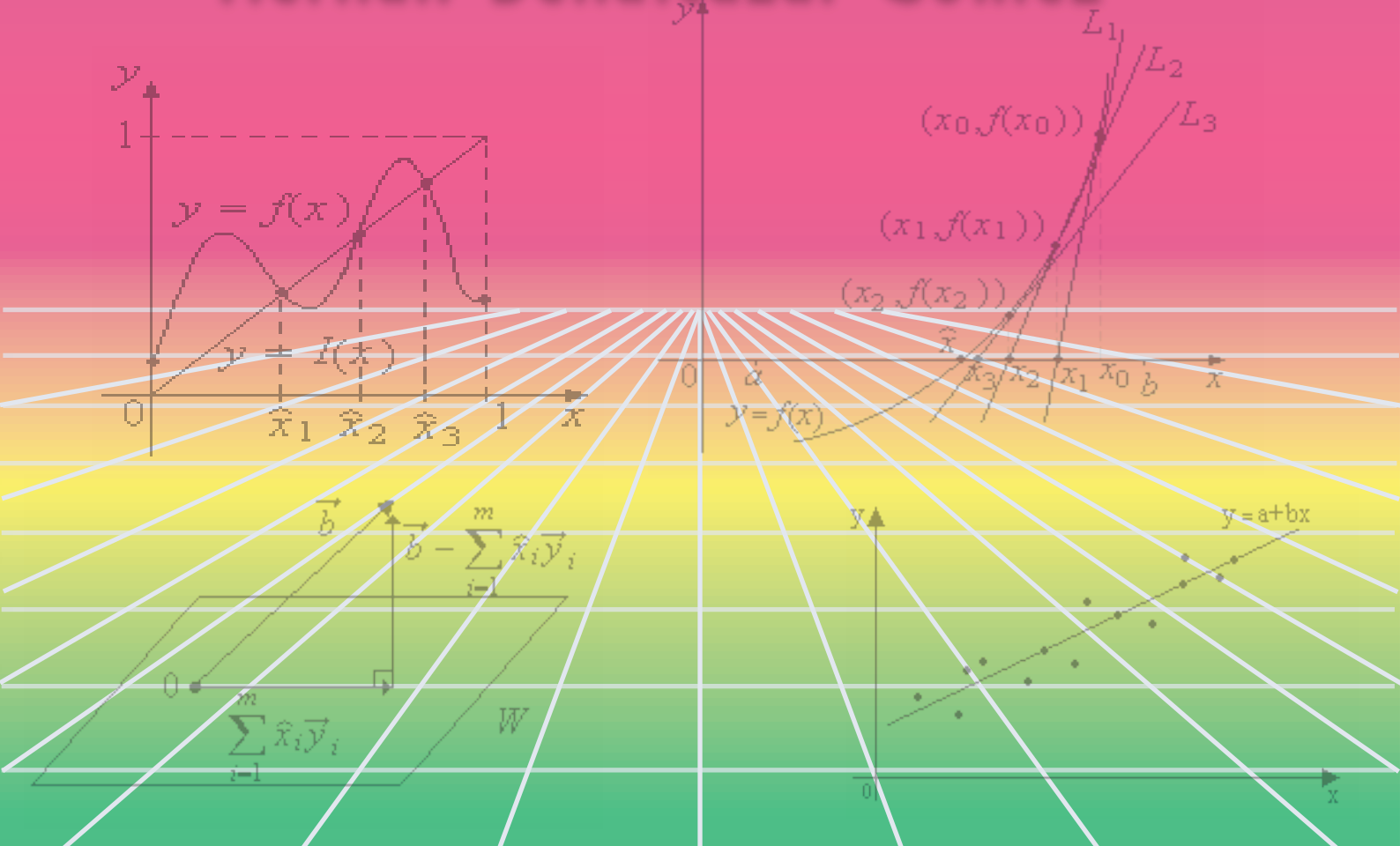
Serie de Matemática Universitaria

ANÁLISIS NUMÉRICO

Hernán Benalcázar Gómez



Análisis Numérico

Hernán Benalcázar Gómez

Quito, noviembre del 2007

Dedicatoria

A mi esposa y a mi hijo.

A mis padres, siempre presentes apoyándome en todos mis proyectos.

Introducción

El análisis numérico es una parte de la matemática y tiene su crecimiento a partir de la década de los cuarenta del siglo pasado, crecimiento que va junto con el de los computadores. Se desarrolla en base a las necesidades de resolver problemas complejos que surgen en las ingenierías, las ciencias físicas, químicas, biológicas, la economía y ciencias sociales, en la industria. En la actualidad, el análisis numérico es parte de la malla curricular de la mayor parte de las carreras de ingeniería y de ciencias fundamentales, y se constituye en la base para la generación de métodos de simulación asistido por computadora ampliamente utilizados en el sector industrial, y últimamente en el ambiental y climático. Los países desarrolados son los que han dado mayor importancia al análisis numérico y a la simulación numérica; en nuestro país es muy poco lo que se hace en matemática y particularmente en análisis numérico.

Este libro es una introducción al análisis numérico. Está destinado a los estudiantes de segundo o tercer años de la carreras de ingeniería y en especial de informática, computación gráfica, de diseño industrial, mecánica, electrónica, química y muy particularmente a los estudiantes de ingeniería matemática de las Escuelas de Ciencias, a los estudiantes de las maestrías en docencia matemática, estadística y optimización, entre otras, a matemáticos e ingenieros interesados en aplicaciones del análisis matemático, del álgebra lineal y de las ecuaciones diferenciales ordinarias. Está basado en las notas que el autor ha impartido en cursos de pregrado y posgrado en varias Universidades y Escuelas Politécnicas del Ecuador.

Los requisitos para el estudio de este libro son los cursos de análisis matemático I y II, de álgebra lineal, como los que se dictan en las Escuelas de Ciencias. Más exactamente se requiere del conocimiento de resultados fundamentales del cálculo diferencial e integral de funciones en una y en varias variables, de las sucesiones y series numéricas, de las sucesiones y series de funciones, de algunos tipos de ecuaciones diferenciales ordinarias de primer orden, y del lado del álgebra lineal, se requiere de conocimientos básicos de los espacios vectoriales, las aplicaciones lineales y matrices, de los sistemas de ecuaciones lineales, de diagonalización de matrices.

El texto contiene once capítulos y un apéndice, cada uno de ellos está dividido en secciones y subsecciones. Al inicio de cada capítulo se presenta el resumen del mismo. Contiene ejemplos y ejercicios resueltos algunos de ellos originales, y una gran cantidad de ejecicios propuestos, una parte de ellos originales, lo que enriquece el material que se ofrece al estudiante. Los resultados numéricos que se presentan en cada uno de los capítulos, en unos casos se han obtenido simplemente con una calculadora de bolsillo, y en otros donde el caso lo amerita, se han elaborado programas en Fortran 77 que han sido corridos en una máquina Pentium V. Más aún, todos los algoritmos propuestos han sido debidamente verificados. Además, en algunos temas y ejercicios se forza al estudiante a que realice sus propios programas y se vuelva un productor de software, más no un consumidor. Al final del capítulo se muestra una amplia bibliografía que van de textos muy elementales a textos muy avanzados y que pueden ser útiles sobre todo para los estudiantes de maestrías que preparan tesis de graduación, como también para que el estudiante de pregrado pueda disponer de otros enfoques que ofrecen muchos libros importantes de análisis numérico que se han publicado.

El primer capítulo está destinado a introducir el lenguaje del análisis numérico y a iniciar en el cálculo aproximado. Se comienza con los elementos del cálculo numérico y de los algoritmos. A continuación se muestran algunos ejemplos de algoritmos y de resolución numérica de problemas elementales. Se

consideran los sistemas de numeración que permiten explicar la representación en punto fijo y en punto flotante. Se estudian los tipos de errores, particularmente los de redondeo, de aproximación (truncamiento y discretización) y la propagación de los mismos. Se introducen las nociones de condicionamiento, estabilidad numérica.

En el segundo capítulo se tratan tres tipos de problemas: la interpolación polinomial, la derivación y la integración numéricas. Todos estos son tratados en el ámbito de los espacios duales, es decir como formas lineales definidas en apropiados espacios vectoriales. Se trata el problema de la existencia del polinomio de interpolación de Lagrange, el error de interpolación. A continuación se estudia la aproximación numérica de derivadas de primero y segundo orden así como de derivadas parciales. Luego, se pasa al estudio de métodos numéricos de integración de funciones de una sola variable, y la aplicación de estos al cálculo de integrales dobles.

El capítulo tres está destinado al cálculo aproximado de series de funciones. Para el efecto, se inicia con una revisión de resultados de las series numéricas y de funciones. Se presta mayor atención a las series de potencias y particularmente a las series de Taylor y su aproximación numérica. Se elaboran algoritmos de las funciones trascendentes más importantes como son las trigonométricas, logaritmo y exponencial, los mismos que son las bases de los algoritmos utilizados en calculadoras de bolsillo y los implementados en los lenguajes de programación como por ejemplo C, C++, Fortran, Delphi, etc. Posteriormente se trata la integración de funciones representadas como series de potencias y se dan aplicaciones. La aproximación de series de Fourier se trata en el capítulo noveno.

En el capítulo cuarto se da respuesta a una pregunta simple: ¿cómo se elaboran las tablas de las funciones de distribución de probabilidades? Se consideran las funciones de probabilidad discretas y continuas. Dentro de las discretas se tratan la binomial y de Poisson. De las continuas se consideran las funciones de distribución tipos gama y beta, normal $\chi i - cuadrada$, t de Student, de Snedekor. Se elaboran algoritmos de cada una de ellas que pueden ser implementados fácilmente en los programas de simulación. Es importante precisar que en muchos textos, sobre todo de métodos de perturbación, se aborda la función error y su aproximación mediante métodos asimptóticos; esta función es muy similar a la función de distribución normal, igualmente se tratan las funciones gama y beta de Euler. En esos textos, para estas funciones no se dan algoritmos completos de aproximación. De las otras funciones continuas de distribución arriba citadas, se ha encontrado escasamente algunos resultados, por lo que el material que aquí se presenta no se ha hallado, al menos en los libros citados en la bibliografía.

El capítulo quinto está destinado al cálculo aproximado de raíces de ecuaciones. Se inicia con la aplicación del teorema de Bolzano a la búsqueda del cambio de signo así como el método de bisección. A continuación, basado en el teorema de Banach del punto fijo se construyen aplicaciones contractivas que están relacionadas con las ecuaciones propuestas y desarrollan algunos métodos de aproximación clásicos. Se trata la convergencia de estos métodos así como dos métodos de aceleración de la convergencia. Se concluye con el estudio de las raíces de polinomios.

Los sistemas de ecuaciones lineales son el objeto del capítulo sexto. Se presentan algunos ejemplos que originan sistemas de ecuaciones lineales, posteriormente se trata los problemas con sistemas de ecuaciones lineales. Para la selección del método numérico es importante tener un conocimiento preciso de las características de la matriz del sistema, es por esto que se presta atención al estudio de algunos tipos de matrices. Luego se focaliza el trabajo en los métodos clásicos de resolución de sistemas de ecuaciones lineales como son: eliminación gaussiana, factorización LU de Crout, factorización L^TL de Choleski. Estos métodos se adaptan particularmente a las matrices tridiagonales. Se concluye con la resolución en norma mínima de sistemas de ecuaciones lineales que tienen una infinidad de soluciones.

En el capítulo séptimo se tratan métodos iterativos de resolución de sistemas de ecuaciones lineales y no lineales. Se consideran primero los sistemas de ecuaciones no lineales. Para el efecto, se revisan algunos resultados de la diferencial de Fréchet y se vuelve a considerar el teorema de Banach del punto fijo, a continuación se trata el método de Newton. Posteriormente, se tratan los métodos de resolución de sistemas de ecuaciones lineales, a saber: el método de Jacobi, Gauss-Seidel y SOR.

El capítulo octavo está destinado al cálculo de los valores y vectores propios. Se inicia con la revisión de algunos resultados fundamentales. Luego se considera la aplicación de los valores y vectores propios

a las cónicas. Por simplicidad, se considera el cálculo de los valores y vectores propios de matrices reales de 3×3. Se considera el método de la potencia para el cálculo del mayor valor propio de una matriz diagonalizable.

Los problemas de mínimos cuadrados se abordan en el capítulo noveno. Se inicia con la resolución de sistemas de ecuaciones lineales en mínimos cuadrados. A continuación se trata el método de Householder que constituye uno de los métodos más importantes para la resolución de sistemas de ecuaciones lineales en mínimos cuadrados así como para el cálculo de vectores propios. Posteriormente se fija la atención en los problemas de ajuste de datos para varios tipos de problemas. Se concluye con los problemas de mínimos cuadrados continuos, particularmente la aproximación numérica de series de Fourier.

En el capítulo décimo se da una breve introducción hacia la teoría de los splines. Básicamente se abordan los splines cúbicos de interpolación y los B-Splines.

Los métodos numéricos para calcular soluciones aproximadas de ecuaciones diferenciales ordinarias tienen lugar en el capítulo décimo primero. Se abordan dos clases de problemas: los de Cauchy de valor inicial y los de valores en la frontera. Para la primera clase de problemas se consideran los métodos de Euler explícitos e implícitos, el método implícito de Crank-Nicolson, todos estos se hallan en la mayor parte de los textos citados en la bibliografía, que no es el caso del método de Petrov-Galerkin que aquí es tratado. Este método se aplica fundamentalmente a problemas de valores en la frontera y se encuentra en textos muy especializados. Se prefirió incluir el método de Petrov-Galerkin y no los ampliamente conocidos métodos de Runge-Kutta, pués estos se los encuentra en la mayor parte de libros de ecuaciones diferenciales y análisis numérico. En la segunda clase de problemas se consideran ecuaciones diferenciales de segundo orden con condiciones de frontera de Dirichlet homogéneas y no homgéneas, de Neumann homogéneas y no homogéneas, y mixtas. Todos estos problemas se aproximan con el método de diferencias finitas. Se concluye con la resolución numérica de un problema no lineal.

Se ha suministrado un apéndice que contiene básicamente una breve revisión de los resultados más importantes de los espacios vectoriales y algunos ejemplos, y, una revisión de los espacios normados y de los espacios con producto interior.

Al escribir este libro se buscó un equilibrio entre abstracción, practicidad, popularidad, simplicidad, novedad, actualidad de métodos de cálculo, lo que condujo a no incluir algunos temas que se consideró muy complejos y surgieron algunas preguntas: ¿por qué no se trató tal o cual tema? ¿por qué unos temas tuvieron mayor atención que otros posiblemente más importantes? ¿Cómo juzgar que temas son trascendentales para un público tan variado? La nueva versión de este libro está ya preparada, se dará mayor atención a temas, que en un principio se consideró muy complejos pero que luego se vió la necesidad de tratarlos, como los siguientes: resolución de sistemas de ecuaciones lineales con los métodos Minres y Gmres, problemas no lineales de ajuste de datos dependientes de varios parámetros, método de integración de Gauss, método de Householder para el cálculo de valores y vectores propios, ampliación de la teoría de splines, resultados de existencia de ecuaciones diferenciales ordinarias y convergencia de los métodos propuestos así como los muy populares métodos de Runge-Kutta. Todos estos temas tendrán también una ampliación de ejemplos.

Mucho agradeceré se me comunique de posibles errores tipográficos y deslices, que por cierto son infaltables a pesar del esfuerzo en controlarlos y eliminarlos.

Mi agradecimiento al señor Darwin Polivio Narváez Vicente que muy responsablemente colaboró y mostró mucha capacidad y profesionalismo en el levantamiento del texto.

.

Hernán Benalcázar Gómez

Profesor de la Escuela de Ciencias

Índice general

Capítulo 1

Cálculo aproximado, algoritmos, errores

Resumen

En este capítulo se realiza un tour corto en los métodos numéricos. Se inicia con la presentación de una metodología para el análisis de problemas y las soluciones aproximadas, la elaboración de algoritmos y algunas nociones de la complejidad de los mismos. A continuación se presentan ejemplos de algoritmos simples así como de algunos problemas elementales que se presentan en el ámbito del álgebra lineal y del análisis matemático, y, métodos simples de resolución numérica. Se hace un corto análisis de los tipos de errores. El uso de instrumentos de cálculo como son las calculadoras de bosillo y los computadores motivan el estudio de la representación en punto flotante, los errores de redondeo y la aritmética en punto flotante, temática que a su vez requiere del análisis de los sistemas de numeración. Luego se realiza un estudio del condicionamiento de funciones de una y varias variables que está relacionado con la amplificación de los errores de redondeo. Particular atención se pone en las operaciones aritméticas, lo que permite establecer una jerarquía en las mismas e identificar que operaciones son las peligrosas y bajo que condiciones y cuales no son peligrosas, lo que constituye una ayuda extremadamente grande cuando se elaboran los algoritmos de cálculo. Mediante algunos ejemplos se analiza el problema de la propagación de los errores así como el de la estabilidad numérica.

1.1. Introducción

Uno de los objetivos importantes del Análisis Numérico es la elaboración de métodos, procedimientos de cálculo y construcción de algoritmos que con la utilización de instrumentos de cálculo como las calculadoras de bolsillo o de instrumentos de cálculo mucho más complejos como los computadores, que requieren de la elaboración de programas computacionales, permitan calcular soluciones exactas o aproximadas de una diversidad de problemas matemáticos de modo que con cualesquiera de estos instrumentos, se deba tener un control sobre los errores cometidos en los cálculos y que los resultados finales sean de calidad.

Por otro lado, los procedimientos de cálculo, los algoritmos numéricos deben ser, en lo posible, los más simples, concisos, de aplicabilidad a una amplia variedad de situaciones. El costo numérico de cada procedimiento o algoritmo y su programa computacional que se construya debe ser, en lo posible, el más pequeño.

La calidad de la solución de un problema dado depende de muchos factores, entre ellos, de los datos de entrada que se requieren para la ejecución del algoritmo, procedimiento o programa computacional construido, así como de los instrumentos de cálculo utilizados, del lenguaje de programación y de la versión del mismo. Es claro que la calidad de la solución depende fuertemente del método numérico empleado y este a su vez depende de dos componentes importantes: el condicionamiento y la estabilidad; y, para problemas cuyas soluciones se aproximan mediante sucesiones, dependen a más de todos los componentes anteriores, de la convergencia.

En este capítulo se tratan algunos elementos de los algoritmos y características de los programas computacionales, los tipos de errores comunes en análisis numérico. Se revisa brevemente los sistemas de numeración entre los que se destacan el binario y el decimal, la representación en punto fijo y punto flotante, los errores de redondeo y la aritmética en punto flotante. Se introducen las nociones elementales de condicionamiento, estabilidad numérica y convergencia que son muy importantes en la construcción de algoritmos, procedimientos de cálculo y de la elaboración de programas computacionales, y que constituyen las bases que deben tenerse siempre presentes para el desarrollo de software en el cálculo científico.

1.2. Cálculo numérico. Algoritmos

Suponemos que un problema (P) ha sido planteado y que requiere de su resolución. Tres situaciones se presentan: la primera en la que la solución del problema (P) podemos encontrarlo directamente y no se requiere del cálculo numérico. La segunda en la que la solución del problema (P) podemos encontrarlo directamente y se requiere de la implementación de un procedimiento de cálculo para aproximar la solución encontrada. La tercera en la que no es posible encontrar directamente la solución y se requiere de un método numérico para aproximar la solución. Son estas dos últimas situaciones que nos interesan. Más aún, en la resolución numérica de un problema matemático (P) se establece la siguiente metodología.

1. Estudio de la existencia de solución del problema (P).
2. Construcción de un método numérico que aproxime la solución del problema (P).
3. Elaboración del respectivo algoritmo o procedimiento de cálculo.
4. Elaboración de un programa o código numérico para el cálculo de la solución aproximada de (P).
5. Realización de pruebas para validar el algoritmo o procedimiento de cálculo y el programa computacional.

Desde el punto de vista práctico, esto es, problemas que surgen en las ciencias y en la industria, la metodología presentada se extiende con la calibración de la solución y luego viene la implementación de la solución. En este curso daremos énfasis fundamentalmente a los puntos 1), 2), 3) y 5) de la metodología precedente.

El punto 4) no lo abordaremos y dejamos al estudiante que elija el lenguaje de programación que le interese para la elaboración de sus propios programas computacionales con los que debe realizar pruebas para verificar resultados mediante la implementación del algoritmo así como verificar la correcta elaboración del programa computacional; o en su defecto, seleccione el paquete de programas de tipo comercial (Matlab, Matemática, etc.) en el que provea la solución del problema planteado con el algoritmo propuesto. Es un error gravísimo el modificar el problema planteado (P) a uno $(\widehat{P})$ cuya solución está implementado en el paquete de programas computacionales. Por otro lado, es también importante el uso de ciertas herramientas informáticas que ayuden a presentar de mejor manera los resultados y permita comprender mejor las soluciones, por ejemplo graficadores para presentar gráficas de curvas 2d, 3d, superficies, flujos, generación de mallas estructuradas y no estructuradas, etc.

El estudio de la existencia de una solución o soluciones del problema (P) es muy importante. Pués en él se deben conocer con precisión las hipótesis con las cuales nuestro problema tiene solución, y bajo que condiciones el problema (P) puede no tener solución. En muchos casos, en el estudio de existencia de soluciones se construye el método que conduce a encontrar la solución de (P). Si el problema no tiene solución, carece de sentido el intentar elaborar un método numérico de solución.

Debido a que los cálculos que se realizan son con números que tienen un número finito de cifras decimales, estos afectan los resultados, por lo que el control de los errores en los cálculos es fundamental, es decir, debemos conocer la precisión con la que obtenemos la solución numérica del problema (P). Este es uno

de los problemas centrales del análisis numérico y que están ligados con las nociones de consistencia y la estabilidad numérica. En el caso en que la solución de (P) se calcula como límite de una sucesión de soluciones de problemas (Pn) más sencillos a resolver, otro de los problemas centrales del análisis numérico es probar o demostrar que las soluciones de esos problemas más sencillos converge a la solución del problema (P), es decir se debe probar la convergencia del método numérico propuesto. La consistencia, estabilidad y convergencia se discutirán más adelante.

Tanto en el estudio de existencia de soluciones como en la elaboración del método numérico se identifican los datos que se requieren para resolver el problema. Una parte de estos datos los conocemos como datos de entrada.

Una vez establecido el método numérico, se pasa enseguida a la elaboración o construcción del algoritmo. En la definición siguiente se establece la noción de algoritmo en su versión la más simple

Definición 1 *Se llama algoritmo a una sucesión finita de operaciones elementales, que organizada como pasos o procedimientos, se describen en forma lógica como calcular la solución de un problema (P) de modo eficaz con datos de entrada dados.*

Un algoritmo contiene los siguientes elementos:

1. **Datos de entrada**: que consisten en valores o datos de partida, los cuales son asignados antes de arrancar la ejecución del algoritmo. Estos datos permiten inicializar el algoritmo para su ejecución.

   Es necesario verificar la lectura correcta de todos los datos de entrada.

   Los datos de entrada dependen obviamente del problema propuesto. Estos pueden ser datos que pertenecen a distintos conjuntos numéricos (enteros, reales, complejos), pueden ser funciones reales como las trigonométricas (seno, coseno, tangente y sus inversas), las funciones exponencial, logarítmica, las funciones hiperbólicas, polinomios, etc, pueden ser datos vectoriales como son los elementos de $\mathbb{R}^n$, pueden ser matrices, etc.

2. **Algoritmo o procedimiento**: constituye la secuencia de todos los pasos o procedimientos de cálculo que se deben ejecutar. Estos deben ser claros, precisos, lógicos. No se deben tener ambigüedades en la descripción de esos pasos o procedimientos. Debe considerarse todas las situaciones posibles que se presenten.

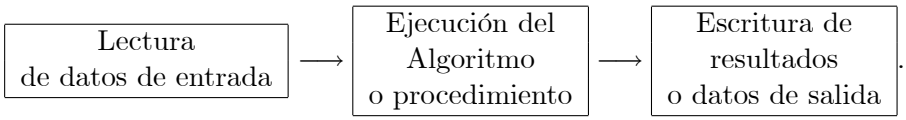
   La ejecución del algoritmo o procedimiento concluye siempre con un número finito de pasos.

3. **Datos de salida**: son una o más cantidades que tiene una relación estrecha con los datos de entrada. Estos resultados están definidos de manera única por los pasos del procedimiento o algoritmo.

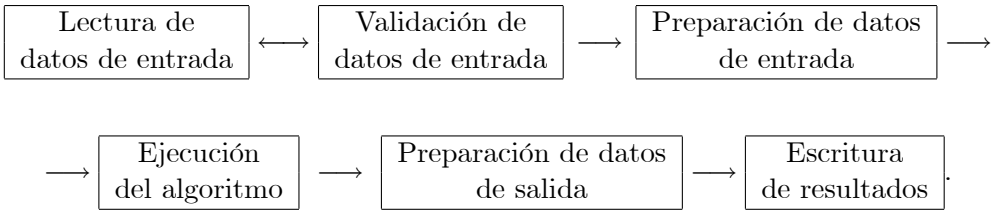
La escritura de un algoritmo contiene los datos de entrada, los datos de salida, y a continuación el procedimiento o la descripción del método a utilizar que constituye el algoritmo propiamente dicho que generalmente se lo expresa en pseodocódigo de modo que facilite la escritura de un programa computacional en cualquier lenguaje de programación.

Más adelante se proponen muchos algoritmos que permiten aclarar todas estas ideas.

En el siguiente esquema se muestra la secuencia de estos tres bloques:



En la práctica estos tres bloques no son suficientes para escribir un programa computacional. Un análisis más detallado de estos tres bloques proponemos en el diagrama siguiente:



En lo posible se busca construir algoritmos que tengan las características siguientes:

1. **Aplicabilidad general**: el algoritmo debe funcionar para una clase de problemas lo más amplia posible, donde las soluciones de un problema específico de la clase resulten solamente por cambios en los datos de entrada.

2. **Simplicidad**: Un algoritmo o procedimiento tiene que ser, en lo posible, simple de programar.

3. **Confiabilidad y seguridad**: el algoritmo no debe ser numéricamente costoso. En lo posible, se debe reducir el número de operaciones elementales a ejecutar. Esto evita que se amplifiquen los errores de redondeo, dando resultados más precisos, y por otro lado, reducen los tiempos de máquina.

   Se debe reducir, en lo posible, el número de variables a utilizar. Igualmente, se debe reducir en lo posible el número de subrutinas o bucles a utilizar, así como la repetición de ciertos cálculos.

   Se deben efectuar tests o pruebas con datos de entrada los más variados a fin de asegurarse que el algoritmo está correctamente elaborado y que los resultados son correctos o muy aceptables. Se buscará, en lo posible, ejemplos que se conozcan las soluciones exactas para compararse con las soluciones numéricas.

   En el estudio de un método numérico y consecuentemente de un algoritmo es importante, siempre que sea posible, determinar el número de operaciones elementales que se realizan para obtener la solución numérica del problema, el número de comparaciones, son menos importantes las reasignaciones . Entenderemos como operaciones elementales a las operaciones aritméticas como la suma, resta, multiplicación, división, raíz n-ésima. Las comparaciones están vinculadas con las relaciones de orden menor que $<$, mayor que $>$, menor o igual que $\leq$, mayor o igual que $\geq$. Prestaremos mayor atención a la determinación del número de operaciones elementales que se requieren para calcular la solución numérica mediante un método o procedimiento está relacionado con la complejidad del algoritmo que análizamos a continuación.

   Complejidad de algoritmos.

   Si para un problema (P) se conocen varios métodos y por lo tanto se pueden proporcionar varios algoritmos, es importante analizar la denominada complejidad del algoritmo. Esta tiene que ver con dos componentes importantes: uno del punto de vista volumen de memoria necesario del instrumento o equipo utilizado para el cálculo, y otro del punto de vista tiempo de máquina que a su vez está relacionado con el número de operaciones elementales (siempre que haya sido posible obtener) que se requieren para calcular la solución. Si se disponen de dos métodos, ¿cómo juzgar que método es mejor? ¿bajo que circunstancias un método es mejor que otro?. Para poder dar respuesta a estas interrograntes debemos estudiar la complejidad de cada algoritmo, esto es, determinar cuánto de memoria se requiere en la ejecución de cada método, el tiempo de máquina requerido para el cálculo de la solución con cada método.

   Cuando un problema (P) puede ser resuelto mediante dos métodos generados por sucesiones de problemas más simples que los notamos $(P_n^{(1)})$ y $(P_n^{(2)})$, el estudio de la convergencia de cada método es importante, esto nos proporcionará un dato que está relacionado con el orden de convergencia, ¿cuál método es mejor?. Para responder a esta interrogante, debemos considerar otro elemento que es la exactitud de la solución numérica que a su vez está relacionada con el orden de convergencia. Desde este punto de vista, el método generado que tenga un orden de convergencia más alto será mejor que el otro, lo que da respuesta a la interrogante.

1.3. Ejemplos de algoritmos y problemas

La metodología arriba propuesta la aplicaremos a algunos ejemplos que proponemos a continuación. Más aún, esta sección está dividida en dos partes: la primera en la que presentamos ejemplos simples de operaciones elementales con vectores y matrices, y luego dos aplicaciones del producto escalar en $\mathbb{R}^2$, y la segunda que está destinada a problemas del análisis matemático como cálculo de valores de funciones polinomiales, funciones con discontinuidad evitable, derivación e integración numérica.

1.3.1. Operaciones elementales con vectores y matrices. Aplicaciones

1. Suma de vectores y producto de escalares por vectores de $\mathbb{R}^n$.

   Sean $\alpha \in \mathbb{R}$, $\overrightarrow{x} = (x_1, \ldots, x_n)$, $\overrightarrow{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. La suma de $\overrightarrow{x}$ con $\overrightarrow{y}$ se nota $\overrightarrow{x} + \overrightarrow{y}$ y se define como

   $$\overrightarrow{x} + \overrightarrow{y} = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

   El producto del escalar α con el vector $\overrightarrow{x}$ se nota $\alpha\overrightarrow{x}$ definido como

   $$\alpha\overrightarrow{x} = \alpha(x_1, \ldots, x_n) = (\alpha x_1, \ldots, \alpha x_n).$$

   Ponemos $\overrightarrow{z} = \overrightarrow{x} + \overrightarrow{y}$ y $\overrightarrow{w} = \alpha\overrightarrow{x}$. A continuación presentamos un algoritmo en el que se calcula $\overrightarrow{z} = \overrightarrow{x} + \overrightarrow{y}$ y $\overrightarrow{w} = \alpha\overrightarrow{x}$.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $\overrightarrow{x} = (x_1, \ldots, x_n)$, $\overrightarrow{y} = (y_1, \ldots, y_n)$.

   Datos de salida: $\overrightarrow{z}$, $\overrightarrow{w}$.

   1. $i = 1, \ldots, n$

      $z_i = x_i + y_i$

      $w_i = \alpha x_i$

      Fin bucle i.

   2. Imprimir $\overrightarrow{z}$, $\overrightarrow{w}$.
   3. Fin.

   Note que los datos son la talla n de los vectores $\overrightarrow{x}$ e $\overrightarrow{y}$ así como sus coordenadas. Observe que las operaciones elementales que intervienen en el cálculo de $\overrightarrow{z}$ son adiciones y en el cálculo de $\overrightarrow{w}$ son productos. Se realizan $2n$ operaciones elementales, y, el proceso de cálculo concluye en exactamente n pasos. No contabilizamos la presentación de resultados y el fin.

   La notación $i = 1, \ldots, n$ significa que para $i = 1$ se realizan los cálculos de $z_1 = x_1 + y_1$ y de $w_1 = \alpha x_1$, a continuación $k = 2$ y se realizan los cálculos $z_2 = x_2 + y_2$ y de $w_2 = \alpha x_2$. Se continua con este proceso hasta $k = n$ con lo que se hacen los cálculos $z_n = x_n + y_n$ y de $w_n = \alpha x_n$.

2. Producto escalar en $\mathbb{R}^n$.

   Sean $\overrightarrow{x} = (x_1, ..., x_n)$, $\overrightarrow{y} = (y_1, ..., y_n)$ dos vectores de $\mathbb{R}^n$. El producto escalar de $\overrightarrow{x}$ con $\overrightarrow{y}$ se nota con $\overrightarrow{x} \cdot \overrightarrow{y}$ o también $\overrightarrow{x}^T \cdot \overrightarrow{y}$ (cuando los vectores $\overrightarrow{x}$ e $\overrightarrow{y}$ se escriben como vectores columna) y se define como sigue:

   $$\overrightarrow{x}^T \cdot \overrightarrow{y} = \overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i.$$

   En el apéndice se resumen algunos resultados de los espacios vectoriales con producto interior.

   Para el cálculo de este producto escalar se requiere de la siguiente información: $n \in \mathbb{Z}^+$ y de los componentes o coordenadas de los vectores $\overrightarrow{x}$, $\overrightarrow{y}$, con lo que el producto escalar que se le denota con p puede calcularse usando el algoritmo que se propone a continuación.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $\vec{x} = (x_1, \cdots, x_n)$, $\vec{y} = (y_1, \cdots, y_n)$.

Datos de salida: p

1. $p = 0$.
2. $k = 1, \ldots, n$

   $p = p + x_k * y_k$.

   Fin bucle k.
3. Imprimir resultdo p.
4. Fin.

Para $n = 4$, $\overrightarrow{x} = (1, 0, -1, 2)$, $\overrightarrow{y} = (5, 2, -2, -3)$, la aplicación del algoritmo da como resultado $p = 1$.

Observe que las operaciones elementales que intervienen en el cálculo de p son adiciones y productos, se realizan $2n$ operaciones elementales, y, el proceso de cálculo de p concluye en exactamente n pasos.

La notación $k = 1, \ldots, n$ significa que para $k = 1$ se realiza el cálculo de $p + x_1 * y_1$ que se asigna a p, a continuación $k = 2$ y se realiza el cálculo $p + x_2 * y_2$ cuyo resultado se asigna nuevamente a p. Se continua con este proceso hasta $k = n$ con lo que se hace el cálculo $p + x_n * y_n$ que se asigna a p. La escritura $p = p + x_k * y_k$ no es una ecuación, en realidad se trata de una asignación del resultado $p + x_k * y_k$ a la variable p. Este tipo de notación será utilizada únicamente en la escritura de los algoritmos.

3. Norma euclídea en $\mathbb{R}^n$.

   Sea $\overrightarrow{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$. La norma euclídea en $\mathbb{R}^n$ se nota con $\| \cdot \|_2$ y se define como:

$$\| \overrightarrow{x} \|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

   En el apéndice se resumen algunos resultados de los espacios normados.

   Para el cálculo de $\| \overrightarrow{x} \|_2$ se requiere de la siguiente información: $n \in \mathbb{Z}^+$, las coordenadas x_i, $i = 1, \ldots, n$, del vector $\overrightarrow{x}$. El siguiente algoritmo permite calcular $\| \overrightarrow{x} \|_2$ que se le nota con N_x.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $\overrightarrow{x} = (x_1, \cdots, x_n)$.

   Datos de salida: N_x

   1. $N_x = 0$.
   2. $i = 1, \ldots, n$

      $N_x = N_x + x_i * x_i$

      Fin bucle i.
   3. $N_x = \sqrt{N_x}$.
   4. Imprimir resultado N_x.
   5. Fin.

   La escritura $N_x = \sqrt{N_x}$, en realidad significa que el cálculo de $\sqrt{N_x}$ se asigna a N_x. Esta notación se utilizará únicamente en la escritura de algoritmos.

   Sean $n = 4$, $\overrightarrow{x} = \left(3, 2, -3, -\sqrt{3}\right)$. La aplicación del algoritmo precedente da como resultado $N_x = 5$.

   Las operaciones elementales que intervienen en el cálculo de N_x son adiciones, productos y una raíz cuadrada, en un total de $2n+1$ operaciones elementales. El proceso de cálculo de N_x concluye luego de $n + 1$ pasos.

4. Suma de matrices reales de $m \times n$.

   Se nota con $M_{m\times n}[\mathbb{R}]$ el espacio vectorial de matrices de $m \times n$ con valores en $\mathbb{R}$. En algunos libros este espacio vectorial se nota como $\mathbb{R}^{mn}$. A una matriz $A \in M_{m\times n}[\mathbb{R}]$ se le nota $A = (a_{ij})_{m\times n}$ y si $m = n$, es decir A es una matriz cuadrada, se escribirá $A = (a_{ij})$.

   Sea $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{m\times n}$. La suma de las matrices A y B está definida como

$$A + B = (a_{ij})_{m\times n} + (b_{ij})_{m\times n} = (a_{ij} + b_{ij})_{m\times n} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ & \vdots & \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

   A esta matriz suma lo denotamos con $C = (c_{ij})_{m\times n}$, esto es, $C = A + B$. El algoritmo para sumar las matrices A y B se muestra a continuación.

   Algoritmo

   Datos de entrada: $m, n \in \mathbb{Z}^+$, $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{m\times n}$.

   Datos de salida: $C = (c_{ij})_{m\times n}$.

   1. $i = 1, \ldots, m$
      $j = 1, \ldots, n$
      $c_{ij} = a_{ij} + b_{ij}$
      Fin bucle j.
      Fin bucle i.
   2. Imprimr $C = (c_{ij})_{m\times n}$.
   3. Fin.

   El cálculo de la matriz C requiere de $m \times n$ adiciones. Note que el índice i es utilizado para indicar las filas, el índice j es utilizado para indicar las columnas. El algoritmo muestra que la matriz C se construye fila a fila, esto es, primera fila, a continuación segunda fila, así sucesivamente. Se deja como ejercicio elaborar un algoritmo de cálculo de C por columnas.

5. Producto de matrices.

   Sean $A = (a_{ij})_{m\times n}$, $B = (b_{jk})_{n\times p}$ matrices reales. El producto de la matriz A con B se nota AB y es la matriz $C = (c_{ik})_{m\times p}$ definida como sigue:

$$C_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk} = a_{i1}b_{1k} + \cdots + a_{in}b_{nk}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, p.$$

   Como se puede apreciar, el elemento c_{ik} es el resultado de las sumas de los productos de los elementos de la fila i de la matriz A con los correspondientes de la columna k de la matriz B. Un algoritmo para calcular $C = AB$ se muestra a continuación.

   Algoritmo

   1. $i = 1, \ldots, m$
      $k = 1, \ldots, p$
      $s = 0.$
      $j = 1, \ldots, n$
      $s = s + a_{ij} \times b_{jk}$
      Fin bucle j.
      $c_{ik} = s$
      Fin bucle k.
      Fin bucle i.

2. Imprimir $C = (c_{ik})_{m\times p}$.

3. Fin.

Este algoritmo concluye en un número finito de pasos, exactamente en $m \times p\,(n+2)$ pasos. Las operaciones elementales que se realizan son sumas y productos. Adicionalmente se hacen $2m \times p$ asignaciones. Note que la escritura $s = s + a_{ij}b_{jk}$ no es una ecuación, se trata de una asignación pués el producto $a_{ij}b_{jk}$ se suma a s y este resultado se almacena en s.

6. Intercambio de dos filas de una matriz.

Sea $A = (a_{ij})_{m\times n} \in M_{m\times n}[\mathbb{R}]$. La matriz $B = (b_{ij})_{m\times n}$ obtenida al intercambiar la fila i con la fila j con $i < j$ se define como $B = E_{i\to j}A$, donde $E_{i\to j} = (e_{pq})_{m\times m}$ se obtiene de la matriz identidad $I = (I_{li})_{m\times m}$ al intercambiar la fila i con la fila j, por lo tanto $E_{i\to j} = (e_{pq})_{m\times m}$ está definida como sigue:

$$e_{ik} = \begin{cases} 0, & \text{si } k \neq j \\ 1, & \text{si } k = j, \end{cases} \qquad e_{jk} = \begin{cases} 0, & \text{si } k \neq j \\ 1, & \text{si } k = i, \end{cases} \qquad k = 1, \ldots, m,$$

y para $p = 1, \ldots, m$ con $p \neq i, j$, $\quad e_{pk} = \begin{cases} 0, & \text{si } p \neq k, \\ 1, & \text{si } p = k \end{cases} \qquad k = 1, \ldots, m.$

Un algoritmo que realiza el producto $E_{i\to j}A$ se muestra a continuación.

Algoritmo

Datos de entrada; $m, n \in \mathbb{Z}^+$, $i, j \in \mathbb{Z}^+$, $A = (a_{ij})_{m\times n}$.

Datos de salida: $B = (b_{pr})_{m\times n}$.

1. Si $m = 1$ continuar en 5).

2. $p = 1, \ldots, m$

si $p \neq i$ y $p \neq j$

$r = 1, \ldots, n$

$b_{pr} = a_{pr}$

Fin bucle r.

Fin bucle p.

3. $r = 1, \ldots, n$

$c = a_{ir}$

$b_{ir} = a_{jr}$

$b_{jr} = c$

Fin bucle r.

4. Imprimir $B = (b_{pr})_{m\times n}$. Continuar en 6).

5. Imprimir mensaje: $m \geq 2$.

6. Fin.

La ejecución de este algoritmo implica la realización de asignaciones y de comparaciones, así el número de comparaciones es $2m+1$ y el número de asignaciones $n\,(m+1)$. Obviamente el algoritmo concluye en un número finito de pasos.

Note que se requiere de la siguiente información: talla de la matriz A, esto es, los enteros positivos m, n, los $m \times n$ coeficientes a_{ij} de A, la fila i, la fila j. Esta última información implica que $m \geq 2$. Si $m = 1$ no se realiza intercambio de filas.

7. Producto de una matriz por un vector.

Sean $A = (a_{ij})_{m\times n} \in M_{m\times n}[\mathbb{R}]$, $\quad \overrightarrow{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$. El producto $A\overrightarrow{x}$ se define como sigue:

$$A\overrightarrow{x} = \begin{bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{bmatrix}.$$

Para elaborar un algoritmo de cálculo del producto de la matriz A por el vector $\overrightarrow{x}$, esto es, $A\overrightarrow{x}$ se requiere de la siguiente información: talla de la matriz A, o sea m, $n \in \mathbb{Z}^+$, de sus componentes a_{ij}, $\ i = 1, \ldots, m,\ \ j = 1, \ldots, n$, y de los componentes o coordenadas $x_i,\ \ i = 1, \ldots, n$ del vector $\overrightarrow{x}$. Con esta información el producto $A\overrightarrow{x}$ puede calcularse con el algoritmo que se propone a continuación.

Algoritmo

Datos de entrada: $m, n \in \mathbb{Z}^+,\ A = (a_{ij})_{m\times n}, \quad \overrightarrow{x} = (x_1, \cdots, x_n)$.

Datos de salida: $\vec{z} = A\vec{x}$.

1. $c = 0$.

2. $i = 1, \ldots, m$

   $j = 1, \ldots, n$

   $c = c + a_{ij} * x_j$

   Fin bucle j.

   $z_i = c$

   $c = 0$.

   Fin bucle i.

3. Imprimir resultado $\overrightarrow{z} = (z_1, \cdots, z_m)$.

4. Fin.

Note que el algoritmo concluye luego de $m \times n$ pasos en los que intervienen productos y adiciones.

8. Vectores colineales. Angulo entre vectores y base ortogonal de $\mathbb{R}^2$.

Sean $\overrightarrow{u}_1 = (a_1, b_1)\,,\ \overrightarrow{u}_2 = (a_2, b_2)$ dos elementos no nulos de $\mathbb{R}^2$. Se considera el siguiente problema: determinar si los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ no son colineales, en tal caso calcular el ángulo que forman dichos vectores y construir una base ortogonal. Elaborar un algoritmo numérico.

Analicemos la existencia de soluciones.

Consideramos en el plano el sistema de coordenadas rectangulares y sean $\overrightarrow{u}_1 = (a_1, b_1)\,,\ \overrightarrow{u}_2 = (a_2, b_2)$ dos vectores no nulos. Se sabe que $\overrightarrow{u}_1$ y $\overrightarrow{u}_2$ son colineales si y solo si sus coordenadas satisfacen la relación $a_1b_2 - a_2b_1 = 0$. Por lo tanto, $\overrightarrow{u}_1$ y $\overrightarrow{u}_2$ no son colineales si y solo si $d = a_1b_2 - a_2b_1 \neq 0$.

El producto escalar de los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ se nota con $\overrightarrow{u}_1 \cdot \overrightarrow{u}_2$ y está definido como $\overrightarrow{u}_1 \cdot \overrightarrow{u}_2 = a_1a_2 + b_1b_2$. La longitud o norma de un vector $\overrightarrow{u} = (a, b) \in \mathbb{R}^2$ se nota $\|\overrightarrow{u}\|$ y se define como

$$\|\overrightarrow{u}\| = (\overrightarrow{u} \cdot \overrightarrow{u})^{\frac{1}{2}} = \sqrt{a^2 + b^2}.$$

Además, la medida del ángulo que forman los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ es el número real $\theta \in [0, \pi]$ definido como

$$\cos(\theta) = \frac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|\,\|\overrightarrow{u}_2\|}$$

y de esta relación

$$\theta = \operatorname{arc\,cos}\left(\frac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|\,\|\overrightarrow{u}_2\|}\right).$$

Recordemos que dos vectores $\overrightarrow{u},\ \overrightarrow{v}$ de $\mathbb{R}^2$ son ortogonales o perpendiculares si y solo si $\overrightarrow{u} \cdot \overrightarrow{v} = 0$. En tal caso escribimos $\overrightarrow{u} \perp \overrightarrow{v}$.

En la figura de la izquierda se muestran los vectores no nulos y no colineales $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ y el ángulo θ que forman dichos vectores. En la figura de la derecha se muestran los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$, la

proyección ortogonal de $\overrightarrow{u}_2$ sobre $\overrightarrow{u}_1$ y el vector $\overrightarrow{c}$ ortogonal a $\overrightarrow{u}_1$, esto es $c \perp \overrightarrow{u}_1$.

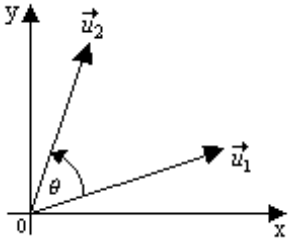


Figura 1

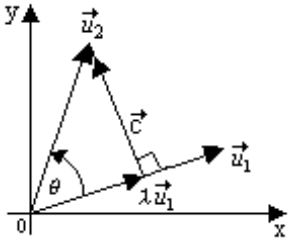


Figura 2

Ponemos $\overrightarrow{v}_1 = \overrightarrow{u}_1$. Para construir una base ortogonal $\{\overrightarrow{v}_1, \overrightarrow{v}_2\}$ consideramos las dos condiciones siguientes:

$$\text{hallar } \lambda \in \mathbb{R} \text{ y } \overrightarrow{c} \in \mathbb{R}^2 \text{ tales que } \begin{cases} \lambda\overrightarrow{u}_1 + \overrightarrow{c} = \overrightarrow{u}_2, \\ \overrightarrow{u}_1 \perp \overrightarrow{c}. \end{cases}$$

Calculemos λ. Multiplicando escalarmente por $\overrightarrow{u}_1$ la primera igualdad, se tiene

$$(\lambda\overrightarrow{u}_1 + \overrightarrow{c}) \cdot \overrightarrow{u}_1 = \overrightarrow{u}_2 \cdot \overrightarrow{u}_1$$

y como el producto escalar es distributivo respecto de la adición de vectores, resulta

$$\lambda\overrightarrow{u}_1 \cdot \overrightarrow{u}_1 + \overrightarrow{c} \cdot \overrightarrow{u}_1 = \overrightarrow{u}_2 \cdot \overrightarrow{u}_1.$$

Tomando en consideración que $\overrightarrow{u}_1 \perp \overrightarrow{c}$ que a su vez es equivalente a $\overrightarrow{u}_1 \cdot \overrightarrow{c} = 0$, se sigue que

$$\lambda\overrightarrow{u}_1 \cdot \overrightarrow{u}_1 = \overrightarrow{u}_2 \cdot \overrightarrow{u}_1.$$

Puesto que $\|\overrightarrow{u}_1\|^2 = \overrightarrow{u}_1 \cdot \overrightarrow{u}_1$ y como el producto escalar es conmutativo, esto es, $\overrightarrow{u}_2 \cdot \overrightarrow{u}_1 = \overrightarrow{u}_1 \cdot \overrightarrow{u}_2$ resulta $\lambda = \dfrac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|^2}$. El número real λ se llama coeficiente de Fourier.

Una vez calculado λ pasamos a determinar el vector $\overrightarrow{c}$. De la igualdad $\lambda\overrightarrow{u}_1 + \overrightarrow{c} = \overrightarrow{u}_2$ se obtiene $\overrightarrow{c}$:

$$\overrightarrow{c} = \overrightarrow{u}_2 - \lambda\overrightarrow{u}_1 = \overrightarrow{u}_2 - \frac{\overrightarrow{u}_1 \cdot \overrightarrow{u}_2}{\|\overrightarrow{u}_1\|^2}\overrightarrow{u}_1.$$

Definimos $\overrightarrow{v}_2 = \overrightarrow{c}$. Así $\overrightarrow{v}_1 \perp \overrightarrow{v}_2$. En la figura siguiente se muestran los vectores $\overrightarrow{v}_1$, $\overrightarrow{v}_2$ tales que $\overrightarrow{v}_1 \perp \overrightarrow{v}_2$.

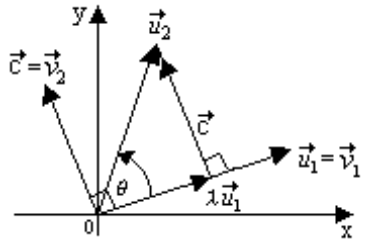


Figura 3

Con todos estos elementos estamos en condiciones de elaborar un algoritmo numérico que permita identificar si dos vectores no nulos son o no colineales. En caso de no ser colineales, calcular el ángulo que forman y obtener una base ortogonal $\{\overrightarrow{v}_1, \overrightarrow{v}_2\}$.

Algoritmo

Datos de entrada: $\overrightarrow{u}_1 = (a_1, b_1)$, $\overrightarrow{u}_2 = (a_2, b_2)$

Datos de salida: Mensaje "vectores colineales", θ, $\overrightarrow{v}_1$, $\overrightarrow{v}_2$.

1. Verificar $a_1 \neq 0$ o $b_1 \neq 0$, y, $a_2 \neq 0$ o $b_2 \neq 0$. Caso contrario $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ son nulos. Continuar en 10)

2. Calcular $d = a_1b_2 - a_2b_1$.

3. Si $d = 0$, continuar en 9).

4. Calcular $p = a_1a_2 + b_1b_2$,

$$n_1 = \left(a_1^2 + b_1^2\right)^{\frac{1}{2}},$$

$$n_2 = \left(a_2^2 + b_2^2\right)^{\frac{1}{2}},$$

$$\theta = \arccos\left(\frac{p}{n_1n_2}\right).$$

5. Poner $\overrightarrow{v}_1 = (a_1, b_1)$.

6. Calcular $\lambda = \dfrac{p}{n_1^2}$,

$$x = a_2 - \lambda a_1,$$

$$y = b_2 - \lambda b_1.$$

7. Poner $\overrightarrow{v}_2 = (x, y)$.

8. Imprimir: ángulo θ, vectores ortogonales $\overrightarrow{v}_1,\ \overrightarrow{v}_2$. Continuar en 11).

9. Imprimir: $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ vectores colineales. Continuar en 11).

10. Imprimir: $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ vectores nulos.

11. Fin.

El número total de operaciones elementales que se realizan en la ejecución de este algoritmo son 22 operaciones, comparaciones 5, asignaciones 4, una evaluación de la función arco coseno. Note que el punto 4) del algoritmo se ejecuta cuando $d \neq 0$.

Verifiquemos el algoritmo con los siguientes datos $\overrightarrow{u}_1 = (3, 1)$, $\overrightarrow{u}_2 = \left(-2, \sqrt{5}\right)$.

Claramente los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ son no nulos. Pasemos a calcular d. Tenemos $d = 3\times\sqrt{5}-(-2)\times 1 = 2 + 3\sqrt{5}$ y $d \neq 0$ con lo que se continua con el cálculo de $p,\ n_1,\ n_2$ y θ. Tenemos

$$\begin{aligned} p &= 3 \times (-2) + 1 \times \sqrt{5} = -6 + \sqrt{5}, \\ n_1 &= \left(3^2 + 1^2\right)^{\frac{1}{2}} = \sqrt{10}, \quad n_2 = \left((-2)^2 + \left(\sqrt{5}\right)^2\right)^{\frac{1}{2}} = 3, \\ \theta &= \arccos\left(\frac{-6 + \sqrt{5}}{3\sqrt{10}}\right) \simeq \arccos\left(-\frac{3{,}763932023}{9{,}48683298}\right) \simeq 1{,}978773429. \end{aligned}$$

Ponemos $\overrightarrow{v}_1 = (3, 1)$.

Calculemos el coeficiente de Fourier λ, y, x e y :

$$\begin{aligned} \lambda &= \frac{p}{n_1^2} = \frac{-6 + \sqrt{5}}{10} \simeq -0{,}3763932023, \\ x &= a_2 - \lambda a_1 \simeq -2 - (-0{,}3763932023) \times 3 = -0{,}870820393, \\ y &= b_2 - \lambda b_1 \simeq \sqrt{5} - (-0{,}3763932023) \times 1 = 2{,}61246118. \end{aligned}$$

El vector $\overrightarrow{v}_2$ está definido como $\overrightarrow{v}_2 = (-0{,}870820393, 2{,}61246118)$.

El símbolo $\simeq$ se utiliza para indicar un valor aproximado.

En la figura siguiente se muestran los vectores $\overrightarrow{u}_1,\ \overrightarrow{u}_2$ y los vectores ortogonales $\overrightarrow{v}_1,\ \overrightarrow{v}_2$.

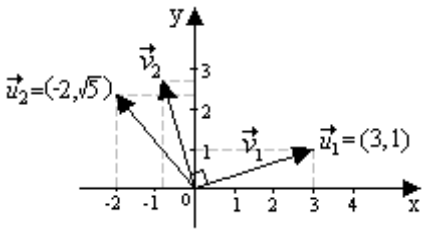


Figura 4

9. En este ejemplo se trata el método de eliminación gaussiana para sistemas de ecuaciones lineales de 3×3. Comencemos observando que los sistemas de tres ecuaciones con tres incógitas más simples de resolver son los sistemas de ecuaciones denominados diagonales, los denominados triangulares superiores y triangulares inferiores que en ese orden se presentan a continuación:

$$\begin{cases} a_1x & & = d_1 \\ & b_2y & = d_2 \\ & & c_3z = d_3 \end{cases}, \quad \begin{cases} a_1x + b_1y + c_1z = d_1 \\ b_2y + c_2z = d_2 \\ c_3z = d_3 \end{cases}, \quad \begin{cases} a_1x & = d_1 \\ a_2x + b_2y & = d_2 \\ a_3x + b_3y + c_3z = d_3, \end{cases}$$

donde $a_i,\ b_i,\ c_i \in \mathbb{R}$ para $i = 1,\ 2,\ 3$, no todos nulos, $d_i \in \mathbb{R}$ para $i = 1,\ 2,\ 3$, donde $x, y, z \in \mathbb{R}$ son las incógnitas del sistema que queremos resolver. Los números reales $a_1,\ b_2,\ c_3$ que figuran en la diagonal de cada uno de los sistemas precedentes se los denomina elementos o coeficientes de la diagonal del sistema de ecuaciones lineales.

En el caso de un sistema de ecuaciones lineales diagonal, la solución es única si y solo si los coeficientes de la diagonal del sistema son no nulos, es decir $a_1 \neq 0,\ b_2 \neq 0,\ c_3 \neq 0$ en cuyo caso la solución es $x = \dfrac{d_1}{a_1},\ y = \dfrac{d_2}{b_2},\ z = \dfrac{d_3}{c_3}$ que lo expresamos como $\left(\dfrac{d_1}{a_1},\ \dfrac{d_2}{b_2},\ \dfrac{d_3}{c_3}\right)$.

Los sistemas de ecuaciones lineales triangulares superiores e inferiores tienen una única solución si y solo si los elementos de la diagonal del sistema son no nulos, esto es, $a_1 \neq 0,\ b_2 \neq 0,\ c_3 \neq 0$.

Ejemplos

1. El sistema de ecuaciones lineales diagonal definido como $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} 2x & = 11 \\ -3y & = 0 \\ 5z = -5, \end{cases}$ tiene como solución $x = \dfrac{11}{2},\ y = -\dfrac{0}{5} = 0,\ z = -\dfrac{5}{5} = -1$, que lo escribimos $(\dfrac{11}{2}, 0, -1)$.

2. Considerar el sistema de ecuaciones lineales definido como sigue: $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} x + 2y + 3z = 11 \\ -y - 2z = 0 \\ 5z = -5. \end{cases}$ Este es un sistema de ecuaciones lineales triangular superior. Para hallar la solución de este sistema, comenzamos por la última ecuación, de la que obtenemos la incógnita $z : z = -\frac{5}{5} = -1$. De la segunda ecuación, se obtiene la incógnita $y : y = -2z = -2 \times (-1) = 2$, y de la primera ecuación, obtenemos $x : x = 11 - 2y - 3z = 11 - 2 \times 2 - 3 \times (-1) = 10$. La solución es $x = 10,\ y = 2,\ z = -1$ que escribimos $(-10,\ 2,\ -1)$.

3. Considerar el sistema de ecuaciones lineales definido por $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} 2x & = 4 \\ 3x + 4y & = 18 \\ -3x + 4y + z = 11. \end{cases}$

Este es un sistema de ecuaciones lineales triangular inferior, cuya solución encontramos resolviendo de la primera a la tercera ecuación. De la primera ecuación obtenemos $x = \frac{4}{2} = 2$. De la segunda ecuación: $y = \frac{1}{4}(18 - 3x) = \frac{1}{4}(18 - 3 \times 2) = 3$, y de la tercera ecuación: $z = 11 + 3x - 4y = 11 + 3 \times 2 - 4 \times 3 = 5$. Así, $x = 2,\ y = 3,\ z = 5$ es la solución que la escribimos $(2,\ 3,\ 5)$.

Pasemos a describir el método de eliminación gaussiana (será tratado con mayor profundidad en el capítulo 6). Para el efecto explicamos mediante tres ejemplos. La idea fundamental en el método

de eliminación gaussiana es transformar el sistema de ecuaciones lineales dado en un sistema de ecuaciones lineales triangular superior, que como hemos visto, esta clase de sistemas son los más simples de resolver.

Ejemplos

1. Resolver el sistema de ecuaciones lineales siguiente: $(x, y, z) \in \mathbb{R}^3$ tal que $\begin{cases} x + 2y + 3z = 7 \\ 2x - y - 2z = 0 \\ 3x - 2y + 5z = 25. \end{cases}$

El procedimiento de la eliminación gaussiana lo dividimos en tres etapas. Las dos primeras que conducen a transformar el sistema de ecuaciones en uno triangular superior; y, la tercera etapa que consiste en resolver el sistema de ecuaciones triangular superior.

a) Primera etapa. Mantenemos fija la primera ecuación. Se trata de eliminar la incógnita x de la segunda y tercera ecuaciones.

Eliminemos x de la segunda ecuación. Para el efecto multiplicamos por $k = -2$ (k se obtiene como el coeficiente de x de la segunda ecuación, dividido para el coeficiente de x de la primera ecuación cambiado de signo) a la primera ecuación y le sumamos el resultado a la segunda ecuación. Obtenemos $\begin{cases} x + 2y + 3z = 7 \\ -5y - 8z = -14 \\ 3x - 2y + 5z = 25. \end{cases}$ Eliminemos x de la tercera ecuación. Para ello multiplicamos por $k = -3$ (k se obtiene como el coeficiente de x de la tercera ecuación, dividido para el coeficiente de x de la primera ecuación cambiado de signo) a la primera ecuación y le sumamos el resultado a la tercera ecuación, resulta $\begin{cases} x + 2y + 3z = 7 \\ -5y - 8z = -14 \\ -8y - 4z = 4. \end{cases}$

b) Segunda etapa. Mantenemos fija la primera y segunda ecuaciones y eliminamos y en la tercera ecuación. Multipliquemos por $k_1 = -\dfrac{-8}{-5} = -\frac{8}{5}$ a la segunda ecuación y el resultado le sumamos a la tercera. k_1 se obtiene como $\begin{cases} x + 2y + 3z = 7 \\ -5y - 8z = -14 \\ \frac{44}{5}z = \frac{132}{5}. \end{cases}$

Note que k_1 se obtiene como el cociente cambiado de signo del coeficiente de y de la tercera ecuación dividido para el coeficiente de y de la segunda ecuación, siempre que este no sea nulo.

c) Tercera etapa: Resolvemos el sistema de ecuaciones triangular superior. Comenzamos con la terera ecuación, obtenemos z: $z = \frac{132}{44} = 3$. De la segunda ecuacion obtenemos y: $y = \frac{14-8z}{5} = \frac{14-8\times 3}{5} = -2$, y de la primera ecuación obtenemos $x = 7 - 2y - 3z = 7 - 2 \times (-2) - 3 \times 3 = 2$. La solución es $x = 2,\ y = -2, z = 3$, que escibimos $(2,\ -2,\ 3)$.

2. Considerar el sistema de ecuaciones lineales siguiente: $(x,\ y,\ z) \in \mathbb{R}^3$ tal que $\begin{cases} 2x + y = -2 \\ 3x + 4y + z = -2 \\ -3x + 4y + z = 4. \end{cases}$ Apliquemos el método de eliminación gaussiana. Mantengamos fija a la primera ecuación, y procedamos a la eliminación de la incógnita x en la segunda y tercera ecuaciones. Multipliquemos a la primera ecuación por $k = -\dfrac{3}{2}$ (coeficiente de x de la segunda ecuación, dividido para el coeficiente de x de la primera ecuación cambiado de signo), el resultado sumamos a la segunda. Obtenemos $\begin{cases} 2x + y = -2 \\ \frac{5}{2}y + z = 1 \\ -3x + 4y + z = 4. \end{cases}$

Sea $k_1 = -\dfrac{-3}{2} = \dfrac{3}{2}$ (k_1 se obtiene dividiendo el coeficiente de x de la tercera ecuación para el coeficiente de x de la primera ecuación, cambiado de signo). Multiplicando a la primera ecuación

por k_1, el resultado sumamos a la tercera ecuación. Tenemos $\begin{cases} 2x+y = -2 \\ \frac{5}{2}y+z=1 \\ \frac{11}{2}y+z=1. \end{cases}$ Para obtener un sistema de ecuaciones triangular superior, mantenemos fijas la primera y segunda ecuaciones del sistema precedente, eliminemos la incógnita y de la tercera ecuación.

Sea $k_2 = -\frac{\frac{11}{2}}{\frac{5}{2}} = -\frac{11}{5}$ (k_2 se obtiene dividiendo el coeficiente de y de la tercera ecuación para el coeficiente de y de la segunda ecuación, cambiado de signo). Multiplicamos a la segunda ecuación por k_2, el resultado sumamos a la tercera, resulta $\begin{cases} 2x+y = -2 \\ \frac{5}{2}y+z=1 \\ -\frac{6}{5}z=-\frac{6}{5}, \end{cases}$ con lo que hemos obtenido un sistema de ecuaciones triangular superior. Determinemos su solución. De la tercera ecuación, obtenemos $z=1$. De la segunda ecuación, se obtiene y : $y=\frac{2}{5}(1-z)=\frac{2}{5}(1-1)=0$. De la primera ecuación se deduce x: $x=\frac{1}{2}(-2-y)=\frac{1}{2}(-2-0)=-1$. La solución del sistema de ecuaciones lineales propuesto es $(-1,\ 0,\ 1)$.

3. Hallar la solución si existe, del sistema de ecuaciones lineales que se propone: $(x,y,z)\in\mathbb{R}^3$ tal que $\begin{cases} y+z=-2 \\ -2x+y-z=6 \\ 5x+y+6z=10. \end{cases}$ Para obtener (siempre que sea posible) un sistema triangular superior, la primera acción que debemos realizar es intercambiar las ecuaciones del modo siguiente: $\begin{cases} 5x+y+6z &= 10 \\ -2x+y-z &= 6 \\ y+z &= -2 \end{cases}$ Mantengamos fija la primera ecuación de este último sistema de ecuaciones lineales. Eliminemos x de la segunda ecuación. Para ello multiplicamos la primera ecuación por $k = -\frac{-2}{5} = \frac{2}{5}$ y el resultado sumamos a la segunda. Obtenemos $\begin{cases} 5x+y+6z=10 \\ \frac{7}{5}y+\frac{7}{5}z=10 \\ y+z=-2. \end{cases}$ Manteniendo fijas las dos primeras ecuaciones, eliminemos y de la tercera ecuación. Multipliquemos por $k_1 = -\frac{\frac{1}{7}}{\frac{5}{}} = -\frac{5}{7}$ a la segunda ecuación y sumemos con la tercera: $\begin{cases} 5x+y+6z=10 \\ \frac{7}{5}y+\frac{7}{5}z=10 \\ 0=-\frac{64}{7}, \end{cases}$ que muestra que la tercera igualdad es contradictoria, es decir que el sistema de ecuaciones propuesto no tiene solución.

1.3.2. Cálculo con funciones

Un polinomio P de grado $\leq n$ con coeficientes reales lo denotamos como sigue:

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k \qquad x\in\mathbb{R},$$

donde $a_k \in \mathbb{R}$ con $k = 0, 1, \cdots, n$ son los coeficientes y $a_n \neq 0$.

En orden de complejidad, los más simples son los polinomios constantes $P(x) = c \quad x \in \mathbb{R}$, con $c \in \mathbb{R}$ fijo. A continuación, los polinomios de grado 1 tienen la forma $P(x) = a + bx$, con $a, b, x \in \mathbb{R}$, a, b fijos y $b \neq 0$. Los polinomios de grado 2 se escriben como $P(x) = a + bx + cx^2$ con $a, b, c, x \in \mathbb{R}$, a, b, c fijos y $c \neq 0$. Los polinomios de grado tres se escriben como $P(x) = a + bx + cx^2 + dx^3$ con $a, b, c, d, x \in \mathbb{R}$, a, b, c, d fijos y $d \neq 0$.

Los polinomios son las funciones reales más simples de calcularse en un asignado dato $x \in \mathbb{R}$. Es claro que los más simples son los polinomios constantes y de grado 1, y realizar cálculos con esta clase de polinomios no presenta dificultad alguna. Nos interesamos en los polinomios de grado ≥ 2 que presentan alguna dificultad con los cálculos pués a medida que el grado del polinomio es más grande, el número de operaciones elementales se incrementa y los resultados pueden no ser suficientemente exactos.

1. Esquema de Hörner.

   Con frecuencia requerimos realizar evaluaciones de polinomios de modo que el número total de operaciones elementales a realizar sea el más pequeño posible y que el resultado sea el más exacto posible. Por razones que veremos más adelante y que están relacionadas con el condicionamiento, debemos evitar el cálculo directo de las potencias de x, de los factoriales, sumas y restas alternadas.

   Consideremos el polinomio $P(x) = a_0 + a_1 x + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k \quad x \in \mathbb{R}$, donde $a_k \in \mathbb{R}$ con $k = 0, 1, \cdots, n$ son los coeficientes y $a_n \neq 0$. Nos interesamos primeramente en el cálculo de $P(x)$ en un asignado $x \in \mathbb{R}$, de modo que se evite el cálculo directo de las potencias de x y el número de operaciones elementales sea el más pequeño posible. Esto se logra si se escribe $P(x)$ en la forma siguiente:

$$\begin{aligned} P(x) &= a_0 + x\left(a_1 + \cdots + a_n x^{n-1}\right) \\ &\vdots \\ &= a_0 + x(a_1 + x(a_2 + x(a_3 + \cdots + x(a_{n-1} + xa_n)\cdots))). \end{aligned}$$

   A esta forma de calcular $P(x)$ se conoce con el nombre de esquema de Hörner. Utilizando esta escritura, podemos elaborar un algoritmo para calcular $P(x)$ en un punto dado $x \in \mathbb{R}$. Note que el proceso de cálculo de $P(x)$ inicia en el término del paréntesis interior $a_{n-1} + xa_n$ y continua sucesivamente al exterior, que hace el proceso de cálculo sea muy práctico en su aplicación. El número de operaciones elementales (sumas y productos) que se requiere para calcular $P(x)$ es a lo más $2n$.

   Note que si $x \in \mathbb{R}$, el cálculo de $x^2 = x \times x$ significa una operación elemental, el cálculo de $x^3 = x^2 \times x$ significa dos operaciones elementales, entonces para el cálculo del polinomio $P(x) = a + bx + cx^2$ se requieren de 5 operaciones (sumas y productos), mientras que si se escribe en la forma $P(x) = a + x(b + cx)$ se requieren únicamente de 4 operaciones (sumas y productos). Para el cálculo del polinomio $P(x) = a + bx + cx^2 + dx^3$ se requieren de 9 operaciones y con el esquema de Hörner se requieren de 6 operaciones y mejora la exactitud del resultado.

   Para elaborar un algoritmo que permita calcular $P(x)$ requerimos de la siguiente información: grado del polinomio $n \in \mathbb{Z}^+$, coefiecientes $a_0, a_1, \ldots, a_n \in \mathbb{R}$ y del dato $x \in \mathbb{R}$. Con estos elementos proponemos el siguiente algoritmo que se conoce con el nombre de esquema de Hörner.

   Algoritmo

   Datos de entrada: $n \in \mathbb{Z}^+$, $a_0, a_1, \ldots, a_n, x \in \mathbb{R}$.

   Datos de salida: x, $P(x)$.

   1. $b = a_n$
   2. $k = 0, 1, \ldots, n-1$

      $j = n - k$

      $z = a_{j-1} + xb$

      $b = z$

Fin bucle k.

3. Imprimir x, $b = P(x)$.
4. Fin.

Note que el cálculo de $P(x)$ concluye en un número finito de pasos. Verifiquemos el algoritmo propuesto. Para el efecto, consideramos el siguiente polinomio que a su vez lo escribimos usando el esquema de Hörner:

$$P(x) = 0{,}5 + 0{,}3x - 0{,}25x^2 - 2{,}56x^3 + 3x^4 = 0{,}5 + x(0{,}3 + x(-0{,}25 + x(-2{,}56 + 3x))) \quad x \in \mathbb{R}.$$

Calculemos $P(x)$ en los puntos $x = 0$, $0{,}5$ y $-0{,}5$. Utilizando la escritura de $P(x)$, tenemos

$$\begin{aligned} P(0.) &= 0{,}5 + 0.(0{,}3 + 0.(-0{,}25 + 0.(-2{,}56 + 3 \times 0.))) = 0{,}5, \\ P(0{,}5) &= 0{,}5 + 0{,}5(0{,}3 + 0{,}5(-0{,}25 + 0{,}5(-2{,}56 + 3 \times 0{,}5))) = 0{,}455, \\ P(-0{,}5) &= 0{,}5 - 0{,}5(0{,}3 - 0{,}5(-0{,}25 - 0{,}5(-2{,}56 - 3 \times 0{,}5))) = 0{,}795. \end{aligned}$$

2. Consideremos la función E de $\mathbb{R}^+$ en $\mathbb{R}$ definida como $E(x) = \displaystyle\sum_{k=0}^{n} \frac{x^k}{(k+1)(k+2)(k+a)} \qquad x \in \mathbb{R}^+,$

donde $a \in \mathbb{R}^+$ fijo.

Dado $x \in \mathbb{R}$, para calcular $E(x)$, primeramente debemos expresar en forma explícita el sumatorio y luego escribirle en forma del esquema de Hörner como a continuación se muestra:

$$\begin{aligned} E(x) &= \frac{1}{1 \times 2 \times a} + \frac{x}{2 \times 3 \times (1+a)} + \frac{x^2}{3 \times 4 \times (2+a)} + \cdots + \frac{x^{n-1}}{n(n+1)(n-1+a)} + \\ &\quad + \frac{x^n}{(n+1)(n+2)(n+a)} \\ &= \frac{1}{2a} + x\left(\frac{1}{2 \times 3(1+a)} + x\left(\frac{1}{3 \times 4(2+a)} + \cdots + x\left(\frac{1}{n(n+1)(n-1+a)} + \right.\right.\right. \\ &\quad \left.\left.\left. + \frac{x}{(n+1)(n+2)(n+a)}\right)\cdots\right)\right). \end{aligned}$$

Note que en la última igualdad se evitan los cálculos directos de las potencias x^k, $k = 2, \ldots, n$, lo que reduce el número de operaciones elementales, facilita la escritura de un algoritmo para su cálculo.

Ponemos $a_k = \dfrac{1}{(k+1)(k+2)(k+a)}$, $k = 0, 1, 2, ..., n$. En el cálculo de a_k intervienen 3 adiciones, 3 productos y una división.

Algoritmo

Datos de entrada: n, a, x.

Datos de salida: x, $E(x)$.

1. $y = \dfrac{1}{(n+1)(n+2)(n+a)}$.
2. $k = 1, ..., n$

   $j = n - k$

   $y = \dfrac{1}{(k+1)(k+2)(k+a)} + y * x.$

   Fin bucle k.
3. Imprimir resultado $y = E(x)$.
4. Fin.

Observe que el cálculo de $E(x)$ concluye en un número finito de pasos.

3. Para cada $n \in \mathbb{Z}^+$ con $n \geq 3$ impar, se considera la función φ_n definida como sigue:

$$\varphi_n(x) = \sum_{k=0}^{n} \frac{(-1)^k x^{\frac{k}{2}}}{k! 3^k} \quad x \geq 0.$$

Se trata de calcular $\varphi_n(x)$ de modo que el número de operaciones elementales sea el más pequeño posible y elaborar un algoritmo de cálculo que permita calcular $\varphi_n(x_k) \quad k = 0, 1, \ldots, m$ en puntos x_k igualmente espaciados en el intervalo $[0, 100]$.

Sigamos la metodología utilizada para resolver problemas. Primeramente debemos constatar que se tienen soluciones. En efecto, la función φ_n está bien definida para todo $x \geq 0$. Además, de la definición de $\varphi_n(x)$ se tiene

$$\varphi_n(x) = 1 - \frac{x^{\frac{1}{2}}}{1! \times 3} + \frac{x}{2! \times 3^2} - \frac{x^{\frac{3}{2}}}{3! \times 3^3} + \cdots + \frac{(-1)^{n-1}}{(n-1)! \times 3^{n-1}} + \frac{(-1)^n}{n! \times 3^n} x^{\frac{n}{2}}.$$

Más adelante veremos que la resta de números positivos muy próximos entre sí es una operación peligrosa pués los errores de redondeo son amplificados, más aún, la realización de sumas y restas alternadas es muy peligrosa ya que los errores de redondeo provocan grandes errores en los datos de salida. Vemos que en el cálculo de $\varphi_n(x)$ debemos realizar este tipo de operaciones, además, se deben calcular los factoriales $k! \quad k = 1, 2, \ldots, n$, las potencias 3^k, $x^{\frac{k}{2}}$.

Puesto que n es impar, $n-1$ es par y en consecuencia $p = \dfrac{n-1}{2}$ es un entero positivo, asociamos todos los términos positivos y todos los términos negativos. Cada grupo contiene exactamente $p+1$ términos. Así

$$\begin{aligned}
\varphi_n(x) &= 1 + \frac{x}{2! \times 3^2} + \cdots + \frac{x^{\frac{n-1}{2}}}{(n-1)! \times 3^{n-1}} - \left[\frac{x^{\frac{1}{2}}}{1! \times 3} + \frac{x^{\frac{3}{2}}}{3! \times 3^3} + \cdots + \frac{x^{\frac{n}{2}}}{n! \times 3^n} \right] \\
&= \sum_{k=0}^{p} \frac{x^k}{(2k)! \times 3^{2k}} - \sum_{k=0}^{p} \frac{x^{k+\frac{1}{2}}}{(2k+1)! \times 3^{2k+1}} \\
&= \sum_{k=0}^{p} \frac{1}{(2k)!} \left(\frac{x}{9}\right)^k - \frac{\sqrt{x}}{3} \sum_{k=0}^{p} \frac{1}{(2k+1)!} \left(\frac{x}{9}\right)^k \quad x \geq 0.
\end{aligned}$$

Definimos

$$\theta_1(x) = \sum_{k=0}^{p} \frac{1}{(2k)!} \left(\frac{x}{9}\right)^k \quad x \geq 0, \qquad \theta_2(x) = \sum_{k=0}^{p} \frac{1}{(2k+1)!} \left(\frac{x}{9}\right)^k \quad x \geq 0.$$

En forma explícita, $\theta_1(x)$ se escribe como sigue:

$$\begin{aligned}
\theta_1(x) &= 1 + \frac{x}{2! \times 9} + \frac{x^2}{4! \times 9^2} + \cdots + \frac{x^{p-1}}{[2(p-1)]! \times 9^{p-1}} + \frac{x^p}{(2p)! \times 9^p} \\
&= 1 + \frac{1}{2}\frac{x}{9}\left(1 + \frac{1}{3 \times 4}\frac{x}{9}\left(1 + \cdots + \frac{1}{(2p-3)(2p-2)}\frac{x}{9}\left(1 + \frac{1}{(2p-1)(2p)}\frac{x}{9}\right)\cdots\right)\right).
\end{aligned}$$

Procediendo en forma similar con $\theta_2(x)$, obtenemos

$$\begin{aligned}
\theta_2(x) &= \frac{1}{1!} + \frac{1}{3!}\frac{x}{9} + \frac{1}{5!}\frac{x^2}{9^2} + \cdots + \frac{1}{(2p-1)!}\frac{x^{p-1}}{9^{p-1}} + \frac{1}{(2p+1)!}\frac{x^p}{9^p} \\
&= 1 + \frac{1}{3!}\frac{x}{9}\left(1 + \frac{1}{4 \times 5}\frac{x}{9}\left(1 + \cdots + \frac{1}{(2p-2)(2p-1)}\frac{x}{9}\left(1 + \frac{1}{(2p)(2p+1)}\frac{x}{9}\right)\cdots\right)\right).
\end{aligned}$$

Si ponemos $y = \dfrac{x}{9}$, $\theta_1(x)$ y $\theta_2(x)$ se escriben como

$$\begin{aligned}
\theta_1(x) &= 1 + \frac{y}{2}\left(1 + \frac{y}{3 \times 4}\left(1 + \cdots + \frac{y}{(2p-3)(2p-2)}\left(1 + \frac{y}{(2p-1)(2p)}\right)\cdots\right)\right), \\
\theta_2(x) &= 1 + \frac{y}{6}\left(1 + \frac{6}{4 \times 5}\left(1 + \cdots + \frac{6}{(2p-2)(2p-1)}\left(1 + \frac{6}{(2p)(2p+1)}\right)\cdots\right)\right).
\end{aligned}$$

La escritura de θ_1 y θ_2 es una variante del esquema de Hörner que evita el cálculo directo de los factoriales $k!\quad k=1,2,\ldots,n$, de las potencias x^k y 3^k. De esta manera reduce significativamente el número de operaciones elementales y permite elaborar un algritmo numérico en forma muy simple. Además,

$$\varphi_n(x)=\theta_1(x)-\frac{\sqrt{x}}{3}\theta_2(x)\qquad x\geq 0,$$

el cálculo de $\varphi_n(x)$ implica una sola resta y no sumas y restas como originalmente se tenía en el cálculo de $\varphi_n(x)$. Note también que los cocientes $\dfrac{x}{9}$ que tanto en $\theta_1(x)$ como en $\theta_2(x)$ se realizan, se evitan con el cálculo de $y=\dfrac{x}{9}$. El número de operaciones elementales que se realizan para calcular $\theta_1(x)$ y $\theta_2(x)$ son a lo más de $2\times(6p)=6(n-1)$, y el cálculo de $\varphi_n(x)$ requiere de a lo más $6(n-1)+5=6n-1$ operaciones elementales.

Si se debe calcular $\varphi_n(x)$ en la forma original se deben realizar a lo más $-1+\dfrac{n}{2}(1+3n)$ operaciones elementales.

Observe que el cálculo de $\dfrac{(\sqrt{x})^k}{k!\times 3^k}\qquad k=3,\ldots,n$ requiere de $3k-2$ operaciones elementales pués $(k-1)!k=k!$ corresponde cada una operación elemental. Análogamente $a^{k-1}a=a^k$ es una operación elemental.

Para $n=7$ se requieren 77 operaciones elementales, mientras que con la forma simplificada se requieren a lo más 41 operaciones elementales. Para $n=15$, en la forma original se requieren aproximadamente 345 operaciones elementales, mientras que en la forma simplificada requieren aproximadamente 89 operaciones elementales.

Cuando n es grande se presenta otras dificultades de cálculo de $n!$, 3^n, x^n, por ejemplo $15!\simeq 1{,}30767436\times 10^{12}$, $\quad 5^{15}\simeq 3{,}051757812\times 10^{10}$.

Finalmente, como se debe calcular $\varphi_n(x_k)$ en puntos igualmente espaciados x_k de $[0,100]$, se define $h=\dfrac{100}{m}$ y $x_k=kh\qquad k=0,1,\ldots,m.$

Con todos estos resultados se propone el siguiente algoritmo de cálculo de $\varphi_n(x_k)\quad k=0,1,\ldots,m$, y, $n\in\mathbb{Z}^+$ con $n\geq 3$ impar.

Algoritmo

Datos de entrada: $m,n\in\mathbb{Z}^+,\ x_k\quad k=0,1,\ldots,m.$

Datos de salida: x_k, $\varphi_n(x_k)\qquad k=0,1,\ldots,m.$

1. Verificar $n\geq 3$ impar. Caso contrario continuar en 6).
2. $h=\dfrac{100}{m}$.
3. Para $x=0$, poner $\varphi_n(0)=1$.
4. Para $k=1,\ldots,m$

   $b=1.$

   $c=1.$

   $x_k=kh$

   $y=\dfrac{x_k}{9}$

   $j=0,1,\ldots,p-1$

   $i=p-j,$

   $b=1+\dfrac{1}{(2i-1)(2i)}\times y\times b,$

   $c=1+\dfrac{1}{(2i)(2i+1)}\times y\times b,$

   Fin de bucle j.

$$\varphi_n(x_k) = b - \frac{\sqrt{x_k}}{3}c.$$

Fin de bucle k.

5. Imprimir $x_k,\ \varphi_n(x_k) \quad k = 0, 1, \ldots, m.$
6. Mensaje: Error de lectura de n.
7. Fin.

Para $n = 3$, la función $\varphi_3(x)$ está definida como

$$\varphi_3(x) = 1 - \frac{\sqrt{x}}{1! \times 3} + \frac{x}{2! \times 3^2} - \frac{\left(\sqrt{x}\right)^3}{3! \times 3^3} \quad x \geq 0.$$

Calculemos $\varphi_3(10)$. Tenemos

$$\begin{aligned} \varphi_3(10) &= 1 - \frac{\sqrt{10}}{1! \times 3} + \frac{10}{2! \times 3^2} - \frac{\left(\sqrt{10}\right)^3}{3! \times 3^3} \\ &= 1 - 1{,}054092553 + 0{,}555555556 - 0{,}1952023247 = 0{,}3062606775. \end{aligned}$$

Para esta cálculo se requieren de 15 operaciones elementales. Note las molestias en la realización de los cálculos en $\varphi_3(x)$. Apliquemos el algoritmo. Ponemos $y = \dfrac{10}{9} \simeq 1{,}111111111$.

$$\begin{aligned} \varphi_3(x) &= 1 + \frac{y}{2} - \frac{\sqrt{10}}{3}\left(1 + \frac{y}{6}\right), \\ \varphi_3(10) &= 1{,}555555556 - \frac{\sqrt{10}}{3} \times 1{,}185185185 = 0{,}306260678. \end{aligned}$$

Se requieren de 9 operaciones lementales.

Para $n = 7$, $\varphi_7(x)$ está definido como

$$\varphi_7(x) = 1 - \frac{\sqrt{x}}{1! \times 3} + \frac{x}{2! \times 3^2} - \frac{\left(\sqrt{x}\right)^3}{3! \times 3^3} + \frac{x^2}{4! \times 3^4} - \frac{\left(\sqrt{x}\right)^5}{5! \times 3^5} + \frac{x^3}{6! \times 3^6} - \frac{\left(\sqrt{x}\right)^7}{7! \times 3^7}$$

que se escribe como

$$\varphi_7(x) = 1 + \frac{y}{2}\left(1 + \frac{y}{12}\left(1 + \frac{y}{30}\right)\right) - \frac{\sqrt{x}}{3}\left(1 + \frac{y}{6}\left(1 + \frac{y}{20}\left(1 + \frac{y}{42}\right)\right)\right)$$

donde $y = \dfrac{x}{9}$. Para $x = 10$, $y = \dfrac{10}{9} \simeq 1{,}111111111$, y aplicando el algoritmo, obtenemos

$$\varphi_7(x) = 1{,}608901082 - 1{,}260426345 = 0{,}348474737.$$

En el siguiente capítulo se tratan las series de potencias, las mismas que se aproximan con sumas finitas, las que a su vez se escriben siguiendo un procedimiento similar al discutido en el ejemplo que acabamos de presentar.

Note que $\varphi(x) = \exp(-\dfrac{\sqrt{x}}{3}) = \sum\limits_{k=0}^{\infty} \dfrac{(-1)^k x^{\frac{k}{2}}}{k! 3^k} \quad x \geq 0$, y la función φ_n es la suma parcial del desarrollo en serie de potencias de φ. Para $x = 10$, los valores que hemos calculado $\varphi_n(10)$ son aproximaciones de $\varphi(10)$:

$$\varphi(10) = \exp(-\frac{\sqrt{10}}{3}) \simeq 0{,}3485085369.$$

4. Este es un ejemplo de una función que posee una discontinuidad evitable.

Se define la función real φ como sigue: $\varphi(x) = \dfrac{\left(1 + x^4\right)^{\frac{1}{3}} - \left(1 - x^4\right)^{\frac{1}{3}}}{x^4} \quad 0 < |x| \leq 1$. Se desea calcular $\varphi(x)$ para $x \in]0, 1[$.

Suponemos que con una calculadora de bolsillo (calculadora hipotética) se tiene $10^{-100} \simeq 0$ pero $10^{-99} \not\simeq 0$. Entonces, para $0 < |x| \leq 10^{-25}$ se tiene $0 < x^4 \leq 10^{-100} \simeq 0$ y no podemos calcular

$\varphi(x)$. Para resolver este inconveniente, aplicamos el binomio de Newton con exponente racional que se define a continuación:

$$(1+a)^r = 1 + ra + \frac{r(r-1)}{2!}a^2 + \frac{r(r-1)(r-2)}{3!}a^3 + \cdots,$$

donde $r \in \mathbb{Q}$ con $r \neq 0,\ |a| < 1$.

Apliquemos el binomio de Newton a $\left(1+x^4\right)^{\frac{1}{3}}$ y $\left(1-x^4\right)^{\frac{1}{3}}$ para $0 < |x| < 1$, tenemos

$$\begin{aligned}
\left(1+x^4\right)^{\frac{1}{3}} &= 1 + \frac{1}{3}x^4 + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)}{2!}x^8 + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)\left(\frac{1}{3}-2\right)}{3!}x^{12} \\
&\quad + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)\left(\frac{1}{3}-2\right)\left(\frac{1}{3}-3\right)}{4!}x^{16} + \frac{\frac{1}{3}\left(\frac{1}{3}-1\right)\left(\frac{1}{3}-2\right)\left(\frac{1}{3}-3\right)\left(\frac{1}{3}-4\right)}{5!}x^{20} \\
&\quad + \cdots \\
&= 1 + \frac{1}{3}x^4 - \frac{1}{9}x^8 + \frac{5}{81}x^{12} - \frac{10}{243}x^{16} + \frac{22}{729}x^{20} + \cdots, \\
\left(1-x^4\right)^{\frac{1}{3}} &= 1 - \frac{1}{3}x^4 - \frac{1}{9}x^8 - \frac{5}{81}x^{12} - \frac{10}{243}x^{16} - \frac{22}{729}x^{20} + \cdots.
\end{aligned}$$

Entonces

$$\left(1+x^4\right)^{\frac{1}{3}} - \left(1+x^4\right)^{\frac{1}{3}} = \frac{2}{3}x^4 + \frac{10}{81}x^8 + \frac{44}{729}x^{16} + \cdots,$$

de donde

$$\varphi(x) = \frac{\left(1+x^4\right)^{\frac{1}{3}} - \left(1-x^4\right)^{\frac{1}{3}}}{x^4} = \frac{2}{3} + \frac{10}{81}x^8 + \frac{44}{729}x^{16} + \cdots \quad 0 < |x| < 1.$$

En esta nueva formulación de la función φ, vemos que se ha eliminado el inconveniente de cálculo que arriba señalamos. En realidad se tiene una discontinuidad evitable en $x = 0$. Tenemos

$$\lim_{x\to 0}\varphi(x) = \frac{2}{3}, \quad \text{luego } \varphi(x) \simeq \frac{2}{3} \text{ si } 0 < |x| \leq 10^{-25}$$

Más aún, si $0 < |x| \leq 10^{-\frac{25}{2}}$, $0 < |x|^8 \leq 10^{-100} \simeq 0$, y $\varphi(x)$ se aproxima como $\varphi(x) \simeq \frac{2}{3}$.

Para x tal que $10^{-\frac{25}{2}} < |x| \leq 10^{-\frac{25}{4}}$, se tiene $10^{-200} < |x|^{16} \leq 10^{-100} \simeq 0$, luego $\varphi(x)$ se aproxima como $\varphi(x) \simeq \frac{2}{3} + \frac{10}{81}x^8$.

Si $10^{-\frac{25}{4}} < |x| \leq 10^{-\frac{100}{24}} \simeq 6{,}812920691 \times 10^{-5}$, entonces $\varphi(x)$ se aproxima como

$$\varphi(x) \simeq \frac{2}{3} + \frac{10}{81}x^8 + \frac{44}{729}x^{16} = \frac{2}{3} + x^8\left(\frac{10}{81} + \frac{44}{729}x^8\right).$$

Para x tal que $10^{-\frac{100}{24}} < |x| \leq 10^{-\frac{100}{32}}$ se tiene $10^{-\frac{400}{3}} < |x|^{32} \leq 10^{-100} \simeq 0$, en cuyo caso

$$\varphi(x) \simeq \frac{2}{3} + \frac{10}{81}x^8 + \frac{44}{729}x^{16} + \frac{718}{19683}x^{24} = \frac{2}{3} + x^8\left(\frac{10}{81} + x^8\left(\frac{44}{729} + \frac{718}{19683}x^8\right)\right).$$

Así sucesivamente.

Para x tal que $10^{-1} < |x| \leq 1$, calculamos $\varphi(x)$ con la expresión que se definió originalmente. Así,

$$\begin{aligned}
\varphi(0{,}1) &= \frac{\left(1+(0{,}1)^2\right)^{\frac{1}{3}} - \left(1-(0{,}1)^4\right)^{\frac{1}{3}}}{(0{,}1)^4} \simeq \frac{1{,}000033332 - 0{,}9999666656}{(0{,}1)^4} \simeq 0{,}6666664, \\
\varphi(0{,}2) &= \frac{\left(1+(0{,}2)^2\right)^{\frac{1}{3}} - \left(1-(0{,}2)^4\right)^{\frac{1}{3}}}{(0{,}1)^4} \simeq \frac{1{,}000533049 - 0{,}999466382}{(0{,}2)^4} \simeq 0{,}666666875.
\end{aligned}$$

Note que si se utiliza el desarrollo de $\varphi(x) = \frac{2}{3} + \frac{10}{81}x^8 + \cdots$, se obtiene $\varphi(0{,}2) \simeq 0{,}6666669827$ que es mucho más exacto que el precedente.

5. En este ejemplo se trata un método de derivación numérica.

Sea f una función real derivable en el intervalo $]a,b[\,,\ x_0 \in]a,b[\,.$ La derivada de f en x_0 se nota $f'(x_0)$ o también $\dfrac{df}{dx}(x_0)$ y se define como

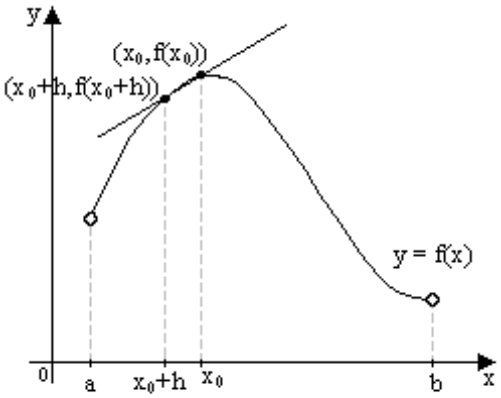
$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h).f(x_0)}{h},$$

siempre que el límite exista. El cociente $\dfrac{f(x_0+h).f(x_0)}{h} \quad h \neq 0$, se llama cociente incremental.

Admitiremos que la función f es derivable en algún intervalo abierto $]a,b[$ de $\mathbb{R}$ y nos proponemos calcular numéricamente $f'(x_0)$. Sea $h \in \mathbb{R}$ con $h \neq 0$ suficientemente pequeño. De la definición de $f'(x_0)$ surge inmediatamente la idea de aproximar $f'(x_0)$ mediante el cociente incremental, esto es,

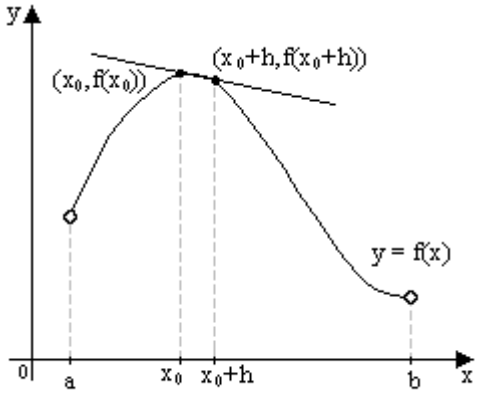
$$f'(x_0) \simeq \frac{f(x_0+h).f(x_0)}{h}.$$

En la figura siguiente se muestra la gráfica de una función f definida en $]a,b[$ y la recta secante que une los puntos $(x_0, f(x_0))$ y $(x_0+h, f(x_0+h))$ en los casos $h<0$ y $h>0$.



h<0

Figura 5



h>0

Figura 6

Ponemos $y_0 = f(x_0)\,,\ y_1 = f(x_0+h)$ y y_0' una aproximación de $f'(x_0)$ definida como

$$y_0' = \frac{y_1 - y_0}{h},$$

y_0' la denominaremos derivada numérica de $f'(x_0)$.

Supongamos que f posee derivada segunda en $]a,b[$. El polinomio de Taylor con resto de f está definido como

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(\xi)\,,$$

con $h \neq 0$ y ξ entre x_0 y x_0+h, entonces

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} - \frac{h}{2!}f''(\xi)\,.$$

La aproximación de $f'(x_0)$ se escribe como

$$y_0' = \frac{y_1 - y_0}{h} + 0(h)\,.$$

Con frecuencia $y_0 = f(x_0)\,,\ y_1 = f(x_0+h)$ no se calculan exactamente, consideraremos $y_0,\ y_1$ aproximaciones de $f(x_0)$ y $f(x_0+h)$ respectivamente.

Ejemplo

Consideremos la función real definida como $f(x) = x^4 \quad x \in \mathbb{R}$. Claramente f es derivable. Calculemos numéricamente $f'(1{,}5)$. Ponemos $x_0 = 1{,}5$. En la tabla siguiente se muestran valores de $h,\ x_0 + h,\ y_0 = f(x_0),\ y_1 = f(x_0 + h),\ f'(x_0) \simeq y_0' = \dfrac{y_1 - y_0}{h}$.

h	$x_0 + h$	y_0	y_1	$f'(x_0) \simeq y_0' = \dfrac{y_1 - y_0}{h}$.
$-0{,}1$	1,4	5,0625	3,8416	12,209
$-0{,}005$	1,495	5,0625	4,995336751	13,4326498
$-0{,}00005$	1,499995	5,0625	5,0624325	13,5
0,05	1,55	5,0625	5,77200625	14,190125
0,0005	1,5005	5,0625	5,069253376	13,506752
0,000005	1,500005	5,0625	5,0625675	13,5

El valor exacto de $f'(1{,}5)$ es 13,5.

En base a la definición de derivada numérica así como al proceso de cálculo seguido en el ejemplo, se propone como ejercicio elaborar el algoritmo correspondiente.

En un capítulo más adelante se verán otros métodos numéricos de cálculo de $f'(x_0)$ y de derivadas de orden superior. Igualmente, se tratará el cálculo aproximado de derivadas parciales.

6. En este ejemplo se considera un método de integración aproximada.

Sean $a, b \in \mathbb{R}$ con $a < b$, u una función real continua definida en $[a, b]$. Se considera el siguiente problema:

$$\text{hallar } I(u) = \int_a^b u(x)\, dx.$$

Como es conocido, la integral definida de una función continua está bien definida. Para el cálculo de $I(u)$ se consideran dos casos: el primero en el que podemos encontrar una función primitiva F de u, esto es, una función F tal que $F'(x) = u(x) \quad \forall x \in [a, b]$ y en consecuencia $I(u) = F(b) - F(a)$. En el segundo caso, no podemos encontrar una función primitiva de u, con lo que el cálculo de $I(u)$ debemos realizarlo en forma aproximada. Para el efecto, elegimos el método conocido como la regla del rectángulo que describimos a continuación.

Sean $m \in \mathbb{Z}^+$ y $\tau(m) = \{x_0 = a,\ x_1, \ldots, x_m = b\}$ una partición de $[a, b]$, esto es, $x_{i-1} < x_i \quad i = 1, \ldots, m$. Ponemos $h_i = x_i - x_{i-1} \quad i = 0, 1, \ldots, m$, y $\widehat{h}$=max$\{h_i \mid i = 1, \ldots, m\}$. Si se elige $h = \dfrac{b-a}{m}$ y $x_i = ih \quad i = 0, 1, \ldots, m$, $\tau(m)$ se dice partición uniforme. Se tiene $h_i = h$ y $\widehat{h} = h$. En general estas particiones son las más comunes.

Se define la función real v_n sobre $[a, b]$ como sigue

$$\begin{cases} v_m(x) = u(t_i) & x \in [x_{i-1}, x_i[\,, \quad i = 1, \ldots, m, \\ v_m(b) = u(b)\,, \end{cases}$$

donde $t_i = x_{i-1} + \dfrac{1}{2} h_i$ es el punto medio del intervalo $[x_{i-1}, x_i]$. La función v_m se le llama interpolante de u.

En la figura siguiente se muestra la gráfica de una función u definida en $[a, b]$, la partición $\tau(m)$

de $[a,b]$ con $m=5$ y la función interpolante v_m de u.

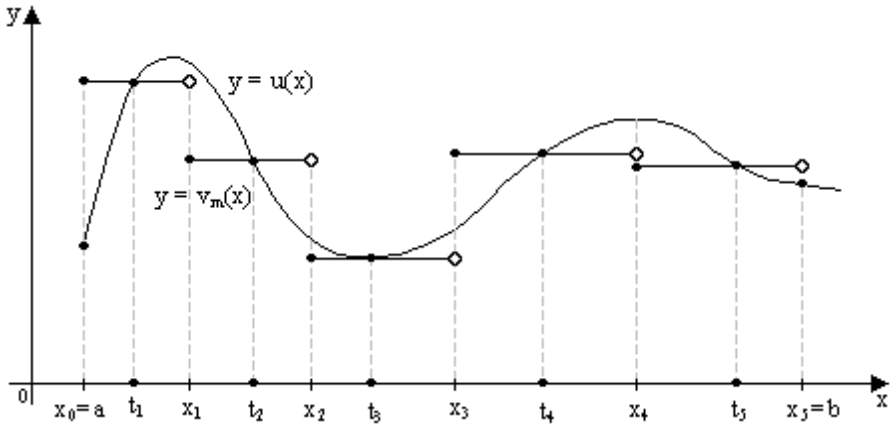


Figura 7

Entonces

$$\begin{aligned} I(v_m) &= \int_a^b v_m(x)\,dx = \sum_{i=1}^{m}\int_{x_{i-1}}^{x_i} v_m(x)\,dx = \sum_{i=1}^{m}\int_{x_{i-1}}^{x_i} u\left(x_{i-1}+\frac{1}{2}h_i\right)dx \\ &= \sum_{i=1}^{m} h_i u\left(x_{i-1}+\frac{1}{2}h_i\right). \end{aligned}$$

La aproximación $I(v_m)$ de $I(u)$ se llama regla del rectángulo. Note que el problema de cálculo de la integral $I(u)$ se le ha tansformado en uno más sencillo que es calcular $I(v_m)=\sum_{i=1}^{m} h_i u\left(x_{i-1}+\frac{1}{2}h_i\right)$

Puesto que la función u se ha discretizado según la partición $\tau(m)$ de $[a,b]$, se tiene un conjunto de puntos $u(a)$, $u(t_i)\quad i=1,\ldots,m$, $u(b)$; en el cálculo de $I(u)$ se comete un error de discretización. En análisis numérico interesa mucho estimar el error de discretización en el cálculo numérico de integrales definidas y particularmente del método de la regla del rectángulo, esto es, estimar $|I(u)-I(v_m)|$ y probar que $I(v_m) \underset{m\to\infty}{\to} I(a)$, en un cápitulo posterior se tratarán todos estos problemas.

Algoritmo

Datos de entrada: $a,b\in\mathbb{R}$, $m\in\mathbb{Z}^+$, función u.

Datos de salida: $I(v_m)$, mensaje.

1. Verificar $a<b$. Caso contrario, continuar en 8).
2. Calcular $h=\dfrac{b-a}{m}$.
3. $j=0,1,\ldots,m$

   $x_j=a+jh$

   Fin de bucle j.
4. $S=0$.
5. $j=1,\ldots,m$

   $t_j=x_{j-1}+0{,}5h$

   $S=S+u(t_j)$

   Fin de bucle j.
6. $I(v_m)=hS$.
7. Imprimir $I(v_m)$. Continuar en 9).

8. Mensaje: $a < b$.

9. Fin.

Como aplicación de la regla del rectángulo consideramos la función u definida como $u(x) = x^3 \quad x \in [1,2]$ y $m = 10$. Se considera la partición uniforme. Se define $h = \dfrac{1}{10} = 0{,}1$, la partición $\tau(10)$ del intervalo $[1,2]$ está definida como $\tau(10) = \{1, 1{,}1, 1{,}2, \ldots, 1{,}9, 1\}$. Entonces

$$\begin{aligned} I(v_{10}) &= \sum_{i=1}^{10} h_i u(t_i) = \sum_{i=1}^{10} h_i u\left(x_{i-1} + \frac{1}{2}h_i\right) = \sum_{i=1}^{10} hu(x_{i-1} + 0{,}05) \\ &= 0{,}1\left[u(1{,}05) + u(1{,}15) + \cdots + u(1{,}95)\right] = 3{,}74625. \end{aligned}$$

El valor exacto es

$$I(u) = \int_1^2 u(x)\,dx = \int_1^2 x^3 dx = \frac{1}{4}x^4\Big|_1^2 = \frac{15}{4} = 3{,}75.$$

Note que $|I(u) - I(v_{10})| = 0{,}00375$.

7. Ecuaciones diferenciales

Sean $T > 0$, f una función real definida en $[0,T] \times \mathbb{R}$. Suponemos que f es continua, más aún, se supone que f satisface la condición de Lipschitz que se indica a continuación

$$\exists M > 0 \text{ tal que } |f(t,y_1) - f(t,y_2)| \leq M|y_1 - y_2| \quad \forall y_1, y_2 \in \mathbb{R} \text{ y } t \in [0,T].$$

Se considera el problema de Cauchy de valor inicial siguiente:

$$\text{hallar } u \in C^1([0,T]) \text{ solución de } \begin{cases} u'(t) = f(t,u(t)) & t \in]0,T[, \\ u(0) = u_0. \end{cases}$$

Por la hipótesis impuesta sobre f, se sabe que dicho problema tiene solución única. En la generalidad de los casos, la función u no puede determinarse explícitamente, esta viene representada como una integral de una función que no puede integrarse con funciones elementales lo que dificulta el cálculo numérico de $u(t) \quad t \in [0,T]$. Frente a estos dos hechos, la idea es aproximar la solución de la ecuación diferencial en forma numérica.

Sean $m \in \mathbb{Z}^+$, $\tau(m) = \{t_0 = 0, t_1, \ldots, t_m = T\}$ una partición de $[0,T]$ donde $t_{j-1} < t_j \quad j = 1, \ldots, m$. Ponemos $h_j = t_{j-1} - t_j \quad j = 1, \ldots, m$ y $\widehat{h} = \max\limits_{j=1,\ldots,m} h_j$. Si se elige una partición uniforme, se tiene $t_j = jh \quad j = 0, 1, \ldots, m$ con $h = \dfrac{T}{m}$.

De la definición de $u'(t) = \lim\limits_{h \to 0} \dfrac{u(t+h) - u(t)}{h}$ se sigue que para h suficientemente pequeño y no nulo,

$$u'(t) \simeq \frac{u(t+h) - u(t)}{h} \quad t \in]0,T[,$$

y como $u'(t) = f(t,u(t))$ entonces

$$\frac{u(t+h) - u(t)}{h} \simeq f(t,u(t)) \quad t \in]0,T[,$$

luego

$$u(t+h) \simeq u(t) + hf(t,u(t)),$$

y en $t = t_j$, se obtiene

$$u(t_{j+1}) \simeq u(t_j) + hf(t_j, u(t_j)) \quad j = 0, 1, \ldots, m-1.$$

Denotamos con u_j un aproximación de $u(t_j) \quad j = 0, 1, \ldots, m$. Consideramos una partición uniforme de $[0,T]$. Se define

$$\begin{cases} u_0 \text{ dado}, \\ u_{j+1} = u_j + hf(t_j, u_j) & j = 0, 1, \ldots, m. \end{cases}$$

que se conoce como esquema numérico de Euler explícito, lo que a su vez da lugar al siguiente algoritmo.

Algoritmo

Datos de entrada: m, función $f(t,u(t))$, u_0, T.

Datos de salida: $t_j,\ u_j,\ j=0,1,\ldots,m$.

1. Poner $h=\dfrac{T}{m}$.
2. $\tau(m)=\{t_j=jh \mid j=0,1,\ldots,m\}$.
3. Para $j=0,1,\ldots,m-1$

   $u_{j+1}=u_j+hf(t_j,u_j)$

   Fin de bucle j.
4. Imprimir resultados: $t_j,\ u_j \quad j=0,1,\ldots,m$.
5. Fin.

Apliquemos el método de Euler explícito al siguiente ejemplo: $\begin{cases} u'(t)=u(t)+t & t\in\,]0,0{,}5[\,, \\ u(0)=0. \end{cases}$

Tenemos $f(t,u(t))=u(t)+t$. En este caso la solución $u(t)$ se determina mediante el conocido método de separación de variables, se obtiene $u(t)=-(t+1)+e^t \quad t\in[0,0{,}5]$.

Sean $m=5,\ h=\dfrac{0{,}5}{5}=0{,}1$ y $\tau(5)=\{0,0{,}1,\ldots,0{,}5\}$ una partición de $[0,0{,}5]$, $u_0=0$. Los resultados del método de Euler explícito se muestran a continuación.

$$\begin{aligned} u_1 &= u_0+h(u_0+t_0)=0+0{,}1(0+0)=0, \\ u_2 &= u_1+h(u_1+t_1)=0+0{,}1(0+0{,}1)=0{,}01, \\ u_3 &= u_2+h(u_2+t_2)=0{,}01+0{,}1(0{,}01+0{,}2)=0{,}031, \\ u_4 &= u_3+h(u_3+t_3)=0{,}03+0{,}1(0{,}031+0{,}3)=0{,}0641, \\ u_5 &= u_4+h(u_4+t_4)=0{,}0641+0{,}1(0{,}0641+0{,}4)=0{,}11051. \end{aligned}$$

En la figura siguiente se muestra la gráfica de la solución $u(t)$ con línea continua y la aproximada se muestra con $*$ sobre la partición de $[0,0{,}5]$.

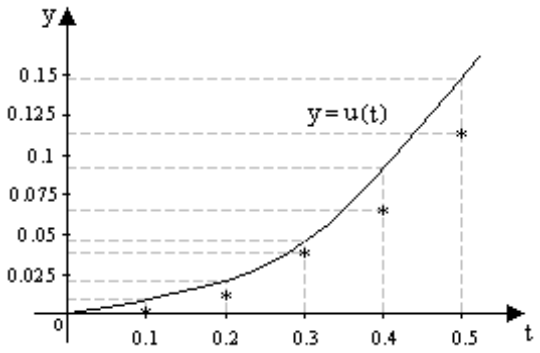


Figura 8

En la tabla siguiente se muestran los valores exactos de u, los calculados u_j sobre la partición $\tau(5)$ así como $|u(t_j)-u_j|$.

j	t_j	$u(t_j)$	u_j	$\lvert u(t_j)-u_j\rvert$
0	0	0	0	0
1	0,1	0,005170918	0	0,005170918
2	0,2	0,021402758	0,01	0,011402758
3	0,3	0,049858808	0,031	0,018858808
4	0,4	0,091824698	0,0641	0,027724698
5	0,5	0,148721271	0,11051	0,038211271

.

Note como se incrementa el error. Esta clase de problemas se abordarán en el último capítulo, donde se hará un análisis de la convergencia de los métodos y en particular de este método de Euler explícito.

1.4. Sistemas de Numeración.

Entre los sistemas de numeración más usados tenemos el sistema decimal o de base 10 cuyas cifras decimales son los enteros comprendidos entre 0 y 9. El número 10 es llamado base de dicho sistema.

Sea $M \in \mathbb{Z}^+$, para indicar su representación decimal escribimos $M = m_n m_{n-1} \dots m_o$, donde $m_i \in \{0, 1, 2, \dots, 9\} \quad \forall i = 0, 1, \dots, n$. A la representación decimal de M le asociamos el polinomio $P(x) = \sum_{k=0}^{n} m_k x^k \quad x \in \mathbb{R}$, donde los coeficientes $m_k \quad k = 0, \cdots, n$ son las cifras decimales del número entero positivo M. Entonces $M = P(10) = \sum_{k=0}^{n} m_k \times 10^k$. Así por ejemplo si $M = 2165$, el polinomio asociado a M está definido como $P(x) = 5 + 6x + x^2 + 2x^3 \quad x \in \mathbb{R}$. En $x = 10$ se tiene

$$P(10) = 5 \times 10^0 + 6 \times 10^1 + 1 \times 10^2 + 2 \times 10^3 = 2165 = M.$$

Otro de los sistemas de numeración más utilizados es el binario o de base 2, cuyas cifras binarias son los dígitos 0 y 1. En este sistema, un número entero positivo A lo representaremos como $A = (a_n a_{n-1} \dots a_0)_2$, donde $a_i \in \{0, 1\} \quad i = 0, \dots, n$. ¿Cuál es la representación de este número A en el sistema decimal? A continuación abordamos el problema de la conversión entre estos dos sistemas de numeración.

Es claro que el número entero 0 se representa como 0 en el sistema decimal y como $(0)_2$ en el sistema binario.

1.4.1. Conversión de binario a decimal y viceversa.

Conversión de binario a decimal.

Sea $A = (a_n a_{n-1} \dots a_0)_2$ un número binario. Para convertir el número A al sistema decimal le asociamos el polinomio $P(x) = \sum_{k=0}^{n} a_k x^k$, donde $a_k \in \{0, 1\} \quad k = 0, \dots, n$ son las cifras binarias del número A, y evaluamos $P(2)$ usando el esquema de Hörner. Así $A = P(2)$ en el sistema de numeración decimal.

Ejemplo

Sea $M = (1101101)_2$. Determinemos M en base 10. Para el efecto, asociamos a M el polinomio $P(x) = 1 + x^2 + x^3 + x^5 + x^6 \quad x \in \mathbb{R}$. Utilizando el esquema de Hörner, se tiene que $P(2) = 109$, por lo tanto $(1101101)_2 = 109$.

Conversión de decimal a binario.

Sea $M \in \mathbb{Z}^+$ en base 10. Supongamos que M tiene la siguiente representación en binario $M = (a_n a_{n-1} \dots a_2 a_1 a_0)_2$ cuyo polinomio asociado es $P(x) = \sum_{k=0}^{n} a_n x^k$ y evaluado en $x = 2$ se expresa como sigue:

$$P(2) = \sum_{k=0}^{n} a_n 2^k = a_0 + a_1 \times 2 + a_2 \times 2^2 + \cdots + a_n \times 2^n.$$

Sea $u \in \mathbb{Z}$. Recordemos que un número entero u se dice par si y solo si existe $j \in \mathbb{Z}$ tal que $u = 2j$, y u se dice impar si y solo si existe $j \in \mathbb{Z}$ tal que $u = 2j + 1$.

Para determinar las cifras binarias $a_0, a_1, \dots, a_n$, procedemos como sigue: $a_1 \times 2 + a_2 + 2^2 + \dots a_n \times 2^n$ es par, entonces

$$M \text{ impar} \Leftrightarrow a_0 = 1, \quad \text{y}, \quad M \text{ par} \Leftrightarrow a_0 = 0.$$

Determinada la cifra a_0, pasamos a determinar la cifra a_1. Definimos $M_1 = \frac{M-a_0}{2} = a_1 + a_2 \times 2 + \ldots + a_n \times 2^{n-1}$. Luego,

$$M_1 \text{ impar} \Leftrightarrow a_1 = 1, \quad \text{y}, \quad M_1 \text{ par} \Leftrightarrow a_1 = 0.$$

De manera análoga a la precedente, definimos $M_2 = \frac{M_1-a_1}{2} = a_2 + a_3 \times 2 + \ldots + a_n \times 2^{n-2}$, entonces

$$M_2 \text{ impar} \Leftrightarrow a_2 = 1, \quad \text{y}, \quad M_2 \text{ par} \Leftrightarrow a_2 = 0.$$

Continuando con este proceso n veces obtenemos las cifras binarias $a_k,\ k = 0, 1, \ldots, n$.

Para determinar n, observamos que si para todo k, $a_n = 1$, entonces $\sum_{k=0}^{n} 2^k = 2^{n+1} - 1 < 2^{n+1}$, y si para $k = 0, 1, \ldots, n-1, \quad a_k = 0$ y $a_n = 1$, entonces $P(2) = 2^n$. Por lo tanto $n \in \mathbb{N}$ debe verificar la desigualdad

$$2^n \leq M < 2^{n+1}.$$

Dado un número entero positivo en el sistema numérico decimal, el procedimiento descrito precedentemente permite obtener las cifras binarias de dicho número. El algoritmo de conversión de decimal a binario es el siguiente:

Algoritmo

Dato de entrada: M.

Dato de salida: $(a_n a_{n-1} \ldots a_0)_2$.

1. Determinar n tal que $2^n \leq M < 2^{n+1}$.
2. $M_0 = M$.
3. $i = 0, 1, \ldots, n-1$

   $$a_i = \begin{cases} 0, & \text{si } M_i \text{ es par,} \\ 1, & \text{si } M_i \text{ es impar.} \end{cases}$$

   $$M_{i+1} = \frac{M_i - a_i}{2}$$

   Fin de bucle i.
4. $a_n = M_n$.
5. Imprimir número binario $(a_n a_{n-1} \ldots a_0)_2$.
6. Fin.

Ejemplos

1. Sea $M = 2412$. Apliquemos el algoritmo precedente. Determinemos el número de cifras requeridas en la representación binaria. Tenemos la desigualdad $2^{11} = 2048 < M < 2^{12} = 4096$, se sigue que $n = 11$. En la siguiente tabla se ilustran los resultados de la aplicación del algoritmo de conversión de decimal a binario del número 2412.

i	0	1	2	3	4	5	6	7	8	9	10	11
M_i	2412	1206	603	301	150	75	37	18	9	4	2	1
a_i	0	0	1	1	0	1	1	0	1	0	0	1

   Por lo tanto $2412 = (100101101100)_2$.
2. Sea $M = 729$. Se tiene que $n = 9$ y la aplicación del algoritmo nos da $729 = (1011011001)_2$.

Caso fraccionario.

Consideramos ahora el caso de la conversión de decimal a binario para números racionales.

Definición 2

i. La serie $\sum_{k=1}^{\infty} a_k 2^{-k}$ *se llama fracción binaria, donde* $a_k \in \{0,1\} \quad \forall k \in \mathbb{Z}^+$.

ii. La fracción binaria se dice finita si $\exists n_0 \in \mathbb{Z}^+$ *tal que* $\forall n \geq n_0,\ a_n = 0$. *De otro modo, la fracción binaria se dice infinita.*

Observación**:** la serie $\sum_{k=1}^{\infty} a_k 2^{-k}$ es convergente, pués $a_k 2^{-k} \leq 2^{-k}, \quad \forall k \in \mathbb{Z}^+$, y

$$\sum_{k=1}^{\infty} a_k 2^{-k} \leq \sum_{k=1}^{\infty} a_k 2^{-k} \leq \sum_{k=1}^{\infty} 2^{-k} = 1.$$

Sea $S = \sum_{k=1}^{\infty} a_k 2^{-k}$ una fracción binaria. En el sistema binario escribimos $S = (0.a_1 a_2 a_3 \ldots)_2$.

Ejemplo

El número $(0{,}00111\ldots)_2$ es una fracción binaria infinita y representa a 0,25, mientras que el número $(0{,}01)_2$ es una fracción binaria finita y también representa a 0,25. Este último es consecuencia del redondeo del primero, que se tratará más adelante.

Dada una fracción binaria finita $(0.a_1 a_2 a_3 \ldots a_n)_2$, asociamos a la misma el polinomio $P(x) = \sum_{k=1}^{n} a_k x^k$ $x \in \mathbb{R}$. Para determinar el valor del número binario en el sistema decimal, calculamos el valor de $P(0{,}5)$ usando el esquema de Hörner. Por tanto $(0.a_1 a_2 \ldots a_n)_2 = P(0{,}5)$ en el sistema decimal.

Ejemplos

1. Si $(0{,}01101)_2$, entonces $P(x) = x^2(1 + x(1 + x^2))$, de donde

$$P(0{,}5) = (0{,}5)^2(1 + 0{,}5(1 + (0{,}5)^2)) = 0{,}40625.$$

2. Sea $b = (0{,}101101)_2$. El polinomio asociado al número b es $P(x) = x + x^3 + x^4 + x^6 \quad x \in \mathbb{R}$. Entonces $b = P(0{,}5) = 0{,}703125$.

Veamos el problema recíproco, es decir la conversión de una fracción decimal a binario.

Primeramente, se debe tener presente que, en general, un número real no admite una representación binaria finita. Por ejemplo, un número real que admite una representación decimal infinita seguramente su representación binaria no es finita. Además, si un número real tiene representación decimal finita no siempre admite representación binaria finita, así los números 0,1 y 0,01 no admiten representación binaria finita pero si periódica.

Sea $b \in \mathbb{R}$ tal que $0 < b < 1$. Fijado $n \in \mathbb{Z}^+$, determinemos las n primeras cifras binarias de b, esto es, determinamos la fracción binaria finita $(0.a_1 a_2 \ldots a_n)_2$ que lo notamos $b_1 = (0.a_1 a_2 \ldots a_n)_2$. Tenemos

$$\begin{aligned} b_1 &= a_1 \times 2^{-1} + a_2 \times 2^{-2} + a_3 \times 2^{-3} + \ldots + a_n \times 2^{-n} \iff \\ 2b_1 &= a_1 + a_2 \times 2^{-1} + a_3 \times 2^{-2} + \ldots + a_n \times 2^{-n+1}, \end{aligned}$$

Entonces $a_2 \times 2^{-1} + a_3 \times 2^{-2} + \ldots + a_n \times 2^{-n+1} < 1$, luego

$$a_1 = 0 \Leftrightarrow 2b < 1, \quad \text{y,} \quad a_1 = 1 \Leftrightarrow 2b \geq 1.$$

Determinada la cifra binaria a_1 pasamos a determinar la cifra binaria a_2. Se define

$$b_2 = 2(2b_1 - a_1) = a_2 + a_3 \times 2^{-2} + \ldots + a_n \times 2^{-n+2}.$$

Razonando como en la parte previa, tenemos

$$a_2 = 0 \Leftrightarrow b_2 < 1, \quad \text{y}, \quad a_2 = 1 \Leftrightarrow b_2 \geq 1.$$

En la k-ésima etapa, tenemos

$$b_k = a_k + a_{k+1}2^{-1} + \ldots + a_n 2^{-n+k},$$

de donde

$$a_k = 0 \Leftrightarrow b_k < 1, \quad \text{y}, \quad a_k = 1 \Leftrightarrow b_k \geq 1.$$

Tenemos el siguiente algoritmo de conversión de decimal a binario.

Algoritmo

Datos de entrada: b, n.

Datos de salida: $(0.a_1a_2 \ldots a_n)_2$.

1. $b_1 = b$.
2. $k = 1, \ldots, n-1$

   $$a_k = \begin{cases} 1, & \text{si } 2b_k \geq 1, \\ 0, & \text{si } 2b_k < 1. \end{cases}$$

   $$b_{k+1} = 2b_k - a_k$$

3. $a_n = \begin{cases} 1, & \text{si } 2b_n \geq 1, \\ 0, & \text{si } 2b_n < 1. \end{cases}$
4. Imprimir la fracción binaria $(0.a_1a_2 \ldots a_n)_2$.
5. Fin.

Ejemplos

1. Sean $b = \dfrac{1}{3}$. Determinemos las primeros cinco cifras binarias de b. Tenemos $n = 5$. En la tabla siguiente se ilustran los resultados de la aplicación del algoritmo precedente.Tenemos,

i	1	2	3	4	5
b_i	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
$2b_i$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$
a_i	0	1	0	1	0.

   Las cinco primeras cifras binarias de b son: $(0.\ 0\ 1\ 0\ 1\ 0)_2$. La fracción binaria finita $(0.\ 0\ 1\ 0\ 1\ 0)_2$ es una aproximación de b.

2. En la tabla siguiente se muestran los resultados de la aplicación del algoritmo de conversión de decimal a binario para $b = 0{,}1$ con 10 cifras binarias:

i	1	2	3	4	5	6	7	8	9	10
b_i	0,1	0,2	0,4	0,8	0,6	0,2	0,4	0,8	0,6	0,2
$2b_i$	0,2	0,4	0,8	1,6	1,2	0,4	0,8	1,6	1,2	0,4
a_i	0	0	0	1	1	0	0	1	1	0.

   Obtenemos $\tilde{b} = (0{,}0001100110)_2$ aproximación de b con 10 cifras binarias.

Observación. Supongamos que $b = \sum_{k=1}^{\infty} a_k 2^{-k}$, y $\widetilde{b} = (0.a_1 \dots a_n)_2$. Sea $0 < Tol < 1$ suficientemente pequeño. Determinemos n tal que $\left|b - \widetilde{b}\right| < Tol$. Sea $x_n = \sum_{k=1}^{n} a_k 2^{-k}$. Entonces

$$b - x_n = \sum_{k=n+1}^{\infty} a_k 2^{-k} \leq 2^{-n} < Tol.$$

Basta elegir n como el más pequeño número entero positivo tal que $2^{-n} < Tol$. Para tal n resulta que la fracción binaria finita $\widetilde{b} = (0.a_1 \dots a_n)_2$ en decimal es $x_n = \sum_{k=1}^{n} a_k 2^{-k}$. A esta fracción binaria finita la denominaremos aproximación de b con una precisión Tol.

Ejemplo

Sean $b = (0{,}001111\dots)_2$, $\widetilde{b} = (0{,}01)_2$, $Tol = 10^{-5}$. Entonces $\left|b - \widetilde{b}\right| = \dfrac{1}{2^n} < Tol = 10^{-5}$. Se tiene $n = 19$ y consecuentemente

$$\widetilde{b} = \sum_{k=1}^{19} a_k 2^{-k} = \frac{1}{4} - \frac{1}{2^{19}} = 0{,}25.$$

Note que $b = (0{,}001111\dots)_2 = \sum_{k=3}^{\infty} 2^{-k} = \dfrac{1}{4} = 0{,}25$, y $\widetilde{b} = (0{,}00111\dots 1)_2 \simeq (0{,}01)_2 = 0{,}25$.

El algoritmo de conversión de decimal a binario para el caso entero puede ser utilizado para determinar las n primeras cifras binarias de un número real $b \in]0, 1[$. En efecto, sea $b \in \mathbb{R}^+$ con $0 < b < 1$. Supongamos que $b = \sum_{k=1}^{\infty} a_k 2^{-k}$ una fracción binaria. Para $n \in \mathbb{Z}^+$ buscamos una aproximación binaria finita de la forma $\widetilde{b} = (0.a_1 a_2 \dots a_n)_2$. Entonces

$$\widetilde{b} = P\left(\frac{1}{2}\right) = \sum_{k=1}^{n} a_k 2^{-k} = a_1 \times 2^{-1} + \dots + a_n \times 2^{-n},$$

de donde

$$2^n \widetilde{b} = a_n + a_{n-1} \times 2 + \dots a_1 \times 2^{n-1}.$$

Puesto que $2^n b$, en general, no es un entero y como

$$\begin{aligned} 2^n b &= a_1 \times 2^{n-1} + \dots + a_{n-1} \times 2 + a_n + a_{n+1} \times 2^{-1} + a_{n+2} \times 2^{-2} + \dots \\ &= 2^n \times \tilde{b} + a_{n+1} \times 2^{-1} + a_{n+2} \times 2^{-2} + \dots, \end{aligned}$$

entonces $M = [2^n b] = 2^n \widetilde{b}$, donde $[\cdot]$ denota la función mayor entero menor o igual que x y para números reales positivos coincide con la parte entera de dicho número. Resulta que $M = (a_1 a_2 \dots a_n)_2$ cuyas cifras binarias pueden ser determinadas por aplicación del algoritmo de conversión de decimal a binario para $M \in \mathbb{Z}^+$ ya descrito anteriormente.

1.4.2. Conversión de decimal a cualquier base y viceversa

Sea $N \in \mathbb{Z}^+$ con $N > 1$. En el sistema de numeración de base N, los dígitos de dicho sistema son $0, 1, \dots, N-1$ si $N \leq 10$, y si $N > 10$ los dígitos de dicho sistema son $0, 1, \dots, 9$ y para las $N - 10$ cifras sucesivas se utilizan otros símbolos, por ejemplo las letras $A, B, C, \dots$ en ese orden.

Sistema	Base	Dígitos
Binario	2	$0, 1$
Octal	8	$0, 1, \dots, 7$
Decimal	10	$0, 1, \dots, 9$
Hexadecimal	16	$0, 1, \dots, 9, A, B, \dots, F.$

Sea $M \in \mathbb{Z}^+$, al número M lo representamos en base N de la manera siguiente: $M = (m_n m_{n-1} \dots m_0)_N$, donde $m_k \in \begin{cases} \{0,1,\dots,N-1\}, & \text{si } N \leq 10, \\ \{0,1,\dots 9,A,B,\dots\}, & \text{si } N > 10. \end{cases}$ A este número entero positivo M representado en el sistema de numeración de base N le asociamos el polinomio real $P(x) = \sum_{k=0}^{n} m_k\, x^k \quad x \in \mathbb{R}$. Entonces M en el sistema de numeración decimal se determina mediante la evaluación del polinomio $P(x)$ en $x = N$, esto es, $M = P(N)$ en base 10.

Ejemplos

1. Sea $M = (7347)_8$. El polinomio asociado a M se escribe como:

$$P(x) = 7 + 4x + 3x^2 + 7x^3 = 7 + x(4 + x(3 + 7x)) \quad x \in \mathbb{R}.$$

   Luego $P(8) = 7 + 8(4 + 8(3 + 7 \times 8)) = 3815$ en base 10. Así $(7347)_8 = 3815$.

2. Sea $M = (3AF)_{16}$. El polinomio asociado es $P(x) = F + Ax + 3x^2$ cuyo valor en $x = 16$ es

$$P(16) = F + A \times 16 + 3 \times 16^2 = 943.$$

   Tenemos $(3AF)_{16} = 943$.

Para elaborar un algoritmo de conversión de base 10 a base $\mathbb{N}$, debemos primeramente revisar el algoritmo de la división de Euclides y las clases residuales.

El algoritmo de la división de Euclides se establece en los siguientes términos: dados $a, b \in \mathbb{Z}^+$, existen $c, r \in \mathbb{N}$ tales que $0 \leq r < b$, y $a = bc + r$. El número natural c se llama cociente y r se llama residuo. Por ejemplo, si $a = 23$, $b = 5$, entonces $23 = 4 \times 5 + 3$, donde $c = 4$ y $r = 3$.

Por otro lado, la congruencia módulo m se define como a continuación se indica: dados $m \in \mathbb{Z}^+$, $a, b \in \mathbb{Z}$, se dice que a y b son congruentes módulo m que se escribe $a \equiv b \bmod(m)$ si y solo si existe $c \in \mathbb{Z}$ tal que $a - b = cm$. Es inmediato verificar que la relación de congruencia módulo m es una relación de equivalencia y que dicha relación define una partición de $\mathbb{Z}$ en clases residuales notadas como $[0]$, $[1], \cdots, [m-1]$ tales que

$$\begin{aligned} [j] &= \{cm + j \,|\, c \in \mathbb{Z}\}, \\ [i] \cap [j] &= \emptyset \text{ para } i \neq j,\ i, j = 0, \dots, m-1, \\ \mathbb{Z} &= \bigcup_{j=0}^{m-1} [j]\,. \end{aligned}$$

El procedimiento descrito para obtener las cifras binarias de un número en base 10 equivale a aplicar el algoritmo de la división de Euclides: $a = 2c + r$, donde $c \in \mathbb{N}$ y $r \in \{0, 1\}$.

Este procedimiento puede extenderse de modo similar a otras bases. En efecto, sea $M \in \mathbb{Z}^+$ expresado en el sistema decimal, $N \in \mathbb{Z}^+$ con $N > 1$ la nueva base. Por el algoritmo de la división de Euclides, se tiene que $M = cN + r$, donde $r \in \mathbb{N}$ tal que $0 \leq r < N$. Las clases residuales módulo N son: $[0], [1], \dots, [N-1]$.

El algoritmo de conversión de base 10 a base N se describe a continuación.

Algoritmo

Datos de entrada: M, N

Dato de salida: $(a_n a_{n-1} \dots a_0)_N$.

1. Determinar n tal que $N^n \leq M < N^{n+1}$.

2. $M_0 = M$.

3. $i = 0, 1, \dots, n-1$

$a_i = r \quad \text{si } M_i \in [r],$

$M_{i+1} = \dfrac{M_i - a_i}{N}$

Fin de bucle i.

4. $a_n = M_n$.

5. Imprimir $M = (a_n a_{n-1} \dots a_0)_N$.

6. Fin.

Ejemplo

1. Sean $M = 35$, $N = 3$. Se verifica inmediatamente que $3^3 \leq 35 < 3^4$, luego $n = 3$ y consecuentemente M tiene cuatro cifras en base 3. Ponemos $M = (a_3 a_2 a_1 a_0)_3$. Debemos determinar las cifras $a_0, a_1, a_2, a_3 \in \{0, 1, 2, \}$. En la siguiente tabla se muestran los resultados de la aplicación del algoritmo de conversión de decimal a base 3.

$$\begin{aligned} M_0 &= M = 35 \in [2], \ a_0 = 2, \qquad M_1 = \frac{M_0 - a_0}{3} = 11 \in [2], \quad a_1 = 2, \\ M_2 &= \frac{M_1 - a_1}{3} = 3 \in [0], \quad a_2 = 0, \qquad M_3 = \frac{M_2 - a_2}{3} = 1 = a_3. \end{aligned}$$

Luego $35 = (1022)_3$.

Relación del número de cifras entre dos sistemas de numeración

La relación del número de cifras para la representación de un número en dos bases distintas lo podemos determinar de la siguiente manera. Sean $a, b \in \mathbb{R}^+$ con $a \neq 1$, $b \neq 1$ dos bases distintas y $M \in \mathbb{Z}^+$ expresado en base 10. Supongamos que M se representa con respecto de estas dos bases como $M = (a_n a_{n-1}, \dots a_0)_a$, y $M = (b_m b_{m-1} \dots b_0)_b$. Entonces $m, n \in \mathbb{N}$ satisfacen las siguientes desigualdades:

$$\begin{aligned} a^n &\leq M < a^{n+1}, \\ b^m &\leq M < b^{m+1}, \end{aligned}$$

y en consecuencia

$$n \leq log_a(M) < n + 1,$$

$$m \leq log_b\,(M) < m + 1.$$

De la relación $\log_b(M) = \log_b(a) \times \log_a(M))$, se sigue que $m \leq \log_b(a) \times \log_a(M) < m + 1$, de donde

$$\frac{m}{\log_a\,(M)} \leq log_b(a) < \frac{m+1}{\log_a\,(M)}.$$

Tomando en consideración que $n < \log_a(M) < n + 1$, se obtiene la siguiente desigualdad

$$\frac{m}{n+1} < \frac{m}{\log_a\,(M)} \leq log_b(a) < \frac{m+1}{\log_a\,(M)} \leq \frac{m+1}{n}.$$

Para M suficientemente grande de modo que $\dfrac{1}{n}$ sea despreciable, se deduce que

$$log_b(a) \simeq \frac{m}{n} \Longleftrightarrow n log_b(a) \simeq m,$$

es decir que M en el sistema de numeración de base a requiere de aproximadamente $n log_b(a)$ cifras en el sistema de numeración de base b.

Para $b = 2$ y $a = 10$ se tiene que $\dfrac{m}{n} \simeq 3$, de donde $m \simeq 3n$. La representación binaria requiere aproximadamente cerca de 3 veces del número de cifras necesarias en la representación decimal. Para $b = 2$ y $a = 16$, se tiene $\dfrac{m}{n} = 4$.

1.5. Representación en punto flotante.

La representación en punto flotante normalizada de un número real no nulo x en el sistema decimal (comúnmente conocida como notación científica) se expresa como $x = \pm a \times 10^p$, donde $a \in [\frac{1}{10}, 1[$, $p \in \mathbb{Z}$. El número a se denomina mantisa de x y el exponente p se denomina característica de x. Este tipo de escritura de los números reales se utiliza en algunos instrumentos de cálculo como por ejemplo las calculadoras de bolsillo, los computadores portátiles. Más precisamente, un número de máquina $d \neq 0$ en una calculadora o en un computador es un número real que tiene su representación en punto flotante normalizada de la forma:

$$d = sign(d) \times a \times 10^p,$$

donde $sign(d)$ denota el signo de d, $a = 0.d_1 \cdots d_m \quad \text{con} \quad d_i \in \{0, 1, \cdots, 9\}, \quad i = 1, \cdots, m, \quad d_1 \neq 0$; y el exponente p, por ejemplo para ciertos tipos de calculadoras de bosillo, pertenece al conjunto $\{-100, -99, \cdots, 98, 99\}$.

Si $b = 0$, entonces $d_i = 0, \quad i = 1, \cdots, m$. La condición $d_1 \neq 0$ asegura que $a \geq 10^{-1}$. Además, para una aplicación numérica, el número de cifras decimales m es, en general, fijo.

Ejemplos

1. El número $x = \dfrac{1685}{3} = 561{,}6666\ldots$ se escribe en punto flotante normalizado $0{,}5616666\ldots \times 10^3$ y como número de máquina en punto flotante (por ejemplo para una calculadora de bolsillo) $0{,}5616666667 \times 10^3$.

2. El número $x = -0{,}0000000001$ como número en punto flotante normalizado (y como número de máquina) se escribe como $-0{,}1 \times 10^{-10}$.

Fijado el número de cifras decimales m, debe notarse que la mantisa más pequeña es $a = 0{,}1$ y no $a = 0{,}0 \cdots 01$ y la mantisa más grande es $a = 0{,}9 \cdots 9$.

En el sistema binario, la representación de números reales es análoga. Un número real $x \neq 0$ tiene una representación binaria normalizada que se escribe como sigue:

$$x = sign(x) \times a \times 2^p,$$

donde $a \in [\frac{1}{2}, 1[$ y $p \in \mathbf{Z}$. El número a se expresa como una serie binaria $\sum_{i=1}^{\infty} a_i 2^{-i}$ y $|p|$ se escribe como un entero en binario, esto es:

$$|p| = (p_n p_{n-1} \ldots p_0)_2 = \sum_{i=0}^{n} p_i \, 2^i.$$

Ejemplos

1. $x = \dfrac{1685}{3} = \left(\dfrac{1685}{3} \times 2^{-10}\right) 2^{10} = \dfrac{1685}{3072} \times 2^{10}.$

2. $-0{,}001 = -\dfrac{2^9}{1000} 2^{-9} = -0{,}512 \times 2^{-9}.$

Una cifra binaria se denomina bit. Si el número de bits asignados para almacenar el número es fijo, un número de máquina tiene una forma de punto flotante binaria normalizada $b = sign(b) \times a \times 2^p$, que puede almacenar exactamente usando los siguientes grupos de bits: $\begin{cases} \text{un bit para el signo de } b, \\ r \text{ bits para el exponente } |p|, \\ m \text{ bits para la mantisa } a. \end{cases}$

Es decir que un número de máquina b en punto flotante binario normalizado requiere para su representación de $m+r+1$ bits, a condición de que dicho número de bits sea fijo. Además,

$$a = (0.\, b_1 \dots b_m)_2, \qquad b_i \in \{0,1\}, \quad i = 1, \dots, m, \quad b_1 \neq 0,$$

$$|p| = (p_{r-1} \dots p_0)_2, \qquad p_j \in \{0,1\}, \quad j = 0, 1, \dots, r-1.$$

La mantisa más pequeña es $a = (0{,}10\dots0)_2 = \frac{1}{2}$ y la más grande es $a = (0{,}11\dots1)_2 = 1 - 2^{-m}$. Para el exponente se tiene que

$$0 \leq |p| \leq 2^r - 1.$$

En consecuencia, el número más grande a representarse en punto flotante binario normalizado es

$$(1 - 2^{-m}) \times 2^{2^r - 1}$$

y el número positivo más pequeño es 2^{-2^r}. Este último se obtiene del modo siguiente:

$$b = (0.b_1 \dots b_m)_2 \times 2^{-q} \geq (0{,}10\dots0) = \frac{1}{2} \times 2^{-q} = 2^{-q-1},$$

donde $q \geq 0$. Como $0 \leq q \leq 2^r - 1$, entonces $-q \geq -2^r + 1$, y en consecuencia $b \geq 2^{-q-1} = 2^{-2^r}$.

Así, un número de máquina en punto flotante binario normalizado satisface la desigualdad:

$$2^{-2^r} \leq |b| \leq \left(1 - 2^{-m}\right) 2^{2^r - 1}.$$

Además, el conjunto de números de máquina es finito.

Por ejemplo,si se tiene $m = 24$, $r = 7$. Entonces $0 \leq p \leq 127$ y $2^{2^r-1} = 2^{127} \simeq 10^{38}$, de modo que

$$10^{-38} \simeq 2^{-2} \leq |b| \leq 2^{2^7 - 1} \simeq 10^{38}.$$

En la actualidad se tiene los siguientes tipos de representación de números reales: simple precisión, doble precisión y doble precisión extendida. El estudiante debe conocer cuantos bits se requieren para la mantisa y cuantos para el exponente. Por ejemplo para ciertos tipos de máquinas se tiene 1 bit para el signo, para doble precisión se tienen 23 bits para la mantisa y 8 bits para el exponente, para doble precisión extendida se tienen 47 bits para la mantisa y 16 bits para el exponente. Si se dispone de 64 bits para representar un número en doble precisión extendida, estos se dispone de la manera siguiente: 47 bits para la mantisa, 1 bit para el signo y 16 bits para el exponente.

Observación.

1. En una calculadora de bolsillo, el número de bits para representar la mantisa así como su equivalente en base 10 es, generalmente fijo. En un computador se tiene alguna flexibilidad para almacenar mantisas de diferentes tallas. Además, la aritmética que utilizan puede ser binaria (base 2), octal (base 8), hexadecimal (base 16).

2. En las calculadoras se almacena el exponente p con un corrimiento de 100, de tal manera que el exponente corrido $E + 100$ es un entero no negativo entre 0 y 199. De este modo se evita utilizar un bit para el signo del exponente. Así por ejemplo:

   $$x = \frac{1685}{3} = 0{,}5616666667 \times 10^3, \quad p + 100 = 103.$$

   En binario, el exponente corrido de r bits tiene una representación de la forma

   $$0 \leq p + p_0 = (b_{r-1} \dots b_0)_2 = \sum_{i=0}^{r-1} b_i 2^i \leq 2^r - 1.$$

   El corrimiento p_0 se le toma como 2^{r-1}, en cuyo caso

   $$-2^{r-1} \leq p \leq 2^r - 1 - p_0 = 2^r - 1 - 2^{r-1} = 2^{r-1} - 1.$$

Los enteros m (para la mantisa), p_0 y r son fijos.

Por ejemplo, el grupo de bits: $-1\,|\;\;101100110\ldots0\;|\;\;0000110$, donde $m = 24,\ r = 7,$ indica que el número es negativo (primer bit -1), la mantisa $a = (0{,}101100110\ldots0)_2 = 0{,}69922$, $p + p_0 = (0000110)_2 = 6$. Puesto que $p_0 = 2^{r-1} = 2^{7-1} = 64$, entonces $p = 6 - p_0 = -58$, $x = -0{,}69922 \times 2^{-58}$. Note que para $r = 7,\ -64 \leq p \leq 63$.

Para el grupo de bits $0\,|\;\;101100110\ldots0\;|\;\;1000010$ se tiene: mantisa $a = (0{,}101100110\ldots0)_2 = 0{,}69922$, exponente con corrimiento: $p + p_0 = (1000010)_2 = 66,\ \ p_0 = 64$ entonces $p = 66 - p_0 = 2$. Luego $x = 0{,}69922 \times 2^2 = 2{,}57688$.

1.6. Tipos de Errores

En la sección 3 hemos visto algunos ejemplos de cálculo de soluciones numéricas de problemas relativamente sencillos que surgen en los ámbitos del álgebra lineal, el análisis matemático, las ecuaciones diferenciales en los que se evidencian los resultados aproximados obtenidos en instrumentos de cálculo como son las calculadoras de bolsillo y los computadores. Con cualquiera de estos instrumentos, los resultados mostrados están sujetos a errores.

En el análisis numérico, uno de los problemas fundamentales es el estudio o análisis del error cometido en cada uno de los métodos de aproximación que se proponen. Interesa establecer la exactitud y precisión en el cálculo de la solución de un problema (P), la minimización de los errores cometidos en el cálculo de la solución aproximada de (P). Se distinguen varios tipos de errores que limitan la exactitud. Estos pueden clasificarse en tres grupos: errores en los datos de entrada o errores inherentes; errores de redondeo y errores de aproximación.

1. **Errores en los datos de entrada o errores inherentes.** Se deben a esquematizaciones hechas para la reducción de términos matemáticos de cierto modelo. Pueden deberse también a errores debidos en las medidas experimentales de una magnitud física o a observaciones de cualquier otra índole (de tipo económico, social, etc.). Pueden tener también su origen como resultados de un cálculo realizado previamente. Nótese que estos errores aparecen antes de iniciar el cálculo de un cierto problema (P). En el estudio que nosotros haremos no nos ocuparemos de este tipo de errores.

2. **Errores de redondeo.** Estos errores son debidos a la necesidad de trabajar con números de máquina. Dependen casi exclusivamente del instrumento de cálculo a disposición. La evaluación rigurosa es, a menudo, muy complicada. Para el cálculo de la solución de ciertos problemas que consideraremos más adelante, los errores de redondeo tienen una influencia enorme que puede arruinar los resultados. Por tanto es de mucha importancia el poder controlarlos.

   A continuación presentamos tres números reales y sus aproximaciones con 8, 16, 24, 32 cifras luego del punto decimal, las mismas que han sido obtenidas con el programa de Matemática. El símbolo $\simeq$ se utilizará en lo sucesivo para indicar aproximación. Tenemos,

$$\begin{aligned}
\frac{805}{111} &\simeq 7{,}25225225, \qquad \frac{805}{111} \simeq 7{,}252252252522522525,\\
\frac{805}{111} &\simeq 7{,}25225225225225225225225,\\
\frac{805}{111} &\simeq 7{,}25225225225225225225225225225225,
\end{aligned}$$

$$\begin{aligned}
\pi &\simeq 3{,}14159265, \qquad \pi \simeq 3{,}1415926535897932,\\
\pi &\simeq 3{,}14159265358979323846,\\
\pi &\simeq 3{,}14159265358979323846264338327950,
\end{aligned}$$

$$\begin{aligned}\sqrt{2} &\simeq 1{,}41421356, \qquad \sqrt{2} \simeq 1{,}4142135623730950,\\ \sqrt{2} &\simeq 1{,}41421356237309504880168872420970,\\ \sqrt{2} &\simeq 1{,}41421356237309504880168872420970.\end{aligned}$$

Note que el número $\dfrac{805}{111}$ es un número racional, su representación decimal es infinita y periódica. Basta conocer el primer período de su representación decimal y con esta puede escribirse el número con el número de cifras que se desee.

Al representar los números reales como aproximaciones con un número determinado de cifras decimales, se comete un error de redondeo que trataremos más adelante.

3. **Errores de aproximación.** Este tipo de errores se dividen en dos grupos: los errores de truncamiento y los errores de discretización.

a) Errores de truncamiento. Consideremos los siguientes ejemplos.

1. Sean (a_n) una sucesión numérica real y $\sum_{n=0}^{\infty} a_n$ una serie que suponemos convergente. Denotamos con S su suma, esto es, $S = \sum_{n=0}^{\infty} a_n$. Con frecuencia, el cálculo exacto de S es muy difícil de obtener, por lo que se recurre al cálculo aproximado. La idea es aproximar la suma S a través de un número finito m de términos, digamos $S_m = \sum_{n=0}^{m} a_n$. Este procedimiento produce un error denominado de truncamiento. La determinación del número de términos necesarios para la aproximación de la solución S es importante, pués evita la ejecución de cálculos que no mejoran la precisión de la solución, y, disminuyen los costos numéricos.

2. Sean (f_n) una sucesión de funciones definidas en un intervalo $[a,b]$ de $\mathbb{R}$ y $\sum_{n=0}^{\infty} f_n$ una serie de funciones que suponemos converge uniformemente en el intervalo $[a,b]$. Se define la función f como sigue: $f(x) = \sum_{n=0}^{\infty} f_n(x) \qquad x \in [a,b]$. Nos interesamos en trazar la gráfica de la función f. En la generalidad de situaciones resulta complicado, y en ocasiones imposible, el cálculo de cada $f(x)$. Una forma de resolver este problema es aproximar cada $f(x)$ con una suma finita de términos de la serie $\sum_{n=0}^{\infty} f_n(x)$, la misma que se elige apropiadamente en función de la precisión que deseamos obtener. Este proceso de aproximación provoca un error denominado de truncamiento. Por otro lado, dado el volumen de cálculo a realizar es conveniente elaborar un algoritmo numérico para calcular cada $f(x)$. Cuando la serie converge rápidamente este es el camino a seguir. Lastimosamente, en ocasiones, el solo hecho de limitar a un número finito de términos no basta sobre todo en el caso de series que convergen lentamente, esto conduce a proponer otro tipo de problema que consiste en la búsqueda de un método para acelerar la convergencia de la serie, una vez logrado esto, se pasa a calcular los valores aproximados de $f(x)$. En un capítulo posterior se estudian esta clase de problemas.

b) Errores de discretización. Consideremos los siguientes ejemplos.

1) Sea v una función real continua en el intervalo cerrado $[a,b]$. Queremos calcular $v(x)$ con $x \in [a,b]$, lastimosamente la función v no es conocida en todo el intervalo $[a,b]$ sino en un conjunto finito de $m+1$ puntos de una partición $\tau(m) = \{x_0 = a,\ x_1, \cdots,\ x_m = b\}$ de $[a,b]$, donde $x_{i-1} < x_i$ $i = 1, \cdots, m$, digamos $S = \{(x_i, v(x_i)) \in \mathbb{R}^2 \mid i = 0, 1, \cdots, m\}$. Este problema (P) se presenta con mucha frecuencia y se le conoce como problema de interpolación. La idea es aproximar $v(x)$ mediante $v_h(x)$ de una función v_h definida en $[a,b]$ que sea mucho más simple de calcular de modo que $v_h(x_i) = v(x_i) \quad i = 0, 1, \cdots, m$, esto conduce a construir la función v_h como $v_h = \sum_{i=0}^{i=m} v(x_i)\varphi_i$, donde $\{\varphi_0, \cdots, \varphi_m\}$ es un conjunto de funciones que se construyen apropiadamente. La función v_h así definida se conoce con el nombre de función interpolante de v. Este proceso produce un error denominado de discretización. En un capítulo posterior se estudia este tipo de problemas.

En la figura siguiente se muestra la gráfica de una función continua v definida en el intervalo $[0,a]$ con $a > 0$, y la de una función interpolante v_h (segmentos de recta) de v. Se muestran también los

puntos de la partición de $\tau(m)$ de $[0,a]$.

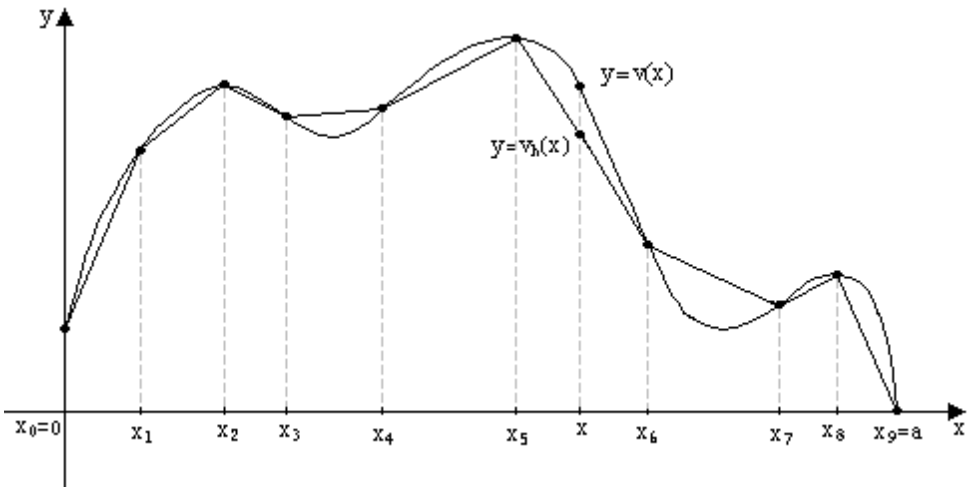


Figura 9

2) Sea $L > 0$. Se considera el siguiente problema:

$$\text{hallar una función } u \in C^2([0,L]) \text{ solución de } \begin{cases} -u''(x) + u(x) = f(x) & \text{si } x \in]0,L[, \\ u(0) = 0, \quad u(L) = 0, \end{cases}$$

donde $f \in C(\ [0,L])$, con $C^k([0,L])$ el espacio de funciones que poseen derivada continua en el intervalo $[0,L]$, $k = 0, 1, \cdots$, y se pone $C([0,L]) = C^0([0,L])$. Este problema es parte de la familia de los denominados problemas unidimensionales de valores en la frontera. Con la hipótesis sobre f, se demuestra que este problema tiene solución única $u \in C^2([0,L])$. Para ciertos tipos de funciones f se puede encontrar soluciones explícitas que no se representan como integrales, para otros tipos de funciones f las soluciones se expresan como integrales de las que no pueden calcularse sus primitivas o que resultan difíciles de calcularse. Por otro lado, se tiene interés en calcular numéricamente la solución $u(x) \quad x \in [0,L]$. Todos estos argumentos nos conducen a resolver el problema de valores en la frontera en forma numérica, es decir, proceder a discretizar dicho problema. Para el efecto, sea $m \in \mathbb{Z}^+$, $\tau(m) = \{x_0 = 0, x_1, \cdots, x_m = L\}$ una partición de $[0,L]$, donde $x_{j-1} < x_j \quad j = 1, \cdots, m$. Ponemos $h_j = x_j - x_{j-1} \quad j = 1, \cdots, m, \quad h = \text{máx}\{h_j \mid i = 1, \cdots, m\}$. En el caso de una partición uniforme, se define $h = \dfrac{L}{m}$ y $x_j = jh \quad j = 0, 1, \cdots, m$. Consideremos una partición uniforme. Se denota con u_j a una aproximación de $u(x_j) \quad j = 0, 1, \cdots, m$. En el capítulo 2 mostraremos que $u''(x)$ se aproxima mediante el cociente denominado diferencias finitas centrales de segundo orden que se indica a continuación:

$$u''(x_j) \simeq \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} \qquad j = 1, \cdots, m-1.$$

Con esta aproximación, el problema propuesto de valores en la frontera se aproxima como

$$\begin{cases} -\dfrac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + u(x_j) \simeq f(x_j) \qquad j = 1, \cdots, m-1, \\ u(x_0) = 0, \qquad u(x_m) = 0, \end{cases}$$

por lo que el problema discreto es el siguiente:

$$\text{hallar } \overrightarrow{u} \;=\; (u_1, \cdots, u_{m-1}) \in \mathbb{R}^{m-1} \quad \text{solución del sistema de ecuaciones lineales}$$
$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + u_j = f(x_j) \qquad j = 1, \cdots, m-1, \\ u_0 = 0, \qquad u_m = 0. \end{cases}$$

Este proceso de discretización del problema de valores en la frontera, produce un error denominado error de discretización.

La estimación de los errores de discretización es fundamental en el Análisis Numérico.

Exacto, inexacto, precisión, imprecisión.

En el lenguaje corriente, los términos exactitud y precisión se usan indistintamente como sinónimos. En el contexto del análisis numérico es importante establecer la diferencia que existe entre estos dos téminos.

El témino exactitud se refiere a que tan cercano está un valor calculado o medido con el verdadero valor; mientras que el término inexacto se define como una desviación del verdadero valor. También se considerará como exacto aquel resultado o método riguroso, conforme a la lógica.

El término precisión se refiere a que tan cercano está un valor individual calculado o medido con cualquier otro; mientras que el término imprecisión se refiere a una magnitud que se aleja una de otra. Se comprenderá como precisión aquello que no deja incertidumbre, determinado rigurosamente.

Por ejemplo, con 9 cifras luego del punto decimal, $\sqrt{3}$ se calcula como 1,732050808. En este caso hablamos de una exactitud de 10^{-9}. Cuando $\sqrt{3}$ se aproxima como 1,7320, 1,7320508, 1,732050808 hablamos de una precisión de 4, 7, 9 cifras luego del punto decimal.

Si $I = \int_0^2 \sqrt{x}dx = \frac{2}{3}(\sqrt{2})^3 \simeq 1{,}88562$, hablamos en este caso del cálculo de I con una exactitud de 10^{-5}, y aplicando la regla del rectángulo siguiente: $I_5 = 0{,}4(\sqrt{0{,}2} + \sqrt{0{,}6} + \sqrt{1{,}0} + \sqrt{1{,}4} + \sqrt{1{,}8})$, resulta $I_5 \simeq 1{,}898667$, hablamos en este caso de un cálculo de I_5 con una precisión de 10^{-6}, y un cálculo aproximado de I con una precisión de 10^{-2}.

Con error de cálculo entenderemos tanto la inexactitud como la imprecisión. Tenemos

$$\text{Verdadero valor} = \text{Valor aproximado} + \text{error},$$

donde el error puede deberse a los errores de redondeo, de aproximación (truncamiento, discretización) o ambos. De manera general, al verdadero valor lo conoceremos como solución exacta y al valor aproximado lo denominaremos solución numérica o también solución aproximada.

De estas observaciones tenemos que los métodos numéricos deben ser suficientemente exactos y precisos.

1.7. Errores de redondeo

Hemos visto que los números de máquina en punto flotante satisfacen la desigualdad:

$$2^{-2^r} \leq |b| \leq \left(1 - \frac{1}{2^m}\right) \times 2^{2^r - 1},$$

donde r, m son enteros positivos. Además, todo número de máquina se escribe en la forma

$$b = (0.a_1 ... a_m)_2 \times 2^{(a_r a_{r-1} ... a_0)_2}.$$

Sea A el conjunto de tales números. Se sigue que el conjunto de números de máquina es finito. El problema que se presenta es el siguiente: ¿cómo aproximar un número $x \notin A$ por un número $y \in A$?

Consideremos lo tres casos siguientes:

1. $x > \left(1 - \frac{1}{2^m}\right) \times 2^{2^r - 1}$.
2. $2^{-2^r} \leq x \leq \left(1 - \frac{1}{2^m}\right) \times 2^{2^r - 1}$.
3. $0 < x < 2^{-2^r}$.

Comenzaremos con el caso 2). Sea $M = \left[2^{-2^r}, \left(1 - \frac{1}{2^m}\right) 2^{2^r - 1}\right] \subset \mathbb{R}$. Se tiene que $A \subset M$.

La métrica usual d en $\mathbb{R}$ está definida como $d(x, y) = |x - y| \quad \forall x, y \in \mathbb{R}$.

Sea $x \in M$, como A es un conjunto cerrado, $\exists \widetilde{x} \in A$ tal que $d(x, A) = d(x, \widetilde{x}) = |x - \widetilde{x}|$, donde la distancia del punto x al conjunto A se define como $d(x, A) = \min\limits_{y \in A} |x - y|$. Resulta que

$$|x - \widetilde{x}| = d(x, A) \leq |x - y| \quad \forall y \in A.$$

Por tanto la aproximación de cualquier número $x \in (M \setminus A)$ por un número notado como $rd(x) \in A$ debe satisfacer la siguiente condición:

$$|x - rd(x)| \leq |x - y| \quad \forall y \in A.$$

El número $rd(x)$ aproximación de x se lo obtiene por redondeo y se denomina redondeado de x.

Ejemplo

Supongamos que nuestro conjunto A está constituido por números reales de la forma $0.a_1a_2a_3a_4 \times 10^p$, donde $a_i \in \{0, 1, \cdots, 9\}$, $p \in \mathbb{Z}$ y $a_1 \neq 0$. Note que la mantisa de los elementos del conjunto A únicamente tienen 4 dígitos, que escribiremos $t = 4$. Entonces $rd(0{,}14285 \times 10^0) = 0{,}1429 \times 10^0$, pués

$$| \, 0{,}14285 \times 10^0 - 0{,}1429 \times 10^0 \, | = 0{,}5000 \times 10^{-4} \leq | \, 0{,}14285 \times 10^0 - y \, | \quad \forall y \in A.$$

De manera similar se obtienen los siguientes resultados

$$\begin{aligned} rd(0{,}8423 \times 10^0) &= 0{,}8423 \times 10^0, \\ rd(3{,}14159 \times 10^0) &= 0{,}3142 \times 10^1, \\ rd(0{,}142842 \times 10^2) &= 0{,}1428 \times 10^2. \end{aligned}$$

En general, para encontrar $rd(x)$ con t dígitos, se procede del modo siguiente: el número $|x| \in (M \setminus A)$ es representado en forma normalizada: $|x| = a \times 10^b$, de modo que $\frac{1}{10} \leq |a| \leq 1$. Sea $a = 0.\alpha_1\alpha_2 \ldots \alpha_t\alpha_{t+1} \ldots$ la representación decimal de a, donde $0 \leq \alpha_i \leq 9 \quad \forall_i = 1, 2, \ldots, \alpha_1 \neq 0$. Definimos

$$\tilde{a} = \begin{cases} 0.\alpha_1\alpha_2 \ldots \alpha_t, & \text{si } 0 \leq \alpha_{t+1} \leq 4, \\ 0, \alpha_1\alpha_2 \ldots \alpha_t + 10^{-t}, & \text{si } 5 \leq \alpha_{t+1} \leq 9. \end{cases}$$

Como $1 \leq \alpha_1 \leq 9$ entonces $|a| \geq 0.\alpha_1 \geq \frac{1}{10} = 0{,}1$. Se pone $sign(x) = \begin{cases} -1, & \text{si } x < 0, \\ 1, & \text{si } x > 0, \end{cases}$ y $rd(x) = sign(x) \times \tilde{a} \times 10^b$.

Definición 3 *El error de redondeo de x se define como $x - rd(x)$. El número no negativo $|x - rd(x)|$ se llama error absoluto.*
El error relativo de x se define mediante la relación:

$$\varepsilon_x = \frac{rd(x) - x}{x} \quad x \neq 0.$$

En algunos textos, el error de redondeo se le denomina error inherente.

Se tiene la siguiente mayoración del error relativo ε_x:

$$\begin{aligned} |\varepsilon_x| &= \left|\frac{rd(x) - x}{x}\right| = \frac{\left|(sign(x)\tilde{a}10^b - sign(x)a{,}10^b\right|}{|sign(x)a \times 10^b|} = \frac{|\tilde{a} - a|}{|a|} \\ &\leq \frac{5 \times 10^{-(t+1)}}{|a|} \leq 5 \times 10^{-(t+1)+1} = 5 \times 10^{-t} = eps. \end{aligned}$$

Luego $|\varepsilon_x| \leq eps = 5 \times 10^{-t}$.

Definición 4 *El número* $eps = 5 \times 10^{-t}$ *se llama precisión de máquina .*

Se tiene que si

$$\frac{rd(x) - x}{x} = \varepsilon_x \Rightarrow rd(x) = x(1 + \varepsilon_x), \text{ con } |\varepsilon_x| \leq eps.$$

El número $rd(x) \in A$ tiene la propiedad:

$$|x - rd(x)| \leq |x - y| \quad \forall y \in A.$$

En el sistema binario, $rd(x)$ está definido de modo análogo. Comenzamos con la escritura de x en la forma. $|x| = a \times 2^b$, donde $a = 0.\alpha_1\alpha_2 \ldots \alpha_t\alpha_{t+1} \ldots, \alpha_i \in \{0, 1\}, \quad i = 1, 2, \ldots,$ y $\alpha_1 = 1$. Se tiene $1 > a \geq \frac{1}{2}$. Se define

$$\tilde{a} = \begin{cases} 0.\alpha_1 \ldots \alpha_t, & \text{si } \alpha_{t+1} = 0, \\ 0.\alpha_1 \ldots \alpha_t + 2^{-t}, & \text{si } \alpha_{t+1} = 1. \end{cases}$$

Entonces $rd(x) = sign(x) \times \tilde{a} \times 2^b$, y $\quad |\varepsilon_x| \leq eps = 2^{-t}$. Resulta que $rd(x) = x(1 + \varepsilon_x)$ con $|\varepsilon_x| \leq eps$.

Ahora analizamos los casos 1) y 3).

Puesto que un número finito de números b son útiles para expresar los exponentes en aritmética de punto flotante, hay desgraciadamente números $x \notin A$ tales que $rd(x) \notin A$. Así, si $t = 4$ y $b = 2$, consideramos los siguientes ejemplos:

1. $rd(0{,}31794 \times 10^{110}) = 0{,}3179 \times 10^{110} \notin A.$
2. $rd(0{,}99997 \times 10^{99}) = 0{,}1 \times 10^{100} \notin A.$
3. $rd(0{,}012345 \times 10^{-99}) = 0{,}1235 \times 10^{-100} \notin A.$
4. $rd(0{,}54321 \times 10^{-115}) = 0{,}5432 \times 10^{-115} \notin A.$

En los ejemplos 1) y 2), el exponente positivo es demasiado grande para almacenarlo en la memoria del computador, en estas condiciones se dice que el exponente está excedido de la capacidad de representación de exponentes en el computador, este caso se lo conoce como exponente en overflow. En el ejemplo 2) existe un overflow solo después de redondear. Situación análoga a la descrita precedentemente se presenta con los ejemplos 3) y 4); en tales casos se dice exponentes en underflow. En caso de underflow u overflow, pueden ser controlados, si por ejemplo se escribe:

$$\begin{aligned} rd(0{,}012345 \times 10^{-99}) &= 0{,}0123 \times 10^{-99} \in A, \\ rd(0{,}54321 \times 10^{-110}) &= 0 \in A, \\ rd(0{,}31794 \times 10^{110}) &= 0{,}3179 \times 10^{15} \times 10^{95}. \end{aligned}$$

Note que los números $0{,}3179 \times 10^{15}$, y 10^{95} pertenecen al conjunto A pero $0{,}3179 \times 10^{15} \times 10^{95} \notin A$. Para estos casos no se satisface la relación $rd(x) = x(1 + \varepsilon) \quad |\varepsilon| \leq eps$. En los computadores digitales estos números x que no pertenecen al conjunto M son tratados como irregularidades o errores en los datos. En el caso de underflow, $rd(x)$ puede ser indicado por 0 o se produce una detención en la ejecución del programa. En el caso de overflow, $rd(x)$ es indicado como un error en x y la inmediata detención del programa en ejecución.

Para evitar estos problemas es necesario incorporar en los programas contraseñas especiales o reescalar los datos de modo apropiado, lo que se traduce en elaborar programas especiales. Por lo dicho precedentemente y por abuso de lenguaje, podemos decir que existe una función $rd : \mathbb{R} \to \mathbb{A}$ definida por $rd(x) = x(1 + \varepsilon) \quad |\varepsilon| \leq eps$.

1.8. Aritmética de punto flotante

Operaciones aritméticas

Hemos denotamos con A el conjunto de números de máquina. Sean $x, y \in A$, en general, $x + y$, $x - y$, $x \times y$, $x/y \quad y \neq 0$, no son números de máquina. Definimos las operaciones aritméticas $\oplus$, $\ominus$, $\otimes$, $\oslash$ llamadas operaciones de punto flotante, como sigue:

$$\begin{aligned} x \oplus y &= rd(x+y), \qquad x \ominus y = rd(x-y), \\ x \otimes y &= rd(x \times y), \qquad x \oslash y = rd(x/y) \quad y \neq 0. \end{aligned}$$

De la definición de la función rd, resulta que

$$\begin{aligned} x \oplus y &= (x+y)(1+\varepsilon_1), \qquad x \ominus y = (x-y)(1+\varepsilon_2), \\ x \otimes y &= (x \times y)(1+\varepsilon_3), \qquad x \oslash y = (x/y)(1+\varepsilon_4), \end{aligned}$$

con $|\varepsilon_i| \leq eps \quad i = 1, 2, 3, 4.$

Las operaciones en punto flotante pueden no ser asociativas o distributivas, así : si $a, b, c \in A$, en general,

$$a \oplus (b \oplus c) \neq (a \oplus b) \oplus c, \qquad a \otimes (b \oplus c) \neq a \oplus b \oplus a \oplus c.$$

Comprobemos con un ejemplo. Sean $t = 5$ el número de cifras de la mantisa, $a = 0{,}21345 \times 10^{-2}$, $b = 0{,}33456 \times 10^{2}$, $c = -0{,}33341 \times 10^{2}$. Con estos datos, verifiquemos que $a \oplus (b \oplus c) \neq (a \oplus b) \oplus c$. Tenemos

$$b \oplus c = rd(0{,}33456 \times 10^{2} - 0{,}33341 \times 10^{2}) = rd(0{,}00115 \times 10^{2}) = 0{,}115 \times 10^{0},$$

luego

$$a \oplus (b \oplus c) = rd(0{,}21345 \times 10^{-2} + 0{,}115 \times 10^{0}) = rd(0{,}11714 \times 10^{0}) = 0{,}11714 \times 10^{0},$$

Calculemos $(a \oplus b) \oplus c$. Para ello calculamos primeramente $a \oplus b$. Tenemos

$$a \oplus b = rd(0{,}21345 \times 10^{-2} + 0{,}33456 \times 10^{2}) = rd(0{,}33458 \times 10^{2}) = 0{,}33458 \times 10^{2},$$

a continuación

$$(a \oplus b) \oplus c = rd(0{,}33458 \times 10^{2} - 0{,}33341 \times 10^{2}) = rd(0{,}00117 \times 10^{2}) = 0{,}117 \times 10^{0},$$

El valor exacto es $a + b + c = 0{,}1171345$. Note que $a \oplus (b \oplus c)$ se aproxima mejor a $a + b + c$.

Con la misma información verifiquemos que $a \otimes (b \oplus c) \neq a \otimes b \oplus a \otimes c$. Calculemos el lado izquierdo. Tenemos $b \oplus c = 0{,}115 \times 10^{0}$

$$\begin{aligned} a \otimes (b \oplus c) &= rd(0{,}21345 \times 10^{-2} * 0{,}115 \times 10^{0}) = rd(0{,}02454675 \times 10^{-2}) \\ &= rd(0{,}2454675 \times 10^{-3}) = 0{,}24547 \times 10^{-3}, \end{aligned}$$

Pasemos a calcular $a \otimes b \oplus a \otimes c$. Entonces

$$a \otimes b = rd(0{,}21345 \times 10^{-2} * 0{,}33456 \times 10^{2}) = rd(0{,}071411832) = 0{,}71412 \times 10^{-1},$$

$$a \otimes c = rd(-0{,}21345 \times 10^{-2} * 0{,}33341 \times 10^{2}) = -rd(0{,}0711663645 \times 10^{0}) = -0{,}71167 \times 10^{-1},$$

Con estos resultados calculemos $a \otimes b \oplus a \otimes c$:

$$a \otimes b \oplus a \otimes c = rd(0{,}71412 \times 10^{-1} - 0{,}71167 \times 10^{-1}) = rd(0{,}00245 \times 10^{-1}) = 0{,}245 \times 10^{-3},$$

Claramente $a \otimes (b \oplus c) = 0{,}24547 \times 10^{-3} \neq a \otimes b \oplus a \otimes c = 0{,}245 \times 10^{-3}$. Valor exacto $a(b + c) = 0{,}0002454675 = 0, 2454675 \times 10^{-3}$.

Expresiones aritméticas y funciones.

Sea E una expresión aritmética. En punto flotante la evaluación de E se nota con $fl(E)$. Sea E una función real definida en un subconjunto I de $\mathbb{R}$. El valor $E(x)$ en $x \in I$ en punto flotante se nota con $E_*(x)$ y se define por

$$E_*(x) = fl(E(x)).$$

Ejemplos

1. Sean $a, b, c \in \mathbb{R}$.
   a) Si $E = a + (b + c)$, entonces $fl(E) = a \oplus (b \oplus c)$.
   b) Si $E = (ab)c$, $fl(E) = (a \otimes b) \otimes c$.
   c) Si $E = a(b + c)$, $fl(E) = a \otimes (b \oplus c)$.

2. Sean $a = 0{,}18 \times 10^2$,$b = 0{,}3596 \times 10^0$,$c = 0{,}1 \times 10^1$,$t = 4$ el número de cifras de la mantisa, calculemos $E = \dfrac{a}{b} + c$ en punto flotante. De la definición de punto flotante de E, tenemos

$$\begin{aligned} fl(E) &= rd([0{,}18 \times 10^2 \oslash 0{,}3596 \times 10^0] \oplus 0{,}1 \times 10^1) = rd(0{,}50055611 \times 10^2 \oplus 0{,}1 \times 10^1) \\ &= rd(0{,}5006 \times 10^2 \oplus 0{,}1 \times 10^1) = 0{,}5106 \times 10^2. \end{aligned}$$

3. Sea $E(x) = sen(x)$ $x \in \mathbb{R}$. Entonces $E_*(x) = fl(sen(x))$ que lo notaremos $sen_*(x)$. Para $t = 5$, y $x = \dfrac{\pi}{6}$,se tiene $sen(\dfrac{\pi}{6}) = 0{,}5$, mientras que $rd(\dfrac{\pi}{6}) = 0{,}5236 \times 10^0$ y en consecuencia $\text{sen}_*(0{,}5236 \times 10^0) = 0{,}5$. En el capítulo 3 se propone un algoritmo de cálculo de sen(x) $x \in \mathbb{R}$.

4. Sea $E(x) = e^x$ $x \in \mathbb{R}$. Entonces $E_*(x) = fl(e^x)$ que lo notaremos e^x_*. Si $x = 0{,}5$ y $t = 5$, entonces $e^{0{,}5} = 1{,}648721171\ldots$, $e^{0{,}5}_* = 0{,}16487 \times 10^1$. En el capítulo 3 se propone un algoritmo de cálculo de e^x $x \in \mathbb{R}$.

5. $fl(\sqrt{x}) = \sqrt{x}^*$, $x \geq 0$. Para $t = 5$, $x = 0{,}14567$, $\sqrt{x} = 0{,}381667395\ldots$, $\sqrt{x}^* = 0{,}38167 \times 10^0$. Más adelante se muestra un algoritmo de cálculo de $\sqrt{x}$ $x > 0$.

Observación. Sean $a, b \in \mathbb{R}$. Las operaciones aritméticas en punto flotante se expresan de la manera siguiente.

$$\begin{aligned} a \oplus b &= rd(rd(a) + rd(b)), \qquad a \ominus b = rd(rd(a) - rd(b)), \\ a \otimes b &= rd(rd(a) * rd(b)), \qquad a \oslash b = rd(rd(a)/rd(b)) \quad rd(b) \neq 0. \end{aligned}$$

Por abuso de lenguaje, a las operaciones en punto flotante las notaremos del mismo que las operaciones aritméticas habituales con números reales.

1.9. Condicionamiento de funciones reales.

La calidad de la solución numérica de un problema (P) depende fuertemente del método numérico empleado y este a su vez depende de dos componentes importantes: el condicionamiento y la estabilidad; y, para problemas cuyas soluciones se aproximan mediante sucesiones, dependen a más de los componentes anteriores, de la convergencia.

En esta sección se introduce la noción de condicionamiento que es muy importantes en la construcción de algoritmos, procedimientos de cálculo y de la elaboración de programas computacionales, y que constituyen las bases que deben tenerse siempre presentes para el desarrollo de software en el cálculo científico. Trataremos primeramente el condicionamiento de funciones reales de una sola variable, a continuación trataremos el condicionamiento de funciones reales en varias variables.

1.9.1. Condicionamiento de funciones reales de una sola variable.

Sea $\varphi : [a,b] \to \mathbb{R}$ una función derivable en $]a,b[$. Ponemos $y = \varphi(x) \quad x \in [a,b]$. Investiguemos como el error absoluto $\triangle x$ de x influye en el cálculo de y, donde $\triangle x = \widetilde{x} - x$ y $\widetilde{x} = rd(x)$. Se pone $\widetilde{y} = \varphi(\widetilde{x})$. Nos interesa determinar la influencia de los errores (redondeo, truncamiento) del dato de entrada x, esto es, de Δx en el dato de salida $y = \varphi(x)$, es decir en $\widetilde{y} = \varphi(\widetilde{x})$ y como medir esa influencia.

Supongamos que la función φ es al menos dos veces derivable en $]a,b[$ y que $\mid \varphi'' \mid$ es acotada en $]a,b[$. Por el desarrollo de Taylor, se tiene $\varphi(\widetilde{x}) = \varphi(x) + \varphi'(x)(\widetilde{x} - x) + \frac{1}{2}(\widetilde{x} - x)^2\varphi''(\xi)$ con ξ entre x y $\widetilde{x}$, y $(\widetilde{x} - x)^2 = (\triangle x)^2 < eps$, lo que implica que $\frac{1}{2}(\widetilde{x} - x)^2 \mid \varphi''(\xi) \mid \simeq 0$.

Definición 5 *El error relativo de y se define mediante la relación*

$$\varepsilon_y = \frac{\widetilde{y} - y}{y} \quad y \neq 0,$$

donde $\widetilde{y} = \varphi(\widetilde{x})$.

Usando el desarrollo de Taylor en primera aproximación, tenemos

$$\triangle y = \widetilde{y} - y = \varphi(\widetilde{x}) - \varphi(x) = \varphi'(x)(\tilde{x} - x) = x\ \varphi'(x)\ \frac{\widetilde{x} - x}{x} = x\varphi'(x)\varepsilon_x \quad x \neq 0.$$

Luego,

$$\varepsilon_y = \frac{\widetilde{y} - y}{y} = \frac{\Delta y}{y} = \frac{x\varphi'(x)}{\varphi(x)}\varepsilon_x, \quad \varphi(x) \neq 0.$$

Note que si $\varphi''(x)$ existe, el término $\frac{1}{2}(\Delta x)^2\ \varphi''(x)$ se redondea por 0 debido a que $(\Delta x)^2$ se redondea por 0.

Definición 6 *El número real $c(x) = \frac{x\varphi'(x)}{\varphi(x)}$ con $\varphi(x) \neq 0$ se llama número de condicionamiento de la función φ en el punto x.*

El número de condicionamiento $c(x)$ indica cuán grande es el error relativo de y ante variaciones del dato de entrada x. Cuando $\mid c(x) \mid > 1$ el error relativo ε_y se amplifica y cuando $\mid c(x) \mid \leq 1$ el error relativo ε_y se contrae.

Definición 7 *Diremos que $y = \varphi(x)$ está bien condicionado si $|c(x)| \leq 1$. En el caso contrario, diremos que $y = \varphi(x)$ está mal condicionado.*

Ejemplos

1. Consideremos la función f definida por $f(x) = e^x \quad x \in \mathbb{R}$. Es conocido que la función f es derivable, y $f'(x) = e^x \quad \forall x \in \mathbb{R}$. El número de condicionamiento de esta función está definido como $c(x) = \frac{xf'(x)}{f(x)} = \frac{xe^x}{e^x} = x \quad \forall x \in \mathbb{R}$. Luego

   $$|c(x)| \leq 1 \Leftrightarrow |x| \leq 1 \Leftrightarrow x \in [-1,1]\,.$$

   Por lo tanto $y = e^x$ está bien condicionado si y solo si $x \in [-1,1]$, en el caso contrario la función f está mal condicionada.
   Definimos $g(x) = \left(e^{\frac{x}{n}}\right)^n \quad x \in \mathbb{R}$. Entonces

   $$c(x) = \frac{xg'(x)}{g(x)} = \frac{x\left[n(e^{\frac{x}{n}})^{n-1}e^{\frac{x}{n}} \cdot \frac{1}{n}\right]}{(e^{\frac{x}{n}})^n} = x.$$

La utilización de $g(x) = (e^{\frac{x}{n}})^n = e^x$ con $n \in \mathbb{Z}^+$, $x \in \mathbb{R}$ es más ventajoso del punto de vista de la elaboración de un algoritmo que permita evaluar e^x.

2. Sea $n \in \mathbb{Z}^+$ y f la función dada por $f(x) = x^n \quad x \in \mathbb{R}$. Para $x \neq 0$, tenemos

$$c(x) = \frac{xf'(x)}{f(x)} = \frac{x(nx^{n-1})}{x^n} = n.$$

Luego $y = x^n$ está bien condicionado si y solo si $n = 1$. Para $n > 1$, la función f está mál condicionada, esto significa que la potencia x^n está mal condicionada cuando $n > 1$. Es por esta razón que se evita el cálculo directo de las potencias. Anteriormente vimos algunos ejemplos de cálculos con polinomios en los que se evitan los cálculos directos de las potencias, de esta manera se mejora el condicionamiento con lo que se logra mejorar los resultados.

3. Considerar la función f definida por $f(x) = x^{\frac{1}{n}}$ con $n \in \mathbb{Z}^+$, $x > 0$. El número de condicionamiento está definido como

$$c(x) = \frac{xf'(x)}{f(x)} = \frac{x\frac{1}{n}x^{\frac{1}{n}-1}}{x^n} = \frac{1}{n}.$$

Resulta que $y = x^{1/n}$ está bien condicionado $\forall x \in \mathbb{R}^+, \quad n \in \mathbb{Z}^+$.

4. Sea $f(x) = sen\,(x) \quad x \in \mathbb{R}$. El número de condicionamiento de esta función está definido como

$$c(x) = \frac{x\cos(x)}{sen(x)} \qquad x \neq k\pi, \;\; k \in \mathbb{Z}.$$

Para el análisis del número de condicionamiento $c(x)$ lo dividimos en dos partes.

a) Puesto que $c(x) = \dfrac{\cos(x)}{\dfrac{sen(x)}{x}}$. Entonces

$$\lim_{x\to 0} c(x) = \frac{\lim\limits_{x\to 0}\cos(x)}{\lim\limits_{x\to 0}\dfrac{sen(x)}{x}} = \frac{1}{1} = 1,$$

que muestra que en $x = 0$ se tiene una discontinuidad evitable. Además,

$$\lim_{x\to \frac{\pi}{2}} c(x) = \frac{\lim\limits_{x\to \frac{\pi}{2}} x\cos(x)}{\lim\limits_{x\to \frac{\pi}{2}} sen(x)} = 0.$$

b) Por otro lado, si escribamos $c(x) = \dfrac{x}{\tan(x)}$. Tenemos

$$|c(x)| \leq 1 \Leftrightarrow |x| \leq |\tan(x)| \quad x \in \left]-\frac{\pi}{2}, 0\right[\cup \left]0, \frac{\pi}{2}\right[.$$

Definimos $g(x) = -x + \tan(x)$. Resulta que $g'(x) = -1 + \sec^2(x)$. Entonces

$$g'(x) = 0 \Leftrightarrow \cos^2(x) = 1 \Leftrightarrow x = 2k\pi, \;\; k \in \mathbb{Z}.$$

Además $g''(x) = 2\sec^2(x)\tan(x)$. Luego, $g''(x) > 0 \quad$ si $x \in \left]0, \frac{\pi}{2}\right[$, y $g''(x) < 0$ si $x \in \left]-\frac{\pi}{2}, 0\right[$. Adicionalmente, g es creciente en $\left]0, \frac{\pi}{2}\right[$, luego $g(x) > g(0) = 0$ con lo cual $\tan(x) > x$.

En conclusión

$$|c(x)| \leq 1 \Leftrightarrow x \in \left]-\frac{\pi}{2}, 0\right[\cup \left]0, \frac{\pi}{2}\right[.$$

Sea $\widetilde{c}(x) = \begin{cases} 1, \text{ si } x = 0, \\ c(x), \text{ si } x \in \left]-\frac{\pi}{2}, 0\right[\cup \left]0, \frac{\pi}{2}\right[. \end{cases}$ Entonces $|\widetilde{c}(x)| \leq 1 \quad \forall x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$, es decir que $\text{sen}(x)$ está bien condicionado en el intervalo $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. Esta propiedad será utilizada para aproximar $\text{sen}(x)$ mediante la serie de Taylor que se verá en el capítulo posterior.

5. Sea $\varphi : \mathbb{R}^+ \to \mathbb{R}$ función dada por $\varphi(x) = y^x$, donde $y > 0$ es fijo. Entonces

$$c(x) = \frac{x}{e^{x\ln(y)}} e^{x\ \ln(y)} \ln(y) = x\ln(y).$$

Luego

$$|c(x)| \leq 1 \Leftrightarrow |x| \leq \frac{1}{|\ln(y)|} \Leftrightarrow x \in \left[\frac{-1}{|\ln(y)|}, \frac{1}{|\ln(y)|}\right] \quad \text{con } y > 0,\ y \neq 1.$$

6. Sea $\varphi : \mathbb{R}^+ \to \mathbb{R}$ la función definida por $\varphi(x) = x^y$ con $y \in \mathbb{R}^+$ fijo. Se tiene

$$c(x) = \frac{x}{e^{y\ln(x)}} e^{y\ln(x)} \times \frac{y}{x} = y.$$

La función φ está bien condicionada si $|y| \leq 1$. Note que si $y \in \mathbb{N}$, $c(x) = y$ fue obtenido anteriormente.

7. Se desea calcular $f = (\sqrt{2}-1)^6$. Se da una aproximación de $\sqrt{2} \simeq 1{,}414$ y seis algoritmos para su cálculo

$$\begin{aligned} f_1 &= (\sqrt{2}-1)^6, \qquad f_2 = \frac{1}{(\sqrt{2}+1)^6}, \\ f_3 &= (3-2\sqrt{2})^3, \qquad f_4 = \frac{1}{(3+2\sqrt{2})^3}, \\ f_5 &= \frac{1}{99+70\sqrt{2}}, \qquad f_6 = 99-70\sqrt{2}. \end{aligned}$$

¿Qué algoritmo está bien condicionado?

Para responder a esta pregunta, primeramente vamos a cálcular los números de condicionamiento asociados con los algoritmos propuestos. Para el efecto denotamos con ε_x el error relativo al dato de entrada $x = \sqrt{2}$.

a) Sea f_1 la función dada por $f_1(x) = (x-1)^6$, entonces

$$f_1'(x) = 6(x-1)^5, \qquad f_1(1{,}414) = (1{,}414-1)^6 = 0{,}005,$$

luego

$$|\varepsilon_{f_1}| = \left|\frac{x}{f_1(x)} f_1'(x)\varepsilon_x\right| = \left|\frac{1{,}414}{0{,}005} \times 6 \times (0{,}414)^5\right| |\varepsilon_x| = 20{,}636\,|\varepsilon_x|\,.$$

b) Sea $f_2(x) = \dfrac{1}{(x+1)^6}$. Entonces $f_2'(x) = -\dfrac{6}{(x+1)^7}$. Resulta $f_2(1{,}414) = 0{,}005$ y

$$|\varepsilon_{f_2}| = \left|\frac{x}{f_2(x)} f_2'(x)\varepsilon_x\right| \left|\frac{1{,}414}{0{,}005}\right| \left(-\frac{6}{(1{,}414+1)^7}\right) |\varepsilon_x| = 3{,}552\,|\varepsilon_x|\,.$$

c) Consideramos la función f_3 definida como $f_3(x) = (3-2x)^3$. Entonces $f_3'(x) = -6(3-2x)^2$. Tenemos $f_3(1{,}414) = 0{,}005$, y

$$|\varepsilon_{f_3}| = \left|\frac{1{,}414}{0{,}005}\right| \left(-6(3-2\times 1{,}414)^2\right) |\varepsilon_x| = 50{,}198\,|\varepsilon_x|\,.$$

d) Tal como en los casos precedentes, sea $f_4(x) = \dfrac{1}{(3+2x)^3} \Longrightarrow f_4'(x) = -\dfrac{6}{(3+2x)^4}$. Se tiene $f_4(1{,}414) = 0{,}005$, y

$$|\varepsilon_{f_4}| = \left|\frac{1{,}414}{0{,}005}\right| \left(-\frac{6}{(3+2\times 1{,}414)^4}\right) |\varepsilon_x| = 1{,}471\,|\varepsilon_x|\,.$$

e) Sea $f_5(x) = \dfrac{1}{99+70x} \Longrightarrow f_5'(x) = -\dfrac{70}{(99+70x)^2}$. Entonces

$$|\varepsilon_{f_5}| = \left|\frac{1{,}414}{0{,}005} \times \frac{-70}{(99+70\times 1{,}414)^2}\right| |\varepsilon_x| = 0{,}5\,|\varepsilon_x|\,.$$

f) De la definición del algoritmo f_6, se sigue que $f_6(x) = 99 - 70x \Longrightarrow f_6'(x) = -70$. Resulta$f_6(1{,}414) = 0{,}020$,

$$|\varepsilon_{f_6}| = \left|\frac{1{,}414}{0{,}020} \times (-70)\right| |\varepsilon_x| = 4949{,}0\,|\varepsilon_x|\,.$$

Comparando los números de condicionamiento de cada una de las funciones, observamos que f_5 tiene el más pequeño número de condicionamiento, esto es, f_5 está bien condicionado, mientras que f_6 tiene el más grande número de condicionamiento, es decir que f_6 está mál condicionado y de hecho es el peor algoritmo que se puede utilizar para calcular el valor aproximado de $(\sqrt{2}-1)^6$. En conclusión, el mejor algoritmo para el cálculo de $(\sqrt{2}-1)^6$ con $\sqrt{2} \simeq 1{,}414$ es $f_5 = \dfrac{1}{99+70\sqrt{2}}$.

1.9.2. Condicionamiento de funciones reales en varias variables

Sean $\Omega \subset \mathbb{R}^n$ un abierto y $\overrightarrow{\varphi}$ una función de Ω en $\mathbb{R}^m$ que suponemos diferenciable en todo punto de Ω, la función $\overrightarrow{\varphi}$ se denomina campo vectorial. Ponemos $\vec{y} = \vec{\varphi}(\vec{x}) = \begin{bmatrix} \varphi_1(\overrightarrow{x}) \\ \vdots \\ \varphi_m(\overrightarrow{x}) \end{bmatrix}$ $\vec{x}^T = (x_1 \ldots, x_n) \in \Omega$, donde $\varphi_1, \cdots, \varphi_m$ son campos escalares diferenciables en Ω.

Para $\vec{x}^T = (x_1 \ldots, x_n) \in \Omega$, ponemos $\widetilde{x}_i = rd(x_i) \quad i = 1, \ldots, n$, y definimos $\widetilde{x}^T = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, $\Delta\overrightarrow{x}^T = \widetilde{x}^T - \overrightarrow{x}^T$. El error relativo de x_i está definido mediante la relación:

$$\varepsilon_{x_i} = \frac{\tilde{x}_i - x_i}{x_i} \qquad x_i \neq 0, \quad i = 1, \ldots, n.$$

Se define $\widetilde{y} = \overrightarrow{\varphi}(\widetilde{x}) = \begin{bmatrix} \varphi_1(\widetilde{x}) \\ \vdots \\ \varphi_m(\tilde{x}) \end{bmatrix}$, y, $\triangle\overrightarrow{y} = \widetilde{y} - \overrightarrow{y} = \begin{bmatrix} \widetilde{y}_1 - y_1 \\ \vdots \\ \widetilde{y}_m - y_m \end{bmatrix} = \begin{bmatrix} \triangle y_1 \\ \vdots \\ \triangle y_n \end{bmatrix}$.

Determinemos el error relativo $\varepsilon_{y_i} = \dfrac{\Delta y_i}{y_i} \quad i = 1, \ldots, m$. Usando el desarrollo de Taylor en primera aproximación, eliminando los desarrollos de orden superior, tenemos

$$\Delta y_i = \widetilde{y}_i - y_i = \varphi_i(\tilde{x}) - \varphi_i(\vec{x}) = \nabla\varphi_i(\vec{x}).\Delta\vec{x} = \sum_{j=1}^{n} (\widetilde{x}_j - x_j)\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}.$$

Para $\vec{x}^T = (x_1 \ldots, x_n) \in \Omega$ tal que $x_j \neq 0, \quad j = 1, \ldots, n$, se tiene la siguiente relación

$$\widetilde{x}_j - x_j = \frac{\widetilde{x}_j - x_j}{x_j} x_j = x_j\varepsilon_{x_j} \qquad x_j \neq 0, \quad j = 1, \ldots, n,$$

y en consecuencia

$$\Delta y_i = \sum_{j=1}^{n} (\widetilde{x}_j - x_j)\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j} = \sum_{j=1}^{n} x_j \frac{\partial\varphi_i(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} \qquad i = 1, \ldots, m.$$

Luego, para $y_i \neq 0 \quad i = 1, \ldots, m$, se tiene

$$\varepsilon_{y_i} = \frac{\Delta y_i}{y_i} = \frac{1}{y_i}\sum_{j=1}^{n} x_j \frac{\partial\varphi_i(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} = \sum_{j=1}^{n} \frac{x_j}{y_i}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} = \sum_{j=1}^{n} \frac{x_j}{\varphi_i(\overrightarrow{x})}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j} \qquad i = 1, \ldots, m.$$

Observamos que cada ε_{y_i} depende de los factores de amplificación $\frac{x_j}{\varphi_i(\overrightarrow{x})}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j}$ de ε_{x_j}, $\quad i=1,\ldots,m$, $j=1,\ldots,n$,

Definición 8 *El conjunto de números reales* $\{C_{ij}(\overrightarrow{x}) \mid i=1,\ldots,m, \quad j=1,\ldots,n\}$, *donde* $C_{ij}(\overrightarrow{x})$ *está definido como*

$$C_{ij}(\overrightarrow{x})=\frac{x_j}{\varphi_i(\overrightarrow{x})}\frac{\partial\varphi_i\,(\overrightarrow{x})}{\partial x_j} \text{ con } \varphi_i(\overrightarrow{x})\neq 0, \quad i=1,\ldots,m, \quad j=1,\ldots,n,$$

se llaman números de condicionamiento de la función $\overrightarrow{\varphi}$ *en* $\overrightarrow{x}\in\Omega$.
La matriz $C(\overrightarrow{x})=(C_{ij}(\overrightarrow{x}))_{m\times n}$ *se llama matriz de condicionamiento de* $\overrightarrow{\varphi}$ *en* $\overrightarrow{x}\in\Omega$.

En el caso en que $m=1$, esto es, φ es un campo escalar, la matriz de condicionamiento de φ en $\overrightarrow{x}\in\Omega$ se identifica con el vector fila $C(\overrightarrow{x})$ definido como

$$C(\overrightarrow{x})=\left(\frac{x_i}{\varphi(\overrightarrow{x})}\frac{\partial\,\varphi}{\partial x_i}(\overrightarrow{x}),\ldots,\frac{x_n}{\varphi(\overrightarrow{x})}\frac{\partial\varphi}{\partial x_n}(\overrightarrow{x})\right),$$

al que lo denominaremos vector de condicionamiento de φ en $\overrightarrow{x}\in\Omega$.

Definimos

$$\overrightarrow{\varepsilon}_{\overrightarrow{y}}=\begin{bmatrix}\varepsilon_{y_1}\\ \vdots\\ \varepsilon_{y_n}\end{bmatrix}=\begin{bmatrix}\displaystyle\sum_{j=1}^{n}\frac{x_j}{\varphi_1(\overrightarrow{x})}\frac{\partial\varphi_1}{\partial x_j}(\overrightarrow{x})\varepsilon_{x_j}\\ \vdots\\ \displaystyle\sum_{j=1}^{n}\frac{x_j}{\varphi_m(\overrightarrow{x})}\frac{\partial\varphi_m(\overrightarrow{x})}{\partial x_j}\varepsilon_{x_j}\end{bmatrix}=C(\overrightarrow{x})\overrightarrow{\varepsilon}_{\overrightarrow{x}},$$

donde $\overrightarrow{\varepsilon}_{\overrightarrow{x}}=\begin{bmatrix}\varepsilon_{x_1}\\ \vdots\\ \varepsilon_{xn}\end{bmatrix}$. Así, $\overrightarrow{\varepsilon}_{\overrightarrow{y}}=C(\overrightarrow{x})\overrightarrow{\varepsilon}_{\overrightarrow{x}}$.

Definición 9 *Se dice que* $\overrightarrow{y}=\overrightarrow{\varphi}(\overrightarrow{x})$ *está bien condicionado en* $\overrightarrow{x}\in\Omega$ *si y solo si* $\mid C_{ij}(\overrightarrow{x})\mid\leq 1$ $\forall i=1,\ldots,m, \quad j=1,\ldots,n$. *En el caso contrario, se dice que* $\overrightarrow{y}=\overrightarrow{\varphi}(\overrightarrow{x})$ *está mal condicionado en* $\overrightarrow{x}\in\Omega$.

Para determinar el condicionamiento de un campo vectorial diferenciable $\overrightarrow{\varphi}$ en $\overrightarrow{x}\in\Omega$ se deben estudiar todos los números de condicionamiento $C_{ij}(\overrightarrow{x}) \quad i=1,\ldots,m, \quad j=1,\ldots,n$. Hacemos notar que solo en pocos casos es posible determinar $\overrightarrow{x}$ tal que $\mid C_{ij}(\overrightarrow{x})\mid\leq 1$. En la generalidad de los casos, es muy difícil y casi imposible determinar $\overrightarrow{x}$ tal que $\mid C_{ij}(\overrightarrow{x})\mid\leq 1$, por lo que se recurre a otros procedimientos para estimar el condicionamiento. Así, en algunos casos el número de condicionamiento C se define por la desigualdad (cociente de Raileygh-Ritz):

$$\frac{\parallel\overrightarrow{\varphi}(\widetilde{x})-\overrightarrow{\varphi}(\overrightarrow{x})\parallel}{\parallel\overrightarrow{\varphi}(\overrightarrow{x})\parallel}\leq C\frac{\parallel\widetilde{x}-\overrightarrow{x}\parallel}{\parallel\overrightarrow{x}\parallel} \qquad \overrightarrow{\varphi}(\overrightarrow{x})\neq 0, \quad \overrightarrow{x}\neq 0,$$

donde $C>0$ es el número de condicionamiento.

Ejemplos

1. Sea φ la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida como $\varphi(x,y)=x+y \quad (x,y)\in\mathbb{R}^2$. Supongamos que para $(x,y)\in\mathbb{R}^2$ se tiene $z=\varphi(x,y)\neq 0$. Determinemos el error relativo de z. Este está dado como sigue:

$$\varepsilon_z=\frac{x}{\varphi(x,y)}\frac{\partial\varphi}{\partial x}(x,y)\ \varepsilon_x+\frac{y}{\varphi(x,y)}\frac{\partial\varphi}{\partial y}(x,y)\ \varepsilon_y=\frac{x}{x+y}\varepsilon_x+\frac{y}{x+y}\varepsilon_y.$$

Los números de condicionamiento de φ en $(x,y) \in \mathbb{R}^2$ están definidos como $C_x = \dfrac{x}{x+y}$, $C_y = \dfrac{y}{x+y}$, y el vector de condicionamiento de φ en $(x,y) \in \mathbb{R}^2$ está definido como $C(x,y) = \left(\dfrac{x}{x+y}, \dfrac{y}{x+y}\right)$.

Analicemos los números de condicionamiento C_x y C_y.

a) Si $\begin{cases} x>0, & y>0, \\ x<0, & y<0, \end{cases}$ o entonces $\mid C_x(x,y) \mid < 1$, $\mid C_y(x,y) \mid < 1$. Luego $z = \varphi(x,y)$ está bien condicionado.

b) Si $x>0$ e $y<0$ tal que $x \neq -y$, entonces al menos uno de los números $|C_x|$ o $|C_y|$ es mayor que 1; en cuyo caso $z = \varphi(x,y)$ está mal condicionado.

En conclusión, la suma de dos números positivos (respectivamente negativos) está bien condicionada, mientras que la suma de dos números uno positivo y otro negativo está mal condicionada, esto equivale a decir que si x, y son números reales positivos, la resta $x-y$ está mal condicionada y consecuentemente la resta de dos números positivos es una operación peligrosa fundamentalmente si $x \neq -y$, $x \simeq -y$.

El resultado que acabamos de obtener se puede extender a sumas de tres o más números reales. Así, sean $x_1, \cdots, x_m \in \mathbb{R}$ y $z = x_1 + \cdots + x_m$. Entonces z está bien condicionado si y solo si $x_i > 0$ $\forall i = 1, \cdots, m$ (respectivamente $x_i < 0 \quad \forall i = 1, \cdots, m$), en el caso contrario tenemos que z está mal condicionado, más áun, las sumas y restas alternadas de números reales positivos está mal condicionada, por lo que este tipo de cálculos son peligrosos ya que amplifican los errores. Para aclarar más estas ideas, sean $x_1, \cdots, x_{2m} \in \mathbb{R}^+$ y $z = x_1 - x_2 + x_3 - x_4 + \cdots + x_{2m-1} - x_{2m}$. Esta suma está mal condicionada, ¿cómo mejorar el resultado? Ecribimos z en la siguiente forma

$$z = x_1 + x_3 + \cdots + x_{2n-1} - x_2 - x_4 - \cdots - x_{2m} = x_1 + x_3 + \cdots + x_{2n-1} - (x_2 + x_4 + \cdots + x_{2m})$$

La sumas $z_1 = x_1 + x_3 + \cdots + x_{2n-1}$, $z_2 = x_2 + x_4 + \cdots + x_{2m}$ están bien condicionadas, luego $z = z_1 - z_2$ con lo que se mejora el resultado. Más adelante se exhiben ejemplos.

2. Sea φ la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida como $\varphi(x,y) = xy \quad (x,y) \in \mathbb{R}^2$. Supongamos que para $(x,y) \in \mathbb{R}^2$ se tiene $z = \varphi(x,y) \neq 0$, esto es $x \neq 0$, $y \neq 0$, entonces,

$$C(x,y) = \left(\frac{x}{\varphi(x,y)}\,\frac{\partial \varphi}{\partial x}(x,y),\ \frac{y}{\varphi(x,y)}\frac{\partial \varphi}{\partial y}(x,y)\right) = \left(\frac{x}{xy}y,\ \frac{y}{xy}x\right) = (1,1).$$

Se tiene que $C_x(x,y) = 1$, $C_y(x,y) = 1$, por lo que el producto de dos números reales no nulos está bien condicionado y por tanto el producto de dos números no es una operación peligrosa.

3. Sean $p, q \in \mathbb{R}$ tales que $p^2 - 4q \geq 0$. Consideramos la ecuación: $x \in \mathbb{R}$ tal que $x^2 + px + q = 0$ cuyas raíces reales son

$$x_1 = \frac{1}{2}\left(-p + \sqrt{p^2-4q}\right), \qquad x_2 = \frac{1}{2}\left(-p - \sqrt{p^2-4q}\right), \quad \text{donde } p^2 - 4q \geq 0.$$

Estas raíces dependen de p y q, lo que nos permite definir las funciones reales φ, θ como

$$\begin{cases} \varphi(p,q) = \dfrac{1}{2}\left(-p + \sqrt{p^2-4q}\right), \\ \theta(p,q) = -\dfrac{1}{2}\left(p + \sqrt{p^2-4q}\right), \end{cases} \quad (p,q) \in \mathbb{R}^2 \text{ tal que } p^2 - 4q \geq 0.$$

La función φ está asociada a la raíz x_1 mientras que la función θ está asociada a la raíz x_2. Estudiemos el condicionamiento de la primera raíz, esto es, el condicionamiento de la función φ. Tenemos

$$\begin{cases} \dfrac{\partial \varphi}{\partial p}(p,q) = \dfrac{-p + \sqrt{p^2-4q}}{2\sqrt{p^2-4q}}, \\ \dfrac{\partial \varphi}{\partial q}(p,q) = -\dfrac{1}{\sqrt{p^2-4q}}, \end{cases} \quad (p,q) \in \mathbb{R}^2 \text{ tal que } p^2 - 4q > 0.$$

Luego

$$\varepsilon_\varphi = \frac{p}{\varphi(p,q)}\frac{\partial\varphi}{\partial p}\varepsilon_p + \frac{q}{\varphi(p,q)}\frac{\partial\varphi}{\partial q}\varepsilon_q = -\frac{p}{\sqrt{p^2-4q}}\varepsilon_p + \frac{p+\sqrt{p^2-4q}}{2\sqrt{p^2-4q}}\varepsilon_q.$$

Los números de condicionamiento de φ están definidos como $\begin{cases} C_p(p,q) = -\dfrac{p}{\sqrt{p^2-4q}}, \\ C_q(p,q) = \dfrac{p+\sqrt{p^2-4q}}{2\sqrt{p^2-4q}}, \end{cases}$ siempre que $(p,q) \in \mathbb{R}^2$ tal que $p^2-4q>0$. Analicemos cada uno de estos números de condicionamiento. Si $q<0$, se tiene

$$\left|\frac{p}{\sqrt{p^2-4q}}\right| < 1, \qquad \left|\frac{p+\sqrt{p^2-4q}}{2\sqrt{p^2-4q}}\right| < 1,$$

con lo cual $x_1 = \varphi(p,q)$ está bien condicionado. Si $q>0$ tal que $p^2+4q>0$, φ está mal condicionado. El número de condicionamiento $\mid C_p(p,q) \mid$ es mucho más grande aún en la situación siguiente: $(p,q) \in \mathbb{R}^2$ tal que $q>0$, $p^2 \simeq 4q$ de modo que $p^2+4q>0$. Esto nos muestra que no es conveniente calcular x_1 con la fórmula arriba propuesta, sino con la que se obtiene del modo siguiente:

$$x_1 = \frac{1}{2}\left(-p+\sqrt{p^2-4q}\right) = \frac{1}{2}\frac{\left(-p+\sqrt{p^2-4q}\right)\left(-p-\sqrt{p^2-4q}\right)}{-p-\sqrt{p^2-4q}} = -\frac{2q}{p+\sqrt{p^2-4q}}.$$

Veamos un ejemplo numérico de esta situación. Consideremos la ecuación $x \in \mathbb{R}$ solución de $x^2+62{,}10x+1=0$. Entonces $x_1 = \frac{1}{2}\left(-62{,}10+\sqrt{(62{,}10)^2-4}\right) = -0{,}1610723\times 10^{-1}$. Efectuemos el cálculo de x_1 con 4 cifras decimales en aritmética de punto flotante, se tiene

$$\begin{aligned} \tilde{x}_1 &= fl(x_1) = 0{,}5\times 10^0 * \left(-0{,}6210\times 10^2 + \sqrt{(0{,}6210\times 10^2) - 0{,}4\times 10^{1^*}}\right) \\ &= 0{,}5\times 10^0 * (-0{,}6210\times 10^2 + 0{,}6206\times 10^2) = -0{,}2\times 10^{-2}. \end{aligned}$$

Utilicemos ahora la nueva escritura de x_1. Obtenemos

$$\widetilde{t_1} = fl(x_1) = -\frac{0{,}2\times 10^1 \ * \ 0{,}1\times 10^1}{0{,}6210\times 10^2 + 0{,}6207\times 10^2} = -\frac{0{,}2\times 10^1}{0{,}1242\times 10^3} = -0{,}1610\times 10^1.$$

Se observa que $\widetilde{t_1}$ es una mejor aproximación de x_1.

Nota. Tomando en consideración el valor absoluto de los números de condicionamiento, de los ejemplos se establece la jerarquía de las operaciones siguientes: la radicación de números reales positivos, la suma de números reales positivos (respectivamente suma de números reales negativos) son consideradas operaciones no peligrosas. A continuación se tiene el producto y cociente de números reales. La potenciación está bien condicionada si el exponente es igual a 1, por este motivo el esquema de Hörner evita el cálculo directo de las potencias. La suma de números reales de signos opuestos es una operación peligrosa ya que al menos un número de condicionamiento es mayor que 1 lo que amplifica los errores. Por esta razón debe evitarse sumas sucesivas con números reales de signos opuestos. De preferencia deben escribirse los algoritmos de modo que se tengan sumas de números positivos y reducir como sea posible las sumas de números con signos opuestos. Igualmente, debe evitarse el cálculo directo de las potencias con exponentes mayores que 1.

1.10. Propagación de los errores.

Ejemplos

1. Sean $a,b,c, \in \mathbb{R}$, y $E = a+b+c$. Se tiene $E = a+(b+c) = (a+b)+c = (a+c)+b$, y por tanto se disponen de tres algoritmos para evaluar E.

Primer algoritmo. Tenemos $E = a + (b + c)$, que puede verse como la composición de funciones siguiente:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \xrightarrow{\varphi_0} \begin{bmatrix} a \\ \\ b+c \end{bmatrix} \xrightarrow{\varphi_1} a + (b + c),$$

donde $\varphi_0 : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$ es la función definida por $\varphi_0(x, y, z) = \begin{bmatrix} x \\ y + z \end{bmatrix}$, y $\varphi_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$ es la función dada por $\varphi_1(u, v) = u + v$.

Luego

$$E = (\varphi_1 \circ \varphi_0)(a, b, c) = \varphi_1(\varphi_0(a, b, c)) = \varphi_1(a, b + c) = a + (b + c).$$

Segundo algoritmo. En este caso $E = (a + b) + c$, que puede expresarse como el resultado de la siguiente composición de funciones:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \xrightarrow{\varphi_0} \begin{bmatrix} a+b \\ \\ c \end{bmatrix} \xrightarrow{\varphi_1} (a + b) + c,$$

donde $\varphi_0 : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ es la función definida por $\varphi_0(x, y, z) = \begin{bmatrix} x + y \\ z \end{bmatrix}$, $\varphi_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$ es la función definida por $\varphi_1(u, v) = u + v$. Se tiene $E = (a + b) + c = (\varphi_1 \circ \varphi_0)(a, b, c)$.

De manera análoga se formula el tercer algoritmo.

Consideremos el segundo algoritmo, esto es $E = (a+b)+c$. Pongamos $\widetilde{E} = fl((a+b)+c)$. Según las operaciones elementales en punto flotante, tenemos: $\eta = fl(a + b) = (a + b)(1 + \varepsilon_1)$,

$$\begin{aligned} \widetilde{E} &= fl(\eta + c) = (\eta + c)(1 + \varepsilon_2) = [(a + b)(1 + \varepsilon_1) + c](1 + \varepsilon_2) \\ &= a + b + c + (a + b)\varepsilon_1 + (a + b + c)\varepsilon_2 + (a + b)\varepsilon_1\varepsilon_2 \\ &= E + (a + b)\varepsilon_1 + y\ \varepsilon_2 + (a + b)\varepsilon_1\ \varepsilon_2. \end{aligned}$$

Luego,

$$\varepsilon_E = \frac{\widetilde{E} - E}{E} = \varepsilon_2 + \frac{a + b}{a + b + c}(1 + \varepsilon_2)\varepsilon_1 \quad \text{si } E = (a + b) + c \neq 0.$$

Puesto que $|\varepsilon_1| \leq eps$, $|\varepsilon_2| \leq eps$, se tiene que $\varepsilon_1\varepsilon_2 \simeq 0$, entonces

$$\varepsilon_E = \varepsilon_2 + \frac{a + b}{a + b + c}\varepsilon_1.$$

Para el primer y tercer algoritmos, procediendo en forma similar al segundo, se obtienen respectivamente los resultados siguientes:

$$\widetilde{\varepsilon}_E = \widetilde{\varepsilon}_2 + \frac{b + c}{a + b + c}\widetilde{\varepsilon}_1, \qquad \widehat{\varepsilon}_E = \widehat{\varepsilon}_2 + \frac{a + c}{a + b + c}\widehat{\varepsilon}_1.$$

Si a, b, c son positivos o todos negativos, los 3 algoritmos están bien condicionados. Mientras que si, por ejemplo, $a < 0$, y, b, c son positivos, la evaluación de y dependerá del algoritmo.

Sean $a = -0{,}33341 \times 10^2$, $b = 0{,}21345 \times 10^{-2}$, $c = 0{,}33456 \times 10^2$. Calculando con 5 cifras decimales de precisión, tenemos

$$\begin{aligned} \frac{a + b}{a + b + c} &= \frac{-0{,}33341 \times 10^2 + 0{,}21345 \times 10^{-2}}{-0{,}33341 \times 10^2 + 0{,}21345 \times 10^{-2} + 0{,}33456 \times 10^2} = -284{,}6, \\ \frac{b + c}{a + b + c} &= 285{,}63, \\ \frac{a + c}{a + b + c} &= 0{,}982. \end{aligned}$$

El algoritmo a elegir es $(a+c)+b$. Observe los resultados siguientes:

$$(a+b)+c = 0{,}117, \qquad (b+c)+a = 0{,}117, \qquad ,(a+c)+b = 0{,}11713.$$

Valor exacto $E = a+b+c = 0{,}1171345$.

2. De manera más general, sean $a_1, \ldots, a_n \in \mathbb{R}$ y $E = \sum_{i=1}^{n} a_i$.

El algoritmo (no eficiente) para la evaluación de E es el siguiente:

Algoritmo

Datos de entrada: n, $a_1, \cdots, a_n$.

Datos de salida: E.

1. $E = a_1$.
2. Para $k = 2, \ldots, n$

   $E = E + a_k$.

   Fin de bucle k
3. Fin.

Como en cada paso del bucle del algoritmo, se suma un dato, este procedimiento se formula usando funciones como sigue:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\varphi_1} \begin{bmatrix} a_1+a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} \xrightarrow{\varphi_2} \begin{bmatrix} a_1+a_2+a_3 \\ a_4 \\ \vdots \\ a_n \end{bmatrix} \longrightarrow \cdots \xrightarrow{\varphi_{n-2}}$$

$$\begin{bmatrix} a_1+a_2+\cdots+a_{n-1} \\ a_n \end{bmatrix} \xrightarrow{\varphi_{n-1}} \sum_{i=1}^{n} a_i,$$

donde $\varphi_1, \varphi_2, ..., \varphi_{n-1}$ se denominan funciones elementales definidas como sigue:

$$\begin{aligned} \varphi_1 &: \mathbb{R}^n \to \mathbb{R}^{n-1}, \quad \varphi_1(a_1, \ldots, a_n) = (a_1+a_2, a_3 \ldots, a_n), \\ \varphi_2 &: \mathbb{R}^{n-1} \to \mathbb{R}^{n-2}, \quad \varphi_2(x_1, \ldots, x_{n-1}) = (x_1+x_2+x_3, \ldots, x_{n-1}), \\ &\vdots \\ \varphi_{n-1} &: \mathbb{R}^2 \to \mathbb{R}, \quad \varphi_{n-1}(u,v) = u+v. \end{aligned}$$

Entonces

$$\begin{aligned} E &= \left(\varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_2 \circ \varphi_1\right)(a_1, \ldots, a_n) = \varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_2\left(\varphi_1(a_1, \ldots, a_n)\right) \\ &= \varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_2(a_1+a_2, a_3, \ldots, a_n) = \varphi_{n-1} \circ \varphi_{n-2} \circ \cdots \circ \varphi_3(a_1+a_2+a_3, a_4, \ldots, a_n) \\ &\vdots \\ &= \varphi_{n-1}\left(\varphi_{n-2}(a_1+a_2+\cdots+a_{n-2}, a_{n-1}, a_n)\right) = \varphi_{n-1}(a_1+a_2+\cdots+a_{n-1}, a_n) \\ &= \sum_{i=1}^{n} a_i. \end{aligned}$$

3. Sean $a, b \in \mathbb{R}$. Supongamos que debemos calcular $E = a^2 - b^2 = (a+b)(a-b)$. Sabemos que $a^2 - b^2 = (a+b)(a-b)$, por lo tanto E puede calcularse de dos maneras: $E = a^2 - b^2$ y $E = (a+b)(a-b)$. Podemos describir estos dos procesos de cálculo mediante funciones reales apropiadas que describan cada operación elemental que se realiza.

Primer procedimiento: el cálculo de $E = a^2 - b^2$ podemos realizarlo mediante la siguiente secuencia de funciones:

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{\varphi_0} \begin{bmatrix} a^2 \\ b^2 \end{bmatrix} \xrightarrow{\varphi_1} a^2 - b^2,$$

donde φ_0 es la función de $\mathbb{R}^2$ en sí mismo definida como $\varphi_0(a,b) = \left(a^2, b^2\right) \quad (a,b) \in \mathbb{R}^2$, φ_1 es la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida por $\varphi_1(u,v) = u - v \quad (u,v) \in \mathbb{R}^2$. Entonces, por la composición de funciones, tenemos

$$E = (\varphi_1 \circ \varphi_0)(a,b) = \varphi_1(a^2, b^2) = a^2 - b^2 \quad \forall (a,b) \in \mathbb{R}^2.$$

Segundo procedimiento: el cálculo de $E = (a+b)(a-b)$ se ejecuta mediante la aplicación de las siguientes funciones

$$\begin{bmatrix} a \\ b \end{bmatrix} \overset{\varphi_0}{\to} \begin{bmatrix} a+b \\ a-b \end{bmatrix} \overset{\varphi_1}{\to} (a+b)(a-b),$$

con φ_0 la función de $\mathbb{R}^2$ en $\mathbb{R}^2$ definida como $\varphi_0(a,b) = (a+b, a-b) \quad (a,b) \in \mathbb{R}^2$, φ_1 función de $\mathbb{R}^2$ en $\mathbb{R}$ definida $\varphi_1(x,y) = x * y \quad \forall (x,y) \in \mathbb{R}^2$. Mediante la composición de funciones se verifica inmediatamente que $E = (\varphi_1 \, o \, \varphi_o)(a,b) \quad \forall (a,b) \in \mathbb{R}^2$.

Observación. Supongamos que un problema consiste en calcular $y_1, \ldots, y_m$ a partir de datos de entrada $x_1, \ldots, x_n$. Ponemos $\overrightarrow{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$. Supongamos que existe una función $\overrightarrow{\varphi} : D \to \mathbb{R}^m$ tal que

$$\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x}) = \begin{bmatrix} \varphi_1(\overrightarrow{x}) \\ \vdots \\ \varphi_m(\overrightarrow{x}) \end{bmatrix} \quad \overrightarrow{x} \in D,$$

donde $D \subset \mathbb{R}^n$, $\varphi_j : D \to \mathbb{R}, \quad j = 1, \ldots, m$.

En cada etapa del cálculo hay un conjunto de números a operarse a partir de datos de entrada x_i, $i = 1, \ldots n$ y cada operación corresponde a la transformación del nuevo conjunto a operarse. Escribamos secuencialmente el conjunto de datos a operarse como un vector.

$$\overrightarrow{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_{n_i}^{(i)} \end{bmatrix} \in R^{n_i},$$

y asociamos la operación elemental con una función.

$$\overrightarrow{\varphi}^{(i)} : D_i \to R^{n_{i+1}}, \qquad D_i \subset R^{n_i},$$

de modo que $\overrightarrow{x}^{(i+1)} = \vec{\varphi}^{(i)}\left(\vec{x}^{(i)}\right)$, donde $\overrightarrow{x}^{(i+1)}$ es el resultado de la transformación del conjunto operado y la función $\varphi^{(i)}$ está definida de modo único salvo permutaciones en las operaciones $\overrightarrow{x}^{(i)}$ y $\overrightarrow{x}^{(i+1)}$.

Dado un algoritmo para el cálculo de $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x})$, la secuencia de operaciones elementales de la descomposición de $\overrightarrow{\varphi}$ en una secuencia de funciones elementales:

$$\begin{aligned} \overrightarrow{\varphi}^{(i)} &: D_i \to D_{i+1} \quad i = 0, 1, \ldots r, \qquad D_i \subset \mathbb{R}^{n_i}; \\ \overrightarrow{\varphi} &= \vec{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)} \circ \cdots \circ \vec{\varphi}^{(0)}, \\ D_0 &= D, \quad D_{r+1} \subset \mathbb{R}^{n_{r+1}} = R^m, \end{aligned}$$

que caracterizan al algoritmo. Así

$$\begin{aligned} \overrightarrow{y} &= \overrightarrow{\varphi}(\overrightarrow{x}) = \left(\overrightarrow{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)} \circ \cdots \circ \overrightarrow{\varphi}^{(0)}\right)(\overrightarrow{x}) = \overrightarrow{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)} \circ \cdots \circ \overrightarrow{\varphi}^{(1)}\left(\overrightarrow{\varphi}^{(0)}(\overrightarrow{x})\right) \\ &= \overrightarrow{\varphi}^{(r)} \circ \overrightarrow{\varphi}^{(r-1)}\left(\overrightarrow{x}^{(r-1)}\right) = \overrightarrow{\varphi}^{(r)}(x^r). \end{aligned}$$

Para mejor comprensión observe los ejemplos 1), 2) y 3) de esta sección.

1.11. Estabilidad numérica. Convergencia.

Sean $D \subset \mathbb{R}^n$ y $\overrightarrow{\varphi} : D \to \mathbb{R}^m$ una función. Ponemos $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x}) \quad \overrightarrow{x} \in D$. Dado un algoritmo de cómputo de $\vec{y} = \varphi(\vec{x})$, digamos

$$\overrightarrow{y} = \overrightarrow{\varphi}_{r-1} \circ \overrightarrow{\varphi}_{r-2} \circ \cdots \circ \overrightarrow{\varphi}_2 \circ \overrightarrow{\varphi}_1(\overrightarrow{x}),$$

en aritmética de punto flotante, errores en los datos de entrada y errores de redondeo en los resultados intermedios perturbarán los mismos y en consecuencia afectarán en el resultado final.

Sea $\overrightarrow{\varepsilon} \in \mathbb{R}^n$ y $\overrightarrow{x}_\varepsilon = \vec{x} + \vec{\varepsilon}$ el dato de entrada perturbado. Sea $\overrightarrow{y}_\varepsilon = \left(\overrightarrow{\varphi}_{r-1} \circ \varphi_{r-2} \circ \cdots \circ \varphi_2 \circ \overrightarrow{\varphi}_1\right)(\overrightarrow{x}_\varepsilon)$. Interesa comparar los resultados obtenidos $\overrightarrow{y} = \overrightarrow{\varphi}(\overrightarrow{x})$, y $\overrightarrow{y}_\varepsilon = \left(\overrightarrow{\varphi}_{r-1} \circ \varphi_{r-2} \circ \cdots \circ \varphi_2 \circ \overrightarrow{\varphi}_1\right)(\overrightarrow{x}_\varepsilon)$, es decir, como los errores de redondeo, de truncamiento afectan en el resultado final mediante la ejecución de la secuencia indicada $\overrightarrow{\varphi}_{r-1} \circ \varphi_{r-2} \circ \cdots \circ \varphi_2 \circ \overrightarrow{\varphi}_1$.

