítulo 2. Aproximando π

- ❒✍ `Cap2_Arquimides.xls` (pág. 46)
  Planilla utilizada para generar el Cuadro 2.2 y explicada en la Figura 2.2 donde se ilustra el método de Arquímides para aproximar π.
- ❒✍ `Cap2_Brent_Salamin.xls` (pág. 53)
  Planilla utilizada para generar el Cuadro 2.3 y explicada en la Figura 2.5 donde se ilustra el método de Brent-Salamín para aproximar π.

[1] *E*xcel es marca registrada de Microsoft Corporation.

[2] La única regla más específica, que se utiliza extensivamente, es que se fuerza la copia absoluta de una celda anteponiendo un signo $ a sus coordenadas.

[3] La plataforma de cálculo vectorial científico *scilab* puede obtenerse gratuitamente en `www.scilab.org`.

❒✍ `Cap2_Arquimides_Aitken.xls` (pág. 56)
Planilla utilizada para generar el Cuadro 2.4 y explicada en la Figura 2.6 donde se ilustra el método de Arquímides con aceleración de Aitken para aproximar π.

❒✍ `Cap2_Cambio_de_Base.xls` (pág. 57)
Planilla utilizada para efectuar cambios de base como se explica en la Sección 2.8.

❒✍ `Cap2_Cuenta_Gotas.xls` (pág. 59)
Planilla explicada en la Figura 2.7 donde se implementa el método de cuenta gotas para aproximar π.

Programas Capítulo 3. Ceros, Interpolación e Integración Numérica

❒✍ `Cap3_Punto_Fijo.sci` (pág. 69)
Este programa se utiliza para generar la Figura 3.1 e ilustra las iteraciones de punto fijo para encontrar la raíz de una función.

❒✍ `Cap3_Biseccion.sci**` (pág. 71)
Este programa se utiliza para generar la Figura 3.2 y el Cuadro 3.2 e ilustra las iteraciones de bisección o encajonamientos sucesivos para encontrar la raíz de una función.

❒✍ `Cap3_Newton_Raphson.sci**` (pág. 74)
Este programa se utiliza para generar la Figura 3.3 y el Cuadro 3.3 e ilustra las iteraciones de Newton-Raphson para encontrar la raíz de una función.

❒✍ `Cap3_Comparacion_Biseccion_Newton_Raphson_Secante.sci**` (pág. 77)
Este programa se utiliza para generar el Cuadro 3.4 de comparación de distintos métodos para aproximar la raíz cuadrada.

❒✍ `Cap3_Polinomio_de_Taylor.sci` (pág. 80)
Este programa se utiliza para generar la Figura 3.4 y el Cuadro 3.1 para ilustrar la aproximación de una función en torno a un punto usando polinomios de Taylor.

❒✍ `Cap3_Polinomio_de_Lagrange.sci` (pág. 83)
Este programa se utiliza para generar el Cuadro 3.5 y la Figura 3.5 e ilustra el uso de polinomios de Lagrange para aproximar una función.

❒✍ `Cap3_Cuadratura.sci**` (pág. 91)
Este programa se utiliza para generar el Cuadro 3.7 donde se comparan distintos métodos de cuadratura para la integral definida.

Programas Capítulo 4. ¿Cómo y por qué resolver sistemas lineales?

❒✍ `Cap4_Tomografia_Computarizada.sci` (pág. 121)
Este programa se utiliza para generar las Figuras 4.3 y 4.4 donde se resuelve el poblema de la tomografía computarizada a través de la resolución de un sistema lineal.

Programas Capítulo 5. ¿Cómo y por qué resolver ecuaciones diferenciales?

❒✍ `Cap5_Metodo_de_Euler_Progresivo.xls` (pág. 128)
Planilla utilizada para generar la Figura 5.1 donde se explica la implementación del método de Euler progresivo.

❒✍ `Cap5_Metodo_de_Euler_Inestabilidad.xls` (pág. 131)
Planilla utilizada para generar la Figura 5.2 explicada en la Figura 5.3.

❒✍ `Cap5_Euler_Progresivo_Malthus.xls` (pág. 134)
Planilla en que se implementa el método de Euler progresivo para estimar la población mundial en el periodo $2000 - 2100$ y de donde se obtiene el Cuadro 5.1.

❒✍ `Cap5_Arbol_de_Feigenbaum.sci**` (pág. 136)
Este programa se utiliza para generar la Figura 5.7 del árbol de bifurcaciones para el modelo logístico discreto.