Definición 10 *De manera general, diremos que un algoritmo es estable numéricamente con respecto de otro si pequeñas variaciones en los datos de entrada producen pequeñas variaciones en los datos de salida. Un algoritmo será inestable si pequeñas variaciones en los datos de entrada producen grandes variaciones en los datos de salida.*

En Análisis Numérico, el estudio de la estabilidad numérica tiene mucha importancia, pués para construir un algoritmo, entre uno de los requerimientos a verificar es el de la estabilidad numérica. Si este requisito no es verificado no puede aceptarse al algoritmo como buen algoritmo y puede ser desechado.

Definición 11 *Sea V un espacio normado provisto de la norma $\| \cdot \|$. Supongamos que la solución S de un problema (P) propuesto en V se aproxima mediante un algoritmo que genera a S_n aproximación de S, $n = 1, 2, \ldots,$ en el sentido siguiente:*

$$\forall \varepsilon > 0,\ \exists n_0 \in \mathbb{Z}^+ \ \ tal\ que\ \ \forall n \geq n_0 \Longrightarrow \| S_n - S \| < \varepsilon,$$

en tal caso diremos que el algoritmo es convergente.

Dado un problema (P) y propuesto un algoritmo de solución, este debe ser bien condicionado y numéricamente estable. Si además el algoritmo genera una sucesión (S_n), debe verificarse que la sucesión (S_n) converge a S. Por lo tanto, la elaboración de un algoritmo implica el estudio del condicionamiento, estabilidad numérica y convergencia.

Notemos que un resultado importante del análisis numérico es el siguiente: si un algoritmo está bien condicionado y es numéricamente estable entonces el algoritmo es convergente. Usaremos la notación:

$$\text{condicionamiento} + \text{estabilidad} \Longrightarrow \text{convergencia}.$$

Cuando hay dos o más formas o métodos de construcción de sucesiones $(S_n^{(1)})$, $(S_n^{(2)})$ que convergen a S, es importante estudiar no solo la convergencia sino el orden de convergencia de cada método, con lo que se puede precisar las bondades y las limitaciones de cada uno de ellos.

Ejemplos

1. Sea $\{a_i \mid i = 1, \ldots, 10000\}$ un conjunto de números reales tales que $a_1 = 1, \ \ a_2 = \dfrac{1}{2}, \ldots, a_{10} = \dfrac{1}{10}$ y para $\ i = 11, \ldots, 10000, \quad \mid a_i \mid \simeq 6 \times 10^{-3}$. Sea $S = \sum_{i=1}^{10000} a_i^2$. Calcular S con una precisión de máquina $eps = 5 \times 10^{-4}$ y que se adapte a la estabilidad numérica.

   Consideremos dos algoritmos para la evaluación de S.

Primer algoritmo. Ponemos $S = 1$, y para $\begin{cases} i = 2, \ldots, 10000 \\ S = S + a_i^2. \end{cases}$ Como para $i = 11, \ldots, 10000$, $|a_i| \simeq 6 \times 10^{-3}$, entonces $a_i^2 \simeq 0{,}36 \times 10^{-4}$ siendo $eps = 5 \times 10^{-4}$ resulta que $a_i^2 \simeq 0$, con lo cual

$$\sum_{i=1}^{10000} a_i^2 = 1 + \frac{1}{2^2} + \ldots + \frac{1}{10^2} = 1{,}5498 = S_1.$$

Segundo algoritmo. Sea $|b_i| = 100\,|a_i| \simeq 6 \times 10^{-1}$, entonces $b_i^2 \simeq 0{,}36 \times 10^o > eps,\ i = 11, \ldots, 10000$. Ponemos

$$S_2 = 1 + \frac{1}{2^2} + \ldots + \frac{1}{10^2} + 10^{-4} \sum_{i=11}^{10000} (100a_i)^2.$$

Ahora bien,

$$\sum_{i=11}^{10000} (100a_i)^2 \simeq 10000 \times 0{,}36 \times 10^o = 0{,}36 \times 10^4.$$

Luego

$$S_2 = 1 + \frac{1}{2^2} + \ldots + \frac{1}{10^2} + 10^{-4} \sum_{i=11}^{10000} (100a_i)^2 \simeq 1{,}9098.$$

El primer algoritmo es numéricamente inestable, el segundo algoritmo es numéricamente estable. La razón de la inestabilidad numérica del primer algoritmo se encuentra en el cálculo de a_i^2 que está mal condicionado, pués

$$\varepsilon_{a_i^2} = \frac{a_i}{a_i^2} \times 2a_i \times \varepsilon_{a_i} = 2\varepsilon_{a_i},$$

donde ε_{a_i} es el error relativo de a_i. Así, el número de condicionamiento de cada a_i^2 es mayor que 1 lo cual amplifica el error relativo y lo vuelve inestable para a_i muy pequeño.

Determinemos el error relativo en porcentaje para cada algoritmo. Se tiene

$$S = \sum_{i=1}^{10000} a_i^2 \simeq 1{,}9094.$$

Para el primer algoritmo

$$|\varepsilon_1| = \left| \frac{S_1 - S}{S} \right| \times 100 = \frac{|1{,}5498 - 1{,}9094|}{1{,}9094} \times 100 = 18{,}8\,\%.$$

Para el segundo algoritmo

$$|\varepsilon_2| = \left| \frac{S_2 - S}{S} \right| \times 100 = \frac{|1{,}9098 - 1{,}9094|}{1{,}9094} = 0{,}00019\,\%.$$

Se observa claramente que el segundo algoritmo es mejor que el primero.

2. Se define la función real φ como $\varphi(x) = \dfrac{\operatorname{senh}(x)}{x} \quad x \neq 0$. Se desea calcular valores de $\varphi(x)$.

Primeramente, para $x \in \mathbb{R}$ tal que $|x| \geq \dfrac{1}{2}$ no se presenta ninguna dificultad en el cálculo de $\varphi(x)$. Obviamente,

$$\lim_{x \to -\infty} \frac{\operatorname{senh}(x)}{x} = \lim_{x \to \infty} \frac{\operatorname{senh}(x)}{x} = \infty.$$

Nos interesamos en calcular $\varphi(x)$ para $x \in \left]-\dfrac{1}{2}, 0\right[\cup \left]0, \dfrac{1}{2}\right[$. Con el uso de una calculadora de bolsillo, se tienen los siguientes resultados: para $x = 10^{-100}$, $\operatorname{senh}\left(10^{-100}\right) = 0$, luego

$$\varphi\left(10^{-100}\right) = \frac{\operatorname{senh}\left(10^{-100}\right)}{10^{-100}} = 0,$$

lo que es falso.

Para $x \in \left]0, \frac{1}{2}\right[$ suficientemente pequeño, podemos suponer $x < eps$, ¿cómo calcular $\varphi(x)$ si $\operatorname{senh}(x) \simeq 0$ y $x \simeq 0$? Sabemos que $\lim_{x\to\infty} \frac{\operatorname{senh}(x)}{x} = 1$. Para responder a la pregunta, recurrimos a la definición de la función seno hiperbólico y por el polinomio de Taylor con resto (véase el apéndice) se tiene para $x \in \mathbb{R}$

$$\operatorname{senh}(x) = \frac{1}{2}\left(e^x - e^{-x}\right) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + E_{2n+1}(x),$$

con $E_m(x)$ el error de aproximación del polinomio de Taylor definido como

$$E_m(x) = \frac{1}{m!}\int_0^x (x-t)^m g^{(m+1)}(t)\,dt = \frac{g^{(m_1)}(c)}{(m+1)!}x^{m+1} \quad 0 \le c < x \le \frac{1}{2},$$

y $g(x) = \operatorname{senh}(x)$, $m \in \mathbb{Z}^+$. Como $g^{(m)}(x) = \begin{cases} \operatorname{senh}(x), & \text{si } m \text{ es par,} \\ \cosh(x), & \text{si } m \text{ es impar,} \end{cases}$ entonces $\left|g^{(m+1)}(x)\right| \le \frac{1}{2}\left(e^{\frac{1}{2}} + e^{-\frac{1}{2}}\right)$ si $x \in \left[0, \frac{1}{2}\right]$ y en consecuencia

$$|E_m(x)| \le \left|\frac{g^{(m+1)}(c)}{(m+1)!}x^{m+1}\right| \le \frac{1}{2}\left(e^{\frac{1}{2}} + e^{-\frac{1}{2}}\right)\frac{x^{m+1}}{(m+1)!}.$$

Por lo tanto, para $x \in \left]0, \frac{1}{2}\right[$ se tiene

$$\varphi(x) = \frac{\operatorname{senh}(x)}{x} = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \cdots + \frac{x^{2n}}{(2n+1)!} + R_{2n}(x) \quad x \in \left]0, \frac{1}{2}\right[,$$

y

$$|R_{2n}(x)| = \left|\frac{E_{2n}(x)}{x}\right| \le \frac{1}{2}\left(e^{\frac{1}{2}} + e^{-\frac{1}{2}}\right)\frac{x^{2n}}{(2n+1)!} \quad x \in \left]0, \frac{1}{2}\right[.$$

Se supone que en una calculadora de bolsillo $10^{-100} \simeq 0$ pero 10^{-99} no se redondea por cero, se tiene

$$\varphi\left(10^{-100}\right) = 1 + R_2\left(10^{-100}\right) \simeq 1 \quad \text{pués} \quad R_2\left(10^{-100}\right) \simeq 0,$$

que es un resultado mucho más apegado a la realidad, pués $\frac{\operatorname{senh}(x)}{x} \longrightarrow 1$ cuando $x \longrightarrow 0$. De la representación de la función φ como polinomio de Taylor con resto arriba indicada, para $x = 10^{-40}$, 10^{-20}, 10^{-10}, se obtienen los siguientes resultados:

$$\begin{aligned}
\varphi\left(10^{-40}\right) &\simeq 1 + \frac{1}{6}\times 10^{-80}, \quad R_2(10^{-40}) \simeq 0,\\
\varphi\left(10^{-20}\right) &\simeq 1 + \frac{1}{6}\times 10^{-40} + \frac{1}{120}\times 10^{-80}, \quad R_4(10^{-40}) \simeq 0,\\
\varphi\left(10^{-10}\right) &\simeq 1 + \frac{1}{6}\times 10^{-20} + \frac{1}{120}\times 10^{-40} + \frac{1}{5040}\times 10^{-60} + \frac{1}{362880}\times 10^{-80}, \quad R_8(10^{-40}) \simeq 0.
\end{aligned}$$

Para $x = 0{,}005$, veamos los siguientes resultados. De la definición de φ, tenemos

$$\varphi(0{,}005) = \frac{\operatorname{senh}(0{,}005)}{0{,}005} \simeq 1{,}000004165.$$

Si utilizamos la definición de seno hiperbólico, se tiene

$$\varphi(x) = \frac{\operatorname{senh}(x)}{x} = \frac{1}{2x}\left(e^x - e^{-x}\right) \quad x \neq 0,$$

y resulta que $\varphi\,(0{,}005) \simeq 1{,}00000418$. Si

$$\varphi\,(x) = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + R_5\,(x) = 1 + \frac{x^2}{6}\left(1 + \frac{x^2}{20}\right) + R_4\,(x)\,,$$

se tiene $\varphi\,(0{,}005) \simeq 1{,}000004167$, y si

$$\varphi\,(x) = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \frac{x^6}{7!} + R_6\,(x) = 1 + \frac{x^2}{6}\left(1 + \frac{x^2}{20}\left(1 + \frac{x^2}{42}\right)\right) + R_6\,(x)\,,$$

se tiene $\varphi\,(0{,}005) \simeq 1{,}000004167$. Con una presición del orden de 10^{-10}, de estos resultados, la mejor aproximación es $\varphi\,(0{,}005) = 1{,}000004167$. Pués

$$|R_6\,(x)| \leq \frac{\left(5 \times 10^{-3}\right)^6}{7!} \simeq 0{,}31002 \times 10^{-17} < 10^{-10}.$$

Resultados similares se obtienen en el caso $x \in \left]-\frac{1}{2}, 0\right[$.

Para $x \in \left]-\frac{1}{2}, 0\right[\cup \left]0, \frac{1}{2}\right[$, se dice que el cálculo de $\varphi\,(x)$ mediante el polinomio de Taylor con error se adapta a la estabilidad numérica y por lo tanto $\varphi\,(x)$ es numéricamente estable frente a las formas de cálculo de $\varphi\,(x)$ como $\varphi\,(x) = \dfrac{\operatorname{senh}(x)}{x}$, o de $\varphi\,(x) = \dfrac{1}{2x}\left(e^x - e^{-x}\right) \quad x \neq 0$.

3. Sean $x_0, x_1, \ldots, x_n$ números de máquina positivos, cuyo número de máquina es ε. Entonces el error de redondeo relativo al calcular $\sum_{k=0}^{n} x_k$ de la manera usual es $(1+\varepsilon)^n - 1 \simeq n\,\varepsilon$.

Sea $S_n = \displaystyle\sum_{i=0}^{n} x_i$ y $\widetilde{S}_n$ el resultado de la suma en el computador. Se tiene $\begin{cases} S_o = x_o \\ S_{k+1} = S_k + x_{k+1}, \end{cases}$ y $\begin{cases} \widetilde{S}_o = x_o \\ \widetilde{S}_{k+1} = fl\left(\widetilde{S}_k + x_{k+1}\right). \end{cases}$ Definimos

$$\varepsilon_{s_k} = \frac{\widetilde{S}_k - S_k}{S_k} \quad \text{y} \quad \varepsilon_k = \frac{\widetilde{S}_{k+1} - (\widetilde{S}_k + x_{k+1})}{\widetilde{S}_k + x_{k+1}}.$$

Entonces

$$\begin{aligned} \varepsilon_{S_{k+1}} &= \frac{\widetilde{S}_{k+1} - S_{k+1}}{S_{k+1}} = \frac{(\widetilde{S}_k + x_{k+1})(1+\varepsilon_k) - (S_k + x_{k+1})}{S_{k+1}} \\ &= \frac{(S_k(1+\varepsilon_{S_k}) + (x_{k+1}))(1+\varepsilon_k) - (S_k + x_{k+1})}{S_{k+1}} \\ &= \varepsilon_k + \varepsilon_{S_k}\left(\frac{S_k}{S_{k+1}}\right)(1+\varepsilon_k). \end{aligned}$$

Puesto que $S_k < S_{k+1}$ y $|\varepsilon_k| \leq \varepsilon$, resulta

$$|\varepsilon_{S_{k+1}}| \leq \varepsilon + |\varepsilon_{S_k}|\,(1+\varepsilon) = \varepsilon + \sigma\,|\varepsilon_{S_k}|\,,$$

donde $\sigma = 1 + \varepsilon$. Se tiene

$$\begin{aligned} |\varepsilon_{So}| &= 0, \\ |\varepsilon_{S_1}| &\leq \varepsilon, \\ |\varepsilon_{s_2}| &\leq \varepsilon + \sigma\varepsilon = \varepsilon(1+\sigma), \\ \left|\varepsilon_{S_3}\right| &\leq \varepsilon + \sigma(\varepsilon + \sigma\varepsilon) = \varepsilon(1 + \sigma + \sigma^2), \\ &\vdots \\ |\varepsilon_{Sn}| &\leq \varepsilon + \sigma\varepsilon + \sigma^2\varepsilon + \ldots + \sigma^{n-1}\varepsilon = \varepsilon(1 + \sigma + \sigma^2 + \ldots + \sigma^{n-1}) \\ &= \varepsilon\frac{\sigma^n - 1}{\sigma - 1} = \varepsilon\frac{(1+\varepsilon)^n - 1}{\varepsilon} = (1+\varepsilon)^n - 1. \end{aligned}$$

Por el binomio de Newton.

$$(1+\varepsilon)^n - 1 = 1 + n\varepsilon + \frac{n(n+1)}{2!}\varepsilon^2 + \ldots + \varepsilon^n - 1 \simeq n\varepsilon.$$

4. Sea $S = \sum_{k=1}^{\infty} a_k$ una serie real convergente. Intentamos aproximar S siguiendo 2 etapas. Primero calculamos la suma parcial $S_n = \sum_{k=1}^{n} a_n$ para n grande y a continuación redondeamos S_n reteniendo una cierta cantidad de dígitos después del punto decimal. Digamos que se han retenido m dígitos. ¿Se puede asegurar que el último dígito es el correcto?.

   Sea $\widetilde{S}_n = rd(S_n)$. Deseamos que $\left|\widetilde{S}_n - S\right| \leq \frac{1}{2} \times 10^{-m}$. Si S_n se ha redondeado correctamente para obtener $\widetilde{S}_n$, entonces $\left|\widetilde{S} - S_n\right| \leq \dfrac{1}{2} \times 10^{-m}$. Pero

   $$\left|\widetilde{S} - S\right| \leq \left|\widetilde{S} - S_n\right| + |S_n - S| \leq \frac{1}{2} \times 10^{-m} + |S_n - S|\,.$$

   Esta desigualdad no se puede mejorar a menos que $S_n = S$, o también $a_k = a$, $k = 1, \ldots, n$, por tanto no se puede lograr $\left|\widetilde{S} - S\right| \leq \dfrac{1}{2} \times 10^{-m}$. Si imponemos que $\left|\widetilde{S} - S\right| < \dfrac{6}{10} \times 10^{-m}$, entonces

   $$\frac{1}{2} \times 10^{-m} + |S_n - S| < \frac{6}{10} \times 10^{-m}.$$

   de donde $|S_n - S| < 10^{-m-1}$. Luego $\left|\sum_{k=n+1}^{\infty} a_k\right| < 10^{-m-1}$.

5. Sea $E(n) = \int_0^1 x^n e^{x-1} dx,\ n = 0, 1, 2, \ldots$.Para cada n, se desea calcular valores aproximados de $E(n)$. Con este propósito se deben elaborar algoritmos para aproximar $E(n)$. Con este ejemplo se obtendrán un algoritmo mal condicionado y otro bien condicionado, numéricamente estable y convergente.

   Primeramente analicemos el problema.

   Sea $f_n(x) = x^n e^{x-1} \quad x \in [0,1], \quad n = 0, 1, \cdots$. Se tiene que $f_n(x) \geq 0 \quad \forall x \in [0,1], \quad \forall n = 0, 1, 2, \ldots,$ y como la integral de una función no negativa es no negativa, se sigue que $E(n) = \int_0^1 f_n(x)dx \geq 0$. Por otra parte, si $0 < x < 1, \quad x^n \underset{n\to\infty}{\longrightarrow} 0$, luego $f_n(x) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall x \in [0,1[$, entonces

   $$\lim_{n\to\infty} E(n) = \lim_{n\to\infty} \int_0^1 x^n e^{x-1} = 0.$$

   En conclusión $E(n)$ es no negativo, $E(n+1) < E(n) \quad$ con $n = 0, 1, \ldots$ que muestra que $E(n)$ es decreciente y acotada por 0.

   Primer algoritmo. Para $n = 0$, se tiene

   $$E(0) = \int_0^1 e^{x-1} dx = e^{x-1}\,|_0^1 = 1 - e^{-1} = 0{,}6321205588\ldots.$$

   Para $n > 0$, aplicamos el método de integración por partes, se obtiene

   $$E(n) = x^n e^{x-1}\,|_0^1 - n\int_0^1 x^{n-1} e^{x-1} dx = 1 - n\ E(n-1), \quad n = 1, 2, \ldots,$$

   Así, se obtiene la ecuación recurrente siguiente: $E(n) = 1 - n\ E(n-1) \quad n = 1, 2, \cdots$ .Conocido un valor aproximado de $E(n-1)$, mediante la ecuación recurrente podemos calcular un valor

aproximado de $E(n)$ lo que nos permite obtener el siguiente algoritmo de cálculo para $E(n)$ para $n = 0, 1, \cdots, N$.

Algoritmo

Datos de entrada: N.

Datos de salida: $E(n)$.

1. $E(0) = 1 - e^{-1} = 0{,}6321205588\ldots$,

2. Para $n = 1, \ldots, N$

$$E(n) = 1 - nE(n-1)$$

Fin de bucle n.

3. Imprimir $E(n)$.

4. Fin.

Con precisión de 5 cifras decimales, los resultados de la aplicación del algoritmo precedente se muestran en la siguiente tabla:

	n	$E(n)$
	0	0,63212
	1	0,36788
	2	0,26424
	3	0,20728
	$\vdots$	
	12	0,05809
$*$	13	0,24478
$*$	14	$-2{,}42688$
$*$	15	37,40316
	$\vdots$	
$*$	20	$-69'478,033{,}14.$

Observe los valores señalados con $*$. El análisis de $E(n)$ muestra que $E(n) \geq 0$ y $E(n) \longrightarrow 0$ cuando $n \to \infty$. A partir de $n = 13$ los resultados son absurdos. Si se realizan los cálculos con un número mayor de cifras decimales, los resultados absurdos se obtienen para n>13.

Note que tomando $\widetilde{E} = rd(E(0)) = 0{,}63212$ como dato de entrada, el error de redondeo en cada iteración es amplificado por $-n$. Estos hechos demuestran que el algoritmo está mal condicionado. Se puede demostrar que el número de condicionamiento de este procedimiento es $C = -n$, y en consecuencia pequeños errores en los datos de entrada provocan grandes errores en los datos de salida, lo que muestra que el algoritmo es inestable numéricamente.

Segundo Algoritmo. Tomando en cuenta que $E(n) > 0, \quad n = 0, 1, \ldots,$ y, $E(n) \longrightarrow 0$ cuando $n \to \infty$, basta elegir n suficientemente grande para obtener $E(n+1) \simeq 0$, entonces

$$0 \simeq E(n+1) = 1 - (n+1)E(n),$$

de donde

$$E(n) = \frac{1}{n+1},$$

y para $n-1, n-2, \ldots, 1$, tenemos

$$E(n-1) = \frac{1 - E(n)}{n}.$$

Para N suficientemente grande, se establece el algoritmo siguiente.

Algoritmo

Datos de entrada: N.

Datos de salida: $E(n)$.

1. $E(N) = \dfrac{1}{N+1}$.

2. Para $n = N-1, \cdots, 1$

$$E(n) = \frac{1 - E(n)}{n}$$

Fin de bucle n.

3. Imprimir $E(n)$.

4. Fin.

Para $N = 20$, en la tabla siguiente se muestran los resultados de la aplicación del algoritmo precedente.

n	$E(n)$
20	0,04762
19	0,04762
18	0,05130
17	0,05277
$\vdots$	
5	0,14553
4	0,17089
3	0,20728
2	0,26424
1	0,36788.

Estos resultados son satisfactorios. Este algoritmo está bien condicionado y los pequños erreres en los datos de entrada provocan pequeños errores en los datos de salida, es decir que el algoritmo es numéricamente estable, no obstante el algoritmo presenta un inconveniente: el número de operaciones que se requiere para calcular $E(n_j)$ a partir de $E(N)$ con una precisión fijada debe ser grande.

Tercer algoritmo. Por la serie de Taylor de e^x se tiene $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad x \in [0,1]$, y por el teorema de la convergencia uniforme y la integración (véase el capítulo 3 donde se tratan las sucesiones y series de funciones), resulta

$$\begin{aligned} E(n) &= \int_0^1 x^n e^{x-1} dx = e^{-1} \int_0^1 x^n e^x dx = e^{-1} \int_0^1 x^n \left(\sum_{k=0}^{\infty} \frac{x^k}{k!} \right) dx \\ &= e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!} \int_0^1 x^{n+k} dx = e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!(n+k+1)}. \end{aligned}$$

Así, $E(n) = e^{-1} \sum_{k=0}^{\infty} \dfrac{1}{k!(n+k+1)} \qquad n = 0, 1, \cdots$. Vemos que $E(n)$ se representa como una serie numérica convergente. Lamentablemente la serie no puede ser evaluada en el computador, necesitamos transformarla en una suma finita $S_{m_0}(n)$. Para el efecto, como la serie $\displaystyle\sum_{k=0}^{\infty} \frac{1}{k!(n+k+1)}$ es convergente, entonces

$$\forall \varepsilon > 0, \ \exists m_0 \in \mathbb{Z}^+ \text{ tal que } \forall m \geq m_0 \Rightarrow \left| \sum_{k=0}^{\infty} \frac{1}{k!(n+k+1)} - \sum_{k=0}^{m} \frac{1}{k!(n+k+1)} \right| < \varepsilon,$$

o bien $\displaystyle\sum_{k=m_0+1}^{\infty} \frac{1}{k!(n+k+1)} < \varepsilon.$

Sea (a_k) una sucesión de números positivos tal que $\sum_{k=1}^{\infty} a_k = 1$. Determinemos m_0 tal que

$$\frac{\frac{1}{k!(n+k+1)}}{a_k} < \varepsilon \quad \text{si} \quad k \geq m_0.$$

Pongamos $a_k = \dfrac{1}{k(k+1)}$. Se tiene $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$. Entonces $\frac{k(k+1)}{k!(n+k+1)} \leq \varepsilon$. Sea $\varepsilon = 10^{-6}$, determinemos m_0 tal que $\frac{1}{(k-1)!} < 10^{-6}$. Esta última desigualdad se verifica para todo $k \geq 11$, luego $m_0 = 11$. Así

$$S_{m_0}(n) = e^{-1} \sum_{k=0}^{11} \frac{1}{k!(n+k+1)}.$$

De este modo $E(n)$ es aproximado con la suma $S_{m_0}(n)$ con una precisión de 10^{-6}. Cada término de la suma está bien condicionada, escribimos $S_{m_o}(n)$ de la manera siguiente:

$$\begin{aligned} S_{mo}(n) &= e^{-1}\left(\frac{1}{n+1} + \frac{1}{1!(n+2)} + \frac{1}{2!(n+3)} + \frac{1}{3!(n+4)} + \ldots + \frac{1}{10!(n+11)} + + \frac{1}{11!(n+12)}\right) \\ &= e^{-1}\left(\frac{1}{n+1} + \frac{1}{n+2} + \frac{1}{2}\left(\frac{1}{n+3} + \frac{1}{3}\left(\frac{1}{n+4} + \ldots + \frac{1}{10}\left(\frac{1}{n+11} + \right.\right.\right.\right. \\ &\quad \left.\left.\left.\left. + \frac{1}{11 \times (n+12)}\right)\ldots\right)\right)\right). \end{aligned}$$

Tenemos así el siguiente algoritmo siguiente:

Algoritmo

Datos de entrada: n.

Datos de salida: $E(n)$.

1. $s = \dfrac{1}{11 \times (n+12)}$
2. Para $k = 1, \cdots, 12$

   $j = 12 - k$

   $s = \dfrac{1}{n+j} + \dfrac{1}{j}s$

   Fin de bucle k.
3. $E(n) = s$.

4-Imprimir E(n).

5. Fin.

En la tabla siguiente se muestran los resultados de la aplicación de este algoritmo.

n	$E(n) \simeq S_{mo}(n)$
1	0,367879
2	0,264241
3	0,207276
4	0,170893
$\vdots$	
100	0,009805
$\vdots$	
500	0,001992
$\vdots$	
1000	0,000998
$\vdots$	
1000000	0,000009999.

Para cada n, $S_{m_0}(n)$ requiere de 59 operaciones elementales. Este algoritmo reune todas las características: condicionamiento, estabilidad, convergencia. Además es fácil de programar y cada $S_{m_o}(n)$ es independiente del cálculo de $S_{m_o}(n-1)$ o de $S_{mo}(n+1)$. En consecuencia este es uno de los mejores algoritmos que puede construirse para aproximar $E(n)$, $\quad n = 1, 2, \ldots$.

6. Ejemplo de un método convergente.

   Consideremos como problema (P) el cálculo de $a^{1/n}$, donde $a \in \mathbb{R}$, $\quad n \in \mathbb{N}$ con $n \geq 2$.

   Notemos primeramente que la raíz $n-$ésima de a está bien definida para todo a si n es impar, y $a \geq 0$ si n es par.

   Supongamos $a \geq 0$, $\quad n \geq 2$. Definimos la función f de $\mathbb{R}^+$ en $\mathbb{R}$, por $\quad f(x) = x^n - a \quad x \in \mathbb{R}^+$. Tenemos la siguiente equivalencia: $f(x) = 0 \Leftrightarrow x = a^{1/n}$, es decir que la ecuación $f(x) = 0$ tiene una única solución $x = a^{1/n} \in \mathbb{R}^+$. Apliquemos el método de Newton cuya interpretación geométrica indicamos a continuación.

   Sea $x_0 \in \mathbb{R}^+$ una aproximación de $a^{1/n}$. La ecuación de la recta tangente L_1 a la gráfica de f en el punto $(x_0,\ f(x_0))$ viene dada por: $\quad y - f(x_0) = f'(x_0)(x - x_0)$. Esta recta corta al eje X en el punto $(x_1, 0)$, en tal caso tenemos $y = 0$, y

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

   Repetimos nuevamente el proceso descrito previamente. La ecuación de la recta tangente L_2 a la gráfica de f en el punto $(x_1,\ f(x_1))$ es: $\ y - f(x_1) = f'(x_1)(x - x_1)$, que corta al eje X en el punto $(x_2, 0)$. Obtenemos entonces

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

   Continuando con este procedimiento k veces, deducimos que

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad k = 0, 1, 2, \ldots, .f'(x_k) \neq 0.$$

   Este último esquema numérico se conoce con el nombre de método de Newton. En el capítulo resolución numérica de ecuaciones no lineales se aborda con más detalle este método.

   En la siguiente figura se ilustran las rectas tangentes L_1, L_2, L_3 a la gráfica de f en los puntos $(x_0,\ f(x_0))$, $(x_1,\ f(x_1))$ y $(x_2,\ f(x_2))$. Note que la gráfica de la función f corta al eje X en $x = a^{1/n}$.

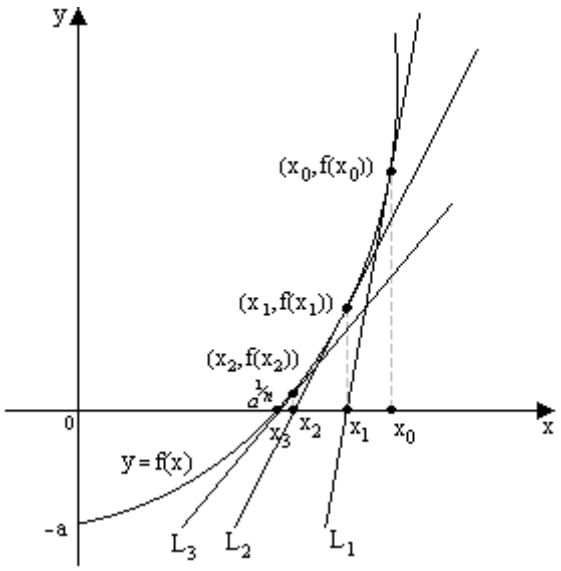


Figura 10

De acuerdo a la definición de la función f y al esquema de Newton arriba obtenido, tenemos

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^n - a}{n\, x_k^{n-1}} = \frac{(n-1)x_k^n + a}{n\, x_k^{n-1}} = \frac{1}{n}\left((n-1)\, x_k + \frac{a}{x_k^{n-1}}\right), \quad k = 0, 1, \ldots,$$

al que nos referiremos como esquema numérico para aproximar $a^{1/n}$.

Si fijamos el número entero m, número de iteraciones a realizar, el procedimiento de cálculo de $a^{1/n}$ es el siguiente:

$$\begin{cases} x_0 > 0 \ \text{ dado}, \\ x_{k+1} = \dfrac{1}{n}\left((n-1)x_k + \dfrac{a}{x_k^{n-1}}\right) \qquad k = 0, 1, \ldots, m. \end{cases}$$

Este algoritmo requiere de la lectura de los siguientes datos de entrada: n, a, m. Como resultado obtendremos el valor aproximado de $a^{1/n}$. Los datos de salida son : a, n, $a^{1/n}$. Se puede probar que dada una aproximación inicial $x_0 > 0$ apropiada de $a^{1/n}$, la sucesión (x_k) generada por el esquema numérico para aproximar $a^{1/n}$, converge efectivamente al valor de $a^{1/n}$. En estas condiciones proponemos un primer algoritmo que es un tanto incompleto como se verá más adelante.

Algoritmo 1.

Datos de entrada: a, n, m.

Datos de salida: $a^{1/n}$.

1. $x = x_0$.
2. Para $k = 0, 1, \ldots, m$

$$x_k = \frac{1}{n}\left((n-1)x + \frac{a}{x^{k-1}}\right)$$
$$x = x_k$$

Fin de bucle k.

3. Imprimir $x = a^{1/n}$.
4. Fin.

Así por ejemplo si $a = 2, \quad n = 2$, el procedimiento para calcular $\sqrt{2}$ es el siguiente:

$$\begin{cases} x_0 > 0 \quad \text{dado}, \\ x_{k+1} = \dfrac{1}{2}\left(x_k + \dfrac{2}{x_k}\right) \qquad k = 0, 1, \ldots \end{cases}$$

Si fijamos el número de iteraciones $m = 5$ y el punto inicial $x_0 = 2$, la aplicación del algoritmo de cálculo de $\sqrt{2}$ nos da los siguientes resultados:

$$\begin{aligned} x_1 &= \frac{1}{2}\left(x_0 + \frac{2}{x_0}\right) = 1{,}5, \qquad x_2 = \frac{1}{2}\left(x_1 + \frac{2}{x_1}\right) = 1{,}4166667, \\ x_3 &= \frac{1}{2}\left(x_2 + \frac{2}{x_2}\right) = 1{,}414215687, \qquad x_4 = \frac{1}{2}\left(x_3 + \frac{2}{x_3}\right) = 1{,}414213563, \\ x_5 &= \frac{1}{2}\left(x_4 + \frac{2}{x_4}\right) = 1{,}414213563. \end{aligned}$$

El valor exacto es $\sqrt{2}$ y el obtenido en una calculadora de bolsillo es 1,414213562 . . . . En este ejemplo vemos las siguientes características: el procedimiento de cálculo descrito en el algoritmo 1 tiene una estructura bien definida, el número de repeticiones de concluye en m pasos, el procedimiento permite aproximar $\sqrt{2}$ con una precisión de 10^{-10}. El algoritmo 1 presenta un inconveniente que es la selección del punto inicial x_0 del que se ha dicho debe ser una aproximación inicial apropiada de $a^{1/n}$. No obstante, para a suficiente grande queda la duda de como elegir dicho punto. Un procedimiento de selección es el siguiente:

a) Para $0 < a < 1$, obtenemos el siguiente esquema numérico:

$$\begin{cases} x_0 = 1 \\ x_{k+1} = \dfrac{1}{n}\left((n-1)x_k + \dfrac{a}{x_k^{n-1}}\right) \end{cases} \quad k = 0, 1, \ldots, m.$$

b) Supongamos que $a > 1$. Sea $j \in \mathbb{N}$ el más pequeño entero tal que $0 < 10^{-jn}a < 1$. Entonces

$$f(x) = x^n - a = 10^{jn}\left(\frac{x^n}{10^{jn}} - \frac{a}{10^{jn}}\right) = 10^{jn}\left(\left(\frac{x}{10^j}\right)^n - \frac{a}{10^{jn}}\right).$$

Ponemos $b = a \times 10^{-jn}$, $t = 10^{-j} \times x$ y definimos $g(t) = t^n - b$. Resulta

$$g(t) = 0 \Longleftrightarrow t = b^{1/n} \Longleftrightarrow 10^{-j}\, x = 10^{-j}a^{1/n} \Longleftrightarrow x = a^{1/n}.$$

Así, $g(t) = 0 \Leftrightarrow x = a^{1/n}$, que muestra que la raíz de la ecuación $g(t) = 0$ es la misma de $f(x) = 0$.

El algoritmo descrito precedentemente puede ser aplicado a condición de reemplazar a por b. Así por ejemplo, sea $a = 36254932{,}65$ y $n = 4$. Resulta que $a = 0{,}3625493265 \times 10^8$, $j = 2$ y $b = 0{,}3625493265$. Entonces

$$\begin{cases} t_0 = 1, \\ t_{k+1} = \dfrac{1}{4}\left(3t_k + \dfrac{0{,}3625493265}{t_k^3}\right) \end{cases} \quad k = 0, 1, \ldots, 5.$$

Tenemos

$$\begin{aligned} t_1 &= 0{,}8406373318, \qquad t_2 = 0{,}783052196, \qquad t_3 = 0{,}7760600163, \\ t_4 &= 0{,}7759643795, \qquad t_5 = 0{,}775964362. \end{aligned}$$

Luego $x = 0{,}775964362 \times 10^2$, con lo cual $a^{1/4}$ se aproxima por 77,5964362 con una precisión $\varepsilon = 10^{-8}$. El valor obtenido de una calculadora de bolsillo es $(36'254,932{,}65)^{1/4} = 77{,}59643619\ldots$.

c) Finalmente, si $a < 0$ y n es impar ponemos $c = -a$ y aplicamos los resultados descritos precedentemente en *a*) y *b*).

Para elaborar un algoritmo completo de cálculo de $a^{1/n}$ introducimos dos variables *indi* e $info$ y que toman los valores 0 y 1. La variable *indi* lo utilizamos para la paridad de n, esto es, n par entonces $indi = 0$, n impar entonces $indi = 1$. La variable $info$ es utilizada para el signo de a, así: $a > 0$ entonces ínfo= 0, $a < 0$ entonces $info = 1$.

Algoritmo 2

Datos de entrada: a, n, m.

Datos de salida: $S = a^{1/n}$.

1. $\begin{cases} \text{Si } n \text{ par, hacer } indi = 0, \\ \text{Si } n \text{ impar, hacer } indi = 1. \end{cases}$
2. $\begin{cases} \text{Si } a > 0, \text{ hacer } info = 0, \\ \text{Si } a < 0, \text{ hacer } info = 1. \end{cases}$
3. Si $indi = 0$ e $info = 1$, Imprimir, "Error". Continuar en 11.
4. Si $indi = 1$ e $info = 1$. Hacer $c = a$. Continuar en 6.
5. Poner $c = a$.
6. Determinar $j \in \mathbb{N}$ el más pequeño tal que $b = 10^{-jn}c < 1$.
7. Poner $x = 1$.
8. Para $k = 1, \ldots, m$
   $x_k = \dfrac{1}{n}\left((n-1)x + \dfrac{b}{x^{n-1}}\right)$
   $x = x_k$.
   Fin bucle k
9. Si $indi = 1$ e $info = 1$. Hacer $x = -x_k$. Poner $S = 10^j x$.
10. Imprimir resultados: S.
11. Fin.

7. Ejemplo de un método numérico impracticable.

   Sean $a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}$. Supongamos que $a_1,\ a_2,\ b_1,\ b_2$ son no nulos y la matriz $A = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$ es invertible. Consideramos el sistema de ecuaciones lineales: $(x, y) \in \mathbb{R}^2$ tal que $\begin{cases} a_1x + b_1y = c_1, \\ a_2x + b_2y = c_2. \end{cases}$ Puesto que A es invertible, este sistema de ecuaciones tiene solución única $(x, y) \in \mathbb{R}^2$. Calculemos esta solución. Para el efecto se disponen de dos métodos. El primero es el conocido método de Cramer cuya solución se calculan como se muestra a continuación

$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}, \qquad y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}},$$

   donde $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$ denota el determinante de la matriz real $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$. El segundo método que consideramos es el de eliminación gaussiana que se indica a continuación: se calcula $k = -\dfrac{a_2}{a_1}$, y se obtiene el sistema de ecuaciones lineales triangular superior siguiente: $\begin{cases} a_1x + b_1y = c_1 \\ (b_2 + kb_1)\, y = c_2 + kc_1, \end{cases}$ cuya solución se calcula como sigue: $\begin{cases} y = \dfrac{c_2 + kc_1}{b_2 + kb_1}, \\ x = \dfrac{1}{a_1}\,(c_1 - b_1y)\,. \end{cases}$ Contabilicemos el número de operaciones que se realizan con cada método. Con el método de Cramer, el cálculo del determinante $\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix} = c_1b_2 - c_2b_1$ implica tres operaciones elementales. Como se deben calcular tres determinantes y dos cocientes, resultan 11 operaciones elementales. Con el método de eliminación gaussiana se tienen las siguientes operaciones elementales: en el cálculo de k se realiza un cociente, el cálculo de y implica 5 operaciones elementales y el de x implica 3 operaciones elementales. En total se requieren de 9 operaciones elementales.

   A juzgar por el número de operaciones elementales, el método de eliminación gaussiana realiza 2 operaciones elementales menos que en el de Cramer. A más de esta razón, el método de eliminación gaussiana es mucho más estable numéricamente. En conclusión, para resolver numéricamente un sistema lineal de dos ecuaciones con dos incógnitas, se debe aplicar el método de eliminación gaussiana.

En lo sucesivo consideraremos sistemas de ecuaciones lineales $A\widehat{x} = \widehat{b}$, donde $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$ y $\overrightarrow{b} \in \mathbb{R}^n$.

Llamamos método directo de resolución del sistema de ecuaciones lineales, un método que conduce a la solución del problema al cabo de un número finito de pasos, o bien en un número finito de operaciones aritméticas (suma, resta, multiplicación y división) que es función de la dimensión del sistema. Para cada método directo estudiado, se debe estimar:

i) El número de operaciones elementales necesarias en la ejecución del algoritmo, es decir que se debe determinar una función N_{oper}: $\mathbb{Z}^+ \to \mathbb{R}$ que a cada $n \in \mathbb{Z}^+$ asocie $N_{oper}\ (n)$.

ii) Precisión del método. Esta precisión depende sobre todo del condicionamiento de la matriz y de la estabilidad del método, es decir que pequeños errores en los datos de entrada provocan pequeños errores en los datos de salida, o lo que es lo mismo, es insensible a la propagación de errores de redondeo.

Supongamos que para resolver el sistema de ecuaciones lineales utilizamos la regla de Cramer:

$$x_i = \frac{\Delta_i}{\det(A)} \qquad i = 1, \ldots, n,$$

donde Δ_i es el determinante de la matriz obtenida al reemplazar la columna i-ésima de A por $\overrightarrow{b}$, $\det(A) \neq 0$.

Estimemos el número de operaciones elementales que se requieren para el cálculo de un determinante de una matriz C de $n \times n$. Para el efecto, determinemos el número de operaciones elementales que se requieren para calcular determinantes de orden 2 y 3.

Para calcular el determinante de una matriz C de 2×2 se efectúan las siguientes operaciones:

$$\text{productos}: \ 2 = 2! \times 1, \qquad \text{adiciones}: \ 1 = 2! - 1,$$

luego $N_{oper}(2) = 3$ operaciones elementales. Si C es una matriz real de 3×3, $C = (C_{ij})_{3x3}$, entonces

$$\det(C) = C_{11} \begin{vmatrix} C_{22} & C_{23} \\ c_{32} & C_{33} \end{vmatrix} - C_{12} \begin{vmatrix} C_{21} & C_{23} \\ C_{31} & C_{32} \end{vmatrix} + C_{13} \begin{vmatrix} C_{21} & C_{22} \\ C_{31} & C_{32} \end{vmatrix},$$

y el número de operaciones elementales se obtiene del modo siguiente: el cálculo de cada determinante de 2×2 requiere de 3 operaciones elementales, de la descomposición precedente, se obtiene

$$\text{multiplicaciones}: 9 = 3 \times 2 + 3, \qquad \text{sumas}: \ 5 = 3 \times 1 + 2,$$

con lo que $N_{oper}(3) = 14$ operaciones elementales.

En general, si C es una matriz de $n \times n$, se tiene

$$\text{multiplicaciones}: \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k), \qquad \text{sumas}: \ n! - 1.$$

El número de operaciones elementales aplicando el método de Cramer es:

$$\begin{aligned} \text{multiplicaciones} &: (n+1) \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k), \\ \text{sumas} &: (n+1)(n! - 1), \\ \text{divisiones} &: n, \end{aligned}$$

con lo cual

$$Noper(n) = n + (n+1)(n! - 1) + (n+1) \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k) = (n+1)! + (n+1) \sum_{j=1}^{n-1} \prod_{k=1}^{j} (n + 1 - k).$$

Así, $Noper(5) = 330$ operaciones elementales, $Noper(6)$=1961 operaciones elementales. Note el tiempo que se requeriría para resolver un sistema de ecuaciones lineales de 5×5 usando una calculadora de bosillo: aproximadamente medio minuto por operación implica aproximadamente 165 minutos el tiempo requerido para resolver dicho sistema de ecuaciones, ¿cuánto tarda usted en resolver un tal sistema? Si despreciamos los $n - 2$ términos del sumatorio, tenemos $N = 2(n + 1)!$ y para $n = 20$, se obtiene $N \simeq 1{,}021818893 \times 10^{20} < Noper(20)$ que muestra que este método es impracticable. Con otros métodos, un sistema de ecuaciones lineales de 5×5 y con el uso de una calculadora de bolsillo y con el tiempo estimado de medio minuto por operación, se requerirá aproximadamente una hora; un sistema de ecuaciones lineales de 20×20 y con el uso de los computadores actuales requerirá de fracciones de segundo.

Por otra parte, las operaciones sumas y restas alternadas incrementan los errores de redondeo, que a su vez deterioran la calidad de la solución. Más aún, cuando n es demasiado grande, a causa de los errores de redondeo, puede provocarse un overflow lo que a su vez provocará una detención en la ejecución del programa. Por estas razones, el cálculo del determinante mediante este procedimiento definitivamente es impracticable, pués es mal condicionado e inestable numéricamente. Consecuentenemente, para el cálculo del determinante de una matriz debe aplicarse otros métodos y algoritmos que son relativamente económicos y fáciles de programarse e implementarse en un PC.

En conclusión, si se utiliza la regla de Cramer para hallar la solución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, del punto de vista numérico es impracticable.

Si A es una matriz invertible, la solución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ tiene una única solución

$$\overrightarrow{x} = A^{-1}\overrightarrow{b},$$

donde $A^{-1} = \frac{1}{\det(A)}(A^D)^t$ y $A^D = [(-1)^{i+j} \text{ menor } (a_{ij})] \;\; i = 1, ..., n, \;\; j = 1, ..., n.$

El cálculo de la matriz A^D implica el cálculo de n^2 determinantes de matrices de $(n-1) \times (n-1)$. Adicionamos a esto el cálculo de $\det(A)$ y a continuación el producto de A^{-1}por $\overrightarrow{b}$. Mediante un razonamiento similar al precedente se puede mostrar que el número de operaciones elementales $Noper(n)$ es muy grande, con lo cual este método es igualmente impracticable. Más aún, si se toma en consideración los errores de redondeo, estos pueden ser muy grandes lo que conducirá a resultados completamente distorsionados. En definitiva, se trata de un método mal condicionado e inestable numéricamente, por lo tanto inutilizable del punto de vista numérico. Más adelante se tratan métodos directos de resolución de sistemas de ecuaciones que son fáciles de aplicarse con un número de operaciones $Noper(n)$ muy razonable.

1.12. Ejercicios

1. Sean T un triángulo cuyos vértices son $\overrightarrow{u}_1 = (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_3 = (x_3, y_3) \in \mathbb{R}^2$ que suponemos son no colineales y distintos, y, un punto dado $\overrightarrow{x} = (a, b) \in \mathbb{R}^2$. Se considera el siguiente problema: determinar si $(a, b) \in T$ o $(a, b) \notin T$. Fundamentar matemáticamente la solución del problema y elaborar un algoritmo numérico. Determine el número de operaciones elementales, asignaciones, comparciones. Realice comprobaciones de su algoritmo.

2. Se considera un triángulo T cuyos vértices son $\overrightarrow{u}_1 (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_3 = (x_3, y_3) \in \mathbb{R}^2$. Suponemos que $\overrightarrow{u}_1$, $\overrightarrow{u}_2$, $\overrightarrow{u}_3$ son distintos y no colineales. Elaborar un algoritmo que permita calcular su perímetro y su área. Recuerde que si $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^2$, la métrica ecuclídea d está definida como $d(\overrightarrow{x}, \overrightarrow{y}) = \|\overrightarrow{x} - \overrightarrow{y}\|$ y $\|\overrightarrow{x}\| = \left(\overrightarrow{x}^T \overrightarrow{x}\right)^{\frac{1}{2}}$. Determine el número de operaciones elementales. Realice comprobaciones de su algoritmo.

3. Sean $\overrightarrow{u}_1 = (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_2 = (x_3, y_3) \in \mathbb{R}^2$ los vértices de un triángulo T. Supongamos que $\overrightarrow{u}_1$, $\overrightarrow{u}_2$, $\overrightarrow{u}_3$ son distintos y no colineales. Elaborar un algoritmo que permita calcular los ángulos interiores del triángulo T y determinar si T es un triángulo rectángulo, isósceles o escaleno. Calcule el número de operaciones elementales y de comprobaciones.

4. Se consideran $\overrightarrow{u}_1 = (x_1, y_1)$, $\overrightarrow{u}_2 = (x_2, y_2)$, $\overrightarrow{u}_3 = (x_3, y_3)$, $\overrightarrow{u}_4 = (x_4, y_4)$ puntos de $\mathbb{R}^2$ dados. Suponemos que dichos puntos son distintos y al menos tres de ellos no son colineales. Elabore un algoritmo que permita identificar si el cuadrilátero es un paralelogramo y en este caso identificar si es un rectángulo. Además, se debe calcular el área de dicho cuadrilátero. Determine el número de operaciones elementales, asignaciones y comprobaciones. Realice pruebas para verificar su algoritmo.

5. Sean $a, b, c, d \in \mathbb{R}$ y $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ una matriz invertible. Obviamente $ad - bc \neq 0$ y $A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

   Ponemos $A^{-1} = \begin{bmatrix} p & r \\ q & s \end{bmatrix}$. Note que $\begin{bmatrix} p & r \\ q & s \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} p \\ q \end{bmatrix}$, $\begin{bmatrix} p & r \\ q & s \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} r \\ s \end{bmatrix}$, o sea $\begin{bmatrix} p \\ q \end{bmatrix}$ es la solución del sistema de ecuaciones $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ y $\begin{bmatrix} r \\ s \end{bmatrix}$ es la solución del sistema $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$de determinan las columnas de A^{-1}. Elabore un algoritmo que resuelva los dos sistemas de ecuaciones lineales de modo que el número de opreaciones elementales sea el más pequeño posible y escriba A^{-1}. Compruebe con las siguientes matrices:

a) $\begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$. **b)** $\begin{bmatrix} 3 & -2 \\ 0 & 8 \end{bmatrix}$. **c)** $\begin{bmatrix} 1 & 2 \\ 5 & 2 \end{bmatrix}$. **d)** $\begin{bmatrix} 1 & \sqrt{2} \\ 5\sqrt{3} & 2\sqrt{5} \end{bmatrix}$.

6. Sean A, B, C matrices reales de 2×2. En cada item se define una matriz D, elabore un algoritmo para calcular la matriz D.

   a) $D = A\,(B + C)$. **b)** $D = AB - C$. **c)** $D = (A - B)\,C + I$ con I la matriz identidad.

   d) $D = C\,(B - A)\,C$. **e)** $D = B\left(I + A + A^2 + A^3 + A^4\right)C$. **f)** $D = B\left(I - A + A^2 - A^3 + A^4\right)C$.

   Compruebe cada algoritmo con las siguientes matrices:

   $A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$, $C = \begin{bmatrix} 3 & 5 \\ -2 & 1 \end{bmatrix}$.

7. **a)** Sea A una matriz real no nula de $m \times m$. Se define $A^{-1} = A$ y $A^{n+1} = A^n A$ para $n \in \mathbb{N}$. Elabore un algoritmo que permita calcular A^n.

   b) Verifique su algoritmo con $n = 3$ y la matriz siguiente $A = \begin{bmatrix} 1 & \frac{1}{2} \\ -\frac{1}{3} & 1 \end{bmatrix}$.

   c) Si $A = \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ \frac{1}{2} & 2 & 0 \\ 1 & 1 & 3 \end{bmatrix}$, aplique su algoritmo y calcule A^3.

8. Aplique el método de eliminación gaussiana con pivoting parcial para hallar, si existe, la solución de cada uno de los sistemas de ecuaciones lineales que se proponen. En caso de calcular la solución, compruebe. De no ser posible, indique si el sistema de ecuaciones tiene infinitas soluciones o ninguna solución.

   a) $\begin{cases} 3x + y - z = 5 \\ x + 2y = 8 \\ y - 2z = -5. \end{cases}$ **b)** $\begin{cases} x + 2y + 3z = -2 \\ -x + y + z = -1 \\ 2x + 3y = -3. \end{cases}$ **c)** $\begin{cases} 4x + y - z = -5 \\ 8x + 2z = -6 \\ x + y = -1. \end{cases}$

   d) $\begin{cases} 0{,}2x + 0{,}3y + 0{,}4z = 0{,}9 \\ -0{,}1x + 0{,}1y + 0{,}2z = 0{,}2 \\ 1{,}1x + 0{,}2y - 2z = -0{,}7. \end{cases}$ **e)** $\begin{cases} y + 2z = -1 \\ 0{,}2x + 1{,}1y + 0{,}3z = 1{,}2 \\ 0{,}3x - y - 2z = -1. \end{cases}$ **f)** $\begin{cases} 2x + 3y - 2z = 66 \\ y + 4z = 90 \\ y + 5z = 45. \end{cases}$

   g) $\begin{cases} 0{,}3x + 0{,}2y + 0{,}5z = 1 \\ 0{,}1x - 0{,}1y = 0 \\ 0{,}2x + 1{,}1y + 0{,}3z = 1{,}1. \end{cases}$ **h)** $\begin{cases} x + 2y + 3z = 2 \\ 1{,}1x + 0{,}5y + 1{,}6z = 2{,}2 \\ -x - 2y - 3z = -2. \end{cases}$ **i)** $\begin{cases} 50x + 20y + 8z = 20{,}6 \\ 30x + 15y + 16z = 15{,}2 \\ 25x + 32y + 40z = 21{,}9. \end{cases}$

   j) $\begin{cases} \frac{1}{2}x + \frac{1}{3}y + \frac{1}{6}z = 11 \\ \frac{1}{6}x + \frac{1}{2}y + z = 21 \\ x + \frac{1}{3}y + \frac{1}{4}z = \frac{37}{2}. \end{cases}$ **k)** $\begin{cases} \frac{1}{4}x + \frac{1}{5}y + \frac{9}{20}z = 27 \\ \frac{1}{5}x + \frac{1}{4}y + \frac{9}{20}z = 27 \\ x + \frac{1}{2}y + \frac{3}{2}z = 90. \end{cases}$

   l) $\begin{cases} \sqrt{2}x + \sqrt{3}y + z = 5 + \sqrt{6} \\ x + \sqrt{2}y + \sqrt{3}z = 4\sqrt{2} + \sqrt{6} \\ -x + \sqrt{3}y + \sqrt{2}z = 3 - \sqrt{2} + 2\sqrt{3}. \end{cases}$ **m)** $\begin{cases} 0{,}8x + 1{,}5y + 2{,}3z = 2{,}4 \\ 1{,}2x + 0{,}8y + 2z = 3{,}6 \\ 1{,}2x - 0{,}4y + 0{,}8z = 3{,}0. \end{cases}$

9. Sean $a, b, c \in \mathbb{R}$ con $a \neq 0$. Se considera la ecuación: hallar $x \in \mathbb{R}$ tal que $ax^4 + bx + c = 0$. Elaborar un algoritmo que permita identificar la existencia de raíces reales y como calcularlas. Verifique su algortimo en los siguientes casos:

   a) $t^4 - 9t^2 + 20 = 0$. **b)** $3t^4 + 7t^2 - 40 = 0$. **c)** $2t^4 + 9t^2 + 4 = 0$.

10. En cada item se define una función u y una partición uniforme $\tau(m)$ del intervalo $[a, b]$ que se indica y $m = 10$. Calcule $I(u) = \int_a^b u(x)\,dx$ y una aproximación $I(v_m)$ de $I(u)$ calculada con la regla del rectángulo. Compare los resultados.

   a) $u(x) = x \quad x \in [0, 10]$. **b)** $u(x) = -3x + 2 \quad x \in [-1, 2]$. **c)** $u(x) = 2x^2 + 5 \quad x \in [-1, 2]$.

   d) $u(x) = x^3 - x^2 + 1 \quad x \in [0, 1]$. **e)** $u(x) = \dfrac{1}{1+x} \quad x \in [0, 1]$. **f)** $u(x) = e^{-x} \quad x \in [0, 4]$

   g) $u(x) = \operatorname{sen}(x) \quad x \in \left[0, \dfrac{\pi}{2}\right]$. **h)** $u(x) = \cos^2(x) \quad x \in [0, \pi]$. **i)** $u(x) = \ln(x) \quad x \in [1, e]$.

   j) $u(x) = \sqrt{1 + x^2} \quad x \in [0, 2]$.

11. En cada item se define una función real φ. Elabore un algoritmo de cálculo de $\varphi(x)$ de modo que el número de operaciones elementales sea el más pequeño posible, contabilice dicho número.

   a) $\varphi(x) = 10 - \dfrac{1}{x^2} - \dfrac{1}{6x^4} - \dfrac{1}{10x^6} - \dfrac{1}{14x^8} - \dfrac{1}{18x^{10}} \quad x > 1.$

   b) $\varphi(x) = 1 + \dfrac{3}{\sqrt{1+x}} - \dfrac{5}{1+x} + \dfrac{7}{(1+x)^{\frac{3}{2}}} - \dfrac{9}{(1+x)^2} + \dfrac{11}{(1+x)^{\frac{5}{2}}} \quad x \geq 0.$

   c) $\varphi(x) = \dfrac{4}{(x^2-3)^{\frac{1}{2}}} + \dfrac{9}{5(x^2-3)^{\frac{3}{2}}} + \dfrac{14}{9(x^2-3)^{\frac{5}{2}}} + \dfrac{19}{14(x^2-3)^{\frac{7}{2}}} + \dfrac{24}{19(x^2-3)^{\frac{9}{2}}} \quad x > \sqrt{3}.$

   d) $\varphi(x) = 2\left(x^2 + \dfrac{4}{3}x^4 + \dfrac{16}{9}x^6 + \dfrac{256}{81}x^8\right)^{\frac{1}{2}} \quad x \in \mathbb{R}.$

   e) $\varphi(x) = 1 + \dfrac{1}{2}\operatorname{sen}(x) - \dfrac{1}{4}\operatorname{sen}^2(x) + \dfrac{1}{8}\operatorname{sen}^3(x) - \dfrac{1}{16}\operatorname{sen}^4(x) \quad x \in \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$.

   f) $\varphi(x) = \dfrac{1}{3} + \dfrac{2}{9}\cos^2(x)\operatorname{sen}(x) - \dfrac{3}{16}\cos^3(x)\operatorname{sen}^2(x) + \dfrac{4}{21}\cos^4(x)\operatorname{sen}^3(x) - \dfrac{5}{26}\cos^5(x)\operatorname{sen}^4(x) \quad x \in \mathbb{R}.$

12. Para $n \in \mathbb{Z}^+$ con $3 \leq n \leq 19$, se define

$$f_n(x) = \sum_{k=0}^{n} \frac{(-1)^k}{(2k+1)!} x^{k+\frac{1}{2}} \quad x \in \left[0, \pi^2\right].$$

   a) Para cada n impar, elabore un algoritmo para calcular valores aproximados de $f_n(x)$ de modo que el número de operaciones elementales sea el más pequeño posible y se eviten sumas y restas alternadas.

   b) Para $n = 19$, $x = \left(\dfrac{\pi}{6}\right)^2$ y una aproximación de $\pi \simeq 3{,}1415926536$, se tiene $f_{19}\left(\dfrac{\pi}{6}\right) = 0{,}5$. Aplique el algoritmo desarrollado en la parte a) precedente y calcule $f_n(x)$ para n, x y la aproximación de π que se indica, obtendrá una aproximación de 0,5

   i) $n = 5$, $\pi \simeq 3{,}1415$, $x = \left(\dfrac{\pi}{6}\right)^2$. **ii)** $n = 7$, $\pi \simeq 3{,}141593$, $x = \left(\dfrac{\pi}{6}\right)^2$.

   iii) $n = 9$, $\pi \simeq 3{,}14159265$, $x = \left(\dfrac{\pi}{6}\right)^2$. **vi)** $n = 11$, $\pi \simeq 3{,}141593$, $x = \left(\dfrac{\pi}{6}\right)^2$.

13. La función real g se define sobre $[0, 10]$ como sigue:

$$g(x) = \sum_{k=0}^{15} \frac{(-1)^k}{k!10^k} x^{k+2} \quad x \in [0, 10]$$

   a) Elabore un algoritmo para calcular valores aproximados de $g(x)$ de modo que se eviten los cálculos directos de $k!$, 10^k, x^{k+2} y sumas y restas alternadas; y, en lo posible que el número de operaciones elementales sea el más pequeño posible.

b) Contabilice el número de operaciones elementales de su algoritmo.

c) Aplique su algoritmo para calcular $g(1)$ y compruebe que obtendrá una aproximación de 0,904837418.

d) Aplique su algoritmo para calcular $g(10)$ y obtendrá una aproximación de 36,78794412.

14. Considerar la función h definida como

$$h(x) = \sum_{k=0}^{5} \frac{x^{2^k}}{(8k)!5^{3k^2}} = x + \frac{x^2}{8!5^3} + \frac{x^4}{16!5^{12}} + \frac{x^8}{24!5^{27}} + \frac{x^{16}}{32!5^{48}} + \frac{x^{32}}{40!5^{75}} \quad x \in \mathbb{R}.$$

a) Con la calculadora de bolsillo calcule $(8k)!$, 5^{3k^2}, $(8k)!5^{3k^2}$ y $\dfrac{1}{(8k)!5^{3k^2}}$ para $k = 0, 1, \ldots, 5$ y analice las dificultades de cálculo y los resultados que obtiene.

b) Utilice el desarrollo de $h(x)$ para calcular $h(20)$ y explique las dificultades de cálculo que se presentan.

c) Sean $x \in \mathbb{R}$ y $u(x) = \dfrac{1}{25 \times \cdots \times 32 \times 5^3} \left(\dfrac{x}{25}\right)^8$. Note que $u(x) = \dfrac{1}{25 \times \cdots \times 32 \times 5^3} \left(\dfrac{x}{25}\right)^8 = \dfrac{1}{125} \times \dfrac{\frac{x}{25}}{25} \times \dfrac{\frac{x}{25}}{26} \times \cdots \times \dfrac{\frac{x}{25}}{32}$.
Calcule $u(20)$.

d) A partir de la escritura de $h(x)$ siguiente:

$$\begin{aligned} h(x) &= x + \frac{x^2}{8!5^3}\left(1 + \frac{1}{9 \times \cdots \times 16 \times 5}\left(\frac{x}{5^4}\right)^2\left(1 + \frac{1}{17 \times \cdots \times 24 \times 5^3}\left(\frac{x}{5^3}\right)^4\right.\right. \\ &\quad \left.\left.\left(1 + \frac{1}{25 \times \cdots \times 32 \times 5^5}\left(\frac{x}{5^2}\right)^8\left(1 + \frac{1}{17 \times \cdots \times 24 \times 5^3}\left(\frac{x}{5}\right)^{16}\right)\right)\right)\right) \end{aligned}$$

y de la observación en la parte c) precedente, exprese $h(x)$ en forma más conveniente y calcule $h(20)$. Explique las dificultades o bondades de cálculo con la nueva escritura de h.

15. Elaborar un logaritmo que permita calcular los valores de $P_m(x)$, $Q_m(x)$ $\quad m = 0, 1, 2, \ldots,$ $x \in [-1, 1]$, si

$$P_m(x) = 1 + \frac{(m!)^2}{(2m)!} \sum_{k=1}^{m} (-1)^k \frac{(2m+2k)!}{(m-k)!(m+k)!(2k)!}\, x^{2k} \quad x \in [-1, 1],$$

$$Q_m(x) = x + \frac{(m!)^2}{(2m+1)!} \sum_{k=1}^{m} (-1)^k \frac{(2m+2k+1)!}{(m-k)!(m+k)!(2k+1)!}\, x^{2k+1} \quad x \in [-1, 1],$$

Los polinomios P_m y Q_m son conocidos como polinomios de Legendre.

16. Se define $v(x) = \sum_{k=0}^{15} \dfrac{1}{k!} \dfrac{1}{(x^2+1)^k} \quad x \in [0, 10]$.

a) Utilice directamente la escritura de $v(x)$ para calcular $v(3)$ y determine el número de operaciones elementales que realiza. Indique las posibles dificultades de cálculo de $v(3)$.

b) Elabore un algoritmo para calcular valores aproximados de $v(3)$ de modo que el número de operaciónes elementales sea el más pequeño posible y contabilice el total de dichas operaciones elementales en el cálculo de $v(x)$ $\quad x \in [0, 10]$. Compare su resultado con el siguiente: $v(3) \simeq 1{,}105170918$.

c) Aplique su algoritmo y calcule $v(10)$ y compruebe su resultado con $v(10) \simeq 1{,}009950167$.

17. Considere la función θ definida como

$$\theta(x) = \frac{1}{2} \sum_{k=0}^{15} \frac{(-1)^k x^k}{(2k+1)!4^k} \quad x \geq 0.$$

a) Utilice sin modificaciones $\theta(x)$ y calcule $\theta\left(\left(\frac{\pi}{3}\right)^2\right) \simeq 0{,}477464829$. ¿Qué dificultades de cálculo se presentan?

b) Elabore un algoritmo que facilite el cálculo de $\theta(x)$ y contabilice el número de operaciones elementales que se realizan. Calcule $\theta\left(\left(\frac{\pi}{3}\right)^2\right)$ y compare con el valor dado en i) precedente.

18. Se da la función f definida como $f(x) = \sum_{k=0}^{11} \frac{x^{2k}}{(2k)!} \quad x \in [0,2]$.

a) Calcule $f\left(\frac{1}{2}\right)$ y compare con $f(0{,}5) \simeq 1{,}127625966$. Contabilice el número de operaciones elementales que realiza.

b) Mejore aún la escritura de $f(x)$ siguiente:

$$f(x) = 1 + \frac{x^2}{2}\left(1 + \frac{x^2}{3\times 4}\left(1 + \frac{x^2}{5\times 6}\left(1 + \cdots + \frac{x^2}{19\times 20}\left(1 + \frac{x^2}{21\times 22}\right)\cdots\right)\right)\right)$$

y calcule $f\left(\frac{1}{2}\right)$. Contabilice el número de operaciones elementales que realiza.

19. Dada la función $g(x) = \sum_{k=0}^{15} \frac{x^{2k}}{(2k+1)!} \quad x \geq 0$. Mediante la elaboración de un algoritmo que facilite el cálculo de $g(x)$, Calcule $g\left(\frac{1}{2}\right)$ y compare con $g\left(\frac{1}{2}\right) \simeq 1{,}042190611$. ¿Cuántas operaciones se requieren para calcular $g(x)$ con y sin el algoritmo?

20. Aplique el esquema de Hörner para calcular $p(x)$ en x que se indica.
a) $p(x) = x^8 + x^7 + x^5 + x^3 + x^2 + x + 1, \quad x = 2.$
b) $p(x) = 5 - x^2 + 10x^3 + 7x^4 - 2x^5, \quad x = -3.$
c) $p(x) = 0{,}5 - 0{,}2x + 0{,}5x^2 + 3{,}25x^3 + 2{,}5x^4, \quad x = 0{,}8.$
d) $p(x) = 3 - 2{,}2x + 1{,}1x^3 - 2{,}8x^4 + 5{,}6x^5, \quad x = 1{,}5.$

21. Considere la función u definida como $u(x) = x^2 \quad x \in [0,2]$. En cada literal se da el número de puntos m de una partición uniforme $\tau(m-1)$ de $[0,2]$. Trace la gráfica de u y de su interpolante v_m utilizada en la regla del rectángulo. Calcule $I(u) = \int_0^2 x^2 dx$ e $I(v_m) = \sum_{j=1}^{m} hu\left(x_{j-1} + \frac{1}{2}jh\right)$.
a) $m = 2$. **b)** $m = 5$. **c)** $m = 9$. **d)** $m = 11$.
Compare los resultados. Para el efecto, calcule $|I(u) - I(v_m)|$ y concluya.

22. Se define la función f definida como $f(x) = \text{sen}(x) \quad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ y $\pi \simeq 3{,}1415926536$. Se define una partición uniforme $\tau(m) = \left\{-\frac{\pi}{2} + ih \mid i = 0, 1, \ldots, m\right\}$ de $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ con $m \in \mathbb{Z}^+$ que en cada literal se define. Trace la gráfica de f y la de su interpolante f_m utilizada en la regla del rectángulo. Calcular $I(f) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \text{sen}(x)\, dx$, $I(f_m) = \sum_{i=1}^{m} hf\left(-\frac{\pi}{2} + \frac{i}{2}h\right)$.
a) $m = 3$. **b)** $m = 5$. **c)** $m = 7$. **d)** $m = 9$.
Para cada m dado en a), b), c), d) calcule $|I(f) - I(f_m)|$ y concluya.

23. Se define la función g como $g(x) = e^x - \ln(x) \quad x \in [1,2]$.
a) Calcule $I(g) = \int_1^2 g(x)\, dx$.
b) Se define una partición $\tau(m) = \{1 + ih \mid i = 0, 1, \ldots, m\}$ con $m \in \mathbb{Z}^+$ que se da en cada caso. Aplique la regla del rectángulo para aproximar $I(g)$, para $m = 3$, $m = 6$, $m = 9$, $m = 12$.
c) Calcule $|I(g) - I(g_m)|$ con $I(g_m)$ calculado en la parte b) precedente. Concluya.

24. Considere la función real f definida como $f(x) = e^x \quad x \in \mathbb{R}$. Se sabe que $f'(-1) = e^{-1}$. Calcule aproximaciones y_0' de la derivada $f'(1)$ para cada h que se indica y calcule $|f'(-1) - y_0'|$. Analice los resultados.

a) $h = -0{,}05$. **b)** $h = -0{,}0005$. **c)** $h = -0{,}000005$. **d)** $h = 0{,}005$. **e)** $h = 0{,}00005$.

f) $h = 0{,}0000005$.

25. Se define la función u como $u(x) = \sqrt{1+x^2} \quad x \in \mathbb{R}$. Tenemos $u'(x) = \dfrac{x}{\sqrt{1+x^2}} \quad x \in \mathbb{R}$. Calcule aproximaciones de la derivada $u'(0) = 0$ para cada h que se indica y estime $|u'(0) - u_0'|$.

a) $h = -0{,}02$. **b)** $h = -0{,}0002$. **c)** $h = -0{,}00002$. **d)** $h = 0{,}002$. **e)** $h = 0{,}00002$.

f) $h = 0{,}0000002$.

26. Considere aproximaciones $v'(x) = \operatorname{sen}(3x) \quad x \in \mathbb{R}$. Se sabe que $v'(x) = 3\cos(3x) \quad x \in \mathbb{R}$. Calcule aproximaciones v_0' de la derivada $v'\left(\dfrac{\pi}{6}\right) = 0$ con $\pi \simeq 3{,}1416926536$, para cada h que se indica.

a) $h = -0{,}04$. **b)** $h = -0{,}0004$. **c)** $h = -0{,}000004$. **d)** $h = 0{,}004$. **e)** $h = 0{,}00004$.

f) $h = 0{,}0000004$.

27. En cada item se define una función v y se dan un punto x_0 y varios valores de h. Calcule aproximaciones v_0' de la derivada $v'(x_0)$ y estime $|v'(x_0) - v_0'|$.

a) $v(x) = 2x^2 - 3 \quad x \in \mathbb{R},\ x_0 = -1,\ h = -0{,}003,\ h = -0{,}0003,\ h = 0{,}00003,\ h = 0{,}000003.$

b) $v(x) = \dfrac{1}{1+x} \quad x > -1,\ x_0 = 0,\ h = -0{,}05,\ h = -0{,}00005,\ h = 0{,}0005,\ h = 0{,}000005.$

c) $v(x) = \cos(x^2) \quad x \in \mathbb{R},\ x_0 = \sqrt{\dfrac{\pi}{2}},\ h = -0{,}001,\ h = -0{,}0001,\ h = 0{,}00001,\ h = 0{,}0000001.$

d) $v(x) = \ln(1+2x) \quad x > -\dfrac{1}{2},\ x_0 = 0,\ h = -0{,}015,\ h = -0{,}000015,\ h = 0{,}00015,\ h = 0{,}0000015.$

e) $v(x) = \left(1+2x^2\right)^{\frac{1}{3}} \quad x \in \mathbb{R},\ x_0 = 2,\ h = -0{,}025,\ h = -0{,}00025,\ h = 0{,}000025,\ h = 0{,}0000025.$

28. La solución del problema de valor inicial siguiente: $\begin{cases} y'(x) + 2y(x) = e^x & 0 < x < 1, \\ y(0) = \dfrac{1}{3}, \end{cases}$ es $y(x) = \dfrac{1}{3}e^x$. Para $m = 10$ y una partición uniforme del intervalo $[0,1]$, aplique el método de Euler explícito y calcule aproximaciones $y_j \quad j = 1, \ldots, 10$ de $y(x_j)$. Trace la gráfica de la función $y(x)$ y represente los puntos $(x_j, y_j) \quad j = 0, 1, \ldots, 10$. Calcule $|y(x_j) - y_j| \quad j = 0, 1, \ldots, 10$ y dé una solución.

29. La solución del problema de valor inicial $\begin{cases} y'(x) = \dfrac{2-y}{1-x^2}x & |x| < 1, \\ y(0) = 1, \end{cases}$ es $y(x) = 2 - \sqrt{1-x^2} \quad |x| < 1$. Para $m = 8$ y una partición uniforme del intervalo $[0, 0{,}9]$, aplique el método de Euler explícito y calcule aproximaciones $y_j \quad j = 0, 1, \ldots, 8$ de $y(x_j)$. Trace las gráficas de la fución $y(x)$ y de los valores calculados $(x_j, y_j) \quad j = 0, 1, \ldots, 8$. Calcule $|y(x_j) - y_j| \quad j = 0, 1, \ldots, 8$ y compare los resultados.

30. Considere el problema de valor inicial $\begin{cases} y'(x) = \dfrac{y(x)}{x} - 1 & 1 < x < 3, \\ y(1) = 2. \end{cases}$ cuya solución es $y(x) = x(2 - \ln(x)) \quad x > 0$. Para $m = 10$ y una partición uniforme del intervalo $[1, 3]$, aplique el método de Euler explícito y calcule aproximaciones $y_j \quad j = 0, 1, \ldots, 10$ de y(x_j). Trace las gráficas de la función $y(x) \quad x \in [1, 3]$ y de los valores calculados $(x_j, y_j) \quad j = 0, 1, \ldots, 10$. Calcule $|y(x_j) - y_j| \quad j = 0, 1, \ldots, 10$ y compare los resultados.

31. Considere el problema de valor inicial $\begin{cases} y'(x) = \dfrac{x^2 + y^2(x)}{3xy(x)} & 1 < x < 1{,}5, \\ y(1) = 2, \end{cases}$ la solución es $y(x) = \sqrt{x^2 + 3x} \quad x > 0$. Para $m = 5$ y una partición uniforme del intervalo $[1, 1{,}5]$, proceda como en el ejercicio precedente.

32. Representar en base 10 los siguientes números:

    a) $(4746)_8$ **b)** $(7412{,}352)_8$ **c)** $(AB98.C31)_{16}$ **d)** $(100{,}001)_3$ **e)** $(1{,}4142)_5$ **f)** $(111{,}0101)_2$

    g) $(0{,}1110111110111011111\ldots)_2$ **h)** $(235{,}3333\ldots)_6$.

33. En cada caso, representar los siguientes números en base 10 a la base b que se indica con 8 cifras de precisión para la parte fraccionaria.

    a) 3,14159, $b = 2$. **b)** 2,718281, $b = 4$. **c)** $\sqrt{2}$, $b = 8$. **d)** $\sqrt{5}$, $b = 5$. **e)** $\dfrac{27}{7}$, $b = 3$.

    f) 1726,00011, $b = 2$. **g)** 135,26 $b = 4$. **h)** 135,42, $b = 8$.

34. Sean $a, b \in \mathbb{N}$ tales que $a \neq b$, $1 < a \leq 10$, $1 < b \leq 10$. Elaborar algoritmos que permitan convertir números positivos en base a a base b y recíprocamente; y, obtener su equivalente en base 10. Sugerencia: considérese $M = (a_n \ldots a_0, a_{-1} \ldots a_{-m})_a$ y $M = (b_p \ldots b_0, b_{-1} \ldots b_{-q})_b$, donde $a_i, a_{-j} \in \{0, 1, \ldots, a-1\}$, $i = 0, 1, \ldots, n$, $j = 1, \ldots, m$, $b_k, b_{-l} \in \{0, 1, \ldots, b-1\}, k = 0, 1, \ldots, p$, $l = 1, \ldots, q$.

35. Sean $a, b, c \in \mathbb{R}^+$.

    a) Considerar las expresiones $u = (a - b)c$ y $v = ac - bc$ con $a \approx b$. Demostrar que u presenta un error relativo menor que v. Verifique con $a = 0{,}6392$, $b = 0{,}6375$ y $c = 0{,}9364$.

    b) Considerar la matriz $A = \begin{bmatrix} a & b \\ c & a \end{bmatrix}$. Estudie el condicionamiento de $\det(A)$.

    c) Si $a = \dfrac{1}{\sqrt{3}}$, $b = 1$, $c = \frac{1}{3}$, $a_1 = rd(a)$, $b_1 = rd(b)$, $c_1 = rd(c)$, estudie la existencia de soluciones de los sistemas de ecuaciones $\begin{cases} ax + by = 1 \\ cx + ay = 0 \end{cases}$ y $\begin{cases} a_1x + b_1y = 1 \\ c_1x + a_1y = 0. \end{cases}$

    d) Sean $a = \dfrac{1}{1500}$, $b = \dfrac{1}{701}$ y $c = \dfrac{7}{22500}$. Si a, b, c se redondean con 8 cifras decimales, estudie la existencia de soluciones de los sistemas de ecuaciones del literal c).

    e) De b), c) y d), ¿qué conclusiones puede obtener?

36. Determinar el número de operaciones elementales para calcular $\det(A)$ si este se calcula usando el método de menores y cofactores cuando A es una matriz de 3×3, de 4×4 y de 5×5. Generalice los resultados.

37. Determinar el número de operaciones elementales que se requieren para calcular A^D la matriz adjunta de A cuando A es una matriz de 3×3, de 4×4 y de 5×5.

38. Usando la aritmética de punto flotante con 3 dígitos, evaluar $f(x) = x^4 - x^3 + 6x^2 - 3x + 0{,}145$ en $x = 4{,}71$.

    a) Aplicar el esquema de Hörner para calcular $f(4{,}71)$.

    b) Determinar el valor exacto de $f(4{,}71)$ y, en cada caso, calcular el error relativo.

    c) Calcular con 3 cifras de precisión $f(-0{,}101)$ directamente y con el esquema de Hörner. Calcular el error relativo.

    d) Calcular con 3 cifras de precisión $f(-0{,}10001)$ directamente y con el esquema de Hörner. Calcular el error relativo.

39. Sea f la función real definida por $f(x) = 2 + \frac{3-x}{x^2-1} \quad |x| \neq 1$.

    a) Calcular f con 2 y 3 cifras en aritmética de punto flotante en $x = 0{,}85$, $x = 0{,}95$, $x = 0{,}99$.

    b) Puesto que $f(x) = \dfrac{x}{x-1} + \dfrac{x-1}{x+1}$, calcule $f(x)$ para los puntos x del literal a).

c) Calcule el valor exacto de $f(x)$ para los puntos x dados en a).

d) Calcule el error relativo de $f(x)$ para los resultados de a) y b).

40. Sea $f(x) = \dfrac{1+x-e^x}{x^2} \quad x \neq 0.$

a) Calcular $\lim_{x\to 0} f(x)$. **b)** Calcular $f(0{,}5 \times 10^{-10})$.

c) Hallar un algoritmo para aproximar $f(x)$ con $|x| \in]0, 10^{-5}]$, y aplique en los puntos $x = 0{,}5\times 10^{-5}$ y $x = 0{,}1 \times 10^{-5}$.

41. Sean $x > 1$ y $n \in \mathbb{N}$. Construya algoritmos que permitan aproximar $\dfrac{x^n}{n!}$ en los siguientes casos:

a) $1 < x < 10$ y $20 < n < 50$. **b)** $x \geq 10$ y $n > 50$.

42. Hallar $\lim_{x\to 0} f(x)$ para las funciones f que se dan a continuación. En cada caso elabore algoritmos que se adapten a la estabilidad numérica en un entorno de cero.

a) $f(x) = \sqrt{x^2+1} - 1$. **b)** $f(x) = \sqrt{x^2+1} - x$. **c)** $f(x) = -x + \operatorname{sen}(x)$. **d)** $f(x) = \dfrac{1}{x+1} - 1$.

e) $f(x) = 1 - \cos(x)$. **f)** $f(x) = \dfrac{e - e^{\cos(x)}}{x^2}, x \neq 0$. **g)** $f(x) = \dfrac{e^x - e^{-x}}{\operatorname{sen}(x)}, x \neq k\pi,\ k \in \mathbb{Z}$.

h) $f(x) = \dfrac{e^x - e^{\operatorname{sen}(x)}}{x^3}, x \neq 0$. **i)** $f(x) = \dfrac{1-\cos(x)}{x},\ x \neq 0$. **j)** $f(x) = \dfrac{e^x - (1+x)}{x^2},\ x \neq 0$.

43. Determinar los números de condicionamiento de las funciones siguientes:

a) $f(x) = \cos(x) \quad x \in \mathbb{R}$. **b)** $f(x) = \tan(x) \quad x \in \left]-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right[$. **c)** $f(x) = \ln(x) \quad x > 0$.

d) $f(x) = 0{,}5 + \dfrac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2} dt \quad x \geq 0$.

e) En los incisos a), b) y c) determinar el más grande subconjunto de $\mathbb{R}$ en el que f está bien condicionado.

f) Pruebe que la función del inciso d) está bien condicionada para todo $x \geq 0$.

44. Sea $\varphi : \mathbb{R}^3 \longrightarrow \mathbb{R}$ la función definida por $\varphi(x, y, z) = \frac{xy}{z}$ con $z \neq 0$.

a) Pruebe que $\varepsilon_\varphi = \varepsilon_x + \varepsilon_y - \varepsilon_z$,donde ε_x, ε_y, ε_z son los errores relativos de x, y, z respectivamente.

b) Determine el error acumulado.

45. Si $\varphi(x, y) = x^y \quad x, y \in \mathbb{R}^+$ y $\varepsilon_x = \varepsilon_y$, pruebe que el error relativo de φ viene dado por

$$\varepsilon_\varphi = \varepsilon_x + (y \ln(x))\varepsilon_y.$$

¿Qué número de condicionamiento influye en el cálculo de φ?.

46. Sea $A = (a_1, \ldots, a_n) \in \mathbb{R}^n$, donde a_i, $i = 1, \ldots, n$ son números de máquina. Sea $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ la función definida por $F(x_1, \ldots, x_n) = \sum_{i=1}^n a_i x_i$. Si $|\varepsilon_{x_i}| \leq eps$, $i = 1, \ldots, n$, pruebe que el error relativo de $F(x_1, \ldots, x_n)$ verifica

$$|\varepsilon_F| \leq eps \text{ si } \begin{cases} a_i > 0, & x_i > 0, \\ a_i < 0, & x_i < 0, \end{cases} \quad i = 1, \ldots n.$$

47. Sea $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ la función definida por $F(a_1, \ldots, a_n) = \frac{1}{n} \sum_{i=1}^n a_i$. Supongamos que $a_i > 0$ $\forall i = 1, \ldots, n$. Proponer un algoritmo de cálculo de $z = F(a_1, \ldots, a_n)$ y estudiar la propagación de los errores. Si el error en cada operación es $\varepsilon_i = \varepsilon$, ¿cuál es el error acumulado?.

48. Sean $(a_1,\ldots,a_n) \in \mathbb{R}^n$ y $\vec{z} = \vec{\varphi}(a_1,\ldots a_n)$, donde $\vec{\varphi}(a_1,\ldots a_n) = \begin{bmatrix} \varphi_1(a_1,\ldots,a_n) \\ \vdots \\ \varphi_n(a_1,\ldots,_n) \end{bmatrix}$, con $\varphi_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1,\ldots,n$ funciones de clase C^1. Supóngase que $\varphi_i(a_1,\ldots a_n) \neq 0$, $i = 1,\ldots,n$. Muestre que los números de condicionamiento de φ_i viene dados por

$$C_i = \frac{a_j}{\varphi_i(a_1,\ldots,a_n)} \frac{\partial \varphi_i}{\partial x_j}(a_1,\ldots,a_n), \quad i,j = 1,\ldots,n.$$

49. Considerar la ecuación $x^2 + 2px - q = 0$, donde $p > 0, q > 0$ y $p >> q$. Sea $x = -p + \sqrt{p^2+q}$ una raíz de la ecuación. Calcule ε_x y demuestre que

$$eps \leq \varepsilon_x = \frac{\Delta x}{x} \leq 3eps.$$

Nota: La notación $p >> q$ significa q es muy pequeño comparado con p o que p es muy grande comparado con q.

50. Se desea calcular $E(x) = \sqrt{1+x} - 1$ para $x = 0{,}0009$ y $x = 0{,}001$ con 4 cifras decimales.

a) Calcule $E(x)$ en dichos puntos. **b)** Utilizando $E(x)$, construya otra expresión que se adapte a la estabilidad numérica y aplique para los puntos dados x. Compare los resultados.

51. Considerar el sistema de ecuaciones lineales $\begin{cases} 0{,}002x + y = 0{,}2 \\ x + y = 1. \end{cases}$ Utilice el método de eliminación gaussiana para determinar:

a) La solución exacta del sistema. **b)** La solución aproximada con 5 cifras decimales. **c)** Intercambie la primera ecuación con la segunda y proceda como en los incisos a) y b). Compare los resultados.

d) El sistema de ecuaciones propuesto es equivalente al siguiente: $\begin{cases} x + 500y = 100 \\ x + y = 1. \end{cases}$ Usando la aritmética de punto flotante con 5 dígitos de precisión, resuelva el sistema de ecuaciones y compare con los resultados precedentes.

52. Sean $a, b \in \mathbb{R}^+$, $m, n \in \mathbb{N}$ tales que $0 \leq m \leq n$. Se desea calcular

$$F(m) = \sum_{k=0}^{m} \binom{n}{k} a^k b^{n-k} \quad m = 0,1,\ldots,n,$$

donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

a) Pruebe que $\binom{n}{m+1} = \frac{n-m}{m+1}\binom{n}{m} \quad m = 0{,}1.\ldots,n-1.$

b) Sea $\varphi(k) = \binom{n}{k} a^k b^{n-k}$. Entonces $F(m) = \sum_{k=0}^{m} \varphi(k)$. Elabore un algoritmo que se adpate a la estabilidad numérica y aplique para $a = 0{,}1$, $b = 0{,}5$, $m = 4, n = 10$.

53. Sea $I(n) = \int_0^1 \frac{x^n}{x+5} dx$, $n = 0,1,2,\ldots$

a) Muestre que $I(n) + 5I(n-1) = \frac{1}{n}$.

b) Considerar el algoritmo: $I(n) = \frac{1}{n} - 5I(n-1) \quad n = 1,2,\ldots,25$. Calcule $I(0)$ e $I(n)$ para $n = 1,2,\ldots,25$.

c) Muestre que $\lim_{n\to\infty} I(n) = 0$.

d) Considere el siguiente algoritmo: $\begin{cases} I(n) = \frac{1}{5n}, \\ I(n-1) = \frac{1}{5}\left(\frac{1}{n} - I(n)\right), n = 25,24,\ldots,1. \end{cases}$ Calcule $I(n)$ e $I(n-1), n = 25,\ldots,1$. Compare con los resultados anteriores.

54. Sea $I(n) = \int_0^1 x^{\frac{n+4}{2}}\, e^x dx \quad n = -4, -3, -2, \ldots$

a) Muestre que $I(n) = e - \left(\frac{n}{2} + 2\right) I(n-2) \quad n = -2, -1, 0, \ldots$.

b) Calcule $I(-4)$.

c) Use el cambio de variable $x = t^2$ y muestre que $I(-3) = e - \int_o^1 e^{t^2} dt$. Utilice el polinomio de Taylor de $e^\alpha \quad \alpha \in \mathbb{R}$ para aproximar $\int_0^1 e^{t^2} dt$ y muestre que

$$\int_0^1 e^{t^2} dt = 1 + \frac{1}{3} + \frac{1}{2!}\frac{1}{5} + \cdots \frac{1}{k!}\frac{1}{2k+1} + E_{k+1},$$

donde E_{k+1} es el error cometido y k es tal que $\dfrac{1}{k!}\,\dfrac{1}{2k+1} < 10^{-6}$.

d) Utilizando el algoritmo dado en a), elabore un programa para calcular $I(n) \quad n = -4, -3, -2, \ldots, 25$.

e) Note que $\lim_{n\to\infty} I(n) = 0$. Establezca el siguiente algoritmo $\begin{cases} I(n) = \dfrac{2e}{n+4}, \\ I(n-1) = \dfrac{2e}{n+3}, \\ I(n-2) = \dfrac{2(e - I(n))}{n+4}. \end{cases}$

f) Elabore un programa para el cálculo de $I(n) \quad n = 25, 24, \ldots, -2$. Compare con los resultados dados en d).

g) Para $\varepsilon = 10^{-6}$, deduzca el siguiente algoritmo

$$I(n) = 2\left[\frac{1}{0!(n+6)} + \frac{1}{1!(n+8)} + \frac{1}{2!(n+10)} + \frac{1}{3!(n+12)} + \cdots + \frac{1}{8!(n+22)}\right].$$

Elabore un programa para el cálculo de $I(n)$, $n = -4, -3, \ldots, 25$. Compare los resultados con los otros algoritmos. ¿Qué concluye?.

h) ¿Por qué no es práctico utilizar la regla de los trapecios para cada n?.

1.13. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.
2. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
3. R. M. Barbolla, M. García, J. Margalef, E. Outerelo, J. L. Pinilla. J. M. Sánchez, Introducción al Análisis Real, Editorial Alambra Universidad, Madrid, 1981.
4. G. Birkhoff, S. Maclane, Algebra Moderna, Cuarta Edición, Editorial Vicens-Vives, Barcelona. 1974.
5. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
6. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
7. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
8. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.
9. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.

10. Francis G. Florey, Fundamentos de Algebra Lineal y Aplicaciones, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1980.

11. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, Pitagora Editrice Bologna, 1983.

12. Waltson Fulks, Cálculo Avanzado, Editorial Limusa, México, 1973.

13. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

14. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.

15. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

16. Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Editorial Society for Industrial and Applied Mathematics, Philadelphia, 1996.

17. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.

18. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.

19. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

20. Rodolfo Luthe, Antonio Olivera, Fernando Schutz, Métodos Numéricos, Editorial Limusa, México, 1986.

21. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

22. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

23. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

24. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

25. A. A. Samarski, Introducción a los Métodos Numéricos, Editorial Mir, Moscú, 1986.

26. Michelle Schatzman, Analyse Numérique, Inter Editions, París, 1991.

27. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

28. Michael Spivak, Calculus, Segunda Edición, Editorial Reverté, Barcelona, 1996.

29. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

30. E. A. Volkov, Métodos Numéricos, Editorial Mir, Moscú, 1990.

Capítulo 2

Interpolación polinomial, derivación e integración numérica

Resumen

En este capítulo nos interesamos en tres temas importantes: la interpolación polinomial, la derivación e integración numéricas. Estos temas los abordamos como formas lineales definidas en apropiados espacios vectoriales reales, es decir, en el ámbito de los espacios duales. Es por esto que iniciamos el estudio de este capítulo con los espacios duales. A continuación posicionamos el problema de la interpolación polinomial y tratamos la existencia del polinomio interpolante de Lagrange, obtenemos una estimación del error de interpolación así como algunos tipos de polinomios de interpolación de Lagrange más usados. Mediante la aplicación de los polinomios de Taylor con error, obtenemos procedimientos de cálculo aproximado de las derivadas primera y segunda de funciones reales, procedimiento que se generaliza al cálculo numérico de derivadas de orden superior. Introducimos los operadores en diferencias finitas y luego se construyen fórmulas de aproximación de derivadas de funciones reales como formas lineales. Estos resultados se extienden para el cálculo de las derivadas parciales primeras, segundas y el laplaciano de funciones reales en dos variables. Posteriormente, tratamos la integración numérica de funciones reales en la que nos limitados a obtener fórmulas de integración numérica conocidas como regla del punto medio, regla de los trapecios y regla de Simpson así como sus generalizaciones y la estimación del error. Estos resultados son aplicados al cálculo numérico de integrales dobles sobre dominios de los tipos I y II, es decir como integrales reiteradas.

2.1. Espacios duales

Los problemas de interpolación polinomial, derivación e integración numérica serán tratados como formas lineales definidas en apropiados espacios funcionales. Para ello comenzamos precisamente con los espacios duales y muy particularmente los espacios vectoriales reales de dimensión finita y sus duales que también son de dimensión finita. Asumimos que el lector tiene algún conocimiento sobre las aplicaciones lineales. En el anexo se resumen algunos resultados importantes, y al final del capítulo se citan algunos textos de álgebra lineal en los que se podrá consultar estos tópicos.

Definición 1 *Sea V un espacio vectorial. Toda aplicación lineal f de V en $\mathbb{R}$ se llama funcional lineal sobre V o también forma lineal en V.*

De la definición se tiene que f es un funcional lineal en V si y solo si satisface las dos condiciones siguientes:

i) f es una función de V en $\mathbb{R}$.

ii) f es lineal, esto es, para todo $\alpha \in \mathbb{R}$, $x, y \in V$, se tiene $f(x+y) = f(x) + f(y), \quad f(\alpha x) = \alpha f(x)$.

La propiedad ii) de la linealidad de f se escribe en una sola, así: si $\alpha, \beta \in \mathbb{R}$, $x, y \in V$, f es lineal si y solo si $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ y esta a su vez es equivalente a $f(\alpha x + y) = \alpha f(x) + f(y)$.

El conjunto de todos los funcionales lineales en V se designa con V^*. Con las operaciones habituales de adición de funciones " $+$ " siguiente:

$$\forall f, g \in V^*, \quad (f+g)(x) = f(x) + g(x) \quad \forall x \in V,$$

y el producto de escalares por funciones " $\cdot$ " :

$$\forall \alpha \in \mathbb{R}, \quad \forall f \in V^*, \quad (\alpha \cdot f)(x) = \alpha f(x) \quad \forall x \in V,$$

el conjunto V^* es un espacio vectorial real denominado espacio dual de V.

Ejemplos

1. Sean $V = \mathbb{R}^2$ y T la función de $\mathbb{R}^2$ en $\mathbb{R}$ definida como $T(x,y) = 2x + y \quad (x,y) \in \mathbb{R}^2$. Entonces T es una forma lineal en $\mathbb{R}^2$, esto es, $T \in \left(\mathbb{R}^2\right)^*$. Efectivamente, sean $\alpha, \beta \in \mathbb{R}$, $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$, entonces
$$\alpha(x_1, y_1) + \beta(x_2, y_2) = (\alpha x_1 + \beta x_2, \alpha y_1 + \beta y_2)$$
y de la definición de la función T se sigue que
$$\begin{aligned} T(\alpha(x_1, y_1) + \beta(x_2, y_2)) &= T(\alpha x_1 + \beta x_2, \alpha y_1 + \beta y_2) = 2(\alpha x_1 + \beta x_2) + (\alpha y_1 + \beta y_2) \\ &= \alpha(2x_1 + y_1) + \beta(2x_2 + y_2) = \alpha T(x_1, y_1) + \beta T(x_2, y_2). \end{aligned}$$

2. Sean $V = C([a,b])$ el espacio de funciones reales continuas en el intervalo cerrado $[a,b]$. Se define la función I de $C([a,b])$ en $\mathbb{R}$ como sigue: $I(u) = \int_a^b u(x)\,dx \quad u \in C([a,b])$. Entonces I es un funcional lineal sobre $C([a,b])$. Pués de las propiedades de la integral de Riemann siguientes:
$$\begin{aligned} \int_a^b (u(x) + v(x))\,dx &= \int_a^b u(x)\,dx + \int_a^b v(x)\,dx \quad u, v \in C([a,b]), \\ \int_a^b \alpha u(x)\,dx &= \alpha \int_a^b u(x)\,dx \quad \alpha \in \mathbb{R}, \ \ u \in C([a,b]), \end{aligned}$$
se deduce la linealidad de I.

3. Sean $V = C^1(]a,b[)$ el espacio de funciones que poseen derivada continua en el intervalo abierto $]a,b[$, $x_0 \in]a,b[$. Se define el funcional F sobre $C^1(]a,b[)$ como a continuación se indica:
$$F(u) = \frac{du}{dx}(x_0) \quad u \in C^1(]a,b[).$$
Por las propiedades de la derivada siguientes:
$$\frac{d}{dx}(u+v)(x_0) = \frac{du}{dx}(x_0) + \frac{dv}{dx}(x_0) \quad u, v \in C^1(]a,b[),$$
$$\frac{d}{dx}(\alpha u)(x_0) = \alpha \frac{du}{dx}(x_0) \quad \alpha \in \mathbb{R}, \ u \in C^1(]a,b[).$$
Se deduce que F es un funcional lineal sobre $C^1(]a,b[)$.

Teorema 1 *Si V es un espacio vectorial real de dimensión n, entonces* $\dim V^* = n = \dim V$.

Demostración. Para el efecto, construiremos un conjunto de funcionales lineales $\{f_1, \ldots, f_n\}$ sobre V y mostraremos que tal conjunto es una base de V^*.

i) Existencia de funcionales lineales sobre V.

Sea $B_v = \{v_1, \ldots, v_n\}$ una base ordenada de V y $x \in V$. Existen $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ tales que $x = \sum_{k=1}^{n} \alpha_k v_k$. Para cada $i = 1, \ldots, n$, se define f_i de V en $\mathbb{R}$ como sigue:

$$f_i(x) = f_i\left(\sum_{k=1}^{n} \alpha_k v_k\right) = \alpha_i \quad x \in V,$$

entonces $f_i \in V^*$. En efecto, sean $x, y \in V$, existen $\alpha_1, \ldots, \alpha_n,\ \lambda_1, \ldots, \lambda_n \in \mathbb{R}$ tales que

$$x = \sum_{k=1}^{n} \alpha_k v_k, \qquad y = \sum_{k=1}^{n} \lambda_k v_k, \qquad x + y = \sum_{k=1}^{n} (\alpha_k + \lambda_k) v_k,$$

de la definición de la función f_i se sigue :

$$\begin{aligned} f_i(x+y) &= f_i\left(\sum_{k=1}^{n} (\alpha_k + \lambda_k) v_k\right) = \alpha_i + \lambda_i = f_i\left(\sum_{k=1}^{n} \alpha_k v_k\right) + f_i\left(\sum_{k=1}^{n} \lambda_k v_k\right) \\ &= f_i(x) + f_i(y). \end{aligned}$$

Sea $\beta \in \mathbb{R}$, entonces

$$\beta x = \beta \sum_{k=1}^{n} \alpha_k v_k = \sum_{k=1}^{n} \beta \alpha_k v_k,$$

luego

$$f_i(\beta x) = f_i\left(\sum_{k=1}^{n} \beta \alpha_k v_k\right) = \beta \alpha_i = \beta f\left(\sum_{k=1}^{n} \alpha_k v_k\right) = \beta f(x).$$

Así, $f_i \in V^* \quad i = 1, \ldots, n$.

ii) Denotamos con $B_* = \{f_1, \ldots, f_n\}$. Probemos que el conjunto B_* es una base de V^*. Para ello mostramos que B_* genera a V^* y es linealmente independiente.

a) Mostramos que B_* genera a V^*. Como $B_* \subset V^*$ se sigue que el subespacio generado por B_*, que se denota con $L(B_*) = \left\{ \sum_{i=1}^{n} \alpha_i f_i \mid \alpha_i \in \mathbb{R},\ i = 1, \ldots, n \right\}$ está contenido en V^*, esto es, $L(B_*) \subset V^*$. Probemos que $V^* \subset L(B_*)$. Sea $f \in V^*$ y $x \in V$. Existen $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ tales que $x = \sum_{k=1}^{n} \alpha_k v_k$. Entonces

$$f(x) = f\left(\sum_{k=1}^{n} \alpha_k v_k\right) = \sum_{k=1}^{n} \alpha_k f(v_k),$$

y de la definición de f_i, se tiene

$$f_i(x) = f_i\left(\sum_{k=1}^{n} \alpha_k v_k\right) = \alpha_i \quad i = 1, \cdots, n,$$

luego

$$f(x) = \sum_{k=1}^{n} f(v_k) f_k(x).$$

Ponemos $\beta_k = f(v_k)$. Resulta

$$f(x) = \sum_{k=1}^{n} \beta_k f_k(x) \quad \forall x \in V,$$

que muestra que f es combinación lineal de los elementos de B_*. Así, $f \in L(B_*)$, o sea $V^* \subset L(B_*)$. En conclusión, $V^* = L(B_*)$.

b) Probemos que B_* es linealmente independiente. Sean $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ y consideremos la combinación lineal nula $\lambda_1 f_1 + \cdots + \lambda_n f_n = 0$. Note que $\lambda_1 f_1(x) + \cdots + \lambda_n f_n$ es un funcional lineal sobre V.

Entonces, $\lambda_1 f_1(x) + \cdots + \lambda_n f_n(x) = 0 \quad \forall x \in V$. Por la definición del funcional f_i, para $x = v_j$ se tiene $f_i(v_j) = \begin{cases} 1, & si \ \ i = j, \\ 0, & si \ \ i \neq j, \end{cases}$ entonces para $x = v_1$, obtenemos

$$0 = \lambda_1 f_1(v_1) + \cdots + \lambda_n f_n(v_1) = \lambda_1,$$

para $x = v_2$, se deduce

$$0 = \lambda_1 f_1(v_2) + \cdots + \lambda_n f_n(v_2) = \lambda_2,$$

así sucesivamente, para $x = v_n$ se obtiene

$$0 = \lambda_1 f_1(v_n) + \cdots + \lambda_n f_n(v_n) = \lambda_n.$$

Consecuentemente,

$$\lambda_1 f_1 + \cdots + \lambda_n f_n = 0 \Rightarrow \lambda_i = 0 \quad i = 1, \ldots, n,$$

que prueba que el conjunto B_* es linealmente independiente.

De i) y ii) se tiene B_* es una base de V^*, por lo tanto $\dim V^* = n$. ■

El conjunto $B_* = \{f_1, \ldots, f_n\}$ se le llama base dual de V^*. Observe que $f_i(v_j) = \begin{cases} 1, & si \ \ i = j, \\ 0, & si \ \ i \neq j. \end{cases}$ Note además que si $f \in V^*$, $B_v = \{v_1, \ldots, v_n\}$ es una base de V y $B_* = \{f_1, \ldots, f_n\}$ la base dual de B_v. En la parte ii) a) precedente se obtuvo para $f \in V^*$ la representación siguiente:

$$f(x) = \sum_{k=1}^{n} f(v_k) f_k(x) \qquad \forall x \in V,$$

que lo denominamos representación de f con respecto de las bases B_v y B_*. Escribiremos

$$f = \sum_{k=1}^{n} f(v_k) f_i \quad \forall f \in V^*.$$

Esta representación de f la utilizaremos en las aplicaciones a los problemas de interpolación polinomial, derivación e integración numérica.

Representación matricial

Sean $B = \{f_1, \ldots, f_n\}$ una base ordenada de V^* y $f \in V^*$. Existen $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ tales que $f = \sum_{i=1}^{n} \alpha_i f_i$.

La matriz de f asociada a la base B viene dada por $[f]_B = (\alpha_1, \ldots, \alpha_n) \in M_{1\times n}[\mathbb{R}]$. En particular, si B_* es la base dual de la base $B_v = \{v_1, \ldots, v_n\}$ de V, se tiene $[f]_{B*} = (f(v_1), \ldots, f(v_n))$.

Ejemplo

Si $V = \mathbb{R}^n$ y $B_v = \{\overrightarrow{e_1}, \ldots, \overrightarrow{e_n}\}$ la base canónica de $\mathbb{R}^n$, $f \in (\mathbb{R}^n)^*$,entonces $[f]_{B_*} = (f(\overrightarrow{e_1}), \ldots, f(\overrightarrow{e_n}))$ y la función f se escribe como sigue:

$$f(\overrightarrow{x}) = f(x_1, \ldots, x_n) = \sum_{i=1}^{n} f(\overrightarrow{e_i}) x_i.$$

Además, el espacio V^* es isomorfo a $M_{1\times n}[\mathbb{R}]$, pués la función Φ de V^* en $M_{1\times n}[\mathbb{R}]$ definida por $\Phi(f) = [f]_B \quad \forall f \in V^*$ con $B = \{f_1, \ldots, f_n\}$ una base ordenada de V^*, es lineal biyectiva, es decir que se trata de un isomorfismo.

Espacio Bidual

Sea V un espacio vectorial de dimensión finita n sobre $\mathbb{R}$, V^* el espacio dual de V. Sea $x \in V$ fijo y h una función de V^* en $\mathbb{R}$ definida como sigue: $h(g) = g(x) \qquad \forall g \in V^*$. Se verifica que h es lineal. Efectivamente, sean g_1, $g_2 \in V^*$ y $\alpha \in \mathbb{R}$. Se tiene

$$\begin{aligned} h(g_1 + g_2) &= (g_1 + g_2)(x) = g_1(x) + g_2(x) = h(g_1) + h(g_2), \\ h(ag_1) &= (\alpha g_1)(x) = \alpha g_1(x) = \alpha h(g_1). \end{aligned}$$

Resulta que h es un funcional lineal definido en V^*, o sea $h \in (V^*)^*$.

Escribimos V^{**} en vez de $(V^*)^*$ y lo denominamos espacio bidual de V.

En el siguiente teorema se establece que si $\dim V = n$, los espacios V y V^{**} son isomorfos.

Teorema 2 *Sea V un espacio vectorial de dimensión finita n sobre el cuerpo $\mathbb{R}$, V^{**} el espacio bidual de V. La aplicación φ de V en V^{**} definida por $\varphi(x) = h$ con $h(g) = g(x) \quad \forall g \in V^*, \forall x \in V$ es un isomorfismo.*

Demostración.

Debemos mostrar que φ es lineal y biyectiva.

i) Probemos que φ es lineal.

Sean $x, y \in V$, $h_1, h_2 \in V^{**}$ tales que $h_1(g) = g(x)$, $h_2(g) = g(y) \quad \forall g \in V^*$. Entonces

$$(h_1 + h_2)(g) = h_1(g) + h_2(g) = g(x) + g(y) = g(x + y).$$

Además, de la definición de la función φ se tiene $\varphi(x) = h_1$, $\varphi(y) = h_2$, $\varphi(x + y) = h_3$ con

$$h_3(g) = g(x + y) = (h_1 + h_2)(g) \quad \forall g \in V^*,$$

luego

$$\varphi(x + y) = h_1 + h_2 = \varphi(x) + \varphi(y).$$

Sea $\alpha \in \mathbb{K}$, de la definición de la función h_1 y de φ se tiene

$$\begin{aligned} h_1(\alpha g) &= \alpha g(x) = \alpha h_1(g) \quad \forall g \in V^*, \\ \varphi(\alpha x) &= \alpha h_1 = \alpha \varphi(x) \quad \forall x \in V. \end{aligned}$$

ii) Mostremos que φ es biyectiva.

Comencemos con la inyectividad de φ. Sea $B_v = \{v_1, \ldots, v_n\}$ una base ordenada de V y $B_* = \{f_1, \ldots, f_n\}$ la base dual de V^*. Probemos que $\ker(\varphi) = \{0\}$..

Sea $x \in \ker(\varphi)$. Entonces $\varphi(x) = 0$. Supongamos que $x \neq 0$. Existen $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ tales que $x = \sum_{i=1}^{n} \alpha_i v_i$ y como $x \neq 0$, existe algún j tal que $\alpha_j \neq 0$. Sea $h \in V^{**}$ definido por $h(g) = g(x)$ $\forall g \in V^*$. En particular, para $g = f_j$ se tiene

$$h(f_j) = f_j(x) = f_j\left(\sum_{i=1}^{n} \alpha_i v_i\right) = \sum_{i=1}^{n} \alpha_i f_j(v_i) = \alpha_j \neq 0.$$

Así, $x \neq 0 \Rightarrow h \neq 0$ o sea $h = 0 \Rightarrow x = 0$, de donde $0 = g(0) = h(g) \quad \forall g \in V^*$, por lo tanto

$$0 = \varphi(x) \Rightarrow x = 0,$$

y en consecuencia φ es inyectiva.

Probemos que φ es sobreyectiva. De la relación entre las dimensiones del núcleo y la imagen o recorrido de una aplicación lineal en espacios de dimensión finita, se tiene

$$\dim(\ker(\varphi)) + \dim(\mathrm{Rec}(\varphi)) = n,$$

y como $\ker(\varphi) = \{0\}$, se sigue que $\dim(\mathrm{Rec}(\varphi)) = n$ y siendo $\mathrm{Rec}(\varphi) \subset V^{**}$ se sigue que $V^{**} = \mathrm{Rec}(\varphi)$. consecuentemente φ es biyectiva. De i) , ii) y iii) se concluye que φ es un isomorfismo. ■

2.2. Interpolación polinomial

En el capítulo primero ya se trataron dos problemas de interpolación polinomial, el primero mediante funciones interpolantes constantes a trozos y el segundo mediante funciones interpolantes que son funciones afines a trozos. En esta sección, ampliamos lo dicho, más aún, nuestra primera tarea será demostrar la existencia de la función interpolante, luego construir de manera general los polinomios de interpolación de Lagrange y se concluye con el estudio del error. Para ello aplicaremos los resultados previos de los espacios duales.

Existencia del polinomio interpolante de Lagrange.

Sean a, $b \in \mathbb{R}$ con $a < b$ y f una función definida en $[a,b]$ en $\mathbb{R}$. Suponemos que el valor numérico de f es únicamente conocido en $n+1$ puntos distintos $x_i \in [a,b] \quad i = 0,\ldots,n$, y sean $y_i = f(x_i), \quad i = 0,\ldots,n$ tales valores. Suponemos que $a = x_0 < x_1 < \ldots < x_n = b$. Esta información se recoge en el conjunto

$$S = \left\{(x_i, y_i) \in \mathbb{R}^2 \mid i = 0,\ldots,n\right\}$$

denominado conjunto de puntos base.

La interpolación en un método de aproximación que permite construir una función φ de $[a,b]$ en $\mathbb{R}$ tal que

$$\varphi(x_i) = f(x_i) = y_i \quad i = 0,1,\ldots,n,$$

y para todo $x \in [a,b]$, $\varphi(x)$ es un valor aproximado de $f(x)$ llamado valor interpolado de $f(x)$. La función φ se llama interpolante de f. Más aún, si $\varepsilon(x)$ denota el error cometido en la interpolación, se tiene

$$f(x) = \varphi(x) + \varepsilon(x) \quad x \in [a,b],$$

la función ε definida sobre $[a,b]$ se llama error de interpolación.

Designamos con $C([a,b])$ el espacio de las funciones reales continuas en $[a,b]$. Este espacio, como ya se ha señalado anteriormente, es de dimensión infinita. Sea $V = \mathbb{K}_n[\mathbb{R}]$ el espacio de polinomios de grado $\leq n$. Se designa con $\tau = \{a = x_0, x_1, \ldots, x_n = b\}$ con $x_{i-1} < x_i \quad i = 1,\ldots,n$ una partición de $[a,b]$. Consideremos el problema siguiente:

$$\text{dado } f \in C([a,b]), \text{ hallar un polinomio } P \in \mathbb{K}_n[\mathbb{R}] \text{ tal que } P(x_i) = f(x_i) \quad i = 0,1,\ldots,n.$$

Denotamos con $B_v = \{v_0, v_1, \ldots, v_n\}$ la base canónica de $\mathbb{K}_n[\mathbb{R}]$, donde $v_0(x) = 1, \ldots, v_n(x) = x^n \;\; x \in \mathbb{R}$.

Definimos $n+1$ funcionales f_i sobre $\mathbb{K}_n[\mathbb{R}]$ como sigue:

$$f_i(P) = P(x_i) \quad i = 0,\ldots,n, \;\; P \in \mathbb{K}_n[\mathbb{R}].$$

Entonces cada funcional f_i es lineal sobre $\mathbb{K}_n[\mathbb{R}]$. Mas aún, $\{f_0,\ldots,f_n\}$ es una base del espacio dual $(\mathbf{K}_n[\mathbb{R}])^*$.

Mostremos que $\{f_0,\ldots,f_n\}$ es linealmente independiente.

Sean $\lambda_0,\ldots,\lambda_n \in \mathbb{R}$ y supongamos que $\lambda_0 f_0 + \ldots + \lambda_n f_n = 0$, esto es

$$\lambda_0 f_0(P) + \ldots + \lambda_n f_n(P) = 0 \quad \forall P \in \mathbb{K}_n[\mathbb{R}],$$

en particular para los elementos de la base B se $\mathbb{K}_n[\mathbb{R}]$, obtenemos el siguiente sistema de ecuaciones:

$$\begin{aligned}
\lambda_0 f_0(v_0) + \cdots + \lambda_n f_n(v_0) = 0 &\Longleftrightarrow \lambda_0 + \cdots + \lambda_n = 0,\\
\lambda_0 f_0(v_1) + \cdots + \lambda_n f_n(v_1) = 0 &\Longleftrightarrow \lambda_0 x_0 + \cdots + \lambda_n x_n = 0,\\
&\vdots\\
\lambda_0 f_0(v_n) + \cdots + \lambda_n f_n(v_n) = 0 &\Longleftrightarrow \lambda_0 x_0^n + \cdots + \lambda_n x_n^n = 0,
\end{aligned}$$

que en forma matricial se expresa como sigue

$$\begin{bmatrix} 1 & \cdots & 1 \\ x_0 & \cdots & x_n \\ \vdots & & \vdots \\ x_0^n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

La matriz $A = \begin{bmatrix} 1 & \cdots & 1 \\ x_0 & \cdots & x_n \\ \vdots & & \vdots \\ x_0^n & \cdots & x_n^n \end{bmatrix}$ se le conoce como matriz de Gram. Como $a = x_0, x_1, \ldots, x_n = b$ son puntos distintos del intervalo $[a, b]$, las columnas de la matriz A son linealmente independientes por lo tanto el rango de la matriz A es $n+1$, con lo que el sistema de ecuaciones:

$$\overrightarrow{\lambda} \in \mathbb{R}^n \text{ tal que } A\overrightarrow{\lambda} = 0,$$

con A la matriz de Gram, tiene una única solución $\overrightarrow{\lambda} = \overrightarrow{0}$. Consecuentemente

$$\lambda_0 f_0 + \ldots + \lambda_n f_n = 0 \Rightarrow \lambda_i = 0 \quad i = 0, 1, \ldots, n,$$

que prueba que el conjunto $\{f_0, \ldots, f_n\}$ es linealmente independiente y siendo

$$\dim \mathbb{K}_n[\mathbb{R}] = \dim (\mathbb{K}_n[\mathbb{R}])^* = n+1,$$

resulta que $\{f_0, \ldots, f_n\}$ es una base de $\mathbb{K}_n[\mathbb{R}]$.

Determinemos una base $B = \{P_0, \ldots, P_n\}$ de $\mathbb{K}_n[\mathbb{R}]$ tal que B_* sea la base dual de B. Esta debe satisfacer la condición $f_i(P_j) = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j, \end{cases}$ y por la definición de cada f_i, se tiene $f_i(P_j) = P_j(x_i)$, de donde $P_j(x_i) = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j, \end{cases}$ que se conoce como condición de interpolación.

Se definen los polinomios $P_0, P_1, \ldots, P_n$ de $\mathbb{K}_n[\mathbb{R}]$ como sigue:

$$\begin{aligned} P_0(x) &= \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)} \quad x \in \mathbb{R}, \\ P_1(x) &= \frac{(x-x_0)(x-x_2)\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)\ldots(x_1-x_n)} \quad x \in \mathbb{R}, \\ &\vdots \\ P_n(x) &= \frac{(x-x_0)(x_0-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})} \quad x \in \mathbb{R}. \end{aligned}$$

Note que en el polinomio P_0 no figura en el numerador el término $x - x_0$, en P_1 no figura el término $x - x_1$, así sucesivamente, en P_n no figura el término $x - x_n$. Además,

$$\begin{aligned} P_0(x_0) &= 1, \quad P_0(x_j) = 0 \quad \text{si } j = 1, \ldots, n, \\ P_1(x_1) &= 1, \quad P_1(x_j) = 0 \quad \text{si } j = 0, 2, \ldots, n, \\ &\vdots \\ P_n(x_n) &= 1, \quad P_n(x_j) = 0 \quad \text{si } j = 0, \ldots, n-1. \end{aligned}$$

Los polinomios $P_0, P_1, \ldots, P_n$ son linealmente independientes, por lo tanto forman una base de $\mathbb{K}_n[\mathbb{R}]$. Estos polinomios se llaman polinomios de interpolación de Lagrange.

Dado $P \in \mathbb{K}_n[\mathbb{R}]$, existen $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$ tales que $P = \sum_{k=0}^{n} \alpha_k P_k$, y para todo $x \in \mathbb{R}$, $P(x) = \sum_{k=0}^{n} \alpha_k P_k(x)$, particularmente para $x = x_j$, $j = 0, \ldots, n$,

$$P(x_j) = \sum_{k=0}^{n} \alpha_k P_k(x_j) = \alpha_j,$$

de donde

$$P(x) = \sum_{k=0}^{n} P(x_k) P_k(x).$$

Sea $f \in C([a,b])$. En el problema de la interpolación polinomial, buscamos un polinomio $P \in \mathbb{K}_n[\mathbb{R}]$ tal que

$$P(x_j) = f(x_j) \quad j = 0, \ldots, n.$$

Definimos

$$\widehat{P}(x_i) = \sum_{k=0}^{n} f(x_k) P_k(x) \quad x \in \mathbb{R}.$$

Se tiene $\widehat{P}(x_i) = f(x_i) \quad i = 0, \ldots, n$, es decir $\widehat{P}$ es el polinomio interpolante de f.

Definición 2 *El operador de interpolación de Lagrange* Π *se define como sigue:*

$$\Pi : \begin{cases} C([a,b]) \to \mathbb{K}_n[\mathbb{R}] \\ f \to \Pi(f), \end{cases}$$

donde $\Pi(f) = \widehat{P} = \sum_{j=0}^{n} f(x_j) P_j$.

Sea $\widehat{x} \in [a,b]$ y $F \in (\mathbb{K}_n[\mathbb{R}])^*$ el funcional definido por $F(P) = P(\widehat{x}) \quad \forall P \in \mathbb{K}_n[\mathbb{R}]$. Para cada $f \in C([a,b])$, de la composición de funciones, se tiene el siguiente resultado:

$$(F \circ \Pi)(f) = F(\Pi(f)) = F\left(\widehat{P}\right) = F\left(\sum_{j=0}^{n} f(x_j) P\right) = \sum_{j=0}^{n} f(x_j) F(P_j) = \sum_{j=0}^{n} f(x_j) P_j(\widehat{x}).$$

Así, $G = F \circ \Pi$ es un funcional lineal sobre $C([a,b])$. Escribiremos

$$G(f) = \sum_{j=0}^{n} f(x_j) P_j(\widehat{x})$$

al valor interpolado de f en el punto arbitrario $\widehat{x} \in [a,b]$.

Este método de interpolación se conoce como interpolación polinomial de Lagrange.

Nota: En la práctica los polinomios de interpolación de Lagrange no son muy utilizados cuando el número de puntos base $(x_i, y_i) \quad i = 0, \ldots, n$ es grande (más aún cuando los x_i son muy cercanos entre sí) ya que el grado del polinomio interpolante de Lagrange $\widehat{P}$ es igualmente grande dando lugar a la presencia de oscilaciones que afectan los resultados. Los polinomios de interpolación de Lagrange más utilizados son de grados $n = 1, 2, 3$ y 4.

En la figura siguiente se muestra la gráfica de una función f defnida en $[a,b]$ (línea continua), que suponemos es no negativa; y, la de un polinomio de interpolación de Lagrange construida sobre una partición $\tau(m)$ de $[a,b]$ (línea cortada).

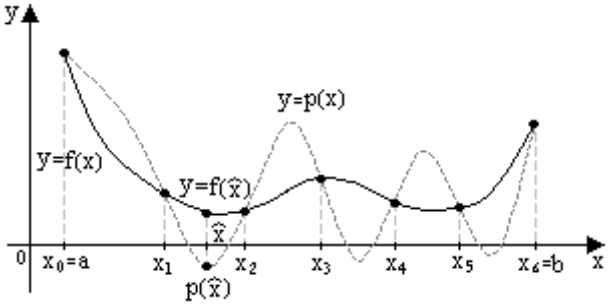


Figura 11

Note que en el punto $\widehat{x} \in [a,b[\,,\ f(\widehat{x}) > 0$ mientras que $p(\widehat{x}) < 0$. Las fuertes oscilaciones de la función p conduce a resultados falsos.

Sean $n \in \mathbb{Z}^+,\ \tau(n) = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$. Se denota con $h_j = x_j - x_{j-1}$ $j = 1, \ldots, n,\ \widehat{h} = \text{máx}\{h_j \mid j = 1, \ldots, n\}$. Cuando $\tau(n)$ es la partición uniforme se tiene $h = \dfrac{b-a}{n}$, $x_j = jh \quad j = 0, 1, \ldots, n,\ \widehat{h} = h$.

Se define la función ω en $[a,b]$ como sigue: $\omega(x) = \prod_{j=0}^{n} (x - x_j) \quad x \in [a,b]$. Claramente ω es un polinomio de grado $n+1$, y $\omega(x_i) = 0 \quad i = 0, 1, \ldots, n$.

Error de interpolación

Teorema 3 *Sean $n \in \mathbb{Z}^+,\ f \in C^{n+1}([a,b]),\ \tau(n) = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$. Entonces, para cada $\widehat{x} \in [a,b]$, existe $\xi \in [a,b]$ tal que*

$$f(\widehat{x}) - p(\widehat{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (\widehat{x} - x_j),$$

donde p es el polinomio de interpolación de Lagrange.

Demostración. Sea p el polinomio de interpolación de Lagrange. Se tiene $p(x_j) = f(x_j) \quad j = 0, 1, \ldots, n$. Además, el grado del polinomio p es n.

Sea $\widehat{x} \in [a,b]$ un punto dado.

i) Si $\widehat{x} = x_j$ para algún j, esto es, $\widehat{x}$ es un punto de la partición $\tau(n)$ entonces $p(x_j) = f(x_j)$ y en este caso $f(\widehat{x}) - p(\widehat{x}) = 0$.

ii) Supongamos que $\widehat{x} \neq x_j \quad \forall j = 0, 1, \ldots, n$. Determinemos una constante k tal que la función θ definida como $\theta(x) = f(x) - p(x) - k\omega(x) \quad x \in [a,b]$, se anule para $x = \widehat{x}$, es decir, $\theta(\widehat{x}) = 0$.

Para los puntos de la partición se tiene $\theta(x_j) = f(x_j) - p(x_j) - k\omega(x_j) = 0 \quad j = 0, 1, \ldots, n$, es decir que la función θ tiene a cada x_j como raíz y como se busca k de modo que $\theta(\widehat{x}) = 0$, entonces θ tiene a $n+2$ raíces en el intervalo $[a,b]$. Por el teorema de Rolle, la derivada θ' tiene $n+1$ raíces en el intervalo $[a,b]$, la derivada segunda θ'' tiene n raíces en el intervalo $[a,b]$, así sucesivamente, $\theta^{(n+1)}$ tiene una raíz en el intervalo $[a,b]$ y sea ξ tal raíz, esto es, $\theta^{(n+1)}(\xi) = 0$. Por otro lado, para cada $x \in [a,b]$ se tiene

$$\begin{aligned}
\theta'(x) &= f'(x) - p'(x) - k\omega'(x), \\
\theta''(x) &= f''(x) - p''(x) - k\omega''(x), \\
&\vdots \\
\theta^{(n+1)}(x) &= f^{(n+1)}(x) - p^{(n+1)}(x) - k\omega^{(n+1)}(x).
\end{aligned}$$

Como p es un polinomio de interpolación de grado n, entonces $p^{(n+1)}(x) = 0$. Además, ω es un polinomio de grado $n+1$, luego $\omega^{(n+1)}(x) = (n+1)!$. Por lo tanto,

$$\theta^{(n+1)}(x) = f^{(n+1)}(x) - k\omega^{(n+1)}(x) = f^{(n+1)}(x) - k(n+1)! \quad x \in [a,b].$$

En particular, para $x = \xi$ se tiene $\theta^{(n+1)}(\xi) = 0$ y en consecuencia

$$0 = \tau^{(n+1)}(\xi) = f^{(n+1)}(\xi) - k(n+1)!$$

de donde $k = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}$ y la θ queda definida como sigue:

$$\theta(x) = f(x) - p(x) - \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x) \quad x \in [a,b].$$

Puesto que la constante k se elige de modo que $\theta(\widehat{x}) = 0$, resulta

$$0 = \theta(\widehat{x}) = f(\widehat{x}) - p(\widehat{x}) - \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(\widehat{x})$$

de donde $f(\widehat{x}) - p(\widehat{x}) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}\omega(\widehat{x})$. ■

El resultado dado en el teorema se conoce como fórmula de error de interpolación de Lagrange, que lo notamos $\xi(\widehat{x})$. Así,

$$\xi(\widehat{x}) = f(\widehat{x}) - p(\widehat{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(\widehat{x}),$$

donde $\xi \in [a,b]$ y $\omega(\widehat{x}) = \prod_{j=0}^{n}(\widehat{x} - x_j)$.

Ejemplos de polinomios de interpolación de Lagrange

1. **Interpolante constante a trozos.**

   Para $n = 0$ se considera $\tau = \left\{\frac{a+b}{2}\right\}$. En este caso, $\mathbb{K}_0[\mathbb{R}]$ es el espacio constituido por todas las funciones constantes en todo $\mathbb{R}$, esto es,

   $$P \in \mathbb{K}_0[\mathbb{R}] \Leftrightarrow P(x) = c \quad \forall x \in \mathbb{R},$$

   para alguna constante $c \in \mathbb{R}$.

   La base de $\mathbb{K}_0[\mathbb{R}]$ está constituida por $B = \{v_0\}$ con $v_0(x) = 1 \ \ \forall x \in \mathbb{R}$ y la base dual B_* de B es $B_* = \{f_0\}$ con $f_0(P) = P(x) \ \ \forall P \in \mathbb{K}_0[\mathbb{R}]$.

   Para $\widehat{x} \in [a,b]$ y $f \in C([a,b])$, de la definición del valor interpolado de f en $\widehat{x} \in [a,b]$ se tiene

   $$G(f) = f\left(\frac{a+b}{2}\right)P_0(\widehat{x}) = f\left(\frac{a+b}{2}\right).$$

   El polinomio interpolante de f es la función p definida en $[a,b]$ como

   $$p(x) = f\left(\frac{a+b}{2}\right) \quad x \in [a,b].$$

   En la figura siguiente se muestran las gráficas de la función f y de su interpolante lagrangeana p.

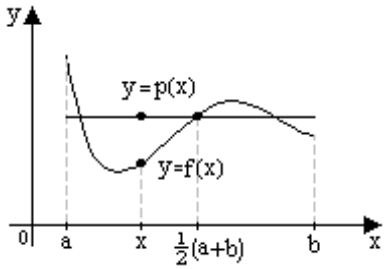


Figura 12

Sea $m \in \mathbb{Z}^+$ y $\tau(m) = \{x_0 = a, x_1, \ldots, x_m = b\}$ con $x_{i-1} < x_i \quad i = 1, \ldots, m$ una partición de $[a,b]$. Se pone $h_i = x_i - x_{i-1}$ la longitud del intervalo $[x_{i-1}, x_i] \quad i = 1, \ldots, m$, y, $\widehat{h} = \text{máx}\{h_i \mid i = 1, \ldots, m\}$. Sea f una función continua en $[a,b]$. La función interpolante p de f está definida como

$$\begin{cases} p(x) = \sum_{i=1}^{m} f(t_i)\,\chi_i(x) & x \in [a,b[\,, \\ p(b) = f(t_m)\,, \end{cases}$$

donde χ_i es la función indicatriz del intervalo $[x_{i-1}, x_i[$ definida como $\chi_i(x) = \begin{cases} 0, & \text{si } x \notin [x_{i-1}, x_i[\,, \\ 1, & \text{si } x \in [x_{i-1}, x_i[\,, \end{cases}$
$i = 1, \ldots, m$, $t_i = x_{i-1} + \dfrac{1}{2}h_i$ el punto medio del intervalo $[x_{i-1}, x_i] \quad i = 1, \ldots, m$.

En la figura siguiente se muestra la gráfica de la función f y de su interpolante p, con $m = 5$

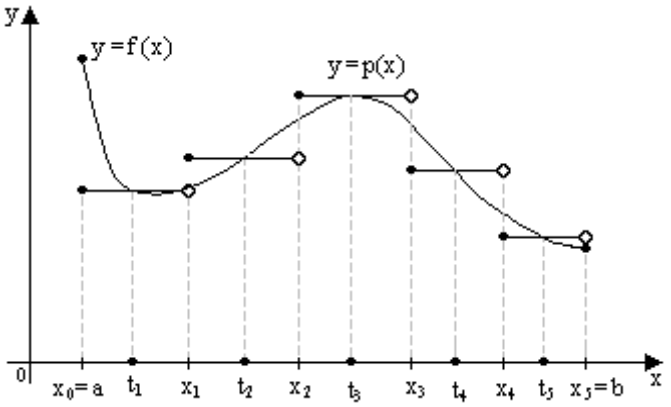


Figura 13

De la fórmula del error de interpolación de Lagrange, sea $f \in C^1([a,b])$ y la partición del intervalo $[a,b]$ se reduce al solo punto medio $\dfrac{a+b}{2}$ de $[a,b]$, se tiene la siguiente estimación del error para $\widehat{x} \in [a,b]$;

$$\xi(\widehat{x}) = f(\widehat{x}) - p(\widehat{x}) = f'(\xi)\left(\widehat{x} - \frac{a+b}{2}\right)$$

para algún $\xi \in [a,b]$.

Sean $m \in \mathbb{Z}^+$, $\tau(m)$ una partición del intervalo $[a,b]$, aplicando el resultado precedente a $\widehat{x} \in [x_{j-1}, x_j] \quad j = 1, \ldots, m$ se tiene

$$\varepsilon_j(\widehat{x}) = f(\widehat{x}) - f(t_j) = f'\left(\xi_j\right)(\widehat{x} - t_j)$$

con $\xi_j \in [x_{j-1}, x_j]$. Luego, si $f \in C^1([a,b])$, existe $M > 0$ tal que $\mid f'(x) \mid \leq M \quad \forall x \in [a,b]$ y en consecuencia

$$\mid f(\widehat{x}) - f(t_j) \mid \leq \mid f'\left(\xi_j\right) \mid\mid \widehat{x} - t_j \mid \leq Mh \underset{h \longrightarrow 0}{\longrightarrow} 0\,.$$

Ejemplo

Supongamos que f es la función definida como $f(x) = 1 + e^x \quad x \in [0,2]$, $m = 10$ y $\tau(10)$ la partición uniforme de $[0,2]$; esto es, $h = \dfrac{2}{10} = 0{,}2$ y $\chi_j = jh = 0{,}2j \quad j = 0, 1, \ldots, 10$, $h_j = h = 0{,}2$ $j = 1, \ldots, 10$. La función interpolante de f está definida como

$$\begin{cases} p(x) = \displaystyle\sum_{j=1}^{10} f(t_j)\,\chi_j(x) & x \in [0,2[\,, \\ p(2) = f(t_{10})\,, \end{cases}$$

donde $t_j = x_{j-1} + \dfrac{1}{2}h_j = x_{j-1} + 0{,}1 \quad j = 1, \ldots, 10$.

Para $x = 0{,}85$ se tiene

$$p(0{,}85) = \sum_{j=1}^{10} f(t_j)\,\chi_j(0{,}85) = 1 + e^{0{,}9},$$

pués $0{,}85 \in [0{,}8, 1[$ para $i = 5$, $t_i = x_{i-1} + 0{,}5 = 0{,}9$, $f(0{,}9) = 1 + e^{0{,}9}$.

2. **Interpolantes afines a trozos.**

Para $n = 1$, se tiene $\tau = \{a = x_0, x_1 = b\}$. Entonces, los polinomios de interpolación de Lagrange son P_0, $P_1 \in \mathbb{K}_1[\mathbb{R}]$ definidos como sigue:

$$P_0(x) = \frac{x - x_1}{x_0 - x_1} \quad x \in \mathbb{R}, \qquad P_1(x) = \frac{x - x_0}{x_1 - x_0} \quad x \in \mathbb{R}.$$

En la figura siguiente se muestran las gráficas de los polinomios P_0, P_1.

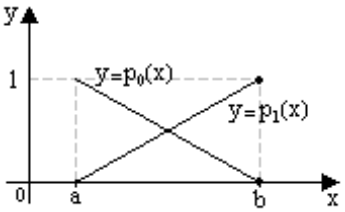


Figura 14

Para $f \in C([a,b])$, por la definición del valor interpolado de f en $\widehat{x} \in [a,b]$ se tiene

$$G(f) = \sum_{k=0}^{1} f(x_k) P_k(x) = f(a) P_0(x) + f(b) P_1(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \qquad x \in [a,b],$$

es el valor interpolado de f en el punto x. Note que el polinomio interpolante de f está dado por

$$\Pi(f) = p(x) = \sum_{k=0}^{n} f(x_k) P_k(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \quad x \in [a,b],$$

que representa la ecuación del segmento de recta que une los puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$.
En la figura siguiente se muestra las gráficas de f y del polinomio interpolante p.

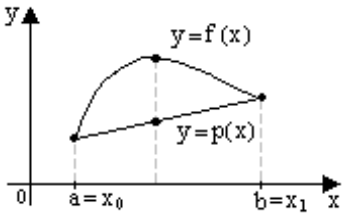


Figura 15

Para $\widehat{x} \in [a,b]$, el error de interpolación $\varepsilon(\widehat{x})$ está definido como

$$\varepsilon(\widehat{x}) = f(\widehat{x}) - p(\widehat{x}) = \frac{f''(\xi)}{2!}(\widehat{x} - a)(\widehat{x} - b),$$

donde $\xi \in [a,b]$ es elegido apropiadamente.

Sean $m \in \mathbb{Z}^+$ y $\tau(m) = \{x_0 = a, x_1, \ldots, x_m = b\}$ con $x_{i-1} < x_i \quad i = 1, \ldots, m$, una partición de $[a,b]$, se pone $h_j = x_j - x_{j-1}$ y $\widehat{h} = \text{Max}\{h_j \mid j = 1, \ldots, m\}$. Apliquemos los resultados precedentes a cada subintervalo $[x_{i-1}, x_i] \quad i = 1, \ldots, m$. Tenemos que la función interpolante v_h de f en el intervalo $[a,b]$ está definida como

$$v(x) = \sum_{i=0}^{m} f(x_i) \varphi_i(x) \quad x \in [a,b],$$

y las funciones $\varphi_0, \varphi_1, \ldots, \varphi_m$ están definidas como sigue:

$$\varphi_0(x) = \begin{cases} -\dfrac{x - x_1}{h_1},\ x \in [a, x_1], \\ 0,\ x \in [a, b] \setminus [a, x_1], \end{cases} \qquad \varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i},\ x \in [x_{i-1}, x_i], \\ -\dfrac{x - x_{i-1}}{h_{i-1}},\ x \in]x_i, x_{i+1}] \\ 0,\ x \in [a, b] \setminus [x_{i-1}, x_i] \end{cases} \quad i = 1, \ldots, m-1,$$

$$\varphi_m(x) = \begin{cases} \dfrac{x - x_{m-1}}{h_m},\ x \in [x_{m-1}, b], \\ 0,\ x \in [a, b] \setminus [x_{m-1}, b]. \end{cases}$$

En las figuras siguientes se muestran las gráficas de $\varphi_0,\ \varphi_1,\ \varphi_5$, donde la partición de $[a, b]$ está constituida por $\tau(5) = \{x_0 = a, x_1, x_2, x_3, x_4, x_5 = b\}$.

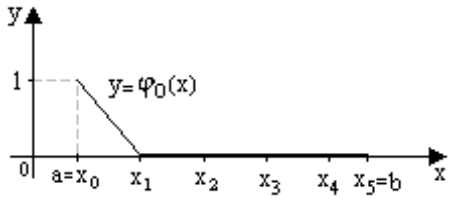


Figura 16

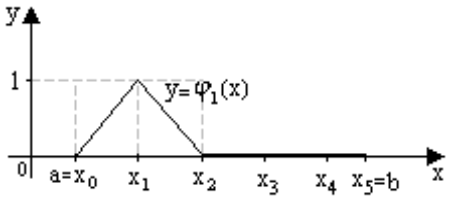


Figura 17

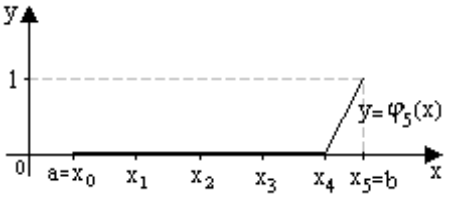


Figura 18

Note que $\varphi_i(x_j) = \begin{cases} 0, & si\ \ i \neq j \\ 1, & si\ \ i = j, \end{cases}$ y $0 \leq \varphi_i(x) \leq 1 \quad x \in [a, b]$. Las funciones $\varphi_2, \ldots, \varphi_{m-1}$ son similares a la función φ_1.

A las funciones $\varphi_0, \ldots, \varphi_m$ se les denomina funciones techo.

En la figura siguiente se muestra la gráfica de una función continua v definida en el intervalo $[0, a]$ con $a > 0$, y la de una función interpolante v_h (segmentos de recta) de v. Se muestran también los

puntos de la partición de $\tau(m)$ de $[0,a]$.

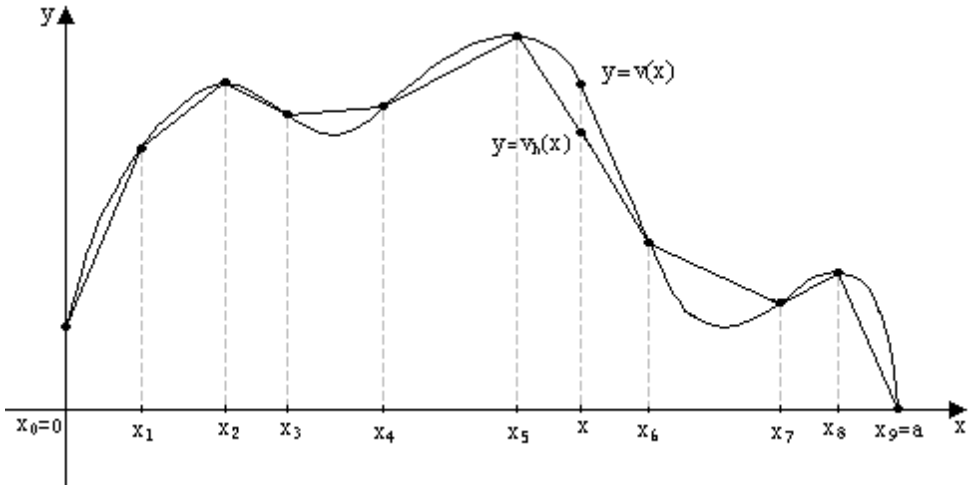


Figura 19

El error de interpolación se estima a partir de la fórmula

$$\varepsilon(\widehat{x}) = f(\widehat{x}) - p(\widehat{x}) = \frac{f''(\xi_j)}{2!}(\widehat{x} - x_{j-1})(\widehat{x} - x_j),$$

donde $\widehat{x} \in [x_{j-1}, x_j]$ y $\xi_j \in [x_{j-1}, x_j]$ elegido apropiadamente. Además,

$$|f(\widehat{x}) - p(\widehat{x})| \leq \frac{h_j^2}{2}|f''(\xi_j)| \leq \frac{M}{2}h^2 \underset{h\longrightarrow 0}{\longrightarrow} 0,$$

donde $M = \underset{x\in[a,b]}{\text{Max}}|f''(x)|$.

3. **Interpolantes cuadráticos a trozos.**

Para $n = 2$, una partición del intervalo $[a,b]$ es $\tau = \{a = x_0,\ x_1,\ x_2 = b\}$. Los polinomios de interpolación de Lagrange $\varphi_0,\ \varphi_1,\ \varphi_2 \in \mathbb{K}_2[\mathbb{R}]$ están definidos como a continuación se indican:

$$\begin{aligned}
\varphi_0(x) &= \frac{(x-x_1)(x-b)}{(a-x_1)(a-b)} \quad x \in \mathbb{R}, \\
\varphi_1(x) &= \frac{(x-a)(x-b)}{(x_1-a)(x_1-b)} \quad x \in \mathbb{R}, \\
\varphi_2(x) &= \frac{(x-a)(x-x_1)}{(b-a)(b-x_1)} \quad x \in \mathbb{R}.
\end{aligned}$$

En la figura siguiente se muestra las gráficas de las funciones $\varphi_0,\ \varphi_1,\ \varphi_2$ restringidas al intervalo $[a,b]$.

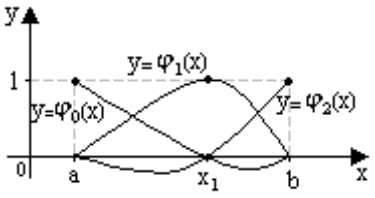


Figura 20

Sea $f \in C([a,b])$. El polinomio interpolante de f está dado por

$$\Pi(f) = p(x) = \sum_{k=0}^{2} f(x_k)\varphi_k(x) = f(a)\varphi_0(x) + f(x_1)\varphi_1(x) + f(b)\varphi_2(x) \quad x \in \mathbb{R}.$$

En la figura siguiente se muestra la gráfica de una función f y de su interpolante p.

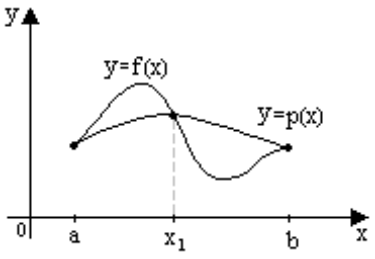


Figura 21

Para $\widehat{x} \in [a,b]$ y $f \in C([a,b])$, el valor interpolado de f está definido como sigue:

$$G(f) = p(\widehat{x}) = f_a \varphi_0(\widehat{x}) + f(x_1)\varphi_1(\widehat{x}) + f(b)\varphi_2(\widehat{x}).$$

Sea $f \in C^3([a,b])$. El error de interpolación se define como

$$\varepsilon(\widehat{x}) = f(\widehat{x}) - p(\widehat{x}) = \frac{f'''(\xi)}{\xi!}(\widehat{x}-a)(\widehat{x}-x_1)(\widehat{x}-b),$$

donde $\xi \in [a,b]$.

Sean $n \in \mathbb{Z}^+$ y $\tau(n) = \{a = x_0, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$. Se pone $h_j = x_j - x_{j-1}$ $j = 1, \ldots, n$ y $\widehat{h} = \text{Max}\{h_j \mid j = 1, \ldots, n\}$. En el caso de una partición uniforme, se tiene $h = \dfrac{b-a}{n}$, $x_j = jh \quad j = 0, 1, \ldots, n$ y $\widehat{h} = h$.

La función interpolante v de f está definida como

$$v(x) = \sum_{i=0}^{n} f(x_i)\psi_i(x) \qquad x \in [a,b],$$

donde $\psi_i \quad i = 0, 1, \ldots, n$ son funciones que se obtienen de $\varphi_0,\ \varphi_1,\ \varphi_2$ aplicadas a cada intervalo $[x_{i-1}, x_{i+1}]$ y que satisfacen las condiciones de interpolación $\psi_i(x_j) = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j. \end{cases}$ A continuación se muestran las gráficas de las tres primeras funciones $\psi_0,\ \psi_1,\ \psi_2$.

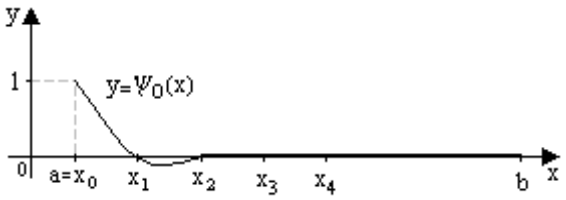


Figura 22

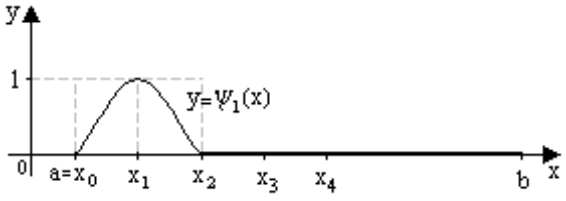


Figura 23

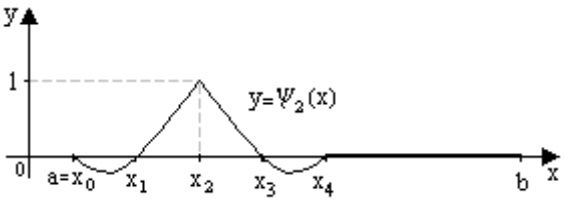


Figura 24

Note que $\psi_0,\ \psi_1,\ \psi_2$ están definidas en $[x_0, x_2]$ como sigue:

$$\psi_0(x) = \begin{cases} \dfrac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, & x \in [x_0, x_2], \\ 0, \text{ en otro caso,} \end{cases}$$

$$\psi_1(x) = \begin{cases} \dfrac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, & \text{si } x \in [x_0, x_2], \\ 0, \text{ en otro caso,} \end{cases}$$

$$\psi_2(x) = \begin{cases} \dfrac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}, & \text{si } x \in [x_0, x_2], \\ \dfrac{(x-x_3)(x-x_4)}{(x_2-x_3)(x_2-x_4)}, & \text{si } x \in]x_2, x_4[, \\ 0, \text{ en otro caso.} \end{cases}$$

Si $f \in C^3([a,b])$, $M = \text{Max}\{|f'''(x)| \mid x \in [a,b]\}$ y $\widehat{x} \in [a,b]$, entonces se tiene la siguiente estimación del error:

$$\varepsilon_j = f(\widehat{x}) - v(\widehat{x}) = \frac{f'''(\xi_j)}{3!}(\widehat{x}-x_{j-1})(\widehat{x}-x_j)(\widehat{x}-x_{j+1})$$

donde $\widehat{x} \in [x_{j-1}, x_{j+1}]$ $\quad j = 1, \ldots, n-1$; y, en consecuencia

$$|f(\widehat{x}) - v(\widehat{x})| \leq \frac{M}{6}\widehat{h}^3 \underset{h \longrightarrow 0}{\longrightarrow} 0\,.$$

2.3. Operadores de diferencias finitas y derivación numérica

Denotamos con $C([a,b])$ es espacio de funciones reales continuas en $[a,b]$. Para $m \in \mathbb{Z}^+$, denotamos con $C^m([a,b])$ es espacio de las funciones reales tales que la derivada m-ésima es continua en $[a,b]$.

Sea $f \in C^2([a,b])$, $h \in \mathbb{R}$ con $h \neq 0$ tal que $\forall x \in]a,b[$, $x+h \in [a,b]$. En general, h es suficientemente pequeño. Las derivadas f' y f'' en $x \in]a,b[$ se definen como

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}, \qquad f''(x) = \lim_{h \to 0} \frac{f'(x+h) - f'(x)}{h}.$$

Para $h \neq 0$ suficientemente pequeño, las derivadas $f'(x)$ y $f''(x)$ se aproximan mediante los siguientes cocientes:

$$f'(x) \simeq \frac{f(x+h) - f(x)}{h}, \text{ y} \qquad f''(x) \simeq \frac{f'(x+h) - f'(x)}{h},$$

más aún, $f'(x)$ y $f''(x)$ se esciben como

$$f'(x) = \frac{f(x+h) - f(x)}{h} + w_1(x,h), \qquad f''(x) = \frac{f'(x+h) - f'(x)}{h} + w_2(x,h),$$

con $|w_1(x,h)| \underset{h \longrightarrow 0}{\longrightarrow} 0$, $|w_2(x,h)| \underset{h \to 0}{\longrightarrow} 0$. Es claro que cuando los residuos $w_1(x,h)$, $w_2(x,h)$ son suficientemente pequeños para h suficientemente pequeño, las dedivadas las podemos aproximar mediante los cocientes incrementales. Los numeradores de estos cocientes dan lugar a las denominadas diferencias finitas y por lo tanto a los operadores en diferencias finitas que a continuación se definen.

1. El operador de diferencia finita hacia adelante se nota y se define como sigue:

$$\Delta f(x) = f(x+h) - f(x).$$

2. El operador de diferencia finita hacia atrás se nota y se define como a continuación se indica:

$$\nabla f(x) = f(x) - f(x-h).$$

3. Operador de diferencia finita central de primer orden se nota y se define del modo siguiente:

$$\delta f(x) = f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right).$$

Aproximación de $f'(x)$.

En el capítulo primero se propuso un método de cálculo de la derivada primera $f'(x) \quad x \in]a,b[$. En esta parte, ampliamos dicho procedimiento de cálculo que incluye el error de aproximación. Además, veremos otros métodos similares de aproximación.

Supongamos que $f \in C^3([a,b])$. Por el desarrollo de Taylor, para $h > 0$ se tiene,

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(\xi_1) \quad \text{con} \quad \xi_1 \in [x, x+h], \\ f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(\xi_2) \quad \text{con} \quad \xi_2 \in [x-h, x], \end{aligned}$$

entonces,

$$\begin{aligned} \frac{\Delta f(x)}{h} &= \frac{f(x+h)-f(x)}{h} = f'(x) + \frac{h}{2!}f''(x) + \frac{h^2}{3!}f'''(\xi_1), \\ \frac{\nabla f(x)}{h} &= \frac{f(x)-f(x-h)}{h} = f'(x) - \frac{h}{2!}f''(x) + \frac{h^2}{3!}f'''(\xi_2), \\ \frac{\partial f(x)}{2h} &= \frac{f(x+h)-f(x-h)}{2h} = f'(x) + \frac{h^2}{3!}\left(f'''(\xi_1) + f'''(\xi_2)\right). \end{aligned}$$

Por hipótesis, f', f'', f''' son acotadas en el intervalo $[a,b]$, luego existen $M_1 > 0$, $M_2 > 0$, $M_3 > 0$ tales que $\mid f'(x) \mid \leq M_1$, $\mid f''(x) \mid \leq M_2$, $\mid f'''(x) \mid \leq M_3 \quad \forall x \in [a,b]$, y $M = \max\{M_1, M_2, M_3\}$, entonces

$$\begin{aligned} \left|\frac{\Delta f(x)}{h} - f'(x)\right| &= \mid \frac{f(x+h)-f(x)}{h} - f'(x) \mid = \mid \frac{h}{2!}f''(x) \mid \leq \frac{M}{2}h \underset{h \longmapsto 0}{\longrightarrow} 0, \\ \left|\frac{\nabla f(x)}{h} - f'(x)\right| &= \mid \frac{f(x)-f(x-h)}{h} - f'(x) \mid = \mid \frac{h}{2!}f''(x) \mid \leq \frac{M}{2}h \underset{h \longmapsto 0}{\longrightarrow} 0, \\ \left|\frac{\partial f(x)}{2h} - f'(x)\right| &= \mid \frac{f(x+h)-f(x-h)}{2h} - f'(x) \mid = \mid \frac{h^2}{3!}\left(f'''(\xi_1) + f'''(\xi_2)\right) \mid \leq \frac{M}{3}h^2 \underset{h \longmapsto 0}{\longrightarrow} 0. \end{aligned}$$

Se observa que las diferencias finitas centrales aproximan mejor la derivada $f'(x)$, es decir que para h suficientemente pequeño y no nulo, el término $\frac{M}{3}h^2$ va a cero más rápidamente que $\frac{M}{2}h$ cuando $h \longmapsto 0$.

Sea $f \in C^2([a,b])$, $x_0 \in]a,b[$ y $h \neq 0$. Con frecuencia se presenta el problema de calcular $f'(x_0)$, con f una función en la que resulta difícil calcular la derivada o que únicamente se conocen los puntos (x_0, y_0), (x_0+h, y_1) y se requiere aproximar $f'(x_0)$. En este último caso se asume que $y_0 = f(x_0)$, $y_1 = f(x_0+h)$.

Se define una aproximación de $f'(x_0)$ como el cociente

$$y_0' = \frac{y_1 - y_0}{h},$$

y se denomina derivada numérica mediante una diferencia finita progresiva. Se tiene la siguiente estimación

$$\left|f'(x_0) - y_0'\right| \leq \frac{M}{2}h \underset{h\to 0}{\longrightarrow} 0.$$

Si $y_0 = f(x_0)$, $y_1 = f(x_0 - h)$, se define una aproximación $f'(x_0)$ como el cociente

$$y_0' = \frac{y_0 - y_1}{h},$$

y se denomina derivada numérica mediante una diferencia finita regresiva. Tenemos la siguiente estimación

$$\left|f'(x_0) - y_0'\right| \leq \frac{M}{2}h \underset{h\to 0}{\longrightarrow} 0.$$

Si $y_0 = f(x_0 - h)$, $y_1 = f(x_0 + h)$, se define y_0' como

$$y_0' = \frac{y_1 - y_0}{h},$$

y se denomina derivada numérica mediante una diferencia finita central. Se tiene

$$\left|f'(x_0) - y_0'\right| \leq \frac{M}{3}h^2 \underset{h\to 0}{\longrightarrow} 0.$$

Ejemplo

Con el propósito de comparar las derivadas numéricas con la derivada de una función, asumimos que f es conocida.

Sea $f(x) = \exp(\operatorname{sen}\sqrt{x}) \quad x > 0$ y $x_0 = 2$. Entonces

$$f'(x) = \frac{\cos\sqrt{x}}{2\sqrt{x}} \exp\left(\operatorname{sen}\sqrt{x}\right) \qquad x > 0,$$

y en consecuencia $f'(2) = \dfrac{\cos\sqrt{2}}{2\sqrt{2}} \exp\left(\operatorname{sen}\sqrt{2}\right) \simeq 0{,}148048539.$

En la tabla siguiente se muestran aproximaciones de $f'(2)$ con diferencias finitas progresivas (cálculos realizados con una calculadora de bolsillo)

h	$x_0 + h$	$y_0 = f(x_0)$	$y_1 = f(x_0 + h)$	$y_0' = \dfrac{y_1 - y_0}{h}$	$\left\|f'(2) - y_0'\right\|$
$-0{,}001$	1,999	2,68522882	2,685080592	0,148228	$1{,}79461 \times 10^{-4}$
$-0{,}000001$	1,999999	2,68522882	2,685228672	0,148000	$4{,}8539 \times 10^{-5}$
0,001	2,001	2,68522882	2,685376689	0,147869	$1{,}79539 \times 10^{-4}$
0,000001	2,000001	2,68522882	2,685228968	0,148	$4{,}8539 \times 10^{-5}$

Note que a medida que $|h|$ se aproxima a cero, y_0' se aproxima a $f'(2)$ y el error $\left|f'(2) - y_0'\right|$ es cada vez más pequeño; sin embargo, con una calculadora de bolsillo, para h suficientemente pequeño y no nulo, se obtienen resultados como los siguientes: si $h = 0{,}00000001$, $x_0 + h = 2{,}00000001$, $y_1 = f(x_0 + h) = 2{,}685228822$, luego

$$y_0' = \frac{y_1 - y_0}{h} = \frac{2{,}685228822 - 2{,}68522882}{0{,}00000001} = 0{,}2,$$

que está muy alejado de $f'(2)$. Esto se debe a que y_0, y_1 son valores aproximados con 9 cifras de precisión que es lo que se obtiene de la calculadora. Para mejorar los resultados se deben calcular en al menos doble precisión, o sea con al menos 16 cifras de precisión que es lo que se obtiene en un computador personal Pentium I o más avanzados.

En la tabla siguiente se muestran aproximaciones de $f'(2)$ mediante el uso de diferencias finitas centrales. Los cálculos son realizados con una calculadora de bolsillo.

h	x_0-h	x_0+h	$y_0=f(x_0+h)$	$y_1=f(x_0+h)$	$y_0'=\dfrac{y_1-y_0}{2h}$	$\lvert f'(2)-y_0'\rvert$
0,001	1,999	2,001	2,685080592	2,685376689	0,1480485	$5{,}39\times10^{-8}$
0,0001	1,9999	2,0001	2,685214014	2,685243623	0,148045	$3{,}539\times10^{-6}$
0,00001	1,99999	2,00001	2,68522734	2,685230301	0,14805	$1{,}461\times10^{-6}$
0,000001	1,999999	2,000001	2,685228672	2,685228968	0,148	$4{,}8539\times10^{-5}$

Note que cuando h es muy pequeño, debido a los errores de redondeo y la representación en un punto fijo, el error tiende a aumentar ¿cuál es el valor de h a elegir para que $\left|f'(x_0)-\dfrac{f(x_0+h)-f(x_0-h)}{h}\right|$ sea muy aceptable? Con una calculadora de bosillo, obtener h para que la aproximación sea suficientemente buena (óptima) no es del todo evidente y depende de cada función f. En un computador personal se deben realizar los cáculos en al menos doble precisión y $\mid h\mid\neq 0$ suficientemente pequeño.

Aproximación de $f''(x)$.

Los operadores de diferencias finitas de orden superior se definen por recurrencia en el sentido de la composición de operadores:

$$\Delta^{k+1}=\Delta^k\circ\Delta,\qquad \nabla^{k+1}=\nabla^k\circ\nabla,\quad \delta^{k+1}=\delta^k\circ\delta\qquad k\in\mathbb{N},$$

donde $\Delta^0=\nabla^0=\delta^0=I$ operador identidad.

Además, podemos construir operadores mixtos como los siguientes: $\Delta\circ\nabla,\ \nabla\circ\Delta,\ \Delta\circ\delta,\ \delta\circ\nabla,\ldots,$ que se escriben simplemente como $\Delta\nabla,\ \nabla\Delta,\ \Delta\delta,\ \delta\nabla,\ldots$. De manera general, si $m,\,n\in\mathbb{N}$, se define $\Delta^m\circ\Delta^n=\Delta^m(\Delta^n)$ que se escribirá $\Delta^m\Delta^n$. De manera similar para las otras combinaciones.

Veamos algunos operadores de segundo orden. Se tiene los siguientes resultados.

1. Diferencia finita progresiva de segundo orden $\Delta^2=\Delta\circ\Delta$. Para toda $f\in C([a,b])$, se tiene

$$\begin{aligned}\Delta^2f(x) &= \Delta(\Delta f(x))=\Delta(f(x+h)-f(x))=f(x+2h)-f(x+h)-(f(x+h)-f(x))\\ &= f(x+2h)-2f(x+h)+f(x),\end{aligned}$$

obviamente, se supone que $h>0$ y $x\in\,]a,b[$ son tales que $x+2h,\,x+h\in[a,b]$.

2. Diferencia finita regresiva de segundo orden $\nabla^2=\nabla\circ\nabla$. Para toda $f\in C([a,b])$, se tiene

$$\begin{aligned}\nabla^2f(x) &= \nabla(\nabla f(x))=\nabla(f(x)-f(x+h))=f(x)-f(x+h)-(f(x+h)-f(x+2h))\\ &= f(x)-2f(x+h)+f(x+2h),\end{aligned}$$

con $x,\,x+2h,\,x+h\in[a,b]$, $h>0$.

3. Diferencia finita central de segundo orden $\delta^2=\delta\circ\delta$. Para toda $f\in C([a,b])$, se tiene

$$\begin{aligned}\delta^2f(x) &= \delta(\delta f(x))=\delta\left(f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right)\right)=f(x+h)-f(x)-(f(x)-f(x-h))\\ &= f(x+h)-2f(x)+f(x-h),\end{aligned}$$

donde $x,\,x+h,\,x-h\in[a,b]$, $h>0$.

4. Operadores mixtos de diferencias finitas de segundo orden: $\nabla\Delta=\nabla\circ\Delta$, $\Delta\nabla=\Delta\circ\nabla$, $\delta\Delta=\delta\circ\Delta$, $\cdots$ .Para toda $f\in C([a,b])$, se tienen las siguientes diferencias finitas de segundo orden:

$$\begin{aligned}\nabla\Delta f(x) &= \nabla(\Delta f(x))=\nabla(f(x+h)-f(x))=f(x+h)-f(x)-(f(x)-f(x-h))\\ &= f(x+h)-2f(x)+f(x-h)=\delta^2f(x),\end{aligned}$$

$$\begin{aligned}\Delta\nabla f(x) &= \Delta(f(x)-f(x-h))=f(x+h)-f(x)-(f(x)-f(x-h))\\ &= f(x+h)-2f(x)+f(x-h)=\delta^2f(x),\end{aligned}$$

$$\begin{aligned}\delta\Delta f(x) &= \delta(f(x+h)-f(x)) = f\left(x+\frac{3h}{2}\right)-f\left(x+\frac{h}{2}\right)-\left[f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right).\right]\\ &= f\left(x+\frac{3h}{2}\right)-2f\left(x+\frac{h}{2}\right)+f\left(x-\frac{h}{2}\right),\end{aligned}$$

donde $x, x+\frac{3h}{2}, x+\frac{h}{2}, x-\frac{h}{2} \in [a,b]$.

Se tiene $\delta^2 = \nabla\Delta = \Delta\nabla$ a la que se le denomina diferencia finita central del segundo orden. Como ejercicio se proponen obtener otras diferencias finitas de segundo orden.

De la definición de derivada segunda de una función f en un punto $x \in]a,b[$, se sigue que la derivada segunda $f''(x)$ se aproxima como sigue:

$$f''(x) \simeq \frac{f'(x+h)-f'(x)}{h}, \qquad f''(x) \simeq \frac{\Delta^2 f(x)}{h^2}, \qquad f''(x) \simeq \frac{\nabla^2 f(x)}{h^2},$$

$$f''(x) \simeq \frac{\delta^2 f(x)}{h^2} = \frac{\Delta\nabla f(x)}{h^2} = \frac{\nabla\Delta f(x)}{h^2}.$$

Supongamos que $f \in C^4([0,L])$. Se tiene

$$\frac{\Delta\nabla f(x)}{h^2} = \frac{\delta^2 f(x)}{h^2} = \frac{\nabla\Delta f(x)}{h} = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} = f''(x) + \frac{h^2}{4!}\left(f^{iv}(\xi_1)+f^{iv}(\xi_2)\right),$$

con $\xi_1 \in [x, x+h]$, $\xi_2 \in [x-h, x]$, consecuentemente

$$\left|\frac{f(x+h)-2f(x)+f(x-h)}{h^2} - f''(x)\right| = \frac{h^2}{4!}\left|\left(f^{iv}(\xi_1)+f^{iv}(\xi_2)\right)\right| \leq \frac{M}{12}h^2 \underset{h\longmapsto 0}{\longrightarrow} 0.$$

Ejemplo

Sea f la función real definida como $f(x) = \exp(\operatorname{sen}\sqrt{x}) \quad x > 0$. Entonces, la derivada segunda de f está definida como

$$f''(x) = \frac{1}{4}\left[\left(\frac{\cos\sqrt{x}}{\sqrt{x}}\right)^2 - \frac{\sqrt{x}\operatorname{sen}\sqrt{x}+\cos\sqrt{x}}{x^{\frac{3}{2}}}\right]\exp\left(\operatorname{sen}\sqrt{x}\right) \qquad x > 0,$$

luego

$$f''(2) = \frac{1}{4}\left[\left(\frac{\cos\sqrt{2}}{\sqrt{2}}\right)^2 - \frac{\sqrt{2}\operatorname{sen}\sqrt{2}+\cos\sqrt{2}}{2^{\frac{3}{2}}}\right]\exp\left(\operatorname{sen}\sqrt{2}\right) \simeq -0{,}3603967624.$$

Aproximemos la derivada segunda mediante la aplicación de diferencias finitas centrales de segundo orden, esto es

$$f''(2) \simeq \frac{f(2+h)-2f(2)+f(2-h)}{h} \quad h \neq 0,$$

y h suficientemente pequeño. Para realizar los cálculos usamos una calculadora de bolsillo.

Sea $h = 0{,}02$. Tenemos $f(2{,}02) = 2{,}688117974$, $f(2) = 2{,}68522882$, $f(1{,}98) = 2{,}682195506$. Luego

$$f'(2) \simeq \frac{f(2+h)-2f(2)+f(2+h)}{h^2} = -0{,}3604.$$

Para $h = 0{,}0001$, se tiene $f(2{,}0001) = 2{,}685243623$, $f(2) = 2{,}68522882$, $f(1{,}999) = 2{,}685214014$. Entonces

$$f'(2) \simeq \frac{f(2{,}0001)-2f(2)+f(1{,}9999)}{(0{,}0001)^2} = -0{,}3.$$

Debido a la representación en punto fijo y a causa de los errores de redondeo se obtiene este resultado que es una aproximación no satisfactoria. Nuevamente, la pregunta es, ¿cómo elegir $h \neq 0$ que nos rinda una buena aproximación de $f''(2)$? Consideremos $h = 0{,}005$. Entonces

$$f'(2) \simeq \frac{f(2{,}005)-2f(2)+f(1{,}995)}{(0{,}005)^2} = \frac{2{,}685964562 - 2\times 2{,}68522882 + 2{,}684484068}{(0{,}005)^2} = -0{,}3604,$$

resultado que obtuvimos anteriormente. La explicación de este hecho es que en una calculadora de bosillo se utliza la representación en punto fijo, y por otro lado, los errores de redondeo y de truncamiento afectan el resultado.

2.3.1. Aproximación de derivadas de funciones reales como formas lineales

Sea $\widehat{x} \in [a,b]$ y $f \in C^m([a,b])$. Se nota $D^m f = \frac{d^m f}{dx^m}(\widehat{x})$ la derivada m-ésima de f en $\widehat{x}$. Entonces D^mes un funcional lineal en $C^m([a,b])$.

Sean $n \in \mathbb{Z}^+$ con $n \geq m$, y $\tau(n) = \{a = x_0, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$. Se pone $h_j = x_j - x_{j-1}$ $j = 1, \ldots, n$ y $\widehat{h} = \text{máx}\{h_j \mid j = 1, \ldots, n\}$. En el caso de una partición uniforme, se tiene $h = \dfrac{b-a}{n}$, $x_j = jh \quad j = 0, 1, \ldots, n$ y $\widehat{h} = h$.

El operador de interpolación de Lagrange de f está definido como

$$\Pi(f) = \sum_{k=0}^{n} f(x_k) P_k$$

y sea $D^m_{num} = D^m \circ \Pi$. Entonces, para todo $f \in C^m([a,b])$ se tiene

$$D^m_{num}(f) = (D^m \circ \Pi)(f) = D^m(\Pi(f)) = D^m\left(\sum_{k=0}^{n} f(x_k) P_k\right) = \sum_{k=0}^{n} f(x_k) D^m P_k = \sum_{k=0}^{n} f(x_k) \frac{d^m P_k}{dx^m}(\widehat{x}).$$

Así, D^m_{num} es un funcional lineal sobre $C^m([a,b])$. Este funcional es la aproximación numérica de la derivada m-ésima de f en el punto $\widehat{x} \in [a,b]$.

Sean $n = m = 1$, entonces $\tau = \{a = x_0,\ x_1 = b\}$, $\widehat{x} \in [a,b]$ y

$$\begin{aligned} P_0(x) &= \frac{x-b}{a-b} \Rightarrow \frac{dP}{dx}(x) = -\frac{1}{b-a}, \\ P_1(x) &= \frac{x-a}{b-a} \Rightarrow \frac{dP}{dx}(x) = \frac{1}{b-a}, \end{aligned}$$

luego

$$D^1_{num}(f) = \sum_{k=0}^{1} f(x_k) \frac{dP_k}{dx}(\widehat{x}) = f(a)\left(-\frac{1}{b-a}\right) + f(b)\left(\frac{1}{b-a}\right) = \frac{f(b) - f(a)}{b-a}.$$

Observamos que la derivada de f en $x = \widehat{x}$ se aproxima como $\dfrac{f(b) - f(a)}{b-a}$, cociente incremental arriba tratado.

Sean $n = 2$, $\tau(2) = \left\{x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b\right\}$ una partición uniforme de $[a,b]$, entonces $h = \frac{b-a}{2}$. Sea $f \in C^2([a,b])$. Para $m = 1$, se tiene

$$D^1_{num}(f) = \sum_{k=0}^{2} f(x_k) \frac{dP_k}{dx}(\widehat{x}) = f(a) \frac{dP_0}{dx}(\widehat{x}) + f\left(\frac{a+b}{2}\right) \frac{dP_1}{dx}(\widehat{x}) + f(b) \frac{dP_2}{dx}(\widehat{x}).$$

En particular, para $\widehat{x} = \frac{a+b}{2}$ se tiene

$$\frac{dP_0}{dx}\left(\frac{a+b}{2}\right) = -\frac{1}{h}, \quad \frac{dP_1}{dx}\left(\frac{a+b}{2}\right) = 0, \quad \frac{dP_2}{dx}\left(\frac{a+b}{2}\right) = \frac{1}{h},$$

$$D^1_{num}(f) = \frac{f(b) - f(a)}{h},$$

que es la aproximación de la derivada mediante una diferencia finita central de primer orden. Escribiremos

$$\delta f(\widehat{x}) = \frac{f(b) - f(a)}{h}.$$

Para $m = 2$, obtenemos

$$D^2_{num}(f) = f(a)\frac{1}{h^2} - f\left(\frac{a+b}{2}\right)\frac{2}{h^2} + f(b)\frac{1}{h^2} = \frac{f(a) - 2f\left(\frac{a+b}{2}\right) + f(b)}{h^2} = \frac{\delta^2 f}{h^2},$$

que corresponde a la aproximación de la derivada segunda mediante una diferencia finita central de segundo orden.

Mediante este proceso podemos construir otras formas lineales que son aproximaciones de las derivadas de una función real.

2.3.2. Aproximación numérica de derivadas parciales primeras, segundas y laplaciano

Sean $\Omega \subset \mathbb{R}^2$ abierto, $(a,b) \in \Omega$, $h,k \in \mathbb{R}$ no nulos tales que $(a+h,b)$, $(a,b+k)$, $(a+h,b+k) \in \Omega$. Sea f una función real continua en Ω. En un punto arbitrario (x,y) de Ω notamos $z = f(x,y)$, a x lo denominamos primera variable, a y lo llamamos segunda variable de la función f.

Se define la derivada parcial de f respecto de x en el punto (a,b) que se nota $\dfrac{\partial f}{\partial x}(a,b)$ y se define como

$$\frac{\partial f}{\partial x}(a,b) = \lim_{h\to 0} \frac{f(a+h,b) - f(a,b)}{h}$$

siempre que el límite exista. De manera similar, la derivada parcial de f respecto de y en el punto (a,b) se nota $\dfrac{\partial f}{\partial y}(a,b)$ y se define como

$$\frac{\partial f}{\partial y}(a,b) = \lim_{k\to 0} \frac{f(a,b+k) - f(a,b)}{k}$$

siempre que el límite exista.

Note que fijado $(a,b) \in \Omega$, se define la función u como $u(x) = f(x,b)$ con $(x,b) \in \Omega$ y la derivada de u en $x = a$ está definida como

$$u'(a) = \lim_{h\to 0} \frac{u(a+h) - u(a)}{h} = \lim_{h\to 0} \frac{f(a+h,b) - f(a,b)}{h} = \frac{\partial f}{\partial x}(a,b).$$

Así, si $u(x) = f(x,b)$ donde $(x,b) \in \Omega$ con b fijo, se tiene $u'(a) = \dfrac{\partial f}{\partial x}(a,b)$.

En forma análoga, fijado $(a,b) \in \Omega$ se define la función v como $v(y) = f(a,y)$ con $(a,y) \in \Omega$. Luego

$$v'(y) = \lim_{k\to 0} \frac{v(b+k) - v(b)}{k} = \lim_{k\to 0} \frac{f(a,b+k) - f(a,b)}{k} = \frac{\partial f}{\partial y}(a,b).$$

Así, si $v(y) = f(a,y)$ donde $(a,y) \in \Omega$ con a fijo, entonces $v'(b) = \dfrac{\partial f}{\partial y}(a,b)$.

Supongamos $f \in C^3(\Omega)$ y $h,k \in \mathbb{R}$ no nulos tales que $(a+h,b)$, $(a-h,b)$, $(a,b+k)$, $(a,b-k) \in \Omega$. De estas dos observaciones y tomando en consideración los métodos de aproximación de la derivada de una función real en un punto, se tienen los siguientes resultados que permiten aproximar $\dfrac{\partial f}{\partial x}(a,b)$:

$$\begin{aligned}
\frac{\Delta f(a,b)}{h} &= \frac{f(a+h,b) - f(a,b)}{h} = \frac{\partial f}{\partial x}(a,b) + \frac{h}{2}\frac{\partial^2 f}{\partial x^2}(\xi_1,b), \\
\frac{\nabla f(a,b)}{h} &= -\frac{f(a-h,b) - f(a,b)}{h} = \frac{\partial f}{\partial x}(a,b) + \frac{h}{2}\frac{\partial^2 f}{\partial x^2}(\xi_2,b), \\
\frac{\delta f(a,b)}{h} &= \frac{f(a+h,b) - f(a-h,b)}{2h} = \frac{\partial f}{\partial x}(a,b) + \frac{h^2}{3!}\left(\frac{\partial^3 f}{\partial x^3}(\xi_1,b) + \frac{\partial^3 f}{\partial x^3}(\xi_2,b)\right),
\end{aligned}$$

donde ξ_1 se encuentra entre a y $a+h$, ξ_2 se encuentra en a y $a-h$, que corresponde a la aplicación de diferencias finitas progresivas, regresivas y centrales, respectivamente.

Resultados similares obtenidos para aproximar $\dfrac{\partial f}{\partial y}(a,b)$ que lo presentamos en la siguiente forma

$$\begin{aligned}
\frac{f(a,b+k) - f(a,b)}{k} - \frac{\partial f}{\partial y}(a,b) &= \frac{k}{2}\frac{\partial^2 f}{\partial y^2}(a,\eta_1), \\
-\frac{f(a,b-k) - f(a,b)}{k} - \frac{\partial f}{\partial y}(a,b) &= \frac{k}{2}\frac{\partial^2 f}{\partial y^2}(a,\eta_2), \\
\frac{f(a,b+k) - f(a,b-k)}{k} - \frac{\partial f}{\partial y}(a,b) &= \frac{k^2}{3!}\left(\frac{\partial^3 f}{\partial y^3}(a,\eta_1) + \frac{\partial^3 f}{\partial y^3}(a,\eta_2)\right),
\end{aligned}$$

con η_1 entre b y $b+k$, η_2 entre b y $b-k$.

i) Para $h \neq 0$ suficientemente pequeño, ponemos $z_0 = f(a,b)$, $z_1 = f(a+h,b)$. El cociente

$$\widetilde{f}_x(a,b) = \frac{z_1 - z_0}{h}$$

es una aproximación de $\dfrac{\partial f}{\partial x}(a,b)$ mediante una diferencia finita prograsiva.

ii) Si $z_0 = f(a,b)$, $z_1 = f(a-h,b)$ con $h \neq 0$ suficientemente pequeño, el cociente

$$\widetilde{f}_x(a,b) = -\frac{z_1 - z_0}{h}$$

es una aproximación de $\dfrac{\partial f}{\partial y}(a,b)$ mediante una diferencia finita regresiva.

iii) Si $h \neq 0$ suficientemente pequeño, $z_1 = f(a+h,b)$, $z_2 = f(a-h,b)$, el cociente

$$\widetilde{f}_x(a,b) = \frac{z_1 - z_2}{h}$$

es una aproximación de $\dfrac{\partial f}{\partial x}(a,b)$ mediante una diferencia finita central.

Obviamente las diferencias finitas centrales son las más utilizadas, por lo tanto $\dfrac{\partial f}{\partial x}(a,b)$, $\dfrac{\partial f}{\partial y}(a,b)$ se aproximan con el uso de diferencias finitas centrales. A menos que se diga lo contrario supondremos que las derivadas parciales son aproximadas mediante el uso de las diferencias finitas centrales.

Ejemplo

Considérese la función f definida como $f(x,y) = x^3y^4 \operatorname{sen}^2(xy) \quad (x,y) \in \mathbb{R}^2$. Se tiene

$$\begin{aligned}
\frac{\partial f}{\partial x}(x,y) &= 3x^2y^4 \operatorname{sen}^2(xy) + x^3y^4 [2y \operatorname{sen}(xy)\cos(xy)] \\
&= (3\operatorname{sen}(xy) + 2xy\cos(xy))\, x^2y^4 \operatorname{sen}(xy) \quad (x,y) \in \mathbb{R}^2, \\
\frac{\partial f}{\partial y}(x,y) &= 4x^3y^3 \operatorname{sen}^2(xy) + x^3y^4 (2x \operatorname{sen}(xy)\cos(xy)) \\
&= (4\operatorname{sen}(xy) + 2xy\cos(xy))\, x^3y^3 \operatorname{sen}(xy) \quad (x,y) \in \mathbb{R}^2.
\end{aligned}$$

Aproximemos $\dfrac{\partial f}{\partial x}(2,3)$ y $\dfrac{\partial f}{\partial y}(2,3)$ mediante diferencias finitas centrales.

Primeramente, calculemos $\dfrac{\partial f}{\partial x}(2,3)$ y $\dfrac{\partial f}{\partial y}(2,3)$. Tenemos

$$\begin{aligned}
\frac{\partial f}{\partial x}(2,3) &= (3\operatorname{sen}(6) + 12\cos(6)) \times 4 \times 81 \operatorname{sen}(6) = -967{,}2107771, \\
\frac{\partial f}{\partial y}(2,3) &= (4\operatorname{sen}(6) + 12\cos(6)) \times 8 \times 27 \operatorname{sen}(6) = -627{,}9434139.
\end{aligned}$$

Calculemos aproximaciones de $\dfrac{\partial f}{\partial x}(2,3)$ mediante diferencias finitas centrales, esto es, calculemos

$$\widetilde{f}_x(2,3) = \frac{f(2+h,3) - f(2-h,3)}{2h},$$

para $h = 0{,}002$, $h = 0{,}00015$, $h = 0{,}000001$.

Para $h = 0{,}002$, se tiene

$$\widetilde{f}_x(2,3) = \frac{f(2{,}002,3) - f(1{,}999,3)}{2 \times 0{,}002} = \frac{48{,}67057704 - 52{,}53621459}{0{,}004} = -967{,}1593875.$$

Para este valor de h, se tiene la siguiente estimación del error:

$$\left|\frac{\partial f}{\partial x}(2,3) - \widetilde{f}_x(2,3)\right| = |-967{,}2107771 - (-967{,}1593875)| = 0{,}0513896.$$

Para $h = 0{,}00015$ se tiene

$$\widetilde{f}_x(2,3) = \frac{f(2{,}00015, 3) - f(1{,}99985, 3)}{2 \times 0{,}00015} = \frac{50{,}44631203 - 50{,}7364752}{0{,}0003} = -967{,}2105667,$$

con lo que el error de aproximación es

$$\left|\frac{\partial f}{\partial x}(2,3) - \widetilde{f}_x(2,3)\right| = |-967{,}2107771 - (-967{,}2105667)| = 0{,}0002104 = 2{,}104 \times 10^{-4}.$$

Para $h = 0{,}000001$, se tiene

$$\widetilde{f}_x(2,3) = \frac{f(2{,}000001, 3) - f(1{,}99999, 3)}{2 \times 0{,}000001} = \frac{50{,}59034995 - 50{,}59243632}{0{,}000002} = --1043{,}185,$$

en consecuencia, se tiene la siguiente estimación del error

$$\left|\frac{\partial f}{\partial x}(2,3) - \widetilde{f}_x(2,3)\right| = |-967{,}2107771 - (-1043{,}185)| = 75{,}9742229.$$

Notamos que para este valor de h, el error ha aumentado significativamente. Esto se debe a los errores de redondeo y de truncamiento en el cálculo de $f(2+h, 3)$ y de $f(2-h, 3)$, operaciones que se realizan en punto fijo con una precisión $\varepsilon = 10^{-9}$. Si se trabaja con doble precisión mejoran los resultados. No incluimos estos resultados y proponemos que los comprueben.

Pasemos ahora al cálculo aproximado de $\dfrac{\partial f}{\partial y}(2,3)$ con diferencias finitas centrales, esto es

$$\widetilde{f}_y(2,3) = \frac{f(2, 3+k) - f(2, 3-k)}{2k} \quad k \neq 0.$$

Para $k = 0{,}05$ se tiene

$$\widetilde{f}_y(2,3) = \frac{f(2, 3{,}05) - f(2, 2{,}95)}{2 \times 0{,}05} = \frac{22{,}97245043 - 84{,}69051462}{0{,}1} = -617{,}1806419,$$

y el error de aproximación

$$\left|\frac{\partial f}{\partial y}(2,3) - \widetilde{f}_y(2,3)\right| = |-627{,}9434139 - (-617{,}1806419)| = 10{,}762772.$$

Se constata que esta aproximación no es aceptable.

Calculemos con $k = 0{,}0015$. Tenemos

$$\widetilde{f}_y(2,3) = \frac{f(2, 3{,}0015) - f(2, 2{,}9885)}{2 \times 0{,}0015} = \frac{49{,}65232798 - 51{,}53612921}{0{,}003} = -627{,}9337433,$$

y el error de aproximación

$$\left|\frac{\partial f}{\partial y}(2,3) - \widetilde{f}_y(2,3)\right| = |-627{,}9434139 - (-627{,}9337433)| = 9{,}6706 \times 10^{-3},$$

con lo que se muestra que esta es una aproximación aceptable.

Sea $k = 0{,}00011$. Entonces

$$\widetilde{f}_y(2,3) = \frac{f(2, 3{,}00011) - f(2, 2{,}99989)}{2 \times 0{,}00011} = \frac{50{,}52225955 - 50{,}66040683}{0{,}00022} = -627{,}9421818,$$

y el error de aproximación

$$\left|\frac{\partial f}{\partial y}(2,3) - \widetilde{f}_y(2,3)\right| = |-627{,}9434139 - (-627{,}9421818)| = 1{,}2321 \times 10^{-3},$$

lo que muestra que es una mejor aproximación que la anterior.

Las derivadas parciales segundas $\frac{\partial^2 f}{\partial x^2}(a,b)$, $\frac{\partial^2 f}{\partial x \partial y}(a,b)$, $\frac{\partial^2 f}{\partial y^2}(a,b)$ pueden ser aproximadas siguiendo la misma metodología empleada para el cálculo aproximado de las derivadas segundas de funciones reales de una sola variable. Así, para $h \neq 0$, $\frac{\partial^2 f}{\partial x^2}(a,b)$ puede aproximarse con diferencias finitas centrales, esto es,

$$\frac{\partial^2 f}{\partial x^2}(a,b) \simeq \widetilde{f}_{xx}(a,b) = \frac{f(a+h,b) - 2f(a,b) + f(a-h,b)}{2h} \quad h \neq 0,$$

h suficientemente pequeño.

De manera similar, tenemos

$$\frac{\partial^2 f}{\partial y^2}(a,b) \simeq \widetilde{f}_{yy}(a,b) = \frac{f(a,b+k) - 2f(a,b) + f(a,b+k)}{2k} \quad k \neq 0,$$

k suficientemente pequeño.

Sea $f \in C^2(\Omega)$ y $(a,b) \in \Omega$. El laplaciano de f en el punto $(a,b) \in \Omega$ se denota $\triangle f(a,b)$ y se define como

$$\triangle f(a,b) = \frac{\partial^2 f}{\partial x^2}(a,b) + \frac{\partial^2 f}{\partial y^2}(a,b),$$

el mismo que puede ser aproximado como sigue:

$$\widetilde{\triangle} f(a,b) = \frac{f(a+h,b) - 2f(a,b) + f(a-h,b)}{2h} + \frac{f(a,b+k) - 2f(a,b) + f(a,b+k)}{2k}$$

con $h \neq 0$, $k \neq 0$ suficientemente pequeño.

Ejemplo

Sea f la función definida como $f(x,y) = \ln\left(x^2+y^2\right) \quad (x,y) \in \mathbb{R}^2$ con $(x,y) \neq (0,0)$. Se tiene

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2}(x,y) &= 2\frac{y^2-x^2}{(x^2+y^2)^2} \quad (x,y) \in \mathbb{R}^2,\ (x,y) \neq (0,0),\\
\frac{\partial^2 f}{\partial y^2}(x,y) &= 2\frac{x^2-y^2}{(x^2+y^2)^2} \quad (x,y) \in \mathbb{R}^2,\ (x,y) \neq (0,0),
\end{aligned}$$

luego

$$\triangle f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} = 0 \quad (x,y) \in \mathbb{R}^2,\ (x,y) \neq (0,0).$$

Las funciones f tales que $\triangle f(x,y) = 0 \quad \forall (x,y) \in \Omega$ se llaman funciones armónicas.

Calculemos valores aproximados de $\widetilde{\triangle} f(1,2)$ aproximación de $\triangle f(1,2)$.

Sea $h = 0{,}002$, $k = 0{,}0015$. Entonces

$$\begin{aligned}
\widetilde{f}_{xx}(1,2) &= \frac{f(1{,}002,2) - 2f(1,2) + f(0{,}998,2)}{2h}\\
&= \frac{1{,}610238392 - 2 \times 1{,}609437912 - 1{,}608638393}{0{,}004} = 0{,}00024,\\
\widetilde{f}_{yy}(1,2) &= \frac{f(1,2{,}0015) - 2f(1,2) + f(1,1{,}9985)}{0{,}003}\\
&= \frac{1{,}610637642 - 3{,}218875825 + 1{,}608237642}{0{,}003} = -1{,}80333 \times 10^{-4},
\end{aligned}$$

luego

$$\widetilde{\triangle} f(1,2) = \widetilde{f}_{xx}(1,2) + \widetilde{f}_{yy}(1,2) = 0{,}00024 - 1{,}80333 \times 10^{-4} = 5{,}966667 \times 10^{-5},$$

es una aproximación $\triangle f(1,2) = 0$.

Para $h = k = 0{,}00012$, tenemos

$$\begin{aligned}
\widetilde{f}_{xx}(1,2) &= \frac{f(1{,}00012, 2) - 2f(1,2) + f(0{,}99988, 2)}{2 \times 0{,}00012} = \frac{3{,}15 \times 10^{-9}}{0{,}00024} = 0{,}000013125, \\
\widetilde{f}_{yy}(1,2) &= \frac{f(1, 2+k) - 2f(1,2) + f(1, 2-k)}{2 \times 0{,}00012} = -\frac{3{,}3 \times 10^{-9}}{0{,}00024} = -0{,}00001375,
\end{aligned}$$

con lo que

$$\widetilde{\triangle} f(1,2) = \widetilde{f}_{xx}(1,2) + \widetilde{f}_{yy}(1,2) = 6{,}25 \times 10^{-7},$$

que es una mejor aproximación de $\widetilde{\triangle} f(1,2)$ precedente. Así, $\triangle f(1,2) \simeq \widetilde{\triangle} f(1,2) = 6{,}25 \times 10^{-7}$.

2.4. Integración numérica

Según la Historia de la Matemática, fue el Cálculo Integral el que primero se desarrolló. Obviamente las primeras funciones que se integraron sobre un intervalo $[a, b]$ fueron las polinomiales. Estas son en realidad las más simples de integrarse. Otras funciones sencillas de integrarse son las funciones trigonométricas seno y coseno. Pronto aparecieron otra clase de funciones continuas f que no se integran mediante funciones elementales, el cálculo de $I(f) = \int_a^b f(x)\,dx$ resulta imposible (en algunos casos es posible mediante la integración de funciones de variable compleja), por lo que dicha integral tendrá que ser aproximada numéricamente. Con este propósito, consideramos un problema más sencillo que es el cálculo de la integral definida de un polinomio interpolante de la función f. Más aún, en esta sección tratamos la fórmula de Newton-Cotes y de esta se desprenden la regla del rectángulo o conocida también como fórmula del punto medio, la regla del trapecio o fórmula del trapecio, la regla de Simpson que son las más utilizadas. Obtenemos estimaciones de errores para cada una de estos métodos y luego se generalizan a particiones regulares del intervalo $[a, b]$ en consideración, lo que da lugar a las reglas generalizadas del rectángulo $R_n(f)$, del trapecio $T_n(f)$ y de Simpson $S_n(f)$. Estas son aplicadas al cálculo de integrales dobles sobre regiones en las que las integrales pueden calcularse como integrales reiteradas, a tales regiones se los denomina del tipo I o II .

La integración de funciones que se representan como series de potencias se estudian en el siguiente capítulo.

2.4.1. Fórmula de Newton-Cotes

Se define el funcional I sobre $C([a, b])$ como sigue: $I(f) = \int_a^b f(x)\,dx \quad \forall f \in C([a, b])$. Entonces I es funcional lineal en $C([a, b])$.

El operador de interpolación de Lagrange Π es lineal. Se define el operador $G = I \circ \Pi$. entonces G es lineal, y

$$G(f) = (I \circ \Pi)(f) = I(\Pi(f)) = I\left(\widehat{P}\right) = I\left(\sum_{k=0}^{n} f(x_k) P_k\right) = \sum_{k=0}^{n} f(x_k) I(P_k) \quad \forall f \in C([a, b]).$$

Así,

$$G(f) = \sum_{k=0}^{n} f(x_k) I(P_k) = \sum_{k=0}^{n} f(x_k) \int_a^b P_k(x)\,dx \quad \forall f \in C([a, b]),$$

que se conoce como la fórmula de Newton-Cotes. De esta fórmula se desprenden algunos resultados que tratamos a continuación.

Fórmula del rectángulo

Para $n=0$, $\tau=\left\{\frac{a+b}{2}\right\}$, una partición del intervalo $[a,b]$ construida únicamente por el punto medio. El polinomio interpolante está definido como $P_0(x)=1 \quad \forall x\in[a,b]$, entonces

$$G(f)=f\left(\frac{a+b}{2}\right)\int_a^b dx=(b-a)\,f\left(\frac{a+b}{2}\right),$$

que se conoce como fórmula del rectángulo para aproximar $I(f)=\int_a^b f(x)\,dx$.

Fórmula de los trapecios

Para $n=1$, $\tau=\{a=x_0,x_1=b\}$ una partición del intervalo $[a,b]$. Una interpolante de la función f está definida como $v(x)=f(a)\,P_0(x)+f(b)\,P_1(x) \quad x\in[a,b]$, donde P_0, P_1 están definidos como

$$P_0(x)=\frac{x-b}{a-b},\quad P_1(x)=\frac{x-a}{b-a}\quad x\in\mathbb{R}.$$

Resulta

$$\begin{aligned} I(P_0) &= \int_a^b P_0(x)\,dx=\int_a^b\frac{x-b}{a-b}dx=\frac{b-a}{2},\\ I(P_1) &= \int_a^b P_1(x)\,dx=\int_a^b\frac{x-a}{b-a}dx=\frac{b-a}{2}. \end{aligned}$$

Consecuentemente

$$G(f)=\sum_{k=0}^{1}f(x_k)\,I(P_k)=f(a)\,\frac{b-a}{2}+f(b)\,\frac{b-a}{2}=\frac{b-a}{2}\,(f(a)+f(b)).$$

Así,

$$G(f)=\frac{b-a}{2}\,(f(a)+f(b)),$$

que se conoce con el nombre de fórmula de los trapecios para aproximar $I(f)=\int_a^b f(x)\,dx$.

Regla de Simpson

Para $n=2$ y $\tau=\left\{a=x_0,x_1=\frac{a+b}{2},x_2=b\right\}$, $h=\frac{b-a}{2}$, τ es una partición de $[a,b]$. Una interpolante de f está definida como $v(x)=f(a)\,P_0(x)+f\left(\frac{a+b}{2}\right)P_1(x)+f(b)\,P_2(x) \quad x\in[a,b]$ donde P_0, P_1, P_2 son los polinomios de interpolación de Lagrange. Tenemos

$$G(f)=\sum_{k=0}^{2}f(x_k)\,I(P_k)=f(a)\,I(P_0)+f\left(\frac{a+b}{2}\right)I(P_1)+f(b)\,I(P_2),$$

con

$$\begin{aligned} I(P_0)&=\int_a^b P_0(x)\,dx=\int_a^b\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}dx=\frac{h}{3},\\ I(P_1)&=\int_a^b P_1(x)\,dx=\int_a^b\frac{(x-x_0)(x-x_1)}{(x_1-x_0)(x_1-x_2)}dx=4\frac{h}{3},\\ I(P_n)&=\int_a^b P_2(x)\,dx=\int_a^b\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}dx=\frac{h}{3}. \end{aligned}$$

Luego,

$$G(f)=\frac{h}{3}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right),$$

que se conoce como la regla de Simpson para aproximar $I(f)=\int_a^b f(x)\,dx$.

2.5. Regla de los trapecios generalizada. Estimación del error

Sea $f \in C([a,b])$ y consideremos como problema (P) el cálculo de la integral $I(f) = \int_a^b f(x)dx$.

Sean $n \in \mathbb{Z}^+$ y $\tau(n) = \{a = x_0, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$. Se pone $h_j = x_j - x_{j-1}$ $j = 1, \ldots, n$ y $\widehat{h} = \text{máx}\{h_j \mid j = 1, \ldots, n\}$. Suponemos que existe $\sigma \geq 1$ tal que $h_j \leq \sigma \dfrac{b-a}{n}$ $j = 1, \ldots, n$. En el caso de una partición uniforme, se tiene $h = \dfrac{b-a}{n}$, $x_j = jh \quad j = 0, 1, \ldots, n$, $\widehat{h} = h$ y $\sigma = 1$.

El polinomio de interpolación de la función f en el j-ésimo subintervalo $[x_{j-1}, x_j]$ de $[a,b]$ es la función afín denotada g_j y definida como sigue:

$$g_j(x) = f(x_{j-1}) + \frac{f(x_j) - f(x_{j-1})}{h_j}(x - x_{j-1}) \quad x \in [x_{j-1}, x_j[, \quad j = 1, \ldots, n,$$

La función interpolante g sobre $[a,b]$ está definida como

$$g(x) = \sum_{j=0}^{n} f(x_j)\, \varphi_j(x) \qquad x \in [a,b],$$

donde $\varphi_j \quad j = 1, \ldots, n$, son las funciones techo antes definidas en los interpolantes afines a trozos. Note que $g(x_i) = f(x_i) \quad i = 0, 1, \ldots, n$. En la figura siguiente se muestra la discretización de $[a,b]$, la gráfica de la función f y de la de su interpolante g.

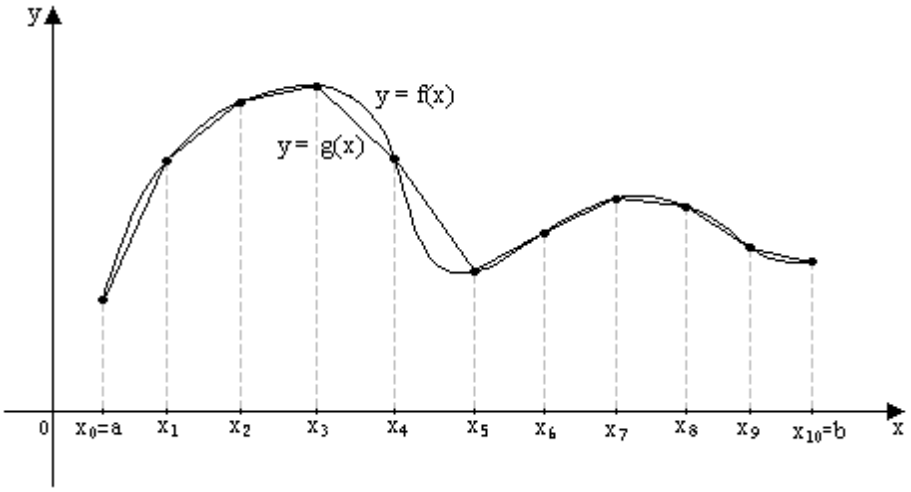


Figura 25

Como problema $(\tilde{P}_n)$ consideramos el siguiente: $I(g) = \int_a^b g(x)dx$. De la definición de la función g, se tiene

$$\begin{aligned} I(g) &= \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} g_j(x)dx = \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} \left(f(x_{j-1}) + \frac{f(x_j) - f(x_{j-1})}{h}(x - x_{j-1}) \right) dx \\ &= \sum_{j=1}^{n} \left(f(x_{j-1}) + \frac{f(x_j) - f(x_{j-1})}{2h}(x - x_{j-1})^2 \,|_{x_{j-1}}^{x_j} \right) = \frac{h}{2} \sum_{j=1}^{n} [f(x_{j-1}) + f(x_j)] \\ &= \frac{h}{2}(f(a) + f(b)) + h \sum_{j=1}^{n-1} f(x_j). \end{aligned}$$

La aproximación que hemos construido se conoce con el nombre de regla de los trapecios generalizada. Escribiremos

$$T_n(f) = \frac{h}{2}(f(a) + f(b)) + h\sum_{j=1}^{n-1} f(x_j),$$

Así,

$$I(f) = \int_a^b f(x)dx \simeq T_n(f) = \frac{h}{2}(f(a) + f(b)) + h\sum_{j=1}^{n-1} f(x_j).$$

Esta aproximación se completa con una estimación del error entre $I(f)$ y $T_n(f)$ que tratamos a continuación.

Se prueba inmediatamente que el funcional T_n de $C([a,b])$ en $\mathbb{R}$ definido como

$$T_n(f) = \frac{h}{2}(f(a) + f(b)) + h\sum_{j=1}^{n-1} f(x_j) \quad \forall f \in C([a,b]),$$

es lineal, es decir que T_n es un elemento del espacio dual de $C([a,b])$.

De la fórmula de los trapecios generalizada para una partición uniforme se observa que para su aplicación se requiere disponer de la siguiente información: extremos del intervalo $[a,b]$ en el que la función f está definida y la propia función f, número de puntos de la partición $\tau(n)$. Con esta información se tiene el siguiente algoritmo de aproximación de una integral definida mediante la regla de los trapecios generalizada.

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $a, b \in \mathbb{R}$, función f.

Datos de salida: n, $T_n(f)$, mensaje.

1. Verificar $a < b$. Caso contrario continuar en 7).
2. Hacer $h = \dfrac{b-a}{n}$.
3. $S = 0$.
4. Para $j = 1, \ldots, n-1$

   $$S = S + f(a + jh),$$

   Fin de bucle j.
5. $T_n(f) = \dfrac{h}{2}(f(a) + f(b)) + hS$..
6. Imprimir n, $T_n(f)$. Continuar en 8).
7. Mensaje: $a < b$.
8. Fin.

Ejemplo

Sea f la función real definida como $f(x) = x^2e^x \quad x \in [0,2]$. Calculemos $I(f) = \int_0^2 f(x)\,dx = \int_0^2 x^2e^x dx$ y aproximemos a $I(f)$ mediante la regla de los trapecios generalizada $T_n(f)$.

Primeramente calculamos el valor exacto de la integral $I(f)$. Aplicando el método de integración por partes, tenemos

$$I(f) = \int_0^2 x^2e^x dx = \left[x^2e^x - 2\ (xe^x - e^x)\right]\Big|_0^2 = 2\left(e^2 - 1\right) \simeq 12{,}7781122.$$

En la tabla siguiente se muestra la aplicación del algoritmo precedente, esto es, la regla de los trapecios generalizada para una partición uniforme del intervalo $[0,2]$.

n	$T_n(f)$	Error=$\mid I(f)-T_n(f)\mid$
10	12,9748064291	$1{,}966942312\times10^{-1}$
20	12,8273508376	$4{,}9238639741\times10^{-2}$
40	12,7904259325	$1{,}2313734634\times10^{-2}$
80	12,7811908863	$3{,}0786884382\times10^{-2}$
160	12,7788818859	$7{,}6968803500\times10^{-4}$
320	12,7783046208	$1{,}9242300412\times10^{-4}$
640	12,7781603037	$4{,}8105813249\times10^{-5}$
1280	12,7781242243	$1{,}2026457201\times10^{-5}$

.

Estimación del error de integración con la fórmula de los trapecios

Primeramente estableceremos una estimación del error de integración con la fórmula de los trapecios y a continuación usaremos este resultado para obtener una etimación del error de integración con la fórmula de los trapecios generalizada.

Sea $f\in C^2([a,b])$. Consideremos el problema $I(f)=\int_a^b f(x)\,dx$. Esta integral se aproxima con la denominada regla de los trapecios $T(f)$ definida como

$$T(f)=\int_a^b P(x)\,dx=\frac{b-a}{2}\left(f(a)-f(b)\right),$$

donde $P(x)=f(a)\,\varphi_0(x)+f(b)\,\varphi_1(x)\quad x\in[a,b]$ es el polinomio de interpolación de f, φ_0, φ_1 son los polinomios de interpolación de Lagrange antes definidos (véase interpolación de Lagrange). El error de interpolación polinomial de Lagrange está defindo como

$$\varepsilon(x)=f(x)-P(x)=\frac{f''(\xi)}{2!}(x-a)(x-b)\quad x\in[a,b],$$

$\xi\in[a,b]$.

El error de integración con la regla de los tapecios se nota ε_f y se define como $\varepsilon_f=I(f)-T(f)$. Nos interesamos en obtener una estimación del error ε_f y en una mayoración de $\mid\varepsilon_f\mid$. De la definición de $I(f)$ y $T(f)$ se sigue que

$$\varepsilon_f=I(f)-T(f)=\int_a^b\left(f(x)-P(x)\right)dx=\int_a^b\frac{f''(\xi)}{2!}(x-a)(x-b)\,dx.$$

Sean $m=\min\limits_{x\in[a,b]}|f''(x)|$, $M=\max\limits_{x\in[a,b]}|f''(x)|$. Entonces

$$m\le\left|f''(\xi)\right|\le M\quad\forall\xi\in[a,b],$$

de donde

$$\frac{m}{2}\left|(x-a)(x-b)\right|\le\frac{|f''(\xi)|}{2}\left|(x-a)(x-b)\right|dx\le\frac{M}{2}\left|(x-a)(x-b)\right|$$

e integrando sobre $[a,b]$, resulta

$$\frac{m}{2}\int_a^b\left|(x-a)(x-b)\right|dx\le\int_a^b\frac{|f''(\xi)|}{2!}\left|(x-a)(x-b)\right|dx\le\frac{M}{2}\int_a^b\left|(x-a)(x-b)\right|dx.$$

Puesto que $\omega(x)=(x-a)(x-b)\le0\quad\forall x\in[a,b]$, entonces

$$\int_a^b\omega(x)\,dx=\int_a^b(x-a)(x-b)\,dx=-\frac{(b-a)^3}{6},$$

con lo que la desigualdad precedente se expresa como

$$\frac{m}{12}(b-a)^3 \leq \int_a^b \frac{|f''(\xi)|}{2}|(x-a)(x-b)|\,dx \leq \frac{M}{12}(b-a)^3,$$

y de esta a su vez se obtiene la siguiente:

$$m \leq \frac{12}{(b-a)^3}\int_a^b \frac{|f''(\xi)|}{2}|(x-a)(x-b)|\,dx \leq M.$$

Por el teorema del valor intermedio (véase en Calculus I de Apostol, página 177), existe $\lambda \in [a,b]$ tal que

$$\left|f''(\lambda)\right| = \frac{12}{(b-a)^3}\int_a^b \frac{|f''(\xi)|}{2}|(x-a)(x-b)|\,dx$$

con lo cual

$$\int_a^b \frac{|f''(\xi)|}{2}|(x-a)(x-b)|\,dx = \frac{(b-a)^3}{12}\left|f''(\lambda)\right|.$$

Consecuentemente

$$|\varepsilon_f| = |I(f)-T(f)| = \left|\int_a^b \frac{f''(\xi)}{2}(x-a)(x-b)\,dx\right| \leq \int_a^b \frac{|f''(\xi)|}{2}|(x-a)(x-b)|\,dx \leq \frac{(b-a)^3}{12}\left|f''(\lambda)\right|.$$

En conclusión, $|\varepsilon_f| \leq \dfrac{(b-a)^3}{12}|f''(\lambda)|$ para algún $\lambda \in [a,b]$. En la práctica resulta difícil obtener $\lambda \in [a,b]$, y como $|f''(\lambda)| \leq M$ se sigue que $|\varepsilon_f| \leq \dfrac{M}{12}(b-a)^3$.

Obtengamos una estimación del error para la fórmula de los trapecios generalizada.

Sea $n \in \mathbb{Z}^+$ y $\tau(n) = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición del intervalo $[a,b]$ con $x_{i-1} < x_i$, $h_i = x_i - x_{i-1}$ $i = 1,\ldots,n$, $\widehat{h} = \max\limits_{i=1,\ldots,n} h_i$. Suponemos que existe $\sigma \geq 1$ tal que $h_i \leq \sigma\dfrac{b-a}{n}$ $i = 1,\ldots,n$. A las particiones que satisfacen esta propiedad se les conoce como particiones regulares. En el caso de una partición uniforme se tiene $h = h_i = \dfrac{b-a}{n}$ $i = 1,2,\ldots,n$, $\sigma = 1$.

La fórmula de los trapecios generalizada está definida como

$$T_n(f) = \frac{h}{2}(f(a)+f(b)) + h\sum_{k=1}^{n-1} f(x_k).$$

Note que el polinomio de interpolación de f de grado 1 en el k-ésimo subintervalo $[x_{k-1},x_k]$ de $[a,b]$ está definido como

$$P_k(x) = f(x_{k-1}) + \frac{f(x_k)-f(x_{k-1})}{h_k}(x-x_{k-1}) \quad x \in [x_{k-1},x_k],$$

consecuentemente

$$T_n(f) = \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k} P_k(x)\,dx = \frac{h}{2}(f(a)+f(b)) + h\sum_{k=1}^{n-1} f(x_k).$$

El error de integración con la fórmula de los trapecios generalizada se nota ε_f y se define como $\varepsilon_f = I(f) - T_n(f)$.

Apliquemos el error de integración con la regla de los trapecios a cada intervalo $[x_{k-1},x_k]$ $k = 1,\ldots,n$. Resulta

$$\begin{aligned}\varepsilon_f &= I(f)-T_n(f) = \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k} f(x)\,dx - \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k} P_k(x)\,dx = \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k}[f(x)-P_k(x)]\,dx \\ &= \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k}\frac{f''(\xi_k)}{2}(x-x_{k-1})(x-x_k)\,dx,\end{aligned}$$

y $\xi_k \in [x_{k-1}, x_k] \quad k = 1, \ldots, n.$

De la estimación del error de integración con la regla de los trapecios antes obtenida, resulta

$$\begin{aligned} |\varepsilon_f| &= |I(f) - T_n(f)| = \left| \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \frac{f''(\xi_k)}{2} (x - x_{k-1})(x - x_k)\, dx \right| \\ &\leq \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \frac{|f''(\xi_k)|}{2} |(x - x_{k-1})(x - x_k)|\, dx \\ &\leq \sum_{k=1}^{n} \frac{(x_k - x_{k-1})^3}{12} |f''(\lambda_k)| = \sum_{k=1}^{n} \frac{h_k^3}{12} |f''(\lambda_k)|. \end{aligned}$$

Puesto que $h_k \leq \sigma \dfrac{b-a}{n} \quad k = 1, 2, \ldots, n$, entonces $h_k^3 \leq h_k \widehat{h}^2 \leq \sigma \dfrac{b-a}{n} \widehat{h}^2$, luego

$$|\varepsilon_f| \leq \sum_{k=1}^{n} \frac{h_k^3}{12} |f''(\lambda_k)| \leq \sigma \frac{b-a}{n} \frac{\widehat{h}^2}{12} \sum_{k=1}^{n} |f''(\lambda_k)|.$$

Además, $m \leq |f''(\lambda_k)| \leq M \quad k = 1, \ldots, n$, de donde

$$m \leq \frac{1}{n} \sum_{k=1}^{n} |f''(\lambda_k)| \leq M,$$

y por el teorema del valor intermedio, existe $\lambda \in [a, b]$ tal que $|f''(\lambda)| = \dfrac{1}{n} \sum\limits_{k=1}^{n} |f''(\lambda_k)|$, con lo cual

$$|\varepsilon_f| \leq \sigma \frac{b-a}{n} \frac{\widehat{h}^2}{12} \sum_{k=1}^{n} |f''(\lambda_k)| = \frac{\sigma(b-a)}{12} \widehat{h}^2 |f''(\lambda)| \leq \frac{\sigma(b-a)}{12} M \widehat{h}^2 \underset{h \to 0}{\longrightarrow} 0,$$

pués $h_k \leq \sigma \dfrac{b-a}{n} \quad k = 1, 2, \ldots, n$, $\widehat{h} = \max\limits_{k=1,\ldots,n} h_k \leq \sigma \dfrac{b-a}{n} \underset{n \longrightarrow \infty}{\longrightarrow} 0$.

En el caso de una partición uniforme se tiene $\sigma = 1$ y $\widehat{h} = h$. entonces

$$|\varepsilon_f| \leq \frac{b-a}{12} M h^2 \underset{h \to 0}{\longrightarrow} 0.$$

En cualquiera de los casos, el método de integración aproximado de los trapecios generalizado es convergente, esto es

$$T_n(f) \underset{n \to \infty}{\longrightarrow} I(f).$$

2.6. Regla de Simpson generalizada

Sean $f \in C([a, b])$, $n \in \mathbb{Z}^+$ con $n > 1$ y $\tau(n) = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición del intervalo $[a, b]$, esto es $x_{i-1} < x_i \quad i = 1, \ldots, n$, $h_i = x_i - x_{i-1}$, $i = 1, \ldots, n$, $\widehat{h} = \max\{h_i \mid i = 1, \ldots, n\}$. En el caso de una partición uniforme, se define $h = \dfrac{b-a}{n}$ y $x_j = a + jh \quad j = 0, 1, \ldots, n$, entonces $\widehat{h} = h$. Suponemos que existe $\sigma \geq 1$ tal que $h_i \leq \sigma \dfrac{b-a}{n} \quad i = 1, \ldots, n$. A las particiones que satisfacen esta propiedad, como ya hemos dicho anteriormente, se les conoce como particiones regulares. En el caso de la partición uniforme se tiene $\sigma = 1$.

Consideramos el k-ésimo intervalo $[x_{k-1}, x_k] \quad k = 1, 2, \ldots, n$ y aplicamos la regla de Simpson a este intervalo. Tenemos

$$I_k = \frac{h_k}{6} \left[f(x_{k-1}) + 4f\left(\frac{x_{k-1} + x_k}{3} \right) + f(x_k) \right].$$

Luego

$$S_n(f) = \sum_{k=1}^{n} I_k = \sum_{k=1}^{n} \frac{h_k}{6}\left[f(x_{k-1}) + 4f\left(\frac{x_{k-1}+x_k}{2}\right) + f(x_k)\right],$$

se llama fórmula de Simpson generalizada.

Se prueba inmediatamente que fijados $n \in \mathbb{Z}^+$ con $n > 1$ y $\tau(n)$ una partición regular del intervalo $[a,b]$, el funcional S_n de $C([a,b])$ en $\mathbb{R}$ definido como $S_n(f) = \sum_{k=1}^{n} \frac{h_k}{6}\left[f(x_{k-1}) + 4f\left(\frac{x_{k-1}+x_k}{2}\right) + f(x_k)\right]$ es lineal.

En el caso de una partición uniforme, se tiene

$$\begin{aligned} S_n(f) &= \sum_{k=1}^{n} I_k = \sum_{k=1}^{n} \frac{h}{6}\left[f(x_{k-1}) + 4f\left(\frac{x_{k-1}+x_k}{2}\right) + f(x_k)\right] \\ &= \frac{h}{6}\sum_{k=1}^{n}\left[f(x_{k-1}) + f(x_k)\right] + \frac{4h}{6}\sum_{k=1}^{n} f\left(\frac{x_{k-1}+x_k}{2}\right) \\ &= \frac{h}{6}(f(a) + f(b)) + \frac{h}{3}\sum_{k=1}^{n-1} f(x_k) + \frac{2}{3}h\sum_{k=1}^{n} f\left(\frac{x_{k-1}+x_k}{2}\right). \end{aligned}$$

De la fórmula de Simpson generalizada para una partición uniforme se observa que para su aplicación se requieren de los siguientes datos: número de puntos de la partición $\tau(n)$, extremos del intervalo $[a,b]$ en el que la función f está definida y la propia función f. Con esta información se tiene el siguiente algoritmo de aproximación de una integral definida mediante el método de Simpson con la fórmula generalizada.

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $a, b \in \mathbb{R}$, función f.

Datos de salida: n, $S_n(f)$, mensaje.

1. Verificar $a < b$. Caso contrario continuar en 9)
2. Hacer $h = \dfrac{b-a}{n}$.
3. $S_1 = 0$.
4. $S_2 = 0$.
5. Para $j = 1, \ldots, n-1$

   $S_1 = S_1 + f(a + jh)$,

   $S_2 = S_2 + f\left(a + \left(j - \frac{1}{2}\right)h\right)$.

   Fin de bucle j.
6. $S_2 = S_2 + f\left(a + \left(n - \frac{1}{2}\right)h\right)$.
7. $S_n(f) = \dfrac{h}{6}(f(a) + f(b)) + \dfrac{h}{3}S_1 + \dfrac{2h}{3}S_2$.
8. Imprimir n, $S_n(f)$. Continuar en 10).
9. Mensaje: $a < b$.
10. Fin.

Ejemplo

Sea f la función real definida como $f(x) = x^2 e^x \quad x \in [0,2]$. Calculemos $I(f) = \int_0^2 f(x)\,dx = \int_0^2 x^2 e^x dx$ y aproximemos a $I(f)$ mediante la fórmula de Simpson generalizada $S_n(f)$.

Aplicando el método de integración por partes, tenemos

$$I(f) = \int_0^2 x^2 e^x dx = \left[x^2 e^x - 2\ (xe^x - e^x)\right]\Big|_0^2 = 2\left(e^2 - 1\right) \simeq 12{,}7781122.$$

Esta integral ya fue calculada en la sección precedente y fue aproximada con la regla de los trapecios generalizada. En la tabla siguiente se muestra la aplicación del algoritmo precedente, esto es, la fórmula de Simpson generalizada para una partición uniforme del intervalo $[0,2]$.

n	$S_n(f)$	$Error :\mid I(f) - S_n(f) \mid$
10	12,7781989738	$8{,}677590441 \times 10^{-5}$
20	12,7781176308	$5{,}4329327987 \times 10^{-6}$
40	12,7781125376	$3{,}3970602331 \times 10^{-7}$
80	12,7781122191	$2{,}123393727516 \times 10^{-8}$
160	12,7781121992	$1{,}3271588273 \times 10^{-9}$

.

Comparando estos resultados con los obtenidos con la fórmula de los trapecios generalizada podemos constatar que con la regla de Simpson generalizada se tiene una convergencia cuadrática mientras que con la de los trapecios generalizada se tiene únicamente una convergencia del tipo lineal. Obviamente que con la fórmula de Simpson generalizada se realizan n evaluaciones adicionales de la función f que las que se realizan en la de los trapecios generalizada.

2.7. Estimación del error en la regla de Simpson

Supongamos $f \in C^4([a,b])$. Ponemos $I(f) = \int_a^b f(x)\,dx$. Para aproximar $I(f)$ aplicamos la regla de Simpson $G(f)$ arriba definida como

$$G(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{b+a}{2}\right) + f(b)\right),$$

y denotamos con ε_f el error de aproximación cometido entre la solución exacta $I(f)$ y su valor aproximado $G(f)$, esto es, $\varepsilon_f = I(f) - G(f)$. Determinemos ε_f.

i) Recordemos que si $f(x)$ es un polinomio de grado 2, entonces $f(x)$ se escribe como

$$f(x) = f(a)\,\varphi_0(x) + f\left(\frac{a+b}{2}\right)\varphi_1(x) + f(b)\,\varphi_2(x) \quad x \in [a,b],$$

donde φ_0, φ_1, φ_2 son los polinomios de interpolación de Lagrange antes definidos (véase la sección interpolación polonomial). Resulta que $I(f) = \int_a^b f(x)\,dx = G(f)$, y

$$\varepsilon_f = \int_a^b f(x)\,dx - G(f) = 0.$$

ii) Mostremos que si f es un polinomio de grado 3, también se tiene $\varepsilon_f = 0$. En efecto, de la fórmula del error de interpolación de Lagrange tenemos

$$\varepsilon(x) = f(x) - P(x) = \frac{f'''(\xi)}{3!}(x-a)\left(x - \frac{a+b}{2}\right)(x-b) \quad x \in [a,b]$$

y $\xi \in\,]a,b[$.

Sea $t = \dfrac{x-a}{h}$ con $h = \dfrac{b-a}{2}$, entonces

$$\varepsilon(x) = h^2 t(t-1)(t-2)\frac{f'''(\xi)}{3!}.$$

Puesto que f es un polinomio de grado 3, $f'''(x)$ es una constante, sea $f'''(x) = c \quad \forall x \in [a,b]$. Resulta

$$\begin{aligned}\int_a^b \varepsilon(x)\,dx &= \int_a^b (f(x) - P(x))\,dx = \int_0^2 h^2 t(t-1)(t-2)\frac{c}{3!}dt \\ &= \frac{h^2 c}{3!}\int_0^2 t(t-1)(t-2)\,dt = \frac{h^2 c}{3!}\left(\frac{1}{4}t^4 - t^3 + t^2\right)\Big|_0^2 = 0.\end{aligned}$$

Por lo tanto

$$\varepsilon_f = I(f) - G(f) = \int_a^b f(x)\,dx - \int_a^b P(x)\,dx = \int_a^b \varepsilon(x)\,dx = 0.$$

iii) Sea $f \in C^4([a,b])$ cualquiera. El resultado que acabamos de obtener en la parte ii) muestra que la fórmula de cuadratura dada por la regla de Simpson es exacta para polinomios de grado 3, por lo que podemos construir un polinomio de interpolación de grado 3 que mejore la precisión de $I(f)$. Busquemos un polinomio P de grado 3 que verifique las siguientes condiciones:

$$P(a) = f(a), \quad P\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right), \quad P(b) = f(b), \quad P'\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right).$$

Sea Q el polinomio de interpolación de f que pasa por los puntos $(a, f(a))$, $\left(\dfrac{a+b}{2}, f\left(\dfrac{a+b}{2}\right)\right)$, $(b, f(b))$, es decir que

$$Q(x) = f(a)\varphi_0(x) + f\left(\frac{a+b}{2}\right)\varphi_1(x) + f(b)\varphi_2(x) \quad x \in [a,b],$$

donde φ_0, φ_1, φ_2 son lo plinomios de interpolación de Lagrange.

Se define $P(x) = Q(x) + \alpha\omega(x) \quad x \in [a,b]$, donde $\alpha \in \mathbb{R}$ se debe determinar por la condición $P'\left(\dfrac{a+b}{2}\right) = f'\left(\dfrac{a+b}{2}\right)$, y ω es la función definida como $\omega(x) = (x-a)\left(x - \dfrac{a+b}{2}\right)(x-b)$ $x \in [a,b]$.

De la definición de P, es claro que

$$P(a) = f(a), \quad P\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right), \quad P(b) = f(b).$$

Derivando la función P, se tiene $P'(x) = Q'(x) + \alpha\omega'(x)$ con

$$\omega'(x) = (x-a)\left(x - \frac{a+b}{2}\right) + (x-a)(x-b) + \left(x - \frac{a+b}{2}\right)(x-b).$$

Entonces

$$\omega'\left(\frac{a+b}{2}\right) = \left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right) = -\frac{(b-a)^2}{4},$$

$$P'\left(\frac{a+b}{2}\right) = Q'\left(\frac{a+b}{2}\right) + \alpha\omega'\left(\frac{a+b}{2}\right) = Q'\left(\frac{a+b}{2}\right) - \frac{\alpha}{4}(b-a)^2.$$

Puesto que $\alpha \in \mathbb{R}$ es tal que $P'\left(\dfrac{a+b}{2}\right) = f'\left(\dfrac{a+b}{2}\right)$, entonces $\alpha \in \mathbb{R}$ satisface la igualdad

$$f'\left(\frac{a+b}{2}\right) = Q'\left(\frac{a+b}{2}\right) - \frac{\alpha}{4}(b-a)^2,$$

lo que a su vez permite elegir α como sigue:

$$\alpha = -4\frac{\left(f'\left(\frac{a+b}{2}\right) - Q'\left(\frac{a+b}{2}\right)\right)}{(b-a)^2} = \frac{4}{(b-a)^2}\left(Q'\left(\frac{a+b}{2}\right) - f'\left(\frac{a+b}{2}\right)\right),$$

entonces

$$P'(x) = Q'(x) + \frac{4}{(b-a)^2}\left(Q'\left(\frac{a+b}{2}\right) - f'\left(\frac{a+b}{2}\right)\right)\omega(x) \quad x \in [a,b].$$

Se verifica inmediatamente que $P'\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right)$.

Determinemos el error de interpolación $\varepsilon(x)$ para este polinomio de interpolación, esto es $\varepsilon(x) = f(x) - P(x) \quad x \in [a,b]$. Con este propósito definimos la función ψ siguiente:

$$\psi(t) = u(x)[f(t) - P(t)] - u(t)[f(x) - P(x)] \quad t \in [a,b],$$

donde $x \in [a,b]$ es fijo y $u(t) = (t-a)\left(t - \frac{a+b}{2}\right)^2(t-b)$.

Puesto que $P(a) = f(a), \quad P\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right), \quad P(b) = f(b)$, se verifica inmediatamente que $\psi(a) = \psi\left(\frac{a+b}{2}\right) = \psi(b) = \psi(x) = 0$. Así, la función ψ tiene cuatro raíces en el intervalo $[a,b]$. Por el teorema de Rolle (véase en Calculus I de Apostol, página 224), $\psi'(t)$ tiene cuatro raíces, pués en $t = \frac{a+b}{2}$ también se anula; $\psi''(t)$ tiene tres raíces, $\psi'''(t)$ tiene dos raíces, $\psi^{iv}(t)$ tiene una raíz y sea $\xi \in [a,b]$ tal que $\psi^{iv}(\xi) = 0$. Puesto que

$$\psi^{iv}(t) = u(x)\left[f^{iv}(t) - P^{iv}(t)\right] - u^{iv}(t)[f(x) - P(x)] \quad t \in [a,b].$$

De la definición del polinomio P se tiene $P^{iv}(t) = 0$, de la definición del polinomio u, $u^{iv}(t) = 4!$. Entonces

$$\begin{aligned} \psi^{iv}(t) &= u(x) f^{iv}(t) - 4!\,(f(x) - P(x)), \\ 0 &= \psi^{iv}(\xi) = u(x) f^{iv}(\xi) - 4!\,(f(x) - P(x)), \end{aligned}$$

de donde

$$\varepsilon(x) = f(x) - P(x) = \frac{f^{iv}(\xi)}{4!}u(x) \quad x \in [a,b].$$

Calculemos el error de integración de$f$ usando la fórmula de cuadratura dada por la regla de Simpson:

$$\varepsilon_f = I(f) - G(f).$$

Como $f \in C^4([a,b])$, sea $M = \max\limits_{x\in[a,b]} \left|f^{iv}(x)\right|$, $m = \min\limits_{x\in[a,b]} \left|f^{iv}(x)\right|$. Entonces

$$\varepsilon_f = \int_a^b \varepsilon(x)\,dx = \int_a^b \frac{f^{iv}(\xi)}{4!}u(x)\,dx.$$

Además, $m \leq \left|f^{iv}(\xi)\right| \leq M$ y ξ depende de x, en consecuencia

$$\frac{u(x)}{4!}m \leq \frac{\left|f^{iv}(\xi)\right|}{4!}|u(x)| \leq \frac{M}{4!}|u(x)|$$

e integrando sobre el intervalo $[a,b]$ se obtiene la siguiente desigualdad

$$\frac{m}{4!}\int_a^b |u(x)|\,dx \leq \int_a^b \frac{\left|f^{iv}(\xi)\right|}{4!}|u(x)\,dx| \leq \frac{M}{4!}\int_a^b |u(x)|\,dx.$$

Para calcular $\int_a^b u(x)\,dx$ realizamos el siguiente cambio de variable: $t = \dfrac{x-a}{h}$ con $h = \dfrac{b-a}{2}$. Tenemos

$$\begin{aligned} u(x) &= (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b) = h^4 t\,(t-1)^2\,(t-2) \\ &= h^4\left(t^4 - 4t^3 + 5t^2 - 2t\right) \quad t \in [0,2]\,. \end{aligned}$$

Luego

$$\int_a^b u(x)\,dx = h^4 \int_0^2 \left(t^4 - 4t^3 + 5t^2 - 2t\right) dt = -\frac{4}{15}h^5,$$

y siendo $u(x) \leq 0 \quad \forall x \in [a,b]$, entonces $|u(x)| = -u(x) \quad x \in [a,b]$ con lo que $\int_a^b |u(x)|\,dx = \dfrac{4}{15}h^5$.

Resulta que

$$\begin{aligned} \frac{m}{4!}\frac{4}{15}h^5 &\leq \int_a^b \frac{\left|f^{iv}(\xi)\right|}{4!}\,|u(x)\,dx| \leq \frac{M}{4!}\frac{4}{15}h^5, \\ m\frac{h^5}{90} &\leq \int_a^b \frac{1}{4!}\left|f^{iv}(\xi)\,u(x)\right| dx \leq M\frac{h^5}{90}, \end{aligned}$$

y de esta desigualdad se obtiene la siguiente:

$$m \leq \frac{90}{h^5}\int_a^b \frac{1}{4!}\left|f^{iv}(\xi)\,u(x)\right| dx \leq M.$$

Aplicando el teorema del valor intermedio, existe $\lambda \in [a,b]$ tal que

$$\left|f^{iv}(\lambda)\right| = \frac{90}{h^5}\int_a^b \frac{1}{4!}\left|f^{iv}(\xi)\,u(x)\right| dx,$$

de donde

$$|\varepsilon_f| = |I(f) - G(f)| = \left|\int_a^b \frac{1}{4!}f^{iv}(\xi)\,u(x)\right| dx \leq \int_a^b \frac{1}{4!}\left|f^{iv}(\xi)\,u(x)\right| dx = \frac{h^5}{90}\left|f^{iv}(\lambda)\right|.$$

Puesto que $h = \dfrac{b-a}{2}$, entonces

$$|\varepsilon_f| \leq \frac{(b-a)^5}{2880}\left|f^{iv}(\lambda)\right| \leq \frac{M}{2880}(b-a)^5\,.$$

Aplicamos este resultado para estimar el error en la aproximación de $I(f)$ mediante la fórmula de Simpson generalizada, que se trata a continuación.

Error de aproximación con la fórmula de Simpson generalizada

Sean $n \in \mathbb{Z}^+$, $\tau(n)$ una partición del intervalo $[a,b]$ con $x_{k-1} < x_k \quad k = 1,\ldots,n$, $h_k = \dfrac{1}{2}(x_k - x_{k-1})$, $\widehat{h} = \max\limits_{k=1,\ldots,n} h_k$. Entonces, para cada $k = 1,\ldots,n$ se tiene

$$\left|\varepsilon_f^{(k)}\right| \leq \frac{(x_k - x_{k-1})^5}{2880}\left|f^{iv}(\lambda_k)\right| \leq \frac{M_k}{2880}(x_k - x_{k-1})^5\,,$$

con $\lambda_k \in [x_{k-1}, x_k]$, $M_k = \max\limits_{x\in[x_{k-1},x_k]}\left|f^{iv}(x)\right|$.

Además,

$$\varepsilon_f = I(f) - S_n(f) = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} f(x)\,dx - \sum_{k=1}^n \int_{x_{k-1}}^{x_k} G_k(f)\,dx = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} (f(x) - G(x))\,dx = \sum_{k=1}^n \varepsilon_f^{(k)},$$

donde $G_k(f) = \dfrac{x_k - x_{k-1}}{6}\left(f(x_{k-1}) + 4f\left(\dfrac{x_{k-1}+x_k}{3}\right) + f(x_k)\right) \quad k = 1, \ldots, n$, es la regla de Simpson aplicada a cada intervalo $[x_{k-1}, x_k]$, y

$$\left|\varepsilon_f^{(k)}\right| = \left|\int_{x_{k-1}}^{x_k} (f(x) - G(f))\, dx\right| \leq \frac{(x_k - x_{k-1})^5}{2880}\left|f^{iv}(\lambda_k)\right| \leq \frac{M_k}{2880}(x_k - x_{k-1})^5.$$

Entonces

$$|\varepsilon_f| = |I(f) - S_n(f)| \leq \sum_{k=1}^{n}\left|\varepsilon_f^{(k)}\right| \leq \sum_{k=1}^{n}\frac{(x_{k-1} - x_k)^5}{2880}\left|f^{iv}(\lambda_k)\right| = \sum_{k=1}^{n}\frac{h_k^5}{90}\left|f^{iv}(\lambda_k)\right| \leq \frac{\widehat{h}^5}{90}\sum_{k=1}^{n}\left|f^{iv}(\lambda_k)\right|.$$

Puesto que $m \leq \left|f^{iv}(\lambda_k)\right| \leq M \quad k = 1, \ldots, n$, se sigue que

$$nm \leq \sum_{k=1}^{n}\left|f^{iv}(\lambda_k)\right| \leq nM$$

y por el teorema del valor intermedio, existe $\lambda \in [a, b]$ tal que

$$\frac{1}{n}\sum_{k=1}^{n}\left|f^{iv}(\lambda_k)\right| = \left|f^{iv}(\lambda)\right|,$$

luego

$$\frac{\widehat{h}^5}{90}\sum_{k=1}^{n}\left|f^{iv}(\lambda_k)\right| = \frac{n}{90}\widehat{h}^5\left|f^{iv}(\lambda)\right| \leq \frac{n}{90}\widehat{h}^5 M.$$

Así,

$$|\varepsilon_f| \leq \frac{n}{90}\widehat{h}^5\left|f^{iv}(\lambda)\right| \leq \frac{n}{90}\widehat{h}^5 M.$$

En el caso de una partición uniforme, $\widehat{h} = h = \dfrac{b-a}{2n}$ se tiene la siguiente estimación del error de integración:

$$|\varepsilon_f| \leq \frac{n}{90}h^5\left|f^{iv}(\lambda)\right| = \frac{n}{90}h^4\frac{b-a}{2n}\left|f^{iv}(\lambda)\right| = \frac{b-a}{2n}h^4\left|f^{iv}(\lambda)\right| \leq \frac{(b-a)M}{180}h^4,$$

y de esta estimación resulta

$$|\varepsilon_f| = \left|\int_a^b f(x)\, dx - S_n(f)\right| \leq \frac{(b-a)M}{180}h^4 \underset{h\to 0}{\longrightarrow} 0.$$

o lo que es lo mismo $\lim\limits_{n\to\infty} S_n(f) = \displaystyle\int_a^b f(x)\, dx$, que muestra que el método de integración mediante la fórmula de Simpson generalizada es convergente.

En el caso de la estimación

$$|\varepsilon_f| = \left|\int_a^b f(x)\, dx - S_n(f)\right| \leq \frac{n}{90}\widehat{h}^5 M,$$

con $\widehat{h} = \max\limits_{k=1,\ldots,n} h_k$, se requiere de una hipótesis suplementaria sobre cada h_k, esto es, la partición del intervalo $[a, b]$ debe ser regular, es decir que existe $\sigma \geq 1$ tal que $h_k \leq \sigma\dfrac{b-a}{2n} \quad k = 1, \cdots, n$, entonces $\widehat{h} \leq \sigma\dfrac{b-a}{2n}$ y en consecuencia

$$|\varepsilon_f| = \left|\int_a^b f(x)\, dx - S_n(f)\right| \leq \sigma\frac{b-a}{190}\widehat{h}^4 M \underset{n\to\infty}{\longrightarrow} 0,$$

que prueba la convergencia del método de integración numérica.

2.8. Integrales dobles

Sea Ω un subconjunto cerrado y acotado de $\mathbb{R}^2$ y $f \in C(\Omega)$. Se desea calcular $I(f) = \iint\limits_{\Omega} f(x,y)dxdy$.

En el caso de dominios sencillos como un disco o un rectángulo y funciones f aparentemente simples, el cálculo de $I(f)$ puede resultar muy dificultoso y en muchas situaciones imposible, más aún, para dominios muy generales y funciones que no se integran mediante funciones elementales, el cálculo de $I(f)$ resulta imposible, por lo que dicha integral tendrá que ser aproximada numéricamente. Con este propósito, consideramos las regiones o dominios Ω de los tipos I y II que se indican a continuación.

1. Sea $[a,b]$ un intervalo cerrado de $\mathbb{R}$. Se dice que Ω es una región o dominio del tipo I si

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \mid \varphi_1(x) \leq y \leq \varphi_2(x),\ x \in [a,b]\right\},$$

donde φ_1, φ_2 son funciones continuas en $[a,b]$ tales que $\varphi_1 \leq \varphi_2$. En la figura siguiente se muestra una región Ω del tipo I.

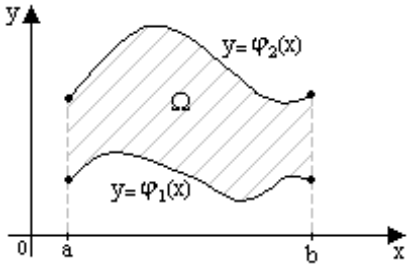


Figura 26

2. Sea $[c,d]$ un intervalo cerrado de $\mathbb{R}$. Se dice que Ω es un dominio o región del tipo II si

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \mid \Psi_1(y) \leq x \leq \Psi_2(y),\ y \in [c,d]\right\},$$

donde $\Psi_1,\ \Psi_2$ son funciones continuas en $[c,d]$ tales que $\Psi_1 \leq \Psi_2$. En la figura siguiente se muestra una región Ω del tipo II.

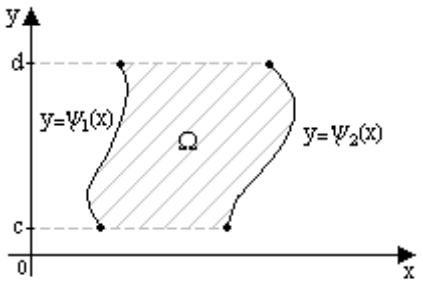


Figura 27

Los dominios Ω muy complejos pueden descomponerse en forma apropiada en subdominios que correspondan a uno de estos tipos precisados, por lo que el cálculo aproximado de $I(f)$ se reduce al cálculo de la integral doble de la función f sobre cada subdominio de la descomposición de Ω que se haya establecido. Por otro lado, para dominios Ω como un rectángulo o regiones del plano del tipo I o II puede aplicarse la regla de los trapecios generalizada, la regla de Simpson generalizada. Nos limitamos a la aplicación de la regla de los trapecios para regiones del tipo I. Para regiones Ω del tipo II se procede en forma muy similar. Igualmente la aplicación de la regla de Simpson generalizada se aplica en forma muy parecida a la de los trapecios generalizada.

Sea $\Omega \in \mathbb{R}^2$ una región del tipo I, esto es,

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \mid \varphi_1(x) \leq y \leq \varphi_2(x) \quad x \in [a,b]\right\},$$

donde φ_1, φ_2 son funciones continuas en $[a,b]$ tales que $\varphi_1(x) \leq \varphi_2(x) \quad \forall x \in [a,b]$.

Sea $f \in C(\Omega)$ e $I(f) = \iint\limits_{\Omega} f(x,y)\,dxdy$. esta integral lo expresamos como una integral reiterada siguiente:

$$I(f) = \int_a^b \left(\int_{y=\varphi_1(x)}^{y=\varphi_2(x)} f(x,y)\,dy \right) dx.$$

Definimos $g(x) = \displaystyle\int_{y=\varphi_1(x)}^{y=\varphi_2(x)} f(x,y)\,dy \quad x \in [a,b]$. Entonces $I(f) = \int_a^b g(x)\,dx$. Apliquemos la fórmula de los trapecios generalizada con una partición uniforme del intervalo $[a,b]$. Para el efecto, sea $n \in \mathbb{Z}^+$. Ponemos $h = \dfrac{b-a}{n}$, $x_j = a + h \quad j = 0,1,\ldots,n$. La regla de los trapecios generalizada para aproximar la integral $I(f) = \int_a^b g(x)\,dx$ se escribe como sigue:

$$T_n(g) = \frac{h}{2}\left(g(a) + g(b)\right) + h\sum_{j=1}^{n-1} g(x_j).$$

Además, de la definición de la función g se tiene

$$\begin{aligned} g(a) &= \int_{\varphi_1(a)}^{\varphi_2(a)} f(a,y)\,dy, \\ g(x_j) &= \int_{\varphi_1(x_j)}^{\varphi_2(x_j)} f(x_j,y)\,dy \quad j = 1,\ldots,n-1, \\ g(b) &= \int_{\varphi_1(b)}^{\varphi_2(b)} f(b,y)\,dy. \end{aligned}$$

Todas estas integrales las expresaremos en la forma

$$I_j(f) = \int_{\varphi_1(x_j)}^{\varphi_2(x_j)} f(x_j,y)\,dy \quad j = 0,1,\ldots,n,$$

las mismas que a su vez pueden ser aproximadas con la regla de los trapecios generalizada como se muestra a continuación.

Sea $m \in \mathbb{Z}^+$. Se deine $h_j = \dfrac{1}{m}\left(\varphi_2(x_j) - \varphi_1(x_j)\right)$ y $y_k = \varphi_1(x_j) + kh_j \quad k = 0,1,\ldots,m$. Entonces

$$T_m^{(j)}(f) = \frac{h_j}{2}\left(f(x_j,\varphi_1(x_j)) + f(x_j,\varphi_2(x_j)) + h_j\sum_{k=1}^{m-1} f(x_j,y_k) \right) \quad j = 0,1,\ldots,n,$$

en consecuencia. $I_j(f) \simeq T_m^{(j)}(f)$, y

$$T_n(g) \simeq \frac{h}{2}\left(T_m^{(0)}(f) + T_m^{(n)}(f)\right) + h\sum_{j=1}^{n-1} T_m^{(j)}(f).$$

Así,

$$I(f) \simeq \frac{h}{2}\left(T_m^{(0)}(f) + T_m^{(n)}(f)\right) + h\sum_{j=1}^{n-1} T_m^{(j)}(f).$$

Ponemos $T_{mn}(f) = \dfrac{h}{2}\left(T_m^{(0)}(f) + T_m^{(n)}(f)\right) + h\sum\limits_{j=1}^{n-1} T_m^{(j)}(f)$ que es la formulación de la regla de los trapecios generalizada para regiones del tipo I. Esta es una forma lineal en $C(\Omega)$.

Mediante un procedimiento similar se establece la formulación de la regla de los trapecios generalizada para regiones del tipo II, la misma que se propone como ejercicio.

Para elaborar el algoritmo para el cálculo aproximado de una integral doble de una función sobre una región Ω del tipo I con la regla de los trapecios generalizada requiere de la siguiente información: intervalo $[a,b]$ y en consecuencia los extremos a y b de dicho intervalo, las funciones continuas φ_1, φ_2 en $[a,b]$ de modo que $\varphi_1(x) \leq \varphi_2(x) \quad x \in [a,b]$, la función continua a integrar f definida en Ω, el número de puntos n de la partición uniforme del intervalo $[a,b]$ que lo llamaremos partición horizontal, el número de puntos m de la partición del intervalo $[\varphi_1(x_j), \varphi_2(x_j)] \quad j = 0,1,\ldots,n$ a la que lo llamaremos particiones verticales.

Algoritmo

Datos de entrada: $m, n \in \mathbb{Z}^+$, $a, b \in \mathbb{R}$, funciones φ_1, φ_2, f.

Datos de salida: $T_{mn}(f)$, mensaje.

1. Verificar $a < b$, caso contrario continuar en 7).
2. $h = \dfrac{b-a}{n}$.
3. $S = 0$.
4. Para $j = 0,\ldots,n$

   $x_j = a + jh$

   $h_j = \frac{1}{m}(\varphi_2(x_j) - \varphi_1(x_j))$

   $S_1 = 0$

   Para $k = 1,\ldots,m-1$

   $y_k = \varphi_1(x_j) + khj$

   $S_1 = S_1 + f(x_j, y_k)$

   Fin de bucle k.

   $S_1 = hjS_1 + \frac{1}{2}hj\,(f(x_j, \varphi_1(x_j)) + f(x_j, \varphi_2(x_j)))$

   Si $j = 0, \quad z_1 = S_1$.

   Si $j = n, \quad z_2 = S_1$.

   Si $0 < j < n$,

   $S = S + S_1$

   Fin de bucle j.
5. $S = \frac{1}{2}h(z_1 + z_2) + hS$.
6. Imprimir $T_{mn}(f) = S$. Continuar en 8).
7. Mensaje: $a < b$.
8. Fin.

Ejemplos

1. Consideremos el problema (P) siguiente: $I = \int_0^1 \left(\int_0^1 ye^{xy}dx\right)dy$. Notemos que I podemos calcularlo exactamente. Pués

$$I = \int_0^1 \left(\int_0^1 ye^{xy}dx\right)dy = \int_0^1 e^{xy}|_0^1 dy = \int_0^1 (e^y - 1)dy = e^y - y\Big|_0^1 = e - 2 \simeq 0{,}718281828.$$

Apliquemos la regla de los trapecios generalizada para aproximar I. Para el efecto, sean $m = 5,\ h_x = \frac{1}{m} = 0{,}2,\ x_j = jh_x,\ j = 0,1,\ldots,5;\ n = 5,\ h_y = \frac{1}{n} = 0{,}2,\ y_k = kh_y,\ k = 0,1,\ldots,5$; y definimos la función g como sigue: $g(y) = \int_0^1 ye^{xy}dx$. Se tiene $I = \int_0^1 g(y)dy$ y utilizando la fórmula de los trapecios generalizada, resulta

$$I(g) = \int_0^1 g(y)dy \simeq \frac{h_y}{2}(g(0) + g(1)) + h_y(g\,(0{,}2) + g\,(0{,}4) + g\,(0{,}6) + g\,(0{,}8)\,.$$

Calculemos $g(y_k)$ para $k = 0,1,\ldots,5$, y aproximemos usando la regla de los trapecios generalizada. Tenemos los siguientes resultados:

$$g(0) = 0,$$

$$g(0{,}2) = \int_0^1 0{,}2e^{0{,}2x}dx \simeq 0{,}1\left[0{,}2\left(1 + 2\left(e^{0{,}04} + e^{0{,}08} + e^{0{,}12} + e^{0{,}16}\right) + e^{0{,}2}\right)\right] \simeq 0{,}2214322787,$$

$$\begin{aligned}
g(0{,}4) &= \int_0^1 0{,}4e^{0{,}4x}dx \simeq 0{,}1\left[0{,}4\left(1 + 2\left(e^{0{,}08} + e^{0{,}16} + e^{0{,}24} + e^{0{,}32}\right) + e^{0{,}4}\right)\right] \simeq 0{,}492086976,\\
g(0{,}6) &= \int_0^1 0{,}6e^{0{,}6x}dx \simeq 0{,}1\left[0{,}6\left(1 + 2\left(e^{0{,}12} + e^{0{,}24} + e^{0{,}36} + e^{0{,}48}\right) + e^{0{,}6}\right)\right] \simeq 0{,}8231051064,\\
g(0{,}8) &= \int_0^1 0{,}8e^{0{,}8x}dx \simeq 0{,}1\left[0{,}8\left(1 + 2\left(e^{0{,}16} + e^{0{,}32} + e^{0{,}48} + e^{0{,}64}\right) + e^{0{,}8}\right)\right] \simeq 1{,}228154301,\\
g(1) &= \int_0^1 e^x dx \simeq 0{,}1\left[1 + 2\left(e^{0{,}2} + e^{0{,}4} + e^{0{,}6} + e^{0{,}8}\right) + e\right] \simeq 1{,}72400562.
\end{aligned}$$

Luego, utilizando la fórmula de los trapecios generalizada, resulta

$$I \simeq 0{,}1(g(0) + g(1)) + 0{,}2[g(0{,}2) + g(0{,}4) + g(0{,}6) + g(0{,}8)] \simeq 0{,}7253562942.$$

En la tabla siguiente se muestran los resultados de la aplicación del algoritmo para diferentes valores de $m = n$, y la estimación del error.

n	m	$T_{mn}\,(f)$	Error: $\lvert I\,(f) - Tmn\,(f)\rvert$
5	5	$7{,}2535629421 \times 10^{-1}$	$7{,}074465748 \times 10^{-3}$
10	10	$7{,}2003851477 \times 10^{-1}$	$1{,}7566863064 \times 10^{-3}$
20	20	$7{,}1872025130 \times 10^{-1}$	$4{,}3842284012 \times 10^{-4}$
40	40	$7{,}1839138732 \times 10^{-1}$	$1{,}0955886528 \times 10^{-4}$
80	80	$7{,}1830921525 \times 10^{-1}$	$2{,}7386787759 \times 10^{-5}$
160	160	$7{,}1828867497 \times 10^{-1}$	$6{,}8465138930 \times 10^{-6}$
320	320	$7{,}1828354008 \times 10^{-1}$	$1{,}7116170327 \times 10^{-6}$
640	640	$7{,}1828225636 \times 10^{-1}$	$4{,}2790354282 \times 10^{-7}$
1280	1280	$7{,}1828193543 \times 10^{-1}$	$1{,}0697584074 \times 10^{-7}$
2560	2560	$7{,}1828185520 \times 10^{-1}$	$2{,}6743957271 \times 10^{-8}$

.

2. Calculemos la integral doble $I = \int_1^2 \int_x^{x^2} \left(x^2 + xy + y^2\right) dxdy$. Para el efecto, primeramente identificamos el tipo de región Ω sobre la que tenemos que integrar la función f definida como $f(x,y) = x^2 + xy + y^2$. Tenemos $\Omega = \left\{(x,y) \in \mathbb{R}^2 \mid x \le y \le x^2 \quad x \in [1,2]\right\}$ que corresponde a una región del tipo I. Ponemos $\varphi_1\,(x) = x$, $\varphi_2\,(x) = x^2 \quad x \in [1,2]$. En la figura siguiente se muestra el

dominio Ω.

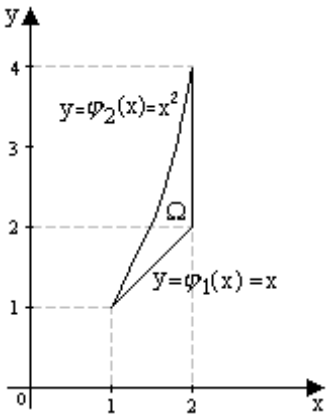


Figura 28

Calculemos I exactamente. Tenemos

$$\begin{aligned} I &= \int_1^2 \int_x^{x^2} (x^2+xy+y^2)dxdy = \int_1^2 \left(\int_x^{x^2} (x^2+xy+y^2)dy \right) dx \\ &= \int_1^2 \left(x^2y + \frac{1}{2}xy^2 + \frac{1}{3}y^3 \right)\Big|_x^{x^2} dx = \int_1^2 \left(x^4 + \frac{1}{2}x^5 + \frac{1}{3}x^6 - x^3 - \frac{1}{2}x^3 - \frac{1}{3}x^3 \right) dx \\ &= \int_1^2 \left(x^4 + \frac{1}{2}x^5 + \frac{1}{3}x^6 - \frac{11}{6}x^3 \right) dx = \left(\frac{1}{5}x^5 + \frac{1}{12}x^6 + \frac{1}{21}x^7 - \frac{11}{24}x^4 \right)\Big|_1^2 = \frac{8923}{840}, \\ I &= 10{,}62261904\ldots \end{aligned}$$

Sea $g(x) = \int_x^{x^2}(x^2+xy+y^2)dy \quad x \in [1,2]$. Entonces $I = \int_1^2 g(x)dx$. Apliquemos el método de los trapecios generalizada con $m = 5$. Sea $h_x = \dfrac{2-1}{5} = 0{,}2$, $x_j = 1 + jh_x = 1 + 0{,}2j$ para $j = 0,1,2,3,4,5$. Luego

$$I \simeq \frac{h_x}{2} \sum_{j=1}^{5} [g(x_{j-1}) + g(x_j)] = \frac{h_x}{2}(g(1)+g(2)) + h_x(g(1{,}2) + g(1{,}4) + g(1{,}6) + g(1{,}8)),$$

donde

$$\begin{aligned} g(1) &= \int_1^1 (x+y+y^2)dy = 0, \qquad g(1{,}2) = \int_{1,2}^{1,41} (1{,}44 + 1{,}2y + y^2)dy, \\ g(1{,}4) &= \int_{1,4}^{1,96} (1{,}96 + 1{,}4y + y^2)dy, \qquad g(1{,}6) = \int_{1,6}^{2,56} (2{,}56 + 1{,}6y + y^2)dy, \\ g(1{,}8) &= \int_{1,8}^{3,24} (3{,}24 + 1{,}8y + y^2)dy, \qquad g(2) = \int_2^4 (4 + 2y + y^2)dy. \end{aligned}$$

Apliquemos nuevamente el método de los trapecios para aproximar $g(x_j),\ j = 1,\ldots,5$.

Sea $n = 5,\ h_j = \dfrac{x_j^2 - x_j}{5}$ y $y_x = x_j + kh_j,\ k = 0,1,\ldots,5$, luego

$$\begin{aligned} g(x_j) &\simeq \frac{h_j}{2} \sum_{k=1}^{5} [f(x_j, y_{x-1}) + f(x_j, y_x)] \\ &= \frac{h_j}{2} \left[f(x_j, y_0) + f(x_j, y_5) \right] + h_j \left[f(x_j, y_1) + f(x_j, y_2) + f(x_j, y_3) + f(x_j, y_y) \right], \end{aligned}$$

donde $f(x_j, y) = x_j^2 + x_j y + y^2 = x_j^2 + y(x_j + y)$.

Para $j = 1$, $h_1 = 0{,}048$, $\quad y_x = 1{,}2 + kh_1 \quad k = 0, \cdots, 5$, los puntos y_x de la partición vertical son $y_x = 1{,}2,\ 1{,}248,\ 1{,}296,\ 1{,}344,\ 1{,}392,\ 1{,}44$. La función f en el punto $(1{,}2, y)$ está definida como:

$$f(1{,}2, y) = 1{,}44 + y(1{,}2 + y),$$

luego para $y = y_x = 1{,}2,\ 1{,}248,\ 1{,}296,\ 1{,}344,\ 1{,}392,\ 1{,}44$, se obtienen los siguientes resultados:

$$\begin{aligned} f(1{,}2, 1{,}2) &= 4{,}32, \quad f(1{,}2,\ 1{,}248) = 4{,}495104, \quad f(1{,}2,\ 1{,}296) = 4{,}674816, \\ f(1{,}2, 1{,}344) &= 4{,}859136, \quad f(1{,}2, 1{,}392) = 5{,}048064, \quad f(1{,}2, 1{,}44) = 5{,}2416, \end{aligned}$$

y por la regla de los trapecios generalizada y la definición de $g(1{,}2)$, resulta

$$g(1{,}2) \simeq 0{,}021 \times 9{,}5116 + 0{,}048 \times 19{,}07712 = 1{,}14518016.$$

Para $j = 2,\ x_2 = 1{,}4,\ h_2 = \dfrac{1{,}96 - 1{,}4}{5} = 0{,}112,\ y_x = 1{,}4 + kh_2 \quad k = 0, \cdots, 5$, los puntos y_x de la partición vertical son $y_x = 1{,}4,\ 1{,}512,\ 1{,}624,\ 1{,}736,\ 1{,}848,\ 1{,}96$. La función f en el punto $(1{,}4, y)$ está definida como:

$$f(1{,}4,\ y) = 1{,}96 + y(1{,}4 + y),$$

y en consecuencia para $y = y_x = 1{,}4,\ 1{,}512,\ 1{,}624,\ 1{,}736,\ 1{,}848,\ 1{,}96$, se tiene

$$\begin{aligned} f(1{,}4, 1{,}4) &= 5{,}88, \quad f(1{,}4, 1{,}512) = 6{,}362944, \quad f(1{,}4, 1{,}624) = 6{,}870976, \\ f(1{,}4, 1{,}736) &= 7{,}404056, \quad f(1{,}4, 1{,}848) = 7{,}962304, \quad f(1{,}4, 1{,}96) = 8{,}5456, \end{aligned}$$

por la regla de los trapecios generalizada y la definición de $g(1{,}4)$, resulta

$$g(1{,}4) \simeq 0{,}056 \times 14{,}4256 + 0{,}112 \times 28{,}60032 = 4{,}01106944.$$

Para $j = 3,\ x_3 = 1{,}6,\ h_3 = \dfrac{2{,}56 - 1{,}6}{5} = 0{,}192; \quad y_x = 1{,}6 + kh_3 \quad k = 0, \cdots, 5$, los puntos de la partición vertical son $y_x = 1{,}6,\ 1{,}792,\ 1{,}984,\ 2{,}176,\ 2{,}368,\ 2{,}56$. La función f en el punto $(1{,}6, y)$ está definida como:

$$f(1{,}6, y) = 2{,}56 + y(1{,}6 + y),$$

y para $y = y_x = 1{,}6,\ 1{,}792,\ 1{,}984,\ 2{,}176,\ 2{,}368,\ 2{,}56$, se obtienen los siguientes resultados

$$\begin{aligned} f(1{,}6, 1{,}6) &= 7{,}68, \quad f(1{,}6, 1{,}792) = 8{,}638464, \quad f(1{,}6, 1{,}984) = 9{,}679656, \\ f(1{,}6, 2{,}176) &= 10{,}776576, \quad f(1{,}6, 2368) = 11{,}956224, \quad f(1{,}6, 2{,}56) = 13{,}2096. \end{aligned}$$

Por la regla de los trapecios generalizada y la definición de $g(1{,}6)$, se obtiene

$$g(1{,}6) \simeq 0{,}096 \times 20{,}8896 + 0{,}192 \times 41{,}04192 = 9{,}88545024.$$

Procediendo como en los casos enteriores, para $j = 4$, $x_4 = 1{,}8$, $h_4 = \dfrac{3{,}24 - 1{,}8}{5} = 0{,}288$, los puntos de la partición vertical son: $y_x = 1{,}8,\ 2{,}088,\ 2{,}376,\ 2{,}664,\ 2{,}952,\ 3{,}24$. La función f en el punto $(1{,}8, y)$ está dada como:

$$f(1{,}8,\ y) = 3{,}24 + y(1{,}8 + y),$$

$$\begin{aligned} f(1{,}8, 1{,}8) &= 9{,}72, \quad f(1{,}8, 2{,}088) = 11{,}358144, \quad f(1{,}8, 2{,}376) = 13{,}162176, \\ f(1{,}8, 2{,}664) &= 15{,}132096, \quad f(1{,}8, 2{,}952) = 17{,}267904, \quad f(1{,}8, 3{,}24) = 19{,}5696, \end{aligned}$$

$$g(1{,}8) \simeq 0{,}144 \times 29{,}2896 + 0{,}288 \times 56{,}92032 = 20{,}61075456.$$

Finalmente, para $j = 5$, $x_5 = 2$, $h_5 = \dfrac{4 - 2}{5} = 0{,}4$, $y_x = 2,\ 2{,}4,\ 2{,}8,\ 3{,}2,\ 3{,}6,\ 4$, y f en el punto $(1{,}2, y)$ está dada como:

$$f(2,\ y) = 4 + y(2 + y),$$

entonces

$$\begin{aligned} f(2,2) &= 12, \qquad f(2,2{,}4) = 14{,}56, \qquad f(2,2{,}8) = 17{,}44, \\ f(2,3{,}2) &= 20{,}64, \qquad f(2,3{,}6) = 24{,}16, \qquad f(2,4) = 28, \end{aligned}$$

y en consecuencia, por la regla de los trapecios generalizada y la definición de $g(2)$ resulta.

$$g(2) \simeq 0{,}2 \times 40 + 0{,}4 \times 76{,}8 = 38{,}72.$$

El valor aproximado de la integral doble $I = \int_1^2 \int_x^{x^2} \left(x^2 + xy + y^2\right) dxdy$ mediante la aplicación de la regla de los trapecios generalizada $T_{mn}(f)$ a la región del tipo I con $m = n$, es:

$$T_{mn}(f) = 0{,}1 \times 38{,}72 + 0{,}2 \times 35{,}6524544 = 11{,}00249088.$$

Este ejemplo pone de manifiesto dos aspectos: el volumen de cálculos a ejecutar y la precisión del cálculo. El primero conduce a la elaboración de un programa computacional y el segundo a una discretización más fina que permita mejorar la precisión. Este segundo punto se lo alcanza con la ejecución del programa computacional para discretizaciones más finas que a la mano son muy largas de ejecutarse. En la tabla siguiente se muestran los resultados de la aplicación del algoritmo.

n	m	$T_{mn}(f)$	Error=$\lvert I(f) - Tmn(f)\rvert$
5	5	11,0024908800	$3{,}7987183238 \times 10^{-1}$
10	10	10,7176023425	$9{,}4983294881 \times 10^{-2}$
20	20	10,6463657751	$2{,}3746727449 \times 10^{-2}$
40	40	10,6285557851	$5{,}9367374871 \times 10^{-3}$
80	80	10,6241032355	$1{,}4841878349 \times 10^{-3}$
160	160	10,6229900948	$3{,}7104717497 \times 10^{-4}$
320	320	10,6227118094	$3{,}2761807254 \times 10^{-5}$
640	640	10,6226422381	$2{,}3190452660 \times 10^{-5}$

Nota: Parecería razonable que con particiones horizontales y verticales muy finas, esto es, que tengan un gran número de puntos y que a su vez sean regulares, se podría aproximar tanto como se quiera la integral de una función continua. Lastimosamente, debido a los errores de redondeo, errores de truncameiento y de aproximación que intervienen en el cálculo de una integral doble, esto no es del todo cierto, pués para particiones con un número elevado de puntos, todos estos tipos de errores intervienen y deterioran los resultados. Por lo tanto, no es recomendable calcular aproximaciones de integrales con particiones regulares que tengan un gran número de puntos. Por este motivo que buscan otros métodos de aproximación que combinen con los métodos estudiados. Uno de estos métodos recomendables es la integración adaptativa que tiene muchas versiones. En la bibliografía se citan algunos textos en los que puede encontrar estos tópicos.

2.9. Ejercicios

1. Sea $T : \mathbb{R}^3 \to \mathbb{R}$ la aplicación lineal definida por $T(x, y, z) = ax + by + cz$ (x,y,z)$\in \mathbb{R}^3$, donde $a, b, c \in \mathbb{R}$ distintos entre sí y no todos nulos.

   a) Determine $\ker(T)$ para las distintas posibilidades de a, b, c e interprete geométricamente el resultado.

   b) Determine $[T]_B$, donde B es la base canónica de $\mathbb{R}^3$.

   c) Probar que todas las aplicaciones lineales de $\mathbb{R}^3$ en $\mathbb{R}$ son de la forma $T(x, y, z) = ax + by + cz$.

   d) Generalizar a) y b) a $\mathbb{R}^n$.

2. Sea $f \in (\mathbb{R}^n)^*$ no nulo.

   a) Pruebe que $0 < \dim(\ker(f)) < n$. **b)** Sea $B_n = \{\overrightarrow{e_1}, \ldots, \overrightarrow{e_n}\}$ la base canónica de $\mathbb{R}^n$, halle $[f]_{B_n}$.

   c) Sea $B = \{\overrightarrow{v_1}, \ldots, \overrightarrow{v_n}\}$ la base de $\mathbb{R}^n$ definida como sigue: $\overrightarrow{v_1} = \overrightarrow{e_1}$, $\overrightarrow{v_2} = \overrightarrow{e_1} + \overrightarrow{e_2}, \ldots, \overrightarrow{v_n} = \overrightarrow{e_1} + \ldots + \overrightarrow{e_n}$. Halle una base dual de B.

3. Para los datos S que en cada item se propone, hallar el polinomio de interpolación de Lagrange $P_h(x)$ y calcular el valor interpolado $P_h(\widehat{x})$ de una función f en el punto $\widehat{x}$ que se indica.

   a) $S = \{(0{,}1,\ 5)\,,\quad (0{,}2,\ 8)\}$, $\widehat{x} = 0{,}16$. **b)** $S = \{(0.,\ 2)\,,\quad (0{,}015,\ 4{,}1)\}$, $\widehat{x} = 0{,}0016$.

   c) $S = \{(-1{,}1,\ 0{,}25)\,,\quad (-1{,}02,\ 2{,}8)\}$, $\widehat{x} = -1{,}08$. **d)** $S = \{(2{,}1,\ 5{,}5)\,,\quad (2{,}25,\ 3.8)\}$, $\widehat{x} = 2{,}19$.

4. Para los datos S que en cada item se propone, hallar el polinomio de interpolación de Lagrange $P_h(x)$ y calcular el valor interpolado $P_h(\widehat{x})$ de una función f en el punto $\widehat{x}$ que se indica.

   a) $S = \{(1,\ 4)\,,\ (1{,}2,\ 5)\,,\ (1{,}5,\ 6{,}5)\}$, $\widehat{x} = 1{,}4$. **b)** $S = \{(-1{,}1,\ 3{,}5)\,,\ (-0{,}8,\ 4.5)\,,\ (-0{,}5,\ 3{,}5)\}$, $\widehat{x} = -0{,}94$. **c)** $S = \{(0{,}1,\ 1.4)\,,\quad (0{,}22,\ 2.5)\,,\quad (0{,}25,\ 1{,}56)\}$, $\widehat{x} = 0{,}145$.

   d) $S = \{(1{,}8,\ -4{,}2)\,,\ (2{,}2,\ \text{-}3.5)\,,\ (2{,}5,\ -4{,}5)\}$, $\widehat{x} = 1{,}995$.

5. Considerar la función f definida en cada item. Calcule $f'(x_0)$ para el punto x_0 que se indica. Calcule aproximaciones de $f'(x_0)$ mediante diferencias finitas centrales de primer orden para cada h que se indica, esto es $y_0' = \dfrac{f(x_0+h) - f(x_0-h)}{2h}$. Estime el error $|f'(x_0) - y_0'|$.

   a) $f(x) = 2x^2 - 5x + 1 \quad x \in \mathbb{R}$, $x_0 = -2$, $h = -0{,}0025$, $h = -0{,}000025$, $h = 0{,}003$, $h = 0{,}00003$.

   b) $f(x) = \dfrac{1}{(x^2+1)^3} \quad x \in \mathbb{R}$, $x_0 = 0$, $h = -0{,}002$, $h = -0{,}0002$, $h = 0{,}0032$, $h = 0{,}000032$.

   c) $f(x) = \sqrt{x^4 - 16} \quad |x| > 4$, $x_0 = 5$, $h = -0{,}0001$, $h = -0{,}00001$, $h = 0{,}00011$, $h = 0{,}000011$.

   d) $f(x) = \operatorname{sen}\left(x^3 + 2\right) \quad x \in \mathbb{R}$, $x_0 = \left(\dfrac{\pi}{3} - 2\right)^{\frac{1}{3}}$, $h = -0{,}004$, $h = -0{,}0004$, $h = 0{,}00041$, $h = 0{,}00001$.

   e) $f(x) = \cos^2\left(\sqrt{x+1}\right) \quad x > -1$, $x_0 = \left(\dfrac{\pi}{2}\right)^2 - 1$, $h = -0{,}0011$, $h = -0{,}00011$, $h = 0{,}0002$, $h = 0{,}00002$. Sugerencia: aproxime π con 9 cifras de precisión.

   f) $f(x) = \ln\left(16 - x^2\right) \quad |x| < 4$, $x_0 = 1$, $h = -0{,}004$, $h = -0{,}0004$, $h = 0{,}0005$, $h = 0{,}00002$.

   g) $f(x) = 2\ln(x) + 3\ln^2(x) + 4\ln^3(x) \quad x > 0$, $x_0 = e$, $h = -0{,}01$, $h = -0{,}0001$, $h = 0{,}001$, $h = 0{,}00001$. Sugerencia: aproxime e con 9 cifras de precisión.

6. Considerar el polinomio P de segundo grado definido como $p(x) = \alpha x^2 + \beta x + \lambda \quad x \in \mathbb{R}$, $\alpha, \beta, \lambda \in \mathbb{R}$ con $\alpha \neq 0$; y, el polinomio de interpolación de Lagrange de P_h definido como $P_h(x) = p(a)\,\varphi_0(x) + p\left(\dfrac{a+b}{2}\right)\varphi_1(x) + p(b)\,\varphi_2(x) \quad x \in [a,b]$, donde φ_0, φ_1, φ_2 son los polinomios de interpolación de Lagrange de segundo grado definidos en $[a,b]$. Se prueba que $P(x) = P_h(x)$ $\forall x \in [a,b]$. En cada item se da un polinomio p de segundo grado y se restringe al intervalo $[a,b]$ que se indica. Hallar P_h y probar que $p(x) = P_h(x) \quad \forall x \in [a,b]$.

   a) $p(x) = x^2 - 1 \quad x \in [-1,1]$. **b)** $p(x) = 2x^2 + 5 \quad x \in [0,2]$. **c)** $p(x) = -x^2 + x + 1 \quad x \in [-1,2]$.

   d) $p(x) = 5x^2 \quad x \in [0,3]$. **e)** $p(x) = -3x^2 + 4x \quad x \in [1,10]$. **f)** $p(x) = 5x^2 + 7x - 1 \quad x \in [2,4]$.

7. En cada item se define una función u. Calcular $u''(x_0)$ en el punto x_0 que se indica. Aproximar $u''(x_0)$ mediante el uso de diferencias finitas centrales de segundo orden para cada $h > 0$ que se da.

   a) $u(x) = -x^3 + x^2 - 1 \quad x \in \mathbb{R}$, $x_0 = 1$, $h = 0{,}01$, $h = 0{,}001$, $h = 0{,}0001$.

   b) $u(x) = \operatorname{sen}^4(x) \quad x \in \mathbb{R}$, $x_0 = \dfrac{\pi}{6}$, $h = 0{,}002$, $h = 0{,}0002$, $h = 0{,}00002$.

   c) $u(x) = x^2 \exp\left(-x^2 + 3\right) \quad x \in \mathbb{R}$, $x_0 = \sqrt{3}$, $h = 0{,}0025$, $h = 0{,}0002$, $h = 0{,}00001$.

   d) $u(x) = x^3\sqrt{x^2 - 2} \quad x \in \mathbb{R}$, $x_0 = \sqrt{2}$, $h = 0{,}005$, $h = 0{,}0025$, $h = 0{,}00005$.

   e) $u(x) = \ln\left(x + \sqrt{1 + x^2}\right) \quad x \in \mathbb{R}$, $x_0 = -\sqrt{8}$, $h = 0{,}03$, $h = 0{,}003$, $h = 0{,}00003$.

   f) $u(x) = \cos\left(\pi x^2\right) \quad x \in \mathbb{R}$, $x_0 = -\dfrac{1}{2}$, $h = 0{,}001$, $h = 0{,}0001$, $h = 0{,}00001$.

8. En cada item se dan los valores $f(a+h)$ y $f(a-h)$ de una cierta función f en $x=a$ y $h\neq 0$. Calcular el valor aproximado de la derivada $f'(a)$.

   a) $f(1{,}005)=2{,}8117482$, $f(0{,}9995)=3{,}114231$.

   b) $f(5{,}00012)=-1{,}252231$, $f(4{,}99988)=0{,}00123112$.

   c) $f(-1{,}11231)=587{,}22314$, $f(-1{,}11211)=495{,}231427$.

   d) $f(-11{,}32145)=10{,}369725$, $f(11{,}32111)=42{,}223583$.

9. En cada literal se dan los valores $f(a+h)$, $f(a)$, $f(a-h)$ de una cierta función f y $h\neq 0$. Calcular el valor aproximado de la derivada segunda $f''(a)$ mediante diferencias finitas centrales.

   a) $f(0{,}0022)=3{,}852224$, $f(0)=1{,}8211253$, $f(-0{,}0022)=2{,}852536$.

   b) $f(1{,}3561)=8{,}923824$, $f(1{,}355)=15{,}162234$, $f(1{,}3490)=25{,}8542321$.

   c) $f(-10{,}4583)=0{,}312112$, $f(-10{,}4572)=4{,}852011$, $f(-10{,}4561)=3{,}2581423$.

   d) $f(20{,}34823)=-13{,}4585252$, $f(20{,}348)=-32{,}4525321$, $f(20{,}34777)=-52{,}85343211$.

10. Sea f una función real continua en $[a,b]$, $h=\dfrac{b-a}{3}$ y $\tau_3=\{a+jh \mid j=0,1,2,3\}$ una partición uniforme de $[a,b]$.

    a) Escriba los polinomios de interpolación de Lagrange φ_0, φ_1, φ_2, φ_3 definidos en $[a,b]$.

    b) Sea $x\in[a,b]$. Escribir el polinomio interpolante $P_h(x)$ de f en $[a,b]$.

    c) Suponga $f\in C^4([a,b])$. Escriba el error de interpolación de Lagrange.

    d) Sea f la función definida como $f(x)=x^3 \quad x\in[0,3]$. Aplique el resultado obtenido en b) y halle $P_h(x)$. Calcule $f(1{,}5)$ y $P_h(1{,}5)$ y verifique que $f(1{,}5)=P_h(1{,}5)$. Demuestre que $f(x)=P_h(x) \quad \forall x\in[0,3]$.

    e) Sea f la función definida como $f(x)=e^x \quad x\in[-1,2]$. Aplique el resultado de la parte b) y halle $P_h(x)$. Calcule $f(-0{,}5)$ y $P_h(-0{,}5)$ así como $f(0{,}5)$ y $\mathrm{P}_h(0{,}5)$.

11. En cada literal se define una función real w en dos variables. Calcular las derivadas parciales $\dfrac{\partial w}{\partial x}(a,b)$, $\dfrac{\partial w}{\partial y}(a,b)$ en el punto $(a,b)\in\mathbb{R}^2$ que se indica. Calcular valores aproximados de $\dfrac{\partial w}{\partial x}(a,b)$ y $\dfrac{\partial w}{\partial y}(a,b)$ mediante el uso de diferencias finitas centrales para cada $h\neq 0$, $k\neq 0$ que se dan.

    a) $w(x,y)=2x^2-xy-y^2 \quad (x,y)\in\mathbb{R}^2$; $a=1$, $b=1$; $h=0{,}02$ y $k=0{,}01$; $h=0{,}002$ y $k=0{,}001$; $h=0{,}0001$ y $k=0{,}0002$.

    b) $w(x,y)=x^3-10xy^2+\dfrac{1}{1+xy} \quad (x,y)\in\mathbb{R}^2$ con $y\neq -\dfrac{1}{x}$; $a=-1$, $b=-1$; $h=k=0{,}01$; $h=k=0{,}0002$; $h=0{,}00005$, $k=0{,}0002$.

    c) $w(x,y)=x\cos(y)+y\cos(x) \quad (x,y)\in\mathbb{R}^2$; $a=\dfrac{\pi}{6}$, $b=-\dfrac{\pi}{3}$; $h=0{,}003$ y $k=0{,}002$; $h=0{,}0003$ y $k=0{,}0002$; $h=0{,}00005$ y $k=0{,}00004$.

    d) $w(x,y)=\ln\left(1+x^2+y^2\right) \quad (x,y)\in\mathbb{R}^2$; $a=\sqrt{2}$, $b=\sqrt{3}$; $h=k=0{,}02$; $h=k=0{,}003$; $h=k=0{,}0004$; $h=k=0{,}00005$.

    e) $w(x,y)=\dfrac{1}{x^2+y^2} \quad (x,y)\in\mathbb{R}^2$ con $x\neq 0$, $y\neq 0$; $a=-1$, $b=1$; $h=k=0{,}003$; $h=k=0{,}0004$; $h=k=0{,}00002$.

    f) $w(x,y)=x\ln(1+y)+y\ln(1+x) \quad (x,y)\in\mathbb{R}^2$ tal que $x>-1$, $y>-1$; $a=b=2$; $h=\dfrac{1}{2}k=0{,}02$; $h=\dfrac{1}{3}k=0{,}001$; $h=\dfrac{1}{4}k=0{,}0001$.

12. Supóngase que f posee derivadas de todos los órdenes en un entorno del punto $x=a$. Se desea calcular valores aproximados $\widetilde{y}'''_a$ de $f'''(a)$. Escriba en forma explícita cada uno de los cocientes que se indican y determine el error de aproximación, donde $h\neq 0$ suficientemente pequeño.

    a) $\widetilde{y}'''_a=\dfrac{\delta\triangle^2 f(a)}{\alpha_1 h^3}$. **b)** $\widetilde{y}'''_a=\dfrac{\delta^3 f(a)}{\alpha_2 h^3}$. **c)** $\widetilde{y}'''_a=\dfrac{\nabla\delta\triangle f(a)}{\alpha_3 h^3}$. **d)** $\widetilde{y}'''_a=\dfrac{\nabla\triangle\delta f(a)}{\alpha_4 h^3}$.

e) $\widetilde{y}_a''' = \dfrac{\delta\triangle\delta f(a)}{\alpha_5 h^3}$, donde $\alpha_i \in \mathbb{R}$ con $\alpha_i \neq 0$ escogido en cada caso apropiadamente, $i = 1,2,3,4,5$. El polinomio de Taylor de grado 3 con error está definido como $f(a+k) = f(a) + kf'(a) + \dfrac{k^2}{2!}f''(a) + \dfrac{k^3}{3!}f'''(a) + \dfrac{k^4}{4!}f^{iv}(\xi)$ con ξ entre a y $a+k$ con $k \neq 0$.

13. Sea v la función definida como $v(x) = \dfrac{x^4}{4} - \dfrac{1}{3}x^3 + \dfrac{1}{2}x^2 \quad x \in [0,4]$. Calcule valores aproximados de $v'''(2)$ mediante los siguientes cocientes:

a) $\dfrac{\delta\triangle\delta v(2)}{\alpha_1 h^3}$. **b)** $\dfrac{\delta\triangle^2 v(2)}{\alpha_2 h^3}$. **c)** $\dfrac{\delta^3 v(2)}{\alpha_3 h^3}$. **d)** $\dfrac{\nabla\delta\nabla v(2)}{\alpha_4 h^3}$. **e)** $\dfrac{\nabla\triangle\delta v(2)}{\alpha_5 h^3}$. **f)** $\dfrac{\triangle\nabla^2 v(2)}{\alpha_6 h^3}$,

donde $h \neq 0$ y $\alpha_i \in \mathbb{R}$ con $\alpha_i \neq 0$ escogido apropiadamente $\quad i = 1, \ldots, 6$.

14. Considerar la función real u definida com $u(x,y) = \sqrt{x^2+y^2} \quad (x,y) \in \mathbb{R}^2$.

a) Hallar las derivadas parciales $\dfrac{\partial u}{\partial x}(x,y)$, $\dfrac{\partial u}{\partial y}(x,y)$, $\dfrac{\partial^2 u}{\partial x \partial y}(x,y)$, el laplaciano $\triangle u(x,y) = \dfrac{\partial^2 u}{\partial x^2}(x,y) + \dfrac{\partial^2 u}{\partial y^2}(x,y)$ con $(x,y) \in \mathbb{R}^2$ tal que $x \neq 0$, $y \neq 0$.

b) Calcular aproximaciones de $\dfrac{\partial u}{\partial x}(a,b)$, $\dfrac{\partial u}{\partial y}(a,b)$, $\dfrac{\partial^2 u}{\partial x \partial y}$ y $\triangle u(a,b)$ mediante diferencias finitas centrales en el punto $\left(\sqrt{2}, \sqrt{2}\right)$ así como en $(4,3)$, con $h \neq 0$, $k \neq 0$ pequeños que usted elige y compare los resultados con los valores exactos.

15. En cada item se define una función que posee derivadas parciales segundas en todo punto $(a,b) \in \mathbb{R}^2$. Calcule valores aproximados $\widetilde{\triangle} u(a,b)$ del laplaciano $\triangle u(a,b)$ en el punto (a,b) y $h \neq 0$, $k \neq 0$ que se indican. Calcule el error de aproximación, esto es, $\left|\triangle u(a,b) - \widetilde{\Delta} u(a,b)\right|$.

a) $f(x,y) = x^3y^4 - x^2y^2 + y^3 \quad (x,y) \in \mathbb{R}^2,\ a = -1,\ b = 1,\ h = k = 0{,}0015$ y $h = k = 0{,}00025$.

b) $f(x,y) = \operatorname{sen}(\pi x)\operatorname{sen}(\pi y) \quad (x,y) \in \mathbb{R}^2,\ a = b = \dfrac{1}{2},\ h = k = 0{,}001$ y $h = k = 0{,}00012$.

c) $f(x,y) = xe^{xy} + ye^x \quad (x,y) \in \mathbb{R}^2,\ a = 0,\ b = 1,\ h = k = 0{,}002$ y $h = k = 0{,}00011$.

d) $f(x,y) = \dfrac{1}{1+x^2+y^2} \quad (x,y) \in \mathbb{R}^2,\ a = 10,\ b = 20,\ h = 0{,}001,\ k = 0{,}002$; y, $h = 0{,}00025$ y $k = 0{,}00012$.

16. Sea v una función que posee derivadas parciales de todos los órdenes en un entorno de $(a,b) \in \mathbb{R}^2$. Se desea calcular valores aproximados de $\dfrac{\partial^2 v}{\partial x \partial y}(a,b)$ mediante cocientes de diferencias finitas que se indican a continuación, donde $h \neq 0$, $k \neq 0$ suficientemente pequeños, y, $\alpha_i \neq 0$ escogidos apropiadamente, $i = 1,2,3,4,5$.

a) $\dfrac{\nabla^2 v(a,b)}{\alpha_1 hk}$. **b)** $\dfrac{\alpha\nabla v(a,b)}{\alpha_2 hk}$. **c)** $\dfrac{\triangle\nabla v(a,b)}{\alpha_3 hk}$. **d)** $\dfrac{\delta^2 v(a,b)}{\alpha_4 hk}$. **e)** $\dfrac{\triangle\delta v(a,b)}{\alpha_5 hk}$.

Estime en cada caso el error de aproximación y analice los resultados.

17. En cada item se define una función real continua f en $[a,b]$. Calcular $I(f) = \int_a^b f(x)\,dx$. Calcular valores aproximados de $I(f)$ con la regla del rectángulo $R_n(f)$ con particiones uniformes $\tau(n)$ con $n = 4$ y luego con $n = 8$. Calcular $|I(f) - R_n(f)|$.

a) $f(x) = x^2 \quad x \in [0,4]$. **b)** $f(x) = \dfrac{2x}{x^2+1} \quad x \in [-1,2]$. **c)** $f(x) = \sqrt{x} \quad x \in [1,9]$.

d) $f(x) = xe^x \quad x \in [-2,2]$. **e)** $f(x) = x\ln(x) \quad x \in [1,e]$. **f)** $f(x) = \arctan(x) \quad x \in [0,1]$.

18. Con cada función f que se define en cada item, calcular $I(f) = \int_a^b f(x)\,dx$ y calcular valores aproximados de dicha integral con la regla de los trapecios $T_n(f)$ con particiones uniformes $\tau(n)$ con $n = 5$ y luego con $n = 10$. Calcular el error $|I(f) - T_n(f)|$.

a) $f(x) = x^2 - 1 \quad x \in [-1,3]$. **b)** $f(x) = \dfrac{1}{x+1} \quad x \in [0,2]$. **c)** $f(x) = \sqrt{2x+1} \quad x \in [0,4]$.

d) $f(x) = 2xe^{x^2} \quad x \in [0,2]$. **e)** $f(x) = x\ln(x) + x^2 \quad x \in [1,e]$. **f)** $f(x) = x\arctan(x) \quad x \in [0,1]$.

19. Con cada función f que se define en cada item, calcular $I(f) = \int_a^b f(x)\,dx$. Aplicar la regla de Simpson $S_n(f)$ para calcular aproximaciones de $I(f)$ con particiones uniformes $\tau(n)$ con $n = 4$, y $n = 8$. Calcule el error $|I(f) - S_n(f)|$.

a) $f(x) = x^2+1 \quad x \in [-1,1]$. **b)** $f(x) = x^3 \quad x \in [-1,1]$. **c)** $f(x) = 3x^3+2x^2-x+1 \quad x \in [0,1]$.

d) $f(x) = (x+1)^{\frac{1}{3}} \quad x \in [0,7]$. **e)** $f(x) = x\,\mathrm{sen}(x) \quad x \in [0,\pi]$. **f)** $f(x) = x\cos^2(x) \quad x \in [0,\pi]$.

20. Sea $f \in C^2([a,b])$, $n \in \mathbb{Z}^+$, $\tau(n) = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$, $h_i = x_i - x_{i-1}$ $i = 1,\ldots,n$. Se supone que $\tau(n)$ es regular, esto es, existe $\sigma \geq 1$ talq ue $h_i \leq \sigma\dfrac{b-a}{n} \quad i = 1,\ldots,n$.
Se define $I(f) = \int_a^b f(x)\,dx$ y $R_n(f) = \sum_{k=1}^{n} h_k f\left(\dfrac{x_{k-1}+h_k}{2}\right)$.

a) Demuestre que $R_n(f)$ es una forma lineal sobre $C([a,b])$ que se conoce como regla de los rectángulos generalizada.

b) Demuestre que existe $\lambda \in [a,b]$ tal que $|I(f) - R_n(f)| \leq \dfrac{h}{2}(b-a)|f'(\lambda)| \leq \dfrac{(b-a)}{2}Mh$, con $M = \underset{x\in[a,b]}{\text{máx}} |f'(x)|$.

21. El área del círculo $C = \left\{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 \leq 4\right\}$ es $a(C) = 4\pi$ (círculo de centro $(0,0)$ y radio $r = 2$). Se define $f(x) = \sqrt{4-x^2} \quad x \in [0,2]$, calcule $I(f) = 4\int_0^2 \sqrt{4-x^2}dx$ y verifique que $a(C) = I(f)$.

a) Aplique la regla del rectángulo generalizado para calcular aproximaciones de $a(C)$ con particiones $\tau(n)$ uniformes con $n = 5$ y $n = 10$. Calcule el error $|a(C) - R_n(f)|$.

b) Aplique la regla de los trapecios generalizada $T_n(f)$ para calcular aproximaciones de $a(C)$ con particiones $\tau(n)$ uniformes con $n = 5$ y $n = 10$. Calcule $|a(C) - T_n(f)|$.

c) Aplique la regla de Simpson generalizada $S_n(f)$ con particiones $\tau(n)$ uniformes con $n = 4$ y $n = 8$. Calcule $|S_n(f) - I(f)|$.

Compare los resultados de a), b) y c).

22. Sea f la función real definida como $f(x) = \dfrac{1}{x^2} \quad x \in \left[\dfrac{1}{4},1\right]$ e $I(f) = \int_{\frac{1}{4}}^1 f(x)\,dx$. Calcule $I(f)$.
Aplique la regla del rectángulo $R_n(f)$, de los trapecios $T_n(f)$, de Simpson $S_n(f)$ generalizadas para calcular aproximaciones de $I(f)$ con particiones $\tau(n)$ regulares que en cada item se indican. Calcule el error con cada método y cada partición.

a) $\tau_1(5) = \{x_0 = 0{,}25,\ x_1 = 0{,}3,\ x_2 = 0{,}4,\ x_3 = 0{,}6,\ x_4 = 0{,}8,\ x_5 = 1\}$.

b) $\tau_2(5) = \{x_0 = 0{,}25,\ x_1 = 0{,}4,\ x_2 = 0{,}55,\ x_3 = 0{,}7,\ x_4 = 0{,}85,\ x_5 = 1\}$.

c) $\tau_1(10) = \{x_0 = 0{,}25,\ x_2 = 0{,}28,\ x_3 = 0{,}32,\ x_4 = 0{,}36,\ x_5 = 0{,}4,\ x_6 = 0{,}5,\ x_7 = 0{,}6,\ x_8 = 0{,}7,$
$x_9 = 0{,}85,\ x_{10} = 1\}$.

d) $\tau_2(10) = \{x_j = 0{,}25 + 0{,}075j \mid j = 0,1,\ldots,10\}$.

e) Compare los resultados obtenidos con la regla del rectángulo generalizada en a) y b), luego en c) y d); concluya. Proceda en forma similar con la regla de los trapecios generalizada en a) y b) luego en c) y d). Concluya

f) Compare los resultados obtenidos con la regla de Simpson generalizada en a) y b); luego en c) y d). Compare estos con los anteriores y concluya.

23. Considere la función u definida como $u(x) = \exp(-10x^2) \quad x \in [-2,2]$.

a) Trace la gráfica de la función u.

b) Aplique la regla de los trapecios generalizada para calcular valores aproximados de $I(u) = \int_{-2}^2 u(x)\,dx$ con cada una de las particiones $\tau_1(6)$, $\tau_2(8)$, $\tau_3(10)$ siguientes:

$\tau_1(6) = \{x_0 = -2,\ x_1 = -1,\ x_2 = -0{,}5,\ x_3 = 0,\ x_4 = 0{,}5,\ x_5 = 1,\ x_6 = 2\}$,

$\tau_2(8) = \left\{ \begin{array}{c} x_0 = -2,\ x_1 = -1,\ x_2 = -0{,}5,\ x_3 = -0{,}25,\ x_4 = 0, \\ x_5 = 0{,}25,\ x_6 = 0{,}5,\ x_7 = 1,\ x_8 = 10 \end{array} \right\}$,

$$\tau_3(10) = \left\{ \begin{array}{c} x_0 = -2,\ x_1 = -1,\ x_2 = -0{,}6,\ x_3 = -0{,}3,\ x_4 = -0{,}1,\ x_5 = 0, \\ x_6 = 0{,}1,\ x_7 = 0{,}3,\ x_8 = 0{,}6,\ x_9 = 1,\ x_{10} = 2 \end{array} \right\}.$$

c) Calcule valores aproximados de $I(u) = \int_{-2}^{2} u(x)\,dx$ con particiones uniformes $\tau(n)$ y $n = 6,\ 8,\ 10$. Compare con los resultados obtenidos en la parte b) precedente.

24. Sean $f, g \in C([a,b])$. Supongamos que $f(x) \leq g(x) \quad \forall x \in [a,b]$,

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \mid f(x) \leq y \leq g(x) \quad x \in [a,b]\right\}.$$

El área de la región Ω está definida como $a(\Omega) = \int_a^b [g(x) - f(x)]\,dx$. En cada item se dan las funciones continuas $f,\ g$ en $[a,b]$. Represente gráficamente la región Ω, calcule $a(\Omega)$ y calcule aproximaciones de $a(\Omega)$ con la regla de Simpson generalizada con una partición uniforme $\tau(n)$ con $n = 5$. Compare los resultados obtenidos.

a) $f(x) = 0,\ g(x) = x \quad x \in [0,4]$. **b)** $f(x) = -1,\ g(x) = x^2 \quad x \in [0,4]$.

c) $f(x) = -x^2 + 1,\ g(x) = x^3 + 1 \quad x \in [0,3]$. **d)** $f(x) = x^2 - 4,\ g(x) = x^2 + x + 1 \quad x \in [0,4]$.

e) $f(x) = x - \dfrac{\pi}{2},\ g(x) = \cos(x) \quad x \in \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$. **f)** $f(x) = x^2,\ g(x) = e^x \quad x \in [0,2]$.

25. En cada literal se define una función f que es impar en el intervalo $[-a,a]$ con $a > 0$ que se indica. Demuestre que $I(f) = 0$ y aplique la regla del rectángulo, trapecios y Simpson generalizadas para calcular aproximaciones de $I(f)$ con particiones uniformes $\tau(n)$ con $n = 5$ y $n = 6$. Analice los resultados.

a) $f(x) = x^3 \quad x \in [-2,2]$. **b)** $f(x) = \operatorname{sen}(x) \quad x \in \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$. **c)** $f(x) = x\sqrt{1-x^2} \quad x \in [-1,1]$.

d) $f(x) = x\cos^2(\pi x) \quad x \in [-1,1]$, **e)** $f(x) = 2\operatorname{sen}^3(\pi x)\cos^2(\pi x) \quad x \in [-1,1]$.

26. Sean $f \in C([a,b])$, $\tau(n)$ una partición regular del intervalo $[a,b]$. Elabore un algoritmo para aproximar $I(f) = \int_a^b f(x)\,dx$ con la regla del rectángulo generalizada.

27. Sea $\Omega \subset \mathbb{R}^2$ una región del tipo II, $f \in C(\Omega)$. Elabore un algoritmo para calcular aproximaciones de $I(f) = \iint\limits_{\Omega} f(x,y)\,dxdy$ con la regla de los trapecios generalizada.

28. **a)** Sea $\Omega \subset \mathbb{R}^2$ una región de tipo I, $f \in C(\Omega)$. Elabore un algoritmo para calcular aproximaciones de $I(f) = \iint\limits_{\Omega} f(x,y)\,dxdy$ con la regla de Simpson generalizada.

b) Suponga ahora $\Omega \subset \mathbb{R}^2$ región de tipo II, $f \in C(\Omega)$. Elabore un algoritmo para calcular aproximaciones de $I(f) = \iint\limits_{\Omega} f(x,y)\,dxdy$ con la regla de Simpson generalizada.

29. En cada item se define una función continua f sobre $\Omega = [a,b] \times [c,d]$ que se indica. Calcule $I(f) = \int_a^b \left(\int_c^d f(x,y)\,dy\right) dx$. Aplique la regla de los trapecios generalizada para calcular aproximaciones de $I(f)$ con $n = m = 5$. Estime el error

a) $f(x,y) = xy \quad (x,y) \in [0,2]$. **b)** $f(x,y) = x^3y + xy^4 \quad (x,y) \in [0,1] \times [0,1]$.

c) $f(x,y) = 10\sqrt{xy} \quad (x,y) \in [0,1] \times [0,1]$. **d)** $f(x,y) = \sqrt{x} + \sqrt{y} \quad (x,y) \in [0,4] \times [0,4]$.

e) $f(x,y) = \dfrac{y}{1+\sqrt{x}} \quad (x,y) \in [0,4] \times [0,2]$. **f)** $f(x,y) = \dfrac{4}{1+xy} \quad (x,y) \in [1,4] \times [1,4]$.

30. Calcular $I(f) = \int_a^b \left(\int_c^d f(x,y)\,dy\right) dx$ para cada función $f \in C(\Omega)$ que se define sobre $\Omega = [a,b] \times [c,d]$. Aplique la regla de Simpson generalizada para calcular aproximaciones de $I(f)$ con $m = n = 5$. Estime el error.

a) $f(x,y) = x^2 + xy \quad (x,y) \in [-1,1] \times [0,1]$. **b)** $f(x,y) = xe^y + ye^x \quad (x,y) \in [0,1] \times [0,1]$.

c) $f(x,y) = \operatorname{sen}(x+y) \quad (x,y) \in \left[0, \dfrac{\pi}{2}\right]$. **d)** $f(x,y) = (x+y)^{\frac{1}{3}} \quad (x,y) \in [0,4]$.

e) $f(x,y) = 2\sqrt{x} - 3\sqrt{y} \quad (x,y) \in [0,1] \times [1,4]$. **f)** $f(x,y) = \dfrac{\sqrt{x}}{\sqrt{y}} \quad (x,y) \in [0,4] \times [1,4]$.

31. En cada item se define una función u sobre $\Omega = [a,b]\times[c,d]$. Calcule $I(f) = \int_c^d \left(\int_a^b u(x,y)\,dx\right) dy$. Calcule aproximaciones de $I(u)$ con $m,\ n$ que se indican; y, estime el error $|I(u).T_{mn}(u)|$.

a) $u(x,y) = x^2+y^2 \quad (x,y)\in[0,1]\times[0,1],\ m=n=5.$

b) $u(x,y) = \sqrt{x+y} \quad (x,y)\in[0,2]\times[1,4],\ m=n=6.$

c) $u(x,y) = ye^{xy} \quad (x,y)\in[0,1]\times[-1,0],\ m=n=5.$

d) $u(x,y) = x^4 \quad (x,y)\in\left[\frac{1}{2},2\right]\times[-1,1],\ m=n=8.$

De modo análogo, calcule aproximaciones de $I(u)$ usando la regla de Simpson generalizada $S_{mn}(u)$ con $m=n=4$; y, estime el error $|I(u)-S_{mn}(u)|$.

32. Sean $V = C([a,b])$ el espacio vectorial real de funciones continuas en $[a,b]$, $n\in\mathbb{Z}^+$, y Δ el funcional definido sobre $C([a,b])$ que en cada item se define. Pruebe que Δ es lineal.

a) $\Delta(f) = \frac{h}{2}(f(a)+f(b)) + h\sum_{j=1}^{n-1} f(x_j)$ (fórmula de los trapecios generalizada), donde $h=\frac{b-a}{n}$ y $x_j = a+jh \quad j=0,1,\ldots,n.$

b) $\Delta(f) = \frac{h}{3}(f(a)+f(b)) + \frac{2}{3}h\sum_{j=1}^{n-1} f(x_{2j}) + \frac{4}{3}h\sum_{j=1}^{n} f(x_{2j-1})$ (fórmula de Simpson generalizada), donde $h=\frac{b-a}{2n}$, $x_j = a+jh \quad j=0,1,\ldots,2n$

33. Aplique la regla de los trapecios generalizada para aproximar las siguientes integrales con una discretización de 10 puntos igualmente espaciados:

a) $\int_0^1 \sqrt{x}dx$. **b)** $\int_0^1 x^{1/4}dx$. **c)** $\int_0^2 xe^{-x}dx$. **d)** $\int_0^{0,5} \text{sen}(x^2)dx$. **e)** $\int_1^2 \frac{\ln x}{x}dx$. **f)** $\int_1^2 \frac{e^x}{x}dx$.

Para los literales a), b) y c) halle el valor exacto de la integral y compare con el valor aproximado.

34. En cada item se define una función f sobre una región Ω. Calclular $I(f) = \iint\limits_\Omega f(x,y)\,dxdy$. Aplicar la regla de los trapecios generalizada $T_{mn}(f)$ para calcular una aproximación de $I(f)$ con $m=n=4$.

a) $f(x,y) = x,y \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid 1-x\le y\le\sqrt{1-x^2},\ \ 0\le x\le 1\right\}.$

b) $f(x,y) = (x-y)^2 \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid x^2\le y\le 1-x^2,\ \ x\in[-1,1]\right\}.$

c) $f(x,y) = \text{sen}(x+y) \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid -\frac{\pi}{`4}\le y\le\frac{\pi}{2},\ \ x\in[0,2]\right\}$

d) $f(x,y) = \frac{x}{y} \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid \frac{1}{x}\le y\le 1+x,\ \ x\in[1,2]\right\}.$

35. En cada item se define una función w sobre una región Ω. Calcular $I(w) = \iint\limits_\Omega w(x,y)\,dxdy$. Aplicar la regla de Simpson generaliza $S_{mn}(w)$ para calcular una aproximación de $I(w)$ con $m=n=4$.

a) $w(x,y) = \frac{1}{6}(x+y)^2 \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid 1+y\le x\le 1+y^2,\ \ y\in[1,2]\right\}.$

b) $w(x,y) = \frac{y}{x+4} \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid y-1\le x\le 1-y^2,\ \ y\in[0,1]\right\}.$

c) $w(x,y) = \frac{4}{1+x^2+y^2} \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid x^2+y^2\le 1,\right\}$, escoja apropiadamente $\Omega_1\subset\Omega$ región del tipo I y proceda con el cálculo. Asimismo, escoja otra región de $\Omega_2\subset\Omega$ del tipo II y proceda con el cálculo. Compare los resultados.

d) $w(x,y) = \cos^2(x-y) \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid x-\frac{\pi}{2}\le y\le x,\ \ y\in\left[0,\frac{\pi}{2}\right]\right\}.$

e) $w(x,y) = x^4+y^4 \quad (x,y)\in\Omega = \left\{(x,y)\in\mathbb{R}^2 \mid 1\le x^2+y^2\le 4,\right\}$. Escoja apropiadamente $\Omega_1\subset\Omega$ del tipo I y calcule $S_{mn}(w)$. De manera similar, escoja $\Omega_2\subset\Omega$ del tipo II y calcule $S_{mn}(w)$. Compare los resultados.

36. Considerar la integral impropia $I = \int_0^\infty x^2 e^{-x^2} dx$, y sea $I_1 = \int_0^1 x^2 e^{-x^2} dx$.

    a) Demuestre que $I = \frac{\sqrt{\pi}}{4} \simeq 0{,}4431134628$.

    b) Muestre que

    $$I_2 = \int_1^\infty x^2 e^{-x^2} dx = \frac{1}{2} \int_0^1 \frac{e^{-\frac{1}{t}}}{t^{5/2}} dt.$$

    c) Aplique la fórmula de los trapecios con $n = 4$ para aproximar I_1, I_2 consecuentemente I (tome en cuenta la singularidad en I_2). Compare el resultado con a).

    d) Repita la parte b) con $n = 10$. Compare el resultado con a) y c).

37. Considerar la integral doble $I = \int_0^1 \left(\int_{x^3}^x (x^2 + y) dy \right) dx$.

    a) Muestre que $I = \dfrac{5}{28}$.

    b) Sean $g(x) = \int_{x^3}^x (x^2 + y) dy$, $x \in [0, 1]$; $h = 0{,}2$ y $x_k = kh$, $k = 0, 1, \ldots, 5$. Calcule $g(x_k)$ y aproxime $g(x_k)$ usando la regla de los trapecios con $m = 5$.

    c) Aplique la regla de los trapecios para aproximar $I = \int_0^1 g(x) dx$ y compare con a).

38. Sea $I = \int_1^2 \int_{x^2}^{x^3} (x^2 + y^2) dx dy$.

    a) Calcule I.

    b) Aplique la regla de los trapecios para aproximar I con $m = n = 5$.

39. Sea $I = \int_{-1}^1 \int_{-1}^1 e^{x^2+y^2} dx dy$.

    a) Pruebe que $I = 4 \left(\int_0^1 e^{x^2} dx \right)^2$.

    b) Utilice la fórmula de los trapecios para aproximar el valor de I con $n = 10$, y luego con $n = 20$.

    c) Utilice la serie de Taylor de e^α y aproxime I mediante una suma finita S_n de modo que

    $$|I - S_n| < 10^{-5},$$

    donde n es el más pequeño número entero positivo que satisface dicha condición. De los resultados de b) y c) ¿qué algoritmo es más costoso numéricamente?.

40. Considerar la integral $I = \int_0^2 \left(\int_0^{2-y} \sqrt{1 + x} dx \right) dy$.

    a) Calcular I.

    b) Sean $g(y) = \int_0^{2-y} \sqrt{1 + x} dx$, $y \in [0, 2]$. Aplique la regla de los trapecios para aproximar I con particiones de 6 puntos igualmente espaciados.

41. Considerar la integral $I = \int_0^4 \left(\int_{-\sqrt{4-y}}^{\frac{y-4}{2}} (xy)^3 \, dx \right) dy$.

    a) Calcule I. **b)** Aproxime I con particiones de 5 puntos igualmente espaciados.

2.10. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Análisis Matemático, Segunda Edición, Editorial Reverté, Barcelona, 1982.
2. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.
3. Tom M. Apostol, Calculus, Volumen 2, Segunda Edición, Editorial Reverté, Barcelona, 1975.
4. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
5. R. M. Barbolla, M. García, J. Margalef, E. Outerelo, J. L. Pinilla. J. M. Sánchez, Introducción al Análisis Real, Editorial Alambra Universidad, Madrid, 1981.

6. Richard H. Bartels, John C. Beatty, Brian A. Barsky, An Introduction to Splines for use in Computer Graphics and Geometric Medeling, Editorial Morgan Kaufmann Publishers, Inc., San Mateo, California, 1987.

7. Jérõme Bastien, Jean-Noël Martin, Introduction à l'Analyse Numérique, Editorial Dunod, París, 2003.

8. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.

9. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.

10. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.

11. Elaine Cohen, Richard F. Riesenfeld, Gershon Elber, Geometric Modeling with Splines, Editorial A. K. Peters, Natick, Massachusetts, 2001.

12. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.

13. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.

14. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.

15. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.

16. Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence, Algebra Lineal, Editorial Publicaciones Cultural, S. A., México, 1982.

17. Waltson Fulks, Cálculo Avanzado, Editorial Limusa, México, 1973.

18. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

19. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

20. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.

21. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.

22. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

23. Rodolfo Luthe, Antonio Olivera, Fernando Schutz, Métodos Numéricos, Editorial Limusa, México, 1986.

24. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

25. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

26. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

27. S. Nikolski, Fórmulas de Cuadratura, Editorial Mir, Moscú, 1990.

28. Ben Noble, James W. Daniel, Algebra Lineal Aplicada, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1989.

29. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

30. A. A. Samarski, Introducción a los Métodos Numéricos, Editorial Mir, Moscú, 1986.

31. Michelle Schatzman, Analyse Numérique, Inter Editions, París, 1991.

32. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

33. Michael Spivak, Calculus, Segunda Edición, Editorial Reverté, Barcelona, 1996.

34. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

35. Gilbert Strang, Algebra Lineal y sus Aplicaciones, Editorial Fondo Educativo Interamericano, México, 1982.

36. E. A. Volkov, Métodos Numéricos, Editorial Mir, Moscú, 1990.

Capítulo 3

Aproximación de series de funciones. Aplicaciones.

Resumen

Muchos problemas en matemáticas conducen a soluciones expresadas mediante series convergentes de funciones, particularmente interesan las series de potencias, que suponemos convergen uniformemente en un cierto intervalo cerrado $[a, b]$ de $\mathbb{R}$. Las series de Fourier se tratarán en el capítulo de mínimos cuadrados. En muy pocos casos se conocen resultados exactos y en la generalidad de los mismos se conocen resultados de convergencia puntual y uniforme. Tanto en el caso de conocer la función suma como en el que se desconoce, interesa calcular valores aproximados de dichas sumas, las mismas que deben ser aproximadas numéricamente.

Este capítulo se inicia con la aproximación de series numéricas. A continuación se tratan las series de potencias a las que se dan mayor atención. Particular interés se da al cálculo aproximado de algunas funciones usuales representadas como series de potencias como son $\text{sen}(x)$, $\cos(x)$, $\text{arcsen}(x)$, $\ln(x)$, $\exp(x)$. Se presentan algunas aplicaciones de las series de potencias como en el caso de la función error, las integrales elípticas. Se pone mucho énfasis en la aplicación de resultados de la consistencia y la estabilidad numérica que nos permitan elaborar algoritmos simples de cálculo.

3.1. Resultados fundamentales de series numéricas convergentes.

Esta sección está destinada a introducir algunos conceptos básicos sobre las series numéricas reales así como presentar algunos resultados importantes sobre los criterios de convergencia. Estos resultados serán de gran utilidad en el cálculo aproximado de series numéricas y series de funciones, y particularmente en las series de potencias y las series de Fourier. El lector que está familiarizado con las series numéricas puede pasar inmediatamente a los métodos de cálculo, aquel que no está familiarizado tendrá la ocasión de tratar este tema en forma resumida. Al final del capítulo se dan algunas observaciones, comentarios y se sugiere una bibliografía especializada para estudios más profundos.

3.1.1. Series numéricas convergentes.

Sea (a_n) una sucesión numérica. A menos que se indique lo contrario, suponemos que las sucesiones numéricas (a_n) están definidas en todo $n \in \mathbb{N}$. La suma $a_0 + a_1 + a_2 + ... + a_n + \cdots$, que se escribe $\sum_{n=0}^{\infty} a_n$ y que se lee suma desde $n = 0$ hasta infinito de a_n, se llama serie numérica. En la serie $\sum_{n=0}^{\infty} a_n$, a_n se llama término general. En el caso de que la sucesión numérica (a_n) está definida para todo $n \in \mathbb{Z}^+$ con $n \geq n_0 \geq 1$, la suma $a_{n_0} + a_{n_0+1} + a_{n_0+2} + \cdots$, se escribirá $\sum_{n=n_0}^{\infty} a_n$.

Sea $n \in \mathbb{N}$. Se define $S_n = \sum_{k=0}^{n} a_k$ y se denomina suma parcial de la serie $\sum_{n=0}^{\infty} a_n$. También se escribirá a la suma parcial $S_n = \sum_{k=0}^{n} a_k$. La sucesión (S_n) se llama sucesión de sumas parciales de la serie $\sum_{n=0}^{\infty} a_n$.

Definición 1 *i) Se dice que la serie $\sum_{n=0}^{\infty} a_n$ es convergente si y solo si la sucesión de sumas parciales (S_n) es convergente, es decir que existe $S \in \mathbb{R}$ tal que $\lim_{n\to\infty} S_n = S$. Escribimos $\sum_{n=0}^{\infty} a_n = S$ y diremos que S es la suma de la serie.*
ii) Diremos que $\sum_{n=0}^{\infty} a_n$ es divergente si y solo si la sucesión de sumas parciales (S_n) es divergente.

Se verifica inmediatamente que si $\sum_{n=0}^{\infty} a_n$ converge, entonces $\lim_{n\to\infty} a_n = 0$. El recíproco, en general, no es cierto como se muestra más adelante con el ejemplo de la serie armónica.

Ejemplos

1. Sea $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$. Observemos primeramente que $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1} \quad \forall n \in \mathbb{Z}^+$. En consecuencia; si $n \in \mathbb{Z}^+$ y $S_n = \sum_{k=1}^{n} \frac{1}{k(k+1)}$ denota la suma parcial, se tiene

$$S_1 = 1 - \frac{1}{2}, \qquad S_2 = \frac{1}{1} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = 1 - \frac{1}{3}, \qquad , S_n = 1 - \frac{1}{n+1} \quad n \in \mathbb{Z}^+.$$

   Resulta $\lim_{n\to\infty} S_n = \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) = 1$, con lo cual la serie $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ es convergente y converge a 1, esto es, $\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$. Note que $\lim_{n\to\infty} \frac{1}{n(n+1)} = 0$, o sea el término general de la serie es convergente.

2. Sea $x \in \mathbb{R}$, $x \neq 0$. Consideramos la serie $\sum_{n=0}^{\infty} x^n$. Esta serie se llama serie geométrica. Para $n \in \mathbb{N}$ se define la suma parcial $S_n = \sum_{k=0}^{n} x^k = 1 + x + ... + x^n$. Multipliquemos a S_n por x. Tenemos

$$xS_n = \sum_{k=0}^{n} x^{k+1} = x + x^2 + ... + x^{n+1},$$

   luego

$$(x-1)\, S_n = xS_n - S_n = x + x^2 + ... + x^{n+1} - (1 + x + ... + x^n) = x^{n+1} - 1.$$

   de donde

$$\begin{aligned} S_n &= \frac{x^{n+1} - 1}{x - 1} = \frac{1}{1-x} - \frac{x^{n+1}}{1-x} \quad \text{si} \quad x \neq 1, \\ S_n &= n + 1 \quad \text{si} \quad x = 1. \end{aligned}$$

   Puesto que $\lim_{n\to\infty} x^{n+1} = 0$ si y solo si $|x| < 1$, se sigue que

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \left(\frac{1}{1-x} - \frac{x^{n+1}}{1-x}\right) = \frac{1}{1-x},$$

   y en consecuencia $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$ si $|x| < 1, \quad x \neq 0$. Si $|x| \geq 1$, la serie $\sum_{n=0}^{\infty} x^n$ es divergente, pués la sucesión (S_n) es divergente y $\lim_{n\to\infty} S_n$ no existe.

3. La serie $\sum_{n=1}^{\infty} \frac{1}{n}$ es divergente. Esta serie se llama serie armónica. Para cada $n \in \mathbb{Z}^+$ se define la suma parcial $S_n = \sum_{k=1}^{n} \frac{1}{k}$. Observe las siguientes sumas parciales:

$$\begin{aligned}
S_2 &= 1 + \frac{1}{2}, \\
S_{2^2} &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} > 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1 + \frac{2}{2}, \\
S_{2^3} &= 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{8} > 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 1 + \frac{3}{2}, \\
S_{2^4} &= 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{16} > 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} > 1 + \frac{4}{2}, \\
&\vdots \\
S_{2^n} &= 1 + \frac{1}{2} + ... + \frac{1}{2^n} > 1 + \frac{n}{2}.
\end{aligned}$$

Luego, $\lim_{n\to\infty} S_{2^n} \geq \lim_{n\to\infty} \left(1 + \frac{n}{2}\right) = \infty$, que muestra que la serie armónica es divergente. Note que $\lim_{n\to\infty} \frac{1}{n} = 0$, o sea el término general de la serie es convergente pero la serie $\sum_{k=1}^{\infty} \frac{1}{k}$ diverge.

Definición 2 *Sea* $\sum_{n=0}^{\infty} a_n$ *una serie numérica.*

i. Se dice que $\sum_{n=0}^{\infty} a_n$ *converge absolutamente si la serie* $\sum_{n=0}^{\infty} |a_n|$ *converge.*

ii. Se dice que $\sum_{n=0}^{\infty} a_n$ *converge condicionalmente si* $\sum_{n=0}^{\infty} a_n$ *converge pero* $\sum_{n=0}^{\infty} |a_n|$ *diverge.*

1. De la definición de convergencia absoluta se deduce que la serie $\sum_{n=0}^{\infty} |a_n|$ converge, entonces $\sum_{n=0}^{\infty} a_n$ converge. En efecto, sean $S_n = \sum_{k=0}^{n} a_k$ y $\widetilde{S}_n = \sum_{k=1}^{n} |a_k|$. Como

$$\widetilde{S}_n = \sum_{k=0}^{n} |a_k| \leq \sum_{k=0}^{n} |a_k| + |a_{n+1}| = \widetilde{S}_{n+1},$$

la sucesión $\left(\widetilde{S}_n\right)$ es creciente. Además, por hipótesis la serie $\sum_{k=0}^{\infty} |a_k|$ es convergente, entonces $\lim_{n\to\infty} \widetilde{S}_n$ existe y sea $\widetilde{S} = \sum_{k=0}^{\infty} |a_k|$. Así, $\left(\widetilde{S}_n\right)$ es creciente y acotada superiormente, y $\widetilde{S} = \lim_{n\to\infty} S_n = \operatorname{Sup}_{n\in\mathbb{Z}^+} \widetilde{S}_n$. Puesto que

$$|S_n| = \left|\sum_{k=0}^{n} a_k\right| \leq \sum_{k=0}^{n} |a_k| = \widetilde{S}_n \leq \widetilde{S} \quad \forall n \in \mathbb{Z}^+,$$

luego $\operatorname{Sup}_{n\in\mathbb{Z}^+} |S_n| \leq \widetilde{S}$, o sea $\left|\sum_{k=0}^{\infty} a_k\right| \leq \widetilde{S}$, es decir que $\sum_{k=0}^{\infty} a_k$ converge.

Ejemplos

1. La serie $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ converge, pero $\sum_{n=1}^{\infty} \frac{1}{n}$ diverge, entonces $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ converge condicionalmente. Más adelante, en las series de potencias se demuestra que $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \ln 2$.

2. La serie $\sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$ converge absolutamente. En efecto, la serie $\sum_{k=0}^{\infty} \frac{1}{k!}$ converge al número $e \simeq 2{,}71828182...$ base de los logaritmos naturales. Luego $\sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$ converge absolutamente pués para cada $n \in \mathbb{Z}^+$, se tiene $|\sum_{k=0}^{n} \frac{(-1)^k}{k!}| \leq . \sum_{k=0}^{n} \frac{1}{k!}$. Además, $\sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = e^{-1}$.

De la definición de serie convergente, se sigue que si $\sum_{n=0}^{\infty} a_n$ es convergente y (S_n) denota la sucesión de sumas parciales de dicha serie, $\lim_{n\to\infty} S_n = S$ si y solo si se verifica la condición

$$\forall \varepsilon > 0, \quad \exists n_0 \in \mathbb{Z}^+ \quad \text{tal que } \forall n \geq n_0 \Rightarrow |S_n - S| < \varepsilon.$$

Tomando en consideración que $S_n = \sum_{k=0}^{n} a_k$, $n = 0, 1, ...$, y $S = \sum_{k=0}^{\infty} a_k$, entonces

$$S - S_n = \sum_{k=0}^{\infty} a_k - \sum_{k=0}^{n} a_k = \sum_{k=n+1}^{\infty} a_k.$$

Luego, $\sum_{k=0}^{\infty} a_k$ es convergente si y solo si se verifica la siguiente condición:

$$\forall \varepsilon > 0, \ \exists n_0 \in \mathbb{Z}^+ \quad \text{tal que } \forall n \geq n_0 \Rightarrow \left| \sum_{k=n+1}^{\infty} a_k \right| < \varepsilon.$$

Se denota con $\mathcal{S}_c$ al conjunto de todas las series convergentes. Sean $\sum_{n=0}^{\infty} a_n$, $\sum_{n=0}^{\infty} b_n \in \mathcal{S}_c$, y $\lambda \in \mathbb{R}$. Se define la adición de series convergentes y el producto de escalares por series convergentes como sigue:

Adición**:** $\sum_{n=0}^{\infty} a_n + \sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} (a_n + b_n)$,

Producto por escalares**:** $\lambda \sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} \lambda a_n$.

Se demuestra fácilmente las dos implicaciones siguientes:

$$\sum_{n=0}^{\infty} a_n, \quad \sum_{n=0}^{\infty} b_n \in \mathcal{S}_c \Rightarrow \sum_{n=0}^{\infty} (a_n + b_n) \in \mathcal{S}_c,$$

y

$$\lambda \in \mathbb{R}, \quad \sum_{n=0}^{\infty} a_n \in S_c \Rightarrow \sum_{n=0}^{\infty} \lambda a_n \in S_c;$$

es decir que el conjunto $\mathcal{S}_c$ con las operaciones de adición y producto por escalares de series convergentes, es un espacio vectorial denominado espacio de series convergentes.

3.1.2. Criterios de convergencia.

Dada una serie numérica $\sum_{n=0}^{\infty} a_n$, se debe determinar si esta es o no convergente. Para el efecto, es preciso familiarizarse con algumos resultados fundamentales que nos permitan decidir si la serie es convergente, divergente o simplemente con un determinado criterio no es posible decidir la convergencia o divergencia y que se requiere de un análisis más fino para deducir la convergencia o divergencia de una serie dada. En esta parte enunciamos sin demostración algunos criterios de convergencia más utilizados y se da un ejemplo en el que se aplique el teorema. Al final del capítulo se cita una amplia bibliografía en la que puede encontrarse las demostraciones de los resultados que damos a continuación (Calculus de Apostol, Volumen 1, Cálculo Avanzado de Fulks, Calculus de Spivak, y otros).

Teorema 1 *(criterio de Leibniz)*
Sea (a_n) *una sucesión numérica decreciente tal que* $\lim_{n\to\infty} a_n = 0$. *Entonces, la serie* $\sum_{n=0}^{\infty} (-1)^n a_n$ *converge.*

Ejemplo

La serie $\sum_{n=0}^{\infty} \frac{(-1)^n}{n!3^n}$ es convergente, pués la sucesión (a_n) cuyo término general está definido como $a_n = \frac{1}{n!3^n}$ es decreciente y se tiene $\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{1}{n!3^n} = 0$. Por el criterio de Leibniz, $\sum_{n=0}^{\infty} \frac{(-1)^n}{n!3^n}$ converge. Se muestra inmediatamente que $\sum_{n=0}^{\infty} \frac{(-1)^n}{n!3^n}$ converge absolutamente y $e^{-\frac{1}{3}} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!3^n}$.

Teorema 2 *(criterio de Cauchy)*
Sea $\sum_{n=0}^{\infty} a_n$ *una serie numérica. Entonces,* $\sum_{n=0}^{\infty} a_n$ *converge si y solo si* $\forall \varepsilon > 0$, $\exists n_0 \in \mathbb{Z}^+$ *tal que* $\forall m, n \in \mathbb{Z}^+$ *con* $m > n > n_0 \Rightarrow |a_{n+1} + a_{n+2} + ... + a_m| < \varepsilon$.

Ejemplo

La serie $\sum_{n=0}^{\infty} \frac{1}{k!}$ es convergente. En efecto, de la definición del factorial de k, esto es $k! = 1 \times 2 \times ... \times k$, se tiene

$$\frac{1}{k!} = \frac{1}{1 \times 2 \times ... \times k} = \frac{1}{1} \times \frac{1}{2} \times \frac{1}{3} \times ... \times \frac{1}{k} < 1 \times \frac{1}{2} \times \frac{1}{2} \times ... \times \frac{1}{2} = \frac{1}{2^{k-1}},$$

de donde para m, $n \in \mathbb{Z}^+$ con $m < n$, obtenemos

$$\begin{aligned} \left| \sum_{k=n+1}^{m+n} \frac{1}{k!} \right| &= \sum_{k=n+1}^{m+n} \frac{1}{k!} < \sum_{k=n+1}^{m+n} \frac{1}{2^{k-1}} = \sum_{k=0}^{m-1} \frac{1}{2^{(k+n+1)-1}} = \frac{1}{2^n} \sum_{k=0}^{m-1} \frac{1}{2^k} = \frac{1}{2^n} \left(2 - 2 \times 2^{-m}\right) \\ &= \frac{1}{2^{n-1}} \left(1 - 2^m\right) < \frac{1}{2^{n-1}} \underset{n\longrightarrow 0}{\longrightarrow} 0. \end{aligned}$$

Así, para $\varepsilon > 0$, $\exists n_0 \in \mathbb{Z}^+$ tal que $\forall n \geq n_0 \Rightarrow \frac{1}{2^{n-1}} < \varepsilon$, y en consecuencia $\left| \sum_{k=n+1}^{m+n} \frac{1}{k!} \right| < \varepsilon$ si $n > m > n_0$, y por el criterio de Cauchy, $\sum_{k=0}^{\infty} \frac{1}{k!}$ converge.

Nota**:** Más adelante se tratará la serie de potencias $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, $x \in \mathbb{R}$. Para $x = 1$, se tiene $e = \sum_{k=0}^{\infty} \frac{1}{k!}$; para $x = -\frac{1}{3}$, se tiene $e^{-\frac{1}{3}} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!3^k}$ que se indicó arriba.

Teorema 3 *(comparación por paso al límite)*
Sean $\sum_{n=0}^{\infty} a_n$, $\sum_{n=0}^{\infty} b_n$ *dos series de términos positivos.*

i) Si $0 < \lim_{n\to\infty} \frac{a_n}{b_n} < \infty$, *entonces ambas series convergen o ambas series divergen.*

ii) Si $\sum_{n=0}^{\infty} b_n$ *es convergente y* $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$, *entonces* $\sum_{n=0}^{\infty} a_n$ *es convergente.*

iii) Si $\sum_{n=0}^{\infty} a_n$ *es divergente y* $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$, *entonces* $\sum_{n=0}^{\infty} b_n$ *es divergente.*

Ejemplos

1. Estudiemos la convergencia de la serie $\sum_{n=1}^{\infty} \frac{e^{\frac{1}{n}}}{n!}$. Primeramente $\sum_{n=1}^{\infty} \frac{1}{n!} = e$. Apliquemos el criterio de

comparación por paso al límite, ponemos $a_n = \frac{e^{\frac{1}{n}}}{n!},\ b_n = \frac{1}{n!} \quad \forall n \in \mathbb{Z}^+$, entonces

$$\frac{a_n}{b_n} = \frac{\frac{e^{\frac{1}{n}}}{n!}}{\frac{1}{n!}} = e^{\frac{1}{n}} \underset{n\longrightarrow\infty}{\longrightarrow} 1,$$

luego, por la parte i) del teorema precedente, la convergencia de la serie $\sum_{n=0}^{\infty} \frac{1}{n!}$ implica la convergencia de la serie $\sum_{n=1}^{\infty} \frac{e^{\frac{1}{n}}}{n!}$. Así, resulta $\sum_{n=1}^{\infty} \frac{e^{\frac{1}{n}}}{n!}$ es convergente.

2. Las series $\sum_{n=0}^{\infty} \frac{1}{n}$ y $\sum_{n=0}^{\infty} \frac{1}{\sqrt{n}}$ divergen. Ponemos $a_n = \frac{1}{n},\ b_n = \frac{1}{\sqrt{n}} \quad \forall n \in \mathbb{Z}^+$, entonces,

$$\frac{a_n}{b_n} = \frac{\frac{1}{n}}{\frac{1}{\sqrt{n}}} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}} \underset{n\longrightarrow\infty}{\longrightarrow} 0.$$

Por la parte iii) del teorema de conparación por paso al límite, la divergencia de la serie armónica implica la divergencia de la serie $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$.

Teorema 4 *(criterio de la integral)*
Sea $\sum_{n=1}^{\infty} a_n$ *una serie de números positivos, donde* (a_n) *es decreciente. Sea* f *una función de* $[1, \infty[$ *en* $\mathbb{R}$ *tal que* $f(n) = a_n,\ n = 1, 2, ...$ *Para* $n \in \mathbb{Z}^+$, *se define* $I_n = \int_1^n f(x)\, dx$. *Entonces, la sucesión* (I_n) *converge si y solo si* $\sum_{n=1}^{\infty} a_n$ *converge, o* (I_n) *diverge si y solo si* $\sum_{n=1}^{\infty} a_n$ *diverge.*

Ejemplo

Sea $p \in \mathbb{R}$ y consideremos la serie $\sum_{n=1}^{\infty} \frac{1}{n^p}$. Estudiemos la convergencia de esta serie. Para ello aplicamos el criterio de la integral. Definimos la función f como sigue: $f(x) = \frac{1}{x^p},\ x \geq 1$. Entonces, para $n = 1, 2, ...,$

$$I_n = \int_1^n f(x)\, dx = \int_1^n \frac{dx}{x^p}.$$

Para $p = 1$, se tiene

$$I_n = \int_1^n \frac{dx}{x} = \ln(x) \,|_1^n = \ln(n) - \ln 1 = \ln(n).$$

Resulta $\lim_{n\to\infty} I_n = \lim_{n\to\infty} \ln(n) = \infty$, y por el criterio de la integral, la serie $\sum_{n=1}^{\infty} \frac{1}{n}$ diverge.

Supongamos $p \neq 1$. Entonces

$$I_n = \int_1^n \frac{dx}{x^p} = \frac{1}{1-p} x^{-p+1}\Big|_1^n = \frac{1}{1-p}\left(n^{-p+1} - 1\right).$$

Como $\lim\limits_{n\to\infty} n^{-p+1} = 0 \Leftrightarrow p > 1$, se deduce

$$\lim_{n\to\infty} I_n = \frac{1}{1-p} \lim_{n\to\infty} \left(n^{-p+1} - 1\right) = \frac{1}{p-1} \Leftrightarrow p > 1.$$

Conclusión**:** por el criterio de la integral, la serie $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converge si y solo si $p > 1$. La serie $\sum_{n=1}^{\infty} \frac{1}{n^p}$ diverge si y solo si $p \leq 1$.

Teorema 5 ***(criterio del cociente)***
Sea $\sum_{n=0}^{\infty} a_n$ *una serie de números positivos y supongamos que* $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = L$. *Entonces*

i) Si $0 \leq L < 1$, *la serie* $\sum_{n=0}^{\infty} a_n$ *converge.*

ii) Si $L = 1$, *el criterio no decide.*

iii) Si $L > 1$, *la serie* $\sum_{n=0}^{\infty} a_n$ *diverge.*

Ejemplos

1. Sea $a > 1$. Estudiemos la convergencia de la serie $\sum_{n=0}^{\infty} \frac{(-1)^n n}{a^n}$. Apliquemos el criterio del cociente. Para el efecto, ponemos $|a_n| = \left|\frac{(-1)^n n}{a^n}\right| \quad n \in \mathbb{Z}^+$. Luego

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{(-1)^{n+1}(n+1)}{a^{n+1}} \times \frac{a^n}{(-1)^n n}\right| = \frac{n+1}{na} = \frac{1}{a}\left(1+\frac{1}{n}\right) \underset{n\longrightarrow\infty}{\longrightarrow} \frac{1}{a},$$

y como $a > 1$, resulta

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = \frac{1}{a} < 1.$$

Por el criterio del cociente $\sum_{n=0}^{\infty} \frac{n}{a^n}$ converge. Mas aún, la serie $\sum_{n=0}^{\infty} \frac{(-1)^n n}{a^n}$ converge absolutamente.

2. Consideremos la serie $\sum_{n=2}^{\infty} \frac{1}{\ln(n)}$. Apliquemos el criterio del cociente. Sea $a_n = \frac{1}{\ln(n)} \quad n \in \mathbb{Z}^+$ con $n \geq 2$, entonces

$$\frac{a_{n+1}}{a_n} = \frac{\ln(n)}{\ln(n+1)} \underset{n\longrightarrow\infty}{\longrightarrow} 1.$$

Como $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = 1$, el criterio del cociente no decide. Sabemos que la serie geométrica $\sum_{n=1}^{\infty} \frac{1}{n}$ es divergente. Sea $b_n = \frac{1}{n} \quad n \in \mathbb{Z}^+$. Apliquemos el criterio de comparación por paso al límite. Tenemos

$$\frac{b_n}{a_n} = \frac{\frac{1}{n}}{\frac{1}{\ln(n)}} = \frac{\ln(n)}{n} \qquad n \in \mathbb{Z}^+.$$

Para calcular el límite apliquemos la regla de L'Hôpital a la función $f(x) = \frac{\ln x}{x}$. Se tiene

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \frac{\ln x}{x} = \lim_{x\to\infty} \frac{\frac{1}{x}}{1} = 0.$$

Así, $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$ y la divergencia de la serie armónica implica la divergencia de la serie $\sum_{n=2}^{\infty} \frac{1}{\ln(n)}$.

3. La serie $\sum_{n=1}^{\infty} \frac{\alpha^n}{n^3}$ con $\alpha > 1$ es divergente, pués si $a_n = \frac{\alpha^n}{n^3} \qquad n \in \mathbb{Z}^+$, entonces

$$\frac{a_{n+1}}{a_n} = \frac{\alpha^{n+1}}{(n+1)^3} \times \frac{n^3}{\alpha^n} = \alpha\left(1-\frac{1}{n+1}\right)^3 \underset{n\longrightarrow\infty}{\longrightarrow} \alpha,$$

luego $\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = \alpha > 1$, y por el criterio del cociente, la serie diverge.

Teorema 6 *(criterio de la raíz)*
Sea $\sum_{n=0}^{\infty} a_n$ *una serie de números no negativos y supongamos que* $\lim_{n\to\infty} a_n^{\frac{1}{n}} = L$. *Entonces*

i) Si $0 \leq L < 1$, *la serie* $\sum_{n=0}^{\infty} a_n$ *converge.*

ii) Si $L = 1$, *el criterio no decide.*

iii) Si $L > 1$, *la serie* $\sum_{n=0}^{\infty} a_n$ *diverge.*

Ejemplos

1. Sea $x \in]-1,1[$. La serie $\sum_{n=0}^{\infty} x^n$ es convergente y $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$. Verifiquemos la convergencia utilizando el criterio de la raíz. Sea $a_n = |x|^n \quad n \in \mathbb{Z}^+$, entonces

$$\lim_{n\to\infty} a_n^{\frac{1}{n}} = \lim_{n\to\infty} (|x|^n)^{\frac{1}{n}} = |x| < 1.$$

   Así, la serie $\sum_{n=0}^{\infty} x^n$ es convergente cuando $|x| < 1$.

2. Consideremos la serie $\sum_{n=1}^{\infty} \frac{1}{n^{\frac{1}{3}}}$. Sea $a_n = \frac{1}{n^{\frac{1}{3}}} \quad n \in \mathbb{Z}^+$, por el criterio de la raíz, tenemos

$$\lim_{n\to\infty} a_n^{\frac{1}{n}} = \lim_{n\to\infty} \frac{1}{n^{\frac{1}{3n}}} = 1.$$

   Este criterio no decide la convergencia o divergencia de la serie.

   Notemos que $n^{\frac{1}{3n}} = \exp\left(\frac{\ln(n)}{3n}\right)$ y en consecuencia

$$\lim_{n\to\infty} n^{\frac{1}{3n}} = \lim_{n\to\infty} \exp\left(\frac{\ln(n)}{3n}\right) = \exp\left(\lim_{n\to\infty} \frac{1}{3n} \ln(n)\right) = e^0 = 1.$$

   Hemos utilizado el resultado siguiente: la función dada por $f(x) = e^x \quad x \in \mathbb{R}$ es continua y (x_n) una sucesión real convergente. Entonces $\lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n) = e^{\lim_{n\to\infty} x_n}$.

   Aplicando el criterio de la integral, resulta que la serie propuesta es divergente.

3.1.3. Cálculo aproximado de series numéricas.

Sean $\sum_{k=1}^{\infty} a_k$ una serie absolutamente convergente de números reales, $S = \sum_{k=1}^{\infty} a_k$ y $S_n = \sum_{k=1}^{n} a_k \quad n \geq 1$. Por definición de serie convergente, dado $\epsilon > 0$, existe $N \in \mathbb{N}$ tal que $|S - S_n| < \epsilon$, $\forall n \geq N$. Desde el punto de vista numérico es importante determinar el más pequeño entero positivo N tal que $\sum_{n=N+1}^{\infty} a_k < \epsilon$, pués reduce el número de téminos a utilizar en el cálculo de la suma S_N que aproxima a S con la precisión fijada ϵ. Por otro lado, resulta difícil determinar dicho entero N. Sin embargo, para las series numéricas absolutamente convergentes resulta útil aplicar los criterios de convergencia (por ejemplo: de la integral, del cociente, de la raíz, de comparación, entre otros), que permiten determinar tal entero N.

Para fijar las ideas, sean $\sum_{k=1}^{\infty} a_k$ una serie convergente de números positivos, (b_k) una sucesión real de números positivos tal que $\sum_{k=1}^{\infty} b_k = 1$ y supongamos que se verfica $\frac{a_k}{b_k} \underset{k\longrightarrow\infty}{\longrightarrow} 0$, lo que muestra que la

sucesión del numerador converge a cero mucho más rápidamente que la del denominador. Entonces, dado $\epsilon > 0$, existe $N \in \mathbb{N}$ tal que $\frac{a_k}{b_k} < \epsilon \quad \forall k \geq N$. De esta desigualdad se sigue que $a_k < \epsilon b_k \quad \forall k \geq N$, luego

$$\sum_{k=N+1}^{\infty} a_k < \epsilon \sum_{k=N+1}^{\infty} b_n < \epsilon \sum_{k=1}^{\infty} b_n = \epsilon.$$

Sea $S_N = \sum_{k=1}^{N} a_k$, entonces $|\sum_{k=1}^{\infty} a_k - S_N| = |\sum_{k=N+1}^{\infty} a_k| < \epsilon$.

Según este criterio si seleccionamos una sucesión positiva (b_k) tal que $\sum_{k=1}^{\infty} b_k = 1$ y $\lim_{k\to\infty} \frac{a_k}{b_k} = 0$, resulta fácil hallar $N \in \mathbb{Z}^+$ que satisfaga la desigualdad $\frac{a_k}{b_k} < \epsilon \quad \forall k \geq N$. Elegimos tal N como el más pequeño entero positivo que verifique $\frac{a_N}{b_N} < \epsilon$. Note que N no es el óptimo, pués depende de la sucesión elegida (b_k). En el caso general, hallar el óptimo N:

$$Min\{N \in \mathbb{N} \,|\, \sum_{k=N+1}^{\infty} a_k < \epsilon\}$$

resulta una tarea difícil.

Para una clase de series numéricas rápidamente convergentes se tendrá N pequeño, para las series numéricas que convergen muy lentamente, el entero positivo N será muy grande con lo que este procedimiento no es muy adecuado para esta clase de series numéricas, pués los errores de redondeo y de truncamiento afectarán seriamente en el resultado.

Para series numéricas rápidamente convergentes, se propone el siguiente algoritmo que permite determinar el más pequeño entero N que verifica la condición $\frac{a_k}{b_k} < \epsilon \quad \forall k \geq N$.

Algoritmo

Datos de entrada: $\epsilon > 0$, sucesiones (a_n), (b_n).

Datos de salida: N.

1. Hacer $k = 1$
2. Si $\frac{a_k}{b_k} < \epsilon$. Continuar en 4).
3. Si $\frac{a_k}{b_k} \geq \epsilon$, hacer $k = k + 1$. Continuar en 2).
4. Imprimir N.
5. Fin.

Determinado el número de términos de S_N, la etapa siguiente es la elaboración de un algoritmo para el cálculo de S_N de modo que se conserven las normas establecidas de condicionamiento y estabilidad numérica.

Ejemplos

1. La serie $\sum_{k=1}^{\infty} \frac{1}{n^p} \quad p > 1$ es convergente, más aún, esta serie converge muy lentamente. Sea $\epsilon > 0$ y apliquemos el criterio de la integral:

$$\int_N^{\infty} \frac{dt}{t^p} = \frac{t^{1-p}}{1-p} \,|_N^{\infty} = \frac{1}{(p-1)N^{p-1}} < \epsilon,$$

que implica $N > \left[\frac{1}{(p-1)\epsilon}\right]^{\frac{1}{p-1}}$, donde $[\cdot]$ denota la función mayor entero menor o igual que. Así por ejemplo, para $p = 2$ se tiene que $N > \left[\frac{1}{\epsilon}\right]$. Note que

$$\frac{3}{2} - \frac{1}{N+1} < \sum_{n=1}^{\infty} \frac{1}{n^2} < 2 - \frac{1}{N}.$$

Elegido $N > \left[\frac{1}{\epsilon}\right]$, se define la suma $S_N = \sum_{n=1}^{N} \frac{1}{n^2}$ que aproxima a $\sum_{n=1}^{\infty} \frac{1}{n^2}$ con una precisión ϵ. Para $\epsilon = 10^{-8}$ se tendrá $N > 10^8$ que es un número muy grande de términos y los errores de redondeo y de truncamiento influirán en el cálculo de S_N, lo que muestra las deficiencias del método. Utilizando las series de Fourier, se prueba que $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

2. Consideremos la serie $\sum_{k=0}^{\infty} \frac{2^k}{k!(2k+1)}$. Este es un ejemplo de una serie numérica rápidamente convergente. En efecto, apliquemos el criterio del cociente y pongamos $a_k = \frac{2^k}{k!(2k+1)}$, entonces

$$\frac{a_{k+1}}{a_k} = \frac{2(2k+1)}{(k+1)(2k+3)} < \frac{2}{k+1} \underset{k\longrightarrow\infty}{\longrightarrow} 0.$$

La serie es convergente. Sea $\epsilon = 10^{-10}$ y $b_k = \frac{1}{2^k}$, entonces $\sum_{k=0}^{\infty} \frac{1}{2^k} = 2$. Luego

$$\frac{a_k}{b_k} = \frac{4^k}{k!(2k+1)} \longrightarrow 0 \text{ cuando } k \to \infty.$$

El más pequeño entero positivo N tal que $\frac{a_k}{b_k} < 10^{-10} \quad \forall k \geq N$ es $N = 23$. Por lo tanto $\sum_{k=24}^{\infty} \frac{2^k}{k!(2k+1)} < 10^{-10}$, con lo que $S_N = \sum_{k=0}^{23} \frac{2^k}{k!(2k+1)}$ aproxima a $\sum_{k=0}^{\infty} \frac{2^k}{k!(2k+1)}$ con una precisión de 10^{-10}. La suma S_N se evalúa del modo siguiente:

$$\begin{aligned} S_N &= 1 + 2\left(\frac{1}{3} + \frac{2}{2}\left(\frac{1}{5} + \frac{2}{3}\left(\frac{1}{7} + \cdots + \frac{2}{22}\left(\frac{1}{2\times 22+1} + \frac{2}{23\times(2\times 23+1}\right)\right)\right)\right)\right) \\ &= 2{,}3644538928\cdots, \end{aligned}$$

resultado en el que están incluidos los errores de redondeo y de truncamiento.

3.2. Sucesiones y series de funciones. Convergencia puntual y uniforme.

Iniciamos esta sección con la convergencia de sucesiones de funciones, tratamos básicamente la convergencia puntual y la convergencia uniforme e introducimos los resultados importantes (sin demostración) sobre la convergencia uniforme y la continuidad, integrabilidad y derivabilidad que serán aplicados en el estudio de las series de funciones, particularmente en las series de potencias y las series de Fourier. El lector que está familiarizado con las sucesiones y series de funciones puede pasar inmediatamente a la aproximación numérica de las series de funciones, aquel que no está familiarizado tendrá la ocasión de tratar este tema en forma resumida. Al final del capítulo se sugiere una bibliografía especializada en estos tópicos.

3.2.1. Sucesiones de funciones

Sea $A \subset \mathbb{R}$, $A \neq \emptyset$. Se denota con $\mathcal{F}(A)$ el espacio vectorial de funciones reales definidas en A.

Definición 3 *Sean $A \subset \mathbb{R}$, $A \neq \emptyset$, $I \subset \mathbb{N}$ con $I \neq \emptyset$. A toda función φ de I en $\mathcal{F}(A)$ se le llama sucesión de funciones.*

Notación**:** si φ es una sucesión de funciones, se tiene $\varphi : \begin{cases} I \to \mathcal{F}(A) \\ n \to \varphi(n) = f_n, \end{cases}$ donde para cada $n \in I$, f_n es una función real definida en A. A la sucesión φ la notaremos (f_n) y diremos sucesión de funciones definida en el conjunto A. El conjunto I se llama conjunto de índices, $f_n(x)$ con $x \in A$ se llama término general de la sucesión (f_n).

En el estudio de las sucesiones de funciones tienen especial interés las sucesiones convergentes y particularmente la convergencia uniforme y sus propiedades.

Convergencia puntual.

Sea (f_n) una sucesión de funciones reales definidas en A. Para cada $x \in A$, $(f_n(x))$ es una sucesión numérica real. Si existe $\lim_{n\to\infty} f_n(x)$, esto es, existe $f(x) \in \mathbb{R}$ tal que $\lim_{n\to\infty} f_n(x) = f(x)$, diremos que $(f_n(x))$ converge a $f(x)$. A este tipo de convergencia la llamaremos convergencia puntual.

Para los puntos $x \in A$ en los que $\lim_{n\to\infty} f_n(x)$ existe, se define una función real f mediante la relación $\lim_{n\to\infty} f_n(x) = f(x)$. Se tiene $\mathrm{Dom}(f) \subset A$. Escribiremos $f_n(x) \underset{n\to\infty}{\to} f(x)$ con $x \in A$. En lo que sigue, supondremos que $\mathrm{Dom}(f) = A$. Tenemos la siguiente definición de convergencia puntual.

Definición 4 *Sea (f_n) una sucesión de funciones reales definidas en A, $f \in \mathcal{F}(A)$. Se dice que (f_n) converge puntualmente a f si y solo si se cumple la siguiente condición:*

$$\forall \varepsilon > 0, \forall x \in A, \quad \exists n_0 \in \mathbb{Z}^+ \text{ tal que } \quad \forall n \geq n_0 \Rightarrow |f_n(x) - f(x)| < \varepsilon.$$

Si la sucesión de funciones (f_n) converge a la función f en el conjunto A, escribiremos $\lim_{n\to\infty} f_n = f$ o de forma equivalente $\lim_{n\to\infty} f_n(x) = f(x) \quad x \in A$.

Si para algún $x \in A$, $\lim_{n\to\infty} f_n(x)$ no existe, diremos que la sucesión $(f_n(x))$ diverge y que la sucesión (f_n) no es convergente en el conjunto A.

En el estudio de sucesiones de funciones, la primera tarea es el análisis de la convergencia puntual. Más adelante veremos la convergencia uniforme y daremos más atención a este tipo de convergencia.

En la convergencia puntual, el elemento $n_0 \in \mathbb{Z}^+$ depende de ε, y de cada punto $x \in A$. En este tipo de convergencia no es posible hallar un $n_0 \in \mathbb{Z}^+$ dependiente únicamente de $\varepsilon > 0$.

Convergencia uniforme.

Definición 5 *Sea (f_n) una sucesión de funciones reales definidas en A, $f \in \mathcal{F}(A)$. Se dice que (f_n) converge uniformemente a f si y solo si se cumple la siguiente condición:*

$$\forall \varepsilon > 0, \ \exists n_0 \in \mathbb{Z}^+ \text{ tal que } \forall x \in A, \ \forall n \in \mathbb{Z}^+, \quad n \geq n_0 \Rightarrow |f_n(x) - f(x)| < \varepsilon.$$

Escribiremos $\lim_{n\to\infty} f_n = f$ uniformemente o también $f_n \underset{n\to\infty}{\to} f$ uniformemente.

Es preciso establecer la diferencia que existe entre la convergencia puntual y la convergencia uniforme. En la convergencia uniforme, el elemento n_0 de $\mathbb{Z}^+$ depende de ε, en general del conjunto A y no de $x \in A$,

el número $n_0 \in \mathbb{Z}^+$ es global para todo $x \in A$, mientras que en la convergencia puntual $n_0 \in \mathbb{Z}^+$ depende de ε y de cada $x \in A$, y, no es posible hallar un $n_0 \in \mathbb{Z}^+$ global para todos los elementos del conjunto A. La convergencia uniforme implica la convergencia puntual, pero el recíproco, en general, no es cierto.

Teorema 7 *Sea (f_n) una sucesión de funciones definidas en el conjunto A que converge puntualmente a la función f definida en A. Para cada $n \in \mathbb{Z}^+$ se define*

$$M_n = \sup_{x \in A} |f_n(x) - f(x)|.$$

Entonces, (f_n) converge uniformemente a f si y solo si (M_n) converge a 0.

Del criterio establecido en el teorema precedente, se sigue que (f_n) no converge uniformemente a f si y solo si $\lim_{n\to\infty} M_n \neq 0$.

En el siguiente teorema se establece el criterio de Cauchy para la convergencia uniforme.

Teorema 8 *Sean $A \subset \mathbb{R}$ con $A \neq \emptyset$, (f_n) una sucesión de funciones definidas en A. Enconces (f_n) converge uniformemente en A a alguna función f si y solo si para $\varepsilon > 0$, existe $n_0 \in \mathbb{Z}^+$ tal que $\forall x \in A$*

$$|f_m(x) - f_n(x)| < \varepsilon \quad \text{si } m, n \in \mathbb{Z}^+ \quad \text{con } m, n \geq n_0.$$

Teorema 9 *Sean $A \subset \mathbb{R}$, con $A \neq \emptyset$, (f_n) y (g_n) sucesiones de funciones reales definidas en A, $f, g \in \mathcal{F}(A)$ y $\lambda \in \mathbb{R}$. Si $\lim_{n\to\infty} f_n = f$ y $\lim_{n\to\infty} g_n = g$ uniformemente, entonces*

i) $(f_n + g_n)$ converge uniformemente a $f + g$.

ii) (λf_n) converge uniformement a λf.

iii) $(|f_n|)$ converge uniformemente a $|f|$.

iv) Si existen $M_1 > 0$, $M_2 > 0$ tales que $\sup_{n\in\mathbb{Z}^+} \sup_{x\in\mathbb{A}} |f_n(x)| \leq M_1$, $\sup_{n\in\mathbb{Z}^+} \sup_{x\in\mathbb{A}} |g_n(x)| \leq M_2$, entonces $(f_n g_n)$ converge uniformement a fg.

v) Si existe $M > 0$ tal que $\inf_{n\in\mathbb{Z}^+} \inf_{x\in\mathbb{A}} |f_n(x)| \geq M$ y $f \neq 0$, entonces $(\frac{1}{f_n})$ converge uniformemente a $\frac{1}{f}$.

Convergencia uniforme y continuidad.

Sean $A \subset \mathbb{R}$, $A \neq \emptyset$, (f_n) una sucesión de funciones reales definidas en A que converge a una función f definida en A, esto es $\lim_{n\to\infty} f_n(x) = f(x)$ $x \in A$. Adicionalmente, suponemos que cada función f_n es continua en todo punto $x \in A$, la pregunta que surge es: ¿la función límite f hereda la continuidad de la sucesión (f_n)?

Sea $x, x_0 \in A$. Si fuese f continua, se tendría $\lim_{x\to x_0} f(x) = f(x_0)$, que en términos del límite de la sucesión de funciones (f_n) la igualdad precedente se expresaría como

$$\lim_{x\to x_0} \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \lim_{x\to x_0} f_n(x).$$

Lastimosamente esta igualdad no siempre se cumple. Para responder a la pregunta, examinemos la sucesión de funciones (f_n) definida como $f_n(x) = \exp(-nx^2)$ $x \in \mathbb{R}$, $n = 1, 2, \ldots$.

Para $x = 0$, $f_n(0) = 1$ consecuentemente $\lim_{n\to\infty} f_n(0) = 1$, y para $x \neq 0$, $\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} e^{-nx^2} = 0$.

La función límite f está definida como $f(x) = \begin{cases} 1, \text{ si } x = 0, \\ 0, \text{ si } x \neq 0. \end{cases}$. Esta función no es continua en $x = 0$.

Cada función f_n es continua en todo $\mathbb{R}$. Resulta

$$\lim_{n\to\infty} \lim_{x\to 0} f_n(x) = \lim_{n\to\infty} \lim_{x\to 0} e^{-nx^2} = \lim_{n\to\infty} 1 = 1,$$

mientras que si $x \neq 0$,

$$\lim_{x\to 0}\lim_{n\to\infty} f_n(x) = \lim_{x\to 0}\lim_{n\to\infty} e^{-nx^2} = \lim_{x\to 0} 0 = 0.$$

Claramente $\lim_{n\to\infty}\lim_{n\to 0} f_n(x) \neq \lim_{n\to 0}\lim_{n\to\infty} f_n(x)$. Este y otros ejemplos muestran que si la sucesión de funciones continuas (f_n) converge puntualmente a una función f, en general, f no hereda la continuidad de cada función f_n. En el siguiente teorema se da una condición para que la función límite f sea continua.

Teorema 10 *Sean $A \subset \mathbb{R}$ con $A \neq \emptyset$ y (f_n) una sucesión de funciones continuas en todo punto $x \in A$. Si (f_n) converge uniformemente a una función límite f definida en A, entonces f es continua en todo punto $x \in A$.*

Este resultado se sintetiza en el siguiente esquema:

$$\begin{cases} f_n \text{ continua en } A,\ n = 1, 2, \ldots, \\ f_n \underset{n\to\infty}{\longrightarrow} f \quad \text{uniformemente}, \end{cases} \Rightarrow \begin{cases} f \text{ continua en A}, \\ \lim_{n\to\infty}\lim_{x\longrightarrow x_0} f_n(x) = f(x_0). \end{cases}$$

Convergencia uniforme e integración.

Sea $[a, b]$ un intervalo cerrado de $\mathbb{R}$. Se considera una sucesión de funciones reales (f_n) definida en $[a, b]$ que converge a una función f definida en el mismo intervalo $[a, b]$. Supongamos que se tiene la convergencia puntual, esto es

$$\lim_{n\to\infty} f_n(x) = f(x) \quad x \in [a, b].$$

Adicionalmente, supongamos que cada función f_n es integrable en $[a, b]$, ¿es la función límite f integrable en $[a, b]$? ¿se verifica $\lim_{n\to\infty}\int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx$? En definitiva, se desea saber las condiciones que se deben verificar para que se cumpla la igualdad siguiente:

$$\lim_{n\to\infty}\int_a^b f_n(x)\,dx = \int_a^b \left(\lim_{n\to\infty} f_n(x)\right) dx,$$

es decir que podamos intercambiar el símbolo de integral con el del límite, o también que una sucesión convergente se pueda integrar término a término.

La convergencia de (f_n) a f así como la integrabilidad de cada función f_n no garantiza, en general, que se verifique la igualdad anterior, como se puede comprobar con el siguiente ejemplo.

Sea (f_n) la sucesión de funciones definida como $f_n(x) = \begin{cases} n, & \text{si } x \in \left[\frac{1}{n}, \frac{2}{n}\right], \\ 0, & \text{si } x \in [0, 2] \setminus \left[\frac{1}{n}, \frac{2}{n}\right] \end{cases}$ $n = 1, 2, \ldots$. Cada función f_n es integrable en $[0, 2]$ (f_n es una función escalonada), y,

$$\int_0^2 f_n(x)\,dx = \int_0^{\frac{1}{n}} f_n(x)\,dx + \int_{\frac{1}{n}}^{\frac{2}{n}} f_n(x)\,dx + \int_{\frac{2}{n}}^2 f_n(x)\,dx = \int_{\frac{1}{n}}^{\frac{2}{n}} n\,dx = 1.$$

Luego $\lim_{n\to\infty}\int_0^2 f_n(x)\,dx = 1$. Por otro lado, $f_n(x) \underset{n\to\infty}{\longrightarrow} 0 \quad x \in [0, 2]$. Se pone $f(x) = 0 \quad x \in [0, 2]$, y se tiene

$$\int_0^2 \left(\lim_{n\to\infty} f_n(x)\right) dx = \int_0^2 f(x)\,dx = 0.$$

Tenemos para este ejemplo se tiene $\lim_{n\to\infty}\int_0^2 f_n(x)\,dx \neq \int_0^2 \left(\lim_{n\to\infty} f_n(x)\right) dx$.

Antes de enunciar el teorema relativo a la convergencia uniforme y la integración revisamos las condiciones que verifican las funciones integrables en $[a, b]$.

Sea u una función acotada en $[a, b]$. Se dice que u es integrable (Riemann integrable) en $[a, b]$ si y solo si

$$\underline{I}(u) = \operatorname*{Sup}_{\varphi\le u}\int_a^b \varphi(x)\,dx = \operatorname*{Inf}_{u\le\theta}\int_a^b \theta(x)\,dx = \overline{I}(u),$$

donde φ y θ son funciones escalonadas en $[a,b]$ tales que $\varphi \leq u \leq \theta$.

Los números reales $\underline{I}(u)$ y $\overline{I}(u)$ se llaman integrales inferior y superior, respectivamente. Se verifica, además que para toda función acotada u definida en $[a,b]$,

$$\int_a^b \varphi(x)\,dx \leq \underline{I}(u) \leq \overline{I}(u) \leq \int_a^b \theta(x)\,dx,$$

donde $\varphi,\ \theta$ son funciones escalonadas definidas en $[a,b]$ tales que $\varphi \leq u \leq \theta$.

Los siguientes enunciados son equivalentes:

i) u es integrable en $[a,b]$.

ii) Para todo $\varepsilon > 0$, existen dos funciones escalonadas θ, φ definidas en $[a,b]$ tales que $\varphi \leq u \leq \theta$ y $\int_a^b (\theta - \varphi)\,dx < \varepsilon$.

Teorema 11 *Sea (f_n) una sucesión real de funciones integrables en $[a,b]$. Supongamos que (f_n) converge uniformemente a una función f definida en $[a,b]$. Entonces,*

i) f es integrable en $[a,b]$.

ii) $\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b \left(\lim_{n\to\infty} f_n(x)\right) dx.$

El resultado del teorema se sintetiza en el siguiente esquema:

$$\begin{cases} f_n \text{ integrable en } [a,b],\ n = 1,2,\ldots, \\ f_n \underset{n\to\infty}{\longrightarrow} f \text{ uniformemente,} \end{cases} \Rightarrow \begin{cases} f \text{ integrable en } [a,b], \\ \lim_{n\to\infty} \int_a^b f_n = \int_a^b f. \end{cases}$$

En este esquema puede verse que la convergencia uniforme es una condición suficiente para que f sea integrable en $[a,b]$.

Teorema 12 *Sea (f_n) una sucesión de funciones continuas en $[a,b]$ que converge uniformemente a f. Entonces*

$$\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx.$$

Convergencia uniforme y derivación.

Sea $A \subset \mathbb{R}$, A abierto $A \neq \emptyset$, (f_n) una sucesión de funciones reales definidas en A que converge puntualmente a una función f definida en A, esto es $\lim_{n\to\infty} f_n(x) = f(x) \quad x \in A$. Supongamos que cada función es derivable en todo punto $x \in A$, ¿la función límite f es derivable en $x \in A$? De ser así, tendríamos

$$\frac{df}{dx}(x) = \lim_{n\to\infty} \frac{df_n}{dx}(x) \qquad x \in A,$$

o sea

$$\frac{d}{dx}\left(\lim_{n\to\infty} f_n(x)\right) = \lim_{n\to\infty} \frac{df_n}{dx}(x) \qquad x \in A,$$

es decir que la sucesión (f_n) puede derivarse término a término. Lamentablemente esta igualdad no siempre se cumple como se ilustra en el ejemplo siguiente.

Sea (f_n) la sucesión de funciones reales definidas como $f_n(x) = \frac{1}{\sqrt{n}} \operatorname{sen}(nx) \quad x \in \mathbb{R},\ n = 1,2,\ldots$. Cada función f_n es derivable y $\dfrac{df_n}{dx}(x) = \sqrt{n}\cos(nx) \quad x \in \mathbb{R}, \quad n = 1,2,\ldots$. Para $x = 2k\pi \quad k \in \mathbb{Z}$, se tiene

$$f_n(2k\pi) = \frac{1}{\sqrt{n}} \operatorname{sen}(2nk\pi) = 0, \qquad \frac{df_n}{dx}(2k\pi) = \sqrt{n}\cos(2kn\pi) = \sqrt{n} \underset{n\to\infty}{\longrightarrow} \infty.$$

Puesto que $\frac{1}{\sqrt{n}}|\text{sen}(nx)| \leq \frac{1}{\sqrt{n}} \underset{n\to\infty}{\longrightarrow} 0 \quad \forall x \in \mathbb{R}$, se pone $f(x) = 0 \quad x \in \mathbb{R}$ y se tiene que $f_n \underset{n\to\infty}{\longrightarrow} f$ uniformemente. Además $\dfrac{df}{dx}(x) = 0$ y en particular $\dfrac{df}{dx}(2k\pi) = 0 \quad \forall k \in \mathbb{Z}$. Resulta, para todo $k \in \mathbb{Z}$,

$$0 = \frac{df}{dx}(2k\pi) = \frac{d}{dx}\left(\lim_{n\to\infty} f_n(2k\pi)\right) \neq \lim_{n\to\infty} \frac{df_n}{dx}(2k\pi) = \infty.$$

Es este ejemplo se muestra que inclusive la convergencia uniforme de la sucesión (f_n) no basta, debemos tener algo más sobre la sucesión $\left(\dfrac{df_n}{dx}\right)$. En el siguiente teorema se proponen las condiciones bajo las cuales se puede derivar término a término.

Teorema 13 *Sean $A \subset \mathbb{R}$, $A \neq \emptyset$ un conjunto abierto, (f_n) una sucesión real de funciones derivables en cada punto $x \in A$ y tales que $\left|\dfrac{df_n}{dx}(x)\right| < \infty, \quad n = 1, 2, \ldots$. Supongamos que $(f_n(x_0))$ converge para algún punto $x_0 \in A$ y que la sucesión (f_n') converge uniformemente a una función g. Entonces,*

i) Existe una función real f definida en A tal que $f_n \underset{n\to\infty}{\longrightarrow} f$ uniformemente.

ii) Para cada $x \in A$, $\dfrac{df}{dx}(x) = g(x)$.

Este resultado se sintetiza en el siguiente esquema:

$$\begin{cases} |\dfrac{df_n}{dx}(x)| < \infty \quad \forall x \in A,\ n = 1, 2, \ldots, \\ x_0 \in A, \quad (f_n(x_0)) \quad \text{convergente}, \\ \dfrac{df_n}{dx} \underset{n\to\infty}{\longrightarrow} g \quad \text{uniformemente}, \end{cases} \Rightarrow \begin{cases} f_n \underset{n\to\infty}{\longrightarrow} f \quad \text{uniformemente}, \\ \dfrac{df}{dx}(x) = g(x) \quad \forall x \in A. \end{cases}$$

3.2.2. Series de funciones.

Sea (f_n) una sucesión de funciones definidas en un subconjunto no vacío A de $\mathbb{R}$. La suma $\sum_{k=1}^{\infty} f_k$ se llama serie de funciones. Para $n \in \mathbb{N}$, se define $S_n = \sum_{k=1}^{n} f_k$,y, $S_n(x) = \sum_{k=1}^{n} f_k(x) \quad x \in A$. La función f_n se llama término general de la serie y S_n se denomina suma parcial de la misma. Además, S_n es una función definida en el conjunto A, y (S_n) es una sucesión de funciones definida sobre A. Para cada $x \in A$, $(S_n(x))$ es una sucesión numérica. Si $\lim_{n\to\infty} S_n(x)$ existe, denotamos al mismo con $S(x)$, esto es, $\lim_{n\to\infty} S_n(x) = S(x) \quad x \in A$, y decimos la serie numérica $\sum_{k=1}^{\infty} f_k(x)$ tiene como suma $S(x)$. Escribimos $\sum_{k=1}^{\infty} f_k(x) = S(x) \quad x \in A$. Se define una función real S en todos los puntos $x \in A$ en los que $\lim_{n\to\infty} S_n(x) = \sum_{k=1}^{\infty} f_k(x)$ existe mediante la relación:

$$S(x) = \lim_{n\to\infty} S_n(x) = \sum_{k=1}^{\infty} f_k(x), \quad x \in A,$$

y diremos que $\sum_{k=1}^{\infty} f_k$ converge puntualmente a S. Se tiene $\text{Dom}(S) \subset A$.

Note que el estudio de la serie de funciones $\sum_{k=1}^{\infty} f_k$ se le ha conducido al estudio de la sucesión de funciones (S_n). La primera tarea es analizar su convergencia puntual, a continuación se debe estudiar la convergencia uniforme así como sus consecuencias, esto es, la convergencia uniforme y continuidad, convergencia uniforme e integración, convergencia uniforme y derivabilidad de dicha serie de funciones.

La convergencia uniforme de la serie y la continuidad tiene que ver con la cuestión relativa al intercambio entre el símbolo de sumatorio con el de límite:

$$\lim_{x\to a}\sum_{k=1}^{\infty} f_k(x) = \sum_{k=1}^{\infty}\lim_{x\to a} f_k(x) = \sum_{k=1}^{\infty} f_k(a) \quad a\in A,$$

resultado que.no siempre es verdadero. La convergencia uniforme de la serie y la integración está relacionada con el intercambio entre el símbolo de sumatorio con el de integración:

$$\int_a^b \left(\sum_{k=1}^{\infty} f_k(x)\right) dx = \sum_{k=1}^{\infty}\int_a^b f_k(x)\,dx \quad a,b\in A \text{ tales que } a<b,$$

intercambio que no siempre es posible. La convergencia uniforme de la serie y la derivabilidad está relacionada, tal como en los casos anteriores, con el intercambio del símbolo de derivación con el de sumatorio:

$$\frac{d}{dx}\left(\sum_{k=1}^{\infty} f_k(x)\right) = \sum_{k=1}^{\infty}\frac{df_k}{dx}(x) \quad x\in A,$$

este resultado no siempre es verdadero. ¿Cuáles son las condiciones suplementarias a la convergencia uniforme que debemos imponer al término general f_k de la serie de funciones para que cada una de las cuestiones citadas siempre sea posible? Estas cuestiones las abordaremos en esta sección.

A continuación proponemos algunos resultados de convergencia puntual y uniforme de series de funciones.

Teorema 14 *Sean $A\subset\mathbb{R}$, $A\neq\emptyset$, y, $\sum_{n=1}^{\infty} f_n$ una serie de funciones definidas en A. Entonces*

i) $\sum_{n=1}^{\infty} f_n$ *converge si y solo si se satisface la siguiente condición:*

$$\forall\varepsilon>0,\ \forall x\in A,\quad \exists p\in\mathbb{Z}^+ \text{ tal que } \forall m\in\mathbb{Z}^+,\qquad |f_{p+1}(x)+\cdots+f_{p+m}(x)|<\varepsilon.$$

ii) $\sum_{n=1}^{\infty} f_n$ *converge uniformemente si y solo si satisface la siguiente condición:*

$$\forall\varepsilon>0,\ \exists\ p\in\mathbb{Z}^+ \text{ tal que } \forall x\in A,\ \forall m\in\mathbb{Z}^+,\qquad |f_{p+1}(x)+\cdots+f_{p+m}(x)|<\varepsilon.$$

En el siguiente teorema se propone la conocida prueba de Weierstrass de la convergencia uniforme de series de funciones.

Teorema 15 *Sea $A\subset\mathbb{R}$ con $A\neq\emptyset$ y (u_n) una sucesión de funciones reales definidas en A. Supongamos que existe $M_n>0$ tal que $|u_n(x)|\leq M_n \quad \forall x\in A,\ n=1,2,\ldots$. Si la serie numérica $\sum_{n=1}^{\infty} M_n$ converge, entonces la serie $\sum_{n=1}^{\infty} u_n$ converge uniformemente en A.*

Ejemplos

1. Consideremos la serie $\sum_{k=1}^{\infty}\frac{1}{(k+1)^2}\operatorname{sen}\left(\frac{k\pi x}{2}\right) \quad x\in\mathbb{R}$. Mostremos que esta serie es convergente. Para ello apliquemos el criterio de Weierstrass. Tenemos

$$\left|\frac{1}{(k+1)^2}\operatorname{sen}\left(\frac{k\pi x}{2}\right)\right|\leq\frac{1}{(k+1)^2}\quad \forall x\in\mathbb{R},\ k=1,2,\ldots$$

Aplicando el criterio de la integral se prueba que la serie numérica $\sum_{n=1}^{\infty}\frac{1}{(n+1)^2}$ es convergente, por la prueba de Weierstrass se sigue que las serie $\sum_{k=1}^{\infty}\frac{1}{(k+1)^2}\operatorname{sen}\left(\frac{k\pi x}{2}\right)$ converge uniformemente en

todo $\mathbb{R}$, lo que define una función S en todo $\mathbb{R}$ dada como

$$S(x)=\sum_{k=1}^{\infty}\frac{1}{(k+1)^2}\operatorname{sen}\left(\frac{k\pi x}{2}\right)\quad x\in\mathbb{R}.$$

2. La serie geométrica $\sum_{k=0}^{\infty}x^k$ converge si y solo si $|x|<1$, en cuyo caso escribimos $\sum_{k=0}^{\infty}x^k=\frac{1}{1-x}\quad |x|<1$.

   Sea $0<a<1$. Para todo $x\in[-a,a]\subset\,]-1,1[$ se tiene $\left|x^k\right|\leq a^k\quad k=0,1,\ldots,\ \sum_{k=0}^{\infty}a^k=\frac{1}{1-a}$, por la prueba de Weierstrass resulta que la sucesión (S_n) definida como $S_n(x)=\sum_{k=0}^{\infty}x^k$ converge uniformemente a

$$S(x)=\sum_{k=0}^{\infty}x^k=\frac{1}{1-x}\quad x\in[-a,a].$$

   La serie $\sum_{k=0}^{\infty}x^k\quad x\in\,]-1,1[$ no converge uniformemente, únicamente se tiene convergencia puntual; esto es, $\sum_{k=0}^{\infty}x^k=\frac{1}{1-x}\quad |x|<1$.

Teorema 16 *Sea $\sum_{n=1}^{\infty}f_n$ una serie de funciones sobre un subconjunto A de $\mathbb{R}$ con $A\neq\emptyset$.*

i) Si $\sum_{n=1}^{\infty}|f_n|$ converge, entonces $\sum_{n=1}^{\infty}f_n$ converge, y, $\left|\sum_{n=1}^{\infty}f_n\right|\leq\sum_{n=1}^{\infty}|f_n|$.

ii) Si $\sum_{n=1}^{\infty}|f_n|$ converge uniformemente, entonces $\sum_{n=1}^{\infty}f_n$ converge uniformemente, y, $\left|\sum_{n=1}^{\infty}f_n\right|\leq\sum_{n=1}^{\infty}|f_n|$.

Ejemplo

Consideremos la serie $\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k}}{(2k+5)!}\quad x\in\mathbb{R}$. Probemos que converge uniformemente sobre cada intervalo $[-r,r]$ con $r>0$. En efecto, para cada $x\in\mathbb{R}$, la serie $\sum_{k=0}^{\infty}\left|\frac{(-1)^k x^{2k}}{(2k+5)!}\right|=\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k+5)!}$ converge, pués por el criterio del cociente, para $x\neq 0$ se tiene

$$\frac{|x|^{2(k+1)}}{(2(k+1)+5)!}\times\frac{(2k+5)!}{|x|^{2k}}=\frac{x^2}{(2k+6)(2k+7)}\underset{k\to\infty}{\longrightarrow}0.$$

Sea $r>0$. Entonces $\frac{x^{2k}}{(2k+5)!}\leq\frac{r^{2k}}{(2k+5)!}\qquad\forall x\in[-r,r],\ k=0,1,\ldots$.El criterio del cociente muestra que la serie $\sum_{k=0}^{\infty}\frac{r^{2k}}{(2k+5)!}$ converge. Por la prueba de Weierstrass, la serie $\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k+5)!}$ converge. Mostremos que la convergencia es uniforme. Sea $\varepsilon>0$. Por el criterio de Cauchy,

$$\exists p\in\mathbb{Z}^+\ \text{ tal que }\ \forall m\in\mathbb{Z}^+\Rightarrow\sum_{k=p+1}^{p+m}\frac{r^{2k}}{(2k+5)!}<\varepsilon,$$

de donde

$$\left|\sum_{k=p+1}^{p+m}\frac{(-1)^k x^{2k}}{(2k+5)!}\right|\leq\sum_{k=p+1}^{p+m}\frac{|x|^{2k}}{(2k+5)!}\leq\sum_{k=p+1}^{p+m}\frac{r^{2k}}{(2k+5)!}<\varepsilon\quad\forall x\in[-r,r]$$

que prueba la convergencia uniforme de la serie $\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k+5)!} \quad \forall x \in [-r,r]$.

Escribamos estos resultados en términos de sucesiones. Denotamos con (u_n), (v_n) las sucesiones definidas por

$$u_n(x) = \sum_{k=0}^{n} \frac{(-1)^k x^{2k}}{(2k+5)!}, \qquad v_n(x) = \sum_{k=0}^{n} \frac{x^{2k}}{(2k+5)!} \quad x \in [-r,r],\ n = 1,2,\ldots.$$

Se tiene $|u_n(x)| \leq v_n(x) \quad x \in [-r,r],\ n = 1,2,\ldots$La convergencia de la serie $\sum_{k=0}^{\infty} \frac{x^{2k}}{(2k+5)!} \quad x \in \mathbb{R}$ implica que $\lim_{n\to\infty} v_n(x)$ existe para todo $x \in [-r,r]$ y por el criterio de Cauchy,

$$\exists p \in \mathbb{Z}^+ \text{ tal que } \forall m,n \in \mathbb{Z}^+ \text{ con } \quad m,n \geq p \Rightarrow |v_m(x) - v_n(x)| < \varepsilon \quad \forall x \in [-r,r].$$

Para $m,n \in \mathbb{Z}^+$ tal que $n > m \geq p$, se tiene $v_n(x) - v_m(x) = \sum_{k=m+1}^{n} \frac{x^{2k}}{(2k+5)!} < \varepsilon \quad \forall x \in [-r,r]$, y de la desigualdad $|u_n(x) - u_m(x)| \leq v_n(x) - v_m(x) < \varepsilon$ si $n > m \geq p$, se deduce $\left| \sum_{k=m+1}^{n} \frac{(-1)^k x^{2k}}{(2k+5)!} \right| < \varepsilon$ si $n > m \geq p,\ x \in [-r,r]$, que son los resultados que hemos obtenido anteriormente.

Tal como en el estudio de las sucesiones de funciones nos interesamos en los problemas de la convergencia uniforme y la continuidad, derivación e integración, esto es, si $\sum_{n=1}^{\infty} f_n$ es una serie de funciones definidas en A que converge uniformemente a una función suma S, se tiene $S(x) = \sum_{n=1}^{\infty} f_n(x) \quad x \in A$. Los resultados precedentes obtenidos en esta sección y los de la sección anterior son aplicados a la sucesión de sumas parciales (S_n) con $S_n = \sum_{k=1}^{n} f_k \quad n = 1,2,\cdots$, y se obtienen los siguientes relativos a la continuidad, integrabilidad y derivabilidad. Así, si f_n es continua en $x_0 \in A,\ n = 1,2,\ldots$, se tiene

$$\lim_{x\to x_0} \sum_{n=1}^{\infty} f_n(x) = \sum_{n=1}^{\infty} \lim_{x\to x_0} f_n(x) = S(x_0).$$

Si f_n es integrable en $A = [a,b]$, $n = 1,2,\ldots$, se tiene

$$\int_a^b S(x)\,dx = \int_a^b \sum_{n=1}^{\infty} f_n(x)\,dx = \sum_{n=1}^{\infty} \int_a^b f_n(x)\,dx.$$

Si A es abierto, $x_0 \in A$ y $\sum_{n=1}^{\infty} f_n(x_0)$ converge, (f_n) es derivable en todo punto de A, y, $\sum_{n=1}^{\infty} \frac{df_n}{dx}$ converge uniformemente a una función g, entonces $\sum_{n=1}^{\infty} f_n$ converge uniformemente a $S = \sum_{n=1}^{\infty} f_n$, y, $\frac{dS}{dx}(x) = g(x) \quad x \in A$, es decir, se tiene

$$\frac{d}{dx}\left(\sum_{n=1}^{\infty} f_n(x)\right) = \sum_{n=1}^{\infty} \frac{df_n}{dx}(x) \quad x \in A.$$

Teorema 17 *Sea $A \subset \mathbb{R}$ con $A \neq \emptyset$, (f_n) una sucesión de funciones continuas en todo punto $x \in A$. Si $\sum_{k=1}^{\infty} f_k$ converge uniformemente a una función S definida en A, entonces S es continua en A, y*

$$\lim_{x\to x_0} \sum_{k=1}^{\infty} f_k(x) = \sum_{k=1}^{\infty} \lim_{x\to x_0} f_k(x) = S(x_0).$$

Teorema 18 *Sean* (f_n) *una sucesión de funciones integrables en* $[a,b]$, (S_n) *la sucesión de sumas parciales de la serie* $\sum_{n=1}^{\infty} f_n$, *esto es,* $S_n = \sum_{k=1}^{n} f_k \;\; n = 1, 2, \ldots$. *Si* (S_n) *converge uniformemente en* $[a,b]$ *a* $S = \sum_{k=0}^{\infty} f_k$, *entonces* S *es integrable, y se tiene*

$$\int_a^b \sum_{k=0}^{\infty} f_k(x)\,dx = \sum_{k=0}^{\infty} \int_a^b f_k(x)\,dx.$$

Para la convergencia uniforme y la derivación de series de funciones consideramos el siguiente ejemplo debido a Weierstrass. La serie $\sum_{k=0}^{\infty} \frac{\cos(3^n x)}{2^n} \quad x \in \mathbb{R}$ converge uniformemente en todo $\mathbb{R}$. Además cada función $u_n(x) = \frac{1}{2^n}\cos(3^n x) \quad n = 1, 2, \ldots$, es derivable en todo $\mathbb{R}$, por lo tanto continua en todo $\mathbb{R}$. Se define la función f como sigue:

$$f(x) = \sum_{k=0}^{\infty} \frac{\cos(3^n x)}{2^n} \quad x \in \mathbb{R}.$$

Resulta que f es continua en todo $\mathbb{R}$. Por otro lado,

$$\frac{du_n}{dx}(x) = -\left(\frac{3}{2}\right)^n \operatorname{sen}(3^n x) \quad n = 1, 2, \ldots, \text{ y,} \quad \left|\frac{du_n}{dx}(x)\right| = \left(\frac{3}{2}\right)^n |\operatorname{sen}(3^n x)| \leq \left(\frac{3}{2}\right)^n \underset{n\to\infty}{\longrightarrow} \infty.$$

No se cumple con la hipótesis de la prueba deWeierstrass. Si $x = 2k\pi \quad k \in \mathbb{Z}^+$, entonces

$$\frac{du_n}{dx}(2k\pi) = \operatorname{sen}(3^n \times 2k\pi) = 0 \quad n = 1, 2, \ldots,$$

y, $\dfrac{df}{dx}(2k\pi) = 0$. Para $x \neq 2k\pi$, la serie $\sum_{k=0}^{\infty} \left(\frac{3}{2}\right)^n \operatorname{sen}(3^n x)$ diverge, luego $\dfrac{df}{dx}(x)$ no existe.

3.3. Series de potencias.

Las series de potencias son series de funciones cuyas sumas parciales $S_n(x) \quad x \in \mathbb{R}$ y $n = 1, 2, \cdots$, son polinomios de grado n, y estos constituyen las funciones con las que se pueden calcularse valores numéricos en forma relativamente simple. Iniciamos esta sección con la convergencia puntual de las series de potencias que se convierte en la determinación del radio de convergencia. A continuación se trata las propiedades de las funciones representadas como series de potencias y relacionamos con la convergencia uniforme y la continuidad, integrabilidad y derivabilidad. Concluimos esta sección con la revisión de algunos resultados de una parte importante de las series de potencias que lo constituyen las series de Taylor.

3.3.1. Series de potencias.

Las series de potencias son series de la forma $\sum_{k=0}^{\infty} a_k (x - x_0)^k$, donde (a_k) es una sucesión numérica real, x,$x_0 \in \mathbb{R}$ con x_0 fijo. El cambio de variable $t = x - x_0$, conduce a la serie $\sum_{k=0}^{\infty} a_k t^k$, por lo que basta estudiar las series de potencias $\sum_{k=0}^{\infty} a_k x^k$. Además, toda serie de potencias $\sum_{k=0}^{\infty} a_k x^k$ converge por lo menos para $x = 0$.

Comenzamos con la convergencia puntual, la convergencia absoluta y la existencia del radio de convergencia de las series de potencias.

Teorema 19 *Sean* $\sum_{k=0}^{\infty} a_k x^k$ *una serie de potencias,* $\beta, \lambda \in \mathbb{R}$ *con* $\beta \neq 0,\ \lambda \neq 0$.

i) Si $\sum_{k=0}^{\infty} a_k \beta^k$ *converge, entonces* $\sum_{k=0}^{\infty} a_k x^k$ *converge absolutamente sobre* $]-|\beta|, |\beta|[$.

ii) Si $\sum_{k=0}^{\infty} a_k \lambda^k$ *diverge, entonces* $\sum_{k=0}^{\infty} a_k x^k$ *diverge para todo* $x \in \mathbb{R}$ *tal que* $|x| > |\lambda|$.

Radio de convergencia

A continuación se define el radio de convergencia R que es muy importante en el análisis de la convergencia de las series de potencias.

Definición 6 *Sea* $\sum_{k=0}^{\infty} a_k x^k$ *una serie de potencias. Se define el conjunto* A *como sigue:*

$$A = \left\{ x \in \mathbb{R} \mid \sum_{k=0}^{\infty} |a_k| \, |x|^k \ \ converge \right\}.$$

Se llama radio de convergencia de la serie $\sum_{k=0}^{\infty} a_k x^k$ *a un número real* $R \geq 0$ *o* $R = \infty$ *que se define como sigue:*

i) Si $A = \{0\}, \quad R = 0$.

ii) Si $A = \mathbb{R}, \quad R = \infty$.

iii) Si $A \neq \{0\}$ *y* $A \neq \mathbb{R}, \quad R = \operatorname{Sup}_{x \in A} |x| > 0$. *El intervalo* $]-R, R[$ *se llama intervalo de convergencia de la serie de potencias* $\sum_{k=0}^{\infty} a_k x^k$.

Teorema 20 *Sean* $\sum_{k=0}^{\infty} a_k x^k$ *una serie de potencias y* $R > 0$ *o* $R = \infty$.

i) Si $R = \infty$, *la serie de potencias converge absolutamente para todo* $x \in \mathbb{R}$.

ii) Si $R > 0$, *la serie de potencias converge absolutamente sobre* $]-R, R[$ *y diverge para* $x \in \mathbb{R}$ *con* $|x| > R$.

En la figura siguiente se ilustra el intervalo de convergencia $]-R, R[$ cuando $0 < R < \infty$, de la serie de potencias $\sum_{k=0}^{\infty} a_k x^k \quad x \in]-R, R[$.

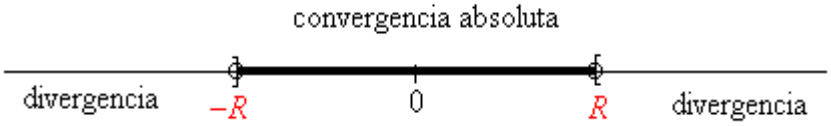


Figura 29

Cabe mencionar que para $x = -R$ o $x = R$, eventualmente se puede tener convergencia absoluta, convergencia condicional, o simplemente la serie puede ser divergente. Más adelante se exhibirá un caso concreto de esta situación.

Para la determinación del radio de convergencia se aplican usualmente los clásicos criterios del cociente

y la raíz Consideremos la serie de potencias $\sum_{k=0}^{\infty} a_k x^k$ y supongamos que $L = \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right|$. Para $x \neq 0$, apliquemos el criterio del cociente. Tenemos

$$\lim_{k\to\infty} \left|\frac{a_{k+1}x^{k+1}}{a_k x^k}\right| = \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| |x| = L\,|x|\,.$$

El radio de convergencia R se elige como sigue: $\begin{cases} i)\ \text{si}\ \ L = 0,\ R = \infty, \\ ii)\ \text{si}\ \ L > 0,\ R = \dfrac{1}{L}, \\ iii)\ \text{si}\ \ L = \infty,\ R = 0. \end{cases}$

Supongamos $L = \lim_{k\to\infty} |a_k|^{\frac{1}{k}}$. Apliquemos el criterio de la raíz:

$$\lim_{k\to\infty} \left(|a_k|\,|x|^k\right)^{\frac{1}{k}} = \lim_{k\to\infty} |x|\,|a_k|^{\frac{1}{k}} = |x|\,L.$$

El radio de convergencia R se elige en forma similar al caso precedente: $\begin{cases} i)\ \text{si}\ \ L = 0,\ R = \infty, \\ ii)\ \text{si}\ L > 0,\ R = \dfrac{1}{L}, \\ iii)\ \text{si}\ \ L = \infty,\ R = 0. \end{cases}$

Ejemplos

1. Consideremos la serie geométrica $\sum_{k=0}^{\infty} x^k$. Se demostró anteriormente que la serie converge si y solo si $|x| < 1$. En tal caso $\sum_{k=0}^{\infty} x^k = \dfrac{1}{1-x}$. Es claro que el radio de convergencia es $R = 1$. Los criterios del cociente y de la raíz confirman que el radio de convergencia es $R = 1$.

2. Más adelante se prueba que $\arctan(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}$, cuyo intervalo de convergencia es $]-1, 1[$. En este ejemplo, para $x = 1$, se obtiene el siguiente resultado: $\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} = \arctan(1) = \frac{\pi}{4}$, que muestra que la serie $\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1}$ converge condiciolnalmente a $\dfrac{\pi}{4}$.

3. La serie de potencias $\sum_{k=0}^{\infty} \dfrac{(-1)^k\, x^{2k}}{k!2^k}$ tiene como radio de convergencia $R = \infty$. En efecto, por el criterio del cociente se tiene

$$L_1 = \lim_{k\to\infty} \frac{1}{(k+1)!2^{k+1}} \times \left(k!2^k\right) = \lim_{k\to\infty} \frac{1}{2\,(k+1)} = 0,$$

luego $R = \infty$. He aquí algunas series numéricas absolutamente convergentes: para $x = 3$ se tiene $\sum_{k=0}^{\infty} \dfrac{(-1)^k\, 3^{2k}}{k!2^k} = e^{-4,5}$, para $x = \sqrt{5}$ se obtiene la serie numérica $\sum_{k=0}^{\infty} \dfrac{(-1)^k}{k!} \left(\dfrac{5}{2}\right)^k = e^{-2,5}$, para $x = \dfrac{\sqrt{2}}{2}$ se tiene $\sum_{k=0}^{\infty} \dfrac{(-1)^k}{k!4^k} = e^{-0,25}$. Se tiene $\sum_{k=0}^{\infty} \dfrac{(-1)^k\, x^{2k}}{k!2^k} = \exp(-\dfrac{x^2}{2}) \quad x \in \mathbb{R}$.

4. La serie de potencias $\sum_{k=0}^{\infty} \dfrac{(-1)^k}{k^2} x^{k+2}$ tiene como radio de convergencia $R = 1$, pués

$$L_2 = \lim_{k\to\infty} \left(\frac{1}{k^2}\right)^{\frac{1}{k}} = \lim_{k\to\infty} \frac{1}{k^{\frac{2}{k}}} = 1,$$

de donde $R = \dfrac{1}{L_2} = 1$. Observe que para $x = 1$ se tiene la serie $\sum_{k=0}^{\infty} \dfrac{(-1)^k}{k^2}$ que converge absolutamente,y para $x = -1$ la serie $\sum_{k=0}^{\infty} \dfrac{1}{k^2}$ también converge absolutamente.

Propiedades de las series de potencias.

Ahora nos interesamos en los problemas de la convergencia uniforme, la continuidad, la derivabilidad e integrabilidad de las funciones que se representan como series de potencias. Más precisamente, consideremos la serie de potencias $\sum_{k=0}^{\infty} a_k x^k$ y supongamos que el radio de convergencia es $R > 0$ o $R = \infty$.

Consideramos en caso $R > 0$. La serie converge sobre el intervalo $]-R, R[$, por lo tanto define una función f de $]-R, R[$ en $\mathbb{R}$ dada por $f(x) = \sum_{k=0}^{\infty} a_k x^k \quad x \in]-R, R[$. Analicemos la convergencia uniforme de la serie de potencias $\sum_{k=0}^{\infty} a_k x^k$ sobre $]-R, R[$. Se define $S_n(x) = \sum_{k=0}^{n} a_k x^k \quad x \in]-R, R[$, $n = 1, 2, \ldots$,y sea $0 < \lambda < R$. Apliquemos los resultados obtenidos anteriormente sobre la convergencia uniforme, continuidad, integrabilidad y derivabilidad de sucesiones de funciones reales.

Primeramente, $\lim_{k\to\infty} S_n(x) = \sum_{k=0}^{\infty} a_k x^k = f(x) \quad x \in]-R, R[$, y como $R > 0$, se tiene $\lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| > 0$, y por la definición del radio de convergencia se tiene $R = \dfrac{1}{\lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right|}$, de modo que $\lambda \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| = \frac{\lambda}{R} < 1$. Cada función S_n, $n = 1, 2, \ldots$, es continua, derivable e integrable sobre $[-\lambda, \lambda]$. Seguidamente, aplicamos la prueba de Weierstrass de la convergencia uniforme. Para $x \in [-\lambda, \lambda]$, se tiene $\left|a_k x^k\right| \leq |a_k| \lambda^k = M_k \quad k = 1, 2, \ldots$.La serie $\sum_{k=0}^{\infty} |a_k| \lambda^k$ converge, pués por el criterio de cociente se tiene

$$\lim_{k\to\infty} \left|\frac{a_{k+1}\lambda^{k+1}}{a_k \lambda^k}\right| = \lambda \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| = \frac{\lambda}{R} < 1.$$

La prueba de Weierstrass muestra que la serie $\sum_{k=0}^{\infty} a_k x^k$ converge uniformemente sobre $[-\lambda, \lambda]$. En consecuencia

$$a_k = \operatorname*{Sup}_{x\in[-\lambda,\lambda]} |S_n(x) - f(x)| = \operatorname*{Sup}_{x\in[-\lambda,\lambda]} \left|\sum_{k=0}^{\infty} a_k x^k\right| \underset{k\longrightarrow\infty}{\longrightarrow} 0,$$

es decir, (S_n) converge uniformemente a f sobre $[-\lambda, \lambda]$. Además, f es continua y por lo tanto integrable en $[-\lambda, \lambda]$. Resulta, para $x_0 \in [-\lambda, \lambda]$,

$$f(x_0) = \lim_{x\to x_0} f(x) = \lim_{x\to x_0} \sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} a \lim_{x\to x_0 k} x^k = \sum_{k=0}^{\infty} a_k x_0^k.$$

Para todo $t \in [-\lambda, \lambda]$,

$$\begin{aligned} \int_{-\lambda}^{t} f(x)\,dx &= \int_{-\lambda}^{t} \sum_{k=0}^{\infty} a_k x^k dx = \sum_{k=0}^{\infty} a_k \int_{-\lambda}^{t} x^k dx = \sum_{k=0}^{\infty} \frac{a_k}{k+1}\left(t^{k+1} - (-\lambda)^{k+1}\right) \\ &= \sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1} - \sum_{k=0}^{\infty} \frac{(-1)^{k+1} a_k}{k+1} \lambda^{k+1}, \end{aligned}$$

ya que las series $\sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1}$ y $\sum_{k=0}^{\infty} \frac{(-1)^{k+1} a_k}{k+1} \lambda^{k+1}$ son convergentes. Así, por el criterio del cociente, para todo $t \in [-\lambda, \lambda]$, $t \neq 0$, se tiene

$$\lim_{k\to\infty} \left|\frac{\frac{a_{k+1}}{k+2} t^{k+2}}{\frac{a_k}{k+1} t^{k+1}}\right| = |t| \lim_{k\to\infty} \left|\frac{(k+1)}{(k+2)} \frac{a_{k+1}}{a_k}\right| = |t| \lim_{k\to\infty} \left(\frac{k+1}{k+2}\right) \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| = |t| \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| \leq \lambda \frac{1}{R} < 1,$$

que muestra que la serie de potencias $\sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1}$ converge sobre $[-\lambda, \lambda]$. En consecuencia

$$\int_{-\lambda}^{t} f(x)\, dx = \sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1} - \sum_{k=0}^{\infty} \frac{(-1)^{k+1} a_k}{k+1} \lambda^{k+1} \quad t \in [-\lambda, \lambda],$$

está bien definida.

Veamos la derivabilidad de f. Tenemos

$$\frac{dS_n}{dx}(x) = \frac{d}{dx}\left(\sum_{k=0}^{n} a_k x^k\right) = \sum_{k=1}^{n} k a_k x^{k-1} \quad x \in [-\lambda, \lambda].$$

Verifiquemos que la serie $\sum_{k=1}^{\infty} k a_k x^{k-1}$ converge uniformemente sobre $[-\lambda, \lambda]$. Primeramente verifiquemos que $\sum_{k=1}^{\infty} k a_k x^{k-1}$ converge sobre $[-\lambda, \lambda]$. En efecto, por el criterio del cociente, se tiene para $x \in [-\lambda, \lambda]$, $x \neq 0$,

$$\lim_{k\to\infty} \left| \frac{(k+1) a_{k+1} x^k}{k a_k x^{k-1}} \right| = |x| \lim_{k\to\infty} \frac{(k+1)}{k} \lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| = |x| \lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| \leq \frac{\lambda}{R} < 1,$$

que prueba la convergencia de $\sum_{k=1}^{\infty} k a_k x^{k-1}$ $x \in [-\lambda, \lambda]$. A continuación verificamos que la convergencia es uniforme. Tenemos

$$\left| k a_k x^{k-1} \right| = k |a_k| |x|^{k-1} \leq k |a_k| \lambda^{k-1} \quad x \in [-\lambda, \lambda], \quad k = 1, 2, \ldots$$

y la serie numérica $\sum_{k=1}^{\infty} k |a_k| \lambda^{k-1}$ converge. Por la prueba de Weierstrass se concluye que la serie de potencias $\sum_{k=1}^{\infty} k a_k x^{k-1}$ converge uniformemente sobre $[-\lambda, \lambda]$. Adicionalmente $S_n(0) = 0 \quad n = 1, 2, \ldots$. Por el teorema de la convergencia uniforme y la derivación se deduce: $\frac{df}{dx}(x) = \sum_{k=1}^{\infty} k a_k x^{k-1} \quad x \in [-\lambda, \lambda]$.

Ejemplo

Si $|x| < 1$, se tiene $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$. El radio de convergencia de la serie $\sum_{k=0}^{\infty} x^k$ es $R = 1$. Si remplazamos x por $-x$ en la serie geométrica, obtenemos $\sum_{k=0}^{\infty} (-1)^k x^k = \frac{1}{1+x}$ cuyo radio de convergencia es también $R = 1$, o sea $\sum_{k=0}^{\infty} (-1)^k x^k = \frac{1}{1+x}$ si $|x| < 1$. Con este ejemplo construimos algunas funciones que se dan a continuación.

1. Para $0 < a < 1$, la serie $\sum_{k=0}^{\infty} a^k$ converge y $\sum_{k=0}^{\infty} a^k = \frac{1}{1+a}$, por la prueba de Weierstrass, la serie $\sum_{k=0}^{\infty} (-1)^k x^k = \frac{1}{1+x}$ con $|x| \leq a$, converge uniformemente. Aplicando el teorema de convergencia uniforme e integración, deducimos el resultado siguiente:

$$\ln(1+x) = \int_0^x \frac{dt}{1+t} = \int_0^x \sum_{k=0}^{\infty} (-1)^k t^k dt = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} x^{k+1} \quad |x| < 1,$$

cuyo radio de convergencia es $R = 1$. Además, para $x = 1$ se obtiene la serie numérica $\sum_{k=0}^{\infty} \frac{(-1)^k}{k+1}$ la misma que converge condicionalmente a $\ln(2)$.

2. Si en el resultado siguiente $\sum_{k=0}^{\infty}(-1)^k x^k = \frac{1}{1+x}$ se remplaza x por x^2, obtenemos la serie de potencias $\sum_{k=0}^{\infty}(-1)^k x^{2k} = \frac{1}{1+x^2}$, y nuevamente, aplicamos el teorema de la convergencia uniforme y la integración sobre el intervalo $[0,x] \subset]-1,1[$ si $x \geq 0$, y, $[x,0] \subset]-1,1[$ si $x \leq 0$. Obtenemos

$$\arctan(x) = \int_0^x \frac{dt}{1+t^2} = \int_0^x \sum_{k=0}^{\infty}(-1)^k t^{2k} dt = \sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}x^{2k+1},$$

resultado que es válido para $|x|<1$, y, para $x=1$, se obtiene el siguiente resultado: $\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1} = \arctan(1) = \frac{\pi}{4}$.

3. Por otro lado, del teorema de la convergencia uniforme y la integración aplicado a la serie $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ en un intervalo $[-a,a] \subset]-1,1[$, obtenemos

$$-\ln(1-x) = \int_0^x \frac{dt}{1-x} = \int_0^x \sum_{k=0}^{\infty} t^k dt = \sum_{k=0}^{\infty}\frac{x^{k+1}}{k+1} \quad \text{si} \quad |x|<1.$$

Por lo tanto,

$$\ln\left(\frac{1+x}{1-x}\right) = \ln(1+x) - \ln(1-x) = \sum_{k=0}^{\infty}\frac{(-1)^k x^{k+1}}{k+1} + \sum_{k=0}^{\infty}\frac{x^{k+1}}{k+1} = 2\sum_{k=0}^{\infty}\frac{x^{2k+1}}{2k+1},$$

o sea $\ln\left(\frac{1+x}{1-x}\right) = 2\sum_{k=0}^{\infty}\frac{x^{2k+1}}{2k+1}$ para $|x|<1$.

4. Puesto que $\frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{1}{(1-x)^2}$ $\forall x \in]-1,1[$, y por otro lado $\frac{d}{dx}\sum_{k=0}^{\infty} x^k = \sum_{k=1}^{\infty} k x^{k-1}$. Como $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ para $x \in]-1,1[$, por el teorema de convergencia uniforme y derivación, tenemos

$$\frac{1}{(1-x)^2} = \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{d}{dx}\left(\sum_{k=0}^{\infty} x^k\right) = \sum_{k=1}^{\infty} k x^{k-1} \text{ para } |x|<1,$$

luego, $\sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}$ $|x|<1$.

5. Procediendo en forma similar a la del ejemplo precedente, la función definida como $\frac{1}{1+x^2} = \sum_{k=0}^{\infty}(-1)^k x^{2k}$ es válida para $|x|<1$. Derivando miembro a miembro, se obtiene

$$-\frac{2x}{(1+x^2)^2} = \frac{d}{dx}\left(\frac{1}{1+x^2}\right) = \frac{d}{dx}\left(\sum_{k=0}^{\infty}(-1)^k x^{2k}\right) = \sum_{k=1}^{\infty}(-1)^k 2k x^{2k-1}.$$

Así, $\sum_{k=1}^{\infty}(-1)^k 2k x^{2k-1} = -\frac{2x}{(1+x^2)^2}$ para $|x|<1$.

Consideremos el caso $R=\infty$. La serie de potencias $\sum_{k=0}^{\infty} a_k x^k$ converge en todo $\mathbb{R}$ y se define una función g dada por $g(x) = \sum_{k=0}^{\infty} a_k x^k$ $x \in \mathbb{R}$. Nuevamente la sucesión de funciones (S_n) definida como $S_n(x) = \sum_{k=0}^{n} a_k x^k$ $x \in \mathbb{R}$, $n = 1,2,\ldots$, es continua, derivable e integrable sobre todo intervalo $[a,b]$ de $\mathbb{R}$. Analicemos la convergencia uniforme de (S_n) en $[0,\lambda]$ con $\lambda > 0$. Como $R=\infty$, se tiene

$\lim_{k\to\infty}\left|\frac{a_{k+1}}{a_k}\right| = 0$ y en consecuencia, $\left|a_k x^k\right| \leq |a_k|\,\lambda^k \quad k = 1,2,\ldots,\ x \in [0,\lambda]$, y la serie numérica $\sum_{k=0}^{\infty} |a_k|\,\lambda^k$ converge, pués

$$\lim_{k\to\infty}\left|\frac{a_{k+1}\lambda^{k+1}}{a_k\lambda^k}\right| = \lambda \lim_{k\to\infty}\left|\frac{a_{k+1}}{a_k}\right| = 0.$$

Por la prueba de Weierstrass, la serie $\sum_{k=0}^{\infty} a_k x^k$ converge uniformemente sobre $|0,\lambda|$. Luego g es continua, derivable e integrable, y se tienen los siguientes resultados:

$$\begin{aligned} g(x_0) &= \lim_{x\to x_0} g(x) = \sum_{k=0}^{\infty} a_k x_0^k \quad \text{para } x_0 \in [0,\lambda], \\ \int_0^t g(x)\,dx &= \int_0^t \sum_{k=0}^{\infty} a_k x^k dx = \sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1} \quad \text{para } t \in [0,\lambda], \\ \frac{dg}{dx}(x) &= \sum_{k=1}^{\infty} k a_k x^{k-1} \quad x \in [0,\lambda]. \end{aligned}$$

La convergencia uniforme de (S_n) sobre $[a,b]$ se deduce inmediatamente de la convergencia uniforme sobre $[0,\lambda]$ con $\lambda = Max\{|a|,|b|\}$.

Ejemplo

Considérese la serie de potencias $\sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+1)^2} x^k \quad x \in \mathbb{R}$. Determinemos el radio de convergencia R. Para el efecto, aplicamos el criterio del cociente:

$$\lim_{k\to\infty}\left|\frac{a_{k+1}}{a_k}\right| = \lim_{k\to\infty}\left[\frac{1}{(k+1)!\,(k+2)^2} k!\,(k+1)^2\right] = \lim_{k\to\infty}\frac{k+1}{(k+2)^2} = 0,$$

entonces $R = \infty$, esto es, la serie converge uniformemente en todo $\mathbb{R}$. Se define

$$f(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(k+1)^2} x^k \quad x \in \mathbb{R}.$$

Calculemos $\frac{df}{dx}(x)$ e $\int_0^t f(x)\,dx$ para $x,t \in \mathbb{R}$. Para todo $\lambda > 0$, se tiene

$$\begin{aligned} \frac{df}{dx}(x) &= \frac{d}{dx}\left(\sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(k+1)^2} x^k\right) = \sum_{k=1}^{\infty} \frac{(-1)^k\,x^{k-1}}{(k-1)!\,(k+1)^2} = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}\,x^k}{k!\,(k+1)^2} \quad x \in [-\lambda,\lambda], \\ \int_0^t f(x)\,dx &= \int_0^t \sum_{k=0}^{\infty} \frac{(-1)^k\,x^k}{k!\,(k+1)^2} dx = \sum_{k=0}^{\infty} \int_0^t \frac{(-1)^k}{k!\,(k+1)^2} x^k dx = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!\,(k+1)^2} t^{k+1} \quad t \in [-\lambda,\lambda]. \end{aligned}$$

Calculemos una aproximación de $f\left(\frac{1}{2}\right)$ con una precisión $\varepsilon = 10^{-3}$. Tenemos $f\left(\frac{1}{2}\right) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(k+1)^2\,2^k}$, y sea $a_k = \frac{1}{k!\,(k+1)^2\,2^k} \quad k = 0,1,\ldots$.La serie $\sum_{k=1}^{\infty} \frac{1}{k\,(k+1)}$ converge a 1. Tenemos $\sum_{k=1}^{\infty} \frac{1}{k\,(k+1)} = 1$, y sea $b_k = \frac{1}{k\,(k+1)} \quad k = 1,2,\ldots$.Por el criterio de comparación, se tiene

$$\frac{a_k}{b_k} = \frac{k\,(k+1)}{k!\,(k+1)^2\,2^k} = \frac{k}{(k+1)!2^k} < 10^{-3} \quad \text{si } k \geq 6.$$

Entonces

$$\sum_{k=7}^{\infty} a_k < \sum_{k=7}^{\infty} 10^{-3} b_k = 10^{-3} \sum_{k=7}^{\infty} b_k < 10^{-3} \sum_{k=1}^{\infty} b_k = 10^{-3}.$$

Luego, si $f_7\left(\frac{1}{2}\right)$ denota una aproximación de $f\left(\frac{1}{2}\right)$ con una precisión ε, se tiene

$$\begin{aligned} f_7\left(\frac{1}{2}\right) &= \sum_{k=0}^{7} \frac{(-1)^k}{k!\,(k+1)^2\,2^k} \\ &= 1 - \frac{1}{1\times 4\times 2} + \frac{1}{2\times 9\times 4} - \frac{1}{6\times 16\times 8} + \frac{1}{24\times 25\times 16} - \frac{1}{120\times 36\times 32} \\ &\quad + \frac{1}{720\times 49\times 64} - \frac{1}{5040\times 64\times 128} \\ &= 1 + \frac{1}{2\times 9\times 4} + \frac{1}{24\times 25\times 16} + \frac{1}{720\times 49\times 64} \\ &\quad - \left(\frac{1}{1\times 4\times 2} + \frac{1}{6\times 16\times 8} + \frac{1}{120\times 36\times 32} + \frac{1}{5040\times 64\times 128}\right) \\ &\simeq 0{,}8886. \end{aligned}$$

Así, $\left|f\left(\frac{1}{2}\right) - f_7\left(\frac{1}{2}\right)\right| < \varepsilon = 10^{-3}$.

3.3.2. Series de Taylor.

Las series de Taylor son series de potencias de la forma $\sum_{k=0}^{\infty} a_k(x-a)^k$, donde los coeficientes a_k están definidos como $a_k = \frac{f^{(k)}(a)}{k!}$ $k = 0, 1, \ldots$, con f una función que posee derivadas de todos los órdenes en $x = a$. Se dice que la serie de Taylor es generada por f en $x = a$. Nos interesamos primeramente en las funciones f que se representan como series de potencias, esto es, $f(x) = \sum_{k=0}^{\infty} a_k x^k$ y determínamos los coeficientes a_k $k = 1, 2, \cdots$; a continuación nos interesa las funciones f que se representan como series de potencias. Asumiremos que $a = 0$, si $a \neq 0$, el cambio de variable $t = x - a$ conduce al caso anterior.

Consideremos la serie de potencias $\sum_{k=0}^{\infty} a_k x^k$ cuyo radio de convergencia es $R > 0$. Se define la función f de $]-R, R[$ en $\mathbb{R}$ como $f(x) = \sum_{k=0}^{\infty} a_k x^k$ $x \in]-R, R[$. Determinemos los coeficientes a_k $k = 1, 2, \cdots$ Se tiene $f(0) = a_0$. Sea $0 < \lambda < R$. Hemos visto que $f'(x)$ existe y es continua en todo punto $x \in [-\lambda, \lambda]$, y $f'(x) = \sum_{k=1}^{\infty} k a_k x^{k-1}$ luego $f'(0) = a_1$. La serie $\sum_{k=1}^{\infty} k a_k x^{k-1}$ converge uniformemente en el intervalo $[-\lambda, \lambda]$. A continuación calculamos la derivada segunda de la función f, tenemos $f''(x) = \sum_{k=2}^{\infty} k(k-1) a_k x^{k-2}$ y de esta obtenemos $f''(0) = 2!a_2$. Nuevamente la serie $\sum_{k=2}^{\infty} k(k-1) a_k x^{k-2}$ converge uniformemente en el intervalo $[-\lambda, \lambda]$, entonces $f'''(x) = \sum_{k=3}^{\infty} k(k-1)(k-2) a_k x^{k-3}$ y $f'''(0) = 3!a_3$. Continuando con este proceso, obtenemos $f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1) \times \ldots \times (k-n+1) a_k x^{k-n}$, y $f^n(0) = n!a_n$. Resulta $a_n = \frac{f^{(n)}(0)}{n!}$ $n = 0, 1, 2, \ldots$, y la función f queda representada como la serie de potencias siguiente $f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$ $x \in]-R, R[$, que se conoce como serie de Taylor de f en un entorno de $x = 0$. Es claro que $f \in C^{\infty}(]-R, R[)$.

Cuando $R = \infty$, la situación es muy similar al caso que acabamos de analizar, esto es, si $f(x) = \sum_{k=0}^{\infty} a_k x^k$ $x \in \mathbb{R}$, entonces $f \in C^{\infty}(\mathbb{R})$, y esta función f se representa como la serie de potencias siguiente:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k \quad x \in \mathbb{R}.$$

Sea $R > 0$ y $f \in C^{\infty}(]-R, R[)$. Consideramos el problema siguiente: expresar f (siempre que sea posible) como una serie de Taylor en el entorno de 0; es decir, $f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$ $x \in]-R, R[$. Si

$f^{(k)}(0) = 0 \ \ \forall k \in \mathbb{Z}^+$, entonces $f(x) = 0 \quad \forall x \in]-R, R[$ y la función f no se representa como una serie de Taylor. Esto se evidencia con el siguiente ejemplo: $f(x) = \begin{cases} \exp(-\frac{1}{x^2}), & \text{si } x \neq 0, \\ 0, & \text{si } x = 0. \end{cases}$ Se demuestra que $f^{(k)}(0) = 0$ existe para $k = 1, 2, \ldots$, luego la función f que se desea representar como una serie de potencias es cero, pero la función f es nula solo en $x = 0$. Se concluye que f no se representa como una serie de potencias.

Sea $f \in C^\infty(]-R, R[)$, el polinomio de Taylor de f en un entorno de $x = 0$ se expresa como

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k + E_n(x) \quad \forall x \in]-R, R[,$$

donde $E_n(x)$ denota el error de aproximación de $f(x)$ en $x = 0$, que se expresa como

$$E_n(x) = \frac{1}{n!} \int_0^x (x-t)^n f^{(n+1)}(t) dt \quad x \in]-R, R[.$$

Si $f^{(n+1)}$ es acotada en un entorno de 0, $|E_n(x)| \leq \frac{M_n}{(n+1)!} |x|^{n+1}$, donde $M_n = \underset{t \in [a-r,\, a+r]}{\text{Sup}} \left| f^{(n+1)}(t) \right|$ con $r > 0$.

Teorema 21 *Sean $f \in C^\infty(]-R, R[)$, $0 < r < R$. Si existe una constante $M > 0$ tal que $| f^{(n+1)}(x) | \leq M^n$, $n = 1, 2, \ldots, x \in [-r, r]$, entonces la serie de Taylor generada por f converge hacia $f(x)$.*

Ejemplos

1. La serie de Taylor $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ $x \in \mathbb{R}$ es generada por e^x en $x = 0$. Esta serie converge absolutamente en todo punto $x \in \mathbb{R}$. El radio de convergencia es $R = \infty$. Se tiene $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ $\forall x \in \mathbb{R}$. Observe las siguientes series numéricas, todas son absolutamente convergentes:

$$e^{-\frac{1}{2}} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! 2^k}, \qquad e^2 = \sum_{k=0}^{\infty} \frac{2^k}{k!}, \qquad e^{\sqrt{2}} = \sum_{k=0}^{\infty} \frac{2^{\frac{k}{2}}}{k!}.$$

Puesto que $\frac{d}{dx}(e^x) = e^x \quad \forall x \in \mathbb{R}$ y como la serie de potencias $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ converge uniformemente sobre todo intervalo cerrado y acotado $[a, b]$ de $\mathbb{R}$, se sigue que

$$\frac{d}{dx}(e^x) = \frac{d}{dx}\left(\sum_{k=0}^{\infty} \frac{x^k}{k!} \right) = \sum_{k=1}^{\infty} \frac{k x^{k-1}}{k!} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x \quad \forall x \in \mathbb{R}.$$

Por otro lado, si se remplaza x por $-x$ en el desarrollo de Taylor de e^x, se tiene la serie de Taylor siguiente: $e^{-x} = \sum_{k=0}^{\infty} \frac{(-1)^k x^k}{k!}$ $\forall x \in \mathbb{R}$. De manera similar, si se remplaza x por $-x^2$ en el desarrollo de e^x, se tiene la serie siguiente: $e^{-x^2} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{k!}$ $\forall x \in \mathbb{R}$. Igualmente $e^{\sqrt{x}} = \sum_{k=0}^{\infty} \frac{x^{\frac{k}{2}}}{k!}$ $\forall x \in [0, \infty[$, $e^{-\frac{x^3}{3}} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{3k}}{k! 3^k}$ $\forall x \in \mathbb{R}$.

La integral de e^{-x^2} no puede calcularse con funciones elementales, sin embargo, si utilizamos el desarrollo de Taylor de dicha función, y los resultados de la convergencia uniforme y la integración, obtenemos la siguiente serie de potencias:

$$\int_0^t e^{-x^2} dx = \int_0^t \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{k!} dx = \sum_{k=0}^{\infty} \int_0^t \frac{(-1)^k x^{2k}}{k!} dx = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k+1}}{k!(2k+1)} \quad \forall t \in [0, \infty[,$$

que es absolutamente convergente. En la siguiente sección mostramos como obtener valores aproximados de series de potencias, particularmente de la función error que se estudia en estadística y en ecuaciones diferenciales ordinarias y en derivadas parciales.

2. La serie de Taylor $\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$ es generada por $\text{sen}(x)$ en un entorno de $x = 0$. Esta serie es absolutamente convergente para todo $x \in \mathbb{R}$ y su radio de convergencia es $R = \infty$. Tenemos

$$\text{sen}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}, \quad x \in \mathbb{R}.$$

De manera similar, la serie de Taylor de $\cos(x)$ en un entorno de $x = 0$, es la siguiente:

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} \quad x \in \mathbb{R}.$$

que converge absolutamente en todo $\mathbb{R}$, y es uniformemente convergente en todo intervalo cerrado y acotado.

Puesto que $(\text{sen}(x))' = \cos(x) \quad \forall x \in \mathbb{R}$, aplicando el resultado de la convergencia uniforme y la derivación para sucesiones de funciones, y, tomando en consideración la convergencia absoluta de dichas series, se sigue que

$$(\text{sen}(x))' = \frac{d}{dx}\left(\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}\right) = \sum_{k=0}^{\infty} \frac{(-1)^k (2k+1) x^{2k}}{(2k+1)!} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} = \cos(x) \quad \forall x \in \mathbb{R}.$$

Por otro lado, si se remplaza x por x^2 en los desarrollos de Taylor de $\text{sen}(x)$ y $\cos(x)$, se obtienen las series de potencias siguientes:

$$\text{sen}(x^2) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{4k+2}}{(2k+1)!}, \quad \text{y} \qquad \cos(x^2) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{4k}}{(2k)!} \quad x \in \mathbb{R}.$$

Igualmente, el desarrollo en serie de potencias de la función real dada por $g(x) = x\,\text{sen}(\sqrt{x})$ $\forall x \in [0, \infty[$, se obtiene del desarrollo de Taylor de $\text{sen}(x)$. Tenemos

$$g(x) = x\,\text{sen}(\sqrt{x}) = x \sum_{k=0}^{\infty} \frac{(-1)^k x^{\frac{2k+1}{2}}}{(2k+1)!} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{\frac{2k+3}{2}}}{(2k+1)!} \quad \forall x \in [0, \infty[.$$

Esta serie de potencias converge absolutamente para todo $x \geq 0$. Además, esta serie converge uniformemente sobre todo intervalo cerrado y acotado $[a, b]$ de $[0, \infty[$, particularmente sobre $[0, t]$; y, si se aplica el teorema de la convergencia uniforme y la integración se obtiene la siguiente serie de potencias absolutamente convergente sobre el intervalo $[0, \infty[$:

$$\int_0^t x\,\text{sen}(\sqrt{x})dx = \int_0^t \sum_{k=0}^{\infty} \frac{(-1)^k x^{\frac{2k+3}{2}}}{(2k+1)!} dx = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} \int_0^t x^{\frac{2k+3}{2}} dx = 2t^{\frac{5}{2}} \sum_{k=0}^{\infty} \frac{(-1)^k t^k}{(2k+1)!(2k+5)}.$$

3.4. Aproximación numérica de series de potencias.

Sean $n \in \mathbb{N}$, $a_k \in \mathbb{R}$ con $k = 0,\ 1, \ldots,\ n$. Como ya se ha dicho anteriormente una función real P de la forma $P(x) = \sum_{k=0}^{n} a_k x^k \quad x \in \mathbb{R}$ se llama función polinomial o simplemete polinomio real P. En la práctica interesan los polinomios de grado mayor o igual que 1. Del punto de vista numérico, son estas funciones las más simples de evaluarse.

Por otro lado algunas funciones como $\exp(x)$, con $x \in \mathbb{R}$, solo conocemos sus valores para algunos puntos $x \in \mathbb{R}$. En muchas situaciones nos es difícil calcular valores de estas funciones fuera de esos datos conocidos

x. La idea fundamental es aproximar $\exp(x)$ mediante polinomios elegidos de modo que el error es un punto dado $x \in \mathbb{R}$, sea tan pequeño como se quiera. Estos polinomios son los denominados polinomios de Taylor de $\exp(x)$. Estos polinomios fueron utilizados para la construcción de algoritmos que actualmente se usan en las calculadoras de bosillo. Otros ejemplos similares a los de la función exponencial son, por ejemplo la función seno, coseno, función error, distribución normal, funciones elípticas, funciones de Bessel, etc.

En general, si f es una función que posee derivadas hasta el orden $n+1$ inclusive en $x = a$ el polinomio de Taylor de f en un entorno de a se escribe como

$$P(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k,$$

y $f(x) = P(x) + E_n(x)$, donde $E_n(x)$ denota el error de aproximación de $f(x)$ mediante $P(x)$, dado por

$$E_n(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)dt.$$

Si $f^{(n+1)}$ es acotada en un entorno de $x = a$, se tiene

$$|E_n(x)| \leq \frac{M}{(n+1)!}|x-a|^{n+1} \underset{n\longrightarrow\infty}{\longrightarrow} 0 \quad \forall x \in [-r+a, r+a].$$

donde $M = \underset{t\in[a-r,\, a+r]}{\text{Sup}} \left|f^{(n+1)}(t)\right|$ con $r > 0$. La serie de Taylor de la función f en un entorno del punto $x = a$ se define como $\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$, y se desea calcular valores de $f(x)$.

En muy pocos casos se puede calcular $f(x)$ exactamente, en la generalidad debemos recurrir a cálculos aproximados.

Sea $\sum_{k=0}^{\infty} a_k x^k$ una serie de potencias que converge en el intervalo $]-R, R[$, donde $R > 0$ designa el radio de convergencia de la serie. Se define la función f de $]-R, R[$ en $\mathbb{R}$ como $f(x) = \sum_{k=0}^{\infty} a_k x^k \quad x \in]-R, R[$. Sea $0 < r < R$, suponemos que la serie de potencias converge uniformemente el intervalo $[-r, r]$ y sea $x \in [-r, r]$, queremos calcular un valor aproximado de $f(x)$ con una precisión $\varepsilon > 0$.

Los cálculos y los resultados que se obtienen en una calculadora de bolsillo o en computador personal, son en general con 10^{-9}, 10^{-12}, 10^{-16}, 10^{-25} cifras de precisión, por esta razón, las estimaciones y cálculos que realizaremos en lo sucesivo con series de potencias son con precisiones como las citadas.

Diremos que una serie de potencias es rápidamente convergente si para un número razonable de términos de la sucesión de sumas parciales se alcanza la exactitud y precisión establecida $\epsilon > 0$ (con fines prácticos $\epsilon = 10^{-9}$, 10^{-10}, etc.).

Consideramos como número razonable de términos $n < 100$ para ϵ del orden 10^{-16}. Es claro que si se aumenta considerablemente la precisión a alcanzar, se requerirán de un número de términos mucho mayor, si por ejemplo ϵ es del orden 10^{-40}, el concepto de número razonable de términos cambiará. Para alcanzar precisiones muy altas del orden 10^{-1000}, 10^{-2000}, etc. se requieren de la elaboración de algoritmos y programas especiales de cálculo que no lo abordaremos en este libro.

Diremos que la serie de potencias converge lentamente si el número de términos de la serie que se requieren para alcanzar la exactitud y precisión deseadas es grande. Lastimosamente, la acumulación de los errores debidos al truncamiento y redondeo influenciarán seriamente y modificarán los resultados. Esto obliga a enfrentar el problema con otro enfoque, es decir, buscar otras representaciones en series de potencias que que converjan rápidamente o en su defecto, obtener la mayor información posible de las propiedades de las funciones que puedan ser aplicadas para simplificar los cálculos aproximados. En este capítulo no utilizaremos los polinomios de Chebyshev y los de Legendre para reducir el número de términos a utilizar.

Con fines prácticos suponemos que ϵ es del orden 10^{-16} o mayor, particularmente 10^{-10}, y para esta precisión supondremos que la serie de potencias es rápidamente convergente en todo el intervalo $[-r, r]$. Ilutremos en este caso, el procedimiento a seguir.

Primeramente, elegimos la serie numérica $\sum_{k=1}^{\infty} b_k$ tal que $b_k > 0 \quad k = 1, 2, \cdots$, y que tiene como suma 1, esto es $\sum_{k=1}^{\infty} b_k = 1$. Denotamos con $S_m(x) = \sum_{k=0}^{m} a_k x^k$, donde $m \in \mathbb{Z}^+$ debe determinarse como el más pequeño entero positivo para el que se verifica la condición siguiente:

$$\mid f(x) - S_m(x) \mid = \left| \sum_{k=m+1}^{\infty} a_k x^k \right| < \varepsilon.$$

Sea $c_k = \dfrac{\mid a_k \mid r^k}{b_k} \quad k = 1, 2, \cdots$, y suponemos que $\lim_{k \longrightarrow \infty} c_k = 0$, esto significa que la serie $\sum_{k=0}^{\infty} \mid a_k \mid r^k$ converge más rápidamente que la serie elegida $\sum_{k=1}^{\infty} b_k$. De la hipótesis $\lim_{k \longrightarrow \infty} c_k = 0$, se sigue que existe $m \in \mathbb{Z}^+$ tal que $\left| \dfrac{a_k r^k}{b_k} \right| < \varepsilon$ si $k \geq m$. Luego $\mid a_k \mid r^k < \varepsilon b_k \quad k \geq m$, de donde

$$\mid f(x) - S_m(x) \mid = \left| \sum_{k=m+1}^{\infty} a_k x^k \right| \leq \sum_{k=m+1}^{\infty} \mid a_k \mid r^k \leq \sum_{k=m+1}^{\infty} \varepsilon b_k \leq \varepsilon \sum_{k=1}^{\infty} b_k = \varepsilon \quad \text{si } k \geq m,$$

que muestra que $S_m(x)$ aproxima a $f(x)$ con una precisión $\varepsilon > 0$. Obviamente el entero m depende de la serie $\sum_{k=1}^{\infty} b_k = 1$ con la que se compara. Algunas de las series usadas son por ejemplo $\sum_{k=1}^{\infty} \dfrac{1}{k(k+1)} = 1$, $\sum_{k=1}^{\infty} \dfrac{1}{2^k} = 1$. La primera converge lentamente, mientras que la segunda converge mucho más rápidamente que la primera.

Ejemplos

1. Consideremos la serie de potencias $\sum_{k=0}^{\infty} \dfrac{x^k}{(2k)!\,(3k+1)}$. Determinemos el radio de convergencia. Aplicando el criterio del cociente resulta para $x \neq 0$,

$$\frac{|x|^{k+1}}{(2\,(k+1))!\,(3k+4)} \frac{(2k)!\,(3k+1)}{|x|^k} = \frac{3k+1}{(2k+2)\,(2k+1)\,(3k+4)} |x| \underset{k \to \infty}{\longrightarrow} 0,$$

es decir que la serie converge absolutamente en todo $\mathbb{R}$. Definimos la función f como sigue:

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{(2k)!\,(3k+1)} x^k \quad x \in \mathbb{R}.$$

Queremos visualizar la gráfica de f en el intervalo $[-5, 10]$. Lastimosamente nos es difícil calcular cada $f(x) \quad x \in [-5, 10]$ y lo haremos en forma aproximada.

Sea $\varepsilon = 10^{-3}$. Determinemos $m \in \mathbb{Z}^+$ el más pequeño posible tal que $S_m(x) = \sum_{k=0}^{m} \dfrac{1}{(2k)!\,(3k+1)} x^k$, y,

$$|f(x) - S_m(x)| = \left| \sum_{k=m+1}^{\infty} \frac{1}{(2k)!\,(3k+1)} x^k \right| \leq \sum_{k=m+1}^{\infty} \frac{10^k}{(2k)!\,(3k+1)} < \varepsilon.$$

Para determinar m aplicamos el criterio de comparación con la serie $\sum_{k=1}^{\infty} \dfrac{1}{k\,(k+1)} = 1$. Ponemos $a_k = \dfrac{10^k}{(2k)!\,(3k+1)}$, $b_k = \dfrac{1}{k\,(k+1)} \quad k = 1, 2, \ldots$, y determinamos $k \in \mathbb{Z}^+$ tales que

$$\frac{a_k}{b_k} = \frac{k\,(k+1)\,10^k}{(2k)!\,(3k+1)} < \varepsilon = 10^{-3}.$$

Para $k = 5$ se tiene $\dfrac{a_5}{b_5} \simeq 5{,}167 \times 10^{-2}$; para $k = 6$ se tiene $\dfrac{a_6}{b_6} \simeq 4{,}615 \times 10^{-3}$; para $k = 7$, $\dfrac{a_7}{b_7} \simeq 2{,}919 \times 10^{-4} < 10^{-3}$. Elegimos $m = 7$ y $S_7(x) = \sum_{k=0}^{7} \dfrac{x^k}{(2k)!\,(3k+1)} \quad x \in [-5, 10]$.

Para $-5 \leq x < 0$, $S_7(x)$ se escribe como:

$$\begin{aligned} S_7(x) &= 1 + \frac{x^2}{4! \times 7} + \frac{x^4}{8! \times 13} + \frac{x^6}{12! \times 19} + \frac{x}{2! \times 4} + \frac{x^3}{6! \times 10} + \frac{x^5}{10! \times 16} + \frac{x^7}{14! \times 22} \\ &= 1 + \frac{x^2}{24}\left(\frac{1}{7} + \frac{x^2}{1680}\left(\frac{1}{13} + \frac{x^2}{225720}\right)\right) + \frac{x}{2}\left(\frac{1}{4} + \frac{x^2}{360}\left(\frac{1}{10} + \frac{x^2}{5040}\left(\frac{1}{16} + \frac{x^2}{528528}\right)\right)\right). \end{aligned}$$

Ponemos $y = x^2$, entonces $S_7(x)$ se escribe como sigue:

$$S_7(x) = 1 + \frac{y}{24}\left(\frac{1}{7} + \frac{y}{1680}\left(\frac{1}{13} + \frac{y}{225720}\right)\right) + \frac{x}{2}\left(\frac{1}{4} + \frac{y}{360}\left(\frac{1}{10} + \frac{y}{5040}\left(\frac{1}{16} + \frac{y}{528528}\right)\right)\right).$$

Ponemos

$$\begin{aligned} y_1 &= 1 + \frac{y}{24}\left(\frac{1}{7} + \frac{y}{1680}\left(\frac{1}{13} + \frac{y}{225720}\right)\right), \\ y_2 &= \frac{x}{2}\left(\frac{1}{4} + \frac{y}{360}\left(\frac{1}{10} + \frac{y}{5040}\left(\frac{1}{16} + \frac{y}{528528}\right)\right)\right), \end{aligned}$$

luego $S_7(x) = y_1 + y_2$.

Para $0 \leq x \leq 10$, $S_7(x)$ se expresa en forma explícita como

$$\begin{aligned} S_7(x) &= 1 + \frac{x}{2! \times 4} + \frac{x^2}{4! \times 7} + \frac{x^3}{6! \times 10} + \frac{x^4}{8! \times 13} + \frac{x^5}{10! \times 16} + \frac{x^6}{12! \times 19} + \frac{x^7}{14! \times 22} \\ &= 1 + \frac{x}{2}\left(\frac{1}{4} + \frac{x}{12}\left(\frac{1}{7} + \frac{x}{30}\left(\frac{1}{10} + \frac{x}{56}\left(\frac{1}{13} + \frac{x}{90}\left(\frac{1}{16} + \frac{x}{132}\left(\frac{1}{19} + \frac{x}{4004}\right)\right)\right)\right)\right)\right). \end{aligned}$$

Sea $m \in \mathbb{Z}^+$, $h = \dfrac{10 - (-5)}{m} = \dfrac{15}{m}$ y $x_j = -5 + jh \quad j = 0, 1, \ldots, m$ puntos igualmente espaciados del intervalo $[-5, 10]$. El algoritmo para el cálculo aproximado de $S_7(x_j) \quad j = 0, 1, \ldots, m$, con una precisión $\varepsilon = 10^{-3}$ se presenta a continuación.

Algoritmo

Datos de entrada: $m \in \mathbb{Z}^+$.

Datos de salida: x_j, $S_7(x_j) \quad j = 0, 1, \ldots, m$.

1. Leer m.
2. $h = \dfrac{15}{m}$.
3. Para $j = 0, 1, \ldots, m$

   $x_j = -5 + jh$

   Si $x_j < 0$

   $y = x_j * x_j$

   $y_1 = 1 + \dfrac{y}{24}\left(\dfrac{1}{7} + \dfrac{y}{1680}\left(\dfrac{1}{13} + \dfrac{y}{225720}\right)\right)$

   $y_2 = \dfrac{x}{2}\left(\dfrac{1}{4} + \dfrac{y}{360}\left(\dfrac{1}{10} + \dfrac{y}{5040}\left(\dfrac{1}{16} + \dfrac{y}{528528}\right)\right)\right)$

   $S_7(x_j) = y_1 + y2.$

   Si $x_j > 0$

   $S_7(x_j) = 1 + \dfrac{x_j}{2}\left(\dfrac{1}{4} + \dfrac{x_j}{12}\left(\dfrac{1}{7} + \dfrac{x_j}{30}\left(\dfrac{1}{10} + \dfrac{x_j}{56}\left(\dfrac{1}{13} + \dfrac{x_j}{90}\left(\dfrac{1}{16} + \dfrac{x_j}{132}\left(\dfrac{1}{9} + \dfrac{x_j}{4004}\right)\right)\right)\right)\right)\right).$

Imprimir x_j, $S_7(x_j)$.

Fin de bucle j.

4. Fin.

Para $x=-2$, se tiene $y=4$, y

$$\begin{aligned} y_1 &= 1+\frac{4}{24}\left(\frac{1}{7}+\frac{4}{1680}\left(\frac{1}{13}+\frac{4}{225720}\right)\right) \simeq 1{,}0238, \\ y_2 &= \frac{-2}{2}\left(\frac{1}{4}+\frac{4}{360}\left(\frac{1}{10}+\frac{4}{5040}\left(\frac{1}{16}+\frac{4}{528528}\right)\right)\right) \simeq -0{,}2511, \end{aligned}$$

luego $S_7(-2)=y_1+y_2 \simeq 1{,}0238-0{,}2511=0{,}7727$.

Para $x=3$, se tiene

$$\begin{aligned} S_7(3) &= 1+\frac{3}{2}\left(\frac{1}{4}+\frac{3}{12}\left(\frac{1}{7}+\frac{3}{30}\left(\frac{1}{10}+\frac{3}{56}\left(\frac{1}{13}+\frac{3}{90}\left(\frac{1}{16}+\frac{3}{132}\left(\frac{1}{19}+\frac{3}{4004}\right)\right)\right)\right)\right)\right) \\ &\simeq 1{,}4325. \end{aligned}$$

En la figura siguiente se muestran 31 puntos de la gráfica de $S_7(x) \quad x \in [-5,10]$ aproximación de $f(x)$ con una precisión $\varepsilon = 10^{-3}$. El conjunto de puntos utilizado para trazar dicha gráfica es calculado con el algoritmo que acabamos de describir.

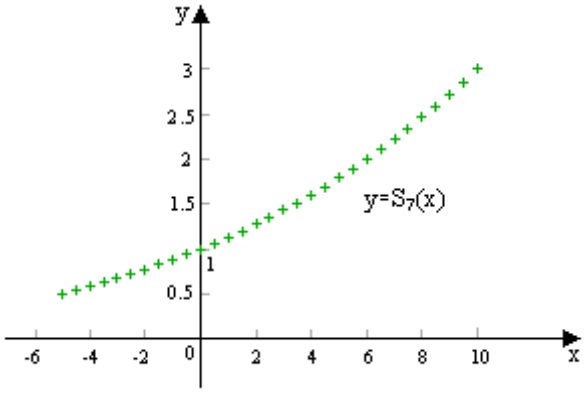


Figura 30

2. Consideramos la serie de potencias:

$$\ln(1+x)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k+1}x^{k+1} \qquad |x|<1$$

Esta serie no es la adecuada para el cálculo de $\ln(a)$ con $a>0$, sin embargo no será de utilidad para explicar algunas dificultades que se presenta.

Supongamos que deseamos calcular el valor aproximado de $\ln(1{,}5)$ con una precisión $\varepsilon = 10^{-10}$. Tenemos $x=0{,}5$. Luego

$$\ln(1{,}5)=\ln(1+0{,}5)=\frac{1}{2}\sum_{k=0}^{\infty}\frac{(-1)^k}{(k+1)\,2^k}.$$

Sabemos que $\sum_{k=1}^{\infty}\frac{1}{k(k+1)}=1$. Ponemos $a_k=\frac{1}{(k+1)\,2^k}$, $b_k=\frac{1}{k(k+1)}$. Entonces

$$c_k=\frac{a_k}{b_k}=\frac{k(k+1)}{(k+1)\,2^k}=\frac{k}{2^k}\underset{k\to\infty}{\longrightarrow}0,$$

luego existe $m\in\mathbb{Z}^+$ tal que $\frac{a_k}{b_k}<\varepsilon$ si $k\geq m$.

Para $m = 40$ se tiene $\dfrac{a_k}{b_k} < \varepsilon = 10^{-10}$ si $k \geq 40$. Se define $S_{40} = \dfrac{1}{2}\sum_{k=0}^{40} \dfrac{(-1)^k}{(k+1)\,2^k}$ y $|\ln(1{,}5) - S_{40}| < \varepsilon = 10^{-10}$.

Si $x = 0{,}8$, se tiene $\ln(1{,}8) = \sum_{k=0}^{\infty} \dfrac{(-1)^k\,(0{,}8)^{k+1}}{k+1}$, y aplicando el mismo criterio de comparación, se tiene que $\dfrac{a_k}{b_k} < \varepsilon = 10^{-10}$ si $k \geq 125$, donde $a_k = \dfrac{(0{,}8)^{k+1}}{k+1}$ y $b_k = \dfrac{1}{k\,(k+1)} \quad k = 1, 2, \ldots$. Se define $S_{125} = \sum_{k=0}^{125} \dfrac{(-1)^k\,(0{,}8)^{k+1}}{k+1}$ y $|\ln(1{,}8) - S_{125}| < \varepsilon = 10^{-10}$.

Cuando $x < 1$ se aproxima a 1, observamos que el número de términos crece enormemente, lo que por una parte dificulta la determinación del número adecuado de términos, por otra parte, se deben calcular sumas con un número muy grande de términos. Estos elementos dificultan la elaboración de un algoritmo de cálculo de $\ln(1+x)$.

3.5. Aproximación de las funciones trigonométricas

Cuando utilizamos un instrumento de cálculo tal como una calculadora de bolsillo o un computador, podemos obtener inmediatamente valores de las funciones trigonométricas seno, coseno, tangente; pero de esto, la pregunta que nos hacemos es ¿cómo con estos instrumentos se calculan valores de estas funciones trigonométricas?, ¿qué método se utiliza para garantizar la precisión de cálculo requerido? Esta sección está destinada a analizar el uso de la serie de Taylor de $\text{sen}(x)$ que nos permitan calcular aproximaciones de esta y de las funciones trigonométricas coseno y tangente.

Aproximación de $\text{sen}(x)$

La serie de Taylor de $sen(x) \quad x \in \mathbb{R}$, viene dada por $\text{sen}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$. Esta serie es absolutamente convergente para todo $x \in \mathbb{R}$, además, es rápidamente convergente. Por otro lado, el número de condicionamiento de $\text{sen}(x)$ está definido por $c(x) = \dfrac{x\cos(x)}{\text{sen}(x)} \quad x \neq 0$, y se probó en el capítulo 1 que $|c(x)| \leq 1$ si $x \in]0, \frac{\pi}{2}]$, con lo que la serie de Taylor será utilizada para aproximar $sen(x) \quad x \in \left[0, \dfrac{\pi}{2}\right]$, mediante una suma finita $S_N(x)$. Determinemos el número entero positivo N, el más pequeño posible, tal que si $\epsilon > 0, \quad |\,\text{sen}\,x - S_N(x)| < \epsilon \quad \forall x \in \left[0, \frac{\pi}{2}\right]$. Para el efecto, primeramente establecemos la siguiente mayoración:

$$\left|\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}\right| \leq \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!} \leq \sum_{k=0}^{\infty} \frac{\left(\frac{\pi}{2}\right)^{2k+1}}{(2k+1)!} \quad x \in \left[0, \frac{\pi}{2}\right],$$

y la convergencia de la última serie es absoluta. A continuación aplicamos el criterio del cociente, ponemos $a_k = \dfrac{\left(\frac{\pi}{2}\right)^{2k+1}}{(2k+1)!} \quad k =, 1, \cdots$ y consideramos una serie numérica convergente de suma 1, elegimos $\sum_{k=1}^{\infty} \frac{1}{2^k} = 1$, ponemos $b_k = \dfrac{1}{2^k} \quad k =, 1, \cdots$ y consideramos la sucesión $\left(\dfrac{a_k}{b_k}\right)$, esto es

$$\frac{a_k}{b_k} = \frac{2^k\left(\frac{\pi}{2}\right)^{2k+1}}{(2k+1)!} \underset{k\longrightarrow\infty}{\longrightarrow} 0\,.$$

De la convergencia se sigue que si $\epsilon = 10^{-10}$, se verifica que $\dfrac{a_k}{b_k} < 10^{-10} \quad \forall k \geq 9$. Para $\epsilon = 10^{-32}$ se verifica que $\dfrac{a_k}{b_k} < 10^{-32} \quad \forall k \geq 20$. Para fijar las ideas, elegimos $\epsilon = 10^{-12}$ y el correspondiente N es $N = 11$. Luego

$$S_{11}(x) = \sum_{k=0}^{11} \frac{(-1)^k x^{2k+1}}{(2k+1)!} = \sum_{k=0}^{5} \frac{x^{4k+1}}{(4k+1)!} - \sum_{k=0}^{5} \frac{x^{4k+3}}{(4k+3)!} = P_1(x) - P_2(x),$$

donde

$$\begin{aligned} P_1(x) &= \sum_{k=0}^{5} \frac{x^{4k+1}}{(4k+1)!} = x\left(1+\frac{x^4}{5!}\left(1+\frac{x^4}{6\times 7\times 8\times 9}\left(1+\frac{x^4}{10\times 11\times 12\times 13}\right.\right.\right. \\ &\quad \left.\left.\left.\left(1+\frac{x^4}{14\times 15\times 16\times 17}\left(1+\frac{x^4}{18\times 19\times 20\times 21}\right)\right)\right)\right)\right) \\ &= x\left(1+\frac{x^4}{120}\left(1+\frac{x^4}{3024}\left(1+\frac{x^4}{17160}\left(1+\frac{x^4}{57120}\left(1+\frac{x^4}{143640}\right)\right)\right)\right)\right), \end{aligned}$$

$$\begin{aligned} P_2(x) &= \sum_{k=0}^{5} \frac{x^{4k+3}}{(4k+3)!} = \frac{x^3}{3!}\left(1+\frac{x^4}{4\times 5\times 6\times 7}\left(1+\frac{x^4}{8\times 9\times 10\times 11}\right.\right. \\ &\quad \left.\left.\left(1+\frac{x^4}{12\times 13\times 14\times 15}\left(1+\frac{x^4}{16\times 17\times 18\times 19}\left(1+\frac{x^4}{20\times 21\times 22\times 23}\right)\right)\right)\right)\right) \\ &= \frac{x^3}{6}\left(1+\frac{x^4}{840}\left(1+\frac{x^4}{7920}\left(1+\frac{x^4}{32760}\left(1+\frac{x^4}{93024}\left(1+\frac{x^4}{212520}\right)\right)\right)\right)\right). \end{aligned}$$

Con esta información podemos construir un algoritmo para el cálculo aproximado de $\text{sen}(x) \quad x \in \left[0, \frac{\pi}{2}\right]$ con una precisión $\epsilon = 10^{-12}$. Requerimos adicionalmente aproximaciones de π. Consideramos las siguientes: con 12 cifras de precisión $\pi \simeq 3{,}14159265359$ por exceso y por defecto $\pi \simeq 3{,}141592653589$, y con 32 cifras de precisión $\pi \simeq 3{,}14159265358979323846264338327950$.

Algoritmo

Dato de entrada: x.

Datos de salida: $x, \text{sen}(x)$.

1. $y_1 = x^2 * x$,
2. $y = y_1 * x$
3. $a_1 = x\left(1+\frac{y}{120}\left(1+\frac{y}{3024}\left(1+\frac{y}{17160}\left(1+\frac{y}{57120}\left(1+\frac{y}{143640}\right)\right)\right)\right)\right)$.
4. $a_2 = \frac{y_1}{6}\left(1+\frac{y}{840}\left(1+\frac{y}{7920}\left(1+\frac{y}{32760}\left(1+\frac{y}{93024}\left(1+\frac{y}{212520}\right)\right)\right)\right)\right)$.
5. $S_{11}(x) = a_1 - a_2$.
6. Imprimir x, $S_{11}(x)$.
7. Fin

Para el cálculo de un solo valor de $\text{sen}(x)$ con $x \in \left[0, \frac{\pi}{2}\right]$ se requieren de 36 operaciones elementales y de 5 asignaciones. En los ejercicios se propone elaborar un algoritmo de cálculo de $\text{sen}(x) \quad x \in \left[0, \frac{\pi}{2}\right]$ y $\epsilon = 10^{-32}$ con el número de términos $N = 21$.

Al algoritmo descrito precedentemente los denominaremos como algoritmo de aproximación o también método de aproximación de $\text{sen}(x)$ con $x \in \left[0, \frac{\pi}{2}\right]$. Lo notaremos $\text{sen}_*(x)$

Para $x \in \mathbb{R} \setminus \left[0, \frac{\pi}{2}\right]$, consideramos los tres casos siguientes: i) $x \in]\frac{\pi}{2}, \pi]$, ii) $x > \pi$, iii) $x < 0$. Para calcular $\text{sen}(x)$ aplicaremos las propiedades de la función $\text{sen}(x)$ de modo que el algoritmo que acabamos de proponer se aplique con ligeras modificaciones.

a) Puesto que $\text{sen}(\pi - x) = \text{sen}(x) \quad \forall x \in \mathbb{R}$; en particular para $x \in]\frac{\pi}{2}, \pi]$ se sigue que $\pi - x \in]0, \frac{\pi}{2}]$, lo que nos permite aproximar $\text{sen}(x)$ mediante la suma $S_{11}(\pi - x) \quad x \in]\frac{\pi}{2}, \pi]$, utilizando el algoritmo arriba descrito.

b) Si $x > \pi$ entonces $y = x - n\pi \in [0, \pi]$, donde $n = [\frac{x}{\pi}]$ y $[\cdot]$ denota la función mayor entero $\leq \frac{x}{\pi}$. Luego

$$\operatorname{sen}(x) = \operatorname{sen}(y + n\pi) = \operatorname{sen}(y)\cos(n\pi) + \operatorname{sen}(n\pi)\cos(y) = \begin{cases} -\operatorname{sen}(y), & \text{si } n \text{ impar}, \\ \operatorname{sen}(y), & \text{si } n \text{ par}. \end{cases}$$

Así, para $x > \pi$, $\operatorname{sen}(x)$ se aproxima utilizando el algoritmo y la parte a) precedente con $y = x - n\pi$; y $\operatorname{sen}(x) = -\operatorname{sen}(y)$ si n es impar, $\operatorname{sen}(x) = \operatorname{sen}(y)$ si n es par.

c) Si $x < 0$, como la función seno es impar, esto es, $\operatorname{sen}(x) = -\operatorname{sen}(-x) \quad \forall x \in \mathbb{R}$, basta cambiar x por $-x$ y utilizar el algoritmo y los resultados de las partes a) y b) precedentes.

Se propone al lector la elaboración completa del algoritmo que permite aproximar $\operatorname{sen}(x) \quad x \in \mathbb{R}$.

Al algoritmo descrito precedentemente así como los resultados obtenidos en a), b) y c) los denominaremos como algoritmo o método de aproximación de $\operatorname{sen}(x) \quad x \in \mathbb{R}$.

Ejemplos

1. Tomando en consideración $pi = 3{,}1415926536$ aproximación de π, calcular una aproximación de $\operatorname{sen}\left(\frac{\pi}{10}\right)$ con una precisión $\varepsilon = 10^{-4}$. Para el efecto aplicamos el algoritmo. Ponemos $x = \frac{\pi}{10} \simeq \frac{pi}{10} = 0{,}3141592654$, $y = x^4 \simeq 0{,}009740909109$. Luego,

$$\begin{aligned} a_1 &= x\left(1 + \frac{y}{120}\left(1 + \frac{y}{3024}\left(1 + \frac{y}{17160}\left(1 + \frac{y}{57120}\left(1 + \frac{y}{143640}\right)\right)\right)\right)\right) \\ &= 0{,}3141847673, \\ a_2 &= \frac{x^3}{6}\left(1 + \frac{y}{780}\left(1 + \frac{y}{7920}\left(1 + \frac{y}{32760}\left(1 + \frac{y}{93024}\left(1 + \frac{y}{212520}\right)\right)\right)\right)\right) \\ &= 0{,}005167772705, \end{aligned}$$

   de donde $S_{11}\left(\frac{\pi}{10}\right) = a_1 - a_2 \simeq 0{,}3090169946$.

   El valor de $\operatorname{sen}\left(\frac{\pi}{10}\right)$ obtenido en una calculadora de bolsillo es $\operatorname{sen}\left(\frac{\pi}{10}\right) \simeq 0{,}3090169944$.

2. Calculemos $\operatorname{sen}(10)$. Para el efecto apliquemos los resultados arriba obtenidos. Ponemos $x = 10$. Se tiene $x > \pi$ entonces $x = 3\pi + 10 - 3\pi$. Ponemos $a = 10 - 3\pi \simeq 0{,}57522220393 \in \left[0, \frac{\pi}{2}\right]$. Luego, $\operatorname{sen}(10) = \operatorname{sen}(3\pi + a) = \operatorname{sen}(3\pi)\cos(a) + \operatorname{sen}(a)\cos(3\pi) = -\operatorname{sen}(a)$. Calculemos una aproximación de $\operatorname{sen}(a)$ con una precisión $\varepsilon = 10^{-10}$.

   Sea $y = a^4 \simeq 0{,}1094818355$ y $a^3 \simeq 0{,}1903296953$. Entonces

$$\begin{aligned} a_1 &= a\left(1 + \frac{y}{120}\left(1 + \frac{y}{3024}\left(1 + \frac{y}{17160}\left(1 + \frac{y}{57120}\left(1 + \frac{y}{143640}\right)\right)\right)\right)\right) \\ &= 0{,}5757468615, \\ a_2 &= \frac{a^3}{6}\left(1 + \frac{y}{840}\left(1 + \frac{y}{7920}\left(1 + \frac{y}{32760}\left(1 + \frac{y}{93024}\left(1 + \frac{y}{212520}\right)\right)\right)\right)\right) \\ &= 0{,}03172575038, \\ \operatorname{sen}_*(a) &= a_1 - a_2 = 0{,}5440211111. \end{aligned}$$

   Así, $\operatorname{sen}(10) = -\operatorname{sen}(a) \simeq -0{,}5440211111$.