❒✍ `Cap5_SIR.xls` (pág. 139)
Planilla en que se implementa el método de Euler progresivo para resolver el modelo de propagación de una epidemia explicado en la Figura 5.8 y conocido como modelo SIR.

❒✍ `Cap5_Metodo_de_Heun.xls` (pág. 143)
Planilla que sirve para resolver el Ejercicio 5.9 donde se pide implementar el método de Heun.

❒✍ `Cap5_Runge-Kutta_Lotka-Volterra.xls` (pág. 147)
Planilla en que se implementa el método de Runge-Kutta de orden 4 para calcular las órbitas del sistema de Lotka-Volterra mostradas en la Figura 5.12. Corresponde a la solución del Ejercicio 5.11.

A.2 Ejemplos de algunos de los algoritmos programados

Aunque no podemos dar aquí las nociones básicas [4] de programación en *scilab* (asignaciones, ciclos `for`, condicionales `if`), el lector puede comparar entre las versiones programadas y los algoritmos tal y como fueron presentados en el texto.

`Cap3_Biseccion_simple.sci**`
(algoritmo de bisección pág. 71)

```
// Método de bisección para aproximar una raíz de f(x)=log(x)-sin(x)
deff('[y]=f(x)','y=log(x)-sin(x)')
a0=1; b0=3;
x=[a0:0.01:b0]';
y=f(x);
a=a0; b=b0;
for i=1:100
  c=(a+b)/2;
  if f(a)*f(c)<0 then
    b=c;
  else
    a=c;
  end
end
raiz=c
```

[4] Ver *Introducción a Scilab* manual básico en español que se puede encontrar en `www.matematicas.unal.edu.co/~hmora/sci.pdf`.

Cap3_Comparacion_Biseccion_Newton_Raphson_Secante.sci**
(algoritmos bisección: pág. 76, Newton: pág. 77 y secante: pág. 77)

```
// Comparación de los métodos de bisección,
// Newton-Raphson y de la secante para aproximar la
// raíz cuadrada de M (Capitulo 3)
M=2;
printf('\nAproximacion de raiz de %d\n\n',M);
printf('  n    bisección  Newton     secante\n');
//número de iteraciones
for niter=[1:4,5:5:20]
// Función f(x)=x^2-M
//Método de bisección
a0=1; b0=M;
a=a0; b=b0;
for i=1:niter
  c=(a+b)/2;
  if c^2>M then
    b=c;
  else
    a=c;
  end
end
raiz_biseccion=c;
//Método de Newton-Raphson
x0=1; xn=x0;
for i=1:niter
  xp=1/2*(xn+M/xn);
  xn=xp;
end
raiz_Newton=xp;
//Método de la secante
x0=1; x1=M;
xn=x0;xnn=x1;
for i=1:niter
  xp=(xn*xnn+M)/(xn+xnn);
  xnn=xn;
  xn=xp;
end
raiz_secante=xp;
printf('%4d %10f %10f %10f\n',niter,raíz_bisección,raíz_Newton,raíz_secante);
end
printf('Raíz exacta es %f\n',sqrt(M));
```

Cap3_Newton_Raphson_simple.sci**
(algoritmo de Newton-Raphson pág. 73)

```
// Método de Newton-Raphson
// para aproximar una raíz de
// la función x^3-x-3

deff('[y]=f(x)','y=x^3-x-3');
deff('[yp]=fp(x)','yp=3*x^2-1');

x0=1;
x=[0.8:0.01:2.6]';
y=f(x);

xn=x0;
for i=1:100
  xp=xn-f(xn)/fp(xn);
  xn=xp;
end
raiz=xp
```

Cap3_Cuadratura.sci**
(algoritmos rectángulos: pág. 88, trapecios: pág. 89 y Simpson: pág. 90)

```
// Comparación de las distintas fórmulas de cuadratura:
// rectángulos, trapecios y Simpson
function func=f(x)
 func=sin(x);
endfunction;
a=0; b=%pi
N=10; h=(b-a)/N;
x=[a:h:b];
suma1=0; suma2=0; suma3=0;
for i=1:N
 suma1=suma1+h*f(x(i)+h/2);
 suma2=suma2+h/2*(f(x(i))+f(x(i+1)));
 suma3=suma3+h/6*(f(x(i))+4*f(x(i)+h/2)+f(x(i+1)));
end;
integral_rectangulos=suma1
integral_trapecios=suma2
integral_Simpson=suma3
```