   El valor $\operatorname{sen}(10)$ obtenido en una calculadora de bolsillo es $\operatorname{sen}(10) \simeq -0{,}5440211109$.

Aproximación de $\cos(x)$

Para aproximar $\cos(x) \quad x \in \mathbb{R}$, utilizamos la siguiente relación: $\cos(x) = \operatorname{sen}\left(\frac{\pi}{2} - x\right) \quad \forall x \in \mathbb{R}$ y aplicamos el método de aproximación de $\operatorname{sen}(x)$ a condición de cambiar x por $\frac{\pi}{2} - x$. Se propone elaborar el algoritmo correspondiente.

Ejemplo

Es conocido que $\cos\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2} \simeq 0{,}8660254038$. Apliquemos el algoritmo de cálculo de $\operatorname{sen}(a)$ con $a \in \left[0, \frac{\pi}{2}\right]$ para calcular una aproximación de $\cos\left(\frac{\pi}{6}\right)$. Sea $a = \frac{\pi}{2} - \frac{\pi}{6} = \frac{\pi}{3} \simeq 1{,}047197551$. Entonces $y = a^4 \simeq 1{,}20258137$, $a^3 = 1{,}148380617$. Luego

$$\begin{aligned} a_1 &= x\left(1 + \frac{y}{120}\left(1 + \frac{y}{3024}\left(1 + \frac{y}{17160}\left(1 + \frac{y}{57120}\left(1 + \frac{y}{143640}\right)\right)\right)\right)\right) \\ &\simeq 1{,}057696227, \\ a_2 &= \frac{a^3}{6}\left(1 + \frac{y}{840}\left(1 + \frac{y}{7920}\left(1 + \frac{y}{32760}\left(1 + \frac{y}{93024}\left(1 + \frac{y}{212520}\right)\right)\right)\right)\right) \\ &\simeq 0{,}1916708232, \\ \operatorname{sen}(a) &\simeq a_1 - a_2 = 1{,}057696227 - 0{,}1916708232 = 0{,}8660254038. \end{aligned}$$

En consecuencia $\cos\left(\frac{\pi}{6}\right) = \operatorname{sen}(a) \simeq 0{,}8660254038$ que es la aproximación obtenida de $\frac{\sqrt{3}}{2}$. En una calculadora de bolsillo se obtiene el siguiente valor $\cos\left(\frac{\pi}{6}\right) = 0{,}8660254038$.

Aproximación de $\tan(x)$

De la definición de la función tangente, se tiene

$$\tan(x) = \frac{\operatorname{sen}(x)}{\cos(x)} \qquad x \in \mathbb{R}\backslash\{\frac{\pi}{2} + k\pi \mid k \in \mathbb{Z}\},$$

luego los valores de $\tan(x) \quad x \in \mathbb{R}\backslash\{\frac{\pi}{2} + k\pi \mid k \in \mathbb{Z}\}$ pueden aproximarse utilizando esta relación y los métodos de aproximación de $\operatorname{sen}(x)$ y $\cos(x)$ arriba tratados.

Ejemplos

1. Es conocido que $\tan\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{3}$. Apliquemos el algoritmo de cálculo de $\operatorname{sen}(x)$ para aproximar $\operatorname{sen}\left(\frac{\pi}{6}\right)$ y $\cos\left(\frac{\pi}{6}\right)$ y así aproximar $\tan\left(\frac{\pi}{6}\right) = \frac{\operatorname{sen}\left(\frac{\pi}{6}\right)}{\cos\left(\frac{\pi}{6}\right)}$.

   Consideremos $\pi \simeq 3{,}141592653$, $x = 0{,}5235987755$, $x^3 \simeq 0{,}1435475771$, $y = x^4 \simeq 0{,}0751613356$. Aplicando el algoritmo de cálculo para aproximar $\operatorname{sen}(x)$, tenemos

   $$\begin{aligned} a_1 &= x\left(1 + \frac{y}{120}\left(1 + \frac{y}{3024}\left(1 + \frac{y}{17160}\left(1 + \frac{y}{57120}\left(1 + \frac{y}{143640}\right)\right)\right)\right)\right) \\ &\simeq 0{,}5239267369, \\ a_2 &= \frac{x^3}{6}\left(1 + \frac{y}{840}\left(1 + \frac{y}{7920}\left(1 + \frac{y}{32760}\left(1 + \frac{y}{93024}\left(1 + \frac{y}{212520}\right)\right)\right)\right)\right) \\ &\simeq 0{,}02392673695. \end{aligned}$$

   En consecuencia

   $$\operatorname{sen}\left(\frac{\pi}{6}\right) \simeq a_1 - a_2 = 0{,}5239267369 - 0{,}02392673695 = 0{,}5.$$

   Con la misma precisión, calculamos una aproximación de $\cos\left(\frac{\pi}{6}\right)$. Ponemos $x = \frac{\pi}{2} - \frac{\pi}{6} = \frac{\pi}{3} \simeq 1{,}047197551$, $x^3 \simeq 1{,}148380617$, $x^4 = 1{,}20258137$. Aplicando nuevamente los desarrollos a_1, a_2 precedentes, se obtiene $a_1 = 1{,}057696227$, $a_2 = 0{,}1916708232$. Luego

   $$\cos\left(\frac{\pi}{6}\right) = \operatorname{sen}\left(\frac{\pi}{6}\right) \simeq a_1 - a_2 = 0{,}8660254038.$$

   Por lo tanto,

   $$\tan\left(\frac{\pi}{6}\right) = \frac{\operatorname{sen}\left(\frac{\pi}{6}\right)}{\cos\left(\frac{\pi}{6}\right)} \simeq \frac{0{,}5}{0{,}8660254038} \simeq 0{,}577350269.$$

Valor obtenido en una calculadora de bolsillo $\tan\left(\frac{\pi}{6}\right) = 0{,}5773502692$.

2. Para $0 < \varepsilon < 10^{-2}$, $\tan\left(\frac{\pi}{2} - \varepsilon\right)$ se puede hacer tan grande como se quiera conforme ε se aproxima a cero.

   Si aproximamos $\tan(x) = \dfrac{\operatorname{sen}(x)}{\cos(x)}$ para $x \in \left[0, \frac{\pi}{2}\right[$, con el algoritmo de cálculo de $\operatorname{sen}(x)$, es crítico para $x = \dfrac{\pi}{2} - \varepsilon$. Veamos esta situación con el siguiente ejemplo: aproximar $\tan(89{,}9995°)$.

   En radianes $89{,}9995\,° = 1{,}5707876\,\text{rad}$. Ponemos $x = 1{,}5707876$, entonces $x^3 \simeq 3{,}875719987$, $y = x^4 \simeq 6{,}087932897$. Calculemos $\operatorname{sen}(89{,}9995\,°) = \operatorname{sen}(1{,}5707876)$ con el algoritmo arriba desarrollado. Tenemos

$$\begin{aligned} a_1 &= x\left(1 + \frac{y}{120}\left(1 + \frac{y}{3024}\left(1 + \frac{y}{17160}\left(1 + \frac{y}{57120}\left(1 + \frac{y}{143640}\right)\right)\right)\right)\right) \\ &\simeq 1{,}650628503, \\ a_2 &= \frac{x^3}{6}\left(1 + \frac{y}{840}\left(1 + \frac{y}{7920}\left(1 + \frac{y}{32760}\left(1 + \frac{y}{93024}\left(1 + \frac{y}{212520}\right)\right)\right)\right)\right) \\ &\simeq 0{,}6506385023, \end{aligned}$$

   luego

$$\operatorname{sen}(1{,}5707876) \simeq a_1 - a_2 = 1{,}000000000.$$

   Calculemos $\cos(89{,}9995\,°) = \cos(1{,}5707876)$. Para el efecto, ponemos $x = \dfrac{\pi}{2} - 1{,}5707876 \simeq 0{,}000008727$. Se obtiene $x^3 \simeq 6{,}646529367 \times 10^{-16}$, $x^4 \simeq 5{,}800426178 \times 10^{-21}$ y aplicando el algoritmo de cálculo de $\operatorname{sen}(x)$ se obtiene: $a_1 \simeq 0{,}000008727$, $a_2 \simeq 1{,}107754895 \times 10^{-16}$. Luego $\cos(1{,}5707876) \simeq 0{,}000008727$ y en consecuencia

$$\tan(1{,}5707876) = \frac{\operatorname{sen}(1{,}5707876)}{\cos(1{,}5707876)} \simeq \frac{1{,}0}{0{,}000008727} \simeq 114586{,}9142.$$

   El valor obtenido en una calculadora de bolsillo es $\tan(1{,}5707876) = 114589{,}7256$. Esta pequeña diferencia se debe a que hemos operado con una precisión de 10^{-9} mientras que en la calculadora, internamente se opera con una precisión de 10^{-12}. Con la versión de Fortran 77, se obtiene en doble precisión el siguiente valor: $\tan(1{,}5707876) = 113924{,}073226171$.

Aproximación de $\operatorname{arcsen}(y)$.

Recordemos que la función f de $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ en $[-1, 1]$ definida como $y = f(x) = \operatorname{sen}(x) \quad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ es biyectiva y su función inversa g está definida como $x = g(y) = \operatorname{arcsen}(y) \quad y \in [-1, 1]$. Se tiene $(g \circ f)(x) = x \quad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Además, f y g son derivables, luego

$$1 = (g \circ f)'(x) = g'(y)\, f'(x) \quad x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$$

de donde $g'(y) = \dfrac{1}{f'(x)}$ con $y = f(x)$. Se tiene

$$f'(x) = \cos(x) = \sqrt{1 - \operatorname{sen}^2(x)} = \sqrt{1 - y^2},$$

luego

$$(\operatorname{arcsen}(y))' = g'(y) = \frac{1}{\sqrt{1 - y^2}} \quad y \in]-1, 1[,$$

e integrando, resulta

$$\operatorname{arcsen}(y) = \int_0^y \frac{dt}{\sqrt{1 - t^2}} \quad y \in]-1, 1[.$$

Por el binomio de Newton:

$$\begin{aligned}\left(1-t^2\right)^{-\frac{1}{2}} &= 1+\left(-\frac{1}{2}\right)\left(-t^2\right)+\frac{\left(-\frac{1}{2}\right)\left(-\frac{1}{2}-1\right)}{2!}\left(-t^2\right)^2+\frac{\left(-\frac{1}{2}\right)\left(-\frac{1}{2}-1\right)\left(-\frac{1}{2}-2\right)}{3!}\left(-t^2\right)^3+\cdots\\ &= 1+\frac{t^2}{2}+\frac{1\times 3}{2!\times 2^2}t^4+\frac{1\times 3\times 5}{3!\times 2^3}t^6+\cdots\\ &= 1+\sum_{k=1}^{\infty}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k}t^{2k} \quad t\in\,]-1,1[\,.\end{aligned}$$

Luego, por el teorema de la convergencia uniforme y la integración, se tiene

$$\begin{aligned}\operatorname{arcsen}(y) &= \int_0^y\left(1+\sum_{k=1}^{\infty}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k}t^{2k}\right)dt\\ &= y+\sum_{k=1}^{\infty}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k\,(2k+1)}y^{2k+1} \quad y\in\,]-1,1[\,.\end{aligned}$$

Para $\varepsilon=10^{-10}$, mediante el criterio de comparación con la serie $\sum_{k=1}^{\infty}\frac{1}{k\,(k+1)}$, se obtiene $m=25$ y en consecuencia $|\operatorname{arcsen}(y)-S_{25}(y)|<\varepsilon=10^{-10}$ si $y\in\left[0,\frac{\sqrt{2}}{2}\right]$, donde

$$S_{25}(y)=y\left(1+\sum_{k=1}^{25}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k\,(2k+1)}y^{2k}\right).$$

Si $y\in\left]\frac{\sqrt{2}}{2},1\right[$, para aproximar $\operatorname{arcsen}(y)$ se requiere de un número mayor de términos. Esto podemos controlarlo del modo siguiente. Sea $x=\operatorname{arcsen}(y)$, entonces

$$y=\operatorname{sen}(x)=\cos\left(\frac{\pi}{2}-x\right)=\sqrt{1-\operatorname{sen}^2\left(\frac{\pi}{2}-x\right)}$$

de donde $1-\operatorname{sen}^2\left(\frac{\pi}{2}-x\right)=y^2$ con lo que $x=\frac{\pi}{2}-\operatorname{arcsen}\left(1-y^2\right)^{\frac{1}{2}}$.

Como $y\in\left]\frac{\sqrt{2}}{2},1\right[$ se sigue que $\left(1-y^2\right)^{\frac{1}{2}}\in\left]0,\frac{1}{2}\right[$.

Ponemos $v=\left(1-y^2\right)^{\frac{1}{2}}$ y $|\operatorname{arcsen}(v)-S_{17}(v)|<\varepsilon=10^{-10}$ con

$$S_{17}(v)=v\left(1+\sum_{k=1}^{17}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k\,(2k+1)}v^{2k}\right).$$

Se define S_ε como sigue:

$$S_\varepsilon=\begin{cases} y\left(1+\sum_{k=1}^{25}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k\,(2k+1)}y^{2k}\right) & \text{si } \ y\in\left[0,\frac{\sqrt{2}}{2}\right]\\ \frac{\pi}{2}-v\left(1+\sum_{k=1}^{17}\frac{1\times 3\times\cdots\times(2k-1)}{k!2^k\,(2k+1)}v^{2k}\right) & \text{si } \ v\in\left[0,\frac{1}{2}\right]\end{cases}$$

con $v=\sqrt{1-y^2}$ $\quad y\in\left]\frac{\sqrt{2}}{2},1\right[$.

Ejemplo

Sabemos que $\operatorname{arcsen}\left(\frac{1}{2}\right) = \frac{\pi}{6} \simeq 0{,}5235987756$.

Apliquemos los resultados obtenidos precedentemente. Ponemos $y = 0{,}5$ y $\varepsilon = 10^{-10}$. Entonces

$$\begin{aligned} S_\varepsilon &= y\left(1 + \sum_{k=1}^{25} \frac{1 \times 3 \times \cdots \times (2k-1)}{k!2^k\,(2k+1)} y^{2k}\right) \\ &= y\left(1 + \frac{y^2}{2}\left(\frac{1}{3} + \frac{3y^2}{y}\left(\frac{1}{5} + \frac{5y^2}{6}\left(\frac{1}{7} + \frac{7y^2}{8}\left(\frac{1}{9} + \cdots + \frac{47y^2}{48}\left(\frac{1}{49} + \frac{49y^2}{50 \times 51}\right)\cdots\right)\right)\right)\right)\right) \\ &= 0{,}5235987755. \end{aligned}$$

Aproximación de $\operatorname{arc\,cos}(x)$

Se propone como ejercicio

Aproximación de $\arctan(y)$

Sea $x = \arctan(y)$. Entonces

$$y = \tan(x) = \frac{\operatorname{sen}(x)}{\cos(x)}, \quad x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[,$$

de donde $sen(x) = \dfrac{y}{\sqrt{1+y^2}}$ y por lo tanto $x = \operatorname{arcsen}\left(\frac{y}{\sqrt{1+y^2}}\right)$. Consecuentemente, $\arctan(y)$ se aproxima utilizando el algoritmo de $\operatorname{arcsen}(z)$, con $z = \dfrac{y}{\sqrt{1+y^2}}$.

Nota: Se recomienda al lector elaborar un programa computacional que permita aproximar las funciones trigonométricas utilizando los algoritmos descritos y comparar los resultados con los proporcionados con los de las calculadoras de bolsillo.

3.6. Aproximación de $\exp(x)$

Sea $x \in \mathbb{R}$, en el primer capítulo se mostró que el número de condicionamiento de $\exp(x)$ es $c(x) = x$, por lo tanto $\exp(x)$ está bien condicionado si $|x| \leq 1$. Por otro lado, en un entorno de $x = 0$, $\exp(x)$ se representa mediante el siguiente desarrollo en serie de potencias $\exp(x) = \sum\limits_{k=0}^{\infty} \frac{x^k}{k!} \quad x \in \mathbb{R}$. Esta serie es absolutamente convergente para todo $x \in \mathbb{R}$ (radio de convergencia $r = +\infty$).

Para $x \in [0,1]$, definimos $S_n(x) = \sum\limits_{k=0}^{n} \frac{x^k}{k!} \quad n = 1, 2, \ldots$. Sea $\epsilon > 0$. Determinemos $n \in \mathbb{N}$ el más pequeño posible tal que $|\exp(x) - S_n(x)| < \epsilon,\ \forall x \in [0,1]$, es decir que $\sum_{k=n+1}^{\infty} \dfrac{x^k}{k!} < \epsilon$. Se tiene

$$\sum_{k=n+1}^{\infty} \frac{x^k}{k!} \leq \sum_{k=n+1}^{\infty} \frac{1}{k!} < \epsilon \quad \forall x \in [0,1].$$

Para determinar $n \in \mathbb{N}$ aplicamos el criterio del cociente. Ponemos $a_k = \dfrac{1}{k!} \quad k = 0, 1, \cdots$, y elegimos la serie convergente $\sum\limits_{k=1}^{\infty} \dfrac{1}{k^p}$ con $1 < p < 2$. Ponemos $b_k = \dfrac{1}{k^p} \quad k = 0, 1, \cdots$, luego

$$\frac{a_k}{b_k} = \frac{k^{p-1}}{(k-1)!} \underset{k\to\infty}{\longrightarrow} 0.$$

Particularmente, para $p = 2$, el criterio para determinar n es el siguiente: $n \in \mathbb{N}$ el más pequeño posible tal que $\frac{n}{(n-1)!} < \epsilon$. Para $\epsilon = 10^{-10}$ se obtiene $n = 16$, para $\epsilon = 10^{-20}$ se obtiene $n = 24$, para $\epsilon = 10^{-32}$ se obtiene $n = 32$.

Para $\epsilon > 0$ dado, $S_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}$ aproxima a e^xcon una precisión ϵ para todo $x \in [0,1]$.

Con el propósito de obtener un algoritmo numéricamente estable, escribimos $S_n(x)$ en una forma anidada:

$$S_n(x) = 1 + x\left(1 + \frac{x}{2}\left(1 + \frac{x}{n-1}\left(1 + \frac{x}{n}\right)\cdots\right)\right),$$

cuyo algoritmo es el siguiente:

Algoritmo

Datos de entrada; n, x.

Datos de salida : x, $\exp(x)$.

1. $b = 1$.
2. $k = 0, \ldots, n-1$

   $$b = 1 + \frac{b * x}{n-k}$$

   Fin de bucle k.
3. Imprimir $\exp(x) = b$.
4. Fin.

Para cada $x \in [0,1]$, la aproximación de $\exp(x)$ dado en el algoritmo requiere de $4\times n$ operaciones elementales y n asignaciones.

Puesto que $\exp(x) = \dfrac{1}{\exp(-x)}$, entonces para $x \in [-1,0[$, $\exp(x)$ se aproxima mediante $\dfrac{1}{S_n(-x)}$ y $S_n(-x)$ se calcula usando el algoritmo precedente.

Para $x \in \mathbb{R}$ tal que $|x| > 1$, $\exp(x)$ está mal condicionado. Dado $\epsilon > 0$, si determinamos n tal que $\dfrac{|x|^n}{(n-1)!} < \epsilon$, resulta que tal n aumenta considerablemente según $|x|$, lo que hace que $S_n(x)$ sea numéricamente costoso y por otro lado, el algoritmo es inestable numéricamente. El remedio a este problema consiste en hacer $y = x - [x]$, donde $[\cdot]$ denota la función mayor entero menor o igual que x. Resulta $y \in]0,1[$, y $\exp(x) = \exp(y)\exp([x])$. Aproximamos $\exp(y)$ mediante $S_n(y)$ si $y > 0$ y $\frac{1}{S_n(y)}$ si $y < 0$. Como el número e base de los logaritmos naturales está dado como la serie $e = \sum_{k=0}^{\infty} \frac{1}{k!}$, se aplica el algoritmo precedente con $x = 1$ y luego $\exp([x])$ se evalúa como una potencia entera. Para $\epsilon = 10^{-10}$, se tiene $S_{16}(1) = \sum_{k=0}^{k=16} \dfrac{1}{k!} = 2{,}7182818284$.

Así, $\exp(x)$ se aproxima como $\begin{cases} S_n(y)\exp([x]), & \text{si } x > 1, \\ \dfrac{1}{S_n(-y)\exp(-[x])}, & \text{si } x < -1. \end{cases}$ Se recomienda al lector elaborar el algoritmo completo para aproximar $\exp(x)$ así como su respectivo programa computacional.

Ejemplos

1. Aplique el algoritmo para calcular una aproximación de $\exp(0{,}4)$ con una precisión $\varepsilon = 10^{-10}$. Se tiene $0 < x < 1$ y en consecuencia

   $$S_{16}(x) = \sum_{k=0}^{16} \frac{x^k}{k!} = 1 + x\left(1 + \frac{x}{2}\left(1 + \frac{x}{3}\left(1 + \ldots + \frac{x}{15}\left(1 + \frac{x}{16}\right)\ldots\right)\right)\right)$$

   en particular para $x = 0{,}4$, se obtiene

   $$S_{16}(x) = 1 + 0{,}4\left(1 + \frac{0{,}4}{2}\left(1 + \frac{0{,}4}{3}\left(1 + \ldots + \frac{0{,}4}{15}\left(1 + \frac{0{,}4}{16}\right)\ldots\right)\right)\right) = 1{,}491824698\ldots$$

El valor de $\exp(0{,}4)$ obtenido en doble precisión es $\exp(0{,}4) \simeq 1{,}491824706533238$ y en una calculadora de bolsillo $\exp(0{,}4) \simeq 1{,}491824698$.

2. Calculemos el valor aproximado de $\exp(22{,}4)$. Tenemos $x = 22{,}4 > 1$, en consecuencia $x = (x - [x]) + [x] = y + [x]$ con $y = x - [x] \in [0, 1]$. Resulta $[22{,}4] = 22$. Luego

$$\exp(22{,}4) = \exp(0{,}4)\exp(22).$$

En el ejemplo 1) previo se calculó el valor aproximado de $\exp(0{,}4)$. Queda por calcular $\exp(22)$. Primeramente $\exp(1) = 2{,}7182818284$. Luego

$$\exp(22) = (2{,}7182818284)^{22} = 3584912833 = 3{,}584912833 \times 10^9,$$

de donde

$$\begin{aligned} \exp(22{,}4) &= \exp(0{,}4)\exp(22) \simeq 1{,}491824698 \times 3{,}584912833 \times 10^9 \\ &= 5{,}348061504 \times 10^9. \end{aligned}$$

El valor de $\exp(22{,}4)$ obtenido en una calculadora de bolsillo es $\exp(22{,}4) \simeq 5{,}348061523 \times 10^9$, y en doble precisión $\exp(22{,}4) \simeq 5{,}34805948262739 \times 10^9$. Note que $\exp(22) \simeq 3{,}584912846131592 \times 10^9$. Debido a que en la calculadora de bolsillo se representan los números en punto fijo, donde se produce mayor error es en el cálculo de $\exp(22) \simeq (2{,}7182818284)^{22}$.

3. Calculemos el valor aproximado de $\exp(-0{,}9)$. Puesto que $\exp(-0{,}9) = \dfrac{1}{\exp(0{,}9)}$, calculamos el valor aproximado de $\exp(0{,}9)$. Aplicando el algoritmo, tenemos

$$S_{16}(x) = \sum_{k=0}^{16} \frac{(0{,}9)^k}{k!} = 1 + 0{,}9\left(1 + \frac{0{,}9}{2}\left(1 + \cdots \frac{0{,}9}{15}\left(1 + \frac{0{,}9}{16}\right)\ldots\right)\right) = 2{,}459603112,$$

de donde

$$S_{16}(0{,}9) = \frac{1}{S_{16}(0{,}9)} = \frac{1}{2{,}459603112} = 0{,}4065696598.$$

El valor obtenido en una calculadora de bolsillo es $\exp(-0{,}9) \simeq 0{,}4065696597$.

3.7. Aproximación de $\ln(x)$

Antes de abordar el problema de la aproximación numérica de $\ln(a)$ con $a > 0$, recordemos algunas propiedades de la función logaritmo.

1. Sean $a, b \in \mathbb{R}^+$, $\ln(ab) = \ln(a) + \ln(b)$.

2. Si $a \in \mathbb{R}^+$, $\ln\left(\dfrac{1}{a}\right) = -\ln(a)$.

3. Si $n \in \mathbb{Z}^+$, $a \in \mathbb{R}^+$, $\ln(a^n) = n\ln(a)$ y $\ln\left(\dfrac{1}{a^n}\right) = -n\ln(a)$.

Por otro lado, sea $a > e$ y $n \in \mathbb{Z}^+$ tal que $1 \leq \dfrac{a}{e^n} < e$, donde $a = e^n \times \dfrac{a}{e^n}$, luego

$$\ln(a) = \ln\left[e^n \times \frac{a}{e^n}\right] = \ln e^n + \ln\left(\frac{a}{e^n}\right) = n + \ln(x),$$

con $x = \dfrac{a}{e^n} \in [1, e]$.

Si $a < 1$ y $n \in \mathbb{Z}^-$ tal que $1 \leq ae^{-n} < e$, entonces $a = e^n \times (ae^{-n})$, de donde

$$\ln(a) = n + \ln\left(ae^{-n}\right) = n + \ln(x),$$

con $x = ae^{-n} \in [1, e]$.

Ejemplos

1. $a = 20{,}11 = e^3 \times \dfrac{20{,}11}{e^3}$ con $x = \dfrac{20{,}11}{e^3} \simeq 1{,}001217945 \in [1, e]$.

2. $a = 145{,}41 = e^4 \times \dfrac{145{,}41}{e^4}$ con $x = \dfrac{145{,}41}{e^4} \simeq 2{,}663277051 \in [1, e]$.

3. $a = 6{,}81 \times 10^{-3} = e^{-5} \times \left(e^5 \times 6{,}81 \times 10^{-3}\right)$, donde $x = e^5 \times 6{,}81 \times 10^{-3} \simeq 1{,}010693613 \in [1, e]$.

De las dos últimas relaciones, si $a > 0$ basta determinar $n \in \mathbb{Z}^+$ tal que $x = ae^n \in [1, e]$ si $a < \dfrac{1}{e}$, $x = \dfrac{a}{e^n} \in [1, e]$ si $a > e$. En cualquiera de los casos, queda calcular $\ln(x)$. Para el efecto, utilizamos el siguiente desarrollo en series de potencias:

$$\ln\left(\frac{1+x}{1-x}\right) = 2x \sum_{k=0}^{\infty} \frac{x^{2k}}{2k+1} \quad \text{si } |x| < 1.$$

Sea $a > 0$. Ponemos $a = \dfrac{1+x}{1-x}$, entonces $x = \dfrac{a-1}{a+1}$ y en consecuencia

$$\ln(a) = 2\frac{a-1}{a+1} \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^{2k}.$$

Para $a = 1$, obviamente $\ln(1) = 0$. Para $a \in [1, e]$, sea $m \in \mathbb{Z}^+$ y $S_m(a) = \sum\limits_{k=0}^{m} \dfrac{1}{2k+1} \left(\dfrac{a-1}{a+1}\right)^{2k}$.

Con fines prácticos elegimos $\varepsilon = 10^{-10}$. Determinemos $m \in \mathbb{Z}^+$ el más pequeño posible tal que

$$\left|\ln(a) - 2\frac{a-1}{a+1} S_m(a)\right| < \varepsilon = 10^{-10},$$

lo que conduce a determinar $m \in \mathbb{Z}^+$ tal que

$$\begin{aligned} \left|\sum_{k=0}^{m} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^{2k} - S_m(a)\right| &= \sum_{k=m+1}^{\infty} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^{2k} \leq \sum_{k=m+1}^{\infty} \frac{1}{2k+1} \left(\frac{e-1}{e+1}\right)^{2k} \\ &< \sum_{k=m+1}^{\infty} \frac{1}{2k+1} \frac{1}{4^k} < \varepsilon = 10^{-10}. \end{aligned}$$

La serie $\sum\limits_{k=1}^{\infty} \dfrac{1}{k(k+1)} = 1$. Ponemos $a_k = \dfrac{1}{2k+1}\dfrac{1}{4^k}$, $\quad b_k = \dfrac{1}{k(k+1)} \quad k = 1, 2, \ldots, \quad C_k = \dfrac{a_k}{b_k} = \dfrac{k(k+1)}{(2k+1)4^k} \underset{k \longrightarrow \infty}{\longrightarrow} 0$, luego, existe $m \in \mathbb{Z}^+$ tal que $C_k = \dfrac{a_k}{b_k} < \varepsilon$ si $k \geq m$. Para $k = 15$ obtenemos $C_{15} = \dfrac{15 \times 16}{31 \times 4^{15}} \simeq 7{,}21 \times 10^{-9}$, para $k = 19$, $C_{19} = \dfrac{19 \times 20}{39 \times 4^{19}} \simeq 3{,}545 \times 10^{-11}$. Elegimos $m = 19$ y definimos

$$\varphi_{19}(a) = \frac{2(a-1)}{a+1} \sum_{k=0}^{19} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^2 \qquad a \in [1, e].$$

Note que si $1 < a \leq \dfrac{e+1}{2}$ entonces $0 < \dfrac{a-1}{a+1} \leq \dfrac{e-1}{e+3}$ de donde $0 < \left(\dfrac{a-1}{a+1}\right)^2 \leq \left(\dfrac{e-1}{e+1}\right)^2 < 0{,}1$. En tal caso, ponemos $a_k = \dfrac{1}{2k+1}\dfrac{1}{10^k}$, $b_k = \dfrac{1}{k(k+1)} \quad k = 1, 2, \ldots$, luego

$$C_k = \frac{a_k}{b_k} = \frac{k(k+1)}{(2k+1)10^k} \underset{k \to \infty}{\longrightarrow} 0,$$

y existe $m \in \mathbb{Z}^+$ tal que $C_k = \dfrac{a_k}{b_k} < \varepsilon = 10^{-10}$ $\quad$ si $k \geq m$.

Para $m = 11$, se tiene $C_{11} = \dfrac{11 \times 12}{23 \times 10^{11}} \simeq 5{,}74 \times 10^{-11}$. Definimos

$$\varphi_{11}(a) = \frac{2(a-1)}{a+1} \sum_{k=0}^{11} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^{2k} \quad a \in \left]1, \frac{e+1}{2}\right].$$

Así,

$$\begin{cases} \varphi_{11}(a) = \dfrac{2(a-1)}{a+1} \displaystyle\sum_{k=0}^{11} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^{2k} & a \in \left]1, \dfrac{e+1}{2}\right], \\ \varphi_{19}(a) = \dfrac{2(a-1)}{a+1} \displaystyle\sum_{k=0}^{19} \frac{1}{2k+1} \left(\frac{a-1}{a+1}\right)^{2k} & a \in \left]\dfrac{e+1}{2}, e\right]. \end{cases}$$

Sea $b_1 = \dfrac{a-1}{a+1}$ y $b = b_1^2$. Entonces

$$\begin{aligned} \varphi_{11}(a) &= 2b_1 \sum_{k=0}^{11} \frac{1}{2k+1} b^k = 2b_1 \left(1 + \frac{b}{3} + \frac{b^2}{5} + \cdots + \frac{b^{10}}{21} + \frac{b^{11}}{23}\right) \\ &= 2b_1 \left(1 + b\left(\frac{1}{3} + b\left(\frac{1}{5} + \cdots + b\left(\frac{1}{21} + \frac{b}{23}\right)\cdots\right)\right)\right). \end{aligned}$$

En forma similar se escribe $\varphi_{19}(a)$.

Un algoritmo para el cálculo de $\ln(a)$ con $a \in [1, e]$ con una precisión $\varepsilon = 10^{-10}$ se propone a continuación

Algoritmo

Datos de entrada: $a \in]1, e[$.

Datos de salida: $\ln(a)$.

1. Si $1 \leq a \leq \dfrac{1+e}{2}$, asignar $n = 11$.
2. 2i $\dfrac{1+e}{2} < a < e$, asignar $n = 19$.
3. $b_1 = \dfrac{a-1}{a+1}$.
4. $b = b_1^2$.
5. $y = \dfrac{1}{2n+1}$.
6. Para $j = 1, \ldots, n$

   $k = n - j$

   $y = \dfrac{1}{2k+1} + b \times y$

   Fin bucle j.
7. $y = 2 \times b_1 \times y$.
8. Imprimir $\ln(a) = y$.
9. Fin.

Ejemplos

1. Calculemos $\ln(2)$. Para el efecto, aplicamos el algoritmo descrito. Ponemos $a = 2$, $b_1 = \dfrac{a-1}{a+1} = \dfrac{1}{3}$, $b = b_1^2 = \dfrac{1}{9}$. Entonces

$$\ln(2) \simeq \varphi_{19}(2) = \frac{2}{3}\left(1 + b\left(\frac{1}{3} + b\left(\frac{1}{5} + \cdots + b\left(\frac{1}{37} + \frac{b}{39}\right)\cdots\right)\right)\right) = 0{,}6931471806.$$

En una calculadora de bolsillo, $\ln(2) = 0{,}6931471806$.

2. Calculemos $\ln(535{,}2)$. Tenemos $\dfrac{535{,}2}{e^6} \simeq 1{,}326628165 \in [1, e]$, luego $535{,}2 = e^6 \dfrac{535{,}2}{e^6}$, de donde

$$\ln(535{,}2) = 6 + \ln \frac{535{,}2}{e^6} = 6 + \ln 1{,}326628165.$$

Sea $a = 1{,}326628165$ entonces $b_1 = \dfrac{a-1}{a+1} = 0{,}1403869213$, $b = b_1^2 = 0{,}01970848767$, luego

$$\varphi_{11}(a) = 2b_1 \left(1 + b\left(\frac{1}{3} + b\left(\frac{1}{5} + \cdots + b\left(\frac{1}{21} + \frac{b}{23}\right)\cdots\right)\right)\right) = 0{,}2826405088.$$

En consecuencia

$$\ln(535{,}2) = 6 + 0{,}2826405088 = 6{,}2826405088.$$

En una calculadora de bolsillo, $\ln(535{,}2) = 6{,}282640509$.

3. Apliquemos el algoritmo y los resultados precedentes para calcular $\ln(0{,}01234)$.
Sea $n \in \mathbb{Z}^+$ tal que $0{,}01234 \times e^n \in [1, e]$. Para $n = 5$ se tiene $x = 0{,}01234 \times e^5 \simeq 1{,}831418383 \in [1, e]$.
Luego

$$a = 0{,}01234 = e^{-5} \times \left(0{,}01234 \times e^5\right) \simeq e^{-5} \times 1{,}831419383,$$

y $\ln(0{,}01234) = -5 + \ln(1{,}83141838)$. Tenemos $x = 1{,}831418383$, $b_1 = \dfrac{x-1}{x+1} = 0{,}2936402433$, $b = b_1^2 = 0{,}08622459249$. Aplicando el algoritmo obtenemos $\ln(x) = 0{,}6050907394$, con lo que

$$\ln(0{,}01234) = -5 + 0{,}605090739 = -4{,}394909261.$$

En una calculadora de bolsillo $\ln(0{,}01234) = -4{,}394909261$.

3.8. Integración de funciones de clase $C^\infty(\mathbb{R})$

Se denota con $C^\infty(\mathbb{R})$ al espacio vectorial de las funciones reales que poseen derivadas de todos los órdenes continuas en todo $\mathbb{R}$. Supongamos que $f \in C^\infty(\mathbb{R})$ se representa mediante una serie de potencias $\displaystyle\sum_{k=0}^{\infty} a_k x^k$ y se desea calcular $I(f) = \displaystyle\int_0^a f(x)dx$, $\quad$ con $a > 0$.

El polinomio de Taylor de f en un entorno de cero viene dado como $P(x) = \displaystyle\sum_{k=0}^{m} \frac{f^{(k)}(0)}{k!} x^k$. Entonces $f(x) = P(x) + E_m(x)$, donde $E_m(x)$ es el error de aproximación en x. Resulta que

$$\begin{aligned} I(f) &= \int_0^a f(x)dx = \sum_{k=0}^{m} \frac{f^{(k)}(0)}{k!} \int_0^a x^k dx + \int_0^a E_m(x)dx \\ &= \sum_{k=0}^{m} \frac{f^{(k)}(0)}{(k+1)!} a^{k+1} + \int_0^a E_m(x)dx, \quad m = 1, 2, 3, \ldots, . \end{aligned}$$

Además $f(x) = \displaystyle\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$, entonces $I(f) = \displaystyle\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{(k+1)!} a^{k+1}$. Sea

$$I_m(f) = \sum_{k=0}^{m} \frac{f^{(k)}(0)}{(k+1)!} a^{k+1} \qquad m = 1, 2, \ldots, .$$

Cada $I_m(f)$ es una aproximación de $I(f)$ por truncamiento. Sea $\epsilon > 0$ la precisión con la que se aproxima $I_m(f)$ y $m \in \mathbb{Z}^+$ tal que $\left|\int_0^a E_m(x)dx\right| < \epsilon$, entonces $I(f)$ puede ser aproximado por $I_m(f) = \sum_{k=0}^{m} \frac{f^{(k)}(0)}{(k+1)!}a^{k+1}$ con una precisión $\epsilon > 0$. La condición $\left|\int_0^a E_m(x)dx\right| < \epsilon$ permite controlar el error de truncamiento.

Ejemplos

1. Sea f la función real definida por: $f(x) = \int_0^x \frac{\text{sen}(t^2)}{t^p}dt \quad x \in [0, \frac{\pi}{2}], \quad p < 3$. Construyamos un algoritmo para calcular los valores aproximados de $f(x)$ y apliquemos a $f\left(\sqrt{\frac{\pi}{6}}\right)$ para $p = -1$. Para el efecto apliquemos el desarrollo de Taylor de $\text{sen}(\alpha)$, tenemos $\text{sen}(\alpha) = \sum_{k=0}^{\infty}(-1)^k \frac{\alpha^{2k+1}}{(2k+1)!}$, y para $\alpha = t^2$ se tiene $\text{sen}(t^2) = \sum_{k=0}^{\infty}(-1)^k \frac{t^{4k+2}}{(2k+1)!}$. Entonces

$$\begin{aligned} f(x) &= \int_0^x \frac{\text{sen}(t^2)}{t^p}dt = \int_0^x \left(t^{-p}\sum_{k=0}^{\infty}(-1)^k \frac{t^{4k+2}}{(2k+1)!}\right)dt \\ &= \int_0^x \left(t^{-p+2}dt + t^{-p}\sum_{k=1}^{\infty}\frac{(-1)^k t^{4k+2}}{(2k+1)!}dt\right) \\ &= \int_0^x t^{-p+2}dt + \int_0^x t^{-p}\sum_{k=1}^{\infty}\frac{(-1)^k t^{4k+2}}{(2k+1)!}dt. \end{aligned}$$

Calculemos el primer término de la última igualdad, tenemos

$$\int_0^x t^{-p+2}dt = \left.\frac{t^{-p+3}}{-p+3}\right|_0^x = \frac{x^{-p+3}}{3-p},$$

de donde $-p+3 > 0$ con lo cual $p < 3$. Por otra parte, la serie $\sum_{k=0}^{\infty}(-1)^k \frac{t^{4k+2}}{(2k+1)!} \quad t \in \left[0, \frac{\pi}{2}\right]$ converge uniformemente sobre $\left[0, \frac{\pi}{2}\right]$, podemos entonces intercambiar el símbolo de sumatoria con el de integral, se tiene

$$f(x) = \frac{x^{3-p}}{3-p} + \sum_{k+1}^{\infty}\frac{(-1)^k}{(2k+1)!}\int_0^x t^{4k+2-p}dt = \sum_{k=0}^{\infty}\frac{(-1)^k}{(2k+1)!(4k+3-p)}x^{4k+3-p}.$$

Observe que si $k = 0$, se debe tener $3 - p > 0$ que implica $p < 3$. Si $p \geq 3$, la integral no es convergente para $x > 0$.

La representación de la función f en serie de potencias no puede ser usada para calcular $f(x)$ en el computador. Necesitamos aproximarle con una suma finita:

$$f_m(x) = \sum_{k=0}^{m}\frac{(-1)^k x^{4k+3-p}}{(2k+1)!(4k+3-p)} \quad x \in \left[0, \frac{\pi}{2}\right].$$

Para $\epsilon = 10^{-6}$ se muestra que $|f(x) - f_m(x)| < \epsilon \ \forall x \in [0, \frac{\pi}{2}]$ y $m \geq 7$. Para $m = 7$, definimos $f_7(x) = x^{3-p}(p_1(x) - p_2(x)) \quad x \in [0, \frac{\pi}{2}]$, donde $p_1(x)$ y $p_2(x)$ son los polinomios obtenidos de f_m con los índices pares e impares, respectivamente. Luego

$$\begin{aligned} p_1(x) &= \frac{1}{3-p} + \frac{x^8}{5!}\left(\frac{1}{11-p} + \frac{x^8}{6\times 7\times 8\times 9}\left(\frac{1}{19-p} + \frac{x^8}{10\times 11\times 12\times 13(27-p)}\right)\right), \\ p_2(x) &= \frac{x^4}{3!}\left(\frac{1}{7-p} + \frac{x^8}{4\times 5\times 6\times 7}\left(\frac{1}{15-p} + \frac{x^8}{8\times 9\times 10\times 11}\left(\frac{1}{23-p} + \frac{x^8}{12\times 13\times 14\times 15(31-p)}\right)\right)\right). \end{aligned}$$

Apliquemos este algoritmo para aproximar $f\left(\sqrt{\frac{\pi}{6}}\right)$ para $p = -1$. Primeramente, notemos que para $p = -1$,

$$f(x) = \int_0^x \frac{\text{sen}\, t^2}{t^{-1}}dt = \int_0^x t\,\text{sen}\, t^2 dt = \frac{1}{2}(1 - \cos x^2).$$

En consecuencia:

$$f\left(\sqrt{\frac{\pi}{6}}\right)=\frac{1}{2}\left(1-\cos\frac{\pi}{6}\right)\simeq 0{,}0669873.$$

Apliquemos el algoritmo. Tomemos en consideración que $x=\sqrt{\frac{\pi}{6}}\simeq 0{,}7236012$, luego

$$p_1\left(\sqrt{\frac{\pi}{6}}\right)\simeq 0{,}2500522,\qquad p_2\left(\sqrt{\frac{\pi}{6}}\right)\simeq 0{,}00571183,$$

entonces

$$f_7\left(\sqrt{\frac{\pi}{6}}\right)=\left(\sqrt{\frac{\pi}{6}}\right)^4\left(p_1\left(\sqrt{\frac{\pi}{6}}\right)-p_2\left(\sqrt{\frac{\pi}{6}}\right)\right)\simeq 0{,}0669873.$$

3.9. Función error

Definición 7 *La función error se nota* err *y se define como sigue:*

$$\text{err}:\begin{cases}[0,\infty[\longrightarrow\mathbb{R}\\ \qquad x\longmapsto \text{err}\,(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.\end{cases}$$

Se sabe que la integral $\int_0^x e^{-t^2}dt\quad x\geq 0$, no puede calcularse con funciones elementales por lo que se debe recurrir a la aproximación numérica de la misma. Por otro lado se demuestra (véase en el siguiente capítulo la función gama de Euler) que $\int_0^\infty e^{-t^2}dt=\frac{\sqrt{\pi}}{2}$ con lo que

$$\text{err}\,(x)\underset{x\to\infty}{\longrightarrow}1.$$

Además, se prueba que la función real u definida como

$$u\,(x,t)=\text{err}\left(\frac{x}{2\sqrt{at}}\right)=\frac{2}{\sqrt{\pi}}\int_0^{\frac{x}{2\sqrt{at}}}e^{-v^2}dv,$$

donde $a>0$ constante, $x\geq 0$, $t>0$, es solución de la ecuación en derivadas parciales del tipo parabólico:

$$\frac{\partial u}{\partial t}(x,t)-a\frac{\partial^2 u}{\partial x^2}(x,t)=0\quad x>0,\ t>0.$$

Esta ecuación aparece en los problemas de transferencia de calor tales como los de conducción inestable; en mecánica de fluidos en los problemas de capa límite térmica, la ecuación de Navier-Stokes para corrientes laminares no estacionarias donde la presión es constante en todo el campo; en los problemas de difusión de contaminantes en el aire así como en los problemas de filtración de contaminantes en el suelo. Es por esto que dedicamos esta sección a la aproximación numérica de la función error.

Sea $\varepsilon>0$. Con fines prácticos elegimos $\varepsilon=10^{-10}$. Determinemos $r>0$ tal que

$$\left|\frac{2}{\sqrt{\pi}}\int_0^\infty e^{-t^2}dt-\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt\right|<\varepsilon\ \text{ si }\ x>r,$$

es decir

$$\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt<\varepsilon\ \text{ si }\ x>r.$$

Apliquemos el criterio de comparación para integrales impropias (este criterio es muy similar al de comparación de series numéricas). Sea $f\,(t)=\frac{2}{\sqrt{\pi}}\exp\left(-t^2\right)\quad t\geq 0$. Puesto que $\int_1^\infty\frac{dt}{t^2}=1$, elegimos $g\,(t)=\frac{1}{t^2}\quad t\geq 1$, y definimos $h\,(t)=\frac{f\,(t)}{g\,(t)}\quad t\geq 1$. Tenemos,

$$h\,(t)=\frac{f\,(t)}{g\,(t)}=\frac{2}{\sqrt{\pi}}t^2\exp\left(-t^2\right)\underset{t\to\infty}{\longrightarrow}0,$$

luego, existe $r > 0$ tal que $\dfrac{f(t)}{g(t)} < \varepsilon$ si $t \geq r$ y de esta desigualdad se sigue que

$$\int_x^\infty f(f)\,dt < \varepsilon \int_x^\infty g(t)\,dt \leq \varepsilon \int_1^\infty g(t)\,dt = \varepsilon \quad \text{si} \quad x \geq r.$$

De la condición $h(t) = \dfrac{f(t)}{g(t)} = \dfrac{2}{\sqrt{\pi}} t^2 \exp\left(-t^2\right) < \varepsilon = 10^{-10}$ determinemos $r > 1$. Para $t = 4$ se tiene $\dfrac{2}{\sqrt{\pi}} \times 16 \exp(-16) \simeq 2{,}032 \times 10^{-6}$, para $t = 5$, se tiene $\dfrac{2}{\sqrt{\pi}} \times 25 \exp(-25) \simeq 3{,}92 \times 10^{-10}$, para $t = 5{,}5$ resulta $\dfrac{2}{\sqrt{\pi}} \times (5{,}5)^2 \exp(-30{,}25) \simeq 2{,}49 \times 10^{-12} < \varepsilon = 10^{-10}$.

Elegimos $r = 5{,}5$. Tenemos $\dfrac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt < \varepsilon$ si $x \geq r = 5{,}5$. Así, la aproximación de la función error se reduce al intervalo $[0, 5{,}5]$. Note que para $t = 6$ se tiene $h(6) \simeq 9{,}42 \times 10^{-15}$, $t = 8$ se tiene $h(8) \simeq 1{,}16 \times 10^{-26}$, $t = 10$ se tiene $h(10) \simeq 4{,}2 \times 10^{-42}$.

Adicionalmente, para $a > 1$ calculemos $\dfrac{2}{\sqrt{\pi}} \int_a^\infty e^{-t^2} dt$ con una precisión $\varepsilon = 10^{-10}$. Tenemos

$$1 = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \left(\int_0^a e^{-t^2} dt + \int_a^\infty e^{-t^2} dt \right),$$

de donde $\operatorname{err}(a) = \dfrac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt = 1 - \dfrac{2}{\sqrt{\pi}} \int_a^\infty e^{-t^2} dt$.

Apliquemos el método de integración por partes

$$\begin{aligned}
\int_a^\infty e^{-t^2} dt &= \int_a^\infty \frac{-2te^{-t^2}}{-2t} dt = -\left. \frac{e^{-t^2}}{2t} \right|_a^\infty - \int_a^\infty \frac{e^{-t^2}}{2t^2} dt \\
&= -\left. \frac{e^{-t^2}}{2t} \right|_a^\infty - \int_a^\infty \frac{-2te^{-t^2}}{-4t^3} dt \\
&= -\left. \frac{e^{-t^2}}{2t} \right|_a^\infty - \left(-\left. \frac{e^{-t^2}}{4t^3} \right|_a^\infty - \int_a^\infty \frac{3e^{-t^2}}{4t^4} dt \right) \\
&= -\left. \frac{e^{-t^2}}{2t} \right|_a^\infty + \left. \frac{e^{-t^2}}{4t^3} \right|_a^\infty + \frac{3}{4} \int_a^\infty \frac{-2te^{-t^2}}{-2t^5} dt \\
&= -\left. \frac{e^{-t^2}}{2t} \right|_a^\infty + \left. \frac{e^{-t^2}}{4t^3} \right|_a^\infty - \left. \frac{3e^{-t^2}}{8t^5} \right|_a^\infty - \frac{15}{8} \int_a^x \frac{e^{-t^2}}{t^6} dt.
\end{aligned}$$

Continuando con este procedimiento n veces, obtenemos

$$\begin{aligned}
\int_a^\infty e^{-t^2} dt = &-\left. \frac{e^{-t^2}}{2t} \right|_a^\infty + \left. \frac{e^{-t^2}}{2^2 t^3} \right|_a^\infty - \left. \frac{1 \times 3 \times e^{-t^2}}{2^3 t^5} \right|_a^\infty + \left. \frac{1 \times 3 \times 5 \times e^{-t^2}}{2^4 t^7} \right|_a^\infty + \cdots + \\
&\left. \frac{(-1)^n \times 1 \times 3 \times \cdots \times (2n-1)\, e^{-t^2}}{2^n t^{2n}} \right|_a^\infty + \varepsilon_n(a),
\end{aligned}$$

con

$$\varepsilon_n(a) = \frac{(-1)^n \times 1 \times 3 \times \cdots \times (2n+1)}{2^n} \int_a^\infty \frac{e^{-t^2}}{t^{2n}} dt.$$

Como para $k = 1, 2, \ldots, n$

$$\left. \frac{e^{-t^2}}{t^{2k}} \right|_a^\infty = \lim_{t \to \infty} \frac{1}{t^{2k} e^{t^2}} - \frac{1}{a^{2k} e^{a^2}} = -\frac{1}{a^{2k} e^{a^2}},$$

entonces

$$\int_a^{\infty} e^{-t^2}dt = \frac{e^{-a^2}}{2a}\left(1-\frac{1}{2a^2}+\frac{1\times 3}{2^2a^4}-\frac{1\times 3\times 5}{2^3a^6}+\cdots+\frac{(-1)^{n+1}\times 1\times 3\times 5\times\cdots\times(2n-1)}{2^n a^{2n}}\right)+\varepsilon_n(a)$$

Estimemos $|\varepsilon_n(a)|$. Tenemos

$$\begin{aligned}|\varepsilon_n(a)| &= \frac{1\times 3\times\cdots\times(2n+1)}{2^n}\left|\int_a^{\infty}\frac{e^{-t^2}}{t^{2n}}dt\right| \leq \frac{1\times 3\times\cdots\times(2n+1)}{2^n}e^{-a^2}\int_a^{\infty}\frac{dt}{t^{2n}}\\ &= \frac{1\times 3\times\cdots\times(2n+1)}{2^n}e^{-a^2}\left[\frac{t^{-2n+1}}{-2n+1}\Big|_a^{\infty}\right] = \frac{1\times 3\times\cdots\times(2n+1)}{2^n(2n-1)a^{2n+1}e^{a^2}}.\end{aligned}$$

Definimos $\theta_n(a) = \frac{e^{-a^2}}{2a}\left(1-\frac{1}{2a^2}+\frac{1\times 3}{2^2a^4}-\frac{1\times 3\times 5}{2^3a^6}+\cdots+\frac{(-1)^{n+1}\times 1\times 3\times 5\times\cdots\times(2n-1)}{2^n a^{2n}}\right)$.

Puesto que:

$$\begin{aligned}\operatorname{err}(a) &= 1-\frac{2}{\sqrt{\pi}}\int_a^{\infty}e^{-t^2}dt = 1-\frac{2}{\sqrt{\pi}}\left[\theta_n(a)+\varepsilon(a)\right]\\ &= 1-\frac{2}{\sqrt{\pi}}\theta_n(a)-\frac{2}{\sqrt{\pi}}\varepsilon_n(a) \quad a>1.\end{aligned}$$

Determinemos un apropiado $n\in\mathbb{Z}^+$ y $a>1$ tal que $|\varepsilon_n(a)|<\varepsilon=10^{-10}$. Lo hacemos por tanteo.

Para $a=3$, tenemos

$$\begin{aligned}|\varepsilon_3(3)| &\leq \frac{1\times 3\times 5\times 7}{2^3\times 5\times 3^7e^9}\simeq 1{,}48\times 10^{-7},\\ |\varepsilon_6(3)| &\leq \frac{1\times 3\times 5\times 7\times 9\times 11}{2^6\times 11\times 3^{13}e^9}\simeq 1{,}443\times 10^{-9},\\ |\varepsilon_{10}(3)| &\leq \frac{1\times 3\times\cdots\times 19}{2^{10}\times 19\times 3^{21}e^9}\simeq 3{,}6\times 10^{-10}.\end{aligned}$$

Como vemos, para $a=3$ no se logra la precisión deseada. Elegimos $a=3{,}5$. Entonces

$$|\varepsilon_6(3{,}5)|\leq\frac{1\times 3\times 5\times 7\times 9\times 11\times 13}{2^6\times 11\times(3{,}5)^{13}e^{12{,}25}}\simeq 7{,}77\times 10^{-11}.$$

Note qie si $a=4$,

$$|\varepsilon_6(4)|\leq\frac{1\times 3\times 5\times 7\times 9\times 11\times 13}{2^6\times 11\times 4^{13}e^{16}}\simeq 3{,}22\times 10^{-13}.$$

Se prueba que $|\varepsilon_6(a)|<\varepsilon=10^{-10}$ para $a\geq 3{,}5$. Los resultados anteriores nos permiten definir la función φ_r siguiente:

$$\varphi_r(x)=\begin{cases}\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt, & \text{si } 0\leq x\leq 3{,}5,\\ 1-\frac{2}{\sqrt{\pi}}\theta_6(x), & \text{si } 3{,}5<x\leq 6,\\ 1, & \text{si } x>6.\end{cases}$$

Nos queda aproximar $\int_0^x e^{-t^2}dt \quad x\in[0,3{,}5]$. Para el efecto, utilizamos el desarrollo de Taylor de $\exp(\alpha)$. Tenemos $\exp(\alpha)=\sum_{k=0}^{\infty}\frac{\alpha^k}{k!}$, haciendo $\alpha=-t^2$ se obtiene $\exp\left(-t^2\right)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}t^{2k}$. Aplicando el teorema de la convergencia uniforme y la integración, se tiene

$$\begin{aligned}\int_0^x e^{-t^2}dt &= \int_0^x\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}t^{2k}dt = \sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\int_0^x t^{2k}dt = \sum_{k=0}^{\infty}\frac{(-1)^k t^{2k+1}}{k!(2k+1)}\\ &= t\sum_{k=0}^{\infty}\frac{(-1)^k}{k!(2k+1)}t^{2k}.\end{aligned}$$

Sera $m \in \mathbb{Z}^+$ y $S_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{k!(2k+1)} \quad x \in [0,3{,}5]$.

Para obtener un algoritmo de cálculo de $\operatorname{err}(x) \quad x \in [0,3{,}5]$, con la precisión fijada $\varepsilon = 10^{-10}$, determinemos $m \in \mathbb{Z}^+$ el más pequeño posible tal que

$$\left| \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{k!(2k+1)} - S_m(x) \right| \leq \sum_{k=m+1}^{\infty} \frac{x^{2k}}{k!(2k+1)} < \varepsilon.$$

Es claro que si $x \in [0,1]$ se requiere menos términos que si $x \in [3,3{,}5]$. Por esta razón consideramos los intervalos $[0,1]$, $]1,2]$, $]2,3]$, $]3,3{,}5]$.

Apliquemos el criterio de comparación. Para $a \in [0,3{,}5]$, ponemos $a_k = \frac{a^{2k}}{k!(2k+1)}$ y elegimos $b_k = \frac{1}{k(k+1)}$ que como es conocido $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$. Definimos $C_k = \frac{a_k}{b_k} \quad k = 1,2,\ldots$, entonces

$$C_k = \frac{k(k+1)}{k!(2k+1)} a^{2k} \underset{k\to\infty}{\longrightarrow} 0,$$

luego existe $m \in \mathbb{Z}^+$ tal que $\frac{a_k}{b_k} < \varepsilon$ si $k \geq m$, y de esta relación se obtiene

$$\left| \sum_{k=m+1}^{\infty} \frac{(-1)^k}{k!(2k+1)} x^{2k} \right| \leq \sum_{k=m+1}^{\infty} \frac{x^{2k}}{k!(2k+1)} < \varepsilon \quad \text{si} \quad x \leq a.$$

El más pequeño $m \in \mathbb{Z}^+$ (no óptimo) se obtiene de la desigualdad $\frac{k(k+1)}{k!(2k+1)} a^{2k} < \varepsilon = 10^{-10}$.

Para $a = 1$, se obtiene $m = 15$, para $a = 2$ resulta $m = 27$, para $a = 3$ se tiene $m = 43$ y para $a = 3{,}5$ es $m = 55$.

Con todos estos resultados, definimos la función φ como sigue:

$$\varphi_\epsilon(x) = \begin{cases} \frac{2x}{\sqrt{\pi}} \sum_{k=0}^{15} \frac{(-1)^k}{k!(2k+1)} x^{2k}, & \text{si} \quad x \in [0,1], \\ \frac{2x}{\sqrt{\pi}} \sum_{k=0}^{27} \frac{(-1)^k}{k!(2k+1)} x^{2k}, & \text{si} \quad x \in]1,2], \\ \frac{2x}{\sqrt{\pi}} \sum_{k=0}^{43} \frac{(-1)^k}{k!(2k+1)} x^{2k}, & \text{si} \quad x \in]2,3], \\ \frac{2x}{\sqrt{\pi}} \sum_{k=0}^{55} \frac{(-1)^k}{k!(2k+1)} x^{2k}, & \text{si} \quad x \in]3,3{,}5], \\ 1 - \frac{2x}{\sqrt{\pi}} \theta_6(x), & \text{si} \quad x \in]3{,}5,6], \\ 1, & \text{si} \quad x > 6. \end{cases}$$

Esta función φ_ϵ aproxima a la función error con una precisión $\varepsilon = 10^{-10}$, tenemos

$$\|\varphi_\epsilon - \operatorname{err}\|_\infty = \max_{x \in [0,\infty[} |\varphi_\epsilon(x) - \operatorname{err}(x)| < \varepsilon,$$

donde $\|\cdot\|_\infty$ denota la norma de Chebyshev (véase en el apéndice los espacios normados).

Sea $m \in \mathbb{Z}^+$ impar. Se pone $n = \frac{m-1}{2}$. Para escribir un algoritmo simple de cálculo asociamos los términos con signo positivo y aquellos con signo negativo. Tenemos

$$\sum_{k=0}^{m} \frac{(-1)^k}{k!(2k+1)} x^{2k} = \sum_{k=0}^{p} \frac{x^{4k}}{(2k)!(4k+1)} - \sum_{k=0}^{p} \frac{x^{2(2k+1)}}{(2k+1)!(4k+3)}$$

y definimos $\psi_1,\ \psi_2,\ \psi_3$ como sigue:

$$\begin{aligned}\psi_1(x) &= \sum_{k=0}^{p}\frac{x^{4k}}{(2k)!\,(4k+1)} = 1+\frac{x^4}{2!\times 5}+\frac{x^8}{4!\times 9}+\frac{x^{12}}{6!\times 13}+\cdots+\frac{x^{4p}}{(2p)!\,(4p+1)}\\ &= 1+\frac{x^4}{2}\left(\frac{1}{5}+\frac{x^4}{3\times 4}\left(\frac{1}{9}+\frac{x^4}{5\times 6}\left(\frac{1}{9}+\cdots+\frac{x^4}{(2p-3)\,(2p-2)}\left(\frac{1}{4p-3}+\right.\right.\right.\right.\\ &\quad \left.\left.\left.\left.+\frac{x^4}{(2p-1)\,(2p)\,(4p+1)}\right)\cdots\right)\right)\right),\end{aligned}$$

$\psi_2(x)=\sum\limits_{k=0}^{p}\dfrac{x^{4k+2}}{(2k+1)!\,(4k+1)}$ y se escribe en forma similar a $\psi_1(x)$.

Finalmente,

$$\begin{aligned}\psi_3(x) &= 1-\frac{2}{\sqrt{\pi}}\theta_6(x)\\ &= 1-\frac{e^{-x^2}}{\sqrt{\pi}x}\left[1+\frac{1}{x^4}\left(\frac{3}{4}+\frac{1}{x^4}\left(\frac{105}{16}+\frac{10395}{64x^4}\right)\right)-\frac{1}{2x^2}\left(1+\frac{1}{x^4}\left(\frac{15}{4}+\frac{945}{16x^4}\right)\right)\right].\end{aligned}$$

Ejemplo

En la tabla siguiente se dan algunos valores aproximados de $\mathrm{err}(x)$ para los valores x que se indican calculados con la función $\varphi_\epsilon(x)$ con una precisión $\varepsilon=10^{-3}$

x	0,2	0,4	0,6	0,8	1,0	1,2	1,5	2,0
$\mathrm{err}(x)$	0,223	0,428	0,604	0,742	0,843	0,910	0,966	0,995

.

3.10. Aproximación numérica de una integral elíptica

En esta sección consideramos la aproximación numérica de una clase de integrales elípticas, más exactamente la aproximación numérica de la integral elíptica incompleta de segunda especie que es a su vez conocida como forma de Legendre para la integral elíptica de segunda especie. Esta integral se define como sigue:

$$E(k)=\int_0^{\frac{\pi}{2}}\sqrt{1-k^2\,\mathrm{sen}^2(t)}dt \qquad \text{para } 0\le k\le 1,$$

y se presenta en el cálculo de la longitud de un arco de la elipse, también aparece en la solución de algunas ecuaciones diferenciales ordinarias. El interés de la aproximación numérica de esta clase de integrales es la de proporcionar de una metodología que puede ser implementada para la aproximación numérica de otros tipos de integrales elípticas, que como se ha dicho aparecen en algunas aplicaciones. Cabe señalar que la integral elíptica incompleta de segunda especie no puede calcularse mediante funciones elementales cuando $k\in]0,1[$.

La función real g definida sobre $[0,\frac{\pi}{2}]\times[0,1]$ como $g(k,t)=\sqrt{1-k^2\,\mathrm{sen}^2(t)}\qquad t\in[0,\frac{\pi}{2}],\ \ k\in[0,1]$, es continua, y la integral $\int_0^{\frac{\pi}{2}}g(k,t)dt$ es dependiente del parámetro $k\in[0,1]$. Por el teorema de la continuidad de integrales dependientes de un parámetro, resulta que la función E definida sobre $[0,1]$ como $E(k)=\int_0^{\frac{\pi}{2}}g(k,t)dt$ es continua sobre $[0,1]$.

Nos interesamos en el cálculo de $E(k)$ cuando $k\in]0,1[$. Para el efecto, representaremos $E(k)$ como una serie de potencias. Primeramente utilizaremos la serie binómica y el teorema de la convergencia uniforme y la integración. La serie de potencias será utilizada para elaborar un algoritmo para aproximar $E(k)$ con una precisión $\epsilon=10^{-6}$ de modo que se adapte a la estabilidad numérica, y, finalmente aplicaremos el algoritmo para aproximar $E(0{,}5)$.

Para $|x| < 1$ y $\alpha \in \mathbb{Q}\mathbb{N}\mathbb{Z}$, la serie binómica está definida como la serie:

$$(1-x)^{\alpha} = 1 - \alpha x + \frac{\alpha(\alpha-1)}{2!}x^2 - \frac{\alpha(\alpha-1)(\alpha-2)}{3!}x^3 + \cdots,$$

y para $\alpha = \frac{1}{2}$, se tiene

$$\begin{aligned}(1-x)^{\frac{1}{2}} &= 1 - \frac{1}{2}x + \frac{1}{2!}\frac{1}{2}\left(-\frac{1}{2}\right)x^2 - \frac{1}{3!}\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)x^3 + \cdots \\ &= 1 - \frac{1}{2}x - \frac{1}{2!}\frac{1}{2^2}x^2 - \frac{1}{3!}\left(\frac{1\times 3}{2^3}\right)x^3 + \cdots\end{aligned}$$

Esta serie es absolutamente convergente para todo $x \in]-1,1[$, por lo que se aplica el teorema de la convergencia uniforme y la integración. Haciendo $x = k^2\operatorname{sen}^2(t)$, se deduce que

$$\begin{aligned}E(k) &= \int_0^{\frac{\pi}{2}} \sqrt{1-k^2\operatorname{sen}^2(t)}dt \\ &= \int_0^{\frac{\pi}{2}} \left(1 - \frac{1}{2}k^2\operatorname{sen}^2(t) - \frac{1}{2!}\frac{1}{2^2}\left(k^2\operatorname{sen}^2(t)\right)^2 - \frac{1}{3!}\left(\frac{1\times 3}{2^3}\right)\left(k^2\operatorname{sen}^2(t)\right)^3 + \cdots\right)dt \\ &= \frac{\pi}{2} - \frac{1}{2}k^2\int_0^{\frac{\pi}{2}}\operatorname{sen}^2(t)dt - \frac{1}{2!}\frac{k^4}{2^2}\int_0^{\frac{\pi}{2}}\operatorname{sen}^4(t)dt - \frac{1}{3!}\frac{1\times 3}{2^3}k^6\int_0^{\frac{\pi}{2}}\operatorname{sen}^6(t)dt + \cdots.\end{aligned}$$

Sea $I(j) = \int_0^{\frac{\pi}{2}}\operatorname{sen}^{2j}(t)dt$, $j = 1, 2, \ldots$. Apliquemos el método de integración por partes. Tenemos

$$\begin{aligned}I(j) &= \int_0^{\frac{\pi}{2}}\operatorname{sen}^{2j}(t)dt = \int_0^{\frac{\pi}{2}}\operatorname{sen}(t)\ \operatorname{sen}^{2j-1}(t)dt \\ &= -\cos(t)\ \operatorname{sen}^{2j-1}(t)|_0^{\frac{\pi}{2}} + (2j-1)\int_0^{\frac{\pi}{2}}\operatorname{sen}^{2j-2}(t)\cos^2(t)dt \\ &= (2j-1)\int_0^{\frac{\pi}{2}}(\operatorname{sen}^{2j-2}(t))\ (1-\operatorname{sen}^2(t))dt = (2j-1)\left[\int_0^{\frac{\pi}{2}}\operatorname{sen}^{2j-2}(t)dt - \int_0^{\frac{\pi}{2}}\operatorname{sen}^{2j}(t)dt\right] \\ &= (2j-1)\left[I(j-1) - I(j)\right],\end{aligned}$$

y de este resultado obtenemos la siguiente fórmula de recursividad

$$I(j) = \frac{2j-1}{2j}I(j-1) \quad j = 1, 2, \ldots$$

Utilizando esta fórmula de recursividad, se obienen los siguientes resultados:

$$\begin{aligned}I(0) &= \int_0^{\frac{\pi}{2}}dt = \frac{\pi}{2}, \quad I(1) = \frac{1}{2}I(0) = \frac{1}{2}\times\frac{\pi}{2}, \quad I(2) = \frac{3}{4}I(1) = \frac{1\times 3}{2^2\times 1\times 2}\frac{\pi}{2}, \\ I(3) &= \frac{5}{6}I(2) = \frac{1\times 3\times 5}{2^3\times 1\times 2\times 3}\frac{\pi}{2}, \cdots, \quad I(j) = \frac{1\times 3\times 5\times\cdots\times(2j-1)}{2^j\times j!}\frac{\pi}{2}.\end{aligned}$$

Remplazando cada uno de estos resultados en la representación de $E(k)$, obtenemos la serie de potencias

$$\begin{aligned}E(k) &= \frac{\pi}{2} - \frac{1}{2}k^2\left(\frac{1}{1\times 2}\frac{\pi}{2}\right) - \frac{1}{2!}\frac{k^4}{2^2}\left(\frac{1\times 3}{2^2\times 2!}\frac{\pi}{2}\right) - \frac{1}{3!}\left(\frac{1\times 3}{2^3}\right)k^6\left(\frac{1\times 3\times 5}{2^3\times 3!}\frac{\pi}{2}\right) - \cdots \\ &= \frac{\pi}{2}\left[1 - \frac{k^2}{2^2} - \frac{1}{2!}\frac{1\times 3}{2^2\times 2^2}k^4 - \frac{1\times 3}{2^3\times 3!}\frac{1\times 3\times 5}{2^3\times 3!}k^6 - \ldots\right] \\ &= \frac{\pi}{2}\left[1 - \frac{k^2}{2^2} - \left(\frac{1\times 3}{2\times 4}\right)^2\frac{k^4}{3} - \left(\frac{1\times 3\times 5}{2\times 3\times 6}\right)^2\frac{k^6}{5} - \ldots\right] \\ &= \frac{\pi}{2}\left[1 - \sum_{j=1}^{\infty}\left(\frac{1\times 3\times\ldots\times(2j-1)}{2\times 4\times\ldots\times 2j}\right)^2\frac{k^{2j}}{2j-1}\right].\end{aligned}$$

En conclusión, la integral elíptica incompleta de segunda especie se representa como la siguiente serie de potencias:

$$E(k) = \frac{\pi}{2}\left[1 - \sum_{j=1}^{\infty}\left(\frac{1\times 3\times\ldots\times(2j-1)}{2\times 4\times\ldots\times 2j}\right)^2 \frac{k^{2j}}{2j-1}\right] \qquad \text{k} \in]0,1[.$$

Para $\varepsilon > 0$, determinemos si es posible, el más pequeño número de términos n tal que $\mid E(k) - E_n(k) \mid < \varepsilon$. Para el efecto aplicamos el criterio de comparación de series. Sean $a_j(k) = \left(\frac{1\times 3\times\ldots\times(2j-1)}{2\times 4\times\ldots\times 2j}\right)^2 \frac{k^{2j}}{2j-1}$, $b_j = \frac{1}{j(j+1)}$. Se tiene que $\sum\limits_{j=1}^{\infty}\frac{1}{j(j+1)} = 1$ y para $0 < k < 1$,

$$\frac{a_j(k)}{b_j} = \frac{j(j+1)}{2j-1}\left(\frac{1\times 3\times\ldots\times(2j-1)}{2\times 4\times\ldots\times 2j}\right)^2 k^{2j} \underset{j\longrightarrow\infty}{\longrightarrow} 0,$$

luego, $\forall\epsilon > 0$, $\exists n \in \mathbb{N}$ tal que $\frac{a_j(k)}{b_j} < \epsilon$ si $j \geq n > 1$, de donde

$$\left|\sum_{j=1}^{\infty} a_j(k) - \sum_{j=1}^{n} a_j(k)\right| = \sum_{j=n+1}^{\infty} a_j(k) < \epsilon.$$

Sea $E_n(k) = \frac{\pi}{2}\left(1 - S_n(k)\right)$, con

$$\begin{aligned} S_n(k) &= \sum_{j=2}^{n}\left(\frac{1\times 3\times\ldots\times(2j-1)}{2\times 4\times\ldots\times 2j}\right)^2 \frac{k^{2j}}{2j-1} \\ &= \frac{k^2}{2^2} + \left(\frac{1\times 3}{2\times 4}\right)^2 \frac{k^4}{3} + \left(\frac{1\times 3\times 5}{2\times 4\times 6}\right)^2 \frac{k^6}{5} + \cdots + \left(\frac{1\times 3\times\cdots\times(2n-1)}{2\times 4\times\cdots\times 2n}\right)^2 \frac{k^{2n}}{2n-1} \\ &= \frac{k^2}{2^2} + \left(\frac{1\times 3}{2\times 4}\right)^2 k^4\left(\frac{1}{3} + \left(\frac{5k}{6}\right)^2\left(\frac{1}{5} + \left(\frac{7k}{8}\right)^2\left(\frac{1}{7} + \cdots + \right.\right.\right. \\ &\quad \left.\left.\left.\left(\frac{(2n-3)k}{2(n-1)}\right)^2\left(\frac{1}{2n-3} + \left(\frac{2n-1}{2n}k\right)^2 \frac{1}{2n-1}\right)\cdots\right)\right)\right). \end{aligned}$$

La escritura de $Sn(k)$ evita el cálculo directo de los coeficientes del sumatorio así como el cálculo directo de las potencias y con esto se reduce significativamente el número de operaciones elementales a realizar. Por otro lado facilita la elaboración de un algoritmo numérico, como el que se propone a continuación.

Algoritmo

Datos de entrada: k, n

Datos de salida: $E_n(k)$.

1. Poner $b = \frac{1}{2n-1}$.

2. $j = 1, \ldots, n-2$

   $m = n + 1 - j$

   $b = \frac{1}{2m-3} + b * \left(\frac{2m-1}{2m}k\right)^2$

   Fin de bucle j.

3. $S_n(k) = \frac{k^2}{4} + \left(\frac{3k^2}{8}\right)^2 * b.$

4. $E_n(k) = \frac{\pi}{2}\left(1 - S_n(k)\right).$

5. Imprimir $E_n(k)$.

6. Fin.

Apliquemos el procedimiento arriba descrito para aproximar $E(0{,}5)$ con una precisión $\varepsilon = 10^{-6}$. Para el efecto, ponemos $k = \dfrac{1}{2}$, entonces

$$\frac{a_j\left(\frac{1}{2}\right)}{b_j} = \frac{j(j+1)}{2j-1}\left(\frac{1\times 3\times\cdots\times(2j-1)}{2\times 4\times\cdots\times 2j}\right)^2\frac{1}{2^{2j}} < 10^{-6} \quad \text{si } j \geq 10.$$

Por lo tanto, $S_{10}\left(\frac{1}{2}\right)$ se expresa como sigue:

$$S_{10}\left(\frac{1}{2}\right) = \frac{1}{16}+\frac{9}{1024}\left(\frac{1}{3}+\left(\frac{5}{12}\right)^2\left(\frac{1}{5}+\left(\frac{7}{16}\right)^2\left(\frac{1}{7}+\left(\frac{9}{20}\right)^2\left(\frac{1}{9}+\right.\right.\right.\right.$$
$$\left.\left.\left.\left.\left(\frac{11}{24}\right)^2\left(\frac{1}{11}+\left(\frac{13}{28}\right)^2\left(\frac{1}{13}+\left(\frac{15}{32}\right)^2\left(\frac{1}{15}+\left(\frac{17}{36}\right)^2\left(\frac{1}{17}+\left(\frac{19}{40}\right)^2\frac{1}{19}\right)\right)\right)\right)\right)\right)\right)\right).$$

Realizando estos cálculos elementales, se obtiene $S_{10}\left(\frac{1}{2}\right) = \frac{1}{16} + 0{,}003284541926$ y el consecuencia

$$E_{10}\left(\frac{1}{2}\right) = \frac{\pi}{2}\left(1-\frac{1}{16}-0{,}003284541926\right) = 1{,}46746221.$$

3.11. Ejercicios

1. Determinar el radio de convergencia de cada una de las series que se proponen en cada item.

   a) $\displaystyle\sum_{n=1}^{\infty}\frac{x^n}{2^n}$. **b)** $\displaystyle\sum_{n=1}^{\infty}\frac{x^n}{n^n}$. **c)** $\displaystyle\sum_{n=1}^{\infty}\frac{x^{n+1}}{n}$. **d)** $\displaystyle\sum_{n=1}^{\infty}\frac{x^n}{n\,(n-1)}$. **e)** $\displaystyle\sum_{n=0}^{\infty}n^2x^n$. **f)** $\displaystyle\sum_{n=0}^{\infty}\frac{n}{(n+1)(n+2)}x^n$.

   g) $\displaystyle\sum_{n=0}^{\infty}\frac{1}{(n!)3^n}x^n$. **h)** $\displaystyle\sum_{n=0}^{\infty}\frac{x^n}{2n+1}$. **i)** $\displaystyle\sum_{n=0}^{\infty}\frac{x^n}{n!2^n}$. **j)** $\displaystyle\sum_{n=0}^{\infty}\frac{nx^n}{3^n}$. **k)** $\displaystyle\sum_{n=0}^{\infty}\frac{(-1)^nx^n}{(n+1)(n+2)(2n+3)}$.

   m) $\displaystyle\sum_{n=0}^{\infty}\frac{(-1)^nx^{2n}}{(2n!)(3n+1)}$. **n)** $\displaystyle\sum_{n=0}^{\infty}\frac{(-1)^{n+1}}{(n+1)2^n}x^n$. **o)** $\displaystyle\sum_{n=3}^{\infty}\frac{1}{(n-1)(n-2)2^n}x^{3n}$. **p)** $\displaystyle\sum_{n=2}^{\infty}\frac{2^n}{(n-1)}x^{\frac{n}{3}}$.

2. Determinar el radio de convergencia de las siguientes series de potencias:

   a) $\operatorname{senh}(x) = \displaystyle\sum_{k=0}^{\infty}\frac{x^{2k+1}}{(2k+1)!}$. **b)** $\cosh(x) = \displaystyle\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!}$.

   c) $(1+x)^{\alpha} = \displaystyle\sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$, $\alpha\in\mathbb{R}$ (serie binómica), $\binom{\alpha}{k} = \dfrac{\alpha\,(\alpha-1)\ldots(\alpha-k+1)}{k!}$ $k = 1, 2, \ldots$.

   d) $\dfrac{1}{1-x} = \displaystyle\sum_{k=0}^{\infty}x^k$. **e)** $\ln(1+x) = \displaystyle\sum_{k=0}^{\infty}\frac{(-1)^k}{k+1}x^{k+1}$.

   f) $\arctan(x) = x-\dfrac{x^3}{3}+\dfrac{x^5}{5}-\dfrac{x^7}{7}+\cdots\dfrac{(-1)^{n-1}}{2n-1}x^{2n-1}+\cdots$

   Utilice d) para obtener e) y f).

   Sean $m\in\mathbb{Z}^+$ y $S_m(x)$ una suma finita con m términos de cada serie. Escriba un algoritmo para calcular $S_m(x)$ de modo que se adapte a la estabilidad numérica, $x\in I$, donde I es el más grande subconjunto de $\mathbb{R}$ en el que la serie converge absolutamente.

3. Se sabe que $e^t = \displaystyle\sum_{k=0}^{\infty}\frac{t^k}{k!}$, $\operatorname{sen}(t) = \displaystyle\sum_{k=0}^{\infty}\frac{(-1)^{2k+1}\,t^{2k+1}}{(2k+1)!}$, $\cos(t) = \displaystyle\sum_{k=0}^{\infty}\frac{(-1)^k\,t^{2k}}{(2k)!}$, $t\in\mathbb{R}$. Aplique el teorema de la convergencia uniforme y la integración para expresar las siguientes integrales en series de potencias, $x>0$.

Para cada función f que se define a continuación, calcular una aproximación $\widetilde{f}(x)$ de $f(x)$ para el punto x que se precisa de modo que $\left|f(x)-\widetilde{f}(x)\right|<10^{-5}$ y el número de operaciones elementales que se requiere en el cálculo de $\widetilde{f}(x)$ sea el más pequeño posible (evite el cálculo directo de los factoriales y las potencias).

a) $f(x)=\int_0^x e^{-t^2}dt, \quad x=0{,}2.$ **b)** $f(x)=\int_0^x \operatorname{sen}\left(t^2\right)dt, \quad x=\sqrt{\frac{\pi}{6}}.$

c) $f(x)=\int_0^x \cos\left(\sqrt{t}\right)dt, \quad x=\dfrac{\pi}{9}.$ **d)** $f(x)=\displaystyle\int_0^x \frac{e^t-1}{t}, \quad x=0{,}1.$

e) $f(x)=\displaystyle\int_0^x \frac{1-\cos(t)}{t^2}dt, \quad x=0{,}1.$ **f)** $f(x)=\displaystyle\int_0^x \frac{\sqrt{t}-\operatorname{sen}(\sqrt{t})}{t^{\frac{3}{2}}}dt, \quad x=0{,}3.$

g) $f(x)=\displaystyle\int_0^x \sqrt{t}(e^{-t^3}-e^{-t^2}+e^{-\sqrt{5}t})dt, \quad x=0{,}2.$

4. Aproximar, en cada caso, la integral con una precisión de 10^{-8}.

a) $\int_0^{\frac{\pi}{2}} \cos\left(x^{\frac{1}{4}}\right)dx.$ **b)** $\int_0^1 t^{\frac{1}{3}}e^{-t^2}dt.$ **c)** $\int_0^{\pi} \operatorname{sen}^2\left(t^{1/2}\right)dt,\ \operatorname{sen}^2(\alpha)=\displaystyle\sum_{n=1}^{\infty}(-1)^{n+1}\frac{2^{2n-1}}{(2n)!}\alpha^{2n},\ \alpha\in\mathbb{R}.$

d) $\displaystyle\int_0^{\pi/2} \frac{dx}{\sqrt{1-\frac{1}{2}\operatorname{sen}^2(x)}}.$ **e)** $\displaystyle\int_0^{\frac{1}{2}} \frac{dx}{(1+x^4)^{1/4}}.$

5. Considerar la integral $I(p)=\int_0^1 \frac{dx}{\sqrt{1+px^4}}$ donde $p\geq 0$.

a) Utilice el binomio de Newton con exponente fraccionario para representar $I(p)$ como una serie de potencias de p.

b) Determine para que valores de $p\geq 0$ la serie de potencias es absolutamente convergente.

c) Para $p\geq 0{,}2$, determine el número más pequeño de términos que se requieren para aproximar $I(p)$ con una precisión de 10^{-4} y aproxime $I(0{,}4)$.

6. Utilice la serie $\dfrac{1}{1-\alpha}=\displaystyle\sum_{k=0}^{\infty}\alpha^k,\ |\alpha|<1$, para elaborar un algoritmo numéricamente estable que permita aproximar $I(p)=\int_0^p \frac{dx}{1+x^4} \quad 0\leq p\leq \frac{1}{2}$, con una precisión $\varepsilon=10^{-10}$.

7. Sea $\alpha\in\mathbb{R}^+$. La función de Bessel de orden α se define mediante la serie

$$f_\alpha(x)=|x|^\alpha\left(1+\sum_{n=1}^{\infty}\frac{(-1)^n x^{2n}}{2^{2n}n!(1+\alpha)(2+\alpha)\cdots(n+\alpha)}\right).$$

a) Estudie la convergencia de la serie.

b) Elabore un algoritmo que permita aproximar $f_\alpha(x)$ con una precisión $\varepsilon>0$.

8. Sean $a\geq 0,\ \ p\in\mathbb{Q}$. El binomio de Newton con exponente fraccionario se expresa mediante el siguiente desarrollo en serie de potencias:

$$(1+a)^p=1+pa+\frac{p(p-1)}{2!}a^2+\frac{p(p-1)(p-2)}{3!}a^3+\cdots$$

Aplique este desarrollo para calcular un valor aproximado $\widetilde{f}(x)$ de $f(x)$ que se define en cada caso, de modo que $\left|f(x)-\widetilde{f}(x)\right|<10^{-4}$ y el número de operaciones elementales para el cálculo de $\widetilde{f}(x)$ sea el más pequeño posible.

a) $f(x)=\displaystyle\int_0^x \frac{dt}{\sqrt{1+\frac{1}{2}t^4}} \quad x=0{,}1.$ **b)** $f(x)=\displaystyle\int_0^x \frac{dt}{\left(1-\frac{1}{4}t^4\right)^{\frac{1}{3}}} \quad x=0{,}5.$

c) $f(x)=\displaystyle\int_0^x \left(1-\frac{1}{4}t^4\right)^{\frac{3}{4}}dt \quad x=0{,}2.$ **d)** $f(x)=\displaystyle\int_0^x \left(1+\frac{1}{5}t^3\right)^{-\frac{2}{3}}dt \quad x=0{,}3.$

e) $f(x) = \int_0^x \left(1+0{,}5t^2\right)^{-\frac{1}{3}} dt \quad x \in [0,1], \quad x = 0{,}5.$ **f)** $f(x) = \int_0^x \left(1+\frac{1}{2}t^3\right)^{\frac{2}{3}} dt \quad x \in [0,1], \; x = 0{,}5.$

g) $f(x) = \int_0^x \left(1+t^3\right)^{\frac{1}{3}} dt \quad x \geq 0, \; x = 0{,}2.$ **h)** $f(x) = \int_0^x \left(1-0{,}2t^2\right)^{\frac{1}{2}} dt \quad x \in [0,1[, \; x = 0{,}3.$

9. Sea $\varepsilon = 10^{-5}$. Aplique el algoritmo para la aproximación de la integral elíptica incompleta de segunda especie en el punto $k = \dfrac{1}{3}$ y $k = 0{,}6$.

10. Aplique el algoritmo de cálculo de $\operatorname{sen}(x)$ en los siguientes casos y una preccisión $\varepsilon = 10^{-9}$.

    a) $\operatorname{sen}(-15{,}2)$. **b)** $\operatorname{sen}\left(\frac{\pi}{4}\right)$. **c)** $\operatorname{sen}\left(\frac{2\pi}{3}\right)$. **d)** $\operatorname{sen}\left(\frac{20}{3}\pi\right)$. **e)** $\operatorname{sen}(125)$.

    Compare con los resultados obtenidos directamente de una calculadora de bolsillo

11. Para $x \in \left[0, \frac{\pi}{2}\right]$ se ha propuesto un algoritmo de cálculo de $\operatorname{sen}(x)$. Elabore un algoritmo de cálculo de $\operatorname{sen}(x) \quad x \in \mathbb{R}$ que incluya los siguientes casos: $x \in \left]\frac{\pi}{2}, \pi\right]$, $x > \pi$ y $x < 0$.

12. Aplique el algoritmo de cálculo de $\exp(x)$ con una precisión de $\varepsilon = 10^{-9}$, en los siguientes casos:

    a) $\exp(0{,}2)$. **b)** $\exp(2{,}5)$. **c)** $\exp(25{,}2)$. **d)** $\exp(-0{,}3)$. **e)** $\exp(-5{,}2)$.

    Compare con los resultados obtenidos directamente de una calculadora de bolsillo.

13. Sean $a > 1$ y $n \in \mathbb{Z}^+$. Elabore un algoritmo de cálculo de $y = a^n$ y verifique en los siguientes casos.

    a) $a = 3{,}14159265$ y $n = 4$. **b)** $a = 2{,}71828184$ y $n = 9$. **c)** $a = \sqrt{2} \simeq 1{,}414213562$ y $n = 10$.

14. Sea f la función real definida como $f(x) = \tan(x) \quad x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. El cálculo de $f^{(k)}(0)$ para $k = 0, 1, \ldots, 11$ da lugar al siguiente desarrollo de Taylor.

$$\begin{aligned} f(x) &= x + \frac{2}{3!}x^2 + \frac{16}{5!}x^5 + \frac{272}{7!}x^7 + \frac{7936}{9!}x^9 + \frac{353792}{11!}x^{11} + \cdots \\ &= x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \frac{62}{2835}x^9 + \frac{1382}{155925}x^{11} + \cdots \end{aligned}$$

    Para $x \in \left[0, \frac{\pi}{10}\right]$ elabore unalgoritmo de cálculo de $\tan(x)$ usando el desarrollo precedente y calcule los siguientes valores:

    a) $\tan(0{,}1)$. **b)** $\tan\left(\frac{\pi}{18}\right)$. **c)** $\tan\left(\frac{\pi}{10}\right)$ y compare con los obtenidos en una calculadora de bolsillo. Estime el error de aproximación.

    Calcule $\tan\left(\frac{\pi}{6}\right)$ con el polinomio de grado 11 y compare con el valor obtenido en una calculadora de bolsillo.

15. La integral elíptica incompleta de primera especie se define como $F(p) = \int_0^{\frac{\pi}{2}} \dfrac{1}{\sqrt{1-p^2\operatorname{sen}^2(\theta)}} d\theta$ $p \in [0,1[$. Esta integral no se calcula con funciones elementales, por lo que se le representa mediante una serie de potencias.

    a) Estudie la continuidad de la función F sobre el intervalo $[0,1[$.

    b) Represente $F(p) \quad p \in]0,1[$, mediante serie de potencias.

    c) Sea $\varepsilon = 10^{-5}$. Construya un algoritmo para la aproximación de la integral elíptica incompleta de primera especie de modo que el número de operaciones elementales sea el más pequeño posible y aplique dichoa algoritmo en los puntos $k = \dfrac{1}{4}$ y $k = 0{,}6$.

16. Se considera la función real h definida como $h(p,x) = \int_0^x \dfrac{1}{(1+p^4t^4)^{\frac{1}{3}}} dt$, donde $p \geq 0$, $x \in [0,1]$. Estudie la función h. Utilice la serie binómica para representar la función h como una serie de potencias. Para $\varepsilon = 10^{-4}$, elabore un algoritmo para la aproximación de $h(p,x)$ de modo el número

de operaciones elemenetales sea el más pequeño entero posible. Aplique el algoritmo para calcular valores aproximados de $h(0{,}5, 0{,}2)$, $h(0{,}5, 0{,}5)$, y, $h(1, 0{,}2)$, $h(1, 1)$.

17. Considerar la integral $I = \int_0^x f(t)dt$, $x > 0$, donde f es la función representada en serie de potencias que en cada caso se define. Calcule I_n aproximación de I para el valor de x que se da de modo que $|I - I_n| < 10^{-4}$.

    a) $f(t) = \sum_{k=0}^{\infty} \dfrac{t^k}{(k+1)k2^k}$, $x = 1$. **b)** $f(t) = \sum_{k=0}^{\infty} \dfrac{(-1)^k t^{2k+1}}{k!(3k+5)}$, $x = 2$.

    c) $f(t) = \sum_{k=0}^{\infty} \dfrac{(-1)^k t^k}{(2k)!}$, $x = 3$. **d)** $f(t) = \sum_{k=0}^{\infty} \dfrac{t^{2k}}{(2k+1)(k+1)5^k}$, $x = 2$.

18. En el siguiente ejercicio

    a) Utilice la serie de Taylor de $\operatorname{sen}(x)$, $x \in \mathbb{R}$, para aproximar la integral $I = \int_0^1 x^{1/2} \operatorname{sen}(x)dx$, mediante sumas finitas con $5, 7, 9, 11$ términos.

    b) Aplique el método de los trapecios para aproximar I con $n = 5, 7, 9, 11$.

    c) Aplique el método de Euler explícito (véase el capítulo 1) para aproximar $u(1)$ solución de la ecución diferencial $\begin{cases} u'(t) = t^{1/2}\operatorname{sen}(t), t \in]0, 1[, \\ u(0) = 0, \end{cases}$ con $n = 5, 7, 9, 11$. Compare los resultados de a), b) precedentes con c).

19. Proceda de manera análoga al ejercicio precedente para aproximar la integral $I = \int_0^1 \cos(x^{1/2})dx$.

20. Considere la integral $I(p) = \int_0^1 x^{-p}e^x dx,\ p > 1$.

    a) Pruebe que $I(p) < \infty, \quad \forall p > 1$.

    b) Utilice la serie de Taylor de e^x y elabore un algoritmo para aproximar $I(p)$ con una precisión $\epsilon = 10^{-6}$.

    c) Aplique el algoritmo para aproximar $I(1{,}1)$, $I(1{,}5)$. ¿Cuántas operaciones elementales se requieren?.

21. La solución en serie de potencias de x de la ecuación de Airy: $y'' = xy,\ x \in \mathbb{R}$, viene dada por

$$\begin{aligned} y(x) &= a_0 \left[1 + \sum_{n=1}^{\infty} \frac{x^{3n}}{(3n)(3n-1)(3n-3)(3n-4) \times \cdots \times 3 \times 2}\right] \\ &\quad + a_1 \left[x + \sum_{n=1}^{\infty} \frac{x^{3n+1}}{(3n+1)(3n)(3n-2)(3n-3) \times \cdots \times 4 \times 3}\right], \end{aligned}$$

    donde a_0, a_1 son constantes reales.

    a) Elaborar un algoritmo que permite aproximar $y(x)$, $x \in [0, 1]$.

    b) Si $a_0 = a_1 = 1$, bosqueje la gráfica de la solución $y(x)$ en puntos igualmente espaciados (tómese por ejemplo $x_k = kh$, con $h = 0{,}2$, $k = 0, 1, \ldots, 5$).

22. La ecuación diferencial de Bessel de orden λ es la ecuación: $x^2y'' + xy' + (x^2 - \lambda^2)y = 0$. La función de Bessel de orden cero de primera clase se representa por

$$J_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2} \left(\frac{x}{2}\right)^{2m}.$$

    La función de Bessel de orden cero de segunda clase se representa por

$$K_0(x) = -\sum_{m=1}^{\infty} \frac{(-1)^m}{(m!)^2} \left(1 + \frac{1}{2} + \cdots + \frac{1}{m}\right) \left(\frac{x}{2}\right)^{2m} + (\ln(x))\, J_0(x).$$

    Se demuestra que estas dos funciones son soluciones de la ecuación de Bessel.

    a) Elaborar un algoritmo que permita aproximar $J_0(x)$ y $K_0(x), x \in]0, 2]$.

    b) Bosquejar las gráficas de $J_0(x)$ y $K_0(x)$, $x \in]0, 2]$ en puntos igualmente espaciados $x_k = 0{,}2k$ $k = 1, \ldots, 10$.

23. Se prueba que la solución de la ecuación en derivadas parciales:

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial u}{\partial t} &= 0 \text{ sobre }]0,T[\times]0,L[, \\ u(0,t) &= u(L,t) = 0, \forall t \in [0,T], \end{aligned}$$

$$u(x,0) = \begin{cases} x, \text{ si } 0 < x < \frac{L}{2}, \\ L - x, \text{ si } \frac{L}{2} \leq x \leq L, \end{cases}$$

donde $c > 0$, $T > 0$, $L > 0$, x es la variable espacial, t es la variable temporal y u es la temperatura; viene dada por

$$u(x,t) = \sum_{k=1}^{\infty} a_k e^{-\lambda_k^2 t} \operatorname{sen}(\frac{\pi k x}{L}) \qquad (x,t) \in [0,T] \times [0,L],$$

donde $\lambda_k = \dfrac{\pi c k}{L}$, $k = 1, 2 \ldots$, y a_k está definido por $a_k = \begin{cases} 0, & \text{si } k \text{ es par}, \\ \dfrac{4L}{\pi^2 k^2}, & \text{si } k = 1, 5, 9, \ldots \\ -\dfrac{4L}{\pi^2 k^2}, & \text{si } k = 3, 7, 11, \ldots \end{cases}$

a) Sean $m, n \in \mathbb{N}$, $h = \frac{L}{m}$, $x_i = ih$, $i = 0, 1, \ldots, m$, $I = \frac{T}{n}$, $t_j = T_j$, $j = 0, 1, \ldots, n$. Elabore un algoritmo para aproximar la solución de $u(x_i, t_j)$, $j = 0, 1, \ldots, n$, $i = 0, 1, \ldots, m$.

b) Supóngase que $c = L = 1$, $T = 2$, $m = 10$, $n = 4$. Trace las gráficas de las soluciones aproximadas a cada instante t_j con $3, 4$ y 5 términos de la serie.

24. En cada uno de los items siguientes se da una función f definida sobre un intervalo $[a, b]$ que se indica y $n \in \mathbb{Z}^+$. Represente $I(f) = \int_a^b f(x)\,dx$ como serie de potencias y aproxime dicha integral con el número de términos que se da, ¿qué precisión logra?

a) $f(x) = e^{x^2}$ $\quad x \in [0,1]$, $\ n = 5$. **b)** $f(x) = \sqrt{1 + x^4}$ $\quad x \in [-1,1]$, $\ n = 4$.

c) $f(x) = \dfrac{e^x}{x}$ $\quad x \in [1,2]$, $\ n = 5$. **d)** $f(x) = \dfrac{\ln x}{x}$ $\quad x \in [1,4]$, $\ n = 5$.

3.12. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Análisis Matemático, Segunda Edición, Editorial Reverté, Barcelona, 1982.
2. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.
3. Tom M. Apostol, Calculus, Volumen 2, Segunda Edición, Editorial Reverté, Barcelona, 1975.
4. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
5. R. M. Barbolla, M. García, J. Margalef, E. Outerelo, J. L. Pinilla. J. M. Sánchez, Introducción al Análisis Real, Editorial Alambra Universidad, Madrid, 1981.
6. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México, 2002.
7. Alan W. Bush, Perturbation Methods for Engineers and Scientists, CRC Press, Boca Raton, 1992.
8. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
9. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
10. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.

11. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.

12. C. H. Edwards, Jr., David E. Penney, Ecuaciones Diferenciales Elementales y Problemas con Condiciones en la Frontera, Tercera Edición, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1993.

13. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.

14. Waltson Fulks, Cálculo Avanzado, Editorial Limusa, México, 1973.

15. Wilfred Kaplan, Donald J. Lewis, Cálculo y Algebra Lineal, Volumen I, Primera Reimpresión, Editorial Limusa, México, 1978.

16. E. J. Hinch, Perturbation Methods, Cambridge University Press, Cambridge, 1991.

17. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.

18. R. Kent Nagle, Edward B. Saff, Arthur David Snider, Ecuaciones Diferenciales y Problemas con Valores en la Frontera, Tercera Edición, Editorial Pearson Educación, México, 2001.

19. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

20. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

21. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

22. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

23. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

24. Bhimsen K. Shivamoggi, Perturbation Methods for Differential Equations, Editorial Birkhäuser, Boston, 2003.

25. Michael Spivak, Calculus, Segunda Edición, Editorial Reverté, Barcelona, 1996.

26. Ferdinand Verhulst, Methods and Applications of Singular Perturbations: Boundary Layers and Multiple Timescale Dynamics, Editorial Springer, New York, 2005.

Capítulo 4

Aproximación de algunas funciones de distribución de probabilidad.

Resumen

La pregunta simple que nos hacemos cuando estudiamos estadística y probabilidades es ¿cómo se elaboran las tablas de datos de algunas de las funciones de distribución estadística? Este capítulo da respuesta a esta interrogante. Se abordan las principales funciones de distribución discretas: la binomial, de Poisson y se establecen criterios basados en el condicionamiento y la estabilidad para elaborar algoritmos de cálculo de estas funciones y de otras discretas. Las principales funciones de distribución continuas como son: las del tipo gama, del tipo beta, la normal, $\chi i-$cuadrada, t de Student, distribución F se aproximan mediante el uso de las series de potencias, del análisis asimptótico en unos casos, y en otros, cuando es posible calcular directamente las integrales, como polinomios. En todos los casos, se aplican los criterios de condicionamiento y de estabilidad numérica. Se debe precisar que en la mayoría de textos de estadística citados en la bibliografía, es muy limitado el tratamiento de los problemas de aproximación numérica de las funciones de distribución estadística. En la mayor parte de libros de métodos perturbación se trata la función error, y fue esta función la que motivó emprender la tarea de construir métodos de aproximación de las funciones de distribución estadística mencionados así como de las funciones gama y beta de Euler. Otras funciones de distribución estadística como la log normal pueden aproximarse fácilmente siguiendo los citerios establecidos con las otras funciones.

4.1. Introducción

El propósito fundamental de este capítulo es el de construir métodos y elaborar algoritmos de cálculo de las principales funciones de distribución en estadística para que puedan incorporarse en los programas de simulación numérica. Las funciones que tratamos son:

1. Discretas: las distribuciones binomial y de Poisson.
2. Continuas: la función gama de Euler y la distribución gama, la función beta y la distribución beta, la distribución normal, χi-cuadrada, t de Student, F de Snedekor.

Estas funciones de distribución estadística se presentan en muchos problemas tales como estimación de parámetros, intervalos de confianza, pruebas de hipótesis, control de calidad, análisis de la varianza, análisis de regresión y correlación lineal y multilineal, en la teoría de colas tales como las líneas de espera, en problemas de econometría, análisis multivariante, en problemas de optimización, etc.

Por otro lado, en la mayoría de textos de probabilidades y estadística, vienen tabulados valores de las funciones de distribución estadística arriba citados, que sin duda alguna, constituye de una gran ayuda, no obstante tienen la desventaja de ser muy limitados y en la automatización de la información, por lo

general, no se disponen a la mano. Por estas razones, es importante contar con algoritmos numéricos para elaborar programas computacionales para calcular valores de las mencionadas funciones de distribución para datos de entrada los más amplios posibles y que superen largamente a los datos proporcionados en las tablas.

Los temas del análisis matemático tales como los métodos de integración por partes y sustitución, las integrales impropias, sucesiones y series numéricas, criterios de convergencia, las sucesiones y series de funciones y la convergencia uniforme e integración y particularmente las series de potencias así como su aproximación numérica tratados en el capítulo anterior, son aplicados a los tipos de funciones de distribución estadística arriba citadas. Además se aplican los resultados de condicionamiento y la estabilidad numérica tratados en el primer capítulo, lo que permite elaborar algoritmos simples de cálculo con la precisión que se desee. Por cuestiones prácticas se ha seleccionado como precisión $\epsilon = 10^{-10}$ y la exactitud de cálculos del orden de 10^{-10}, aún cuando la metodología establecida se adapta fácilmente para $\epsilon > 0$ arbitrario. La metodología que se implementa puede adaptarse en forma inmediata a la aproximación de otras funciones de distribución estadística tanto discretas como continuas.

Al final del capítulo se provee de una amplia bibliografía.

4.2. La distribución de probabilidad binomial

Definición 1 *Una variable aleatoria X tiene una distribución binomial o distribución de Bernoulli basada en n pruebas, con probabilidad de éxito p, si su función de densidad está definida mediante*

$$f(k) = \begin{cases} 0, & si\ k \in \mathbb{Z} - \{0, 1, \ldots, n\}, \\ \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}, & si\ k = 0, 1, \ldots, n, \end{cases}$$

donde $p \in [0, 1]$, $q \in= 1 - p$ y $\left(\begin{array}{c} n \\ k \end{array} \right)$ denota el coeficiente binomial definido por $\left(\begin{array}{c} n \\ k \end{array} \right) = \dfrac{n!}{k!(n-k)!}$.

De la definición de los coeficientes binomiales se deduce que

$$\left(\begin{array}{c} n \\ k+1 \end{array} \right) = \frac{n-k}{k+1} \left(\begin{array}{c} n \\ k \end{array} \right),$$

consecuentemente

$$f(k+1) = \left(\begin{array}{c} n \\ k+1 \end{array} \right) p^{k+1} q^{n-(k+1)} = \frac{n-k}{k+1} \left(\begin{array}{c} n \\ k \end{array} \right) \frac{p}{q} p^k q^{n-k} = \frac{n-k}{k+1} r f(k),$$

con $r = \frac{p}{q}, \quad p \neq 1$.

Esta última relación nos permite elaborar un algoritmo para calcular $F(k)$, con $0 \leq k \leq n,\ k \in \mathbb{Z}$.

Algoritmo

Datos de entrada: p, k, n.

Datos de salida: $x, F(k)$.

1. $q = 1 - p$.
2. $r = \dfrac{p}{q}$.
3. $S_1 = q^n$.
4. $S = S_1$.
5. $k = 0, 1, \ldots, x - 1$

$$S_1 = \frac{n-x}{x+1} r S_1$$

$$S = S + S_1$$

Fin de bucle k.

6. $F(x) = S$.

7. Fin.

Este algoritmo presenta algunos inconvenientes por lo que debe tomarse en consideración otras alternativas en base a las propiedades de la función de distribución se se verán a continuación.

Sea $r = \frac{p}{q}$ con $p \neq 1$. De la definición de la función F se establece la siguiente escritura anidada:

$$\begin{aligned} F(k) &= \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} = q^n \sum_{j=0}^{n} \binom{n}{j} r^j \\ &= q^n \left(1 + nr\left(1 + \frac{n-1}{2} r\left(1 + \frac{n-2}{3} r\left(1 + \cdots + \frac{n-k-2}{k-1} r\right)\left(1 + \frac{n-k-1}{k}\right)\right)\right)\ldots\right) \end{aligned}$$

Por otro lado, para cada $k = 0, 1, \ldots, n$ se tiene

$$1 = (p+q)^n = \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} = \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} + \sum_{j=k+1}^{n} \binom{n}{j} p^j q^{n-j},$$

que permite obtener una forma alternativa de cálculo de $F(k)$:

$$\begin{aligned} F(k) &= 1 - \sum_{j=k+1}^{n} \binom{n}{j} p^j q^{n-j} = 1 - \sum_{j=0}^{n-k-1} \binom{n}{j+k+1} p^{j+k+1} q^{n-j-k-1} \\ &= 1 - p^{k+1} q^{n-k-1} \sum_{j=0}^{n-k-1} \binom{n}{j+k+1} r^j. \end{aligned}$$

Cuando k es aproximadamente $\leq \frac{n}{2}$ se utiliza la primera forma anidada de cálculo de $F(k)$, y si $\frac{n}{2} < k \leq n$ se utiliza su forma alternativa de cálculo de $F(k)$. Si $\frac{n}{2} < k \leq n$, la primera forma de cálculo de $F(k)$ contiene más términos que la segunda lo que incrementa los costos numéricos y si $0 \leq k \leq \frac{n}{2}$, la forma alternativa contiene más términos que la primera. Además, si n es grande, q^n está mal condicionado, mientras que q^{n-k-1} es mucho mejor que q^n.

Para la primera forma de cálculo de $F(k)$ se propone el algoritmo que se da a continuación. Se propone como ejercicio escribir $F(k)$ en forma anidada así como elaborar el respectivo algoritmo numérico.

Algoritmo

Datos de entrada: p, k, n.

Datos de salida: x, $F(k)$.

1. $q = 1 - p$.

2. $r = \frac{p}{q}$.

3. $S = 1$.

4. $j = 1, \ldots, k$

$$i = k + 1 - j$$

$$S = 1 + \frac{n-k-j}{i} r S$$

Fin de bucle i.

Fin de bucle j.

5. $S = q^n S$.

6. Fin.

Tomando en consideración las condiciones $0 \leq k \leq \frac{n}{2}$, y $\frac{n}{2} < k \leq n$, se propone como ejercicio la elaboración completa del algoritmo de cálculo de $F(k)$ así como su respectivo programa computacional. Los resultados compárelos con los provistos en tablas de textos de estadística.

4.3. Distribución de Poisson

Definición 2 *Una variable aleatoria X tiene una distribución binomial o distribución de Poisson de media $\lambda > 0$ si su función de densidad está definida por*

$$f(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{Z}^+.$$

La función de distribución está definida mediante

$$F(\lambda, m) = \sum_{k=0}^{m} f(k) = e^{-\lambda}\sum_{k=0}^{m}\frac{\lambda^k}{k!}, \ \ m \in \mathbb{Z}^+,$$

donde $\lambda > 0$ es fijo.

Puesto que

$$\lim_{m\to\infty}\sum_{k=0}^{m}\frac{\lambda^k}{k!} = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda},$$

entonces $\forall\varepsilon > 0$, $\exists n \in \mathbb{Z}^+$ tal que

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} - \sum_{k=0}^{m}\frac{\lambda^k}{k!} = \sum_{k=m+1}^{\infty}\frac{\lambda^k}{k!} < \varepsilon \quad \forall m \geq n.$$

Por otro lado,

$$1 = e^{-\lambda}\left(\sum_{k=0}^{n}\frac{\lambda^k}{k!} + \sum_{k=n+1}^{\infty}\frac{\lambda^k}{k!}\right) = e^{-\lambda}\sum_{k=0}^{n}\frac{\lambda^k}{k!} + e^{-\lambda}\sum_{k=n+1}^{\infty}\frac{\lambda^k}{k!},$$

de donde

$$F(\lambda, n) = 1 - e^{-\lambda}\sum_{k=n+1}^{\infty}\frac{\lambda^k}{k!}.$$

Para elaborar un algoritmo de cálculo de $F(\lambda, n)$ con una precisión $\varepsilon > 0$ para $\lambda > 0$ y $0 \leq m \leq n$, determinemos una condición sobre el parámetro λ, n y ε. Para el efecto, apliquemos el criterio de comparación de series numéricas.

Sean $a_k = \frac{\lambda^k}{k!}$, $b_k = \frac{1}{k(k+1)}$. Entonces $\sum_{k=1}^{\infty} b_k = 1$, y

$$\frac{a_k}{b_k} = \frac{k+1}{(k-1)!}\lambda^k \underset{k\to\infty}{\longrightarrow} 0.$$

Luego, existe n tal que $\frac{k+1}{(k-1)!}\lambda^k < \varepsilon$ si $k \geq n$. Para k suficientemente grande, se puede considerar la desigualdad.

$$\frac{\lambda^k}{k!} < \varepsilon \text{ si } k \geq n,$$

y tomando logaritmos en la misma, tenemos

$$k\ln(\lambda)-\sum_{j=1}^{k}\ln(j)<\ln(\varepsilon)\Longleftrightarrow\sum_{j=1}^{k}\ln(j)-k\ln(\lambda)>-\ln(\varepsilon).$$

La determinación de n mediante esta última expresión resulta ser numéricamente costosa. Con el propósito de obtener una expresión práctica de cálculo del más pequeño entero positivo n que satisfaga la desigualdad precedente, utilizamos la desigualdad:

$$\sum_{j=1}^{k}\ln(j)\leq\int_{1}^{k}\ln(t)\,dt=k\ln(k)-k+1,$$

donde la integral se calcula utilizando el método de integración por partes. Entonces

$$k\ln(k)-k+1-k\ln(\lambda)>-\ln(\varepsilon)$$

y de esta desigualdad se obtiene la siguiente:

$$k\ln\left(\frac{k}{\lambda e}\right)>-1-\ln(\varepsilon),$$

donde $e=2{,}71828182\ldots=\sum\limits_{k=0}^{\infty}\dfrac{1}{k!}$ es la base de los logaritmos naturales. Sea

$$n=\text{mín}\left\{k\in\mathbb{Z}^{+}\mid k\ln\left(\frac{k}{\lambda e}\right)>-1-\ln\varepsilon\right\},$$

donde $\lambda>0$ y $0<\varepsilon<1$ son fijos.

Si $\dfrac{n}{2}\leq m\leq n$, el cálculo de $F(\lambda,m)$ se vuelve numéricamente costoso. Para disminuir el costo numérico, utilizamos las siguiente relación:

$$1=e^{-\lambda}\left(\sum_{k=0}^{m}\frac{\lambda^{k}}{k!}+\sum_{k=m+1}^{n}\frac{\lambda^{k}}{k!}+\sum_{k=n+1}^{\infty}\frac{\lambda^{k}}{k!}\right)$$

y como $e^{-\lambda}\sum\limits_{k=n+1}^{\infty}\dfrac{\lambda^{k}}{k!}<\varepsilon$, despreciando este último término, $F(\lambda,m)$ se aproxima mediante

$$1-e^{-\lambda}\sum_{k=m+1}^{n}\frac{\lambda^{k}}{k!}=1-e^{-\lambda}\frac{\lambda^{m+1}}{(m+1)!}\left(1+\frac{\lambda}{m+2}\left(1+\frac{\lambda}{m+3}\left(1+\cdots+\frac{\lambda}{n-1}\right)\left(1+\frac{\lambda}{n}\right)\cdots\right)\right).$$

Con fines prácticos $\varepsilon=10^{-10}$, $\ln(\varepsilon)\backsimeq 23{,}1$, y $n=\text{mín}\left\{k\in\mathbb{Z}^{+}|k\ln\left(\dfrac{k}{e\lambda}\right)>22{,}1\right\}$. Por ejemplo para $\lambda\in\,]0,1]$ es $n=14$, para $\lambda=10$ es $n=45$ y para $\lambda=30$ se tiene $n=102$. Esta información nos permite definir la función $\tilde{F}(\lambda,m)$ para las distintas alternativas como a continuación se indica:

$$\tilde{F}(\lambda,m)=\begin{cases}e^{-\lambda}\sum\limits_{k=0}^{m}\dfrac{\lambda^{k}}{k!},\ \ 0\leq m\leq 14,m\in\mathbb{N},\ 0<\lambda\leq 1.\\ 1,\text{ si }m>14.\end{cases}$$

$$\tilde{F}(\lambda,m)=\begin{cases}e^{-\lambda}\sum\limits_{k=0}^{m}\dfrac{\lambda^{k}}{k!},\quad\text{si}\ \ 0\leq m\leq\dfrac{n}{2},\\ 1-e^{-\lambda}\sum\limits_{k=m+1}^{n}\dfrac{\lambda^{k}}{k!},\ \ \text{si}\ \ \dfrac{n}{2}<m\leq n,\ \lambda>1.\\ 1,\text{ si }\ \ m>n.\end{cases}$$

Se tiene que $\tilde{F}(\lambda,m)$ es una aproximación de $F(\lambda,m)$ con una precisión ε.

Para $\lambda > 1$ y m grande, los términos λ^{m+1} y $(m+1)!$ son muy grandes lo que puede causar problemas al momento de su cálculo en el computador. Par evitar etas molestias, se calcula como sigue:

$$\frac{\lambda^{m+1}}{(m+1)!} = \exp\left((m+1)\ln(\lambda) - \sum_{j=1}^{m+1}\ln(j)\right),$$

que mejora la estabilidad numérica. Además, para evitar el cálculo directo de los factoriales y de las potencias escribimos en forma anidada como a continuación se indica

$$\sum_{k=0}^{m}\frac{\lambda^k}{k!} = 1 + \frac{\lambda}{1}\left(1 + \frac{\lambda}{2}\left(1 + \cdots + \frac{\lambda}{m-1}\left(1 + \frac{\lambda}{m}\right)\cdots\right)\right),$$

$$\sum_{k=m+1}^{n}\frac{\lambda^k}{k!} = \frac{\lambda^{m+1}}{(m+1)!}\left(1 + \frac{\lambda}{m+2}\left(1 + \frac{\lambda}{m+3}\left(1 + \cdots + \frac{\lambda}{n-1}\left(1 + \frac{\lambda}{n}\right)\cdots\right)\right)\right)$$

que mejoran la estabilidad numérica.

En las aplicaciones prácticas de la distribución de Poisson tal como en la teoría de colas, el valor de λ está en el intervalo $]0, 30]$.

En resumen $F(\lambda, m)$ se aproxima numéricamente con una precisión $\varepsilon = 10^{-10}$ por $\tilde{F}(\lambda, m)$ definidos a continuación

1. Si $0 < \lambda \leq 1$, entonces
$$\tilde{F}(\lambda, m) = \begin{cases} e^{\cdot\lambda}\left(1 + \frac{\lambda}{1}\left(1 + \frac{\lambda}{2}\left(1 + \cdots + \frac{\lambda}{m-1}\left(1 + \frac{\lambda}{m}\right)\cdots\right)\right)\right), & \text{si } 0 \leq m \leq 14, \\ 1, \text{ si } m > 14. \end{cases}$$
2. Si $\lambda > 1$, entonces
$$\tilde{F}(\lambda, m) = \begin{cases} e^{-\lambda}\left(1 + \frac{\lambda}{1}\left(1 + \frac{\lambda}{2}\left(1 + \cdots + \frac{\lambda}{m-1}\left(1 + \frac{\lambda}{m}\right)\cdots\right)\right)\right), & \text{si } 0 \leq m \leq \frac{n}{2}, \\ 1 - \frac{\lambda^{m+1}}{(m+1)!}\left(1 + \frac{\lambda}{m+2}\left(1 + \frac{\lambda}{m+3}\left(1 + \cdots + \frac{1}{n-1}\left(1 + \frac{\lambda}{n}\right)\cdots\right)\right)\right), \\ \quad \text{si } \frac{n}{2} < m \leq n, \\ 1, \quad \text{si } m > n. \end{cases}$$
donde $n = \text{mín}\left\{k \in \mathbb{Z}^+ | k\ln\left(\frac{k}{\lambda e}\right) > 22{,}1\right\}$, $\frac{\lambda^{m+1}}{(m+1)!} = \exp\left[(m+1)\ln(\lambda) - \sum_{j=1}^{m+1}\ln(j)\right]$.

Ejercicio

Se propone la elaboración del algoritmo respectivo de cálculo de $\tilde{F}(\lambda, m)$ y su programa computacional. Compare los resultados con los proporcionados en las tablas de textos de estadística y probabilidades.

4.4. Función gama de Euler.

Definición 3 *La función gama de Euler se define como sigue:*

$$\Gamma : \begin{cases} \mathbb{R}^+ \to \mathbb{R}^+ \\ p \longmapsto \Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt. \end{cases}$$

Las propiedades mas importantes de la función gama se enuncian en el siguiente teorema.

Teorema 1

i. *La integral* $\int_0^\infty t^{p-1}e^{-t}dt$ *converge para todo* $p \in \mathbb{R}^+$ *y diverge para todo* $p \leq 0$.

ii. $\Gamma(1) = 1 \quad y \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

iii. *Para todo* $p \in \mathbb{R}^+$, $\Gamma(p+1) = p\Gamma(p)$. *En particular, si* $p = n \in \mathbb{Z}^+$, $\Gamma(n+1) = n!$.

iv. *Para todo* $n \in \mathbb{N}$, $\Gamma\left(n + \frac{1}{2}\right) = \dfrac{(2n-1)(2n-3) \times \ldots \times 1}{2^n}\sqrt{\pi}$.

v. *Sea* $p \in \mathbb{R}^+ - \mathbb{N}$. *Entonces* $\Gamma(p) = (p-1)(p-2)\cdots(p-n)\Gamma(p-n)$, *donde* $n = [p]$ *y* $[p]$ *denota el mayor entero menor menor o igual que* p.

vi. *La función gama es continua sobre* $\mathbb{R}^+$. *Además,*

$$\Gamma(\epsilon) \underset{\epsilon \to 0^+}{\longrightarrow} +\infty \qquad y \qquad \Gamma(p) \underset{p \to +\infty}{\longrightarrow} +\infty.$$

Demostración.

i) Sea $p \in \mathbb{R}^+$. Entonces

$$\Gamma(p) = \int_0^1 t^{p-1}e^{-t}dt + \int_1^\infty t^{p-1}e^{-t}dt.$$

Ponemos $I_1 = \int_0^1 t^{p-1}e^{-t}dt, \qquad I_2 = \int_1^\infty t^{p-1}e^{-t}dt.$

Si $p \geq 1$, la función $t \to t^{p-1}e^{-t}$de $[0,1]$ en $\mathbb{R}$ es continua, con lo cual I_1 existe.

Sea $0 < p < 1$. Puesto que

$$\int_0^1 t^{p-1}dt = \lim_{r \to 0}\int_r^1 t^{p-1}dt = \lim_{r \to 0}\frac{1}{p}t^p\Big|_r^1 = \frac{1}{p}\lim_{r \to 0}(1 - r^p) = \frac{1}{p},$$

se sigue que para $0 < r < 1$,

$$0 < \int_r^1 t^{p-1}e^{-t}dt < \infty.$$

Luego.

$$I_1 = \int_0^1 t^{p-1}e^{-t}dt = \lim_{r \to 0}\int_r^1 t^{p-1}e^{-t}dt \leq \lim_{r \to 0} e^{-r}\int_r^1 t^{p-1}dt = \frac{1}{p}.$$

En consecuencia, I_1 existe para todo $p \in \mathbb{R}^+$.

Mostremos la existencia de I_2. Puesto que $\int\limits_1^\infty \dfrac{dx}{x^2} = 1$, resulta

$$\lim_{t \to \infty}\frac{t^{p-1}e^{-t}}{\frac{1}{t^2}} = \lim_{t \to \infty} t^{p+1}e^{-t} = 0,$$

que es una consecuencia de la aplicación de la regla de L'Hópital. Por el criterio del cociente para integrales impropias, se deduce que I_2 existe.

Por lo tanto, $\Gamma(p) = \int\limits_0^\infty t^{p-1}e^{-t}dt$ está bien definida para todo $p > 0$.

Si $p = 0$, de la desigualdad $e^{-1}t^{-1} \leq t^{-1}e^{-t} \leq t^{-1} \quad \forall t \in]0,1[$, se sigue que $I_1 = +\infty$, pués $\int_0^1 t^{-1}dt = \ln t\big|_0^1 = +\infty$.

Si $p < 0$, la integral $I_1 = \int_0^1 t^{p-1}e^{-t}dt$ diverge. Pués

$$\int_0^1 \frac{dt}{t} = \ln t\big|_0^1 = +\infty,$$

por el criterio del cociente: para integrales impropias, se tiene

$$\lim_{t\to 0}\frac{\frac{1}{t}}{t^{p-1}e^{-t}}=\lim_{t\to 0}t^{-p}e^{+t}=0,$$

y de este resultado se obtiene la conclusión.

ii) Para $p=1$, se tiene

$$\Gamma(1)=\int_0^\infty e^{-t}dt=1.$$

Sea $p=\frac{1}{2}$. Mostremos primeramente que $\int_{-\infty}^{\infty}e^{-t^2}dt=\sqrt{\pi}$. En efecto, sea $I=\int_{-\infty}^{\infty}e^{-t^2}dt$. Entonces

$$I^2=\left(\int_{-\infty}^{\infty}e^{-t^2}dt\right)\left(\int_{-\infty}^{\infty}e^{-x^2}dt\right)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-(t^2+x^2)}dtdx.$$

Sea $r>0$ y $\overline{B(0,r)}$ el disco cerrado de centro 0 y radio r. Utilizando coordenadas polares: $\begin{cases} t=\rho\cos\theta, \\ x=\rho\,\mathrm{sen}\,\theta, \end{cases}$, donde $0\le\theta\le 2\pi,\ \ 0\le\rho\le r$, se tiene,

$$\iint\limits_{\overline{B(o,r)}}e^{-(t^2+x^2)}dtdx=\int_0^{2\pi}\int_0^r\rho e^{-\rho^2}d\rho d\theta=\int_0^{2\pi}d\theta\int_0^r\rho e^{-\rho^2}d\rho=2\pi\left(-\frac{1}{2}e^{-\rho^2}\Big|_0^r\right)=\pi(1-e^{-r}),$$

luego

$$I^2=\lim_{r\to\infty}\iint\limits_{\overline{B(0,r)}}e^{-(t^2+x^2)}dtdx=\lim_{r\to\infty}\pi(1-e^{-r})=\pi,$$

con lo que $I=\sqrt{\pi}$.

Pasemos a probar que $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$. Por definición de la función gama, se tiene

$$\Gamma\left(\frac{1}{2}\right)=\int_0^\infty t^{-\frac{1}{2}}e^{-t}dt.$$

Efectuando la sustitución $t=x^2$ en la integral indefinida precedente, resulta

$$\Gamma\left(\frac{1}{2}\right)=2\int_0^\infty e^{-x^2}dx=\int_{-\infty}^{\infty}e^{-x^2}dx=\sqrt{\pi}.$$

iii) Sea $p\in\mathbb{R}^+$. De la definición de la función gama, se tiene

$$\Gamma(p+1)=\int_0^\infty t^pe^{-t}dt.$$

Utilizando el método de integración por partes: $u=t^p$, $dv=e^{-t}dt$, se sigue que

$$\Gamma(p+1)=-t^pe^{-t}\big|_0^\infty+p\int_0^\infty t^{p-1}e^{-t}dt,$$

y mediante la aplicación de la regal de L'Hôpital para evaluar $\lim\limits_{t\to\infty}t^pe^{-t}=0$, se obtiene

$$\Gamma(p+1)=p\Gamma(p).$$

Si $p=n\in\mathbb{Z}^+$, por inducción se prueba que $\Gamma(n+1)=n!$.

iv) Sea $n \in \mathbb{N}$. Entonces, por la propiedad iii), se deduce que

$$\begin{aligned}\Gamma\left(n+\frac{1}{2}\right) &= \Gamma\left(\left(n-\frac{1}{2}\right)+1\right) = \left(n-\frac{1}{2}\right)\Gamma\left(n-\frac{1}{2}\right)\\ &= \left(n-\frac{1}{2}\right)\Gamma\left(\left(n-\frac{3}{2}\right)+1\right)\\ &= \left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\Gamma\left(n-\frac{3}{2}\right)\\ &\vdots\\ &= \left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\times\cdots\times\frac{1}{2}\Gamma\left(\frac{1}{2}\right).\end{aligned}$$

Por la propiedad ii), $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, entonces

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{(2n-1)(2n-3)\times\cdots\times 1}{2^n}\sqrt{\pi}.$$

v) Sea $p \in \mathbb{R}^+\backslash\mathbb{N}$. Denotemos con n el mayor entero menor o igual que p, entonces $p-n \in]0,1[$. Utilizando la propiedad iii) se deduce

$$\begin{aligned}\Gamma(p) &= \Gamma((p-1)+1) = (p-1)\Gamma(p-1) = (p-1)(p-2)\Gamma(p-2)\\ &\vdots\\ &= (p-1)(p-2)\cdots(p-n)\Gamma(p-n).\end{aligned}$$

vi) La demostración de la continuidad de la función Γ requiere de argumentos que están fuera del alcance de estas notas. (véase el Análisis Matemático de Apostol, el Cálculo Avanzado de Fulks).

Para $\epsilon > 0$, por la propiedad iii) se tiene

$$\Gamma(1+\epsilon) = \epsilon\Gamma(\epsilon) \Longrightarrow \Gamma(\epsilon) = \frac{\Gamma(1+\epsilon)}{\epsilon},$$

y por la continuidad de Γ se sigue que

$$\lim_{\epsilon\to 0^+}\Gamma(\epsilon) = \lim_{\epsilon\to 0^+}\frac{\Gamma(1+\epsilon)}{\epsilon} = +\infty.$$

Como $n! \underset{n\to\infty}{\longrightarrow} +\infty$, se sigue que $\Gamma(n) \underset{n\to\infty}{\longrightarrow} +\infty$. ■

4.4.1. Definición de $\Gamma(p)$ para $p < 0$ y no entero

Puesto que $\Gamma(p+1) = p\Gamma(p)$, entonces

$$\Gamma(p) = \frac{\Gamma(p+1)}{p}, \quad p > 0.$$

Si $-1 < p < 0$, entonces $0 < p+1 < 1$, por lo tanto $\dfrac{\Gamma(p+1)}{p}$ está bien definido, en cuyo caso definimos

$$\Gamma(p) = \frac{\Gamma(p+1)}{p}, -1 < p < 0.$$

Definido $\Gamma(p)$ para $p \in]-1,0[$, podemos definir $\Gamma(p)$ para $p \in]-2,-1[$ del modo siguiente:

$$\Gamma(p) = \frac{\Gamma(p+2)}{p(p+1)},$$

pues si $-2 < p < -1$ entonces $-1 < p+1 < 0$ y $0 < p+2 < 1$ con lo cual $\Gamma(p+1) = \dfrac{\Gamma(p+2)}{p+1}$ y

$$\Gamma(p) = \frac{\Gamma(p+1)}{p} = \frac{\Gamma(p+2)}{p(p+1)}.$$

Continuando con este proceso, si $n \in \mathbb{Z}^+$ y $-n < p < -n+1$, se define $\Gamma(p)$ como sigue:

$$\Gamma(p) = \frac{\Gamma(p+n)}{p(n+1)\ldots(p+n-1)} = \frac{\Gamma(p+n)}{\prod_{j=1}^{n}(n+j-1)}.$$

Note que $0 < p+n < 1$ y $\Gamma(p+n)$ está bien definido

Ejemplos

1. $\Gamma(3) = \int_0^\infty t^2 e^{-t} dt = 2$.
2. $\Gamma\left(\frac{5}{2}\right) = \Gamma\left(\dfrac{3}{2}+1\right) = \dfrac{3}{2}\Gamma\left(\dfrac{3}{2}\right) = \dfrac{3}{2}\Gamma\left(\frac{1}{2}+1\right) = \dfrac{3}{2} \times \frac{1}{2}\Gamma\left(\dfrac{1}{2}\right) = \dfrac{3}{4}\sqrt{\pi}$.

   Observe que $\Gamma\left(\frac{5}{2}\right) = \int_0^\infty t^{\frac{3}{2}} e^{-t} dt$.
3. Para calcular $\Gamma(-2{,}5)$ procedemos como a continuación se indica

$$\begin{aligned}
\Gamma(-0{,}5) &= \frac{\Gamma(0{,}5)}{-0{,}5} = -2\sqrt{\pi}, \\
\Gamma(-1{,}5) &= \frac{\Gamma(-0{,}5)}{-1{,}5} = \frac{4}{3}\sqrt{\pi}, \\
\Gamma(-2{,}5) &= \frac{\Gamma(-1{,}5)}{-2{,}5} = -\frac{8}{15}\sqrt{\pi}.
\end{aligned}$$

   Por otro lado,

$$\Gamma(-2{,}5) = \frac{\Gamma(0{,}5)}{(-2{,}5)(-1{,}5)(-0{,}5)} = -\frac{8}{15}\sqrt{\pi}.$$

4.5. Aproximación numérica de $\Gamma(p)$.

Sea $0 < p < 1$ fijo y $a > 0$. Entonces

$$\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt = \int_0^a t^{p-1} e^{-t} dt + \int_a^\infty t^{p-1} e^{-t} dt.$$

Ponemos

$$\begin{aligned}
f(a) &= \int_0^a t^{p-1} e^{-t} dt, \\
g(a) &= \int_a^\infty t^{p-1} e^{-t} dt.
\end{aligned}$$

Sea $\epsilon > 0$. Con fines prácticos $\epsilon = 10^{-10}$ con lo que $\Gamma(p)$ será aproximado con 10 cifras decimales de precisión. Para el efecto, aproximemos $f(a)$ y $g(a)$ con una precisión $\dfrac{\varepsilon}{2}$ que precisaremos más adelante.

Aproximemos primeramente $g(a)$. Integrando por partes $k+1$ veces, obtenemos

$$\begin{aligned}
g(a) &= a^{p-1}e^{-a} + (p-1)a^{p-2}e^{-a} + (p-1)(p-2)a^{p-3}e^{-a} + \cdots + \\
&\quad (p-1)(p-2)\cdots(p-k)a^{p-(k+1)}e^{-a} + (p-1)\cdots(p-(k+1))\int_a^\infty t^{p-(k+2)}e^{-t}dt.
\end{aligned}$$

Sean

$$\begin{aligned}\theta_k(a) &= e^{-a}a^{p-1}\left[1+\frac{p-1}{a}+\cdots+\frac{(p-1)(p-2)\cdots(p-k)}{a^k}\right],\\ \theta(a) &= (p-1)(p-2)\ldots(p-(k+1))\int_a^\infty t^{p-(k+2)}e^{-t}dt.\end{aligned}$$

Entonces $g(a)=\theta_k(a)+\theta(a)$. Determinemos a y k tales que $|\theta(a)|<\frac{\epsilon}{2}$, luego

$$|g(a)-\theta_k(a)|=|\theta(a)|<\frac{\epsilon}{2}.$$

Puesto que

$$\begin{aligned}|\theta(a)| &= |(p-1)(p-2)\ldots(p-(k+1))|\int_a^\infty t^{p-(k+2)}e^{-t}dt \le \left(\prod_{j=1}^{k+1}(j-p)\right)e^{-a}\int_a^\infty t^{p-(k+2)}dt\\ &= \left(\prod_{j=1}^{k+1}(j-p)\right)\frac{e^{-a}}{(k+1-p)a^{k+1-p}}.\end{aligned}$$

Cuando $p\to 0$, se tiene

$$|\theta(a)|\le\frac{(k+1)!}{(k+1)e^a a^{k+1}}=\frac{k!}{e^a a^{k+1}}.$$

Como $\dfrac{a^k}{k!}\underset{k\to\infty}{\longrightarrow}0$, la sucesión $\left(\dfrac{k!}{a^k}\right)$ no converge a 0; más aún dicha sucesión es divergente, pero si $k\in\mathbb{N}$ es fijo,

$$\frac{k!}{e^a a^{k+1}}\underset{a\to\infty}{\longrightarrow}0.$$

Notemos que

$$\frac{k!}{a^k}=\frac{1}{a}\times\cdots\times\frac{k}{a}<1 \text{ si } \frac{k}{a}\le 1,$$

en cuyo caso, de la igualdad

$$\frac{k!}{e^a a^{k+1}}=\frac{1}{ae^a}\frac{k!}{a^k},$$

obtenemos las dos relaciones siguientes:

$$\frac{1}{ae^a}<10^{-10} \text{ y } \frac{k!}{a^k}\simeq 10^{-5}$$

Para $a=12$, tenemos $\dfrac{1}{12e^{12}}\simeq 5{,}12\times10^{-7}$ y $\dfrac{12!}{12^{12}}\simeq 5{,}37\times10^{-5}$, con lo que $\dfrac{k!}{e^a a^{k+1}}<\dfrac{\epsilon}{2}$ si $a\ge 12$ y $k=12$, luego $|\theta(a)|<\dfrac{\epsilon}{2}$. Así

$$|g(a)-\theta_1(a)|<\frac{\epsilon}{2} \quad \text{si} \quad a\ge 12.$$

Escribamos $\theta_k(a)$ de modo que sea numéricamente estable

$$\begin{aligned}\theta_k(a) &= e^{-a}a^{p-1}\left[1+\frac{p-1}{a}+\cdots+\frac{(p-1)(p-2)\ldots(p-k)}{a^k}\right]\\ &= e^{-a}a^{p-1}\left[1-\frac{1-p}{a}+\frac{(1-p)(2-p)}{a^2}-\frac{(1-p)(2-p)(3-p)}{a^3}+\right.\\ &\quad+\frac{(1-p)(2-p)(3-p)(4-p)}{a^4}+\cdots-\frac{(1-p)(2-p)\ldots(k-1-p)}{a^{k-1}}+\\ &\quad\left.+\frac{(1-p)(2-p)\ldots(k-p)}{a^k}\right].\end{aligned}$$

Sean $\theta_1(a)$, $\theta_2(a)$ las sumas de los términos positivos y de los negativos, respectivamente de $\theta_k(a)$, esto es

$$\begin{aligned}\theta_1(a) &= 1+\frac{(1-p)(2-p)}{a^2}+\frac{(1-p)(2-p)(3-p)(4-p)}{a^4}+\cdots+\frac{(1-p)(2-p)\ldots(k-p)}{a^k}\\ &= 1+\frac{(1-p)(2-p)}{a^2}\left(1+\frac{(3-p)(4-p)}{a^2}\right.\\ &\quad\left.\left(1+\cdots+\frac{(k-3-p)(k-2-p)}{a^2}\left(1+\frac{(k-1-p)(k-p)}{a^2}\right)\ldots\right)\right),\end{aligned}$$

$$\begin{aligned}\theta_2(a) &= \frac{1-p}{a}+\frac{(1-p)(2-p)(3-p)}{a^3}+\cdots+\frac{(1-p)(2-p)\ldots(k-1-p)}{a^{k-1}}\\ &= \frac{1-p}{a}\left(1+\frac{(2-p)(3-p)}{a^2}\left(1+\frac{(4-p)(5-p)}{a^2}\right.\right.\\ &\quad\left.\left.\left(1+\cdots+\frac{(k-3-p)(k-4-p)}{a^2}\left(1+\frac{(k-2-p)(k-1-p)}{a^2}\right)\cdots\right)\right)\right).\end{aligned}$$

Para $k=12$, el número de términos de $\theta_k(a)$ dentro del corchete es 13 y el número de términos positivos es 7 y de los negativos es 6, es decir que $\theta_1(a)$ tiene 7 términos y $\theta_2(a)$ tiene 6 términos. La escritura anidada de $\theta_1(a)$ y $\theta_2(a)$ asegura la estabilidad numérica y además es fácil de programar. Luego

$$\theta_k(a)=e^{-a}a^{p-1}\left(\theta_1(a)-\theta_2(a)\right).$$

Aproximemos $f(a)$. Puesto que $e^{-t}=\sum_{k=0}^{\infty}\frac{(-1)^k t^k}{k!}$, se sigue que

$$\begin{aligned}f(a) &= \int_0^a t^{p-1}e^{-t}dt=\int_0^a t^{p-1}\left(\sum_{k=0}^{\infty}\frac{(-1)^k t^k}{k!}\right)dt=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\int_0^a t^{p+k-1}dt\\ &= \sum_{k=0}^{\infty}\frac{(-1)^k}{k!(k+p)}a^{k+p}.\end{aligned}$$

La última serie converge absolutamente (demuestre). Sean $m\in\mathbb{N}$ y $f_m(a)=\sum_{k=0}^{m}\frac{(-1)^k a^{k+p}}{k!(k+p)}$ tales que $|f(a)-f_m(a)|<\frac{\epsilon}{2}$. Para el efecto, aplicamos el criterio de comparación para series reales. Ponemos $a_k=\frac{a^{k+p}}{k!(k+p)}$, $b_k=\frac{1}{k(k+1)}$. Entonces $\sum_{k=1}^{\infty}b_k=1$, y

$$\frac{a_k}{b_k}=\frac{k(k+1)a^{k+p}}{k!(k+p)}<\frac{(k+1)a^{k+1}}{k!}<\frac{\epsilon}{2}.$$

Para $a=12$, se prueba que para todo $k\geq 55$ se tiene $\frac{(k+1)a^{k+1}}{k!}<\frac{\epsilon}{2}$ es decir que $\frac{a_k}{b_k}<\frac{\epsilon}{2}$ si $k\geq 56$, con lo cual $m=56$ y

$$f_m(a)=\sum_{k=0}^{56}\frac{(-1)^k a^{k+p}}{k!(k+p)}.$$

Para elaborar un algoritmo numéricamente estable, escribamos $f_m(a)$ de la manera siguiente:

$$\begin{aligned}f_m(a) &= a^p\sum_{k=0}^{m}\frac{(-1)^k a^k}{k!(k+p)}=a^p\left[\frac{1}{p}+a\left(\frac{a}{2!(2+p)}-\frac{1}{1+p}\right)+\frac{a^3}{3!}\left(\frac{a}{4(4+p)}-\frac{1}{3+p}\right)+\right.\\ &\quad\left.+\frac{a^5}{5!}\left(\frac{a}{6(6+p)}-\frac{1}{5+p}\right)+\cdots+\frac{a^{m-1}}{(m-1)!}\left(\frac{a}{m(m+p)}-\frac{1}{m-1+p}\right)\right].\end{aligned}$$

Ponemos $c_k = \dfrac{a}{2k(2k+p)} - \dfrac{1}{2k-1+p}, k = 1, 2, \ldots, \dfrac{m}{2}$. Entonces

$$\begin{aligned} f_m(a) &= a^p \left(\frac{1}{p} + c_1 a + c_2 \frac{a^3}{3!} + c_3 \frac{a^5}{5!} + \cdots + c_{\frac{m}{2}} \frac{a^{m-1}}{(m-1)!} \right) \\ &= a^p \left(\frac{1}{p} + a \left(c_1 + \frac{a^2}{2 \times 3} \left(c_2 + \frac{a^2}{4 \times 5} \left(c_3 + \cdots + \right. \right. \right. \right. \\ &\quad \left. \left. \left. \left. \frac{a^2}{(m-4)(m-3)} \right) \left(c_{\frac{m}{2}-1} + \frac{a^2}{(m-2)(m-1)} c_{\frac{m}{2}} \right) \right) \right) \right). \end{aligned}$$

Por lo tanto $\Gamma(p)$ se aproxima mediante $\Gamma_a(p)$ definido por

$$\Gamma_a(p) = f_m(a) + e^{-a} a^{p-1} (\theta_1(a) + \theta_2(a)),$$

donde $f_m(a), \theta_1(a), \theta_2(a)$ definidos precedentemente y $\epsilon = 10^{-10}$ para el cual $m = 56, a = 12$.

El algoritmo para calcular $\Gamma_a(p)$, $0 < p < 1$, con una precisión $\epsilon = 10^{-10}$ es el siguiente:

Algoritmo

Dato de entrada: p.

Dato de salida: $p, \Gamma_a(p)$.

1. Leer p y verificar que $p \in]0, 1[$.
2. Hacer $m = 56, \ a = 12$.
3. Calcular $f_m(a)$.
4. Para $k = 12$, calcular $\theta_1(a)$ y $\theta_2(a)$.
5. Calcular $\Gamma_a(p) = f_m(a) + e^{-a} a^{p-1} (\theta_1(a) - \theta_2(a))$.
6. Fin.

Nota: para el cálculo de $f_m(a)$ se debe elaborar un algoritmo tipo esquema de Hörner. De manera similar $\theta_1(a)$ y $\theta_2(a)$ requieren de la elaboración de los respectivos algoritmos para su cálculo. Se propone como ejercicio la elaboración de algoritmos para el cálculo de $f_m(a)$, $\theta_1(a)$ y $\theta_2(a)$ utilizando su escritura anidada de modo que se eviten los cálculos directos de los factoriales y de las potencias. En la siguiente sección necesitaremos nuevamente la escritura anidada de $f_m(a)$, $\theta_1(a)$ y $\theta_2(a)$ para aproximar la función de distribución gama.

De las propiedades de la función gama, de la definición de $\Gamma(p)$ para $p \in \mathbb{R}^- - \mathbb{Z}^-$, así como de la aproximación de $\Gamma(p)$ mediante $\Gamma_a(p)$, $0 < p < 1$, se presenta el siguiente algoritmo de cálculo de $\Gamma(p)$.

Algoritmo

Dato de entrada: p.

Dato de salida: $p, \Gamma(p)$.

1. Si $p \in \mathbb{Z}^+, \ \Gamma(p) = (p-1)!$.
2. Si $p = n + \dfrac{1}{2}, \Gamma(p) = \dfrac{\sqrt{\pi}}{2^n} \prod_{j=1}^{n} (2j-1)$.
3. Si $0 < p < 1, \ \Gamma(p) = \Gamma_a(p)$.
4. Si $p > 1, n = [p], \Gamma(p) = \Gamma(p-n) \prod_{j=1}^{n} (p-j)$.

([ ] denota la función mayor entero menor o igual que, $p \in \mathbb{R}^+ - \mathbb{N}$, $p - n \in]0, 1[$, $\Gamma(p-n) \simeq \Gamma_a(p-n)$).

5. Si $p < 0,\ p \notin \mathbb{Z}^-$, $\Gamma(p) = \dfrac{\Gamma(p+n)}{\prod_{j=1}^{n}(p+j-1)}$,

(donde $n = |[p]-1|$, $p+n \in]0,1[$, $\Gamma(p+n) \simeq \Gamma_a(p+n)$).

Nota: para la elaboración de un programa computacional, el siguiente indicador indi es de utilidad:

indi $= 1$, si p es un entero positivo.

indi $= 2$, si p es un real positivo de la forma $n + \dfrac{1}{2}$, $n \in \mathbb{N}$.

indi $= 3$, si p es un real tal que $p \in]0,1[$.

indi $= 4$, si p es un real tal que $p > 1$, $p \notin \mathbb{N}$.

indi $= 5$, si p es un real negativo tal que $p \notin \mathbb{Z}^-$.

4.6. Distribución de probabilidad de tipo gama

Definición 4 *Una variable aleatori X tiene una distribución del tipo gama si su función de densidad está definida por*

$$f(t) = \begin{cases} 0, & si\ t \leq 0, \\ \dfrac{t^{p-1}e^{-\frac{t}{\beta}}}{\beta^p \Gamma(p)}, & si\ t \in]0,\infty[\,, \end{cases}$$

donde p, $\beta \in \mathbb{R}^+$ y Γ denota la función gama.
La función de distribución de probabilidad del tipo gama está definida por

$$F_\beta(x) = \frac{1}{\beta^p \Gamma(p)} \int_0^x t^{p-1} e^{-\frac{t}{\beta}} dt, \quad x \geq 0.$$

Cuando $p = 1$, la distribución gama coincide con la distribución exponencial:

$$F_\beta(x) = 1 - e^{-\frac{x}{\beta}} \quad x \geq 0.$$

Para $p = \dfrac{n}{2}$, $n \in \mathbb{Z}^+$ y $\beta = 2$, la distribución gama coincide con la distribución χ^2 con n grados de libertad. Esta distribución se estudiará más adelante.

Cuando $p = n \in \mathbb{Z}^+$ y $\beta = \dfrac{1}{n\mu}$ con $\mu > 0$, la distribución gama se conoce con el nombre de distribución de Erlang de parámetros (n, μ). Cuando $p = m + 1$, $m \in \mathbb{N}$ y $\beta = 1$, la distribución gama se llama distribución exponencial potencial.

Utilizando el cambio de variable $v = \dfrac{t}{\beta}$, se tiene

$$F_\beta(x) = \frac{1}{\beta^p \Gamma(p)} \int_0^{\beta x} (v\beta)^{p-1} e^{-v} \beta dv = \frac{1}{\Gamma(p)} \int_0^{\beta x} v^{p-1} e^{-v} dv,$$

lo que nos conduce a estudiar la función F_p definida por

$$F_p(x) = \frac{1}{\Gamma(p)} \int_0^x t^{p-1} e^{-t} dt,\ x \geq 0,$$

que se conoce con el nombre de distribución gama.

Aproximación de $F_p(x)$, $x, p \in]0,\infty[$.

Para escribir un algoritmo de aproximación de $F_p(x)$, consideramos los tres casos siguientes:

1. $p \in \mathbb{Z}^+$,
2. $0 < p < 1$,
3. $p > 1$, $p \notin \mathbb{Z}^+$.

Caso 1. Si p es un entero positivo, entonces $\Gamma(p) = (p-1)!$ y de la definición de la función $F_p(x)$, se sigue que

$$F_p(x) = \frac{1}{(p-1)!}\int_0^x t^{p-1}e^{-x}dx \quad x \geq 0.$$

Para $p = 1$ se tiene que $F_1(x) = 1 - e^{-x}$, $x \geq 0$.

Si $p > 1$, integrando por partes $p-1$ veces, se tiene

$$\begin{aligned} F_p(x) &= \frac{1}{(p-1)!}\left(-x^{p-1}e^{-x} - (p-1)\,x^{p-2}e^{-x} - (p-1)\,(p-2)\,x^{p-3}e^{-x} - \cdots - \right. \\ &\quad \left. (p-1)(p-2)\ldots \times 2e^{-x} + (p-1)(p-2)\ldots \times 1 - (p-1)(p-2) \times \cdots \times 1 \times e^{-x}\right) \\ &= 1 - e^{-x}\left(1 + x + \frac{x^2}{2!} + \cdots + \frac{x^{p-2}}{(p-2)!} + \frac{x^{p-1}}{(p-1)!}\right) = 1 - e^{-x}\sum_{k=0}^{p-1}\frac{x^k}{k!}. \end{aligned}$$

Así,

$$F_p(x) = 1 - e^{-x}\sum_{k=0}^{p-1}\frac{x^k}{k!} \quad x \geq 0.$$

Note que, por la regla de L'Hôpital, $\lim\limits_{x\to\infty}\dfrac{x^k}{k!}e^{-x} = 0$, luego

$$\lim_{x\to\infty} e^{-x}\sum_{k=0}^{p-1}\frac{x^k}{k!} = 0,$$

Para $\varepsilon = 10^{-10}$, determinemos una condición sobre x y p tal que $F_p(x)$ sea calculado con una precisión ε. Esta condición es $x^{p-1}e^{-x} \leq 10^{-10}$, de donde $x - (p-1)\ln(x) \geq 10\ln(10)$.

Para $p = 1$, definimos

$$F(1,x) = \begin{cases} 1 - e^{-x}, & \text{si } x \leq 10\ln(10), \\ 1, & \text{si } x > 10\ln(10). \end{cases}$$

Para $p > 1$, definimos

$$F(p,x) = \begin{cases} 1 - \exp\left(-x + \ln\left(\sum\limits_{k=0}^{p-1}\dfrac{x^k}{k!}\right)\right), & \text{si } x - (p-1)\ln(x) \leq 10\ln(10), \\ 1, & \text{si } x - (p-1)\ln(x) > 10\ln(10). \end{cases}$$

En esta última escritura de $F(p,x)$ se evita que el término $e^{-x}\sum\limits_{k=0}^{p-1}\dfrac{x^k}{k!}$ se redondee por 0. Además $\sum\limits_{k=0}^{p-1}\dfrac{x^k}{k!}$ tiene que escribirse de forma anidada para evitar el cálculo directo de las potencias y de los factoriales..

Caso 2.- Supongamos ahora que $0 < p < 1$.

Recordemos que si $0 < p < 1$, $\Gamma(p)$ se aproxima mediante

$$\Gamma_a(p) = f_m(a) + e^{-a}a^{p-1}(\theta_1(a) - \theta_2(a)),$$

donde $m = 56$ y $a = 12$, $f_m(a)$, $\theta_1(a)$ y $\theta_2(a)$ están definidos en la sección precedente.

La escritura anidada de $f_m(a)$, $\theta_1(a)$ y $\theta_2(a)$ serán utilizados para aproximar $F_p(x)$ del modo siguiente: si $0 \leq x \leq 12$, entonces $F_p(x)$ se aproxima mediante

$$F_m(x) = \frac{f_m(x)}{\Gamma_a(p)}.$$

Si $x > 12$, entonces $F_p(x)$ se aproxima mediante

$$F_m(x) = \frac{f_m(x) + \theta(a) - \theta(x)}{\Gamma_a(p)},$$

donde $\theta(t) = e^{-t}t^{p-1}(\theta_1(t) - \theta_2(t))$, $t \in [12, \infty[$.

Por otro lado, como $\Gamma(p)$ converge para todo $p > 0$, en particular para $0 < p < 1$ se sigue que dado $\varepsilon > 0$, existe $R > 0$ tal que para todo $x \geq R$, se tiene

$$\left| \int_0^\infty t^{p-1}e^{-t}dt - \int_0^x t^{p-1}e^{-t}dt \right| = \int_x^\infty t^{p-1}e^{-t}dt < \varepsilon.$$

Para $\varepsilon = 10^{-10}$ se deduce que $\int\limits_R^\infty t^{p-1}e^{-t}dt < 10^{-10}$ si $R \backsimeq 24$. En consecuencia,

$$F_m(x) = \begin{cases} \dfrac{f_m(x)}{\Gamma_a(p)}, & \text{si } 0 \leq x \leq 12, \\ \dfrac{f_m(x) + \theta(12) - \theta(x)}{\Gamma_a(p)}, & \text{si } 12 < x \leq 24, \\ 1, & \text{si } x > 24. \end{cases}$$

Para completar el algoritmo debe tomarse en cuenta lo siguiente: para cada $x \in [0, 12]$, $f_m(x)$ debe calcularse con $m = 56$ que se obtuvo cuando $x = 12$, pero para $0 < x < 12$ se requerirán menos términos para lograr la misma precisión. En la tabla siguiente se ilustran algunos subintervalos de $[0, 12]$ con sus respectivos valores de m.

x	1	2	3	4	5	6	7	8	9	10	11	12
m	16	21	25	29	33	37	40	43	47	50	53	56

.

Para simplificar la selección de m para $x \in [0, 12]$, la siguiente relación puede ser útil:

$$j = 1, \ldots, 12, \ x = j, \ m = 16 + 4(j - 1).$$

Si $x = j$ entonces $m = 16 + 4\,(j - 1)$ para $j = 1,\ 2, \ldots, 12$.

Caso 3.- Consideremos ahora el caso $p > 1,\ \ p \notin \mathbb{Z}^+$.

Sea $q = p - n$ con $n = [p]$ el mayor entero menor o igual que p, entonces $q \in]0, 1[$. Utilizando el método de integración por partes $n - 1$ veces, tenemos

$$\begin{aligned} F_p(x) &= \frac{1}{\Gamma(p)} \left(\int_0^x t^{q-1}e^{-t}dt \right) \left(\prod_{j=1}^n (p - j) \right) - x^{p-1}e^{-x} - (p-1)x^{p-2}e^{-x} \\ &\quad -(p-1)(p-2)x^{p-3}e^{-x} - \cdots - (p-1)(p-2)\cdots(p-(n-1)x^{p-n}e^{-x}). \end{aligned}$$

Además, $\Gamma(p) = \Gamma(q) \prod\limits_{j=1}^n (p - j)$.

Sean

$$F_q(x) = \frac{1}{\Gamma(p)} \left(\int_0^x t^{q-1}e^{-t}dt \right) \prod_{j=1}^n (p - j) = \frac{1}{\Gamma(q)} \int_0^x t^{q-1}e^{-t}dt,$$

$$\begin{aligned} Q(x) &= -\frac{1}{\Gamma(p)}\left(x^{p-1}e^{-x}+(p-1)x^{p-2}e^{-x}+(p-1)(p-2)x^{p-3}e^{-x}+\cdots\right. \\ &\quad \left.+x^{p-n}e^{-x}\prod_{j=1}^{n-1}(p-j)\right) \\ &= -\frac{x^q}{q}e^{-x}\left(1+\frac{x}{q+1}+\frac{x^2}{(q+1)(q+2)}+\cdots\right. \\ &\quad \left.+\frac{x^{n-2}}{(q+n-2)\cdots(q+1)}+\frac{x^{n-1}}{(q+n-1)\cdots(q+1)}\right) \\ &= -\frac{x^q}{q}e^{-x}\left(1+\frac{x}{q+1}\left(1+\frac{x}{q+2}\left(1+\cdots+\frac{x}{q+n-2}\left(1+\frac{x}{q+n-1}\right)\cdots\right)\right)\right). \end{aligned}$$

entonces

$$F_p(x) = F_q(x) + Q(x), \;\; x \geq 0.$$

El algoritmo para evaluar $F_q(x)$ con $q \in]0,1[$, $x \geq 0$, está descrito en la parte 2) precedente. Describimos el algoritmo para evaluar $Q(x)$.

Algoritmo

Datos de entrada: n, q, x.

Datos de salida: $Q(x)$.

1. $b = 1$.

2. $\left[\begin{array}{rcl} k & = & 1, \cdots, n-1 \\ & & J = n-k \\ & & b = 1 + \dfrac{bx}{q+j} \end{array}\right.$

   Fin de bucle k.

3. $Q(x) = -\dfrac{x^q e^{-x}}{q}$.

4. Fin.

En resumen, para $p > 0$ y $x \geq 0$, $F_p(x)$ se calcula (aproxima) con una precisión $\varepsilon = 10^{-10}$ del modo siguiente.

Si $p \in \mathbb{Z}^+$,

Si $p = 1$, $F(1,x) = \begin{cases} 1 - e^{-x}, & \text{si } x \leq 10\ln(10), \\ 1, & \text{si } x > 10\ln(10), \end{cases}$

Si $p > 1$, $F(p,x) = \begin{cases} 1 - \exp\left(-x + \ln\left(\sum\limits_{k=0}^{p-1}\dfrac{x^k}{k!}\right)\right), & \text{si } x - (p-1)\ln(x) \leq 10\ln(10), \\ 1, & \text{si } x - (p-1)\ln(x) > 10\ln(10). \end{cases}$

Si $0 < p < 1$, $F_m(x) = \begin{cases} \dfrac{f_m(x)}{\Gamma_a(p)}, & \text{si } 0 \leq x \leq 12, \\ \dfrac{f_m(x) + \theta(12) - \theta(x)}{\Gamma_a(p)}, & \text{si } 12 < x \leq 24, \\ 1, & \text{si } x > 24. \end{cases}$

Si $p > 1$, $p \notin \mathbb{Z}^+$, $F_p(x) = F_q(x) + Q(x)$.

4.7. Función beta. Aproximación de la función beta $B(p,q)$, $p>0$, $q>0$.

Definición 5 *Sean $p,\ q\in\mathbb{R}^+$. La función beta denotada $B(p,q)$ se define por*

$$B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt.$$

Algunas propiedades fundamentales de la función beta se proponen en el teorema siguiente.

Teorema 2

i) La función beta $B(p,q)$ está bien definida si $p>0,\ q>0$ y diverge en cualquier otro caso. Además, la función beta es continua sobre $\mathbb{R}^+\times\mathbb{R}^+$.

ii) Para todo $p,\ q\in\mathbb{R}^+,\ B(p,q)=B(q,p)$.

iii) Para todo $p,\ q\in\mathbb{R}^+$,

$$\begin{aligned} B(p,q) &= 2\int_0^{\frac{\pi}{2}}\operatorname{sin}^{2p-1}(\theta)\cos^{2q-1}(\theta)d\theta,\\ B(p,q) &= \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \end{aligned}$$

iv) Para todo $r>0$,

$$\begin{aligned} B\left(\frac{r+1}{2},\frac{1}{2}\right) &= 2\int_0^{\frac{\pi}{2}}\operatorname{sen}^r(\theta)d\theta,\\ B\left(\frac{1}{2},\frac{r+1}{2}\right) &= 2\int_0^{\frac{\pi}{2}}\cos^r(\theta)d\theta. \end{aligned}$$

En particular, si $r=n\in\mathbb{Z}^+$, se tiene

$$2\int_0^{\frac{\pi}{2}}\operatorname{sin}^n(\theta)d\theta=2\int_0^{\frac{\pi}{2}}\cos^n(\theta)d\theta=\begin{cases}\dfrac{1\times3\times\ldots\times(n-1)}{2\times4\times\ldots\times n}\dfrac{\pi}{2}, & \text{si } n \text{ es par,}\\ \dfrac{2\times4\times\ldots\times(n-1)}{1\times3\times\ldots\times n}, & \text{si } n \text{ es impar.}\end{cases}$$

v) Para todo $p\in\,]0,1[$,

$$B(p,1-p)=\Gamma(p)\Gamma(1-p)=\frac{\pi}{\operatorname{sen}(\pi p)}.$$

Demostración.

i) Sean $p,q\in\mathbb{R}^+$. Si $p\geq1$, $q\geq1$, la función $t\mapsto t^{p-1}(1-t)^{q-1}$ de $[0,1]$ en $\mathbb{R}$ es continua, por lo tanto $B(p,q)$ está bien definida.

Supongamos que $0<p<1$, $0<q<1$. La función $t\to t^{p-1}(1-t)^{q-1}$ de $]0,1[$ en $\mathbb{R}$ es discontinua en 0 y 1.

Sean $I_1=\int_0^{1/2}t^{p-1}(1-t)^{q-1}dt$, $I_2=\int_{1/2}^1 t^{p-1}(1-t)^{q-1}dt$, entonces $B(p,q)=I_1+I_2$. Mostremos la existencia de I_1 y I_2. Se tiene que

$$\begin{aligned} I_1 &= \int_0^{\frac{1}{2}}t^{p-1}(1-t)^{q-1}dt\leq\int_0^{\frac{1}{2}}t^{p-1}dt=\frac{1}{2^p p},\\ I_2 &= \int_{\frac{1}{2}}^1 t^{p-1}(1-t)^{q-1}dt\leq\int_{\frac{1}{2}}^1(1-t)^{q-1}dt=\frac{1}{2^q q}, \end{aligned}$$

que prueba que I_1 e I_2 existen.

En consecuencia $B(p,q)$ está bien definida si $p>0$, $q>0$.

Para probar que $B(p,q)$ diverge en cualquier otro caso, admitamos que $p\leq 0$ y $q\geq 0$. Entonces

$$I_1=\int_0^{\frac{1}{2}} t^{-1}(1-t)^{q-1}dt\geq\int_0^{\frac{1}{2}} t^{-1}\left(\frac{1}{2}\right)^{q-1}dt=\frac{1}{2^{q-1}}\int_0^{\frac{1}{2}}\frac{dt}{t}=\frac{1}{2^{q-1}}\ln|_0^{1/2}=+\infty.$$

ii) Sean $p,q\in\mathbb{R}^+$ y $x=1-t$, $t\in]0,1[$. Entonces

$$B(p,q)=\int_0^1(1-x)^{p-1}x^{q-1}dx=B(q,p).$$

iii) Sean $p,q\in\mathbb{R}^+$ y $t=\operatorname{sen}^2(\theta)$, $\theta\in]0,\frac{\pi}{2}[$. Entonces

$$\begin{aligned}B(p,q) &= \int_0^1 t^{p-1}(1-t)^{q-1}dt=\int_0^{\frac{\pi}{2}}(\operatorname{sen}^2(\theta))^{p-1}(1-\operatorname{sen}^2(\theta))^{q-1}2\operatorname{sen}(\theta)\cos(\theta)d\theta\\ &= 2\int_0^{\frac{\pi}{2}}\operatorname{sen}^{2p-1}(\theta)\cos^{2q-1}(\theta)d\theta.\end{aligned}$$

Sea $t=x^2,\ x\in]0,\infty[$. Entonces

$$\begin{aligned}\Gamma(p) &= \int_0^\infty t^{p-1}e^{-t}dt=2\int_0^\infty x^{2p-1}e^{-x^2}dx,\\ \Gamma(q) &= 2\int_0^\infty x^{2q-1}e^{-x^2}dx.\end{aligned}$$

Luego

$$\Gamma(p)\Gamma(q)=4\left(\int_0^\infty x^{2p-1}e^{-x^2}dx\right)\left(\int_0^\infty y^{2q-1}e^{y^2}dy\right)=4\int_0^\infty\int_0^\infty x^{2p-1}y^{2q-1}e^{-(x^2+y^2)}dxdy.$$

Utilizando coordenadas polares: $\begin{cases} x=\varrho\cos\varphi,\\ y=\varrho\operatorname{sen}\varphi,\end{cases}$ $\varrho\geq 0$, $\varphi\in]0,\frac{\pi}{2}[$, se deduce que

$$\begin{aligned}\Gamma(p)\Gamma(q) &= 4\int_0^{\frac{\pi}{2}}\int_0^\infty \varrho^{2p-1}\varrho^{2q-1}\cos^{2p-1}(\varphi)\operatorname{sen}^{2q-1}(\varphi)\ \varrho^{-\rho^2}\varrho d\varrho d\varphi\\ &= 4\left(\int_0^{\frac{\pi}{2}}\cos^{2p-1}(\varphi)\operatorname{sen}^{2q-1}(\varphi)d\varphi\right)\left(\int_0^\infty \varrho^{2(p+q)-1}e^{-\varrho^2}d\varrho\right)\\ &= 2\left(\int_0^\infty v^{p+q-1}e^{-v}dv\right)\left(\int_0^{\frac{\pi}{2}}\cos^{2p-1}(\varphi)\operatorname{sen}^{2q-1}(\varphi)d\varphi\right)\\ &= \Gamma(p+q)B(p,q).\end{aligned}$$

iv) Sea $r>0$. Por la propiedad iii), haciendo $p=\frac{r+1}{2}$, $q=\frac{1}{2}$, se tiene

$$B\left(\frac{r+1}{2},\frac{1}{2}\right)=2\int_0^{\frac{\pi}{2}}\operatorname{sen}^{2\left(\frac{r+1}{2}\right)-1}(\theta)\cos^{2\left(\frac{1}{2}\right)-1}(\theta)d\theta=2\int_0^{\frac{\pi}{2}}\operatorname{sen}^r(\theta)d\theta,$$

Además

$$B\left(\frac{r+1}{2},\frac{1}{2}\right)=\frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{r+2}{2}\right)},$$

de donde

$$\int_0^{\frac{\pi}{2}}\operatorname{sen}^r(\theta)d\theta=\frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{2\Gamma\left(\frac{r+2}{2}\right)}=\frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r+2}{2}\right)}.$$

Si $r = n \in \mathbb{Z}^+$, utilizando las propiedades de la función gama se obtiene la conclusión.

Mediante la sustitución $\theta = \dfrac{\pi}{2} - \varphi$, $\varphi \in [0, \pi/2]$ se obtiene $\int_0^{\pi/2} \operatorname{sen}^r(\theta) d\theta = \int_0^{\pi/2} \cos^r(\theta) d\theta$.

Note que $B(\frac{1}{2}, \frac{1}{2}) = \Gamma^2\left(\frac{1}{2}\right)$ y $B\left(\frac{1}{2}, \frac{1}{2}\right) = 2\int_0^{\pi/2} d\theta = \pi$. Luego $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

v) Sea $p \in]0, 1[$. Entonces, para todo $n \in \mathbb{Z}^+$, se tiene

$$B(p, n+1) = \int_0^1 t^{p-1}(1-t)^n dt = \frac{\Gamma(p)\Gamma(n+1)}{\Gamma(n+p+1)}.$$

Por otro lado, si $t = \frac{x}{n}$ entonces

$$B(p, n+1) = \int_0^n \left(\frac{x}{n}\right)^{p-1} \left(1 - \frac{x}{n}\right)^n \frac{dx}{n} = \frac{1}{n^p} \int_0^n x^{p-1} \left(1 - \frac{x}{n}\right)^n dx,$$

Luego

$$\int_0^n x^{p-1} \left(1 - \frac{x}{n}\right)^n dx = \frac{n^p \Gamma(p)\Gamma(n+1)}{\Gamma(n+p+1)}.$$

Utilizando el binomio de Newton, no es difícil demostrar que $\lim\limits_{n\to\infty} \left(1 - \frac{x}{n}\right)^n = e^{-x} \quad x \in \mathbb{R}$, entonces

$$\begin{aligned} \Gamma(p) &= \int_0^\infty x^{p-1} e^{-x} dx = \int_0^\infty x^{p-1} \lim_{n\to\infty} \left(1 - \frac{x}{n}\right)^n dx = \lim_{n\to\infty} \int_0^n x^{p-1} \left(1 - \frac{x}{n}\right)^n dx \\ &= \lim_{n\to\infty} \frac{n^p \Gamma(p)\Gamma(n+1)}{\Gamma(n+p+1)} = \Gamma(p) \lim_{n\to\infty} \frac{n^p n!}{\Gamma(n+p+1)}, \end{aligned}$$

con lo cual $\lim\limits_{n\to\infty} \frac{n^p n!}{\Gamma(n+p+1)} = 1$.

El límite $\int_0^\infty x^{p-1} e^{-x} dx = \lim\limits_{n\to\infty} \int\limits_0^n x^{p-1} \left(1 - \frac{x}{n}\right)^n dx$ es consecuencia del teorema de convergencia de Tannery para integrales de Riemann (véase el Análisis de Apostol, página 365).

Como $\Gamma(n+p+1) = (n+p)(n+p-1)\ldots p\Gamma(p)$, se sigue que

$$1 = \lim_{n\to\infty} \frac{n^p n!}{(n+p)(n+p-1) \times \cdots \times p\,\Gamma(p)},$$

de donde

$$\Gamma(p) = \lim_{n\to\infty} \frac{n^p n!}{(n+p)(n+p-1)\ldots p},$$

que es la definición de Gauss de la función gama.

Por otra parte,

$$\Gamma(1-p) = \lim_{n\to\infty} \frac{n^{1-p} n!}{(n+1-p)(n-p)\ldots(1-p)},$$

luego

$$\begin{aligned} \Gamma(p)\Gamma(p-1) &= \lim_{n\to\infty} \frac{n(n!)^2}{(n+1-p)(n^2-p^2)((n-1)^2-p^2)\cdots(1-p^2)p} \\ &= \lim_{n\to\infty} \frac{n}{n+1-p} \lim_{n\to\infty} \frac{1}{p(1-p^2)\left(1-\frac{p^2}{2^2}\right)\cdots\left((1-\frac{p^2}{n^2}\right)} \\ &= \lim_{n\to\infty} \frac{1}{p(1-p^2)\left(1-\frac{p^2}{2^2}\right)\cdots\left(1-\frac{p^2}{n^2}\right)} \\ &= \lim_{n\to\infty} \frac{1}{p\prod\limits_{j=1}^{n}\left(1-\frac{p^2}{j^2}\right)}. \end{aligned}$$

La función f definida por $f(x)=\operatorname{sen}(\pi x)\quad x\in\mathbb{R}$ se anula en $x=k\in\mathbb{Z}$, es decir que el conjunto de todas las raíces de la ecuación $f(x)=0$ es $\mathbb{Z}$. La función real g definida por

$$g(x)=x\prod_{j=1}^{\infty}\left(1-\frac{x}{j}\right)\left(1+\frac{x}{j}\right)=x\prod_{j=1}^{\infty}\left(1-\frac{x^2}{i^2}\right)\qquad x\in\mathbb{R},$$

es tal que $g(x)=0$ si y solo si $x=j\in\mathbb{Z}$; esto es, las funciones f y g tienen el mismo conjunto de ceros. Con estos argumentos se demuestra que

$$\operatorname{sen}(\pi x)=\pi x\prod_{j=1}^{\infty}\left(1-\frac{x^2}{j^2}\right),$$

que es la representación factorial de Weierstrass de $\operatorname{sen}(\pi x)$. Por lo tanto

$$\Gamma(p)\,\Gamma(p-1)=\lim_{n\to\infty}\frac{1}{p\prod\limits_{j=1}^{n}\left(1-\frac{p^2}{j^2}\right)}=\frac{\pi}{\operatorname{sen}(\pi p)}.$$

■

Aproximación de la función beta $B(p,q)$, $p>0$, $q>0$

Sean $p,q\in\mathbb{R}^+$.

1. Si $p,q\in\mathbb{Z}^+$, de las propiedades establecidas para las funciones gama y beta, se tiene

$$B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}=\frac{(p-1)!(q-1)!}{(p+q-1)!}.$$

2. Si p,q son tales que $p+q=1$, de la propiedad v) de la función beta, obtenemos

$$B(p,1-p)=\Gamma(p)\,\Gamma(1-p)=\frac{\pi}{\operatorname{sen}(\pi p)}.$$

3. Si $0<p<1$, $0<q<1$, y $p+q\neq 1$, entonces

$$B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt=\int_0^{\frac{1}{2}} t^{p-1}(1-t)^{q-1}dt+\int_{\frac{1}{2}}^1 t^{p-1}(1-t)^{q-1}dt.$$

Sea $x=1-t\quad t\in\left[\frac{1}{2},1\right]$, resulta que

$$\int_{\frac{1}{2}}^1 t^{p-1}(1-t)^{q-1}dt=\int_0^{\frac{1}{2}} x^{q-1}(1-x)^{p-1}dx,$$

consecuentemente

$$B(p,q)=\int_0^{\frac{1}{2}} t^{p-1}(1-t)^{q-1}dt+\int_0^{\frac{1}{2}} t^{q-1}(1-t)^{p-1}dt.$$

Sea $g(t)=(1-t)^{\alpha-1}$, donde $0<\alpha<1$, $t\in]-1,1[$. Representemos la función g mediante una serie de Taylor en un entorno de cero. Tenemos

$$\begin{aligned}
g(0) &= 1,\\
g'(0) &= 1-\alpha,\\
g''(0) &= (1-\alpha)(2-\alpha)\\
&\vdots\\
g^{(k)}(0) &= \prod_{j=1}^{k}(j-\alpha),\ \ \forall k\in\mathbb{Z}^+.
\end{aligned}$$

Entonces

$$g(t) = 1 + \sum_{k=1}^{\infty} \frac{g^{(k)}(0)}{k!} t^k = 1 + \sum_{k=1}^{\infty} \frac{(1-\alpha)(2-\alpha)\cdots(k-\alpha)}{k!} t^k.$$

Esta serie es absolutamente convergente para todo $t \in]-1,1[$ y converge uniformemente sobre todo conjunto $[-a,a]$, $0 < a < 1$ (demuestre!). Sean

$$\begin{aligned} A_1(p,q) &= \int_0^{1/2} t^{p-1}(1-t)^{q-1}dt, \\ A_2(p,q) &= \int_0^{1/2} t^{q-1}(1-t)^{p-1}dt, \end{aligned}$$

es decir que $B(p,q) = A_1(p,q) + A_2(p,q)$.

Para $\alpha = q$, obtenemos

$$\begin{aligned} A_1(p,q) &= \int_0^{1/2} t^{p-1}(1-t)^{q-1}dt = \int_0^{1/2} t^{p-1}\left(1 + \sum_{k=1}^{\infty} \frac{(1-q)(2-q)\ldots(k-q)}{k!} t^k\right) dt \\ &= \int_0^{1/2} t^{p-1}dt + \sum_{k=1}^{\infty} \frac{(1-q)(2-q)\ldots(k-q)}{k!} \int_0^{1/2} t^{p+k-1}dt \\ &= \frac{t^p}{p}\Big|_0^{1/2} + \sum_{k=1}^{\infty} \frac{(1-q)(2-q)\ldots(k-q)}{k!} \frac{t^{p+k}}{p+k}\Big|_0^{1/2} \\ &= \frac{1}{p2^p} + \sum_{k=1}^{\infty} \frac{(1-q)(2-q)\ldots(k-q)}{k!(k+p)2^{p+k}} \\ &= \frac{1}{2^p}\left(\frac{1}{p} + \sum_{k=1}^{\infty} \frac{(1-q)(2-q)\ldots(k-q)}{k!(k+p)2^k}\right). \end{aligned}$$

De manera similar obtenemos

$$A_2(p,q) = \frac{1}{2^q}\left(\frac{1}{q} + \sum_{k=1}^{\infty} \frac{(1-p)(2-p)\ldots(k-p)}{k!(k+q)2^k}\right).$$

Para construir un algoritmo bien condicionado y numéricamente estable, aproximemos $A_1(p,q)$ y $A_2(p,q)$ mediante sumas finitas con $m+1$ términos $\theta_1(p,q)$ y $\theta_2(p,q)$ respectivamente tales que si $\varepsilon > 0$,

$$\begin{aligned} |A_1(p,q) - \theta_1(p,q)| &< \frac{\varepsilon}{2}, \\ |A_2(p,q) - \theta_2(p,q)| &< \frac{\varepsilon}{2}. \end{aligned}$$

Sea

$$\theta_1(p,q) = \frac{1}{2^p}\left(\frac{1}{p} + \sum_{k=1}^{m} \frac{(1-q)(2-q)\cdots(k-q)}{k!(k+p)2^k}\right).$$

Con fines prácticos $\varepsilon = 10^{-10}$.Apliquemos el criterio del cociente.

Como $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$, entonces si

$$a_k = \frac{(1-q)(2-q)\cdots(k-q)}{k!(k+p)2^k}, \qquad b_k = \frac{1}{k(k+1)},$$

se sigue que

$$\begin{aligned} \frac{a_k}{b_k} &= \frac{(1-q)(2-q)\cdots(k-q)k(k+1)}{k!(k+p)2^k} < \frac{k!k(k+1)}{k!k2^k} \\ &= \frac{k+1}{2^k} < \frac{1}{2}10^{-10} \quad \text{si } k \geq 41. \end{aligned}$$

Si escogemos $m=41$, tenemos

$$\theta_1(p,q)=\frac{1}{2^p}\left(\frac{1}{p}+\sum_{k=1}^{41}\frac{(1-q)(2-q)\cdots(k-q)}{k!(k+p)2^k}\right),$$

que en forma anidada se escribe

$$\begin{aligned}\theta_1(p,q) &= \frac{1}{2^p}\left(\frac{1}{p}+\frac{1-q}{2}\left(\frac{1}{1+p}+\frac{2-q}{2\times 2}\left(\frac{1}{2+p}+\frac{3-q}{3\times 2}\left(\frac{1}{3+p}+\cdots+\right.\right.\right.\right.\\ &\left.\left.\left.\left.\frac{40-q}{40\times 2}\right)\left(\frac{1}{40+p}+\frac{41-q}{40\times 2\times(41+p)}\right)\cdots\right)\right)\right).\end{aligned}$$

Cambiando p por q, se obtiene una escritura anidada de $\theta_2(p,q)$.

Se propone como ejercicio elaborar un algoritmo que permita calcular $\theta_1(p,q)$ y $\theta_2(p,q)$.

Finalmente $B(p,q)$ se aproxima mediante $\theta_1(p,q)+\theta_2(p,q)$ con una precisión $\varepsilon=10^{-10}$.

Nota: Puesto que $B(p,q)=\dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$. Si para aproximar $B(p,q)$ se utiliza el algoritmo para aproximar $\Gamma(p),\Gamma(q)$ y $\Gamma(p+q)$, resulta que este es numéricamente más costoso que la aproximación mediante $\theta_1(p,q)+\theta_2(p,q)$. Además este último es muy simple de programar.

4. Supongamos que al menos uno de los dos parámetros p, q es mayor o igual que 1, además $p,q\notin\mathbb{Z}^+$.

   Sean $m=[p]$, $n=[q]$ donde $[\cdot]$ denota la función mayor entero menor o igual que y $r=m+n$, entonces $p-m$, $q-n\in]0,1[$.

   Luego

$$\begin{aligned}B(p,q) &= \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}=\frac{((p-1)\cdots(p-n)\,\Gamma(p-m))\,((q-1)\cdots(q-n)\Gamma(q-n))}{(p+q-1)\cdots(p+q-r)\,\Gamma(p+q-r)}\\ &= \frac{\prod_{j=1}^{m}(p-j)\prod_{k=1}^{n}(q-k)}{\prod_{i=1}^{m+n}(p+q-i)}\frac{\Gamma(p-n)\,\Gamma(q-m)}{\Gamma(p+q-r)},\end{aligned}$$

   donde

$$\begin{aligned}p+q &= (p-n)+n+(q-m)+m=(p-n)+(q-m)+r,\\ p+q-r &= (p-n)+(q-m),\end{aligned}$$

   consecuentemente

$$B(p-m,q-n)=\frac{\Gamma(p-m)\Gamma(q-n)}{\Gamma(p+q-n-m)}=\frac{\Gamma(p-m)\Gamma(q-n)}{\Gamma(p+q-r)},$$

   y

$$B(p,q)=\frac{\left(\prod_{j=1}^{m}(p-j)\right)\left(\prod_{k=1}^{n}(q-k)\right)}{\prod_{i=1}^{m+n}(p+q-i)}B(p-m,q-n),$$

   con $B(p-m,\,q-n)$ que se aproxima mediante $\theta_1\,(p-m,q-n)+\theta_2\,(p-m,q-n)\,.$

En resumen,

1. Si $p,q\in\mathbb{Z}^+$,

$$B(p,q)=\frac{(p-1)!(q-1)!}{(p+q-1)!}.$$

2. Si $p, q \in \mathbb{R}^+$ tales que $p+q=1$,
$$B(p,1-p) = \frac{\pi}{\operatorname{sen}(\pi p)}.$$

3. Si $0<p<1,\ \ 0<q<1$, tales que $p+q \neq 1$, entonces
$$B(p,q) \simeq \theta_1(p,q) + \theta_2(p,q).$$

4. Si $p \geq 1$ o $q \geq 1$ y $p, q \notin \mathbb{Z}^+$, entonces
$$B(p,q) \simeq \frac{\left(\prod_{j=1}^{m}(p-j)\right)\left(\prod_{k=1}^{n}(q-k)\right)}{\prod_{i=1}^{m+n}(p+q-i)}(\theta_1(p-m,q-n)+\theta_2(p-m,q-n)),$$
y $m=[p]$, $n=[q]$.

4.8. Distribución beta.

Definición 6 *Se dice que una variable aleatoria X tiene una distribución de probabilidad beta con parámetros p y q si y solo si la función de densidad de X está definida mediante:*
$$f(t) = \begin{cases} \dfrac{t^{p-1}(1-p)^{q-1}}{B(p,q)}, & t \in [0,1], \\ 0, & si\ t \in \mathbb{R} \setminus [0,1], \end{cases}$$
donde $p,\ q \in \mathbb{R}^+$ y $B(p,q)$ denota la función beta en p y q.
La función de distribución beta está definida por
$$F(p,q,x) = \frac{1}{B(p,q)}\int_0^x t^{p-1}(1-p)^{q-1}dt, \quad x \in [0,1].$$

Proposición 3 *Para todo $x \in [0,1]$, se tiene*
$$F(p,q,x) = 1 - F(q,p,1-x).$$

Demostración. Puesto que
$$1 = \frac{1}{B(p,q)}\int_0^1 t^{p-1}(1-t)^{q-1}dt,$$
se sigue que para todo $x \in [0,1]$,
$$\begin{aligned} 1 &= \frac{1}{B(p,q)}\left(\int_0^x t^{p-1}(1-t)^{q-1}dt + \int_x^1 t^{p-1}(1-t)^{q-1}dt\right) \\ &= F(p,q,x) + \frac{1}{B(p,q)}\int_x^1 t^{p-1}(1-t)^{q-1}dt. \end{aligned}$$
Utilizando el cambio de variable $u = 1-t$, se deduce que
$$\begin{aligned} F(p,q,x) &= 1 - \frac{1}{B(p,q)}\int_{1-x}^0 (1-u)^{p-1}u^{q-1}(-du) \\ &= 1 - \frac{1}{B(q,p)}\int_0^{1-x} u^{q-1}(1-u)^{p-1}du = 1 - F(q,p,1-x). \end{aligned}$$
■

Proposición 4 *Sea* $p, q \in]0,1[$ *fijos,* $f(x) = \int_0^x t^{p-1}(1-t)^{q-1}dt,\ x \in [0,1]\,,\ \varepsilon > 0$. *Existen dos funciones* $P_1,\ P_2$ *tales que*

$$\begin{aligned} |f(x) - P_1| &< \varepsilon \quad si\ x \in \left[0, \frac{1}{2}\right], \\ |f(x) - P_2| &< \varepsilon \quad si\ x \in \left]\frac{1}{2}, 1\right[. \end{aligned}$$

Demostración. En la sección precedente se mostró que la serie de Taylor de la función $g(t) = (1-t)^{\alpha-1}$, $t \in [0, x[$ y $0 < \alpha < 1$ está dada por

$$g(t) = 1 + \sum_{k=1}^{\infty} \frac{(1-\alpha)(2-\alpha)\cdots(k-\alpha)}{k!} t^k,$$

la cual es uniformemente convergente sobre $\left[0, \frac{1}{2}\right]$. Entonces, para $x \in \left[0, \frac{1}{2}\right]$ se tiene

$$\begin{aligned} f(x) &= \int_0^x t^{p-1}(1-t)^{q-1}dt = \int_0^x t^{p-1}\left(1 + \sum_{k=1}^{\infty} \frac{(1-q)\cdots(k-q)}{k!} t^k\right) dt \\ &= \frac{x^p}{p} + \sum_{k=1}^{\infty} \frac{(1-q)\cdots(k-q)}{k!\,(k+p)} t^{p+k}. \end{aligned}$$

Ahora bien, para todo $x \in \left[0, \frac{1}{2}\right]$ se tiene

$$\sum_{k=1}^{\infty} \frac{(1-q)\cdots(k-q)}{k!(k+p)} x^{p+k} \leq \sum_{k=1}^{\infty} \frac{k!}{k!k} \frac{1}{2^{p+k}} = \frac{1}{2^p} \sum_{k=1}^{\infty} \frac{1}{k2^k}.$$

La serie $\sum\limits_{k=1}^{\infty} \frac{1}{k2^k}$ es convergente. Sean $a_k = \frac{1}{k2^k}$, $b_k = \frac{1}{k(k+1)}$ entonces $\sum\limits_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$, y

$$\frac{a_k}{b_k} = \frac{k+1}{2^k} \underset{k\to\infty}{\longrightarrow} 0,$$

Luego, existe $m \in \mathbb{Z}^+$ tal que $\frac{a_k}{b_k} < \varepsilon$ si $k \geq m$, con lo cual

$$\frac{1}{2^p} \sum_{k=m+1}^{\infty} \frac{1}{k\,2^k} < \varepsilon.$$

Definimos

$$P_1(x) = \frac{x^p}{p} + \sum_{k=1}^{m} \frac{(1-q)\cdots(k-q)}{k!(k+p)} x^{k+p} \quad \text{si } x \in \left[0, \frac{1}{2}\right],$$

entonces

$$|f(x) - P_1(x)| = \sum_{k=m+1}^{\infty} \frac{(1-q)\cdots(k-q)}{k!\,(k+p)} x^{k+p} \leq \frac{1}{2^p} \sum_{k=m+1}^{\infty} \frac{1}{k\,2^k} < \varepsilon.$$

Aplicando la proposición precedente, definimos

$$P_2(x) = 1 - P_1(x), \ \text{ si } x \in \left]\frac{1}{2}, 1\right[,$$

entonces

$$|f(x) - P_2(x)| < \varepsilon, \quad \forall x \in \left]\frac{1}{2}, 1\right[.$$

■

1. Sean $p, q \in]0,1[$.

   En el caso en que $p, q \in]0,1[$, la proposición precedente es utilizada para construir un algoritmo numéricamente estable. Definimos

$$P_1(p,q,x) = \frac{x^p}{B(p,q)}\left(\frac{1}{p} + \sum_{k=1}^{m} \frac{(1-q)\cdots(k-q)}{k!\,(k+p)} x^k\right) \qquad x \in \left[0, \frac{1}{2}\right],$$

   donde m es tal que $|f(x) - P_1(x)| < \varepsilon$ si $x \in \left[0, \frac{1}{2}\right]$, y $P_1(x)$ definida en la proposición precedente

$$P_2(p,q,x) = 1 - P_1(q,p,1-x), \;\; x \in \left]\frac{1}{2}, 1\right].$$

   Sea $x \in \left[0, \frac{1}{2}\right]$. A medida que x se aproxima a $\frac{1}{2}$, $P_1(x)$ requiere de un número mayor de términos para alcanzar la precisión requerida. Sea $\varepsilon = 10^{-10}$ y dividamos al intervalo $\left[0, \frac{1}{2}\right]$ en cinco subintervalos de igual longitud $[x_{j-1}, x_j]$, donde $x_j = 0{,}1j$, $\; j = 1,2,3,4,5$. Entonces

$$m_j = 12 + 7(j-1), \quad j = 1, \cdots, 5.$$

   es tal que $\sum_{k=m_j+1}^{\infty} \frac{x_j^k}{k} < \varepsilon, \;\; j = 1, \cdots 5.$

   Como es habitual, $P_1(p,q,x)$ se escribe en forma anidada.

   Se propone como ejercicio la elaboración de un algoritmo para calcular $P_1(p,q,x)$ y $P_2(p,q,x)$. El algoritmo para el cálculo de $B(p,q)$ está descrito en la sección precedente.

2. Sean $p, q \in]1, \infty]$.

   Sean $k = [q] - 1$ y $r = q - (k+1) \in [0,1[$, donde $[\cdot]$ denota la función mayor entero menor o igual que. Integrando por partes k veces, tenemos

$$\begin{aligned}
\int_0^x t^{p-1}(1-t)^{q-1}dt &= \frac{t^p}{p}(1-t)^{q-1}|_0^x + \frac{q-1}{p}\int_0^x t^p(1-t)^{q-2}dt \\
&= \frac{x^p}{p}(1-x)^{q-1} + \frac{(q-1)}{p}\left(\frac{x^{p+1}}{p+1}(1-x)^{q-2} + \frac{q-2}{p+1}\int_0^x t^{p+1}(1-t)^{q-3}dt\right) \\
&= \frac{x^p}{p}(1-x)^{q-1} + \frac{(q-1)}{p(p+1)}x^{p+1}(1-x)^{q-2} + \frac{(q-1)(q-2)}{p(p+1)(p+2)}x^{p+2}(1-x)^{q-3} \\
&\quad + \cdots + \frac{(q-1)(q-2)\cdots(q-(k-1))}{p(p+1)(p+2)\cdots(p+k-1)}x^{p+k-1}(1-x)^{q-k} \\
&\quad + \frac{(q-1)(q-2)\cdots(q-k)}{p(p+1)\cdots(p+k-1)}\int_0^x t^{p+k-1}(1-t)^{q-(k+1)}dt.
\end{aligned}$$

   Identificamos dos casos: $q = k+1 \in \mathbb{Z}^+$, $p > 1$; y, $q > 1, \;\; q \notin \mathbb{Z}^+$, $p > 1$.

   Consideremos el primer caso: $q = k+1 \in \mathbb{Z}^+, \;\; p > 1$. Entonces

$$\begin{aligned}
\int_0^x t^{p+1}(1-t)^{q-1}dt &= \frac{x^p}{p}(1-x)^{q-1} + \frac{q-1}{p(p+1)}x^{p+1}(1-x)^{q-2} + \frac{(q-1)(q-2)}{p(p+1)(p+2)}x^{p+2}(1-x)^{q-3} \\
&\quad + \cdots + \frac{(q-1)(q-2)\cdots(q-(k-1))}{p(p+1)\cdots(p+k-1)}x^{p+k-1}(1-x)^{q-k} \\
&\quad + \frac{(q-1)(q-2)\cdots(q-k)}{p(p+1))\cdots(p+k-1)}\frac{x^{p+k}}{p+k}.
\end{aligned}$$

   Por otro lado,

$$B(p,q) = \frac{\Gamma(p)\,\Gamma(k+1)}{\Gamma(p+k+1)} = \frac{k!\,\Gamma(p)}{(p+k)(p+k-1)\cdots p\Gamma(p)} = \frac{k!}{(p+k)(p+k-1)\cdots p}.$$

Luego

$$\begin{aligned} F(p,q,x) &= \frac{1}{B(p,q)}\int_0^x t^{p-1}(1-t)^{q-1}dt \\ &= x^{p+k}+\frac{p+k}{1!}x^{p+k-1}(1-x)+\frac{(p+k)(p+k-1)}{2!}x^{p+k-2}(1-x)^2+\cdots+ \\ &\quad \frac{(p+1)(p+2)\cdots(p+k)}{k!}x^p(1-x)^k \\ &= x^{p+k}\left(1+\frac{(p+k)}{1}y\left(1+\frac{p+k-1}{2}y\left(1+\frac{p+k-2}{3}y\left(1+\cdots+\right.\right.\right.\right. \\ &\quad \left.\left.\left.\left.\frac{p+2}{k-1}y\right)\left(1+\frac{p+1}{k}y\right)\right)\right)\right), \end{aligned}$$

donde $y=\frac{1-x}{x}$, $x>0$.

Note que este último desarrollo es válido cualesquiera que sea $p>0$ y $q=k+1$ un entero mayor que 1.

Para $q=2$, $p>1$, se tiene

$$F(p,2,x)=x^{p+1}(1+(p+1)y).$$

Para $q=3$, $p>1$,

$$F(p,3,x)=x^{p+2}(1+(p+2)\ y(1+\frac{p+1}{2}y)),$$

donde $y=\dfrac{1-x}{x}$, $0<x\leq 1$.

En el caso en que $q>0$ y $p=k+1$ un entero mayor que 1 se utiliza la relación

$$F(p,q,x)=1-F(q,p,1-x)$$

y $F(q,p,1-x)$ se calcula mediante el algoritmo arriba descrito a condición de cambiar p por q y x por $1-x$.

Consideremos ahora el segundo caso: $q>1$, $q\notin\mathbb{Z}^+$.

Definimos

$$\begin{aligned} P_3(p,q,x) &= \frac{x^p}{p}\left((1-x)^{q-1}+\frac{q-1}{p+1}\frac{x}{1-x}(1-x)^{q-1}+\frac{(q-1)(q-2)}{(p+1)(p+2)}\frac{x^2}{(1-x)^2}(1-x)^{q-1}\right. \\ &\quad \left.+\cdots+\frac{(q-1)(q-2)\cdots(q-k+1)}{(p+1)(p+2)\cdots(p+k-1)}\frac{x^{k-1}}{(1-x)^{k-1}}(1-x)^{q-1}\right) \\ &= \frac{x^p}{p}(1-x)^{q-1}\left(1+\frac{q-1}{p+1}\frac{x}{1-x}+\frac{(q-1)(q-2)}{(p+1)(p+2)}\left(\frac{x}{1-x}\right)^2+\cdots+\right. \\ &\quad \left.\frac{(q-1)(q-2)\cdots(q-k+1)}{(p+1)(p+2)\cdots(p+k-1)}\left(\frac{x}{1-x}\right)^{k-1}\right) \\ &= \frac{x^p}{p}(1-x)^{q-1}\left(1+\frac{q-1}{p+1}y\left(1+\frac{q-2}{p+2}y\left(1+\cdots+\right.\right.\right. \\ &\quad \left.\left.\left.\frac{q-k+2}{p+k-2}y\right)\left(1+\frac{q-k+1}{p+k-1}y\right)\cdots\right)\right), \end{aligned}$$

donde $y=\frac{x}{1-x}$, $0\leq x<1$, y.

$$P_4(p,q,x)=\frac{(q-1)(q-2)\cdots(q-k)}{p(p+1)\cdots(p+k-1)}\int_0^x t^{p+k-1}(1-t)^{q-(k+1)}dt,$$

entonces

$$F(p,q,x)=\frac{1}{B(p,q)}\left(P_3(p,q,x)+P_4(p,q,x)\right).$$

Para $x \in \left[0, \frac{1}{2}\right]$, la integral $\int_0^x t^{p+k-1}(1-t)^{q-(k+1)}dt$ se calcula utilizando el algoritmo descrito en i) y si $x \in \left]\frac{1}{2}, 1\right]$ se utiliza la relación

$$F(p,q,x) = 1 - F(p,q,1-x) = 1 - \frac{1}{B(p,q)}\left(P_3\left(q,p,1-x\right) + P_4\left(q,p,1-x\right)\right),$$

y a continuación $P_4(q,p,1-x)$ se calcula en la parte precedente.

3. Finalmente, si $p = 1$ tenemos

$$F(1,q,x) = 1 - (1-x)^q, \quad x \in [0,1].$$

Si $q = 1$,

$$F(p,1,x) = x^p, \quad x \in [0,1].$$

Se propone como ejercicio la elaboración de un algoritmo completo que permite calcular (aproximar) valores de la distribución beta para $p > 0, q > 0$ y $x \in [0,1]$.

4.9. Distribución normal.

Definición 7 *Una variable aleatoria X tiene una distribución de probabilidad normal de media $\mu \in \mathbb{R}$ y varianza $\sigma > 0$ si su función densidad está dada por*

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \quad t \in \mathbb{R}.$$

La función de distribución está definida por

$$N(\mu,\sigma,x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \quad x \in \mathbb{R}.$$

Utilizando el cambio de variable $z = \frac{t-\mu}{\sigma}$, se tiene

$$N(\mu,\sigma,\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt.$$

En lo que sigue consideraremos la función φ definida por

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \quad x \in \mathbb{R}.$$

que corresponde a $N(0,1,x)$.

Se probó que $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$ y utilizando el cambio de variable $t = \sqrt{2}x$ se prueba que $\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$. Por otro lado, si $f(t) = e^{-\frac{t^2}{2}} \quad t \in \mathbb{R}$, se tiene que $f(-t) = f(t) \quad \forall t \in \mathbb{R}$, es decir que f es una función par. En consecuencia

$$\varphi(x) = 0{,}5 + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt.$$

Utilizando la serie de potencias de e^{α} :

$$e^{\alpha} = \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} \quad \alpha \in \mathbb{R},$$

que converge absolutamente para todo $\alpha \in \mathbb{R}$ y es uniformemente convergente sobre todo intervalo cerrado y acotado de $\mathbb{R}$ y haciendo $\alpha = -\frac{t^2}{2}$, tenemos

$$e^{-\frac{t^2}{2}} = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{k!\, 2^k},$$

luego

$$\begin{aligned}\varphi(x) &= 0{,}5 + \frac{1}{\sqrt{2\pi}}\int_0^x \left(\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\,2^k}t^{2k}\right)dt = 0{,}5 + \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\,2^k}\int_0^x t^{2k}dt \\ &= 0{,}5 + \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\,2^k(2k+1)}x^{2k+1}.\end{aligned}$$

La última serie de potencias es absolutamente convergente para todo $x \in \mathbb{R}$.

Para aproximar $\varphi(x)$ mediante una suma finita, se debe tomar en cuenta que el número de términos depende de x. Más adelante volveremos a tratar esta serie.

Sea $\varepsilon > 0$. Con el propósito de elaborar un algoritmo numéricamente estable y económico, determinemos $r > 1$ tal que

$$\left|\frac{1}{\sqrt{2\pi}}\int_0^{\infty} e^{-\frac{t^2}{2}}dt - \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt\right| < \varepsilon \quad \text{si } x \geq r,$$

es decir que

$$\frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\frac{t^2}{2}}dt < \varepsilon \quad \text{si } x \geq r.$$

Apliquemos el criterio de comparación para integrales impropias. Como $\int_1^{\infty}\frac{dt}{t^2} = 1$ y haciendo $g(t) = \frac{1}{t^2}$, se tiene

$$\frac{f(t)}{g(t)} = \frac{t^2}{\sqrt{2\pi}\,e^{\frac{t^2}{2}}} \underset{t\to\infty}{\longrightarrow} 0,$$

luego, existe $r > 1$ tal que

$$\frac{t^2}{\sqrt{2\pi}\,e^{\frac{t^2}{2}}} < \varepsilon \quad \text{si} \quad t \geq r.$$

Para $\varepsilon = 10^{-10}$, esta última desigualdad se verifica para $r = 8$, es decir que

$$\frac{f(t)}{g(t)} < 10^{-10} \quad \text{si } t \geq 8,$$

en consecuencia

$$\frac{1}{\sqrt{2\pi}}\int_8^{\infty} e^{-\frac{t^2}{2}}dt < 10^{-10}.$$

Definimos

$$\varphi_r(x) = \begin{cases} 0, & \text{si } x \leq -8, \\ \varphi(x), & \text{si } -8 < x < 8, \\ 1, & \text{si } x \geq 8. \end{cases}$$

Para todo $a, x \in]-8, 8[$, tenemos

$$\varphi(x) = 0{,}5 + \frac{1}{\sqrt{2\pi}}\left(\int_0^a e^{-\frac{t^2}{2}}dt + \int_a^x e^{-\frac{t^2}{2}}dt\right) = \varphi(a) + \frac{1}{\sqrt{2\pi}}\int_a^x e^{-\frac{t^2}{2}}dt.$$

Además, para todo $x \in]-8, 8[$,

$$1 = \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^x e^{-\frac{t^2}{2}}dt + \int_x^8 e^{-\frac{t^2}{2}}dt + \int_8^{\infty} e^{-\frac{t^2}{2}}dt\right) \simeq \varphi(x) + \frac{1}{\sqrt{2\pi}}\int_x^8 e^{-\frac{t^2}{2}}dt + \frac{1}{\sqrt{2\pi}}\int_8^{\infty} e^{-\frac{t^2}{2}}dt,$$

y como $\frac{1}{\sqrt{2\pi}}\int_8^{\infty} e^{-\frac{t^2}{2}}dt < 10^{-10}$, despreciando este último término, resulta que

$$1 = \varphi_r(x) + \frac{1}{\sqrt{2\pi}}\int_x^8 e^{-\frac{t^2}{2}}dt,$$

de donde

$$\varphi_r(x) = 1 - \frac{1}{\sqrt{2\pi}}\int_x^8 e^{-\frac{t^2}{2}}dt \quad \text{si } x \in]-8, 8[\,.$$

Para calcular $\int_a^x e^{-\frac{t^2}{2}}dt$ con una precisión $\varepsilon = 10^{-10}$, utilicemos el método de integración por partes. Tenemos

$$\begin{aligned}\int_a^x e^{-\frac{t^2}{2}}dt &= \int_a^x \frac{-t\,e^{-\frac{t^2}{2}}}{-t}dt = -\frac{e^{-\frac{t^2}{2}}}{t}\bigg|_a^x - \int_a^x \frac{e^{-\frac{t^2}{2}}}{t^2}dt = -\frac{e^{-\frac{t^2}{2}}}{t}\bigg|_a^x - \int_a^x \frac{-te^{-\frac{t^2}{2}}}{-t^3}dt \\ &= -\frac{e^{-\frac{t^2}{2}}}{t}\bigg|_a^x - \left(-\frac{e^{-\frac{t^2}{2}}}{t^3}\bigg|_a^x - 3\int_a^x \frac{e^{-\frac{t^2}{2}}}{t^4}dt\right) = e^{-\frac{t^2}{2}}\left(-\frac{1}{t}+\frac{1}{t^3}\right)\bigg|_a^x + 3\int_a^x \frac{e^{-\frac{t^2}{2}}}{t^4}dt.\end{aligned}$$

Continuando con este procedimiento k veces, obtenemos

$$\int_a^x e^{-\frac{t^2}{2}}dt = \theta_1(t)|_a^x + \theta_2(x),$$

donde

$$\theta_1(t) = \frac{e^{-\frac{t^2}{2}}}{t}\left(-1+\frac{1}{t^2}-\frac{1\times 3}{t^4}+\frac{1\times 3\times 5}{t^6}+\cdots+(-1)^{k+1}\frac{1\times 3\times 5\times\cdots\times(2k-1)}{t^{2k}}\right),$$

$$\theta_2(x) = (-1)^{k+1}1\times 3\times 5\times\cdots\times(2k+1)\int_a^x \frac{e^{-\frac{t^2}{2}}dt}{t^{2k+2}}.$$

Entonces, para $x > a > 0$, tenemos la siguiente estimación:

$$\begin{aligned}|\theta_2(x)| &\leq e^{-\frac{a^2}{2}}\left[1\times 3\times 5\times\cdots\times(2k+1)\right]\int_a^x t^{-(2k+2)}dt \\ &= \frac{1\times 3\times 5\times\cdots\times(2k+1)}{2k+1}e^{-\frac{a^2}{2}}\left(\frac{1}{a^{2k+1}}-\frac{1}{x^{2k+1}}\right) \\ &< \frac{1\times 3\times 5\times\cdots\times(2k+1)}{2k+1}\,\frac{e^{-\frac{a^2}{2}}}{a^{2k+1}} < \varepsilon = 10^{-10}.\end{aligned}$$

Para $a = 4{,}7$ y $k = 10$ se tiene $|\theta_2(x)| < 10^{-10}$.

Por otro lado, $\theta_1(8) < 10^{-10}$. En consecuencia

$$\frac{1}{\sqrt{2\pi}}\int_x^8 e^{-\frac{t^2}{2}}dt = \frac{1}{\sqrt{2\pi}}\left(\theta_1(t)|_x^8 + \theta_2(x)\right) = \frac{1}{\sqrt{2\pi}}\left(\theta_1(8) - \theta_1(x) + \theta_2(x)\right),$$

con lo cual $\frac{1}{\sqrt{2\pi}}\int_x^8 e^{-\frac{t^2}{2}}dt$ se aproxima mediante $-\frac{1}{\sqrt{2\pi}}\theta_1(x)$ con una precisión $\varepsilon = 10^{-10}$, y

$$\varphi_r(x) \simeq 1 - \frac{1}{\sqrt{2\pi}}\int_x^8 e^{-\frac{t^2}{2}}dt$$

se aproxima mediante $1 + \frac{1}{\sqrt{2\pi}}\theta_1(x)$ para $x \in [4{,}7,\ 8[$.

Escribamos $\theta_1(x)$ en forma anidada de modo que se adapte a la estabilidad numérica. Tenemos

$$\begin{aligned}\theta_1(x) &= \frac{e^{-\frac{t^2}{2}}}{x}\left(\frac{1}{x^2}\left(1+\frac{3\times 5}{x^4}\left(1+\frac{7\times 9}{x^4}\left(1+\frac{11\times 13}{x^4}\left(1+\frac{11\times 17}{x^4}\right)\right)\right)\right)\right) \\ &\quad -\left(1+\left(\frac{1\times 3}{x^4}\left(1+\frac{5\times 7}{x^4}\left(1+\frac{9\times 11}{x^4}\left(1+\frac{13\times 15}{x^4}\left(1+\frac{17\times 19}{x^4}\right)\right)\right)\right)\right)\right),\end{aligned}$$

cuyo algoritmo es el siguiente:

Algoritmo

Datos de entrada: $x \in [4{,}7,\ 8[$.

Datos de salida: $x \in [4{,}7,\ 8[$.

1. $y = x^4$.

2. $b_1 = 1$.

3. $b_2 = 1$.

4. $\left[\begin{array}{ll} j = & 1, \ldots, 4 \\ & k = 6 - j \\ & b_1 = 1 + \frac{(4k-1)(4k-3)}{y} b_1 \\ & b_2 = 1 + \frac{(4k-3)(4k-5)}{y} b_2 \end{array}\right.$

   Fin de bucle j.

6. $\theta_1(x) = \frac{e^{-\frac{x^2}{2}}}{x}\left(\frac{b_2}{x^2} - \frac{3b_1}{y} - 1\right) = -\frac{e^{-\frac{x^2}{2}}}{x}\left(1 + \frac{3b_1}{y} - \frac{b_2}{x^2}\right)$.

7. Imprimir $\theta_1(x)$.

8. Fin.

Por lo tanto $\varphi(x)$ se aproxima mediante la función $\Phi(x)$ definida a continuación:

$$\Phi(x) = \begin{cases} 0, \quad \text{si } x \leq -8, \\ -\frac{1}{\sqrt{2\pi}}\theta_1(x), \quad \text{si } x \in]-8, -4{,}7[\,, \\ 0{,}5 - \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{m} \frac{(-1)^k x^{2k+1}}{k!\, 2^k (2k+1)}, \quad \text{si } x \in [-4{,}7, 0]\,, \\ 0{,}5 + \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{m} \frac{(-1)^k x^{2k+1}}{k!\, 2^k (2k+1)}, \quad \text{si } x \in]0, 4{,}7]\,, \\ 1 + \frac{1}{\sqrt{2\pi}}\theta_1(x), \ \text{si } x \in]4{,}7, 8[\,, \\ 1, \quad \text{si } x \geq 8. \end{cases}$$

Queda por determinar m tal que para todo $x \in [-4{,}7,\ 4{,}7]$ se verifique

$$\left| \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{k!\, 2^k (2k+1)} - \sum_{k=0}^{m} \frac{(-1)^k x^{2k+1}}{k!\, 2^k (2k+1)} \right| < \epsilon = 10^{-10}.$$

En la siguiente tabla se muestran los valores de m_j para $x_j = 1, 2, 3, 4, 4{,}7$.

x_j	1	2	3	4	4,7
m_j	11	23	31	43	55

Sean $m_1 = \frac{m_{j-1}}{2}$ y

$$\begin{aligned} S_{m_j} &= \sum_{k=0}^{m_J} \frac{(-1)^k\, x^{2k+1}}{k!\, 2^k (2k+1)} \\ &= \sum_{k=0}^{m_1} \frac{x}{(2k)!\,(4k+1)} \left(\frac{x^2}{2}\right)^{2k} - \sum_{k=0}^{m_1} \frac{x}{(2k+1)!(4k+3)} \left(\frac{x^2}{2}\right)^{2k+1} \\ &= S_1(x) - S_2(x) \quad x \in [-4{,}7, 4{,}7]\,. \end{aligned}$$

La escritura anidada de $S_1(x)$ y $S_2(x)$ garantizan la estabilidad numérica.

Se propone como ejercicio elaborar un algoritmo completo para calcular $\Phi(x)$ (valores aproximados de $\varphi(x)$). Así mismo, elabore un programa computacional y los resultados numéricos compare con los datos provistos en las tablas de la distribución normal proporcionados en los libros de probabilidad y estadística.

4.10. Distribución χi- cuadrada

Definición 8 *Una variable aleatoria X que tiene una función de distribución de probabilidad definida por*

$$f(t) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} \qquad t > 0,\ n = 1, 2, 3, \ldots,$$

se dice que tiene una distribución χi-cuadrada con n grados de libertad.
La función de distribución χ-cuadrada está definida mediante

$$F(n, x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \qquad x \geq 0,\ n = 1, 2, \ldots.$$

La distribución χi-cuadrada es un caso particular de la distribución tipo gama (véase la distribución tipo gama) cuando $p = \frac{1}{n}$ y $\beta = \frac{1}{2}$. Esta función es muy importante en estadística y probabilidades.

Utilizando el cambio de variable $t = 2u$, obtenemos

$$F(n, x) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{x}{2}} u^{\frac{n}{2}-1} e^{-u} du.$$

Dados $n \in \mathbb{Z}^+$ y $x \geq 0$, para calcular o aproximar $F(n, x)$ consideramos cuatro casos.

1. Si $n = 1$, utilizando el cambio de variable $t = \frac{u^2}{2}$, tenemos

$$F(1, x) = \frac{1}{\Gamma\left(\frac{1}{2}\right)} \int_0^{\frac{x}{2}} t^{-\frac{1}{2}} e^{-t} dt = \sqrt{2\pi} \int_0^{\sqrt{x}} e^{-\frac{t^2}{2}} dt.$$

La integral $\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}} dt$ no puede calcularse mediante funciones elementales, lo que nos conduce a aproximarla numéricamente. En la sección relativa a la distribución normal se dió una técnica de aproximación de dicha integral que la escribimos inmediatamente a continaución.

$$\tilde{F}(1, x) = \begin{cases} 0, & \text{si } x \leq 0, \\ \sqrt{\frac{2}{\pi}} \sum_{k=0}^{m} \frac{(-1)^k}{k!\, 2^k} \frac{x^{k+\frac{1}{2}}}{2k-1}, & \text{si } x \in]0, 22], \\ 1 + \sqrt{\frac{2}{\pi}}\theta_1(x), & \text{si } x \in]22, 64[, \\ 1, & \text{si } x \geq 64, \end{cases}$$

donde

$$\begin{aligned} \theta_1\left(\sqrt{x}\right) = \frac{e^{-\frac{x}{2}}}{\sqrt{x}} \Bigg[\frac{1}{x}\Bigg[\left(1 + \frac{3\times 5}{x^2}\left(1 + \frac{7\times 5}{x^2}\left(1 + \frac{11\times 13}{x^2}\left(1 + \frac{15\times 17}{x^2}\right)\right)\right)\right) \\ - \left(1 + \frac{1\times 3}{x^2}\left(1 + \frac{5\times 7}{x^2}\left(1 + \frac{9\times 11}{x^2}\left(1 + \frac{13\times 15}{x^2}\left(1 + \frac{17\times 19}{x^2}\right)\right)\right)\right)\right)\Bigg]\Bigg], \end{aligned}$$

Además,

$$\sum_{k=0}^{m} \frac{(-1)^k}{k!\, 2^k} \frac{x^{k+\frac{1}{2}}}{2k+1} = \sqrt{x}\left(\sum_{k=0}^{m_1} \frac{x^{2k}}{(2k)!\, 2^{2k}(4k+3)} - x\sum_{k=0}^{m_1} \frac{x^{2k}}{(2k+1)!2^{2k+1}(4k+3)}\right),$$

y cada sumatorio se escribe en forma anidada, donde $m_1 = \frac{m+1}{2}$ con m impar y para $x \in [x_{j-1}, x_j]$, $j = 1, \ldots, 5$, m y x_j están dados en la siguiente tabla ($x_0 = 0$):

x_j	2	4	9	16	22
m	11	21	31	43	55

.

2. Si $n = 2$, entonces

$$F(2,x) = \frac{1}{\Gamma(1)} \int_0^{\frac{x}{2}} e^{-t} dt = 1 - e^{-\frac{x}{2}},\ x \geq 0.$$

Sean $n > 2$. Integrando por partes k veces $\int_0^{\frac{x}{2}} t^{\frac{n}{2}-1} e^{-t} dt$, obtenemos

$$\begin{aligned} F(n,x) &= \frac{e^{-\frac{x}{2}}}{\Gamma\left(\frac{n}{2}\right)} \left[-\left(\frac{x}{2}\right)^{\frac{n}{2}-1} - \left(\frac{n}{2}-1\right)\left(\frac{x}{2}\right)^{\frac{n}{2}-2} - \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\left(\frac{x}{2}\right)^{\frac{n}{2}-3} - \cdots - \right. \\ &\quad \left. \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{n}{2}-k\right)\left(\frac{x}{2}\right)^{\frac{n}{2}-(k+1)} \right] + \\ &\quad \frac{\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{n}{2}-(k+1)\right)}{\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{x}{2}} t^{\frac{n}{2}-(k+2)} e^{-t} dt. \end{aligned}$$

3. Si $n = 2k+4$, $x = 0, 1, 2, \ldots$. Entonces

$$\begin{aligned} \Gamma\left(\frac{n}{2}\right) &= \Gamma\left(\frac{2k+4}{2}\right) = (k+1)!, \\ \int_0^{\frac{x}{2}} t^{\frac{n}{2}-(k+2)} e^{-t} dt &= \int_0^{\frac{x}{2}} e^{-t} dt = 1 - e^{-\frac{x}{2}}, \end{aligned}$$

y reemplazando en la desarrollo precedente de $F(n,k)$, tenemos

$$\begin{aligned} F(2k+4,x) &= 1 - e^{-\frac{x}{2}} \left[1 + \frac{1}{(k+1)!}\left(\frac{x}{2}\right)^{k+1} + \frac{k+1}{(k+1)!}\left(\frac{x}{2}\right)^{k} + \frac{k(k+1)}{(k+1)!}\left(\frac{x}{2}\right)^{k-1} + \cdots + \frac{x}{2} \right] \\ &= 1 - e^{-\frac{x}{2}} \sum_{j=0}^{k+1} \frac{1}{j!}\left(\frac{x}{2}\right)^{j}. \end{aligned}$$

Así, si $n = 2k+4$, $k = 0, 1, 2, \ldots$,

$$F(2k+4,x) = 1 - e^{-\frac{x}{2}} \sum_{j=0}^{k+1} \frac{1}{j!}\left(\frac{x}{2}\right)^{j} \quad x \geq 0.$$

Por otro lado,

$$\begin{aligned} 1 &= \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} t^{\frac{n}{2}-1} e^{t} dt = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\int_0^{\frac{x}{2}} t^{\frac{n}{2}-1} e^{-t} dt + \int_{\frac{x}{2}}^{\infty} t^{\frac{n}{2}-1} e^{-t} dt \right) \\ &= F(2k+4,x) + \frac{1}{\Gamma(k+2)} \int_{\frac{x}{2}}^{\infty} t^{k+1} e^{-t} dt \\ &= 1 - e^{-\frac{x}{2}} \sum_{j=0}^{k+1} \frac{1}{j!}\left(\frac{x}{2}\right)^{j} + \frac{1}{(k+1)!} \int_{\frac{x}{2}}^{\infty} t^{k+1} e^{-t} dt, \end{aligned}$$

de donde

$$\int_{\frac{x}{2}}^{\infty} t^{k+1} e^{-t} dt = (k+1)!\ e^{-\frac{x}{2}} \sum_{j=0}^{k+1} \frac{1}{j!}\left(\frac{x}{2}\right)^{j} \quad x \geq 0.$$

Así por ejemplo

$$\int_1^{\infty} t^3 e^{-t} dt = 2\, e^{-1} \sum_{j=0}^{3} \frac{1}{j!} = \frac{4\, e^{-1}}{3}.$$

4. Supongamos que $n = 2k+3$, $k = 0, 1, 2, \ldots$. Entonces

$$\Gamma\left(\frac{n}{2}\right) = \Gamma\left(k + \frac{3}{2}\right) = \frac{(2k+1)(2k-1) \times \cdots \times 1}{2^{k+1}} \sqrt{\pi},$$

$$\frac{\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{n}{2}-(k+1)\right)}{\Gamma\left(\frac{n}{2}\right)}=\frac{\left(k+\frac{1}{2}\right)\left(k-\frac{1}{2}\right)\times\cdots\times\frac{1}{2}}{\Gamma\left(k+\frac{3}{2}\right)}=\frac{1}{\sqrt{\pi}},$$

$$\int_0^{\frac{x}{2}} t^{\frac{n}{2}-(k+2)}e^{-t}dt=\int_0^{\frac{x}{2}} t^{-\frac{1}{2}}e^{-t}dt=\sqrt{2}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt.$$

Esta última integral se aproxima como en la parte 1).

En consecuencia,

$$\begin{aligned}
F(2k+3,x) &= \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt-\frac{e^{-\frac{x}{2}}}{\Gamma\left(\frac{n}{2}\right)}\left(\left(\frac{x}{2}\right)^{\frac{n}{2}}+\left(\frac{n}{2}-1\right)\left(\frac{x}{2}\right)^{\frac{n}{2}-2}+\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\right.\\
&\quad \left.\left(\frac{x}{2}\right)^{\frac{n}{2}-3}+\cdots\div\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{n}{2}-k\right)\left(\frac{x}{2}\right)^{\frac{n}{2}-(k+1)}\right)\\
&= \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt-e^{-\frac{x}{2}}\left(\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{k+\frac{1}{2}}+\frac{k+\frac{1}{2}}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{k-\frac{1}{2}}+\right.\\
&\quad \left.\frac{\left(k+\frac{1}{2}\right)\left(k-\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{k-\frac{3}{2}}+\cdots+\frac{\left(k+\frac{1}{2}\right)\left(k-\frac{1}{2}\right)\times\cdots\times\frac{3}{2}}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{\frac{1}{2}}\right)\\
&= \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt-\sqrt{\frac{x}{2}}e^{-\frac{x}{2}}\left(\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{k}+\frac{k+\frac{1}{2}}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{k-1}+\right.\\
&\quad \left.\frac{\left(k+\frac{1}{2}\right)\left(k-\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{x}{2}\right)^{k-2}+\cdots+\frac{\left(k+\frac{1}{2}\right)\left(k-\frac{1}{2}\right)\times\cdots\times\frac{3}{2}}{\Gamma\left(\frac{n}{2}\right)}\right)\\
&= \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt-\sqrt{\frac{x}{2}}e^{-\frac{x}{2}}\left(\frac{2}{\sqrt{\pi}}+\frac{2^2}{1\times3\sqrt{\pi}}\frac{x}{2}+\frac{2^3}{1\times3\times5\sqrt{\pi}}\left(\frac{x}{2}\right)^2+\cdots+\right.\\
&\quad \left.\frac{2^k}{1\times3\times\cdots\times(2k-1)\sqrt{\pi}}\left(\frac{x}{2}\right)^{k-1}+\frac{2^{k+1}}{1\times3\times\cdots\times(2k+1)\sqrt{\pi}}\left(\frac{x}{2}\right)^{k}\right).
\end{aligned}$$

Ponemos

$$\begin{aligned}
\theta(x) &= 1+\frac{2}{1\times3}\frac{x}{2}+\frac{2^2}{1\times3\times5}\left(\frac{x}{2}\right)^2+\cdots+\frac{2^{k-1}}{1\times3\times\cdots\times(2k-1)}\left(\frac{x}{2}\right)^{k-1}\\
&\quad +\frac{2^k}{1\times3\times\cdots\times(2k-1)}\left(\frac{x}{2}\right)^{k}\\
&= 1+\frac{x}{3}\left(1+\frac{x}{5}\left(1+\frac{x}{7}\left(1+\cdots+\frac{x}{2k-1}\left(1+\frac{x}{2k+1}\right)\cdots\right)\right)\right).
\end{aligned}$$

Resulta que

$$F(2k+3,x)=\frac{\sqrt{2}}{\pi}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt-\sqrt{\frac{2x}{\pi}}e^{-\frac{x}{2}}\theta(x),\quad x\geq0.$$

En resumen,

$$\begin{aligned}
F(1,x) &= \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt \quad x\geq0,\\
F(2,x) &= 1-e^{-\frac{x}{2}} \quad x\geq0,\\
F(2k+4,x) &= 1-e^{-\frac{x}{2}}\sum_{j=0}^{k+1}\frac{1}{j!}\left(\frac{x}{2}\right)^j \quad x\geq0,\ k=0,1,2,\ldots,\\
F(2k+3,x) &= \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt-\sqrt{\frac{2x}{\pi}}e^{-\frac{x}{2}}\theta(x),\ x\geq0, k=0,1,2,\ldots,
\end{aligned}$$

donde la integral $\int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}dt$ se aproxima como en 1).

4.11. Distribución t de Student

Definición 9 *Una variable aleatoria T se dice que tiene una distribución de probabilidad t de Student con n grados de libertad si su función densidad está definida por*

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \quad t \geq 0.$$

La función de distribución t de Student está definida mediante:

$$F(n,k) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{x} \left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}} dt, \quad x \in \mathbb{R}.$$

Cuando $n = 1$, tenemos

$$F(1,x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{dt}{1+t^2} = \frac{1}{2} + \frac{1}{\pi}\arctan(x), \quad x \in \mathbb{R},$$

que es conocida con el nombre de distribución de Cauchy.

Sean $n \in \mathbb{Z}^+$ y $h(n,t) = \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$, $t \in \mathbb{R}$. Entonces $h(-t) = h(t) \quad \forall t \in \mathbb{R}$, luego

$$F(n,x) = 0{,}5 + \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \int_0^x h(n,t)dt, \ \ x \in \mathbb{R},$$

y si $x < 0$, $\int\limits_0^x h(n,t)dt = -\int\limits_0^{-x} h(n,t)dt$, se sigue que

$$F(n,x) = \begin{cases} 0{,}5 - \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \int_0^{-x} h(n,t)dt, & \text{si } x < 0, \\ 0{,}5 + \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \int_0^{x} h(n,t)dt, & \text{si } \ x > 0. \end{cases}$$

Por otro lado,

$$1 = 0{,}5 + \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \left(\int_0^x h(n,t)dt + \int_x^\infty h(n,t)dt\right) \quad \forall x \in \mathbb{R},$$

de donde

$$F(n,x) = 1 - \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \int_x^\infty h(n,t)dt \quad \forall x \in \mathbb{R}.$$

Sea $\varepsilon > 0$ ($\varepsilon = 10^{-10}$). Ponemos

$$G(n,x) = \int_0^x h(n,t)dt \quad x \in \mathbb{R},\ n = 2,3,\ldots.$$

En lo que sigue, nos ocuparemos de aproximar $G(n,x)$ con una precisión ε. Para $n \in \mathbb{Z}^+$ fijo, $n > 1$, $\Gamma\left(\frac{n+1}{2}\right)$ y $\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)$ se calculan con una precisión ε.

Para obtener una relación que ligue x con n de modo que

$$\int_x^\infty h(n,t)dt < 10^{-10},$$

aplicamos el criterio de comparación para integrales impropias. Para el efecto, sea $g(t) = \frac{1}{t^2}$, $t \geq 1$. Entonces $\int_1^\infty g(t)dt = 1$, y

$$\frac{h(n,t)}{g(t)} = t^2\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \underset{t\to\infty}{\longrightarrow} 0, \quad n > 1,$$

luego existe $x > 1$ tal que $t^2\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} < 10^{-10}$, si $t \geq x$, $n > 1$.

Por ejemplo para $n = 11$, esta última relación se verifica si $t \geq 40$. Para $n = 40$ se tiene $t \geq 12$, para $n = 140$ obtenemos $t \geq 8$.

Dado $n \in \mathbb{Z}^+$ con $n > 1$, notamos con

$$\hat{x}_n = Min\left\{t > 0 \mid t^2\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} < 10^{-10}\right\},$$

el término $\int_{\hat{x}_n}^\infty h(n,t)dt$ será despreciado, pués por el criterio del cociente se tiene

$$\int_{\hat{x}_n}^\infty h(n,t)dt < 10^{-10}\int_{\hat{x}_n}^\infty \frac{dt}{t^2} \leq 10^{-10}\int_1^\infty \frac{dt}{t^2} = 10^{-10},$$

consecuentemente $F(n,x)$ se aproxima por 1 si $x \geq \hat{x}_n$.

La aproximación de $G(n,x)$ se limitará al intervalo $]0,\hat{x}_n[$.

Sea $u = \arctan(\frac{t}{\sqrt{n}})$ entonces $\tan(u) = \frac{t}{\sqrt{n}}$y para $t = x$, notaremos $y_n(x) = \arctan(\frac{x}{\sqrt{n}})$. Entonces

$$\tan(y_n(x)) = \frac{x}{\sqrt{n}} = \frac{\operatorname{sen}(y_n(x))}{\cos(y_n(x))},$$

de donde

$$\begin{aligned}\operatorname{sen}(y_n(x)) &= \frac{x}{\sqrt{n+x^2}},\\ \cos(y_n(x)) &= \left(\frac{n}{n+x^2}\right)^{1/2}.\end{aligned}$$

Además

$$\begin{aligned}G(n,x) &= \int_0^x h(n,t)dt = \int_0^{y_n(x)} \frac{\sqrt{n}\sec^2(u)du}{(1+\tan^2(u))^{\frac{n+1}{2}}} = \sqrt{n}\int_0^{y_n(x)} \cos^{n-1}(u)du\\ &= \sqrt{n}\int_0^{y_n(x)} \cos(t)\cos^{n-2}(t)dt \quad n \geq 2.\end{aligned}$$

Integrando por partes, tenemos

$$\begin{aligned}G(n.x) &= \sqrt{n}\left(\operatorname{sen}(y_n(x))\cos^{n-1}(y_n(x)) + (n-2)\int_0^{y_n(x)} \cos^{n-3}(t)(1-\cos^2(t))dt\right)\\ &= \sqrt{n}\left(\operatorname{sen}(y_n(x))\cos^{n-2}(y_n(x)) + (n-2)\int_0^{y_n(x)} \cos^{n-3}(t)dt\right) - (n-2)G(n,k)\end{aligned}$$

donde

$$G(n,k) = \sqrt{n}\left(\frac{(\operatorname{sen}(y_n(x))\cos^{n-2}(y_n(x))}{n-1} + \frac{n-2}{n-1}\int_0^{y_n(x)} \cos^{n-3}(t)dt\right) \quad n \geq 3.$$

Esta fórmula recursiva será utilizada para obtener una expresión general de $F(n,x)$ $n \geq 3$.

Para $n = 2$ tenemos

$$G(2,x) = \sqrt{2}\int_0^{y_2(x)} \cos(u)du = \sqrt{2}\,\mathrm{sen}(y_2(x)) = \sqrt{2}\frac{x}{\sqrt{2+x^2}},$$

$$F(2,x) = 0{,}5 + \frac{\Gamma\left(\frac{3}{2}\right)}{\sqrt{2\pi}\Gamma(1)}\sqrt{2}\frac{x}{\sqrt{2+x^2}} = 0{,}5 + \frac{1}{2}\frac{x}{\sqrt{2+x^2}} \quad x \geq 0.$$

Puesto que

$$\int_0^{y_n(x)} \cos^{n-3}(t)dt = \frac{\mathrm{sen}(y_n(x))\cos^{n-4}(y_n(x))}{n-3} + \frac{n-4}{n-3}\int_0^{y_n(x)} \cos^{n-5}(t)dt,$$

entonces

$$\begin{aligned} G(n,x) &= \sqrt{n}\,\mathrm{sen}(y_n(x))\left(\frac{\cos^{n-2}(y_n(x))}{n-1} + \frac{n-2}{(n-1)(n-3)}\cos^{n-4}(y_n(x))\right) + \\ &\quad + \frac{\sqrt{n}(n-2)(n-4)}{(n-1)(n-3)}\int_0^{y_n(x)} \cos^{n-5}(t)dt. \end{aligned}$$

Continuando con este procedimiento k veces, tenemos

$$\begin{aligned} G(n,x) &= \sqrt{n}\,\mathrm{sen}(y_n(x))\left(\frac{\cos^{n-2}(y_n(x))}{n-1} + \frac{n-2}{(n-1)(n-3)}\cos^{n-4}(y_n(x)) + \right. \\ &\quad \left. + \frac{(n-2)(n-4)}{(n-1)(n-3)(n-5)}\cos^{n-6}(y_n(x)) + \cdots + \frac{(n-2)(n-4)\times\cdots\times(n-2k+2)}{(n-1)(n-3)\times\cdots\times(n-2k+1)}\cos^{n-2k}(y_n(x))\right) \\ &\quad + \frac{\sqrt{n}(n-2)(n-4)\times\cdots\times(n-2k)}{(n-1)(n-3)\times\cdots\times(n-2k+1)}\int_0^{y_n(x)} \cos^{n-2k-1}(t)dt. \end{aligned}$$

Consideramos dos casos: n par y n impar.

1. Supongamos que n es impar; esto es, $n = 2k+1$, $k = 1,2,3,\ldots$. Entonces $n-2k-1 = 0$,

$$\int_0^{y_n(x)} dt = y_n(x),$$

$$\frac{\sqrt{n}(n-2)(n-4)\times\cdots\times(n-2k)}{(n-1)(n-3)\times\cdots\times(n-2k+1)} = \frac{\sqrt{n}(2k-1)(2k-3)\times\cdots\times 1}{2k(2k-2)\times\cdots\times 2} = \frac{\sqrt{n}}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)},$$

$$\begin{aligned} G(n,x) &= \frac{\sqrt{n}}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}y_n(x) + \sqrt{n}\,\mathrm{sen}(y_n(x))\left(\frac{\cos^{n-2}(y_n(x))}{n-1} + \frac{n-2}{(n-1)(n-3)}\cos^{n-4}(y_n(x)) + \right. \\ &\quad + \frac{(n-2)(n-4)}{(n-1)(n-3)(n-5)}\cos^{n-6}(y_n(x)) + \cdots + \\ &\quad \left. \frac{(n-2)(n-4)\times\cdots\times(n-2k+2)}{(n-1)(n-3)\times\cdots\times(n-2k+1)}\cos^{n-2k}(y_n(x))\right) \end{aligned}$$

Para $n = 2k+1$, se tiene la siguiente expresión para el cálculo de $F(2k+1,x)$ $\quad k = 1,2,\ldots$:

$$\begin{aligned} F(2k+1,x) &= 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{n}}) + \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\mathrm{sen}(y_n(x))\left(\frac{\cos^{n-2}(y_n(x))}{n-1}\right. \\ &\quad + \frac{n-2}{(n-1)(n-3)}\cos^{n-4}(y_n(x)) + \frac{(n-2)(n-4)}{(n-3)(n-5)}\cos^{n-6}(y_n(x)) + \cdots + \\ &\quad \left. \frac{(n-2)(n-4)\times\cdots\times 3}{(n-1)(n-3)\times\cdots\times 4}\cos^3(y_n(x)) + \frac{(n-2)(n-4)\times\cdots\times 1}{(n-1)(n-3)\times\cdots\times 2}\cos(y_n(x))\right). \end{aligned}$$

Reemplazando $\text{sen}(y_n(x)) = \frac{x}{\sqrt{n+x^2}}$, $\cos(y_n(x)) = \left(\frac{n}{n+x^2}\right)^{1/2}$ y tomando en cuenta que $\frac{n-2j}{2} = \frac{2k+1-2j}{2} = k-j+\frac{1}{2}$, $j = 1,2,\ldots,k$, se tiene

$$\begin{aligned} F(2k+1,x) &= 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{n}}) + \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\frac{x}{\sqrt{n+x^2}}\left(\frac{1}{n+1}\left(\frac{n}{n+x^2}\right)^{k-\frac{1}{2}} + \right. \\ &+\frac{n-2}{(n-1)(n-3)}\left(\frac{n}{n+x^2}\right)^{k-\frac{3}{2}} + \frac{(n-2)(n-4)}{(n-1)(n-3)(n-5)}\left(\frac{n}{n+x^2}\right)^{k-\frac{5}{2}} + \cdots + \\ &\left.\frac{(n-2)(n-4)\times\cdots\times 3}{(n-1)(n-3)\times\cdots\times 4}\left(\frac{n}{n+x^2}\right)^{\frac{3}{2}} + \frac{(n-2)(n-4)\times\cdots\times 1}{(n-1)(n-3)\times\cdots\times 2}\left(\frac{n}{n+x^2}\right)^{1/2}\right). \end{aligned}$$

Sea $z = \frac{n}{n+x^2}$. Entonces

$$\begin{aligned} F(2k+1,x) &= 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{n}}) + \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\frac{\sqrt{n}x}{n+x^2}\left(\frac{z^{k-1}}{2k} + \frac{2k-1}{2k(2k-2)}z^{k-2} + \right. \\ &+\frac{(2k-1)(2k-3)}{2k(2k-2)(2k-u)}z^{k-3} + \cdots + \frac{(2k-1)(2k-3)\times\cdots\times 3}{2k(2k-2)\times\cdots\times 4}z \\ &\left.+\frac{(2k-1)(2k-3)\times\cdots\times 1}{2k(2k-2)\times\cdots\times 2}\right) \\ &= 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{n}}) + \frac{\sqrt{n}}{\pi}\frac{x}{n+x^2}\left(1 + \frac{2}{3}z + \frac{2\times 4}{3\times 5}z^2 + \frac{2\times 4\times 6}{1\times 3\times 5\times 7}z^3 + \cdots + \right. \\ &\left.\frac{2\times 4\times\cdots\times 2(k-2)}{1\times 3\times 5\times\cdots\times(2k-3)}z^{k-2} + \frac{2\times 4\times\cdots\times 2(k-1)}{1\times 3\times 5\times\cdots\times(2k-1)}z^{k-1}\right). \end{aligned}$$

Por ejemplo, para $x \geq 0$ se tiene: $z = \frac{n}{n+x^2}$, y

$$F(3,x) = 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{3}}) + \frac{\sqrt{3}}{\pi}\frac{x}{3+x^2},$$

$$F(5,x) = 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{5}}) + \frac{\sqrt{5}}{\pi}\frac{x}{5+x^2}\left(1 + \frac{2}{3}z\right),$$

$$F(7,x) = 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{7}}) + \frac{\sqrt{7}}{\pi}\frac{x}{7+x^2}\left(1 + \frac{2}{3}z\left(1 + \frac{4}{5}z\right)\right),$$

$$F(9,x) = 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{9}}) + \frac{\sqrt{9}}{\pi}\frac{x}{9+x^2}\left(1 + \frac{2}{3}z\left(1 + \frac{4}{5}z\left(1 + \frac{6}{7}z\right)\right)\right),$$

así sucesivamente.

2. Supongamos que $n = 2k+2$, $k = 0,1,2,\ldots$. Entonces $n-2k-1 = 1$,

$$\int_0^{y_n(x)} \cos^{n-2k-1}(t)dt = \int_0^{y_n(x)} \cos(t)dt = \text{sen}(y_n(x)),$$

y como

$$\frac{\sqrt{n}(n-2)(n-4)\times\cdots\times(n-2k)}{(n-1)(n-3)\times\cdots\times(n-2k+1)} = \sqrt{n}\frac{2k(2k-2)\times\cdots\times 2}{(2k+1)(2k-1)\times\cdots\times 3} = \sqrt{\pi n}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)},$$

se obtiene

$$\begin{aligned} G(n,x) &= \sqrt{\pi n}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\text{sen}(y_n(x)) + \sqrt{n}\,\text{sen}(y_n(x))\left(\frac{\cos^{n-2}(y_n(x))}{n-1} + \frac{n-2}{(n-1)(n-3)}\cos^{n-4}(y_n(x))\right. \\ &+\frac{(n-2)(n-4)}{(n-1)(n-3)(n-5)}\cos^{n-6}(y_n(x)) + \cdots + \\ &\left.\frac{(n-2)(n-4)\times\cdots\times(n-2k+2)}{(n-1)(n-3)\times\cdots\times(n-2k+1)}\cos^{n-2k}(y_n(x))\right). \end{aligned}$$

Resulta que

$$\begin{aligned} F(2k+2,x) &= 0{,}5+\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}G(n,x) \\ &= 0{,}5+\frac{1}{2}\operatorname{sen}(y_n(x))+\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\operatorname{sen}(y_n(x))\left(\frac{\cos^{n-2}(y_n(x))}{n-1}+\frac{n-2}{(n-1)(n-3)}\right. \\ &\quad \cos^{n-4}(y_n(x))+\frac{(n-2)(n-4)}{(n-1)(n-3)(n-5)}\cos^{n-6}(y_n(x))+\cdots+ \\ &\quad \left.\frac{(n-2)(n-4)\times\cdots\times(n-2k+2)}{(n-1)(n-3)\times\cdots\times(n-2k+1)}\cos^{n-2k}(y_n(x))\right), \end{aligned}$$

$$\begin{aligned} F(2k+2,x) &= 0{,}5+\frac{x}{2\sqrt{n+x^2}}+\frac{\Gamma\left(k+\frac{3}{2}\right)}{\sqrt{\pi}\Gamma(k+1)}\frac{x}{\sqrt{n+x^2}}\left(\frac{1}{2k+1}\left(\frac{n}{n+x^2}\right)^k+\right. \\ &\quad \frac{2k}{(2k+1)(2k-1)}\left(\frac{n}{n+x^2}\right)^{k-1}+\frac{2k(2k-2)}{(2k+1)(2k-1)(2k-3)}\left(\frac{n}{n+x^2}\right)^{k-2}+\cdots+ \\ &\quad \left.\frac{2k(2k-2)\times\cdots\times 4}{(2k+1)(2k-1)\times\cdots\times 3}\times\frac{n}{n+x^2}\right), \end{aligned}$$

$$\begin{aligned} F(2k+2,x) &= 0{,}5+\frac{x}{2\sqrt{n+x^2}}\left(1+\frac{1}{2}\frac{n}{n+x^2}+\frac{3}{2^3}\left(\frac{n}{n+x^2}\right)^2+\frac{5\times 3}{3\times 2^4}\left(\frac{n}{n+x^2}\right)^3\right. \\ &\quad +\cdots+\frac{(2k-5)(2k-7)\times\cdots\times 3\times 1}{(k-2)!\ 2^{k-2}}\left(\frac{n}{n+x^2}\right)^{k-2}+ \\ &\quad \frac{(2k-3)(2k-5)\times\cdots\times 3\times 1}{(k-2)!\ 2^{k-1}}\left(\frac{n}{n+x^2}\right)^{k-1}+ \\ &\quad \left.\frac{(2k-1)(2k-3)\times\cdots\times 1}{k!\ 2^k}\left(\frac{n}{n+x^2}\right)^k\right). \end{aligned}$$

Haciendo $z=\frac{n}{n+x^2}$, tenemos

$$\begin{aligned} F(2k+2,x) &= 0{,}5+\frac{1}{2}\frac{x}{\sqrt{n+x^2}}\left(1+\frac{z}{2}+\frac{3}{2^3}z^2+\frac{5\times 3}{3\times 2^4}z^3+\cdots+\right. \\ &\quad \frac{(2k-5)(2k-7)\times\cdots\times 3\times 1}{(k-2)!\,2^{k-2}}z^{k-2}+\frac{(2k-3)(2k-5)\times\cdots\times 3\times 1}{(k-1)!\,2^{k-1}}z^{k-1}+ \\ &\quad \left.+\frac{(2k-1)(2k-3)\times\cdots\times 1}{k!\,2^k}z^k\right) \end{aligned}$$

Por ejemplo,

$$\begin{aligned} F(4,x) &= 0{,}5+\frac{1}{2}\frac{x}{\sqrt{4+x^2}}\left(1+\frac{1}{2}\frac{4}{4+x^2}\right)=0{,}5+\frac{1}{2}\frac{x}{\sqrt{4+x^2}}\left(1+\frac{2}{4+x^2}\right) \\ F(6,x) &= 0{,}5+\frac{1}{2}\frac{x}{\sqrt{6+x^2}}\left(1+\frac{1}{2}\frac{6}{6+x^2}+\frac{3}{2^3}\left(\frac{6}{6+x^2}\right)^2\right) \\ &= 0{,}5+\frac{1}{2}\frac{x}{\sqrt{6+x^2}}\left(1+\frac{1}{2}\frac{6}{6+x^2}\left(1+\frac{3}{4}\frac{6}{6+x^2}\right)\right) \\ F(8,x) &= 0{,}5+\frac{1}{2}\frac{x}{\sqrt{8+x^2}}\left(1+\frac{1}{2}\frac{8}{8+x^2}\left(1+\frac{3}{4}\frac{8}{8+x^2}\left(1+\frac{5}{2}\frac{8}{8+x^2}\right)\right)\right). \end{aligned}$$

En resumen

$F(1,x)=\frac{1}{2}+\frac{1}{\pi}\arctan(x)$, $x\in\mathbb{R}$,

$F(2,x) = 0{,}5 + \frac{1}{2}\frac{x}{\sqrt{2+x^2}}$, $x \geq 0$.

Para $n = 2k+1$, $k = 1,2,3,\ldots$ y $x \geq 0$, $z = \frac{n}{n+x^2}$,

$$\begin{aligned} F(2k+1,x) &= 0{,}5 + \frac{1}{\pi}\arctan(\frac{x}{\sqrt{n}}) + \frac{\sqrt{n}}{\pi}\frac{x}{n+x^2}\left(1 + \frac{2}{3}z + \frac{2\times 4}{3\times 5}z^2 + \frac{2\times 4\times 6}{1\times 3\times 5\times 7}z^3 + \cdots + \right. \\ &\left. \frac{2\times 4\times\cdots\times 2(k-2)}{1\times 3\times 5\times\cdots\times(2k-3)}z^{k-2} + \frac{2\times 4\times\cdots\times 2(k-1)}{1\times 3\times\cdots\times(2k-1)}z^{k-1}\right) \end{aligned}$$

Para $n = 2k+2$, $k = 1,2,3,\ldots, x \geq 0$, $z = \frac{n}{n+x^2}$,

$$\begin{aligned} F(2k+2,x) &= 0{,}5 + \frac{1}{2}\frac{x}{\sqrt{n+x^2}}\left(1 + \frac{1}{2}z + \frac{3}{2^3}z^2 + \frac{5\times 3}{3\times 2\times 2^3}z^3 + \cdots + \right. \\ &\left. \frac{(2k-3)(2k-5)\times\cdots\times 3\times 1}{(k-1)!\,2^{k-1}}z^{k-1} + \frac{(2k-1)(2k-3)\times\cdots\times 1}{k!\,2^k}z^k\right) \end{aligned}$$

Además $x \in\,]-\hat{x}_n, \hat{x}_n[$, donde $n \in \mathbb{Z}^+$ y

$$\hat{x}_n = \text{mín}\left\{t > 0 | t^2\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} < 10^{-10}\right\},$$

$F(n,x)$ se aproxima por 0 si $x \leq -\hat{x}_n$ y se aproxima por 1 si $x \geq \hat{x}_n$.

Por otro lado, para $x \in\,]-\hat{x}_n, x_n[$, $F(n,x)$ se escribe en forma anidada.

Se recomienda al lector elaborar un algoritmo para el cálculo de $F(n,x)$ así como su respectivo programa computacional. Los resultados del programa deben compararse con las tablas de la distribución t de Student proporcionados en los libros de probabilidad y estadística.

4.12. Distribución F (de Snedekor)

Definición 10 *Sean Y, Z variables aleatorias independientes que tienen distribuciones χ-cuadrada con m y n grados de libertad, respectivamente. La variable aleatoria*

$$X = \frac{\frac{y}{m}}{\frac{z}{n}}$$

tiene una distribución F definida por

$$F(m,n,x) = m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{-\frac{1}{2}}(n+mt)^{-\frac{m+n}{2}}dt \quad x \geq 0.$$

Para $m = 1$, se tiene

$$F(1,n,x) = \frac{n^{\frac{n}{2}}\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{-\frac{1}{2}}(n+t)^{-\frac{n+1}{2}}dt \quad x \geq 0.$$

Efectuando el cambio de variable $t = u^2$, se obtiene

$$F(1,n,x) = \frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \int_0^{\sqrt{x}} \left(1 + \frac{u^2}{n}\right)^{-\frac{n+1}{2}} du,$$

que tiene la forma de la distribución t de Student. La función

$$G(n,\sqrt{x}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \int_0^{\sqrt{x}} (1 + \frac{u^2}{n})^{-\frac{n+1}{2}}du \quad x \geq 0,$$

se aproxima utilizando el algoritmo de aproximación de la distribución t de Student. Luego

$$F(1,n,x) = 2\,G(n,\sqrt{x}) \quad x \geq 0,\ \ n = 1,2,\cdots .$$

Si $m = 2$, tenemos

$$F(2,n,x) = \frac{2\,\Gamma\left(\frac{n+2}{2}\right)}{n\,\Gamma\left(\frac{n}{2}\right)} \int_0^x \left(1+\frac{2}{n}t\right)^{-\frac{n+2}{2}} dt = 1 - \left(1+\frac{2x}{n}\right)^{-\frac{n}{2}} \quad x \geq 0,\ n = 1,2,\cdots .$$

Sean $m, n \in \mathbb{Z}^+$ con $m > 2$. Entonces

$$F(m,n,x) = \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{\frac{m}{2}-1} \left(1+\frac{mt}{n}\right)^{-\frac{m+n}{2}} dt.$$

Utilizando el cambio de variables $u = \frac{mt}{n}$, tenemos

$$\begin{aligned} F(m,n,x) &= \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{mx}{n}} \left(\frac{nu}{m}\right)^{\frac{m}{2}-1} (1+u)^{-\frac{m+n}{2}} \frac{n}{m} du \\ &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{mx}{n}} t^{\frac{m}{2}-1} (1+t)^{-\frac{m+n}{2}} dt \quad x \geq 0. \end{aligned}$$

Para $m, n \in \mathbb{Z}^+$ tal que $m > 2$, definimos

$$I(m,n,\alpha) = \int_0^\alpha t^{\frac{m}{2}-1}(1+t)^{-\frac{m+n}{2}} dt \quad \alpha \geq 0,$$

y mediante el método de integración por partes, obtenemos

$$I(m,n,\alpha) = -\frac{\alpha^{\frac{m}{2}-1}(1+\alpha)^{-\frac{m-2+n}{2}}}{\frac{m-2+n}{2}} + \frac{\frac{m}{2}-1}{\frac{m-2+n}{2}} \times I(m-2,n,x).$$

Esta fórmula recursiva la aplicaremos sucesivamente para obtener una expresión que nos permita describir un algoritmo de cálculo de $F(m,n,\alpha)$. Así,

$$\begin{aligned} I(m-2,n,\alpha) &= -\frac{\alpha^{\frac{m-2}{2}-1}(1+\alpha)^{-\frac{m-4+n}{2}}}{\frac{m-4+n}{2}} + \frac{\frac{m-2}{2}-1}{\frac{m-4+n}{2}} \times I(m-4,n,\alpha), \\ I(m-4,n,\alpha) &= -\frac{\alpha^{\frac{m-4}{2}-1}(1+\alpha)^{-\frac{m-6+n}{2}}}{\frac{m-6+n}{2}} + \frac{\frac{m-4}{2}-1}{\frac{m-6+n}{2}} \times I(m-6,n,\alpha). \end{aligned}$$

Entonces

$$\begin{aligned} I(m,n,\alpha) &= -\frac{\alpha^{\frac{m}{2}-1}(1+\alpha)^{-\frac{m-2+n}{2}}}{\frac{m-2+n}{2}} - \frac{\left(\frac{m}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)} \alpha^{\frac{m-2}{2}-1}(1+\alpha)^{-\frac{m-4+n}{2}} \\ &\quad - \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\left(\frac{m-6+n}{2}\right)} \alpha^{\frac{m-4}{2}-1}(1+\alpha)^{-\frac{m-6+n}{2}} + \\ &\quad \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\left(\frac{m-4}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\left(\frac{m-6+n}{2}\right)} \times I(m-6,n,\alpha). \end{aligned}$$

Continuando con este procedimiento k veces, obtenemos

$$\begin{aligned} I(m,n,\alpha) &= \frac{\alpha^{\frac{m}{2}-1}(1+\alpha)^{-\frac{m-2+n}{2}}}{\frac{m-2+n}{2}} - \frac{\frac{m}{2}-1}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)} \alpha^{\frac{m-2}{2}-1}(1+\alpha)^{-\frac{m-4+n}{2}} \\ &\quad - \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\left(\frac{m-6+n}{2}\right)} \alpha^{\frac{m-4}{2}-1}(1+\alpha)^{-\frac{m-6+n}{2}} - \cdots - \\ &\quad \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\times\cdots\times\left(\frac{m-2k+4}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\times\cdots\times\left(\frac{m-2k+n}{2}\right)} \alpha^{\frac{m-2k+2}{2}-1}(1+\alpha)^{-\frac{m-2k+n}{2}} + \\ &\quad \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\times\cdots\times\left(\frac{m-2k+2}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\times\cdots\times\left(\frac{m-2k+n}{2}\right)} \times I(m-2k,n,\alpha). \end{aligned}$$

1. Si $\frac{m-2k}{2}-1=0$, entonces $m=2k+2$, $k=1,2,3,\ldots$,

$$\begin{aligned} I(m-2k,n,\alpha) &= I(2,n,\alpha)=\int_0^\alpha (1+t)^{-\frac{m-2k+n}{2}}dt=\int_0^\alpha (1+t)^{-\frac{n+2}{2}}dt \\ &= \left.\frac{(1+t)^{-\frac{n+2}{2}+1}}{-\frac{n+2}{2}+1}\right|_0^\alpha = \frac{1}{\frac{n}{2}}\left(1-(1+\alpha)^{-\frac{n}{2}}\right). \end{aligned}$$

Además

$$\begin{aligned} \frac{\Gamma\left(\frac{m+2}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} &= \frac{\Gamma\left(\frac{n}{2}+k+1\right)}{k!\,\Gamma\left(\frac{n}{2}\right)}=\frac{\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\frac{n}{2}\Gamma\left(\frac{n}{2}\right)}{k!\,\Gamma\left(\frac{n}{2}\right)} \\ &= \frac{\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\frac{n}{2}}{k!}, \end{aligned}$$

y para $j=1,2,\ldots,k$,

$$\begin{aligned} &\frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\times\cdots\times\left(\frac{m-2j+4}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\times\cdots\times\left(\frac{m-2j+2+n}{2}\right)\left(\frac{m-2j+n}{2}\right)} \\ &= \frac{k(k-1)\times\cdots\times(k-j+2)}{\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\left(\frac{n}{2}+k-j+2\right)\left(\frac{n}{2}+k-j+1\right)} \\ &= \frac{k!}{(k-j+3)!\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\left(\frac{n}{2}+k-j+2\right)\left(\frac{n}{2}+k-j+1\right)}. \end{aligned}$$

$$\begin{aligned} \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\times\cdots\times\left(\frac{m-2j+4}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\times\cdots\times\left(\frac{m-2j+2+n}{2}\right)\left(\frac{m-2j+n}{2}\right)} &= \frac{k(k-1)\times\cdots\times(k-j+2)}{\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\left(\frac{n}{2}+k-j+2\right)\left(\frac{n}{2}+k-j+1\right)} \\ &= \frac{k!}{(k-j+3)!\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\left(\frac{n}{2}+k-j+2\right)\left(\frac{n}{2}+k-j+1\right)}. \end{aligned}$$

Por lo tanto, si $m=2k+2$, $k=0,1,2,\ldots$, tomando en cuenta el desarrollo de $I(m,n,\alpha)$ y el cálculo de los coeficientes, obtenemos

$$\begin{aligned} F(m,n,\alpha) &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\times I(m,n,\alpha)=\frac{\left(\frac{n}{2}+k\right)\left(\frac{n}{2}+k-1\right)\times\cdots\times\frac{n}{2}}{k!}\times I(m,n,\alpha) \\ &= 1-(1+\alpha)^{-\frac{n}{2}}\left(1+ny+\frac{n(n+2)}{2!}y^2+\frac{n(n+2)(n+4)}{3!}y^3+\cdots+\right. \\ &\quad \left.\frac{n(n+2)\times\cdots\times(n+2(k-1))}{k!}y^k\right), \end{aligned}$$

donde $\alpha=\dfrac{mx}{n}$ y $y=\dfrac{\alpha}{2(1+\alpha)}$.

En conclusión, si $m=2k+2$, $k=0,1,2,\ldots$, $n=1,2,3,\ldots$, $\alpha=\frac{mx}{n}$, $x\geq 0$, $y=\frac{\alpha}{2(1+\alpha)}$, entonces

$$F(2k+2,n,\alpha)=1-(1+\alpha)^{-\frac{n}{2}}\left(1+ny\left(1+\tfrac{n+2}{2}y\left(1+\cdots+\tfrac{n+2(k-2)}{k-1}y\left(\tfrac{n+2(k-1)}{k}y\right)\cdots\right)\right)\right).$$

Ejemplos

1. Si $m=10$, $n=5$, $x=4{,}74$, se tiene

$$F(m,n,\alpha)=1-(1+\alpha)^{-\frac{n}{2}}\left(1+ny\left(1+\tfrac{n+2}{2}y\left(1+\tfrac{n+4}{3}y\left(1+\tfrac{n+6}{4}y\right)\right)\right)\right),$$

$\alpha=9{,}48$, $y=0{,}4522900763$, $F(10,\,5,\,9{,}48)=0{,}950104214$.

2. Si $n=32$ y $x=2{,}14$, obtenemos

$$F(10,\,32,\,0{,}66875)=0{,}9497430676.$$

2. Si $m = 2k+1$, $k = 1, 2, \ldots$, entonces

$$I(m-2k,n,x) = \int_0^\alpha t^{\frac{m-2k}{2}-1}(1+t)^{-\frac{m-2k+n}{2}}dt = \int_0^\alpha t^{-\frac{1}{2}}(1+t)^{-\frac{n+1}{2}}dt = \frac{2}{\sqrt{n}}\int_0^{\sqrt{n\alpha}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt.$$

En la sección precedente se describió un procedimiento de cálculo de la función de distribución t de Student:

$$T(n,x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\int_{-\infty}^x \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt \quad x \in \mathbb{R},\ n = 1, 2, \ldots.$$

Dicho procedimiento se centró en calcular $\int_0^x \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt \quad x \geq 0$. Esos resultados serán utilizados para calcular valores de $F(2k+1,n,x)$ con $k = 0, 1, 2, \ldots$, $n = 1, 2, \ldots$, $x \geq 0$. Para el efecto, definimos

$$F_1(n,\beta) = 2\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\int_0^\beta \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt \quad \beta \geq 0,\ n = 1, 2, \ldots.$$

Si $n = 2k+1$, $k = 0, 1, 2, \ldots$, $\beta \geq 0$, $z = \frac{n}{n+\beta^2}$, entonces

$$\begin{aligned} F_1(n,\beta) &= \frac{2}{\pi}\arctan\left(\frac{\beta}{\sqrt{n}}\right) + \frac{2\sqrt{n}}{\pi}\frac{\beta}{n+\beta^2}\left(1 + \frac{2}{1\times 3}z + \frac{2\times 4}{1\times 3\times 5}z^2 + \frac{2\times 4\times 6}{1\times 3\times 5\times 7}z^3 + \cdots + \right. \\ &\left. \frac{2\times 4\times\cdots\times 2(k-2)}{1\times 3\times 5\times\cdots\times(2k-3)}z^{k-2} + \frac{2\times 4\cdots\times 2(k-1)}{1\times 3\times\cdots\times(2k-1)}z^{k-1}\right). \end{aligned}$$

Si $n = 2k+2$, $k = 0, 1, 2, \ldots$,

$$\begin{aligned} F_1(n,\beta) &= \frac{\beta}{\sqrt{n+\beta^2}}\left(1 + \frac{1}{2}z + \frac{3}{2\times 1\times 2^2}z^2 + \frac{3\times 5}{3\times 2\times 1\times 2^3}z^3 + \cdots + \right. \\ &\left. \frac{1\times 3\times\cdots\times(2k-3)}{(k-1)!\,2^{k-1}}z^{k-1} + \frac{1\times 3\times\cdots\times(2k-1)}{k!\,2^2}z^k\right). \end{aligned}$$

Volvamos al cálculo de $F(2k+1,n,x)$. Comencemos con el análisis del término que contiene $I(m-2k,n,x)$. Tenemos

$$\begin{aligned} &\frac{\Gamma\left(\frac{m+n1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\times\cdots\times\left(\frac{m-2k+2}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\times\cdots\times\left(\frac{m-2k+n}{2}\right)}\frac{2}{\sqrt{n}}\int_0^{\sqrt{mx}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt \\ = &\frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\int_0^{\sqrt{mx}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}dt = F_1(n,\sqrt{mx}). \end{aligned}$$

Luego

$$F(2k+1,n,\alpha) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \times I(m,n,\alpha) = G(m,n,\alpha) + F_1(n,\sqrt{mx}),$$

donde

$$\begin{aligned} G(m,n,\alpha) &= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(-\frac{\alpha^{\frac{m}{2}-1}(1+\alpha)^{-\frac{m-2+n}{2}}}{\frac{m-2+n}{2}} - \frac{\left(\frac{m}{2}-1\right)\alpha^{\frac{m-2}{2}-1}(1+\alpha)^{-\frac{m-4+n}{2}}}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)}\right. \\ &- \frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\left(\frac{m-6+n}{2}\right)}\alpha^{\frac{m-4}{2}-1}(1+\alpha)^{-\frac{m-6+n}{2}} - \cdots - \\ &\left.\frac{\left(\frac{m}{2}-1\right)\left(\frac{m-2}{2}-1\right)\times\cdots\times\left(\frac{m-2k+4}{2}-1\right)}{\left(\frac{m-2+n}{2}\right)\left(\frac{m-4+n}{2}\right)\times\cdots\times\left(\frac{m-2k+n}{2}\right)}\alpha^{\frac{m-2k+2}{2}-1}(1+\alpha)^{-\frac{m-2k+n}{2}}\right). \end{aligned}$$

Teniendo presente que $m = 2k+1$, $k = 1, 2, \ldots$, $G(m, n, \alpha)$ se escribe en la forma siguiente:

$$\begin{aligned} G(m,n,\alpha) &= -\alpha^{\frac{1}{2}}(1+\alpha)^{-\frac{n+1}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{\alpha^{k-1}(1+\alpha)^{-k+1}}{\frac{n+2k-1}{2}} + \right. \\ &\quad \frac{\left(k-\frac{1}{2}\right)\alpha^{k-2}(1+\alpha)^{-k+2}}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)} + \frac{\left(k-\frac{1}{2}\right)\left(k\frac{3}{2}\right)(1+\alpha)^{-k+3}}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)\left(\frac{n+2k-5}{2}\right)} + \cdots + \\ &\quad \left. \frac{\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)\times\cdots\times\frac{3}{2}}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)\times\cdots\times\frac{n+1}{2}} \right). \end{aligned}$$

Sea $y = \frac{\alpha}{1+\alpha}$, la expresión anterior se escribe como

$$\begin{aligned} G(m,n,\alpha) &= -\sqrt{y}(1+\alpha)^{-\frac{n}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{\frac{n+2k-1}{2}} y^{k-1} + \right. \\ &\quad \frac{\left(k-\frac{1}{2}\right)}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)} y^{k-2} \frac{\left(k-\frac{1}{2}\right)\left(k\frac{3}{2}\right)}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)\left(\frac{n+2k-5}{2}\right)} y^{k-3} + \cdots + \\ &\quad \left. \frac{\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)\times\cdots\times\frac{3}{2}}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)\times\cdots\times\frac{n+1}{2}} \right). \end{aligned}$$

Para obtener una forma práctica de cálculo de $G(m, n, \alpha)$ debemos expresaar de modo conveniente todos los coeficientes. Para el efecto, obervemos que

$$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{1}{\frac{n+2k-1}{2}} = \frac{\Gamma\left(\frac{n+2k-1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} = \frac{\frac{n+2k-3}{2}\times\cdots\times\frac{n+1}{2}\Gamma\left(\frac{n+1}{2}\right)}{\left(k-\frac{1}{2}\right)\times\cdots\times\frac{3}{2}\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{n}{2}\right)},$$

$$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{k-\frac{1}{2}}{\left(\frac{m+2k-1}{2}\right)\left(\frac{m+2k-3}{2}\right)} = \frac{\Gamma\left(\frac{n+2k-3}{2}\right)}{\Gamma\left(\frac{2k-1}{2}\right)\Gamma\left(\frac{n}{2}\right)} = \frac{\frac{n+2k-5}{2}\times\cdots\times\frac{n+1}{2}}{\left(k-\frac{3}{2}\times\cdots\times\frac{3}{2}\Gamma\left(\frac{3}{2}\right)\right)} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)},$$

$$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)\left(\frac{n-2k-5}{2}\right)} = \frac{\Gamma\left(\frac{n+2k-5}{2}\right)}{\Gamma\left(\frac{2k-3}{2}\right)\Gamma\left(\frac{n}{2}\right)} = \frac{\frac{n+2k-7}{2}\times\cdots\times\frac{n+1}{2}}{\left(k-\frac{5}{2}\right)\times\cdots\times\frac{3}{2}\Gamma\left(\frac{3}{2}\right)} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)},$$

$\vdots$

$$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)\times\cdots\times\frac{3}{2}}{\left(\frac{n+2k-1}{2}\right)\left(\frac{n+2k-3}{2}\right)\times\cdots\times\frac{n+1}{2}} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{n}{2}\right)} = \frac{2}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Por lo tanto,

$$\begin{aligned} G(m,n,\alpha) &= -\frac{2}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sqrt{y}\left(1+\alpha^{-\frac{n}{2}}\right)\left(1+\frac{n+1}{y}\left(1+\frac{n+3}{5}y\,(1+\cdots+\right.\right. \\ &\quad \left.\left.\frac{n+2k-7}{2k-5}\left(1+\frac{n+2k-5}{2k-3}y\left(1+\frac{n+2k-3}{2k-1}y\right)\right)\cdots\right)\right)\right) \end{aligned}$$

que es una expresión muy fácil de programar.

Debemos notar que si $n = 2j$, $j = 1, 2, \ldots$, entonces

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{\left(j-\frac{1}{2}\right)\left(j-\frac{3}{2}\right)\times\cdots\times\frac{1}{2}\sqrt{\pi}}{(j-1)!} = \frac{\left(j-\frac{1}{2}\right)}{j-1}\frac{\left(j-\frac{3}{2}\right)}{j-2}\times\cdots\times\frac{1}{2}\sqrt{\pi},$$

y si $n = 2j+1$, $j = 0, 1, 2, \ldots$, entonces

$$\begin{aligned} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} &= \frac{\Gamma(j+1)}{\Gamma(j+\frac{1}{2})} = \frac{j!}{\left(j-\frac{1}{2}\right)\left(j-\frac{3}{2}\right)\times\cdots\times\frac{1}{2}\sqrt{\pi}} \\ &= \frac{j}{j-\frac{1}{2}}\times\frac{j-1}{j-\frac{3}{2}}\times\cdots\times\frac{1}{\frac{1}{2}}\times\frac{1}{\sqrt{\pi}}. \end{aligned}$$

Ejemplo

Si $m = 9$, $n = 15$, $x = 2{,}59$. Se tiene $k = 4$, $j = 7$,

$$\begin{aligned} F_1(n,\beta) &= \frac{1}{2}\arctan(\frac{\beta}{\sqrt{15}}) + \frac{2\sqrt{15}}{\pi}\frac{\beta}{15+\beta^2}\left(1+\frac{2z}{3}\left(1+\frac{4z}{5}\left(1+\frac{6z}{7}\right.\right.\right. \\ &\quad \left.\left.\left.\left(1+\frac{8z}{9}\left(1+\frac{10z}{11}\left(1+\frac{12z}{13}\right)\right)\right)\right)\right)\right), \end{aligned}$$

donde $\beta = \sqrt{mx}$, $z = \frac{n}{n+\beta^2}$. Entonces $\beta = 4{,}828043082$, $z = 0{,}3915426782$, $x = \frac{mx}{n} = 1{,}554$, $y = \frac{\alpha}{1+\alpha} = 0{,}608457322$, $F_1(n,\beta) = 0{,}9997786188$.

$$\begin{aligned} G(m,n,\alpha) &= -\frac{2}{\sqrt{\pi}}\sqrt{y}\frac{\Gamma(8)}{\Gamma\left(\frac{15}{2}\right)}(1+\alpha)^{-\frac{15}{2}}\left(1+\frac{16y}{3}\left(1+\frac{18y}{5}\left(1+\frac{20y}{7}\right)\right)\right) \\ &= -0{,}04961972164. \end{aligned}$$

$$F(9,15,2{,}59) = G(m,n,\alpha) + F_1(n,\beta) = 0{,}9501588972.$$

4.13. Ejercicios

1. Aplique la función gama de Euler para calcular las integrales siguientes

   a) $\int_0^\infty \frac{e^{-t}}{t^{1/4}}dt$. **b)** $\int_0^\infty t^{1/2}e^{-t^4}dt$. **c)** $\int_0^\infty \frac{e^{-t^2}}{t^{1/3}}dt$. **d)** $\int_0^\infty t^2e^{-t^3}dt$. **e)** $\int_0^\infty t\,e^{-\sqrt{t}}dt$.

   f) $\int_0^\infty t^3e^{-t^{1/3}}dt$. **g)** $\int_0^\infty x^me^{-ax^{1/n}}dx$, donde $a \in \mathbb{R}^+$, $m,n \in \mathbb{Z}^+$.

   h) $\int_0^\infty x^{1/m}e^{-ax^n}dx$, donde $a \in \mathbb{R}^+, m,n \in Z^+$.

2. Sea $p \in \mathbb{R}^+$. Demuestre que $\Gamma(p) = \int_0^1 \left(\ln\frac{1}{x}\right)^{p-1} dt$.

3. Utilice el resultado del ejercicio 2) para calcular las siguientes integrales.

   a) $\int_0^1 \frac{dx}{\left(\ln(\frac{1}{x})\right)^{1/3}}$. **b)** $\int_0^1 \frac{dx}{\left(\ln(\frac{1}{x})\right)^{1/4}}$. **c)** $\int_0^1 (\ln(x))^2\,dx$. **d)** $\int_0^1 (\ln(x))^6\,dx$.

   e) $\int_0^1 (\ln(x))^{2k}\,dx$, donde $k \in \mathbb{Z}^+$. **f)** $\int_0^1 \frac{dx}{\left(\ln(\frac{1}{x})\right)^{1/m}}$, donde $m \in \mathbb{Z}^+$.

4. Sean $\alpha, p \in \mathbb{R}^+$. Demuestre que $\Gamma(p) = \alpha^p \int_0^\infty t^{p-1}e^{-\alpha t}dt$.

5. Calcular las integrales siguientes

   a) $\int_0^1 x^{1/2}(\ln(x))^3dx$. **b)** $\int_0^1 x^{1/3}(\ln(x))^4dx$. **c)** $\int_0^1 x^2(\ln(x))^{1/5}dx$.

   d) $\int_0^1 x^p(\ln(x))^m dx$, donde $p \in \mathbb{R}^+$, $m \in \mathbb{Z}^+$.

6. Sea $p \in \mathbb{R}^+$. Demostrar que

$$\frac{d\Gamma(p)}{dp} = \int_0^\infty t^{p-1}e^{-t}(\ln(t))dt.$$

7. Calcular los términos de la sucesión $(\Gamma(-n+\frac{1}{2}))$, donde $n \in \mathbb{Z}^+$.

8. Sea $g : [1,2] \to \mathbb{R}$ la función definida por

$$g(p) = 2[(2-\sqrt{\pi})p(-3+p) + 4{,}5 - 2\sqrt{\pi}].$$

a) La función g es una interpolante de $\Gamma(p)$ con $p \in [1,2]$. Calcule $g(1)$, $g(1{,}5)$, $g(2)$ y compare con $\Gamma(1)$, $\Gamma\left(\frac{1}{2}\right)$ y $\Gamma(2)$.

b) Utilizando la función g bosqueje la gráfica de $\Gamma(p)$, $p \in [1,2]$.

c) Tomando en cuenta que $\Gamma(p) \underset{p\to 0}{\longrightarrow} \infty$, $\Gamma(p) \underset{p\to\infty}{\longrightarrow} \infty$; y de la información proporcionada en a) y b), bosqueje la gráfica de $\Gamma(p)$, $p \in \mathbb{R}^+$.

d) Bosqueje la gráfica de $\Gamma(p)$ para $p \in \mathbb{R}^- \setminus \mathbb{Z}^-$.

9. Aplique las propiedades de la función beta para calcular las integrales siguientes.

   a) $\int_0^{\frac{\pi}{2}} \operatorname{sen}^2(\theta)\cos^3(\theta)d\theta$. **b)** $\int_0^{\frac{\pi}{2}} \operatorname{sen}(\theta)\cos^5(\theta)d\theta$. **d)** $\int_0^{\frac{\pi}{2}} \operatorname{sen}^4(\theta)\cos(\theta)\,d\theta$.

   e) $\int_0^{\frac{\pi}{2}} \operatorname{sen}^5(\theta)\cos^4(\theta)\,d\theta$. **f)** $\int_0^{\frac{\pi}{2}} \operatorname{sen}^9(\theta)\cos(\theta)\,d\theta$. **g)** $\int_0^{\frac{\pi}{2}} \operatorname{sen}^{10}(\theta)\,d\theta$. **h)** $\int_0^{\frac{\pi}{2}} \cos^9(\theta)\,d\theta$.

10. Sean $p, q \in \mathbb{R}^+$. Utilizar la transformación $t = \dfrac{x}{1+x}$ para demostrar que

$$B(p,q) = \int_0^\infty \frac{x^{p-1}}{(1+x)^{p+q}}dx.$$

11. Sean $p, q \in \mathbb{R}^+$.

    a) Demostrar que el área de la región S limitada por la curva de ecuación $x^{\frac{2}{p}} + y^{\frac{2}{q}} = 1$, $x \geq 0$, $y \geq 0$ y los ejes coordenados viene dada por:

$$a(S) = \frac{pq}{2(p+q)} \frac{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}{\Gamma\left(\frac{p+q}{2}\right)} = \frac{pq}{2(p+q)} B\left(\frac{p}{2}, \frac{q}{2}\right).$$

    b) Calcule $a(S)$ para p, q en los casos siguientes: $p = q = 1$, $p = q = 2$, $p = q = 3$ (arco de asteroide).

12. Calcule las integrales siguientes en términos de la función beta y luego en términos de la función gama.

    a) $\int_0^1 x^7(1-x)^8dx$. **b)** $\int_0^1 x^{1/2}(1-x)^4dx$. **c)** $\int_0^1 x^3(1-x)^{1/2}dx$. **d)** $\int_0^a x^2(a-x)^5dx$, $a > 0$.

    e) $\int_0^a x^m(a-x)^n dx$, $a > 0$, $m, n \in \mathbb{Z}^+$. **f)** $\int_0^a x^m(a^n - x^n)^{\frac{1}{2}}dx$, donde $a > 0$, $m, n \in \mathbb{Z}^+$.

    g) $\int_0^2 \dfrac{x^{1/2}dx}{(4-x^2)^{1/5}}$.

13. Calcular las integrales siguientes en términos de la función gama.

    a) $\int_0^1 \sqrt{1-x^6}\,dx$. **b)** $\int_0^1 (1-x^4)^{1/5}dx$. **c)** $\int_0^1 (1-x^8)^{-\frac{1}{3}}dx$.

    d) $\int_0^1 (1-x^{2k})^{1/n}dx$, $k, n \in \mathbb{Z}^+$, $n \geq 2$. **e)** $\int_0^1 (1-x^m)^{-1/n}dx$, $m, n \in \mathbb{Z}^+$, $n > 1$.

14. En muchos casos se requieren valores de la distribución normal con una precisión $\varepsilon = 10^{-3}$. Establezca las modificaciones necesarias para generar un algoritmo que permita calcular valores de dicha función de distribución con $\varepsilon = 10^{-3}$.

15. Se requieren calcular valores de la distribución gama con una precisión $\varepsilon = 10^{-3}$. Establezca las modificaciones necesarias para generar un algoritmo que permita calcular valores de dicha función de distribución con $\varepsilon = 10^{-3}$. Calcule algunos de ellos y verifique sus resultados con los dados en los textos de Estadística y Probabilidades.

16. Se desea calcular valores de la distribución χ-cuadrada con una precisión $\varepsilon = 10^{-3}$. Establezca las modificaciones necesarias para generar un algoritmo que permita calcular valores de dicha función de distribución con $\varepsilon = 10^{-3}$. Calcule algunos de ellos y verifique sus resultados con los dados en los textos de Estadística y Probabilidades.

17. Establezca las modificaciones necesarias para generar un algoritmo que permita calcular valores de la función de distribución t de Student con $\varepsilon = 10^{-3}$. Calcule algunos de ellos y verifique sus resultados con los dados en los textos de Estadística y Probabilidades.

18. Establezca las modificaciones necesarias para generar un algoritmo que permita calcular valores de la función de distribución F de Snedekor con $\varepsilon = 10^{-3}$. Calcule algunos de ellos y verifique sus resultados con los dados en los textos de Estadística y Probabilidades.

4.14. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Análisis Matemático, Segunda Edición, Editorial Reverté, Barcelona, 1982.

2. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.

3. Tom M. Apostol, Calculus, Volumen 2, Segunda Edición, Editorial Reverté, Barcelona, 1975.

4. R. M. Barbolla, M. García, J. Margalef, E. Outerelo, J. L. Pinilla. J. M. Sánchez, Introducción al Análisis Real, Editorial Alambra Universidad, Madrid, 1981.

5. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.

6. Alan W. Bush, Perturbation Methods for Engineers and Scientists, CRC Press, Boca Raton, 1992.

7. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.

8. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.

9. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.

10. John E. Freund, Ronald E. Walpole, Estadística Matemática con Aplicaciones, Cuarta Edición, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1990.

11. Waltson Fulks, Cálculo Avanzado, Editorial Limusa, México, 1973.

12. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

13. Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Editorial Society for Industrial and Applied Mathematics, Philadelphia, 1996.

14. E. J. Hinch, Perturbation Methods, Cambridge University Press, Cambridge, 1991.

15. William W. Hines, Douglas C. Montgomery, Probabilidad y Estadística para Ingeniería y Administración, Compañía Editorial Continental, México, 1986.

16. Erwin Kreyszig, Introducción a la Estadística Matemática, Editorial Limusa, México, 1981.

17. L. Lebart, A. Morineau, J.-P. Fénelon, Tratamiento Estadístico de Datos, Editorial Marcombo Boixareu Editores, Barcelona, 1985.

18. Thomas M. Little, F. Jackson Hills, Métodos Estadísticos para la Investigación en la Agricultura, Editorial Trillas, México, 2002.

19. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

20. William Mendenhall, Dennis D. Wackerly, Richard L. Scheaffer, Estadística Matemática con Aplicaciones, Segunda Edición, Grupo Editorial Iberoamérica, México, 1994.

21. Paul L. Meyer, Probabilidad y Aplicaciones Estadísticas, Editorial Fondo Educativo Interamericano, México, 1973.

22. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

23. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

24. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

25. J. W. Schmidt, R. E, Taylor, Análisis y Simulación de Sistemas Industriales, Editorial Trillas, México, 1979.

26. Stephen P. Shao, Estadística para Economistas y Admistradores de Empresas, Editorial Herrero Hermanos, México, 1967.

27. Bhimsen K. Shivamoggi, Perturbation Methods for Differential Equations, Editorial Birkhäuser, Boston, 2003.

28. Fausto I. Toranzos, Estadística, Editorial Kapelusz, Buenos Aires, 1962.

Capítulo 5

Resolución Numérica de Ecuaciones no Lineales

Resumen

En este capítulo se tratan problemas primeramente de existencia de soluciones de ecuaciones no lineales en una sola variable. El punto de partida lo constituye el teorema de Bolzano con el que se genera el algoritmo de separación de las raíces y el método de bisección. A continuación se trata el teorema de Banach del punto fijo que asegura la existencia del punto fijo de aplicaciones contractivas definidas en intervalos cerrados y acotados de $\mathbb{R}$, lo que conduce a su vez a construir aplicaciones contractivas en intervalos cerrados y acotados donde están localizada (aislada) una sola raíz de la ecuación $f(x) = 0$. De este modo se generan algunos métodos iterativos que permiten calcular en forma aproximada la o las raíces de dicha ecuación. Entre los métodos más importantes citamos el de punto fijo, punto fijo modificado, Newton Raphson, Newton modificado, secantes, regula-falsi. Por otro lado, interesa conocer la rapidez con la que se aproxima la solución y comparar los diferentes métodos. Con esta información se plantean métodos de aceleración de la convergencia, básicamente se desarrollan dos: el método Δ^2 de Aitken y el método de Steffensen. Se consideran métodos para determinar las raíces de multiplicidad. Se concluye con el estudio de las ecuaciones algebraicas, es decir ecuaciones con funciones polinomiales o lo que es lo mismo el cálculo de las raíces de polinomios. Damos prioridad a los polinomios con coeficientes reales y nos centramos en el cálculo de las raíces reales; para el efecto, la primera tarea es localizar las raíces para en una segunda etapa proceder al cálculo de las mismas.

5.1. Introducción

En la actualidad se pone mucha atención el problema de la contaminación ambiental, particularmente del agua, pués en el futuro se debe proteger mucho más a este recurso. A continuación describimos brevemente un modelo matemático de control de la calidad del agua propuesto por Streeter y Phelps (1925) ampliamente utilizado (véase G. Kiely, volumen II, R. Banks)

Los microorganismos que requieren de oxígeno para su crecimiento se llama aeróbicos y aquellos que no lo requieren se llaman anaeróbicos. En el caso de los microorganismos aeróbicos, el oxígeno debe estar disponible en forma de oxígeno libre disuelto. Los microorganismos que pueden crecer en presencia de oxígeno se llaman aeróbicos obligados. Cuando un nutriente entra en una corriente de agua, los microorganismos aeróbicos consumen el oxígeno disuelto al efectuar la descomposición del nutriente, de este modo, el nutriente ejerce una demanda sobre la disponibilidad de oxígeno disuelto.

Los nutrientes disueltos causan contaminación cuando entran en una corriente de agua en cantidades suficientes para destruir la capacidad de autopurificación de esta; esto es, si los nutrientes disueltos entran al agua con una tasa tal que el oxígeno disuelto se gaste más rápidamente de lo que puede reponer, el agua se desoxigena. En estas condiciones, ningún aeróbico obligado (desde los microorganismos hasta los

peces) podrá sobrevivir y los contaminantes orgánicos se acumularán en el agua dando lugar a los procesos anaeróbicos que producirán sustancias malolientes de los contaminantes y el agua quedará contaminada.

Uno de los parámetros de calidad del agua y aguas residuales es la demanda bioquímica de oxígeno (DBO) que se define (Gerard Kiely, Vol.II, página 413) como la cantidad de oxígeno que necesitan los organismos vivientes (aeróbicos) en la fase de estabilización de la materia orgánica de las aguas y aguas residuales.

La DBO es una medida de su poder para causar contaminación y se produce cuando la demanda de oxígeno (DO) sobrepasa a la cantidad de oxígeno disponible.

La prueba de DBO estima el oxígeno gastado en la descomposición biológica de una muestra residual y es un simulación de laboratorio del proceso microbiano de autopurificación. Es importante el conocimiento preciso de la concentración de oxígeno disuelto en el agua para la prueba de DBO que es útil como indicador del estado de contaminación de una corriente de agua. La prueba DBO consiste en el proceso de laboratorio siguiente.

En una muestra de los residuos se diluye una mezcla con una población mixta adecuada de microorganismos. Se mide la concentración de oxígeno disuelto DO al instante $t = 0$. Esta mezcla se incuba a una temperatura fija ($T = 20^oC$) y luego de cierto tiempo ($t = 5$ días, $t = 15$ días, $t = 21$ días) se mide nuevamente la concentración de oxígeno disuelto $DO\,(t)$. El cambio $DO\,(0) - DO\,(t)$ mide la cantidad de oxígeno no utiizado en ese tiempo por los microorganismos al procesar nutrientes de la muestra de agua residual. Los más usuales son DBO_5 para $t = 5$ días, DBO_{15} para $t = 15$ días, DBO_{21} para $t = 21$ días. La primera prueba de este género fue propuesta en 1913.

El modelo más sencillo se establece en los términos siguientes: la taza de descomposición de materia orgánica es proporcional a la cantidad de materia orgánica disponible, esto es,

$$\frac{dL}{dt} = -k_1 L,$$

donde L es la demanda bioquímica de oxígeno remanente en $\frac{mg}{l}$, $k_1 > 0$ es el coeficiente de velocidad de desoxigenación de DBO en días^{-1}.

Al instante $t = 0$, la DBO inical del efluente en el punto de vertido a un curso de agua se le nota L_0. Se tiene

$$\left\{ \begin{array}{rcll} \frac{dL}{dt} & = & -k_1 L & t \in \,]0, T]\,, \\ L\,(0) & = & L_0, & \end{array} \right.$$

cuya solución es $L\,(t) = L_0 e^{-k_1 t} \quad t \geq 0$.

El modelo de Streeter y Phelps establece que

$$\frac{dDO}{dt} = k_1 L_1 - k_2 DO = k_1 L_0 e^{-k_1 t} - k_2 DO,$$

con DO el déficit de oxígeno disuelto, $k_2 > 0$ es la velocidad de reaireación atmosférica medida en día^{-1}.

La solución de la ecuación diferencial precedente es

$$DO\,(t) = \frac{k_1 L_0}{k_2 - k_1}\left(e^{-k_1 t} - e^{-k_2 t}\right) + d_0 e^{-k_2 t} \qquad t \geq 0,$$

donde d_0 es el déficit de oxígeno disuelto en $t = 0$. La función $DO\,(t) \quad t \geq 0$ representa el déficit de oxígeno disuelto saturado en cualquier instante.

Supóngase $d_0 = 0{,}7\ \frac{mg}{L}$, $k_1 = 0{,}25$ día^{-1}, $L_0 = 25\frac{mg}{L}$, $t = 4$ días y $DO\,(4) = 9{,}2\frac{mg}{L}$, se tiene

$$9{,}2 = 0{,}7e^{-4k_2} + \frac{0{,}25 \times 25}{k_2 - 0{,}25}\left(e^{-4\times 0{,}25} - e^{-4k_2}\right),$$

y de esta, se obtiene la siguiente ecuación:

$$9{,}2 = 0{,}7e^{-4k_2} + \frac{6{,}25}{k_2 - 0{,}25}\left(e^{-1} - e^{-4k_2}\right),$$

para la que no existe una fórmula que permita calcular k_2, consecuentemente se debe recurrir a métodos numéricos iterativos para calcular una solución aproximada, siempre que esta exista.

Esta clase de problemas son muy comunes en aplicaciones de la matemática.

Posición del problema

Sean $I \subset \mathbb{R}$ con $I \neq \emptyset$ un conjunto cerrado y f una función real de I en $\mathbb{R}$. Consideramos el problema siguiente

$$\text{hallar } \widehat{x} \in I, \text{ si existe, tal que } \ f(\widehat{x}) = 0.$$

Más precisamente, asignada la función f definida en I y en consecuencia la ecuación $f(x) = 0$, se trata de estudiar si dicha ecuación tiene o no solución en I, esto es, estudiar si existe al menos un $\widehat{x} \in I$ tal que $f(\widehat{x}) = 0$; y en el caso en que exista solución, interesa como calcular $\widehat{x}$ o como aproximar $\widehat{x}$, mediante una sucesión $(x_n) \subset I$ tal que

$$x_n \underset{n\to\infty}{\longrightarrow} \widehat{x} \ \text{ y } \ f(x_n) \underset{n\to\infty}{\longrightarrow} f(\widehat{x}) = 0.$$

Definición 1 *Asignada la ecuación $f(x) = 0$, un elemento $\widehat{x} \in I$ tal que $f(\widehat{x}) = 0$ se denomina cero de f o raíz de la ecuación $f(x) = 0$.*

Para un número limitado de funciones reales pueden darse métodos directos de resolución de la ecuación $f(x) = 0$. Así por ejemplo.

1. Sean $a, b \in \mathbb{R}$ con $a \neq 0$ y f la función real definida por $f(x) = ax + b \quad x \in \mathbb{R}$. Entonces

$$f(x) = 0 \Leftrightarrow ax + b = 0 \Leftrightarrow x = -\frac{b}{a},$$

   $\widehat{x} = -\frac{b}{a}$ es la raíz de $f(x) = 0$ o cero de f ya que $f\left(-\frac{b}{a}\right) = 0$.

2. Sean $a, b, c \in \mathbb{R}$ con $a \neq 0$ y $f(x) = ax^2 + bx + c \quad x \in \mathbb{R}$. La ecuación

$$f(x) = 0 \Leftrightarrow ax^2 + bx + c = 0,$$

   tiene solución en $\mathbb{R}$ si y solo si $d = b^2 - 4ac \geq 0$; en cuyo caso las raíces de la ecuación vienen dadas como:

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \quad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

   Si $d = b^2 - 4ac < 0$, la ecuación $ax^2 + bx + c = 0$ tiene dos raíces complejas, una conjugada de la otra.

3. Sea f la función real definida por $f(x) = -\frac{1}{2} + \operatorname{sen} x$. Entonces,

$$f(x) = 0 \Leftrightarrow \operatorname{sen} x = \frac{1}{2} \Leftrightarrow x \in \left\{\frac{\pi}{6} + 2k\pi \mid k \in \mathbb{Z}\right\} \cup \left\{\frac{5\pi}{6} + 2k\pi \mid k \in \mathbb{Z}\right\}.$$

4. Las siguientes son ecuaciones que igualmente se resuelven fácilmente en el conjunto $\mathbb{R}$: $2^x = \frac{1}{64}$, $\log_3(x) = 243$.

Para las ecuaciones como las que a continuación se indican, no es posible determinar un método directo que permita calcular las raíces exactas, únicamente es posible resolverlas de manera aproximada y es éste el objetivo de este capítulo.

1. $x - \cos(x) = 0 \quad x \in \mathbb{R}$.
2. $\arctan(x) = \frac{1}{1+x^2} \quad x \in \mathbb{R}$.

3. $x^4 + 5x^3 - x^2 + 1 = 0 \quad x \in \mathbb{R}$.

4. $4 - x^2 - e^{-3x} = 0 \quad x \in \mathbb{R}$.

5. Sean $n \in \mathbb{Z}^+$, $c = \frac{1}{n}\int_0^1 e^{-t^2}dt$. Ponemos $x_0 = 0$ y definimos

$$g_j(x) = c - \int_{x_{j-1}}^{x} e^{-t^2}dt \quad x \in [x_{j-1}, 1],\ j = 1, \ldots, n.\ \ g_j(x) = 0, \quad j = 1, \ldots, n.$$

Nota: Sea P un polinomio de grado 3, esto es $P(x) = a + bx + cx^2 + dx^3 \quad x \in \mathbb{R}$ con $a, b, c, d \in \mathbb{R}$. Mediante transformaciones adecuadas, la ecuación $P(x) = 0$ puede resolverse directamente mediante las denominadas fórmulas de Cardano. De manera similar, si $P(x) = a + bx + cx^2 + dx^3 + ex^4 \quad x \in \mathbb{R}$ es un polinomio de grado 4 con coeficientes en $\mathbb{R}$, mediante el método de Euler pueden calcularse directamente las raíces (reales o complejas) de la ecuación $P(x) = 0$ (véase H. Hall y Knight, Kurosh, Kostrikin).

Para polinomios de grado $n \geq 5$ no existen métodos directos de cálculo de las raíces de la ecuación $P(x) = 0$ con la excepción de casos muy particulares como por ejemplo los que se citan a continuación:

$$P(x) = x^5 - 32, \quad P(x) = x\left(x^2 + 1\right)\left(x^2 - 1\right), \quad P(x) = (x - 1)^5.$$

Las raíces reales de polinomios de grado 3 o 4 con coeficientes reales serán aproximadas mediante sucesiones.

5.2. Separación de las raíces.

En lo sucesivo supondremos que f es una función real definida en un subconjunto I de $\mathbb{R}$ y en consecuencia tendremos asignada la ecuación $f(x) = 0$ en el conjunto I.

La primera tarea para el estudio de la ecuación $f(x) = 0$ es la existencia de soluciones. Para el efecto consideramos dos procedimientos: el método gráfico y el algoritmo de búsqueda del cambio de signo.

Definición 2 *Una raíz $\widehat{x} \in I$ de la ecuación $f(x) = 0$ se dice separada en un intervalo $[a, b] \subset I$ si este intervalo contiene únicamente a la raíz $\widehat{x}$.*

i. Método gráfico

a) Si la función f puede ser graficada sin dificultad, la separación de las raíces se obtiene observando los intervalos en los cuales la gráfica de f corta al eje x.

Por lo general este procedimiento es limitado ya que la construcción de la gráfica conduce al estudio de la función f, estudio que puede resultar mas complicado que resolver la ecuación. En efecto, si f es derivable en I, para determinar los subconjuntos de I en los que f es creciente, decreciente, se deben resolver las inecuaciones $f'(x) > 0$, $f'(x) < 0$ y la ecuación $f'(x) = 0$ que pueden ser más complejas que la ecuación $f(x) = 0$. Si f'' existe en I, se deben determinar los subconjuntos de I en los que f es cóncava, convexa y determinar los puntos de inflexión de f, lo que conduce a calcular f'' y resolver las inecuaciones $f''(x) > 0$, $f''(x) < 0$ y la ecuación $f''(x) = 0$ que pueden resultar más difíciles que la ecuación $f(x) = 0$. Más adelante se exhiben ejemplos con estas características.

Por otro lado, las imprecisiones en el trazado de la gráfica pueden conducir a falsas interpretaciones.

b) Si la ecuación $f(x) = 0$ puede escribirse como $g(x) - h(x) = 0$, donde g, h son funciones definidas en el conjunto I cuyas gráficas pueden trazarse fácilmente, entonces la ecuación $f(x) = 0$ se transforma en determinar los puntos $x \in I$ tales que $g(x) = h(x)$. Las raíces de f se separan observando los intervalos en los cuales sus correspondientes gráficas se cortan. Este procedimiento es también limitado.

Ejemplos

1. Considerar la ecuación: $x \in \mathbb{R}$ tal que $x - \cos(x) = 0$. Se tiene $\cos(x) = x$. Ponemos $g(x) = \cos(x)$, $h(x) = x \quad x \in \mathbb{R}$. En la figura siguiente se muestran las gráficas de estas dos funciones. Se observa que dichas gráficas se cortan en un punto. La ecuación propuesta tiene una raíz $\widehat{x} \in \left[0, \frac{\pi}{2}\right]$.

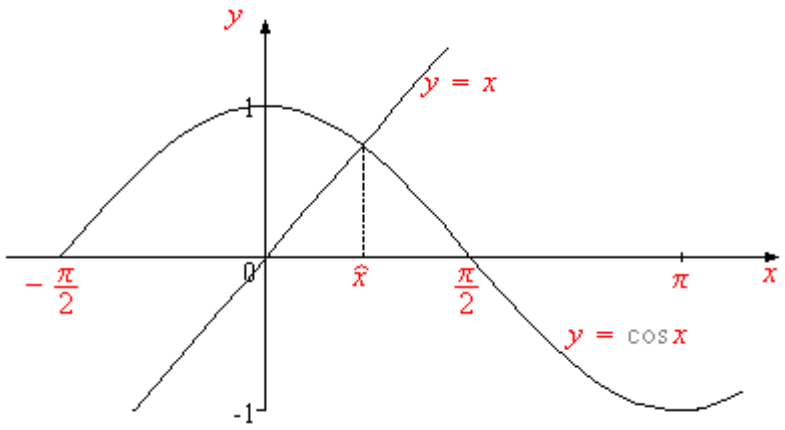


Figura 31

2. La ecuación $-x^2 + 1 - \tan(x) = 0$ puede escribirse en la forma $\tan(x) = -x^2 + 1$.
   Sean $g(x) = -x^2 + 1$, $h(x) = \tan(x) \quad x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. Las gráficas de g y h se cortan en un punto cuya abscisa $\widehat{x} \in [0, 1]$.

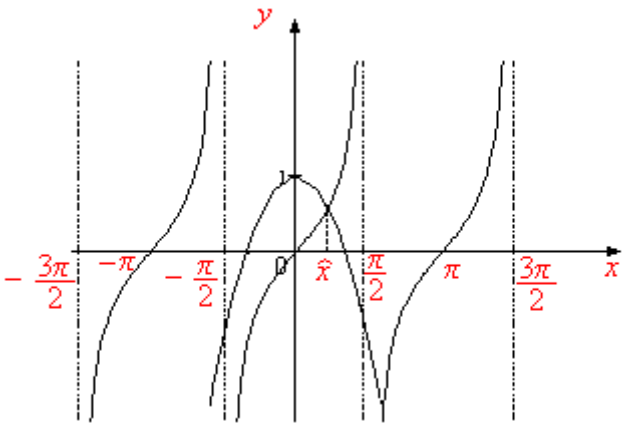


Figura 32

En el dibujo se observa que otra raíz está localizada en el intervalo $\left]\frac{\pi}{2}, \pi\right]$, ¿es justa esta aseveración? De ser así, ¿existen otras raíces para $x > \frac{\pi}{2}$? De acuerdo al gráfico no podemos dar respuesta inmediata. Requerimos de un análisis más fino para determinar, si existe o no, otras raíces de dicha ecuación.

3. Considerar la ecuación en $\mathbb{R}$ siguiente: $x^3 - 12x - 1 = 0$. Ponemos $f(x) = x^3 - 12x - 1 \quad x \in \mathbb{R}$. Entonces $f'(x) = 3x^2 - 12$. Luego,

$$\begin{aligned} f'(x) &> 0 \Leftrightarrow 3(x+12)(x-2) > 0 \Leftrightarrow x \in]-\infty, -2[\cup]2, \infty[, \\ f'(x) &\leq 0 \Leftrightarrow x \in [-2, 2]. \end{aligned}$$

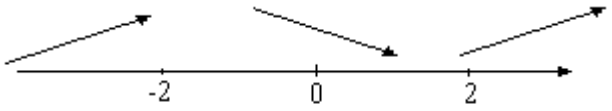


Figura 33

En $x=-2$ se tiene un máximo local y en $x=2$ se tiene un mínimo local. Además, $f''(x)=6x$. Se tiene

$$f''(x)>0 \Leftrightarrow x>0, \quad f''(x)<0 \Leftrightarrow x<0$$

En $x=0$ se tiene un punto de inflexión.

La gráfica de f siguiente muestra que tiene tres ceros $\widehat{x}_1, \widehat{x}_2, \widehat{x}_3$ localizados en los intervalos $[-4,-3]$, $[-1,0]$, $[3,4]$. Observe que la gráfica presenta imprecisiones.

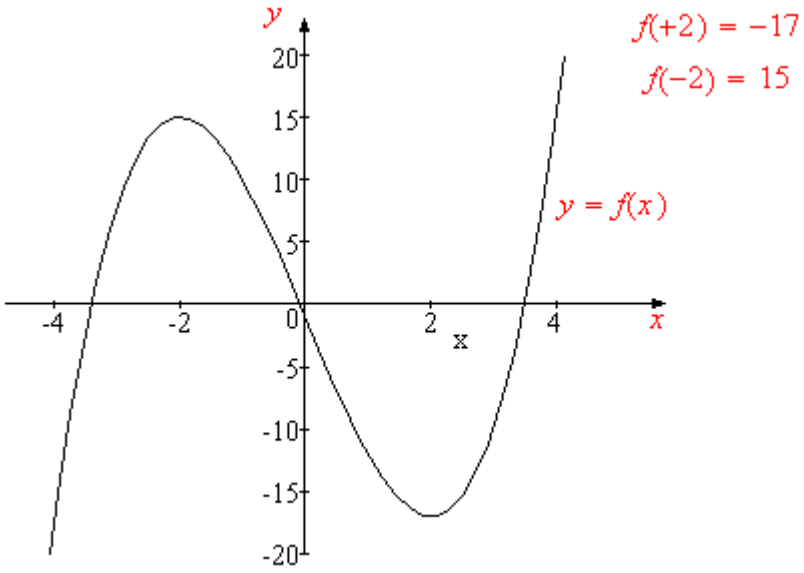


Figura 34

ii. Algoritmo de búsqueda del cambio de signo

Teorema 1 ***(de Bolzano)***
Sea f una función de $[a,b]$ en $\mathbb{R}$ continua en $[a,b]$. Si $f(a)f(b)<0$, existe $\widehat{x}\in[a,b]$ tal que $f(\widehat{x})=0$.

El teorema de Bolzano afirma que si la función continua f es tal que $f(a)$ y $f(b)$ tiene signos opuestos, la ecuación $f(x)=0$ tiene al menos una raíz $\widehat{x}\in[a,b]$. Además, este teorema garantiza la existencia de al menos una raíz de la ecuación $f(x)=0$.

En la práctica se tienen funciones continuas en las que $f(a)f(b)>0$ y sin embargo la ecuación $f(x)=0$ tiene solución en $[a,b]$ como lo prueba el siguiente ejemplo: $f(x)=x^2-1 \quad x\in[-2,2]$. Se tiene $f(-2)=f(2)=3$

En las gráficas que se muestran a continuación se presentan los dos casos.

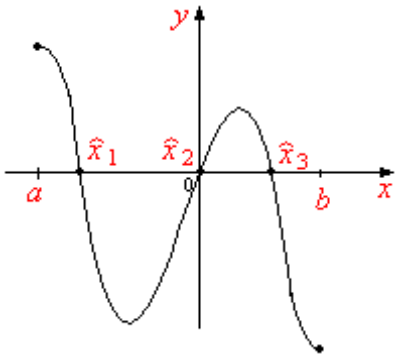


Figura 35

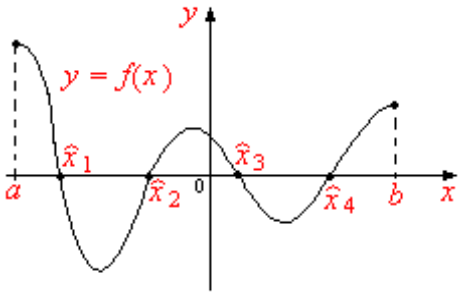


Figura 36

El algoritmo de búsqueda del cambio de signo se basa en el teorema de Bolzano y tiene dos propósitos: determinar la existencia de soluciones de la ecuación $f(x) = 0$ y separar las mismas. Describimos a continuación dicho algoritmo.

Sean $n \in \mathbb{Z}^+$ y $h = \frac{b-a}{n}$; h se denomina paso.

1. Calculamos $f(a)$.

   Si $f(a) = 0$ entonces a es una raíz de f y continuar en el punto 2).

2. Calculamos $f(a+h)$.

   Si $f(a) \neq 0$ y $f(a) f(a+h) = 0$ entonces $a+h$ es una raíz de f. Continuar en el punto 3).

   Si $f(a) f(a+h) < 0$ entonces existe $\widehat{x} \in]a, a+h[$ tal que $f(\widehat{x}) = 0$, es decir que f tiene un cero en el intervalo $]a, a+h[$. Continuar en 3).

   Si $f(a) f(a+h) > 0$ entonces f no tiene ceros en el intervalo $[a, a+h]$ o el paso h es demasiado grande. Continuar en 3).

3. Calculamos $f(a+2h)$. Entonces,

   Si $f(a+2h) = 0 \Rightarrow a+2h$ es una raíz de $f(x) = 0$.

   Si $f(a+h) \cdot f(a+2h) < 0 \Rightarrow \exists \widehat{x} \in]a+h, a+2h[$ tal que $f(\widehat{x}) = 0$.

   Si $f(a+h) \cdot f(a+2h) > 0$, no tiene raíces reales en $[a+h, a+2h]$, o h es demasiado grande.

Este procedimiento continua hasta llegar al extremo b del intervalo $[a, b]$.

Una vez localizada una raíz en un cierto subintervalo $[x_{j-1}, x_j]$ con $x_j = a + jh$, $j =, \ldots, n$, conviene asegurarse que no hay otras raíces. Para ello, se elige un entero $n_1 > n$ y se repite el procedimiento

anterior con el nuevo paso $h_1 = \frac{b-a}{n_1}$. Igualmente se repite el procedimiento en el caso en que no haya sido localizada ninguna raíz.

En la figura siguiente se ilustra esta situación: $f(x_3) * f(x_4) < 0$ con lo que existe al menos una raíz $\widehat{x}_1 \in [x_3, x_4]$. Situación similar se presenta en los otros intervalos.

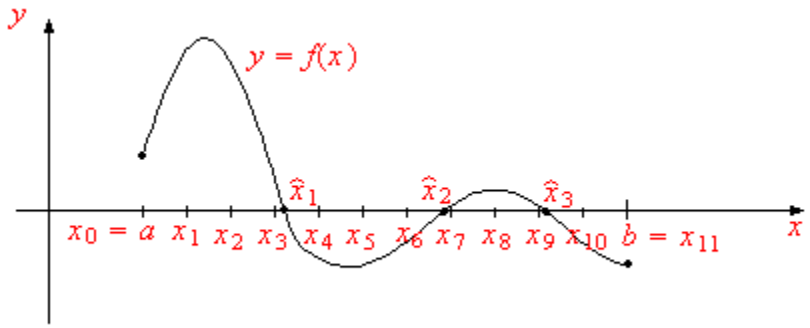


Figura 37

Observe en la gráfica que para el paso h seleccionado se han separado las tres raíces, cosa que no sucede para la gráfica de la función f siguiente.

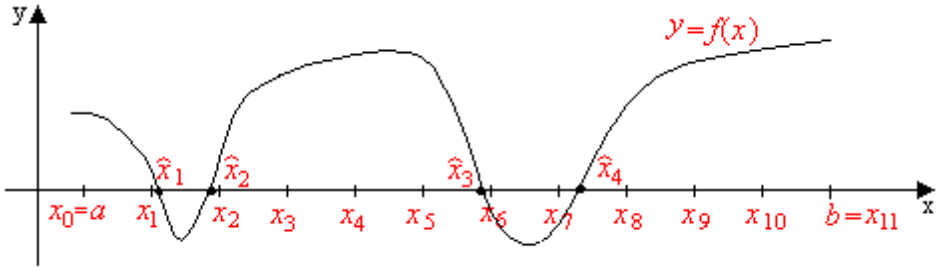


Figura 38

Al seleccionar un paso $h_1 = \frac{h}{2}$ se detectan raíces $\widehat{x}_1, \widehat{x}_2 \in [x_1, x_2]$.

Debemos notar que si f tiene una raíz de multiplicidad par, el cambio de signo no es detectado ya que si $\widehat{x}$ es una raíz de multiplicidad par y $\widehat{x} \in]x_j, x_{j+1}[$ se tiene $f(x_j) \times f(x_{j+1}) > 0$. Este tipo de problemas serán abordados en la sección 5**.**

De lo dicho precedentemente, se desprende el siguiente algoritmo de búsqueda del cambio de signo.

Algoritmo

Datos de entrada: a, b extremos del intervalo $[a, b]$, función f.

Datos de salida: xi, xd extremos del intervalos $[xi, xd]$, mensajes.

1. Leer n y hacer $h = \frac{b-a}{n}$.
2. $xi = a$.
3. $xi > b$, fin del procedimiento. Continuar en 8)**.**
4. $f(xi) = 0$; Mensaje: "xi es raíz de $f(x) = 0$".

   $xd = xi + h$.

5. $xd > b$; Mensaje: "f no tiene raíces reales en $[a, b]$ o f tiene raíces de multiplicidad par o h es demasiado grande".

6. $f(xd) = 0$; Mensaje: "xd es raíz de f".

   $xi = xd + h$. Continuar en 3).

7. $f(xi) \times f(xd) < 0$; Mensaje: "f tiene una raíz en $[xi, xd]$".

   $xi = xd$. Continuar en 5).

8. Fin.

Ejemplo

Sea f la función definida por $f(x) = x^3 - x - 1 \quad x \in \mathbb{R}$. Hallar el cambio de f en el intervalo $[-2, 2]$.

Sea $n = 10$. Entonces $h = \frac{b-a}{n} = 0{,}4$, $x_k = -2 + kh \quad k = 0, 1 \ldots, 10$. Escribimos $f(x) = -1 + x\left(-1 + x^2\right)$. En la tabla siguiente se muestran los resultados de la aplicación del algoritmo de búsqueda del cambio de signo.

k	xi	xd	$yi = f(xi)$	$yd = f(xd)$	Signo$(yi \times yd)$
0	-2	$-1{,}6$	-7	$-3{,}496$	$+$
1	$-1{,}6$	$-1{,}2$	$-3{,}496$	$-1{,}528$	$+$
2	$-1{,}2$	$-0{,}8$	$-1{,}528$	$-0{,}712$	$+$
3	$-0{,}8$	$-0{,}4$	$-0{,}712$	$-0{,}664$	$+$
4	$-0{,}4$	0	$-0{,}664$	$-1{,}0$	$+$
5	0	0,4	$-1{,}0$	$-1{,}336$	$+$
6	0,4	0,8	$-1{,}336$	$-1{,}288$	$+$
7	0,8	1,2	$-1{,}288$	$-0{,}472$	$+$
8	1,2	1,6	$-0{,}472$	1,492	$-$
9	1,6	2	1,492	5	$+$

El algoritmo de la búsqueda del cambio de signo muestra que f tiene una raíz real localizada en el intervalo [1,2, 1,6]. El estudio de la función f muestra que la ecuación $f(x) = 0$ tiene una única raíz localizada en el intervalo antes precisado.

5.3. Método de bisección

Sea f una función real, continua en $[a, b]$ y consideramos la ecuación $f(x) = 0$. Supongamos que el algoritmo de búsqueda del cambio de signo muestra que existe una raíz $\widehat{x} \in [\alpha, \beta]$. Más aún, suponemos que dicha raíz ha sido separada en dicho intervalo.

Entre los métodos más usados para el cálculo aproximado de $\widehat{x}$ es el conocido método de bisección. Su aplicación radica en dos hechos importantes: el algoritmo es siempre convergente y porque es fácilmente programable. No obstante, el método tiene la desventaja de requerir un número bastante grande de iteraciones para aproximar $\widehat{x}$ con una precisión ε fijada.

Describimos el método de bisección.

Sea $c_1 = \frac{\alpha+\beta}{2}$ el punto medio del intervalo $[\alpha, \beta]$. Ponemos $x_0 = \alpha$, $y_0 = \beta$.

Si $f(c_1) = 0$ entonces c_1 es una raíz de $f(x) = 0$.

Si $f(c_1) \neq 0$, consideramos los intervalos $[\alpha_1, c_1]$, $[c_1, \beta]$ y controlamos si $f(\alpha) f(c_1) < 0$ o $f(c_1) f(\beta) < 0$. En la figura siguiente se muestra la gráfica de una función f que tiene una raíz localizada en el intervalo

$[\alpha, \beta]$.

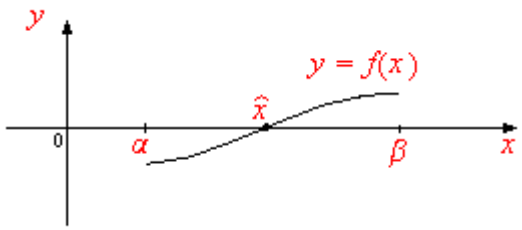


Figura 39

Supongamos que se verifica $f(c_1) f(\alpha) < 0$ (véase las gráficas de la función f y la posición de $\widehat{x} \in [\alpha, \beta]$ raíz de la ecuación $f(x) = 0$), lo que significa que la raíz $\widehat{x}$ pertence al intervalo $[\alpha, c_1]$. A este intervalo lo notamos $[x_1, y_1]$, donde $x_1 = \alpha$, $y_1 = c_1$.

Sea $c_2 = \frac{x_1+y_1}{2}$ el punto medio del intervalo $[x_1, y_1]$. Si $f(c_2) = 0$ entonces c_2 es una raíz de la ecuación $f(x) = 0$. Si $f(c_2) \neq 0$, nuevamente consideramos los intervalos $[x_1, c_2]$ y $[c_2, y_1]$ y controlamos el signo de $f(x_1) f(c_2)$. Con referencia de la posición $\widehat{x} \in [\alpha, \beta]$, se tiene $f(x_1) f(c_2) > 0$ lo que significa que $\widehat{x} \in [c_2, y_1]$. Notamos a este intervalo $[x_2, y_2]$ con $x_2 = c_2$, $y_2 = y_1$.

Sea $c_3 = \frac{x_2+y_2}{2}$ el punto medio del intervalo $[x_2, y_2]$. Calculamos $f(c_3) = 0$. En el caso contrario, controlamos el signo de $f(x_2) f(x_3)$. Observamos en la gráfica que $f(x_2) f(c_3) < 0$ que implica $\widehat{x} \in [x_2, c_3]$. Ponemos $x_3 = x_2$, $y_3 = c_3$.

En la gráfica que se muestra a continuación se visualizan los puntos del intervalo $[\alpha, \beta]$ que se obtienen mediante este procedimiento.

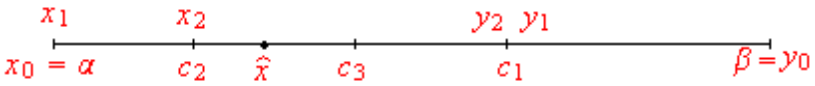


Figura 40

Este proceso repetimos n veces. Así, obtenemos el intervalo $[x_n, y_n]$, donde $f(x_n) f(y_n) < 0$. Como cada subintervalo $[x_n, y_n]$ de $[\alpha, \beta]$ $n = 0, 1, \ldots$ se divide en dos subintervalos de igual longitud, la longitud del intervalo $[x_n, y_n]$ es

$$y_n - x_n = \frac{\beta - \alpha}{2^n}.$$

Por otro lado los extremos izquierdos de los intervalos $[x_n, y_n]$ con $n = 1, 2, \ldots$, forman una sucesión monótona creciente, o sea $x_n \leq x_{n+1}$ y como $\alpha \leq x_n < \beta$, la sucesión (x_n) es acotada. Consecuentemente (x_n) es creciente y acotada, por lo tanto convergente. Sea $x = \lim_{n \longrightarrow \infty} x_n$.

Los extemos derechos de los intervalos $[x_n, y_n]$ forman una sucesión y_n decreciente: $y_{n+1} \leq y_n$ $n = 1, 2, \ldots$, y $\alpha < y_n \leq \beta$, con lo cual (y_n) es acotada. Así (y_n) es decreciente y acotada que implica (y_n) convergente. Sea $y = \lim_{n \longrightarrow \infty} y_n$. Como $\lim_{n \to \infty} \frac{\beta - \alpha}{2^n} = 0$, se sigue que

$$0 = \lim_{n \to \infty} \frac{\beta - \alpha}{2^n} = \lim_{n \to \infty} (y_n - x_n).$$

Luego

$$0 = \lim_{n \to \infty} (y_n - x_n) = \lim_{n \to \infty} y_n - \lim_{n \to \infty} x_n = y - x,$$

de donde $x = y$.

Por hipótesis f es continua en $[\alpha, \beta]$, entonces las sucesiones $(f(x_n))$ y $(f(y_n))$ son convergentes. Entonces

$$f(x) = f\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} f(x_n) = f(y).$$

Además,

$$\begin{aligned} x_n &< \widehat{x} < y_n \quad n = 1, 2, \ldots, \\ x &= \lim_{n\to\infty} x_n \leq \widehat{x} \leq \lim_{n\to\infty} y_n = y, \end{aligned}$$

y como $x = y$, entonces $\widehat{x} = x = y$ y $f(x) = f(\widehat{x}) = 0$. Así,

$$\begin{aligned} x_n &\underset{n\to\infty}{\longrightarrow} \widehat{x} \quad \text{y} \quad f(x_n) \underset{n\to\infty}{\longrightarrow} f(\widehat{x}) = 0, \\ y_n &\underset{n\to\infty}{\longrightarrow} \widehat{x} \quad \text{y} \quad f(y_n) \underset{n\to\infty}{\longrightarrow} f(\widehat{x}) = 0, \end{aligned}$$

que prueba que el método de bisección es convergente.

Teorema 2 *Sea f una función real, continua en $[a, b]$. Supongamos que $f(a) f(b) < 0$. Entonces, el método de bisección genera una sucesión (c_n) que converge a $\widehat{x}$ raíz de la ecuación $f(x) = 0$ y tal que*

$$|c_n - \widehat{x}| \leq \frac{b-a}{2^n} \quad n = 1, 2, \cdots.$$

Demostración. Para cada $n \in \mathbb{Z}^+$, tenemos $y_n - x_n = \dfrac{b-a}{2^n}$ y $\widehat{x} \in [x_n, y_n]$. Puesto que $c_n = \dfrac{x_n + y_n}{2}$ $n = 1, 2, \cdots$, se sigue que

$$0 \leq |c_n - \widehat{x}| \leq \frac{1}{2}(y_n - x_n) \leq \frac{b-a}{2^{n+1}} \quad n = 1, 2, \cdots.$$

Luego

$$0 \leq \lim_{n\to\infty} |c_n - \widehat{x}| \leq \lim_{n\to\infty} \frac{b-a}{2^n} = 0,$$

de donde $\lim\limits_{n\to\infty} c_n = \widehat{x}$, consecuentemente $0 = f(\widehat{x}) = \lim_{n\to\infty} f(c_n)$. Sea $\varepsilon > 0$. Del teorema precedente se tiene

$$|c_n - \widehat{x}| \leq \frac{b-a}{2^n} \quad n = 1, 2, \ldots,$$

y como $\lim\limits_{n\to\infty} \dfrac{b-a}{2^n} = 0$, existe $N_0 \in \mathbb{Z}^+$ tal que $\forall n \geq N_0 \Longrightarrow \dfrac{b-a}{2^n} < \varepsilon$. En particular, para $n = N_0$ se tiene $\dfrac{b-a}{2^{N_0}} < \varepsilon$. Luego

$$|c_{N_0} - \widehat{x}| \leq \frac{b-a}{2^{N_0}} < \varepsilon.$$

■

Para elaborar el algoritmo del método de bisección, queda determinar el número máximo de iteraciones $N_{\text{máx}}$. Para $\varepsilon = 10^{-t}$ con $t \in \mathbb{Z}^+$ (por ejemplo $10^{-4}, 10^{-6}, 10^{-8}, \cdots$), se tiene $\frac{b-a}{2^{N_0}} < 10^{-t}$. Como la función logaritmo natural es creciente, tomando logaritmos en ambos lados de esta desigualdad, resulta

$$\ln\left(\frac{b-a}{2^{N_0}}\right) < \ln\left(10^{-t}\right) \Longleftrightarrow \ln(b-a) - N_o \ln(2) < \ln\left(10^{-t}\right) \Longleftrightarrow N_0 > \frac{\ln\left[(b-a)\,10^t\right]}{\ln 2}.$$

El número máximo de iteraciones $N_{\text{máx}}$ elegimos como como sigue:

$$N_{\text{máx}} = \left[\frac{\ln\left[(b-a)\,10^t\right]}{\ln 2}\right] + 1,$$

donde $[\cdot]$ denota la función mayor entero menor o igual que. Así, $|C_{N_{\text{máx}}} - \widehat{x}| < 10^{-t}$.

Debe considerarse el hecho siguiente: puede verificarse que $|c_n - \widehat{x}| < \varepsilon$ pero $|f(c_n)| > \varepsilon$ para $n < N_{\text{máx}}$. En este caso el proceso debe continuar hasta lograr, en lo posible, $|f(c_n)| \leq \varepsilon$ para $n \leq N_{\text{máx}}$. Esto corresponde al denominado control vertical de la raíz.

Separada la raíz $\widehat{x}$ de la ecuación $f(x) = 0$ mediante el algoritmo de búsqueda del cambio de signo; esto es, dado $[a, b]$ intervalo en el que está localizada la única raíz $\widehat{x}$ de $f(x) = 0$ y dado $\varepsilon = 10^{-t}$ la precisión con la que $\widehat{x}$ será aproximada, el método de bisección se resume en el siguiente algoritmo.

Algoritmo

Datos de Entrada*:* a, b extremos del intervalo $[a, b]$, función f, $\varepsilon = 10^{-t}$.

Datos de Salida*:* $\widehat{x}$, n, $N_{\text{máx}}$.

1. Calcular $N_{\text{máx}} = \left[\dfrac{\ln\left[(b-a)\,10^t\right]}{\ln 2}\right] + 1.$

2. Poner $yi = f(a)$.

3. Para $n = 1, \ldots, N_{\text{máx}}$.

4. $c = \dfrac{a+b}{2}$.

5. $y = f(c)$.

6. Si $y = 0$, continuar en 10).

7. Si $\dfrac{b-a}{2} < \varepsilon$ y $|y| < \varepsilon$, continuar en 10).

8. Si $yi * y > 0$, entonces $a = c, yI = y$.

9. Si $yi * y < 0$, entonces $b = c$.

10. Imprimir $\widehat{x} = c$ raíz de $f(x) = 0$, iteración $n, N_{\text{máx}}$.

11. Fin.

Nota: Si $\widehat{x}$ ha sido separada utilizando el método de búsqueda del cambio de signo, $\widehat{x} \in [xi, xd] \subset [a, b]$. Previo al punto 1). del algoritmo de bisección de $x = xd$. Una vez calculado $\widehat{x}$, no finalizar el programa, se designa $xi = x$ y continua la ejecución del programa en la parte correspondiente a la búsqueda del cambio de signo en el resto del intervalo $[a, b]$. Note además que

$$N_{\text{máx}} = \left[\frac{\ln\left[(xd - xi)\,10^t\right]}{\ln 2}\right] + 1.$$

Ejemplos

1. Consideremos la ecuación: $x \in \mathbb{R}$ tal que $e^x - x^2 = 0$. Aproximemos la raíz de la misma con una precisión $\varepsilon = 10^{-2}$.

   Ponemos $g(x) = e^x$, $h(x) = x^2 \quad x \in \mathbb{R}$. En la figura siguiente se muestran las gráficas de las funciones g y h. Además, tales gráficas se cortan en el punto de abscisa $\widehat{x} \in [-1, 0]$ que muestra que

la ecuación dada tiene una solución separada en el intervalo $[-1, 0]$.

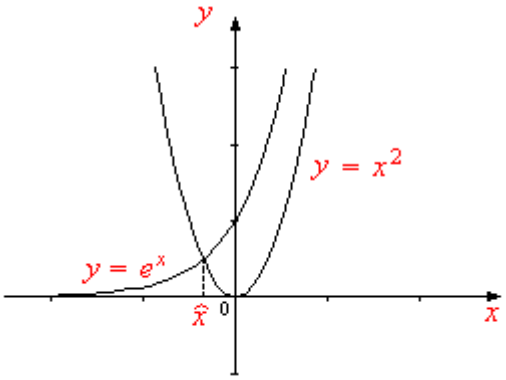


Figura 41

Para $\varepsilon = 10^{-2}$ y $xi = -1$, $xd = 0$, se tiene

$$N_{\text{máx}} = \left[\frac{\ln\left[(xd - xi) \times 10^2\right]}{\ln(2)}\right] + 1 = \left[\frac{\ln(10^2)}{\ln(2)}\right] + 1 = 6.$$

En la tabla siguiente se recogen los datos de la aplicación del algoritmo del método de bisección, donde f es la función real definida como $f(x) = e^x - x^2 \quad x \in \mathbb{R}$.

n	a	b	c	$f(a) = yi$	$f(c) = y$	signo$(yi * y)$
1	$-1.$	0.	$-0{,}5$	$-0{,}632$	0,357	$-$
2	$-1.$	$-0{,}5$	$-0{,}75$	$-0{,}632$	$-0{,}901$	$+$
3	$-0{,}75$	$-0{,}5$	$-0{,}625$	$-0{,}901$	0,145	$-$
4	$-0{,}75$	$-0{,}625$	$-0{,}6875$	$-0{,}901$	0,302	$-$
5	$-0{,}75$	$-0{,}6875$	$-0{,}71875$	$-0{,}901$	$-0{,}0292$	$+$
6	$-0{,}71875$	$-0{,}6575$	$-0{,}703125$	$-0{,}029$	$6{,}51 \times 10^{-4}$	$-$

La raíz aproximada de $\widehat{x}$ con una precisión $\varepsilon = 10^{-2}$ y $N_{\text{máx}} = 6$ es $c = -0{,}703125$. Note que para $n = 6$, se tiene $|f(c_6)| < \varepsilon$.

2. Aproximemos las raíces de la ecuación $2\ln(x + 4) - x^2 = 0$ con una precisión $\varepsilon = 10^{-3}$.

   Sean $f(x) = 2\ln(x + 4) - x^2$, $g(x) = 2\ln(x + 4)$ y $h(x) = x^2$. El ,método gráfico muestra que la ecuación $f(x) = 0$ tiene dos raíces separadas en los intervalos $[-2, -1]$ y $[1, 2]$. Sean $\widehat{x}_1 \in [-2, -1]$, $\widehat{x}_2 \in [1, 2]$ las raíces de la ecuación $f(x) = 0$.

   En la figura siguiente se muestran las gráficas de las funciones g y h así como los puntos de corte de las mismas cuyas abscisas son las raíces de la ecuación $f(x) = 0$.

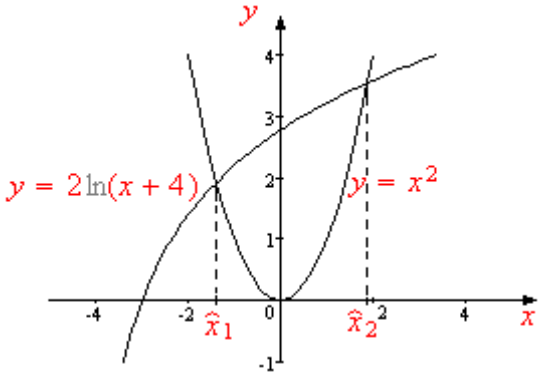


Figura 42

Aproximemos $\widehat{x}_1 \in [-2,-1]$ utilizando el método de bisección con una precisión $\varepsilon = 10^{-3}$. Se tiene $xi = -2$, $xd = -1$,

$$N_{\text{máx}} = \left[\frac{\ln\left((xd - xi)\,\varepsilon\right)}{\ln(2)}\right] + 1 = 9.$$

En la tabla se muestran los resultados de la aplicación del algoritmo del método de bisección.

n	a	b	$c = \frac{a+b}{2}$	$yi = f(a)$	$y = f(c)$	sign$(yi * y)$
1	$-2.$	$-1.$	$-1{,}5$	$-2{,}614$	$-0{,}417$	$+$
2	$-1{,}5$	$-1.$	$-1{,}25$	$-0{,}417$	$0{,}461$	$-$
3	$-1{,}5$	$-1{,}25$	$-1{,}375$	$-0{,}417$	$0{,}0395$	$-$
4	$-1{,}5$	$-1{,}375$	$-1{,}4375$	$-0{,}417$	$-0{,}184$	$+$
5	$-1{,}4375$	$-1{,}375$	$-1{,}40625$	$-0{,}184$	$-0{,}071$	$+$
6	$-1{,}40625$	$-1{,}375$	$-1{,}390625$	$-0{,}0713$	$-0{,}0156$	$+$
7	$-1{,}390625$	$-1{,}375$	$-1{,}3828125$	$-0{,}0156$	$0{,}0123$	$-$
8	$-1{,}390625$	$-1{,}3828125$	$-1{,}38671875$	$-0{,}0156$	$-0{,}001776$	$+$
9	$-1{,}38671875$	$-1{,}3828125$	$-1{,}384765625$	$-0{,}001776$	$0{,}00513$	$-$

Raíz $\widehat{x}_1 \simeq -1{,}385$, $f(-1{,}385) = 0{,}0043$. Note que $|f(c_9)| > \varepsilon$.

5.4. Desarrollo de métodos iterativos

Sean $E \subset \mathbb{R}$ con $E \neq \emptyset$ y f una función real definida en E. Consideramos el problema (P) siguiente:

$$\text{hallar } \widehat{x} \in E, \text{ si existe, tal que } f(\widehat{x}) = 0.$$

Suponemos que el problema (P) tiene al menos una solución $\widehat{x} \in E$ y que no existe un método directo de cálculo de $\widehat{x}$ por lo que debemos recurrir a los métodos aproximados. Estos métodos, en la generalidad, son iterativos y tienen la forma siguiente.

Se comienza con $x_0 \in E$. Las aproximaciones sucesivas x_n, $n = 1, 2, \ldots$ de $\widehat{x}$ se logra mediante una función iterativa Φ de $\widetilde{E}$ en $\widetilde{E}$ tal que $x_{n+1} = \Phi(x_n)$ $n = 0, 1, 2, \ldots$, donde $\widetilde{E} \subset E$, $\widetilde{E} \neq \emptyset$. De este modo se construye una sucesión $(x_n) \subset \widetilde{E}$. Algunas cuestiones surgen de esta construcción: ¿es (x_n) convergente a la raíz $\widehat{x}$ de $f(x) = 0$? ¿qué condiciones ha de verifiacar la función de iteración Φ para que (x_n) sea convergente a la raíz $\widehat{x}$? ¿cómo seleccionar el conjunto $\widetilde{E}$? ¿cómo construir una función de iteración Φ de modo que la sucesión (x_n) converja a $\widehat{x}$ raíz de $f(x) = 0$?

En la mayoría de casos la función de iteración Φ aparece por la propia formulación del problema.

Esta sección está destinada a la construcción de funciones de iteración Φ ligadas a métodos numéricos de resolución de ecuaciones no lineales muy conocidos en la literatura. Para ello introduciremos las denominadas funciones o aplicaciones contractivas y un teorema muy importante en análisis, a saber, el teorema de Banach del punto fijo.

Definición 3 *Sean $E \subset \mathbb{R}$, $E \neq \emptyset$ y T de E en E una función. Se dice que T es una aplicación contractiva en E si y solo si satisface la siguiente propiedad:*

$$\exists k, \quad 0 \leq k < 1 \text{ tal que } |T(x) - T(y)| \leq k\,|x - y| \qquad \forall x, y \in E.$$

La constante k de la definición precedente es independiente de x e y.

Como consecuencia inmediata de la definición se tiene que toda aplicación contractiva es uniformemente continua. El recíproco, en general, no es cierto.

Sea T una aplicación contractiva y $\varepsilon > 0$. De la definición se sigue que existe k, $0 \leq k < 1$ tal que $\forall x, y \in E$,

$$|T(x) - T(y)| \leq k\,|x - y| < \varepsilon.$$

Elegimos $\delta = \dfrac{\varepsilon}{k}\,(k \neq 0)$. Entonces

$$|x-y| < \delta \Rightarrow |T(x) - T(y)| < \varepsilon.$$

Observe que $\delta > 0$ es independiente de x e y. Además, si $k = 0$ se deduce que T es constante en E.

Por otro lado, si T es contractiva se tiene

$$|T(x) - T(y)| \leq k\,|x-y| \quad \forall x, y \in E.$$

ya que $0 \leq k < 1$, esto es, $|T(x) - T(y)| \leq k\,|x-y| \quad \forall x,\ y \in E$, pero puede suceder esto último sin ser contractiva se verá en un ejemplo propuesto más adelante.

Ejemplos

1. Sea $E = [1, 0]$ y $T : E \to E$ la función definida por $T(x) = \frac{1}{4}x^2$. Entonces T es contractiva. En efecto, sean $x,\ y \in E$, entonces

$$|T(x) - T(y)| = \left|\frac{1}{4}x^2 - \frac{1}{4}y^2\right| = \frac{1}{4}\,|(x+y)(x-y)|\,|x+y|\,|x-y|\,.$$

   Como $x, y \in [1, 0]$, $0 \leq x + y \leq 2$, luego

$$|T(x) - T(y)| = \frac{1}{4}\,|x+y|\,|x-y| \leq \frac{1}{2}\,|x-y|\,.$$

   La constante $k = \frac{1}{2}$, T es constractiva.

2. Sean $E = \mathbb{R}$, $a \in \mathbb{R}$ y T la función real definida por $T(x) = ax \quad x \in \mathbb{R}$. Entonces, para todo $x,\ y \in \mathbb{R}$ se tiene

$$|T(x) - T(y)| = |ax - ay| = |a|\,|x-y|\,.$$

   La función T será contractiva si y solo si $|a| < 1$.

3. Sean $E = [0, \infty[$ y G la función de E en E definida por $G(x) = \sqrt{1+x^2}$. La función G no es contractiva. Efectivamente, para todo $x,\ y \in [0, \infty[$ se tiene

$$|G(x) - G(y)| = \left|\sqrt{1+x^2} - \sqrt{1+y^2}\right| = \frac{\left|x^2 - y^2\right|}{\sqrt{1+x^2} + \sqrt{1+y^2}} = \frac{x+y}{\sqrt{1+x^2} + \sqrt{1+y^2}}\,|x-y|\,.$$

   Como $x \leq \sqrt{1+x^2}$, $y \leq \sqrt{1+y^2}$, se sigue que $\dfrac{x+y}{\sqrt{1+x^2} + \sqrt{1+y^2}} \leq 1$. Luego,

$$|G(x) - G(y)| \leq |x-y| \quad \forall x, y \in [0, \infty[\,.$$

   Si $y = 0$, se tiene

$$|G(x) - G(0)| = \frac{x}{1 + \sqrt{1+x^2}}\,|x-0| \quad \forall x, y \in [0, \infty[\,.$$

   Ponemos $k(x) = \dfrac{x}{1 + \sqrt{1+x^2}}$. Resulta $k(x) \underset{x \to 0^+}{\longrightarrow} 0$, $k(x) \underset{x \to +\infty}{\longrightarrow} 1$, luego $0 < k(x) < 1$ $\forall x \in [0, \infty[$, con lo cual $|G(x) - G(y)| = k(x)\,|x-0|$. De la última igualdad, observamos que no es posible encontrar una constante k con $0 < k < 1$ independiente de x. Este ejemplo pone de manifiesto que se verifica la desigualdad $|G(x) - G(y)| \leq |x-y| \quad \forall x, y \in [0, \infty[$, sin ser G contractiva.

Definición 4 *Sean $E \subset \mathbb{R}$, $E \neq \emptyset$ y T de E en E una función. Un punto $\widehat{x} \in E$ se dice un punto fijo de T si verifica la condición $T(\widehat{x}) = \widehat{x}$.*

Ejemplos

1. Sean $E = [1,2]$ y T de E en E la aplicación definida por $T(x) = \frac{1}{2}\left(x + \frac{2}{x}\right) \quad x \in [1,2]$. El punto $x = \sqrt{2}$ es un punto fijo de T. Pués

$$T\left(\sqrt{2}\right) = \frac{1}{2}\left(\sqrt{2} + \frac{2}{\sqrt{2}}\right) = \sqrt{2}.$$

Note que para todo $x \in [1,2]$, $T(x) \in [1,2]$.

2. Sea $f : [0,1] \to [0,1]$ una función continua en $[0,1]$. Existe $\widehat{x} \in [0,1]$ tal que $f(\widehat{x}) = \widehat{x}$. Efectivamente, sea g la función real definida por $g(x) = x - f(x) \quad x \in [0,1]$. Resulta que g es continua.

i. Si $x = 0$, $g(0) = -f(0)$.

Si $f(0) = 0$, resulta que $\widehat{x} = 0$ es un punto fijo de f.

Supongamos $f(0) \neq 0$. Como $f(0) \in [0,1]$, se tiene $g(0) = -f(0) < 0$.

ii. Si $x = 1$, $g(1) = 1 - f(1)$.

Si $f(1) = 1$, es $\widehat{x} = 1$ un punto fijo de f.

Supongamos $f(1) \neq 1$. Puesto que $f(1) \in [0,1]$ se tiene $g(1) > 0$. Luego, $g(0)\,g(1) < 0$ y por el teorema de Bolzano, existe $\widehat{x} \in [0,1]$ tal que $0 = g(\widehat{x}) = \widehat{x} - f(\widehat{x})$, de donde $f(\widehat{x}) = \widehat{x}$, esto es, $\widehat{x}$ es un punto fijo de f.

El resultado de este ejemplo tiene la siguiente interpretación geométrica. Consideremos las gráficas de las funciones f y de la identidad I en $[0,1]$. Un punto fijo de f es la abscisa del punto de intersección de la gráfica de f y de la función identidad I en el intervalo$[0,1]$. En los gráficos siguientes se muestran tres situaciones.

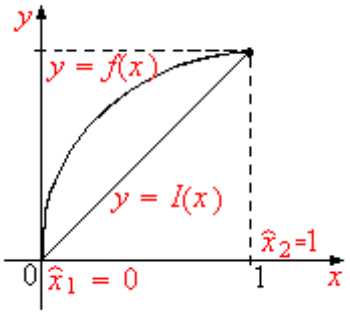


Figura 43

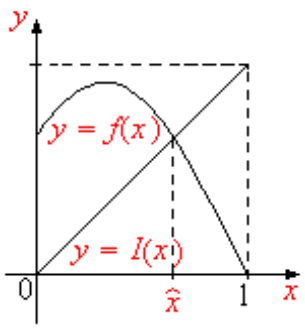


Figura 44

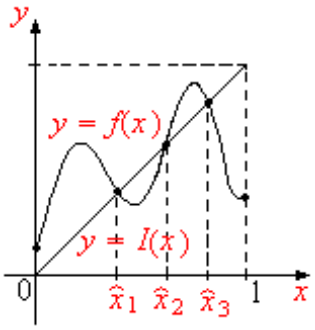


Figura 45

Teorema 3 ***(De Banach del punto fijo)***
Sean $E \subset \mathbb{R}$ con $R \neq \emptyset$ y E cerrado, T de E en E una aplicación contractiva en E. Entonces, existe un único $\widehat{x} \in E$ tal que $T(\widehat{x}) = \widehat{x}$.

Demostración. La demostración de este teorema la dividimos en dos partes. La primera que corresponde a la existencia del punto fijo $\widehat{x}$ de T y la segunda a la unicidad.

i) Existencia. Por hipótesis T es contractiva, entonces existe k, $0 \leq k < 1$ tal que

$$|T(x) - T(y)| \leq k\,|x - y| \qquad \forall x, y \in E. \tag{1}$$

Sea $x_0 \in E$. Definimos la sucesión $(x_n) \subset E$ como sigue

$$\begin{aligned} x_1 &= T(x_0) \\ x_2 &= T(x_1) \\ &\vdots \\ x_{n+1} &= T(x_n), \quad n = 0, 1, \ldots \end{aligned}$$

Mostremos que la sucesión (x_n) es una sucesión de Cauchy en E.

Sean $m, n \in \mathbb{Z}^+$ con $m > n$ y sea $p \in \mathbb{Z}^+$ tal que $= n + p$. Entonces, por la desigualdad triangular, se tiene

$$|x_n - x_m| = |x_n - x_{n+p}| \leq |x_n - x_{n+1}| + |x_{n+1} - x_{n-2}| + \cdots + |x_{n+p-1} - x_{n+p}|. \tag{2}$$

Por la definición de (x_n) y por (1) se tiene

$$\begin{aligned} |x_n - x_{n+1}| &= |T(x_{n-1}) - T(x_n)| \leq k\,|x_{n-1} - x_n| = k\,|T(x_{n-2}) - T(x_{n-1})| \\ &\leq k^2\,|x_{n-2} - x_{n-1}| \\ &\vdots \\ &\leq k^n\,|x_0 - x_1|. \end{aligned}$$

Luego,

$$|x_n - x_{n+1}| \leq k^n\,|x_0 - x_1| \qquad \forall n \in \mathbb{Z}^+. \tag{3}$$

Aplicando (3) en (2), resulta

$$|x_n - x_m| \leq k^n\,|x_0 - x_1| + k^{n+1}\,|x_0 - x_1| + \cdots + k^{n+p}\,|x_0 - x_1| = k^n\,|x_0 - x_1|\,(1 + k + \cdots + k^p)$$

$$\leq k^n\,|x_0 - x_1|\left(1 + k + \cdots + k^p + k^{p+1} + \cdots\right). \tag{4}$$

Sea $S_p(k) = 1 + k + \cdots + k^p$. Entonces

$$(1 - k)\,S_p(k) = S_p(k) - kS_p(k) = 1 + k + \cdots k^p - \left(k + k^2 + \cdots + k^{p+1}\right) = 1 - k^{p+1}.$$

de donde

$$S_p(k) = \frac{1-k^{p+1}}{1-k} = \frac{1}{1-k} - \frac{k^{p+1}}{1-k}.$$

Como $0 \leq k < 1$, $\lim\limits_{p\to\infty} k^{p+1} = 0$. Luego

$$\lim_{p\to\infty} S_p(k) = \lim_{p\to\infty}\left(\frac{1}{1-k} - \frac{k^{p+1}}{1-k}\right) = \frac{1}{1-k} - \lim_{p\to\infty}\frac{\kappa^{p-1}}{1-k} = \frac{1}{1-k},$$

con lo cual

$$\sum_{p=0}^{\infty} k^p = \lim_{p\to\infty} S_p(k) = \frac{1}{1-k}. \tag{5}$$

Remplazando (5) en (4), se obtiene

$$|x_n - x_m| \leq k^n |x_0 - x_1| \sum_{p=0}^{\infty} k^p = \frac{k^n}{1-k} |x_0 - x_1|. \tag{6}$$

Puesto que $\lim\limits_{n\to\infty} k^n = 0$ se sigue que $\forall \varepsilon > 0$, $\exists n_0 \in \mathbb{Z}^+$ tal que $\forall n \geq n_0$

$$k^n < \frac{1-k}{|x_0 - x_1|}\varepsilon \tag{7}$$

De (6) y (7) resulta

$$|x_n - x_m| \leq \frac{k^n}{1-k} |x_0 - x_1| < \varepsilon \quad \text{si } m, n \geq n_0,$$

es decir que (x_n) es una sucesión de Cauchy en E y por hipótesis E es cerrado, entonces la sucesión (x_n) tiene límite en E; esto es, existe $\widehat{x} \in E$ tal que $\lim\limits_{n\to\infty} x_n = \widehat{x}$.

Puesto que T es contractiva, T es uniformemente continua y por lo tanto continua. Luego

$$\lim_{n\to\infty} T(x_n) = T\left(\lim_{n\to\infty} x_n\right) = T(\widehat{x}).$$

Además, $x_{n+1} = T(x_n)$ y $\lim_{n\to\infty} x_{n+1} = \widehat{x}$, resulta que

$$T(\widehat{x}) = \lim_{n\to\infty} T(x_n) = \lim_{n\to\infty} x_{n+1} = \widehat{x}.$$

Así, $T(\widehat{x}) = \widehat{x}$ o sea $\widehat{x} \in E$ es un punto fijo de T.

ii) Unicidad Probemos que $\widehat{x} \in E$ tal que $T(\widehat{x}) = \widehat{x}$ es único. Para el efecto, supongamos que existe $y \in E$ tal que $T(y) = y$. Mostremos que $y = \widehat{x}$.

Como T es contractiva, se tiene

$$|\widehat{x} - y| = |T(\widehat{x}) - T(y)| \leq k |\widehat{x} - y|,$$

de donde $|\widehat{x} - y|(1-k) \leq 0$, y siendo $0 \leq k < 1$, entonces $1 - k > 0$ y en consecuencia $|\widehat{x} - y| \leq 0$. Como el valor absoluto es no negativo, la única posibilidad es $|\widehat{x} - y| = 0 \Leftrightarrow y = \widehat{x}$. ■

Observaciones

1. El teorema de Banach del punto fijo asegura la existencia de un único punto fijo $\widehat{x} \in E$ e la plicación contractiva T definida en el conjunto cerrado E de $\mathbb{R}$.

2. Note que la métrica usual d de $\mathbb{R}$ está definida por $d(x, y) = | x - y |$ $\forall x, y \in \mathbb{R}$, admás $(\mathbb{R}, d)$ es un espacio métrico completo, esto es, toda sucesión de Cauchy en $\mathbb{R}$ es convergente en $\mathbb{R}$. Como $E \subset \mathbb{R}$, $E \neq \emptyset$, el par (E, d) es un espacio métrico y siendo E cerrado, se prueba que toda sucesión de Cauchy en E es convergente en E, con lo cual (E, d) es un espacio métrico completo.

El conjunto $E =]0,1] \subset \mathbb{R}$ no es cerrado. La sucesión $(x_n) \subset E$ con $x_n = \frac{1}{n}$ $n = 1,2,\ldots$, es una sucesión de Cauchy en E que no es convergente en E, pues $\lim_{n\to\infty} x_n = 0 \notin E$. Luego (E,d) es un espacio métrico que no es completo.

En los textos de Análisis, el teorema de Banach del punto fijo se enuncia como sigue: Sea (E,d) un espacio métrico completo y T de E en E una aplicación contractiva. Entonces, existe un único $\widehat{x} \in E$ tal que $T(\widehat{x}) = \widehat{x}$.

El enunciado y la prueba del teorema de Banach del punto fijo que hemos dado está particularizado a subconjuntos cerrados E de $\mathbb{R}$ $(E \neq \emptyset)$ provistos de la métrica usual de $\mathbb{R}$.

La demostración del teorema de Banach del punto fijo para espacios métricos completos muy generales (E,d) es muy similar a la aquí propuesta con la salvedad que los valores absolutos $|x-y| = d(x,y)$ $x,y \in \mathbb{R}$ se remplazan simplemente por $d(x,y)$ con d la métrica en el conjunto E.

3. En la demostración del teorema de Banach del punto fijo se muestra una manera de calcular el punto fijol $\widehat{x} \in E$. Pues se parte de un punto arbitrario $x_0 \in E$ y se construye la sucesión $(x_n) \subset E$ tal que $x_{n+1} = T(x_n)$ $n = 0,1,\ldots$. Entonces $\widehat{x} = \lim_{n\to\infty} x_n$ es el punto fijo de T. De este hecho se desprende que podemos aproximar el funto fijo $\widehat{x}$ con una precisión $\varepsilon > 0$.

4. La construcción de la sucesión (x_n) tiene la siguiente interpretación geométrica. Sea $E \subset \mathbb{R}$, $E \neq \emptyset$ y E cerrado, T de E en E una aplicación contractiva en E e I la aplicación identidad en E. Por el teorema de Banach del punto fijo, las gráficas de las funciones T e I se cortan en el punto $(\widehat{x}, T(\widehat{x})) = (\widehat{x}, I(\widehat{x}))$. La abscisa $\widehat{x} \in E$ es el punto fijo de T.

   En las dos figuras que se muestran a continuación se exhiben los puntos x_n de la sucesión (x_n) con $x_{n+1} = T(x_n)$, $n = 0,1,\ldots$. Se parte de $x_0 \in E$. Se calcula $x_1 = T(x_0)$, se proyecta $T(x_0)$ sobre la recta $y = x$, como $x_1 = I(x_1)$ se obtiene en el eje X el punto x_1. Nuevamente se calcula $x_2 = T(x_1)$ y se proyecta $T(x_1)$ sobre la recta $y = x$, se obtiene en el eje x el punto $x_2 = I(x_2)$. El proceso continua.

   En la gráfica que se muestra a continuación se tiene una sucesión creciente que converge a $\widehat{x}$

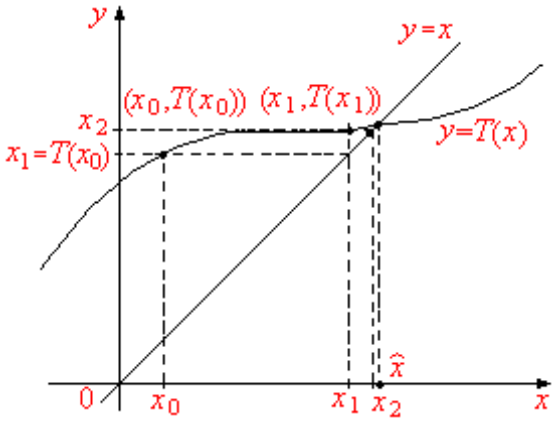


Figura 46

En la gráfica siguiente se muestran puntos de una sucesión que oscila entorno al punto fijo $\widehat{x}$.

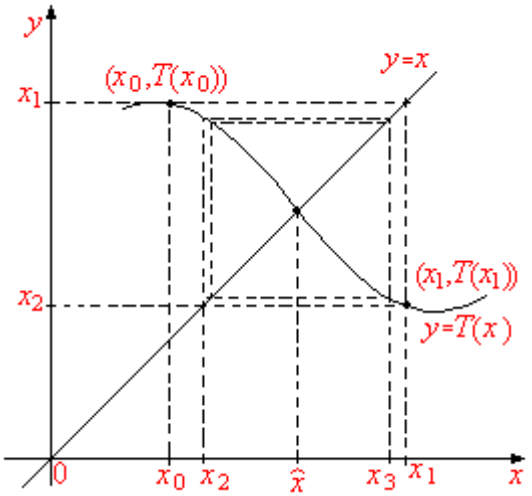


Figura 47

En la gráfica siguiente se muestra una función T que no es contractiva en E y una sucesión (x_n) que, en general, no es convergente.

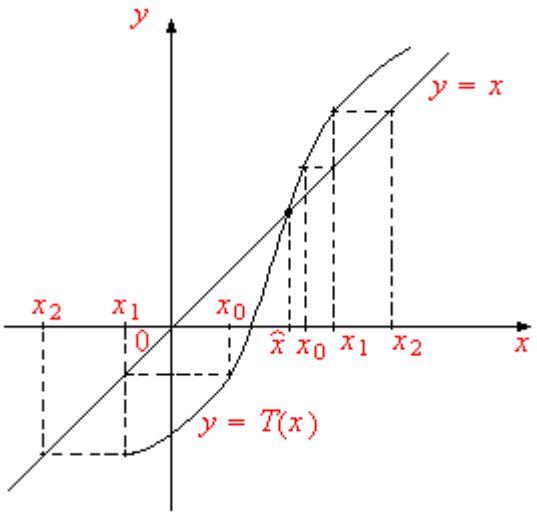


Figura 48

Ejemplos

1. Sean $E = [1,2]$ y T de E en E la aplicación definida por $T(x) = \frac{1}{2}\left(x + \frac{2}{x}\right) \quad x \in [1,2]$. Entonces, T es contractiva en E. Además $\widehat{x} = \sqrt{2}$ es el punto fijo de T. Aproximamos $\widehat{x} = \sqrt{2}$ mediante x_n, $n = 0, 1, \ldots$, con $x_{n+1} = T(x_n)$.

   Sea $x_0 = 2 \in E$. Entonces

$$\begin{aligned} T(x_0) &= x_1 = \frac{1}{2}\left(x_0 + \frac{2}{x_0}\right) = \frac{1}{2}\left(2 + \frac{2}{2}\right) = 1{,}5, \\ T(x_1) &= x_2 = \frac{1}{2}\left(x_1 + \frac{2}{x_1}\right) = \frac{1}{2}\left(1{,}5 + \frac{2}{1{,}5}\right) = 1{,}4166667, \\ T(x_2) &= x_3 = \frac{1}{2}\left(x_2 + \frac{2}{x_2}\right) = 1{,}414215686, \\ T(x_3) &= x_4 = \frac{1}{2}\left(x_3 + \frac{2}{x_3}\right) = 1{,}414213562, \\ &\vdots \end{aligned}$$

Observamos que 1,414213 ... es una aproximación de $\sqrt{2}$.

2. Sea $E = [0,1]$. La aplicación T de $[0,1]$ en $[0,1]$ definida por $T(x) = \cos(x)$ es contractiva en $[0,1]$. Por tanto, existe $\widehat{x} \in [0,1]$ tal que $\widehat{x} = T(\widehat{x}) = \cos(\widehat{x})$, de donde $\widehat{x} - \cos(\widehat{x}) = 0$. El punto $\widehat{x} \in [0,1]$ solución de la ecuación $x - \cos(x) = 0$.

3. Sea $E = [0,\infty[$ y $T(x) = (x+1)^{\frac{1}{3}} \quad x \in [0,\infty[$. Resulta que T es contractiva en $[0,\infty[$. Luego, existe $\widehat{x} \in [0,\infty[$ tal que $\widehat{x} = T(\widehat{x}) = (\widehat{x}+1)^{\frac{1}{3}}$, de donde $\widehat{x}^3 - \widehat{x} - 1 = 0$. El punto fijo $\widehat{x}$ es solución de la ecuación $x^3 - x - 1 = 0$.

   Sea $f(x) = x^3 - x - 1 \quad x \in [0,\infty[$. La gráfica de f muestra que la ecuación $f(x) = 0$ tiene una sola raíz $\widehat{x} \in [0,\infty[$.Más aún dicha raíz está separada en $[1,2]$. Determinemos el valor aproximado de $\widehat{x}$. En la tabla siguiente se muestran algunos valores de la sucesión (x_n) con $x_0 = 1$ y $x_{n+1} = T(x_n)$, $n = 0, 1, \ldots$

x	1	1,2599	1,3123	1,3224	1,3243	1,3246	1,3247
$T(x)$	1,2599	1,3123	1,3224	1,3243	1,3246	1,3247	1,3247

   Con $n = 7$ iteraciones se tiene $|x_n - \widehat{x}| < 5 \times 10^{-4}$. Luego $x_7 \simeq 1{,}3242$ es una aproximación de la raíz $\widehat{x}$ de $f(x) = 0$, $f(x_7) \simeq 0$.

5.4.1. Método de punto fijo

Sean $I \subset \mathbb{R}$, con $I \neq \emptyset$ y f una función real definida en I. Consideramos la ecuación $f(x) = 0$. Supongamos que este problema tiene almenos una solución $\widehat{x} \in I$, y que esta ha sido separada en el intervalo $[a,b] \subset I$.

Si $f(x) = x - T(x) \quad x \in [a,b]$. Entonces $f(x) = 0 \Leftrightarrow T(x) = x$. Además, si T es contractiva en $[a,b]$, existe $\widehat{x} \in [a,b]$ tal que $T(\widehat{x}) = \widehat{x}$ o bien $f(\widehat{x}) = \widehat{x} - T(\widehat{x}) = 0$, es decir que $\widehat{x}$ es la raíz de la ecuación $f(x) = 0$. La dificultad radica en la selección de la función T y del intervalo $[a,b]$ en el que está localizada la raíz $\widehat{x}$ de $f(x) = 0$ de modo que la imagen directa $T([a,b]) = [a,b]$ y T contractiva en $[a,b]$.

Supuesto que la función T de $[a,b]$ en $[a,b]$ es contractiva. El método de iteración de punto fijo se basa en la construcción de la sucesión (x_n) siguiente: elegimos $x_0 \in [a,b]$ y definimos $x_{n+1} = T(x_n) \quad n = 0, 1, \ldots$. El algoritmo del método de punto fijo es el siguiente:

Algoritmo

Datos de entrada: $N_{\text{máx}}$ número máximo de iteraciones, $\varepsilon > 0$ la precisión, T función de $[a,b]$ en $[a,b]$.

Datos de salida: iteración n, valor aproximado x_n de $\widehat{x}$.

1. Leer $x_0 \in [a,b]$ y poner $x = x_0$.

2. Para $n = 0, \ldots, N_{\text{máx}}$,

3. Calcular $y = T(x)$.

4. Si $|x-y|<\varepsilon$. Continuar en 6).

5. $x=y$.

6. Si $n<N_{\text{máx}}$, imprimir "raíz aproximada y de $\widehat{x}$ en n iteraciones con una precisión $\varepsilon>0$". Continuar en 8).

7. Imprimir "raíz aproximada y de $\widehat{x}$ en $N_{\text{máx}}$ iteraciones con precisión $\varepsilon>0$".

8. Fin.

Ejemplos

1. Encontrar todos los valores de $x\in\mathbb{R}$ tales que $\sin(x)=x^2-1$.

   Definimos $f(x)=x^2-1-\sin(x)$ y sean $\varphi_1(x)=\sin(x)$, $\varphi_2(x)=x^2-1$ $x\in\mathbb{R}$. En la figura siguiente se muestran las gráficas de las funciones φ_1 y φ_2 que se cortan en dos puntos. El método gráfico muestra que f tiene dos raíces localizadas en los intervalos $[-1,0]$ y $[1,2]$.

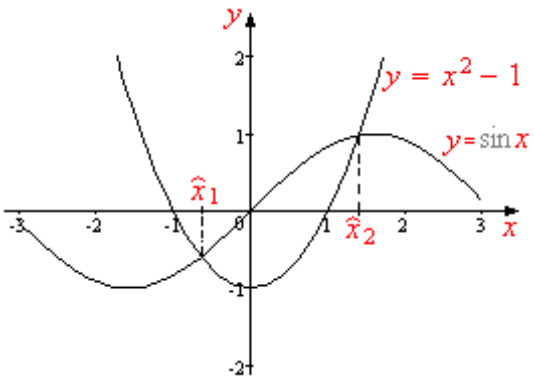


Figura 49

   De la igualdad $\sin(x)=x^2-1$ se sigue que $x^2=1+\sin(x)$. Luego $x=\pm\sqrt{1+\sin(x)}$. Definimos $T(x)=\sqrt{1+\sin(x)}$ $x\in[1,2]$. La aplicación T es contractiva (véase el teorema que se enuncia a continuación), y para $x_0\in[1,2]$ dado, se define $x_{n+1}=T(x_n)$ $n=0,1,\ldots$.En las dos tablas siguientes se muestran los resultados de la aplicación del algoritmo de punto fijo para dos puntos iniciales distintos en $[1,2]$.

a).

n	x_n
0	1
1	1.3570
2	1.4061
3	1.4094
4	1.4096
5	1.4096

b).

n	x_n
0	$\frac{\pi}{2}\simeq 1{,}5708$
1	1.4142
2	1.4199
3	1.4096
4	1.4096

   Raíz aproximada $\widetilde{x}_2=1{,}4096$ de $\widehat{x}_2$.

   Pongamos $T_1(x)=-\sqrt{1+\sin(x)}$ $x\in[-1,0]$. La aplicación T es contractiva en $[-1,0]$. Para $x_0=-0{,}5$, en 20 iteraciones se tiene $\widetilde{x}_1=-0{,}6367$.

2. Encontrar el menor valor $x\in\mathbb{R}^+$ para el cual $\tan^2(x)=2x+1$.

   Sean $f(x)=2x+1-\tan^2(x)$ y $\varphi_1(x)=\tan^2(x)$, $\varphi_2(x)=2x+1$ para $x\in\mathbb{R}^+$. El método gráfico muestra que la menor raíz positiva de la ecuación $f(x)=0$ está localizada en el intervalo $\left[\frac{\pi}{4},\frac{\pi}{2}\right[$. Notemos además que $\varphi_1(1)\simeq 2{,}43$, $\varphi_2(1)=3$ y $\varphi_1(1{,}2)\simeq 6{,}62$, $\varphi_2(1{,}2)=3{,}4$, luego $\widehat{x}\in[1,1{,}2]$.

   Elegir una función que sea contractiva resulta difícil.

Teorema 4 *Sea T de $[a,b]$ en $[a,b]$ una función derivable en $[a,b]$. Si $|T'(x)| \leq k < 1 \quad \forall x \in [a,b]$, entonces T es contractiva en $[a,b]$. Consecuentemente, existe $\widehat{x} \in [a,b]$ tal que $T(\widehat{x}) = \widehat{x}$.*

Demostración. Sean $x_1, x_2 \in [a,b]$ con $x_1 < x_2$. Entonces

$$|T(x_1) - T(x_2)| = \left|T'(x)(x_1 - x_2)\right| = \left|T'(x)\right| |x_1 - x_2| \leq k\,|x_1 - x_2|,$$

donde $x\,[a,b]$. Así, existe k, $0 \leq k < 1$ tal que

$$|T(x_1) - T(x_2)| \leq k\,|x_1 - x_2| \quad \forall x_1, x_2 \in [a,b],$$

que prueba que T es contractiva. Por el teorema de Banach del punto fijo, existe $\forall \widehat{x} \in [a,b]$ tal que $T(\widehat{x}) = \widehat{x}$. ■

Ejemplo

Sea $T(x) = -\sqrt{1+\sin(x)} \quad x \in [-1,0]$. Entonces

$$T'(x) = -\frac{\cos(x)}{\sqrt{1+\sin(x)}} \quad x \in [-1,0].$$

Se deduce que

$$0 < \frac{\cos(-1)}{2} \leq \frac{\cos(x)}{2} \leq T(x) \leq \frac{\cos(x)}{2\sqrt{1-\sin(1)}} \leq \frac{1}{2\sqrt{1-\sin(1)}}.$$

Luego $k = \dfrac{1}{2\sqrt{1-\sin(1)}} < 1$.

Nota: El método de punto fijo es muy limitado en las aplicaciones por la dificultad de seleccionar una aplicación contractiva, sin embargo ofrece una metodología para construir aplicaciones contractivas como se verá ás adelante.

5.4.2. Método de punto fijo modificado

Definición 5 *Sean $E \subset \mathbb{R}$ con $E \neq 0$ y f una función real definida en E. Se dice que f es Lipschisiana o que satisface la condición de Lipschitz si y solo si $\exists k > 0$ tal que $|f(x) - f(y)| \leq k\,|x-y| \quad \forall x, y \in E$. La constante k se llama constante de Lipschitz.*

Ejemplos

1. Sea f la función real definida por $f(x) = |x|$. Se tiene que f es lipschisiana. En efecto,para $x, y \in \mathbb{R}$,

   $$|f(x) - f(y)| = ||x| - |y|| \leq |x - y|.$$

   La constante de Lipschitz es $k = 1$.

2. La función f de $[0,\infty[$ en $[0,\infty[$ definida por $f(x) = \sqrt{1+x^2}$ es lipschisiana. Esta función se trató anteriormente, se probó que f no es contractiva y que

   $$|f(x) - f(y)| \leq |x - y| \quad \forall x, y \in [0,\infty[,$$

   que prueba que f es lipschisiana.

Teorema 5 *Sean $c, r \in \mathbb{R}$ con $r > 0$ y g uan función real definida en $[c-r, c+r]$. Si g es lipschisiana con constante $0 \leq k < 1$ tal que $|g(c) - c| \leq (1-k)\,r$, entonces*

i) Para todo $x \in [c-r, c+r]$, $g(x) \in [c-r, c+r]$.

ii) g tiene un único punto fijo $\widehat{x} \in [c-r, c+r]$.

Demostración. Por hipótesis g es lipschisiana con constante $0 \leq k < 1$. Entonces

$$|g(x) - g(y)| \leq k|x - y| \quad \forall x, y \in [c - r, c + r].$$

Además, $|g(c) - c| \leq (1 - k)r$ y por la desigualdad triangular, se tiene para $x \in [c - r, c + r]$,

$$\begin{aligned} |g(x) - c| &= |g(x) - g(c) + g(c) - c| \leq |g(x) - g(c)| + |g(c) - c| \leq k|x - c| + (1 - k)r \\ &\leq kr + (1 - k)r = r. \end{aligned}$$

Así, $|g(x) - c| \leq r \quad \forall x \in [c - r, c + r]$ que implica $g(x) \in [c - r, c + r]$. En otros términos, la imagen directa $g([c - r, c + r]) = [c - r, c + r]$.

Sea $\widetilde{g} : [c - r, c + r] \to [c - r, c + r]$ la función definida por

$$\widetilde{g}(x) = g(x) \quad \forall x \in [c - r, c + r].$$

Entonces $\widetilde{g}$ es contractiva en $[c - r, c + r]$. Pues existe $k, 0 \leq k < 1$ tal que

$$|\widetilde{g}(x) - \widetilde{g}(y)| = |g(x) - g(y)| \leq k|x - y| \quad \forall x, y \in [c - r, c + r].$$

Por otro lado, $[c - r, c + r]$ es un conjunto cerrado. Por el teorema de Banach del punto fijo, existe un único $\widehat{x} \in [c - r, c + r]$ tal que $\widetilde{g}(\widehat{x}) = \widehat{x} = g(\widehat{x})$, es decir que $\widehat{x}$ es el único punto fijo de g. ■

Sean $I \subset \mathbb{R}$ con $I \neq \emptyset$ y f una función real definida en I. Consideramos el problema (P) siguiente:

$$\text{hallar } \widehat{x} \in I, \text{ si existe, tal que } f(\widehat{x}) = 0. \tag{P}$$

Supongamos que el problema (P) tiene solución, esto es, existe al menos un $\widehat{x} \in I$ tal que $f(\widehat{x}) = 0$ y que dicha raíz ha sido separada; o sea existe $[a, b] \subset I$ en el que $\widehat{x}$ es la sola raíz de la ecuación $f(x) = 0$.

Como se ha dicho anteriormente, el problema radica en construir una función T que sea contractiva en $[a, b]$ tal que el punto fijo $\widehat{x}$ de T es la raíz $\widehat{x}$ de $f(x) = 0$ y recíprocamente. El siguiente teorema muestra como construir tal función T utilizando f.

Teorema 6 *Sea f una función real continua en $[a, b]$ tal que $f(a) f(b) < 0$ y sea $\widehat{x} \in [a, b]$ el único número real tal que $f(\widehat{x}) = 0$. Además, suponemos que*

$$\begin{aligned} 0 &< \alpha = \inf_{\substack{x,y \in [a,b] \\ x \neq y}} \left| \frac{f(x) - f(y)}{x - y} \right|, \\ \beta &= \sup_{\substack{x,y \in [a,b] \\ x \neq y}} \left| \frac{f(x) - f(y)}{x - y} \right|, \qquad \alpha < \beta. \end{aligned}$$

Entonces, existe una función T definida en $[a, b]$ tal que T es lipschisiana con constante $0 \leq k < 1$ y $T(\widehat{x}) = \widehat{x}$.

Demostración. Por hipótesis f es continua en $[a, b]$ y $f(a) f(b) < 0$. Por el teorema de Bolzano, existe $\widehat{x} \in [a, b]$ tal que $f(\widehat{x}) = 0$. Adicionalmente $\widehat{x} \in [a, b]$ es único.

Sea $m \in \mathbb{R}$. Definimos la función T en $[a, b]$ como sigue: $T(x) = x - mf(x) \quad x \in [a, b]$. Se tiene

$$T(\widehat{x}) = \widehat{x} - mf(\widehat{x}) = 0,$$

esto es, $\widehat{x}$ es un punto fijo de T, pues $f(\widehat{x}) = 0$.

Determinemos una constante m para la cual T sea lipschisiana en $[a, b]$ de constante $0 \leq k < 1$.

Sean $x, y \in [a, b]$ con $x < y$. Entonces

$$\begin{aligned} |T(x) - T(y)| &= |x - mf(x) - (y - mf(y))| = |x - y - m(f(x) - f(y))| \\ &= \left| (x - y)\left(1 - m\frac{f(x) - f(y)}{x - y}\right) \right| = |x - y| \left| 1 - m\frac{f(x) - f(y)}{x - y} \right|. \end{aligned}$$

Buscamos una constante $\widehat{m}$ tal que

$$\left|1 - m\frac{f(x) - f(y)}{x - y}\right| \leq k < 1 \quad \forall x, y \in [a, b] \text{ con } x \neq y.$$

Primeramente, la hipótesis $0 < \alpha = \inf_{\substack{x,y\in[a,b]\\ x\neq y}} \left|\frac{f(x)-f(y)}{x-y}\right|$ implica que los cocientes $\frac{f(x) - f(y)}{x - y}$ conservan el signo para todo $x, y \in [a, b]$ con $x \neq y$.

Definimos $\widehat{m} = \begin{cases} -\frac{1}{\beta}, & \text{si } \frac{f(x) - f(y)}{x - y} < 0 \quad \forall x, y \in [a, b] \quad \text{con} \quad x \neq y, \\ \frac{1}{\beta}, & \text{si } \frac{f(x) - f(y)}{x - y} > 0 \quad \forall x, y \in [a, b] \quad \text{con} \quad x \neq y. \end{cases}$

i. Si $\widehat{m} > 0$ entonces $\widehat{m} = \frac{1}{\beta}$. Se tiene, $\forall x, y \in [a, b]$ con $x \neq y$,

$$\frac{f(x) - f(y)}{x - y} \leq \beta = \frac{1}{\widehat{m}} \Longrightarrow \widehat{m}\frac{f(x) - f(y)}{x - y} \leq 1,$$

de donde $0 \leq 1 - \widehat{m}\frac{f(x) - f(y)}{x - y}$ $\forall x, y \in [a, b]$ con $x \neq y$.

Además, $\forall x, y \in [a, b]$ con $x \neq y$,

$$\alpha \leq \left|\frac{f(x) - f(y)}{x - y}\right| = \frac{f(x) - f(y)}{x - y} \Longrightarrow \widehat{m}\alpha \leq \widehat{m}\frac{f(x) - f(y)}{x - y},$$

con lo cual

$$0 \leq 1 - \widehat{m}\frac{f(x) - f(y)}{x - y} \leq 1 - \widehat{m}\alpha = 1 - \frac{\alpha}{\beta} = k < 1.$$

ii. Si $\widehat{m} < 0$ entonces $\widehat{m} = -\frac{1}{\beta}$. Para todo $x, y \in [a, b]$ con $x \neq y$, se tiene

$$-\frac{f(x) - f(y)}{x - y} \leq \beta = -\frac{1}{\widehat{m}} \Longrightarrow \widehat{m}\frac{f(x) - f(y)}{x - y} \leq 1,$$

de donde

$$0 \leq 1 - \widehat{m}\frac{f(x) - f(y)}{x - y}.$$

como

$$\alpha \leq \left|\frac{f(x) - f(y)}{x - y}\right| = -\frac{f(x) - f(y)}{x - y} \quad \forall x, y \in [a, b] \quad \text{con} \quad x \neq y,$$

se sigue que

$$\widehat{m}\alpha \geq -\widehat{m}\frac{f(x) - f(y)}{x - y} \quad \forall x, y \in [a, b] \quad \text{con} \quad x \neq y,$$

consecuentemente

$$0 \leq 1 - \widehat{m}\frac{f(x) - f(y)}{x - y} \leq 1 + \widehat{m}\alpha = 1 - \frac{\alpha}{\beta} = k < 1.$$

De i) y ii) se concluye que

$$\left|1 - \widehat{m}\frac{f(x) - f(y)}{x - y}\right| \leq 1 - \frac{\alpha}{\beta} = k < 1 \quad \forall x, y \in [a, b] \quad \text{con} \quad x \neq y,$$

con lo cual

$$|T(x) - T(y)| \leq k\,|x - y| \quad \forall x, y \in [a, b],$$

que prueba que $T(x) = x - \widehat{m}f(x)$ $x \in [a, b]$ es lipschisiana con constante $0 \leq k < 1$. ■

Interpretación Geométrica

Sea f una función real continua en $[a,b]$ tal que $f(a)\,f(b)<0$ y sea $\widehat{x}\in[a,b]$ la única raíz de la ecuación $f(x)=0$. Sea aún

$$\begin{aligned}\beta &= \sup_{\substack{x,y\in[a,b]\\ x\neq y}}\left|\frac{f(x)-f(y)}{x-y}\right|,\\ \widehat{m} &= \begin{cases} -\dfrac{1}{\beta}, \text{ si } \dfrac{f(x)-f(y)}{x-y}<0 \quad \forall x,y\in[a,b] \quad \text{con} \quad x\neq y,\\ \dfrac{1}{\beta}, \text{ si } \dfrac{f(x)-f(y)}{x-y}>0 \quad \forall x,y\in[a,b] \quad \text{con} \quad x\neq y.\end{cases}\\ T(x) &= x-\widehat{m}f(x) \quad x\in[a,b].\end{aligned}$$

Por el teorema precedente, T es lipschisiana de constante $0\leq k<1$ y $T(\widehat{x})=\widehat{x},\ \ f(\widehat{x})=0$.

Sea $x_0\in[a,b]$ y $x_{n+1}=T(x_n)\quad n=0,1,\ldots$. La interpretación geométrica del método iterativo

$$\begin{cases} x_0\in[a,b]\\ x_{n+1}=T(x_n) \quad n=0,1,\ldots\end{cases}$$

es la siguiente: se trata de aproximar la raíz $\widehat{x}$ de la ecuación $f(x)=0$ mediante la sucesión de puntos de intersección de las rectas de ecuación

$$y=\frac{x-x_n}{\widehat{m}}+f(x_n) \quad n=0,1,\ldots$$

con el eje X. Así la recta L_1 que pasa por $(x_0,f(x_0))$ y tiene pendiente $\dfrac{1}{\widehat{m}}$ con $x_0\in[a,b]$ viene dada por

$$y-f(x_0)=\frac{1}{\widehat{m}}(x-x_0).$$

Entonces, $y=0\Longleftrightarrow x=x_0-\widehat{m}f(x_0)$. Ponemos $x_1=x_0-\widehat{m}f(x_0)$.

La recta L_2 que pasa por $(x_1,f(x_1))$ y que tiene pendiente $\dfrac{1}{\widehat{m}}$ es

$$y-f(x_1)=\frac{1}{\widehat{m}}(x-x_1),$$

con lo cual $y=0\Longleftrightarrow x=x_1-\widehat{m}f(x_1)$. Ponemos $x_2=x_1-\widehat{m}f(x_1)$.

Continuando con este proceso obtenemos una sucesión (x_n) que converge a la raíz $\widehat{x}$ de $f(x)=0$.

En el gráfico que se muestra continuación se ilustra este procedimiento.

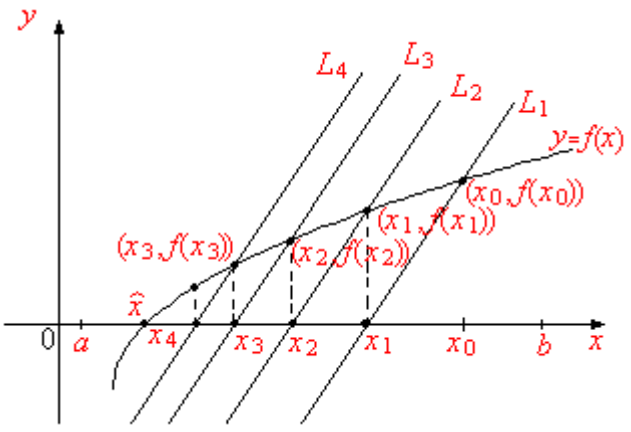


Figura 50

Algoritmo

El proceso iterativo $\begin{cases} x_0 \in [a,b] \text{ aproximación inicial,} \\ x_{n+1} = T(x_n) \quad n = 0, 1, \cdots . \end{cases}$ se llama método iterativo de punto fijo modificado. El problema principal de este método es calcular la constante $\widehat{m}$.

Supongamos que la función continua f cambia de signo en el intervalo $[a,b]$, esto es, $f(a) f(b) < 0$ y que la longitud $b-a$ del intervalo $[a,b]$ sea suficientemente pequeña para que la aplicación $T(x) = x - mf(x)$ sea contractiva en $[a,b]$ con $m \in \mathbb{R}$. En estas condiciones elegimos m como sigue

$$m = \frac{b-a}{f(b) - f(a)}$$

con lo cual

$$T(x) = x - \frac{b-a}{f(b) - f(a)} f(x) \quad x \in [a,b].$$

El método iterativo de punto fijo modificado queda en la forma

$$\begin{cases} x_0 \in [a,b] \text{ aproximación inicial,} \\ x_{n+1} = x_n - \frac{b-a}{f(b)-f(a)} f(x_n) \quad n = 0, 1, 2, \cdots . \end{cases}$$

Estimemos el número máximo de iteraciones que se requieren para aproximar $\widehat{x}$ con una precisión $\varepsilon = 10^{-t}$ con $t \in \mathbb{Z}^+$ (por ejemplo $\varepsilon = 10^{-2}, 10^{-5}, 10^{-6} \ldots$). En la demostración del teorema de Banach del punto fijo se dedujo la desigualdad

$$|x_n - x_m| \leq \frac{k^n}{1-k} |x_0 - x_1| \qquad m, n \in \mathbb{Z}^+, 0 < k < 1.$$

Dejando fijo n y considerando que $\widehat{x} = \lim\limits_{m \to \infty} x_m$, se tiene por la continuidad de la función valor absoluto

$$\begin{aligned} \lim_{m\to\infty} |x_n - x_m| &\leq \lim_{m\to\infty} \frac{k^n}{1-k} |x_0 - x_1| \Leftrightarrow \left| x_n - \lim_{m\to\infty} \right| \leq \frac{k^n}{1-k} |x_0 - x_1| \Leftrightarrow \\ |x_n - \widehat{x}| &\leq \frac{k^n}{1-k} |x_0 - x_1| . \end{aligned}$$

Puesto que $x_0, x_1 \in [a,b]$, $|x_0 - x_1| \leq b - a$, se sigue que

$$|x_n - \widehat{x}| \leq \frac{b-a}{1-k} k^n,$$

y como

$$\lim_{n\to\infty} \frac{b-a}{1-k} k^n = \frac{b-a}{1-k} \lim_{n\to\infty} k^n = 0,$$

entonces para $\varepsilon = 10^{-t}$, $\exists n_0 \in \mathbb{Z}^+$ tal que $\forall n \geq n_0 \Rightarrow \dfrac{b-a}{1-k} k^n < \varepsilon$.

Para $n = n_0$ podemos asumir que $\dfrac{b-a}{1-k} k^{n_0} \leq \varepsilon$. Tomando logaritmos en ambos miembros de esta última desigualdad, obtenemos

$$n_0 \ln(k) \leq \ln\left(\frac{\varepsilon(1-k)}{b-a} \right),$$

y como $0 < k < 1$, $\ln(k) < 0$. Luego

$$n_0 \geq \frac{\ln\left(\frac{\varepsilon(1-k)}{b-a} \right)}{\ln(k)}.$$

El número máximo de iteraciones $N_{\text{máx}}$ elegimos como sigue:

$$N_{\text{máx}} = \left[\frac{\ln \frac{(\varepsilon(1-k))}{b-a}}{\ln(k)} \right] + 1,$$

con $[\cdot]$ la función mayor entero menor o igual que.

Así, $|x_n - \widehat{x}| < \varepsilon = 10^{-t} \quad n = 1, 2, \ldots, N_{\text{máx}}$.

Por el teorema 5, la constante $k = 1 - \frac{\alpha}{\beta}$ resulta difícil de estimar.

Algoritmo

Datos de entrada: a, b extremos de $[a, b]$, función f, $\varepsilon = 10^{-t}$, $N_{\text{máx}}$.

Datos de salida: Valor aproximado $\widetilde{x}$ de $\widehat{x}$, n número de iteraciones.

1. Leer $x_0 \in [a, b]$ y poner $x = x_0$.
2. Calcular $m = \frac{b-a}{f(b)-f(a)}$.
3. Para $n = 0, \ldots, N_{\text{máx}}$.
4. Calcular $y = x - mf(x)$.
5. Si $|x - y| < \varepsilon$. Continuar en 7).
6. $x = y$.
7. Si $n < N_{\text{máx}}$, imprimir "raíz aproximada $\widetilde{x} = y$ en n iteraciones". Continuar en 9).
8. Imprimir "raíz aproximada $\widetilde{x} = y$ en $N_{\text{máx}}$ iteraciones".
9. Fin.

Ejemplos

1. Calcular el valor aproximado de $\sqrt[4]{2}$ con 6 cifras decimales de precisión.

   Sea $f(x) = x^4 - 2$ con $x > 0$. Entonces

   $$f(x) = 0 \Leftrightarrow x^4 - 2 = 0 \Rightarrow x = \sqrt[4]{2}.$$

   La función f tiene a $\sqrt[4]{2}$ como raíz localizada en el intervalo $[1, 1{,}5]$. Además, $f(1) = -1$, $f(1{,}5) = 5{,}0625$ y $f(1) \cdot f(1{,}5) < 0$. Tenemos $a = 1, b = 1{,}5$. Luego

   $$m = \frac{b-a}{f(b) - f(a)} = \frac{1{,}5 - 1}{f(1{,}5) - f(1)} = 0{,}1231.$$

   El esquema numérico es el siguiente:

   $$\begin{cases} x_0 \in [1, 1{,}5] \quad \text{aproximació inicial,} \\ x_{n+1} = x_n - 0{,}1231 f(x_n) \quad n = 0, 1, \cdots. \end{cases}$$

   Tomando $x_0 = 1$, en la tabla de la izquierda se muestran los resultados de la aplicación del esquema numérico. En la tabla de la derecha se muestra la aplicación del mismo esquema numérico pero con otra aproximación inicial $x_0 = 1{,}3$.

n	x_n
0	1,0
1	1,1231
2	1,173446
3	1,186241
4	1,188688
⋮	⋮
10	1,189207102
11	1,189207113
12	1,189207115,

n	x_n
0	1,3
1	1,194614
2	1,190106
3	1,189367
4	1,189233
5	1,189212
6	1,18907
7	1,189207

   .

   El valor aproximado de $\sqrt[4]{2}$ con una precisión de 6 cifras decimales es 1,189207. Valor obtenido en una calculadora de bolsillo $\widetilde{x} = 1{,}18927115\ldots$

2. Encontrar todas las raíces reales de la ecuación $x^3 + 2{,}1x^2 + 32{,}17x - 23{,}205 = 0$.

Solución: pongamos $f(x) = x^3 + 2{,}1x^2 + 32{,}17x - 23{,}205$ y escribamos f utilizando el esquema de Hörner. Se tiene

$$f(x) = -23{,}205 + x(-32{,}17 + x(2{,}1 + x)).$$

Busquemos las raíces en el intervalo $[-10, 10]$ (en la sección 6 se precisará como determinar los intervalos en los que están localizadas las raíces reales de polinomios). Sea $n = 50$, entonces $h = \frac{20}{50} = 0{,}4$. Aplicamos el algoritmo de búsqueda de cambio de signo. Tenemos

x	$f(x)$
-10	$-491{,}5$
$-9{,}6$	$-405{,}6$
$\vdots$	$\vdots$
$-7{,}2$	$-55{,}9$
$-6{,}8$	$-21{,}8$
$-6{,}4$	$6{,}555$
$-6{,}0$	$29{,}4$
$\vdots$	$\vdots$

x	$f(x)$
$-1{,}2$	$16{,}7$
$-0{,}8$	$3{,}4$
$-0{,}4$	$-10{,}1$
0	$-23{,}205$
$\vdots$	$\vdots$

x	$f(x)$
$4{,}4$	$-38{,}9$
$4{,}8$	$-18{,}6$
$5{,}2$	$6{,}9$
$5{,}6$	$38{,}1$
$\vdots$	$\vdots$
10	$865{,}1$

La ecuación $f(x) = 0$ tiene 3 raíces localizadas (separadas) en los intervalos $[6{,}8, -6{,}4]$, $[-0{,}8, -0{,}4]$, $[4{,}8, 5{,}2]$.

Observación: La constante $\widehat{m}$ está definida por (véase teorema 5)

$$\widehat{m} = \begin{cases} -\frac{1}{\beta}, \text{ si } \frac{f(x)-f(y)}{x-y} < 0 \quad \forall x, y \in [a,b] \quad \text{con } x \neq y, \\ \frac{1}{\beta}, \text{ si } \frac{f(x)-f(y)}{x-y} > 0 \quad \forall x, y \in [a,b] \text{ con } x \neq y. \end{cases}$$

con $\beta = \underset{\substack{x,y\in[a,b] \\ x\neq y}}{\text{Sup}} \left|\frac{f(x)-f(y)}{x-y}\right|$ es equivalente a la siguiente selección

$$\widehat{m} = \frac{\text{sign}\left(\frac{f(x')-f(y')}{x'-y'}\right)}{\underset{\substack{x,y\in[a,b] \\ x\neq y}}{\text{Sup}} \left|\frac{f(x)-f(y)}{x-y}\right|},$$

donde $x', y' \in [a,b]$ con $x' \neq y'$ los puntos en los que $\left|\frac{f(x)-f(y)}{x-y}\right|$ $x \neq y$, alcanza el valor extremo.

Si la función f es derivable en cada subintervalo en el que está separada cada raíz de $f(x) = 0$, se sigue que

$$f(x_1) - f(x_2) = f'(x)(x_1 - x_2) \text{ con } x \text{ entre } x_1 \text{ y } x_2,$$

de donde

$$\frac{f(x_1) - f(x_2)}{x_1 - x_2} = f'(x), \quad x_1 \neq x_2.$$

Resulta que si $f'(x) \neq 0 \ \ \forall x \in [a,b]$,

$$\widehat{m} = \frac{\text{sign}(f'(\overline{x}))}{\underset{x\in[a,b]}{\text{Sup}} |f'(x)|},$$

con $\overline{x} \in [a,b]$ en el que $|f'(x)|$ alcanza el valor extremo. Entonces, la aplicación contractiva T está definida por

$$T(x) = x - \widehat{m} f(x) \qquad x \in [a,b].$$

Calculemos las raíces de $f(x) = 0$ utilizando $\widehat{m}$ el estimado con $f'(x)$. Tenemos

$$f'(x) = 3x^2 + 4{,}2x - 32{,}17 = -32{,}17 + x(4{,}2 + 3x).$$

i. Cálculo de $\widehat{x}_1 \in [6{,}3, -6{,}4]$.

Se tiene $f'(-6{,}8) = 79{,}99$, $f'(-6{,}4) = 63{,}83$. Luego

$$\operatorname*{Sup}_{x\in[-6{,}8,-6{,}4]} \left|f'(x)\right| = 79{,}99,$$

con lo cual $\widehat{m_1} = \frac{1}{79{,}99} \simeq 0{,}013$, $T(x) = x - 0{,}013 f(x) \quad x \in [6{,}3, -6{,}4]$.

Esquema numérico: $\begin{cases} x_0 \in [-6{,}8, -6{,}4] \text{ punto inicial,} \\ x_{n+1} = T(x_n) \quad n = 0, 1, \cdots. \end{cases}$ En la tabla siguiente se muestran los resultados de la aplicación del esquema numérico precedente:

n	x_n
0	$-6{,}8$
1	$-6{,}517$
2	$-6{,}502$
3	$-6{,}500\ldots$
4	$-6{,}50003$
5	$-6{,}500000$

La solución en el intervalo $[6{,}3, -6{,}4]$ es**:** $\widehat{x}_1 = -6{,}5$.

ii. Cálculo de $\widehat{x}_2 \in [-0{,}8, -0{,}4]$.

Se tiene $f'(-0{,}8) = -33{,}61$, $f'(-0{,}4) = -33{,}37$, $\operatorname*{Sup}_{x\in[-0{,}8,-0{,}4]} |f'(x)|$, $\widehat{m_2} = \frac{-1}{33{,}61} \simeq -0{,}03$ y $T(x) = x + 0{,}03 f(x) \quad x \in [-0{,}8, -0{,}4]$. En la tabla siguiente se muestran los resultados de la aplicación del algoritmo de punto fijo modificado:

n	x_n
0	$-0{,}8$
1	$-0{,}699$
2	$-0{,}700008$
3	$-0{,}699999\ldots$
4	$-0{,}7$

La solución en el intervalo $[-0{,}8, -0{,}4]$ es $\widehat{x}_2 = -0{,}7$.

iii. Cálculo de $\widehat{x}_3 \in [4{,}8, 5{,}2]$. Estimemos la constante $\widehat{m}_3$. Obtenemos $f'(4{,}8) = 57{,}11$, $f'(50{,}2) = 70{,}79$, luego

$$\begin{aligned} \widehat{m_3} &= \frac{1}{\operatorname*{Sup}_{x\in[4{,}8,5{,}2]} |f'(x)|} = \frac{1}{7079}, \\ \widehat{m_3} &\simeq 0{,}0141, \\ T(x) &= x - 0{,}0141 f(x) \quad x \in [4{,}8, 5{,}2]. \end{aligned}$$

El esquema numérico es el siguiente:

$$\begin{cases} x_0 = 4{,}8, \\ x_{n+1} = T(x_n) \quad n = 1, 2, \cdots, \end{cases}$$

En la tabla que se muestra a continuación se resumen los resultados obtenidos en la ejecución del algoritmo precedente.

n	x_n
0	4,8
1	5,063
2	5,098
3	5,0998
4	5,0999
5	5,09999
6	5,09999
7	5,1

La solución es $\widehat{x}_3 = 5{,}1$.

Calculemos ahora las raíces $\widehat{x}_1$, $\widehat{x}_2$, $\widehat{x}_3$ pero esta vez utilizamos $\widehat{m}$ dado por

$$\widehat{m} = \frac{b-a}{f(b) - f(a)}.$$

i. $\widehat{m_1} = \frac{0{,}4}{f(-6{,}4)-f(-6{,}8)} = \frac{0{,}4}{6{,}555+21{,}8} = 0{,}0141$. Note la diferencia entre este valor $\widehat{m}_1$ y el calculado con la derivada ($\widehat{m}_1 \approx 0{,}013$). Se tiene $T(x) = x - 0{,}0141 f(x)$. En la tabla siguiente se muestran los términos de la sucesión (x_n), con $x_{n+1} = T(x_n) \quad n = 0, 1, \cdots$, con x_0 dado:

n	x_n
0	$-6{,}8$
1	$-6{,}493$
2	$-6{,}4996$
3	$-6{,}4999$
4	$-6{,}49999$
5	$-6{,}499999$
6	$-6{,}5$

Raíz $\widehat{x}_1 = -6{,}5$.

ii. Para el cálculo de $\widehat{x}_2$, obtenemos

$$\widehat{m_2} = \frac{0{,}4}{f(-0{,}4) - f(-0{,}8)} = -0{,}0296, \quad T(x) = x + 0{,}0296 f(x) \quad x \in [-0{,}8, -0{,}4].$$

Procediendo como en el caso precedente, tenemos los siguientes resultados:

n	x_n
0	$-0{,}8$
1	$-0{,}7000\ldots$
2	$-0{,}7000019$
3	$-0{,}7$

Raíz $\widehat{x}_2 = 0{,}7$.

iii. Cálculo de $\widehat{x}_3$. Se tiene $\widehat{m_3} = \frac{0{,}4}{f(5{,}2)-f(4{,}8)} = 0{,}0157 \quad T(x) = x - 0{,}0157 f(x) \quad x \in [4{,}8, 5{,}2]$. Como en los casos precedentes, se obtienen los siguientes resultados:

n	x_n
0	4,8
1	5,093
2	5,1004
3	5,09997
4	5,100001
5	$5{,}099999\ldots$
6	5,1

$\widehat{x}_3 = 5{,}1$.

3. Encontrar el menor $\widehat{x} \in \mathbb{R}^+$ tal que $\tan^2(\widehat{x}) = 2\widehat{x} + 1$.

Solución: sean $f(x) = 2x + 1 - \tan^2(x)$, $\varphi_1(x) = \tan^2(x)$, $\varphi_2(x) = 2x + 1$. El método gráfico muestra que la menor raíz positiva de $f(x) = 0$ está localizada en el intervalo $\left[\frac{\pi}{4}, \frac{\pi}{2}\right[$ (véase gráfico adjunto).

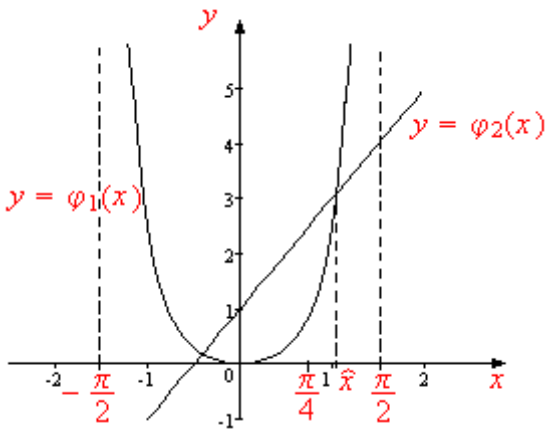


Figura 51

Más aún, $\varphi_1(1{,}2) \simeq 6{,}62$, $\varphi_2(1{,}2) \simeq 3{,}4$. Luego $\widehat{x} \in [1, 1{,}2]$,

$$\begin{aligned} f(1) &\simeq 0{,}57, f(1{,}2) \simeq -3{,}22, \\ \widehat{m} &= \frac{b-a}{f(b)-f(a)} \simeq -0{,}0528, \end{aligned}$$

y la función contractiva está dada por

$$T(x) = x + 0{,}0528 f(x) \quad x \in [1, 1{,}2].$$

Con $x_0 = 1$, la sucesión (x_n) con $x_{n+1} = T(x_n) \quad n = 0, 1, \ldots$ converge a $\widehat{x}$. Con $n = 30$ iteraciones $\widehat{x} \simeq 1{,}05486$ con una precisión $\varepsilon = 10^{-3}$.

5.4.3. Método de Newton-Raphson

Sea $I \subset \mathbb{R}$, $I \neq \emptyset$ y f una función real definida en I. Consideramos el problema (P) siguiente:

$$\text{hallar } \widehat{x} \in I, \text{ si existe, talque} \quad f(\widehat{x}) = 0. \tag{P}$$

Suponemos nuevamente que el problema (P) tiene solución; es decir, existe al menos un $\widehat{x} \in I$ tal que $f(\widehat{x}) = 0$ y que $\widehat{x}$ ha sido separada, o sea, existe $[a, b] \subset I$ en el que $\widehat{x}$ es la única raíz de $f(x) = 0$ allí localizada.

El método iterativo de punto fijo presenta los siguientes inconvenientes:

i) A partir de f, elegir una función contractiva T.

ii) Supuesto que se ha seleccionado una aplicación contractiva T y dado $x_0 \in [a, b]$, la sucesión (x_n) definida por $x_{n+1} = T(x_n)$, $n = 0, 1, \ldots$ converge muy lentamente.

En el método de punto fijo modificado, la construcción de la aplicación contractiva T es relativamente sencilla si la longitud del intervalo $[a, b]$ en el que está localizada la raíz es muy pequeña. En tal caso, si $f(a) f(b) < 0$,

$$T(x) = x - \frac{b-a}{f(b)-f(a)} f(x) \quad x \in [a, b]$$

es contractiva. La constante $m = \frac{b-a}{f(b)-f(a)}$ no es la óptima.

La sucesión (x_n) puede converger lentamente para $x_0 \in [a, b]$ dado y $x_{n+1} = T(x_n) \quad n = 0, 1, \ldots$. Note que no se requiere que f sea derivable. Para acelerar la convergencia se utilizará este método en el denominado método se Steffensen que será abordado en una sección posterior de este capítulo.

En el caso en que la función f sea derivable en $[a,b]$, uno de los métodos más utilizados para aproximar la solución $\widehat{x}$ de la ecuación $f(x)=0$ el método de Newton del que nos ocuparemos en esta sección.

En la demostración del teorema 5 se propuso la búsqueda de una constante m tal que

$$\left|1-m\frac{f(x)-f(y)}{x-y}\right|\leq k<1 \quad \forall x,y\in[a,b] \quad \text{con} \quad x\neq y.$$

Se ha supuesto que los cocientes $\dfrac{f(x)-f(y)}{x-y}$ $x,y\in[a,b]$ con $x\neq y$ conservan el signo de la constante $\widehat{m}$ dada por

$$\widehat{m}=\frac{\operatorname{sign}\left(f'(\overline{x})\right)}{\underset{x\in[a,b]}{\operatorname{Sup}}\ |f'(x)|},$$

con $\overline{x}\in[a,b]$ en el que $|f'(x)|$ alcanza el valor extremo (véase ejemplo 2 de la sección 4.2). Si en vez de buscar una constante global $\widehat{m}$ en $[a,b]$, se busca m que dependa de cada punto $x\in[a,b]$, esto es, para cada par de puntos $x_1,x_2\in[a,b]$ con $x_1<x_2$,

$$\begin{aligned} f(x_1)-f(x_2) &= f'(x)(x_1-x_2) \text{ con } x\in[x_1,x_2], \\ f'(x) &= \frac{f(x_1)-f(x_2)}{x_1-x_2}, \end{aligned}$$

se busca m que dependa de $x\in[a,b]$ y que satisfaga la condición

$$\left|1-m(x)\frac{f(x_1)-f(x_2)}{x_1-x_2}\right|\leq k<1 \quad x\in[a,b],$$

se tiene igualmente una aplicación contractiva. Si se pone $m=\dfrac{1}{f'(x)}$, con $f'(x)\neq 0$, entonces

$$T(x)=x-\frac{1}{f'(x)}f(x) \quad x\in[a,b]$$

es contractiva en $[a,b]$.

La función T es la función de iteración del método de Newton siguiente:

$$\begin{cases} x_0\in[a,b] \text{ dado}, \\ x_{n+1}=T(x_n)=x_n-\dfrac{f(x_n)}{f'(x_n)} \quad n=0,1,\ldots \end{cases}$$

Interpretación geométrica

Sea $x_0\in[a,b]$. La ecuación de la recta tangente a la gráfica de f en el punto $(x_0,f(x_0))$ es

$$L(x)=f'(x_0)(x-x_0)+f(x_0).$$

La recta L corta al eje X en el punto $(x_1,0)$, esto es

$$L(x)=0\Longleftrightarrow f'(x_0)(x-x_0)+f(x_0)=0\Longleftrightarrow x=x_0-\frac{f(x_0)}{f'(x_0)}.$$

Pongamos $x_1=x$. Notemos que x_1 es un valor aproximado de $\widehat{x}$.

La ecuación de la recta tangente a la gráfiica de f que pasa por el punto $(x_1,f(x_1))$ es

$$L_1(x)=f'(x_1)(x-x_1)+f(x_1),$$

que corta al eje X en $(x_2,0)$, en cuyop caso $L_1(x)=0$, de donde

$$x=x_1-\frac{f(x_1)}{f'(x_1)}.$$

Sea $x_2 = x$. Entonces x_2 es un valor aproximado de $\widehat{x}$.

Continuando con este procedimiento n veces, tenemos

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad n = 0, 1, \ldots,$$

donde x_{n+1} es un valor aproximado de la raíz $\widehat{x}$ de $f(x) = 0$.

De esta construcción podemos definir la función de iteración del método de Newton siguiente

$$T(x) = x - \frac{f(x)}{f'(x)} \quad x \in [a, b], \quad f'(x) \neq 0.$$

En la figura que se muestra a continuación se exhibe el procedimiento que acabamos de describir.

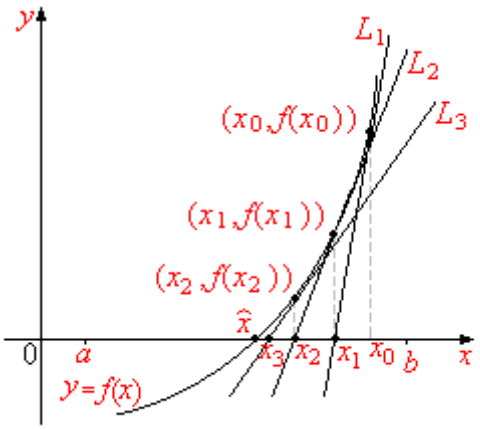


Figura 52

Otra forma de obtener la función de iteración del método de Newton es la siguiente. El desarrollo de Taylor en un entorno de $\widehat{x}$ raíz de la ecuación $f(x) = 0$, está dado por $0 = f(x) + f'(x)(x - \widehat{x})$ de donde

$$\widehat{x} = x - \frac{f(x)}{f'(x)} \qquad f'(x) \neq 0, \quad x \in [a, b].$$

Ponemos

$$T(x) = x - \frac{f(x)}{f'(x)} \qquad f'(x) \neq 0, \quad x \in [a, b],$$

y se tiene la función de iteración del método de Newton.

Observación

Supongamos que f tenga dos raíces $\widehat{x}_1, \widehat{x}_2$ en el intervalo $[a, b]$ y que f posea derivada segunda continua en $]a, b[$.

Sea $[\widehat{x}_1 - r, \widehat{x}_1 + r]$ un entorno cerrado de $\widehat{x}_1$ y $x_0 \in [\widehat{x}_1 - r, \widehat{x}_1 + r]$ tal que $f''(x_0) = 0$, es decir $(x_0, f(x_0))$ es un punto de inflexión de f. Entonces, el método de Newton, en general, no converge a $\widehat{x}_1$ sino a $\widehat{x}_2$ o

bien diverge. Véase la figura siguiente:

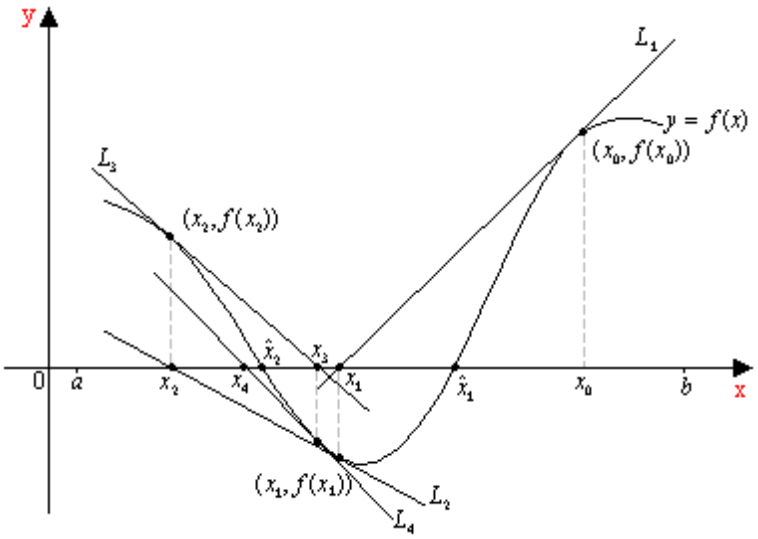


Figura 53

Supongamos que $f''(\widehat{x}) = 0$, es decir que $(\widehat{x}, f(\widehat{x})) = (\widehat{x}, 0)$ es un punto de inflexión de f y que $\widehat{x}$ es la sola raíz localizada en $[a, b]$.

Sea $x_0 \in [a, b]$. La sucesión (x_n) con $x_{n+1} = x_n - \dfrac{f(x_n)}{f''(x_n)}$ $n = 0, 1, \ldots$ puede ser divergente.

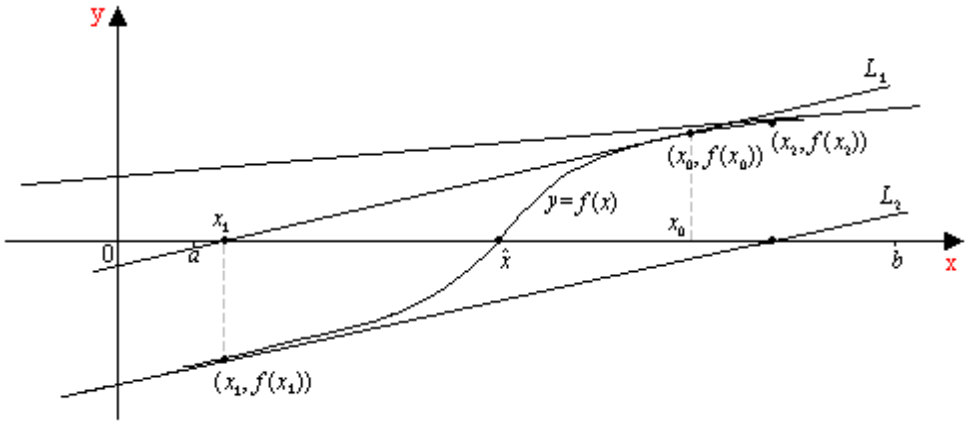


Figura 54

Estos dos problemas ponen de manifiesto que el método de Newton-Raphson, en general, no converge a $\widehat{x}$.

Establezcamos las condiciones que la función f y el punto inicial x_0 han de verificar para que el método de Newton-Raphson sea convergente.

Teorema 7 *Supongamos que $f \in C^2([a, b])$ tal que $f'(x) \neq 0$, $f''(x) \neq 0$ $\forall x \in [a, b]$. Si $x_0 \in [a, b]$ es una aproximación inicial de $\widehat{x}$ tal que $f(x_0) f''(x_0) > 0$, entonces la sucesión (x_n) generada por la función de iteración T converge a $\widehat{x}$.*
El punto x_0 se denomina extremo de Fourier.

Demostración. Notemos que $\widehat{x} \in [a, b]$ es la única raíz de $f(x) = 0$ allí localizada. Se tiene $f(a) f(b) < 0$. Sin pérdida de generalidad podemos suponer que $f(a) < 0$ y $f(b) > 0$; y que $f'(x) > 0$, $f''(x) > 0$

$\forall x \in [a,b]$; esto es, f es estrictamente creciente y convexa.

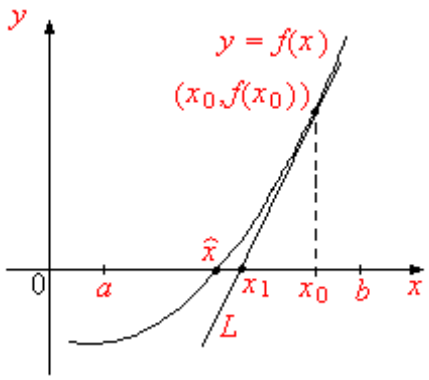


Figura 55

Sea $x_0 \in [a,b]$ y supongamos que $f(x_0) f''(x_0) > 0$. Como $f''(x_0) > 0$, se sigue que $f(x_0) > 0$ y par todo $n\mathbb{Z}^+$, $x_n > \widehat{x}$ y $f(x_n) > 0$. Probemos por inducción.

i) $x_0 > \widehat{x}$. Puesto que $f(\widehat{x}) = 0$ y f es estrictamente creciente, se tiene

$$f(x_0) - f(\widehat{x}) = f(x_0) > 0,$$

con lo cual $x_0 > \widehat{x}$.

ii) Supongamos que $x_n > \widehat{x}$ para $n \geq 0$. Probemos que $x_{n+1} > \widehat{x}$ y $f(x_{n+1}) > 0$. Como $f \in C^2([a,b])$. Por el desarrollo de Taylor, se tiene

$$0 = f(\widehat{x}) = f(x_n) + f'(x_n)(\widehat{x} - x_n) + \frac{1}{2!} f''(x_n)(\widehat{x} - x_n)^2.$$

Puesto que $f''(x_n) > 0$, entonces $\frac{1}{2!} f''(x_n)(\widehat{x} - x_n)^2 > 0$. Para que la igualdad precedente tenga lugar; debemos tener

$$f(x_n) + f'(x_n)(\widehat{x} - x_n) < 0.$$

Además, por hipótesis $f'(x_n) > 0$. Luego

$$\frac{f(x_n)}{f'(x_n)} + \widehat{x} - x_n < 0,$$

y multiplicando por -1, obtenemos

$$-\frac{f(x_n)}{f'(x_n)} + x_n - \widehat{x} > 0.$$

Entonces

$$\begin{aligned} x_{n+1} &= T(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}, \\ x_{n+1} - \widehat{x} &= x_n - \frac{f(x_n)}{f'(x_n)} - \widehat{x} > 0, \end{aligned}$$

de donde $x_{n+1} - \widehat{x} > 0$ que implica $x_{n+1} > \widehat{x}$. Por ser f creciente, $f(x_{n+1}) - f(\widehat{x}) = f(x_{n+1}) > 0$.

Mostremos que (x_n) converge a $\widehat{x}$. Para ello probemos que (x_n) es una sucesión decreciente y acotada. Como $T(x) = x - \frac{f(x)}{f'(x)}$ $x \in [a,b]$, $f'(x) > 0$, resulta que T es derivable en $[a,b]$ pués $f \in C^2([a,b])$. Por el teorema del valor medio, tenemos

$$T(x_n) - T(\widehat{x}) = T'(t_n)(x_n - \widehat{x}) \quad \text{con } \widehat{x} < t_n < x_n.$$

Como $x_n > \widehat{x}$, $x_{n+1} > \widehat{x}$ y por definición de (x_n), se tiene

$$T(x_n) - T(\widehat{x}) = x_{n+1} - \widehat{x} > 0,$$

luego

$$T'(x_n)(x_n - \widehat{x}) > 0 \Rightarrow T'(x_n) > 0 \text{ ya que } x_n > \widehat{x}.$$

Además, $f(x_n) > 0$, $f'(x_n) > 0$, $\frac{f(x_n)}{f'(x_n)} > 0$, y

$$T(x_n) = x_n - \frac{f(x_n)}{f'(x_n)} \Rightarrow \frac{f(x_n)}{f'(x_n)} = x_n - T(x_n) > 0,$$

de donde

$$x_n - x_{n+1} > 0 \Rightarrow x_n > x_{n+1},$$

pues $x_{n+1} = T(x_n)$.

Así, $(x_n) \in c[a,b]$ y (x_n) decreciente. Luego (x_n) es convergente.

Para completar la prueba, mostremos que $\lim_{n\to\infty} x_n = \widehat{x}$.

Como $x_{n+1} < x_n$ se sigue que $x_{n+1} - \widehat{x} < x_n - \widehat{x}$. Luego

$$0 < \frac{x_{n+1} - \widehat{x}}{x_n - \widehat{x}} < 1.$$

Pero

$$\begin{aligned} x_{n+1} - \widehat{x} &= T(x_n) - \widehat{x} = T(x_n) - T(\widehat{x}) = T'(t_n)(x_n - \widehat{x}), \\ 0 &< T'(x_n) = \frac{x_{n+1} - \widehat{x}}{x_n - \widehat{x}} < 1 \end{aligned}$$

que prueba que T es contractiva. ■

Observaciones

1. De la condición $f(x_0) f''(x_0) > 0$ resulta que $f(x_0) > 0$ y $f''(x_0) > 0$ o $f(x_0) < 0$ y $f''(x_0) < 0$.

   i. Si $f(x_0)$ y $f''(x_0) > 0$, f es convexa. Por hipótesis del teorema 6, $f'(x) \neq 0 \ \ \forall x \in [a,b]$, f' mantiene el mismo signo en $[a,b]$ dando como resultado que f es estaricamente decreciente en $[a,b]$ o f es estrictamente creciente en $[a,b]$. En las dos figuras que se muestran a continuación se presentan estas dos situaciones

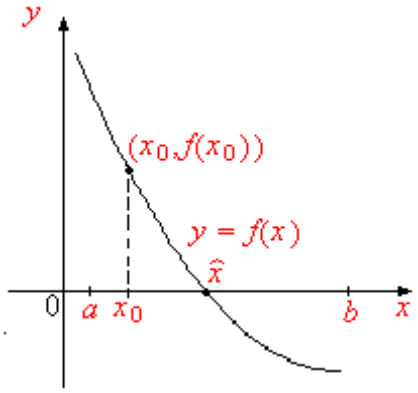


Figura 56

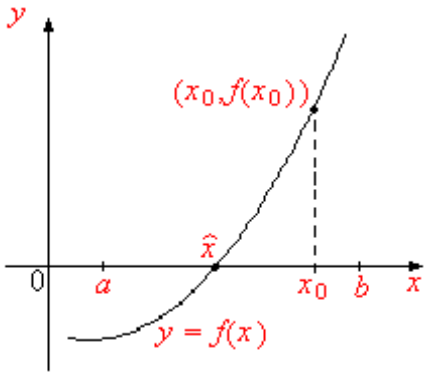


Figura 57

ii. Si $f(x_0) < 0$ y $f''(x_0) < 0$. Resulta que f es cóncava, estrictamente creciente o estrictamente decreciente.

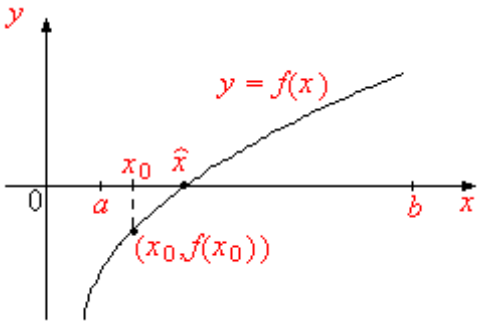


Figura 58

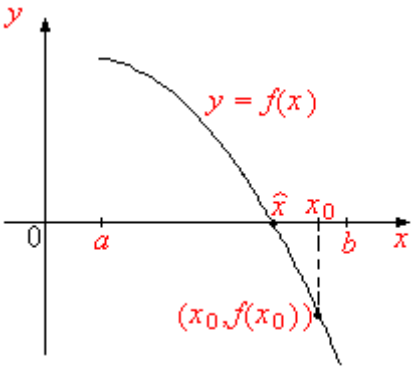


Figura 59

2. Si $f(x_0) f''(x_0) < 0$ entonces $f(x_0) > 0$ y $f''(x_0) < 0$ o $f(x_0) < 0$ y $f''(x_0) > 0$.

   Supongamos $f(x_0) > 0$ y $f''(x_0) < 0$. Entonces f es convexa. Puede suceder que $x_1 = T(x_0) \notin$

$[a,b]$.

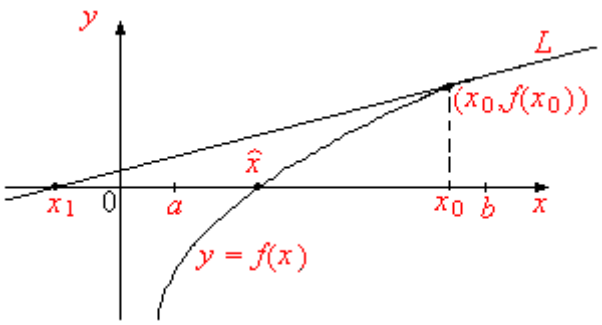


Figura 60

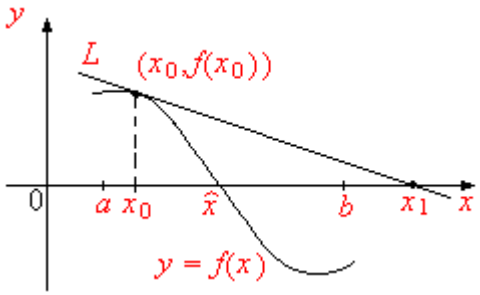


Figura 61

Teorema 8 *Supongamos que* $f \in C^2([a,b])$ *y* $\widehat{x} \in [a,b]$ *la única raíz de la ecuación* $f(x)=0$ *allí localizada. Si* $f'(\widehat{x}) \neq 0$, *existe* $r>0$ *tal que la sucesión* (x_n) *generada por el método de Newton converge a* $\widehat{x}$ *para todo* $x_0 \in [\widehat{x}-r, \widehat{x}+r]$ *aproximación inicial de* $\widehat{x}$.

Demostración. Mostremos que la función de iteración $T(x) = x - \frac{f(x)}{f'(x)}$ $x \in [a,b]$ es lipschisiana en $[\widehat{x}-r, \widehat{x}+r]$ para algún $r>0$ y constante $0<k<1$.

Por hipótesis $f \in C^2([a,b])$ entonces f' es continua en $[a,b]$. Además $f'(\widehat{x}) \neq 0$. Existe $\delta > 0$ tal que $f'(x) \neq 0$ para $x \in [\widehat{x}-\delta, \widehat{x}+\delta] \subset [a,b]$. Definimos

$$T(x) = x - \frac{f(x)}{f'(x)} \quad x \in [\widehat{x}-\delta, \widehat{x}+\delta] .$$

Como f y f' son continuas en $[a,b]$, entonces T es continua en $[\widehat{x}-\delta, \widehat{x}+\delta]$. Por otra parte,

$$T'(x) = 1 - \frac{[f'(x)]^2 - f(x) f''(x)}{(f'(x))^2} = \frac{f(x) f''(x)}{(f'(x))^2} \quad \forall x \in [\widehat{x}-\delta, \widehat{x}+\delta] .$$

Resulta $T \in C^1([\widehat{x}-\delta, \widehat{x}+\delta])$. Puesto que $f(\widehat{x})=0$ entonces $T'(x) = \frac{f(\widehat{x}) f''(\widehat{x})}{(f'(\widehat{x}))^2} = 0$. Luego

$$\lim_{x \to \widehat{x}} T'(x) = T'(x) = 0.$$

Sea $\varepsilon > 0$, existe $r>0$ tal que $\forall x \in [\widehat{x}-\delta, \widehat{x}+\delta]$ con $|x-\widehat{x}| < r \Rightarrow |T'(x) - T'(\widehat{x})| < \varepsilon$. En particular para $0<\varepsilon<1$, existe $r>0$ tal que $r<\delta$ y

$$\left|T'(x)\right| < \varepsilon < 1 \quad \forall x \in [\widehat{x}-r, \widehat{x}+r] ,$$

y por el teorema del valor medio

$$T(x) - T(\widehat{x}) = T'(t)(x-\widehat{x}) \quad \text{con } t \text{ entre } x \text{ y } \widehat{x}.$$

Entonces

$$|T(x) - T(\widehat{x})| = |T(x) - \widehat{x}| = \left|T'(t)\right| |x - \widehat{x}| < \varepsilon |x - \widehat{x}| \leq \varepsilon r < r.$$

Luego

$$|T(x) - \widehat{x}| < r \Leftrightarrow \widehat{x} - r < T(x) < \widehat{x} + r \ \forall x \in [\widehat{x} - r, \widehat{x} + r].$$

Resulta que la imagen directa $T([\widehat{x} - r, \widehat{x} + r]) = [\widehat{x} - r, \widehat{x} + r]$ y $|T'(x)| \leq \varepsilon < 1 \ \ \forall x \in [\widehat{x} - r, \widehat{x} + r]$ que muesta que T es contractiva. Por el teorema de Banach del punto fijo, para todo $x_0 \in [\widehat{x} - r, \widehat{x} + r]$, la sucesión (x_n) con $x_{n+1} = T(x_n) \ \ n = 0, 1, \ldots$ converge a $\widehat{x}$. ■

Ejemplos

1. Sean $a > 0$, $n \in \mathbb{Z}^+$. Calculemos $\sqrt[n]{a}$.

   Sea f de $[0, \infty[$ en $\mathbb{R}$ la función definida por $f(x) = x^n - a \quad x \in [0, \infty[$.

   i) Si $0 < a < 1$ entonces $f(0) = -a < 0$, $f(1) = 1 - a > 0$. Por el teorema de Bolzano, existe $\widehat{x} \in [0, 1]$ tal que $f(\widehat{x}) = 0$, esto es,

   $$f(\widehat{x}) = 0 \Leftrightarrow \widehat{x}^n - a = 0 \Rightarrow \widehat{x} = \sqrt[n]{a}.$$

   Además, $f'(x) = nx^{n-1}$, $f''(x) = n(n-1)x^{n-2}$. Se tiene

   $$f'(x) > 0, \ f''(x) > 0 \quad \forall x \in]0, \infty[.$$

   La función de iteración del método de Newton está definida por

   $$T(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^n - a}{nx^{n-1}} = \frac{nx^n - x^n + a}{nx^{n-1}} = \frac{1}{n}\left[(1-n)x + \frac{a}{x^{n-1}}\right] \quad x \in [0, 1].$$

   Sea $x_0 = 1$, x_0 es el extremo de Fourier, pues $f(x_0) f''(x_0) > 0$. La sucesión (x_m) generada por la función de iteración T es convergente a $a^{\frac{1}{n}}$.

   ii) Si $a = 1$. Se tiene $\widehat{x} = 1$.

   iii) Si $a > 1$, entonces $f(1) = 1 - a < 0$, $f(a) = a^n - a > 0$. Por el teorema de Bolzano, existe $\widehat{x} \in [1, a]$ tal que $f(\widehat{x}) = 0$. La función de iteración está definida por

   $$T(x) = \frac{1}{n}\left[(n-1)x + \frac{a}{x^{n-1}}\right] \quad x \in [1, a].$$

   Para valores de $a > 1$ suficientemente grandes, resulta difícil elegir $x_0 \in [1, a]$ de modo que la sucesión (x_m) generada por la función de iteración T converja rápidamente.

   Sean $a > 1$ y $j \in \mathbb{Z}^+$ el mas pequeño entero tal que $10^{-nj}a < 1$. Ponemos $b = 10^{-nj}a$ y definimos $g(t) = t^n - b$. Por i)**.** existe $\widehat{t} \in [0, 1]$ tal que $g(\widehat{t}) = 0 \Leftrightarrow \widehat{t} = b^{\frac{1}{n}} = \left(10^{-nj}a\right)^{\frac{1}{n}} = 10^{-j}a^{\frac{1}{n}}$ de donde $a^{\frac{1}{n}} = 10^j\widehat{t}$. En estas condiciones, con $t_0 = 1$, la sucesión (t_m) definida por $t_{m+1} = \frac{1}{n}\left[(n-1)t_m + \frac{b}{t_m^{n-1}}\right] \quad m = 0, 1, \ldots$ converge a $\widehat{t}$ y en consecuencia $a^{\frac{1}{n}} = 10^j \lim_{m\to\infty} t_m = 10^j\widehat{t}$.

   Nota: Un hecho importante del análisis matemático es probar que todo número real es límite de una sucesión de números racionales. Para $a \in \mathbb{Q}^+$ que no sea potencia n-ésima de $c \in \mathbb{Q}^+$, la función de iteración

   $T(x) = \frac{1}{n}\left[(n-1)x + \frac{a}{x^{n-1}}\right]$ $x > 0$ proporciona una forma de construir sucesiones de números racionales que convergen a $a^{\frac{1}{n}} \notin \mathbb{Q}^+$. Así por ejemplo

   i. Para $a = 2$, $n = 2$, $x_0 = 2$ y $x_{m+1} = \frac{1}{2}\left(x_m + \frac{2}{x_m}\right)$. La sucesión (x_m) converge a $\sqrt{2} \notin \mathbb{Q}^+$.

   ii. Para $a = 2$, $n = 5$, $x_0 = 3$ y $x_{m+1} = \frac{1}{5}\left(4x_m + \frac{2}{x_m^4}\right)$, (x_m) converge a $\sqrt[5]{2} \notin \mathbb{Q}^+$.

   iii. Para $a = 5$, $n = 3$, $x_0 = 3$ y $x_{m+1} = \frac{1}{3}\left(2x_m + \frac{5}{x_m^2}\right)$, (x_m) converge a $\sqrt[3]{5} \notin \mathbb{Q}^+$.

2. Encontrar las raíces de la ecuación $e^{-\frac{x}{4}}(2-x)-1=0$.

   Ponemos $f(x)=e^{-\frac{x}{4}}(2-x)-1$. Entonces

$$f(x)=0 \Leftrightarrow e^{-\frac{x}{4}}(2-x)-1=0 \Leftrightarrow e^{-\frac{x}{4}}=\frac{1}{2-x} \quad x\neq 0.$$

   El método gráfico muestra que la ecuación $f(x)=0$ tiene única raíz localizada en el intervalo $[0,1]$, como se puede observar en la figura siguiente.

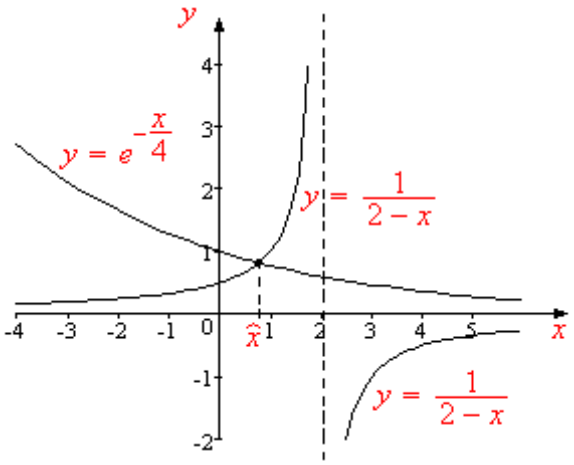


Figura 62

   Puesto que

$$\begin{aligned} f(x) &= e^{-\frac{x}{4}}(2-x)-1, \\ f'(x) &= -\frac{1}{4}e^{-\frac{x}{4}}(6-x), \\ f''(x) &= \frac{1}{16}e^{-\frac{x}{4}}(10-x). \end{aligned}$$

   Sea $x_0=0$. Entonces $f(0)\cdot f''(0)>0$ con lo cual x_0 es el extremo de Fourier. La función de iteración del método de Newton está dada por

$$T(x)=x-\frac{f(x)}{f'(x)}=x+\frac{4\left(2-x-e^{-\frac{x}{4}}\right)}{6-x} \quad x\in[0,1].$$

   Luego, el esquema numérico es $x_{n+1}=T(x_n)\ \ n=0,1,\ldots$. En la tabla siguiente se muestran los resultados de la aplicación del método de Newton:

n	x_n
0	0
1	0,666667
2	0,780646
3	0,783594
4	0,783596
5	0,783596

   Sea $x_0=1$. Se tiene $f(1)f''(1)<0$, x_0 no es el extremo de Fourier, sin embargo el método converge. A continuación se muestran los resultados de la aplicación del método de Newton:

n	x_n
0	1
1	0,772779
2	0,783570
3	0,783596
4	0,783596

Si equivocadamente se elige el punto $x_0 = 8 \notin [0,1]$, se tiene $T(x_0) = 34{,}778$, $T(x_1) = 869{,}153$, $T(x_2) = 1{,}079 \times 10^{32}$, $T(x_3)$ overflow. Así (x_n) diverge.

3. Hallar las raíces de la ecuación $e^x(x - 2{,}4055) + 3 = 0$. Como en los ejemplos anteriores definimos la función real f como sigue: $f(x) = e^x(x - 2{,}4055) + 3$. Entonces $f'(x) = e^x(x - 1{,}4055)$, $f''(x) = e^x(x - 0{,}4055)$. El estudio de la función f muestra que la ecuación $f(x) = 0$ tiene dos raíces ubicadas en los intervalos $[0,1]$ y $[1,3]$ como se puede observar en la figura que se muestra a continuación.

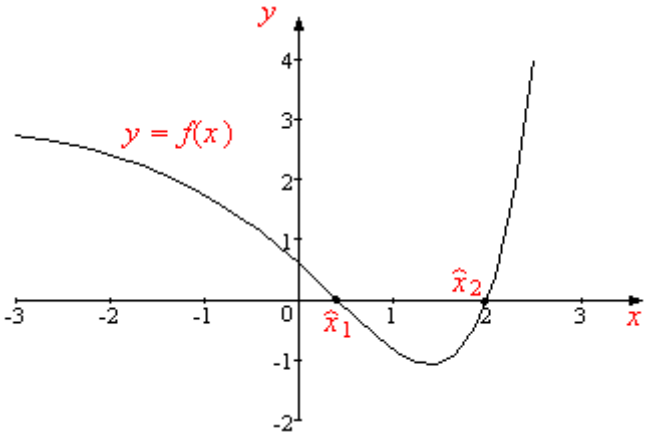


Figura 63

Además, f es convexa si $x > 0{,}4055$, cóncava si $x < 0{,}4055$.

La función de iteración del método de Newton está definida por

$$\begin{aligned}\Phi(x) &= x - \frac{f(x)}{f'(x)} = x - \frac{e^x(x - 2{,}4055) + 3}{e^x(x - 1{,}4055)} \\ &= \frac{2{,}4055 + x(-2{,}4055 + x) - 3e^{-x}}{x - 1{,}4055} \qquad x \neq 1{,}4055.\end{aligned}$$

Calculemos los valores aproximados de las raíces $\widehat{x}_1$, $\widehat{x}_2$ de la ecuación $f(x) = 0$. Para ello vamos a elegir un punto x_0 que sea, en unos casos, el extremo de Fourier, y en otros que no lo sea.

i. Sea $x_0 = 0$. Entonces $f(0) f''(0) < 0$. El punto x_0 no es el extremo de Fourier, sin embargo el método converge. A continuación se muestran los resultados del método de Newton:

n	x_n
0	0
1	0,422981
2	0,405428
3	0,405430
4	0,405430
5	0,405430.

Valor aproximado de $\widehat{x}_1$: 0,405430 (precisión $\varepsilon = 10^{-6}$).

ii. Sea $x_0 = 1{,}3$. Se tiene $f(1{,}3) f''(1{,}3) < 0$ con lo cual x_0 no es el extremo de Fourier. Los resultados son los siguientes.

n	x_n
0	1,3
1	−1,428954
2	1,636376
3	2,437981
4	2,152753.

Valor aproximado de $\widehat{x}_2$ con una precisión $\varepsilon = 10^{-6}$: 2,152753.

Si en i) se elige $x_0 = -1{,}5$ o $x_0 = -2$ para aproximar $\widehat{x}_1$, la sucesión (x_n) no converge a $\widehat{x}_1$ sino a $\widehat{x}_2$.

iii. Sea $x_0 = 1{,}4$. Se tiene $f(x_0) f''(x_0) < 0$. Con este punto inicial la sucesión (x_n) es divergente.

n	x_n
0	1,4
1	$-46{,}911$
2	$1{,}4659 \times 10^{19}$
3	$Overflow.$

Situación análoga si se toma $x_0 = -1{,}3$.

iv. Para $x_0 = 2{,}4$, $f(x_0) f''(x_0) > 0$, o sea x_0 es el extremo de Fourier. Se tiene

n	x_n
0	2,4
1	2,131871
2	2,018683
3	1,999645
4	1,999148
5	1,999148.

Para $\varepsilon = 10^{-6}$ el valor aproximado de $\widehat{x}_3$ es 1,999148.

5.4.4. Método de Newton modificado

El método de Newton-Raphson requiere que en cada paso se evalúe $f'(x)$ que en muchos casos puede resultar laborioso.

Sea x_0 un punto inicial para el que $f'(x) \neq 0$ y sea $\alpha = \dfrac{1}{f'(x_0)}$. La función de iteración Φ definida por $\Phi(x) = x - \alpha f(x)$ se denomina método de Newton modificado.

Hemos supuesto que una función f tiene un cero $\widehat{x}$ separado en $[a, b]$ y que f' existe en $[a, b]$. La sucesión (x_n) generada por el método de Newton modificado esta definida por

$$\begin{cases} x_0 \in [a, b] \text{ aproximación inicial,} \\ x_{n+1} = \Phi(x_n) = x_n - \alpha f(x_n) \quad n = 0, 1, \ldots \end{cases}$$

Si x_0 es el extremo de Fourier, la sucesión (x_n) generada por $x_{n+1} = \Phi(x_n) \quad n = 0, 1, \ldots$, converge a la sola raíz $\widehat{x}$ de $f(x) = 0$.

Ejemplo

Hallar las raíces reales de la ecuación $x^3 \sin(x) - 1 = 0$ en el intervalo $[0, \pi]$. Para el efecto, sea $f(x) = x^3 \sin(x) - 1 = 0 \;\; x \in [0, \pi]$. Entonces

$$f(x) = 0 \Leftrightarrow x^3 \sin x - 1 = 0 \Leftrightarrow \sin x = \frac{1}{x^3} \; x \neq 0$$

El método gráfico muestra dos raíces de $f(x) = 0$ ubicadas en los intervalos $\left[1, \frac{\pi}{2}\right]$ y $[3, \pi]$.

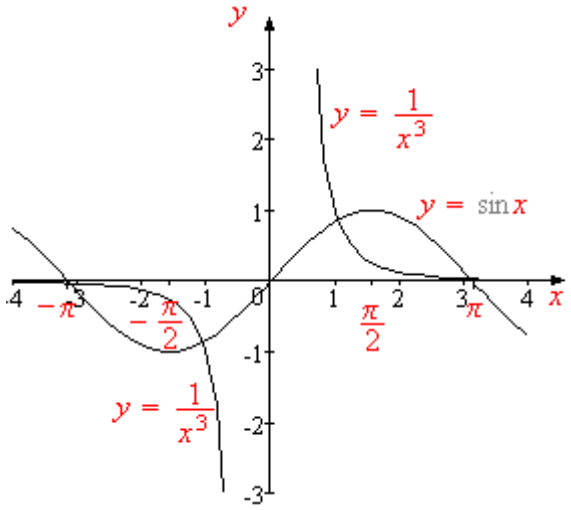


Figura 64

Note que $f(1) = -1{,}158$, $f\left(\frac{\pi}{2}\right) = 1{,}8757$, $f(3) = 1{,}81$, $f(\pi) = -1$. Además,

$$\begin{aligned} f'(x) &= x^2(3\sin x + x\cos x), \\ f''(x) &= x\left[\left(6 - x^2\sin x + 6x\cos x\right)\right]. \end{aligned}$$

Sea $x_0 = \frac{\pi}{2} \simeq 1{,}5708$. Entonces $f(x_0) f''(x_0) > 0$, x_0 es el extremo de Fourier. Ponemos

$$\alpha = \frac{1}{f'(x_0)} = \frac{1}{7{,}4022} \simeq 0{,}1351.$$

La función de iteración Φ del método de Newton modificado está dada por $\Phi(x) = x - 0{,}1351 f(x)$. Se tiene

$$\begin{cases} x_0 = \frac{\pi}{2}, \\ x_{n+1} = \Phi(x_n) \quad n = 0, 1, \cdots. \end{cases}$$

En 18 iteraciones se logra $\widehat{x}_1 = 1{,}278283055$.

Para el cálculo de la segunda raíz elegimos $x_0 = \pi$. Entonces $f'(\pi) = -\pi^3$, resulta

$$\begin{aligned} \alpha &= \frac{1}{f'(\pi)} = -\frac{1}{\pi^3} \simeq -0{,}03225, \\ \Phi(x) &= x + 0{,}03225 f(x). \end{aligned}$$

Par $n = 10$ se tiene $\widehat{x}_2 = 3{,}072589665$.

5.4.5. Método de las secantes

Sea $f \in C([a,b])$ tal que $f(a) f(b) < 0$. Por el teorema de Bolzano, existe $\widehat{x} \in [a,b]$ tal que $f(\widehat{x}) = 0$. Supongamos que $\widehat{x}$ es la única raíz allí localizada. Para evitar el problema de la evaluación de la derivada en el método de Newton, podemos obtener una variable de éste. Por definición

$$f'(x_{n-1}) = \lim_{x \to x_{n-1}} \frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}}.$$

Con esta aproximación de la derivada $f'(x_{n-1})$, el método de Newton se expresa en la siguiente forma:

$$x_n = x_{n-1} - \frac{1}{\frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}}} f(x_{n-1})$$

o bien

$$x_n = x_{n-1} - \frac{x_{n-1} - x_{n-2}}{f(x_{n-1}) - f(x_{n-2})} f(x_{n-1}) \quad n = 2, 3, \ldots \qquad (*)$$

La aproximación de la raíz $\widehat{x}$ de la ecuación $f(x) = 0$ utilizando la fórmula **(*)** se llama método de las secantes.

Interpretación geométrica

Sean $x_0, x_1 \in [a, b]$ tales que $f(x_0) f(x_1) < 0$. La ecuación de la recta que pasa por los puntos $(x_0, f(x_0))$, $(x_1, f(x_1))$ viene dada por

$$L_1(x) - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_1).$$

Luego,

$$L_1(x) = 0 \Leftrightarrow x = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1).$$

Sea $x_2 = x$. La ecuación de la recta que pasa por $(x_1, f(x_1))$, $(x_2, f(x_2))$ está definida por

$$\begin{aligned} L_1(x) - f(x_2) &= \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_2), \\ L_2(x) &= 0 \Leftrightarrow x = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2). \end{aligned}$$

Sea $x_3 = x$. Continuando con este procedimiento, obtenemos

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) f(x_{n-1})} f(x_n) \quad n = 1, 2, \cdots.$$

En la gráfica siguiente se muestra el procedimiento previamente descrito.

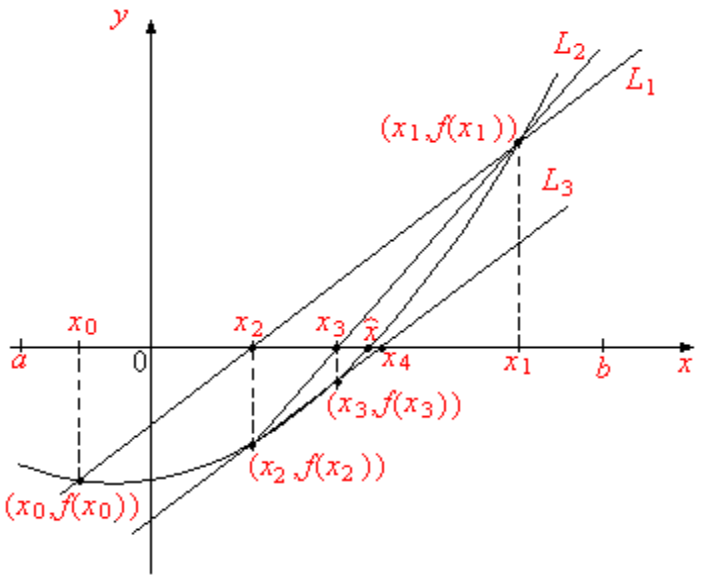


Figura 65

Ejemplo

Hallar la raíz positiva de la ecuación $e^{\sin(x)} - \frac{2}{1+x^2} = 0$.

Sea $f(x) = e^{\sin(x)} - \frac{2}{1+x^2}$. Entonces

$$f'(c) = e^{\sin(x)} \cos(x) + \frac{4x}{(1 + x^2)^2}.$$

Este es un ejemplo de una función f cuyo estudio de f conduce a resolver las inecuaciones $f'(x) > 0$, $f'(x) < 0$ y a la ecuación $f'(x) = 0$ más complicadas que la ecuación $f(x) = 0$. Puesto que

$$f(x) = 0 \Leftrightarrow e^{\sin(x)} - \frac{2}{1+x^2} = 0 \Leftrightarrow e^{\sin(x)} = \frac{2}{1+x^2}.$$

Sean $\varphi_1(x) = e^{\sin(x)}$, $\varphi_2(x) = \frac{2}{1+x^2}$. En la figura que se muestra a continuación se exhiben las gráficas de φ_1 y φ_2.

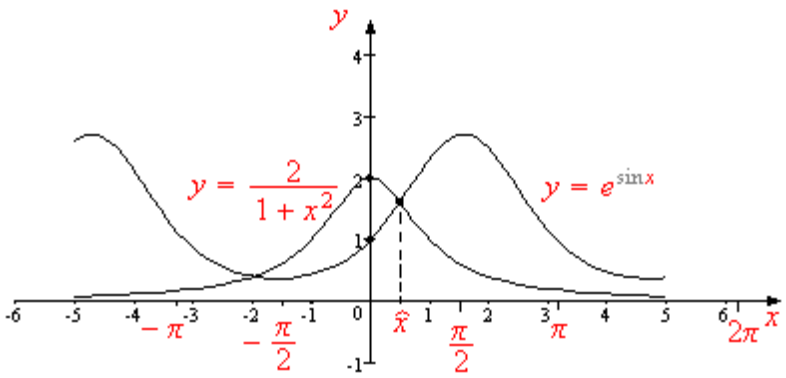


Figura 66

La búsqueda del cambio de signo con un paso $h = 0{,}2$ en el intervalo $[0, \infty[$ muestra que $f(x) = 0$ tiene una sola raíz en $[0, \infty[$ ubicada en el intervalo $[0{,}4, 0{,}6]$:

x	0	0,2	0,4	0,6
$f(x)$	-1	$-0{,}7$	$-0{,}25$	0,29

Ponemos $x_0 = 0{,}4$, $x_1 = 0{,}6$, $y_0 = f(0{,}4) = -0{,}24802$, $y_1 = f(0{,}6) = 0{,}28823$. En la tabla siguiente se muestran los resultados de la aplicación del método de las secantes:

n	x_n
0	0,4
1	0,6
2	0,49250
3	0,49435
4	0,49438
5	0,49438.

Valor aproximado de $\widehat{x}$ con una precisión $\varepsilon = 10^{-5}$: 0,49438.

Algoritmo

Datos de entrada: aproximaciones iniciales $x_0, x_1 \in [a, b]$, $\varepsilon = 10^{-t}$, $N_{\text{máx}}$ número máximo de iteraciones, función f.

Datos de salida: $\widehat{x}$, n número de iteraciones.

1. $y_0 = f(x_0)$.
2. $y_1 = f(x_1)$.
3. $n = 2, \ldots, N_{\text{máx}}$.
4. $x = x_1 - \dfrac{x_1 - x_0}{y_1 - y_0} y_0$.
5. Si $|x - x_1| < \varepsilon$, continuar en 7).

6. $x_0 = x_1$.

   $y_0 = y_1$.

   $x_1 = x$.

   $y_1 = f(x)$.

7. Si $n < N_{\text{máx}}$, imprimir x, n. Continuar en 9).

8. Si $n = N_{\text{máx}}$, imprmir $x, f(x)$.

9. Fin.

5.4.6. Método regula-falsi

Sean $I \subset \mathbb{R}$ con $I \neq \emptyset$ y f una función real definida en I. Consideramos el problema (P) siguiente:

$$\text{hallar } \widehat{x} \in I, \text{ si existe tal que } f(\widehat{x}) = 0.$$

Supongamos que (P) tiene solución y que mediante la aplicación del algoritmo de búsqueda del cambio de signo se ha separado una única raíz $\widehat{x} \in [a,b] \subset I$. El método de las secantes no puede ser escrito en la forma

$$\begin{cases} x_0 \in [a,b] \text{ aproximación inicial,} \\ x_{n+1} = \varphi(x_n), \quad n = 0, 1, \ldots, \end{cases}$$

donde φ es una función de iteración sobre $[a,b]$.

El método regula-falsi es una combinación del método de las secantes y un análogo al método de bisección. Se le conoce también como método de las cuerdas o de la falsa posición. Sean $a_n, x_n \in [a,b]$ tales que $f(x_n) f(a_n) < 0 \quad n = 0, 1, \ldots$, donde a_n, x_n son determinados en cada paso de modo que solo en uno de los intervalos $[x_n, a_n]$ o $[a_n, x_n]$ está localizada la única raíz $\widehat{x}$ de $f(x) = 0$.

Para definir a_{n+1}, x_{n+1} consideramos la ecuación de la recta que pasa por los puntos $(x_n, f(x_n))$ y $(a_n, f(a_n))$:

$$L(x) = f(x_n) + \frac{f(x_n) - f(a_n)}{x_n - a_n}(x - x_n).$$

Luego

$$L(x) = 0 \Longleftrightarrow x = x_n - \frac{x_n - a_n}{f(x_n) - f(a_n)} f(x_n).$$

Sea $u_n = x$. Puesto que $f(x_n) \cdot f(a_n) < 0$ entonces $f(x_n) - f(a_n) \neq 0$ con lo cual

$$u_n = \frac{x_n - a_n}{f(x_n) - f(a_n)} f(x_n),$$

esta bien definido. Se tiene $a_n < u_n < x_n$ o $x_n < u_n < a_n$. En las figuras que se muestran a continuación se presentan estos dos casos.

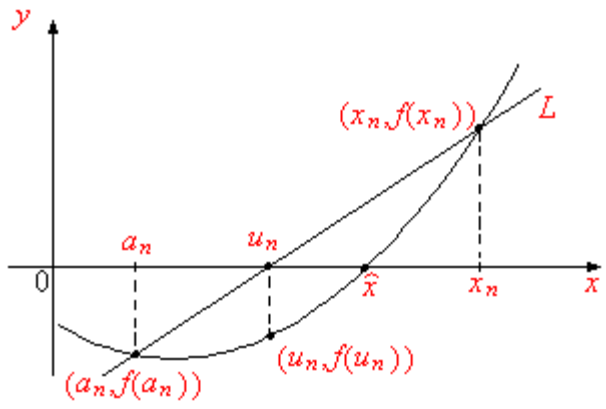


Figura 67

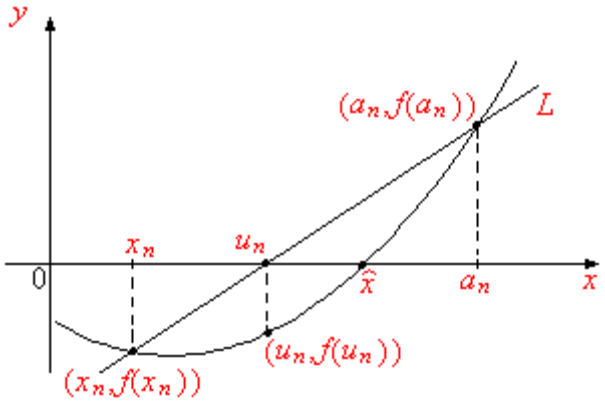


Figura 68

Supongamos que se tenga $a_n < u_n < x_n$. se tienen las tres situaciones siguientes.

i) Si $f(u_n) = 0$ entonces $\widehat{x} = u_n$ es la raíz buscada y el procedimiento concluye.

ii) Si $f(u_n) \cdot f(x_n) < 0$ entonces $a_{n+1} = u_n, x_{n+1} = x_n$.

iii) Si $f(u_n) \cdot f(x_n) > 0$ entonces $a_{n+1} = a_n, x_{n+1} = u_n$.

El procedimiento que acabamos de describir tampoco puede expresarse en la forma:

$$\begin{cases} x_0 \in [a,b] \text{ aproximación inicial,} \\ x_{n+1} = \varphi(x_n) \quad n = 0, 1, \ldots, \end{cases}$$

donde φ es una función de iteración definida en $[a,b]$. Sin embargo, si $x_0 \in [a,b]$ es el extremo de Fourier, esto es, supuesto que $f''(x_0)$ existe, $f(x_0) f''(x_0) > 0$; se define

$$\varphi(x) = x - \frac{x - x_0}{f(x) - f(x_0)} f(x) \quad x \in [a,b] \text{ con } f(x) f(x_0) < 0,$$

con lo cual φ es una función de iteración definida en un subconjunto E de $[a,b]$ en el que puede hacerse φ una aplicación contractiva, resultando que la sucesión (x_n) generada por

$$\begin{cases} x_0 \in [a,b] \text{ extremo de Fourier,} \\ x_{n+1} = \varphi(x_n) \quad n = 1, 2, \ldots, \end{cases}$$

con $x_1 \in [a,b]$ tal que $f(x_1) f(x_0) < 0$, converge a $\widehat{x}$ raíz de $f(x) = 0$.

Note que si x_0 es el extremo de Fourier, se tiene $f(x_0) f''(x_0) > 0$. En el primer caso se tiene que f es convexa y en el segundo f es cóncava.

Supongamos que $x_0 < x_1$ $(f(x_0) f(x_1) < 0)$. Si f es convexa se tiene $f(u_1) < 0$, luego $\widehat{x} \in [x_0, u_1]$ y $a_2 = x_0, x_2 = u_1$.

Si f es cóncava, se tiene $f(u_1) > 0$, luego $\widehat{x} \in [x_0, u_1]$ y $a_2 = x_0, x_2 = u_1$.

En la figura de la izquierda se muestra el primer caso (f es convexa) y en la de la derecha se muestra el

caso en que f es cóncava.