Cap5_Arbol_de_Feigenbaum_simple.sci**
(pág. 138)

```
function [ ]=arbol_Feigenbaum()
// Árbol de Feigenbaum obtenido del
// modelo logístico discreto P(n+1)=sigma*P(n)*(P(n)-M)
// N: número de iteraciones, M: población máxima

N=1000; v=200; M=100; eps=0.01;n_equi=0;
s_inicio=1.8;s_fin=3.0;s_paso=0.001;

for sigma=s_inicio:s_paso:s_fin
 p=1.01*M;
 for i=1:N-1
  p(i+1)=p(i)+sigma/100*p(i)*(M-p(i))
 end
 xsigma=zeros(N-v:N)'+sigma;
 plot2d(xsigma,p(N-v:N),0,"011"," ",[s_inicio 0 s_fin 140]);
end

endfunction
```

*.sci

```
//Aquí hay espacio para que usted ensaye sus propias líneas de programa.
```

Bibliografía

[1] Aguayo, J. *Cálculo Integral y Series*. J.C. Sáez Editor, Santiago, 2011.
[2] Bailey, D., Borwein, P., Plouffe, S. *On the rapid computation of various polylogarithmic constants*. Manuscript, 1996.
[3] Beckman, P. *The History of Pi*. The Golem Press, Boulder, Colorado, 1971.
[4] Berggren, L., Borwein J., Borwein P. *Pi: A Source Book*. Springer-Verlag. New York, 1997.
[5] Cajori, F. *A History of Mathematics*. MacMillan and Co. Londres, 1926.
[6] Clawson, C.C. *Misterios matemáticos. Magia y belleza de los números*, Ed. Diana, México, 1999.
[7] Deuflhard, P., Hohmann, A. *Numerical Analysis in Modern Scientific Computing, An Introduction*, Text in Applied Mathematics, 2da edición, Springer, New-York, 2003.
[8] Delahaye, J.-P. *Obsession de* π, Pour la Science 231, Enero 1997.
[9] Ekeland, I. *Le Chaos*, Flammarion, 1995.
[10] Eymard P., Lafon J.-P. *Autour du nombre* π, Actualités Scientifiques et Industrieles, 1143, Hermann, Paris, 1999.
[11] Feynman, R. *El Carácter de las Leyes Físicas*, Editorial Universitaria, Santiago, 1972.
[12] Gajardo, P. *Modelando Fenómenos de Evolución*. J.C. Sáez Editor, Santiago, 2011.
[13] Gil, O. *Excursiones por el Álgebra Lineal y sus Aplicaciones*. J.C. Sáez Editor, Santiago, 2011.
[14] Gourévitch, B. *Les algorithmes compte-gouttes*. 2009, De la página *L´univers de Pi* `http://www.pi314.net`
[15] Héron, B., Issard-Roch, F., Picard, C. *Analyse Numérique, Excercises et problèmes corrigés*, Dunod, Paris, 1999.
[16] Hofbauer, J., Sigmund, K. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998.
[17] Lacourly, N. *Estadística Multivariada*. J.C. Sáez Editor, Santiago, 2011.
[18] Lewin, R. *Introducción al Álgebra*. J.C. Sáez Editor, Santiago, 2011.
[19] Maor, E. *e: the Story of a Number*, Princeton University Press, Princeton, 1994.
[20] Melvin, H. *A course in Numerical Analysis*, Harper International Edition, New York, 1968.
[21] Moler, C. B. *Numerical Computing with Matlab*. SIAM, Philadelphia, 2004.
[22] Perelman, Y. *Álgebra Recreativa*, Editorial Mir, Moscú, 1969.
[23] Rabinowitz, S., Wagon, S. *A Spigot Algorithm for the Digits of* π, The American Mathematical Monthly, 1995, 102(3), 195–203.
[24] Rappaz, J., Picasso, M. *Introduction àl Ánalyse Numérique*, Presses polytechniques et universitaires romandes, Lausanne, 2004.
[25] Simmons, G.F. *Ecuaciones Diferenciales. Con aplicaciones y notas históricas*, Segunda Edición, McGraw-Hill, Madrid, 1993.
[26] Stewart, I. *Les algorithmes compte-gouttes*, Pour la Science, 1995, 215, 104–107.
[27] Stewart, I. *L´univers des nombres*. Collection Bibliothèque pour La Science. Belin, Paris, 2000.
[28] Stewart, I. *¿Juega Dios a los dados? La nueva matemática del caos*, Dracontos, Barcelona, 2007.
[29] Trefethen, L. N. *The definition of Numerical Analysis*, SIAM News, November 1992.
[30] Utreras, F. *Análisis Numérico I*, Apuntes del Departamento de Ingeniería Matemática, Universidad de Chile, Santiago, 1983.

Índice de figuras

Índice de cuadros

Índice de Términos