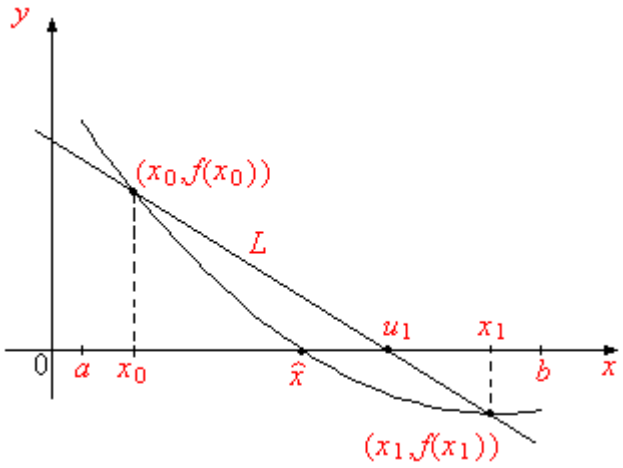


Figura 69

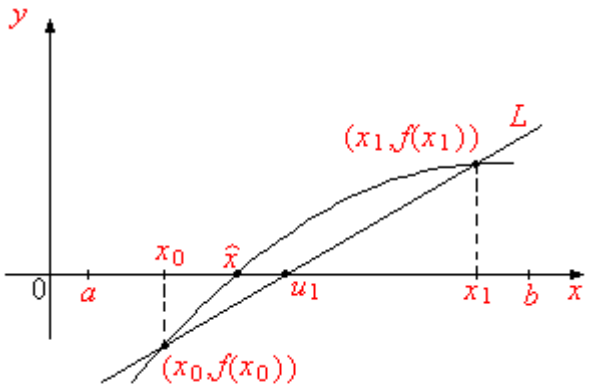


Figura 70

Algoritmo

Datos de Entrada: a, b extremos del intervalo $[a, b]$, precisión $\varepsilon > 0$, función f, número máximo de iteraciones $N_{\text{máx}}$.

Datos de Salida: n número de iteraciones, $\widetilde{x}$ aproximación de $\widehat{x}$, $f(\widetilde{x})$.

1. $x_0 = a$.

2.. $x_1 = b$.

3. Para $n = 2, \ldots, N_{\text{máx}}$.

4. $u = x_1 - \dfrac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1)$ // $f(x_0) \cdot f(x_1) < 0$.

5. $|u - x_1| < \varepsilon$ continuar en 9).

6. $f(u) = 0$ continuar en 9).

7. $f(x_0) f(u) < 0$ entonces $x_1 = u$. Continuar en 4).

8. $f(x_0) f(u) > 0$ entonces $x_0 = u, x_1 = b$. Continuar en 4).

9. Si $n < N_{\text{máx}}$, $\widetilde{x} = u$. Continuar en 11).

10. "$\widehat{x} = u$ en n iteraciones". Continuar en 11).

11. Fin.

Ejemplos

1. Hallar las raíces reales de la ecuación $x^3 - 3x^2 + 2x - 6 = 0$.

   Sea $f(x) = x^3 - 3x^2 + 2x - 6 = 0$. Con un paso $h = 0{,}5$, el algoritmo de búsqueda de cambio de signo, muestra que la ecuación $f(x) = 0$ tiene una sola raíz localizada en el intervalo $[2{,}5, 3{,}5]$.

   Sean $x_0 = 2,5$, $x_1 = 3,5$. Entonces $f(2{,}5) = -4{,}125$, $f(3{,}5) = 7{,}125$. Los resultados de la aplicación del algoritmo precedente se muestran en la tabla siguiente:

n	x_n
2	2,86667
3	3,09889
4	2,99272
5	2,99962
6	3,00607
7	2,99999
8	2,99999
9	3,0.

2. Encontrar las dos mas pequeñas raíces positivas de la ecuación $x^2 |\sin(x)| - 4 = 0$.

   Ponemos $f(x) = x^2 |\sin(x)| - 4 = 0 \quad x \in [0, \infty[$. Para $h = 0{,}2$, el algoritmo de búsqueda de cambio de signo muestra la existencia de una raíz localizada en el intervalo $[3{,}2, 3{,}6]$ y otra en $[6{,}2, 6{,}4]$. En la tabla de la izquierda se muestran los resultados de la aproximación de $\widehat{x}_1 \in [3{,}2, 3{,}6]$ y a la derecha, de $\widehat{x}_2 \in [6{,}2, 6{,}4]$.

$x_0 = 3{,}4, x_1 = 3{,}5$

n	x_n
2	3,47522
3	3,52215
4	3,47846
5	3,47851
6	3,47856
7	3,47851
8	3,47851

$x_0 = 6{,}2, x_1 = 6{,}4$

n	x_n
2	6,30204
3	6,35202
4	6,37650
5	6,38849
6	6,38155
7	6,38156
8	6,38157
9	6,38156

5.5. Convergencia. Convergencia acelerada

En esta sección estudiamos el orden de convergencia de los métodos que hemos tratado previamente, es decir que determinaremos la rapidez con la que la sucesión (x_n) generada por el método numérico utilizado converge a la raíz $\widehat{x}$ de la ecuación $f(x) = 0$.

Sea $I \subset \mathbb{R}$, $I \neq \emptyset$, f una función real definida en I. Suponemos que existe una única raíz $\widehat{x} \in [a, b] \subset I$ de $f(x) = 0$.

Definición 6 *Sea (x_n) una sucesión que converge a $\widehat{x}$ raíz de la ecuación $f(x)=0$.*

i. Ponemos $\varrho_n = x_n - \widehat{x}$ $n=0,1,\ldots$, ϱ_n se llama error de la n-ésima iteración.

ii. Si existen $p>0, c>0$ tales que $\lim\limits_{n\to\infty} \dfrac{|\varrho_{n+1}|}{|\varrho_n|^p} = c$, entonces (x_n) se dice convergente a $\widehat{x}$ de orden p, con un error asimptótico constante c.

iii. Un método numérico se dice convergente de orden p si la sucesión (x_n) generada por tal método converge a la raíz $\widehat{x}$ es de orden p.

Para $p=1$, (x_n) se dice convergente a $\widehat{x}$ de orden 1 o que el método converge linealmente. Para $p=2$, (x_n) se dice convergente a $\widehat{x}$ de orden 2 o que el método converge cuadráticamente.

Supongamos que la sucesión (x_n) es generada por una función de iteración $\varphi \in C^{p+1}([a,b])$ para $p \in \mathbb{Z}^+$, esto es, (x_n) está generada por el siguiente esquema numérico:

$$\begin{cases} x_0 \in [a,b] & \text{aproximación inicial,} \\ x_{n+1} = \varphi(x_n) & n=0,1,\ldots. \end{cases}$$

Para determinar el orden de convergencia podemos utilizar el polinomio de Taylor con resto entorno a la raíz $\widehat{x} \in [a,b]$.

Sea $x_n \in [a,b]$ $n=0,1,\ldots$ Supongamos que $\varphi^{(k)}(\widehat{x}) = 0$ para $k=1,2,\ldots,p-1$ pero $\varphi^{(p)}(\widehat{x}) \neq 0$. Entonces

$$\varphi(x_n) = \sum_{k=0}^{p} \frac{\varphi^{(k)}(\widehat{x})}{k!}(x_n - \widehat{x})^k + E_p(x_n, \widehat{x}),$$

con $E_p(x_n,\widehat{x}) \underset{n\to\infty}{\to} 0$.

Puesto que $\varphi(\widehat{x}) = \widehat{x}$, $\varphi(x_n) = x_{n+1}$ $n=0,1,\ldots$, y $\varphi^{(k)}(\widehat{x}) = 0$ para $k=1,\ldots,p-1$, se tiene

$$x_{n+1} = \varphi(\widehat{x}) + \frac{\varphi^{(p)}(\widehat{x})}{p!}(x_n - \widehat{x})^p + E_p(x_n, \widehat{x}),$$

de donde

$$\frac{x_{n+1} - \widehat{x}}{(x_n - \widehat{x})^p} = \frac{\varphi^{(p)}(\widehat{x})}{p!} + \frac{1}{(x_n - \widehat{x})^p} E_p(x_n, \widehat{x}).$$

De la formula integral del error $E_p(x_n,\widehat{x})$ y de la forma de Lagrange del resto, se tiene

$$E_p(x_n,\widehat{x}) = \frac{1}{p!}\int_{\widehat{x}}^{x_n} (x_n - t)^p \varphi^{(p+1)}(t)\,dt = \frac{1}{(p+1)!}\varphi^{(p+1)}(c_n)(x_n - \widehat{x})^{p+1},$$

para c_n en el intervalo cerrado que une x_n con $\widehat{x}$, $\lim\limits_{n\to\infty} c_n = \widehat{x}$. Resulta que

$$\frac{\varrho_{n+1}}{\varrho_n^p} = \frac{\varphi^{(p)}(\widehat{x})}{p!} + \frac{x_n - \widehat{x}}{(p+1)!}\varphi^{(p+1)}(c_n).$$

Tomando en cuenta que $\lim\limits_{n\to\infty} x_n = \widehat{x}$ y de la continuidad de $\varphi^{(p+1)}$, se sigue que

$$\lim_{n\to\infty} \frac{\varrho_{n+1}}{\varrho_n^p} = \frac{\varphi^{(p)}(\widehat{x})}{p!} + \frac{1}{(p+1)!}\lim_{n\to\infty}(x_n - \widehat{x})\,\varphi^{(p+1)}\left(\lim_{n\to\infty} c_n\right) = \frac{\varphi^{(p)}(\widehat{x})}{p!}.$$

Así, si $\varphi \in C^{p+1}([a,b])$, la sucesión (x_n) generada por φ convergente a $\widehat{x}$ es de orden p.

Si $\varphi \in C^p([a,b])$ y si se supone que $E_p(x,\widehat{x}) = 0\left(|x-\widehat{x}|^{p+1}\right)$, esto es, $E_p(x,\widehat{x}) \underset{x\to\widehat{x}}{\to} 0$, se tiene

$$\frac{\varrho_{n+1}}{\varrho_n^p} = \frac{\varphi^{(p)}(\widehat{x})}{p!} + 0(|x_n - \widehat{x}|),$$

con lo cual

$$\lim_{n\to\infty}\frac{\varrho_{n+1}}{\varrho_n^p}=\frac{\varphi^{(p)}(\widehat{x})}{p!}.$$

Nota: Existen muchas sucesiones (x_n) que son generadas por esquemas numéricos que no se expresan mediante funciones de iteración φ. Un ejemplo de esta clase de sucesiones son las que provienen del método de las secantes.

A continuación estudiamos el orden de convergencia de los métodos numéricos que hemos tratado.

1. **Método del punto fijo**

   En este método y en los que siguen, suponemos que $\widehat{x}\in[a,b]$ es la única raíz de la ecuación $f(x)=0$.

   Supóngase que $\varphi\in C^1([a,b])$ y (x_n) la sucesión definida por $\begin{cases} x_0\in[a,b] & \text{aproximación inicial,}\\ x_{n+1}=\varphi(x_n) & n=0,1,\ldots.\end{cases}$
   Supongamos además que existe $k, 0\le k<1$ tal que $|\varphi'(x)|\le k \quad \forall x\in[a,b]$. Entonces

   $$\varrho_{n+1}=x_{n+1}-\widehat{x}=\varphi(x_n)-\varphi(\widehat{x})=\varphi'(c_n)(x_n-\widehat{x})=\varphi'(c_n)\varrho_n,$$

   con c_n entre x_n y $\widehat{x}$. Luego

   $$\frac{\varrho_{n+1}}{\varrho_n}=\varphi'(c_n).$$

   Como $\varphi(x_n)\to\varphi(\widehat{x})=\widehat{x}$, se tiene $\lim_{n\to\infty}\varphi'(c_n)=\varphi'(\widehat{x})$. Resulta

   $$\lim_{n\to\infty}\frac{\varrho_{n+1}}{\varrho_n}=\lim_{n\to\infty}\varphi'(c_n)=\varphi'(\widehat{x}).$$

   Si $\varphi'(\widehat{x})\neq 0$, el método de punto fijo converge linealmente.

   Si $\varphi'(\widehat{x})=0$, se puede tener un orden de convergencia más elevado. Así, si $\varphi\in C^2([a,b])$ tal que $\varphi''(\widehat{x})\neq 0$, entonces

   $$\varphi(x)=\varphi(\widehat{x})+\varphi'(\widehat{x})\cdot(x-\widehat{x})+\frac{1}{2!}\varphi''(c_x)(x-\widehat{x})^2 \quad x\in[a,b],$$

   con c_x entre x y $\widehat{x}$.

   Para $x=x_n$, se tiene $\varphi(x_n)=x_{n+1}$, $\varphi(\widehat{x})=\widehat{x}$, $\varphi'(\widehat{x})=0$, luego

   $$\begin{aligned} x_{n+1} &= \widehat{x}+\varphi'(\widehat{x})(x_n-\widehat{x})+\frac{1}{2!}\varphi''(c_n)(x_n-\widehat{x})^2,\\ \frac{\varrho_{n+1}}{\varrho_n} &= \frac{1}{2!}\varphi''(c_n),\end{aligned}$$

   con lo cual

   $$\lim_{n\to\infty}\frac{\varrho_{n+1}}{\varrho_n}=\frac{1}{2!}\varphi''(\widehat{x}).$$

   En este caso el método numérico converge cuadráticamente.

2. **Método de punto fijo modificado**

   En este método la función de iteración φ está definida por $\varphi(x)=x-\widehat{m}f(x) \quad x\in[a,b]$ con $f\in C([a,b])$, donde $\widehat{m}$ está definido por

   $$\begin{aligned} \widehat{m} &= \begin{cases} -\frac{1}{\beta}, \text{ si } \frac{f(x)-f(y)}{x-y}<0 \quad \forall x,y\in[a,b], \ x\neq y,\\ \frac{1}{\beta}, \text{ si } \frac{f(x)-f(y)}{x-y} \quad \forall x,y\in[a,b], \ x\neq y,\end{cases}\\ \beta &= \sup_{\substack{x,y\in[a,b]\\ x\neq y}}\left|\frac{f(x)-f(y)}{x-y}\right|.\end{aligned}$$

φ es contractiva, y el esquema numérico está dado por

$$\begin{cases} x_0 \in [a,b] \text{ aproximación inicial,} \\ x_{n+1} = \varphi(x_n) \quad n = 0,1,\ldots \end{cases}$$

Se tiene

$$\begin{aligned} x_{n+1} - \widehat{x} &= \varphi(x_n) - \varphi(\widehat{x}) = x_n - \widehat{m} f(x_n) - (\widehat{x} - \widehat{m} f(\widehat{x})) = x_n - \widehat{x} - \widehat{m}(f(x_n) - f(\widehat{x})), \\ \varrho_{n+1} &= \varrho_n - \widehat{m}[f(x_n) - f(\widehat{x})] \quad n = 1,2,\ldots \end{aligned}$$

Supongamos $x_n \neq \widehat{x} \quad n = 0,1,\ldots$, entonces

$$\begin{aligned} \frac{\varrho_{n+1}}{\varrho_n} &= 1 - \widehat{m}\frac{f(x_n) - f(\widehat{x})}{\varrho_n} = 1 - \widehat{m}\frac{f(x_n) - f(\widehat{x})}{x_n - \widehat{x}} \\ \left|\frac{\varrho_{n+1}}{\varrho_n}\right| &= \left|1 - \widehat{m}\frac{f(x_n) - f(\widehat{x})}{x_n - \widehat{x}}\right| \leq 1 - \frac{\alpha}{\beta} < 1, \end{aligned}$$

donde $0 < \alpha = \underset{\substack{x,y\in[a,b] \\ x\neq y}}{\text{Inf}} \left|\frac{f(x)-f(y)}{x-y}\right|$. En consecuencia,

$$\lim_{n\to\infty} \left|\frac{\varrho_{n+1}}{\varrho_n}\right| = 1 - \frac{\alpha}{\beta},$$

con lo cual el método de punto fijo modificado es de orden 1, o sea la sucesión (x_n) converge a $\widehat{x}$ linealmente.

3. **Método de Newton-Raphson**

Supongamos que $f \in C^3([a,b])$. La función de iteración del método de Newton-Raphson está dada por

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \quad f'(x) \neq 0 \quad x \in [a,b].$$

Entonces

$$\varphi'(x) = 1 - \frac{(f'(x))^2 - f(x) f''(x)}{[f'(x)]^2} = \frac{f(x) f''(x)}{[f'(x)]^2} \quad x \in [a,b].$$

Para $x = \widehat{x}$, se tiene $f(\widehat{x}) = 0$ y

$$\varphi'(\widehat{x}) = \frac{f(\widehat{x}) f''(\widehat{x})}{[f'(\widehat{x})]^2} = 0.$$

Calculemos la derivada segunda de φ (esta existe ya que $f \in C^3$), obtenemos

$$\varphi''(x) = \frac{f(x) f'(x) f'''(x) + [f'(x)]^2 f''(x) - 2f(x)[f''(x)]^2}{[f'(x)]^4} \quad x \in [a,b].$$

Para $x = \widehat{x}$, resulta

$$\varphi''(\widehat{x}) = \frac{f(\widehat{x}) f'(\widehat{x}) f'''(\widehat{x}) + [f'(\widehat{x})]^2 f''(\widehat{x}) - 2f(\widehat{x})[f''(\widehat{x})]^2}{[f'(\widehat{x})]^4} = \frac{f''(\widehat{x})}{[f'(\widehat{x})]^2} \neq 0.$$

Luego,

$$\lim_{n\to\infty} \frac{\varrho_{n+1}}{\varrho_n} = \frac{f''(\widehat{x})}{[f'(\widehat{x})]^2}, \quad f'(\widehat{x}) \neq 0, \quad f''(\widehat{x}) \neq 0.$$

El método de Newton es de segundo orden.

Si $f \in C^2([a,b])$ y $f''(\widehat{x}) \neq 0$, se prueba que

$$\varrho_{n+1} = x_{n+1} - \widehat{x} = \frac{1}{2}\frac{f''(c_n)}{f'(x_n)}(x_n - \widehat{x})^2 = \frac{1}{2}\frac{f''(c_n)}{f'(c_n)}\varrho_n^2,$$

con c_n entre x_n y $\widehat{x}$. Resulta

$$\frac{\varrho_{n+1}}{\varrho_n^2} = \frac{1}{2}\frac{f''(c_n)}{f'(x_n)}, \quad n = 1,2,\ldots$$

4. **Método de la secantes**

Sean $f \in C^2([a,b])$, $x_0, x_1 \in [a,b]$ tales que $f(x_0) f(x_1) < 0$. En el método de las secantes se construye una sucesión (x_n) definida por

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \quad n = 2, 3, \ldots$$

Supongamos que para todo n, $x_n \neq \widehat{x}$. Entonces

$$x_{n+1} - \widehat{x} = x_n - \widehat{x} - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) = (x_n - \widehat{x}) \left[1 - \frac{f(x_n)}{x_n - \widehat{x}} \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \right]. \quad (1)$$

Definimos

$$L(x) = f(x_n) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} (x - x_n) \quad x \in [a,b].$$

Se tiene, $L(x_n) = f(x_n)$ y $L(x_{n-1}) = f(x_{n-1})$, es decir que L es un interpolante de f. Luego,

$$f(x) = L(x) + \varepsilon(x) \quad x \in \left[\widetilde{a}, \widetilde{b} \right],$$

con $\widetilde{a} = \text{mín}\{x_{n-1}, x_n\}$, $\widetilde{b} = \text{máx}\{x_{n-1}, x_n\}$ y $\varepsilon(x)$ el error de interpolación en x (error de interpolación polinomial de Lagrange). Para $x = \widehat{x}$, se tiene

$$0 = f(\widehat{x}) = L(\widehat{x}) + \varepsilon(\widehat{x}).$$

Se prueba que $\varepsilon(x) = \frac{1}{2}(x - x_{n-1})(x - x_n) f''(\eta) \quad x \in \left[\widetilde{a}, \widetilde{b} \right] \quad \eta \in \left[\widetilde{a}, \widetilde{b} \right]$. Así,

$$f(x) = f(x_n) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} (x - x_n) + \frac{1}{2}(x - x_{n-1}) f''(\eta),$$

y para $x = \widehat{x}$, se tiene

$$0 = f(\widehat{x}) = f(x_n) + \frac{f(x_n) f(x_{n-1})}{x_n - x_{n-1}} (\widehat{x} - x_n) + \frac{1}{2}(\widehat{x} - x_{n-1})(\widehat{x} - x_n) f''(\eta),$$

de donde

$$\frac{f(x_n)}{x_n - \widehat{x}} \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} - 1 = \frac{1}{2} \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f''(\eta). \quad (2)$$

Además,

$$f(x_n) - f(x_{n-1}) = f'(t_n)(x_n - x_{n-1}),$$

con t_n entre x_n y x_{n-1}. Luego

$$x_{n+1} - \widehat{x} = -\frac{1}{2}(x_n - \widehat{x})(x_{n-1} - \widehat{x}) \frac{f''(\eta)}{f'(t_n)},$$

consecuentemente

$$\varrho_{n+1} = -\frac{1}{2} \varrho_n \varrho_{n-1} \frac{f''(\eta)}{f'(t_n)}.$$

Por hipótesis $f \in C^2([a,b])$, entonces

$$0 < \frac{1}{2} \lim_{n \to \infty} \left| \frac{f''(\eta)}{f'(t_n)} \right| = \frac{1}{2} \left| \frac{f''(\eta_{\widehat{x}})}{f'(\widehat{x})} \right| \leq c,$$

se sigue que

$$\left| \varrho_{n+1} \right| \leq c \left| \varrho_n \right| \left| \varrho_{n-1} \right|.$$

Sea $E_n = c|\varrho_n|$. Multplicando por c en la desigualdad precedente, resulta

$$E_{n+1} \leq E_n E_{n-1} \quad n = 2, 3, \ldots$$

Supongamos que

$$\begin{aligned} E_0 &\leq \lambda, \quad E_1 \leq \lambda \text{ con } 0 < \lambda < 1. \text{ Luego} \\ E_2 &\leq E_1 E_0 \leq \lambda^2, \\ E_3 &\leq E_2 E_1 \leq \lambda^2 \cdot \lambda = \lambda^3, \\ E_4 &\leq E_3 E_2 \leq \lambda^3 \cdot \lambda^2 = \lambda^5, \\ E_5 &\leq E_4 E_3 \leq \lambda^5 \cdot \lambda^3 = \lambda^8, \\ &\vdots \\ E_k &\leq \lambda^{ak}, \end{aligned}$$

donde (a_k) es la sucesión de Fibonacci definida por

$$a_0 = a_1 = 1, \quad a_{k+1} = a_k + a_{k-1} \quad k \geq 1. \tag{3}$$

La ecuación (3) es una ecuación en diferencias homogénea de segundo orden,

$$a_{k+1} - a_k - a_{k-1} = 0 \quad k = 1, 2, \ldots \tag{4}$$

cuya ecuación característica es $\alpha^2 - \alpha - 1 = 0$ y cuya solución es

$$\alpha_1 = \frac{1 - \sqrt{5}}{2}, \quad \alpha_2 = \frac{1 + \sqrt{5}}{2}.$$

Las soluciones de (4) son: $c_1 + c_2, \quad c_1\alpha_1 + c_2\alpha_2, \quad c_1\alpha_1^2 + c_2\alpha_2^2, \ldots$.Como $c_1 + c_2 = a_1$ y $c_1\alpha_1 + c_2\alpha_2 = a_2$, se tiene

$$\begin{cases} c_1 + c_2 = 1, \\ c_1 \frac{1-\sqrt{5}}{2} + c_2 \frac{1+\sqrt{5}}{2} = 1, \end{cases}$$

de donde

$$c_1 = \frac{1 + \sqrt{5}}{2\sqrt{5}}, \quad c_2 - \frac{1 - \sqrt{5}}{2\sqrt{5}}$$

$$a_k = c_1\alpha_1^{k-1} + c_2\alpha_2^{k-1} = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k}{\sqrt{5}} = \frac{\alpha_2^k - \alpha_1^k}{\sqrt{5}},$$

que es conocida como la fórmula de Binet. Consecuentemente

$$E_k \leq \lambda^{ak} = \lambda^{\frac{1}{\sqrt{5}}\alpha_2^k} \lambda^{-\frac{1}{\sqrt{5}}\alpha_1^k},$$

y como $|\alpha_1| = \left|\frac{1-\sqrt{5}}{2}\right| < 1$, se sigue que

$$\lambda^{-\frac{1}{\sqrt{5}}\alpha_1^k} \leq \beta \quad k = 2, 3, \ldots$$

de donde

$$E_k \leq \beta\lambda^{\frac{1}{\sqrt{5}}\alpha_2^k} \quad k = 2, 3, \ldots$$

Si

$$\lim_{n\to\infty} \frac{|x_{n+1} - \widehat{x}|}{|x_n - \widehat{x}|^p} = c \Leftrightarrow \lim_{n\to\infty} \frac{\left|\varrho_{n+1}\right|}{|\varrho_n|^p} = c,$$

se sigue que para $\varepsilon > 0$, $\widetilde{c} = c + \varepsilon$ y

$$\begin{aligned} \left|\varrho_{n+1}\right| &\leq \widetilde{c}\,|\varrho_n|^p = \widetilde{c}\,|\varrho_n| \left|\varrho_{n-1}\right|, \\ E_{n+1} &\leq \widetilde{c}E_n^p = \widetilde{c}E_n E_{n-1} \leq c\left(\lambda^{\frac{1}{\sqrt{5}}\alpha_2}\right)^{\alpha_2^k}, \end{aligned}$$

con lo cual $p = \frac{1+\sqrt{5}}{2} \simeq 1{,}618$.

El método de las secantes es de orden $p \simeq 1{,}618$.

El método de las secantes es uno de los métodos de interpolación para calcular las raíces de ecuaciones. En cada paso del método de las secantes requiera la evaluación adicional de la función f. Dos pasos del método de las secantes es algo más costoso que un paso del método de Newton. Además, dos pasos del método de las secantes conducen a un método de orden $p \simeq 1{,}618\ldots$ que hace que converja localmente mas rápidamente que el método de Newton.

La ventaja de este método radica en que no requiere del cálculo de la derivada de la función f, pero requiere, como se ha dicho, de una evaluación adecional de la función f en cada paso.

5. **Método regula-falsi**

Este método es también uno de los métodos de interpolación para calcular las raíces de la ecuación $f(x) = 0$. En el siguiente teorema se considera una variante de este método.

Teorema 9 *Sean $f \in C^2([a,b])$ y $\widehat{x} \in [a,b]$ la única raíz de la ecuación $f(x) = 0$. Supongamos que $x_1 < y$ tales que $f(x_1)\,,\ f(y) > 0$ y $f''(x) \geq 0 \quad \forall x \in [a,b]$. Entonces*
i. La sucesión (x_n) generada por la función de iteración φ definida por

$$\varphi(x) = x - \frac{y-x}{f(y)-f(x)} f(x) \qquad x \in [x_1, y[$$

converge a $\widehat{x}$.
ii. (x_n) converge linealmente.

Demostración.

i. La ecuación de la recta L que pasa por $(x_1, f(x_1))\,,\quad (y, f(y))$ esta dada por

$$L(x) = f(x_1) + \frac{f(y) - f(x_1)}{y - x_1}(x - x_1) \qquad x \in [x_1, y]\,.$$

Se tiene $L(x_1) = f(x_1)\,,\quad L(y) = f(y)$ con lo cual L es una interpolante de f, más exactamente, tal como en el caso del método de las secantes, L es un caso particular de polinomio de interpolación de Lagrange (L es un polinomio de grado 1). De la fórmula de interpolación con error $f(x) = L(x) + \varepsilon(x) \quad x \in [x_1, y]\,,$ con $\varepsilon(x)$ el error de interpolación en el punto x, definimos

$$F(x) = f(x) - L(x) + k(x - x_1)(x - y) \qquad x \in [x_1, y]\,,$$

donde k es una constante a determinarse por la condición $f(\overline{x}) = 0$ para $\overline{x} \in [x_1, y]\,.$

Puesto que$L(x_1) = f(x_1)\,,\quad L(y) = f(y)$, se sigue que $F(x_1) = F(y) = F(\overline{x}) = 0$, entonces F tiene tres raíces en $[x_1, y]$. Además F es derivable, y

$$F'(x) = f'(x) - L'(x) + k(2x - x_1 - y) \qquad x \in [x_1, y]\,.$$

Por el teorema de Rolle, existen $\eta_1, \eta_2 \in [x_1, y]$ tales que $\eta_1 < \eta_2$ y

$$F'(\eta_1) = F'(\eta_2) = 0.$$

Por otro lado,

$$F''(x) = f''(x) - L''(x) + 2k \quad x \in [x_1, y]$$

y como $L''(x) = 0$, se tiene

$$F''(x) = f''(x) + 2k \quad x \in [x_1, y]\,.$$

Nuevamente, por el teorema de Rolle, existe $\eta \in [\eta_1, \eta_2]$ tal que

$$F''(\eta) = 0,$$

entonces

$$0 = F''(\eta) = f''(\eta) + 2k \Rightarrow k = -\frac{f''(\eta)}{2}.$$

Consecuentemente

$$F(x) = f(x) - L(x) - \frac{1}{2}(\overline{x} - x_1)(x - y) f''(\eta) \quad x \in [x_1, y]$$

y para $x = \overline{x} \in [x_1, y]$ se tiene

$$0 = F(\overline{x}) = f(\overline{x}) - L(\overline{x}) - \frac{1}{2}(\overline{x} - x)(\overline{x} - y) f''(\eta),$$

$$f(\overline{x}) - L(\overline{x}) = \frac{1}{2}(\overline{x} - x_1)(\overline{x} - y) f''(\eta).$$

Por hipótesis $f''(\eta) \geq 0, \quad x_1 \leq \overline{x} \leq y \Rightarrow \overline{x} - x_1 \geq 0, \quad \overline{x} - y \leq 0$, entonces

$$\frac{1}{2}(\overline{x} - x_1)(\overline{x} - y) f''(\eta) \leq 0,$$

y de esta desigualdad se deduce que

$$f(\overline{x}) - L(\overline{x}) \leq 0 \Leftrightarrow f(\overline{x}) \leq L(\overline{x}).$$

En particular, si $L(u_1) = 0$ (La recta L corta al eje X, véase la figura) entonces

$$u_1 = x_1 - \frac{y - x_1}{f(y) - f(x_1)} f(x_1)$$

y por la desigualdad previa

$$f(u_1) \leq L(u_1) = 0 \Rightarrow f(u_1) \leq 0.$$

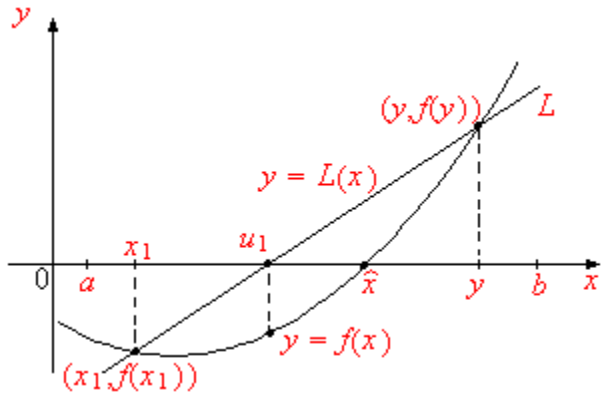


Figura 71

Si $f(u_1) = 0$, u_1 es la raíz $\widehat{x}$ de la ecuación $f(x) = 0$ y el proceso concluye. Supongamos $f(u_1) < 0$. Como por hipótesis $f(y) > 0$, $f(u_1) f(y) < 0$, consecuentemente $\widehat{x} \in [u_1, y]$. Ponemos $x_2 = u_1$.

El proceso anterior se repite con el intervalo $[x_2, y]$. De este modo se construye una sucesión (x_n) creciente y acotada por y. Luego (x_n) es convegente a $\widehat{c}$, eso es,

$$\widehat{c} = \lim_{n \to \infty} x_n = \operatorname*{Sup}_{n \in \mathbb{Z}^+} x_n.$$

Por hipótesis f es continua en $[a, b]$ y por construcción de (x_n), $x_n \leq \widehat{c}$, $n = 1, 2, \ldots$, $f(x_n) < 0$, $n = 1, 2, \ldots$. Entonces

$$f(\widehat{c}) = f\left(\lim_{n \to \infty} x_n\right) = \lim_{n \to \infty} f(x_n) \leq 0.$$

Si $f(\widehat{c}) < 0$ y como $f(y) > 0$, $f(\widehat{c}) f(y) < 0$ luego $\widehat{x} \in [\widehat{c}, y]$ con $\widehat{c} < \widehat{x}$. Existe $u \in]\widehat{c}, y[$ tal que $f(u) \leq 0$ con lo que u es un término de la sucesión (x_n) y en consecuencia $u \leq \widehat{c}$ lo que constituye una contradicción con $u \in]\widehat{c}, y[$. Así, $f(\widehat{c}) = 0$ o sea $\widehat{c} = \widehat{x}$ la raíz de $f(x) = 0$.

Observe que

$$x_{n+1} = \varphi(x_n) = x_n - \frac{y - x_n}{f(y) - f(x_n)} f(x_n) \quad n = 1, 2, \ldots$$

ii. Puesto que $f \in C^2([a,b])$, entonces $\varphi \in C^2([a,b])$. Tenemos

$$\varphi'(x) = 1 - \frac{[f(y) - f(x)][(y-x) f'(x) - f(x)] + (y-x) f(x) f'(x)}{[f(y) - f(x)]^2}.$$

Como $f(\widehat{x}) = 0$, se tiene

$$\begin{aligned} \varphi'(\widehat{x}) &= 1 - \frac{[f(y) - f(\widehat{x})][(y-\widehat{x}) f'(\widehat{x}) - f(\widehat{x})] + (y-\widehat{x}) f(\widehat{x}) f'(\widehat{x})}{[f(y) - f(\widehat{x})]^2} \\ &= 1 - \frac{f(y)(y-\widehat{x}) f'(\widehat{x})}{[f(y)]^2} = 1 - \frac{y-\widehat{x}}{f(y)} f'(\widehat{x}) = 1 - \frac{y-\widehat{x}}{f(y) - f(\widehat{x})} f'(\widehat{x}). \end{aligned}$$

Por el teorema del valor medio, existe $\eta \in [\widehat{x}, y]$ tal que

$$f(y) - f(\widehat{x}) = f'(\eta)(y - \widehat{x}) \Rightarrow \frac{f(y) - f(\widehat{x})}{y - \widehat{x}} = f'(\eta),$$

luego

$$\varphi'(\widehat{x}) = 1 - \frac{f'(\widehat{x})}{f'(\eta)}.$$

Puesto que $f''(x) > 0$, f' es creciente y $f' > 0$. Entonces $f'(\widehat{x}) < f'(\eta)$ de donde $\frac{f'(\widehat{x})}{f'(\eta)} < 1$ y $0 < \varphi(\widehat{x}) < 1$, o sea φ es contractiva en $[x_1, y]$. Por lo tanto

$$\lim_{n\to\infty} \frac{\varrho_{n+1}}{\varrho_n} = \lim_{n\to\infty} \frac{x_{n+1} - \widehat{x}}{x_n - \widehat{x}} = \varphi'(\widehat{x}).$$

El método regula-falsi converge linealmente. ■

Análisis del error para métodos iterativos

Hemos definido el orden de un método iterativo con el número real $p > 0$ tal que

$$\lim_{n\to\infty} \frac{\left|\varrho_{n+1}\right|}{|\varrho_n|^p} = \lim_{n\to\infty} \frac{|x_{n+1} - \widehat{x}|}{|x_n - \widehat{x}|^p} = c > 0,$$

o sea la sucesión (x_n) generada por el método iterativo converge a $\widehat{x}$ de orden p, con una constante de error asimptótico $c > 0$.

Método	**Orden**
Iteración de punto fijo modificado	1
Newton - Raphson para raíces simples	2
Newton modificado	1
Secantes	1,618
Regula - falsi	1

Sean $I \subset \mathbb{R}$, $I \neq \emptyset$ y f una funicón real definida en I. Suponemos que existe una única raíz $\widehat{x}$ de $f(x) = 0$ localizada en un intervalo $[a,b] \subset I$. Supongamos que para el cálculo aproximado de $\widehat{x}$ se utilizan dos métodos iterativos cuyas sucesiones generadas por dichos métodos son (x_n) y (t_n) respectivamente. Pongamos $\varrho_n = x_n - \widehat{x}$, $E_n = t_n - \widehat{x}$ $\quad n = 1, 2, \ldots$ los errores cometidos en cada iteración por cada algoritmo. Para simplificar, supongamos que el primer método es de primer orden y el segundo método es de segundo orden. Entonces

$$\lim_{n\to\infty} \left|\frac{\varrho_{n+1}}{\varrho_n}\right| = c_1, \lim_{n\to\infty} \frac{|E_{n+1}|}{E_n^2} = c_2 > 0,$$

con $0 < c_1 < 1$.

Para n suficientemente grande,

$$\left|\frac{\varrho_{n+1}}{\varrho_n}\right| \simeq c_1 \Longrightarrow \left|\varrho_{n+1}\right| \simeq c_1 \left|\varrho_n\right|,$$

$$\frac{|E_{n+1}|}{E_n^2} \simeq c_2 \Longrightarrow |E_{n+1}| \simeq c_2 |E_n|^2.$$

Para el primer método, se tiene

$$|\varrho_n| \simeq c_1 \left|\varrho_{n-1}\right| \simeq c_1^2 \left|\varrho_{n-2}\right| \simeq \ldots \simeq c_1^n |\varrho_0|, \tag{1}$$

y para el segundo, obtenemos

$$\begin{aligned} |E_n| &\simeq c_2 |E_{n-1}|^2 \simeq c_2 \left(c_2 E_{n-2}^2\right)^2 = c_2^3 |E_{n-2}|^4 \simeq c_2^3 \left(c_2 |E_{n-2}|^2\right)^4 \\ &= c_2^7 |E_{n-2}|^8 \simeq \ldots \simeq c_2^{2^n-1} |E_0|^{2^n}. \end{aligned} \tag{2}$$

Con la finalidad de comparar la rapidez de convergencia de estos métodos, supongamos que $0 < \lambda < 1$ con $\lambda = |x_0 - \widehat{x}|$, $\lambda = |\varrho_0|$, $\lambda = |E_0|$ y sea $\varepsilon = 5 \times 10^{-m}$ con $m \in \mathbb{Z}^+$ la precisión con la que es aproximada $\widehat{x}$ para los dos métodos. Determinemos el número mínimo de iteraciones para el cual la raíz $\widehat{x}$ es aproximada con la precisión ε.

Para el primer método, tenemos:

$$|\varrho_n| \simeq c_1^n |\varrho_0| = c_1^n \lambda \leq 5 \times 10^{-m},$$

de donde

$$c_1^n \lambda \leq 5 \times 10^{-m} \Rightarrow c_1^n \leq \frac{5 \times 10^{-m}}{\lambda},$$

$$\begin{aligned} n \ln c_1 &\leq \ln 5 - m \ln 10 - \ln \lambda \\ n &\geq \frac{-\ln 5 + m \ln 10 + \ln \lambda}{-\ln c_1} = \frac{\ln\left(\frac{\lambda}{5}\right) + m \ln 10}{-\ln c_1}. \end{aligned}$$

Sea

$$N_0 = \left[-\frac{\ln\left(\frac{\lambda}{5}\right) + m \ln 10}{\ln c_1}\right] + 1. \tag{3}$$

Para el segundo método, de (2) se sigue que

$$|E_n| \simeq c_2^{2^n-1} |E_0|^{2^n} = c_2^{2^n-1} \lambda^{2^n} \leq 5 \times 10^{-m},$$

de donde

$$\begin{aligned} (c_2\lambda)^{2^n} &\leq 5 \times 10^{-m}, \\ 2^n \ln(c_2\lambda) &\leq \ln(5c_2) - m \ln 10. \end{aligned} \tag{4}$$

Al menor número entero positivo n que verifica (4) designémosle con N_1, es decir que N_1 es tal que

$$2^{N_1} \ln(c_2\lambda) \leq \ln(5c_2) - m \ln 10.$$

Exhibamos mediante un ejemplo que $N_1 < N_0$.

Ejemplo

Considerar la ecuación $-x + \ln^2(x) = 0$.

Sea $f(x) = -x + \ln^2(x) \quad x > 0$, y $\varepsilon = 5 \times 10^{-8}$. La ecuación $f(x) = 0$ tiene una raíz localizada en el intervalo $[0{,}4, 0{,}5]$. Aproximemos $\widehat{x}$ con los métodos regula falsi y Newton-Raphson. Se tiene los siguientes resultados:

Método regula-falsi

n	x_n
0	0,4
1	0,6
2	0,5129111
3	0,49794886
⋮	⋮
11	0,4948664
12	0,4948664

Método de Newton-Raphson

n	t_n
0	0,4
1	0,4787588
2	0,4943978
3	0,4948660
4	0,4948664

Tenemos $\lambda = |\varrho_0| = |E_0| = |0{,}4 - 0{,}4948664| \simeq 0{,}1$.

Para el método regula falsi,

$$c_1 = \frac{|x_3 - \widehat{x}|}{|x_2 - \widehat{x}|} = \frac{|0{,}4979486 - 0{,}4948664|}{|0{,}5129111 - 0{,}4948664|} \simeq 0{,}17.$$

Entonces

$$N_0 = \left[-\frac{\ln\left(\frac{\lambda}{5}\right) + m \ln 10}{\ln c_1} \right] + 1 = \left[-\frac{\ln\left(\frac{0{,}1}{5}\right) + 8 \ln 10}{\ln 0{,}17} \right] + 1 = [12{,}60341292] + 1 = 13.$$

Esto significa que a partir de $N_0 = 13$ se logra la precisión deseada. Si en el método regula - falsi se considera un número máximo de iteraciones $N_{\text{máx}}$ tal que $N_{\text{máx}} < N_0$, no se logrará la precisión deseada. Debemos tener $N_{\text{máx}} \geq N_0$ para lograr la precisión requerida.

Para el método de Newton - Raphson, tenemos

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \Rightarrow \frac{1}{2}\varphi''(\widehat{x}) = -\frac{[-2 + 3 \ln \widehat{x}] \ln \widehat{x}}{\widehat{x}\left(-\widehat{x} + 2 \ln \widehat{x}\right)^2},$$

con lo cual

$$c_2 = \lim_{n \to \infty} \frac{|x_{n+1} - \widehat{x}|}{|x_n - \widehat{x}|^2} = \frac{1}{2}\varphi''(\widehat{x}) \simeq 1{,}4723.$$

Entonces

$$\begin{aligned}
2^n \ln(c_2 \lambda) &\leq \ln(5c_2) - m \ln 10, \\
2^n \ln(1{,}4723 \times 0{,}1) &\leq \ln(5 \times 1{,}4723) - 8 \ln 10, \\
-1{,}915759289 \times 2^n &\leq -16{,}42441703, \\
n \ln 2 &\geq \ln\left(\frac{16{,}42441703}{1{,}915759289}\right) = \ln 8{,}573319792, \\
n &\geq \frac{\ln(8{,}573319792)}{\ln 2} \simeq 3{,}099.
\end{aligned}$$

Se tiene $N_1 = 4$. La precisión deseada se logra a partir de $N_1 = 4$, o lo que es lo mismo $|x_n - \widehat{x}| \leq 5 \times 10^{-8}$ para $n \geq 4$.

Si $N_{\text{máx}}$ denota el número máximo de iteraciones. Para $N_{\text{máx}} \geq N_1$.

El método de Newton modificado es un algoritmo de primer orden. Para $N_0 = 13$ se logra la precisión deseada. Sea $x_0 = 0{,}4$. La función de iteración φ del método de Newton modificado etá definida por

$$\varphi(x) = x - \frac{1}{f'(x_0)} f(x) \quad x \in [0{,}4, 0{,}5].$$

Como $f(x) = -x + \ln^2(x) \Rightarrow f'(x) = -1 + \frac{2}{x}\ln(x)$, entonces

$$f'(0{,}4) = -1 + \frac{2}{0{,}4}\ln(0{,}4) = -5{,}58145366,$$

$$\varphi(x) = x + 0{,}1792 f(x) = x + 0{,}1792\left(-x + \ln^2(x)\right).$$

En la tabla siguiente se muestran los resultados de la aplicación de este método.

n	x_n
0	0,4
1	0,478774296
2	0,49018866
3	0,493437619
4	0,494424147
5	0,497289704
6	0,494823648
⋮	⋮
11	0,494866289
12	0,4948663754
13	0,4948664023

Se tiene $f(x_{13}) = 0{,}000000047 \leq 5 \times 10^{-8} = \varepsilon$.

Convergencia acelerada

Sean $I \subset \mathbb{R}$, $I \neq \emptyset$ y f una función real definida en I. Suponemos que existe una raíz $\widehat{x}$ de $f(x) = 0$ separada en el intervalo $[a, b] \subset I$ y sea (x_n) una sucesión convergente a $\widehat{x}$. Los métodos de aceleración de la convergencia transforman la sucesión (x_n) en sucesiones (t_n) que convergen mas rapidamente que $[x_n]$. En general , los métodos de aceleración de la convergencia utilizan métodos de orden 1 en los que intervienen únicamente la función f y no su derivada. Los más conocidos son Δ^2 de Aitken y el método de Steffensen.

Método Δ^2 de Aitken

Sea (x_n) una sucesión que converge a $\widehat{x}$ raíz de la ecuación $f(x) = 0$. Suponemos que (x_n) converge linealmente. Entonces existe $0 < c < 1$ tal que

$$\lim_{n\to\infty}\left|\frac{x_{n+1} - \widehat{x}}{x_n - \widehat{x}}\right| = c,$$

y sea $E_n = x_n - \widehat{x}$ para $n = 1, 2, \ldots$. Se tiene $x_{n+1} - \widehat{x} = c(x_n - \widehat{x})$. Entonces c y $\widehat{x}$ pueden ser determinados utilizando las ecuaciones

$$\begin{cases} x_{n+1} - \widehat{x} = c(x_n - \widehat{x}), \\ x_{n+2} - \widehat{x} = c(x_{n+1} - \widehat{x}), \end{cases}$$

cuyas soluciones son

$$\begin{aligned} c &= \frac{x_{n+2} - x_{n+1}}{x_{n+1} - x_n}, \\ \widehat{x} &= x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} \quad \text{con} \quad x_{n+2} - 2x_{n+1} + x_n \neq 0. \end{aligned}$$

Definimos

$$t_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}. \tag{1}$$

El método de Δ^2 de Aitken se basa en la suposición de que la sucesión (t_n) converge más rápidamente que (x_n).

Teorema 10 *Sea (x_n) una sucesión convergente a $\widehat{x}$ y $x_n \neq \widehat{x} \quad \forall n \in \mathbb{Z}^+$. Supongamos que existen una constante k, $0 < k < 1$ y una sucesión $[\delta_n]$ tales que*

$$\begin{aligned} x_{n+1} - \widehat{x} &= (k + \delta_n)(x_n - \widehat{x}), \\ \lim_{n\to\infty} \delta_n &= 0. \end{aligned}$$

Entonces (t_n) dada por (1) etá bien definida y $\lim_{n\to\infty} \dfrac{t_n - \widehat{x}}{x_n - \widehat{x}} = 0.$

Demostración. Sea $E_n = x_n - \widehat{x}$. Entonces $E_{n+1} = (k + \delta_n) E_n$. Luego

$$\begin{aligned} x_{n+2} - 2x_{n+1} + x_n &= x_{n+2} - \widehat{x} - 2(x_{n+1} - \widehat{x}) - x_n - \widehat{x} = E_{n+2} - 2E_{n+1} + E_n \\ &= (k + \delta_{n+1}) E_{n+1} - 2(k + \delta_n) E_n + E_n \\ &= (k + \delta_n + 1)(k + \delta_n) E_n - 2(k + \delta_n) E_n + E_n \\ &= E_n \left[(k-1)^2 + (\delta_n + \delta_{n+1}) k + \delta_n (\delta_{n+1} - 2) \right], \end{aligned}$$

además

$$x_{n+1} - x_n = x_{n+1} - \widehat{x} - (x_n - \widehat{x}) = E_{n+1} - E_n = E_n [k - 1 + \delta_n].$$

Puesto que $x_{n+2} - 2x_{n+1} + x_n \neq 0$, $E_n = x_n - \widehat{x} \neq 0$, $0 < k < 1$ y por hipótesis $\lim_{n\to\infty} \delta_n = 0$, Entonces

$$(\delta_n + \delta_{n+1}) k + \delta_n (\delta_{n+1} - 2) \underset{n\to\infty}{\to} 0,$$

de donde para n suficientemente grande

$$\frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} = \frac{E_n^2 (k - 1 + \delta_n)^2}{E_n \left((k-1)^2 + \mu_n \right)} = E_n \frac{(k-1) + \delta_n^2}{(k-1)^2 + \mu_n} \underset{n\to\infty}{\to} 0,$$

con $\mu_n = (\delta_n + \delta_{n+1}) k + \delta (\delta_{n+1} - 2) \underset{n\to\infty}{\to} 0.$

Por lo tanto

$$\lim_{n\to\infty} t_n = \lim_{n\to\infty} \left(x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} \right) = \lim_{n\to\infty} x_n - \lim_{n\to\infty} \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} = \widehat{x}.$$

La sucesión (t_n) es convergente a $\widehat{x}$. Así, (t_n) está bien definida.

Por otro lado

$$\begin{aligned} t_n - \widehat{x} &= x_n - \widehat{x} - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} = E_n - \frac{E_n^2 (k - 1 + \delta_n)^2}{E_n \left((k-1)^2 + \mu_n \right)} \\ &= E_n \left(1 - \frac{(k - 1 + \delta_n)^2}{(k-1)^2 + \mu_n} \right), \end{aligned}$$

de donde

$$\begin{aligned} \frac{t_n - \widehat{x}}{x_n - \widehat{x}} &= \frac{t_n - \widehat{x}}{E_n} = 1 - \frac{(k - 1 + \delta_n)^2}{(k-1)^2 + \mu_n}, \\ \lim_{n\to\infty} \frac{t_n - \widehat{x}}{x_n - \widehat{x}} &= 1 - \lim_{n\to\infty} \frac{(k - 1 + \delta_n)^2}{(k-1)^2 + \mu_n} = 1 - \frac{(k-1)^2}{(k-1)^2} = 0. \end{aligned}$$

■

Sea $\varepsilon > 0$. Puesto que $\lim_{n\to\infty} \dfrac{t_n - \widehat{x}}{x_n - \widehat{x}} = 0$, existe $n_0 \in \mathbb{Z}^+$ tal que $\forall n \geq n_0 \Rightarrow |t_n - \widehat{x}| < |x_n - \widehat{x}| \, \varepsilon$.

La última desigualdad muestra que la sucesión (t_n) converge mas rapidamente que $[x_n]$.

Ejemplo

Hallar $x \in \mathbb{R}$ tal que $e^x - 0{,}5x + 1 = 0$. Pongamos $f(x) = e^x - 0{,}5x + 1$. Entonces

$$f(x) = 0 \Longleftrightarrow e^x = -0{,}5x - 1.$$

El método gráfico muestra que la ecuación $f(x) = 0$ tiene una única raíz $\widehat{x}$ localizada o separada en el intervalo $[-2{,}5, -2{,}0]$. Aproximemos la raíz $\widehat{x}$ con el método de punto fijo modificado y luego aceleramos la convergencia con el método de Δ^2 de Aitken.

La función de iteración del método de punto fijo modificado está definida por

$$\varphi(x) = x - mf(x) \quad x \in [-2{,}5, -2{,}0],$$

donde $m = \frac{b-a}{f(b)-f(a)} = \frac{-2+2{,}5}{f(-2)-f(-2{,}5)} = 1{,}6488$,

$$\varphi(x) = x - 1{,}6488\,(1 + 0{,}5x + e^x) \quad x \in [-2{,}5, -2{,}0].$$

La tabla que se muestran a continuación se exhiben los resultado de la aplicación del método de punto fijo modificado.

n	x_n
0	$-2{,}0$
1	$-2{,}223140815$
2	$-2{,}217696668$
3	$-2{,}217715177$
4	$-2{,}217715105$
5	$-2{,}217715106$
6	$-2{,}21771506$

A continuación se muestran los términos de (t_n) con el método de Δ^2 de Aitken.

$$\begin{aligned} t_1 &= x_1 - \frac{(x_2 - x_1)^2}{x_3 - 2x_2 - x_1} = -2{,}217715144, \\ t_2 &= x_2 - \frac{(x_3 - x_2)^2}{x_4 - 2x_3 - x_2} = -2{,}217715105, \\ t_3 &= x_3 - \frac{(x_4 - x_3)^2}{x_5 - 2x_4 - x_3} = -2{,}217715106. \end{aligned}$$

Valor aproximado de $\widehat{x}$: $-2{,}217715106$.

Método de Steffensen

En el método de Δ^2 de Aitken, para inicializar el proceso se requieren de x_1, x_2, x_3. Con esta información se calcula t_1. A continuación se calcula x_4 y con los precedentes x_2, x_3 se calcula t_2, luego se calcula x_5 y con este se obtiene t_3, así sucesivamente.

El método de Steffensen toma ventaja de la construcción de la sucesión (t_n) cuando la sucesión (x_n) está generada por una función de iteración φ. El método de Steffensen es recomendado para métodos de orden 1.

Pongamos $y_n = \varphi(x_n)$, $z_n = \varphi(y_n)$. Se tiene

$$t_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} = x_n - \frac{(y_n - x_n)^2}{z_n - 2y_n + x_n} = x_n - \frac{(\varphi(x_n) - x_n)^2}{\varphi(\varphi(x_n)) - 2\varphi(x_n) + x_n} \quad n = 0, 1, \ldots \quad (1)$$

La sucesión (t_n) construida mediante el esquema (1) se conoce como método de Steffensen.

El esquema nuérico dado por (1) conduce a una nueva función de iteración ψ definida

$$\psi(x) = x - \frac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x} = \frac{x\varphi(\varphi(x)) - \varphi^2(x)}{\varphi(\varphi(x)) - 2\varphi(x) + x} \quad x \in [a, b],$$

donde $\widehat{x} \in [a,b]$ es la raíz de la ecuación $f(x) = 0$.

El esquema numérico (1) se escribe entonces

$$\begin{cases} x_0 \in [a,b] \text{ aproxiamación inicial,} \\ x_{n+1} = \psi(x_n) \quad n = 0,1,\ldots. \end{cases}$$

Las funciones de iteración φ y ψ, por lo general, tienen el mismo punto fijo, más precisamente, tenemos el siguiente teorema.

Teorema 11 *Sea $\widehat{x} \in [a,b]$. Entonces $\psi(\widehat{x}) = \widehat{x} \Rightarrow \varphi(\widehat{x}) = \widehat{x}$. Recíprocamente, si $\varphi(\widehat{x}) = \widehat{x}$ y $\varphi'(\widehat{x}) \neq 1 \Rightarrow \psi(\widehat{x}) = \widehat{x}$.*

Demostración. Por la definción de la función de iteración ψ, se tiene

$$\psi(x) = x - \frac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x} \quad x \in [a,b],$$

de donde

$$[\psi(x) - x][\varphi(\varphi(x)) - 2\varphi(x) + x] = (\varphi(x) - x)^2 \quad x \in [a,b].$$

Entonces, si $\psi(\widehat{x}) = \widehat{x}$ se tiene

$$[\psi(\widehat{x}) - \widehat{x}][\varphi(\varphi(\widehat{x})) - 2\varphi(\widehat{x}) + \widehat{x}] = (\varphi(\widehat{x}) - \widehat{x})^2$$

$$0 = (\varphi(\widehat{x}) - \widehat{x})^2 \Rightarrow \varphi(\widehat{x}) = \widehat{x}.$$

Recíprocamente, supongamos que $\varphi(\widehat{x}) = \widehat{x}$ y $\varphi'(\widehat{x}) \neq 1$. Entonces

$$\begin{aligned} \psi(\widehat{x}) &= \lim_{x\to\widehat{x}} \psi(x) = \lim_{x\to\widehat{x}} x - \frac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x} = \widehat{x} - \lim_{x\to\widehat{x}} \frac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x} \\ &= \widehat{x} - \lim_{x\to\widehat{x}} \frac{2(\varphi(x) - x)(\varphi'(x) - 1)}{\varphi'(\varphi(x))\varphi'(x) - 2\varphi'(x) + 1} = \widehat{x} - \frac{2(\varphi(\widehat{x}) - \widehat{x})(\varphi'(\widehat{x}) - 1)}{\varphi'(\varphi(\widehat{x}))\varphi'(\widehat{x}) - 2\varphi'(\widehat{x}) + 1} \\ &= \widehat{x}. \end{aligned}$$

ya que $\varphi(\widehat{x}) = \widehat{x}$.

Note que

$$\varphi'(\varphi(\widehat{x}))\varphi'(\widehat{x}) - 2\varphi'(\widehat{x}) + 1 = \varphi'(\widehat{x})\varphi'(\widehat{x}) - 2\varphi'(\widehat{x}) + 1 = \left(\varphi'(\widehat{x}) - 1\right)^2,$$

y por hipótesis $\varphi'(\widehat{x}) \neq 1, (\varphi'(\widehat{x}) - 1)^2 \neq 0$. ■

Algoritmo

Datos de Entrada: a, b extremos del intervalo $[a,b]$. precisión $\varepsilon > 0$, $N_{\text{máx}}$ número máximo de iteraciones, función f.

Datos de Salida: $\widehat{x}$ raíz, $y = f(\widehat{x})$, n número de iteraciones.

1. Leer $x_0 \in [a,b]$ y poner $x = x_0$.
2. Para $n = 1,\ldots,N_{\text{máx}}$
3. $y_1 = \varphi(x)$.
4. $y_2 = \varphi(y_1)$.
5. $t = x - \dfrac{(y_1 - x_0)^2}{y_2 - 2y_1 + x}$.
6. Si $|t - x| < \varepsilon$. Continuar en 8).
7. Si $|t - x| > \varepsilon, x = t$. Continuar en 3).

8. Si $n < N_{\text{máx}}$, imprimir $\widehat{x} = t, y = f(t)$. Continuar en 10).

9. Si $n > N_{\text{máx}}$, imprimir $\widehat{x} = t, y = f(t)$.

10. Fin.

Ejemplo

Sea $f(x) = \sin(x)\cosh\left(\dfrac{x}{1+\sqrt{x}}\right) - 1$. Hallar los ceros de f para $x \in [0,10]$.

La aplicación del agoritmo de búsqueda del cambio de signo en $[0,10]$ con un paso $h = 0{,}4$ muestra que f tiene cuatro ceros localizados en los intervalos $[0{,}8, 1{,}2]$, $[2, 2{,}4]$, $[6{,}4, 6{,}8]$, $[9{,}2, 9{,}6]$.

Calculemos las dos primeras raíces con el método regula-falsi (método de orden 1). En la tabla de la izquierda se muestran los resultados de la aplicación de este método para la aproximación de $\widehat{x}_1 \in [0{,}8, 1{,}2]$ raíz de $f(x) = 0$ y en la de la derecha para $\widehat{x}_2 \in [2, 2{,}4]$.

n	x_n
1	0,8
2	1,0838480
3	1,0698168
4	1,0564792
5	1,0681479
6	1,0681382
7	1,0681286
8	1,0681382

n	x_n
1	2,4
2	2,39602478
3	2,397285172
4	2,397287862
5	2,397287868

Apliquemos el método de Steffensen: $\begin{cases} \psi(x) = x - \dfrac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x}, \\ t_{n+1} = \psi(t_n) \quad n = 1, 2, \ldots. \end{cases}$

Cálculo de $\widehat{x}_1 \in [0{,}8, 1{,}2]$. Se tienen los siguientes resultados de la plicación del método de Steffensen:

$t_1 = 0{,}8,$

$t_2 = t_1 - \dfrac{(\varphi(t_1) - t_1)^2}{\varphi(\varphi(t_1)) - 2\varphi(t_1) + t_1} = 0{,}8 - \dfrac{(\varphi(0{,}8) - 0{,}8)^2}{\varphi(\varphi(0{,}8)) - 2\varphi(0{,}8) + 0{,}8} = 1{,}0678892,$

$t_3 = \psi(t_2) = 1{,}068137133,$

$t_4 = \psi(t_3) = 1{,}068138455,$

$t_5 = \psi(t_4) = 1{,}068138463.$

Cálculo de $\widehat{x}_2 \in [2, 2{,}4]$:

$t_1 = 2, \quad y = 2{,}4,$

$t_2 = \psi(t_1) = 2{,}397289196,$

$t_3 = \psi(t_2) = 2{,}397287868.$

Nota: Para método de orden 1, el método de Steffensen es de orden 2.

5.6. Raíces de multiplicidad

Sean $I \subset \mathbb{R}$ con $I \neq \emptyset$ y f una función real definida en I. Consideramos el problema (P) siguiente:

$$\text{hallar } \widehat{x} \in I, \text{ si existe, solución de } f(x) = 0. \tag{P}$$

Definición 7 *Se dice que $\widehat{x} \in I$ es una raíz de multiplicidad $m \geq 2$ de la ecuación $f(x) = 0$ si existe una función g definida en I tal que $f(x) = (x - \widehat{x})^m g(x)$ $x \in I$, con $g(\widehat{x}) \neq 0$.*

Ejemplos

1. Sea f la función definida por $f(x) = x(x-2)^2$ $x \in \mathbb{R}$. Entonces $\widehat{x} = 2$ es una raíz de multiplicidad 2. Note que $g(x) = x$ y $g(2) = 2$. Además $\widehat{x} = 0$ es una raíz real simple de $f(x) = 0$.
2. Sea f la función real dada por $f(x) = \left(x^2 + 5\right)(x+1)^3$ $x \in \mathbb{R}$. Entonces $\widehat{x} = -1$ es la única raíz real de multiplicidad 3. Se tiene $g(x) = x^2 + 5$ con $g(-1) = 6$.

La aplicación del algoritmo de búsqueda de cambio de signo, en general, no da resultados positivos si la ecuación $f(x) = 0$ tiene raíces de multiplicidad $m \geq 2$. En el ejemplo 1), la aplicación de este algoritmo no separa la raíz $\widehat{x} = 2$, mientras que en el ejemplo 2), si lo separa pero no detecta que sea de multiplicidad 3.

En el ejemplo 1). La raíz $\widehat{x} = 2$ es de multiplicidad $m = 2$ (par). y en el ejemplo 2) la raíz $\widehat{x} = -1$ es de multiplicidad $m = 3$ (impar).

Supongamos que $\widehat{x}$ es una raíz de multiplicidad $m \geq 2$ de la ecuación $f(x) = 0$. Entonces, existe una función g definida en I tal que

$$f(x) = (x - \widehat{x})^m g(x) \quad x \in I, g(\widehat{x}) \neq 0.$$

Si $f \in C^2([a,b])$, donde $\widehat{x} \in [a,b] \subset I$ y $m \geq 2$, entonces

$$f'(x) = m(x-\widehat{x})^{m-1} g(x) + (x-\widehat{x})^m g'(x) = (x-\widehat{x})^{m-1}\left[mg(x) + (x-\widehat{x})g'(x)\right] \quad x \in [a,b].$$

Se tiene $f'(\widehat{x}) = 0$. Ponemos

$$v(x) = mg(x) + (x-\widehat{x})g'(x) \quad x \in [a,b].$$

Resulta $v(\widehat{x}) = mg(\widehat{x}) \neq 0$.

Puesto que $f \in C^2([a,b])$, entonces las funciones g y v son continuas en $[a,b]$ y como $g(\widehat{x}) \neq 0$, $v(\widehat{x}) \neq 0$, existe $r > 0$ tal que $g(x) \neq 0$, $v(x) \neq 0$ $\forall x \in [\widehat{x} - r, \widehat{x} + r] \subset [a,b]$.

Definimos

$$\begin{aligned} u(x) &= \frac{f(x)}{f'(x)} = \frac{(x-\widehat{x})^m g(x)}{(x-\widehat{x})^{m-1}(mg(x) + (x-\widehat{x})g'(x))} \\ &= (x-\widehat{x})\frac{g(x)}{v(x)} \qquad x \in [\widehat{x}-r, \widehat{x}+r] \setminus \{\widehat{x}\}, \end{aligned}$$

con $\frac{g(x)}{v(x)} \neq 0$ $\forall x \in [\widehat{x} - r, \widehat{x} + r]$.

Como $\lim\limits_{x \to \widehat{x}} u(x) = 0$, la función u tiene una discontinuidad evitable. Definimos

$$\widetilde{u}(x) = \begin{cases} 0, & \text{si } x = \widehat{x}, \\ u(x) & \text{si } x \in [\widehat{x}-r, \widehat{x}+r] \setminus \{\widehat{x}\}. \end{cases}$$

Entonces, $\widetilde{u}(\widehat{x}) = 0 = f(\widehat{x})$, es decir que $\widetilde{u}(x) = 0$ tiene la misma raíz que $f(x)$ en el entorno $[\widehat{x} - r, \widehat{x} + r]$. La raíz $\widehat{x}$ es una raíz simple de $\widetilde{u}(\widehat{x}) = 0$.

El método deNewton-Raphson es efectivo para raíces simples. Podemos aplicar este método a la función u.

La función de iteración φ está dada por

$$\varphi(x) = x - \frac{u(x)}{u'(x)} \quad x \in [\widehat{x}-r, \widehat{x}+r] \setminus \{\widehat{x}\}.$$

Como $u(x) = \dfrac{f(x)}{f'(x)}$, se sigue que

$$u'(x) = \frac{(f'(x))^2 + f(x) f''(x)}{(f'(x))^2},$$

con lo cual

$$\varphi(x) = x - \frac{u(x)}{u'(x)} = x - \frac{f(x) f'(x)}{[f'(x)]^2 - f(x) f''(x)} \quad x \in [\widehat{x} - r, \widehat{x} + r] \setminus \{\widehat{x}\}.$$

El esquema numérico para la aproximación de la raíz $\widehat{x}$ es el siguiente:

$$\begin{cases} x_0 \in [\widehat{x} - r, \widehat{x} + r] \setminus \{\widehat{x}\} & \text{aproximación inicial,} \\ x_{n+1} = \varphi(x_n) \quad n = 0, 1, \ldots \end{cases}$$

Por otro lado, supongamos que $\forall x \in [a, b]$, $g(x) > 0$ y $f'(x) \neq 0 \ \ \forall x \in [a, b] \setminus \{\widehat{x}\}$. Definimos

$$w(x) = (f(x))^{\frac{1}{m}} \qquad x \in [a, b].$$

Como $f(x) = (x - \widehat{x})^m g(x) \quad x \in [a, b]$ con $g(\widehat{x}) \neq 0$, se tiene

i. Si m es par, $(x - \widehat{x})^m \geq 0 \ \ \forall x \in [a, b]$ y siendo $g(x) > 0$, resulta que $f(x) \geq 0 \quad \forall x \in [a, b]$. Luego

$$w(x) = (f(x))^{\frac{1}{m}} = [(x - \widehat{x})^m g(x)]^{\frac{1}{m}} = |x - \widehat{x}| (g(x))^{\frac{1}{m}} \qquad x \in [a, b].$$

ii. Si m es impar,

$$w(x) = (x - \widehat{x}) (g(x))^{\frac{1}{m}} \qquad x \in [a, b].$$

De i) y ii) se sigue que $w(\widehat{x}) = 0$, esto es, la función w tiene $\widehat{x}$ como cero simple en $[a, b]$. Apliquemos el método de Newton. La función de iteración Φ está definida por

$$\Phi(x) = x - \frac{u(x)}{u'(x)} = x - \frac{(f(x))^{\frac{1}{m}}}{\frac{1}{m}[f(x)]^{\frac{1}{m}-1} f'(x)} = x - m\frac{f(x)}{f'(x)} \qquad x \in [a, b] \setminus \{\widehat{x}\}.$$

El método de Newton para ceros de multiplicidad $m \geq 2$ con m dado, se expresa como

$$\Phi(x) = x - m\frac{f(x)}{f'(x)} \quad x \in [a, b], x \neq \widehat{x}.$$

El esquema numérico es el siguiente:

$$\begin{cases} x_0 \in [a, b] \setminus \{\widehat{x}\} & \text{aproximación inicial,} \\ x_{n+1} = \Phi(x_n) & n = 0, 1, \ldots. \end{cases}$$

Nota: Se pueden implementar fácilmente los otros métodos que han sido estudiados anteriormente.

Ejemplos

1. Hallar las raíces reales positivas de la ecuación: $x^4 - 8{,}6x^3 - 35{,}51x^2 + 464{,}4x - 998{,}46 = 0$.

   La búsqueda del cambio de signo en el intervalo $[0, \infty[$ muestra que la función f asociada a la ecuación dada tiene una raíz en el intervalo $[7, 8]$ y posiblemente una raíz de multiplicidad en un entorno de $x = 4$ (se presume por la observación de los valores de $f(x)$ en un entorno de $x = 4$). Por otro lado, el estudio de la función f muestra que la ecuación $f(x) = 0$ tiene una raíz en el intervalo $]-\infty, 0]$. Ponemos

   $$\begin{aligned} f(x) &= x^4 - 8{,}6x^3 - 35{,}51x^2 + 464{,}4x - 998{,}46, \\ u(x) &= \frac{f(x)}{f'(x)} \quad f'(x) \neq 0. \end{aligned}$$

Entonces

$$\begin{aligned} u(4) &= \frac{f(4)}{f'(4)} = \frac{-3{,}44}{23{,}52} < 0, \\ u(5) &= \frac{f(5)}{f'(5)} = \frac{-14{,}23}{-35{,}7} > 0. \end{aligned}$$

La función u tiene una raíz en el intervalo $[4,5]$, esta raíz de la ecuación $f(x)=0$. De este modo confirmamos que $f(x)=0$ tiene una raíz de multiplicidad 2.

Nota: La búsqueda del cambio de signo se aplicó a la función u en el intervalo $[0,7]$.

Para la aproximación de la raíz $\widehat{x} \in [4,5]$ se utilizan las dos funciones de iteración φ y Φ.

i. Con la función de iteración $\varphi(x) = x - \dfrac{f(x) f'(x)}{[f'(x)]^2 - f(x) f''(x)}$ se tiene

n	x_n
0	4
1	4,3081
2	4,3081
3	4,3000

La raíz $\widehat{x} = 4{,}3$ de $f(x)=0$ es de multiplicidad $m=2$.

ii. Como $m=2$, con la función de iteración

$$\Phi(x) = x - 2\frac{f(x)}{f'(x)},$$

se tiene

n	x_n
0	4
1	4,29082
2	4,29998
3	4,29998
4	4,3000

iii. Con $x_0 = 7$, la raíz simple localizada en el intervalo $[7,8]$, se aproxima con el método de Newton. Con 5 iteraciones se tiene $\widehat{x}_1 = 7{,}34847$.

2. Hallar las raíces de la ecuación: $x^2 - 2xe^{-x} + e^{-2x} = 0$. Sea $f(x) = x^2 - 2xe^{-x} + e^{-2x} = (x - e^x)^2$ $x \in \mathbb{R}$.

Entonces,

$$f(x) = 0 \Longleftrightarrow (x-e^x)^2 = 0 \Longleftrightarrow x - e^x = 0 \Longleftrightarrow x = e^{-x}.$$

La ecuación $f(x) = 0$ tiene una raíz en el intervalo $[0,1]$. Calculemos $f'(x)$. Se tiene $f'(x) = 2(x-e^x)(1+e^{-x})$. Definimos

$$u(x) = \frac{f(x)}{f'(x)} \quad f'(x) \neq 0.$$

Con la ayuda de la función u, separamos la raíz de la ecuación $f(x)=0$. Tenemos $u(0) < 0$, $u(1) > 0$, luego existe $\widehat{x} \in]0,1[$ tal que $f(\widehat{x}) = 0$. Para la aproximación de $\widehat{x}$ utilizamos el método de Newton.

i. Sea

$$\varphi(x) = x - \frac{u(x)}{u'(x)} = x - \frac{f(x) f'(x)}{[f'(x)]^2 - f(x) f''(x)}.$$

En la tabla siguiente se muestran los resultados de la aplicación del esquema numérico: $\begin{cases} x_0 \in [0,1] \\ x_{n+1} = \varphi(x_n) \quad n = 0,1,\ldots \end{cases}$

n	x_n
0	0
1	0,666667
2	0,568769
3	0,567144
4	0,567144

$\widehat{x} \simeq 0{,}567144$ con una precisión $\varepsilon = 10^{-6}$.

ii. Sea

$$\Phi(x) = x - 2\frac{f(x)}{f'(x)}.$$

Los resultados de la aplicación del esquema numérico

$$\begin{cases} x_0 \in [0,1] \\ x_{n+1} = \Phi(x_n) \quad n = 0,1,\ldots \end{cases}$$

se muestran a continuación.

n	x_n
0	0,0
1	0,5
2	0,566311
3	0,567143
4	0,567143

Valor aproximado de $\widehat{x}$ con una precisión $\varepsilon = 10^{-6}$: 0,567143.

Teorema 12

i) Si $f \in C^1([a,b])$ y si la ecuación $f(x) = 0$ tiene un cero simple $\widehat{x} \in [a,b]$, entonces $f'(\widehat{x}) \neq 0$.

ii) Si $f \in C^m([a,b])$ para $m \geq 2$ y si la ecuación $f(x) = 0$ tiene un cero de multiplicidad m, entonces $f^{(m)}(\widehat{x}) \neq 0$.

Demostración.

i) Si la ecuación $f(x) = 0$ tiene una raíz simple en $x = \widehat{x}$, entones $f(\widehat{x}) = 0$ y existe una función g tal que

$$f(x) = (x - \widehat{x})\, g(x) \text{ con } g(\widehat{x}) \neq 0.$$

Por hipótesis $f \in C^1([a,b])$, entonces $g \in C^1([a,b])$ y

$$f'(x) = (x - \widehat{x})\, g'(x) + g(x),$$

de donde

$$f'(x) = g(\widehat{x}) \neq 0.$$

ii) Sean $m \geq 2$ y $f \in C^m([a,b])$. Si la ecuación $f(x) = 0$ tiene una raíz de multiplicidad m, existe una función $g \in C^m([a,b])$ tal que

$$f(x) = (x - \widehat{x})^m\, g(x) \text{ con } g(\widehat{x}) \neq 0.$$

Resulta que

$$f^m(x) = \sum_{k=0}^{m} \binom{m}{k} g^{(k)}(x) \left[(x - \widehat{x})^m\right]^{(m-k)},$$

de donde $\binom{m}{k} = \frac{m!}{k!(m-k)!}$, $[(x-\widehat{x})^m]^{(m-k)}$ denota la derivada de $(x-\widehat{x})^m$ de orden $m-k$. Entonces

$$f^{(m)}(\widehat{x}) = m!g(\widehat{x}) \neq 0.$$

Observe que si $m \geq 2$, $g(x) = \frac{f(x)}{(x-\widehat{x})^m} \quad x \neq \widehat{x}$, y

$$\lim_{x\to\widehat{x}} g(x) = g(\widehat{x}).$$

■

Ejemplo

Considere la función f definida por $f(x) = e^x - \frac{1}{2}x^2 - x - 1$. Entonces, la ecuación $f(x) = 0$ tiene una raíz de multiplicidad 3 en $x = 0$. Sea

$$g(x) = \frac{e^x - \frac{1}{2}x^2 - x - 1}{x^3} \quad x \neq 0.$$

Aplicando la regla de L'Hôpital, obtenemos

$$\lim_{x\to 0} g(x) = \lim_{x\to 0} \frac{e^x - \frac{1}{2}x^2 - x - 1}{x^3} = \lim_{x\to 0} \frac{e^x - x - 1}{3x^2} = \lim_{x\to 0} \frac{e^x - 1}{6x} = \lim_{x\to 0} \frac{e^x}{6} = \frac{1}{6}.$$

Además,

$$f(x) = (x-0)^3 \frac{e^x - \frac{1}{2}x^2 - x - 1}{x^3} = x^3 g(x).$$

Se tiene

$$\begin{aligned} f'''(x) &= \sum_{k=0}^{3} \binom{3}{k} g^{(k)}(x) \left[x^3\right](3-k) = 6g(x) + 18xg'(x) + 9x^2 g''(x) + x^3 g'''(x), \\ f'''(0) &= 6g(0) = 3!g(0) = 3!\frac{1}{6} = 1. \end{aligned}$$

Nota: El método de Newton para ceros de multiplicidad $m \geq 2$ es de orden 1; o sea

$$\lim_{n\to\infty} \left| \frac{x_{n+1} - \widehat{x}}{x_n - \widehat{x}} \right| = c \quad \text{con } 0 < c < 1.$$

5.7. Raíces reales de polinomios

Sean $n \in \mathbb{N}$, $a_k \in \mathbb{R} \quad k = 0, 1, \ldots, n$ con $a_n \neq 0$. Una función real P de forma

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k \quad x \in \mathbb{R},$$

se llama polinomio de grado n con coeficientes reales a_k. El polinomio nulo P_0 definido por $P_0(x) = 0 \quad \forall x \in \mathbb{R}$ no se le asigna grado alguno.

En lo sucesivo consideraremos polinomios reales de grado $n \geq 1$ con coeficientes en $\mathbb{R}$ y diremos simplemente P polinomio de grado n sobrentendiéndose que $n \geq 1$ y sus coeficientes son reales.

Definición 8 *Sea P un polinomio de grado n. La ecuación*

$$P(x) = 0 \Leftrightarrow a_0 + a_1 x + \cdots + a_n x^n = 0,$$

se llama ecuación algebraica.

Teorema 13 ***(teorema fundamental del álgebra)***
Sea P un polinomio de grado n. Entones, la ecuación $P(x) = 0$ tiene exactamente n raíces reales o complejas incluidas las de multiplicidad.

Demostración. La demostración de este teorema está fuera del alcance de estas notas, se encuentra en, por ejemplo, Churchill, páginas 145-146. ■

Ejemplos

1. Sean $x_1, \ldots, x_n \in \mathbb{R}$ con $x_i \neq x_j, \ \ i, j = 1, \ldots, n$. El polinomio $P(x) = \prod_{j=1}^{n} (x - x_j)$ tiene exactamente n raíces reales.
2. El polinomio $P(x) = (x+5)(x-2)^3 \left(x^2+1\right)^2$ tiene ocho raíces: $\widehat{x}_1 = -5$ es una raíz real simple, $\widehat{x}_2 = 2$ es una raíz real de multiplicidad 2 y $\widehat{x}_3 = i, \ \widehat{x_4} = -i$ son raíces complejas de multiplicidad 2.

Sea P un polinomio de grado n. Si la ecuación $P(x) = 0$ tiene una raíz compleja z_1, existe $z_2 \in \mathbb{C}$ tal que $z_2 = \overline{z_1}$ raíz de la ecuación $P(x) = 0$, donde $\overline{z_1}$ denota el número complejo conjugado de z_1.

En el estudio de una ecuación algebraica dada $P(x) = 0$ interesa los siguientes aspectos:

i. Determinar los intervalos en los que se encuentran localizadas loa raíces reales positivas y las raíces reales negativas.

ii. El número de raíces reales simples y de multiplicidad y como calcularlas.

iii. Los discos en los cuales se localizan las raíces complejas.

iv. El número de raíces complejas simples y de multiplicidad y como calcularlas.

En esta sección daremos especial atención a los aspectos i) y ii) Los aspectos iii) y iv) no serán abordados.

Supongamos que $P(x) = \sum_{k=0}^{n} a_k x^k$ y $a_n > 0$. Sea $x > 0$. Entonces

$$P(x) = x^n \left(\frac{a_0}{x^n} + \frac{a_1}{x^{n-1}} + \cdots + a_n \right).$$

Si n es par, resulta que $P(x) \underset{|x| \to +\infty}{\longrightarrow} \infty$ Si n es impar, se tiene $P(x) \underset{x \to -\infty}{\longrightarrow} -\infty$, $P(x) \underset{x \to +\infty}{\longrightarrow} \infty$. Por lo tanto, si n es impar, la ecuación $P(x) = 0$ tiene al menos una raíz real. En la siguiente tabla se muestra el número de raíces reales y complejas según el grado del polinomio P.

grad (P)	Número de raíces reales	Número de raíces complejas
1	1	0
2	0, 2	2, 0
3	1, 3	2, 0
4	0, 2, 4	4, 2, 0
5	1, 3, 5	4, 2, 0
6	0, 2, 4, 6	6, 4, 2, 0
7	1, 3, 5, 7	6, 4, 2, 0
$\vdots$	$\cdots$	$\cdots$

Supongamos que $P(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$, grad $(P) \geq 2$ y $P(x) = 0$ tiene $\zeta_1, \ldots, \zeta_{n-2}$ raíces reales. Entonces ζ_{n-1} y ζ_n son raíces complejas con $\zeta_{n-1} = \overline{\zeta_n}$. Pongamos $\zeta_n = a + ib$. Entonces

$$\left(x - \zeta_{n-1}\right)\left(x - \zeta_n\right) = x^2 - \left(\zeta_{n-1} + \zeta_n\right)x + \zeta_{n-1}\zeta_n,$$

resulta que

$$\begin{aligned}\zeta_{n-1}+\zeta_n &= 2R_e(\zeta_n)=2a,\\ \zeta_{n-1}\cdot\zeta_n &= |\zeta_n|^2=a^2+b^2,\end{aligned}$$

con lo cual

$$\left(x-\zeta_{n-1}\right)\left(x-\zeta_n\right)=x^2+2ax+a^2+b^2,$$

y en consecuencia

$$P(x)=\left(x^2-2ax+a^2+b^2\right)\prod_{j=1}^{n-2}\left(x-\zeta_j\right).$$

Si $\text{grad}(P)\geq 4$ y P tiene cuatro raíces complejas simples, el razonamiento anterior muestra que P puede escribirse en la forma

$$P(x)=\left(x^2-2ax+a^2+b^2\right)\left(x^2-2\widetilde{a}_1x+\widetilde{a}_1^2+\widetilde{b}_1^2\right)Q(x),$$

con $z_1=a+ib$, $z_2=\widetilde{a}_1+i\widetilde{b}_1$, $\text{grad}(Q)=n-4$ y $P(z_1)=P(z_2)=0$.

5.7.1. Fronteras superior e inferior de las raíces de la ecuación $P(x)=0$

Sea P un polinomio de grado n. Como P está definido en todo $\mathbb{R}$, necesitamos, en términos de los coeficientes de P, seleccionar los intervalos en los que se debe aplicar el algoritmo de búsqueda del cambio de signo. A continuación se establecen criterios para seleccionar tales intervalos.

Teorema 14 *Sea* $P(x)=\sum_{k=0}^{n}a_kx^k$ *un polinomio de grado* n. *Entonces, para todo* $x\in\mathbb{C}$ *tal que* $|x|>1$ *y* $0<k\leq\frac{|a_n|(|x|-1)}{\underset{k=0,\ldots,n-1}{\text{Max}}|a_n|}$ *se tiene*

$$|a_nx^n|>k\left|\sum_{k=0}^{n-1}a_kx^k\right|.$$

Demostración. Sea $A=\underset{k=0,\ldots,n-1}{\text{Max}}|a_k|$. Entonces, para $|x|\neq 1$, se tiene

$$\left|\sum_{k=0}^{n-1}a_kx^k\right|\leq\sum_{k=0}^{n-1}|a_k||x|^k\leq A\sum_{k=0}^{n-1}|x|^k=A\frac{|x|^n-1}{|x|-1}<A\frac{|x|^n}{|x|-1}.$$

Luego

$$\left|\sum_{k=0}^{n-1}a_kx^k\right|<\frac{A}{|x|-1}|x|^n\qquad\text{para }|x|\neq 1.$$

Sea $x\in\mathbb{C}$ tal que $|x|>1$. Si $k\in\mathbb{R}$ es tal que

$$0<k\leq\frac{|a_n|(|x|-1)}{A}$$

entonces

$$0<\frac{k}{|a_n|}\leq\frac{|x|-1}{A}\Leftrightarrow\frac{A}{|x|-1}\leq\frac{|a_n|}{k}$$

con lo cual

$$\begin{aligned}\frac{|a_n|}{k}|x|^n &\geq \frac{A}{|x|-1}|x|^n>\left|\sum_{k=0}^{n-1}a_kx^k\right|\\ |a_nx|^n &> k\left|\sum_{k=0}^{n-1}a_kx^k\right|\quad\text{si }|x|>1.\end{aligned}$$

■

Teorema 15 *Sea* $P(x)=\sum_{k=0}^{n} a_k x^k$ *un polinomio de grado* n *y* $A=\underset{k=0,\ldots,n-1}{\text{Max}}|a_k|$.
Si $\widehat{x}_i$, $i=1,\ldots,n$, *son las raíces (reales o complejas) de la ecuación* $P(x)=0$, *entonces* $|\widehat{x}_i|<1+\frac{A}{|a_n|}$ $i=1,\ldots,n$.

Demostración. Por el teorema anterior, si $k=1$ entonces $\frac{|a_n|(|x|-1)}{A}\geq 1$ para $|x|>1$, y en consecuencia

$$|a_n x^n|>\left|\sum_{k=0}^{n-1} a_k x^k\right| \quad \text{si } |x|>1.$$

Luego

$$\begin{aligned}|a_n x^n|-\left|\sum_{k=0}^{n-1} a_k x^k\right| &\geq |a_n x^n|-\sum_{k=0}^{n-1}|a_k|\,|x|^k \geq \left|a_n x^k\right|-A\sum_{k=0}^{n-1}|x|^k \\ &= |a_n|\,|x|^n - A\frac{|x|^n-1}{|x|-1} > |a_n|\,|x|^n - A\frac{|x|^n}{|x|-1} \\ &= |x|^n\left(|a_n|-\frac{A}{|x|-1}\right).\end{aligned}$$

Sea $x\in\mathbb{C}$ tal que $|x|>1$ y $|x|^n\left(|a_n|-\frac{A}{|x|-1}\right)\geq 0 \Rightarrow |a_n|-\frac{A}{|x|-1}\geq 0$. Resulta que

$$|P(x)|\geq |a_n x^n|-\left|\sum_{k=0}^{n-1} a_k x^k\right| > |x|^n\left(|a_n|-\frac{A}{|x|-1}\right)\geq 0 \qquad |x|>1.$$

Más aún $|a_n|-\frac{A}{|x|-1}\geq 0 \Rightarrow |x|\geq 1+\frac{A}{|a_n|}$. Así, $|P(x)|>0$ si $|x|\geq 1+\frac{A}{|a_n|}=R$. Consecuentemente, $|\widehat{x}_i|<R$ $i=1,\ldots,n$. ■

Denotamos con $\overline{B(0,R)}$ el disco cerrado de centro 0 y radio $R=1+\frac{A}{|a_n|}$.

Para $|x|>R$ se tiene $|P(x)|>0$ con lo que en el exterior de $\overline{B(0,R)}$ no se encuentra localizada ninguna raíz de $P(x)=0$. Todas las raíces de la ecuación $P(x)=0$ están localizadas en $\overline{B(0,R)}$, esto es,

$$\widehat{x}_i\in\overline{B(0,R)} \quad i=1,\ldots,n.$$

Ejemplos

1. Consideremos el polinomio $P(x)=8x^8-x^6+16x^4+x^3-5x^2+3x+1$.

   Sea $A=\underset{k=0,n-1}{\text{Max}}|a_k|=\text{Max}\{1,3,5,1,16,1\}=16$. Entonces

   $$R=1+\frac{A}{a_n}=1+\frac{16}{8}=3.$$

   Todas las raíces de la ecuación $P(x)=0$ se localizan en el disco cerrado $\overline{B(0,3)}$.

2. Todas las raíces de la ecuación $2x^5-x^4-3x^3+x-3=0$ están localizadas en el disco cerrado $B(0,R)$ con

   $$R=1+\frac{\underset{k=0,4}{\text{Max}}|a_n|}{|a_n|}=1+\frac{3}{2}=\frac{5}{2}.$$

Teorema 16 *Sea* $P(x)=\sum_{k=0}^{n} a_k x^k$ *un polinomio de grado* n *con* $a_0\neq 0$ *y* $\widehat{x}_i \quad i=1,\ldots,n$ *las raíces de la ecuación* $P(x)=0$. *Entonces*

$$\widehat{x}_i > \frac{|a_0|}{|a_0|+\underset{k=1,\ldots,n}{\text{Max}}|a_k|}=r, \qquad i=1,\ldots,n.$$

Demostración. Puesto que $P(x)=\sum_{k=0}^{n} a_k x^k = x^n \sum_{k=0}^{n} a_k x^{k-n}$ para $x\neq 0$. Sea $y=\frac{1}{x} \quad x\neq 0$ y $Q(y)=a_0 yn + a_1 y^{n-1}+\cdots+a_n$. Resulta

$$P\left(\frac{1}{4}\right)=\frac{1}{y^n}Q(y) \Rightarrow Q(y)=y^n P\left(\frac{1}{4}\right).$$

Para $|y|>1+\frac{\underset{k=1,\ldots,n}{\text{Max}}|a_k|}{|a_0|}$ se tiene $y^n P\left(\frac{1}{4}\right)>0$, entonces

$$\frac{1}{|x|}>1+\frac{\underset{k=1,\ldots,n}{\text{Max}}|a_k|}{|a_0|} \Rightarrow P(x)>0,$$

de donde

$$|x|<\frac{|a_0|\underset{k=1,\ldots,n}{\text{Max}}|a_k|}{|a_0|} \Rightarrow P(x)>0.$$

Consecuentemente, $|\widehat{x}_i|>r=\frac{|a_0|}{|a_0|+\underset{k=1,\ldots,n}{\text{Max}}|a_k|}$. ■

Conclusión: si $\widehat{x}_i, \quad i=1,\ldots,n$ son las raíces reales o complejas de la ecuación algebraica $P(x)=0$, entonces $\widehat{x}_i \in \overline{B(0,R)}-\overline{B(0,r)}$, con

$$R=1+\frac{\underset{k=0,\ldots,n-1}{\text{Max}}|a_k|}{|a_n|}, \quad \text{y} \quad r=\frac{|a_0|}{|a_0|+\underset{k=1,\ldots,n}{\text{Max}}|a_k|}, \qquad a_0\neq 0,$$

R es la frontera superior y r es la frontera inferior en las que están localizadas todas las raíces:

$$r<|\widehat{x}_i|<R \qquad i=1,\ldots,n.$$

En particular, las raíces reales se encuentran localizadas en los intervalos $[-R,-r]$ y $[r,R]$.

Las raíces positivas de $P(x)=0$ pertenecen a $[r,R]$ y las negativas a $[-R,r]$.

En el siguiente teorema se establece una mejor estimación de la frontera superior de las raíces reales.

Teorema 17 ***(de Lagrange)***
Sea $P(x)=\sum_{k=0}^{n} a_k x^k$ *un polinomio de grado* n. *Supongamos que* $a_n>0$ *y* $k<n$ *el mayor de los índices para los que* $a_k<0$. *Entonces, la frontera superior de las raíces positivas de la ecuación* $P(x)=0$ *es el número real* $R=1+\sqrt[k]{\frac{B}{a_n}}$, *donde* $B=\text{Max}\{|a_k| \mid a_k<0\}$.

Demostración. Sea $x>1$ y $Q(x)$ el polinomio que se obtiene de P al sustituir todos los coeficientes no negativos $a_{n-1},\ldots a_{k+1}$ por cero y cada uno de los coeficientes restantes $a_k,\ldots,a_0$ se sustituyen por $-B$, donde $B=\text{Max}\{|a_k| \mid a_k<0\}$.

Como $P(x)=a_0+a_1x+\cdots+a_nx^n$, entonces

$$\begin{aligned} Q(x) &= a_nx^n-Bx^k-Bx^{k-1}-\cdots-B=a_nx^n-B\frac{x^{k+1}-1}{x-1}>a_nx^n-B\frac{x^{k+1}}{x-1} \\ &= \frac{x^{k+1}}{x-1}\left(a_nx^{n-k-1}(x-1)-B\right)>\frac{x^{k+1}}{x-1}\left[a_n(x-1)^k-B\right] \qquad x\neq 1. \end{aligned}$$

Para $x > 1$ tal que $a_n (x-1)^k - B \geq 0 \Rightarrow x \geq 1 + \left(\frac{B}{a_n}\right)^{\frac{1}{k}}$ se tiene $P(x) > Q(x) \geq 0$.

Luego, $\widehat{x}_i < 1 + \left(\frac{B}{a_n}\right)^{\frac{1}{k}}$ con $\widehat{x}_i > 0$. ■

Si todos los coeficientes de P son positivos, para $x \geq 0$, $P(x) > 0$, es decir que la ecuación $P(x) = 0$ no tiene raíces reales positivas.

Ejemplos

1. Sea $P(x) = 8x^8 - x^6 + 16x^4 + x^3 - 5x^2 + 3x + 1$.

   Observamos que el mayor de los índices $k < 8$ para los que $a_k < 0$ es $k = 6$. Además, los índices para los que $a_k < 0$ son 6 y 2. Entonces $b = \text{Max}\{|a_k| \mid a_k < 0\} = \text{Max}\{1, 5\} = 5$. Resulta que

$$\begin{aligned} Q(x) &= 8x^8 - 5\left(x^6 + x^5 + \cdots + 1\right), \\ R &= 1 + \left(\frac{B}{a_8}\right)^{\frac{1}{6}} = 1 + \left(\frac{5}{8}\right)^{\frac{1}{6}} \simeq 1{,}925. \end{aligned}$$

2. Considerar el polinomio $P(x) = 3x^6 + 2x^5 + 8x^4 - x^3 - 10x^2 - 60$.

   Los coeficientes negativos son $a_3 = -1$, $a_2 = -10$, $a_0 = -60$, el mayor de lo índices de estos coeficientes es $k = 3$. Además $B = \text{Max}\{|a_k| \mid a_k < 0\} = \text{Max}\{1, 10, 60\} = 60$. Luego

$$\begin{aligned} Q(x) &= 3x^6 - 60\left(x^3 + x^2 + 1\right) \\ R &= 1 + \left(\frac{B}{a_6}\right)^{\frac{1}{3}} = 1 + \frac{1}{60}^{\frac{1}{3}} \simeq 3{,}72. \end{aligned}$$

   Si $R' = 1 + \frac{\underset{k=0,\ldots,5}{\text{Max}}|a_k|}{|a_6|} = 1 + \frac{60}{3} = 21$. Claramente $R < R'$.

Observación

Sean $0 < r < R$ las fronteras inferior y superior respectivamente de las raíces positivas de la ecuación $P(x) = 0$. El algoritmo de búsqueda del cambio de signo se aplica en el intervalo $[r, R]$ y para las negativas, el algoritmo se lo aplica en el intervalo $[-R, -r]$.

Ejemplo

Hallemos todas las raíces de la ecuación $3x^4 - 5{,}4x^3 + 3{,}11x^2 - 9x - 3{,}15 = 0$.

Sea $P(x) = 3x^4 - 5{,}4x^3 + 3{,}11x^2 - 9x - 3{,}15$. Determinemos un conjunto en el que están localizadas todas las raíces reales y complejas. Como $R = 1 + \left(\frac{B}{a_n}\right)^{\frac{1}{k}}$, donde $a_n > 0$, k el mayor de los índices para los que $a_k < 0$, $B = \text{Max}\{|a_k| \mid a_k < 0\}$. Se tiene $a_4 = 3$, $k = 3$ pues $a_3 = -5{,}4$, $a_1 = -9$, $a_0 = 3{,}15$, $B = 9$. Luego

$$R = 1 + \left(\frac{9}{3}\right)^{\frac{1}{3}} \simeq 2{,}44225.$$

Todas las raíces reales o complejas están localizadas en le disco cerrado $\overline{B(0, 2{,}5)}$ $(2{,}5 > 1 + 3^{\frac{1}{3}} \simeq 2{,}44225)$.

La aplicación del algoritmo de búsqueda del cambio de signo con un paso $h = 0{,}5$ muestra que $P(x) = 0$ tiene dos raíces reales localizadas en los intervalos $[-0{,}5, 0]$ y $[2{,}0, 2{,}5]$.

Note que $r = \frac{|a_0|}{|a_0| + \underset{k=1,\ldots,n}{\text{Max}}|a_k|} = \frac{3{,}15}{3{,}15+9} \simeq 0{,}25926$. En el intervalo $[-r, r]$ no existen raíces de $P(x) = 0$.

Además, si $u(x) = \frac{P(x)}{P'(x)}$, la aplicación del algritmo de búsqueda del cambio de signo muestra que u no tiene raíces reales múltiples. Consecuentemente, la ecuación propuesta tiene dos raíces reales y dos raíces complejas una conjugada de la otra.

i. Cálculo de $\widehat{x}_i \in [-0{,}5, -0{,}25]$ con una precisión $\varepsilon = 10^{-4}$. Apliquemos el método de Newton:

$$\varphi(x) = x - \frac{P(x)}{P'(x)} \quad x \in [-0{,}5, -0{,}25].$$

Escribamos $P(x)$ y $P'(x)$ usando el esquema de Hörner:

$$\begin{aligned} P(x) &= -3{,}15 + x(-9 + x(3{,}11 + x(-5{,}4 + 3x))), \\ P'(x) &= 12x^3 - 16{,}2x^2 + 6{,}22x - 9 = -9 + x(6{,}22 + x(-16{,}2 + 12x)). \end{aligned}$$

El esquema numérico es el siguiente: $\begin{cases} x_0 \in [-0{,}5, -0{,}25] \text{ aproximación inicial,} \\ x_{n+1} = \varphi(x_n) \quad n = 0, 1, \ldots, \end{cases}$ con $|x_{n+1} - x_n| < 10^{-4}$.

Sea $x_0 = -0{,}5$, entonces

$$\begin{aligned} x_1 &= -0{,}5 - \frac{P(-0{,}5)}{P'(-0{,}5)} = -0{,}5 - \frac{8{,}63}{-17{,}66} = -0{,}01133, \\ x_2 &= -0{,}01133 - \frac{P(-0{,}01133)}{P'(-0{,}01133)} = -0{,}01133 - \frac{-3{,}04767}{-9{,}07254} = -0{,}34725, \\ x_3 &= x_2 - \frac{P(x_2)}{P'(x_2)} = -0{,}34725 - \frac{0{,}61996}{-13{,}61574} = -0{,}30172, \\ x_4 &= x_3 - \frac{P(x_3)}{P'(x_3)} = -0{,}30172 - \frac{0{,}02172}{-12{,}68007} = -0{,}300002, \\ x_5 &= -0{,}3 \end{aligned}$$

Se tiene $|x_5 - x_4| < 10^{-4}$. La raíz negativa de $P(x) = 0$ es $\widehat{x}_1 = -0{,}3$.

Mediante un procedimiento análogo, con una aproximación inicial $x_0 = 2$ y cuatro iteraciones se calcula la raíz $\widehat{x}_2 = 2{,}1$

Se tiene $P(x) = (x + 0{,}3)(x - 2{,}1)\,Q(x)$ con $Q(x) = ax^2 + bx + c$. Se verifica fácilmente que $Q(x) = 3x^2 + 5$. Así $P(x) = (x + 0{,}3)(x - 2{,}1)\left(3x^2 + 5\right)$. Las raíces complejas son $\widehat{x}_3 = \frac{\sqrt{15}}{3}i$, $\widehat{x}_4 = -\frac{\sqrt{15}}{3}i$.

5.8. Ejercicios

1. Para las ecuaciones que en cada item se propone, separar las raíces utilizando el método gráfico y el algoritmo de búsqueda del cambio de signo. Aplique el método de bisección para aproximar la o las raíces, si existen, con una precisión $\varepsilon = 10^{-2}$.

   a) $x^2 + 2x - 3 = 0$. **b)** $x^3 - x - 1 = 0$. **c)** $x - 3^{-x} = 0$. **d)** $e^x - x^2 - 4x + 2 = 0$.

   e) $e^x + 2^{-x} - \dfrac{1}{x} = 0 \qquad x > 0$. **f)** $e^{|x|+1} - \sin(x) = 0$. **g)** $\cos(x) + x^2 + 2 = 0$.

2. Aplicar los métodos de punto fijo modificado, Newton modificado y regula. falsi para aproximar $\sqrt[3]{2}$ con una precisión $\varepsilon = 10^{-3}$. Para los tres métodos elija el mismo punto inicial x_0. Compare el número de iteraciones que se requieren para aproximar $\sqrt[3]{2}$ con la precisión ε.

3. Sea $a \in \mathbb{Q}$ tal que $0 < a < 1$ y a no es potencia cuarta de ningún número racional.

   a) Construya las funciones de iteración de los métodos: punto fijo modificado, Newton - Raphson, Newton modificado y regula - falsi para aproximar $\sqrt[4]{a}$.

   b) Para cada función de iteración φ del inciso **a.**, sea $x_0 = 1$ y $x_{n+1} = \varphi(x_n)$ $n = 0, 1, \ldots$. ¿Es (x_n) convergente a $\sqrt[4]{a}$?

4. En cada inciso, determine un intervalo $[a, b]$ en el cual la función de iteración Φ dada tenga un punto fijo. Estime el número $N_{\text{máx}}$ de iteraciones necesarias para obtener una precisión del punto fijo de 10^{-4}.

a) $\Phi(x) = \frac{\sqrt{3}}{3}e^{\frac{x}{2}}$. **b)** $\Phi(x) = 5^{-x}$. **c)** $\Phi(x) = \frac{1}{3}\left(2 - e^x + x^2\right)$. **d)** $\Phi(x) = \frac{5}{x^2} + 2$.

e) $\Phi(x) = \frac{x}{2} + \frac{1}{x}$. **f)** $\Phi(x) = \left(x^3 + 1\right)^{\frac{1}{3}}$.

Escriba la ecuación para la cual la raíz es punto fijo de Φ.

5. Sean $a, b \in \mathbb{R}$, $n \in \mathbb{Z}^+, n \geq 2$. Demostrar que la ecuación $x^n + ax + b = 0$ tiene a lo más dos raíces rales si n es par y tres raíces reales si n es impar.

   Si n es par, ¿qué condiciones han de verificar a y b para que la ecuación $x^n + ax + b = 0$ tenga dos raíces reales?

   Si n es impar, ¿qué condiciones han de verificar a y b para que la ecuación $x^n + ax + b = 0$ tenga tres raíces reales?

6. Sean $a, b \in \mathbb{R}$, $n \in \mathbb{Z}^+$ con $n \geq 3$. Estudiar la ecuación $x^n + ax^2 + b = 0$.

7. Encontrar todas las raíces de la ecuación $e^{0,2x^2} - 5x - 2 = 0$. Aplique los métodos de Newton-Raphson y de las secantes. La precisión $\varepsilon = 10^{-3}$.

8. Hallar la mas pequeña raíz positiva de la ecuación $2 - e^x \cos(x) = 0$ con una precisión $\varepsilon = 10^{-7}$, aplicando los métodos de Steffensen, donde φ es la función de iteración del método de Newton modificado; y, el algoritmo que se describe a continuación:

$$\begin{aligned} y &= x_n - \frac{f(x_n)}{f'(x_n)} \\ x_{n+1} &= y - \frac{f(y)}{f'(x_n)} \quad n = 0, 1, 2, \ldots \end{aligned}$$

9. Hallar la más grande raíz negativa de la ecuación $e^{-x}\sin(x) - 1 = 0$. Aplique los métodos de las secantes y de Steffensen donde la función de iteración φ viene dada por el método de regula-falsi. (precisión $\varepsilon = 10^{-6}$). Escriba en cada caso el algoritmo correspondiente.

10. Escriba un algoritmo que permita aproximar la raíz $\widehat{x} \in [a, b]$ de $f(x) = 0$ de tal manera que cada aproximación de $\widehat{x}$ se obtenga intercambiando el método de bisección y de Newton modificado, así sucesivamente.

11. Encontrar todas las raíces reales de la ecuación $x^3 - 0{,}6x^2 - 18{,}63x + 34{,}992 = 0$. ¿Existe alguna raíz de multiplicidad? Aplique el métodod de Newton - Raphson para determinar la raíz simple si esta existe y/o un método para determinar raíces de multiplicidad.

12. Dar un método localmente convergente para determinar el punto fijo $\widehat{x} = \sqrt[5]{2}$ de $\Phi(x) = x^5 + x - 2$.

13. Sea $\varepsilon = 10^{-5}$. Para la ecuación que se da en cada inciso aplicar el método que se propone para aproximar la o las raíces de la misma con la precisión ε. Escriba el respectivo algoritmo. Nota**:** si la ecuación tiene una infinidad de raíces, calcule todas aquellas que están localizadas en el intervalo $[-3, 6]$.

    a) $x - e^{-x} = 0$, método de punto fijo.

    b) $e^x + 2^{-x} + 2\cos(x) - 6 = 0$, método de punto fijo modificado.

    c) $e^x - x^2 + 3x - 2 = 0$, método de Newton - Raphson.

    d) $x^2 + 10\cos(x) = 0$, método de Newton modificado.

    e) $4\cos(x) - e^x = 0$, método de las secantes.

    f) $\ln\left(x^2 + 1\right) - e^{0,4x}\cos(\pi x) = 0$, método de regula - falsi.

    g) $x^3 - 3{,}23x^2 - 5{,}54x + 9{,}84 = 0$, método de punto fijo modificado.

    h) $e^{x\sin(x)} + 0{,}5x = 0$, método de las secantes.

14. Calcular las cuatro primeras raíces positivas de la ecuación $\tan(x) - 2x = 0$.

15. Calcular todas las raíces reales de las ecuaciones que se dan a continuación. Analice el caso de posibles raíces de multiplicidad.

    a) $x^4 - 7{,}223x^3 + 13{,}447x^2 - 0{,}672x - 10{,}223 = 0$. **b)** $5x^3 + 4{,}5x^2 - 34{,}8x + 25{,}2 = 0$.

    c) $x^4 - 2{,}2x^3 - 5{,}03x^2 + 6{,}864x + 9{,}7344 = 0$. **d)** $4x^3 - 22{,}8x^3 - 34{,}2x - 64{,}8 = 0$.

16. Aplique el método de Steffensen para calcular las raíces de las ecuaciones:

    a) $e^{0{,}5x^2} - 2x^2 - 3 = 0$. **b)** $(x-1)^2 - \log\left(2x^2+3\right) = 0$.

17. Sea $I \subset \mathbb{R}$ con $I \neq \emptyset$ y f una función real definida en I. Supongamos que existe $\widehat{x} \in [a,b] \subset I$ tal que $f(\widehat{x}) = 0$ y que $f \in C^2([a,b])$. Se define la sucesión (x_n) como sigue:

$$\begin{cases} x_0 \in [a,b], \\ y = x_n - \frac{f(x_n)}{f'(x_n)}, \quad f'(x_n) \neq 0 \\ x_{n+1} = y - \frac{f(y)}{f'(x_n)} \quad n = 0,1,\ldots \end{cases}$$

    a) Dé una interpretación geométrica de este esquema numérico.

    b) Si (x_n) es convergente, pruebe que (x_n) converge a $\widehat{x}$ al menos cúbicamente.

    c) Escriba el algoritmo correspondiente y aplique a las dos ecuaciones que se proponen a continuación.

    i) $2e^x - x - 5 = 0$. ii) $\cos(x) - x^2 + 1 = 0 \quad x \in [-\pi, \pi]$.

18. Sean $I \subset \mathbb{R}$ con $I \neq \emptyset$, f de I en $\mathbb{R}$ una función definida en I y $\widehat{x} \in I$ la única raíz de la ecuación $f(x) = 0$. Se define la función de iteración φ de I en $\mathbb{R}$ como $\varphi(x) = x - f(x) \quad x \in I$. Sea $x_0 \in I$ y (x_n) la sucesión que se define a continuación:

$$\begin{aligned} y_n &= \varphi(x_n) \\ z_n &= \varphi(y_n) \\ x_{n+1} &= x_n - \frac{(y_n - x_n)^2}{z_n - 2y_n + x_n} \quad n = 0,1,\ldots \end{aligned}$$

    a) Muestre que

$$x_{n+1} = x_n - \frac{[f(x_n)]^2}{f(x_n) - f(x_n - f(x_n))} \quad n = 0,1,\ldots$$

    Este esquema numérico se conoce como método casi Newton.

    b) Dé una interpretación geométrica del esquema numérico.

    c) Probar que $[x_n]$ converge a $\widehat{x}$ cuadráticamente.

    d) Sea $a \in \mathbb{R}$. La ecuación $\left(x^2+1\right)(x-a) = 0 \Longleftrightarrow x^3 - ax^2 + x - a = 0$ tiene a $\widehat{x} = a$ como la única raíz real de la ecuación $f(x) = 0$, donde $f(x) = x^3 - ax^2 + x - a$. Sea $a = 1{,}3521$ y $x_0 = 1$. Muestre que la sucesión (x_n) generada por el esquema numérico dado en a) converge a $\widehat{x} = 1{,}3521$.

19. Sea $a > 0$ y $\widehat{x} = \sqrt{a + \sqrt{a + \sqrt{a + \ldots}}}$. Sea φ de $\mathbb{R}^+$ en $\mathbb{R}$ la función definida por $\varphi(x) = \sqrt{a+x}$, $x > -a$. Se define (x_n) como sigue: $\begin{cases} x_0 = 0, \\ x_{n+1} = \varphi(x_n) \quad n = 0,1,\ldots \end{cases}$

    a) Pruebe que $\lim\limits_{n\to\infty} x_n = \widehat{x}$ y que $\widehat{x} = \frac{1+\sqrt{1+4a}}{2}$.

    b) Sea $\varepsilon = 10^{-2}, a = 2$. Aproxime $\widehat{x}$ con una precisión ε.

20. Sea $a > 0$. Se desea calcular $\widehat{x} = \frac{1}{a}$ sin usar la división. Para el efecto se define la función $f(x) = \frac{1}{x} - a$ para $x > 0$. Utilice el método de Newton para construir una sucesión (x_n) convergente a $\widehat{x}$.

21. Considerar el método de Newton de dos pasos siguiente

$$\begin{cases} x_0 \in [a,b], \\ y_n = x_n - \dfrac{f(x_n)}{f'(x_n)} \\ x_{n+1} = y_n - \dfrac{f(y_n)}{f'(x_n)} \quad n = 0,1,\ldots \end{cases}$$

a) Encuentre una función de iteración Φ sobre $[a,b]$ tal que $x_{n+1} = \Phi(x_n) \;\; n = 0,1,\cdots$.

b) Si (x_n) converge a $\widehat{x}$ muestre que

$$\lim_{n\to\infty} \frac{x_n - \widehat{x}}{(y_n - \widehat{x})(x_n - \widehat{x})} = \frac{f''(\widehat{x})}{f'(\widehat{x})}.$$

c) Pruebe que la convergencia es cúbica:

$$\lim_{n\to\infty} \frac{|x_{n+1} - \widehat{x}|}{|x_n - \widehat{x}|^3} = \frac{1}{2}\left(\frac{f''(\widehat{x})}{f'(\widehat{x})}\right)^2.$$

22. En cada inciso se define una función f. considere la ecuación $f(x) = 0$. Aplique los métodos de aproximación de raíces de multiplicidad para calcular la raíz de $f(x) = 0$.

a) $f(x) = x^2 - 2xe^{-x} + e^{-2x}$.

b) $f(x) = \sin^2[\pi(3x+2)] - \sin[\pi(3x+2)] - 1 \quad x \in \left[-\frac{\pi}{6}, \frac{\pi}{4}\right]$.

c) $f(x) = x^3 - 3\sqrt{2}x^2 + 6x - 2\sqrt{2}$.

23. Sean $\varepsilon = 10^{-6}$, $p = 0{,}8$ y $\varphi(x) = 0{,}5 + \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_0^x e^{-\frac{t^2}{2}}\,dt$, $\;x \geq 0$, la función de distribución normal.

Construya un método que permita aproximar $\widehat{x} > 0$ tal que $\varphi(\widehat{x}) = 0{,}8$ con una precisión ε. [Sugerencia: aplique la serie de Taylor de e^{α} y adecuadas sumas finitas de una serie de potencias].

24. Sea f una función real dependiente de un parámetro c. Escribiremos $t = f(x,c)$. Suponga que $\dfrac{\partial f}{\partial c}$ es continua.

Se dispone de un conjunto de datos experimentales $S = \left\{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1,\ldots,n\right\}$ y se asume que cada $y_i = f(x_i,c) + r_i(c)$, donde $r_i(c)$ denota el error en la observación y_i, $i = 1,\ldots,n$. En el método de mínimos cuadrados se considera el problema siguiente:

$$\min_{c\in\mathbb{R}} \sum_{i=1}^{n} r_i^2(c).$$

Se define $E(c) = \displaystyle\sum_{i=1}^{n} r_i^2(c) = \sum_{i=1}^{n} (y_i - f(x_i,c))^2$.

a) Elaborar un algoritmo para aproximar $\widehat{c} \in \mathbb{R}$ tal que $E(\widehat{c}) = \min_{c\in\mathbb{R}} E(c)$.

b) Se considera la siguiente información experimental:

$$S = \{(1,1{,}35), (1{,}5, 0{,}498), (2, 0{,}183), (2{,}2, 0{,}123)\}$$

Aplique el método de mínimos cuadrados para calcular la constante $\widehat{c} > 0$ tal que $f(t) = 10e^{-\widehat{c}t}$.

c) Se considera el siguiente conjunto de datos

$$S = \{(0{,}26, 5), (0{,}785, 5), (0{,}5, 8{,}7), (1{,}05, 8{,}7)\}$$

Aplique el método de mínimos cuadrados para calcular la constante $\widehat{c}$ tal que $f(t) = 10\sin(ct)$.

5.9. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Análisis Matemático, Segunda Edición, Editorial Reverté, Barcelona, 1982.
2. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.
3. Tom M. Apostol, Calculus, Volumen 2, Segunda Edición, Editorial Reverté, Barcelona, 1975.
4. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
5. Robert B. Banks, Growth and Diffusion Phenomena, Mathematical Frameworks and Applications, Editorial Springer-Verlag, Berlín, 1994.
6. R. M. Barbolla, M. García, J. Margalef, E. Outerelo, J. L. Pinilla. J. M. Sánchez, Introducción al Análisis Real, Editorial Alambra Universidad, Madrid, 1981.
7. G. Birkhoff, S. Maclane, Algebra Moderna, Cuarta Edición, Editorial Vicens-Vives, Barcelona. 1974.
8. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.
9. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
10. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
11. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial McGraw-Hill, México, 1981.
12. Ruel V. Churchill, James Ward Brown, Variable Compleja y Aplicaciones, Cuarta Edición, Editorial McGraw-Hill, Madrid, 1986.
13. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.
14. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.
15. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.
16. Waltson Fulks, Cálculo Avanzado, Editorial Limusa, México, 1973.
17. A. Kurosh, Cours D'Algèbre Supérieure, Editions Mir, Moscou, 1973.
18. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.
19. H. Hall, S.R. Knight, Algebra Superior, Unión Tipográfica Editorial Hispano-Americana, México, 1977.
20. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.
21. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.
22. Gerard Kiely , Ingeniería Ambiental, Volumen II, Editorial McGraw-Hill, Madrid, 1999.
23. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

24. A. Kurosh, Cours D'Algèbre Supérieure, Editions Mir, Moscou, 1973.

25. A. I. Kostrikin, Introducción al Algebra, Editorial Mir, Moscú, 1978.

26. Peter Linz, Theoretical Numerical Analysis, Editorial Dover Publications, Inc., New York, 2001.

27. Rodolfo Luthe, Antonio Olivera, Fernando Schutz, Métodos Numéricos, Editorial Limusa, México, 1986.

28. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

29. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

30. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

31. J. M. Ortega, W. C. Rheinbolodt, Iterative Solution of Nonlinear Equatios in Several Variables, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.

32. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

33. A. A. Samarski, Introducción a los Métodos Numéricos, Editorial Mir, Moscú, 1986.

34. Michelle Schatzman, Analyse Numérique, Inter Editions, París, 1991.

35. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

36. M. Sibony, J. Cl. Mardon, Analyse Numérique I, Systèmes Linéaires et non Linéaires, Editorial Hermann, París, 1984.

37. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

38. E. A. Volkov, Métodos Numéricos, Editorial Mir, Moscú, 1990.

Capítulo 6

Resolución numérica de sistemas de ecuaciones lineales

Resumen

El objetivo de este capítulo es presentar algunos métodos numéricos muy conocidos y prácticos para encontrar soluciones de sistemas de ecuaciones lineales. Primeramente se presentan algunos ejemplos donde surgen los sistemas de ecuaciones lineales. A continuación se examinan tres tipos de problemas con sistemas de ecuaciones lineales: sistemas de ecuaciones que poseen solución única, sistemas de ecuaciones que poseen infinitas soluciones, sistemas de ecuaciones que no tienen solución. Para el estudio de la existencia de soluciones de sistemas de ecuaciones lineales y el método numérico a elegir es importante el reconocimiento del tipo de matriz de dicho sistema, por lo que tratamos algunos tipos de matrices. Pasamos luego a la resolución numérica de los sistemas de ecuaciones lineales. Consideramos los sistemas más simples a resolver como son los triangulares superiores y los inferiores. A continuación tratamos el método de eliminación gaussiana y la implementación del pivoting parcial y total. Se consideran métodos para el cálculo del determinante de una matriz y el cálculo de la matriz inversa. Se trata el método de factorización LU de Crout. Cuando las matrices son simétricas, definida positivas se implementa el método de factorización L^TL de Choleski. Se dan aplicaciones a las matrices tridiagonales y a los sistemas de ecuaciones que tienen una infinidad de soluciones. Se concluye este capítulo con un análisis del condicionamiento de una matriz.

Las soluciones en mínimos cuadrados de sistemas de ecuaciones, que en general, no tienen solución serán tratados en el capítulo de mínimos cuadrados.

Para la aproximación de soluciones de grandes sistemas de ecuaciones lineales que poseen solución única cuya matriz del sistema tiene estructura de matriz en banda, se utilizan métodos iterativos. Algunos de estos métodos se tratan en el capítulo de métodos iterativos.

Al final del capítulo se precisa una amplia bibliografía sobre todos estos temas.

6.1. Problemas que conducen a la resolución de sistemas de ecuaciones lineales.

La resolución numérica de sistemas de ecuaciones lineales surgen en modelos matemáticos de la mayor parte de las ciencias tales como la física, la química, la biología, la economía, la psicología, la medicina, las diferentes ramas de la ingeniería, en los problemas ambientales, en estadística, en optimización y muy particularmente en investigación de operaciones y en el cálculo científico. Es por esto que se deben disponer de algoritmos numéricos y de programas computacionales listos a ser implementados en una variedad de situaciones.

A continuación presentamos algunos problemas clásicos que conducen a la resolución de sistemas de ecuaciones lineales.

6.1.1. Problemas de mínimos cuadrados discreto.

Supongamos que se dispone de un conjunto de n pares de datos experimentales

$$S = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, ..., n\}.$$

Se desea encontrar un polinomio P de grado 3:

$$P(x) = a + bx + cx^2 + dx^3 \quad x \in \mathbb{R}, \tag{1}$$

de modo que P se ajuste de la mejor manera al conjunto de datos S.

El polinomio P queda perfectamente bien definido si se conocen todos sus coeficientes a, b, c, d. Estos coeficientes son calculados mediante el denominado método de mínimos cuadrados discreto que describimos a continuación.

Denotemos con r_i el residuo en cada medición, esto es,

$$y_i = P(x_i) + r_i = a + bx_i + cx_i^2 + dx_i^3 + r_i, \quad i = 1, ..., n.$$

En forma matricial, el conjunto de ecuaciones precedente, se escribe como

$$\begin{bmatrix} y_1 \\ \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} r_1 \\ \vdots \\ \vdots \\ r_n \end{bmatrix}. \tag{3}$$

El residuo en cada medición depende de los coeficientes a, b, c, d del polinomio P. Definimos los vectores $\overrightarrow{X}$, $\overrightarrow{Y}$, $\overrightarrow{r}(\overrightarrow{x})$ como sigue:

$$\overrightarrow{X} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \qquad \overrightarrow{Y} = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, \qquad \overrightarrow{r}(\overrightarrow{x}) = \begin{bmatrix} r_1(\overrightarrow{x}) \\ \vdots \\ \vdots \\ r_n(\overrightarrow{x}) \end{bmatrix},$$

y la matriz A siguiente: $A = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$. El sistema de ecuaciones (3) se transforma en el siguiente

$$\overrightarrow{Y} = A\overrightarrow{X} + \overrightarrow{r}(\overrightarrow{x}),$$

de donde

$$\overrightarrow{r}(\overrightarrow{x}) = \overrightarrow{y} - A\overrightarrow{x}. \tag{4}$$

El problema de hallar el "mejor polinomio" que se ajusta al conjunto de datos S se expresa como sigue:

$$\text{hallar } \widehat{x}^T = (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d}) \in \mathbb{R}^4, \text{ si existe, talque } \|\overrightarrow{r}(\widehat{x})\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^4}{\text{Min}} \|\overrightarrow{r}(\widehat{x})\|^2, \tag{5}$$

o de modo equivalente

$$\|\overrightarrow{y} - A\widehat{x}\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^4}{\text{Min}} \|\overrightarrow{y} - A\widehat{x}\|^2. \tag{6}$$

Este problema se conoce como método de mínimos cuadrados y se demostrará que conduce a resolver el sistema de ecuaciones

$$A^T A\overrightarrow{x} = A^T\overrightarrow{y}, \tag{7}$$

donde A^T denota la matriz transpuesta de A.

Otros problemas semejantes al descrito se presentan en la aproximación en mínimos cuadrados continuos, en regresión lineal y multilineal, ajuste de datos, entre otros.

La formulación algebraica del método de mínimos cuadrados fue publicada por vez primera por Legendre en 1805.

6.1.2. Aproximación de un problema de valores de frontera.

Sea $L > 0$. Se denota con $C^2([0, L])$ el conjunto de funciones que poseen derivadas segundas continuas en $[0, L]$. Consideramos el problema siguiente: dadas dos funciones f, q continuas en $[0, L]$, hallar una función $u \in C^2([0, L])$ solución de

$$\begin{cases} -u''(x) + q(x)u(x) = f(x) & x \in]0, L[\,, \\ u(0) = u(L) = 0. \end{cases} \tag{8}$$

Supondremos que la función q satisface la condición

$$q(x) \geq 0 \quad \forall x \in [0, L[\,.$$

Se demuestra que este problema tiene solución única. Desafortunadamente su solución exacta puede determinarse en muy pocos casos, lo que conduce a calcularla de manera aproximada. Para el efecto, aplicamos el método de diferencias finitas que describimos brevemente a continuación.

Este problema se encuentra en muchas aplicaciones en ingeniería, por ejemplo, en la flexión de una viga fija en los extremos y sujeta a una carga $f(x)$, $x \in [0, L[$, en problemas de transferencia de calor y de masa, en problemas de contaminación ambiental.

Sea $n \in \mathbb{Z}^+$. Dividimos el intervalo $[0, L[$ en n subintervalos de longitud $h = \frac{L}{n}$. Ponemos $x_k = kh$, $k = 0, 1, ..., n$. El conjunto de puntos $\{x_0 = 0, x_1, ..., x_n = L\}$ se llaman nodos de discretización.

Sea u_k una aproximación de $u(x_k)$ que escribimos $u_k \simeq u(x_k)$, $k = 0, 1, ..., n$. Entonces, para $x = 0$, se tiene $0 = u(0) = u_0$, y para $x = L$, $0 = u(L) = u_n$. El polinomio de Taylor con error permite escribir los desarrollos siguientes:

$$\begin{aligned} u(x_{k+1}) &= u(x_k + h) = u(x_k) + hu'(x_k) + \frac{h^2}{2!}u''(x_k) + o(h^3), \\ u(x_{k-1}) &= u(x_k - h) = u(x_k) - hu'(x_k) + \frac{h^2}{2!}u''(x_k) + o(h^3). \end{aligned}$$

Sumando miembro a miembro obtenemos

$$u(x_{k+1}) + u(x_{k-1}) = 2u(x_k) + h^2u''(x_k) + o(h^3),$$

de donde

$$u''(x_k) = \frac{u(x_{k+1}) - 2u(x_k) + u(x_{k-1})}{h^2} + o(h),$$

y en consecuencia, la derivada segunda $u''(x_k)$ se aproxima mediante el cociente

$$\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2}, \quad k = 1, ..., n-1$$

que se denomina diferencia finita central de segundo orden (véase el capítulo 2).

La ecuación diferencial en cada punto x_k, $k = 1, ..., n-1$ se escribe

$$\begin{cases} -u''(x_k) + q(x_k)u(x_k) = f(x_k) & k = 1, ..., n-1, \\ u_0 = u_n = 0, \end{cases} \tag{9}$$

y al remplazar la derivada segunda por la diferencia finita central de segundo orden, la ecuación diferencial precedente se aproxima como

$$\begin{cases} -\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + q(x_k)u_k = f(x_k) & k = 1, ..., n-1, \\ u_0 = u_n = 0. \end{cases} \tag{10}$$

o lo que es lo mismo

$$\begin{cases} -\frac{u_2 - 2u_1}{h^2} + q(x_1)u_1 = f(x_1), \\ \vdots \\ -\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + q(x_2)u_2 = f(x_2), \\ \vdots \\ -\frac{2u_{n-1} + u_{n-1}}{h^2} + q(x_{n-1})u_{n-1} = f(x_{n-1}), \end{cases}$$

pués para $k = 1$ se ha considerado $u_0 = 0$ y para $k = n - 1$ se tiene $u_n = 0$, que en forma matricial se escribe

$$\frac{1}{h^2}\begin{bmatrix} 2+h^2q(x_1) & -1 & 0 & \cdots & 0 \\ -1 & 2+h^2q(x_2) & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -1 \\ 0 & \cdots & -1 & & 2+h^2q(x_{n-1}) \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-2}) \\ f(x_{n-1}) \end{bmatrix}.$$

Definimos la matriz A como

$$A = \begin{bmatrix} 2+h^2q(x_1) & -1 & 0 & \cdots & 0 \\ -1 & 2+h^2q(x_2) & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -1 \\ 0 & \cdots & -1 & & 2+h^2q(x_{n-1}) \end{bmatrix},$$

y los vectores $\overrightarrow{u}^T = (u_1, ..., u_{n-1})$, $\overrightarrow{b}^T = (h^2f(x_1), ..., h^2f(x_{n-1})) \in \mathbb{R}^{n-1}$.

La matriz A es tridiagonal, simétrica, definida positiva. El sistema de ecuaciones (10) se escribe entonces

$$A\overrightarrow{u} = \overrightarrow{b}. \tag{11}$$

A continuación se indican otros problemas que se discretizan mediante el método de diferencias finitas, elementos finitos, volúmenes finitos.

1. Resolución numérica de ecuaciones en derivadas parciales como la ecuación de Laplace que modela los flujos saturados incompresibles, la ecuación de conducción del calor y la ecuación de propagación de ondas.
2. Resolución numérica de ecuaciones integrales como las que provienen de la representación de soluciones de ecuaciones en derivadas parciales mediante funciones o núcleos de Green.

6.1.3. Trazado de una curva suave a partir de observaciones experimentales.

Supóngase que se dispone de un conjunto de n pares de datos experimentales de $\mathbb{R}^2$ siguiente:

$$S = \left\{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, ..., n\right\},$$

tales que $a = x_1 < x_2 < .... < x_n = b$.

El trazado de una curva suave que pase por todos los puntos de S, es un problema de interpolación que consiste en hallar una función f de $[a, b]$ en $\mathbb{R}$ que posee una cierta regularidad tal que

$$f(x_i) = y_i, \quad i = 1, ..., n,$$

de modo que dado $x \in [a, b]$ podamos calcular $f(x)$.

La construcción de la función f permite trazar una curva suave que pasa por todos ellos. Una estrategia es hallar un polinomio f cuya gráfica para por todos los puntos del conjunto S. Este problema se conoce como interpolación polinomial. Lastimosamente, cuando el número de puntos es grande se presentan oscilaciones lo que provoca muchas imprecisiones en los cálculos. Otra estrategia es utiliza tipos especiales de funciones denominada splines. Las gráficas de estas funciones no necesariamente pasan por todos los puntos, es decir que, en general, no se interpolan.

Esta clase de problemas se presentam fundamentalmente en computación gráfica, diseño geométrico asistido por computadora, en la robótica, etc.

Un spline cúbico ajusta una curva suave (de clase $C^2([a,b])$) al conjunto de puntos S.

Para fijar las ideas, se consideran los splines cúbicos con condiciones de frontera naturales que a continuación se definen. Denotamos con $\mathcal{P}_3$ al espacio vectorial de polinomios de grado ≤ 3. Dado el conjunto S, se busca una función f de al menos clase $C^2([a,b])$ que cumpla con las siguientes condiciones:

i) $f_{/_{[x_{i-1},x_i]}} \in \mathcal{P}_3$ que se le denota S_i, $i = 1, ..., n-1$.

ii) $S_i(x_i) = Y_i \quad i = 1, ..., n$.

iii) $S_{i+1}(x_{i+1}) = S_i(x_{i+1}), \quad i = 1, ..., n-1$.

iv) $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1}), \quad i = 1, ..., n-1$.

v) $S''_{i+1}(x_{i+1}) = S''_i(x_{i+1}) \quad i = 1, ..., n-1$.

vi) $S''(a) = S''(b) = 0$.

Para la construcción de la función f consecuentemente de S_i que es un polinomio de grado ≤ 3 en cada subintervalo $[x_{i-1}, x_i]$, $i = 1, ..., n$ ponemos $h_i = x_{i+1} - x_i$, $i = 1, ..., n-1$, y

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad i = 1, ..., n-1.$$

Notemos que las condiciones iii), ii), v) establecen la continuidad de la función f, f' y f'' en todo $[a,b]$.

De ii) se deduce

$$y_i = S_i(x_i) = a_i, \quad i = 1, ..., n-1. \tag{1}$$

Se define

$$a_n = f(x_n) = y_n.$$

Por iii), se tiene

$$\begin{cases} S_{i+1}(x_{i+1}) = a_{i+1}, \\ S_i(x_{i+1}) = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \end{cases}$$

de donde

$$a_{i+1} = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \quad i = 1, ..., n-1. \tag{2}$$

Las derivadas $S'_i(x)$ y $S'_{i+1}(x)$ están definidas como sigue:

$$\begin{aligned} S'_i(x) &= b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2, \\ S'_{i+1}(x) &= b_{i+1} + 2c_{i+1}(x - x_{i+1}) + 3d_{i+1}(x - x_{i+1})^2, \end{aligned}$$

y por la condición iv) tenemos

$$\begin{aligned} S'_{i+1}(x_{i+1}) &= b_{i+1}, \\ S'_i(x_{i+1}) &= b_i + 2c_i h_i + 3d_i h_i^2 \quad i = 1, ..., n-1, \end{aligned}$$

con lo cual $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1})$ implica

$$b_{i+1} = b_i + 2c_i h_i + 3d_i h_i^2 \quad i = 1, ..., n-1. \tag{3}$$

Se define $b_n = f'(x_n)$.

Las derivadas $S''_i(x)$ y $S''_{i+1}(x)$ están definidas como sigue:

$$\begin{aligned} S''_i(x) &= 2c_i + 6d_i(x - x_i), \\ S''_{i+1}(x) &= 2c_{i+1} + 6d_{i+1}(x - x_{i+1}). \end{aligned}$$

Entonces,

$$\begin{aligned} S''_{i+1}(x_{i+1}) &= 2c_{i+1}, \\ S''_i(x_{i+1}) &= 2c_i + 6d_i h_i, \end{aligned}$$

y de la condición v) obtenemos

$$2c_{i+1} = 2c_i + 6d_i h_i,$$

o bien

$$c_{i+1} = c_i + 3d_i h_i \quad i = 1, ..., n-1. \tag{4}$$

de donde

$$d_i = \frac{c_{i+1} - c_i}{3h_i} \quad i = 1, ..., n-1. \tag{5}$$

Remplazando d_i en (2), obtenemos

$$a_{i+1} = a_i + b_i h_i + c_i h_i^2 + \frac{c_{i+1} - c_i}{3h_i} h_i^3 = a_i + b_i h_i + c_i h_i^2 + \frac{1}{3}(2c_i + c_{i+1}) h_i^2,$$

$$a_{i+1} = a_i + b_i h_i + \frac{1}{3}(2c_i + c_{i+1}) h_i^2 \quad i = 1, ..., n-1. \tag{6}$$

Remplazando d_i en (3)

$$b_{i+1} = b_i + 2c_i d_i + 3\frac{c_{i+1} - c_i}{3h_i} h_i^2 = b_i + 2c_i h_i + (c_{i+1} - c_i) h_i.$$

$$b_{i+1} = b_i + h_i (c_{i+1} + c_i) \qquad i = 1, ..., n-1. \tag{7}$$

Por otro lado, de (6)

$$b_i = \frac{a_{i+1} - a_i}{h_i} - \frac{c_{i+1} + 2c_i}{3} h_i \quad i = 1, ..., n-1. \tag{8}$$

y disminuyendo en 1 el índice de la igualdad precedente, se obtiene

$$b_{i-1} = \frac{a_i - a_{i-1}}{h_{i-1}} - \frac{c_i + 2c_{i-1}}{3} h_{i-1} \quad i = 2, ..., n, \tag{9}$$

y en (7)

$$b_i = b_{i-1} + h_{i-1}(c_i + c_{i-1}), \quad i = 2, ..., n. \tag{10}$$

Remplazando (8) y (9) en (10), se deduce

$$\frac{a_{i+1} - a_i}{h_i} - \frac{c_{i+1} + 2c_i}{3} h_i = \frac{a_i - a_{i-1}}{h_{i-1}} - \frac{c_i + 2c_{i-1}}{3} h_{i-1} + h_{i-1}(c_i + c_{i-1}),$$

de donde

$$\frac{a_{i+1} - a_i}{h_i} - \frac{a_i - a_{i-1}}{h_{i-1}} = \frac{c_{i+1} + 2c_i}{3} h_i - \frac{c_i + 2c_{i-1}}{3} h_{i-1} + (c_i + c_{i-1}) h_{i-1}.$$

Puesto que

$$\begin{aligned} \frac{3}{h_i}(a_{i+1} - a_i) &= \frac{3}{h_{i-1}}(a_i - a_{i-1}) = c_{i+1} + 2c_i - (c_i + 2c_{i-1}) h_i + 3(c_i + c_{i-1}) h_{i-1} \\ &= c_{i-1} h_{i-1} + 2(h_{i-1} + h_i) c_i + c_{i+1} h_i. \end{aligned}$$

De (1) se tiene

$$\frac{3}{h_i}(y_{i+1} - y_i) - \frac{3}{h_{i-1}}(y_i - y_{i-1}) = c_{i-1} h_{i-1} + 2(h_{i-1} + h_i) c_i + c_{i+1} h_i \quad i = 2, ..., n-1,$$

que a su vez puede escribirse como

$$(h_{i-1}, 2(h_{i-1} + h_i), h_i) \begin{bmatrix} c_{i-1} \\ c_i \\ c_{i+1} \end{bmatrix} = \frac{3}{h_i}(y_{i+1} - y_i) - \frac{3}{h_{i-1}}(y_i - y_{i-1}) \quad i = 2, ..., n-1.$$

Por otro lado, se tiene

$$c_n = \frac{1}{2} f''(x_n) = 0,$$

y

$$0 = f''(x_1) = S_1''(x_1) = 2c_1,$$

de donde $c_1 = 0$.

Se definen $\overrightarrow{c}^T = (c_1, ..., c_n) \in \mathbb{R}^n$ con $c_1 = c_n = 0$, la matriz A siguiente:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ h_1 & 2(h_1+h_2) & h_2 & 0 & \cdots & \cdots & 0 \\ 0 & h_2 & 2(h_2+h_3) & h_3 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \ddots & \vdots & \vdots \\ \vdots & & & \ddots & & h_{n-2} & \vdots \\ 0 & \cdots & \cdots & \cdots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix},$$

y el vector $\overrightarrow{b}^T \in \mathbb{R}^n$:

$$\overrightarrow{b} = \begin{bmatrix} 0 \\ \frac{3}{h_2}(y_3 - y_2) - \frac{3}{h_1}(y_2 - y_1) \\ \\ \frac{3}{h_{n-1}}(y_n - y_{n-1}) - \frac{3}{h_{n-2}}(y_{n-1} - y_{n-2}) \\ 0 \end{bmatrix}.$$

En consecuencia, se obtiene el siguiente sistema de ecuaciones lineales

$$A\overrightarrow{c} = \overrightarrow{b}.$$

Una vez calculado $\overrightarrow{c}$, de (5) se obtiene $d_1, ..., d_{n-1}$ y de (8) se obtiene $b_1, ..., b_{n-1}$.

Problemas de optimización.

Mencionamos brevemente otros problemas que requieren de la resolución numérica de sistemas de ecuaciones lineales en la resolución de problemas de optimización como en: programación lineal, programación cuadrática, programación dinámica, control optimal, optimización de funciones convexas con o sin restricciones, problemas de grafos y redes, y de manera más general en el análisis combinatorio.

6.2. Problemas con sistemas de ecuaciones lineales.

Un sistema de ecuaciones lineales es un conjunto de ecuaciones de la forma

$$\begin{cases} a_{11}x_1 + ... + a_{1n}x_n = b_1 \\ \vdots \\ a_1x_1 + ... + a_{mn}x_n = b_m, \end{cases} \tag{12}$$

donde $x_1, ..., x_n$ son las incógnitas cuyos valores queremos determinar, y $a_{ij} \quad i = 1, ..., n$, b_i, $i = 1, ..., m$ son constantes reales conocidas. Pongamos

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}, \quad \overrightarrow{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

entonces A es una matriz de $m \times n$, esto es, $A \in M_{m\times n}[\mathbb{R}]$; $\overrightarrow{b} \in \mathbb{R}^m$ y $\overrightarrow{x} \in \mathbb{R}^n$. El sistema de ecuaciones (12) se expresa en forma matricial como $A\overrightarrow{x} = \overrightarrow{b}$.

En lo sucesivo consideraremos el problema (P) siguiente: hallar, si existe, $\overrightarrow{x} \in \mathbb{R}^n$ solución del sistema de ecuaciones lineales

$$A\overrightarrow{x} = \overrightarrow{b}.$$

Consideremos tres clases de problemas.

Problema I

Suponemos que $m < n$, es decir que tenemos más incógnitas que ecuaciones. Esta clase de sistemas de ecuaciones poseen infinitas soluciones o ninguna solución.

Cuando el sistema de ecuaciones lineales tiene infinitas soluciones, se dice que el sistema es sobredeterminado. Si el rango de la matriz A es m, esto es, $R(A) = m$, el sistema de ecuaciones posee infinitas soluciones y en tal caso consideramos el problema siguiente denominado solución del sistema de ecuaciones lineales en norma mínima:

$$\|\widehat{x}\|^2 = \underset{\substack{\overrightarrow{x}\in\mathbb{R}^n \\ A\overrightarrow{x}=\overrightarrow{b}}}{\text{Min}} \|\overrightarrow{x}\|^2, \tag{13}$$

esto es, entre todas las soluciones $\overrightarrow{x} \in \mathbb{R}^n$ del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, seleccionamos una que posea norma mínima que lo notamos con $\widehat{x}$.

Ejemplos

1. La ecuación $x+y+z+w = 1$ posee infinitas soluciones. Esta ecuación se escribe en forma matricial como
$$(1,1,1,1)\begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} = 1.$$
Note que la matriz $A = (1,1,1,1)$ es un vector fila y su rango es $R(A) = 1$, $b = 1$. La solución en norma mínima es $\widehat{x} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

2. El sistema de ecuaciones lineales $\begin{cases} 3x - 2y + 5z = 2 \\ 8x + y - 3z = 3 \end{cases}$ con $(x, y, z) \in \mathbb{R}^3$, tiene infinitas soluciones. La matriz A, los vectores $\overrightarrow{b}$ y $\overrightarrow{x}$ son
$$A = \begin{bmatrix} 3 & -2 & 5 \\ 8 & 1 & -3 \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \overrightarrow{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$
El rango de la matriz A es 2.

3. El sistema de ecuaciones lineales $\begin{cases} x - 2y + 3z = -1 \\ -4x + 8y - 12z = 0 \end{cases}$ con $(x, y, z) \in \mathbb{R}^3$, no tiene solución. Pués si multiplicamos por -4 a la primera ecuación, obtenemos el sistema de ecuaciones siguiente:
$\begin{cases} -4x + 8y - 12z = 4 \\ -4x + 8y - 12z = 0, \end{cases}$ que es un sistema contradictorio.

Problema II

Supongamos que $m > n$. En este caso, el sistema de ecuaciones lineales tiene más ecuaciones que incógnitas. Esta clase de ecuaciones tienen, por lo general, solución única o ninguna solución.

Denotemos con A_j la j-ésima columna de la matriz A. Si el vector $\overrightarrow{b}$ pertenece el espacio generado por las columnas de A:
$$\overrightarrow{b} \in \left\{ \sum_{j=1}^{n} \alpha_j A_j \mid \alpha_i \in \mathbb{R}, \quad j = 1, ..., n \right\},$$
entonces, el sistema de ecuaciones lineales posee una única solución $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$.

Si $\overrightarrow{b} \notin \left\{ \sum_{j=1}^{n} \alpha_j A_j \mid \alpha_j \in \mathbb{R}, \quad j = 1, ..., n \right\}$, el sistema de ecuaciones lineales no tiene solución. En este caso, consideraremos el problema siguiente: hallar $\widehat{x} \in \mathbb{R}^n$ tal que

$$\|A\widehat{x} - \overrightarrow{b}\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^n}{\text{Min}} \|A\overrightarrow{x} - \overrightarrow{b}\|^2. \tag{14}$$

Este problema se conoce como solución en mínimos cuadrados.

Note que no se pretende resolver el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, de hecho este sistema no tiene solución. En realidad planteamos un problema de mínimos cuadrados análogo al presentado en la sección precedente. En efecto, como el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ no tiene solución, definimos el residuo $\overrightarrow{r}(\overrightarrow{x}) \in \mathbb{R}^m$ como

$$\overrightarrow{r}(\overrightarrow{x}) = A(\overrightarrow{x}) - \overrightarrow{b}, \qquad \overrightarrow{x} \in \mathbb{R}^n.$$

El problema de mínimos cuadrados consiste en determinar en vector $\widehat{x} \in \mathbb{R}^n$ que minimice $\|\overrightarrow{r}(\overrightarrow{x})\|^2$ cuando $\overrightarrow{x}$ recorre todo $\mathbb{R}^n$, o sea

$$\|\overrightarrow{r}(\widehat{x})\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^n}{\text{Min}} \|\overrightarrow{r}(\overrightarrow{x})\|^2.$$

Este problema se abordará con más detalle en el capítulo de mínimos cuadrados.

Ejemplo

Considérese los datos de la tabla siguiente:

x	y	z
0	1	5
1	1.5	10.9
2	2	17.1
3	2.5	23
4	3	30

Con estos datos se desea encontrar una función real f de la forma

$$z = f(x, y) = a + bx + cy \qquad x, y \in \mathbb{R}.$$

Se establece el sistema de ecuaciones lineales siguiente:

$$\begin{cases} a & & & + & c & = & 5 \\ a & + & b & + & 1{,}5 & = & 10{,}9 \\ a & + & 2b & + & 2c & = & 17{,}1 \\ a & + & 3b & + & 2{,}5c & = & 23 \\ a & + & 4b & + & 3c & = & 30. \end{cases}$$

Este sistema de ecuaciones no tiene solución. Poniendo

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1{,}5 \\ 1 & 2 & 2 \\ 1 & 3 & 2{,}5 \\ 1 & 4 & 3 \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} 5 \\ 10{,}5 \\ 17{,}1 \\ 23 \\ 30 \end{bmatrix}, \quad \overrightarrow{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

el residuo es

$$\overrightarrow{r}(\overrightarrow{x}) = \overrightarrow{r}(a, b, c) = A\overrightarrow{x} - \overrightarrow{b},$$

con lo cual

$$\|\overrightarrow{r}(\widehat{a}, \widehat{b}, \widehat{c})\|^2 = \underset{(a,b,c) \in \mathbb{R}^3}{\text{Min}} \|\overrightarrow{r}(a, b, c)\|^2.$$

Problema III

Cosideramos sistemas de ecuaciones lineales que tienen igual número de ecuaciones que incógnitas, esto es, $m = n$. Encontramos tres clases de sistemas: aquellos que tienen solución única denominados sistemas de ecuaciones lineales consistente. Aquellos sistemas que tienen infinitas soluciones denominados sobredeterminados y aquellos que no tienen solución llamados inconsistentes.

Denotamos con T_A la aplicación lineal asociada a la matriz A, esto es:

$$T_A : \left\{ \begin{array}{ccc} \mathbb{R}^n & \rightarrow & \mathbb{R}^n \\ \overrightarrow{x} & \rightarrow & T_A(\overrightarrow{x}) = A\overrightarrow{x}. \end{array} \right.$$

El núcleo de T_A se define como el conjunto

$$\ker(T_A) = \{\overrightarrow{x} \in \mathbb{R}^n \mid T_A(\overrightarrow{x}) = 0\} = \{\overrightarrow{x} \in \mathbb{R}^n \mid A\overrightarrow{x} = 0\}.$$

El rango de T_A

$$R(T_A) = \{T_A(\overrightarrow{x}) \mid \overrightarrow{x} \in \mathbb{R}^n\} = \{A\overrightarrow{x} \mid \overrightarrow{x} \in \mathbb{R}^n\}.$$

El resultado fundamental del álgebra lineal que caracteriza a las aplicaciones lineales en espacios de dimensión finita es la relación que se establece entre las dimensiones del núcleo y del rango que están ligadas por la siguiente fórmula:

$$\dim \ker(T_A) + \dim R(T_A) = n.$$

Entonces, el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ tiene solución única si y solo si una de las propiedades siguientes se verifica:

i) $\ker(T_A) = \{0\}$.

ii) $R(T_A) = \mathbb{R}^n$.

iii) A es una matriz invertible.

iv) $\det(A) \neq 0$.

La propiedad *i)* significa que T_A es inyectiva. La propiedad *ii)* muestra que T_A es sobreyectiva. La propiedad *iii)* significa que T_A es biyectiva y $T_A^{-1} = T_{A^{-1}}$. Además, en el caso en que una de estas propiedades se verifique, las columnas de la matriz A son linealmente independientes. De manera similar, las filas de la matriz A son linealmente independientes.

A la solución única del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ lo notamos con $\overrightarrow{x} = A^{-1}\overrightarrow{b}$, donde A^{-1} denota la matriz inversa de A.

Si el sistema de ecuaciones no tiene solución, abordaremos el problema de mínimos cuadrados siguiente:

$$\text{hallar } \widehat{x} \in \mathbb{R}^n \text{ tal que } \| A\widehat{x} - \widehat{b} \|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^n}{\text{Min}} \| A\widehat{x} - \widehat{b} \|^2 .$$

Si el sistema de ecuaciones lineales tiene infinitas soluciones, trataremos el problema de norma mínima siguiente:

$$\text{hallar } \widehat{x} \in \mathbb{R}^n \text{ tal que } \| \widehat{x} \|^2 = \underset{\substack{\overrightarrow{x} \in \mathbb{R}^n \\ A\widehat{x} = \widehat{b}}}{\text{Min}} \| \widehat{x} \|^2 .$$

En este capítulo nos ocuparemos de la resolución numérica de estos tres problemas. Particularmente, para los sistemas cuadrados de ecuaciones lineales utilizaremos los métodos directos. Los métodos iterativos y las soluciones en mínimos cuadrados se tratarán más adelante en capítulos separados.

Observación: Si A es una matriz de $n \times n$ invertible, cuando notamos a la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ con $\overrightarrow{x} = A^{-1}\overrightarrow{b}$, donde A^{-1} denota la matriz inversa de A, lo único que queremos indicar es que nuestro sistema de ecuaciones tiene solución única. Esto no quiere decir que debemos calcular la la matriz inversa A^{-1} para hallar su solución $\overrightarrow{x}$. Del punto de vista numérico esto no se hace, es por ello que se buscan métodos para resolver el sistema de ecuaciones que evitan el cálculo de la matriz inversa A^{-1}.

6.3. Algunos tipos de matrices importantes.

En esta sección tratamos principalmente los siguientes tipos de matrices: simétricas definidas positivas, monótonas, estrictamente diagonalmente dominantes, normales, y ortogonales.

Las relaciones de orden $\leq$ y $<$ en el espacio de matrices $M_{n\times n}[\mathbb{R}]$ se definen a continuación. Sean $A=(a_{ij})$, $B=(b_{ij})$ dos matrices de $M_{n\times n}[\mathbb{R}]$. Escribiremos

$$\begin{aligned} A &\leq B \Leftrightarrow a_{ij}\leq b_{ij} \quad i,j=1,...,n,\\ A &< B \Leftrightarrow a_{ij}< b_{ij} \quad i,j=1,...,n. \end{aligned}$$

De acuerdo a las relaciones de orden $\leq$ y $<$ definidas en $M_{n\times n}[\mathbb{R}]$, escribiremos

$$\begin{aligned} A &\geq 0 \Leftrightarrow a_{ij}\geq 0 \quad i,j=1,...,n,\\ A &> 0 \Leftrightarrow a_{ij}> 0 \quad i,j=1,...,n. \end{aligned}$$

En forma similar se definen las relaciones de orden $\leq$, y $<$ en $\mathbb{R}^n$, esto, es, si $\overrightarrow{x}^T=(x_1,...,x_n)$, $\overrightarrow{y}^T=(y_1,...,ny_n)$ son dos elementos de $\mathbb{R}^n$, escribiremos

$$\begin{aligned} \overrightarrow{x} &\leq \overrightarrow{y} \Leftrightarrow x_i\leq y_i \quad i=1,...,n,\\ \overrightarrow{x} &< \overrightarrow{y} \Leftrightarrow x_i< y_i \quad i=1,...,n,\\ \overrightarrow{x} &\geq 0 \Leftrightarrow x_i\geq 0 \quad i=1,...,n,\\ \overrightarrow{x} &> 0 \Leftrightarrow x_i> 0 \quad i=1,...,n. \end{aligned}$$

El producto escalar en $\mathbb{R}^n$ de dos vectores columna $\overrightarrow{x}^T=(x_1,...,x_n)$, $\overrightarrow{y}^T=(y_1,...,y_n)$ s denota $\langle\overrightarrow{x},\overrightarrow{y}\rangle$, o también $\overrightarrow{x}^T\overrightarrow{y}$ o $\overrightarrow{x}\cdot\overrightarrow{y}$ y se define como

$$\langle\overrightarrow{x},\overrightarrow{y}\rangle=\overrightarrow{x}^T\overrightarrow{y}=\overrightarrow{x}\cdot\overrightarrow{y}=\sum_{i=1}^{n}x_iy_i.$$

La norma asociada al producto escalar $\langle\cdot,\cdot\rangle$ se nota $\|\cdot\|$ y se define como

$$\|\overrightarrow{x}\|=\left(\overrightarrow{x}^T\overrightarrow{x}\right)^{\frac{1}{2}}=\left(\sum_{i=1}^{n}x_i^2\right)^{\frac{1}{2}}\qquad\forall\overrightarrow{x}\in\mathbb{R}^n,$$

con $\overrightarrow{x}^T=(x_1,...,x_n)$.

6.3.1. Matrices simétricas definidas positivas.

Sea $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ una matriz simétrica, esto es, $A=A^T$, donde A^T denota la matriz transpuesta de A.

Definición 1 *Consideramos la forma cuadrática q de $\mathbb{R}^n$ en $\mathbb{R}$ definida por $q(\overrightarrow{x})=\overrightarrow{x}^TA\overrightarrow{x}$ $\forall\overrightarrow{x}\in\mathbb{R}^n$.*

i) Se dice que la forma cuadrática q es definida positiva si $q(\overrightarrow{x})>0\quad\forall\overrightarrow{x}\in\mathbb{R}^n,\quad\overrightarrow{x}\neq 0$.

ii) Se dice que la forma cuadrática q es semi-definida positiva si $q(\overrightarrow{x})\geq 0\quad\forall\overrightarrow{x}\in\mathbb{R}^n$.

iii) Se dice que q es definida negativa si $-q$ es definida positiva.

iv) Se dice que q es semi-definida negativa si $-q$ es semi-definida positiva.

Definición 2

i) Diremos que A es definida positiva si la forma cuadrática q es definida positiva.

ii) Diremos que A es semi-definida positiva si la forma cuadrática q es semi-definida positiva.

iii) Diremos que A es definida negativa (semi-definida negativa) si la forma cuadrática q es definida negativa (resp. semi-definida negativa).

Ejemplos

1. Sean $\lambda_1, ..., \lambda_n \in \mathbb{R}$ y $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ la matriz definida como

$$\begin{aligned} a_{ii} &= \lambda_i, \quad i = 1, ..., n, \\ a_{ij} &= 0, \quad i.j = 1, ..., n, \ \ i \neq j. \end{aligned}$$

   La matriz A se llama matriz diagonal y se le denota como $A = diag(\lambda_1, ..., \lambda_n)$. Se tiene que A es simétrica. Además, A es definida positiva si y solo si $\lambda_i > 0, \quad i = 1, ..., n$.

2. Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ una matrtiz no singular. Las matrices $B = A^T A$ y $C = AA^T$ son simétricas, definidas positivas. En efecto, B es simétrica. Pués, $B^T = \left(A^T A\right)^T = A^T \left(A^T\right)^T$ y como $\left(A^T\right)^T = A$, se sigue que $B^T = A^T A = B$. Además, como A no es singular, se tiene $A\overrightarrow{x} = 0 \Leftrightarrow \overrightarrow{x} = 0$. Luego, para $\overrightarrow{x} \in \mathbb{R}^n$ con $\overrightarrow{x} \neq 0$,

$$\overrightarrow{x}^T B\overrightarrow{x} = \overrightarrow{x}^T A^T \overrightarrow{x} = \left(A\overrightarrow{x}\right)^T A\overrightarrow{x} = \|A\overrightarrow{x}\|^2 > 0,$$

   que prueba que B es definida positiva.

   Así, B es simétrica, definida positiva. En forma similar se muestra que C es simétrica, definida positiva.

Consecuencias

Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$.

1. Si A es simétrica, definida positiva, A es no singular.

2. Si A es simétrica, definida positiva, entonces la función $\langle \cdot, \cdot \rangle$ de $\mathbb{R}^n$ en $\mathbb{R}$ definida por

$$\langle A\overrightarrow{x}, \overrightarrow{x} \rangle = \overrightarrow{x}^T A\overrightarrow{x} \quad \forall \overrightarrow{x} \in \mathbb{R}^n,$$

   es un producto escalar en $\mathbb{R}^n$ y la norma asociada a este producto se nota

$$\|\overrightarrow{x}\|_A = \left(\overrightarrow{x}^T A\overrightarrow{x}\right)^{\frac{1}{2}} \qquad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

Teorema 1 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$. Las siguiente proposiciones son equivalentes.*

i) *A es simétrica, definida positiva.*

ii) *Para toda matriz $B \in M_{n\times n}[\mathbb{R}]$ no singular, B^TAB es simétrica definida positiva.*

iii) *Todos los valores propios de A son positivos.*

iv) *A^{-1} es simétrica, definida positiva.*

v) *$a_{ii} > 0 \quad i = 1, ..., n.$*

vi) *$\det(A) > 0$ y $\det(A_k) > 0, \quad k = 1, ..., n$, donde A_k es la matriz de $k \times k$ obtenida de A con las k primeras filas y columnas de A.*

vii) *Se define $A^0 = I$, $A^{m+1} = A^mA$ y $A^{-m} = \left(A^{-1}\right)^m$ para $m \in \mathbb{N}$. Se tiene A^m simétrica, definida positiva, para todo $m \in \mathbb{Z}$.*

viii) *Existe una matriz triangular inferior L no singular tal que $A = LL^T$.*

Demostración. Son resultados conocidos del álgebra lineal. Las demostraciones y más detalles sobre este tema puede encontrar en los textos de Algebra Lineal citados en la bibliografía. ■

6.3.2. Matrices monótonas y diagonalmente dominantes.

Definición 3 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$. Se dice que A es monótona si para $\overrightarrow{x} \in \mathbb{R}^n$, $A\overrightarrow{x} \geq 0 \Longrightarrow \overrightarrow{x} \geq 0$.*

Ejemplo

Sea $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ y $\overrightarrow{x}^T = (x_1, x_2, x_3) \in \mathbb{R}^3$. Entonces,

$$A\overrightarrow{x} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{bmatrix} \geq 0$$

es decir

$$\begin{cases} 2x_1 & -x_2 & & \geq & 0 \\ -x_1 & +2x_2 & -x_3 & \geq & 0 \\ & -x_2 & +2x_3 & \geq & 0 \end{cases}$$

Multiplicando por 2 a la primera desigualdad y sumando con la segunda, obtenemos $3x_1 - x_3 \geq 0 \quad (*)$. De manera similar, multiplicando por 2 a la tercera desugaldad y sumando con la segunda, obtenemos $-x_1 + 3x_3 \geq 0 \quad (**)$. Multiplicando por 3 a (**) y sumando (*) deducimos $8x_3 \geq 0 \Rightarrow x_3 \geq 0$. De (*), se tiene

$$3x_1 \geq x_3 \geq 0 \Rightarrow x_1 \geq 0.$$

Como $-x_1 + 2x_2 - x_3 \geq 0 \Rightarrow 2x_2 \geq x_1 + x_3 \geq 0 \Rightarrow x_2 \geq 0$.

Luego, $A\overrightarrow{x} \geq 0 \Rightarrow \overrightarrow{x} \geq 0$, es decir A es monótona.

Teorema 2 *Sea $A \in M_{n\times n}[\mathbb{R}]$. Entonces, A es monótona si y solo si $A^{-1} \geq 0$.*

Demostración. Supongamos que $A^{-1} \geq 0$. Mostremos que A es monótona. En efecto, sea $\overrightarrow{x} \in \mathbb{R}^n$ y supongamos $A\overrightarrow{x} \geq 0$. Entonces,

$$A^{-1}(A\overrightarrow{x}) \geq A^{-1} \times 0 = 0,$$

de donde

$$\left(A^{-1}A\right)\overrightarrow{x} \geq 0 \Leftrightarrow \overrightarrow{x} \geq 0.$$

Así,

$$A\overrightarrow{x} \geq 0 \Rightarrow \overrightarrow{x} \geq 0.$$

Recíprocamente, supongamos que A es monótona, probemos que A es invertible y que $A^{-1} \geq 0$.

Sea $\overrightarrow{x} \in \mathbb{R}^n$ tal que $A\overrightarrow{x} = 0$. Por ser A monótona, se tiene $\overrightarrow{x} \geq 0$.

Por otro lado, $A(-\overrightarrow{x}) = 0 \Rightarrow -\overrightarrow{x} \geq 0$ o $\overrightarrow{x} \leq 0$. Consecuentemente, $A\overrightarrow{x} = 0 \Rightarrow \overrightarrow{x} = 0$, es decir que $\ker(T_A) = \{0\}$, donde T_A es la aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^n$ definida por $T_A(\overrightarrow{x}) = A\overrightarrow{x}$, que prueba que A es invertible.

Ponemos $A^{-1} = [B_1, ..., B_n]$ con B_j la j-ésima columna de A^{-1} y sea $\left\{\overrightarrow{e}_1^T, ..., \overrightarrow{e}_n^T\right\}$ la base canónica de $\mathbb{R}^n$. Puesto que

$$AA^{-1}\overrightarrow{e}_j = \overrightarrow{e}_j \geq 0 \Rightarrow A^{-1}\overrightarrow{e}_j \geq 0 \quad j = 1, ..., n,$$

pero $A^{-1}\overrightarrow{e}_j = B_j \geq 0$. Luego $A^{-1} \geq 0$. ■

Definición 4 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$.*

i) Se dice que A es estrictamente diagonalmente dominante si y solo si $|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$, $i = 1, ..., n$.

ii) Se dice que A es diagonalmente dominante si y solo si $|a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$, $i = 1, ..., n$.

Ejemplos

1. La siguiente es una matriz estrictamente diagonalmente dominante: $A = \begin{bmatrix} 4 & -1 & 1 & 0 \\ 2 & 5 & 1 & 1 \\ 3 & 2 & 7 & -1 \\ 0 & 1 & 4 & 6 \end{bmatrix}$.

2. La matriz $A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & -1 & 2 \end{bmatrix}$ es diagonalmente dominante.

3. Sean $\tau_k > 0$, $k = 1, ..., m$, $h_i > 0$, $i = 1, ..., n+1$; $A = (a_{ij})$, $B = (b_{ij})$ las matrices que se definen a continuación:

$$\begin{cases} a_{ii} = \frac{1}{h_i} + \frac{1}{h_{i+1}}, & i = 1, ..., n, \\ a_{ii-1} = -\frac{1}{h_i}, & i = 2, ..., n, \\ a_{ii+1} = -\frac{1}{h_{i+1}}, & i = 1, ..., n-1, \\ a_{ij} = 0 \quad \text{si} & |i-j| > 1, \end{cases} \qquad \begin{cases} b_{ii} = \frac{h_i}{3} + \frac{h_{i+1}}{3}, & i = 1, ..., n, \\ b_{i,i-1} = \frac{h_i}{6}, & i = 2, ..., n, \\ b_{ii+1} = \frac{h_{i+1}}{6}, & i = 1, ..., n-1, \\ b_{ij} = 0 \quad \text{si} & |i-j| > 1. \end{cases}$$

Las matrices A y B son tridiagonales con A diagonalmente dominante y B estrictamente diagonalmente dominante. Las matrices $B + \frac{\tau_k}{2}A \quad k = 1, ..., m$, son estrictamente diagonalmente dominantes. Esta clase de matrices surgen en la discretización de ecuaciones en derivadas parciales del tipo parabólico siguiente:

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f.$$

Teorema 3 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$.*

i) Si A es estrictamente diagonalmente dominante, A es no singular.

ii) Si A es estrictamente diagonalmente dominante y simétrica con $a_{ii} > 0$ $i = 1, ..., n$; A es definida positiva.

Demostración. i) Supongamos que A es singular, entonces $\ker(T_A) = \{0\}$ donde T_A es la aplicación lineal definida por $T_A(\overrightarrow{x}) = A\overrightarrow{x}, \quad \forall \overrightarrow{x} \in \mathbb{R}^n$. Sea $\overrightarrow{x}^T = (x_1, ..., x_n) \in \ker(T_A)$ con $\overrightarrow{x} \neq \overrightarrow{0}$ y $k \in \{1, ..., n\}$ tal que $|x_k| = \underset{i=1,...,n}{\text{Max}} |x_i|$. Se tiene $x_k \neq 0$ y como $\overrightarrow{x} \in \ker(T_A), \quad T_A(\overrightarrow{x}) = A\overrightarrow{x} = 0$ y en consecuencia la k-ésima ecuación se escribe

$$a_{k1}x_1 + ... + a_{kk-1}x_k + a_{kk}x_k + a_{kk+1}x_{k+1} + ... + a_{kn}x_n = 0,$$

de donde

$$a_{kk}x_k = -\sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}x_j,$$

y de esta igualdad, tomando el valor absoluto, se tiene

$$|a_{kk}|\,|x_k| = |a_{kk}x_k| = |\sum_{\substack{j=1 \\ j\neq k}}^{n} a_k x_j\,| \leq \sum_{\substack{j=1 \\ j\neq k}} |a_{kj}|\,|x_j|\,,$$

y de esta desigualdad se obtiene la siguiente:

$$|a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|\,\frac{|x_j|}{|x_k|}.$$

Puesto que $|x_j| \leq |x_k| \quad j = 1, ..., n, \ \frac{|x_j|}{|x_k|} \leq 1$. Luego $|a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|\,,$ que contradice la hipótesis A es estrictamente diagonalmente dominante, esto es, $|a_{kk}| > \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| \quad k = 1, ..., n.$

ii) Se propone como ejercicio. ■

Teorema 4 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$. Supóngase que $a_{ii} > 0 \quad i = 1, ..., n, \ a_{ij} \leq 0$ para $i, j = 1, ..., n, \quad i \neq j$; y, A es estrictamente diagonalmente dominante, entonces A es monótona.*

Demostración. Sea D la matriz diagonal definida por $D = \text{diag}(a_{11}, ..., a_{nn})$. Puesto que $a_{ii} > 0, \ i = 1, ..., n, \ D$ es invertible y

$$D^{-1} = diag\left(\frac{1}{a_{11}}, ..., \frac{1}{a_{nn}}\right).$$

Se define $B = I - D^{-1}A = (b_{ij})$. Entonces

$$\begin{aligned} b_{ii} &= 0 \quad i = 1, ..., n, \\ b_{ij} &= -\frac{a_{ij}}{a_{ii}} \geq 0 \quad i, j = 1, ..., n, \ i \neq j. \end{aligned}$$

Como A es estrictamente diagonalmente dominante, A es invertible. De la igualdad $B = I - D^{-1}A$ se sigue que $A = D(I - B)$ o bien $D^{-1}A = I - B$ que muesta que $I - B$ es invertible. Además,

$$A^{-1} = (I - B)^{-1} D^{-1}.$$

Por otro lado, A es estrictamente diagonalmente dominante, entonces

$$a_{ii} > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| = -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} \quad i = 1, ..., n,$$

de donde

$$1 > \sum_{\substack{j=1 \\ j\neq i}} -\frac{a_{ij}}{a_{ii}} = \sum_{\substack{j=1 \\ j\neq i}}^{n} b_{ij},$$

que muestra que la matriz $I - B$ es estrictamente diagonalmente dominante.

Sea $m \in \mathbb{Z}^+$, no es difícil probar que

$$\lim_{m\to\infty} (I-B)^{-1} B^{m+1} = 0.$$

Se define $S_m = \sum_{k=0}^{m} B^k$. Entonces

$$\begin{aligned} S_m - BS_m &= I - B^{m+1} \Longleftrightarrow (I-B) S_m = I - B^{m+1} \\ S_m &= (I-B)^{-1} \left(I - B^{m+1}\right) = (I-B)^{-1} - (I-B)^{-1} B^{m+1}. \end{aligned}$$

Luego

$$\sum_{k=0}^{\infty} b^k = \lim_{m\to\infty} S_m = (I-B)^{-1} - \lim_{m\to\infty} (I-B)^{-1} B^{m+1} = (I-B)^{-1}.$$

Así,

$$A^{-1} = (I-B)^{-1} D^{-1} = \sum_{k=0}^{\infty} B^k D^{-1} \geq 0.$$

■

6.3.3. Matsrices normales y ortogonales.

Definición 5 *Sea $Q = (q_{ij}) \in M_{n\times n}[\mathbb{R}]$.*

i) Se dice que Q es una matriz normal si $QQ^T = Q^T Q$.

ii) Se dice que Q es una matriz ortogonal si $QQ^T = Q^T Q = I$.

Ejemplos

1. Toda matriz simétrica A es una matriz normal. En efecto, como $A = A^T$ se sigue que
$$A^2 = AA = AA^T = A^T A.$$

2. Sea $A \in M_{n\times n}[\mathbb{R}]$ y $Q = A^T A$. Entonces Q es una matriz normal. Pués
$$Q^T = \left(A^T A\right)^T = A^T \left(A^T\right)^T = A^T A = Q,$$
que muestra que Q es una matriz simétrica. En consecuencia, Q es una matriz normal. Note que $Q \in M_{n\times n}[\mathbb{R}]$. De manera similar, la matriz $Q = AA^T$ es simétrica luego Q es una matriz normal. Note que $Q \in M_{m\times m}[\mathbb{R}]$.

3. Sea $\theta \in \mathbb{R}$ y $Q(\theta) = \begin{bmatrix} \cos(\theta) & -\operatorname{sen}(\theta) \\ \operatorname{sen}(\theta) & \cos(\theta) \end{bmatrix}$. Entonces
$$Q(\theta)^T = \begin{bmatrix} \cos(\theta) & \operatorname{sen}(\theta) \\ -\operatorname{sen}(\theta) & \cos(\theta) \end{bmatrix}.$$
Como $\operatorname{sen}^2(\theta) + \cos^2(\theta) = 1$, resulta
$$\begin{aligned} Q(\theta) &= Q(\theta)^T = \begin{bmatrix} \cos(\theta) & -\operatorname{sen}(\theta) \\ \operatorname{sen}(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \cos(\theta) & \operatorname{sen}(\theta) \\ -\operatorname{sen}(\theta) & \cos(\theta) \end{bmatrix} \\ &= \begin{bmatrix} \cos^2(\theta) + \operatorname{sen}^2(\theta) & \\ & \operatorname{sen}^2(\theta) + \cos^2(\theta) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$
Así, $Q(\theta) Q(\theta)^T = I$. De manera similar se obtiene $Q(\theta)^T Q(\theta) = I$. Por lo tanto $Q(\theta)$ es una matriz ortogonal.

La matriz $Q(\theta)$ se llama matriz de rotación.

4. Toda matriz de permutación P es una matriz ortogonal. pues $P^T = P^{-1}$ o bien $PP^T = P^TP = I$.

5. Toda matriz ortogonal es una matriz normal, pero el recíproco, en general, no es cierto. Para ello considérese la matriz $A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & 0 \\ 3 & 3 & 1 \end{bmatrix}$ y $Q = A^TA$. Resulta que Q es una matriz simétrica, por lo tanto Q es una matriz normal. Además

$$\begin{aligned} Q &= A^TA = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 \\ 2 & 2 & 0 \\ 3 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 14 & 2 \\ 14 & 14 & 2 \\ 2 & 2 & 2 \end{bmatrix}, \\ QQ^T &= \begin{bmatrix} 14 & 14 & 2 \\ 14 & 14 & 2 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} 14 & 14 & 2 \\ 14 & 14 & 2 \\ 2 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 396 & 396 & 60 \\ 396 & 396 & 396 \\ 60 & 396 & 12 \end{bmatrix} \neq I. \end{aligned}$$

6. Matriz de Householder. Sea $\overrightarrow{u} \in \mathbb{R}^n$ tal que $\|\overrightarrow{u}\| = 1$. La matriz

$$H = I - 2\overrightarrow{u}\,\overrightarrow{u}^T$$

es ortogonal. Esta matriz H se conoce como matriz de Householder, quien la propuso en 1958. Mostremos que H es ortogonal. En efecto,

$$\begin{aligned} HH^T &= \left(I - 2\overrightarrow{u}\,\overrightarrow{u}^T\right)\left(I - 2\overrightarrow{u}\,\overrightarrow{u}^T\right)^T = \left(I - 2\overrightarrow{u}\,\overrightarrow{u}^T\right)\left(I^T - 2\left(\overrightarrow{u}\,\overrightarrow{u}^T\right)^T\right) \\ &= \left(I - 2\overrightarrow{u}\,\overrightarrow{u}^T\right)\left(I - 2\overrightarrow{u}\,\overrightarrow{u}^T\right) = I - 2I\left(\overrightarrow{u}\,\overrightarrow{u}^T\right) - 2\left(\overrightarrow{u}\,\overrightarrow{u}^T\right)I + 4\left(\overrightarrow{u}\,\overrightarrow{u}^T\right)\left(\overrightarrow{u}\,\overrightarrow{u}^T\right). \end{aligned}$$

Puesto que $I\left(\overrightarrow{u}\,\overrightarrow{u}^T\right) = \overrightarrow{u}\,\overrightarrow{u}^T$, $\left(\overrightarrow{u}\,\overrightarrow{u}^t\right)I = \overrightarrow{u}\,\overrightarrow{u}^T$, y

$$1 = \|\overrightarrow{u}\|^2 = \overrightarrow{u}^T\overrightarrow{u},$$

entonces

$$HH^T = I - 4\overrightarrow{u}\,\overrightarrow{u}^TI + 4\overrightarrow{u}\left(\overrightarrow{u}^T\overrightarrow{u}\right)\overrightarrow{u}^T = I - 4\overrightarrow{u}\,\overrightarrow{u}^T + 4\overrightarrow{u}\,\overrightarrow{u}^T = I.$$

Además, la matriz H es simétrica. Pués,

$$H^T = \left(I - 2\overrightarrow{u}\,\overrightarrow{u}^T\right)^T = I - 2\overrightarrow{u}\,\overrightarrow{u}^T = H.$$

En consecuencia,

$$H^2 = H^TH = HH^T = I.$$

La matriz de Householder H es vital para el desarrollo del método de factorización QR de Householder que se utiliza en la resolución de sistemas de ecuaciones lineales (véase el capítulo de mínimos cuadrado, método de Householder) y el cálculo de valores y vectores propios (véase el capítulo de valores y vectores propios). En la descomposición de Householder, Q es una matriz ortogonal que se construye con la matrices H y R es una matriz triangular superior. Más adelante se verá esta factorización.

Observación.

De la definición de matriz ortogonal se desprende inmediatamente que si Q es tal matriz, Q es invertible y que $Q^{-1} = Q^T$. Como consecuencia de este último resultado, se deduce que la matriz de Householder es invertible, y,

$$H^{-1} = H^T = H.$$

Teorema 5 *Sea $Q \in M_{n\times n}[\mathbb{R}]$. Entonces Q es normal si y solo si*

$$\left\|Q^T\overrightarrow{x}\right\| = \|Q\overrightarrow{x}\| \qquad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

Demostración. Supongamos que Q es normal. Entonces $QQ^T = Q^TQ$. Luego, para todo $\overrightarrow{x} \in \mathbb{R}^n$,

$$\left\|Q^T\overrightarrow{x}\right\|^2 = \left(Q^T\overrightarrow{x}\right)^T Q^T\overrightarrow{x} = \overrightarrow{x}^T\left(Q^T\right)^T Q^T\overrightarrow{x} = \overrightarrow{x}^T QQ^T\overrightarrow{x} = \overrightarrow{x}Q^TQ\overrightarrow{x} = \left(Q\overrightarrow{x}\right)^T Q\overrightarrow{x} = \left\|Q\overrightarrow{x}\right\|^2.$$

Tomando en cuenta que la norma es no negativa, se sigue que

$$\left\|Q^T\overrightarrow{x}\right\| = \left\|Q\overrightarrow{x}\right\| \quad \forall\overrightarrow{x} \in \mathbb{R}^n.$$

Recíprocamente, supongamos que $\left\|Q^T\overrightarrow{x}\right\| = \left\|Q\overrightarrow{x}\right\| \quad \forall\overrightarrow{x} \in \mathbb{R}^n$. Se tiene

$$\begin{aligned}\left\|Q^T\overrightarrow{x}\right\|^2 &= \left\|Q\overrightarrow{x}\right\|^2 \Leftrightarrow \left(Q^T\overrightarrow{x}\right)^T\left(Q^T\overrightarrow{x}\right) = \left(Q\overrightarrow{x}\right)^T Q\overrightarrow{x} \Leftrightarrow \overrightarrow{x}^T QQ^T\overrightarrow{x} = \overrightarrow{x}^T Q^TQ\overrightarrow{x} \\ &\Leftrightarrow \overrightarrow{x}^T\left(QQ^T - Q^TQ\right)\overrightarrow{x} = 0.\end{aligned}$$

Así, $\overrightarrow{x}^T\left(QQ^T - Q^TQ\right)\overrightarrow{x} = 0 \quad \forall\overrightarrow{x} \in \mathbb{R}^n$, de donde

$$QQ^T - Q^TQ = 0 \Leftrightarrow QQ^T = Q^TQ.$$

■

Teorema 6 *Sean Q_1, Q_2 dos matrices ortogonales. Entonces Q_1Q_2 es una matriz ortogonal.*

Demostración. Si Q_1, Q_2 son matrices ortogonales, se tiene

$$\begin{aligned}Q_1Q_1^T &= Q_1^TQ_1 = I, \\ Q_2Q_2^T &= Q_2^TQ_2 = I.\end{aligned}$$

Luego,

$$\left(Q_1Q_2\right)^T Q_1Q_2 = Q_2^TQ_1^TQ_1Q_2 = Q_2^T\left(Q_1^TQ_1\right)Q_2 = Q_2^TIQ_2 = Q_2^TQ_2 = I.$$

De manera similar se prueba que $Q_1Q_2\left(Q_1Q_2\right)^T = I$. Por lo tanto Q_1Q_2 es una matriz ortogonal. ■

Nota: Si Q_1, Q_2 son matrices normales, en general, Q_1Q_2 no es una matriz normal. Exhibimos dos ejemplos, uno en el que el resultado es verdadero y otro en el que el resultado es falso.

1. Consideremos Q_1, Q_2 las matrices $Q_1 = \begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix}$, $Q_2 = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}$. Entonces

$$\begin{aligned}Q_1Q_1^T &= \begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & 1 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = 10I, \\ Q_1Q_1^T &= \begin{bmatrix} 3 & 1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix} = 10I.\end{aligned}$$

   Luego, $Q_1^TQ_1 = Q_1Q_1^T$, es decir, Q_1 es una matriz normal. De modo similar, tenemos

$$Q_2Q_2^T = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = 5I = Q_2^TQ_2,$$

   que muestra que Q_2 es normal.

   Ahora,

$$Q_1Q_2 = \begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -7 \\ 7 & 1 \end{bmatrix},$$

   y

$$\begin{aligned}\left(Q_1Q_2\right)^T Q_1Q_2 &= \begin{bmatrix} 1 & 7 \\ -7 & 1 \end{bmatrix}\begin{bmatrix} 1 & -7 \\ 7 & 1 \end{bmatrix} = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} = 50I, \\ \left(Q_1Q_2\right)\left(Q_1Q_2\right)^T &= \begin{bmatrix} 1 & -7 \\ 7 & 1 \end{bmatrix}\begin{bmatrix} 1 & 7 \\ -7 & 1 \end{bmatrix} = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} = 50I.\end{aligned}$$

   Resulta que Q_1Q_2 es una matriz normal.

2. Sean $Q_1 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$, $Q_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$. Entonces

$$\begin{aligned} Q_1 Q_1^T &= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 2I, \\ Q_1^T Q_1 &= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 2I, \\ Q_2 Q_2^T &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = Q_2^T Q_2. \end{aligned}$$

Sea $A = Q_1 Q_2 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. Luego

$$\begin{aligned} AA^T &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \\ A^T A &= \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}. \end{aligned}$$

Claramente $AA^T \neq A^T A$. El producto de dos matrices normales, no es en general, una matriz normal como acabamos de comprobar.

Teorema 7 *Sea $Q \in M_{n\times n}[\mathbb{R}]$. Las tres proposiciones siguientes son equivalentes:*

i) $Q^T Q = I$,

ii) $(Q\overrightarrow{x})^T Q\overrightarrow{y} = \overrightarrow{x}^T \overrightarrow{y} \quad \forall \overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n$,

iii) $\|Q\overrightarrow{x}\| = \|\overrightarrow{x}\| \quad \forall \overrightarrow{x} \in \mathbb{R}^n$.

Demostración. i) $\Rightarrow$ ii.) Supongamos que $Q^T Q = I$ sean $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n$. Entonces

$$(Q\overrightarrow{x})^T Q\overrightarrow{y} = \overrightarrow{x}^T Q^T Q\overrightarrow{y} = \overrightarrow{x}^T I\overrightarrow{y} = \overrightarrow{x}^T \overrightarrow{y}.$$

ii) $\Rightarrow$ iii.) Sean $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n$. Si $(Q\overrightarrow{x})^T Q\overrightarrow{y} = \overrightarrow{x}^T \overrightarrow{y}$, en particular para $\overrightarrow{x} = \overrightarrow{y}$, se tiene

$$\|Q\overrightarrow{x}\|^2 = (Q\overrightarrow{x})^T Q\overrightarrow{x} = \overrightarrow{x}^T \overrightarrow{x} = \|\overrightarrow{x}\|^2,$$

de donde

$$\|Q\overrightarrow{x}\| = \|\overrightarrow{x}\| \quad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

iii) $\Rightarrow$ i.) Si $\|Q\overrightarrow{x}\| = \|\overrightarrow{x}\| \quad \forall \overrightarrow{x} \in \mathbb{R}^n$, se sigue que:

$$\begin{aligned} \|Q\overrightarrow{x}\|^2 &= \|\overrightarrow{x}\|^2 \Leftrightarrow (Q\overrightarrow{x})^T Q\overrightarrow{x} = \overrightarrow{x}^T \overrightarrow{x} \Leftrightarrow \overrightarrow{x}^T Q^T Q\overrightarrow{x} = \overrightarrow{x}^T I\overrightarrow{x} \\ &\Leftrightarrow \overrightarrow{x}^T \left(Q^T Q - I\right) \overrightarrow{x} = 0 \quad \forall \overrightarrow{x} \in \mathbb{R}^n, \end{aligned}$$

de donde $Q^T Q = I$. ■

Teorema 8 *Sea $Q \in M_{n\times n}[\mathbb{R}]$. Entonces, Q es ortogonal si y solo si se satisfacen las dos condiciones siguientes:*

i) Q es invertible.

ii) $\|Q\overrightarrow{x}\| = \|\overrightarrow{x}\| \quad \forall \overrightarrow{x} \in \mathbb{R}^n$.

Demostración. Supongamos que Q es ortogonal. Entonces $Q^T Q = QQ^T = I$, de donde $Q^T = Q^{-1}$, esto es, Q es una matriz invertible.

Sea $\overrightarrow{x} \in \mathbb{R}^n$. Entonces

$$\|Q\overrightarrow{x}\|^2 = (Q\overrightarrow{x})^T Q\overrightarrow{x} = \overrightarrow{x}^T Q^T Q\overrightarrow{x} = \overrightarrow{x}^T I\overrightarrow{x} = \overrightarrow{x}^T\overrightarrow{x} = \|\overrightarrow{x}\|^2,$$

con lo cual

$$\|Q\overrightarrow{x}\| = \|\overrightarrow{x}\| \qquad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

Así, si Q es ortogonal, se tiene $i) \Rightarrow ii)$.

Recíprocamente, supongamos que se satisfacen i) y ii). Mostremos que Q es ortogonal. Por i) se tiene que Q es invertible y por ii)

$$\|Q\overrightarrow{x}\| = \|\overrightarrow{x}\| \qquad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

Por el teorema precedente se deduce que $Q^TQ = I$ que implica $Q^T = Q^{-1}$. Luego

$$QQ^T = QQ^{-1} = I = Q^TQ,$$

es decir que Q es una matriz ortogonal. ■

6.4. Métodos directos de resolución de sistemas de ecuaciones lineales.

En lo sucesivo consideraremos sistemas de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, donde $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$ y $\overrightarrow{b} \in \mathbb{R}^n$. Como ya se indicó en el capítulo 1, llamamos método directo de resolución del sistema de ecuaciones lineales, un método que conduce a la solución del problema al cabo de un número finito de pasos, o bien en un número finito de operaciones aritméticas elementales (suma, resta, multiplicación , división y raíz cuadrada) que es función de la dimensión n del sistema.

En cada método directo, se debe estimar: el número de operaciones elementales necesarias en la ejecución del algoritmo, y, la exactitud y precisión del método.

i) El número de operaciones elementales necesarias en la ejecución del algoritmo es una función que depende de la dimensión n de la matriz cuadrada A, a esta función se le denota $Noper$: $\mathbb{Z}^+ \to \mathbb{R}$ que a cada $n \in \mathbb{Z}^+$ asocia $Noper\ (n)$ que expresa el número total de operaciones elementales.

ii) La exactitud y precisión del método dependen sobre todo del condicionamiento de la matriz y de la estabilidad del método, es decir que pequeños errores en los datos de entrada provocan pequeños errores en los datos de salida, o lo que es lo mismo, es insensible a la propagación de errores de redondeo. En el apéndice se muestra el número de condicionamineto.

Para cada método estudiado se debe elaborar un algoritmo numérico, en lo posible, el más óptimo.

En este capítulo se tratarán los métodos clásicos de resolución de sistemas de ecuaciones lineales siguientes:

1.- Método de eliminación gaussiana simple, con pivoting parcial y con pivoting total.

2.- Método de facturación LU.

3.- Método de Choleski.

El método de Householder será estudiado en el capítulo de mínimos cuadrados, sin embargo, en este capítulo introducimos algunos resultados acerca de las matrices ortogonales.

Por otro lado, utilizando el método de eliminación gaussiana se propone un algoritmo de cálculo de la matriz inversa de A y otro para el cálculo del determinante de A. Adicionalmente, par matrices $A = (a_{ij}) \in M_{m\times n}[\mathbb{R}]$ no nulas, $\overrightarrow{b} \in \mathbb{R}^m$, se consideran sistemas de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, y se buscan soluciones en mínimos cuadrados si el sistema de ecuaciones no tiene solución; y, en norma mínima si el sistema de ecuaciones posee infinitas soluciones.

La idea general de los métodos directos de resolución de sistemas de ecuaciones lineales es la factorización de la matriz $A \in M_{n\times n}[\mathbb{R}]$ en la forma LR**,** donde L es una matriz elegida apropiadamente, R una matriz triangular superior. Además, L y R son matrices invertibles. Entonces

$$A\overrightarrow{x} = \overrightarrow{b} \Leftrightarrow LR\overrightarrow{x} = \overrightarrow{b}.$$

Sea $\overrightarrow{y} = R\overrightarrow{x}$, entonces $L\overrightarrow{y} = \overrightarrow{b}$. Así,

$$A\overrightarrow{x} = \overrightarrow{b} \Leftrightarrow \begin{cases} L\overrightarrow{y} = \overrightarrow{b}, \\ R\overrightarrow{x} = \overrightarrow{y}, \end{cases} \tag{15}$$

que muestra que si la matriz A se factora en la forma LR, la resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ es equivalente a los siguientes:

$$L\overrightarrow{y} = \overrightarrow{b} \tag{16}$$

y a continuación

$$R\overrightarrow{x} = \overrightarrow{y} \tag{17}$$

Además, como L, R son invertibles, se tiene

$$\overrightarrow{y} = L^{-1}\overrightarrow{b}$$

y

$$\overrightarrow{x} = R^{-1}\overrightarrow{y}.$$

Luego

$$\overrightarrow{x} = R^{-1}\overrightarrow{y} = R^{-1}\left(L^{-1}\overrightarrow{b}\right) = \left(R^{-1}L^{-1}\right)\overrightarrow{b} = (LR)^{-1}\overrightarrow{b} = A^{-1}\overrightarrow{b},$$

es decir que la solución $\overrightarrow{x}$ del par de sistemas de ecuaciones (16) y (17) es la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ y recíprocamente. Por otro lado, los sistemas de ecuaciones 16) y 17) son muy sencillos de resolver. Más aún, comenzaremos nuestro estudio de solución de sistemas de ecuaciones lineales de los tipos (16) y (17) denominados triangulares inferiores y superiores, respectivamente. A continuación se describen los métodos de factorización de matrices A en la forma LR.

6.4.1. Sistemas de ecuaciones lineales triangulares superiores e inferiores.

Definición 6 *Sean* $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, \ldots, b_n) \in \mathbb{R}^n$.

i) Se dice que A es una matriz triangular superior si y solo si los coeficientes de A satisfacen la siguiente condición:

$$a_{ij} = 0 \quad \text{para } i > j, \quad i = 2, ...n, \;\; j = 1, ...n. \tag{18}$$

ii) Si A es una matriz triangular superior, se dice que el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ es triangular superior.

iii) Se dice que A es una matriz triangular inferior si y solo si los coefiecientes de A satisfacen la condición siguiente:

$$a_{ij} = 0 \quad \text{para } \;\; j > i, \quad i = 1, ..., n, \;\; j = 2, ...n. \tag{19}$$

iv) Si A es una matriz triangular inferior, el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ se llama sistema triangular inferior.

Ejemplos

1. La matriz A siguiente es triangular superior. $A = \begin{bmatrix} 3 & 5 & 0 & 1 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & -3 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix}$.

2. El siguiente es un sistema de ecuaciones triangular superior: $\left\{\begin{array}{rrrrcr} 2x+ & 3y+ & 5z- & w & = & 1 \\ & y- & z+ & w & = & 2 \\ & & 3z- & w & = & -1 \\ & & & 5w & = & 20. \end{array}\right.$

   Note que la matriz A del sistema de ecuaciones es $A = \begin{bmatrix} 2 & 3 & 5 & -1 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 5 \end{bmatrix}$ que es triangular superior.

3. El siguiente es un sistema de ecuaciones lineales triangular inferior: $\left\{\begin{array}{rrrrcr} 4x & & & & = & 2 \\ 3x & -2y & & & = & -1 \\ x & +2y & +3z & & = & 0 \\ -x & -2y & -2z & -w & = & -2 \end{array}\right.$

   Observe que la matriz A del sistema de ecuaciones es $A = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 3 & -2 & 0 & 0 \\ 1 & 2 & 3 & 0 \\ -1 & -2 & -2 & -1 \end{bmatrix}$ que es triangular inferior.

Resolución numérica.

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ una matriz triangular superior, $\overrightarrow{b} \in \mathbb{R}^n$. Consideramos el sistema de ecuaciones lineales triangular superior:

$$A\overrightarrow{x} = \overrightarrow{b},$$

o en forma explícita

$$\left\{\begin{array}{rcl} a_{11}x_1 + \cdots + a_{1n}x_n & = & b_1 \\ & \vdots & \\ a_{nn}x_n & = & b_n. \end{array}\right.$$

Los valores x_i, $i = 1, ..., n$, de la solución $\overrightarrow{x} = (x_1, ...x_n)$, siempre que esta exista, se obtienen mediante el procedimiento de "vuelta atrás", siguiente:

$$\begin{array}{rcll} x_n & = & \frac{b_n}{a_{nn}} & a_{nn} \neq 0, \\ x_{n-1} & = & \frac{b_{n-1}-a_{n-1n}x_n}{a_{n-1,n-1}}, & a_{n-1,n-1} \neq 0, \\ & \vdots & & \\ x_j & = & \frac{1}{a_{jj}}\left(b_j - \sum_{k=j+1}^{n} a_{jk}x_k\right), & a_{jj} \neq 0, \\ & \vdots & & \\ x_1 & = & \frac{1}{a_{11}}\left(b_1 - \sum_{k=2}^{n} a_{1k}x_k\right), & a_{11} \neq 0. \end{array}$$

Del cálculo de $x_1, ..., x_n$, se deduce que el sistema de ecuaciones lineales triangular superior tiene solución única si y solo si $a_{ii} \neq 0$, $i = 1, ..., n$.

Para la resolución de este sistema de ecuaciones se requieren ejecutar los números siguientes de operaciones elementales:

$$\begin{array}{rcl} \text{Productos} & : & 0 + 1 + ... + n - 1 = \dfrac{n(n-1)}{2}, \\ \text{Adiciones} & : & 0 + 1 + ... + n - 1 = \dfrac{n(n-1)}{2}, \\ \text{Divisiones} & : & n. \end{array}$$

El número total de operaciones para el cálculo de $\overrightarrow{x}$ solución de $A\overrightarrow{x} = \overrightarrow{b}$ es

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

Ejemplo

Resolver el sistema de ecuaciones $\left\{\begin{array}{rrrrcr} -3x & +2y & -z & +u & = & -1 \\ & 0{,}5y & +z & -u & = & 3 \\ & & -4z & +u & = & 0 \\ & & & 2u & = & 1. \end{array}\right.$ Aplicando el procedimiento de "vuelta atrás", tenemos

$$\begin{aligned} u &= \frac{1}{2} = 0{,}5, \\ z &= \frac{0 - 1 \times 0{,}5}{-4} = 0{,}125, \\ y &= \frac{3 - 0{,}125 - (-1) \times 0{,}5}{0{,}5} = 6{,}75, \\ x &= \frac{-1 - 2 \times 6{,}75 - (-1) \times 0{,}125 - 0{,}5}{-3} = 4{,}625. \end{aligned}$$

La solución es $\overrightarrow{x}^T = (4{,}625, 6{,}75, 0{,}125, 0{,}5)$.

Algoritmo

El procedimiento de "vuelta atrás" descrito precedentemente permite elaborar el siguiente algoritmo:

Datos de entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$.

Datos de salida: $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$, Mensaje: "Matriz A singular".

1. Si $a_{nn} \neq 0$, $x_n = \frac{b_{nn}}{a_{nn}}$.

   Caso contrario, imprimir mensaje. Continuar en 4).

2. Para $j = n-1, ..., 1$

   Si $a_{jj} \neq 0.$,

   $S = 0.$

   $k = j+1,\ n$

   $S = S + a_{jk}x_k$

   $x_j = \frac{(b_j - S)}{a_j}$

   Fin de bucle k.

   Caso contrario, imprimir mensaje. Continuar en 4)

   Fin de bucle j.

3. Imprimir $\overrightarrow{x}^T = (x_1, ..., x_n)$. Continuar en 5).

4. Mensaje: matriz singular.

5. Fin.

Tratamos continuación los sistemas de ecuaciones lineales triangulares inferiores.

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ una matriz triangular inferior, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$. Consideremos el sistema de ecuaciones triangular inferior $A\overrightarrow{x} = \overrightarrow{b}$, o escrito en forma explícita

$$\left\{\begin{array}{lcl} a_{11}x_1 & = & b_1, \\ & \vdots & \\ a_{n1}x_1 \quad +\ldots+ a_{nn}x_n & = & b_n, \end{array}\right.$$

cuya solución $\overrightarrow{x}^T = (x_1, ..., x_n)$, siempre que esta exista, puede calcularse con el procedimiento siguiente:

$$\begin{cases} x_1 = \frac{b_1}{a_{11}} & , \quad a_{11} \neq 0, \\ x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=1}^{j-1} a_{jk} x_k \right), & a_{jj} \neq 0, j = 2, ..., n. \end{cases}$$

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$.

Datos de salida: $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$, Mensaje: "Matriz A singular".

1. Si $a_{11} \neq 0$, $x_1 = \frac{b_1}{a_{11}}$

   Caso contrario, imprimir mensaje. Continuar en 5).

2. Para $j = 2, ..., n$

   Si $a_{jj} \neq 0.$,

   $S = 0.$

   $k = 1, ..., j-1$

   $S = S + a_{jk} x_k$

   $x_j = \frac{(b_j - S)}{a_{jj}}$

   Fin de bucle k.

   Caso contrario, imprimir mensaje. Continuar en 5).

   Fin de bucle j.

4. Imprimir $\overrightarrow{x}^T = (x_1, ..., x_n)$. Continuar en 6).

5. Mensaje: matriz singular.

6. Fin.

El número de operaciones elementales, para hallar la solución de un sistema de ecuaciones triangular inferior es:

$$Noper\,(n) = n^2.$$

Debe notarse que si A es una matriz triangular superior o inferior singular, el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ puede tener infinitas soluciones o ninguna solución. Se propone como ejercicio el estudio de estas dos situaciones y la implementación correspondiente en los algoritmos descritos precedentemente.

Nota**:** Si $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ es una matriz triangular superior, los elementos de la diagonal principal de A; a_{jj}, $j = 1, ..., n$, se llaman pivotes de la matriz triangular, y si $\overrightarrow{b} \in \mathbb{R}^n$, a_{jj} se llaman pivotes del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

Consecuencias

El cálculo del determinante de una matriz $A \in M_{n\times n}[\mathbb{R}]$ triangular superior se establece en el siguiente teorema.

Teorema 9 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ una matriz triangular superior. Entonces,* $\det(A) = \prod_{j=1}^{n} a_{ij}$.

Demostración. Procedemos por inducción.

Si $n = 1$, $A = (a_{11})$ y en consecuencia $\det(A) = a_{11}$.

La hipótesis inductiva establece que si $A=(a_{ij})\in M_{(n-1)\times(n-1)}[\mathbb{R}]$, entonces $\det(A)=\prod_{j=1}^{n-1}a_{jj}$.

Probemos que es cierto para $n\geq 1$. Sea $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ una matriz triangular superior.

Desarrollamos el determinante de A por elementos de la primera columna. Se tiene

$$\det(A)=a_{11}A_{11}+...+a_{n1}A_{n1},$$

donde $A_{i1},\ \ i=1,...,n$, es el cofactor de a_{i1}, $i=1,...,n$. Puesto que A es triagular superior, $a_{21}=...=a_{n1}=0$, y

$$a_{11}A_{11}=a_{11}(-1)^2\det(A_{n-1})=a_{11}\det(A_{n-1}),$$

donde $\det(A_{n-1})$ es el determinante de la matriz obtenida de A al eliminar la primera fila y la primera columna. Por la hipótesis inductiva, $\det(A_{n-1})=\prod_{j=2}^{n}a_{jj}$. Luego

$$\det(A_{n-1})=a_{11}\prod_{j=2}^{n}a_{jj}=\prod_{j=1}^{n}a_{jj}.$$

■

Un resultado similar se obtiene cuando $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ es una triangular inferior. Se tiene $\det(A)=\prod_{j=1}^{n}a_{jj}$.

En el siguiente teorema se establece un método para el cálculo de la matriz inversa de una matriz triangular superior invertible.

Teorema 10 *Sea $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ una matriz triangular superior invertible. Entonces, la j-ésima columna de A^{-1} es solución del sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{e_j}$, donde $\{\overrightarrow{e_1},...,\overrightarrow{e_n}\}$ es la base canónica de $\mathbb{R}^n$, y $\overrightarrow{e_j}^T=(0,...,1,...,0)$. Además, A^{-1} es triangular superior.*

Demostración. Sea $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ una matriz triangular superior invertible. Pongamos $A^{-1}=\left[A_1^{(-1)},...,A_n^{(-1)}\right]$, donde $A_j^{(-1)}$ denota la j-ésima columna de A^{-1}.

Sea $\overrightarrow{x}$ la solución del sistema de ecuaciones lineales

$$A\overrightarrow{x}=\overrightarrow{e_j},\quad j=1,...,n.$$

Multiplicando por A^{-1}, se tiene

$$A^{-1}A\overrightarrow{x}=A^{-1}\overrightarrow{e_j},$$

y tomando en cuenta que $AA^{-1}=A^{-1}A=I$, se sigue que

$$\overrightarrow{x}=\left[A_1^{(-1)},...,A_n^{(-1)}\right]\overrightarrow{e_j}=A_j^{(-1)}.$$

Mostremos que A^{-1} es triangular superior.

Sean $\overrightarrow{b}\in\mathbb{R}^n$ y $\overrightarrow{x}$ la solución del sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{b}$. Entonces

$$\begin{cases} x_n=\dfrac{b_n}{a_{nn}} \\ \quad\vdots \\ x_1=\dfrac{1}{a_{11}} & \left(b_1-\sum_{j=2}^{n}a_{1j}x_j\right)\end{cases}$$

En particular, si $\overrightarrow{b}=\overrightarrow{e_j}\quad j=1,...,n$, se tiene: para $j=1$, $\overrightarrow{b}^T(1,0,...,0)$, $x_2=0,...,x_n=0$, $x_1=\frac{b_1}{a_{11}}$

$$A_1^{(-1)}=\begin{bmatrix}\frac{1}{a_{11}}\\ \vdots\\ 0\end{bmatrix}.$$

Continuando con este procedimiento, para $j = n$ se tiene

$$\begin{cases} x_n = \dfrac{1}{a_{nn}}, \\ \quad \vdots \\ x_1 = \dfrac{1}{a_{11}} \quad \left(0 - \sum\limits_{j=2}^{n} a_{1j}x_j\right), \end{cases}$$

con lo cual $A_n^{(-1)} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ de donde $A^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & -\frac{1}{a_{11}}\frac{a_{12}}{a_{22}} & & -\frac{1}{a_{11}}\sum\limits_{a=2}^{n} a_{1j}x_j \\ 0 & \frac{1}{a_{22}} & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{a_{nn}} \end{bmatrix}$ que muestra que A^{-1} es triangular superior. ■

Un resultado similar se establece para matrices triangulares inferiores invertibles, esto es, si $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ es triangular inferior invertible, A^{-1} es triangular inferior y cada columna de A^{-1} es solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{e_j}$, $j = 1, ..., n$.

Si se toma en consideración que cada columna de A^{-1} es solución de sistema de ecuaciones triangular superior $A\overrightarrow{x} = \overrightarrow{e_j}$, el número de operaciones elementales para calcular $\overrightarrow{x}$ es n^2. Luego, para calcular A^{-1} se requieren de n^3 operaciones elementales, esto es,

$$Noper\,(n) = n^3.$$

El procedimiento descrito en el teorema precedente para el cálculo de A^{-1} es práctico pero no óptimo. Si se observa la matriz A^{-1}, es claro que se puede elaborar un algoritmo óptimo y reducir el número de operaciones elementales. Un algoritmo de cálculo de A^{-1}, no óptimo, basado el el teorema precedente, es el siguiente:

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+, A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$.

Datos de salida: Mensaje 1: "A no es triangular superior". Mensaje 2: "A no es invertible", $B = A^{-1} = (b_{ij})$.

1. Para $i = 2, ..., n$,

   Para $j = 1, ..., n$

   Si $i > j$ y $|a_{ij}| > 0$, imprimir mensaje 1, continuar en 5)

. Fin de bucle j.

Fin de bucle i.

2. Para $i = 1, ..., n$

   Si $|a_{ii}| \leq 0$, imprimir mensaje 2, continuar en 5).

Fin de bucle i.

3. Para $j = 1, ..., n$

   Resolver el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{e_j}$.

   Para $i = 1, ..., n$,

   $b_{ij} = x_i$

Fin de bucle i.

Fin de bucle j.

4. Imprimir $A^{-1} = B$.

5. Fin

Ejercicios

1. Mejorar el algoritmo precedente en en punto 3.

2. Elaborar un algoritmo, el mejor posible, para calcular A^{-1}. Estimar $Noper\,(n)$.

6.5. Operaciones elementales con matrices.

La matriz identidad $I = (e_{pq}) \in M_{n\times n}[\mathbb{R}]$ está definida como $e_{pq} = \begin{cases} 1, & \text{si } p = q, \ \ p, q = 1, ..., n, \\ 0, & \text{si } \ p \neq q. \end{cases}$

Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ una matriz no nula.

i) Intercambio de dos filas de A

Sea $F_{i,j} = (f_{pq}) \in M_{n\times n}[\mathbb{R}]$ la matriz obtenida de I al intercambiar la fila i con la fila j, $\ i < j$, esto es,

$$f_{pq} = \begin{cases} 1, \ \text{si} \ \ p = q, \ \ p, q = 1, ..., n, \ \ p \neq i, j \ \ \text{o} \ \ p = j, \ \ q = i, \ \ \text{o} \ \ p = i, q = j. \\ 0, \ \text{en otro caso.} \end{cases}$$

La matriz $A_{i,j}$ que se obtiene de A al intercambiar la fila i con la fila j de la matriz A esta definida como

$$A_{i,.j} = F_{i,j}A.$$

Ejemplo

Sean $A = \begin{bmatrix} 2 & 3 & 5 & 8 \\ -1 & -3 & -4 & 2 \\ 0 & 1 & 0 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$ y $F_{2,3} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Entonces

$$A_{2,3} = F_{2,3}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 5 & 8 \\ -1 & -3 & -4 & 2 \\ 0 & 1 & 0 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 5 & 8 \\ 0 & 1 & 0 & 1 \\ -1 & -3 & -4 & 2 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

Observe que la tercera fila de A ha remplazado a la segunda fila de A y esta ha ocupado la tercera fila. Note que la matriz $F_{i,j}$ es no singular y $\det(F_{i,j}) = -1$.

ii) Intercambio de dos columnas de A.

Sea $C_{i,j} = (C_{pq}) \in M_{n\times n}[\mathbb{R}]$ la matriz definida como

$$c_{pq} = \begin{cases} 1, \ \text{si} \ \ p = q, \ \ p, q = 1, ..., n, \ \ \text{o} \ \ p = i, \ \ q = j, \ \ \text{o} \ \ p = j, q = i. \\ 0, \quad \text{en otro caso.} \end{cases}$$

La matriz B_{ij} que se obtiene de A al intercambiar la columna i con la j de la matriz A está definida como:

$$B_{i,j} = AC_{i,j}, \quad i < j.$$

Ejemplo

Sean $A = \begin{bmatrix} 2 & 4 & 6 & 8 \\ -3 & -6 & -9 & -12 \\ 5 & 10 & 15 & 20 \\ 3 & 6 & 9 & 12 \end{bmatrix}$ y $C_{14} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$. Entonces

$$B_{1,4} = AC_{1,4} = \begin{bmatrix} 2 & 4 & 6 & 8 \\ -3 & -6 & -9 & -12 \\ 5 & 10 & 15 & 20 \\ 3 & 6 & 9 & 12 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 8 & 4 & 6 & 2 \\ -12 & -6 & -9 & -3 \\ 20 & 10 & 15 & 5 \\ 12 & 0 & 9 & 3 \end{bmatrix}.$$

Observe que la matriz $B_{1,4}$ se obtiene al intercambiar la primera con la cuarta columna de A. La matriz $C_{i,j}$ es no singular y $\det(C_{i,j}) = -1$.

iii) Producto de una constante α por una fila o columna de A.

Sean $1 \leq p \leq n, \quad \alpha \in \mathbb{R}$. Se define $C_p = (c_{ij}) \in M_{n\times n}[\mathbb{R}]$ la matriz definida como

$$c_{ij} = \begin{cases} \alpha, & \text{si } \ i = j = p, \\ 1, & \text{si } \ i = j;\ i, j = 1, ..., n, \quad i \neq p, \ \ j \neq p, \\ 0, & \text{en otro caso.} \end{cases}$$

El producto de la constante α por la fila p de A se define como $A_p = C_pA$, o sea

$$A_p = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & \alpha & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1n} \\ \vdots & & & & \vdots \\ a_{p1} & \cdots & \cdots & \cdots & a_{pn} \\ \vdots & & & & \vdots \\ a_{n1} & \cdots & \cdots & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1n} \\ \vdots & & & & \vdots \\ \alpha a_{p1} & \cdots & \cdots & \cdots & \alpha a_{pn} \\ \vdots & & & & \vdots \\ a_{n1} & \cdots & \cdots & \cdots & a_{nn} \end{bmatrix}.$$

El producto de la constante α por la columna p de A se define como $B_p = AC_p$. Para $\alpha \neq 0$, la matriz C_p es invertible.

Ejemplo

Sean $A = \begin{bmatrix} 1 & 6 & 3 \\ 5 & 7 & 7 \\ 8 & 9 & 10 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Entonces

$$A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 6 & 3 \\ 5 & 7 & 7 \\ 8 & 9 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 100 & 120 & 140 \\ 8 & 9 & 10 \end{bmatrix}.$$

Note que la segunda fila de A_2 se obtiene al multiplicar la constante $\alpha = 20$ con la segunda fila de A.

$$B_2 = \begin{bmatrix} 1 & 6 & 3 \\ 5 & 7 & 7 \\ 8 & 9 & 10 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 40 & 3 \\ 5 & 120 & 7 \\ 8 & 180 & 10 \end{bmatrix}.$$

Observe que la segunda columna de B_2 se obtiene al multiplicar la constante $\alpha = 20$ con los elementos de la seguna columna de A.

iv) Producto de α por la fila i y suma del resultado con la fila j de A.

Sean $1 \leq i \leq j \leq n, \quad \alpha \in \mathbb{R}$. Se definen $K_{i,j} = (k_{pq}) \in M_{n\times n}[\mathbb{R}]$ como sigue

$$k_{pq} = \begin{cases} 1, & \text{si } \ p = q, \ \ p, q = 1, ..., n, \\ a, & \text{si } \ p = i, \ \ q = j, \\ 0, & \text{en otro caso,} \end{cases}$$

y

$$\widetilde{A}_{i,j} = K_{i,j}A.$$

La matriz $\widetilde{A}_{i,j}$ se obtiene al multiplicar la fila i de A por α y el resultado se suma con la fila j de A.

Ejemplo

Sean $A = \begin{bmatrix} 2 & 8 & -1 & 2 \\ 5 & 10 & 15 & -20 \\ 3 & 6 & 9 & 12 \\ 1 & -1 & 0 & 1 \end{bmatrix}$, $K_{1,3} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \alpha & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Entonces

$$\widetilde{A}_{1,3} = K_{1,3}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \alpha & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 8 & -1 & 2 \\ 5 & 10 & 15 & -20 \\ 3 & 6 & 9 & 12 \\ 1 & -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 8 & -1 & 2 \\ 5 & 10 & 15 & -20 \\ 2\alpha+3 & 8\alpha+6 & -\alpha+9 & 2\alpha+12 \\ 0 & - & 0 & 1 \end{bmatrix}.$$

Note que la tercera fila de $\widetilde{A}_{1,3}$ es el resultado de multiplicar la primera fila de A por α y luego sumar con la tercera fila.

Sea $K_{2,4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \alpha & 0 & 1 \end{bmatrix}$. Entonces

$$\widetilde{A}_{2,4} = K_{2,4}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \alpha & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 8 & -1 & 2 \\ 5 & 10 & 15 & -20 \\ 3 & 6 & 9 & 12 \\ 1 & -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 8 & -1 & 2 \\ 5 & 10 & 15 & -20 \\ 3 & 6 & 9 & 12 \\ 5\alpha+1 & 10\alpha+-1 & 15\alpha & -20\alpha+1 \end{bmatrix}.$$

Observe que la cuarta fila de $\widetilde{A}_{2,4}$ es el resultado de multiplicar la segunda fila de A por α y este resultado sumar con la cuarta fila de A.

6.6. Método de eliminación gaussiana.

Sean $A = (a_{ij})\, M_{n\times n}[\mathbb{R}]$ no nula, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$. Consideramos el sistema de ecuaciones lineales

$$A\overrightarrow{x} = \overrightarrow{b}. \tag{1}$$

La matriz ampliada $\widetilde{A} = (\widetilde{a}_{ij})$ se define como

$$\widetilde{a}_{ij} = \begin{cases} a_{ij}, & \text{si } \ i = 1, .., n \quad j = 1, ..., n, \\ b_i, & \text{si } \ i = 1, ..., n, \quad j = n+1. \end{cases} \tag{2}$$

Claramente $\widetilde{A} \in M_{n\times n}[\mathbb{R}]$. Esta matriz $\widetilde{A}$ se le representa como $\widetilde{A} = \left[A \mid \overrightarrow{b}\right]$, que en forma explícita se escribe

$$\widetilde{A} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & b_n \end{array}\right], \tag{3}$$

y ponemos $\widetilde{A}_0 = \widetilde{A}$.

En la resolución de los sistemas de ecuaciones lineales solo los elementos de la matriz A y los componentes del vector $\overrightarrow{b}$ intervienen. La idea general del método del eliminación gaussiana es la siguiente: partiendo de la matriz ampliada $\widetilde{A}_0$, construir en $n-1$ etapas una matriz $\widetilde{A}_{n-1}$ de la forma

$$\left[\begin{array}{ccc|c} a_{11}^{(n-1)} & \cdots & a_{1n}^{(n-1)} & b_1^{(n-1)} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right]$$

que es equivalente al sistema de ecuaciones lineales triangular superior

$$\left\{\begin{array}{rcl} a_{11}^{(n-1)}x_1 + ... \quad +a_{1n}^{(n-1)}x_n & = & b_1^{(n-1)} \\ & \vdots & \\ a_{nn}^{(n-1)}x_n & = & b_n^{(n-1)} \end{array}\right. \tag{4}$$

que como se ha dicho, es muy sencillo de resolverlo.

Comenzamos con el método de eliminación gaussiana sin pivoting que se explica a continuación. Debemos enfatizar en esta parte que no estamos muy interesados en la precisión de la solución y lo que queremos es describir los elementos que intervienen en el algoritmo de la eliminación gaussiana.

6.6.1. Eliminación gaussiana sin pivoting.

Con el propósito de presentar las ideas fundamentales del método de eliminación gaussiana, consideramos primeramente en ejemplo.

Sean $A = \begin{bmatrix} 1 & 5 & -1 & 0 \\ 2 & 2 & 0 & 0 \\ -2 & 1 & -1 & 4 \\ 3 & 6 & 2 & 7 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} -3 \\ 2 \\ -1 \\ 7 \end{bmatrix}$. Formamos la matriz ampliada $\widetilde{A} = \left[A \mid \overrightarrow{b}\right]$, esto es,

$$\widetilde{A} = \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 2 & 2 & 0 & 0 & 2 \\ -2 & 1 & -1 & 4 & -1 \\ 3 & 6 & 2 & 7 & 7 \end{array}\right].$$

Para transformar la matriz $\widetilde{A}$ en $n-1$ etapas (en este caso 3 etapas) en una matriz de la forma

$$\left[\begin{array}{ccc|c} a_{11}^{(3)} & \cdots & a_{14}^{(3)} & b_1^{(3)} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & a_{44}^{(3)} & b_4^{(3)} \end{array}\right]$$

utilizamos las operaciones elementales con matrices, y particularmente, la de multiplicar a una fila por una constante y el resultado sumar a otra fila.

Etapa 1

Si $a_{11} \neq 0$, se define $k_{i1} = -\dfrac{a_{i1}}{a_{11}}$, $i = 2, 3, 4$, y K_1 es la matriz siguiente:

$$K_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ k_{21} & 1 & 0 & 0 \\ k_{31} & 0 & 1 & 0 \\ k_{41} & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{bmatrix}.$$

Resulta que K_1 es una matriz triangular inferior invertible. Ponemos

$$\widetilde{A} = K_1\widetilde{A}_0 = K_1\left[A \mid \overrightarrow{b}\right] = \left[K_1A \mid K_1\overrightarrow{b}\right].$$

Entonces

$$\widetilde{A}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{bmatrix}\left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 2 & 2 & 0 & 0 & 2 \\ -2 & 1 & -1 & 4 & -1 \\ 3 & 6 & 2 & 7 & 7 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 0 & -8 & 2 & 0 & 8 \\ 0 & 11 & -3 & 4 & -7 \\ 0 & -9 & 5 & 7 & 16 \end{array}\right].$$

Esta matriz $\widetilde{A}_1$ lo notamos $\left[A_1 \mid \overrightarrow{b}_1\right]$ o también$\left(a_{ij}^{(1)}\right)$, esto es, $\widetilde{A} = \left[A_1 \mid \overrightarrow{b}_1\right] = \left(a_{ij}^{(1)}\right)$, donde $A_1 = K_1 A$ y $\overrightarrow{b} = K_1 \overrightarrow{b}$.

Note que la construcción de la matriz K_1 conduce a que en la matriz $\widetilde{A}_1 = K_1\widetilde{A}$ se hagan ceros los elementos de la primera columna bajo el primer elemento de la diagonal de A.

Etapa 2

Si $a_{22}^{(1)} \neq 0$, se define $k_{i2} = -\dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}}$, $i = 3, 4$. En nuestro caso particular $a_{22}^{(1)} = -8$.

Se define $K_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & k_{34} & 1 & 0 \\ 0 & k_{42} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{11}{8} & 1 & 0 \\ 0 & -\frac{9}{8} & 0 & 1 \end{bmatrix}$. La matriz K_2 es triangular inferior invertible.

Se define

$$\widetilde{A}_2 = K_2\widetilde{A}_1 = K_2\left[A_1 \mid \overrightarrow{b}_1\right] = \left[K_2A_1 \mid K_2\overrightarrow{b}_1\right].$$

Entonces

$$\widetilde{A}_2 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{11}{8} & 1 & 0 \\ 0 & -\frac{9}{8} & 0 & 1 \end{array}\right] \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 0 & -8 & 2 & 0 & 8 \\ 0 & 11 & -3 & 4 & -7 \\ 0 & -9 & 5 & 7 & 16 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 0 & -8 & 2 & 0 & 8 \\ 0 & 0 & -\frac{1}{4} & 4 & 4 \\ 0 & 0 & \frac{11}{4} & 7 & 7 \end{array}\right].$$

Ponemos $\widetilde{A}_2 = \left[A_2 \mid \overrightarrow{b}_2\right] = \left(a_{ij}^{(2)}\right)$, donde $A_2 = K_2A_1$, $\quad \overrightarrow{b}_1 = K_1\overrightarrow{b}_1$.

Observe que la construcción de la matriz K_2 hace que la matriz aumentada $\widetilde{A}_2 = K_2\widetilde{A}_1$ conserve los ceros de la primera columna bajo el elemento de la diagonal y se hagan ceros los elementos de la segunda columna bajo el elemento de la diagonal de la matriz A_1.

Etapa 3

Si $a_{33}^{(2)} \neq 0$, se define $k_{i3} = -\dfrac{a_{i3}^{(2)}}{a_{33}^{(2)}}$, $i = 4$. En nuestro caso $a_{33}^{(2)} = -\frac{1}{4}$.

Sea $K_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & k_{43} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 11 & 1 \end{bmatrix}$. La matriz K_3 es triangular inferior invertible. Ponemos

$$\widetilde{A}_3 = K_2\widetilde{A}_2 = K_3\left[A_2 \mid \overrightarrow{b}_2\right] = \left[K_3A_2 \mid K_3\overrightarrow{b}_2\right].$$

Luego

$$\widetilde{A}_3 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 11 & 1 \end{array}\right] \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 0 & -8 & 2 & 0 & 8 \\ 0 & 0 & -\frac{1}{4} & 4 & 4 \\ 0 & 0 & \frac{11}{4} & 7 & 7 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 0 & -8 & 2 & 0 & 8 \\ 0 & 0 & -\frac{1}{4} & 4 & 4 \\ 0 & 0 & 0 & 51 & 51 \end{array}\right].$$

Ponemos $\widetilde{A}_3 = \left[A_3 \mid \overrightarrow{b}_3\right]$, es decir que $A_3 = K_3A_2$ y $\overrightarrow{b}_3 = K_3\overrightarrow{b}_2$ que se indican a continuación:

$$A_3 = \begin{bmatrix} 1 & 5 & -1 & 0 \\ 0 & -8 & 2 & 0 \\ 0 & 0 & -\frac{1}{4} & 4 \\ 0 & 0 & 0 & 51 \end{bmatrix}, \qquad \overrightarrow{b}_3 = \begin{bmatrix} -3 \\ 8 \\ 4 \\ 51 \end{bmatrix}.$$

La matriz A_3 es triangular superior que lo notamos R. Entonces

$$\begin{aligned} R &= A_3 = K_3A_2 = K_3K_2A_1 = K_3K_2K_1A, \\ \overrightarrow{b}_3 &= K_3\overrightarrow{b}_2 = K_3K_2\overrightarrow{b}_1 = K_3K_2K_1\overrightarrow{b}. \end{aligned}$$

Puesto que $K_1,\ K_2,\ K_3$ son matrices triangulares inferiores invertibles, el producto $K_3K_2K_1$ es una matriz triangular invertible. Ponemos $E = K_3K_2K_1$. Entonces E^{-1} es una matriz triangular inferior, y $R = EA,\quad \overrightarrow{b}_3 = E\overrightarrow{b}$ de donde $A = E^{-1}R,\quad \overrightarrow{b} = E^{-1}\overrightarrow{b}_3$. Consecuentemente,

$$A\overrightarrow{x} = \overrightarrow{b} \Leftrightarrow E^{-1}R\overrightarrow{x} = E^{-1}\overrightarrow{b}_3 \Longleftrightarrow R\overrightarrow{x} = \overrightarrow{b}_3,$$

es decir que la solución $\overrightarrow{x} \in \mathbb{R}^4$ de $A\overrightarrow{x} = \overrightarrow{b}$ es la solución de $R\overrightarrow{x} = \overrightarrow{b}_3$ y recíprocamente. Además, es sistema de ecuaciones $R\overrightarrow{x} = \overrightarrow{b}_3$ es triangular superior que en forma explícita se escribe:

$$\begin{cases} x_1 & +5x_2 & -5x_3 & & = & -3 \\ & -8x_2 & +2x_3 & & = & 8 \\ & & -\frac{1}{4}x_3 & +4x_4 & = & 4 \\ & & & 51x_4 & = & 51, \end{cases}$$

cuya solución es $\overrightarrow{x}^T = (2, -1, 0, 1)$.

Conclusión: Si en cada etapa del proceso de eliminación gaussiana, no existe intercambio de filas o columnas, y los elementos pivotes a_{11}, $a_{22}^{(1)}$, $a_{33}^{(2)}$, $a_{44}^{(3)}$ son no nulos, la matriz A se factora en la forma LR, donde $L = E^{-1}$ es triangular inferior, R triangular superior. Además, el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ es equivalente al triangular superior $R\overrightarrow{x} = \overrightarrow{b}_3$.

Generalicemos las ideas expuestas en el ejemplo.

Etapa 1

Supongamos que $a_{11} \neq 0$. Ponemos $a_{11}^{(0)} = a_{11}$. Se definen

$$k_{i1} = -\frac{a_{i1}}{a_{11}}, \quad i = 2, ..., n,$$

y sea K_1 la matriz siguiente:

$$K_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ k_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & 0 & \cdots & 1 \end{bmatrix}.$$

Definimos $\widetilde{A}_1 = K_1\widetilde{A}$. Entonces

$$\widetilde{A}_1 = K_1\widetilde{A} = K_1\left[A \mid \overrightarrow{b}\right] = \left[K_1A \mid K_1\overrightarrow{b}\right] = \left[A_1 \mid \overrightarrow{b}_1\right],$$

esto es

$$\begin{aligned} \widetilde{A}_1 &= \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ k_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & 0 & \cdots & 1 \end{array}\right] \left[\begin{array}{cccc|c} a_{11} & \cdots & \cdots & a_{1n} & b_1 \\ a_{21} & \cdots & \cdots & a_{2n} & b_2 \\ \vdots & & & \vdots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} & b_n \end{array}\right] \\ &= \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & k_{21}a_{12} + a_{22} & \cdots & k_{21}a_{1n} + a_{2n} & k_{21}b_1 + b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & k_{n1}a_{12} + a_{n1} & \cdots & k_{n1}a_{1n} + a_{nn} & k_{n1}b_1 + b_n \end{array}\right]. \end{aligned}$$

Note el efecto de la construcción de la matriz K_1 en $\widetilde{A}_1 = K_1\widetilde{A}$ al obtener en esta última ceros bajo el elemento de la diagonal en la primera columna.

Etapa 2

Ponemos $\widetilde{A}_2 = \left(a_{ij}^{(1)}\right)$ y supongamos que $a_{22}^{(1)} \neq 0$. Se define

$$k_{i2} = -\frac{a_{i2}^{(1)}}{a_{22}^{(1)}} \quad i = 3, ..., n,$$

y sea $K_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & k_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & k_{n2} & 0 & \cdots & 1 \end{bmatrix}$. Definimos $\widetilde{A}_{12} = K_2\widetilde{A}_1$. Entonces

$$\widetilde{A}_2 = K_2\widetilde{A}_1 = K_2\left[A_1 \mid \overrightarrow{b}_1\right] = \left[K_2A_1 \mid K_2\overrightarrow{b}_1\right] = \left[A_2 \mid \overrightarrow{b}_2\right],$$

$$\widetilde{A}_2 = \left[\begin{array}{ccccc|c} 1 & 0 & 0 & \cdots & 0 & \\ 0 & 1 & 0 & \cdots & 0 & \\ 0 & k_{32} & 1 & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & k_{n2} & 0 & \cdots & 1 & \end{array}\right] \left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & a_{21}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} & b_3^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right]$$

$$= \left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ 0 & 0 & k_{32}a_{23}^{(1)} + a_{33}^{(1)} & \cdots & k_{32}a_{2n}^{(1)} + a_{3n}^{(1)} & k_{32}b_2^{(1)} + b_3^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & k_{n2}a_{23}^{(1)} + a_{n3}^{(1)} & \cdots & k_{n2}a_{2n}^{(1)} + a_{nn}^{(1)} & k_{n2}b_2^{(1)} + b_n^{(1)} \end{array}\right].$$

Ponemos $\widetilde{A}_2 = \left(a_{ij}^{(2)}\right)$.

Etapa j-ésima

Supongamos $a_{jj}^{(j-1)} \neq 0 \quad j = 1, ..., n-1$. Se define

$$k_{ij} = -\frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}} \quad i = j+1, ..., n, \qquad K_j = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & k_{j+1,j} & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & k_{n,j} & 0 & \cdots & 1 \end{bmatrix},$$

y $\widetilde{A}_j = K_j\widetilde{A}_{j-1}$. Entonces

$$\widetilde{A}_j = K_j\widetilde{A}_{j-1} = K_j\left[A_{j-1} \mid \overrightarrow{b}_{j-1}\right] = \left[K_jA_{j-1} \mid K_j\overrightarrow{b}_{j-1}\right] = \left[A_j \mid \overrightarrow{b}_j\right],$$

donde $\widetilde{A}_j$ tiene la forma siguiente:

$$\widetilde{A}_j = \left[\begin{array}{ccccccc|c} a_{11}^{(j)} & a_{12}^{(j)} & \cdots & a_{1j}^{(j)} & \cdots & \cdots & a_{1n}^{(j)} & b_1^{(j)} \\ \vdots & & & \vdots & & & \vdots & \vdots \\ 0 & \cdots & \cdots & a_{jj}^{(j)} & a_{jj+1}^{(j)} & & a_{jn}^{(j)} & b_j^{(j)} \\ \vdots & & & 0 & a_{j+1,j+1}^{(j)} & & a_{j+1,n}^{(j)} & b_{j+1}^{(j)} \\ \vdots & & & \vdots & & & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & a_{n,j+1}^{(j)} & & a_{n,n}^{(j)} & b_n^{(1)} \end{array}\right].$$

Etapa n-1

Para $j = n-1$, si $a_{n-1,n-1}^{(n-1)} \neq 0$, se define $k_{n,n-1} = -\dfrac{a_{nn-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}}$, $K_{n-1} = \begin{bmatrix} 1 & \cdots & \cdots & 0 & 0 \\ \vdots & & & \vdots & \vdots \\ \vdots & & & \vdots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \\ 0 & \cdots & \cdots & k_{n,n-1} & 1 \end{bmatrix}$,

$$\widetilde{A}_{n-1} = K_{n-1}\widetilde{A}_{n-2} = \left[K_{n-1}A_{n-2} \mid K_{n-1}\overrightarrow{b}_{n-2}\right] = \left[A_{n-1} \mid \overrightarrow{b}_{n-1}\right],$$

donde

$$\widetilde{A}_{n-1} = \left[\begin{array}{ccc|c} a_{11}^{(n-1)} & \cdots & a_{1n}^{(n-1)} & b_1^{(n-1)} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right].$$

La matriz A_{n-1} es triangular superior, lo notaremos con R.

Suponemos $a_{nn}^{(n-1)} \neq 0$. Se tiene

$$\begin{aligned} R &= A_{n-1} = K_{n-1}A_{n-2} = \cdots = K_{n-1}K_{n-2}\cdots K_1 A, \\ \overrightarrow{b}_{n-1} &= K_{n-1}\overrightarrow{b}_{n-2} = \cdots = K_{n-1}K_{n-2}\cdots K_1\overrightarrow{b}. \end{aligned}$$

Las matrices $K_1, ..., K_{n-1}$ son triangulares invertibles. Luego, el producto $K_{n-1}K_{n-2}\cdots K_1$ es triangular inferior invertible que lo notamos E. Entonces, E^{-1} es triangular inferior, y

$$R = EA, \qquad \overrightarrow{b}_{n-1} = E\overrightarrow{b},$$

de donde $A = E^{-1}R$, $\overrightarrow{b} = E^{-1}\overrightarrow{b}_{n-1}$. Notando $L = E^{-1}$, se tiene $A = LR$, $\overrightarrow{b} = L\overrightarrow{b}_{n-1}$.

Así, si todos los elementos pivotes $a_{jj}^{(j-1)} \neq 0$, $j = 1, ..., n-1$ con $a_{11}^{(0)} = a_{11}$, la matriz A se factora en la forma LR, con L una matriz triangular inferior invertible, R matriz triangular superior invertible. Luego,

$$A\overrightarrow{x} = \overrightarrow{b} \Leftrightarrow LR\overrightarrow{x} = L\overrightarrow{b}_{n-1} \Leftrightarrow R\overrightarrow{x} = \overrightarrow{b}_{n-1},$$

es decir que la solución $\overrightarrow{x}$ de $A\overrightarrow{x} = \overrightarrow{b}$ es solución de $R\overrightarrow{x} = \overrightarrow{b}_{n-1}$ y recíprocamente.

El sistema de ecuaciones $R\overrightarrow{x} = \overrightarrow{b}$ es triangular superior que en forma explícita se escribe:

$$\left\{\begin{array}{rcl} a_{11}^{(n-1)}x_1 + \cdots + a_{1n}^{(n-1)}x_n &=& b_1^{(n-1)} \\ & \vdots & \\ a_{nn}^{(n-1)}x_n &=& b_n^{(n-1)}, \end{array}\right.$$

cuya solución se calcula mediante el algoritmo de resolución de sistemas de ecuaciones triangulares superiores.

Algoritmo de la eliminación gaussiana sin pivoting

El procedimiento descrito precedentemente para transformar la matriz A en una triangular superior R y el vector $\overrightarrow{b}$ en $\overrightarrow{b}_{n-1}$ se recoge en el siguiente algoritmo denominado algoritmo de eliminación gaussiana sin pivoting para la resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

Algoritmo

Datos de Entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$.

Datos de Salida: Mensaje 1: "Error: matriz nula". Mensaje 2: "Pivote nulo". Solución $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$.

1. $S = 0$.

2. Para $i = 1, ..., n$

   Para $j = 1, ..., n$ Verificación de no nulidad de la matriz A.

   Si $|a_{ij}| \leq 0$,

   $S = S + j$

   Fin de bucle j.

   Fin de bucle i.

3. Si $S = n^2$, imprimir mensaje 1: "Error: matriz nula". Continuar en 9).

4. Para $i = 1, ..., n$ Matriz ampliada.

   $a_{i,n+1} = b_i$.

   Fin de bucle i.

5. Para $j = 1, ..., n-1$

   Si $a_{jj} \neq 0$ entonces Transformación en un sistema triangular superior.

   Para $i = j + 1, ..., n$

   $$k_{ij} = -\frac{a_{ij}}{a_{jj}}$$

   Para $r = j + 1, ..., n + 1$

   $$a_{ir} = k_{ij} a_{jr} + a_{ir}$$

   Fin de bucle r.

   Fin de bucle i

   Caso contrario, continuar en 8).

   Fin de bucle j.

6. Resolver el sistema triangular superior $R\overrightarrow{x} = \overrightarrow{b}_{n-1}$.

7. Escribir $\overrightarrow{x}$. Continuar en 9).

8. Escribir mensaje 2: "Pivote nulo".

9. Fin.

Número de operaciones elementales.

En la primera etapa se ejecutan las siguientes operaciones elementales:

$$\text{divisiones}: \ n-1, \quad \text{adiciones}: \ n \times (n-1), \quad \text{productos}: \ n \times (n-1).$$

En la segunda etapa se realizan las siguientes operaciones elementales:

$$\text{divisiones}: \ n-2, \quad \text{adiciones}: \ (n-2) \times (n,2), \quad \text{productos}: \ (n-2) \times (n-2).$$

En la j-ésima etapa, se tiene

$$\text{divisiones}: \ n-j, \quad \text{adiciones}: \ (n-j)(n-j+1), \quad \text{productos}: \ (n-j)(n-j+1).$$

En la última etapa se ejecutan las siguientes operaciones elementales:

$$\text{divisiones}: \ 1, \quad \text{adiciones}: \ 2, \quad \text{productos}: \ 2.$$

Total de operaciones elementales:

$$\begin{aligned}
\text{divisiones} &: \quad 1+2+...+n-1=\frac{(n-1)}{2},\\
\text{adiciones} &: \quad 2+...+n(n-1)=\sum_{j=1}^{n-1}(n-j)(n-j+1),\\
\text{productos} &: \quad 2+...+n(n-1)=\sum_{j=1}^{n-1}(n-j)(n-j+1).
\end{aligned}$$

Se tiene

$$\begin{aligned}
\sum_{j=1}^{n-1}(n-j)(n-j+1) &= \sum_{j=1}^{n-1}\left[n^2+n-(2n+1)j+j^2\right]\\
&= \left(n^2+n\right)(n-1)-(2n+1)\frac{n(n-1)}{2}+\frac{1}{6}(n-1)n(2n-1)\\
&= \frac{n(n-1)(n+1)}{3}.
\end{aligned}$$

La resolución del sistema de ecuaciones triangular superior involucra n^2 operaciones elementales. Consecuentemente, el número total de operaciones elementales que se requieren para resolver el sistema de ecuaciones lineales con el método de eliminación gaussiana se denota con $Noper(n)$ dado por:

$$Noper(n)=\frac{n(n-1)}{2}+2\frac{n(n-1)(n+1)}{3}+n^2=\frac{n}{6}\left(4n^2+9n-7\right).$$

Así, para $n=3$, $Noper(3)=28$, $Noper(4)=62$, $Noper(5)=115$.

6.6.2. Eliminación gaussiana con pivoting.

En el método de eliminación gaussiana sin pivoting descrito precedentemente, si el elemento $a_{jj}^{(j-1)}$ es nulo, el proceso se detiene $\left(j=1,...,n-1,\ \text{con}\ a_{11}^{(0)}=a_{11}\right)$. Una mejora al algoritmo es buscar en los elementos de la columna $a_{ij}^{(j-1)}$, $i=j+1,...,n$, aquel elemento no nulo y efectuar el intercambio de la fila i-ésima con la fila j-ésima, $i>j$.

En la práctica se muestra que este intercambio no basta pués no mejora la precisión de la solución. Para mejorar la precisión de la solución es preciso introducir la estrategia del pivoting parcial y total.

Pivoting parcial

Sean $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ con $A\neq 0$, con $\overrightarrow{b}^T=(b_1,...,b_n)\in\mathbb{R}^n$; y consideramos el sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{b}$. Construimos la matriz ampliada $\widetilde{A}=\left[A\mid\overrightarrow{b}\right]$ y ponemos $\widetilde{A}_0=\widetilde{A}$.

Etapa 1

Sea r el entero positivo tal que $1\leq r\leq n$ y $|a_{r1}|=\underset{i=1,...,n}{\text{Max}}|a_{i1}|$.

Si $a_{r1}=0$ entonces $a_{i1}=0$, $i=1,...,n$, con loque la matriz A es singular. El proceso de eliminación gaussiana concluye.

Si $|a_{r1}|>0$, intercambiamos las filas $i=1$ con la fila $i=r$ de la matriz $\widetilde{A}_0$. Este proceso se realiza mediante la operación elemental entre matrices que notamos

$$B_0=F_{1,r}\widetilde{A}_0=(b_{ij}),$$

donde F_{1r} es la matriz obtenida de la matriz identidad al intercambiar la fila $i = 1$ con la fila $i = r$. A continuación definimos $k_{i1} = -\frac{b_{i1}}{b_{11}} \quad i = 2, ..., n$, la matriz K_1 siguiente:$K_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ k_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ k_{n1} & 0 & \cdots & 1 \end{bmatrix}$, y

$$\widetilde{A}_1 = K_1 B_0 = \left(a_{ij}^{(1)}\right).$$

Etapa 2

Sea r el entero positivo tal que $2 \leq r \leq n$ y $\left|a_{r2}^{(1)}\right| = \underset{i=2,...,n}{\text{Max}} \left|a_{i2}^{(1)}\right|$.

Si $a_{r2}^{(1)} = 0$ entonces $a_{i2}^{(1)} = 0 \quad i = 2, ..., n$, con lo que la matriz A es singular. El proceso de eliminación gaussiana concluye.

Si $\left|a_{r2}^{(1)}\right| > 0$, intercambiamos las filas $1 = 2$ con la fila $i = r$ de la matriz $\widetilde{A}_1$. Para el efecto, definimos

$$B_1 = F_{2,r}\widetilde{A}_1 = \left(b_{ij}^{(1)}\right),$$

donde $F_{2,1}$ es la matriz obtenida de la matriz identidad al intercambiar la fila 2 con la fila r, y a continuación definimos $k_{i2} = -\frac{b_{i2}^{(1)}}{b_{22}^{(1)}} \quad i = 3, ..., n, \qquad K_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & k_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & k_{n2} & 0 & \cdots & 1 \end{bmatrix}$, y sea

$$\widetilde{A}_2 = K_2 B_1 = \left(a_{ij}^{(2)}\right).$$

Este proceso continua hasta la etapa $j = n - 1$.

Etapa n-1

Sea r es entero positivo tal que $n - 1 \leq r \leq n$ y $\left|a_{r,n-1}^{(n-2)}\right| = \underset{i=n-1,n}{\text{Max}} \left|a_{i,n-1}^{(n-2)}\right|$.

Si $a_{r,n-1}^{(n-2)} = 0$, la matriz A es singular. El procesos de eliminación gaussiana concluye.

Si $\left|a_{r,n-1}^{(n-2)}\right| > 0$, intercambiamos la fila $i = n - 1$ con r, definimos $B_{n-2} = F_{n-1,n}\widetilde{A}_{n-2} = \left(b_{ij}^{(n-1)}\right)$, $k_{n,n-1} = -\frac{b_{n,n-1}^{(n-2)}}{b_{n-1,n-1}}, \quad K_{n-1} = \begin{bmatrix} 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \ddots & & \vdots & 0 \\ \vdots & & \ddots & \vdots & 0 \\ 0 & \vdots & \vdots & 1 & \vdots \\ 0 & \cdots & \cdots & k_{n,n-1} & 1 \end{bmatrix}$, y sea

$$\widetilde{A}_{n-1} = K_{n-1} B_{n-2} = \left(a_{ij}^{(n-1)}\right).$$

Resulta

$$\widetilde{A}_{n-1} = K_{n-1}\widetilde{B}_{n-2} = K_{n-2}F_{n-1,n}\widetilde{A}_{n-2} = \cdots = K_{n-2}F_{n-1,n} \cdots K_1 F_{1,r}\widetilde{A}.$$

Cada matriz $K_i \quad i = 1, ..., n-1$ y cada $F_{i,r}, \quad i = 1, ..., n-1, \quad r \in \{i, ..., n\}$, son invertibles.

Ponemos $E = K_{n-1}F_{n-1,n}...K_1F_{1,r}$. Entonces

$$\widetilde{A}_{n-1} = E\widetilde{A} = E\left[A \mid \overrightarrow{b}\right] = \left[EA \mid E\overrightarrow{b}\right]$$

y $\widetilde{A}_{n-1} = \left[A_{n-1} \mid \overrightarrow{b}\, n-1\right]$, donde A_{n-1} es una matriz triangular superior que se le nota R.

Luego,

$$\begin{cases} A_{n-1} = EA \\ \overrightarrow{b}_{n-1} = E\overrightarrow{b}. \end{cases}$$

Así, $A = E^{-1}R$. Note que E^{-1} es una matriz triangular inferior.

Si $a_{n,n}^{(n-1)} \neq 0$, la matriz $R = A_{n-1}$ es invertible y en consecuencia A se factora en la forma LR, donde $L = E^{-1}$, y,

$$A\overrightarrow{x} = \overrightarrow{b} \Leftrightarrow A_{n-1}\overrightarrow{x} = \overrightarrow{b}_{n-1}.$$

El sistema de ecuaciones lineales $A_{n-1}\overrightarrow{x} = \overrightarrow{b}_{n-1}$ es triangular superior, cuyo método de resolución ha sido descrito anteriormente.

Para la elaboración del algoritmo de eliminación gaussiana con pivoting parcial debemos tener en cuenta, en cada etapa, el proceso de intercambio de la fila i con la fila r de la matria $\widetilde{A}_{j-1}$, $j = 1, ..., n-1$.

Antes de proponer el algoritmo, exhibimos un ejemplo que muestre el proceso descrito en el método de eliminación gaussiana con pivoting parcial.

Ejemplo

Sean $A = \begin{bmatrix} 1 & 5 & -1 & 0 \\ 2 & 2 & 0 & 0 \\ -2 & 1 & -1 & 4 \\ 3 & 6 & 2 & 7 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} -3 \\ 2 \\ -1 \\ 7 \end{bmatrix}$, y consideremos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$.

La solución de este ejemplo fue determinada con el método de eliminación gaussiana sin pivoting. La matriz ampliada $\widetilde{A}$ está dada por $\widetilde{A} = \left[\begin{array}{cccc|c} 1 & 5 & -1 & 0 & -3 \\ 2 & 2 & 0 & 0 & 2 \\ -2 & 1 & -1 & 4 & -1 \\ 3 & 6 & 2 & 7 & 7 \end{array}\right]$.

Etapa 1

Selección del pivoting: $3 = |a_{41}| = \underset{i=1,2,3,4}{\text{Max}} |a_{i,1}|$, $r = 4$. Intercambio de las filas 1 con la 4 de $\widetilde{A}$. Se tiene $B_1 = \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 2 & 2 & 0 & 0 & 2 \\ -2 & 1 & -1 & 4 & -1 \\ 1 & 5 & -1 & 0 & -3 \end{array}\right]$. Se definen las matrices siguientes: $K_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{2}{3} & 1 & 0 & 0 \\ \frac{2}{3} & 0 & 1 & 0 \\ -\frac{1}{3} & 0 & 0 & 1 \end{bmatrix}$,

$$\widetilde{A}_1 = K_1B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{2}{3} & 1 & 0 & 0 \\ \frac{2}{3} & 0 & 1 & 0 \\ -\frac{1}{3} & 0 & 0 & 1 \end{bmatrix} \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 2 & 2 & 0 & 0 & 2 \\ -2 & 1 & -1 & 4 & -1 \\ 1 & 5 & -1 & 0 & -3 \end{array}\right] = \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & -2 & -\frac{4}{3} & -\frac{14}{3} & -\frac{8}{3} \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & 3 & -\frac{5}{3} & -\frac{7}{3} & -\frac{16}{3} \end{array}\right].$$

Etapa 2

Selección del pivoting: $5 = \left|a_{32}^{(1)}\right| = \underset{i=3,4}{\text{Max}} \left|a_{i,2}^{(1)}\right|$, $r = 5$. Intercambio de las filas 2 con 3 de $\widetilde{A}_1$. Se tiene $\left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & -2 & -\frac{4}{3} & -\frac{14}{3} & -\frac{8}{3} \\ 0 & 3 & -\frac{5}{3} & -\frac{7}{3} & -\frac{16}{3} \end{array}\right]$. Se definen $K_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{5} & 1 & 0 \\ 0 & -\frac{3}{5} & 0 & 1 \end{bmatrix}$,

$$\widetilde{A}_2 = K_2 B_2 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{5} & 1 & 0 \\ 0 & -\frac{3}{5} & 0 & 1 \end{array}\right] \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & -2 & -\frac{4}{3} & -\frac{14}{3} & -\frac{8}{3} \\ 0 & 3 & -\frac{5}{3} & -\frac{7}{3} & -\frac{16}{3} \end{array}\right] = \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & 0 & -\frac{6}{5} & -\frac{6}{5} & -\frac{6}{5} \\ 0 & 0 & -\frac{28}{15} & -\frac{113}{15} & -\frac{113}{15} \end{array}\right].$$

Etapa 3

Selección del pivoting. $\frac{28}{5} = \left|a_{43}^{(2)}\right| = \underset{i=4}{\text{Max}} \left|a_{i3}^{(2)}\right|$, $r = 4$. Intercambiamos las filas 3 y 4 de $\widetilde{A}_2$. Tenemos $B_3 = \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & 0 & -\frac{28}{15} & -\frac{113}{15} & -\frac{113}{15} \\ 0 & 0 & -\frac{6}{5} & -\frac{6}{5} & -\frac{6}{5} \end{array}\right]$. Se definen las matrices K_3 y $\widetilde{A}_3$ como sigue: $K_3 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{9}{14} & 1 \end{array}\right]$, y

$$\widetilde{A}_3 = K_3 B_3 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{9}{14} & 1 \end{array}\right] \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & 0 & -\frac{28}{15} & -\frac{113}{15} & -\frac{113}{15} \\ 0 & 0 & -\frac{6}{5} & -\frac{6}{5} & -\frac{6}{5} \end{array}\right] = \left[\begin{array}{cccc|c} 3 & 6 & 2 & 7 & 7 \\ 0 & 5 & \frac{1}{3} & \frac{26}{3} & \frac{11}{3} \\ 0 & 0 & -\frac{28}{15} & -\frac{113}{15} & -\frac{113}{15} \\ 0 & 0 & 0 & 51 & 51 \end{array}\right].$$

De la matriz ampliada $\widetilde{A}_3$ se establece el sistema de ecuaciones lineales triangular superior siguiente:

$$\left\{\begin{array}{rrrrcc} 3x_1 & +6x_2 & +2x_3 & +7x_4 & = & 7 \\ & 5x_2 & +\frac{1}{3}x_3 & +\frac{26}{3}x_4 & = & \frac{11}{3} \\ & & -\frac{28}{15}x_3 & -\frac{113}{15}x_4 & = & -\frac{113}{15} \\ & & & 51x_4 & = & 51. \end{array}\right.$$

cuya solución es $\overrightarrow{x}^T = (2, -1, 0, 1)$.

Algoritmo de eliminación gaussiana con pivoting parcial

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij})\, M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$.

Datos de salida: Mensaje 1: "Error: matriz nula". Mensaje 2: "Matriz singular". Solución $\overrightarrow{x}^T = (x_1, ..., x_n)\, \mathbb{R}^n$.

1. $S = 0$.

2. Para $i = 1, ..., n$ Verificación de no nulidad de la matriz A.

   Para $j = 1, ..., n$

   Si $|a_{ij}| \leq 0$,

   $S = S + j$

   Fin de bucle j.

   Fin de bucle i.

3. Si $S = n^2$. Continuar en 8).

4. Para $i = 1, ..., n$ Matriz ampliada.

   $a_{i,n+1} = b_i$

Fin de bucle i.

5. Para $j = 1, ..., n-1$

$r = j$

Para $i = j+1, ..., n$ Selección del pivoting parcial en la etapa j, y de la fila r.

Si $|a_{ij}| > |a_{ij}|$,

$r = i$

Fin de bucle i.

Para $p = j, ..., n+1$ Intercambio de la fila j con la fila r.

$t = a_{jp}$

$a_{jp} = a_{rp}$

$a_{rp} = t$

Fin de bucle p.

Si $|a_{jj}| \leq 0$, Continuar en 9).

Si $|a_{jj}| > 0$, entonces

Para $i = j+1, ..., n$ Transformación del sistema en un triangular superior

$$k_{ij} = -\frac{a_{ij}}{a_{jj}}$$

Para $p = j+1, ..., n+1$

$a_{ip} = k_{ij} a_{jp} + a_{ir}$

Fin de bucle p

Fin de bucle i.

Fin de bucle j.

6. Resolver el sistema triangular superior $R\overrightarrow{x} = \overrightarrow{b}_{n-1}$.

7. Escribir $\overrightarrow{x}^T = (x_1, ..., x_n)$. Continuar en 10).

8. imprimir mensaje 1. Continuar en 10).

9. Imprimir mensaje 2.

10. Fin.

Pivoting total.

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$ y $\overrightarrow{b}^T = (b_1, ..., b_n)\,\mathbb{R}^n$. Consideramos el sistema de ecuaciones lineales

$$A\overrightarrow{x} = \overrightarrow{b}.$$

Etapa 1

Sean $r, s \in \mathbb{Z}^+$ tales que $1 \leq r \leq n$, $1 \leq s \leq n$, $|a_{rs}| = \underset{\substack{i=1,...,n \\ j=1,...,n}}{\text{Max}} |a_{ij}|$.

Si $a_{rs} = 0$, la matriz A es nula y el proceso concluye. Supongamos que $a_{rs} \neq 0$. Intercambiamos las filas 1 con r y luego las columnas 1 con s. Este proceso se realiza mediante la operación entre matrices que notamos

$$B_0 = F_{1,r} A C_{1,s} = (b_{ij}),$$

donde $F_{1,r}$ es la matriz obtenida de la matriz identidad al intercambiar la fila 1 con la fila r, y $C_{1,s}$ es la matriz que se obtiene de la matriz identidad al intercambiar la columna 1 con la columna s.

Nótese que el intercambio de columnas provoca un intercambio de las incógnitas x_1 con x_s.

A continuación se procede como en el caso del pivoting parcial.

Se definen $k_{i1} = -\frac{b_{i1}}{b_{11}} \quad i = 2, ..., n$ (b_{11} es el elemento a_{rs}), $K_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ k_{21} & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ k_{n1} & 0 & \cdots & 1 \end{bmatrix}$, y

$$\begin{aligned} A_1 &= K_1 B_0 = \left(a_{ij}^{(1)}\right), \\ \overrightarrow{b}_1 &= K_1 F_{1,r} \overrightarrow{b}. \end{aligned}$$

Debe observarse que no se está utilizando la matriz ampliada $\widetilde{A}$.

Etapa 2

Sean $r, s \in \mathbb{Z}^+$ tales que $2 \leq r \leq n$, $2 \leq s \leq n$, $\left|a_{rs}^{(1)}\right| = \underset{\substack{i=2,...,n \\ j=2,...,n}}{\text{Max}} \left|a_{ij}^{(1)}\right|$.

Si $a_{rs}^{(1)} = 0$, la matriz A es singular. Concluir el proceso.

Si $a_{rs}^{(1)} \neq 0$, intercambiamos la fila 2 con la fila r, y la columna 2 con la columna s de A_1, o sea

$$B_1 = F_{2,1} A_1 C_{2,s} = \left(b_{ij}^{(1)}\right),$$

donde $F_{2,r}$ es la matriz obtenida de la matriz identidad al intercambiar la fila 2 con la fila r; y, $C_{2,s}$ es la matriz obtenida también de la identidad al intercambiar la columna 2 con la s. Note que este último intercambio provoca el intercambio de las incógnitas x_2 con x_s.

Se definen $A_2 = K_2 B_1$, y, $\overrightarrow{b}_2 = K_2 F_{2r} \overrightarrow{b}_1$,donde

$$K_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & k_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & k_{n2} & 0 & \cdots & 1 \end{bmatrix} \qquad \text{con} \qquad k_{i2} = -\frac{b_{ij}^{(1)}}{b_{22}^{(1)}} \quad i = 3, ..., n.$$

Continuando con este proceso hasta la etapa $n-1$, se obtiene

$$B_{n-2} = F_{n-1,n} A_{n-2} C_{n-1,n} = \left(b_{ij}^{(n-2)}\right),$$

con $F_{n-1,n}$, $C_{n-1,n}$ matrices con similares significados que en las etapas 1 y 2.

Se definen $A_{n-1} = K_{n-1} B_{n-2}$, y , $\overrightarrow{b}_{n-1} = K_{n-1} F_{n-1,n} \overrightarrow{b}_{n-2}$, donde

$$K_{n-1} = \begin{bmatrix} 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \\ 0 & \cdots & -\dfrac{b_{n,n-1}}{b_{n-1,n-1}^{(n-2)}} & 1 \end{bmatrix},$$

siempre que $b_{n-1,n-1}^{(n-2)} \neq 0$.

La matriz A_{n-1} es triangular superior. Se tiene

$$\begin{aligned} A_{n-1} &= K_{n-1}B_{n-1} = K_{n-1}F_{n-1,n}A_{n-2}C_{n-1,n} \\ &\vdots \\ &= K_{n-1}F_{n-1,n}\cdots K_2B_1C_{2,s}\cdots C_{n-1,n} \\ &= K_{n-1}F_{n-1,n}\cdots K_1F_{1,r}AC_{1,s}\cdots C_{n-1,n}. \end{aligned}$$

Ponemos

$$\begin{aligned} E &= K_{n-1}F_{n-1,n}\cdots K_1F_{1,r}, \\ C &= C_{1,s}\cdots C_{n-1,n}, \end{aligned}$$

entonces

$$A_{n-1} = EAC.$$

Las matrices E y C son invertibles. Resulta que

$$AC = E^{-1}A_{n-1}.$$

Poniendo $L = E^{-1}$, $R = A_{n-1}$, se tiene la siguiente factorización $AC = LR$.

Por otro lado,

$$\overrightarrow{b}_{n-1} = K_{n-1}F_{n-1,n}\overrightarrow{b}_{n-2} = \cdots = K_{n-1}F_{n-1,n}\cdots K_1F_{1,r}\overrightarrow{b} = E\overrightarrow{b}.$$

Además,

$$A\overrightarrow{x} = \overrightarrow{b} \iff E^{-1}RC^{-1}\overrightarrow{x} = E^{-1}\overrightarrow{b}_{n-1} \iff RC^{-1}\overrightarrow{x} = \overrightarrow{b}_{n-1} \iff \overrightarrow{x} = CR^{-1}\overrightarrow{b}_{n-1}.$$

Sea $\overrightarrow{y} = R^{-1}\overrightarrow{b}_{n-1}$, entonces $R\overrightarrow{y} = \overrightarrow{b}_{n-1}$. En el método de eliminación gaussiana con pivoting total, se resuelve primeramente el sistema de ecuaciones lineales triangular superior $R\overrightarrow{y} = \overrightarrow{b}_{n-1}$ y luego $\overrightarrow{x} = C\overrightarrow{y}$, pués se deben recuperar las variable originales.

Ejemplo

Consideremos nuevamente el ejemplo propuesto en el método de eliminación gaussiana con pivoting parcial.

Sean $A = \begin{bmatrix} 1 & 5 & -1 & 0 \\ 2 & 2 & 0 & 0 \\ -2 & 1 & -1 & 4 \\ 3 & 6 & 2 & 7 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} -3 \\ 2 \\ -1 \\ 7 \end{bmatrix}$

Etapa 1

Selección del pivoting total. Observemos que $7 = |a_{44}| = \underset{\substack{i=1,\ldots,4 \\ j=1,\ldots,4}}{\text{Max}} |a_{ij}|$. Tenemos $r = 4$, $s = 4$. Intercambiamos la fila 1 con la 4 y luego la columna 1 con la 4 en la matriz A. Resulta

$$B_0 = \begin{bmatrix} 7 & 6 & 2 & 3 \\ 0 & 2 & 0 & 2 \\ 4 & 1 & -1 & -2 \\ 0 & 5 & -1 & 1 \end{bmatrix}, \quad C_{1,4} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Denotamos con $\overrightarrow{d}$ al vector que se obtiene de $\overrightarrow{b}$ al intercambiar la fila 1 con la 4, esto es, $\overrightarrow{d}^T = (7, 2, -1, -3)$.

Se definen $K_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{4}{7} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, y

$$A_1 = K_1 B_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{4}{7} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 7 & 6 & 2 & 3 \\ 0 & 2 & 0 & 2 \\ 4 & 1 & -1 & -2 \\ 0 & 5 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 6 & 2 & 3 \\ 0 & 2 & 0 & 2 \\ 0 & -\frac{17}{7} & -\frac{15}{7} & -\frac{26}{7} \\ 0 & 5 & -1 & 1 \end{bmatrix},$$

$$\overrightarrow{b}_1 = K_1 \overrightarrow{d} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{4}{7} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 2 \\ -1 \\ -3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2 \\ -5 \\ -3 \end{bmatrix}.$$

Etapa 2

Seleccionamos el pivoting. Tenemos

$$S = \left|a_{4,2}^{(1)}\right| = \max_{\substack{i=2,3,4 \\ j=2,3,4}} \left|a_{ij}^{(1)}\right|, \quad r = 4, \quad s = 2.$$

Intercambiamos la fila 2 con la 4 de A_1. Como $s = 2$ y $j = 2$, no se requiere de intercambio de columnas, en este caso se tiene la matriz identidad.

Igualmente, intercambiamos la fila 2 con la 4 de $\overrightarrow{b}_1$. Tenemos

$$B_1 = \begin{bmatrix} 7 & 6 & 2 & 3 \\ 0 & 5 & -1 & 1 \\ 0 & -\frac{17}{7} & -\frac{15}{7} & -\frac{26}{7} \\ 0 & 2 & 0 & 2 \end{bmatrix}, \quad \overrightarrow{d}_1 = \begin{bmatrix} 7 \\ -3 \\ -5 \\ 2 \end{bmatrix}, \quad C_{2,2} = I.$$

Sean $K_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{17}{35} & 1 & 0 \\ 0 & -\frac{2}{5} & 0 & 1 \end{bmatrix}$,

$$A = K_2 B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{17}{35} & 1 & 0 \\ 0 & -\frac{2}{5} & 0 & 1 \end{bmatrix} \begin{bmatrix} 7 & 6 & 2 & 3 \\ 0 & 5 & -1 & 1 \\ 0 & -\frac{17}{7} & -\frac{15}{7} & -\frac{26}{7} \\ 0 & 2 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 7 & 6 & 2 & 3 \\ 0 & 5 & -1 & 1 \\ 0 & 0 & -\frac{92}{35} & -\frac{113}{35} \\ 0 & 0 & \frac{2}{5} & \frac{8}{5} \end{bmatrix},$$

$$\overrightarrow{b}_2 = K_2 \overrightarrow{d}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{17}{35} & 1 & 0 \\ 0 & -\frac{2}{5} & 0 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ -3 \\ -5 \\ 2 \end{bmatrix} = \begin{bmatrix} 7 \\ -3 \\ -\frac{226}{35} \\ \frac{16}{5} \end{bmatrix}.$$

Etapa 3

Seleccionamos el pivoting. Tenemos

$$\frac{113}{35} = \left|a_{3,4}^{(2)}\right| = \max_{\substack{i=3,4 \\ j=3,4}} \left|a_{ij}^{(2)}\right|, \quad r = 3, \quad s = 4.$$

Intercambiamos la columna 3 con la 4.

$$B_2 = \begin{bmatrix} 7 & 6 & 3 & 2 \\ 0 & 5 & 1 & -1 \\ 0 & 0 & -\frac{113}{35} & -\frac{92}{35} \\ 0 & 0 & \frac{8}{5} & \frac{2}{5} \end{bmatrix}, \quad \overrightarrow{d}_2 = \overrightarrow{b}_2, \quad C_{3,4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Definimos $K_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{56}{113} & 1 \end{bmatrix}$,

$$A_3 = K_3 B_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{56}{113} & 1 \end{bmatrix} \begin{bmatrix} 7 & 6 & 3 & 2 \\ 0 & 5 & 1 & -1 \\ 0 & 0 & -\frac{113}{35} & -\frac{92}{35} \\ 0 & 0 & \frac{8}{5} & \frac{2}{5} \end{bmatrix} = \begin{bmatrix} 7 & 6 & 3 & 2 \\ 0 & 5 & 1 & -1 \\ 0 & 0 & -\frac{113}{35} & -\frac{92}{35} \\ 0 & 0 & 0 & -\frac{102}{113} \end{bmatrix},$$

$$\overrightarrow{b}_3 = K_3 \overrightarrow{d}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{56}{113} & 1 \end{bmatrix} \begin{bmatrix} 7 \\ -3 \\ -\frac{226}{35} \\ \frac{16}{5} \end{bmatrix} = \begin{bmatrix} 7 \\ -3 \\ -\frac{226}{35} \\ 0 \end{bmatrix}.$$

El sistema de ecuaciones triangular superior es: $\begin{cases} 7y_1 + 6y_2 + 3y_3 + 2y_4 = 7 \\ 5y_2 + y_3 - y_4 = -3 \\ -\frac{113}{35}y_3 - \frac{92}{35}y_4 = -\frac{226}{35} \\ -\frac{102}{113}y_4 = 0, \end{cases}$ cuya solución es $\overrightarrow{y}^T = (1, -1, 2, 0)$.

Como $C = C_{1,4} I C_{3,4}$, se sigue que

$$\begin{aligned} \overrightarrow{x} &= C\overrightarrow{y} = C_{1,4} I C_{3,4} \overrightarrow{y} \\ &= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 1 \end{bmatrix}. \end{aligned}$$

Se define el vector $V = (S_1, S_2, ..., S_{n-1})$, donde $S_j \in \{1, ..., n\}$, $j = 1, ..., n-1$. El componente S_1 significa que, en la primera etapa, se intercambian la columna 1 con la S_1; el componente S_2 significa que, en la segunda etapa, se intercambian las columnas 2 con la S_2, así sucesivamente. Si en la etapa j, $S_j = j$ no hay intercambio.

En el ejemplo tenemos $V = (4, 2, 4)$. Se han realizado los siguientes intercambios:

primera etapa	:	la columna 1 con la 4.
segunda etapa	:	permanece invariante: $j = 2$, $S_2 = 2$.
tercera etapa	:	la columna 3 con la 4.

Para recuperar las variables originales hacemos referencia a los componentes de V y de la solución Y. Ponemos $\overrightarrow{y}_3 = \overrightarrow{y} = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 0 \end{bmatrix}$. Como el tercer componente de V es 4, se realizó el intercambio de la columna 3 con la 4, en $\overrightarrow{y}$ realizamos este intercambio, tenemos $\overrightarrow{y}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 2 \end{bmatrix}$. El segundo componente de V es 2, no hay intercambio, ponemos $\overrightarrow{y}_1 = \overrightarrow{y}_2$. El primer componente de V es 4, se realizó el intercambio de la columna 1 con la 4, en $\overrightarrow{y}_1$ se realiza este intercambio, tenemos $\overrightarrow{x} = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 1 \end{bmatrix}$. la solución.

Ejercicio

Elaborar un algoritmo para la resolución numérica de un sistema de ecuaciones lineales mediante el método de eliminación gaussiana con pivoting total.

Observación

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$, $\overrightarrow{b}^T = (b_1, ..., B_n) \in \mathbb{R}^n$. Se considera el sistema de ecuaciones lineales:

$$A\overrightarrow{x} = \overrightarrow{b}.$$

Supongamos que en la k-ésima etapa de la eliminación gaussiana con pivoting total (parcial), se tiene

$$A = \begin{bmatrix} a_{11}^{(k)} & \cdots & a_{1k}^{(k)} & a_{1k+1}^{(k)} & \cdots & a_{1,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & a_{kk}^{(k)} & a_{kk+1}^{(k)} & \cdots & a_{k,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & a_{n,k}^{(k)} & a_{nk+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}, \qquad \overrightarrow{b}_k = \begin{bmatrix} b_1^{(k)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{bmatrix}.$$

Para realizar la nueva etapa, debemos seleccionar el pivoting:

$$\left|a_{rs}^{(k)}\right| = \underset{\substack{i=k,...,n \\ j=k,...,n}}{\text{Max}} \left|a_{ij}^{(k)}\right|, \text{ con } \quad k \leq r \leq n, \ k \leq s \leq n.$$

Si $a_{rs}^{(k)} = 0$, entonces $a_{ij}^{(k)} = 0$, $i = k, ..., n, \quad j = k, ..., n$, con lo cual la matriz A es singular; y, el sistema de ecuaciones tiene infinitas soluciones o ninguna solución.

i) El sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ no tiene solución si en la k-ésima etapa $a_{rs}^{(k)} = 0 = \underset{\substack{i=k,...,n \\ j=k,...,n}}{\text{Max}} \left|a_{ij}^{(k)}\right|$ y si algún $b_i^{(k)} \neq 0$ para algún $i = k., , , n$.

ii) El sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ tiene infinitas soluciones si en la k-ésima etapa $a_{rs}^{(k)} = 0 = \underset{\substack{i=k,...,n \\ j=k,...,n}}{\text{Max}} \left|a_{ij}^{(k)}\right|$ y $b_i^{(k)} = 0$ par todo $i = k, ..., n$.

En este análisis no se ha considerado los errores de redondeo.

Se deja como ejercicio la elaboración de un algoritmo para la resolución numérica de un sistema de ecuaciones lineales mediante el método de eliminación gaussiana con pivoting parcial o total, y que en el caso de ser la matriz A singular, permita identificar si el sistema tiene infinitas soluciones o ninguna solución.

6.6.3. Cálculo de la matriz inversa A^{-1} y del determinante de la matriz A.

Cálculo de la matriz inversa.

Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$. El método de eliminación gaussiana con pivoting parcial o total puede ser utilizado para calcular la matriz inversa A^{-1} de A, siempre que A^{-1} exista.

Sean $\left\{\overrightarrow{e}_1^T, ..., \overrightarrow{e}_n^T\right\}$ la base canónica de $\mathbb{R}^n$ y B_j la j-ésima columna de A^{-1}. La matriz identidad se nota I. Puesto que $AA^{-1} = A^{-1}A = I$, entonces $AA^{-1}\overrightarrow{e}_j = I\overrightarrow{e}_j = \overrightarrow{e}_j$, y $A^{-1}\overrightarrow{e}_j = B_j, \quad j = 1, ..., n$, se sigue que $AB_j = \overrightarrow{e}_j, \quad j = 1, ..., n,$ es decir que B_j es la solución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{e}_j, \quad j = 1, ..., n$.

Ejemplo

Sea $A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \\ -1 & 0 & 1 \end{bmatrix}$. Hallemos A^{-1}. Para el efecto, apliquemos el método de eliminación gaussiana al sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{e}_j, \quad j = 1, 2, 3$, con $\overrightarrow{e}_1^T = (1, 0, 0)$, $\overrightarrow{e}_2^T = (0, 1, 0)$, $\overrightarrow{e}_3^T = (0, 0, 1)$.

Cálculo de la primera columna de A^{-1} : se considera el sistema de ecuaciones

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

que es equivalente, vía eliminación gaussiana, al siguiente

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix},$$

cuya solución es $B_1^T = \left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right)$.

Cálculo de la segunda columna de A^{-1} : se considera el sistema de ecuaciones

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

que mediante el método de eliminación gaussiana, se obtiene

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix},$$

cuya solución es $B_2^T = \left(-\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}\right)$.

Cálculo de la tercera columna de A^{-1}. Esta es solución del sistema de ecuaciones lineales

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

y procediendo en forma similar a los precedentes, obtenemos

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

cuya solución es $B_3^T = \left(-\frac{5}{6}, \frac{1}{6}, \frac{1}{6}\right)$.

Luego

$$A^{-1} = [B_1, B_2, B_3] = \begin{bmatrix} \frac{1}{2} & -\frac{1}{3} & -\frac{5}{6} \\ -\frac{1}{2} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{2} & -\frac{1}{3} & \frac{1}{6} \end{bmatrix}.$$

Observe que la parte común de los sistemas triangulares superiores que se obtienen para el cálculo de las respectivas columnas B_1, B_2, B_3.

Cálculo de $\det(A)$.

Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$. El método de eliminación gaussiana con pivoting parcial o total permite calcular el determinante de la matriz A. Si aplicamos el método de eliminación gaussiana con pivoting parcial a la matariz A, obtenemos $\begin{cases} B_j = F_{j,S_j} A_{j-1}, & j = 1, ..., n, \\ A_j = K_j B_j, \end{cases}$ donde $A_0 = A$ y F_j, S_j es la matriz obtenida de la identidad al intercambiar la fila j con la $s_j \in \{j, ..., n\}$. Si $s_j = j$, no existe intercambio de filas, en tal caso $F_{j,s_j} = I$ matriz identidad.

Denotamos con m el número de intercambios de filas, es decir, m es el número de matrices $F_{j,s_j} \neq I$.

Por otro lado, K_j es la matriz definida como $K_j = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & k_{j+1,j} & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & k_{n,j} & 0 & \cdots & 1 \end{bmatrix}$ con $K_{i,j} = -\dfrac{b_{ij}^{(j)}}{b_{jj}^{(j)}}$, $b_{jj}^{(j)} \neq 0, \quad i = j+1, ..., n.$

Se tiene que $A_{n-1} = \left(a_{ij}^{(n-1)}\right)$ es triangular superior. Luego $\det(A_{n-1}) = \prod_{i=1}^{n} a_{ii}^{(n-1)}$, $\det(K_j) = 1, \quad j = 1, ..., n-1$, $\det\left(F_{j,S_j}\right) = \begin{cases} 1, & si \quad F_{j,S_j} = I, \\ -1, & si \quad F_{j,S_j} \neq I. \end{cases}$ Como

$$A_{n-1} = K_{n-1} \times F_{n-1,S_{n-1}} \times ... \times K_1 F_{1,S_1} A,$$

se sigue que

$$\begin{aligned} \det(A_{n-1}) &= \det(K_{n-1}) \times \det\left(F_{n-1,S_{n-1}}\right) \times ... \times \det(K_1) \times \det(F_{1,S_1}) \det(A) \\ &= \det\left(F_{n-1,S_{n-1}}\right) \times ... \times \det(F_{1,S_1}) \det(A) \\ &= (-1)^m \det(A), \end{aligned}$$

de donde

$$\det(A) = (-1)^m \det(A_{n-1}) = (-1)^m \prod_{i=1}^{n} a_{ii}^{(n-1)}.$$

Si utilizamos el método de eliminación gaussiana con pivoting total, se tiene

$$A_{n-1} = K_{n-1} F_{n-1,n} ... K_1 F_{1,r} A C_{1,S_1} ... C_{n-1,n},$$

donde $A_{n-1} = \left(a_{ij}^{(n-1)}\right)$ es una matriz triangular superior, K_i y F_{j,s_j} son matrices del tipo descrito en el pivoting parcial y $C_{j,S_j} = I$, es decir que no existe intercambio de columnas.

Denotamos con m_1 el número de intercambio de filas, esto es, el número de matrices $F_{j,S_j} \neq I$; y m_2 el número de intercambios de columnas, es decir que m_2 es el número de matrices $C_{j,S_j} \neq I$. Entonces

$$\det(A) = (-1)^{m_1+m_2} \prod_{j=1}^{n} a_{jj}^{(n-1)}.$$

6.7. Método de Choleski.

Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ una matriz simétrica, definida positiva, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$. Se considera el problema siguiente:

$$\text{hallar } \overrightarrow{x} \in \mathbb{R}^n \text{ solución de } A\overrightarrow{x} = \overrightarrow{b}.$$

Por ser la matriz A simétrica, definida positiva, A es no singular y en consecuencia (1) admite una única solución $\widehat{x} \in \mathbb{R}^n$. Por otra parte, existe una matriz $L = (l_{ij}) \in M_{n\times n}[\mathbb{R}]$ triangular inferior tal que $A = LL^T$.

El sistema de ecuaciones (1) es equivalente a los siguientes: $\begin{cases} L\overrightarrow{y} = \overrightarrow{b}, \\ L^T\overrightarrow{x} = \overrightarrow{y}. \end{cases}$

Primeramente se resuelve el sistema $L\overrightarrow{y} = \overrightarrow{b}$. Calculado el vector $\overrightarrow{y}$, se resuelve a continuación el sistema de ecuaciones $L^T\overrightarrow{x} = \overrightarrow{y}$, que permite calcular $\overrightarrow{x} \in \mathbb{R}^n$ solución de (1).

Describamos el algoritmo de factorización de la matriz A en la forma LL^T. Como $A = LL^T = (a_{ij})$, se sigue de la definición de producto de dos matrices que

$$a_{ij} = \sum_{k=1}^{n} l_{ik} l_{kj}^t, \quad i.j = 1, ..., n,$$

donde $L^T = \left(l_{ij}^t\right)$ y $l_{ij}^t = l_{ji}, \quad i, j = 1, ..., n.$

La igualdad (4) así como la definición de la matriz triangular inferior $L = (l_{ij})$ serán utilizados sucesivamente par construir cada columna de la matriz L.

Primera columna**:** $j = 1$. Se tiene

$$a_{i1} = \sum_{k=1}^{n} l_{ik} l_{k1}^t = \sum_{k=1}^{n} l_{ik} l_{1k} = l_{i1} l_{11}, \quad i = 1, ..., n,$$

ya que $l_{1k} = 0$ para $k = 2, ..., n$. Así,

$$a_{i1} = l_{i1} l_{11}, \quad i = 1, ..., n.$$

Para $i = 1$, se tiene $a_{11} = l_{11}^2 \Rightarrow l_{11} = \sqrt{a_{11}}$ ($l_{11} = -\sqrt{a_{11}}$ no es posible porque A es simétrica, definida positiva).

Para $i = 2, ..., n$, se tiene$l_{i1} = \frac{a_{i1}}{l_{11}}$

Segunda columna: $j = 2$. Tenemos

$$a_{i2} = \sum_{k=1}^{n} l_{ik} l_{k2}^t = \sum_{k=1}^{n} l_{ik} l_{2k} = l_{i1} l_{21} + l_{i2} l_{22}, \quad i = 2, ..., n,$$

pués $l_{2k} = 0$ para $k = 3, ..., n$.

Los elementos $l_{i1}, \ \ i = 1, ..., n$ son calculados en la etapa precedente. Debemos calcular $l_{i2}, \ \ i = 2, ..., n$.

Para $i = 2$, se tiene $a_{22} = l_{21}^2 + l_{22}^2 \Rightarrow l_{22} = \sqrt{a_{22} - l_{21}^2}$ y para $i = 3, ..., n$, se obtiene $l_{i2} = \frac{a_{i2} - l_{i1} l_{21}}{l_{22}}$.

j-ésima columna: $1 < j \leq n$.

Supongamos conocidas las $j - 1$ columnas de L. La j-ésima columna de L se determina como sigue:

$$a_{ij} = \sum_{k=1}^{n} l_{ik} l_{kj}^t = \sum_{k=1}^{j} l_{ik} l_{jk}, \quad i = j, ..., n.$$

Note que $l_{ij+1} = 0, ..., \quad l_{in} = 0.$

Para $i = j$,

$$a_{jj} = \sum_{k=1}^{j} l_{jk} l_{jk} = \sum_{k=1}^{j-1} l_{jk}^2 + l_{jj}^2$$

de donde

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2} \tag{9}$$

y para $i = j + 1, ..., n$, se tiene

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik} l_{jk} + l_{ij} l_{jj}$$

con lo cual

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}} \tag{10}$$

Note que $a_{jj} > \sum_{k=1}^{j-1} l_{jk}^2$, luego $l_{jj} > 0$. Además $l_{jj} \leq \sqrt{a_{jj}}$ para $j = 1, ..., n$.

Hacemos notar que en la práctica, dada una matriz simétrica A, el algoritmo de Choleski permite identificar si A es definida positiva o no por lo que en cada etapa del algoritmo de Choleski se verifica si

$$a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 > 0,$$

y en consecuencia $l_{jj} > 0, \quad j = 1, ..., n$.

En el procedimiento de factorización de A en la forma LL^T descrito, no se consideran los errores de redondeo.

Por otro lado, en el algoritmo se asume que la matriz A es simétrica. En realidad, la primera tarea es verificar que la matriz A sea simétrica.

Con todos estos elementos se establece el siguiente algoritmo de factorización de Choleski.

Algoritmo de factorizacion LL^T

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+, \quad A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$,

Datos de salida: Mensaje 1: "La matriz no es simétrica". Mensaje 2: "La matriz A no es definida positiva". Matriz $L = (l_{ij})$

1. Verificar que la matriz A es simétrica.
2. Si $a_{11} > 0$, entonces

   $l_{11} = \sqrt{a_{11}}$,

   caso contrario, continuar en 6).
3. Para $i = 2, ..., n$

   $$l_{i1} = \frac{a_{i1}}{l_{11}}$$

   Fin de bucle i
4. Para $j = 2, ..., n$

   $S = 0$

   Para $k = 1, ..., j-1$

   $S = S + l_{jk} \times l_{jk}$

   Si $a_{jj} - S > 0$ entonces

   $l_{jj} = \sqrt{a_{jj} - S}$,

   Caso contrario, continuar en 6).

   Fin de bucle k.

   Para $i = j + 1, ..., n$

   $S = 0$

   Para $k = 1, ..., j-1$

   $S = S + l_{ik} \times l_{jk}$

$$l_{ij} = \frac{(a_{ij} - S)}{l_{jj}}$$

Fin de bucle k

Fin de bucle i.

Fin de bucle j.

5. Escribir $L = (l_{ij})$. Continuar en 7).

6. Escribir mensaje 2: "La matriz no es definida positiva".

7. Fin.

Observación**:** Para $j = n$, se tiene

$$a_{in} = \sum_{k=1}^{n-1} l_{ik} l_{kn}^t + l_{in} l_{kn}^t = \sum_{k=1}^{n-1} l_{ik} l_{kk} + l_{in} l_{nn},$$

y para $i = n$,

$$a_{nn} = \sum_{k=1}^{n-1} l_{kk}^2 + l_{nn}^2 \Rightarrow l_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2}.$$

Para $j = n$, el bucle $i = n+1, ..., n$ no se realiza.

Ejercicio

Elaborar un algoritmo para verificar si una matriz $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ es no simétrica. Si $A \neq A^T$, imprimir mensaje "La matriz A no es simétrica". Concluir.

Para la resolución numérica de un sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ mediante el método de Choleski, se propone el siguiente algoritmo en el que se supone se realiza la factorización LL^T descrito en el algoritmo precedente.

Algoritmo del método de Choleski

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$.

Datos de salida: Mensaje: "A no es simétrica, definida positiva". Solución $\overrightarrow{x}^T = (x_1, ..., x_n)$.

1. Aplicar el algoritmo de la factorización LL^T.

   Si A no es simétrica, definida positiva, continuar en 5).

2. Resolver el sistema de ecuaciones $L\overrightarrow{y} = \overrightarrow{b}$.

3. Resolver el sistema de ecuaciones $L^T\overrightarrow{x} = \overrightarrow{y}$.

4. Escribir la solución $\overrightarrow{x}^T = (x_1, ..., x_n)$. Continuar en 6).

5. Escribir mensaje: "A no es simétrica, definida positiva".

6. Fin.

Ejemplo

Hallar la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, donde $A = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 4 & 1 \\ 2 & 4 & 9 & 2 \\ 3 & 1 & 2 & 14 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 3 \\ 2 \\ 3 \\ 11 \end{bmatrix}$.

Apliquemos el método de Choleski. Observamos primeramente que la matriz A es simétrica, esto es $A = A^T$. Pasemos al algoritmo de factorización de Choleski.

Etapa 1 $(j = 1)$. Tenemos los siguientes resultados

$$l_{11} = \sqrt{a_{11}} = \sqrt{1} = 1,$$

$$\begin{aligned} \text{Para } i &= 2,3,4: \ l_{i1} = \frac{a_{i1}}{l_{11}}, \\ l_{21} &= \frac{a_{21}}{l_{11}} = \frac{1}{1} = 1, \\ l_{31} &= \frac{a_{31}}{l_{11}} = \frac{2}{1} = 2, \\ l_{41} &= \frac{a_{41}}{l_{11}} = \frac{3}{1} = 3. \end{aligned}$$

Etapa 2 $(j = 2)$

$$\begin{aligned} l_{22}\sqrt{a_{22} - l_{21}^2} &= \sqrt{2-1} = 1, \\ \text{Para } i &= 3,4: \ l_{i,2} = \frac{a_{i2} - l_{i1}l_{21}}{l_{22}}, \\ l_{32} &= \frac{a_{32} - l_{31}l_{21}}{l_{22}} = \frac{4 - 2 \times 1}{1} = 2, \\ l_{42} &= \frac{a_{42} - l_{41}l_{21}}{l_{22}} = \frac{1 - 3 \times 1}{1} = -2. \end{aligned}$$

Etapa 3 $(j = 3)$

$$\begin{aligned} l_{33} &= \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{9 - 4 - 4} = 1, \\ \text{Para } i &= 4: \quad l_{i3} = \frac{a_{i3} - l_{i1}l_{31} - l_{i2}l_{32}}{l_{33}}, \\ l_{43} &= \frac{a_{43} - l_{41}l_{31} - l_{42}l_{32}}{l_{33}} = \frac{2 - 3 \times 2 - (-2) \times 2}{1} = 0. \end{aligned}$$

Etapa 4 $(j = 4)$

$$l_{44} = \sqrt{a_{44} - l_{41}^2 - l_{42}^2 - l_{43}^2} = \sqrt{14 - 9 - 4 - 0} = 1.$$

En consecuencia L y L^T son las matrices $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 2 & 1 & 0 \\ 3 & -2 & 0 & 1 \end{bmatrix}$, $L^T = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 0 & 1 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Se verifica inmediatamente que $A = LL^T$.

Pasamos a la resolución de los sistemas de ecuaciones $L\overrightarrow{y} = \overrightarrow{b}$ y $L^T\overrightarrow{x} = \overrightarrow{y}$.

El sistema de ecuaciones $L\overrightarrow{y} = \overrightarrow{b}$ es el siguiente: $\begin{cases} y_1 & & & & = 3 \\ y_1 & +y_2 & & & = 2 \\ 2y_1 & +2y_2 & +y_3 & & = 3 \\ 3y_1 & -2y_2 & & +y_4 & = 11, \end{cases}$ cuya solución es $\overrightarrow{y}^T = (3, -1, -1, 0)$.

El sistema de ecuaciones $L^T\overrightarrow{x} = \overrightarrow{y}$ es el siguiente: $\begin{cases} x_1 & +x_2 & +2x_3 & +3x_4 & = & 3 \\ & x_2 & +2x_3 & -2x_4 & = & -1 \\ & & x_3 & & = & -1 \\ & & & x_4 & = & 0, \end{cases}$ cuya solución es $\overrightarrow{x}^T = (4, 1, -1, 0)$.

Número de operaciones elementales.

Asumimos que la raíz cuadrada de un número real no negativo es una operación elemental. Con esta suposición, determinemos el número de operaciones elementales (adiciones, productos, divisiones, raíces cuadradas) necesarias para la construcción de la matriz triangular inferior $L = (l_{ij})$. Tenemos

$j = 1$,

$$\begin{aligned} \text{raíz cuadrada:} &\quad 1, \\ \text{divisiones:} &\quad n-1, \\ \text{productos:} &\quad 0, \\ \text{adiciones:} &\quad 0, \end{aligned}$$

$j = 2$,

$$\begin{aligned} \text{raíz cuadrada} &: 1, \\ \text{divisiones} &: n-2, \\ \text{productos} &: 1+n-2 = n-1, \\ \text{adiciones} &: 1+n-2 = n-1, \end{aligned}$$

$j = n-1$,

$$\begin{aligned} \text{raíz cuadrada} &: 1, \\ \text{divisiones} &: 1, \\ \text{productos} &: n-2+n-2 = 2(n-2), \\ \text{adiciones} &: n-2+n-2 = 2(n-2), \end{aligned}$$

$j = n$,

$$\begin{aligned} \text{raíz cuadrada} &: 1, \\ \text{productos} &: n-1, \\ \text{adiciones} &: n-1. \end{aligned}$$

Luego, en todas las etapas se realizan las siguientes operaciones elementales:

$$\begin{aligned} \text{raíces cuadradas} &: n \\ \text{divisiones} &: n-1+n-2+...+1 = \sum_{j=1}^{n-1}(n-j) = \frac{n(n-1)}{2}, \\ \text{productos} &: 0+n-1+...+2(n-2)+m-1 = \sum_{j=1}^{n}(n-j+1)(j-1) = \frac{n\left(n^2-1\right)}{6} \\ \text{adiciones} &: 0+n-1+...+2(n-2)+n-1 = \sum_{j=1}^{n}(n-j+1)(j-1) = \frac{n\left(n^2-1\right)}{6} \end{aligned}$$

Total de operaciones elementales requeridas para la factorización de la matriz A:

$$N = n + \frac{n\left(n^2-1\right)}{6} + \frac{n\left(n^2-1\right)}{6} + \frac{n(n-1)}{2} = \frac{n}{6}\left(2n^2+3n+1\right).$$

Para la resolución de los sistemas de ecuaciones lineales $L\overrightarrow{y} = \overrightarrow{b}$ y $L^T\overrightarrow{x} = \overrightarrow{y}$ se requieren de $2n^2$ operaciones elementales. Entonces

$$Noper(n) = 2n^2 + \frac{n}{6}\left(2n^2+3n+1\right) = \frac{n}{6}\left(2n^2+15n+1\right).$$

Así, para $n = 4$, $Noper(4) = 62$, para $n = 5$, $Noper(5) = 105$. Para $n \geq 5$, el número de operacione elementales requerido para resolver un sistema de ecuaciones lineales con el método de Choleski es inferior al utilizado en el método de eliminación gaussiana $\left(Noper(n) = \frac{n}{6}\left(4n^2+9n-7\right)\right)$.

Cálculo del determinante

Si $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ es una matriz simétrica, definida positiva, existe una matriz $L=(l_{ij})$ triangular inferior tal que $A=LL^T$. Entonces, el determinante de L denotado $\det(L)$ se calcula como el producto de los elementos de la diagonal, esto es,

$$\det(L)=\prod_{i=1}^{n} l_{ii}.$$

Por otro lado,

$$\det(A)=\det\left(LL^T\right)=\det(L)\det\left(L^T\right)=[\det(L)]^2$$

de modo que

$$\det(A)=\left(\prod_{i=1}^{n} l_{ii}\right)^2=\prod_{i=1}^{n} l_{ii}^2.$$

El número de operaciones elementales que se requieren para la factorización de la matriz A es:

$$N=\frac{n}{6}\left(2n^2+3n+1\right),$$

el número de operaciones elementales para el cálculo de $\prod_{i=1}^{n} l_{ii}^2$ es n^2 productos. Luego, el número de operaciones elementales requeridas para el cálculo de $\det(A)$ es

$$Noper(n)=\frac{n}{6}\left(2n^2+3n+1\right)+n^2=\frac{n}{6}\left(2n^2+9n+1\right).$$

6.8. Método de Crout.

Sean $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ no nula, $\overrightarrow{b}^T=(b_1,...,b_n)\in\mathbb{R}^n$. Consideramos el problema siguiente: hallar $\overrightarrow{x}\in\mathbb{R}^n$, si existe, solución de

$$A\overrightarrow{x}=\overrightarrow{b}. \tag{1}$$

En el método de factorización de Crout se buscan, si existen, dos matrices $L=(l_{ij})$, $U=(u_{ij})$ se $M_{n\times n}[\mathbb{R}]$ tales que

$$\begin{aligned} l_{ij} &= 0 \quad si \quad j>i, \quad i=1,...,n, \ \ j=2,...,n,\\ u_{ij} &= 0 \quad si \quad i>j, \quad i=2,...,n, \ \ j=1,...,n,\\ u_{ii} &= 1, \quad i=1,...,n \end{aligned}$$

y

$$A=LU. \tag{2}$$

En definitiva, la matriz L es triangular inferior, la matriz U es triangular superior y cuyos elementos de la diagonal son todos 1. De la factorización de A, es sistema de ecuaciones (1) es equivalente a los siguientes:

$$A\overrightarrow{x}=\overrightarrow{b}\Leftrightarrow LU\overrightarrow{x}=\overrightarrow{b}\Leftrightarrow\begin{cases} L\overrightarrow{y}=\overrightarrow{b},\\ U\overrightarrow{x}=\overrightarrow{y}. \end{cases}$$

El primero $L\overrightarrow{y}=\overrightarrow{b}$ es un sistema triangular inferior, y el segundo $U\overrightarrow{x}=\overrightarrow{b}$ es un sistema triangular superior. Por lo tanto, queda describir un algoritmo para determinar las matrices L y U.

De la definición de producto de matrices, se tiene

$$a_{ij}=\sum_{k=1}^{n} l_{ik}u_{kj} \quad i=1,...,n, \ \ j=1,...,n. \tag{3}$$

Etapa 1 $(j = 1)$. De la definición de las matrices L y U, se tiene

$$a_{i1} = \sum_{k=1}^{n} l_{ik}u_{k1} = l_{i1}u_{11} = l_{i1} \quad i = 1, ..., n.$$

Así,

$$l_{i1} = a_{i1}, \quad i = 1, ..., n. \tag{4}$$

Para $i = 1$,

$$a_{1j} = \sum_{k=1}^{n} l_{1k}u_{kj} = l_{11}u_{1j},$$

de donde

$$u_{1j} = \frac{a_{1j}}{l_{11}}, \quad l_{11} \neq 0, \quad j = 2, ..., n. \tag{5}$$

Etapa 2 $(j = 2)$

$$a_{i2} = \sum_{k=1}^{n} l_{ik}u_{k2} = l_{i1}u_{12} + l_{i2}u_{22} = l_{i1}u_{12} + l_{i2},$$

de donde

$$l_{i2} = a_{i2} - l_{i1}u_{12} \quad i = 2, ..., n.$$

Para $i = 2$,

$$\begin{aligned} a_{2j} &= \sum_{k=1}^{n} l_{2k}u_{kj} = l_{21}u_{1j} + l_{22}u_{2j} \\ u_{2j} &= \frac{a_{2j} + l_1u_{1j}}{l_{22}} \quad si \;\; l_{22} \neq 0, \;\; j = 3, ..., n. \end{aligned}$$

Note que en la primera etapa se construye la primera columna de L y luego la primera fila de U. En la segunda etapa se construye la segunda columna de L y a continuación la segunda fila de U. Continuando con este procedimiento, para $1 \leq j \leq n$ fijo,

$$a_{ij} = ss_{k=1}^{n} l_{ik}u_{kj} = l_{i1}u_{1j} + l_{i2}u_{2j} + ... + l_{ij}u_{ii} = \sum_{k=1}^{j-1} l_{ik}u_{kj} + l_{ij}$$

de donde

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \quad i = j, ..., n. \tag{6}$$

Para $1 \leq i \leq n - 1$ fijo, se tiene

$$\begin{aligned} a_{ij} &= ssk = 1^n l_{ik}u_{kj} = l_{i1}u_{1j} + l_{i2}u_{2j} + ... + l_{ii}u_{ij}, \\ u_{ij} &= \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}}{l_{jj}}, \quad l_{jj} \neq 0, \quad j = i + 1, ..., n. \end{aligned}$$

Antes de proceder a la descripción del algoritmo de factorización de A mediante el método de Crout, consideremos un ejemplo.

Ejemplo

Hallar la solución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, con A la matriz y $\overrightarrow{b}$ el vector que se dan a continuación. $A = \begin{bmatrix} 2 & -2 & 0 & 2 & 0 \\ 1 & 0 & -1 & 2 & 0 \\ -1 & 2 & 1 & 4 & -2 \\ 2 & -3 & 0 & 1 & 1 \\ 0 & -1 & 1 & 0 & -1 \end{bmatrix}$, $\overrightarrow{b} \begin{bmatrix} 0 \\ 0 \\ 0 \\ -2 \\ 1 \end{bmatrix}$. Apliquemos el método de Crout.Para ello, comencemos con el pricedimiento de factorización LU descrito precedentemene.

Etapa 1 $(j = 1)$. En esta etapa se determinan los elementos de la primera columna de A :

$$\begin{aligned} l_{1i} &= a_{i1} \quad i = j, ..., n \\ l_{11} &= a_{11} = 2, \\ l_{21} &= a_{21} = 1, \\ l_{31} &= a_{31} = -1, \\ l_{41} &= a_{41} = 2, \\ l_{51} &= a_{51} = 0. \end{aligned}$$

Inmediatamente, se pasa a la construcción de los elementos de la primera fila de U :

$$l_{11} \neq 0, \quad u_{1k} = \frac{a_{1k}}{l_{11}}, \quad k = j + 1, ..., n$$

Resulta,

$$\begin{aligned} u_{12} &= \frac{a_{12}}{l_{11}} = \frac{-2}{2} = -1, \\ u_{13} &= \frac{a_{13}}{l_{11}} = \frac{0}{2} = 0, \\ u_{14} &= \frac{a_{14}}{l_{11}} = \frac{2}{2} = 1, \\ u_{15} &= \frac{a_{15}}{l_{11}} = \frac{0}{2} = 0. \end{aligned}$$

Etapa 2 $(j = 2)$. Se tiene $l_{i2} = a_{i2} - l_{i1}u_{12}, \quad i = j, ..., n$, con lo que se construye la segunda columna de L. Resulta,

$$\begin{aligned} l_{22} &= a_{22} - l_{21}u_{12} = 0 - 1 \times (-1) = 1, \\ l_{32} &= a_{32} - l_{31}u_{12} = 2 - (-1) \times (-1) = 1, \\ l_{42} &= a_{42} - l_{41}u_{12} = -3 - 2 \times (-1) = -1, \\ l_{52} &= a_{52} - l_{51}u_{12} = -1 - 0 \times (-1) = -1. \end{aligned}$$

Se pasa inmediantemente a la obtención de los elementos de la segunda fila de U. Para $k = j + 1, ..., n$ o sea $k = 3, 4, 5$,

$$u_{jk} = \frac{a_{jk} - l_{j1}u_{1k}}{l_{jj}}.$$

Luego,

$$\begin{aligned} u_{23} &= \frac{a_{23} -_{21} u_{13}}{l_{22}} = \frac{-1 - 1 \times 0}{1} = 1, \\ u_{24} &= \frac{a_{24} - l_{21}u_{14}}{l_{22}} = \frac{2 - 1 \times 1}{1} = 1, \\ u_{25} &= \frac{a_{25} - l_{21}u_{15}}{l_{22}} = \frac{0 - 1 \times 0}{1} = 0. \end{aligned}$$

Etapa 3 $(j = 3)$. Se construye la tercera columna de L. Se tiene

$$l_{i3} = a_{i3} - l_{i1}u_{13} - l_{i2}u_{23}, \quad i = j, ..., n,$$

es decir que

$$\begin{aligned} l_{33} &= a_{33} - l_{31}u_{13} - l_{32}u_{23} = 1 - (-1) \times 0 - 1 \times (-1) = 2, \\ l_{43} &= a_{43} - l_{41}u_{13} - l_{42}u_{23} = 0 - 2 \times 0 - (-1)(-1) = -1, \\ l_{53} &= a_{53} - l_{51}u_{13} - l_{52}u_{23} = 1 - 0 \times 0 - (-1)(-1) = 0. \end{aligned}$$

Construimos la tercera fila dae U. Tenemos $l_{jj} \neq 0$, y

$$u_{jk} = \frac{a_{jk} - l_{j1}u_{1k} - l_{j2}u_{2k}}{l_{jj}} \quad k = j + 1, ..., n,$$

con lo cual

$$\begin{aligned} u_{34} &= \frac{a_{34} - l_{31}u_{14} - l_{32}u_{24}}{l_{33}} = \frac{4 - (-1) \times 1 - 1 \times 1}{2} = 2, \\ u_{35} &= \frac{a_{35} - l_{31}u_{15} - l_{32}u_{25}}{l_{33}} = \frac{-2 - (-1) \times 0 - 1 \times 0}{2} = -1. \end{aligned}$$

Etapa 4 $(j = 4)$. Se construye la cuarta columna de L.

$$l_{i4} = a_{i4} - l_{i1}u_{14} - l_{i2}u_{24} - l_{i3}u_{34}, \quad i = j, ..., n,$$

$$\begin{aligned} l_{44} &= a_{44} - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34} = 1 - 2 \times 1 - (-1) \times 1 - (-1) \times 2 = 2, \\ l_{54} &= a_{54} - l_{51}u_{14} - l_{52}u_{24} - l_{53}u_{34} = 0 - 0 \times 1 - (-1) \times 1 - 0 \times 2 = 1. \end{aligned}$$

Inmediatamente se construye la cuarta fila de U.

$$\begin{aligned} u_{4k} &= \frac{a_{4k} - \sum_{r=1}^{j-1} l_{4r}u_{rk}}{l_{44}}, \quad k = j + 1, ..., n, \\ u_{45} &= \frac{a_{45}l_{41}u_{15} - l_{42}u_{25} - l_{43}u_{35}}{l_{44}} = \frac{1 - 2 \times 0 - (-1) \times 0 - (-1) - (-1)}{2} = 0. \end{aligned}$$

Etapa 5 $(j = 5)$. Note que $n = 5$, o sea $j = n$. En esta etapa se construye únicamente el elemento l_{nn}. Se tiene

$$l_{i5} = a_{i5} - \sum_{r=1}^{j-1} l_{ir}u_{r5} \quad i = j, ..., n,$$

como $j = n = 5, \quad i = 5$. Entonces

$$\begin{aligned} l_{55} &= a_{55} - l_{51}u_{15} - l_{52}u_{25} - l_{53}u_{35} - l_{54}u_{45} \\ &= -1 - 0 \times 0 - (-1) \times 0 - 0 \times (-1) - 1 \times 0 = -1. \end{aligned}$$

Obtenemos

$$L = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 2 & 0 & 0 \\ 2 & -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 & -1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

El sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ es equivalente a los dos siguientes: $L\overrightarrow{y} = \overrightarrow{b}$ y $U\overrightarrow{x} = \overrightarrow{y}$.

Comencemos con la resolución del sistema triangular inferior $L\overrightarrow{y} = \overrightarrow{b}$. Tenemos

$$\begin{cases} 2y_1 & & & & & = 0 \\ y_1 & +y_2 & & & & = 0 \\ -y_1 & +y_2 & +2y_3 & & & = 0 \\ 2y_1 & -y_2 & -y_3 & 2y_4 & & = -2 \\ & -y_2 & & +y_4 & -y_5 & = 1, \end{cases}$$

cuya solución es $\overrightarrow{y}^T = (0, 0, 0, -1, -2)$.

Concluimos con la resolución numérica del sisteam de ecuaciones triangular superior $U\overrightarrow{x} = \overrightarrow{y}$ siguiente:

$$\begin{cases} x_1 & -x_2 & & +x_4 & & = & 0 \\ & x_2 & -x_3 & +x_4 & & = & 0 \\ & & x_3 & +2x_4 & -x_5 & = & 0 \\ & & & x_4 & & = & -1 \\ & & & & x_5 & = & -2. \end{cases}$$

La solución es $\overrightarrow{x}^T = (2, 1, 0, -1, 2)$.

Algoritmo de factorización LU

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$.

Datos de salida: L, U.

1. Para $i = 1, ..., n$

   $l_{i1} = a_{i1}$.

   Fin de bucle i.

2. Si $l_{11} \neq 0$,

   Para $k = 2, ..., n$

   $$u_{1k} = \frac{a_{1k}}{l_{11}}$$

   Fin de bucle k.

   Caso contrario, continuar en 5).

3. Para $j = 2, ..., n$

   Para $i = j, ..., n$

   $S = 0$

   Para $k = 1, ..., j-1$

   $S = S + l_{ik} \times u_{kj}$

   $l_{ij} = a_{ij} - S$

   Fin de bucle k.

   Si $l_{jj} \neq 0$,

   Para $k = j+1, ..., n$

   $S = 0$

   Para $i = 1, ..., j-1$

   $S = S + l_{ki} \times u_{ij}$

   $$u_{jk} = \frac{a_{jk} - S}{l_{jj}}$$

   Fin de bucle i.

   Fin de bucle k.

   Caso contrario, continuar en 5)

   Fin de bucle j.

4. Imprimir $L.U$. Continuar en 6).

5. Imprimir mensaje: "Matriz singular o no se factora en la forma LU"

6. Fin.

Una vez que se ha procedido a la factoración de la matriz A en la forma LU, se pasa inmediatamente a la resolución del sistema de ecuaciones lineales que se recoje en el algoritmo de Crout.

Algoritmo de Crout

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$.

Datos de salida: Mensaje : "Matriz singular". Solución $\overrightarrow{x}^T = (x_1, ..., x_n)$.

1. Aplicar el algoritmo de factorización LU.

   Si A es singular, continuar en 5).

2. Resolver el sistema de ecuaciones $L\overrightarrow{y} = \overrightarrow{b}$.

3. Resolver el sistema de ecuaciones $U\overrightarrow{x} = \overrightarrow{y}$.

4. Escribir la solución $\overrightarrow{x}^T = (x_1, ..., x_n)$. Continuar en 6).

5. Escribir mensaje: "A es singular o no se factora en la forma LU".

6. Fin.

Observaciones

1. Si $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ se factora en la forma LU, esta es única.

   En efecto, supongamos que A admite otra factorización L_1U_1, donde L_1 es triangular inferior, U triangular superior

   Si A es no singular entonces L, L_1, U, U_1 son no singulares, y

   $$LU = L_1U_1 \Leftrightarrow L_1^{-1}L = U_1U^{-1}.$$

   Como U y U_1 son triangulares superiores, U_1U^{-1} es triangular superior que posee unos en la diagonal. Por otro lado L, L_1 son triangulares inferiores, $L_1^{-1}L$ es triangular inferior. Para que se tenga la igualdad $L_1^{-1}L = U_1U^{-1}$ debe ser $U_1U_{-1} = I$, de donde $U_1 = U$, y $L_1^{-1}L = I$ implica $L = L_1$. Así, la factorización es única.

2. La matriz $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ no se factora en la forma LU. La matriz A es claramente no singular.

   Supongamos que $A = LU$ con $L = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix}$, $U = \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix}$. Entonces

   $$A = LU = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{11}u_{12} \\ l_{21} & l_{21}u_{12} + l_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

   Luego,

   $$l_{11} = 0, \quad 1 = l_{11}u_{12} = 0 \times u_{12} = 0 \quad \text{que es absurdo.}$$

3. Si una matriz singular A se factora en la forma LU, esta no es única. Por ejemplo

   $$\begin{aligned} A &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ \lambda & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda} & 0 \\ 0 & 1 \end{bmatrix} \quad \forall \lambda \in \mathbb{R}, \quad \lambda \neq 0, \\ B &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha} & 0 \\ \alpha & \alpha^3 - \alpha \end{bmatrix} \quad \forall \alpha \in \mathbb{R}, \quad \alpha \neq 0. \end{aligned}$$

6.9. Sistemas de ecuaciones lineales con matrices tridiagonales.

Definición 7 *Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ tal que $A \neq 0$. Se dice que A es tridiagonal si*

$$a_{ij} = 0 \quad para \quad |i-j| > 1, \quad i,j = 1,...,n.$$

Como hemos visto, essta clase de matrices se presentaron en la discretización de problemas de valores de frontera $1d$, en la construcción de splines de interpolación. Además, aparecen en la discretización mediante diferencias finitas y elementos finitos de muchos problemas del tipo

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right) + v\frac{\partial u}{\partial x} + qu = f, \\ \quad + \text{Condición inicial}, \\ \quad + \text{Condiciones de frontera}, \end{cases}$$

donde f, p, q, v son funciones dadas que cumplen cierta regularidad en $[0,L]\times[0,T]$, con $L > 0$, $T > 0$.

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ con $A \neq 0$ y A tridiagonal, $\overrightarrow{b}^T = (b_1,...,b_n) \in \mathbb{R}^n$. Se considera el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$.

De la definición de matriz tridiagonal, la matriz A tiene la forma

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}$$

y el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, en forma explícita se escribe:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 & & & & = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 & & & & = b_2 \\ & & & & \vdots \\ & & & & \vdots \\ & & a_{nn-1}x_{n-1} + a_{nn}x_n & & = b_n. \end{cases}$$

Esta clase de problemas involucra el almacenamiento adecuado de los datos y una simplificación del algoritmo de resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

Comenzamos con el almacenamiento de los datos.

Para almacenar los elementos a_{ij}, $i,j = 1,...,n$ de A se requieren n^2 espacios de memoria, que para n grande, n^2 puede ser muy significativo. Por ejemplo para $n = 1000$, $n^2 = 1{,}000{,}000$. Es claro que en la actualidad una cifra como esta es muy modesta frente a la capacidad de almacenamiento de datos que poseen los modernos equipos de computación. Sin embargo, por grande que sea esta capacidad de almacenamiento, es preciso tratar de optimizar el espacio de memoria utilizado. Con este propósito, para almacenar los datos de una matriz tridiagonal, únicamente se requieren de aquellos elementos a_{ij} para los que $|i-j| \leq 1$, $i,j = 1,...,n$, y no todos los n^2 elementos de la matriz A.

Se define la matriz $B = (b_{ik}) \in M_{n\times 3}[\mathbb{R}]$ del modo siguiente:

$$\begin{aligned} b_{11} &= b_{n3} = 0, \\ b_{i1} &= a_{ii-1}, \quad i = 2,...,n, \\ b_{i2} &= a_{ii}, \quad i = 1,...,n, \\ b_{i3} &= a_{ii+1}, \quad i = 1,...,n-1, \end{aligned}$$

es decir que la matriz B es de la forma

$$B = \begin{bmatrix} 0 & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} \\ a_{32} & a_{33} & a_{34} \\ & \vdots & \\ a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ a_{nn-1} & a_{nn} & 0 \end{bmatrix}.$$

Observamos que los elementos de la diagonal principal de A está localizados en la segunda columna de B. Los elementos de la diagonal superior adyacente a la principal de A están localizados en la tercera columna de B, los elementos de la diagonal inferior adyacente a la principal de A están localizados en la primera columna de B.

Note que la matriz B requiere de $3n$ espacios de memoria. Por ejemplo para $n = 1000$, solo se requieren de 3000 espacios de memoria. Adicionalmente, el tiempo de máquina y la precisión de la solución, son dos situaciones importantes a considerar. En cuanto se refiere a la precisión de la solución, posponemos el análisis correspondiente.

En lo que se refiere al tiempo de máquina utilizado en la resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, se buscan algoritmos cuyos tiempos de máquina, sean en lo posible, los más pequeños. Se logra este objetivo utilizando la hipótesis A es una matriz tridiagonal y elaborando algoritmos que eviten realizar cálculos innecesarios con elementos $a_{ij} = 0$ para $|i-j| > 1,\ i,j = 1,...,n$. El arreglo de elementos de A en la matriz B conduce al propósito antes precisado.

Para esta clase de matrices tridiagonales y con hipótesis suplementarias sobre la matriz A, proponemos tres algoritmos de resolución numérica de los sistemas de ecuaciones lineales.

Eliminación gaussiana sin pivoting.

Supongamos que $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ es una matriz tridiagonal y estrictamente diagonalmente dominante; $\overrightarrow{b}^T = (b_1,...,b_n) \in \mathbb{R}^n$. Consideramos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$.

Con la hipótesis A es estrictamente diagonalmente dominante, debemos mostrar que las matrices $A_1 = \left(a_{ij}^{(1)}\right), ..., A_{n-1} = \left(a_{ij}^{(n-1)}\right)$ correspondientes a cada etapa de la eliminación gaussiana, son estrictamente diagonalmente dominantes.

Etapa 1. Como A es estrictamente diagonalmente dominante, se tiene $|a_{11}| > |a_{12}|, \quad |a_{22}| > |a_{21}| + |a_{23}|$. Resulta $|a_{11}| > 0$. Sean $k_{21} = -\dfrac{a_{21}}{a_{11}}$, y

$$a_{22}^{(1)} = k_{21}a_{12} + a_{22} = -\frac{a_{21}}{a_{11}}a_{12} + a_{22} = -a_{21}\frac{a_{12}}{a_{11}} + a_{22}.$$

1. Como $|a_{11}| > |a_{12}|$ se sigue que $\left|\dfrac{a_{12}}{a_{11}}\right| < 1$. Luego $\left|a_{21}\frac{a_{12}}{a_{11}}\right| < |a_{21}|$. Entonces

$$|a_{22}| > |a_{21}| + |a_{23}| > \left|a_{21}\frac{a_{12}}{a_{11}}\right| + |a_{22}|$$

de donde

$$|a_{22}| - \left|a_{21}\frac{a_{12}}{a_{11}}\right| > |a_{23}|.$$

Utilizando la desigualdad $||x| - |y|| \leq |x-y|,\ \forall x, y \in \mathbb{R}$. Se tiene

$$\left|a_{22}^{(1)}\right| = \left|a_{22} - a_{21}\frac{a_{12}}{a_{11}}\right| \geq |a_{22}| - \left|a_{21}\frac{a_{12}}{a_{11}}\right| > |a_{23}|.$$

Así,

$$\left|a_{22}^{(1)}\right| > \left|a_{23}^{(1)}\right| = |a_{23}|\,.$$

Como las otras filas de la matriz A quedan inalteradas, se tiene

$$|a_{ii}| > |a_{ii-1}| + |a_{ii+1}|\,, \quad i = 3, ..., n.$$

Se pone $a_{ij}^{(1)} = a_{ij}$ para $i = 1, 3, ..., n,\ j = 1, ..., n,\ a_{23}^{(1)} = a_{23},\ a_{2j}^{(1)} = 0,\ j = 1, 4, ..., n$, y $A_1 = \left(a_{ij}^{(1)}\right)$. Consecuentemente, $A_1 = \left(a_{ij}^{(1)}\right)$ es estrictamente diagonalmente dominante.

Etapa 2. Como $a_{22}^{(1)} \neq 0$. Se define $k_{32} = -\dfrac{a_{32}^{(1)}}{a_{22}^{(1)}}$, y

$$a_{33}^{(2)} = k_{32}a_{23}^{(1)} + a_{33}^{(1)} = -\frac{a_{32}^{(1)}}{a_{22}^{(1)}}a_{23}^{(1)} + a_{33}^{(1)} = -a_{32}^{(1)}\frac{a_{23}^{(1)}}{a_{22}^{(1)}} + a_{33}^{(1)}\,.$$

Puesto que $\left|a_{22}^{(1)}\right| > \left|a_{23}^{(1)}\right|$ y razonando en forma similar a la realizada en la etapa 1, se muestra que $\left|a_{33}^{(2)}\right| > \left|a_{34}^{(2)}\right|$. Denotamos con $A_2 = \left(a_{ij}^{(2)}\right)$, donde

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} \quad i = 1, ..., n,\ i \neq 3,\ j = 1, ..., n, \\ a_{32}^{(2)} &= 0,\ a_{34}^{(2)} = a_{34}^{(1)},\ a_{3j}^{(2)} = 0,\ j = 4, ..., n. \end{aligned}$$

Continuando con este proceso llegamos a la etapa $n-1$ siguiente.

Etapa $n-1$. Se tiene

$$\left|a_{n-1,n-1}^{(n-2)}\right| > \left|a_{n-1,n}^{(n-2)}\right| = |a_{n-1,n}| \geq 0.$$

Se define

$$k_{n.n-1} = -\frac{a_{nn-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}},$$

y

$$a_{nn}^{(n-1)} = k_{n,n-1}a_{n-1,n}^{(n-2)} + a_{nn}^{(n-2)}.$$

Mostremos que $a_{nn}^{(n-1)} \neq 0$. Como $\left|a_{n-1,n-1}^{(n-2)}\right| > \left|a_{n-1,n}^{(n-2)}\right|$ se sigue $\left|\dfrac{a_{n-1,n}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}}\right| < 1$ y

$$\left|a_{n,n-1}^{(n-2)}\frac{a_{n-1,n}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}}\right| < \left|a_{nn-1}^{(n-2)}\right| < \left|a_{nn}^{(n-2)}\right|.$$

Luego,

$$\left|a_{nn}^{(n-1)}\right| = \left|-\frac{a_{n,n-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}}a_{n-1,n}^{(n-2)} + a_{n,n}^{(n-2)}\right| \geq \left|a_{nn}^{(n-2)}\right| - \left|a_{nn-1}^{(n-2)}\frac{a_{n-1,n}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}}\right| > 0.$$

En términos de los elementos de la matriz B y el vector $\overrightarrow{b}$, el procedimiento de eliminación gaussiana para la resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ se escribe en los siguientes términos.

Etapa 1. Se pone $c = -\dfrac{b_{21}}{b_{12}}$. Luego $b_{22} = cb_{13} + b_{22}, b_2 = cb_1 + b_2$.

Etapa 2. Se pone

$$c = -\frac{b_{31}}{b_{22}}, \qquad b_{32} = cb_{23} + b_{32}, \qquad b_3 = cb_2 + b_3.$$

Continuando con este procedimiento, en la etapa $n-1$, se tiene

Etapa $n-1$

1.
$$c=-\frac{b_{n1}}{b_{n-1,2}},\quad b_{n2}=cb_{n-1,3}+b_{n2},\quad b_n=cb_{n-1}+b_n.$$

La resolución del sistema de ecuaciones se describe en el siguiente procedimiento:

$$x_n=\frac{b_n}{b_{n^2}},$$

y para $i=n-1,...,1$,

$$x_i=\frac{(b_i-b_{i3}x_{i+1})}{b_{i,2}}.$$

Se establece el siguiente algoritmo de resolución del sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{b}$. Se asume que la matriz A es tridiagonal con lo que los elementos de A son almacenados en la matriz B.

Algoritmo

Datos de entrada: $n\in\mathbb{Z}^+,\quad B=(b_{ij})\in M_{n\times 3}[\mathbb{R}],\quad \overrightarrow{b}^T=(b_1,...,b_n)\in\mathbb{R}^n$.

Datos de salida: Mensaje: "A no es estrictamente diagonalmente dominante", solución $\overrightarrow{x}^T=(x_1,...,x_n)$.

1. Para $i=1,...,n$,

   Si $|b_{i2}|\leq|b_{i1}|+|b_{i3}|$, continuar en 6).

   Fin de bucle i.

2. Para $j=1,...,n-1$

   $c=-\dfrac{b_{j+1,1}}{b_{j2}}$ Proceso de eliminación gaussiana

   $b_{j+1,2}=c\times b_{j3}+b_{j+1,2}$

   $b_{j+1}=c\times b_j+b_{j+1}$

   Fin de bucle j.

3. $x_n=\dfrac{b_n}{b_{n2}}$ Resolución del sistema triangular superior

4. Para $j=n-1,...,1$

   $x_j=\dfrac{(b_j-b_{j,3}\times x_{j+1})}{b_{j2}}$

   Fin de bucle j.

5. Imprimir $\overrightarrow{x}^T=(x_1,...,x_n)$. Continuar en 7).

6. Imprimir mensaje: "A no es estrictamente diagonalmente dominante".

7. Fin.

Nota**:** Se puede probar que el número total de operaciones elementales en la resolución del sistema de ecuaciones $A\overrightarrow{x}=\overrightarrow{b}$ es $Noper^{(c)}(n)=8n-7$.

Ejemplos

1. Sean $A = \begin{bmatrix} 5 & -1 & 0 & 0 & 0 \\ -2 & 4 & 1 & 0 & 0 \\ 0 & 1 & 3 & 1 & 0 \\ 0 & 0 & 2 & 5 & 2 \\ 0 & 0 & 0 & 3 & 4 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 11 \\ -8 \\ 0 \\ 1 \\ -5 \end{bmatrix}$. Consideramos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ que en forma explícita se escribe

$$\begin{cases} 5x_1 & -x_2 & & & & = & 11 \\ -2x_1 & +4x_2 & +x_3 & & & = & -8 \\ & x_2 & +3x_3 & +x_4 & & = & 0 \\ & & 2x_3 & +5x_4 & +2x_5 & = & 1 \\ & & & 3x_4 & +4x_5 & = & -5. \end{cases}$$

Apliquemos el algoritmo precedente. Primeramente, la matriz A es tridiagonal, consecuentemente los elementos de A son dispuestos en la siguiente matriz:

$$B = \begin{bmatrix} 0 & 5 & -1 \\ -2 & 4 & 1 \\ 1 & 3 & 1 \\ 2 & 5 & 2 \\ 3 & 4 & 0 \end{bmatrix}.$$

1. Observamos que la matriz A es estrictamente diagonalmente dominante, lo que equivale a observar que en cada fila de B se satisface la condición

$$|b_{j2}| > |b_{j1}| + |b_{j3}|, \quad j = 1, 2, 3, 4, 5.$$

2. Pasemos al proceso de eliminación gaussiana descrito en el punto 2) del algoritmo precedente. Para $j = 1, 2, 3, 4$. Se tiene

$j = 1$,

$$\begin{aligned} c &= -\frac{b_{21}}{b_{12}} = -\frac{-2}{5} = \frac{2}{5}, \\ b_{22} &= c \times b_{13} + b_{22} = \frac{2}{5} \times (-1) + 4 = \frac{18}{5}, \\ b_2 &= c \times b_1 + b_2 = \frac{2}{5} - 11 + (-8) = -\frac{18}{5}, \end{aligned}$$

$j = 2$,

$$\begin{aligned} c &= -\frac{b_{31}}{b_{22}} = -\frac{1}{\frac{18}{5}} = -\frac{5}{18}, \\ b_{3,2} &= c \times b_{23} + b_{32} = -\frac{5}{18} \times 1 + 3 = \frac{49}{18}, \\ b_3 &= c \times b_2 + b_3 = -\frac{5}{18} \times \left(-\frac{18}{5}\right) + 0 = 1, \end{aligned}$$

$j = 3$,

$$\begin{aligned} c &= -\frac{b_{41}}{b_{32}} = -\frac{2}{\frac{49}{18}} = -\frac{36}{49}, \\ b_{42} &= c \times b_{33} + b_{42} = -\frac{36}{49} \times 1 + 5 = \frac{109}{49}, \\ b_4 &= c \times b_3 + b_4 = -\frac{36}{49} \times 1 + 1 = \frac{13}{49}, \end{aligned}$$

$j = 4$,

$$\begin{aligned} c &= -\frac{b_{51}}{b_{42}} = -\frac{3}{\frac{209}{49}} = -\frac{147}{209}, \\ b_{52} &= c \times b_{43} + b_{52} = -\frac{147}{209} \times 2 + 4 = \frac{542}{209}, \\ b_5 &= c \times b_4 + b_5 = -\frac{147}{209} \times \frac{13}{49} + (-5) = -\frac{53116}{10241}. \end{aligned}$$

3. Pasamos a la resolución del sistema de ecuaciones triangular superior:

$$\begin{aligned} x_5 &= \frac{b_5}{b_{52}} = \frac{-\frac{53116}{10241}}{\frac{542}{209}} = -2, \\ x_4 &= \frac{b_4 - b_{43}x_5}{b_{42}} = \frac{\frac{13}{49} - 2\times(-2)}{\frac{209}{49}} = 1, \\ x_3 &= \frac{b_3 - b_{33}x_4}{b_{32}} = \frac{1 - 1\times 1}{\frac{49}{18}} = 0, \\ x_2 &= \frac{b_2 - b_{23}x_3}{b_{22}} = \frac{-\frac{18}{5} - 1\times 0}{\frac{18}{5}} = -1, \\ x_1 &= \frac{b_1 - b_{13}x_2}{b_{12}} = \frac{11 - (-1)\times(-1)}{5} = 2. \end{aligned}$$

La solución del sistema de ecuaciones lineales es $\overrightarrow{x}^T = (2, -1, 0, 1, -2)$.

2. Sean $A = \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ -2 & 6 & 2 & -1 & 0 \\ 1 & -2 & 7 & 2 & -1 \\ 0 & 1 & -2 & 6 & 2 \\ 0 & 0 & 1 & -2 & 4 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 0 \\ 9 \\ 14 \\ 3 \\ 0 \end{bmatrix}$. Consideramos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ Apliquemos el algoritmo precedente.

Etapa 1

$$A_1 = K_1A_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{2}{5} & 1 & 0 & 0 & 0 \\ -\frac{1}{5} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ -2 & 6 & 2 & -1 & 0 \\ 1 & -2 & 7 & 2 & -1 \\ 0 & 1 & -2 & 6 & 2 \\ 0 & 0 & 1 & -2 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & -\frac{12}{5} & \frac{36}{5} & 2 & -1 \\ 0 & 1 & -2 & 6 & 2 \\ 0 & 0 & 1 & -2 & 4 \end{bmatrix},$$

$$\overrightarrow{b}_1 = K_1\overrightarrow{b} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{2}{5} & 1 & 0 & 0 & 0 \\ -\frac{1}{5} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 9 \\ 14 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 9 \\ 14 \\ 3 \\ 0 \end{bmatrix}.$$

Etapa 2

$$A_2 = K_2A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{6}{17} & 1 & 0 & 0 \\ 0 & -\frac{5}{34} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & -\frac{12}{5} & \frac{36}{5} & 2 & -1 \\ 0 & 1 & -2 & 6 & 2 \\ 0 & 0 & 1 & -2 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & 0 & \frac{132}{17} & \frac{28}{17} & -1 \\ 0 & 1 & -\frac{38}{17} & \frac{209}{34} & 2 \\ 0 & 0 & 1 & -2 & 4 \end{bmatrix},$$

$$\overrightarrow{b}_2 = K_2\overrightarrow{b}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{6}{17} & 1 & 0 & 0 \\ 0 & -\frac{5}{34} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 9 \\ 14 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 9 \\ \frac{292}{17} \\ \frac{57}{34} \\ 0 \end{bmatrix}.$$

Etapa 3

$$A_3 = K_3A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{19}{66} & 1 & 0 \\ 0 & 0 & -\frac{17}{132} & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & 0 & \frac{132}{17} & \frac{28}{17} & -1 \\ 0 & 1 & -\frac{38}{17} & \frac{209}{34} & 2 \\ 0 & 0 & 1 & -2 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & 0 & \frac{132}{17} & \frac{28}{17} & -1 \\ 0 & 0 & 0 & \frac{437}{66} & \frac{113}{66} \\ 0 & 0 & 0 & -\frac{73}{33} & \frac{545}{132} \end{bmatrix},$$

$$\overrightarrow{b}_3 = K_3\overrightarrow{b}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{19}{66} & 1 & 0 \\ 0 & 0 & -\frac{17}{132} & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 9 \\ \frac{292}{17} \\ \frac{57}{34} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 9 \\ \frac{292}{17} \\ \frac{437}{66} \\ -\frac{73}{33} \end{bmatrix}.$$

Etapa 4

$$A_4 = K_4A_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{146}{437} & 1 \end{bmatrix} \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & 0 & \frac{132}{17} & \frac{28}{17} & -1 \\ 0 & 0 & 0 & \frac{437}{66} & \frac{113}{66} \\ 0 & 0 & 0 & -\frac{73}{33} & \frac{545}{132} \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 2 & -1 & 0 & 0 \\ 0 & \frac{34}{5} & \frac{8}{5} & -1 & 0 \\ 0 & 0 & \frac{132}{17} & \frac{28}{17} & -1 \\ 0 & 0 & 0 & \frac{437}{66} & \frac{113}{66} \\ 0 & 0 & 0 & 0 & \frac{8217}{1748} \end{bmatrix},$$

$$\overrightarrow{b}_4 = K_4\overrightarrow{b}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{146}{437} & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 9 \\ \frac{292}{17} \\ \frac{437}{66} \\ -\frac{73}{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 9 \\ \frac{292}{17} \\ \frac{437}{66} \\ 0 \end{bmatrix}.$$

La solución del sistema de ecuaciones triangular superior es $\overrightarrow{x} = (0, 1, 2, 1, 0)^T$. Adicionalmente para la resolución del sistema de ecuaciones lineales se realizaron 47 operaciones elementales para transformar el sistema en uno triangular superior y 19 operaciones para resolver este último sistema. La resolución de este sistema de ecuaciones requiere de 66 operaciones elementales.

Método de Choleski

Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ tridiagonal, simétrica, definida positiva, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^n$. Consideramos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$.

La hipótesis A es una matriz simétrica, definida positiva implica la existencia de una matriz triangular inferior $L = (l_{ij}) \in M_{n\times n}[\mathbb{R}]$ tal que $A = LL^T$. En inmediato verificar que

$$l_{ij} = 0 \quad si \quad |i - j| > 1, \;\; i, j = 1, ..., n,$$

esto es, L es una matriz de la forma

$$L = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ 0 & l_{32} & l_{33} & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & l_{nn-1} & l_{nn} \end{bmatrix}.$$

Como se ha dicho, los coeficientes de la matriz A se disponen en la matriz $B = (b_{jk}) \in M_{n\times 3}[\mathbb{R}]$.

De la estructura de la matriz L, se sigue que los coeficientes de interés l_{ij} son únicamente l_{ii}, $i = 1, ..., n$, $l_{i,i-1}$, $i = 2, ..., n$; lo que conduce a definir una matriz $C = (c_{ij}) \in M_{n\times 2}[\mathbb{R}]$ siguiente:

$$\begin{aligned} c_{11} &= 0, \\ c_{i1} &= l_{ii-1}, \quad i = 2, ..., n, \\ c_{i2} &= l_{ii}, \quad i = 1, ..., n, \end{aligned}$$

es decir que C es la matriz de la forma siguiente: $C = \begin{bmatrix} 0 & l_{11} \\ l_{21} & l_{22} \\ l_{31} & l_{33} \\ \vdots & \vdots \\ l_{nn-1} & l_{nn} \end{bmatrix}$.

Con la hipótesis A es una matriz tridiagonal, realizamos las simplificaciones en el algoritmo de Choleski. Tenemos el procedimiento siguiente.

1. $l_{11} = \sqrt{a_{11}}$.
2. $l_{21} = \dfrac{a_{12}}{l_{11}}$.
3. Para $j = 2, ..., n-1$

   $l_{jj} = \sqrt{a_{jj} - l_{jj-1}^2}$

   $l_{j+1,j} = \dfrac{a_{jj+1}}{l_{jj}}$

   Fin de bucle j
4. $l_{nn} = \sqrt{a_{nn} - l_{nn-1}^2}$.

En términos de los elementos de las matrices B y C, el procedimiento precedente se escribe de la manera que a continuación se indica.

1. $c_{12} = \sqrt{b_{12}}$
2. $c_{21} = \frac{b_{13}}{c_{12}}$
3. Para $j = 2, ..., n-1$

   $cj2 = \sqrt{b_{j2} - c_{j1}^2}$

   $c_{j+1,1} = \dfrac{b_{j3}}{c_{j2}}$

   Fin de bucle j.
4. $c_{n2} = \sqrt{b_{n2} - c_{n1}^2}$.

Por otro lado, la resolución del sistema de ecuaciones triangular inferior $L\overrightarrow{y} = \overrightarrow{b}$ se describe en el siguiente procedimiento:

1. $y_1 = \frac{b_1}{L_{11}} = \frac{b_1}{c_{12}}$.

2. Para $j = 2, ..., n$

$$y_i = \frac{b_j - L_{jj-1}y_{j-1}}{L_{jj}} = \frac{b_j - c_{j1}y_{j-1}}{c_{j2}}.$$

Fin de bucle j.

La resolución del sistema de ecuaciones triangular superior $L^T\overrightarrow{x} = \overrightarrow{y}$ se expresa en el siguiente procedimiento.

1. $x_n = \frac{y_n}{L_{nn}} = \frac{y_n}{c_{n2}}$.

2. Para $j = n-1, ..., 1$

$$x_j = \frac{y_j - L_{jj-1}x_{j-1}}{L_{jj}} = \frac{y_j - c_{j1}x_{j-1}}{c_{j2}}.$$

Fin de bucle j.

Ejemplo

Apliquemos el algoritmo precedente al sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, donde

$$A = \begin{bmatrix} 4 & 2 & 0 & 0 & 0 & 0 \\ 2 & 5 & -2 & 0 & 0 & 0 \\ 0 & -2 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 10 & -6 & 0 \\ 0 & 0 & 0 & -6 & 5 & 1 \\ 0 & 0 & 0 & 0 & 1 & 5 \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 16 \\ 1 \\ 0 \end{bmatrix}.$$

Primeramente, los elementos de la matriz A se disponen en la matriz B siguiente:

$$B = \begin{bmatrix} 0 & 4 & 2 \\ 2 & 5 & -2 \\ -2 & 2 & 1 \\ 1 & 10 & -6 \\ -6 & 5 & 1 \\ 1 & 5 & 0 \end{bmatrix}.$$

En el método de Choleski (sin considerar los errores de redondeo) y para matrices tridiagonales, si se tiene $A^T = A$, y

$$a_{jj} > L_{jj-1}^2 \quad j = 1, 2, ..., n,$$

la matriz A es simétrica, definida positiva.

Apliquemos el pricedimiento de resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ arriba descrito. Tenemos:

1. $L_{11}\sqrt{b_{12}} = \sqrt{4} = 2$,

2. $c_{21} = L_{21} = \frac{b_{13}}{c_{12}} = \frac{2}{2} = 1$,

3. Para $j = 2, 3, 4, 5$

$j = 2$, $c_{22} = \sqrt{b_{22} - c_{21}^2} = \sqrt{5-1} = 2$,

$c_{31} = \frac{b_{23}}{c_{22}=-\frac{2}{2}} = -1$,

$j = 3$, $c_{32} = \sqrt{b_{32} - c_{31}^2} = \sqrt{2-1} = 1$,

$c_{41} = \frac{b_{33}}{c_{32}} = \frac{1}{1} = 1$,

$j=4,\ \ c_{42}=\sqrt{b_{42}-c_{41}^2}=\sqrt{10-1}=1,$

$c_{51}=\frac{b_{43}}{c_{42}}=-\frac{6}{3}=-2,$

$j=5,\ \ c_{52}=\sqrt{b_{52}-c_{51}^2}=\sqrt{5-4}=1,$

$c_{61}=\frac{b_{53}}{c_{52}}=\frac{1}{1}=1,$

4. $c_{62}=\sqrt{b_{62}-c_{61}^2}=\sqrt{5-1}=2.$

Los elementos de la matriz $L=(l_{ij})$ se disponen en la matriz C siguiente:

$$C=\begin{bmatrix} 0 & l_{11} \\ l_2 & l_{22} \\ l_{32} & l_{33} \\ l_{43} & l_{44} \\ l_{54} & l_{55} \\ l_{65} & l_{66} \end{bmatrix}=\begin{bmatrix} 0 & 2 \\ 1 & 2 \\ -1 & 1 \\ 1 & 3 \\ -2 & 1 \\ 1 & 2 \end{bmatrix}.$$

El sistema de ecuaciones $L\overrightarrow{y}=\overrightarrow{b}$ en términos de la matriz C tiene la forma siguiente:

$$\begin{array}{llcl} c_{12}y_1 & & = & b_1 \\ c_{21}y_1 & +c_{22}y_2 & = & b_2 \\ c_{31}y_2 & +c_{32}y_3 & = & b_3 \\ c_{41}y_3 & +c_{42}y_4 & = & b_4 \\ c_{51}y_4 & +c_{52}y_5 & = & b_5 \\ c_{61}y_5 & c_{62}y_6 & = & b_6 \end{array} \Leftrightarrow \left\{ \begin{array}{llcl} 2y_1 & & = & 2 \\ y_1 & +2y_2 & = & 1 \\ -y_2 & +y_3 & = & 4 \\ y_3 & +3y_4 & = & 16 \\ -2y_4 & +y_5 & = & -7 \\ y_5 & +2y_6 & = & 1 \end{array} \right.$$

cuya solución es $\overrightarrow{y}^T=(1,0,4,4,1,0)$.

El sistema de ecuaciones $L^T\overrightarrow{x}=\overrightarrow{y}$ expresado en términos de la matriz C tiene la forma siguiente:

$$\left\{ \begin{array}{llcl} c_{12}x_1 & +c_{21}x_2 & = & y_1 \\ c_{22}x_2 & +c_{31}x_3 & = & y_2 \\ c_{32}x_3 & +c_{41}x_4 & = & y_3 \\ c_{42}x_4 & +c_{51}x_5 & = & y_4 \\ c_{52}x_5 & +c_{61}x_6 & = & y_5 \\ & c_{62}x_6 & = & y_6 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{llcl} 2x_1 & x_2 & = & 1 \\ 2x_2 & -x_3 & = & 0 \\ x_3 & +x_4 & = & 4 \\ 3x_4 & -2x_5 & = & 4 \\ x_5 & +x_6 & = & 1 \\ & 2x_6 & = & 0 \end{array} \right.$$

cuya solución es $\overrightarrow{x}^T=(0,1,2,2,1,0)$.

Observación

Se propone como ejercicio la elaboración de un algoritmo completo para la resolución del sistema de ecuaciones $A\overrightarrow{x}=\overrightarrow{b}$, donde $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ es tridiagonal, simétrica, definida positiva.

El número total de operaciones elementales en la resolución del sistema de ecuaciones $A\overrightarrow{x}=\overrightarrow{b}$, mediante el algoritmo arriba descrito es

$$Noper^{(c)}(n)=10n-7,\ \ n\in\mathbb{Z}^+,$$

es decir que $Noper > Noper = 8n-7$. El método de eliminación gaussiana es mucho mejor que el método de Choleski, visto respecto del número de operaciones.

Método de Crout para matrices tridiagonales

Sean $A=(a_{ij})\in M_{n\times n}[\mathbb{R}]$ una matriz tridiagonal no nula, $\overrightarrow{b}^T=(b_1,...,b_n)\in\mathbb{R}^n$. Consideramos el sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{b}$.

En el método de Crout se busca (si existen) una factorización de la matriz A en la forma LU, es decir, $A = LU$, donde

$$L = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ 0 & l_{32} & l_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & l_{nn} \end{bmatrix}, \quad U = \begin{bmatrix} 1 & u_{12} & 0 & \cdots & 0 \\ 0 & 1 & u_{23} & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & u_{n-1,n} \\ 0 & \cdots & \cdots & \cdots & 1 \end{bmatrix}.$$

Como se ha dicho anteriormente, los elementos de interés de la matriz A se guardan en la matriz B. De acuerdo a la estructura que presentan las matrices L y U, se definen las matrices $\widetilde{L} = \left(\widetilde{l}_{ik}\right)$, $\widetilde{U} = (\widetilde{u}_{ik})$ de $M_{n\times 2}[\mathbb{R}]$ siguientes

$$\begin{aligned} \widetilde{l}_{11} &= 0 \\ \widetilde{l}_{i1} &= l_{ii-1}, \quad i = 2, ..., n, \\ \widetilde{l}_{i2} &= l_{ii}, \quad i = 1, ..., n, \\ \widetilde{u}_{i1} &= 1, \quad i = 1, ..., n, \\ \widetilde{u}_{i2} &= u_{ii+1}, \quad i = 1, ..., n-1, \\ \widetilde{u}_{n2} &= 0, \end{aligned}$$

o sea, las matrices $\widetilde{L}$, $\widetilde{U}$ tienen la forma

$$\widetilde{L} = \begin{bmatrix} 0 & l_{11} \\ l_{21} & l_{22} \\ \vdots & \vdots \\ l_{nn-1} & l_{nn} \end{bmatrix}, \quad \widetilde{U} = \begin{bmatrix} 1 & u_{12} \\ 1 & u_{23} \\ \vdots & \vdots \\ 1 & u_{n-1,n} \\ 1 & 0 \end{bmatrix}.$$

El algoritmo de factorización LU para matrices tridiagonales se reduce al siguiente:

1. $l_{11} = a_{11}$
2. $u_{12} = \dfrac{a_{12}}{l_{11}}$.
3. Para $i = 2, ..., n-1$

   $l_{i,i-1} = a_{ii-1}$

   $l_{ii} = a_{ii} - l_{i,i-1}u_{i-1,i}$

   $u_{ii+1} = \dfrac{a_{ii+1}}{l_{ii}}$.

   Fin de bucle i.

4. $l_{n,n-1} = a_{nn-1}$.
5. $l_{nn} = a_{nn} - l_{nn-1}u_{n-1,n}$.

Este algoritmo en términos de las matrices B, $\widetilde{L}$, $\widetilde{U}$ se expresa en los siguientes términos.

1. $\widetilde{l}_{12} = b_{12}$.
2. $\widetilde{u}_{12} = \dfrac{b_{13}}{\widetilde{l}_{12}}$.
3. Para $i = 2, ..., n-1$

   $\widetilde{l}_{i1} = b_{i1}$

$$\widetilde{l}_{i2} = b_{i2} - \widetilde{l}_{i1}\widetilde{u}_{i-1,2}$$

$$\widetilde{u}_{i2} = \frac{b_{i3}}{\widetilde{l}_{i2}}.$$

Fin de bucle i.

4. $\widetilde{l}_{n1} = b_{n1}$.
5. $\widetilde{l}_{n2} = b_{n2} - \widetilde{l}_{n1} * \widetilde{u}_{n-1,2}$.

El sistema de ecuaciones triangular inferior $L\overrightarrow{y} = \overrightarrow{b}$ expresado en términos de la matriz $\widetilde{L}$ es:

$$\begin{cases} \widetilde{l}_{12}y_1 & & = b_1 \\ \widetilde{l}_{21}\widetilde{y}_1 & \widetilde{l}_{22}y_2 & = b_2 \\ \widetilde{l}_{31}\widetilde{y}_2 & +l_{32}y_3 & = b_3 \\ & & \vdots \\ l_{n1}y_{n-1} & +l_{n2}y_n & = b_n, \end{cases}$$

cuya solución es

1. $y_1 = \dfrac{b_1}{\widetilde{l}_{12}}$.
2. Para $i = 2, ..., n$

$$y_i = \frac{(b_i - l_{i1}y_{i-1})}{l_{i2}}.$$

Fin de bucle i.

El sistema de ecuaciones triangular superior $U\overrightarrow{x} = \overrightarrow{y}$ expresado en términos de la matriz $\widetilde{U}$ es

$$\begin{cases} x_1 & +u_{12}x_2 & = & y_1 \\ x_2 & +u_{22}x_2 & = & y_2 \\ & & \vdots & \\ & x_n & = & y_n. \end{cases}$$

cuya solución es:

1. $x_n = y_n$.
2. Para $i = n-1, ..., 1$

$$x_i = y_i - u_{i2}x_{i+1}.$$

Fin de bucle i.

El número de operaciones elementales para la resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ mediante el método de Crout, con A matriz tridiagonal, es

$$Noper\,(n) = 8n - 7, \;\; n \in \mathbb{Z}^+.$$

Se observa que el número de operaciones elementales para la resolución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, con A una matriz tridiagonal, estrictamente diagonalmente dominante, mediante los métodos de eliminación gaussiana y Crout, coinciden.

Ejemplo

Considerar el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, con

$$A = \begin{bmatrix} 1 & 2 & 0 & 0 & 0 \\ -2 & 0 & 6 & 0 & 0 \\ 0 & 4 & 13 & 3 & 0 \\ 0 & 0 & -3 & -11 & -2 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} -3 \\ 4 \\ 2 \\ -42 \\ -18 \end{bmatrix}.$$

La matriz B está definida como

$$B = \begin{bmatrix} 0 & 1 & 2 \\ -2 & 0 & 6 \\ 4 & 13 & 3 \\ -3 & -11 & -2 \\ 1 & -2 & 0 \end{bmatrix}.$$

Comenzamos con la contrucción (si existen) de las matrices $\widetilde{L}$ y $\widetilde{U}$ usando el algoritmo arriba presentado.

1. $\widetilde{l}_{12} = b_{12} = 1$.
2. $\widetilde{u}_{12} = \dfrac{b_{13}}{\widetilde{l}_{12}} = \frac{1}{1} = 1$,
3. Para $i = 2, 3, 4$

   $i = 2$, $\widetilde{l}_{21} = b_{21} = -2$,
   $\widetilde{l}_{22} = b_{22} - \widetilde{l}_{21}\widetilde{u}_{12} = 0 - (-2) - 1 = 2$,
   $\widetilde{u}_{22} = \dfrac{b_{23}}{\widetilde{l}_{22}} = \frac{6}{2} = 3$,

   $i = 3$, $\widetilde{l}_{31} = b_{31} = 4$
   $\widetilde{l}_{32} = b_{32} - \widetilde{l}_{31}\widetilde{u}_{22} = 13 - 4 \times 3 = 1$,
   $\widetilde{u}_{32} = \dfrac{b_{33}}{\widetilde{l}_{32}} = \frac{3}{1} = 3$,

   $i = 4$, $\widetilde{l}_{41} = b_{41} = -3$,
   $\widetilde{l}_{42} = b_{42} - \widetilde{l}_{41\times}\widetilde{u}_{32} = -11 - (-3) \times 3 = -2$,
   $u_{42} = \dfrac{b_{43}}{\widetilde{l}_{42}} = -\frac{-2}{-2} = 1$,
4. $\widetilde{l}_{51} = b_{51} = 1$
5. $\widetilde{l}_{52} = b_{52} - b_{51} \times \widetilde{u}_{42} = -2 - 1 \times 1 = -3$.

Así,

$$\widetilde{L} = \begin{bmatrix} 0 & 1 \\ -2 & 2 \\ 4 & 1 \\ -3 & -2 \\ 1 & -3 \end{bmatrix}, \quad \widetilde{U} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 3 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

La solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, es por lo tanto, $\overrightarrow{x}^T = (-2, -1, 0, 2, 10)$.

Observación

Los métodos de eliminació gaussiana, Choleski y Crout descritos para la resolución del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, con $A \in M_{n\times n}[\mathbb{R}]$ matriz tridiagonal más hipótesis suplementarias sobre A, pueden aplicarse, en general, a matrices $A = (a_{ij})$ en banda, con longitud de banda $lb = 1, 2, ...$, de modo que $lb < n$ y lb no muy grande $\left(lb < \frac{n}{2}\right)$, donde

$$a_{ij} = 0 \quad \text{si} \quad |i - j| > lb, \quad i,\, j = 1, ..., n.$$

Cuando n es grande y $lb < n$ es grande, se debe pensar en guardar los datos a_{ij} con $|i - j| \leq lb$, i, $j = 1, ..., n$, en arreglos (matrices) o archivos adecuados y con estos arreglos o archivos elaborar algoritmos adecuados de resolución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

Si n es muy grande, $lb < n$ es pequeño con respecto de n, en este caso es recomendable los métodos iterativos que serán presentados más adelante.

6.10. Resolución de un sistema de ecuaciones lineales en norma mínima.

Sean $A=(a_{ij})\in M_{m\times n}[\mathbb{R}]$ tal que $R(A)=m<n$, y $\overrightarrow{b}^T=(b_1,...,b_n)\in\mathbb{R}^n$. El sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{b}$ posee una infinidad de soluciones. Denotamos con S el conjunto de todas estas soluciones, esto es,

$$S=\left\{\overrightarrow{x}\in\mathbb{R}\,|A\overrightarrow{x}=\overrightarrow{b}\right\}.$$

Consideramos el problema (P) siguiente: hallar $\widehat{x}\in S$, si existe, tal que

$$\|\widehat{x}\|^2=\underset{\overrightarrow{x}\in S}{\text{Min}}\|\widehat{x}\|^2,$$

o lo que es lo mismo $\|\widehat{x}\|^2\leq\|\overrightarrow{x}\|^2\quad\forall\overrightarrow{x}\in S$, que a su vez puede escribirse como

$$\|\widehat{x}\|^2=\underset{\substack{\overrightarrow{x}\in\mathbb{R}^n\\A\overrightarrow{x}=\overrightarrow{b}}}{\text{Min}}\|\overrightarrow{x}\|^2.$$

Este es un problema de extremos condicionados en el que se busca minimizar la función g definida por

$$g(\overrightarrow{x})=\overrightarrow{x}^T\overrightarrow{x},\quad\overrightarrow{x}\in\mathbb{R}^n,$$

sujeta a la restricción $A\overrightarrow{x}=\overrightarrow{b}$. El método de los multiplicadores de Lagrange proporciona una condición necesaria de extremo. Sea $\overrightarrow{\lambda}^T=(\lambda_1,...,\lambda_m)\in\mathbb{R}^m$ con $\overrightarrow{\lambda}\neq0$ y definimos la finción Φ de $\mathbb{R}^n$ en $\mathbb{R}$ como

$$\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)=g(\overrightarrow{x})+\overrightarrow{\lambda}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right)=\overrightarrow{x}^T\overrightarrow{x}+\overrightarrow{\lambda}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right),\quad\overrightarrow{x}\in\mathbb{R}^n.$$

Las componentes de $\overrightarrow{\lambda}$: $\lambda_1,...,\lambda_m$ se llaman multiplicadores de Lagrange y $\overrightarrow{\lambda}$ se llama vector multiplicador de Lagrange. Las condiciones necesarias de extremo establecen que

$$\nabla_{\overrightarrow{x}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)=0,\qquad\nabla_{\overrightarrow{\lambda}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)=0,$$

donde $\nabla_{\overrightarrow{x}},\nabla_{\overrightarrow{\lambda}}$ denotan los operadores gradiente con respecto de $\overrightarrow{x}$ y con respecto de $\overrightarrow{\lambda}$, respectivamente.

Para el efecto, determinemos la derivada direccional de Φ con respecto de $\overrightarrow{x}$ y de $\overrightarrow{\lambda}$ según las direcciones $\overrightarrow{y}\in\mathbb{R}^n$ y $\overrightarrow{\alpha}\in\mathbb{R}^m$ que se escriben $D_{\overrightarrow{y}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)$ y $D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)$, esto es,

$$\begin{aligned}D_{\overrightarrow{y}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)&=\lim_{t\to0}\frac{\Phi\left(\overrightarrow{x}+t\overrightarrow{y},\overrightarrow{\lambda}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)}{t},\\D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)&=\lim_{t\to0}\frac{\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}+t\overrightarrow{\alpha}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)}{t}.\end{aligned}$$

Comencemos con el cálculo de la derivada direccional $D_{\overrightarrow{y}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)$ en $\overrightarrow{x}\in\mathbb{R}^n$, $\overrightarrow{\lambda}\in\mathbb{R}^m$ según la dirección $\overrightarrow{y}\in\mathbb{R}^n$.

Sean $t\neq0$, $\overrightarrow{y}\in\mathbb{R}^n$ con $\overrightarrow{y}\neq0$. Entonces,

$$\begin{aligned}\Phi\left(\overrightarrow{x}+t\overrightarrow{y},\overrightarrow{\lambda}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)&=(\overrightarrow{x}+t\overrightarrow{y})^T(\overrightarrow{x}+t\overrightarrow{y})+\overrightarrow{\lambda}^T\left(A(\overrightarrow{x}+t\overrightarrow{y})-\overrightarrow{b}\right)\\&\quad-\left(\overrightarrow{x}^T\overrightarrow{x}+\overrightarrow{\lambda}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right)\right)\\&=t\overrightarrow{x}^T\overrightarrow{y}+t\overrightarrow{y}^T\overrightarrow{x}+t^2\overrightarrow{y}^T\overrightarrow{y}+t\overrightarrow{\lambda}^TA\overrightarrow{y}.\end{aligned}$$

Puesto que $\overrightarrow{y}^T\overrightarrow{x}=\overrightarrow{x}^T\overrightarrow{y}$, y $\overrightarrow{y}^T\overrightarrow{y}=\|\overrightarrow{y}\|^2$, resulta

$$\Phi\left(\overrightarrow{x}+t\overrightarrow{y},\overrightarrow{\lambda}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)=2t\overrightarrow{x}^T\overrightarrow{y}+t^2\|\overrightarrow{y}\|^2+t\overrightarrow{\lambda}^TA\overrightarrow{y}=t\left(2\overrightarrow{x}^T\overrightarrow{y}+\overrightarrow{\lambda}^TA\overrightarrow{y}+t\|\overrightarrow{y}\|^2\right).$$

Luego,

$$\begin{aligned} D_{\overrightarrow{y}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) &= \lim_{t\to 0}\frac{\Phi\left(\overrightarrow{x}+t\overrightarrow{y},\overrightarrow{\lambda}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)}{t} = \lim_{t\to 0}\left(2\overrightarrow{x}^T\overrightarrow{y}+\overrightarrow{\lambda}^T A\overrightarrow{y}+t\left\|\overrightarrow{y}\right\|^2\right) \\ &= 2\overrightarrow{x}^T\overrightarrow{y}+\overrightarrow{\lambda}^T A\overrightarrow{y} = \left(2\overrightarrow{x}^T+\overrightarrow{\lambda}^T A\right)\overrightarrow{y}. \end{aligned}$$

Como la derivada direccional $D_{\overrightarrow{y}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)$ existe en toda dirección $\overrightarrow{y}$ y es continua en $\overrightarrow{x}$, $\overrightarrow{\lambda}$, entonces

$$D_{\overrightarrow{y}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = \left(\nabla_{\overrightarrow{x}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)\right)^T\overrightarrow{y},$$

es decir

$$\left(\nabla_{\overrightarrow{x}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)\right)^T\overrightarrow{y} = \left(2\overrightarrow{x}^T+\overrightarrow{\lambda}^T A\right)\overrightarrow{y},$$

de donde

$$\nabla_{\overrightarrow{x}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = 2\overrightarrow{x}+A^T\overrightarrow{\lambda}.$$

Calculemos la derivada direccional $D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)$ en $\overrightarrow{x}\in\mathbb{R}^n$, $\overrightarrow{\lambda}\in\mathbb{R}^m$ según la dirección $\overrightarrow{\alpha}\in\mathbb{R}^m$. Para el efecto, sean $\overrightarrow{x}\in\mathbb{R}^n$, $\overrightarrow{\alpha}\in\mathbb{R}^m$ con $\overrightarrow{\alpha}\neq 0$, y $t\neq 0$. Entonces

$$\begin{aligned} \Phi\left(\overrightarrow{x},\overrightarrow{\lambda}+t\overrightarrow{\alpha}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) &= \overrightarrow{x}^T\overrightarrow{x}+\left(\overrightarrow{\lambda}+t\overrightarrow{\alpha}\right)^T\left(A\overrightarrow{x}-\overrightarrow{b}\right)-\left(\overrightarrow{x}^T\overrightarrow{x}+\overrightarrow{\lambda}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right)\right) \\ &= t\overrightarrow{\alpha}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right). \end{aligned}$$

Luego,

$$\begin{aligned} D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) &= \lim_{t\to 0}\frac{\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}+t\overrightarrow{\alpha}\right)-\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)}{t} = \lim_{t\to 0}\overrightarrow{\alpha}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right) \\ &= \overrightarrow{\alpha}^T\left(A\overrightarrow{x}-\overrightarrow{b}\right) = \left(A\overrightarrow{x}-\overrightarrow{b}\right)^T\overrightarrow{\alpha}. \end{aligned}$$

Por lo tanto,

$$D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = \left(A\overrightarrow{x}-\overrightarrow{b}\right)^T\overrightarrow{\alpha}.$$

De la existencia de $D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)$ en $\overrightarrow{\alpha}\in\mathbb{R}^m$ y la continuidad en $\overrightarrow{\lambda}$, se sigue que

$$D_{\overrightarrow{\alpha}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = \left(\nabla_{\overrightarrow{\lambda}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right)\right)^T\overrightarrow{\alpha},$$

de donde

$$\nabla_{\overrightarrow{\lambda}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = A\overrightarrow{x}-\overrightarrow{b}.$$

Consecuentemente, las condiciones necesarias de extremo

$$\begin{cases} \nabla_{\overrightarrow{x}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = 2\overrightarrow{x}+A^T\overrightarrow{b} = 0, \\ \nabla_{\overrightarrow{\lambda}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = A\overrightarrow{x}-\overrightarrow{b} = 0. \end{cases}$$

Note que $\nabla_{\overrightarrow{\lambda}}\Phi\left(\overrightarrow{x},\overrightarrow{\lambda}\right) = 0$ es equivalente a introducir la restricción $A\overrightarrow{x} = \overrightarrow{b}$.

De la ecuación $2\overrightarrow{x}+A^T\overrightarrow{\lambda} = 0$, obtenemos $\overrightarrow{x} = -\frac{1}{2}A^T\overrightarrow{\lambda}$ con lo cual

$$0 = A\overrightarrow{x}-\overrightarrow{b} = A\left(-\frac{1}{2}A^T\overrightarrow{\lambda}\right)-\overrightarrow{b} = -\frac{1}{2}AA^T\overrightarrow{\lambda}-\overrightarrow{b}$$

de donde

$$AA^T\overrightarrow{\lambda} = -2\overrightarrow{b}.$$

Como $R(A) = m$ y $AA^T \in M_{m\times m}[\mathbb{R}]$, se tiene que el rango de AA^T es también m, esto es, $R\left(AA^T\right) = m$, con lo cual AA^T es invertible. Luego

$$\overrightarrow{\lambda} = -2\left(AA^T\right)^{-1}\overrightarrow{b}$$

y

$$\overrightarrow{x} = -\frac{1}{2}A^T\overrightarrow{\lambda} = -\frac{1}{2}A^T\left[-2\left(AA^T\right)^{-1}\overrightarrow{b}\right] = A^T\left(AA^T\right)^{-1}\overrightarrow{b}.$$

Así,

$$\widehat{x} = A^T\left(AA^T\right)^{-1}\overrightarrow{b} \text{ es solución de } \underset{\overrightarrow{x}\in S}{\text{Min}}\|\overrightarrow{x}\|^2.$$

Verifiquemos que $\widehat{x} \in S$ y que $\|\widehat{x}\|^2 \leq \|\overrightarrow{x}\|^2 \quad \forall\overrightarrow{x} \in S$. Se tiene

$$A\widehat{x} = A\left(A^T\left(AA^T\right)^{-1}\overrightarrow{b}\right) = AA^T\left(AA^T\right)^{-1}\overrightarrow{b} = I\overrightarrow{b} = \overrightarrow{b},$$

donde $I \in M_{m\times m}[\mathbb{R}]$ es la identidad. Así, $A\overrightarrow{x} = \overrightarrow{b}$ que muestra $\widehat{x} \in S$.

Sea $\overrightarrow{x} \in S$. Entonces $A\overrightarrow{x} = \overrightarrow{b}$. Se propone como ejercicio mostrar que $\|\widehat{x}\|^2 \leq \|\overrightarrow{x}\|^2 \quad \forall\overrightarrow{x} \in S$.

Algoritmo.

Para el cálculo de $\widehat{x} = A^T\left(AA^T\right)^{-1}\overrightarrow{b}$ se utiliza el algoritmo siguiente que evita la inversión directa de la matriz AA^T.

Sea $\overrightarrow{z} = \left(AA^T\right)^{-1}\overrightarrow{b}$ entonces $AA^T\overrightarrow{z} = \overrightarrow{b}$. Luego $\overrightarrow{x} = A^T\overrightarrow{z}$.

1. Calcular $B = AA^T$.
2. Aplicar el método de eliminación gaussiana para resolver el sistema de ecuaciones lineales $B\overrightarrow{z} = \overrightarrow{b}$.
3. Calcular $\widehat{x} = A^T\overrightarrow{z}$.

Ejemplos

1. Sean $a_1, ..., a_n \in \mathbb{R}$ con $a_i \neq 0, \quad i = 1, ..., n, \ b \neq 0$. Consideramos la ecuación

$$a_1x_1 + ... + a_nx_n = b.$$

Esta ecuación admite una infinidad de soluciones. Resolvemos el problema en norma mínima. Ponemos $A = (a_1, ..., a_n)$, $\overrightarrow{b} = b$, $\overrightarrow{x}^T(x_1, ..., x_n) \in \mathbb{R}^n$. Entonces

$$A\overrightarrow{x} = \overrightarrow{b} \Leftrightarrow a_1x_1 + ... + a_nx_n = b.$$

Se tiene

$$B = AA^T = (a_1, ..., a_n)\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \sum_{i=1}^{n} a_i^2.$$

La solución de la ecuación $Bz = b$ es $z = \dfrac{b}{\sum\limits_{i=1}^{n} a_i^2}$. Luego

$$\widehat{x} = A^Tz = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}\frac{b}{\sum\limits_{i=1}^{n} a_i^2} = \begin{bmatrix} \frac{a_1b}{\sum\limits_{i=1}^{n} a_i^2} \\ \vdots \\ \frac{a_nb}{\sum\limits_{i=1}^{n} a_i^2} \end{bmatrix}.$$

Por ejemplo la solución de la ecuación $x + 2y + z = 1$ es$\widehat{x}^T = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{6}\right)$.

2. Consideramos el sistema de ecuaciones lineales $\begin{cases} 3x+y+z=1 \\ -x+y+2z=0. \end{cases}$ Tenemos $A=\begin{bmatrix} 3 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix}$, $\overrightarrow{b}=\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Luego $A^T=\begin{bmatrix} 3 & -1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$,

$$B=AA^T=\begin{bmatrix} 3 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix}\begin{bmatrix} 3 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}=\begin{bmatrix} 11 & 0 \\ 0 & 6 \end{bmatrix}.$$

La solución del sistema de ecuaciones $B\overrightarrow{z}=\overrightarrow{b}$:

$$\begin{bmatrix} 11 & 0 \\ 0 & 6 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}=\begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

es $\overrightarrow{z}^T=\left[\frac{1}{11},\frac{1}{6}\right]$.

Entonces $\widehat{x}=A^T\overrightarrow{z}=\begin{bmatrix} 3 & -1 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} \frac{1}{11} \\ \frac{1}{6} \end{bmatrix}=\begin{bmatrix} \frac{7}{66} \\ \frac{17}{66} \\ \frac{28}{66} \end{bmatrix}$.

6.11. Condicionamiento.

Sean $A\in M_{n\times n}[\mathbb{R}]$ una matriz invertible, $\overrightarrow{b}\in\mathbb{R}^n$. Consideramos el problema (P) siguiente: hallar $\overrightarrow{x}\in\mathbb{R}^n$ solución del sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{b}$.

En general, los coeficientes de la matriz A y los componentes del vector $\overrightarrow{b}$ son redondeados antes de ingresar al computador. Esto hace que no tratemos el problema (P) sino el sistema de ecuaciones lineales siguiente, llamado problema $(\widetilde{\mathrm{P}})$: $\widetilde{A}\widetilde{x}=\widetilde{b}$, de donde $\widetilde{A}$ es la matriz obtenida de A por redondeo de sus coeficientes, $\widetilde{b}$ es obtenido de $\overrightarrow{b}$ por redondeo de sus componentes.

Un método aproximado para resolver $(\widetilde{\mathrm{P}})$ es el de eliminación gaussiana. Usando este método, hallamos un vector $\widetilde{x}$ que, en general, es diferente de $\overrightarrow{x}$. La pregunta que nos ponemos es: ¿cómo influencian estas perturbaciones en la solución del problema?.Consideramos dos casos:

Los coeficientes de la matriz A son números de máquina y perturbamos el vector $\overrightarrow{b}$.

Los componentes de vector $\overrightarrow{b}$ son números de máquina y perturbamos los coeficientes de la matriz A.

Supongamos que $\overrightarrow{x}+\Delta\overrightarrow{x}$ es la solución del sistema de ecuaciones lineales $A(\overrightarrow{x}+\Delta\overrightarrow{x})=\overrightarrow{b}+\Delta\overrightarrow{b}$.

Las normas de matrices que utilizaremos a continuación son submultiplicativas, véase en el apéndice normas en $\mathbb{R}^n$ y normas de matrices. La norma $\|\cdot\|$ en $\mathbb{R}^n$ que utilizaremos, es la norma euclídea y la norma de matrices submultiplicativa es cualesquiera.

Puesto que $A\overrightarrow{x}=\overrightarrow{b}$, se tiene entonces $A\Delta\overrightarrow{x}=\Delta\overrightarrow{b}$, o bien $\Delta\overrightarrow{x}=A^{-1}\Delta\overrightarrow{b}$. Resulta que $\|\Delta\overrightarrow{x}\|\leq\||A^{-1}|\|\left\|\Delta\overrightarrow{b}\right\|$.

Como $A\overrightarrow{x}=\overrightarrow{b}$, entonces $\left\|\overrightarrow{b}\right\|=\|A\overrightarrow{x}\|\leq\||A|\|\,\|\overrightarrow{x}\|$. Por lo tanto el error relativo de $\overrightarrow{x}$ definido por la relación $\frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|}$ se mayora por

$$\frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|}\leq\frac{\||A^{-1}|\|\left\|\Delta\overrightarrow{b}\right\|}{\|\overrightarrow{x}\|}\leq\||A^{-1}|\|\,\||A|\|\frac{\left\|\Delta\overrightarrow{b}\right\|}{\left\|\overrightarrow{b}\right\|}.$$

El condicionamiento de la matriz A se nota con $cond(A)$ y se define como sigue: $cond(A)=\||A^{-1}|\|\,\||A|\|$. El condicionamiento de una matriz es muy importante en la resolución de los sistemas de

ecuaciones lineales así como en el cálculo de los valores y vectores propios. La calidad de las soluciones numéricas de los sistemas de ecuaciones lineales está ligado al condicionamiento de la matriz.

1. Puesto que $I = AA^{-1}$, entonces $1 \geq |||A||| \left|\left|\left|A^{-1}\right|\right|\right| = cond\,(A)$. El condicionamiento de A mide la sensibilidad del error relativo de la solución del sistema de ecuaciones a cambios o perturbaciones en el vector $\overrightarrow{b}$. Se dice que el sistema de ecuaciones lineales está mal condicionado si la matriz A está mal condicionada, es decir que el número $cond\,(A)$ es muy grande.

2. Sea $A \in M_{n\times n}[\mathbb{R}]$ una matriz invertible, $\widetilde{A}$ la matriz perturbada de A. Consideramos los sistemas de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, y, $\widetilde{A}\widetilde{x} = \overrightarrow{b}$. Tenemos entonces que

$$0 = A\overrightarrow{x} - \widetilde{A}\widetilde{x} = \left(A - \widetilde{A}\right)\overrightarrow{x} + \widetilde{A}\left(\overrightarrow{x} - \widetilde{x}\right),$$

con lo cual

$$\widetilde{A}\left(\overrightarrow{x} - \widetilde{x}\right) = \left(\widetilde{A} - A\right)\overrightarrow{x}.$$

Notamos $\Delta\overrightarrow{x} = \overrightarrow{x} - \widetilde{x}$, $\Delta A = \widetilde{A} - A$. Por la igualdad precedente, se tiene $\widetilde{A}\Delta\overrightarrow{x} = \Delta A\overrightarrow{x}$. Para estimar el error relativo $\frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|}$, probemos primeramente el teorema siguiente:

Teorema 11 *Sea $B \in M_{n\times n}[\mathbb{R}]$ tal que $|||B||| < 1$. Entonces $(I+B)^{-1}$ existe, y*

$$\left|\left|\left|(I+B)^{-1}\right|\right|\right| \leq \frac{1}{1-|||B|||}.$$

Demostración. Sea $T_B : \mathbb{R}^n \to \mathbb{R}^n$ la aplicación lineal definida por $T_B = (\overrightarrow{x}) = (I+B)\overrightarrow{x}$. Probemos que el núcleo de la transformación T, esto es $\ker(T_B) = \{\overrightarrow{x} \in \mathbb{R}^n \mid T_B(\overrightarrow{x}) = \overrightarrow{0}\}$ se reduce al vector nulo, es decir $\ker(T_B) = \{\overrightarrow{0}\}$ y de esta igualdad se deduce que T_B es inyectiva. Utilizando la desigualdad

$$|\|\overrightarrow{x}\| - \|\overrightarrow{y}\|| \leq \|\overrightarrow{x} - \overrightarrow{y}\|, \quad \forall\overrightarrow{x},\ \overrightarrow{y} \in \mathbb{R}^n,$$

y poniendo $\overrightarrow{y} = -B\overrightarrow{x}$, para $\overrightarrow{x} \neq 0$ se tiene

$$\begin{aligned} \|T_B(\overrightarrow{x})\| &= \|(I+B)\overrightarrow{x}\| = \|\overrightarrow{x} + B\overrightarrow{x}\| \geq \|\overrightarrow{x}\| - \|B\overrightarrow{x}\| \\ &\geq \|\overrightarrow{x}\| - |||B|||\,\|\overrightarrow{x}\| = \|\overrightarrow{x}\|(1 - |||B|||) > 0, \end{aligned}$$

pues $|||B||| < 1$ y $\|\overrightarrow{x}\| > 0$. Luego $\ker(T_B) = \{0\}$, es decir que T_B es invertible o sea T_B^{-1} existe y como la matriz asociada a T_B^{-1} relativa a la base canónica de $\mathbb{R}^n$ es $(I+B)^{-1}$. Se tiene entonces la existencia de $(I+B)^{-1}$.

Probemos que $\left|\left|\left|(I+B)^{-1}\right|\right|\right| \leq \frac{1}{1-|||B|||}$. Sea $C = (I+B)^{-1}$. Entonces

$$\begin{aligned} 1 &= |||I||| = |||(I+B)C||| = |||C + BC||| \geq |||C||| - |||BC||| \\ &\geq |||C||| - |||B|||\,|||C||| = |||C|||\,(1 - |||B|||), \end{aligned}$$

de donde $|||C||| \leq \frac{1}{1-|||B|||}$. ■

Teorema 12 *Sea $A \in M_{n\times n}[\mathbb{R}]$ no singular, $B = A(I+F)$, donde $F \in M_{n\times n}[\mathbb{R}]$ tal que $|||F||| < 1$ y $\overrightarrow{b} \in \mathbb{R}^n$. Sean $\overrightarrow{x}, \Delta\overrightarrow{x} \in \mathbb{R}^n$ las soluciones respectivamente de $A\overrightarrow{x} = \overrightarrow{b}$, y, $B(\overrightarrow{x} + \Delta\overrightarrow{x}) = \overrightarrow{b}$. Se tiene entonces las siguientes estimaciones:*

i) $\dfrac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leq \dfrac{|||F|||}{1-|||F|||}.$

ii) $\dfrac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leq \dfrac{cond\,(A)}{1 - cond\,(A)\,\frac{|||B-A|||}{|||A|||}}$, *si* $cond\,(A) = \frac{|||B-A|||}{|||A|||} < 1.$

Demostración. Por definición de $\overrightarrow{x}$ y $\Delta\overrightarrow{x}$, tenemos

$$\begin{aligned} B\Delta\overrightarrow{x} &= \overrightarrow{b} - B\overrightarrow{x} = \overrightarrow{b} - BA^{-1}\overrightarrow{b} = \left(I - BA^{-1}\right)\overrightarrow{b} = \left(I - BA^{-1}\right)AA^{-1}\overrightarrow{b} = (A-B)\,A^{-1}\overrightarrow{b} \\ &= (A-B)\,\overrightarrow{x}. \end{aligned}$$

Resulta que B es una perturbación de A.

i. Por el teorema precedente, la matriz $I+F$ es no singular y por hipótesis A es igualmente no singular, se tiene entonces que $B = A(I+F)$ es no singular. Además $\Delta\overrightarrow{x} = B^{-1}(A-B)\,\overrightarrow{x}$. Entonces $\frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leq \left|\left|\left|B^{-1}(A-B)\right|\right|\right|$. Pero

$$\begin{aligned} B^{-1}(A-B) &= (I+F)^{-1}A^{-1}(A-B) = (I+F)^{-1}\left(I - A^{-1}B\right) = (I+F)^{-1}(I-(I+F)) \\ &= -(I+F)^{-1}F, \end{aligned}$$

$$\left|\left|\left|B^{-1}(A-B)\right|\right|\right| = \left|\left|\left|(I+F)^{-1}F\right|\right|\right| \leq \left|\left|\left|(I+F)^{-1}\right|\right|\right| |||F||| \leq \frac{|||F|||}{1-|||F|||}.$$

Así $\frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leq \frac{|||F|||}{1-|||F|||}$.

ii. Puesto que $F = A^{-1}B - I = A^{-1}(B-A)$, entonces $|||F||| \leq \left|\left|\left|A^{-1}\right|\right|\right| |||B-A|||$. Resulta que

$$\begin{aligned} \frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} &\leq \frac{|||F|||}{1-|||F|||} \leq \frac{\left|\left|\left|A^{-1}\right|\right|\right| |||B-A|||}{1-|||A^{-1}||| \, |||B-A|||} \\ &= \frac{\left|\left|\left|A^{-1}\right|\right|\right| |||A|||}{1-|||A^{-1}||| \, |||A||| \, \frac{|||B-A|||}{|||A|||}} \frac{|||B-A|||}{|||A|||} = \frac{cond(A)}{1-cond(A)\,\frac{|||B-A|||}{|||A|||}} \frac{|||B-A|||}{|||A|||}. \end{aligned}$$

Observe que si B es una perturbación de A tal que $|||B-A||| = \frac{1}{|||A^{-1}|||}$, entonces $cond(A) = \frac{|||A|||}{|||B-A|||}$.

■

Consideremos nuevamente el problema (P): $A\overrightarrow{x} = \overrightarrow{b}$.

Se perturban los datos A y $\overrightarrow{b}$ respectivamente por ΔA y $\Delta\overrightarrow{b}$, lo que da un resultado perturbado $\Delta\overrightarrow{x}$ de $\overrightarrow{x}$. Se tiene entonces $(A+\Delta A)(\overrightarrow{x}+\Delta\overrightarrow{x}) = \overrightarrow{b} + \Delta\overrightarrow{b}$. Por el teorema precedente, si $1 - cond(A)\,\frac{|||B-A|||}{|||A|||} > 0$, y como B es una perturbación de A, esto es, $B - A = \Delta A$, entonces $1 - \left|\left|\left|A^{-1}\right|\right|\right| |||\Delta A||| > 0$ de donde $|||\Delta A||| < \frac{1}{|||A^{-1}|||}$. Se prueba entonces que

$$\frac{\|\Delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leq \frac{cond(A)}{1-cond(A)\,\frac{|||\Delta A|||}{|||A|||}} \left(\frac{\left\|\Delta\overrightarrow{b}\right\|}{\left\|\overrightarrow{b}\right\|} + \frac{|||\Delta A|||}{|||A|||}\right).$$

Ejemplo

Considerar el sistema de ecuaciones $\begin{cases} \varepsilon x + 1000y = 1 \\ x + 1000y = 2, \end{cases}$ donde $\varepsilon > 0$, $\varepsilon \neq 1$. Apliquemos el método de eliminación gaussiana sin pivoting. Tenemos $\begin{cases} \varepsilon x + 1000y = 1 \\ \left(1 - \frac{1}{\varepsilon}\right) 1000y = 2 - \frac{1}{\varepsilon} = 2 - \varepsilon^{-1}, \end{cases}$ de donde

$$\begin{aligned} y &= \frac{2-\varepsilon^{-1}}{(1-\varepsilon^{-1})\,1000} \\ x &= \frac{1-1000y}{\varepsilon} = \frac{1-1000\left(\frac{2-\varepsilon^{-1}}{(1-\varepsilon^{-1})1000}\right)}{\varepsilon} = \frac{1-\varepsilon^{-1}-2+\varepsilon^{-1}}{\varepsilon(1-\varepsilon^{-1})} = -\frac{1}{\varepsilon-1} = \frac{1}{1-\varepsilon}. \end{aligned}$$

Así, para $\varepsilon > 0$ y $\varepsilon \neq 1$, la solución del sistema de ecuaciones es $\left\{ \begin{array}{c} x = \frac{1}{1-\varepsilon}, \\ y = \frac{2-\varepsilon^{-1}}{(1-\varepsilon^{-1})} 10^{-3}. \end{array} \right.$ Para $\varepsilon = 10^{-4}$, calculamos la solución $(\widetilde{x}, \widetilde{y})$ con tres cifras de precisión:

$$\begin{aligned} y &= \frac{2-\varepsilon^{-1}}{(1-\varepsilon^{-1})\,10^3} = \frac{2-10^4}{(1-10^4)\,10^3} = \frac{-9998}{-9999\times 10^3} \simeq 0{,}999\times 10^{-3}, \\ x &= \frac{1-10^3 y}{\varepsilon} \simeq \frac{1-10^3\times\left(0{,}999\times 10^{-3}\right)}{10^{-4}} = \frac{1-0{,}999}{10^{-4}} = \frac{10^{-3}}{10^{-4}} = 10. \end{aligned}$$

Apliquemos ahora el pivoting parcial. Tenemos

$$\left\{ \begin{array}{c} x+1000y = 2 \\ \varepsilon x + 1000y = 1, \end{array} \right. \Longleftrightarrow \left\{ \begin{array}{c} x+1000y = 2, \\ 1000\,(1-\varepsilon)\,y = 1-2\varepsilon, \end{array} \right.$$

de donde $y = \frac{1-2\varepsilon}{(1-\varepsilon)10^3}$, con lo que $x = 2-1000y = 2-10^3\left(\frac{1-2\varepsilon}{(1-\varepsilon)10^3}\right) = 2-\frac{1-2\varepsilon}{1-\varepsilon}$.

Para $\varepsilon = 10^{-4}$, tenemos

$$\begin{aligned} y &= \frac{1-2\varepsilon}{10^3\,(1-\varepsilon)} = \frac{1-2\times 10^{-4}}{10^3\,(1-10^{-4})} = \frac{0{,}9998}{999{,}9} \simeq 10^{-3}, \\ x &= 2-10^3 y = 2-10^3\times 10^{-3} \simeq 2. \end{aligned}$$

Luego $x \simeq 2, \quad y \simeq 10^{-3}$.

La solución exacta es

$$\begin{aligned} x &= \frac{1}{1-\varepsilon} = \frac{1}{1-10^{-4}} = \frac{1}{0{,}9999} = 1{,}000111 \simeq 1{,}001, \\ y &= \frac{2\varepsilon-1}{(\varepsilon-1)\,10^3} = \frac{0{,}0002-1}{(0{,}0001-1)\,10^3} = \frac{0{,}9998}{999{,}9} \simeq 0{,}9999\times 10^{-3}. \end{aligned}$$

En resumen, la solución del sistema de ecuaciones lineales con $\varepsilon = 10^{-4}$ se indica a continuación

$$\begin{aligned} \text{Sin pivoting} \;&:\quad x \simeq 10, \quad y \simeq 0{,}999\times 10^{-3}, \\ \text{Con pivoting} \;&:\quad x \simeq 2, \quad y \simeq 10^{-3} \\ \text{Exacta} \;&:\quad x = 1{,}0001, \quad y \simeq 0{,}9999\times 10^{-3}. \end{aligned}$$

Vemos que la solución con pivoting parcial es más próxima de la solución exacta. Este fenómeno se debe al condicionamiento de la matriz A del sistema, esto es, si $A = \begin{bmatrix} \varepsilon & 1000 \\ 1 & 1000 \end{bmatrix}$, tenemos $\|A\|_1 = 10001$, $A^{-1} = \begin{bmatrix} \frac{1}{\varepsilon-1} & -\frac{1}{\varepsilon-1} \\ -\frac{10^{-3}}{\varepsilon-1} & \frac{\varepsilon\times 10^{-3}}{\varepsilon-1} \end{bmatrix}$, y $\left\|A^{-1}\right\|_1 = \frac{1}{1-\varepsilon}$. El condicionamiento de la matriz A está definido como

$$cond\,(A) = \|A\|_1 \left\|A^{-1}\right\|_1 = \frac{10001}{1-\varepsilon}$$

luego $cond\,(A) \underset{\varepsilon\longrightarrow 0}{\longrightarrow} 10001, \quad cond\,(A) \underset{\varepsilon\longrightarrow 1}{\longrightarrow} -\infty$.

El número de condicionamiento depende de la norma de matriz elegida. Este número de condicionamiento es bastante grande lo que nos dice que la matriz A está mal condicionada. En esta clase de problemas es preciso resolver es sistema de ecuaciones lineales sea con el empleo del pivoting parcial o bien el pivoting total que mejora la precisión de la solución. Es importante también trabajar con más precisión: doble precisión y doble precisión extendida.

6.12. Ejercicios

1. Halle la solución de cada sistema de ecuaciones triangular superior.

a) $\begin{cases} 3x-2y+z=0 \\ 2y+5z=-1 \\ -4z=8. \end{cases}$ **b)** $\begin{cases} 10x+3y-5z=4 \\ 8y+15z=-1 \\ 5z=-3. \end{cases}$ **c)** $\begin{cases} -x+2y-3z-2w=2 \\ -y+z=-3 \\ z+8w=0 \\ -4w=3. \end{cases}$

d) $\begin{cases} x_1+x_2+x_3+x_4=0 \\ -2x_2+x_3-x_4=1 \\ \frac{1}{3}x_3+x_4=-1 \\ \frac{1}{5}x_4=1. \end{cases}$ **e)** $\begin{cases} x_1+2x_2+2x_3-x_4=0 \\ -2x_2+5x_3-x_4=1 \\ \frac{2}{7}x_3+\frac{12}{5}x_4=-1 \\ \frac{3}{5}x_4=1. \end{cases}$

2. Halle la solución de cada sistema de ecuaciones triangular inferior.

a) $\begin{cases} \frac{1}{5}x=-3 \\ x+3y=0 \\ 2x+y-6z=0. \end{cases}$ **b)** $\begin{cases} 3x=0 \\ 5x+4y=-10 \\ -2x-3y-4z=2. \end{cases}$ **c)** $\begin{cases} 2{,}3x=4{,}8 \\ 1{,}5x-2y=1{,}5 \\ -0{,}5x+3{,}2y-0{,}8z=2{,}3 \end{cases}$

d) $\begin{cases} 8x=72 \\ 3x+9y=93 \\ -5x+2y-z=15. \end{cases}$ **e)** $\begin{cases} 10x=20{,}2 \\ -2x-y=30{,}5 \\ 4x+7y+2z=90{,}3 \\ 5x-2{,}3y-z+4{,}5w=185. \end{cases}$

f) $\begin{cases} \frac{2}{3}x=-1 \\ 1{,}5x-\frac{1}{4}y=0 \\ 3{,}2x+4{,}8y+z=0 \\ x+1{,}5y+3z-4{,}5w=1. \end{cases}$ **f)** $\begin{cases} 0{,}25x=-1 \\ 1{,}5x-0{,}4y=0 \\ 8{,}2x-0{,}8y+2{,}2z=0 \\ 2{,}5x+3{,}5y-3{,}2z-4{,}8w=1. \end{cases}$

3. Con cada matriz triangular superior invertible A que se propone, aplique el método de eliminación gaussiana para calcular A^{-1}

a) $A=\begin{bmatrix} 3 & -2 & 1 \\ 0 & 2 & 5 \\ 0 & 0 & -4 \end{bmatrix}$. **b)** $A=\begin{bmatrix} 10 & 3 & -5 \\ 0 & 8 & 15 \\ 0 & 0 & 5 \end{bmatrix}$. **c)** $A=\begin{bmatrix} -1 & 2 & -3 & -2 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & -4 \end{bmatrix}$.

d) $A=\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & -2 & 1 & -1 \\ 0 & 0 & \frac{1}{3} & 1 \\ 0 & 0 & 0 & \frac{1}{5} \end{bmatrix}$. **e)** $A=\begin{bmatrix} -1 & 1 & 0 & -1 \\ 0 & -2 & \frac{1}{2} & -1 \\ 0 & 0 & -\frac{5}{3} & 3 \\ 0 & 0 & 0 & -5 \end{bmatrix}$.

4. Aplique el método de eliminación gaussiana para calcular A^{-1} con cada matriz triangular inferior invertible que en cada item se da.

a) $A=\begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 1 & 3 & 0 \\ 2 & 1 & -6 \end{bmatrix}$. **b)** $A=\begin{bmatrix} 2 & 0 & 0 \\ 5 & 4 & 0 \\ -2 & -3 & -4 \end{bmatrix}$. **c)** $A=\begin{bmatrix} 2{,}3 & 0 & 0 \\ 1{,}5 & -2 & 0 \\ -0{,}5 & 3{,}2 & -0{,}8 \end{bmatrix}$.

d) $A=\begin{bmatrix} 10 & 0 & 0 & 0 \\ -2 & -1 & 0 & 0 \\ 4 & 7 & 2 & 0 \\ 5 & -2{,}3 & -1 & 4{,}5 \end{bmatrix}$. **e)** $A=\begin{bmatrix} -10 & 0 & 0 & 0 \\ -2 & -5 & 0 & 0 \\ 4 & 0 & -2 & 0 \\ 0 & -2{,}3 & 0 & -1 \end{bmatrix}$.

5. Con cada matriz A que se da calcule $\det(A)$. Para el efecto, aplique el método de eliminación gaussiana y transforme la matriz A en una triangular superior.

a) $A = \begin{bmatrix} 1 & 0 & -2 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$. **b)** $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 3 & 0 \\ 2 & 1 & 5 \end{bmatrix}$. **c)** $A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 3 & 1 & 0 \\ -1 & 1 & 0 & -1 \\ 2 & -2 & 2 & 0 \end{bmatrix}$.

d) $A = \begin{bmatrix} -2 & 1 & 0 & 1 \\ 0 & -1 & 1 & 2 \\ -3 & 1 & 2 & 4 \\ -1 & -7 & 1 & 4 \end{bmatrix}$. **e)** $A = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 3 & 2 & -3 & -1 \\ -4 & -4 & 2 & 0 \\ 2 & 0 & -4 & 1 \end{bmatrix}$. **f)** $A = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 2 & 5 & 0 & 0 \\ 3 & -3 & 7 & 1 \\ 4 & 1 & -2 & 1 \end{bmatrix}$

6. En cada item se propone un conjunto S. Calcule $\widehat{x} \in S$ tal que $\|\widehat{x}\|^2 = \min\limits_{\overrightarrow{x} \in S} \|\overrightarrow{x}\|^2$.

   a) $S = \left\{(x,y) \in \mathbb{R}^2 \mid 2x + 3y = 0\right\}$. **b)** $S = \left\{(x,y) \in \mathbb{R}^2 \mid 4x - \frac{1}{4}y = -1\right\}$.

   c) $S = \left\{(x,y,z) \in \mathbb{R}^3 \mid x + 2y - z = 4\right\}$. **d)** $S = \left\{(x,y,z) \in \mathbb{R}^3 \mid 2x + 3y - z = 5\right\}$.

   e) $S = \left\{(x,y,z,w) \in \mathbb{R}^4 \mid 2x + 3y - 2z + w = -2\right\}$.

   f) $S = \left\{(x,y,z,w) \in \mathbb{R}^4 \mid x - 2y + z + 2w = 10\right\}$.

7. En cada item se da un sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$. Aplicar el método de Crout para factorar A en la forma $A = LU$. Verifique el resultado. Resuelva el sistema de ecuaciones lineales equivalente $\begin{cases} L\overrightarrow{y} = \overrightarrow{b} \\ U\overrightarrow{x} = \overrightarrow{y} \end{cases}$. Verifique que $\overrightarrow{x}$ es solución del sistema de ecuaciones propuesto. Contabilice el número de operaciones elementales que realiza.

   a) $\begin{cases} x \quad - 2z = 8 \\ 2x + y - z = 15 \\ 3x - y - 10z = 27, \end{cases}$ $\overrightarrow{x}^T = (4,5,-2)$. **b)** $\begin{cases} 2x + 2y - 2z = 10 \\ 3x + 4y - 7z = 2 \\ 4x + 4y + 4z = 60, \end{cases}$ $\overrightarrow{x}^T = (3,7,5)$.

   c) $\begin{cases} 3x - 3y + 6z = -18 \\ x + y + 8z = -2 \\ -x \quad - 4z = 3, \end{cases}$ $\overrightarrow{x}^T = (1,5,-1)$.

   d) $\begin{cases} x + 2y + 3z + 4w = 34 \\ x + 4y + 7z + 10w = 62 \\ x + 4y + 10z + 16w = 74 \\ x + 4y + 10z + 20w = 74, \end{cases}$ $\overrightarrow{x}^T = (10,6,4,0)$.

   d) $\begin{cases} 4x + \quad + 20w = 48 \\ y - z + 3w = 5 \\ -2y + 5z - 12w = -22 \\ 3x + 3y - 2z + 21w = 44, \end{cases}$ $\overrightarrow{x}^T = (-3,-2,2,3)$.

   f) $\begin{cases} -2x_1 + 2x_2 + 8x_3 - 6x_4 \quad = 10 \\ -x_1 + 2x_2 \quad + 5x_4 = 1 \\ 3x_2 + 16x_3 + 28x_4 = 16 \\ 5x_2 + 26x_3 + 44x_4 = 26 \\ 2x_1 - 2x_2 + 32x_3 + 45x_4 + 2x_5 = 33, \end{cases}$ $\overrightarrow{x}^T = (-1,0,1,0,-1)$.

   g) $\begin{cases} 2x_1 + 4x_2 - 6x_3 + 8x_4 + 16x_5 = 26 \\ 2x_1 + 5x_2 - 8x_3 + 8x_4 + 15x_5 = 24 \\ 2x_1 + 2x_2 - x_3 + 8x_4 + 20x_5 = 33 \\ 2x_1 + 7x_2 - 10x_3 + 11x_4 + 20x_5 = 32 \\ 2x_1 + 5x_2 + 5x_3 + 10x_4 + 24x_5 = 48, \end{cases}$ $\overrightarrow{x}^T = (2,1,1,1,1)$.

8. En cada literal se da una matriz A. Pruebe que A no se factora en la forma LU, con L una matriz triangular inferior, U matriz triangular superior tal que $u_{ii} = 1, \quad i = 1, \ldots, n$.

   a) $A = \begin{bmatrix} 0 & 2 \\ 3 & 0 \end{bmatrix}$. **b)** $A = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 2 & 3 & 0 \end{bmatrix}$. **c)** $A = \begin{bmatrix} 1 & 2 & 1 \\ -3 & -4 & -5 \\ 2 & -1 & 7 \end{bmatrix}$.

d) $A=\begin{bmatrix} 2 & 1 & 1 \\ -1 & -1 & 1 \\ 4 & 3 & -1 \end{bmatrix}$. **e)** $A=\begin{bmatrix} 1 & 2 & 1 & 1 \\ 3 & 2 & -3 & -1 \\ -4 & -4 & 2 & 0 \\ 2 & 0 & -4 & 1 \end{bmatrix}$

9. Aplicar el método de Choleski para factorar la matriz A del sistema de ecuaciones $A\overrightarrow{x}=\overrightarrow{b}$ en la forma $A=L^TL$. Resuelva el sistema de ecuaciones equivalente $\begin{cases} L^T\overrightarrow{y}=\overrightarrow{b} \\ L\overrightarrow{x}=\overrightarrow{y} \end{cases}$. Compare con el vector $\overrightarrow{x}$ que se propone. Contabilice el número de operaciones elementales que realiza.

a) $\begin{cases} x+2y+3z=10 \\ 2x+5y+5z=19 \\ 3x+5y+11z=33, \end{cases}$ $\overrightarrow{x}^T=(2,1,2)$. **b)** $\begin{cases} 4x+6y+8z=-8 \\ 6x+10y+12z=-10 \\ 8x+12y+80z=-336, \end{cases}$ $\overrightarrow{x}^T=(5,2,-5)$

c) $\begin{cases} x+y+z+w=5 \\ x+5y+5z+5w=9 \\ x+5y+14z+14w=9 \\ x+5y+14z+30w=25, \end{cases}$ $\overrightarrow{x}^T=(4,1,-1,1)$.

d) $\begin{cases} 16x_1+\quad\quad\quad +12x_4=-100 \\ \quad\quad x_2-2x_3+3x_4=15 \\ \quad -2x_2+13x_3-3x_4=-15 \\ 12x_1+3x_2-3x_3+20x_4=-20, \end{cases}$ $\overrightarrow{x}^T=(-10,0,0,5)$.

e) $\begin{cases} 4x_1+2x_2\quad\quad\quad -4x_5=52 \\ 2x_1+2x_2+3x_3+5x_4+5x_5=52 \\ \quad\quad 3x_2+25x_3+39x_4+53x_5=135 \\ \quad\quad 5x_2+39x_3+65x_4+85x_5=217 \\ -4x_1+5x_2+53x_3+85x_4+122x_5=231, \end{cases}$ $\overrightarrow{x}^T=(12,7,3,1,0)$.

f) $\begin{cases} 4x_1+4x_2+2x_3+4x_4+4x_5+2x_6=0 \\ 4x_1+5x_2\quad\quad +7x_4+5x_5+3x_6=-3 \\ x_1\quad\quad +14x_3-4x_4\quad\quad -x_6=5 \\ 4x_1+7x_2-4x_3+22x_4+11x_5+5x_6=-14 \\ 2x_1+5x_2\quad\quad +11x_4+13x_5+5x_6=-1 \\ 2x_1+3x_2-x_3+5x_4+5x_5+28x_6=-1, \end{cases}$ $\overrightarrow{x}^T=(1,-1,0,-1,1,0)$.

10. En cada item se da una matriz A. Determine A^TA. Aplique el método de eliminación gaussiana para determinar los rangos $R(A)$ y $R\left(A^TA\right)$ de las matrices A y A^TA. Compruebe que $R(A)=R\left(A^TA\right)$ es el que se indica.

a) $A=\begin{bmatrix} 4 & 0 & 1 \\ 1 & 2 & -3 \\ 0 & -1 & 2 \end{bmatrix}$, $R(A)=3$. **b)** $A=\begin{bmatrix} 1 & 1 & 5 \\ -1 & 3 & 3 \\ 0 & -1 & -2 \end{bmatrix}$, $R(A)=2$.

c) $A=\begin{bmatrix} 1 & 2 & 3 & 1 \\ 2 & 1 & 1 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix}$, $R(A)=3$. **d)** $A=\begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{bmatrix}$, $R(A)=2$.

e) $A=\begin{bmatrix} 2 & 1 & -2 \\ 3 & 2 & 2 \\ 5 & 4 & 3 \end{bmatrix}$, $R(A)=3$. **f)** $A=\begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & -1 \\ 3 & -2 & -1 \end{bmatrix}$, $R(A)=3$.

g) $A=\begin{bmatrix} 2 & 1 & 1 \\ -1 & -1 & 1 \\ 4 & 3 & -1 \end{bmatrix}$, $R(A)=2$. **h)** $A=\begin{bmatrix} 1 & 2 & 1 & 1 \\ 3 & 2 & -3 & -1 \\ -4 & -4 & 2 & 0 \\ 2 & 0 & -4 & 1 \end{bmatrix}$, $R(A)=3$.

i) $A=\begin{bmatrix} -2 & 1 & 0 & 1 \\ 0 & -1 & 1 & 2 \\ -3 & 1 & 2 & 4 \\ -1 & -7 & 1 & 4 \end{bmatrix}$, $R(A)=4$. **j)** $A=\begin{bmatrix} 6 & 0 & 0 \\ 2 & 5 & 0 \\ 3 & -3 & 7 \\ 4 & 1 & -2 \end{bmatrix}$, $R(A)=3$.

11. En cada item se propone un sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$. Estudie a la matriz A del sistema para determinar si es estrictamente diagonalmente dominante, simétrica, definida positiva, monótona, etc. Aplique el método de eliminación gaussiana, el de factorización de Crout LU y siempre que sea posible el de Choleski L^TL y halle la solución del sistema. Contabilice el número de operaciones elementales que realiza con cada método.

a) $\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 3 \\ 2 \end{bmatrix}, \quad \overrightarrow{x}^T = (5, 8, 8, 5)\,.$

b) $\begin{bmatrix} 4 & 1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & 1 & 4 & -1 \\ 0 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 11 \\ 6 \\ 17 \\ 12 \end{bmatrix}, \quad \overrightarrow{x}^T = (2, 3, 4, 2)\,.$

c) $\begin{bmatrix} 5 & 1 & 1 & 0 & 0 \\ -1 & 5 & 1 & -1 & 0 \\ -1 & 1 & 6 & 1 & 1 \\ 0 & 0 & -1 & 6 & 1 \\ 0 & 0 & 0 & 1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} == \begin{bmatrix} 16 \\ 14 \\ 22 \\ 10 \\ 8 \end{bmatrix}, \quad \overrightarrow{x}^T = (2, 3, 3, 2, 1)$

d) $\begin{bmatrix} 3 & -1 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 11 \\ 4 \\ 1 \\ 4 \\ 11 \end{bmatrix}, \quad \overrightarrow{x}^T = (5, 4, 3, 4, 5)\,.$

e) $\begin{bmatrix} 5 & 2 & 1 & 0 & 0 \\ 2 & 5 & 1 & 0 & 0 \\ 1 & 1 & 5 & 2 & 1 \\ 0 & 0 & 2 & 5 & 1 \\ 0 & 0 & 0 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -12 \\ -9 \\ 1 \\ 7 \\ 11 \end{bmatrix}, \quad \overrightarrow{x}^T = (-2, -1, 0, 1, 2)\,.$

f) $\begin{bmatrix} 4 & -1 & -\frac{1}{2} & 0 & 0 & 0 \\ -1 & 5 & -1 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & -1 & 6 & -1 & -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & -1 & 7 & -1 & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & -1 & 8 & -1 \\ 0 & 0 & 0 & -\frac{1}{2} & -1 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 29 \\ 21 \\ 13 \\ 19 \\ 45 \\ 79 \end{bmatrix}, \quad \overrightarrow{x}^T = (10, 8, 6, 6, 8, 10)\,.$

12. Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ que satisface las dos condiciones siguientes: $a_{ij} = 0$ si $|i - j| > 2$ para $i, j = 1, \ldots, n$ y que $a_{ii} > |a_{i\,i-2}| + |a_{i\,i-1}| + |a_{i\,i+1}| + |a_{i\,i+2}|\,, \quad i = 1, \ldots, n,\ \overrightarrow{b} \in \mathbb{R}^n$.

a) Demuestre que el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ tiene una única solución.

b) Defina una matriz $\widetilde{A}$ de $n \times 5$ de modo que contenga la información relevante de la matriz A.

c) Aplique el método de eliminación gaussiana y la definición de $\widetilde{A}$ para elaborar un algoritmo para hallar la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

d) Aplique el método de factorización LU y $\widetilde{A}$ para escribir un algoritmo para hallar la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$. Defina matrices $\widetilde{L}$ y $\widetilde{U}$ apropiadas de modo que se reduzca significativamente el número de elementos a almacenar.

e) Suponga adicionalmente que A es simétrica, ¿es A definida positiva? En caso de ser, aplique la factorización de Choleski L^TL y la definición de $\widetilde{A}$ para elaborar un algoritmo que permita calcular la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

f) Considere el sistema de ecuaciones lineales siguiente:

$$\begin{bmatrix} 4 & -1 & -1 & 0 & 0 \\ -1 & 5 & -1 & -1 & 0 \\ -1 & -1 & 5 & -1 & -1 \\ 0 & -1 & -1 & 4 & -1 \\ 0 & 0 & -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -5 \\ 10 \\ 25 \\ 4 \\ -4 \end{bmatrix}$$

Verifique las hipótesis de la matriz A. Definida $\widetilde{A}$ y aplique sus algoritmos para hallar la solución de dicho sistema y compare con $\overrightarrow{x}^T = (2, 5, 8, 5, 3)$.

13. Considere el sistema de ecuaciones lineales siguiente:

$$\begin{bmatrix} 4 & 2 & -2 & 0 & 0 \\ 2 & 10 & 2 & -3 & 0 \\ -2 & 2 & 18 & 3 & -4 \\ 0 & -3 & 3 & 18 & 3 \\ 0 & 0 & -4 & 3 & 18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -6 \\ 21 \\ 96 \\ 66 \\ 7 \end{bmatrix}.$$

a) Demuestre que la matriz A de este sistema es simétrica, definida positiva y estrictamente diagonalmente dominante.

b) Aplique los métodos de eliminación gaussiana, factorización LU de Crout y de Choleski para hallar la solución de tal sistema. Contabilice con cada método el número de operaciones elementales. Compare la solución con $\overrightarrow{x}^T = (0, 2, 5, 3, 1)$.

14. Considere el conjunto S que se define. Halle $\widehat{x} \in S$ tal que $\|\widehat{x}\|^2 = \underset{\overrightarrow{x} \in S}{\text{mín}} \|\overrightarrow{x}\|^2$.

a) $S = \left\{ \overrightarrow{x} = (x, y, z) \in \mathbb{R}^3 \mid \begin{cases} x + y - z = 1 \\ 2x - y + z = 2. \end{cases} \right\}$.

b) $S = \left\{ \overrightarrow{x} = (x, y, z) \in \mathbb{R}^3 \mid \begin{cases} 2x - y + z = 0 \\ x \quad - 2z = 1. \end{cases} \right\}$

c) $S = \left\{ \overrightarrow{x} = (x, y, z) \in \mathbb{R}^3 \mid \begin{cases} x \quad - z = 2 \\ 2y + 3z = -1. \end{cases} \right\}$.

d) $S = \left\{ \overrightarrow{x}^T = (x, y, z, w) \in \mathbb{R}^4 \mid A\overrightarrow{x} = \overrightarrow{b} \right\}$, donde $A = \begin{bmatrix} 2 & 1 & 0 & 1 \\ -1 & 0 & 1 & -1 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

e) $S = \left\{ \overrightarrow{x}^T = (x, y, z, w) \in \mathbb{R}^4 \mid A\overrightarrow{x} = \overrightarrow{b} \right\}$, con $A = \begin{bmatrix} -1 & 1 & -1 & 1 \\ 2 & 0 & 1 & 2 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

6.13. Lecturas complementarias y bibliografía

1. Owe Axelsson, Iterative Solution Methods, Editorial Cambridge University Press, Cambridge, 1996.
2. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
3. Richard H. Bartels, John C. Beatty, Brian A. Barsky, An Introduction to Splines for use in Computer Graphics and Geometric Medeling, Editorial Morgan Kaufmann Publishers, Inc., San Mateo, California, 1987.
4. Jérõme Bastien, Jean-Noël Martin, Introduction à L'Analyse Numérique, Editorial Dunod, París, 2003.
5. Abraham Berman, Robert J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia,1994.
6. Rajendra Bhatia, Matrix Analysis, Editorial Springer-Verlag, New York, 1997.

7. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.

8. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.

9. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.

10. P. G. Ciarlet, Introduction á L'Analyse Numérique Matricielle et á l'Optimisation, Editorial Masson, París, 1990.

11. Elaine Cohen, Richard F. Riesenfeld, Gershon Elber, Geometric Modeling with Splines, Editorial A. K. Peters, Natick, Massachusetts, 2001.

12. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.

13. James W. Demmel, Applied Numerical Linear Algebra, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.

14. J. E. Dennis, Jr., Robert B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1996.

15. V. N. Faddeva, Métodos de Cálculo de Algebra Lineal, Editorial Paraninfo, Madrid, 1967.

16. Francis G. Florey, Fundamentos de Algebra Lineal y Aplicaciones, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1980.

17. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.

18. Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence, Algebra Lineal, Editorial Publicaciones Cultural, S. A., México, 1982.

19. Noel Gastinel, Análisis Numérico Lineal, Editorial Reverté, S. A., Barcelona, 1975.

20. M. K. Gavurin, Conferencias sobre los Métodos de Cálculo, Editorial Mir, Moscú, 1973.

21. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

22. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.

23. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

24. I. N. Herstein, J. Winter, Algebra Lineal y Teoría de Matrices, Grupo Editorial Iberoamericana, México, 1989.

25. Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Editorial Society for Industrial and Applied Mathematics, Philadelphia, 1996.

26. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.

27. Franz E. Hohn, Algebra de Matrices, Editorial Trillas, México, 1979.

28. Roger A. Horn, Charles R. Johnson, Matrix Analysis, Editorial Cambridge University Press, Cambridge, 1999.

29. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.

30. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

31. Roland E. Larson, Bruce H. Edwards, Introducción al Algebra Lineal, Editorial Limusa, Noriega Editores, México, 1995.

32. P. Lascaux, R. Théodor, Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur, Tome 1, Editorial Masson, París, 1986.

33. P. Lascaux, R. Théodor, Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur, Tome 2, Editorial Masson, París, 1987.

34. Charles L. Lawson, Richard J. Hanson, Solving Least Squares Problems, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1995.

35. L. Lebart, A. Morineau, J.-P. Fénelon, Tratamiento Estadístico de Datos, Editorial Marcombo Boixareu Editores, Barcelona, 1985.

36. Peter Linz, Theoretical Numerical Analysis, Editorial Dover Publications, Inc., New York, 2001.

37. Rodolfo Luthe, Antonio Olivera, Fernando Schutz, Métodos Numéricos, Editorial Limusa, México, 1986.

38. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

39. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

40. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

41. Ben Noble, James W. Daniel, Algebra Lineal Aplicada, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1989.

42. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

43. Fazlollah Reza, Los Espacios Lineales en la Ingeniería, Editorial Reverté, S. A., Barcelona, 1977.

44. A. A. Samarski, Introducción a los Métodos Numéricos, Editorial Mir, Moscú, 1986.

45. Michelle Schatzman, Analyse Numérique, Inter Editions, París, 1991.

46. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

47. M. Sibony, J. Cl. Mardon, Analyse Numérique I, Systèmes Linéaires et non Linéaires, Editorial Hermann, París, 1984.

48. Helmuth Späth, One Dimensional Spline Interpolation algorithms, Editorial A. K. Peters, Wellesley, Massachusetts, 1995.

49. G. W. Stewart, Matrix Algotithms, Volume I: Basic Decomposition, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

50. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

51. Gilbert Strang, Algebra Lineal y sus Aplicaciones, Editorial Fondo Educativo Interamericano, México, 1982.

52. V. Voïévodine, Principes Numériques D'Algèbre Linéaire, Editions Mir, Moscú, 1976.

53. E. A. Volkov, Métodos Numéricos, Editorial Mir, Moscú, 1990.

54. David S. Watkins, Fundamentals of Matrix Computations, Editorial John Wiley & Sons, New York, 1991

Capítulo 7

Métodos iterativos

Resumen

En este capítulo se introducen dos amplios temas de sistemas de ecuaciones lineales y no lineales. Se comienza con los sistemas de ecuaciones no lineales. Nos limitados a la aplicación del método de Newton. Este a su vez requiere del conocimiento de la diferencial de Fréchet y sus propiedades, de las aplicaciones contractivas y por supuesto del teorema de Banach del punto fijo. A continuación volvemos a tratar los sistemas de ecuaciones lineales, pero esta vez con la mira de los métodos iterativos más conocidos como son el método de Jacobi, de Gauss-Seidel y SOR.

7.1. Diferencial de Fréchet. Propiedades

En esta sección definiremos la diferencial de Fréchet, sus propiedades más importantes y daremos algunas aplicaciones.

Definición 1 *Sean V, W dos espacios normados con normas $\|\cdot\|_V$, $\|\cdot\|_W$, Ω un abierto no vacío de V, F una función de Ω en W.*
i. Se dice que F es diferenciable en $a \in \Omega$ si y solo si existe $T_a \in L(V,W)$ tal que

$$F(a+h) - F(a) = T_a(h) + R(a,h), \quad con \quad \lim_{h\to 0} \frac{\|R(a,h)\|_W}{\|h\|_V} = 0.$$

En tal caso se dice que T_a es la diferencial de F en a que se le denota $Df(a)$.
ii. Se dice que F es diferenciable en Ω si F es diferenciable en cada punto de Ω.

La aplicación $T_a \in L(V,W)$ se le denomina aplicación diferencial de F en a, o diferencial de Fréchet de F en a. Note que si $\alpha, \beta \in \mathbb{R}$, $h_1, h_2 \in V$ se tiene $T_a(\alpha h_1 + \beta h_2) = \alpha T_a(h_1) + \beta T_a(h_2)$, es decir la linealidad de T_a. Por otro lado, T_a es continua, esto es, existe $M_a > 0$ tal que $\|T_a(x)\|_W \leq M_a \|x\|_V \quad \forall x \in V$, donde

$$M_a = \|T_a\|_{L(V,W)} = \sup_{\|x\|_V \leq 1} \|T_a(x)\|_W = \|Df(a)\|_{L(V,W)}.$$

Si los espacios vectoriales V, W son de dimensiones finitas m y n respectivamente, $\beta_V = \{v_1, \ldots, v_m\}$, $\beta_W = \{w_1, \ldots, w_n\}$ son las bases canónicas de V y W, la matriz asociada a la aplicación lineal T_a relativa a las bases β_V y β_W se le nota $[T_a]_{\beta_V}^{\beta_W}$. Esta matriz, por abuso de lenguaje, se le nota $DF(a)$ de modo que $T_a(x) = DF(a)\,x \quad \forall x \in V$.

Particularmente, si $V = \mathbb{R}^m$, $W = \mathbb{R}^n$, las bases canónicas de $\mathbb{R}^m$ y $\mathbb{R}^n$ las designamos con $\beta_m = \{\overrightarrow{e}_1, \ldots, \overrightarrow{e}_m\}$, $\beta_n = \left\{\overrightarrow{f}_1, \ldots, \overrightarrow{f}_n\right\}$ y las normas son la euclídea en cada espacio, el espacio $L(\mathbb{R}^m, \mathbb{R}^n)$ es isomorfo al espacio de matrices $M_{n\times m}[\mathbb{R}]$, es decir que cada elemento de $L(\mathbb{R}^m, \mathbb{R}^n)$ se identifica con una matriz apropiada de $M_{n\times m}[\mathbb{R}]$.

Ejemplos

1. Sea Ω un intervalo abierto de $\mathbb{R}$, F una función de Ω en $\mathbb{R}^n$. Supongamos $F=(f_1,\ldots,f_n)$ donde las funciones $f_i \;\; i=1,\ldots,n$ son funciones reales definidas en Ω y derivables en $a\in\Omega$. Se supone que $\mathbb{R}$ está provisto de la norma $|\cdot|$ y $\mathbb{R}^n$ de la norma euclídea $\|\cdot\|$. La existencia de $f_i'(a) \quad i=1,\ldots,n$ implica
$$f_i(a+h)-f_i(a)=f_i'(a)\,h+R_i(a,h)$$
con $R_i(a,h)\underset{h\to 0}{\longrightarrow}0 \quad i=1,\ldots,n$, lo que conduce a definir $T_a\in L(\mathbb{R},\mathbb{R}^n)$ como $T_a(h)=(f_1'(a)\,h,\ldots,f_n'(a)\,h) \quad \forall h\in\mathbb{R}$. Claramente T_a es lineal continua. Además,
$$F(a+h)-F(a)=T_a(h)+R(a,h)$$
donde $R(a,h)=(R_1(a,h),\ldots,R_n(a,h))$, y $\lim\limits_{h\to 0}\dfrac{\|R(a,h)\|}{|h|}=0$. Resulta que la representación matricial de T_a es $DF(a)=(f_1'(a),\ldots,f_n'(a))$. Note que
$$F(a+h)-F(a)=T_a(h)+R(a,h)\Longleftrightarrow f_i(a+h)-f_i(a)=f_i'(a)\,h+R_i(a,h) \quad i=1,\ldots,n.$$

2. Sean Ω un abierto de $\mathbb{R}^n$, F una función de Ω en $\mathbb{R}$, esto es, F un campo escalar definido en Ω. Sea $\overrightarrow{a}\in\mathbb{R}$. Recordemos que $\dfrac{\partial F}{\partial x_i}(\overrightarrow{a})$ está definido como sigue
$$\frac{\partial F}{\partial x_i}(\overrightarrow{a})=\lim_{h\to 0}\frac{F(\overrightarrow{a}+h\overrightarrow{e}_i)-F(\overrightarrow{a})}{h},$$
siempre que el límite exista. Además, el gradiente de F en a se define como
$$\nabla F(\overrightarrow{a})=\left(\frac{\partial F}{\partial x_1}(\overrightarrow{a}),\ldots,\frac{\partial F}{\partial x_n}(\overrightarrow{a})\right).$$
Se define $T_a\in L(\mathbb{R}^n,\mathbb{R})$ como sigue: $T_{\overrightarrow{a}}\left(\overrightarrow{h}\right)=\left\langle\nabla F(\overrightarrow{a}),\overrightarrow{h}\right\rangle \quad \forall\overrightarrow{h}\in\mathbb{R}^n$, donde $\langle\cdot,\cdot\rangle$ denota el producto escalar en $\mathbb{R}^n$. La linealidad y la continuidad de $T_{\overrightarrow{a}}$ se verifican inmediatamente. Se tiene
$$\left|T_{\overrightarrow{a}}\left(\overrightarrow{h}\right)\right|\leq\|\nabla F(\overrightarrow{a})\|\left\|\overrightarrow{h}\right\| \quad \forall\overrightarrow{h}\in\mathbb{R}^n.$$
La representación matricial de $T_{\overrightarrow{a}}$ respecto de la base canónica de $\mathbb{R}^n$ es $\nabla F(\overrightarrow{a})$. Resulta
$$F\left(\overrightarrow{a}+\overrightarrow{h}\right)-F(\overrightarrow{a})=\left\langle\nabla F(\overrightarrow{a}),\overrightarrow{h}\right\rangle+R\left(\overrightarrow{a},\overrightarrow{h}\right)$$
con $R\left(\overrightarrow{a},\overrightarrow{h}\right)\underset{\overrightarrow{h}\to 0}{\longrightarrow}0$. Luego, la diferencial de Fréchet está definida como
$$T_{\overrightarrow{a}}\left(\overrightarrow{h}\right)=\left\langle\nabla F(\overrightarrow{a}),\overrightarrow{h}\right\rangle \;\; \forall\overrightarrow{h}\in\mathbb{R}^n.$$

3. Sean Ω un abierto de $\mathbb{R}^m$, F una función de Ω en $\mathbb{R}^n$, es decir, F es un campo vectorial. Ponemos $F^T=(f_1,\ldots,f_n)$ al vector transpuesto de F donde $f_i \quad i=1,\ldots,n$ es un campo escalar que suponemos diferenciable en $\overrightarrow{a}\in\Omega$, esto es,
$$f_i\left(\overrightarrow{a}+\overrightarrow{h}\right)-f_i(\overrightarrow{a})=\left\langle\nabla f_i(\overrightarrow{a}),\overrightarrow{h}\right\rangle+R_i\left(\overrightarrow{a},\overrightarrow{h}\right)$$
con $\lim\limits_{\overrightarrow{h}\to 0}\dfrac{\left|R_i\left(\overrightarrow{a},\overrightarrow{h}\right)\right|}{\left\|\overrightarrow{h}\right\|_m}=0$. Se define $T_{\overrightarrow{a}}\left(\overrightarrow{h}\right)=\begin{bmatrix}\left\langle\nabla f_1(\overrightarrow{a}),\overrightarrow{h}\right\rangle\\ \vdots\\ \left\langle\nabla f_n(\overrightarrow{a}),\overrightarrow{h}\right\rangle\end{bmatrix} \quad \forall\overrightarrow{h}\in\mathbb{R}^n$. Entonces $T_{\overrightarrow{a}}\in L(\mathbb{R}^m,\mathbb{R}^n)$ pués es lineal continua. La matriz de $T_{\overrightarrow{a}}$ asociada a las bases canónicas de $\mathbb{R}^m$ y

$\mathbb{R}^n$ es la matriz jacobiana

$$DF\left(\overrightarrow{a}\right)=\left[\begin{array}{c}\dfrac{\partial f_1}{\partial x_1}\left(\overrightarrow{a}\right)\cdots\dfrac{\partial f_1}{\partial x_m}\left(\overrightarrow{a}\right)\\ \vdots\\ \dfrac{\partial f_n}{\partial x_1}\left(\overrightarrow{a}\right)\cdots\dfrac{\partial f_n}{\partial x_m}\left(\overrightarrow{a}\right)\end{array}\right]\in M_{n\times m}\left[\mathbb{R}\right].$$

Resulta

$$F\left(\overrightarrow{a},\overrightarrow{h}\right)-F\left(\overrightarrow{a}\right)=\left[\begin{array}{c}\left\langle\nabla f_1\left(\overrightarrow{a}\right),\overrightarrow{h}\right\rangle\\ \vdots\\ \left\langle\nabla f_n\left(\overrightarrow{a}\right),\overrightarrow{h}\right\rangle\end{array}\right]+\left[\begin{array}{c}R_1\left(\overrightarrow{a},\overrightarrow{h}\right)\\ \vdots\\ R_n\left(\overrightarrow{a},\overrightarrow{h}\right)\end{array}\right]=DF\left(\overrightarrow{a}\right)\overrightarrow{h}+R\left(\overrightarrow{a},\overrightarrow{h}\right)$$

donde $\lim\limits_{\overrightarrow{h}\to 0}\dfrac{\left\|R\left(\overrightarrow{a},\overrightarrow{h}\right)\right\|_n}{\left\|\overrightarrow{h}\right\|_m}=0$, que muestra que la diferencial de Fréchet es el operador $T_{\overrightarrow{a}}$ definido como

$$T_a\left(\overrightarrow{h}\right)=DF\left(\overrightarrow{a}\right)\overrightarrow{h}\quad\forall\overrightarrow{h}\in\mathbb{R}^n.$$

4. Sea $\Omega=\mathbb{R}^2$ provisto de la norma euclídea $\|\cdot\|$, f la función real definida en $\mathbb{R}^2$ como sigue: $f\left(x,y\right)=\left\{\begin{array}{l}\dfrac{x^2-y^2}{x^2+y^2},\text{ si }\left(x,y\right)\neq\left(0,0\right)\\ \quad 0,\text{ si }\left(x,y\right)=\left(0,0\right).\end{array}\right.$

Probemos que f no es diferenciable en $\overrightarrow{0}=\left(0,0\right)$. Para el efecto, supongamos lo contrario, es decir que existe una aplicación lineal $T_{\overrightarrow{0}}\in L\left(\mathbb{R}^2,\mathbb{R}\right)$ tal que

$$f\left(\overrightarrow{0}+\overrightarrow{h}\right)-f\left(\overrightarrow{0}\right)=T_{\overrightarrow{0}}\left(\overrightarrow{h}\right)+R\left(\overrightarrow{0},\overrightarrow{h}\right)\text{ con }\lim_{\overrightarrow{h}\to\overrightarrow{0}}\frac{\left|R\left(\overrightarrow{0},\overrightarrow{h}\right)\right|}{\left\|\overrightarrow{h}\right\|}=0.$$

Ponemos $\overrightarrow{h}=\left(a,b\right)\in\mathbb{R}^2$ con $\overrightarrow{h}\neq\overrightarrow{0}$. De la definición de f, resulta

$$f\left(\overrightarrow{0}+\overrightarrow{h}\right)-f\left(\overrightarrow{0}\right)=\frac{a^2-b^2}{a^2+b^2}=T_{\overrightarrow{0}}\left(\overrightarrow{h}\right)+R\left(\overrightarrow{0},\overrightarrow{h}\right).$$

La representación matricial de $T_{\overrightarrow{0}}$ respecto de la base canónica de $\mathbb{R}^2$ lo notamos $A=\left(\alpha,\beta\right)$, es decir que

$$T_{\overrightarrow{0}}\left(\overrightarrow{h}\right)=\left(\alpha,\beta\right)\left[\begin{array}{c}a\\ b\end{array}\right]=\alpha a+\beta b,$$

luego

$$\begin{aligned}\frac{a^2-b^2}{a^2+b^2}&=\alpha a+\beta b+R\left(\overrightarrow{0},\overrightarrow{h}\right)\Longrightarrow R\left(\overrightarrow{0},\overrightarrow{h}\right)=\frac{a^2-b^2}{a^2+b^2}-\alpha a-\beta b\\ \lim_{\overrightarrow{h}\to\overrightarrow{0}}\frac{R\left(\overrightarrow{0},\overrightarrow{h}\right)}{\left\|\overrightarrow{h}\right\|}&=\lim_{\overrightarrow{h}\to\overrightarrow{0}}\frac{1}{\sqrt{a^2+b^2}}\left(\frac{a^2-b^2}{a^2+b^2}-\alpha a-\beta b\right)\\ &=\lim_{\overrightarrow{h}\to\overrightarrow{0}}\left(\frac{a^2-b^2}{a^2+b^2}-\alpha\frac{a}{\sqrt{a^2+b^2}}-\beta\frac{b}{\sqrt{a^2+b^2}}\right).\end{aligned}$$

Pero $\lim\limits_{(a,b)\to(0,0)}\left(\alpha\dfrac{a}{\sqrt{a^2+b^2}}+\beta\dfrac{b}{\sqrt{a^2+b^2}}\right)=0$, mientras que $\lim\limits_{(a,b)\to(0,0)}\dfrac{a^2-b^2}{\left(a^2+b^2\right)^{\frac{3}{2}}}$ no existe, luego $\lim\limits_{\overrightarrow{h}\to\overrightarrow{0}}\dfrac{\left|R\left(\overrightarrow{0},\overrightarrow{h}\right)\right|}{\left\|\overrightarrow{h}\right\|}$ no existe, en contradicción con lo supuesto.

Propiedades de la diferencial.

Sean V, W espacios normados con normas $\|\cdot\|_V$ y $\|\cdot\|_W$, Ω un abierto de V y F una función de Ω en W.

Teorema 1 *Si F es diferenciable en $a \in \Omega$, entonces F es continua en a.*

Demostración. Debemos mostrar que $\lim\inf_{x\to a} F(x) = F(a)$. Por hipótesis F es Fréchet diferenciable en $a \in \Omega$, entonces existe $T_a \in L(V,W)$ tal que $F(a+h) - F(a) = T_a(h) + R(a,h)$ con $\lim_{h\to 0} \dfrac{\|R(a,h)\|_W}{\|h\|_V} = 0$ y de la existencia del límite, para $\varepsilon > 0$, existe $\delta > 0$ tal que $\|h\|_V < \delta \Longrightarrow \|R(a,h)\|_W < \varepsilon \|h\|_V$ luego

$$\begin{aligned} \|F(a+h) - F(a)\| &\leq \|T_a(h)\| + \|R(a,h)\|_W \leq \|T_a\|_{L(V,W)} \|h\|_V + \|R(a,h)\|_W \\ &\leq \left(\|T_a\|_{L(V,W)} + \varepsilon\right) \|h\|_V \quad \text{si} \quad \|h\|_V < \delta. \end{aligned}$$

Como $\left(\|T_a\|_{L(V,W)} + \varepsilon\right)\|h\| \underset{h\to 0}{\longrightarrow} 0$, se sigue que $\lim_{h\to 0} (F(a+h) - F(a)) = 0$ de donde $F(x) \underset{x\to a}{\longrightarrow} F(a)$. ■

Teorema 2 *Sean F, G funciones diferenciables en $a \in \Omega$. Entonces*

i) $F+G$ es diferenciable en a y $D(F+G)(a) = DF(a) + DG(a)$.

ii) λF es diferenciable en a y $D(\lambda F)(a) = \lambda DF(a) \quad \forall \lambda \in \mathbb{R}$.

Demostración. Se propone como ejercicio. ■

En el siguiente teorema se propone la conocida regla de la cadena.

Teorema 3 *Sean U, V, W espacios normados con normas $\|\cdot\|_U$, $\|\cdot\|_V$, $\|\cdot\|_W$, Ω un abierto de U, F una función de Ω en V diferenciable en $a \in \Omega$, G una función de V en W diferenciable en $F(a)$. Entonces $G \circ F$ es diferenciable en a y $D(G\circ F)(a) = DG(F(a))\, DF(a)$.*

Demostración. Por la diferenciabilidad de F en $a \in \Omega$, existe $S_a \in L(V,W)$ tal que $F(a+h) - F(a) = S_a(h) + R_F(a,h)$ con $\lim_{h\to 0} \dfrac{\|R_F(a,h)\|_V}{\|h\|_U} = 0$, de donde $\|R_F(a,h)\|_V \underset{\|h\|_V \to 0}{\longrightarrow} 0$.

Sea $y = F(a)$. Por la diferenciabilidad de G en y, existe $T_y \in L(V,W)$ tal que

$$G(y+k) - G(y) = T_y(k) + R_G(y,k)$$

con $\lim_{k\to 0} \dfrac{\|R_G(a,h)\|_W}{\|k\|_V} = 0$ o bien $\|R_G(y,k)\|_W \underset{\|k\|_V \to 0}{\longrightarrow} 0$.

Por lo tanto, de la definición de composición de funciones, se tiene

$$(G\circ F)(a+h) - (G\circ F)(a) = G(F(a+h)) - G(F(a))$$

y como $F(a+h) = F(a) + S_a(h) + R_F(a,h)$, se tiene $G(F(a+h)) = G(F(a) + S_a(h) + R_F(a,h))$.

Ponemos $k = S_a h + R_F(a,h)$. Entonces

$$\|k\|_V = \|S_a(h) + R_F(a,h)\|_V \leq \|S_a\|_{L(U,V)} \|h\|_V + \|R_F(a,h)\|_V \underset{\|h\|_V}{\longrightarrow} 0.$$

Resulta

$$\begin{aligned} G(F(a+h)) &= G(y+k) = G(y) + T_y(k) + R_G(y,k) \\ &= G(F(a)) + T_y(S_a(h) + R_F(a,h)) + R_G(y,k). \end{aligned}$$

Por la linealidad de T_y, se tiene

$$T_y(S_a(h) + R_F(a+h)) = T_y(S_a(h)) + T_y(R_F(a,h))$$

y de esta igualdad, se obtiene

$$(G\circ F)(a+h)=F(F(a))+T_y(S_a(h))+T_y(R_F(a,h))+R_G(y,k)$$

de donde

$$(G\circ F)(a+h)-(G\circ F)(a)=(T_y\circ T_a)(h)+T_y(R_F(a,h))+R_G(y,k).$$

Es claro que $T_y\circ T_a\in L(U,W)$. Probemos que $\lim\limits_{h\to 0}\inf\dfrac{\|T_y(R_F(a,h))+R_G(y,k)\|_W}{\|h\|_v}=0.$

Por la desigualdad triangular, se tiene

$$\begin{aligned}\|T_y(R_F(a,h))+R_G(y,k)\|_W &\leq \|T_y(R_F(a,h))\|_W+\|R_G(y,k)\|_W\\ &\leq \|T_y\|_{L(V,W)}\|R_F(a,h)\|_V+\|R_G(y,k)\|_W\end{aligned}$$

y como

$$\|R_G(y,k)\|_W=\frac{\|R_G(y,k)\|_W}{\|k\|_V}\ \|k\|_V\leq\frac{\|R_G(y,k)\|_W}{\|k\|_V}\left(\|S_a\|_{L(U,V)}\|h\|_V+\|R_F(a,h)\|_V\right)\underset{\|h\|_V\to 0}{\longrightarrow}0.$$

Note que $\dfrac{\|R_G(y,k)\|_W}{\|k\|_V}\underset{k\to 0}{\longrightarrow}0$, consecuentemente

$$\|T_y(R_F(a,h))+R_G(y,k)\|_W\leq\|T_y\|_{L(V,W)}\|R_F(a,h)\|_V+\|R_G(y,k)\|_W\underset{\|h\|_V\to 0}{\longrightarrow}0$$

y de la existencia de este límite resulta que $G\circ F$ es diferenciable en a. Además la diferencial de Fréchet de $G\circ F$ en a esta definido como $DG(F(a))\,DF(a)$. ■

Una aplicación de la regla de la cadena se la da para probar la fórmula de los incrementos finitos de Lagrange que se propone a continuación.

Definición 2 *Sea $\Omega\subset V$. Se dice que Ω es convexo si $\forall\alpha\in[0,1]$, $\forall x,y\in\Omega$, se tiene $\alpha x+(1-\alpha)y\in\Omega$.*

Teorema 4 *Sean V, W espacios normados provistos de las normas $\|\cdot\|_V$ y $\|\cdot\|_W$, Ω un abierto convexo de V, y F una función diferenciable en todo punto $a\in\Omega$. Entonces*

$$F(x)-F(a)=\int_0^1 D(F(a+\theta(x-a))d\theta(x-a)\quad\forall x\in\Omega.$$

Demostración. Sean $a\in\Omega$ y $G(\theta)=a+\theta(x-a)\quad\forall\theta\in[0,1]$. Se tiene $G(0)=a$, $G(1)=x\quad$ y $G'(\theta)=x-a$.

Sea H la función de $[0,1]$ en W definida como $H(\theta)=(F\circ G)(\theta)=F(G(\theta))\quad\forall\theta\in[0,1]$. Resulta ■

Teorema 5 ***Demostración.*** *$H(0)=F(a)$, $H(1)=F(x)$, y por hipótesis F es Fréchet diferenciable en todo punto $a\in\Omega$, por la regla de la cadena, se tiene*

$$H'(\theta)=DF(G(\theta))G'(\theta)=DF(G(\theta))(x-a)$$

de donde

$$F(x)-F(a)=\int_0^1 H'(\theta)d\theta=\int_0^1 DF(a+\theta(x-a))d\theta(x-a).$$

■

7.2. Aplicaciones contractivas y lipschisianas.

Sean V un espacio vectorial de dimensión finita provisto de la norma $\|\cdot\|$, E un subconjunto cerrado de V. El conjunto E con la métrica definida como $d(x,y) = \| x-y \| \quad \forall x,y \in E$, es un espacio métrico completo, esto es, toda sucesión de Cauchy en E es convergente en el espacio normado V. Como $E \subset V$, $E \neq \emptyset$, el par (E,d) es un espacio métrico y siendo E cerrado, se prueba que toda sucesión de Cauchy en E es convergente en E, con lo cual (E,d) es un espacio métrico completo.

El conjunto $E =]0,1] \subset \mathbb{R}$ no es cerrado. La sucesión $(x_n) \subset E$ con $x_n = \frac{1}{n} \;\; n = 1,2,\ldots$, es una sucesión de Cauchy en E que no es convergente en E, pues $\lim\limits_{n\to\infty} x_n = 0 \notin E$. Luego (E,d) es un espacio métrico que no es completo.

En esta sección tratamos una clase de funciones denomimadas contractivas y lipschisianas definidas de E en E.

Definición 3 *Sean $E \subset V$, $E \neq \emptyset$ y T de E en E una función. Se dice que T es una aplicación contractiva en E si y solo si satisface la siguiente propiedad:*

$$\exists k,\ 0 \leq k < 1 \text{ tal que } \| T(x) - T(y) \| \leq k \| x-y \| \qquad \forall x,y \in E.$$

La constante k de la definición precedente es independiente de x e y.

Definición 4 *Sean $E \subset V$, $E \neq \emptyset$ y T de E en V una función. Se dice que T es una aplicación lischisiana en E si y solo si satisface la siguiente propiedad:*

$$\exists k > 0 \text{ tal que } \| T(x) - T(y) \| \leq k \| x-y \| \qquad \forall x,y \in E.$$

La constante k de la definición precedente es independiente de x e y.

Como consecuencia inmediata de la definición se tiene que toda aplicación contractiva es uniformemente continua. El recíproco, en general, no es cierto.

Sea T una aplicación contractiva y $\varepsilon > 0$. De la definición se sigue que existe k, $0 \leq k < 1$ tal que $\forall x,y \in E$,

$$\| T(x) - T(y) \| \leq k \| x-y \| < \varepsilon.$$

Elegimos $\delta = \dfrac{\varepsilon}{k}$ $(k \neq 0)$. Entonces $\| x-y \| < \delta \Rightarrow \| T(x) - T(y) \| < \varepsilon$. Observe que $\delta > 0$ es independiente de x e y. Además, si $k = 0$ se deduce que T es constante en E.

Por otro lado, si T es contractiva se tiene $\| T(x) - T(y) \| \leq \| x-y \| \quad \forall x,y \in E$.ya que $0 \leq k < 1$, pero puede suceder esto último sin ser contractiva se verá en un ejemplo propuesto más adelante.

Teorema 6 *Sea T de E en E una función Fréchet diferenciable y lipschisiana. Entonces $\| DT(x) \| = \sup\limits_{\|h\|\leq 1} \| T_x(h) \| \leq k < 1 \;\; \forall x \in E$.*
Si el conjunto E es convexo y $\| DT(x) \| \leq k < 1 \;\; \forall x \in E$. Entonces T es lipschisiana.

Demostración. Por hipótesis T es Fréchet diferenciable en $a \in E$, en consecuencia existe $DT(a) \in L(V)$ tal que $T(a+h) - T(a) - DT(a)(\mathrm{h}) = R(a,h)$ con $\| R(a,h) \| \leq \varepsilon \| h \| \longrightarrow 0$ con $\varepsilon > 0$ arbitrario. Por otro lado, T es lipschisiana, luego existe $k \in [0,1[$ tal que

$$\| T(x) - T(y) \| \leq k \| x-y \| \quad \forall x,y \in E.$$

Entonces

$$\begin{aligned} \| \quad DT(a) \|_{L(V)} &= \sup_{\|h\|\leq 1} \| T(a+h) - T(a) - R(a,h) \| \\ \leq & \sup_{\|h\|\leq 1} (\| T(a+h) - T(a) \| + \| R(a,h) \|) \\ \leq & \sup_{\|h\|\leq 1} (k \| h \| + \varepsilon \| h \|) = k + \varepsilon \end{aligned}$$

De la arbitrariedad de $\varepsilon > 0$ se sigue que $\| DT(a) \|_{L(V)} = \sup_{\|h\|\leq 1} \| T_x(h) \| \leq k < 1 \;\; a \in E.$

Por la fórmula de los incrementos finitos de Lagrange, se tiene

$$F(x) - F(y) = \int_0^1 D(F(y + \theta(x-y))d\theta(x-y) \quad \forall x, y \in \Omega,$$

luego

$$\begin{aligned} \| \quad F(x) - F(y) \|_V &\leq \| \int_0^1 D(F(y + \theta(x-y))d\theta(x-y) \|_V \\ \leq & \| \int_0^1 D(F(y + \theta(x-y))d\theta \|_{L(V)} \| x - y \|_V \\ \leq & \int_0^1 \| D(F(y + \theta(x-y)) \|_{L(V)} \, d\theta \, \| x - y \|_V \\ \leq & \int_0^1 k d\theta \, \| x - y \|_V = k \, \| x - y \|_V \qquad \forall x, y \in V, \end{aligned}$$

que muestra que F es lipschisiana. Note que se requiere de la convexidad de Ω. ■

Definición 5 *Sean $E \subset V$, $E \neq \emptyset$ y T de E en E una función. Un punto $\widehat{x} \in E$ se dice un punto fijo de T si verifica la condición $T(\widehat{x}) = \widehat{x}$.*

Teorema de Banach del punto fijo.

El teorema de Banach del punto fijo es uno de los resultados importantes del análisis no lineal, que se aplica en la resolución de sistemas ecuaciones lineales, ecuaciones en derivadas parciales del tipo no lineal, etc. En esta sección extendemos los resultados obtenidos en el capítulo 5.

Teorema 7 ***(De Banach del punto fijo)***
Sean $E \subset V$ con $E \neq \emptyset$ y E cerrado, T de E en E una aplicación contractiva en E. Entonces, existe un único $\widehat{x} \in E$ tal que $T(\widehat{x}) = \widehat{x}$.

Demostración. La demostración de este teorema la dividimos en dos partes. La primera que corresponde a la existencia del punto fijo $\widehat{x}$ de T y la segunda a la unicidad.

Existencia. Por hipótesis T es contractiva, entonces existe $k, \;\; 0 \leq k < 1$ tal que

$$\| T(x) - T(y) \| \leq k \, \| x - y \| \quad \forall x, y \in E.$$

Sea $x_0 \in E$. Definimos la sucesión $(x_n) \subset E$ como sigue

$$x_1 = T(x_0), \qquad x_2 = T(x_1), \cdots, x_{n+1} = T(x_n) \qquad n = 0, 1, \ldots$$

Mostremos que la sucesión (x_n) es una sucesión de Cauchy en E. Sean $m, n \in \mathbb{Z}^+$ con $m > n$ y sea $p \in \mathbb{Z}^+$ tal que $= n + p$. Entonces, por la desigualdad triangular, se tiene

$$\| x_n - x_m \| = \| x_n - x_{n+p} \| \leq \| x_n - x_{n+1} \| + \| x_{n+1} - x_{n-2} \| + \cdots + \| x_{n+p-1} - x_{n+p} \| .$$

Por la definición de (x_n) se tiene

$$\begin{aligned} \| \quad & x_n - x_{n+1} \| = \| T(x_{n-1}) - T(x_n) \| \leq k \| x_{n-1} - x_n \| \\ = \quad & k \| T(x_{n-2}) - T(x_{n-1}) \| \leq k^2 \| x_{n-2} - x_{n-1} \| \\ & \vdots \\ \leq \quad & k^n \| x_0 - x_1 \| . \end{aligned}$$

Luego, $\| x_n - x_{n+1} \| \leq k^n \| x_0 - x_1 \| \quad \forall n \in \mathbb{Z}^+$. Aplicando el resultado que acabamos de obtener, se obtiene

$$\begin{aligned} \| \quad & x_n - x_m \| \leq k^n \| x_0 - x_1 \| + k^{n+1} \| x_0 - x_1 \| + \cdots + k^{n+p} \| x_0 - x_1 \| \\ = \quad & k^n \| x_0 - x_1 \| (1 + k + \cdots + k^p) \\ & \vdots \\ \leq \quad & k^n \| x_0 - x_1 \| \left(1 + k + \cdots + k^p + k^{p+1} + \cdots\right). \end{aligned}$$

Sea $S_p(k) = 1 + k + \cdots + k^p$. Entonces

$$(1-k) S_p(k) = S_p(k) - k S_p(k) = 1 + k + \cdots k^p - \left(k + k^2 + \cdots + k^{p+1}\right) = 1 - k^{p+1}.$$

de donde

$$S_p(k) = \frac{1 - k^{p+1}}{1-k} = \frac{1}{1-k} - \frac{k^{p+1}}{1-k}.$$

Como $0 \leq k < 1$, $\lim\limits_{p\to\infty} k^{p+1} = 0$. Luego

$$\lim_{p\to\infty} S_p(k) = \lim_{p\to\infty} \left(\frac{1}{1-k} - \frac{k^{p+1}}{1-k} \right) = \frac{1}{1-k} - \lim_{p\to\infty} \frac{\kappa^{p-1}}{1-k} = \frac{1}{1-k},$$

con lo cual $\sum_{p=0}^{\infty} k^p = \lim_{p\to\infty} S_p(k) = \frac{1}{1-k}$.

Por lo tanto $\| x_n - x_m \| \leq k^n \| x_0 - x_1 \| \sum_{p=0}^{\infty} k^p = \frac{k^n}{1-k} \| x_0 - x_1 \|$. Puesto que $\lim\limits_{n\to\infty} k^n = 0$ se sigue que $\forall \varepsilon > 0$, $\exists n_0 \in \mathbb{Z}^+$ tal que $\forall n \geq n_0, \quad k^n < \frac{1-k}{|x_0 - x_1|}\varepsilon$. Luego

$$\| x_n - x_m \| \leq \frac{k^n}{1-k} \| x_0 - x_1 \| < \varepsilon \quad \text{si} \quad m, n \geq n_0,$$

es decir que (x_n) es una sucesión de Cauchy en E y por hipótesis E es cerrado, entonces la sucesión (x_n) tiene límite en E; esto es, existe $\widehat{x} \in E$ tal que $\lim\limits_{n\to\infty} x_n = \widehat{x}$.

Puesto que T es contractiva, T es uniformemente continua y por lo tanto continua. Luego

$$\lim_{n\to\infty} T(x_n) = T\left(\lim_{n\to\infty} x_n\right) = T(\widehat{x}).$$

Además, $x_{n+1} = T(x_n)$, y $\lim_{n\to\infty} x_{n+1} = \widehat{x}$,resulta que $T(\widehat{x}) = \lim_{n\to\infty} T(x_n) = \lim_{n\to\infty} x_{n+1} = \widehat{x}$. Así, $T(\widehat{x}) = \widehat{x}$ o sea $\widehat{x} \in E$ es un punto fijo de T.

Unicidad. Probemos que $\widehat{x} \in E$ tal que $T(\widehat{x}) = \widehat{x}$ es único. Para el efecto, supongamos que existe $y \in E$ tal que $T(y) = y$. Mostremos que $y = \widehat{x}$. Como T es contractiva, se tiene

$$\| \widehat{x} - y \| = \| T(\widehat{x}) - T(y) \| \leq k \| \widehat{x} - y \|,$$

de donde $\| \widehat{x} - y \| (1-k) \leq 0$, y siendo $0 \leq k < 1$, entonces $1 - k > 0$ y en consecuencia $\| \widehat{x} - y \| \leq 0$. Como el valor absoluto es no negativo, la única posibilidad es $\| \widehat{x} - y \| = 0 \Leftrightarrow y = \widehat{x}$. ■

Observaciones

1. El teorema de Banach del punto fijo asegura la existencia de un único punto fijo $\widehat{x} \in E$ de la plicación contractiva T definida en el conjunto cerrado E de V.

2. En los textos de Análisis, el teorema de Banach del punto fijo se enuncia como sigue: Sea (E,d) un espacio métrico completo y T de E en E una aplicación contractiva. Entonces, existe un único $\widehat{x} \in E$ tal que $T(\widehat{x}) = \widehat{x}$.

   La demostración del teorema de Banach del punto fijo para espacios métricos completos muy generales (E,d) es muy similar a la aquí propuesta con la salvedad que la métrica $d(x,y) = \| x-y \|$ $x, y \in V$ se remplazan simplemente por $d(x,y)$ con d la métrica en el conjunto E.

3. En la demostración del teorema de Banach del punto fijo se muestra una manera de calcular el punto fijol $\widehat{x} \in E$. Pués se parte de un punto arbitrario $x_0 \in E$ y se construye la sucesión $(x_n) \subset E$ tal que $x_{n+1} = T(x_n)$ $n = 0, 1, \ldots$. Entonces $\widehat{x} = \lim\limits_{n\to\infty} x_n$ es el punto fijo de T. De este hecho se desprende que podemos aproximar el funto fijo $\widehat{x}$ con una precisión $\varepsilon > 0$.

7.3. Resolución numérica de sistemas de ecuaciones no lineales

Sean $n \geq 2$, $\Omega \subset \mathbb{R}^n$ con $\Omega \neq \emptyset$ y $\overrightarrow{\mathbf{F}}$ una función de Ω en $\mathbb{R}^n$. Pongamos $\overrightarrow{\mathbf{F}} = (f_1, \ldots, f_n)^T$ donde cada f_i, $i = 1, \ldots, n$, es una función de Ω en $\mathbb{R}$. Consideramos el problema **(P)** siguiente:

$$\text{hallar } \widehat{x} \in \Omega, \text{ si existe, tal que } \overrightarrow{\mathbf{F}}(\widehat{x}) = 0. \tag{P}$$

Note que la ecuación $\mathbf{F}(\overrightarrow{x}) = \overrightarrow{0}$ es equivalente al sistema de ecuaciones:

$$\begin{cases} f_1(\overrightarrow{x}) = 0 \\ \vdots \\ f_n(\overrightarrow{x}) = 0. \end{cases}$$

El teorema de Bolzano no tiene validez para funciones de $\Omega \subset \mathbb{R}^n$ en $\mathbb{R}^n$. Asumimos que el problema **(P)** tiene solución, esto es, asumimos la existencia de al menos una solución $\widehat{x} \in \Omega$ tal que $\overrightarrow{\mathbf{F}}(\widehat{x}) = 0$.

1. **Método de punto fijo**

   Supongamos que Ω es cerrado y que la función $\overrightarrow{\mathbf{F}}$ se expresa en la forma

   $$\overrightarrow{\mathbf{F}}(\overrightarrow{x}) = \overrightarrow{x} - \overrightarrow{\mathbf{G}}(\overrightarrow{x}) \qquad \overrightarrow{x} \in \Omega,$$

   con $\overrightarrow{\mathbf{G}}$ una aplicación contractiva en Ω. Entonces,

   $$\overrightarrow{\mathbf{F}}(\overrightarrow{x}) = 0 \Leftrightarrow \overrightarrow{\mathbf{G}}(\widehat{x}) = \widehat{x},$$

   es decir que $\widehat{x} \in \Omega$ es un punto fijo de $\overrightarrow{\mathbf{G}}$.

   La sucesión $(\overrightarrow{x}_m)$ definida por

   $$\begin{cases} \overrightarrow{x}_0 \in \Omega, \\ \overrightarrow{x}_{m+1} = \mathbf{G}(\overrightarrow{x}_m) \qquad m = 0, 1, \ldots \end{cases}$$

   converge a $\widehat{x}$ (teorema de Banach del punto fijo).

   Sea $\varepsilon > 0$ $\left(\varepsilon = 10^{-4}, 10^{-5}, \ldots\right)$ la precisión con la que se desea aproximar $\widehat{x}$ y $N_{\text{máx}}$ el número máximo de iteraciones. Se tiene el siguiente algoritmo de punto fijo para aproximar soluciones de sistemas de ecuaciones no lineales.

   Algoritmo

   Datos de entrada: ε precisión, $N_{\text{máx}}$ número máximo de iteraciones, funciones $g_1, \ldots, g_n$.

   Datos de salida: n número de iteraciones, $\overrightarrow{y}$ solución aproximada, $\overrightarrow{\mathbf{F}}(\overrightarrow{y})$.

   1. $\overrightarrow{x} = \overrightarrow{x}_0$

2. Par $k = 0, 1, \ldots, N_{\text{máx}}$
3. $\quad \overrightarrow{y} = \overrightarrow{\mathbf{G}}(\overrightarrow{x})$
4. $\quad$ Si $\|\overrightarrow{x} - \overrightarrow{y}\| < \varepsilon$ continuar en 6)**.**
5. $\quad \overrightarrow{x} = \overrightarrow{y}$
6. Si $k < N_{\text{máx}}$, imprimir $k, \overrightarrow{y}, \overrightarrow{\mathbf{F}}(\overrightarrow{y})$. Continuar en 8).
7. Si $k = N_{\text{máx}}$, imprimir $\overrightarrow{y}, \mathbf{F}(\overrightarrow{y})$.
8. Fin.

Nota: La norma $\|\cdot\|$ en $\mathbb{R}^n$ que se considera aquí es la norma euclídea definida como sigue:

$$\|\overrightarrow{a}\| = \left(\sum_{i=1}^{n} a_i^2\right)^{\frac{1}{2}} \text{ con } \overrightarrow{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n.$$

2. **Método de Newton**

Supongamos que $\overrightarrow{\mathbf{F}} \in C^1(\Omega)$ y $\widehat{x} \in \Omega$ tal que $\overrightarrow{\mathbf{F}}(\overrightarrow{x}) = 0$.

Por el desarrollo de Taylor en un entorno de $\widehat{x}$, se tiene

$$0 = \overrightarrow{\mathbf{F}}(\widehat{x}) = \mathbf{F}(\overrightarrow{x}) + D\overrightarrow{\mathbf{F}}(\overrightarrow{x})(\widehat{x} - \overrightarrow{x}) + 0\left(\|\widehat{x} - \overrightarrow{x}\|^2\right),$$

de donde $D\overrightarrow{\mathbf{F}}(\overrightarrow{x})$ es la matriz jacobiana definida por:

$$D\overrightarrow{\mathbf{F}}(\overrightarrow{x}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}(\overrightarrow{x}) & \cdots & \dfrac{\partial f_1}{\partial x_n}(\overrightarrow{x}) \\ \vdots & & \vdots \\ \dfrac{\partial f(n)}{\partial x_1}(\overrightarrow{x}) & \cdots & \dfrac{\partial f(n)}{\partial x_n}(\overrightarrow{x}) \end{bmatrix} \qquad \overrightarrow{x} \in \Omega.$$

Suponemos que la matriz $D\overrightarrow{\mathbf{F}}(\overrightarrow{x})$ no es singular en todo punto $\overrightarrow{x} \in \Omega$.

Si se desprecia el término $0\left(\|\widehat{x} - \overrightarrow{x}\|^2\right)$ en el desarrollo de Taylor precedente, se tiene

$$\overrightarrow{\mathbf{F}}(\overrightarrow{x}) + D\overrightarrow{\mathbf{F}}(\overrightarrow{x})(\widehat{x} - \overrightarrow{x}) = 0,$$

y siendo $D\overrightarrow{\mathbf{F}}(\overrightarrow{x})$ no singular, se sigue que

$$\begin{aligned} D\overrightarrow{\mathbf{F}}(\overrightarrow{x})(\widehat{x} - \overrightarrow{x}) &= -\overrightarrow{\mathbf{F}}(\overrightarrow{x}) \\ \widehat{x} - \overrightarrow{x} &= -\left(D\overrightarrow{\mathbf{F}}(\overrightarrow{x})\right)^{-1}\overrightarrow{\mathbf{F}}(\overrightarrow{x}) \end{aligned}$$

de donde

$$\widehat{x} = \overrightarrow{x} - \left(D\overrightarrow{\mathbf{F}}(\overrightarrow{x})\right)^{-1}\overrightarrow{\mathbf{F}}(\overrightarrow{x})$$

lo que nos permite definir la función de iteración φ :

$$\varphi(\overrightarrow{x}) = \overrightarrow{x} - \left(D\overrightarrow{\mathbf{F}}(\overrightarrow{x})\right)^{-1}\overrightarrow{\mathbf{F}}(\overrightarrow{x}) \qquad \overrightarrow{x} \in \Omega,$$

con $D\overrightarrow{\mathbf{F}}(\overrightarrow{x})$ matriz no singular.

El método de Newton para aproximar la raíz $\widehat{x} \in \Omega$ de $\mathbf{F}(\overrightarrow{x}) = 0$ es el siguiente:

$$\begin{cases} \overrightarrow{x}_0 \in \Omega \text{ una aproximación inicial de } \widehat{x}, \\ \overrightarrow{x}_{m+1} = \varphi(\overrightarrow{x}_m) \qquad m = 0, 1, \ldots \end{cases}$$

Si φ es una aplicación contractiva en Ω, por el teorema de Banach del punto fijo, la sucesión $(\overrightarrow{x}_m)$ generada por el método de Newton converge a $\widehat{x}$.

Puesto que

$$\begin{aligned} \overrightarrow{x}_{m+1} &= \varphi(\overrightarrow{x}_m) = \overrightarrow{x}_m - \left(D\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m)\right)^{-1}\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m), \\ \left(D\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m)\right)^{-1}\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m) &= \overrightarrow{x}_m - \overrightarrow{x}_{m+1}, \\ D\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m)(\overrightarrow{x}_m - \overrightarrow{x}_{m+1}) &= \overrightarrow{\mathbf{F}}(\overrightarrow{x}_m). \end{aligned}$$

Ponemos $\overrightarrow{y} = \overrightarrow{x}_m - \overrightarrow{x}_{m+1} \Rightarrow \overrightarrow{x}_{m+1} = \overrightarrow{x}_m - \overrightarrow{y}$. Se tiene

$$D\mathbf{F}(\overrightarrow{x}_m)\overrightarrow{y} = \overrightarrow{\mathbf{F}}(\overrightarrow{x}_m),$$

que es un sistema de ecuaciones lineales con $\overrightarrow{y}$ el vector incógnita. Este sistema de ecuaciones lineales puede ser resuelto utilizando los métodos de eliminación gaussiana con pivoting, factorización LU, Choleski, dependiento de las propiedades de la matriz jacobiana $D\overrightarrow{\mathbf{F}}(\overrightarrow{x})$.

Debe advertirse que el cálculo directo de $\left(D\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m)\right)^{-1}$ no se realiza.

Dado $\overrightarrow{x}_m$, para calcular la nueva aproximación $\overrightarrow{x}_{m+1}$ de $\widehat{x}$, resolvemos el sistema de ecuaciones lineales

$$D\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m)\overrightarrow{y} = \overrightarrow{\mathbf{F}}(\overrightarrow{x}_m),$$

y una vez calculado $\overrightarrow{y}$, se tiene $\overrightarrow{x}_{m+1} = \overrightarrow{x}_m - \overrightarrow{y}$; y, se repite el procedimiento hasta considerar $\overrightarrow{x}_m$ tal que $\overrightarrow{\mathbf{F}}(\overrightarrow{x}_m) \simeq 0$ sea satisfactorio.

Sea $\varepsilon > 0$ la precisión con la que se aproxima $\widehat{x}$ y $N_{\text{máx}}$ el número máximo de iteraciones. El esquema numérico generado por el método de Newton se presenta a continuación.

Algoritmo

Datos de Entrada: $\varepsilon, N_{\text{máx}}$, funciones $f_1, \ldots, f_n$, $\dfrac{\partial f_i}{\partial f_j} \quad i, j = 1, \ldots, n$

Datos de Salida: n número de iteraciones, $\vec{x}$, $\tilde{\mathbf{F}}(\vec{x})$.

1. $\vec{x} = \vec{x}_0$
2. Para $k = 1, \ldots, N_{\text{máx}}$
3. Resolver el sistema de ecuaciones $D\tilde{\mathbf{F}}(\vec{x})\,\vec{y} = \tilde{\mathbf{F}}(\vec{x})$.
4. Si $\|\vec{y}\| < \varepsilon$. Continuar en 6).
5. $\vec{x} = \vec{x} - \vec{y}$
6. Si $k < N_{\text{máx}}$, imprimir: n, $\vec{x}$, $\tilde{\mathbf{F}}(\vec{x})$. Continuar en 8).
7. Si $k = N_{\text{máx}}$, imprimir: $N_{\text{máx}}$, $\vec{x}$, $\tilde{\mathbf{F}}(\vec{x})$.
8. Fin

Ejemplo

Resolver el sistema de ecuaciones no lineales $\begin{cases} x^2 - y^2 = 1 \\ (x+3)^2 + 4(y-3)^2 = 4. \end{cases}$

Asociemos a este sistema de ecuaciones no lineales la función siguiente:

$$\tilde{\mathbf{F}}(x, y) = \left(x^2 - y^2 - 1, (x+3)^2 + 4(y-3)^2 - 4\right) \quad (x, y) \in \mathbb{R}.$$

Entonces

$$\tilde{\mathbf{F}}(x, y) = (0, 0) \Longleftrightarrow \begin{cases} x^2 - y^2 = 1, \\ (x+3)^2 + 4(y-3)^2 = 4. \end{cases}$$

El conjunto de puntos $\left\{(x, y) \in \mathbb{R} \,|x^2 - y^2 = 1\right\}$ representa una hipérbola y la ecuación el conjunto de puntos $\left\{(x, y) \in \mathbb{R} \,|\, (x+3)^2 + 4(y-3)^2 = 4\right\}$ representa una elipse.

En la figura siguiente se muestran los graficos de la hipérbola y de la elipse.

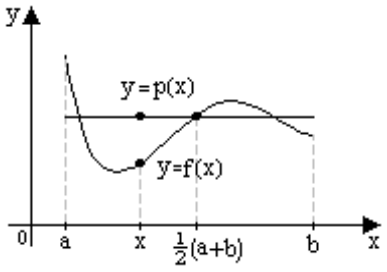


Figura 72

Las gráficas de la hipérbola y de la elipse se cortan en dos puntos. Por lo tanto, el sistema de ecuaciones $\tilde{\mathbf{F}}(x,y)=(0,0)$ tiene dos soluciones $\widehat{X}_1=(\widehat{x}_1,\widehat{y}_1)$, $\widehat{X}_2=(\widehat{x}_2,\widehat{y}_2)$.

La matriz jacobiana de $\tilde{\mathbf{F}}$ está definida como

$$D\tilde{\mathbf{F}}(x.y)=\begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix}=\begin{bmatrix} 2x & -2y \\ 2(x+3) & 8(y-3) \end{bmatrix} \quad (x,y)\in\mathbb{R}.$$

i. Apliquemos el método de Newton para calcular una aproximación de $\widehat{X}_1=(\widehat{x}_1,\widehat{y}_1)$.

Sea $\vec{x}_0=\begin{bmatrix} -2{,}5 \\ 2{,}5 \end{bmatrix}$. Entonces $D\tilde{\mathbf{F}}(\vec{x}_0)\,\vec{y}=\tilde{\mathbf{F}}(\vec{x}_0)$, resolviendo este sistema de ecuaciones y tomando en cuenta que $\vec{x}_1=\vec{x}_0-\vec{y}$, se obtiene $\vec{x}_1=\begin{bmatrix} -2{,}11 \\ 1{,}91 \end{bmatrix}$. Continuando con la ejecución del método de Newton, se obtienen los siguientes resultados:

$$\vec{x}_2=\begin{bmatrix} -2{,}28453 \\ 2{,}051495 \end{bmatrix}, \qquad \vec{x}_3=\begin{bmatrix} -2{,}29376 \\ 2{,}064322 \end{bmatrix}, \qquad \vec{x}_4=\begin{bmatrix} -2{,}29385 \\ 2{,}06440 \end{bmatrix}.$$

Con una precisión $\varepsilon=10^{-5}$, una aproximación de la solución es $\widehat{X}_1=\begin{bmatrix} -2{,}29385 \\ 2{,}06440 \end{bmatrix}$.

ii. Apliquemos el método de Newton para calcular una aproximación de $\widehat{X}_2=(\widehat{x}_2,\widehat{y}_2)$.

Sea $\vec{x}_0=\begin{bmatrix} -4{,}2 \\ 3{,}8 \end{bmatrix}$. Los resultados de la aplicación del algoritmo generado por el método de Newton se muestra a continuación.

$$\vec{x}_1=\begin{bmatrix} -4{,}00444 \\ 3{,}87333 \end{bmatrix}, \qquad \vec{x}_2=\begin{bmatrix} -3{,}99476 \\ 3{,}86757 \end{bmatrix}, \qquad \vec{x}_3=\begin{bmatrix} -3{,}99473 \\ 3{,}86754 \end{bmatrix},$$

$$\widehat{X}_1=\begin{bmatrix} -3{,}99473 \\ 3{,}86754 \end{bmatrix}.$$

Observación

De la ecuación $x^2-y^2=1$ se deduce $x=\pm\sqrt{1+y^2}$. Pongamos $x=-\sqrt{1+y^2}$ en la ecuación $(x+3)^2+4(y-3)^2=4$. Obtenemos

$$\begin{aligned} \left(-\sqrt{1+y^2}+3\right)^2+4(y-3)^2 &= 4, \\ 5y^2-24y+42 &= 6\sqrt{1+y^2}, \end{aligned}$$

de donde

$$y^4-9{,}6y^3+38{,}4y^2-80{,}64y+69{,}12=0.$$

Sea

$$\begin{aligned} P(y) &= 69{,}12 - 80{,}64y + 38{,}4y^2 - 9{,}6y^3 + y^4 \\ &= 69{,}12 + y\left(-80{,}64 + y\left(38{,}4 + y\left(-9{,}6 + y\right)\right)\right). \end{aligned}$$

Determinemos las fronteras inferior y superior donde están localizadas las raíces positivas de la ecuación $P(y) = 0$.

Se tiene

$$\begin{aligned} R &= 1 + (80{,}64)^{\frac{1}{3}} \simeq 5{,}3203 < 5{,}5 \\ r &= \frac{1}{1 + \dfrac{\underset{k=1,\ldots n}{\text{máx}} |a_k|}{|a_0|}} = \frac{1}{1 + \frac{80{,}64}{60{,}92}} = 0{,}46 < 0{,}5. \end{aligned}$$

Buscamos las raíces de $P(y) = 0$ en el intervalo $[0{,}5, 5{,}5]$.

Con un paso $h = 0{,}5$, la aplicación del algoritmo de búsqueda del cambio de signo muestra que existen dos raíces localizadas en los intervalos $[2, 2{,}5]$ y $[3{,}5, 4]$. Además, $P(2) = 0{,}64$, $P(2{,}5) = -3{,}42$, $P(3{,}5) = -4{,}26$, $P(4) = 2{,}56$.

i. Cálculo de $\widehat{y}_1 \in [2, 2{,}5]$. La función de iteración del método de Newton está dada por

$$\varphi(y) = y - \frac{P(y)}{P'(y)}$$

Sea $y_0 = 2$. Entonces

$$\begin{aligned} y_1 &= \varphi(2) = 2{,}0625, \\ y_2 &= \varphi(y_1) = \varphi(2{,}0625) = 2{,}064402, \\ y_3 &= \varphi(y_2) = \varphi(2{,}064402) = 2{,}064404, \end{aligned}$$

ii. Cálculo de $\widehat{y}_2 \in [3{,}5, 4]$.

Sea $y_0 = 4$,

$$\begin{aligned} y_1 &= \varphi(y_0) = \varphi(4) = 3{,}882353, \\ y_2 &= \varphi(y_1) = \varphi(3{,}882353) = 3{,}867753, \\ y_3 &= \varphi(y_2) = \varphi(3{,}867753) = 3{,}867541. \end{aligned}$$

Con una precisión $\varepsilon = 10^{-6}$, $\widehat{y}_1 = 2{,}064404$, $\widehat{y}_2 = 3{,}867541$. Resulta

$$\begin{aligned} P(y) &= (y - \widehat{y}_1)(y - \widehat{y}_2)\left(y^2 + by + c\right) \\ &= (y - 2{,}064404)(y - 3{,}867541)\left(y^2 + by + c\right) \\ &= y^4 + (b - 5{,}93194)y^3 + (c - 5{,}93194b + 7{,}98415)y^2 + \\ &\quad (7{,}98415b - 5{,}93194c)y + 7{,}98415c. \end{aligned}$$

con lo cual

$$\begin{cases} b - 5{,}93194 = -9{,}6 \\ c - 5{,}93194b + 7{,}98415 = 38{,}4 \\ 7{,}98415b - 5{,}93194c = -80{,}64 \\ 7{,}98415c = 69{,}12 \end{cases} \Rightarrow \begin{array}{c} b = -3{,}66806, \\ c = 8{,}65714. \end{array}$$

Luego

$$y^2 + by + c = 0 \Longleftrightarrow y^2 - 3{,}66806y + 8{,}65714 = 0.$$

Obtenemos

$$\begin{aligned} y_3 &= 1{,}83403 - 2{,}30076i, \\ y_4 &= \overline{y_3} = 1{,}83403 + 2{,}30076i. \end{aligned}$$

¿Raíces Múltiples? Acabamos de calcular todas las raíces reales o complejas de la ecuación $P(y) = 0$.

Si únicamente hubiesemos calculado las raíces reales $\widehat{y}_1, \widehat{y}_2$, las dos raíces restantes podían ser complejas o una real de multiplicidad 2. Despejamos esta duda aplicando los métodos descritos en la aproximación de raíces de multiplicidad.

Definimos

$$u(y) = \frac{P(y)}{P'(y)} \quad P'(y) \neq 0.$$

La aplicación del algoritmo de búsqueda del cambio de signo aplicado a la función u muestra la existencia de dos raíces $\widehat{y}_1 \in [2, 2{,}5]$, $\widehat{y}_2 \in [3{,}5, 4]$. ¿Qué ocurre? Explique.

Puesto $x = -\sqrt{1+y^2}$se deducen $\widehat{x}_1 = -\sqrt{1+\widehat{y}_1^2}$ y $\widehat{x}_2 = -\sqrt{1+\widehat{y}_2^2}$.

7.4. Métodos iterativos de resolución de sistemas de ecuaciones lineales

Para terminar este capítulo, en contraposición con los métodos directos de resolución de sistemas de ecuaciones lineales presentamos dos métodos iterativos, el de Jacobi y el de SOR.

7.4.1. Métodos de Jacobi y Gauss-Seidel

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$, $\overrightarrow{b}^T = (b_1, \cdots, b_n) \in \mathbb{R}^n$. Consideramos el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ que en forma explícita se escribe como sigue:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + \cdots + a_{mn}x_n &= b_1. \end{aligned}$$

El método que vamos a describir requiere, en principio, que los coeficientes $a_{11}, a_{22}, \ldots, a_{nn}$ que conforman la diagonal de la matriz A sean distintos de cero. Veremos más adelante que se puede hacer en caso de que esto no suceda.

El método de Jacobi consiste en despejar x_1 de la primera ecuación, x_2 de la segunda ecuación, etc.:

$$\begin{aligned} x_1 &= -\frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \cdots - \frac{a_{1n}}{a_{11}}x_n + \frac{b_1}{a_{11}}, \\ x_2 &= -\frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_3 - \cdots - \frac{a_{2n}}{a_{22}}x_n + \frac{b_2}{a_{22}}, \end{aligned}$$

en general

$$x_i = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}}x_j + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n.$$

Sea el vector $\overrightarrow{X}^{(0)}$, un vector cualquiera: $\overrightarrow{X}^{(0)} = \begin{pmatrix} x_1^{(0)} \\ \vdots \\ x_n^{(0)} \end{pmatrix}$, en base a las ecuaciones precedente se obtiene el vector: $\overrightarrow{X}^{(1)} = \begin{pmatrix} x_1^{(1)} \\ \vdots \\ x_n^{(1)} \end{pmatrix}$, donde

$$x_i^{(1)} = -\sum_{j=1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(0)} + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n.$$

Con los valores obtenidos para $x_i^{(1)}$ y las ecuaciones en (1) se obtiene $\overrightarrow{X}^{(2)}$ y, así sucesivamente , se obtienen $\overrightarrow{X}^{(3)}$, $\overrightarrow{X}^{(4)}, \ldots$, mediante la fórmula recurrente:

$$x_j^{(k+1)} = -\sum_{j=1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n.$$

El procedimiento anterior se expresa en forma matricial como sigue: la matriz A del sistema se descompone en la forma

$$A = D - L - U,$$

donde D es una matriz diagonal, L una matriz triangular inferior y U una matriz triangular superior:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ -a_{21} & 0 & 0 \\ -a_{31} & -a_{32} & 0 \end{pmatrix} - \begin{pmatrix} 0 & -a_{12} & -a_{13} \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{pmatrix}.$$

El sistema $A\overrightarrow{x} = \overrightarrow{b}$ toma entonces la forma:

$$\begin{aligned} (D - L - U)\overrightarrow{x} &= \overrightarrow{b} \\ D\overrightarrow{x} - (L + U)\overrightarrow{x} &= \overrightarrow{b} \\ D\overrightarrow{x} &= (L + U)\overrightarrow{x} + \overrightarrow{b} \\ \overrightarrow{x} &= D^{-1}(L + U)\overrightarrow{x} + D^{-1}\overrightarrow{b}. \end{aligned}$$

así,

$$A\overrightarrow{x} = \overrightarrow{b} \Longleftrightarrow \overrightarrow{x} = D^{-1}(L + U)\overrightarrow{x} + D^{-1}\overrightarrow{b}.$$

1. Claramente, la matriz D es no singular. Si notamos $T = D^{-1}(L + U)$ y $C = D^{-1}\overrightarrow{b}$, la ecuación $\overrightarrow{x} = D^{-1}(L + U)\overrightarrow{x} + D^{-1}\overrightarrow{b}$ toma la forma: $\overrightarrow{x} = T\overrightarrow{x} + \overrightarrow{c}$, y la fórmula recurrente del método De Jacobi arriba formulada se expresa como:

$$\overrightarrow{X}^{(k+1)} = T\overrightarrow{X}^{(k)} + \overrightarrow{c}, \quad k = 0, 1, 2, 3, \ldots.$$

   La validez de $\overrightarrow{X}^{(k)}$ como solución aproximada del sistema $A\overrightarrow{x} = \overrightarrow{b}$ está garantizada por los siguientes resultados.

   Recordemos que si A es una matriz cuadrada, su radio espectral $\rho(A)$ es el máximo de los valores absolutos de sus valores propios.

Sea A una matriz no singular. Para cualquier vector $\overrightarrow{X}^{(0)}$ en $\mathbb{R}^n$, la sucesión de vectores $\left(\overrightarrow{X}^{(k)}\right)_k$ definida por

$$\overrightarrow{X}^{(k+1)} = T\overrightarrow{X}^{(k)} + \overrightarrow{c}, \quad k = 1, 2, 3, \ldots$$

converge a la solución del sistema $A\overrightarrow{x} = \overrightarrow{b}$, si y solo si $\rho(A) < 1$.

Sea A una matriz cuadrada y $\overrightarrow{x}$ un vector (columna) de $\mathbb{R}^n$, se tiene $A\overrightarrow{x} \in \mathbb{R}^n$ y tiene sentido hablar de la norma (en $\mathbb{R}^n$) $\|A\overrightarrow{x}\|_\infty$. El máximo de estas normas cuando $\overrightarrow{x}$ recorre todos los vectores de norma 1 se llama la norma $\|\cdot\|_\infty$ de la matriz A, y está definida como

$$\|A\|_\infty = \max_{\|\overrightarrow{x}\|_\infty \leq 1} \|A\overrightarrow{x}\|_\infty .$$

Se tiene que

$$\rho(A) \leq \|A\|_\infty .$$

Ejemplo

Si $A = \begin{pmatrix} 1 & -1 & 0 \\ 2 & 3 & -2 \\ 0 & 1 & 1 \end{pmatrix}$, se tiene $\rho(A) \leq \|A\|_\infty = 7$.

Sea A una matriz no singular. Si $\|T\| < 1$, la sucesión de vectores $\left(\overrightarrow{X}^{(k)}\right)_k$ converge a la solución $\overrightarrow{X}$ del sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ para cualquier $\overrightarrow{X}^{(0)} \in \mathbb{R}^n$. Además se satsface:

$$\begin{aligned} \left\|\overrightarrow{x} - \overrightarrow{X}^{(k)}\right\| &\leq \|T\|^k \left\|\overrightarrow{X}^{(0)} - \overrightarrow{x}\right\| \quad k = 1, 2, \ldots, \\ \left\|\overrightarrow{x} - \overrightarrow{X}^{(k)}\right\| &\leq \frac{\|T\|^k}{1 - \|T\|} \left\|\overrightarrow{X}^{(1)} - \overrightarrow{X}^{(0)}\right\| \quad k = 1, 2, \ldots . \end{aligned}$$

La última desigualdad nos proporciona una cota para el error de aproximación de $\overrightarrow{X}^{(k)}$ a la solución $\overrightarrow{x}$.

Ejemplo

Consideremos el sistema el sistema de ecuaciones lineales siguiente:

$$\begin{aligned} 10x_1 - x_2 + 2x_3 &= 6 \\ -x_1 + x_2 - x_3 + 3x_4 &= 25 \\ 2x_1 - x_2 + 10x_3 - x_4 &= -11 \\ 3x_2 - x_3 + 8x_4 &= 15 \end{aligned}$$

Apliquemos el método de Jacobi. Partiendo de: $\overrightarrow{X}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, se obtiene en la iteración $k = 10$ el vector $\overrightarrow{X}^{(10)} = \begin{pmatrix} 1{,}0001 \\ 1{,}9998 \\ -0{,}9998 \\ 0{,}9998 \end{pmatrix}$. Por otra parte, la solución exacta es $\overrightarrow{x} = \begin{pmatrix} 1 \\ 2 \\ -1 \\ 1 \end{pmatrix}$. Un método que, generalmente, produce una convergencia más rápida de la sucesión $\left(\overrightarrow{X}^{(k)}\right)$ consiste en utilizar los valores de $x_1^{(k)}, x_2^{(k)}, \ldots, x_{j-1}^{(k)}$ para calcular $x_i^{(k)}$, en lugar de $x_1^{(k-1)}, \ldots, x_{j-1}^{(k-1)}$. Así:

$$\begin{aligned} x_1^{(k)} &= -\frac{a_{12}}{a_{11}} x_2^{(k-1)} - \frac{a_{13}}{a_{11}} x_3^{(k-1)} - \cdots - \frac{a_{1n}}{a_{11}} x_n^{(k-1)} + \frac{b_1}{a_{11}}, \\ x_2^{(k)} &= -\frac{a_{21}}{a_{22}} x_1^{(k)} - \frac{a_{23}}{a_{22}} x_3^{(k-1)} - \cdots - \frac{a_{2n}}{a_{22}} x_n^{(k-1)} + \frac{b_2}{a_{22}}, \\ x_3^{(k)} &= -\frac{a_{31}}{a_{33}} x_1^{(k)} - \frac{a_{32}}{a_{33}} x_2^{(k)} - \cdots - \frac{a_{32}}{a_{33}} x_4^{(k-1)} - \cdots - \frac{a_{3n}}{a_{33}} x_n^{(k-1)} + \frac{b_3}{a_{33}}, \end{aligned}$$

en general

$$x_i^{(k)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k-1)} + \frac{b_i}{a_{ii}}.$$

En forma matricial este método tiene la forma

$$(D-L)\,\overrightarrow{X}^{(k)} = U\overrightarrow{X}^{(k-1)} + \overrightarrow{b}\,,$$

y considerando que la matriz $D-L$ es no singulare, la ecuación precedente es quivalente a

$$\overrightarrow{X}^{(k)} = (D-L)^{-1}\,U\overrightarrow{X}^{(k-1)} + (D-L)^{-1}\,\overrightarrow{b} \quad k = 1, 2, \cdots,$$

o también

$$\overrightarrow{X}^{(k)} = T\overrightarrow{X}^{(k-1)} + \overrightarrow{c},$$

con $T = (D-L)^{-1}\,U$ y $\overrightarrow{c} = (D-L)^{-1}\,\overrightarrow{b}$. Este método se conoce como el método de Gauss-Seidel.

En el ejemplo anterior, con el método de Gauss-Seidel se obtiene para la iteración $k = 5$ prácticamente la solución exacta:

$$\overrightarrow{X}^{(5)} = \begin{pmatrix} 1{,}0001 \\ 2{,}0000 \\ -1{,}0000 \\ 1{,}000 \end{pmatrix}.$$

Tanto en el métod de Jacobi como en el de Gauss-Seidel se requiere que los términos de la diagonal de la matriz A sean no nulos: $a_{11} \neq 0, a_{22} \neq 0, \ldots, a_{nn} \neq 0$.

En caso de que esto no suceda, se reordenan las ecuaciones para conseguir este objetivo:

El sistema:

$$\begin{aligned} a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3, \end{aligned}$$

es equivalente a

$$\begin{aligned} a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned}$$

Si una reordenación no es posible de tal manera que el coeficiente de cada x_i en la i-ésima ecuación sea distinto de cero la matriz A tiene determinante nulo, lo que implica que el sistema no tiene solución o tiene infinitas soluciones.

7.4.2. Método SOR (Successive Over-Relaxation)

Sea el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, donde A es una matriz cuadrada no singular. Si $\widetilde{X}$ es una aproximación de la solución del sistema,

$$\overrightarrow{r} = \overrightarrow{b} - A\widetilde{X}$$

se llama el vector residual de $\widetilde{X}$. El objetivo es hallar una sucesión de soluciones aproximadas de tal manera que la sucesión de los vectores residuales converja a 0.

Con relación al método de Gauss-Seidel consideremos la solución aproximada de $A\overrightarrow{x} = \overrightarrow{b}$:

$$\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_i^{(k-1)}, \ldots, x_n^{(k-1)}\right)$$

y notemos por $R_i^{(k)} = \left(r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)}\right)$ a su vector residual. Se tiene que

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} - a_{ij}x_i^{(k-1)}.$$

Como en el método de Gauss-Seidel

$$X_i^{(k)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}x_j^{(k)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}x_j^{(k-1)} + \frac{b_j}{a_{ii}},$$

se tiene para $i = 1, \ldots, n$

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}$$

o

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}.$$

Modificando la última ecuación a:

$$x_i^{(k)} = x_i^{(k-1)} + \omega\frac{r_{ii}^{(k)}}{a_{ii}}, \quad i = 1, \ldots, n$$

se puede demostrar que para ciertas elecciones de ω positivo la convergencia del vector $\overrightarrow{X}^{(k)} = \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix}$ al vector solución de $A\overrightarrow{x} = \overrightarrow{b}$ es significativamente más rápida. Para fines de cálculo es conveniente expresar la relación :

$$x_i^{(k)} = x_i^{(k-1)} + w\frac{r_{ii}^{(k)}}{a_{ii}}$$

en forma equivalente a la sigiente:

$$x_i^{(k)} = (1-\omega)\, x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right]$$

o lo que es lo mismo

$$a_{ii}x_i^{(k)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j x_j^{(k)} = (1-\omega)\, a_{ii}x_i^{(k-1)} - \omega\sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} + \omega b_i,$$

que en forma matricial corresponde a

$$(D - \omega L)\,\overrightarrow{X}^{(k)} = [(1-\omega)\, D + \omega U]\,\overrightarrow{X}^{k-1} + \omega\,\overrightarrow{b},$$

de donde

$$\overrightarrow{X}^{(k)} = (D - \omega L)^{-1}\,[(1-\omega)\, D + \omega U]\,\overrightarrow{X}^{(k-1)} + \omega\,(D - \omega L)^{-1}\,\overrightarrow{b}.$$

Así llegamos a la forma familiar

$$\overrightarrow{X}^{(k)} = T\overrightarrow{X}^{(k-1)} + \overrightarrow{c}.$$

Ejemplos

El sistema de ecuaciones: $\begin{cases} 4x_1 + 3x_2 & = 24 \\ 3x_1 + 4x_2 - x_3 & = 30 \\ -x_2 + 4x_3 & = -24, \end{cases}$ tiene la solución exacta $\overrightarrow{x} = \begin{pmatrix} 3 \\ 4 \\ -5 \end{pmatrix}$. Por el método de Gauss-Seidel se obtiene en la iteración 7 la solución aproximada $\overrightarrow{X}^{(7)} = \begin{pmatrix} 3{,}0134110 \\ 3{,}9888241 \\ -5{,}0027940 \end{pmatrix}$, mientras que con el método SOR se obtiene para $\omega = 1{,}25$, $\overrightarrow{X}^{(7)} = \begin{pmatrix} 3{,}000094 \\ 4{,}0002586 \\ -5{,}0003486 \end{pmatrix}$. Terminamos esta sección con algunos resultados que justifican las afirmaciones anteriores sobre el método SOR.

Iniciamos con algunas definiciones y notaciones previas. Recordemos que una matriz cuadrada A es definida positiva si $\overrightarrow{x}^t A\overrightarrow{x} > 0$ para todo vector columna $\overrightarrow{x} \neq \overrightarrow{0}$.

Los métodos de aproximación anteriores se resumen en una expresión de la forma

$$\overrightarrow{X}^{(k)} = T\overrightarrow{X}^{(k-1)} + \overrightarrow{c}.$$

Usaremos las notaciones T_J, T_G y T_W para indicar que la matriz T se refiere al método de Jacobi, Gauss-Seidel o SOR respectivamente.

Consideremos una matriz tridiagonal siguiente:

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & a_{n-1,n} \\ 0 & \dots & \dots & 0 & a_{n,n-1} & a_{nn} \end{pmatrix}$$

Se tienen los siguientes resultados.

1. Si A es una matriz definida positiva y $0 < \omega < 2$, la sucesión $\left(\overrightarrow{X}^{(k)}\right)$ del método SOR converge para cualquier elección de $\overrightarrow{X}^{(0)}$.

Si, además, A es tridiagonal, entonces $\rho(T_G) = [\rho(T_J)]^2 < 1$ y la elección óptima para w es

$$w = \frac{2}{1 + \sqrt{1 - [\rho(T_J)]^2}}.$$

Con este valor de w, $\rho(T_W) = w - 1$.

Ejercicio resuelto

a) Sea A una matriz tridagonal de $n \times n$ estrictamente diagonalmente dominante, $\overrightarrow{b} \in \mathbb{R}^n$ y $1 < \omega < 2$. Elaborar un algoritmo para calcular la solución aproximada usando el método S.O.R.

b) Aplique su algoritmo para hallar la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, donde

$$A = \begin{bmatrix} 3 & 2 & 0 & 0 \\ 1 & 4 & 2 & 0 \\ 0 & 2 & 5 & 2 \\ 0 & 0 & 3 & 5 \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix},$$

1. $Tol = 10^{-2}$, $\omega = 1{,}5$ y $\overrightarrow{x}_0^T = (0,0,0,0)$.

   Solución

   a) Puesto que el método S.O.R. viene dado por

   $$(D - \omega L)\overrightarrow{x}^{(k+1)} = [(1-w)D + \omega U]\overrightarrow{x}^{(k)} + \omega\overrightarrow{b},$$

   donde $A = -L + D - U$ es la descomposición habitual antes indicada. Como A es tridiagonal, tenemos

   $$L = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ -a_{21} & 0 & 0 & \dots & 0 \\ 0 & -a_{32} & 0 & \dots & 0 \\ \vdots & & c & \ddots & \vdots \\ 0 & \dots & \dots & -a_{n,n-1} & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & -a_{12} & 0 & \dots & 0 \\ 0 & 0 & -a_{23} & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & -a_{n-1,n} \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}$$

   $$D - \omega L = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ \omega a_{21} & a_{22} & & \dots & 0 \\ 0 & \omega a_{32} & a_{33} & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & \omega a_{n,n-1} & a_{nn} \end{bmatrix},$$

$$(D-\omega L)\,\overrightarrow{x}^{(k+1)} = \begin{bmatrix} a_{11}x_1^{(k+1)} \\ \omega a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} \\ \omega a_{32}x_2^{(k+1)} + a_{33}x_3^{(k+1)} \\ \vdots \\ \omega a_{n,n-1}x_{n-1}^{(k+1)} + a_{nn}x_n^{(k+1)} \end{bmatrix}$$

$$(1-\omega)\,D+\omega U = \begin{bmatrix} (1-\omega)\,a_{11} & -wa_{12} & 0 & \dots & 0 \\ 0 & (1-\omega)\,a_{22} & -wa_{23} & \dots & 0 \\ 0 & 0 & (1-\omega)\,a_{33} & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & \dots & (1-\omega)\,a_{nn} \end{bmatrix}$$

$$[(1-\omega)\,D+\omega U]\,\overrightarrow{x}^{(k)} = \begin{bmatrix} (1-\omega)\,a_{11}x_1^{(k)} - \omega a_{12}x_2^{(k)} \\ (1-\omega)\,a_{22}x_2^{(k)} - \omega a_{23}x_3^{(k)} \\ (1-\omega)\,a_{33}x_3^{(k)} - \omega a_{34}x_4^{(k)} \\ \vdots \\ (1-\omega)\,a_{n-1,n-1}x_{n-1}^{(k)} - \omega a_{n-1,n}x_n^{(k)} \\ (1-\omega)\,a_{nn}x_n^{(k)} \end{bmatrix}$$

Se tiene

$$\begin{bmatrix} a_{11}x_1^{(k+1)} \\ \omega a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} \\ \omega a_{32}x_2^{(k+1)} + a_{33}x_3^{(k+1)} \\ \vdots \\ \omega a_{n,n-1}x_{n-1}^{(k+1)} + a_{nn}x_n^{(k+1)} \end{bmatrix} = \begin{bmatrix} (1-\omega)\,a_{11}x_1^{(k)} - \omega a_{12}x_2^{(k)} \\ (1-\omega)\,a_{22}x_2^{(k)} - \omega a_{23}x_3^{(k)} \\ (1-\omega)\,a_{33}x_3^{(k)} - \omega a_{34}x_4^{(k)} \\ \vdots \\ (1-\omega)\,a_{n-1,n-1}x_{n-1}^{(k)} - \omega a_{n-1,n}x_n^{(k)} \\ (1-\omega)\,a_{nn}x_n^{(k)} \end{bmatrix} + \begin{bmatrix} \omega b_1 \\ \omega b_2 \\ \omega b_3 \\ \vdots \\ \vdots \\ \omega b_n \end{bmatrix},$$

de donde obtenemos el siguiente esquema numérico

$$\begin{aligned} x_1^{(k+1)} &= \frac{(1-w)\,a_{11}x_1^{(k)} - \omega a_{12}x_2^{(k)} + \omega b_1}{a_{11}} \\ x_2^{(k+1)} &= \frac{-\omega a_{21}x_1^{(k+1)} + (1-w)\,a_{22}x_2^{(k)} - \omega a_{23}x_3^{(k)} + \omega b_2}{a_{22}} \\ x_3^{(k+1)} &= \frac{-\omega a_{32}x_2^{(k+1)} + (1-w)\,a_{33}x_3^{(k)} - \omega a_{34}x_4^{(k)} + \omega b_3}{a_{33}} \\ &\vdots \\ x_{n-1}^{(k+1)} &= \frac{-\omega a_{n-1,n-2}x_{n-2}^{(k+1)} + (1-w)\,a_{n-1,n-1}x_{n-1}^{(k)} - \omega a_{n-1,n}x_n^{(k)} + \omega b_{n-1}}{a_{n-1,n-1}} \\ x_n^{(k+1)} &= \frac{-\omega a_{n.n-1}x_{n-1}^{(k+1)} + (1-w)\,a_{nn}x_n^{(k)} + \omega b_n}{a_{nn}} \\ k &= 0, 1, \dots . \end{aligned}$$

La solución "exacta" es: $\overrightarrow{x}^T = (0{,}446154, -0{,}6692307,\ 0{,}615384,\ -0{,}369223)$.

Ponemos $\overrightarrow{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$. Primera iteración, aplicando el esquema numérico, obtenemos

$$\overrightarrow{x}^{(1)} = \begin{bmatrix} 0 \\ -0{,}375 \\ 0{,}525 \\ -0{,}4725 \end{bmatrix}, \quad \left\| \overrightarrow{x}^{(1)} - \overrightarrow{x}^{(0)} \right\|_{\infty} > Tol,$$

continuamos con la segunda iteración, volvemos a plicar el esquema numérico, tenemos

$$\overrightarrow{x}^{(2)} = \begin{bmatrix} 0{,}375 \\ -0{,}721875 \\ 0{,}754125 \\ -0{,}4424625 \end{bmatrix}, \quad \left\| \overrightarrow{x}^{(2)} - \overrightarrow{x}^{(1)} \right\|_{\infty} > Tol,$$

a continuación realizamos la tercera iteración. obtenemos

$$\overrightarrow{x}^{(3)} = \begin{bmatrix} 0{,}534375 \\ -0{,}780046875 \\ 0{,}656413125 \\ -0{,}3695675625 \end{bmatrix}, \quad \left\| \overrightarrow{x}^{(3)} - \overrightarrow{x}^{(2)} \right\|_{\infty} > Tol,$$

luego

$$\overrightarrow{x}^{(4)} = \begin{bmatrix} 0{,}512859375 \\ -0{,}669631172 \\ 0{,}595312678 \\ -0{,}350997629 \end{bmatrix}, \quad \left\| \overrightarrow{x}^{(4)} - \overrightarrow{x}^{(3)} \right\|_{\infty} > Tol,$$

$$\overrightarrow{x}^{(5)} = \begin{bmatrix} 0{,}41320148 \\ -0{,}64161947 \\ -0{,}597913926 \\ -0{,}362623719 \end{bmatrix}, \quad \left\| \overrightarrow{x}^{(5)} - \overrightarrow{x}^{(4)} \right\|_{\infty} > Tol.$$

Continuado con la ejecución del esquema numérico, en la iteración 7 se verifica que $\left\| \overrightarrow{x}^{(7)} - \overrightarrow{x}^{(6)} \right\|_{\infty} < Tol$, y concluimos con el procedimiento de cálculo.

7.5. Ejercicios

1. Sea Ω un intervalo de $\mathbb{R}$, f una función de Ω en $M_{2\times 2}[\mathbb{R}]$ definida como $f(x) = \begin{bmatrix} x^2 & 1-x \\ x & 2x^2+3 \end{bmatrix}$ $x \in \Omega$. Pruebe que f es Fréchet diferenciable en $a \in \Omega$ y la diferencial de Fréchet T_a está definida como $T_a(h) = \begin{bmatrix} 2ah & -h \\ h & 4ah \end{bmatrix}$ $\forall h \in \mathbb{R}$.

2. Sea Ω un intervalo de $\mathbb{R}$, f una función de Ω en $M_{2\times 2}[\mathbb{R}]$ con $f(x) = (a_{ij}(x))$ $x \in \Omega$ y a_{ij} funciones reales derivables en todo punto $a \in \Omega$. Demuestre que f es Fréchet diferenciable y que la diferencial T_a está definida como $T_a(h) = (a_{ij}(a)\, h)$ $\forall h \in \mathbb{R}$.

3. Considere la función real definida en todo $\mathbb{R}^2$ como $f(x,y) = \begin{cases} \dfrac{x^2+y^2}{x^2-y^2}, \text{ si } x \neq y \\ 0, \text{ si } x = y. \end{cases}$. demuestre que f no es diferenciable en $(0,0)$.

4. Considere la función real definida en todo $\mathbb{R}^2$ como $f(x,y) = \begin{cases} \dfrac{x-y}{x^2+y^2}, \text{ si } (x,y) \neq (0,0) \\ 0, \text{ si } (x,y) = (0,0). \end{cases}$. demuestre que f no es diferenciable en $(0,0)$.

5. Demuestre que la función v de $\mathbb{R}$ en $\mathbb{R}^3$ definida como $v(t) = \left(|t-1|, t, 2t^2\right) \quad t \in \mathbb{R}$ no es diferenciable en $t = 1$.

6. Sean $A \in M_{n\times n}[\mathbb{R}]$ una matriz simétrica, definida positiva, $\overrightarrow{b} \in \mathbb{R}^n$. Se define el funcional J sobre $\mathbb{R}^n$ como sigue: $j(x) = \frac{1}{2}\overrightarrow{x}^T A\overrightarrow{x} + \overrightarrow{b}^T\overrightarrow{x} \quad \forall \overrightarrow{x} \in \mathbb{R}^n$. Demuestre que J es Fréchet diferenciable y la diferencial de Fréchet está definida como $\langle DJ(\overrightarrow{x}), \overrightarrow{v}\rangle = \left\langle A\overrightarrow{x} + \overrightarrow{b}, \overrightarrow{v}\right\rangle \quad \forall \overrightarrow{v} \in \mathbb{R}^n$.

7. El espacio de matrices $M_{2\times 2}[\mathbb{R}]$ está provisto de la norma submultiplicativa $\|\cdot\|$. En cada item se define una función F de $M_{2\times 2}[\mathbb{R}]$ en si mismo. Pruebe que F es Fréchet diferenciable y determine la diferencial de Fréchet $T_A \in L(M_{2\times 2}[\mathbb{R}], M_{2\times 2}[\mathbb{R}])$, con $A \in M_{2\times 2}[\mathbb{R}]$.

   a) $F(X) = \frac{1}{2}\left(X + X^T\right) \quad X \in M_{2\times 2}[\mathbb{R}]$.

   b) $F(X) = \frac{1}{2}\left(X - X^T\right) \quad X \in M_{2\times 2}[\mathbb{R}]$.

   c) $F(X) = B^{-1}XB \quad X, \quad B \in M_{2\times 2}[\mathbb{R}]$ con B invertible fija.

   d) $F(X) = X^TBX \quad X, B \in M_{2\times 2}[\mathbb{R}]$ con B fija.

   e) $F(X) = X^TX + BX \quad X, B \in M_{2\times 2}[\mathbb{R}]$ con B fija.

8. Sean Ω abierto del espacio normado V y F una función de Ω en el espacio normado W. Pruebe que si F es Fréchet diferenciable, la diferencial $T_a \in L(V, W)$ de F es única.

9. Sean Ω abierto del espacio normado V, F y G funciones de Ω en W diferenciables en $a \in \Omega$. Demostrar $F + G$ diferenciable en a y $D(F+G)(a) = DF(a) + DG(a)$

10. Sea V un espacio provisto del producto escalar $< \cdot, \cdot >$ y norma asociada $\|\cdot\|$. Sea F el funcional definido como$F(x) = \| x \|^2 \quad \forall x \in V$. Demuestre que F es Fréchet diferenciable.

11. Sean $\Omega \subset V$ abierto, con V un espacio provisto del producto escalar $< \cdot, \cdot >$ y norma asociada $\|\cdot\|$. En cada item se define un funcional. Pruebe que es diferenciaciable en Ω.

    a) $F(x) =< x, y > \quad x, y \in V \quad$ con y fijo.

    b) $F(x) =< G(x), y > \quad x, y \in V \quad$ con y fijo, G función de Ω en V, diferenciable en Ω.

    c) $F(x) =\| G(x) \|^2 \quad x \in \Omega$, G función de Ω en V, diferenciable en Ω.

12. Sea V un espacio normado de dimensión finita con $\|\cdot\|$ su norma y $T \in \mathcal{L}(V)$ tal que $\|T\| < 1$. Sea $g : V \to V$ una aplicación definida por $g(x) = T(x) + c$ con $c \in V$ fijo. Demostrar que la sucesión (x_n) definida por

$$\begin{cases} x_0 \in V \\ x_{n+1} = g(x_n) \quad n = 0, 1, \ldots \end{cases}$$

    converge a un punto fijo $\widehat{x}$ de g. [Sugerencia: Pruebe que g es lipschisiana].

13. Sea (E, d) un espacio métrico, $g : E \to E$ una aplicación continua que posee un punto fijo u. Sean $r > 0$ y $0 < k < 1$ tales que $d(g(x), g(y)) \leq kd(x, y) \quad \forall x, y \in B(u, r)$. Demostrar que para todo $x_0 \in B(u, r)$, la sucesión (x_n) definida por $x_{n+1} = g(x_n) \quad n = 0, 1, \ldots$ es tal que $(x_n) \subset B(u, r)$ y $\lim\limits_{n\to\infty} x_n = u$.

    [Sugerencia: pruebe por inducción que $x_n \in B(u, r)$].

14. Sea $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ una función diferenciable con continuidad en todo punto de $\mathbb{R}^n$ tal que

$$\sup_{\vec{x}\in\mathbb{R}^n} \|D\Phi(\vec{x})\| \leq k < 1.$$

    Demuestre que Φ es contractiva y que para todo $\vec{x}_0 \in \mathbb{R}^n$, la sucesión $(\vec{x}_m)$ generada por $\vec{x}_{m+1} = \Phi(\vec{x}_m) \quad m = 0, 1, \ldots$ converge a $\widehat{x}$ punto fijo de Φ.

    Pruebe que $\|\vec{x}_{m+1} - \widehat{x}\| \leq k\|\vec{x}_m - \widehat{x}\| \leq k^{m+1}\|\vec{x}_0 - \widehat{x}\|, \quad m = 1, 2, \ldots$.

15. En los ejercicios siguientes, calcular, si existen, las soluciones de los sistemas de ecuaciones no lineales que se indican utilizando el método de Newton. [Sugerencia: cada sistema está formado por ecuaciones de hipérbolas, parábolas, elipses, circunferencias, identifíquelas].

a) $\begin{cases} 2x^2 - y = 1 \\ x^2 - y^2 = 1. \end{cases}$ **b)** $\begin{cases} 2x - y^2 = 3 \\ 4x^2 + \dfrac{y^2}{9} = 1. \end{cases}$ **c)** $\begin{cases} 4x^2 - y^2 = 1 \\ x^2 + y = 2. \end{cases}$

d) $\begin{cases} (x-3)^2 + (y-2)^2 = 9 \\ \dfrac{(x-1)^2}{4} + \dfrac{(y-1)^2}{9} = 1. \end{cases}$ **e)** $\begin{cases} x^2 + (y-2)^2 = 16 \\ \dfrac{(x+1)^2}{2} - \dfrac{(y-1)^2}{5} = 1. \end{cases}$

16. En los ejercicios siguientes, calcular, si existen, las soluciones de los sistemas de ecuaciones no lineales que se indican utilizando el método de Newton. Para el efecto defina una función vectorial asociada al sistema de ecuaciones no lineales, calcule la matriz Jacobiana, y seleccione un vector $\overrightarrow{x}_0$ y mediante la aplicación del método de Newton, calcule una aproximación $\overrightarrow{x}_1$, continue con el procedimiento. Calcule $\| \overrightarrow{x}_{k+1} - \overrightarrow{x}_k \|$ $k = 0, 1, 2, 3$, y analice los resultados.

a) $\begin{cases} 2x^2 - y + z^2 = 1 \\ 3x + 2y - 5z = -1 \\ x - y^2 + 2z^2 = 2. \end{cases}$ **b)** $\begin{cases} 2x - y^2 + 2z = 3 \\ 2x - y - z^2 = -3 \\ 4x^2 + \dfrac{y^2}{9} + z^2 = 5. \end{cases}$ **c)** $\begin{cases} 4x^2 - y^2 - z^2 = 9 \\ 2x - y + 3z = 5 \\ x^2 + y^2 + z^2 = 16. \end{cases}$

d) $\begin{cases} 2x + 5y + z = 10 \\ 2x - y^2 - z = -3 \\ x^2 - y^2 + 3z^2 = 5. \end{cases}$ **e)** $\begin{cases} 2x^2 - 5y + 2z = 10 \\ 2x^2 - y - z^2 = -3 \\ x^2 - \dfrac{y^2}{9} + 3z^2 = 5. \end{cases}$

17. Considerar el sistema de ecuaciones no lineales

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} x_1^3 \\ x_2^3 \\ x_3^3 \\ x_4^3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}. \tag{P}$$

a) Aplique el método de Newton con una aproximación inicial $\vec{X}_0 = (0, 0, 0, 0)^T$ y el métodod de factorización LU para aproximar la solución $\widehat{X} = (x_1, x_2, x_3, x_4)^T$ de dicho sistema con una precisión de 10^{-2}.

b) Generalice el problema (P) para $n \geq 3$ y elabore un algoritmo numérico para aproximar la solución de dicho problema.

18. Considerar la matriz $A = \begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}$.

a) Pruebe que A es una matriz definida positiva.

b) Calcule $\rho(T_J)$ para el sistema $A\overrightarrow{x} = \overrightarrow{b}$, con $\overrightarrow{b} = \begin{pmatrix} 24 \\ 30 \\ -24 \end{pmatrix}$.

c) Encuentre el valor óptimo de ω cuando se utiliza el método SOR para encontrar soluciones aproximadas del sistema $A\overrightarrow{x} = \overrightarrow{b}$.

19. En cada item se da un sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$. Aplicar el método de Gauss-Seidel para calcular solucione aproximadas Verifique que $\overrightarrow{x}$ es solución del sistema de ecuaciones propuesto. Contabilice el número de operaciones elementales que realiza.

a) $\begin{cases} x \quad\quad - 2z = 8 \\ 2x + y - \ z \ = 15 \\ 3x - y - 10z = 27, \end{cases}$ $\overrightarrow{x}^T = (4, 5, -2)$. **b)** $\begin{cases} 2x + 2y - 2z = 10 \\ 3x + 4y - 7z = 2 \\ 4x + 4y + 4z = 60, \end{cases}$ $\overrightarrow{x}^T = (3, 7, 5)$.

c) $\begin{cases} 3x - 3y + 6z = -18 \\ x + y + 8z = -2 \\ -x - 4z = 3, \end{cases}$ $\overrightarrow{x}^T = (1, 5, -1)$.

d) $\begin{cases} x + 2y + 3z + 4w = 34 \\ x + 4y + 7z + 10w = 62 \\ x + 4y + 10z + 16w = 74 \\ x + 4y + 10z + 20w = 74, \end{cases}$ $\overrightarrow{x}^T = (10, 6, 4, 0)$.

d) $\begin{cases} 4x + 20w = 48 \\ y - z + 3w = 5 \\ -2y + 5z - 12w = -22 \\ 3x + 3y - 2z + 21w = 44, \end{cases}$ $\overrightarrow{x}^T = (-3, -2, 2, 3)$.

f) $\begin{cases} -2x_1 + 2x_2 + 8x_3 - 6x_4 = 10 \\ -x_1 + 2x_2 + 5x_4 = 1 \\ 3x_2 + 16x_3 + 28x_4 = 16 \\ 5x_2 + 26x_3 + 44x_4 = 26 \\ 2x_1 - 2x_2 + 32x_3 + 45x_4 + 2x_5 = 33, \end{cases}$ $\overrightarrow{x}^T = (-1, 0, 1, 0, -1)$.

g) $\begin{cases} 2x_1 + 4x_2 - 6x_3 + 8x_4 + 16x_5 = 26 \\ 2x_1 + 5x_2 - 8x_3 + 8x_4 + 15x_5 = 24 \\ 2x_1 + 2x_2 - x_3 + 8x_4 + 20x_5 = 33 \\ 2x_1 + 7x_2 - 10x_3 + 11x_4 + 20x_5 = 32 \\ 2x_1 + 5x_2 + 5x_3 + 10x_4 + 24x_5 = 48, \end{cases}$ $\overrightarrow{x}^T = (2, 1, 1, 1, 1)$.

20. Aplicar el método SOR para aproximar la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$. Compare con el vector $\overrightarrow{x}$ que se propone. Contabilice el número de operaciones elementales que realiza.

a) $\begin{cases} x + 2y + 3z = 10 \\ 2x + 5y + 5z = 19 \\ 3x + 5y + 11z = 33, \end{cases}$ $\overrightarrow{x}^T = (2, 1, 2)$. **b)** $\begin{cases} 4x + 6y + 8z = -8 \\ 6x + 10y + 12z = -10 \\ 8x + 12y + 80z = -336, \end{cases}$ $\overrightarrow{x}^T = (5, 2, -5)$

c) $\begin{cases} x + y + z + w = 5 \\ x + 5y + 5z + 5w = 9 \\ x + 5y + 14z + 14w = 9 \\ x + 5y + 14z + 30w = 25, \end{cases}$ $\overrightarrow{x}^T = (4, 1, -1, 1)$.

d) $\begin{cases} 16x_1 + 12x_4 = -100 \\ x_2 - 2x_3 + 3x_4 = 15 \\ -2x_2 + 13x_3 - 3x_4 = -15 \\ 12x_1 + 3x_2 - 3x_3 + 20x_4 = -20, \end{cases}$ $\overrightarrow{x}^T = (-10, 0, 0, 5)$.

e) $\begin{cases} 4x_1 + 2x_2 - 4x_5 = 52 \\ 2x_1 + 2x_2 + 3x_3 + 5x_4 + 5x_5 = 52 \\ 3x_2 + 25x_3 + 39x_4 + 53x_5 = 135 \\ 5x_2 + 39x_3 + 65x_4 + 85x_5 = 217 \\ -4x_1 + 5x_2 + 53x_3 + 85x_4 + 122x_5 = 231, \end{cases}$ $\overrightarrow{x}^T = (12, 7, 3, 1, 0)$.

f) $\begin{cases} 4x_1 + 4x_2 + 2x_3 + 4x_4 + 4x_5 + 2x_6 = 0 \\ 4x_1 + 5x_2 + 7x_4 + 5x_5 + 3x_6 = -3 \\ x_1 + 14x_3 - 4x_4 - x_6 = 5 \\ 4x_1 + 7x_2 - 4x_3 + 22x_4 + 11x_5 + 5x_6 = -14 \\ 2x_1 + 5x_2 + 11x_4 + 13x_5 + 5x_6 = -1 \\ 2x_1 + 3x_2 - x_3 + 5x_4 + 5x_5 + 28x_6 = -1, \end{cases}$ $\overrightarrow{x}^T = (1, -1, 0, -1, 1, 0)$.

21. En cada item se propone un sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$. Estudie a la matriz A del sistema para determinar si es estrictamente diagonalmente dominante, simétrica, definida positiva, monótona, etc. Aplique los métodos de Gauss-Seidel y SOR, y, con cada uno de ellos halle la solución aproximada del sistema de ecuaciones. Contabilice el número de operaciones elementales que realiza.

a) $\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 3 \\ 2 \end{bmatrix}$, $\overrightarrow{x}^T = (5,8,8,5)$.

b) $\begin{bmatrix} 4 & 1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & 1 & 4 & -1 \\ 0 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 11 \\ 6 \\ 17 \\ 12 \end{bmatrix}$, $\overrightarrow{x}^T = (2,3,4,2)$.

c) $\begin{bmatrix} 5 & 1 & 1 & 0 & 0 \\ -1 & 5 & 1 & -1 & 0 \\ -1 & 1 & 6 & 1 & 1 \\ 0 & 0 & -1 & 6 & 1 \\ 0 & 0 & 0 & 1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} == \begin{bmatrix} 16 \\ 14 \\ 22 \\ 10 \\ 8 \end{bmatrix}$, $\overrightarrow{x}^T = (2,3,3,2,1)$

d) $\begin{bmatrix} 3 & -1 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 11 \\ 4 \\ 1 \\ 4 \\ 11 \end{bmatrix}$, $\overrightarrow{x}^T = (5,4,3,4,5)$.

e) $\begin{bmatrix} 5 & 2 & 1 & 0 & 0 \\ 2 & 5 & 1 & 0 & 0 \\ 1 & 1 & 5 & 2 & 1 \\ 0 & 0 & 2 & 5 & 1 \\ 0 & 0 & 0 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -12 \\ -9 \\ 1 \\ 7 \\ 11 \end{bmatrix}$, $\overrightarrow{x}^T = (-2,-1,0,1,2)$.

f) $\begin{bmatrix} 4 & -1 & -\frac{1}{2} & 0 & 0 & 0 \\ -1 & 5 & -1 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & -1 & 6 & -1 & -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & -1 & 7 & -1 & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & -1 & 8 & -1 \\ 0 & 0 & 0 & -\frac{1}{2} & -1 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 29 \\ 21 \\ 13 \\ 19 \\ 45 \\ 79 \end{bmatrix}$, $\overrightarrow{x}^T = (10,8,6,6,8,10)$.

22. Sea $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ que satisface las dos condiciones siguientes: $a_{ij} = 0$ si $|i-j| > 2$ para $i,j = 1,\ldots,n$ y que $a_{ii} > |a_{i\,i-2}| + |a_{i\,i-1}| + |a_{i\,i+1}| + |a_{i\,i+2}|$, $\quad i = 1,\ldots,n$, $\overrightarrow{b} \in \mathbb{R}^n$.

a) Demuestre que el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ tiene una única solución.

b) Aplique los métodos de Jacobi, Gauss-Seidel y SOR y exprese la sucesion $(\overrightarrow{x}_k)$ en forma explícita para cada método y elabore un algoritmo que permita calcular la solución aproximada del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

c) Considere el sistema de ecuaciones lineales siguiente:

$$\begin{bmatrix} 4 & -1 & -1 & 0 & 0 \\ -1 & 5 & -1 & -1 & 0 \\ -1 & -1 & 5 & -1 & -1 \\ 0 & -1 & -1 & 4 & -1 \\ 0 & 0 & -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -5 \\ 10 \\ 25 \\ 4 \\ -4 \end{bmatrix}$$

Verifique las hipótesis de la matriz A. Aplique sus algoritmos para hallar la solución de dicho sistema y compare con $\overrightarrow{x}^T = (2,5,8,5,3)$.

23. Considere el sistema de ecuaciones lineales siguiente:

$$\begin{bmatrix} 4 & 2 & -2 & 0 & 0 \\ 2 & 10 & 2 & -3 & 0 \\ -2 & 2 & 18 & 3 & -4 \\ 0 & -3 & 3 & 18 & 3 \\ 0 & 0 & -4 & 3 & 18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -6 \\ 21 \\ 96 \\ 66 \\ 7 \end{bmatrix}.$$

a) Demuestre que la matriz A de este sistema es simétrica, definida positiva y estrictamente diagonalmente dominante.

b) Aplique los métodos de Jacobi, Gauss-Seidel y SOR para hallar la solución aproximada de tal sistema. Contabilice con cada método el número de operaciones elementales. Compare la solución con $\overrightarrow{x}^T = (0, 2, 5, 3, 1)$.

7.6. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Análisis Matemático, Segunda Edición, Editorial Reverté, Barcelona, 1982.
2. Tom M. Apostol, Calculus, Volumen 2, Segunda Edición, Editorial Reverté, Barcelona, 1975.
3. Owe Axelsson, Iterative Solution Methods, Editorial Cambridge University Press, Cambridge, 1996.
4. N. Bakhvalov, Métodos Numéricos, Editorial Paraninfo, Madrid, 1980.
5. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.
6. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
7. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
8. P. G. Ciarlet, Introduction á l'Analyse Numérique Matricielle et á l'Optimisation, Editorial Masson, París, 1990.
9. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
10. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.
11. James W. Demmel, Applied Numerical Linear Algebra, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.
12. J. E. Dennis, Jr., Robert B. Schnabel, Numerical Methods for Unconslrained Optimization and Nonlinear Equations, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1996.
13. V. N. Faddeva, Métodos de Cálculo de Algebra Lineal, Editorial Paraninfo, Madrid, 1967.
14. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.
15. A. Kurosh, Cours D'Algèbre Supérieure, Editions Mir, Moscou, 1973.
16. Noel Gastinel, Análisis Numérico Lineal, Editorial Reverté, S. A., Barcelona, 1975.
17. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.
18. Anne Greenbaum, Iterative Methods for Solving Linear Systems, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.
19. Wolfgang Hackbusch, Iterative Solution of Large Sparse Systems of Equations, Editorial Springer-Verlag, New York, 1994.
20. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

21. Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Editorial Society for Industrial and Applied Mathematics, Philadelphia, 1996.

22. Franz E. Hohn, Algebra de Matrices, Editorial Trillas, México, 1979.

23. Roger A. Horn, Charles R. Johnson, Matrix Analysis, Editorial Cambridge University Press, Cambrisge, 1999.

24. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

25. P. Lascaux, R. Théodor, Analyse Numérique Matricielle Appliquée à L' Art de L' Ingénieur, Tome 1, Editorial Masson, París, 1986.

26. P. Lascaux, R. Théodor, Analyse Numérique Matricielle Appliquée à L' Art de L' Ingénieur, Tome 2, Editorial Masson, París, 1987.

27. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

28. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

29. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

30. Ben Noble, James W. Daniel, Algebra Lineal Aplicada, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1989.

31. J. M. Ortega, W. C. Rheinbolodt, Iterative Solution of Nonlinear Equatios in Several Variables, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.

32. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

33. Werner C. Rheinboldt, Methods for Solving Systems of Nonlinear Equations, Second Edition, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

34. M. Sibony, J. Cl. Mardon, Analyse Numérique I, Sustèmes Linéaires et non Linéaires, Editorial Hermann, París, 1984.

35. G. W. Stewart, Matrix Algotithms, Volume II: Eingensystems, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

36. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

37. Gilbert Strang, Algebra Lineal y sus Aplicaciones, editorial Fondo Educativo Interamericano, México, 1982.

38. V. Voïévodine, Principes Numériques D' Algèbre Linéaire, Editions Mir, Moscú, 1976.

39. David S. Watkins, Fundamentals of Matrix Computations, Editorial John Wiley&Sons, New York, 1991.

Capítulo 8

Valores y Vectores Propios

Resumen

Este capítulo se inicia con una revisión de resultados importantes sobre el problema de valores y vectores propios. La primera aplicación que se da es a la geometría analítica plana, más precisamente a las formas cuadráticas y ecuaciones cuadráticas. Luego se considera el cálculo de valores y vectores propios de matrices de 3×3 que se presenta con mucha frecuencia, sobre todo en los problemas de optimización de funciones reales en tres variables independientes. Se concluye este capítulo con el método de la potencia.

8.1. Introducción

En lo sucesivo consideraremos matrices reales de $n \times n$ aunque algunos resultados aparezcan dentro del campo de los números complejos. Nos interesamos fundamentalmente en el caso real.

Definición 1 *Sea $A \in M_{n\times n}[\mathbb{R}]$. Un escalar $\lambda \in \mathbb{R}$ se denomina valor propio de la matriz A si y solo si existe $\overrightarrow{x} \in \mathbb{R}^n$ no nulo tal que $A\overrightarrow{x} = \lambda\overrightarrow{x}$. El vector $\overrightarrow{x}$ se llama vector propio de A asociado al valor propio λ.*

Los términos de valor y vector propio que aquí hemos definido, en muchos textos se los encuentran como valor y vector característico, eigenvalor y eigenvector, autovalor y autovector.

Un valor proprio λ puede ser cero, esto es, $\lambda = 0$ pero un vector propio $\overrightarrow{x}$ no puede ser $\overrightarrow{0}$. En el caso en que $\lambda = 0$, tenemos $A\overrightarrow{x} = \overrightarrow{0}$ para algún $\overrightarrow{x} \in \mathbb{R}^n$ con $\overrightarrow{x} \neq \overrightarrow{0}$ lo que significa que $\overrightarrow{x} \in \ker(A)$, donde $\ker(A)$ denota el núcleo de la matriz A que se define como $\ker(A) = \{\overrightarrow{x} \in \mathbb{R}^n \mid A\overrightarrow{x} = \overrightarrow{0}\}$.

Sea λ un valor propio de A, denotamos con $S_\lambda = \{\overrightarrow{x} \in \mathbb{R}^n \mid A\overrightarrow{x} = \lambda\overrightarrow{x}\} \cup \{\overrightarrow{0}\}$. Se prueba inmediatamente que el conjunto S_λ es un subespacio real que lo denominamos subespacio asociado al valor propio λ o simplemente subespacio propio de λ.

Sea λ un valor propio de A y $\overrightarrow{x} \in \mathbb{R}^n$ un vector propio asociado a λ. De la definición de valor y vector propio de A se tiene

$$A\overrightarrow{x} = \lambda\overrightarrow{x} \Longleftrightarrow (A - \lambda I)\overrightarrow{x} = \overrightarrow{0}.$$

El sistema de ecuaciones homogéneo $(A - \lambda I)\overrightarrow{x} = \overrightarrow{0}$ tiene soluciones no triviales ($\overrightarrow{x} \neq \overrightarrow{0}$) si y solo si la matrix $A - \lambda I$ es singular, que a su vez es equivalente a que $\det(A - \lambda I) = 0$. Esta última ecuación se llama ecuación característica. Se define

$$p(\lambda) = \det(A - \lambda I) \quad \lambda \in \mathbb{R},$$

y se le denomina polinomio característico.

El conjunto de todos los valores propios de la matriz A se denota $\sigma(A)$ y se le denomina espectro de la matriz A, esto es,

$$\sigma(A) = \{\lambda \in \mathbb{R} \mid \exists \overrightarrow{x} \in \mathbb{R}^n \backslash \{\overrightarrow{0}\} \text{ tal que }, \ A\overrightarrow{x} = \lambda \overrightarrow{x}\}.$$

Ejemplos

1. Los espacios $M_{1\times 2}[\mathbb{R}]$ y $\mathbb{R}^2$ son isomorfos. Sea $A = (1,2)$, calculemos $|||A|||_2$ (véase el apéndice, normas de matrices). Tenemos $A \in M_{1\times 2}[\mathbb{R}]$ y $A \in \mathbb{R}^2$, en este último caso escribimos $\overrightarrow{A} = (1,2)$. Resulta $\left\|\overrightarrow{A}\right\|_2 = \sqrt{1^2+2} = \sqrt{5}$. Por otro lado,

$$A^tA = \begin{bmatrix} 1 \\ 2 \end{bmatrix}(1,2) = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

El polinomio característico de A^TA está definido como $p(\lambda) = \begin{vmatrix} 1-\lambda & 2 \\ 2 & 4-\lambda \end{vmatrix} = \lambda(\lambda-5)$, luego $p(\lambda) = 0 \Longleftrightarrow \lambda = 0$ y $\lambda = 5$; $\lambda = 0$ y $\lambda = 5$ son los valores propios de A^TA. Resulta

$$|||A|||_2 = \sup_{\|\overrightarrow{x}\|_2 \leq 1} \|A\overrightarrow{x}\|_2 = (\text{máx}\{0,5\})^{\frac{1}{2}} = \sqrt{5}.$$

2. Sea $A = \begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$. Calculemos $|||A|||_2$. Para el efecto, calculelos A^TA y luego hallamos los valores propios de esta matriz. Tenemos

$$A^TA = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}.$$

El polinomio característico $p(\lambda)$ de la matriz A^TA está definido como

$$p(\lambda) = \det\left(A^TA - \lambda I\right) = \begin{vmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{vmatrix} = \lambda^2 - 7\lambda + 6 = (\lambda-1)(\lambda-6).$$

Entonces $p(\lambda) = 0 \Longleftrightarrow \lambda = 1$ o $\lambda = 6$. Los valores propios son $\lambda = 1$, y $\lambda = 6$. Por lo tanto $|||A|||_2 = \sqrt{6}$.

3. Sea A una matriz real de $n \times n$ triangular superior (respectivamente triangular inferior), digamos $A = (a_{ij})$ con $a_{ij} = 0$ si $i > j$. Entonces

$$p(\lambda) = \det(A - \lambda I) = \prod_{i=1}^{n}(a_{ii} - \lambda) \quad \lambda \in \mathbb{R},$$

luego

$$p(\lambda) = 0 \Longleftrightarrow \lambda_1 = a_{11}, \cdots, \lambda_n = a_{nn}.$$

El sistema de ecuaciones $(A - \lambda_1 I)\overrightarrow{x} = \overrightarrow{0}$ se escribe en forma explícita como

$$\begin{cases} a_{12}x_1 + \cdots + a_{1n}x_n = 0, \\ (a_{22} - \lambda_1)x_2 + \cdots + a_{2n}x_n = 0, \\ \vdots \\ (a_{nn} - \lambda_1)x_n = 0, \end{cases}$$

y cualquier solución no nula de este sistema es un vector propio asociado al valor propio λ_1. Así sucesivamente

$$\begin{cases} (a_{11} - \lambda_n)x_1 + \cdots + a_{1n}x_n = 0, \\ (a_{22} - \lambda_n)x_2 + \cdots + a_{2n}x_n = 0, \\ \vdots \\ (a_{n-1n} - \lambda_n)x_n = 0. \end{cases}$$

Este sistema de ecuaciones lineales consta de $n-1$ ecuaciones, y cualquier solución no nula de este sistema es un vector propio asociado al valor propio λ_n.

Se tiene $\sigma(A) = \{a_{ii} \mid i = 1, \cdots, n\}$. Como se puede apreciar, estos son lo problemas más simples de cálculo de valores y vectores propios

Definición 2 *Sean $A, B \in M_{n\times n}[\mathbb{R}]$. Se dice que las matrices A, B son semejantes si y solo si existe una matriz invertible P tal que $B = P^{-1}AP$.*

Dos matrices que son semejantes tienen exactamente los mismos valores propios. Efectivamente, sea λ un valor propio de la matriz A y $\overrightarrow{x}$ un vector porpio asociado a λ, entonces $A\overrightarrow{x} = \lambda\overrightarrow{x}$, luego

$$\overrightarrow{0} = (A-\lambda I)\overrightarrow{x} = (PBP^{-1} - \lambda I)\overrightarrow{x} = P(B-\lambda I)P^{-1}\overrightarrow{x} = (B-\lambda I)\overrightarrow{y} = \overrightarrow{0},$$

con $\overrightarrow{y} = P^{-1}\overrightarrow{x}$. Así, $p(\lambda) = \det(A-\lambda I) = \det(B-\lambda I)$.

Sea $A \in M_{n\times n}[\mathbb{R}]$ tal que $A^T = A$, es decir que la matriz A es simétrica, entonces todos sus valores propios son reales. Además, si $\lambda \in \mathbb{R}$ es un valor propio de multiplicidad $2 \leq k \leq n$, entonces $\dim(S_\lambda) = k$. Por otro lado, si λ_1, λ_2 son valores propios de A tales que $\lambda_1 \neq \lambda_2$, los vectores propios asociados $\overrightarrow{x}_1$, $\overrightarrow{x}_2$ son ortogonales, es decir que si $A\overrightarrow{x}_1 = \lambda_1\overrightarrow{x}_2$, $A\overrightarrow{x}_2 = \lambda_2\overrightarrow{x}_2$, $\lambda_1 \neq \lambda_2 \Longrightarrow \overrightarrow{x}_1 \perp \overrightarrow{x}_2$.

Localización de los valores propios

Sea $A \in M_{n\times n}[\mathbb{R}]$ y $|||\cdot|||$ una norma submultipilicativa en $M_{n\times n}[\mathbb{R}]$. Entonces, si λ un valor propio de la matriz A y $\overrightarrow{x}$ un vector propio asociado a λ, entonces $A\overrightarrow{x} = \lambda\overrightarrow{x}$, y en consecuencia

$$\mid \lambda \mid\mid\mid \overrightarrow{x} \mid\mid\mid = \mid\mid\mid \lambda\overrightarrow{x} \mid\mid\mid = \mid\mid\mid A\overrightarrow{x} \mid\mid\mid \leq \mid\mid\mid A \mid\mid\mid \mid\mid\mid \overrightarrow{x} \mid\mid\mid,$$

y siendo $\overrightarrow{x}$ un vector porpio, se tiene $\mid\mid \overrightarrow{x} \mid\mid \neq 0$, y de esta desigualdad se sigue que $\mid \lambda \mid \leq \mid\mid\mid A \mid\mid\mid$.

Tenemos (véase el apéndice, normas de matrices),

$$\begin{aligned}
\mid \lambda \mid &\leq |||A|||_1 = \sup_{\|\overrightarrow{x}\|_1 \leq 1} \|A\overrightarrow{x}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|, \\
\mid \lambda \mid &\leq \sup_{\|\overrightarrow{x}\|_\infty \leq 1} \|A\overrightarrow{x}\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|, \\
\mid \lambda \mid &\leq |||A|||_2 = \sup_{\|\overrightarrow{x}\|_2 \leq 1} \|A\overrightarrow{x}\|_2 = \left(\max_{i=1,\ldots,n} |\lambda_i|\right)^{\frac{1}{2}},
\end{aligned}$$

donde $\lambda_1, \ldots \lambda_n$ son los valores propios de A^TA, y A^T denota la matriz transpuesta de A.

Ejemplos

1. Sea $A = \begin{bmatrix} 2 & -20 & 200 \\ 0 & 5 & 100 \\ 1 & -1 & 10 \end{bmatrix}$, entonces $\mid \lambda \mid \leq |||A|||_1 = 310$, $\mid \lambda \mid \leq |||A|||_\infty = 222$. Note que la estimación $\mid \lambda \mid \leq |||A|||_2$ requiere del cálculo de los valores propios de la matriz A^TA. Como la matriz A es real de 3×3, el polinomio característico $p(\lambda)$ es de grado 3 y tiene coeficientes reales, al menos un valor propio de A es real. De las estimaciones anteriores, se tiene $\lambda \in [-222, 222]$ que es un intervalo demasiado grande para localizar a esta raíz real. Este ejemplo muestra que se requieren de estimaciones más finas.

Teorema 1 *(de Gershgorin) Sea $A \in M_{n\times n}[\mathbb{R}]$. Los valores propios λ de la matriz A están localizados en la unión de los discos $\{\lambda \in \mathbb{C} \mid \mid \lambda - a_{kk} \mid \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj} \mid\}$. Estos discos se llaman discos de Gershgorin.*
Además, si la unión de k discosde Gershgorin son disjuntos unos de otros, entonces la unión contiene exactamente k valores propios de la matriz A.

Demostración. Sea λ un valor propio de la matriz A y $\overrightarrow{x}$ un vector porpio asociado a λ, entonces $A\overrightarrow{x} = \lambda\overrightarrow{x}$. Elegimos $\overrightarrow{x}$ tal $\| \overrightarrow{x} \|_\infty = x_k = 1$. Explícitamente, la fila k de $A\overrightarrow{x} = \lambda\overrightarrow{x}$ es la siguiente:

$$a_{k1}x_1 + \cdots + a_{kk}x_k + \cdots + a_{kn}x_n = \lambda x_k$$

de donde

$$\mid a_{kk} - \lambda \mid = \mid a_{k1}x_1 + \cdots + a_{kn}x_n \mid \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj} \mid\mid x_j \mid \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} \mid a_{kj} \mid$$

Para la prueba de la segunda parte se propone como ejercicio. ■

Ejemplo

Sea $A = \begin{bmatrix} 2 & -20 & 200 \\ 0 & 5 & 100 \\ 1 & -1 & 10 \end{bmatrix}$. Los radios de los discos de Gershgorin están definidos como sigue: $r_1 = \mid -20 \mid + 200 = 220$, $r_2 = 100$, $r_3 = 2$, luego los discos de Gershgorin son

$$B(2, 220) = \{\lambda \in \mathbb{C} \mid\mid \lambda - 2 \mid \leq 220\},$$

$$B(5, 100) = \{\lambda \in \mathbb{C} \mid\mid \lambda - 5 \mid \leq 100\},$$

$$B(10, 2) = \{\lambda \in \mathbb{C} \mid\mid \lambda - 10 \mid \leq 2\},$$

con lo $\lambda \in B(2, 220) \cup B(5, 100) \cup B(10, 2)$.

8.2. Formas cuadráticas y ecuaciones cuadráticas en $\mathbb{R}^2$.

Sean $a, b, c, d \in \mathbb{R}$ tales que $c \neq 0$ y $|a| + |b| > 0$. Consideramos el subconjunto C de $\mathbb{R}^2$ definido como

$$C = \left\{(x, y) \in \mathbb{R}^2 \mid ax^2 + by^2 + cxy = d\right\}.$$

Se trata de determinar si $C = \emptyset$ o $C \neq \emptyset$. En el caso $C \neq \emptyset$, determinar el tipo de conjunto que C representa, esto es, si es un punto, una recta, dos rectas, una cónica (circunferencia, elipse, parábola, hipérbola), encontrar la ecuación canónica de dicha cónica, o de la recta o de las rectas; y, representar gráficamente el conjunto C. Proponer un algoritmo.

Sigamos la metodología propuesta en la resolución de problemas.

Analicemos la existencia de soluciones, es decir determinemos si $C = \emptyset$ o $C \neq \emptyset$. Para el efecto definimos la forma cuadrática Q como sigue:

$$Q(x, y) = ax^2 + cxy + by^2 = (x, y) \begin{bmatrix} a & \frac{1}{2}c \\ \frac{1}{2}c & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (x, y) \in \mathbb{R}^2.$$

La matriz A de la forma cuadrática Q relativa a la base canónica $B_1 = \left\{\overrightarrow{i}, \overrightarrow{j}\right\}$ de $\mathbb{R}^2$ está definida como $A = \begin{bmatrix} a & \frac{1}{2}c \\ \frac{1}{2}c & b \end{bmatrix}$. La hipótesis $c \neq 0$ y $|a| + |b| > 0$ implica $A \neq 0$ y claramente A es simétrica, esto es, $A = A^T$.

Calculemos los valores propios de A, es decir, determinamos $\lambda \in \mathbb{R}$ tal que $\det(A - \lambda I) = 0$. Se tiene

$$\begin{aligned} \det(A-\lambda I) &= 0 \Longleftrightarrow \begin{vmatrix} a-\lambda & \frac{1}{2}c \\ \frac{1}{2}c & b-\lambda \end{vmatrix} = 0 \Longleftrightarrow (a-\lambda)(b-\lambda) - \frac{1}{4}c^2 = 0 \\ &\Longleftrightarrow \lambda^2 - (a+b)\lambda + ab - \frac{1}{4}c^2 = 0 \end{aligned}$$

es decir que los valores propios de A son soluciones de la ecuación de segundo grado:

$$t \in \mathbb{C} \text{ tal que } t^2 + \alpha t + \beta = 0,$$

donde $\alpha, \beta \in \mathbb{R}$, cuya solución se determina con la conocida fórmula

$$t = \frac{-\alpha \pm \sqrt{\alpha^2 - 4\beta}}{2}.$$

Sea $\gamma = \alpha^2 - 4\beta$. Si $\gamma \geq 0$, las raíces son reales; y si $\gamma < 0$, las raíces son complejas.

Puesto que la matriz A es simétrica, las raíces de la ecuación:

$$\lambda^2 - (a+b)\lambda + ab - \frac{1}{4}c^2 = 0$$

son reales. En efecto, el discriminante de esta ecuación es no negativo, pués

$$\gamma = (a+b)^2 - 4\left(ab - \frac{1}{4}c^2\right) = a^2 + 2ab + b^2 - 4ab + c^2 = (a-b)^2 + c^2.$$

Por hipótesis $c \neq 0$ y $|a| + |b| > 0$, que significa que al menos dos de estos números son no nulos, luego $\gamma = (a-b)^2 + c^2 \geq 0$. Entonces

$$\begin{aligned} \lambda_1 &= \frac{1}{2}(a+b) - \frac{1}{2}\sqrt{(a-b)+c^2}, \\ \lambda_2 &= \frac{1}{2}(a+b) + \frac{1}{2}\sqrt{(a-b)+c^2}, \end{aligned}$$

son los valores propios de la matriz A.

Determinemos los vectores propios asociados a λ_1 y λ_2, es decir, hallamos las soluciones de los sistemas de ecuaciones $A\overrightarrow{x} = \lambda_1 \overrightarrow{x}$ y $A\overrightarrow{x} = \lambda_2 \overrightarrow{x}$, que es equivalente al sistema de ecuaciones

$$(a - \lambda I)\overrightarrow{x} = 0 \Longleftrightarrow \begin{cases} (a-\lambda)x + \frac{1}{2}cy = 0 \\ \frac{1}{2}cx + (b-\lambda)y = 0. \end{cases}$$

Para $\lambda = \lambda_1$ obtenemo $\overrightarrow{u} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ tal que $\|\overrightarrow{u}\| = 1$ y para $\lambda = \lambda_2$ obtenemos $\overrightarrow{v} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$ con $\|\overrightarrow{v}\| = 1$. Estos dos vectores propios de A son ortogonales, esto es, $\overrightarrow{u} \cdot \overrightarrow{v} = 0$. Escribimos $\overrightarrow{u} \perp \overrightarrow{v}$. Consecuentemente, $\{\overrightarrow{u}, \overrightarrow{v}\}$ forman una base B_2 de $\mathbb{R}^2$. Ponemos $B_2 = \{\overrightarrow{u}, \overrightarrow{v}\}$.

Note que

$$\begin{aligned} Q(\overrightarrow{u}) &= \overrightarrow{u}^T A \overrightarrow{u} = \overrightarrow{u}^T(\lambda_1 \overrightarrow{u}) = \lambda_1 \|\overrightarrow{u}\|^2 = \lambda_1, \\ Q(\overrightarrow{v}) &= \overrightarrow{v}^T A \overrightarrow{v} = \overrightarrow{v}^T(\lambda_2 \overrightarrow{v}) = \lambda_2 \|\overrightarrow{v}\|^2 = \lambda_2, \end{aligned}$$

pués $\|\overrightarrow{u}\| = 1$ y $\|\overrightarrow{v}\| = 1$.

La forma bilineal simétrica F está definida como

$$F(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x}^T A \overrightarrow{y} \quad \forall \overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^2.$$

Se tiene $Q(\overrightarrow{x}) = F(\overrightarrow{x}, \overrightarrow{x}) \quad x \in \mathbb{R}^2$, y, $F(\overrightarrow{u}, \overrightarrow{u}) = \lambda_1$, $F(\overrightarrow{v}, \overrightarrow{v}) = \lambda_2$, $F(\overrightarrow{u}, \overrightarrow{v}) = F(\overrightarrow{v}, \overrightarrow{u}) = 0$. Así, la matriz de la aplicación bilineal simétrica F relativa a la base $B_2 = \{\overrightarrow{u}, \overrightarrow{v}\}$ está definida como

$$[F]_{B_2} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

a la que notamos D, esto es, $D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$.

La matriz de cambio de base de B_2 a B_1 está definida como $P = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix}$ y se verifica que $P^{-1}AP = D$.
La matriz P es ortogonal, esto es, $P^{-1} = P^T$.

La forma cuadrática Q referida a la base $B_2 = \{\overrightarrow{u}, \overrightarrow{v}\}$ se escribe como

$$Q(s,t) = (s,t) \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = \lambda_1 s^2 + \lambda_2 t^2 \quad (s,t) \in \mathbb{R}^2,$$

y

$$C = \left\{(x,y) \in \mathbb{R}^2 \mid ax^2 + cxy + by^2 = d\right\} = \left\{(s,t) \in \mathbb{R}^2 \mid \lambda_1 s^2 + \lambda_2 t^2 = d\right\}.$$

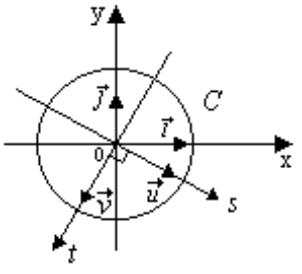
Puesto que $\lambda_1, \lambda_2 \in \mathbb{R}$, se presenta los siguientes casos: $\lambda_1, \lambda_2 \in \mathbb{R}^+$; $\lambda_1\lambda_2 \in \mathbb{R}^-$; $\lambda_1 \in \mathbb{R}^+$ y $\lambda_2 \in \mathbb{R}^-$; $\lambda_1 \in \mathbb{R}^-$ y $\lambda_2 \in \mathbb{R}^+$; $\lambda_1 = 0, \lambda_2 \neq 0$, $\lambda_1 \neq 0$, $\lambda_2 = 0$.

1. Supongamos $\lambda_1, \lambda_2 \in \mathbb{R}^+$.
   a) Si $d < 0$ entonces $C = \emptyset$.
   b) Si $d = 0$ entonces $C = \{(0,0)\}$.
   c) Si $d > 0$ entonces $C \neq \emptyset$.
   En el caso $\lambda_1 = \lambda_2 = \lambda$ entonces la ecuación

$$(s,t) \in \mathbb{R}^2 \text{ tal que } \lambda_1 s^2 + \lambda_2 t^2 = d \Longleftrightarrow (s,t) \in \mathbb{R}^2 \text{ tal que } s^2 + t^2 = \frac{d}{\lambda},$$

   corresponde a la ecuación de una circunferencia de centro $(0,0)$ y radio $r = \sqrt{\frac{d}{\lambda}}$.

   En la figura siguiente se muestra el conjunto C, o sea una circunferencia de radio $r = \sqrt{\frac{d}{\lambda}}$.



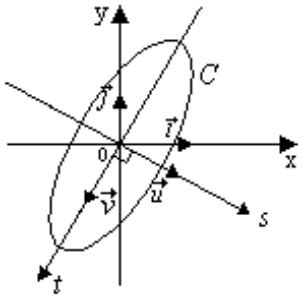
$\lambda_1 = \lambda_2 = \lambda$
$d > 0$

Figura 73

   En el caso $\lambda_1 \neq \lambda_2$, la ecuación

$$\begin{aligned} (s,t) &\in \mathbb{R}^2 \text{ tal que } \lambda_1 s^2 + \lambda_2 t^2 = d \Longleftrightarrow \\ (s,t) &\in \mathbb{R}^2 \text{ tal que } \left(\frac{s}{\sqrt{\frac{d}{\lambda_1}}}\right)^2 + \left(\frac{t}{\sqrt{\frac{d}{\lambda_2}}}\right)^2 = 1 \end{aligned}$$

que representa a una elipse de centro $(0,0)$. En la figura siguiente se muestra este conjunto



$\lambda_1 > \lambda_2 > 0$
$d > 0$

Figura 74

2. Supongamos $\lambda_1, \lambda_2 \in \mathbb{R}^-$.

   a) Si $d > 0$, resulta que $C = \emptyset$.

   b) Si $d = 0$, entonces $C = \{(0,0)\}$.

   c) Si $d < 0$, multiplicando por -1 a la ecuación: $(s,t) \in \mathbb{R}^2$ tal que $\lambda_1 s^2 + \lambda_2 t^2 = d$ se obtiene $(\lambda_1)\, s^2 + (-\lambda_2)\, t^2 = -d$ que ha sido analizado en el caso 1) parte c) precedente.

   Supongamos $\lambda_1 \in \mathbb{R}^+$ y $\lambda_2 \in \mathbb{R}^-$.

   a) Si $d = 0$, la ecuación

$$(s,t) \in \mathbb{R}^2 \text{ tal que } \lambda_1 s^2 + \lambda_2 t^2 = 0 \Longleftrightarrow (s,t) \in \mathbb{R}^2 \text{ tal que } t^2 = -\frac{\lambda_1}{\lambda_2} s^2.$$

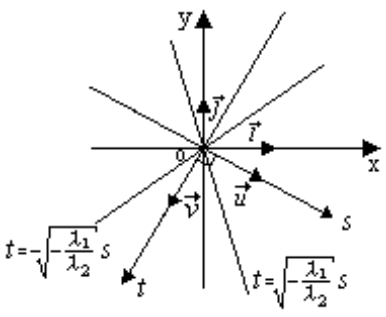
   De la última relación se obtienen las dos siguientes

$$(s,t) \in \mathbb{R}^2 \text{ tal que } \begin{cases} t = -\sqrt{-\dfrac{\lambda_1}{\lambda_2}}\, s, \\ t = \sqrt{-\dfrac{\lambda_1}{\lambda_2}}\, s. \end{cases}$$

   Consecuentemente

$$\begin{aligned} C &= \left\{(s,t) \in \mathbb{R}^2 \mid \lambda_1 s^2 + \lambda_2 t^2 = 0\right\} \\ &= \left\{(s,t) \in \mathbb{R}^2 \mid t = -\sqrt{-\frac{\lambda_1}{\lambda_2}}\, s\right\} \cup \left\{(s,t) \in \mathbb{R}^2 \mid t = \sqrt{-\frac{\lambda_1}{\lambda_2}}\, s\right\}. \end{aligned}$$

   En la figura siguiente se muestran los vectores $\overrightarrow{u}$, $\overrightarrow{v}$ y las rectas de ecuaciones $t = -\sqrt{-\dfrac{\lambda_1}{\lambda_2}}\, s$, $t = \sqrt{-\dfrac{\lambda_1}{\lambda_2}}\, s$ con $s \in \mathbb{R}$.



$\lambda_1 > 0,\ \lambda_2 < 0$
$\lambda_1 > |\lambda_2|$

Figura 75

b) Si $d \neq 0$, la ecuación

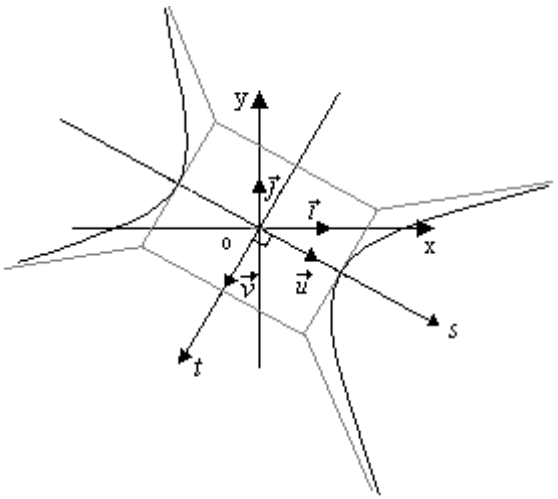
$$(s,t) \in \mathbb{R}^2 \text{ tal que } \lambda_1 s^2 + \lambda_2 t^2 = d$$

representa a una hipérbola.

En la figura siguiente se muestran los vectores $\overrightarrow{u}$, $\overrightarrow{v}$ y la hipérbola $C = \left\{(s,t) \in \mathbb{R}^2 \mid \lambda_1 s^2 + \lambda_2 t^2 = d\right\}$ con $d > 0$.

Note que la ecuación

$$\begin{aligned} (s,t) &\in \mathbb{R}^2 \text{ tal que } \lambda_1 s^2 + \lambda_2 t^2 = d \Longleftrightarrow \\ (s,t) &\in \mathbb{R}^2 \text{ tal que } \left(\frac{s}{\sqrt{\frac{d}{\lambda_1}}}\right)^2 - \left(\frac{t}{\sqrt{-\frac{d}{\lambda_2}}}\right)^2 = 1 \end{aligned}$$



$\lambda_1 > 0,\ \lambda_2 < 0$
$d > 0$
$|\lambda_2| > \lambda_1$

Figura 76

Si $\lambda_1 = 0$ y $\lambda_2 \neq 0$, entonces el conjunto C se escribe como sigue

$$C = \left\{(s,t) \in \mathbb{R}^2 \mid \lambda_2 t^2 = d\right\} = \left\{(s,t) \in \mathbb{R}^2 \mid t^2 = \frac{d}{\lambda_2}\right\}.$$

Se tiene los siguientes casos: $\dfrac{d}{\lambda_2} < 0$, $d = 0$, $\dfrac{d}{\lambda_2} > 0$.

a) En el caso $\dfrac{d}{\lambda_2} < 0$, la ecuación $t^2 = \dfrac{d}{\lambda_2}$ es contradictoria con lo que $C = \emptyset$.

b) En el caso $d = 0$, el conjunto C se expresa como sigue

$$C = \left\{(s,t) \in \mathbb{R}^2 \mid t = 0\right\} = \left\{s\,(1,0) \mid s \in \mathbb{R}\right\},$$

es decir que en el sistema de referencia $\{\overrightarrow{u}, \overrightarrow{v}\}$, C representa una recta que es paralela a $\overrightarrow{u}$. En la figura siguiente se muestra este conjunto.

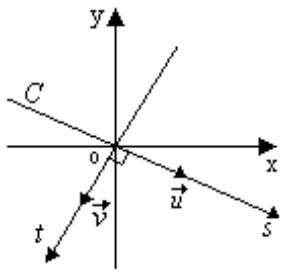


Figura 77

En el caso $\frac{d}{\lambda_2} > 0$, se tiene la ecuación $t^2 = \frac{d}{\lambda_2}$, de donde $t = -\sqrt{\frac{d}{\lambda_2}}$, o, $t = \sqrt{\frac{d}{\lambda_2}}$ con lo que

$$\begin{aligned} C &= \left\{(s,t) \in \mathbb{R}^2 \mid t^2 = \frac{d}{\lambda_2}\right\} \\ &= \left\{(s,t) \in \mathbb{R}^2 \mid t = -\sqrt{\frac{d}{\lambda_2}}\right\} \cup \left\{(s,t) \in \mathbb{R}^2 \mid t = \sqrt{\frac{d}{\lambda_2}}\right\}. \end{aligned}$$

Se define $L_1 = \left\{\left(s, -\sqrt{\frac{d}{\lambda_2}}\right) \mid s \in \mathbb{R}\right\}$, $L_2 = \left\{\left(s, \sqrt{\frac{d}{\lambda_2}}\right) \mid s \in \mathbb{R}\right\}$. Los conjuntos L_1 y L_2 representan rectas paralelas al vector $\overrightarrow{u}$. En la figura siguiente se ilustran los conjuntos L_1, L_2.

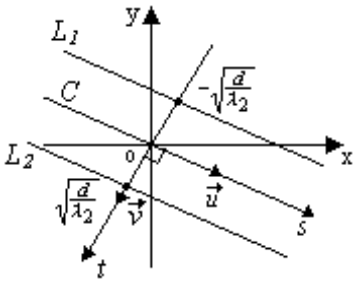


Figura 78

El caso $\lambda_1 \neq 0$ y $\lambda_2 = 0$ se analiza en forma parecida al caso 4).

Con todo el análisis realizado sabemos en que condiciones $C = \emptyset$, y en cuáles $C \neq \emptyset$. En este último caso podemos identificar si se trata de una circunferencia, elipse, hipérbola o simplemente rectas; y, estamos en condiciones de proponer un algoritmo que permita identificar todos estos casos.

Algoritmo

Datos de entrada: $a, b, c, d \in \mathbb{R}$.

Datos de salida: Mensaje 1 : $C \neq \emptyset$; Mensaje 2 : $C = \{(0,0)\}$; Mensaje 3 : Datos no cumplen con la hipótesis; λ_1, λ_2, $\overrightarrow{u}$, $\overrightarrow{v}$.

1. Si $c = 0$, o $a = 0$ y $b = 0$. Continuar en 13)

2. Calcular $u = \frac{1}{2}(a+b)$

$$\alpha = \frac{1}{2}\sqrt{(a-b)^2 + c^2}$$

$$\lambda_1 = u - \alpha$$

$$\lambda_2 = u + \alpha$$

3. Resolver el sistema de ecuaciones $\begin{cases} (a-\lambda_2)\,x + \frac{1}{2}cy = 0 \\ \frac{1}{2}cx + (b-\lambda_1)\,y = 0. \end{cases}$.

Obtener $\overrightarrow{u} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ tal que $\|\overrightarrow{u}\| = 1$.

Resolver el sistema de ecuaciones $\begin{cases} (a-\lambda_2)\,x + \frac{1}{2}cy = 0 \\ \frac{1}{2}cx + (b-\lambda_1)\,y = 0. \end{cases}$

Obtener $\overrightarrow{v} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$ tal que $\|\overrightarrow{v}\| = 1$.

4. Graficar sistema de coordenadas rectangulares $x,\ y$, y respecto de $B_2 = \{\overrightarrow{u}, \overrightarrow{v}\}$ el sistema de coordenas $s,\ t$.

5. Si, $\lambda_1 > 0,\ \lambda_2 > 0$.

   Si $d < 0$, continuar en 11).

   Si $d = 0$, continuar en 12).

   Si $d > 0$

   Si $\lambda_1 \neq \lambda_2$

   Calcular $p = \sqrt{\dfrac{d}{\lambda_1}},\ q = \sqrt{\dfrac{d}{\lambda_2}}$.

   Graficar la elipse $\left\{(s,t) \in \mathbb{R}^2 \mid \left(\dfrac{s}{p}\right)^2 + \left(\dfrac{t}{q}\right)^2 = 1\right\}$.

   Continuar en 14).

   Si $\lambda_1 = \lambda_2$

   Calcular $r = \sqrt{\dfrac{d}{\lambda_1}}$.

   Graficar la circunferencia $\left\{(s,t) \in \mathbb{R}^2 \mid s^2 + t^2 = r^2\right\}$.

   Continuar en 14).

6. Si, $\lambda_1 < 0,\ \lambda_2 < 0$.

   Si $d > 0$, continuar en 11).

   Si $d = 0$, continuar en 12).

   Si $d < 0$.

   Si $\lambda_1 \neq \lambda_2$

   Calcular $p = \sqrt{\dfrac{d}{\lambda_1}},\ q = \sqrt{\dfrac{d}{\lambda_2}}$.

   Graficar la elipse $\left\{(s,t) \in \mathbb{R}^2 \mid \left(\dfrac{s}{p}\right)^2 + \left(\dfrac{t}{q}\right)^2 = 1\right\}$.

   Continuar en 14).

   Si $\lambda_1 = \lambda_2$

   Calcular $r = \sqrt{\dfrac{d}{\lambda_1}}$.

   Graficar la circunferencia $\left\{(s,t) \in \mathbb{R}^2 \mid s^2 + t^2 = r^2\right\}$.

   Continuar en 14).

7. Si, $\lambda_1 < 0,\ \lambda_2 > 0$.

   Si $d > 0$,

   Calcular $p = \sqrt{-\dfrac{d}{\lambda_1}},\ q = \sqrt{\dfrac{d}{\lambda_2}}$.

Graficar la hipérbola $\left\{(s,t) \in \mathbb{R}^2 \mid \left(\frac{s}{p}\right)^2 - \left(\frac{t}{q}\right)^2 = -1\right\}$.

Continuar en 14).

Si $d < 0$,

Calcular $p = \sqrt{\frac{d}{\lambda_1}}$, $q = \sqrt{-\frac{d}{\lambda_2}}$.

Graficar la hipérbola $\left\{(s,t) \in \mathbb{R}^2 \mid \left(\frac{s}{p}\right)^2 - \left(\frac{t}{q}\right)^2 = 1\right\}$.

Continuar en 14).

Si $d = 0$,

Calcular $m = \sqrt{-\frac{\lambda_1}{\lambda_2}}$.

Graficar $C_1 = \left\{(s,t) \in \mathbb{R}^2 \mid t = -ms\right\}$,

$C_2 = \left\{(s,t) \in \mathbb{R}^2 \mid t = ms\right\}$.

Continuar en 14).

8. Si, $\lambda_1 > 0,\ \lambda_2 < 0$.

Si $d > 0$,

Calcular $p = \sqrt{\frac{d}{\lambda_1}}$, $q = \sqrt{-\frac{d}{\lambda_2}}$.

Graficar la hipérbola $\left\{(s,t) \in \mathbb{R}^2 \mid \left(\frac{s}{p}\right)^2 - \left(\frac{t}{q}\right)^2 = 1\right\}$.

Continuar en 14).

Si $d < 0$,

Calcular $p = \sqrt{-\frac{d}{\lambda_1}}$, $q = \sqrt{\frac{d}{\lambda_2}}$.

Graficar la hipérbola $\left\{(s,t) \in \mathbb{R}^2 \mid \left(\frac{s}{p}\right)^2 + \left(\frac{t}{q}\right)^2 = -1\right\}$.

Continuar en 14).

Si $d = 0$,

Calcular $m = \sqrt{-\frac{\lambda_1}{\lambda_2}}$.

Graficar $C_1 = \left\{(s,t) \in \mathbb{R}^2 \mid t = -ms\right\}$,

$C_2 = \left\{(s,t) \in \mathbb{R}^2 \mid t = ms\right\}$.

Continuar en 14).

9. Si $\lambda_1 = 0$

Calcular $p = \frac{d}{\lambda_2}$.

Si $p < 0$, continuar en 11)

Si $d = 0$,

Graficar la recta $C = \{s\,(1,0) \mid s \in \mathbb{R}\}$.

Continuar en 14).

Si $p > 0$,

Graficar las recta $L_1 = \left\{\left(s, -\sqrt{p}\right) \mid s \in \mathbb{R}\right\}$,

$$L_2 = \left\{\left(s, \sqrt{p}\right) \mid s \in \mathbb{R}\right\}.$$

Continuar en 14).

10. Si $\lambda_2 = 0$

Calcular $p = \dfrac{d}{\lambda_1}$.

Si $p < 0$, continuar en 11)

Si $d = 0$,

Graficar la recta $C = \{t\,(0,1) \mid t \in \mathbb{R}\}$.

Continuar en 14).

Si $p > 0$,

Graficar las recta $L_1 = \left\{\left(-\sqrt{p}, t\right) \mid t \in \mathbb{R}\right\}$,

$$L_2 = \left\{\left(\sqrt{p}, t\right) \mid t \in \mathbb{R}\right\}.$$

Continuar en 14).

11. Mensaje 1 : $C = \emptyset$. Continuar en 14).

12. Mensaje 2 : $C = \{(0,0)\}$. Continuar en 14).

13. Mensaje 3 : Datos no cumplen con la hipótesis.

14. Fin

El algoritmo concluye en un número finito de pasos. Note que se realizan las 4 operaciones aritméticas y raíz cuadrada. Se realizan algunas comparaciones.

Para verificar el algoritmo proponemos tres ejemplos.

1. Consideremos el subconjunto C de $\mathbb{R}^2$ definido como

$$C = \left\{(x,y) \in \mathbb{R}^2 \mid 4x^2 + 2xy + 4y^2 = 1\right\}.$$

Tenemos $a = 4$, $b = 4$, $c = 2$ y $d = 1$. Claramente $c \neq 0$, $|a| + |b| > 0$. Según el algoritmo (punto 2), pasamos a calcular los valores porpios de la matriz $A = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$. Tenemos $u = \dfrac{1}{2}(a+b) = 4$, $\alpha = \dfrac{1}{2}\sqrt{(a-b)^2 + c^2} = 1$.

Los valores propios de A son $\lambda_1 = u - \alpha = 3$ y $\lambda_2 = u + \alpha = 5$.

Continuando con el algoritmo (punto 3), determinamos los vectores propios de A asociados a λ_1 y λ_2, esto es, determinamos $\overrightarrow{x} = (x,y) \in \mathbb{R}^2$ tal que $\begin{cases} (a-\lambda)\,x + \dfrac{1}{2}cy = 0, \\ \dfrac{1}{2}cx + (b-\lambda)\,y = 0. \end{cases}$

Con $\lambda_1 = 3$ se tiene $\begin{cases} -x + y = 0 \\ x + y = 0 \end{cases} \Leftrightarrow x + y = 0 \Leftrightarrow y = -x.$

Luego $\overrightarrow{x} = (x, y) = (x, -x) = x\,(1, -1) \quad x \in \mathbb{R}$.

Los vectores propios normalizados de A son (punto 4 del algoritmo) $\overrightarrow{u} = \dfrac{1}{\sqrt{2}}(1, -1)$ y $\overrightarrow{v} = \dfrac{1}{\sqrt{2}}(1, 1)$. se verifica inmediatamente que $\overrightarrow{u}$ y $\overrightarrow{v}$ son ortogonales, esto es, $\overrightarrow{u} \cdot \overrightarrow{v} = 0$. Puesto que $\lambda_1 > 0$, $\lambda_2 > 0$ y $d > 0$ (punto 5 del algortimo) entonces

$$p = \sqrt{\frac{d}{\lambda_1}} = \sqrt{\frac{1}{3}} = \frac{\sqrt{3}}{3} \simeq 0{,}577, \qquad q = \sqrt{\frac{d}{\lambda_2}} = \sqrt{\frac{1}{5}} = \frac{\sqrt{5}}{5} \simeq 0{,}447.$$

Entonces $C = \left\{ (s,t) \in \mathbb{R}^2 \mid \left(\dfrac{5}{\frac{\sqrt{3}}{3}} \right)^2 + \left(\dfrac{t}{\frac{\sqrt{5}}{5}} \right)^2 = 1 \right\}$ es una elipse.

En la figura siguiente se muestran los vectores ortogonales $\overrightarrow{u}$, $\overrightarrow{v}$ y la elipse C.

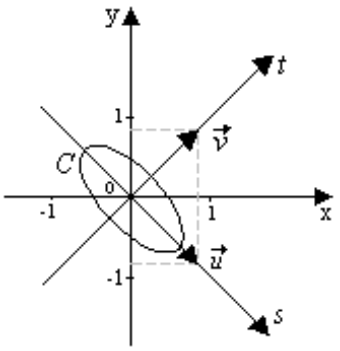


Figura 79

2. Sean $d \in \mathbb{R}^2$ y $C = \left\{ (x, y) \in \mathbb{R}^2 \mid 2x^2 + 4xy + 2y^2 = d \right\}$. Tenemos $a = 2$, $b = 2$, $c = 4$. Se verifica inmediatamente que $|a| + |b| > 0$ y $c \neq 0$ (punto 1 del algoritmo).

Calculemos los valores propios de la matriz $a = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ (punto 2 del algoritmo). Tenemos

$$\begin{aligned} u &= \frac{1}{2}(a + b) = 2, \quad \alpha = \frac{1}{2}\sqrt{(a-b)^2 + c^2} = 2, \\ \lambda_1 &= u - \alpha = 0, \quad \lambda_2 = u + \alpha = 4. \end{aligned}$$

Determinemos los vectores propios de la matriz A asociados a los valores propios λ_1 y λ_2.

Con $\lambda_1 = 0$, hallar $\overrightarrow{x} = (x, y) \in \mathbb{R}^2$ solución de $\begin{cases} (2 - \lambda_1)\,x + 2y = 0, \\ 2x + (2 - \lambda_1)\,y = 0. \end{cases}$ La solución de este sistema conduce a la ecuación $x + y = 0$, luego $\overrightarrow{x} = (x, y) = (x, -x) = x\,(1, -1) \quad x \in \mathbb{R}$.

Con $\lambda_2 = 0$, hallar $\overrightarrow{x} = (x, y) \in \mathbb{R}^2$ solución de $\begin{cases} (2 - \lambda_2)\,x + 2y = 0, \\ 2x + (2 - \lambda_2)\,y = 0. \end{cases}$

Resulta $\begin{cases} -2x + 2y = 0 \\ 2x - 2y = 0 \end{cases} \iff x = y$. Luego

$$\overrightarrow{x} = (x, y) = (x, x) = x\,(1, 1) \quad x \in \mathbb{R}.$$

Ponemos $\overrightarrow{u} = \dfrac{1}{\sqrt{2}}(1, -1)$, $\overrightarrow{v} = \dfrac{1}{\sqrt{2}}(1, 1)$. Los vectores $\overrightarrow{u}$ y $\overrightarrow{v}$ son vectores propios de A, esto es, $A\overrightarrow{u} = \lambda_1 \overrightarrow{u}$ y $A\overrightarrow{v} = \lambda_2 \overrightarrow{v}$; y, $\overrightarrow{u} \perp \overrightarrow{v}$.

Siguiendo con el algoritmo (punto 9) se tiene

$$C = \left\{ (s,t) \in \mathbb{R}^2 \mid \lambda_1 s^2 + \lambda_2 t^2 = d \right\} = \left\{ (s,t) \in \mathbb{R}^2 \mid 4b^2 = d \right\}.$$

Si $d<0$, la ecuación $4t^2=d$ es absurda, luego $C=\emptyset$.
Si $d=0$, la ecuación $4t^2=0\Rightarrow t=0$, luego $c=\left\{(s,t)\in\mathbb{R}^2 \mid t=0\right\}=\left\{s\,(1,0) \mid s\in\mathbb{R}\right\}$.
Si $d>0$, la ecuación $t^2=\dfrac{d}{4}$ tiene dos raíces $t_1=-\dfrac{\sqrt{d}}{2}$ y $t_2=\dfrac{\sqrt{d}}{2}$ con lo que

$$C=\left\{(s,t)\in\mathbb{R}^2 \mid t^2=\frac{d}{4}\right\}=\left\{(s,t)\in\mathbb{R}^2 \mid t=-\frac{\sqrt{d}}{2}\right\}\cup\left\{(s,t)\in\mathbb{R}^2 \mid t=\frac{\sqrt{d}}{2}\right\}.$$

Ponemos $L_1=\left\{(s,t)\in\mathbb{R}^2 \mid t=-\dfrac{\sqrt{d}}{2}\right\}=\left\{\left(s,-\dfrac{\sqrt{d}}{2}\right) \mid s\in\mathbb{R}\right\}$, $L_2=\left\{\left(s,\dfrac{\sqrt{d}}{2}\right) \mid s\in\mathbb{R}\right\}$.
En la figura de la izquierda se representa el conjunto $C=\{s\,(1,0) \mid s\in\mathbb{R}\}$ y en el de la derecha se representa el conjunto $C=L_1\cup L_2$.

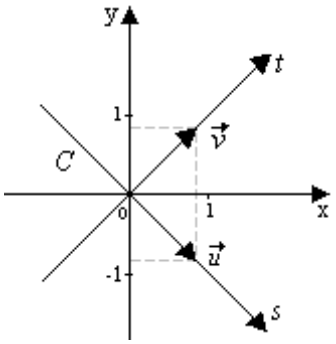


Figura 80

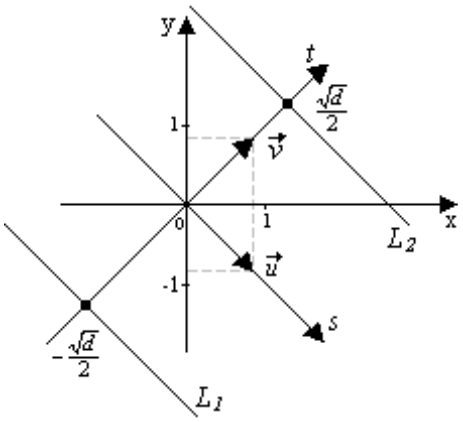


Figura 81

Note que $C=\left\{(x,y)\in\mathbb{R}^2 \mid 2x^2+4xy+2y^2=d\right\}=\left\{(x,y)\in\mathbb{R}^2 \mid (x+y)^2=\dfrac{d}{2}\right\}$.
Si $d<0$, resulta $C=\emptyset$. Si $d=0$, se obtiene $C=\left\{(x,y)\in\mathbb{R}^2 \mid x+y=0\right\}$ que representa una recta de ecuación cartesiana $y=-x$. Si $d>0$, se tiene $C=\left\{(x,y)\in\mathbb{R}^2 \mid x+y=-\dfrac{\sqrt{2d}}{2}\right\}\cup\left\{(x,y)\in\mathbb{R}^2 \mid x+y=\dfrac{\sqrt{2d}}{2}\right\}$.Las ecuaciones cartesianas de las rectas son:

$$(x,y)\in\mathbb{R}^2 \text{ tal que } x+y=-\frac{\sqrt{2d}}{2},\qquad (x,y)\in\mathbb{R}^2 \text{ tal que } x+y=\frac{\sqrt{2d}}{2}.$$

8.3. Valores y vectores propios de matrices de 3×3

Sea $A=(a_{ij})\in M_{3\times 3}[\mathbb{R}]$. Determinemos $\lambda\in\mathbb{R}$ y $\overrightarrow{x}\in\mathbb{R}^3$ no nulo tal que $A\overrightarrow{x}=\lambda\overrightarrow{x}$. En este caso el polinomio característico está definido como sigue:

$$\begin{aligned} p(\lambda) &= \det(A-\lambda I)=\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13}\\ a_{21} & a_{22}-\lambda & a_{23}\\ a_{31} & a_{32} & a_{33-}\lambda \end{vmatrix}\\ &= (a_{11}-\lambda)\begin{vmatrix} a_{22}-\lambda & a_{23}\\ a_{32} & a_{33-}\lambda \end{vmatrix}-a_{12}\begin{vmatrix} a_{21} & a_{23}\\ a_{31} & a_{33-}\lambda \end{vmatrix}+a_{13}\begin{vmatrix} a_{21} & a_{22}-\lambda\\ a_{31} & a_{32} \end{vmatrix}. \end{aligned}$$

Los valores propios se obtienen como solución de la ecuación $p(\lambda)=0$. Como el polinomio característico es de grado 3, existe al menos una raíz real la misma que puede ser calculada como solución aproximada

mediante el método de Newton. Los vectores propios son solución del sistema de ecuaciones lineales

$$\begin{bmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Ejemplo

1. Consideramos la matriz $A = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}$. Apliquemos los resultados obtenidos, el algoritmo de búsqueda del cambio de signo y el método de Newton para calcular todos los valores propios de A.

   Primeramente, la matriz A es simétrica luego sus valores propios son reales, tenemos $|\lambda| \leq |||A|||_\infty = 4$. Así $\lambda \in [-4, 4]$. Una estimación más fina la obtenemos si determinamos los discos de Gershgorin. Tenemos $|\lambda - 2| \leq 2$, $|\lambda - 2| \leq 1$, $|\lambda - 2| \leq 1$ entonces $|\lambda - 2| \leq 2 \Longleftrightarrow \lambda \in [0, 4]$.

   Determinemos el polinomio característico:

$$p(\lambda) = \det(A - \lambda I) = \begin{vmatrix} 2-\lambda & -1 & 1 \\ -1 & 2-\lambda & 0 \\ 1 & 0 & 2-\lambda \end{vmatrix} = -\lambda^3 + 6\lambda^2 - 10\lambda + 4 \quad \lambda \in \mathbb{R}.$$

   Para este ejemplo los valores propios se calculan fácilmente y así podemos comparar con los métodos a utilizar. Se tiene

$$p(\lambda) = (2-\lambda)\left(\lambda^2 - 4\lambda + 2\right) = 0 \Leftrightarrow \begin{cases} \lambda_1 = 2, \\ \lambda_2 = 2 - \sqrt{2}, \\ \lambda_3 = 2 + \sqrt{2}. \end{cases}$$

   Sea $h = 0{,}5$. La aplicación del algoritmo de búsqueda del cambio de signo en el intervalo $[-0{,}3, 4]$ nos da los resultdos que se indican en la tabla siguiente:

x	$p(x)$
$-0{,}3$	$7{,}567$
$0{,}2$	$2{,}232$
$0{,}7$	$-0{,}403$
$1{,}2$	$-1{,}088$
$1{,}7$	$-0{,}573$
$2{,}2$	$0{,}392$
$2{,}7$	$1{,}057$
$3{,}2$	$0{,}672$
$3{,}7$	$-1{,}513$
$4{,}0$	$-4{,}0$

   Como se observa, la aplicación del algoritmo de búsqueda del cambio de signo muestra la existencia de tres raíces reales: $C_1 \in [0{,}2, 0{,}7]$, $C_2 \in [1{,}7, 2{,}2]$, $C_3 \in [3{,}3, 3{,}7]$.

   En realidad el algoritmo de búsqueda del cambio de signo se aplica en el intervalo $[0, 4]$.

   Apliquemos el método de Newton para calcular cada una de estas raíces. Tenemos

$$\begin{aligned} p(x) &= -x^3 + 6x^2 - 10x + 4 = 4 + x(-10 + x(6 - x)), \\ p'(x) &= -3x^2 + 12x - 10 = -10 + x(12 - 3x), \end{aligned}$$

   y el esquema de Newton está definido como

$$\begin{cases} x_0 \ \text{dado}, \\ x_{n+1} = x_n - \dfrac{p(x_n)}{p'(x_n)} \quad n = 0, 1, \ldots, N_{\text{máx}}. \end{cases}$$

Cálculo de $C_1 \in [0{,}2, 0{,}7]$. Ponemos x_0 el punto medio del intervalo, esto es $x_0 = 0{,}45$. En la tabla siguiente se encuentran los resultados de la aplicación del método de Newton

Iteración	x_j	$p(x_j)$
0	0,45	0,623875
1	0,569803	0,065021
2	0,585522	0,0010563
3	0,585786	0,0000003
4	0,585786	$2{,}33 \times 10^{-14}$

.

La raíz es $C_1 = 0{,}5857864376$.

Procediendo en forma similar a la precedente, elegimos x_0 el punto medio del intervalo $[1{,}7, 2{,}2]$, es decir $x_0 = 1{,}95$. En la tabla siguiente se muestran los resultados obtenidos de la aplicación del método de Newton.

Iteración	x_j	$p(x_j)$
0	1,95	$-0{,}099875$
1	2,000125	0,0002509
2	1,9999999	$-0{,}39 \times 10^{-11}$
3	2,0	0.

.

La raíz es $C_2 = 2$.

Sea x_0 el punto medio del intervalo $[3{,}3, 3{,}7]$: $x_0 = 3{,}45$. Entonces,

Iteración	x_j	$p(x_j)$
0	3,45	$-0{,}148625$
1	3,415496	$-0{,}00513764$
2	3,41421530	$-0{,}000006965$
3	3,414213562	$-1{,}28 \times 10^{-11}$

.

La raíz es $C_3 = 3{,}414213562$.

Calculemos los vectores propios de A asociados a los valores propios $\lambda_1 = 2-\sqrt{2},\ \lambda = 2, \lambda_3 = 2+\sqrt{2}$. Tenemos

$$(A - \lambda I)\overrightarrow{x} = \overrightarrow{0} \Leftrightarrow \begin{cases} (2-\lambda)x & -y & +z = 0, \\ -x + (2-\lambda)y & & = 0, \\ x + & +(2-\lambda)z = 0. \end{cases}$$

Para $\lambda_1 = 2 - \sqrt{2}$ se tiene $\begin{cases} \sqrt{2}x & -y & +z = 0 \\ -x + \sqrt{2}y & & = 0 \\ x + & +\sqrt{2}z = 0. \end{cases}$

Aplicando el método de eliminación gaussiana se obtiene el siguiente sistema $\begin{cases} \sqrt{2}x & -y & +z = 0 \\ & \frac{1}{\sqrt{2}}y + \frac{1}{\sqrt{2}}z = 0, \end{cases}$

de donde

$$S_{\lambda_1} = \left\{ \overrightarrow{x}^T = (x, y, z) \in \mathbb{R}^3 \mid (A - \lambda_1 I)\overrightarrow{x} = \overrightarrow{0} \right\} = \left\{ z\left(-\sqrt{2}, -1, 1\right) \mid z \in \mathbb{R} \right\}.$$

Para $\lambda = 2$ se obtiene el sistema de ecuaciones $\begin{cases} & -y + z = 0 \\ -x & = 0 \\ x & = 0, \end{cases}$ luego

$$S_{\lambda_2} = \left\{ \overrightarrow{x} \in \mathbb{R}^3 \mid (A - 2I)\overrightarrow{x} = \overrightarrow{0} \right\} = \{ y(0, 1, 1) \mid y \in \mathbb{R} \}$$

Se deja como ejercicio determinar S_{λ_3}.

8.4. Método de las Potencias

En muchas situaciones no estamos interesados en calcular todos los valores propios y todos los vectores propios, sino algunos de ellos, por ejemplo el más grande o el más pequeño en valor absoluto de los valores propios con sus respectivos vectores propios. Este método puede adaptarse en forma apropiada para calcular otros valores y vectores propios.

Suponemos que la matriz A es diagonalizable, es decir que existe una matriz invertible P tal que $D = P^{-1}AP$, donde $D = (\lambda_1, \cdots, \lambda_n)$ una matriz diagonal con los valores propios de la matriz A en su diagonal. Consecuentemente xisten $\overrightarrow{x}_1, \cdots, \overrightarrow{x}_n$ vectores propios asociados a $\lambda_1, \cdots, \lambda_n$, respectivamente. Tenemos $A\overrightarrow{x}_i = \lambda_i \overrightarrow{x}_i \quad i = 1, \cdots, n$. Adicionalmente, suponemos que $\mid \lambda_n \mid \leq \cdots \leq \mid \lambda_1 \mid$, y asumimos que el conjunto $\{\overrightarrow{x}_1, \cdots, \overrightarrow{x}_n\}$ es una base de $\mathbb{R}^n$. Al valor propio λ_1 lo denominaremos valor propio dominante de la matriz A.

Sea $\overrightarrow{x} \in \mathbb{R}^n$ no nulo. Existen $\beta_1, \cdots, \beta_n \in \mathbb{R}$ tales que $\overrightarrow{x} = \sum_{i=1}^n \beta_i \overrightarrow{x}_i$, luego

$$\begin{aligned}
\overrightarrow{y}_1 &= A\overrightarrow{x} = A\sum_{i=1}^n \beta_i \overrightarrow{x}_i = \sum_{i=1}^n \beta_i A\overrightarrow{x}_i = \sum_{i=1}^n \beta_i \lambda_i \overrightarrow{x}_i, \\
\overrightarrow{y}_2 &= A^2\overrightarrow{x} = A\sum_{i=1}^n \beta_i \lambda_i \overrightarrow{x}_i = \sum_{i=1}^n \beta_i \lambda_i^2 \overrightarrow{x}_i, \\
&\vdots \\
\overrightarrow{y}_m &= A^m\overrightarrow{x} = A\sum_{i=1}^n \beta_i \lambda_i^{m-1} \overrightarrow{x}_i = \sum_{i=1}^n \beta_i \lambda_i^m \overrightarrow{x}_i.
\end{aligned}$$

La última igualdad se expresa como

$$\overrightarrow{y}_m = A^m\overrightarrow{x} = \lambda_1^m \sum_{i=1}^n \beta_i \left(\frac{\lambda_i}{\lambda_1}\right)^m \overrightarrow{x}_i.$$

Por otro lado, de la relación $\mid \lambda_n \mid \leq \cdots \leq \mid \lambda_1 \mid$ se tiene $\mid \frac{\lambda_n}{\lambda_1} \mid^m \leq \cdots \leq \mid \frac{\lambda_2}{\lambda_1} \mid^m \leq 1$, de donde

$$\lim_{m\longrightarrow\infty} \mid \frac{\lambda_k}{\lambda_1} \mid^m = 0 \quad \text{si} \quad \left|\frac{\lambda_k}{\lambda_1}\right| < 1 \quad k = 2, \cdots, n.$$

El principio del método de la potencia está en la relación de cada valor propio con el valor propio dominante de la matriz A, es decir, de la razón $\left|\frac{\lambda_k}{\lambda_1}\right| < 1 \quad k = 2, \cdots, n$. Tenemos

$$\begin{aligned}
\lim_{m\longrightarrow\infty} A^m\overrightarrow{x} &= \lim_{m\longrightarrow\infty} \lambda_1^m \sum_{i=1}^n \beta_i \left(\frac{\lambda_i}{\lambda_1}\right)^m \overrightarrow{x}_i = \lim_{m\longrightarrow\infty} \lambda_1^m \left(\beta_1 \overrightarrow{x}_1 + \sum_{i=2}^n \beta_i \left(\frac{\lambda_i}{\lambda_1}\right)^m \overrightarrow{x}_i\right) \\
&= \lim_{m\longrightarrow\infty} \lambda_1^m \beta_1 \overrightarrow{x}_1, \quad \text{si} \quad \left|\frac{\lambda_k}{\lambda_1}\right| < 1 \quad k = 2, \cdots, n.
\end{aligned}$$

Es claro que $\lim_{m\longrightarrow\infty} \lambda_1^m = 0$ si y solo si $\mid \lambda_1 \mid < 1$, $\lim_{m\longrightarrow\infty} \lambda_1^m = 1$ si y solo si $\lambda_1 = 1$, $\lim_{m\longrightarrow\infty} \lambda_1^m$ no existe si y solo si $\mid \lambda_1 \mid > 1$ o $\lambda_1 = -1$.

Elegimos $\overrightarrow{x}$ de modo $\beta_1 \neq 0$.

i) Si $\mid \lambda_1 \mid > \mid \lambda_2 \mid$, se tiene

$$\overrightarrow{y}_m = \lambda_1^m \left(\beta_1 \overrightarrow{x}_1 + \sum_{i=2}^n \beta_i \left(\frac{\lambda_i}{\lambda_1}\right)^m \overrightarrow{x}_i\right) \Longleftrightarrow \frac{\overrightarrow{y}_m}{\lambda_1^m} = \beta_1 \overrightarrow{x}_1 + \sum_{i=2}^n \beta_i \left(\frac{\lambda_i}{\lambda_1}\right)^m \overrightarrow{x}_i,$$

luego

$$\lim_{m\longrightarrow\infty}\frac{\overrightarrow{y}_m}{\lambda_1^m}=\lim_{m\longrightarrow\infty}\left(\beta_1\overrightarrow{x}_1+\sum_{i=2}^{n}\beta_i\left(\frac{\lambda_i}{\lambda_1}\right)^m\overrightarrow{x}_i\right)=\beta_1\overrightarrow{x}_1.$$

Este límite es a su vez equivalente a los siguientes, expresados en términos de sus componentes:

$$\lim_{m\longrightarrow\infty}\frac{y_i^m}{\lambda_1^m}=\beta_1 x_i^{(1)}\quad i=1,\cdots n,$$

con $\overrightarrow{x}_1=(x_1^{(1)},\cdots,x_n^{(1)})$, $\overrightarrow{y}_m=(y_1^m,\cdots,y_n^m)$. Se define $q_i^m=\dfrac{y_i^m}{y_i^{m-1}}\quad y_i^{m-1}\neq 0$, y de la existencia del límite precedente se tiene $\lim\limits_{m\longrightarrow\infty}q_i^m=\lambda_1$.

ii) Supongamos $\mid\lambda_1\mid=\cdots=\mid\lambda_k\mid$ con $2\leq k<n$. Para comprender mejor la situación, supongamos $k=2$ y que $\lambda_1=\lambda_2$, entonces

$$\overrightarrow{y}_m=\lambda_1^m\left(\beta_1\overrightarrow{x}_1-\beta_2\overrightarrow{x}_2+\sum_{i=3}^{n}\beta_i\left(\frac{\lambda_i}{\lambda_1}\right)^m\overrightarrow{x}_i\right)\Longleftrightarrow\frac{\overrightarrow{y}_m}{\lambda_1^m}=\beta_1\overrightarrow{x}_1-\beta_2\overrightarrow{x}_2+\sum_{i=2}^{n}\beta_i\left(\frac{\lambda_i}{\lambda_1}\right)^m\overrightarrow{x}_i.$$

Se consideran dos casos: $m=2N$, y $m=2N+1$ para $N\in\mathbb{Z}^+$. Se muestra que $\lim\limits_{m\longrightarrow\infty}\dfrac{y_i^{m+2}}{y_i^m}=\lambda_1^2$ siempre que $y_i^m\neq 0$.

En la práctica, el método de la potencia se aplica del siguiente modo.

1. A menos que se tenga una buena estimación del vector $\overrightarrow{x}$, elegimos $\overrightarrow{x}=\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$. Calculamos $\overrightarrow{y}_1=A\overrightarrow{x}$, sea $p_1=\parallel\overrightarrow{y}_1\parallel_\infty$ y se define $\overrightarrow{z}_1=\dfrac{1}{\parallel\overrightarrow{y}_1\parallel_\infty}\overrightarrow{y}_1$.

2. Calculamos $\overrightarrow{y}_2=A\overrightarrow{z}_1$, $p_2=\parallel\overrightarrow{y}_2\parallel_\infty$ y $\overrightarrow{z}_2=\dfrac{1}{\parallel\overrightarrow{y}_2\parallel_\infty}\overrightarrow{y}_2$.

El proceso continua m veces. Cuando el número de iteraciones aumenta, p_m se aproxima al más grande valor propio en valor absoluto y $\overrightarrow{z}_m=\dfrac{1}{\parallel\overrightarrow{y}_m\parallel_\infty}\overrightarrow{y}_m$ se aproxima al vector propio asociado al valor propio dominante.

Ejemplos

1. Consideremos la matriz $A=\begin{bmatrix}2&0&1\\0&0&2\\0&2&1\end{bmatrix}$. Tenemos $\mid\lambda\mid\leq\mid\mid\mid A\mid\mid\mid=3$. Además, los discos de Gershgorin muestran que $\lambda\in[-2,3]$. Por otro lado, el polinomio característico está definido como

$$p(\lambda)=\det(A-\lambda I)=-\lambda^3+3\lambda+2\lambda-8\quad\lambda\in\mathbb{R},$$

luego, $p(\lambda)=0\Longleftrightarrow\lambda_3=\dfrac{1-\sqrt{5}}{2},\quad\lambda_2=\dfrac{1+\sqrt{5}}{2},\quad\lambda_1=2.$

El vector propio asociado al valor propio dominante es $\overrightarrow{x}_1=\begin{bmatrix}1\\0\\0\end{bmatrix}$.

Apliquemos el método de las potencias para aproximar al valor dominante y consecuentemente al vector propio asociado.

Sea $\overrightarrow{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Primera iteración: calculamos $\overrightarrow{y}_1 = A\overrightarrow{x} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix}$,

$$p_1 = \| \overrightarrow{y}_1 \|_\infty = 3,$$

$$\overrightarrow{z}_1 = \frac{1}{p_1}\overrightarrow{y}_1 = \frac{1}{3}\begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{2}{3} \\ 1 \end{bmatrix}.$$

Segunda iteración: calculamos $\overrightarrow{y}_2 = A\overrightarrow{z}_1 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{2}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ \frac{7}{3} \end{bmatrix}$,

$$p_2 = \| \overrightarrow{y}_2 \|_\infty = 3,$$

$$\overrightarrow{z}_2 = \frac{1}{p_2}\overrightarrow{y}_2 = \frac{1}{3}\begin{bmatrix} 3 \\ 2 \\ \frac{7}{3} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{2}{3} \\ \frac{7}{9} \end{bmatrix}.$$

Tercera iteración: calculamos $\overrightarrow{y}_3 = A\overrightarrow{z}_2 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{2}{3} \\ \frac{7}{9} \end{bmatrix} = \begin{bmatrix} \frac{25}{9} \\ \frac{7}{9} \\ \frac{19}{9} \end{bmatrix}$,

$$p_3 = \| \overrightarrow{y}_3 \|_\infty = \frac{25}{9} = 2{,}77777778,$$

$$\overrightarrow{z}_3 = \frac{1}{p_3}\overrightarrow{y}_3 = \frac{9}{3}\begin{bmatrix} \frac{25}{9} \\ \frac{7}{9} \\ \frac{19}{9} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{7}{25} \\ \frac{19}{25} \end{bmatrix}.$$

Cuarta iteración: calculamos $\overrightarrow{y}_4 = A\overrightarrow{z}_3 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{7}{25} \\ \frac{19}{25} \end{bmatrix} = \begin{bmatrix} \frac{69}{25} \\ \frac{14}{25} \\ \frac{33}{25} \end{bmatrix}$,

$$p_4 = \| \overrightarrow{y}_4 \|_\infty = \frac{69}{25} = 2{,}76,$$

$$\overrightarrow{z}_4 = \frac{1}{p_4}\overrightarrow{y}_4 = \frac{69}{25}\begin{bmatrix} \frac{69}{25} \\ \frac{14}{25} \\ \frac{33}{25} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{14}{69} \\ \frac{33}{69} \end{bmatrix}.$$

Quinta iteración: calculamos $\overrightarrow{y}_5 = A\overrightarrow{z}_4 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{14}{69} \\ \frac{33}{69} \end{bmatrix} = \begin{bmatrix} \frac{171}{69} \\ \frac{66}{69} \\ \frac{61}{69} \end{bmatrix}$,

$$p_5 = \| \overrightarrow{y}_5 \|_\infty = \frac{171}{69} = 2{,}47826087,$$

$$\overrightarrow{z}_5 = \frac{1}{p_5}\overrightarrow{y}_5 = \frac{69}{171}\begin{bmatrix} \frac{69}{25} \\ \frac{14}{25} \\ \frac{33}{25} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{14}{171} \\ \frac{33}{171} \end{bmatrix}.$$

Note que a medida que se realizan más iteraciones los valores de p_m se aproximan al valor propio dominante $\lambda_1 = 2$. Igualmente, $\overrightarrow{z}_m$ se aproxima al vector propio $\overrightarrow{x}_1$ asociado a λ_1.

2. Consideremos la matriz $A = \begin{bmatrix} 3 & 7 & 9 \\ 7 & 4 & 3 \\ 9 & 3 & 8 \end{bmatrix}$.

Esta matriz A es simétrica, por lo tanto tiene tres valores propios reales. Apliquemos el procedimiento arriba descrito.

Sea $\overrightarrow{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Primera iteración: calculamos $\overrightarrow{y}_1 = A\overrightarrow{x} = \begin{bmatrix} 3 & 7 & 9 \\ 7 & 4 & 3 \\ 9 & 3 & 8 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 19 \\ 14 \\ 20 \end{bmatrix}$,

$$p_1 = \| \overrightarrow{y}_1 \|_\infty = 20,$$

$$\overrightarrow{z}_1 = \frac{1}{p_1}\overrightarrow{y}_1 = \frac{1}{20}\begin{bmatrix} 19 \\ 14 \\ 20 \end{bmatrix} = \begin{bmatrix} 0{,}95 \\ 0{,}7 \\ 1 \end{bmatrix}.$$

Segunda iteración: calculamos $\overrightarrow{y}_2 = A\overrightarrow{z}_1 = \begin{bmatrix} 3 & 7 & 9 \\ 7 & 4 & 3 \\ 9 & 3 & 8 \end{bmatrix}\begin{bmatrix} 0{,}95 \\ 0{,}7 \\ 1 \end{bmatrix} = \begin{bmatrix} 16{,}95 \\ 12{,}45 \\ 18{,}65 \end{bmatrix}$,

$$p_2 = \| \overrightarrow{y}_2 \|_\infty = 18{,}65,$$

$$\overrightarrow{z}_2 = \frac{1}{p_2}\overrightarrow{y}_2 = \frac{1}{1865}\begin{bmatrix} 16{,}75 \\ 12{,}45 \\ 18{,}65 \end{bmatrix} = \begin{bmatrix} 0{,}8981233244 \\ 0{,}6675603217 \\ 1 \end{bmatrix}.$$

Tercera iteración: calculamos

$$\overrightarrow{y}_3 = A\overrightarrow{z}_2 = \begin{bmatrix} 3 & 7 & 9 \\ 7 & 4 & 3 \\ 9 & 3 & 8 \end{bmatrix} \begin{matrix} 0{,}95 \\ 0{,}7 \\ 1 \end{matrix} \begin{bmatrix} 0{,}8981233244 \\ 0{,}6675603217 \\ 1 \end{bmatrix} = \begin{bmatrix} 16{,}36729223 \\ 11{,}95710456 \\ 18{,}08579089 \end{bmatrix},$$

$$p_3 = \| \overrightarrow{y}_3 \|_\infty = 18{,}08579089.$$

En la tercera iteración, una aproximación del valor propio dominante $\lambda_1 \simeq 18{,}085$. En 10 iteraciones se obtiene $\lambda_1 \simeq 18{,}0138$ y una aproximación del vector propio $\overrightarrow{x}_1 \simeq \begin{bmatrix} 0{,}90217 \\ 0{,}66059 \\ 1 \end{bmatrix}$.

Cálculo del más pequeño valor propio en valor absoluto.

El método de las potencias se emplea directamente para calcular el más pequeño valor propio en valor absoluto de una matriz invertible.

Sean $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$ matriz invertible, $\lambda \in \mathbb{R}$ un valor propio de A y $\overrightarrow{x} \in \mathbb{R}^n$ un vector propio asociado a λ. Entonces $A\overrightarrow{x} = \lambda\overrightarrow{x}$. Multiplicando por la matriz inversa A^{-1} y considerando que $A^{-1}A = I$, se tiene

$$\overrightarrow{x} = \lambda A^{-1}\overrightarrow{x} \iff A^{-1}\overrightarrow{x} = \frac{1}{\lambda}\overrightarrow{x},$$

es decir que $\dfrac{1}{\lambda}$ es el valor propio de la matriz A^{-1}.

Ponemos $B = A^{-1}$ y $\xi = \dfrac{1}{\lambda}$. El método de la potencia se aplica a $B\overrightarrow{x} = \xi\overrightarrow{x}$ lo que permite aproximar al valor propio dominante $\xi_1 = \dfrac{1}{\lambda_n}$ y de esta relación se tiene $\lambda_n = \dfrac{1}{\xi_1}$.

Lastimosamente, la aplicación de este método requiere del cálculo de la matriz inversa A^{-1} de A, y cuando la dimensión de la matriz A es grande, los cálculos que se realizan en el método de la potencia son muy grandes lo que hace que este método no sea muy práctico.

8.5. Ejercicios

1. En cada item se define un subconjunto $C \in \mathbb{R}^2$. Realice una transformación de coordenadas para representar C en un nuevo sistema de modo que no aparezca el término xy.

   Determine que conjunto representa C ($\emptyset$, un punto, recta o rectas, cónicas). Represente C si $C \neq 0$. Estime el número de operaciones elementales.

   a) $C = \left\{(x,y) \in \mathbb{R}^2 \mid 2x^2 + 2xy + 2y^2 = -3\right\}$. **b)** $C = \left\{(x,y) \in \mathbb{R}^2 \mid 3x^2 - xy + 4y^2 = 5\right\}$.

   c) $C = \left\{(x,y) \in \mathbb{R}^2 \mid x^2 - 2xy + 2y^2 = 2\right\}$. **d)** $C = \left\{(x,y) \in \mathbb{R}^2 \mid 2xy + 3y^2 = 0\right\}$.

   e) $C = \left\{(x,y) \in \mathbb{R}^2 \mid 5x^2 + 2xy = -1\right\}$. **f)** $C = \left\{(x,y) \in \mathbb{R}^2 \mid 2x^2 - 6xy + 2y^2 = 1\right\}$.

   g) $C = \left\{(x,y) \in \mathbb{R}^2 \mid 3x^2 + 2xy + 2y^2 = 1\right\}$. **h)** $C = \left\{(x,y) \in \mathbb{R}^2 \mid 5x^2 + 4xy - 5y^2 = 1\right\}$.

   i) $C = \left\{(x,y) \in \mathbb{R}^2 \mid -2x^2 + 6xy + y^2 = 1\right\}$. **j)** $C = \left\{(x,y) \in \mathbb{R}^2 \mid -x^2 + 4xy - 2y^2 = 1\right\}$.

2. Con cada matriz triangular superior invertible A que se propone, determine los valores y vectores propios

   a) $A = \begin{bmatrix} 3 & -2 & 1 \\ 0 & 2 & 5 \\ 0 & 0 & -4 \end{bmatrix}$. **b)** $A = \begin{bmatrix} 10 & 3 & -5 \\ 0 & 8 & 15 \\ 0 & 0 & 5 \end{bmatrix}$. **c)** $A = \begin{bmatrix} -1 & 2 & -3 & -2 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 8 \\ 0 & 0 & 0 & -1 \end{bmatrix}$.

   d) $A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & -2 & 1 & -1 \\ 0 & 0 & \frac{1}{3} & 1 \\ 0 & 0 & 0 & \frac{1}{5} \end{bmatrix}$. **e)** $A = \begin{bmatrix} -1 & 1 & 0 & -1 \\ 0 & -2 & \frac{1}{2} & -1 \\ 0 & 0 & -\frac{5}{3} & 3 \\ 0 & 0 & 0 & -5 \end{bmatrix}$.

3. Determine los valores y vectores propios de cada una de las matrices que se dan.

   a) $A = \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 1 & 3 & 0 \\ 2 & 1 & -6 \end{bmatrix}$. **b)** $A = \begin{bmatrix} 4 & 0 & 0 \\ 5 & 4 & 0 \\ -2 & -3 & 4 \end{bmatrix}$. **c)** $A = \begin{bmatrix} 2{,}3 & 0 & 0 \\ 1{,}5 & -2 & 0 \\ 0 & 3{,}2 & -0{,}8 \end{bmatrix}$.

   d) $A = \begin{bmatrix} 10 & 0 & 0 & 0 \\ -2 & -1 & 0 & 0 \\ 4 & 7 & 2 & 0 \\ 5 & -2{,}3 & -1 & 4{,}5 \end{bmatrix}$. **e)** $A = \begin{bmatrix} -10 & 0 & 0 & 0 \\ -2 & -5 & 0 & 0 \\ 4 & 0 & -2 & 0 \\ 0 & -2{,}3 & 0 & -1 \end{bmatrix}$

4. Determine los valores y vectores propios de cada una de las matrices simétricas que se dan. Aplique el método de Newton para el cálculo de las raíces del polinomio característico. Determine también los discos de Gershgorin.

   a) $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -2 & 3 \\ 0 & 3 & 0 \end{bmatrix}$. **b)** $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. **c)** $A = \begin{bmatrix} 0 & 0 & -2 \\ 0 & 4 & 0 \\ -2 & 0 & 0 \end{bmatrix}$.

d) $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$. **e)** $A = \begin{bmatrix} 0 & -3 & -2 \\ -3 & 0 & -3 \\ -2 & -3 & 0 \end{bmatrix}$. **f)** $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

g) $A = \begin{bmatrix} 4 & 0 & -2 \\ 0 & 4 & -3 \\ -2 & -3 & 4 \end{bmatrix}$. **h)** $A = \begin{bmatrix} 2 & 1{,}5 & 0 \\ 1{,}5 & -2 & 3{,}2 \\ 0 & 3{,}2 & 0 \end{bmatrix}$. **i)** $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 4 \end{bmatrix}$.

5. Aplique el método de la potencia para aproximar los valores y vectores propios de cada una de las matrices que se dan. Realice 5 iteraciones y compare con el valor propio dominante exacto. Determine en cada caso los discos de Gershgorin.

a) $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 2 & 3 \\ 0 & 3 & 0 \end{bmatrix}$. **b)** $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. **c)** $A = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 4 & 0 \\ -2 & 0 & 0 \end{bmatrix}$.

d) $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. **e)** $A = \begin{bmatrix} 0 & -3 & -2 \\ -3 & 0 & -3 \\ -2 & -3 & 0 \end{bmatrix}$. **f)** $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

g) $A = \begin{bmatrix} 4 & 0 & -2 \\ 0 & 4 & -3 \\ -2 & -3 & 4 \end{bmatrix}$. **h)** $A = \begin{bmatrix} 2 & 1{,}5 & 0 \\ 1{,}5 & -2 & 3{,}2 \\ 0 & 3{,}2 & 0 \end{bmatrix}$. **i)** $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 4 \end{bmatrix}$.

8.6. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Calculus, Volumen 2, Segunda Edición, Editorial Reverté, Barcelona, 1975.
2. Owe Axelsson, Iterative Solution Methods, Editorial Cambridge University Press, Cambridge, 1996.
3. N. Bakhvalov, Metodos Numéricos, Editorial Paraninfo, Madrid, 1980.
4. Jérõme Bastien, Jean-Noël Martin, Introduction à L' Analyse Numérique, Editorial Dunod, París, 2003.
5. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.
6. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
7. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
8. P. G. Ciarlet, Introduction á L' Analyse Numérique Matricielle et á L' Optimisation, Editorial Masson, París, 1990.
9. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
10. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.
11. B. P. Demidowitsch, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.
12. James W. Demmel, Applied Numerical Linear Algebra, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.
13. V. N. Faddeva, Métodos de Cálculo de Algebra Lineal, Editorial Paraninfo, Madrid, 1967.

14. Francis G. Florey, Fundamentos de Algebra Lineal y Aplicaciones, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1980.

15. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.

16. Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence, Algebra Lineal, Editorial Publicaciones Cultutal, S. A., México, 1982.

17. Noel Gastinel, Análisis Numérico Lineal, Editorial Reverté, S. A., Barcelona, 1975.

18. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

19. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.

20. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

21. Nicholas J. Higham, Accuracy and Stability of Numerical Algorithms, Editorial Society for Industrial and Applied Mathematics, Philadelphia, 1996.

22. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.

23. Franz E. Hohn, Algebra de Matrices, Editorial Trillas, México, 1979.

24. Roger A. Horn, Charles R. Johnson, Matrix Analysis, Editorial Cambridge University Press, Cambrisge, 1999.

25. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.

26. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

27. Roland E. Larson, Bruce H. Edwards, Introducción al Algebra Lineal, editorial Limusa, Noriega Editores, México, 1995.

28. P. Lascaux, R. Théodor, Analyse Numérique Matricielle Appliquée à L' Art de L' Ingénieur, Tome 1, Editorial Masson, París, 1986.

29. P. Lascaux, R. Théodor, Analyse Numérique Matricielle Appliquée à L' Art de L' Ingénieur, Tome 2, Editorial Masson, París, 1987.

30. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

31. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

32. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

33. Ben Noble, James W. Daniel, Algebra Lineal Aplicada, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1989.

34. Beresford N. Parlett, The Symmetric Eingenvalue Problem, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

35. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

36. Fazlollah Reza, Los Espacios Lineales en la Ingeniería, Editorial Reverté, S. A., Barcelona, 1977.

37. Michelle Schatzman, Analyse Numérique, Inter Editions, París, 1991.

38. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

39. M. Sibony, J. Cl. Mardon, Analyse Numérique I, Sustèmes Linéaires et non Linéaires, Editorial Hermann, París, 1984.

40. G. W. Stewart, Matrix Algotithms, Volume II: Eingensystems, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

41. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

42. Gilbert Strang, Algebra Lineal y sus Aplicaciones, editorial Fondo Educativo Interamericano, México, 1982.

43. V. Voïévodine, Principes Numériques D' Algèbre Linéaire, Editions Mir, Moscú, 1976.

Capítulo 9

Mínimos Cuadrados

Resumen

El tratamiento de datos experimentales tiene lugar en este capítulo. Se tratan dos tipos de problemas: los discretos y los continuos. En el caso discreto, se comienza con el planteamiento del problema de mínimos cuadrados. A continuación se trata el método de Householder que constituye uno de los más importantes métodos de resolución de sistemas de ecuaciones lineales. Se considera el ajuste de datos de algunos problemas que conducen a resolver sistemas de ecuaciones lineales en mínimos cuadrados. A continuación se trata el ajuste de datos en los que hay que determinar un parámetro. Posteriormente, se trata el problema de mínimos cuadrados continuos y se da aplicaciones a la aproximación de las series de Fourier.

9.1. Introducción

En muchas observaciones científicas se deben determinar los valores de ciertas constantes $a_1, \ldots, a_n$. Sin embargo, determinar o medir dichas constantes resulta muy difícil y por lo general imposible. En tales casos, el método indirecto siguiente es aplicado: en vez de observar los a_i resulta más fácil tomar una muestra de una cantidad medible "y" la cual depende de los a_i y de las mediciones experimentales que denotamos por x, esto es,

$$y = f(x, a_1, \ldots, a_n).$$

Con el propósito de determinar a_i, se realizan experimentos bajo m condiciones diferentes $x_1, \ldots, x_m$, obteniéndose m resultados diferentes: $y_k = f(x_k, a_1, \ldots, a_n) \quad k = 1, \ldots, m$. En general, al menos $m \geq n$ experimentos deben ejecutarse con el propósito de determinar $a_i \quad i = 1, \ldots, n$. Además, estos valores a_i deben satisfacer la relación precedente.

Si $m > n$, $y_k = f(x_k, a_1, \ldots, a_n) \quad k = 1, \ldots, m$ forman un sistema sobredeterminado para los parámetros desconocidos $a_1, \ldots, a_n$ que usualmente no tiene solución porque las cantidades observadas y_i están perturbadas por errores de medición. Consecuentemente, en vez de encontrar una solución exacta de dicho sistema, el problema se traduce en encontrar una mejor aproximación posible aplicando la conocida técnica de mínimos cuadrados. Este método fue publicado por primera vez por Legendre en 1805. Esta clase de problemas se los conoce como ajuste de datos.

La función f es conocida. Esta función se elige siguiendo varios criterios:

1. Se tiene un modelo matemático gobernado, por ejemplo, por ecuaciones diferenciales ordinarias y que tiene como solución la función f que depende de n parámetros $a_i \quad i = 1, \ldots, n$ que deben determinarse de la información experimental existente.$S = \{(x_i, y_i) \mid i = 1, \ldots, n\}$.
2. En base al conjunto de datos experimentales $S = \{(x_i, y_i) \mid i = 1, \ldots, n\}$ representados gráficamente y por alguna información suplementaria se intuye que siguen un comportamiento del tipo $y = f(x, a_1, \ldots, a_n)$.

3. Dado el conjunto de datos $S = \{(x_i, y_i) \mid i = 1, \ldots, n\}$ se fija directamente la función f así como los parámetros a determinar $a_1, \ldots, a_n$, esto es $y = f(x, a_1, \ldots, a_n)$.

   En todos los casos suponemos que

   $$y_k = f(x_k, a_1, \ldots, a_n) + r_k \quad k = 1, \ldots, m,$$

   donde cada r_k es la perturbación del dato experimentas (x_k, y_k). Esta perturbación depende de los parámetros $a_1, \ldots, a_n$ que deben determinarse. Escribimos

   $$r_k(a_1, \ldots, a_n) = y_k - f(x_k, a_1, \ldots, a_n) \quad k = 1, \ldots, m.$$

   El método de mínimos cuadrados consiste en elegir la mejor opción de los parámetros $a_1, \ldots, a_n$ en el sentido que precisamos a continuación: sea $\Omega \subset \mathbb{R}^n$ abierto, se define la función E de Ω en $\mathbb{R}$ como sigue:

   $$E(a_1, \ldots, a_n) = \sum_{k=1}^{m} r_k^2(a_1, \ldots, a_n) = \sum_{k=1}^{m} (y_k - f(x_k, a_1, \ldots, a_n))^2 \quad (a_1, \ldots, a_n) \in \Omega,$$

   con lo que $E(a_1, \ldots, a_n) \geq 0 \quad \forall (a_1, \ldots, a_n) \in \Omega$; y se considera el problema siguiente:

   $$\text{hallar } (\widehat{a}_1, \ldots, \widehat{a}_n) \in \Omega \text{ tal que } E(\widehat{a}_1, \ldots, \widehat{a}_n) = \min_{(a_1, \ldots, a_n) \in \Omega} E(a_1, \ldots, a_n).$$

   Así, la mejor elección de los parámetros $a_1, \ldots, a_n$ es la solución del problema de minimización: $\min_{(a_1, \ldots, a_n) \in \Omega} E(a_1, \ldots, a_n)$.

   En la generalidad de los casos se supone que la función f es de clase C^1 en Ω, es decir que $\dfrac{\partial f}{\partial a_i}$ $i = 1, \ldots, n$, son continuas en Ω.

   Para hallar un punto $(\widehat{a}_1, \ldots, \widehat{a}_n) \in \Omega$ en el que la función E alcanza su mínimo, aplicamos las condiciones necesarias de extremo, esto es:

   $$\nabla E(a_1, \ldots, a_n) = \overrightarrow{0} \Leftrightarrow \begin{cases} \dfrac{\partial E}{\partial a_1}(a_1, \ldots, a_n) = 0, \\ \vdots \\ \dfrac{\partial E}{\partial a_n}(a_1, \ldots, a_n) = 0, \end{cases}$$

   con lo que obtenemos un sistema de ecuaciones lineal o no dependiendo exclusivamente de la función f. La solución de dicho sistema nos provee de un punto crítico $(\widehat{a}_1, \ldots, \widehat{a}_n)$ que puede ser en el que E alcanza su mínimo.

   Puesto que para cada $i = 1, \ldots, m$, se tiene

   $$\begin{aligned} \frac{\partial E}{\partial a_i}(a_1, \ldots, a_n) &= \sum_{k=1}^{m} 2(y_k - f(x_k, a_1, \ldots, a_n)) \left(-\frac{\partial f}{\partial a_i}(x_k, a_1, \ldots, a_n) \right) \\ &= -2 \sum_{k=1}^{m} \left[y_k \frac{\partial f}{\partial a_i}(x_k, a_1, \ldots, a_n) - \frac{\partial f}{\partial a_i}(x_k, a_1, \ldots, a_n) f(x_k, a_1, \ldots, a_n) \right], \end{aligned}$$

   se sigue que el sistema de ecuaciones precedente está definido como sigue:

   $$\begin{cases} \sum_{k=1}^{m} \left[y_k \dfrac{\partial f}{\partial a_1}(x_k, a_1, \ldots, a_n) - \dfrac{\partial f}{\partial a_1}(x_k, a_1, \ldots, a_n) f(x_k, a_1, \ldots, a_n) \right] = 0 \\ \vdots \\ \sum_{k=1}^{m} \left[y_k \dfrac{\partial f}{\partial a_n}(x_k, a_1, \ldots, a_n) - \dfrac{\partial f}{\partial a_n}(x_k, a_1, \ldots, a_n) f(x_k, a_1, \ldots, a_n) \right] = 0. \end{cases}$$

En el caso en que la función f es lineal respecto de las variables (parámetros a determinar) $a_1, \ldots, a_n$, el sistema de ecuaciones precedente es lineal. Dos ejemplos de funciones lineales nos proporcionan los polinomios de grado m :

$$f(x, a_0, \ldots, a_n) = a_0 + a_1 x + \cdots + a_n x^m \quad x \in \mathbb{R}, \quad (a_0, \ldots, a_n) \in \mathbb{R}^{n+1},$$

y los polinomios trigonométricos:

$$\begin{aligned} f(x, a_1, \ldots, a_m) &= \sum_{k=1}^{m} a_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \\ f(x, a_0, \ldots, a_m) &= \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \\ f(x, a_0, a_1, \ldots, a_m, b_1, \ldots, b_m) &= \frac{a_0}{2} + \sum_{k=1}^{m} \left[a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)\right], \end{aligned}$$

donde $L > 0$, $a_0, \ldots, a_m$, $b_1, \ldots, b_m$ son los coeficientes de Fourier.

En el caso de los polinomios trigonométricos, el cálculo aproximado de los coeficientes de Fourier ya fue establecido en el capítulo de la aproximación numérica de las series de funciones. En este capítulo mostraremos que los coeficientes que figuran en los polinomios trigonométricos son efectivamente los coeficientes de Fourier.

En el caso de los polinomios de grado m focalizaremos nuestra atención a los polinomios de grado 1, 2 y 3; esto es,

$$\begin{aligned} f(x, a, b) &= a + bx \quad x \in \mathbb{R}, \\ f(x, a, b, c) &= a + bx + cx^2 \quad x \in \mathbb{R}, \\ f(x, a, b, c, d) &= a + bx + cx^2 + dx^3 \quad x \in \mathbb{R}. \end{aligned}$$

En el caso en que f no es lineal respecto de las variables $a_1, \ldots, a_n$, el sistema de ecuaciones a resolver es no lineal. En la generalidad de los casos se tiene f función de clase C^2 en Ω, y dicho sistema se resuelve en forma aproximada con el método de Newton, la linealización del sistema de ecuaciones se resuelve mediante el método de eliminación gaussiana. En pocos casos se conocen de métodos particulares para resolver dicho sistema cuando dependen de 1, 2, 3 parámetros. Más adelante trataremos algunos de estos ejemplos que tienen muchas aplicaciones.

Los problemas de mínimos cuadrados que acabamos de describir se extienden a funciones en dos o más variables, así a polinomios de grado $1, 2, \ldots$, en dos variables independientes x, y como los que se describen a continuación:

$$\begin{aligned} p(x, y) &= a + bx + cy \quad (x, y) \in \mathbb{R}^2, \\ p(x, y) &= a + bx + cy + dxy \quad (x, y) \in \mathbb{R}^2, \\ p(x, y) &= a + bx + cy + dxy + ex^2 + dy^2 \quad (x, y) \in \mathbb{R}^2, \end{aligned}$$

que requieren, para el cálculo de las constantes que figuran en cada clase de polinomios, de un conjunto de datos $S = \left\{(x_k, y_k, z_k) \in \mathbb{R}^3 \mid k = 1, \ldots, n\right\}$. Otra clase de funciones son las conocidas funciones lineales y afines en n variables del tipo

$$z = f(x_1, \ldots, x_n) = \sum_{k=1}^{n} a_k x_n \quad (x_1, \ldots, x_n) \in \mathbb{R}^n$$

y

$$z = f(x_1, \ldots, x_n) = a_0 + \sum_{k=1}^{n} a_k x_n \quad (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

que requieren, para el cálculo de las constantes que figuran en cada una de esta clase de funciones, de un conjunto de datos

$$S = \left\{ \overrightarrow{x}_k = \left(x_1^{(k)}, \ldots, x_n^{(k)}, z_k \right) \in \mathbb{R}^{n+1} \mid k = 1, \ldots, m \right\}.$$

Comenzaremos con la resolución de los sistemas de ecuaciones lineales en mínimos cuadrados. Estos sistemas de ecuaciones provienen, en general, de problemas de ajuste de datos lineales. A continuación tratamos el ajuste de datos de datos polinomial. Luego tratamos el ajuste de datos de funciones afines del tipo

$$f(x_1, \ldots, x_n) = a_0 + \sum_{k=1}^{n} a_k x_n \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Concluimos con ejemplos de problemas de ajuste de datos de funciones no lineales.

9.2. Soluciones de sistemas de ecuaciones lineales en mínimos cuadrados.

Sean m, $n \in \mathbb{Z}^+$ con $m \geq n$, $A = (a_{ij}) \in M_{m\times n}[\mathbb{R}]$ no nula y de rango $\mathcal{R}(A) = n$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^m$. Consideramos el problema siguiente:

$$\text{hallar } \overrightarrow{x} \in \mathbb{R}^n \text{ solución del sistema de ecuaciones lineales } \quad A\overrightarrow{x} = \overrightarrow{b}.$$

Estos sistemas de ecuaciones se caracterizan por tener más ecuaciones que incógnitas. Estos sistemas, como hemos visto, surgen en la determinación de ciertos parámetros $x_1, ..., x_n$ que deben calcularse a partir de una información experimental y que corresponden a un modelo del tipo lineal.

Ponemos $A = [A_1, ..., A_n]$, donde A_j es la j-ésima columna de A y sea

$$W = L(A_1, ..., A_n) = \left\{ \sum_{j=1}^{n} \alpha_j A_j \mid \alpha_j \in \mathbb{R}, \ \ j = 1, ..., n \right\},$$

el espacio constituido por todas las combinaciones lineales de $A_1, ..., A_n$. Por definición, la dimensión de W es el rango de A y que por hipótesis, $\mathcal{R}(A) = n$. Luego $\dim W = n$. Entonces, el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ tiene solución si y solo si $\overrightarrow{b} \in W$. Esta situación se presenta en muy pocos casos. En la práctica, los sistemas de ecuaciones arriba propuesto, en general, no tienen solución.

Se define $\overrightarrow{r}(\overrightarrow{x}) = A\overrightarrow{x} - \overrightarrow{b}$, $\quad \overrightarrow{x} \in \mathbb{R}^n$. El vector $\overrightarrow{r}(\overrightarrow{x})$ se llama residuo. Proponemos un problema alterno denominado problema en mínimos cuadrados (P_a) que se indica a continuación:

$$\text{hallar, si existe, } \widehat{x} \in \mathbb{R}^n \text{ tal que } \quad \|\overrightarrow{r}(\widehat{x})\|^2 \leq \|\overrightarrow{r}(\overrightarrow{x})\|^2 \quad \forall \overrightarrow{x} \in \mathbb{R}^n,$$

o lo que es lo mismo

$$\left\| A\widehat{x} - \overrightarrow{b} \right\|^2 \leq \left\| A\overrightarrow{x} - \overrightarrow{b} \right\| \quad \forall \overrightarrow{x} \in \mathbb{R}^n,$$

que a su vez es equivalente al siguiente:

$$\text{hallar, si existe, } \widehat{x} \in \mathbb{R}^n \text{ tal que } \quad \left\| A\widehat{x} - \overrightarrow{b} \right\|^2 = \underset{\overrightarrow{x}\in\mathbb{R}^n}{\text{Min}} \left\| A\overrightarrow{x} - \overrightarrow{b} \right\|^2,$$

donde $\|\cdot\|$ es la norma euclídea en $\mathbb{R}^m$.

Este problema lo enfrentamos de dos maneras. En la primera, probamos la existencia de $\widehat{x} \in \mathbb{R}^n$ mediante métodos del análisis matemático como minimización de un cierto funcional definido en $\mathbb{R}^n$. En la segunda, probamos la existencia de $\widehat{x} \in \mathbb{R}^n$ mediante métodos netamente del álgebra lineal utilizando la ortogonalidad.

1. **Minimización de un funcional cuadrático definido en $\mathbb{R}^n$.**

   Sea J de $\mathbb{R}^n$ en $\mathbb{R}$ el funcional definido por $J(\overrightarrow{x}) = \left\|A\overrightarrow{x} - \overrightarrow{b}\right\|^2$, $\quad \overrightarrow{x} \in \mathbb{R}^n$. Se sabe que J es un funcional convexo. Hallemos la derivada de Gâteaux de J, esto es, la derivada direccional de J en $\overrightarrow{x} \in \mathbb{R}^n$ según la dirección $\overrightarrow{y}$, denotada $D_{\overrightarrow{y}}J(\overrightarrow{x})$ y definida como

$$D_{\overrightarrow{y}}J(\overrightarrow{x}) = \lim_{t\to 0}\frac{J(\overrightarrow{x}+t\overrightarrow{y}) - J(\overrightarrow{x})}{t}.$$

   De la definición del producto escalar en $\mathbb{R}^n$, se tiene

$$J(\overrightarrow{x}) = \left\|A\overrightarrow{x} - \overrightarrow{b}\right\|^2 = \left(A\overrightarrow{x} - \overrightarrow{b}\right)^T\left(A\overrightarrow{x} - \overrightarrow{b}\right) \qquad \overrightarrow{x} \in \mathbb{R}^n.$$

   Sean $\overrightarrow{x}$, $\overrightarrow{y} \in \mathbb{R}^n$ fijos, $t \neq 0$. Entonces,

$$\begin{aligned} J(\overrightarrow{x}+t\overrightarrow{y}) - J(\overrightarrow{x}) &= \left(A(\overrightarrow{x}+t\overrightarrow{y}) - \overrightarrow{b}\right)^T\left(A(\overrightarrow{x}+t\overrightarrow{y}) - \overrightarrow{b}\right) - \left(A\overrightarrow{x} - \overrightarrow{b}\right)^T\left(A\overrightarrow{x} - \overrightarrow{b}\right) \\ &= 2t\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y} + t^2\left(A\overrightarrow{y}\right)^T A\overrightarrow{y}. \end{aligned}$$

   Se ha utilizado el hecho que $A\overrightarrow{x} - \overrightarrow{b}$, $A\overrightarrow{x} \in \mathbb{R}^m$, y, $\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y} = \left(A\overrightarrow{y}\right)^T\left(A\overrightarrow{x} - \overrightarrow{b}\right)$, pués el producto escalar en $\mathbb{R}^m$ es conmutativo. Luego, para $t \neq 0$ se tiene

$$\frac{J(\overrightarrow{x}+t\overrightarrow{y}) - J(\overrightarrow{x})}{t} = 2\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y} + t\left(A\overrightarrow{y}\right)^T A\overrightarrow{y},$$

   de donde

$$\begin{aligned} D_{\overrightarrow{y}}J(\overrightarrow{x}) &= \lim_{t\to 0}\frac{J(\overrightarrow{x}+t\overrightarrow{y}) - J(\overrightarrow{x})}{t} = \lim_{t\to 0}\left(2\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y}\right) + t\left(A\overrightarrow{y}\right)^T A\overrightarrow{y} \\ &= 2\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y}. \end{aligned}$$

   Así, la derivada de Gâteaux de J en $\overrightarrow{x}$ según la dirección $\overrightarrow{y}$ está definida como

$$D_{\overrightarrow{y}}J(\overrightarrow{x}) = 2\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y}.$$

   Por otro lado, un resultado muy conocido del Análisis Matemático sobre las derivadas direccionales es que si $D_{\overrightarrow{y}}J(\overrightarrow{x})$ es continuo en $\overrightarrow{x} \in \mathbb{R}^n$ en toda dirección $\overrightarrow{y} \in \mathbb{R}^n$, entonces

$$D_{\overrightarrow{y}}J(\overrightarrow{x}) = \left(\nabla J(\overrightarrow{x})\right)^T\overrightarrow{y},$$

   donde $\nabla J(\overrightarrow{x})$ denota el gradiente de J en $\overrightarrow{x}$ definido por $\left(\nabla J(\overrightarrow{x})\right)^T = \left(\frac{\partial J}{\partial x_1}(\overrightarrow{x}), ..., \frac{\partial J}{\partial x_n}(\overrightarrow{x})\right)$. Luego,

$$\left(\nabla J(\overrightarrow{x})\right)^T\overrightarrow{y} = 2\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\overrightarrow{y} \qquad \forall\overrightarrow{y} \in \mathbb{R}^n,$$

   con lo cual

$$\nabla J(\overrightarrow{x}) = \left(A\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T A\right)^T = 2A^T\left(A\overrightarrow{x} - \overrightarrow{b}\right)^T.$$

   Las condiciones necesarias de extremo implican

$$\nabla J(\overrightarrow{x}) = 0 \Leftrightarrow A^T\left(A\overrightarrow{x} - \overrightarrow{b}\right) = 0 \Leftrightarrow A^TA\overrightarrow{x} = A^T\overrightarrow{b}.$$

   El sistema de ecuaciones lineales $A^TA\overrightarrow{x} = A^T\overrightarrow{b}$, se llama sistema de ecuaciones normales. Note que la matriz A^TA es una matriz simétrica y en consecuencia es una matriz normal.

Por hipótesis $R(A) = n$. Luego $R\left(A^T\right) = n$ y $R\left(A^T A\right) = n$. Como $A^T A \in M_{n\times n}[\mathbb{R}]$ y $R\left(A^T A\right) = n$, se sigue que $A^T A$ es invertible y en consecuencia

$$A^T A\overrightarrow{x} = A^T\overrightarrow{b} \Leftrightarrow \overrightarrow{x} = \left(A^T A\right)^{-1} A^T\overrightarrow{b}.$$

Ponemos

$$\widehat{x} = \left(A^T A\right)^{-1} A^T\overrightarrow{b}.$$

Probemos que $\left\|A\widehat{x} - \overrightarrow{b}\right\|^2 \leq \left\|A\overrightarrow{x} - \overrightarrow{b}\right\|^2 \quad \forall\overrightarrow{x} \in \mathbb{R}^n$. Sea $\overrightarrow{x} \in \mathbb{R}^n$. Entonces

$$\begin{aligned}\|\overrightarrow{r}(\overrightarrow{x})\|^2 &= (\overrightarrow{r}(\overrightarrow{x}))^T \overrightarrow{r}(\overrightarrow{x}) = [A(\overrightarrow{x} - \widehat{x}) + \overrightarrow{r}(\overrightarrow{x})]^T [A(\overrightarrow{x} - \widehat{x}) + \overrightarrow{r}(\overrightarrow{x})] \\ &= [A(\overrightarrow{x} - \widehat{x})]^T A(\overrightarrow{x} - \widehat{x}) + (\overrightarrow{r}(\widehat{x}))^T \overrightarrow{r}(\widehat{x}) = \|A(\overrightarrow{x} - \widehat{x})\|^2 + \|\overrightarrow{r}(\widehat{x})\|^2,\end{aligned}$$

donde el producto $(A(\overrightarrow{x} - \widehat{x}))^T \overrightarrow{r}(\widehat{x}) = 0$, pues $A^T\left(A\widehat{x} - \overrightarrow{b}\right) = 0$ y en consecuencia

$$[A(\overrightarrow{x} - \widehat{x})]^T \overrightarrow{r}(\widehat{x}) = [A(\overrightarrow{x} - \widehat{x})]^T \left(A\widehat{x} - \overrightarrow{b}\right) = (\overrightarrow{x} - \widehat{x})^T A^T\left(A\widehat{x} - \overrightarrow{b}\right) = 0.$$

Por lo tanto,

$$\|\overrightarrow{r}(\overrightarrow{x})\|^2 = \|A(\overrightarrow{x} - \widehat{x})\|^2 + \|\overrightarrow{r}(\widehat{x})\|^2$$

y como $\|A(\overrightarrow{x} - \widehat{x})\|^2 \geq 0$, se sigue que $\|\overrightarrow{r}(\widehat{x})\|^2 \leq \|\overrightarrow{r}(\overrightarrow{x})\|^2 \quad \forall\overrightarrow{x} \in \mathbb{R}^n$, es decir

$$\left\|A\widehat{x} - \overrightarrow{b}\right\|^2 = \underset{\overrightarrow{x}\in\mathbb{R}^n}{\text{Min}} \left\|A\overrightarrow{x} - \overrightarrow{b}\right\|^2.$$

A la solución $\widehat{x}$ del problema (P_a) lo denominaremos solución en mínimos cuadrados.

2. **Proyección ortogonal.**

Ponemos $A = [A_1, ..., A_n]$, con A_j la j-ésima columna de la matriz A. Sea $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$. Entonces

$$A\overrightarrow{x} = \sum_{j=1}^{n} x_j A_j \in W.$$

El ortogonal de W, por definición, es el conjunto notado $W^\perp$ y definido como sigue:

$$\begin{aligned}W^\perp &= \{\overrightarrow{y} \in \mathbb{R}^m \mid \langle\overrightarrow{y}, A\overrightarrow{x}\rangle = 0 \quad \forall\overrightarrow{x} \in \mathbb{R}^n\} = \left\{\overrightarrow{y} \in \mathbb{R}^m \mid (A\overrightarrow{x})^T\overrightarrow{y} = 0 \quad \forall\overrightarrow{x} \in \mathbb{R}^n\right\} \\ &= \left\{\overrightarrow{y} \in \mathbb{R}^m \mid \overrightarrow{x}^T A^T\overrightarrow{y} = 0 \quad \forall\overrightarrow{x} \in \mathbb{R}^m\right\}.\end{aligned}$$

Luego

$$\overrightarrow{y} \in W^\perp \Leftrightarrow \overrightarrow{y} \in \ker\left(A^T\right) = \left\{\overrightarrow{y} \in \mathbb{R}^m \mid A^T\overrightarrow{y} = 0\right\}.$$

Se tiene la siguiente suma directa $\mathbb{R}^m = W \oplus W^\perp$, y de esta, para cada $\overrightarrow{b} \in \mathbb{R}^m$, existe un único $\widehat{x} \in \mathbb{R}^n$ y $\widehat{y} \in W^\perp$ tales que $\begin{cases} A\widehat{x} \perp \widehat{y}, \\ \overrightarrow{b} = A\widehat{x} + \widehat{y}, \end{cases}$ de donde $\widehat{y} = \overrightarrow{b} - A\widehat{x}$. En la figura siguiente se ilustran el subespacio W y su ortogonal $W^\perp$, el vector $A\widehat{x} \in W$ y el vector ortogonal $\overrightarrow{y} \in W^\perp$.

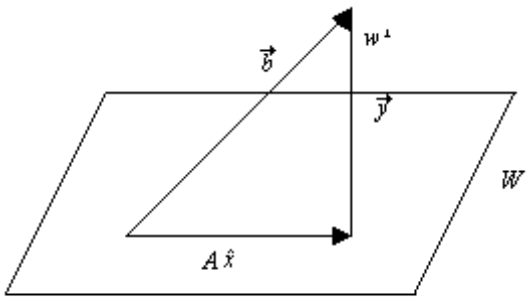


Figura 82

Sea $\overrightarrow{x} \in \mathbb{R}^n$. Entonces $A\overrightarrow{x} \in W$ y en consecuencia

$$(A\overrightarrow{x})^T \widehat{y} = 0 \Leftrightarrow (A\overrightarrow{x})^T \left(\overrightarrow{b} - A\widehat{x}\right) = 0 \Leftrightarrow \overrightarrow{x}^T A^T \left(\overrightarrow{b} - A\widehat{x}\right) = 0 \Leftrightarrow \overrightarrow{x}^T \left(A^T\overrightarrow{b} - A^T A\widehat{x}\right) = 0,$$

de donde

$$A^T A\widehat{x} = A^T\overrightarrow{b}.$$

Resulta que $\|\widehat{y}\|^2 \leq \|\overrightarrow{y}\|^2 \quad \forall \overrightarrow{y} \in W^{\perp}$, o bien

$$\left\|\overrightarrow{b} - A\widehat{x}\right\|^2 = \underset{\overrightarrow{x}\in\mathbb{R}^n}{\text{Min}} \left\|\overrightarrow{b} - A\overrightarrow{x}\right\|^2.$$

El resultado que acabamos de obtener se le conoce como proyección de un vector $\overrightarrow{b} \in \mathbb{R}^m$ con $\overrightarrow{b} \notin W$, sobre el subespacio cerrado W de $\mathbb{R}^m$. El vector $\widehat{x} \in \mathbb{R}^n$ se le conoce como solución en mínimos cuadrados.

Observaciones

1. El sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$, con $m, n \in \mathbb{Z}^+$, $m \geq n$, $A = (a_{ij}) \in M_{m\times n}[\mathbb{R}]$ no nula y de rango $\mathcal{R}(A) = n$, $\overrightarrow{b}^T = (b_1, ..., b_n) \in \mathbb{R}^m$, puede ser resuelto directamente aplicando el método de factorización QR de Householder, que se expone más adelante.

2. Supóngase que $A \in M_{n\times n}[\mathbb{R}]$ matriz invertible y consideremos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, donde $\overrightarrow{b} \in \mathbb{R}^n$ dado, cuya solución en mñinimos cuadrados es $\widehat{x} = \left(A^T A\right)^{-1} A^T\overrightarrow{b}$. Como A es invertible, se sabe que A^T es también invertible y en consecuencia existe $\left(A^T\right)^{-1}$ tal que $\left(A^T\right)^{-1} A^T = I$. Luego

$$\widehat{x} = \left(A^T A\right)^{-1} A^T\overrightarrow{b} = A^{-1}\left(A^T\right)^{-1} A^T\overrightarrow{b} = A^{-1}\overrightarrow{b},$$

que muestra que la solución en mínimos cuadrados coincide con la solución del sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

3. En la práctica, la solución en mínimos cuadrados se calcula como sigue: del sistema de ecuaciones normal $A^T A\overrightarrow{x} = A^T\overrightarrow{b}$, se definen $B = A^T A$, $\overrightarrow{c} = A^T\overrightarrow{b}$, con lo que dicho sistema se escribe como $B\overrightarrow{x} = \overrightarrow{c}$, sistema de ecuaciones lineales que puede ser resuelto mediante el método de eliminación gaussiana, de Choleski o de factorización LU, dependiendo de las características de la matriz B.

4. Sean $m, n \in \mathbb{Z}^+$ con $m \leq n$, $V = \mathbb{R}^n$, $\{\overrightarrow{y}_1, \ldots, \overrightarrow{y}_m\} \subset \mathbb{R}^n$ tal que$\{\overrightarrow{y}_1, \ldots, \overrightarrow{y}_m\}$ linealmente independiente,

$$W = \left\{\sum_{i=1}^{m} x_i\overrightarrow{y}_i \mid x_i \in \mathbb{R},\ i = 1, \ldots, m\right\}.$$

Se tiene que W es un subespacio de $\mathbb{R}^n$ de dimensión m. Del resultado arriba establecido de la proyección de un vector $\overrightarrow{b} \in \mathbb{R}^n$ con $\overrightarrow{b} \notin W$, sobre el subespacio cerrado W, existe $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_m) \in \mathbb{R}^m$ tal que

$$\left\|\overrightarrow{b} - \sum_{i=1}^{m} \widehat{x}_i\overrightarrow{y}_i\right\|^2 = \underset{\overrightarrow{x}=(x_1,\ldots,x_m)\in\mathbb{R}^m}{\text{Min}} \left\|\overrightarrow{b} - \sum_{i=1}^{m} x_i\overrightarrow{y}_i\right\|^2.$$

Además, se prueba que

$$\left\langle \overrightarrow{b} - \sum_{i=1}^{m} \widehat{x}_i\overrightarrow{y}_i, \overrightarrow{w} \right\rangle = 0 \quad \forall \overrightarrow{w} \in W.$$

En la figura siguiente se ilustra la solución en mínimos cuadrados.

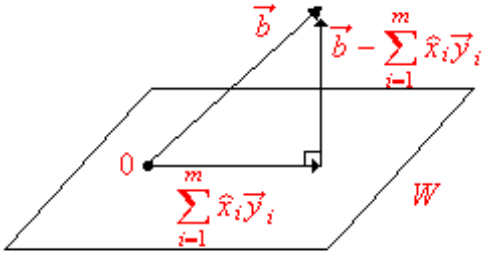


Figura 83

9.3. Método de Householder y mínimos cuadrados

Sea $A \in M_{n\times n}[\mathbb{R}]$. Supongamos que las columnas de A son ortogonales, entonces el rango de A es n, además A es invertible. En efecto, escribamos la matriz A en la forma $A = [A_1, \ldots, A_n]$, donde A_j es la j-ésima columna de A, luego $A^T = \begin{bmatrix} A_1^T \\ \vdots \\ A_n^T \end{bmatrix}$ y

$$A^T A = \begin{bmatrix} A_1^T \\ \vdots \\ A_n^T \end{bmatrix} [A_1, \ldots, A_n] = \begin{bmatrix} A_1^T A_1 & \ldots & A_1^T A_n \\ & \vdots & \\ A_n^T A_1 & \ldots & A_n^T A_n \end{bmatrix}$$

Como las columnas de A son ortogonales se sigue que $A_i^T A_j = 0$ si $i \neq j$, y $A_i^T A_i = \|A_i\|^2 = 1$.

Luego $A^T A = I$ y de esta relación resulta que $A^T = A^{-1}$.

Consideremos el sistema de ecuaciones lineales $A\vec{x} = \vec{b}$ con $\vec{b} \in \mathbb{R}^n$ dado. Entonces

$$\vec{x} = A^{-1}\vec{b} = A^T\vec{b}.$$

El siguiente resultado constituye la base del algoritmo de Householder para la resolución de sistemas de ecuaciones lineales y factorización de una matriz A en la forma QR.

Teorema 1 ***(de Householder)***
Sea $\vec{v} \in \mathbb{R}^n$ con $\vec{v} \neq 0$. Existe una matriz ortogonal H y $\alpha \in \mathbb{R}$ tales que

$$H\vec{v} = \alpha\vec{e}_1,$$

de donde $\vec{e}_1^T = (1, 0, ..., 0)$ es el primer vector de la base canónica de $\mathbb{R}^n$.

Demostración. Sea $\vec{u} \in \mathbb{R}^n$ tal que $\|\vec{u}\| = 1$. La matriz de Householder

$$H = I - 2\vec{u}\,\vec{u}^T$$

es simétrica y ortogonal.

Sea $\vec{v} \in \mathbb{R}^n$ con $\vec{v} \neq 0$. Mostremos que existe $\vec{u} \in \mathbb{R}^n$ tal que $\|\vec{u}\| = 1$ y $H\vec{v} = \alpha\vec{e}_1$. En efecto, sea $\alpha \in \mathbb{R}$ tal que $\|\vec{v}\| = |\alpha|$. Entonces

$$H\vec{v} = \left(I - 2\vec{u}\,\vec{u}^T\right)\vec{v} = \vec{v} - 2\vec{u}\,\vec{u}^T\vec{v},$$

y como $H\overrightarrow{v} = \alpha\overrightarrow{e}_1$, se sigue que

$$\begin{aligned} \overrightarrow{v} - 2\overrightarrow{u}\,\overrightarrow{u}^T\overrightarrow{v} &= \alpha\overrightarrow{e}_1 \\ 2\overrightarrow{u}\,\overrightarrow{u}^T\overrightarrow{v} &= \overrightarrow{v} - \alpha\overrightarrow{e}_1. \end{aligned}$$

Sea $p = 2\overrightarrow{u}^T\overrightarrow{v}$. Se tiene

$$p\overrightarrow{u} = \overrightarrow{v} - \alpha\overrightarrow{e}_1$$

de donde

$$\|\overrightarrow{v} - \alpha\overrightarrow{e}_1\| = \|p\overrightarrow{u}\| = |p|\,\|\overrightarrow{u}\| = |p|\,.$$

Para evitar que $p = 0$, elegimos α del modo siguiente:

$$\overrightarrow{v} - \alpha\overrightarrow{e}_1 = \begin{bmatrix} v_1 - \alpha \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

entonces

$$\alpha = \begin{cases} -\operatorname{sign}(v_1)\,\|\overrightarrow{v}\|\,, & si \quad v_1 \neq 0, \\ -\|\overrightarrow{v}\|\,, & si \quad v_1 = 0. \end{cases}$$

Supongamos primeramente que $v_1 \neq 0$. Se tiene $\alpha = -\operatorname{sign}(v_1)\,\|\overrightarrow{v}\|$, luego

$$\|\overrightarrow{v} - \alpha\overrightarrow{e}_1\|^2 = \|\overrightarrow{v} + \operatorname{sign}(v_1)\,\|\overrightarrow{v}\|\,\overrightarrow{e}_1\|^2 = (v_1 + \operatorname{sign}(v_1)\,\|\overrightarrow{v}\|)^2 + \sum_{k=2}^{n} v_k^2.$$

Si $v_1 > 0$, $\operatorname{sign}(v_1) = 1$, y

$$v_1 + \operatorname{sign}(v_1)\,\|\overrightarrow{v}\| = v_1 + \|\overrightarrow{v}\|\,.$$

Si $v_1 < 0$, $\operatorname{sign}(v_1) = -1$, y

$$v_1 + \operatorname{sign}(v_1)\,\|\overrightarrow{v}\| = v_1 - \|\overrightarrow{v}\| = -\left(-v_1 + \|\overrightarrow{v}\|\right) = -\left(|v_1| + \|\overrightarrow{v}\|\right).$$

Por lo tanto

$$\begin{aligned} \|\overrightarrow{v} - \alpha\overrightarrow{e}_1\| &= (|v_1| + \|\overrightarrow{v}\|)^2 + \sum_{k=2}^{n} v_k^2 = v_1^2 + 2\,|v_1|\,\|\overrightarrow{v}\| + \|\overrightarrow{v}\|^2 + \sum_{k=2}^{n} v_k^2 \\ &= 2\,|v_1|\,\|\overrightarrow{v}\| + \|\overrightarrow{v}\|^2 + \sum_{k=1}^{n} v_k^2 = 2\,|v_1|\,\|\overrightarrow{v}\| + 2\,\|\overrightarrow{v}\|^2\,. \end{aligned}$$

Si $v_1 = 0$, $\alpha = -\|\overrightarrow{v}\|$. Entonces

$$\|\overrightarrow{v} - \alpha\overrightarrow{e}_1\|^2 = \|\overrightarrow{v} + \|\overrightarrow{v}\|\,\overrightarrow{e}_1\|^2 = \|\overrightarrow{v}\|^2 + \sum_{k=2}^{n} v_k^2 = \|\overrightarrow{v}\|^2 + \sum_{k=1}^{n} v_k^2 = 2\,\|\overrightarrow{v}\|^2\,.$$

Consecuentemente, de la igualdad $p\overrightarrow{u} = \overrightarrow{v} - \alpha\overrightarrow{e}_1$, se sigue que

$$\begin{aligned} \overrightarrow{u} &= \frac{\overrightarrow{v} - \alpha\overrightarrow{e}_1}{p} = \frac{\overrightarrow{v} - \alpha\overrightarrow{e}_1}{\|\overrightarrow{v} - \alpha\overrightarrow{e}_1\|} = \frac{\overrightarrow{v} - \alpha\overrightarrow{e}_1}{\left(2\,\|\overrightarrow{v}\|^2 + 2\,|v_1|\,\|\overrightarrow{v}\|\right)^{\frac{1}{2}}} \quad si \quad v_1 \neq 0, \\ \overrightarrow{u} &= \frac{\overrightarrow{v} - \alpha\overrightarrow{e}_1}{\sqrt{2}\,\|\overrightarrow{v}\|} \quad si \quad v_1 = 0. \end{aligned}$$

La matriz H queda definida como

$$H = I - 2\overrightarrow{u}\,\overrightarrow{u}^T = I - \frac{1}{\|\overrightarrow{v}\|^2 + |v_1|\,\|\overrightarrow{v}\|}\left(\overrightarrow{v} - \alpha\overrightarrow{e}_1\right)\left(\overrightarrow{v} - \alpha\overrightarrow{e}_1\right)^T \text{ si } v_1 \neq 0,$$

y

$$H = I - \frac{1}{\|\overrightarrow{v}\|^2}\left(\overrightarrow{v} - \alpha\overrightarrow{e}_1\right)\left(\overrightarrow{v} - \alpha\overrightarrow{e}_1\right)^T, \text{ si } v_1 = 0.$$

Si ponemos

$$\begin{aligned} r &= \begin{cases} \|\overrightarrow{v}\|^2 + |v_1|\,\|\overrightarrow{v}\|, & si \quad v_1 \neq 0, \\ 2\|\overrightarrow{v}\|^2, & si \quad v_1 = 0, \end{cases} \\ \overrightarrow{w} &= \overrightarrow{v} - \alpha\overrightarrow{e}_1, \end{aligned}$$

entonces

$$H = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T.$$

■

Observación

i) Se puede escribir $H = I - 2\dfrac{(s\overrightarrow{u})(s\overrightarrow{u})^t}{s^2}$; el cálculo explícito de $\overrightarrow{u}$ no es necesario.

Ponemos $r = \dfrac{s^2}{2}$ y $\overrightarrow{w} = s\overrightarrow{u}$ entonces $H = I - \dfrac{1}{r}\overrightarrow{w}\overrightarrow{w}^T$ donde $\overrightarrow{w}$ y r se determinan por el sistema:

$$\begin{cases} |\alpha| = \|\overrightarrow{v}\|, \\ \overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_1, \\ r = \alpha(\alpha - v_1). \end{cases}$$

ii) El signo de α se tomará el opuesto al de v_1

iii) Para calcular $H\overrightarrow{v}$ podemos proceder de la siguiente manera:

$$H\overrightarrow{v} = \left(I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T\right)\overrightarrow{v} = \overrightarrow{v} - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T\overrightarrow{v} = \overrightarrow{v} - \frac{1}{r}\left(\overrightarrow{w}^T\overrightarrow{v}\right)\overrightarrow{w},$$

cuyos cálculos sucesivos son:

$$\begin{aligned} q &= \overrightarrow{w}^T\overrightarrow{v}, \\ p &= \frac{q}{r}, \\ H\overrightarrow{v} &= \overrightarrow{v} - p\overrightarrow{w}. \end{aligned}$$

Algoritmo de Householder

Sea $A \in M_{n\times n}[\mathbb{R}]$ invertible, $\overrightarrow{b} \in \mathbb{R}^n$ dado y consideramos el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

Aplicaremos el lema 2, m veces para triangularizar la matriz A. Ponemos $A^{(1)} = A$, $\overrightarrow{b}^{(1)} = \overrightarrow{b}$. Generaremos $m-1$ ecuaciones equivalentes a la ecuación original propuesta: $A^{(k)}\overrightarrow{x} = \overrightarrow{b}^{(k)} \quad k = 2, \ldots, n$ donde $A^{(k)}$ y $\overrightarrow{b}^{(k)}$ son de la forma:

$$A^{(k)} = \left[\begin{array}{cccc|ccc} a_{11}^{(2)} & \ldots & \ldots & a_{1k-1}^{(2)} & a_{1k}^{(2)} & \ldots & a_{1n}^{(2)} \\ & a_{22}^{(3)} & \ldots & a_{2k-1}^{(3)} & \vdots & & \vdots \\ & & \ddots & \vdots & \vdots & & \vdots \\ 0 & & & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \ldots & a_{k-1,n}^{(k)} \\ \hline & & & & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ & & 0 & & \vdots & & \vdots \\ & & & & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{array}\right], \quad \overrightarrow{b}^{(k)} = \left[\begin{array}{c} b_1^{(2)} \\ \vdots \\ b_{k-1}^{(k)} \\ \hline b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{array}\right]$$

que de manera abreviada se escribe:

$$A^{(k)} = \left[\begin{array}{c|c} A_{11}^{(k)} & A_{12}^{(k)} \\ \hline 0 & A_{22}^{(k)} \end{array}\right], \quad \overrightarrow{b}^{(k)} = \left[\begin{array}{c} \overrightarrow{c}^{(k)} \\ \hline \overrightarrow{d}^{(k)} \end{array}\right]$$

Para pasar de $A^{(k)}$ a $A^{(k+1)}$ y de $\overrightarrow{b}^{(k)}$ a $\overrightarrow{b}^{(k+1)}$, buscamos una matriz ortogonal elemental (matrix de Householder) $H^{(k)}$ tal que $H^{(k)}A^{(k)}$ es una matriz cuyas primeras k columnas forman una matriz triangular de la forma de $A_{11}^{(k)}$.

Tomamos un vector $\overrightarrow{u}^{(k)}$ tal que $u_1^{(k)} = u_2^{(k)} = \cdots = u_{k-1}^{(k)} = 0$ y notamos con $\widetilde{u}^{(k)}$ el vector $(u_k, \ldots, u_n) \in \mathbb{R}^{n-k+1}$ y

$$\widetilde{H}^{(k)} = I_{n-k+1} - 2\widetilde{u}^{(k)}\widetilde{u}^{(k)T}.$$

La matriz $H\,(k) = I - 2u^{(k)}u^{(k)T}$ es tal que $H^{(k)}A^{(k)}$ deja fijas las $k-1$ filas y columnas de $A^{(k)}$, además $H^{(k)}\overrightarrow{b}^{(k)}$ no modifica las $k-1$ primeros elementos de $\overrightarrow{b}^{(k)}$. Luego $H^{(k)}A^{(k)}$ se escribe:

$$\begin{aligned} H^{(k)}A^{(k)} &= \left[\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & \widetilde{H}^{(k)} \end{array}\right]\left[\begin{array}{c|c} A_{11}^{(k)} & A_{12}^{(k)} \\ \hline 0 & A_{22}^{(k)} \end{array}\right] = \left[\begin{array}{c|c} A_{11}^{(k)} & A_{12}^{(k)} \\ \hline 0 & \widetilde{H}^{(k)}A_{22}^{(k)} \end{array}\right] \\ H^{(k)}\overrightarrow{b}^{(k)} &= \left[\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & \widetilde{H}^{(k)} \end{array}\right]\left[\begin{array}{c} \overrightarrow{c}^{(k)} \\ \hline \overrightarrow{d}^{(k)} \end{array}\right] = \left[\begin{array}{c} \overrightarrow{c}^{(k)} \\ \hline \widetilde{H}^{(k)}\overrightarrow{d}^{(k)} \end{array}\right]. \end{aligned}$$

Como $\det\left(A^{(k)}\right) = \left[\prod_{i=1}^{k-1} a_{ii}^{(i+1)}\right] \det\left(A_{22}^{(k)}\right) \neq 0$, los elementos de la primera columna de $A_{22}^{(k)}$ no son nulos, consecuentemente se puede aplicar el lema2 con $\overrightarrow{v} = \left(a_{kk}^{(k)}, \ldots, a_{nk}^{(k)}\right) \in \mathbb{R}^{n-k+1}$ para anular todos salvo el primero; cuyo algoritmo de cálculo es:

$$\begin{aligned} a_{kk}^{(k-1)} &= -sign\left(a_{kk}^{(k)}\right)\left(\sum_{i=k}^{n}\left[a_{ik}^{(k)}\right]^2\right)^{\frac{1}{2}}, \\ r^{(k)} &= a_{kk}^{(k+1)}\left(a_{kk}^{(k+1)} - a_{kk}^{(k)}\right), \\ w_k^{(k)} &= a_{kk}^{(k)} - a_{kk}^{(k+1)}, \\ w_i^{(k)} &= a_{ik}^{(k)} \quad i = k+1, \ldots, n \end{aligned}$$

que permiten determinar $H^{(k)}$ cuyo cálculo de $\widetilde{H}^{(k)}A_{22}^{(k)}$ es el siguiente:

$$\begin{aligned} q_j^{(k)} &= \sum_{i=k}^{n} w_i^{(k)}a_{ij}^{(k)}, \\ p_j^{(k)} &= \frac{q_i^{(k)}}{r^{(k)}} \quad j = k+1, \ldots, n. \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - p_j^{(k)}w_i, \quad i = k+1, \ldots, n \end{aligned}$$

Para el cálculo de $\widetilde{H}^{(k)}\overrightarrow{b}^{(k)}$ se emplea el siguiente algoritmo;

$$\begin{aligned} q^{(k)} &= \sum_{i=k}^{n} w_i^{(k)}b_i^{(k)}, \\ p^{(k)} &= \frac{q^{(k)}}{r^{(k)}}, \\ b_i^{(k+1)} &= b_i^{(k)} - p^{(k)}w_i^{(k)} \quad i = k, \ldots, n. \end{aligned}$$

Finalmente, ponemos $H^{(k)}A^{(k)} = A^{(k+1)}$ y $H^{(k)}\overrightarrow{b}^{(k)} = \overrightarrow{b}^{(k+1)}$.

La matriz $A^{(n)}$ es una matriz triangular superior que permite resolver fáclmente el sistema: $A^{(n)}\overrightarrow{x} = \overrightarrow{b}^{(n)}$.

Observación

1. El método de Housholder permite calcular $\det(A)$. Pués de la factorización siguiente:

$$A^{(n)} = H^{(n-1)}H^{(n-2)}\dots H^{(1)}A^{(1)},$$

y como $A^{(1)} = A$ y $\det\left(H^{(k)}\right) = -1 \quad \forall k = 1, \cdots, n$, se sigue que

$$\det(A) = (-1)^{n-1}\det\left(A^{(n)}\right) = (-1)^{n-1}\prod_{i=1}^{n} a_{ii}^{(n)}.$$

2. Para mejorar la estabilidad del método, podemos proceder como sigue: en la k-ésima etapa, en lugar de transformar la k-ésima columna de $A^{(k)}$, se elige entre las últimas columnas de $A^{(k)}$ aquel elemento que hace: $a_{ij}^{(k)} = \sum_{i=k}^{n}\left(a_{ij}^{(k)}\right)^2$ máximo. Se permuta aquella con la k-ésima columna, si en $k = L$ se alcanza tal máximo, entonces

$$a_{kk}^{(k+1)} = -sign\left(a_{kk}^{(k)}\right)\sqrt{a_L^{(k)}}.$$

Ejemplos

1. Sea $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Construyamos la matriz de Householder y factoremos la matriz en la forma $A = QR$ con Q una matriz ortogonal y R una matriz triangular superior.

   Sea $\overrightarrow{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, se tiene $\alpha = -1$,

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

   Además $r = \alpha(\alpha - v_1) = -1(-1-2) = 2$. Luego

$$\begin{aligned} H &= I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 2 \\ 0 \end{bmatrix}(2,0) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \\ A^{(1)} &= HA = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} = R. \end{aligned}$$

   Obtenemos $A = H^TR = QR \quad \text{con } Q = H^T = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$.

2. Sea $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Construyamos la matriz de Householder y factoremos la matriz en la forma $A = QR$ con Q una matriz ortogonal y R una matriz triangular superior.

   Sea $\overrightarrow{v} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, se tiene $v_1 = 1$, $\alpha = -1$,

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}.$$

Además $r = \alpha(\alpha - v_1) = -1(-1-1) = 2$. Luego

$$H^{(1)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}(2,0,0) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A^{(1)} = H^{(1)}A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Segunda etapa

Definimos $\overrightarrow{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Procediendo tal como en la primera etapa obtenemos $\widetilde{H}^{(2)} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$,

$$H^{(2)} = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & \widetilde{H}^{(2)} \end{array}\right] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$A^{(2)} = H^{(2)}A^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = R.$$

De la definición de las matrices precedentes obtenemos $R = H^{(2)}A^{(1)} = H^{(2)}H^{(1)}A \Rightarrow A = QR$ con $Q = \left(H^{(2)}H^{(1)}\right)^T$, y

$$H^{(2)}H^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

3. Sea $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. Construyamos la matriz de Householder y factoremos la matriz en la forma $A = QR$ con Q una matriz ortogonal y R una matriz triangular superior.

Sea $\overrightarrow{v} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, se tiene $\|v\| = 1$, $v_1 = 0$, $\alpha = -1$,

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Además $r = \alpha(\alpha - v_1) = 1$. Luego

$$H^{(1)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}(1,0,1) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

$$A^{(2)} = H^{(1)}A = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Segunda etapa

Definimos $\overrightarrow{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Procediendo tal como en la primera etapa obtenemos

$$H^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$A^{(3)} = H^{(2)}A^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = R.$$

Consecuentemente $R = H^{(2)}A^{(2)} = H^{(2)}H^{(1)}A \Rightarrow A = QR$ con $Q = \left(H^{(2)}H^{(1)}\right)^T$, y

$$H^{(2)}H^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

4. Sea $A = \begin{bmatrix} 2 & -1 & 4 \\ 2 & 0 & -1 \\ 1 & 3 & 1 \end{bmatrix}$. Construyamos una matriz triangular superior aplicando el algoritmo que acabamos de describir.

   i) Sea $\overrightarrow{v} = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$ obtenido como la primera columna de la matriz A, entonces $\|\overrightarrow{v}\| = \sqrt{9} = 3$. Determinemos α, $\overrightarrow{w}$ y r. Tenemos

$$\begin{aligned} \alpha &= |\alpha|\, sign\,(\alpha) = \|\overrightarrow{v}\|\, sign\,(\alpha) = -\|\overrightarrow{v}\| = -3, \quad \text{pués } sign\,(\alpha) = -sign\,(v_1) = -1, \\ \overrightarrow{w} &= \overrightarrow{v} - \alpha\overrightarrow{e}_1 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} - (-3)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 1 \end{bmatrix}, \\ r &= \alpha\,(\alpha - v_1) = -3\,(-3-2) = 15, \end{aligned}$$

   y la matriz de Householder está definida como sigue:

$$H^{(1)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{15}\begin{bmatrix} 5 \\ 2 \\ 1 \end{bmatrix}(5,2,1) = \begin{bmatrix} -\frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \\ -\frac{2}{3} & \frac{11}{15} & -\frac{2}{15} \\ -\frac{1}{3} & -\frac{2}{15} & \frac{14}{15} \end{bmatrix},$$

   entonces

$$A^{(2)} = H^{(1)}A^{(1)} = \begin{bmatrix} -\frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \\ -\frac{2}{3} & \frac{11}{15} & -\frac{2}{15} \\ -\frac{1}{3} & -\frac{2}{15} & \frac{14}{15} \end{bmatrix} \begin{bmatrix} 2 & -1 & 4 \\ 2 & 0 & -1 \\ 1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -\frac{1}{3} & -\frac{7}{3} \\ 0 & \frac{4}{15} & -\frac{53}{15} \\ 0 & \frac{47}{15} & -\frac{4}{15} \end{bmatrix}.$$

   ii) Continuado con el algoritmo, elegimos el vector $\overrightarrow{v} = \begin{bmatrix} \frac{4}{15} \\ \frac{47}{15} \end{bmatrix}$ obtenido de la segunda columna de la matriz $A^{(2)}$, luego $\|\overrightarrow{v}\| = \frac{1}{15}\sqrt{2225} = \frac{1}{3}\sqrt{89} \simeq 3{,}1446603777$. Determinemos α, $\overrightarrow{w}$ y r. Tenemos

$$\begin{aligned} \alpha &= -\frac{1}{3}\sqrt{89} = -3{,}1446603777, \\ \overrightarrow{w} &= \begin{bmatrix} \frac{4}{15} \\ \frac{47}{15} \end{bmatrix} + 3{,}1446603777\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3{,}4113270440 \\ 3{,}1333333333 \end{bmatrix}, \\ r &= \alpha\,(\alpha - v_1) = -3{,}1446603777\left(-3{,}1446603777 - \frac{4}{15}\right) = 10{,}7274649895, \\ \frac{1}{r} &= 0{,}09322186682. \end{aligned}$$

En esta estapa la matriz de Householder está definida como sigue:

$$\begin{aligned} H^{(2)} &= I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &\quad -0{,}09322186682 \begin{bmatrix} 0 \\ 3{,}4113270440 \\ 3{,}1333333333 \end{bmatrix} (0, 3{,}4113270440, 3{,}1333333333) \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0{,}0847998304 & -0{,}9963980072 \\ 0 & -0{,}9963980072 & 0{,}0847998304 \end{bmatrix}. \end{aligned}$$

luego, calculamos la matriz $A^{(3)} = H^{(2)}A^{(2)}$, tenemos

$$\begin{aligned} A^{(3)} &= H^{(2)}A^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0{,}0847998304 & -0{,}9963980072 \\ 0 & -0{,}9963980072 & 0{,}0847998304 \end{bmatrix} \begin{bmatrix} -3 & -\frac{1}{3} & -\frac{7}{3} \\ 0 & \frac{4}{15} & -\frac{53}{15} \\ 0 & \frac{47}{15} & -\frac{4}{15} \end{bmatrix} \\ &= \begin{bmatrix} -3 & -\frac{1}{3} & -\frac{7}{3} \\ 0 & -3{,}1446603774 & 0{,}56533322027 \\ 0 & 0 & 3{,}4979930040 \end{bmatrix}, \end{aligned}$$

que es la matriz triangular superior buscada. Note que

$$A^{(3)} = H^{(2)}A^{(2)} = H^{(2)}H^{(1)}A^{(1)} = H^{(2)}H^{(1)}A = QA$$

con $Q = H^{(2)}H^{(1)}$ matriz ortogonal. Ponemos $R = A^{(3)}$ y se tiene $A = Q^T R$.

5. Sean $A = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 4 & 1 \\ 2 & 4 & 0 & 2 \\ 3 & 1 & 2 & 14 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$. Apliquemos el método de Householder para resolver el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$. Para el efecto, primeramente construimos una matriz triangular superior $A^{(4)} = H^{(3)}H^{(2)}H^{(1)}A$, a continuación obtenemos el vector $\overrightarrow{b}^{(4)} = H^{(3)}H^{(2)}H^{(1)}\overrightarrow{b}$. Entonces $A\overrightarrow{x} = \overrightarrow{b} \iff A^{(4)}\overrightarrow{x} = \overrightarrow{b}^{(4)}$, este último sistema de ecuaciones lineales es triangular superior.

i) De acuerdo al algoritmo antes descrito, elegimos $\overrightarrow{v} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \end{bmatrix}$ que es la primera columna de la matriz A, luego $\|\overrightarrow{v}\| = \sqrt{15} = 3{,}8729833462$. A continuación determinamos $\alpha = -3{,}8729833462$, y

$$r = \alpha\,(\alpha - v_1) = -\sqrt{15}\left(-\sqrt{15} - 1\right) = 18{,}8729833462, \quad \frac{1}{r} = 0{,}05298579259,$$

y defiinos el vector $\overrightarrow{w}$ como sigue:

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\,\overrightarrow{e}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \end{bmatrix} + \sqrt{15}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1+\sqrt{15} \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 4{,}8729833462 \\ 1 \\ 2 \\ 3 \end{bmatrix}.$$

La matriz de Householder $H^{(1)}$ está definida como

$$H^{(1)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - 0{,}05298579259 \begin{bmatrix} 4{,}8729833462 \\ 1 \\ 2 \\ 3 \end{bmatrix} (4{,}8729833462, 1, 2, 3)$$

$$= \begin{bmatrix} -0{,}2581988897 & -2{,}5819889747 & -0{,}5163977795 & -0{,}7745966692 \\ -2{,}5819889747 & 0{,}9470142064 & -0{,}1059715872 & -0{,}1589573807 \\ -0{,}5163977795 & -0{,}1059715872 & -0{,}7880568256 & -0{,}3179147615 \\ -0{,}7745966692 & -0{,}1589573807 & -0{,}3179147615 & 0{,}52312785769 \end{bmatrix}.$$

y en consecuencia

$$A^{(2)} = H^{(1)}A^{(1)} = \begin{bmatrix} -3{,}8729833462 & -3{,}6147844565 & -3{,}0983866769 & -12{,}9099444873 \\ 0{,}0 & 1{,}052987936 & 2{,}9537442846 & -2{,}2649289679 \\ 0{,}0 & 2{,}1059715872 & -2{,}0925114308 & -4{,}5298579359 \\ 0{,}0 & -1{,}8410426192 & -1{,}1387671462 & 4{,}2052130962 \end{bmatrix}.$$

ii) Tomando en consideración la segunda columna de la matriz $A^{(2)}$, elegimos el vector $\overrightarrow{v} = \begin{bmatrix} 0 \\ 1{,}052987936 \\ 2{,}1059715872 \\ -1{,}8410426192 \end{bmatrix}$ y calculamos su norma, tenemos $\|\overrightarrow{v}\| = 2{,}9888682362 \simeq 3$ con lo que $\alpha = -2{,}9888682362$. Calculemos r y r^{-1} :

$$\begin{aligned} r &= \alpha(\alpha - v_2) = -2{,}9888682362\,(-2{,}9888682362 - 1{,}052987936) = 12{,}08056912496, \\ \frac{1}{r} &= 0{,}082777557055. \end{aligned}$$

Determinamos el vector $\overrightarrow{w}$:

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_2 = \begin{bmatrix} 0 \\ 1{,}052987936 \\ 2{,}1059715872 \\ -1{,}8410426192 \end{bmatrix} + 2{,}9888682362 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}0 \\ 4{,}04185402978 \\ 2{,}1059715872 \\ -1{,}8410426192 \end{bmatrix},$$

y con este pasamos al cálculo de la matriz de Householder $H^{(2)}$:

$$H^{(2)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & -0{,}3523025139 & -0{,}7046050279 & 0{,}6159664708 \\ 0{,}0 & -0{,}7046050279 & 0{,}632871905277 & 0{,}32094377398 \\ 0{,}0 & 0{,}6159664708 & 0{,}3209437773988 & 0{,}71943060871 \end{bmatrix}$$

entonces

$$A^{(3)} = H^{(2)}A^{(2)} = \begin{bmatrix} -3{,}8729833462 & -3{,}6147844565 & -3{,}0983866769 & -12{,}9099444873 \\ 0{,}0 & -2{,}9888682362 & -0{,}2676598420 & 6{,}5799711169 \\ 0{,}0 & 0 & -3{,}7709949958 & 0{,}07869747780 \\ 0{,}0 & 0 & 0{,}32956498588 & 0{,}176409012875 \end{bmatrix}$$

iii) Continuando con el método de Householder, de la tercera columna de la matriz $A^{(3)}$, elegimos el vector $\overrightarrow{v} = \begin{bmatrix} 0{,}0 \\ 0{,}0 \\ -3{,}7709949958 \\ 0{,}32956498588 \end{bmatrix}$ y calculamos su norma, obtenemos $\|\overrightarrow{v}\| = 3{,}78528178726$ y con este valor tenemos $\alpha = 3{,}78528178726$. Calculemos r y r^{-1} :

$$\begin{aligned} r &= \alpha(\alpha - v_3) = 3{,}78528178726\,(3{,}78528178726 + 3{,}7709949958) = 28{,}6026368867 \\ \frac{1}{r} &= 0{,}04596181153. \end{aligned}$$

Calculamos el vector correspondiente $\overrightarrow{w}$:

$$\overrightarrow{w} = v - \alpha \overrightarrow{e}_3 = \begin{bmatrix} 0{,}0 \\ 0{,}0 \\ -3{,}7709949958 \\ 0{,}32956498588 \end{bmatrix} - 3{,}78528178726 \begin{bmatrix} 0{,}0 \\ 0{,}0 \\ 1{,}0 \\ 0{,}0 \end{bmatrix} = \begin{bmatrix} 0{,}0 \\ 0{,}0 \\ -7{,}5562767831 \\ 0{,}3285649859 \end{bmatrix}$$

y la matriz de Householder queda definida como

$$H^{(3)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T = \begin{bmatrix} 1{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 1{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & -0{,}99622569938 & 0{,}086800667518 \\ 0{,}0 & 0{,}0 & 0{,}086800667518 & 0{,}99622569938 \end{bmatrix}.$$

Con esta matriz, calculamos la matriz $A^{(4)}$ siguiente:

$$A^{(4)} = H^{(3)}A^{(3)} = \begin{bmatrix} -3{,}8729833462 & -3{,}6147844565 & -3{,}0983866769 & -12{,}9099444873 \\ 0{,}0 & -2{,}9888682362 & -0{,}2676598420 & 6{,}5799711169 \\ 0{,}0 & 0 & 3{,}7852817872 & -0{,}06308802978 \\ 0{,}0 & 0 & 0 & 0{,}1825741858 \end{bmatrix}.$$

Puesto que $\overrightarrow{b}^{(4)} = H^{(3)}H^{(2)}H^{(1)}\overrightarrow{b}$, se sigue que

$$\overrightarrow{b}^{(2)} = H^{(1)}\overrightarrow{b} = \begin{bmatrix} 0{,}516397779494 \\ -0{,}09924150898 \\ -0{,}19848301796 \\ -1{,}29772452694 \end{bmatrix}, \quad \overrightarrow{b}^{(3)} = H^{(2)}\overrightarrow{b}^{(2)} = \begin{bmatrix} 0{,}516397779494 \\ -0{,}6245396314 \\ -0{,}4721848668 \\ -1{,}0584544077 \end{bmatrix},$$

$$\overrightarrow{b}^{(4)} = H^{(3)}\overrightarrow{b}^{(3)} = \begin{bmatrix} 0{,}516397779494 \\ -0{,}6245396314 \\ 0{,}37852817872 \\ -1{,}0954451150 \end{bmatrix}.$$

Resolvamos el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$. Tenemos $A^{(4)}\overrightarrow{x} = \overrightarrow{b}^{(4)}$:

$$\begin{bmatrix} -3{,}8729833462 & -3{,}6147844565 & -3{,}0983866769 & -12{,}9099444873 \\ 0{,}0 & -2{,}9888682362 & -0{,}2676598420 & 6{,}5799711169 \\ 0{,}0 & 0 & 3{,}7852817872 & -0{,}06308802978 \\ 0{,}0 & 0 & 0 & 0{,}1825741858 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix}$$

$$= \begin{bmatrix} 0{,}516397779494 \\ -0{,}6245396314 \\ 0{,}37852817872 \\ -1{,}0954451150 \end{bmatrix}$$

y la solución de este sistema triangular superior es $\overrightarrow{x} = \begin{bmatrix} 32 \\ -13 \\ 0 \\ -6 \end{bmatrix}$.

9.3.1. Número de operaciones elementales

Para pasar de $A^{(k)}$ a $A^{(k+1)}$ se requiere las siguientes operaciones elementales:

$n-k+1$	elevaciones al cuadrado (multiplicaciones),
$n-k$	adiciones,
1	raíz cuadrada para calcular $a_{kk}^{(k+1)}$,
1	sustracción,
1	multiplicación para calcular $r^{(k)}$,
1	sustracción para calcular $w_i^{(k)}$, $\quad i=k,\ldots,n$,
$n-k+1$	multiplicaciones para calcular $q_j^{(k)}$,
$n-k$	adiciones,
1	división para calcular $q_i^{(k)}$,
$n-k+1$	multiplicaciones para calcular $a_{ij}^{(k+1)}$, $\quad i=k,\ldots,n$,
$n-k+1$	sustracciones

Para pasar de $\overrightarrow{b}^{(k)}$ a $\overrightarrow{b}^{(k+1)}$, se requieren de:

$n-k+1$	multiplicaciones para calcular $q^{(k)}$,
$n-k$	adiciones,
1	división para calcular $p^{(k)}$,
$n-k+1$	multiplicaciones para calcular $b_i^{(k+1)}$, $\quad i=k,\ldots,n$,
$n-k+1$	sustracciones,

Para pasar del sistema $A^{(k)}\overrightarrow{x}=\overrightarrow{b}^{(k)}$ al sistema $A^{(k+1)}\overrightarrow{x}=\overrightarrow{b}^{(k+1)}$ se requieren de:

$2n^2-4nk+5n+2k^2-5k+4$	multiplicaciones,
$2n^2-4nk+4n+2k^2-4k+3$	adiciones o sustracciones,
$n-k+1$	divisiones,
1	raíz cuadrada,

Por lo tanto, para la triangularización de A se necesitan

$(n-1)\left[\dfrac{2n^2}{3}+\dfrac{13n}{6}+4\right]$	multiplicaciones,
$(n-1)\left[\dfrac{2n^2}{3}+\dfrac{5n}{3}+3\right]$	adiciones o sustracciones,
$(n-1)\dfrac{n+2}{2}$	divisiones,
$n-1$	raíces cuadradas,

Para la resolución del sistema triangular $A^{(n)}\overrightarrow{x}=\overrightarrow{b}^{(n)}$ se necesitan de n^2 operaciones elementales.

El número total de operaciones en el método de Householder es:

$$T_H=\frac{4n^3}{3}+4n^2+\frac{14n}{3}-9.$$

Metodo de Householder

El método de Householder puede aplicarse para resolver problemas de aproximación con el método de mínimos cuadrados.

Sea $A\in M_{m\times n}[\mathbb{R}]$ con $m\geq n$ y $\overrightarrow{b}\in\mathbb{R}^m$. Consideramos el sistema de ecuaciones $A\overrightarrow{x}=\overrightarrow{b}$ y el problema en mínimos cuadrados:

$$\text{hallar } \widehat{x}\in\mathbb{R}^n \text{ tal que } \left\|A\widehat{x}-\overrightarrow{b}\right\|=\min_{\overrightarrow{x}\in\mathbb{R}^n}\left\|A\overrightarrow{x}-\overrightarrow{b}\right\|.$$

Apliquemos el método de ortogonalización de House holder para resolver el sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$.

Ponemos $A^{(0)} = A$ y $\overrightarrow{b}^{(0)} = \overrightarrow{b}$. Mediante el método de Householder construimos matrices ortogonales $Q_i \in M_{m\times m}[\mathbb{R}]$, matrices $A^{(i)}$ tales que $A^{(i)} = Q_i A^{(i-1)}$ y $\overrightarrow{b}^{(i)} = Q_i \overrightarrow{b}^{(i-1)} \quad i = 1, \ldots, n.$

Sea $Q = Q_{n-1}Q_{n-2}\ldots Q_1$ entonces $QA^{(n)} = \begin{bmatrix} R \\ 0 \end{bmatrix}$, donde $R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix}$, $0 \in M_{m\times m}[\mathbb{R}]$.

Ponemos $\overrightarrow{h} = \overrightarrow{b}^{(n)} = Q\overrightarrow{b}$ y $\overrightarrow{h} = \begin{bmatrix} \overrightarrow{h}_1 \\ \overrightarrow{h}_2 \end{bmatrix}$ con $\overrightarrow{h}_1 \in \mathbb{R}^n$ y $\overrightarrow{h}_2 \in \mathbb{R}^{m-n}$.

Puesto que la matriz Q es ortogonal, se tiene $\|Q\overrightarrow{u}\| = \|\overrightarrow{u}\| \quad \forall \overrightarrow{u} \in \mathbb{R}^m$ y en consecuencia

$$\left\|A\overrightarrow{x} - \overrightarrow{b}\right\| = \left\|Q\left(A\overrightarrow{x} - \overrightarrow{b}\right)\right\| = \left\|QA\overrightarrow{x} - Q\overrightarrow{b}\right\| = \left\|A^{(n)}\overrightarrow{x} - \overrightarrow{h}\right\|,$$

y como

$$A^{(n)}\overrightarrow{x} - \overrightarrow{h} = \begin{bmatrix} R \\ 0 \end{bmatrix}\overrightarrow{x} - \begin{bmatrix} \overrightarrow{h}_1 \\ \overrightarrow{h}_2 \end{bmatrix} = \begin{bmatrix} R\overrightarrow{x} - \overrightarrow{h}_1 \\ -\overrightarrow{h}_2 \end{bmatrix},$$

con lo que

$$\left\|A^{(n)}\overrightarrow{x} - \overrightarrow{h}\right\| = \left\|\begin{bmatrix} R\overrightarrow{x} - \overrightarrow{h}_1 \\ -\overrightarrow{h}_2 \end{bmatrix}\right\| = \left(\left\|R\overrightarrow{x} - \overrightarrow{h}_1\right\|^2 + \left\|\overrightarrow{h}_2\right\|^2\right)^{\frac{1}{2}}$$

y de esta igualdad, se sigue que $\left\|A^{(n)}\overrightarrow{x} - \overrightarrow{h}\right\|$ tendrá norma mínima si se elige $\overrightarrow{x}$ como solución del sistema de ecuaciones lineales $R\overrightarrow{x} = \overrightarrow{h}_1$, de donde $\overrightarrow{x} = R^{-1}\overrightarrow{h}_1$.

Note que la matriz R tiene inversa si y solo si las columnas de A son linealmente independientes, esto es, $\dim R(A) = n$.

Teorema 2 *Sea $A \in M_{m\times n}[\mathbb{R}]$ tal que $\dim \mathcal{R}(A) = n$ con $m \geq n$. Entonces A puede factorarse en un producto de la forma $A = Q\widetilde{R}$, donde $Q \in M_{m\times m}[\mathbb{R}]$ es una matriz ortogonal y $\widetilde{R} = \begin{bmatrix} R \\ 0 \end{bmatrix}$ con $R \in M_{n\times n}[\mathbb{R}]$ una matriz triangular superio invertible.*
Si $m = n$, $A = QR$.

Demostración. Basta aplicar el método de Householder. ■

Observación

Sea $A \in M_{n\times n}[\mathbb{R}]$ con $\dim \mathcal{R}(A) = n$, $\overrightarrow{b} \in \mathbb{R}^n$ y consideramos el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$. Ponemos $A = QR$. La solución del sistema de ecuaciones en mínimos cuadrados viene dada como

$$\begin{aligned} \overrightarrow{x} &= \left(A^TA\right)^{-1}A^T\overrightarrow{b} = \left[(QR)^TQR\right]^{-1}(QR)^T\overrightarrow{b} \\ &= \left[R^TQ^TQR\right]^{-1}R^TQ^T\overrightarrow{b} = \left(R^TR\right)^{-1}R^TQ^T\overrightarrow{b} \\ &= R^{-1}\left(R^T\right)^{-1}R^TQ^T\overrightarrow{b} = R^{-1}Q^T\overrightarrow{b}. \end{aligned}$$

Así, $\overrightarrow{x} = R^{-1}Q^T\overrightarrow{b} \Leftrightarrow R\overrightarrow{x} = Q^T\overrightarrow{b}$, además dicha solución $\overrightarrow{x}$ es única.

Ejemplo

Consideremos el sistema de ecuaciones lineales $\begin{cases} x - 2y + z = 1 \\ 2x + 5y + z = 2 \\ 4y - z = 3 \\ 2x + 3y + z = 4 \end{cases}$. Calculemos la solución en mínimos cuadrados aplicando el método de Householder.

Ponemos $A = \begin{bmatrix} 1 & -2 & 1 \\ 2 & 5 & 1 \\ 0 & 5 & -1 \\ 2 & 3 & 1 \end{bmatrix}$, $\overrightarrow{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$. Determinaremos una matriz ortogonal Q y una matriz triangular superior invertible R tales que $QA = \begin{bmatrix} R \\ 0 \end{bmatrix}$ y $\overrightarrow{h} = Q\overrightarrow{b} = \begin{bmatrix} \overrightarrow{h}_1 \\ \overrightarrow{h}_2 \end{bmatrix}$ con $\overrightarrow{h}_1 \in \mathbb{R}^n$, $\overrightarrow{h}_2 \in \mathbb{R}^{m-n}$.

Ponemos $A^{(1)} = A$ y $\overrightarrow{b}^{(1)} = \overrightarrow{b}$. Apliquemos el método de Householder.

1. $A^{(2)} = Q^{(1)}A^{(1)}$ con $Q^{(1)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T$ y $\alpha = -\|\overrightarrow{v}\| \, sign\,(v_1)$, $r = \alpha\,(\alpha - v_1)$, $\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_1$.

   Sea $\overrightarrow{v} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \end{bmatrix}$ vector obtenido de la primera columna de la matriz A. Calculamos su norma: $\|\overrightarrow{v}\| = 3$, y en consecuencia $\alpha = -3$. Puesto que $v_1 = 1$, se sigue que $r = -3\,(-3-1) = 12$. Con esta información pasamos a calcular el vector $\overrightarrow{w}$. Tenemos

$$\overrightarrow{w} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \end{bmatrix} + 3\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ 0 \\ 8 \end{bmatrix},$$

   con lo que la matriz de Householder está definida como

$$Q^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \frac{1}{12}\begin{bmatrix} 4 \\ 2 \\ 0 \\ 2 \end{bmatrix}(4,2,0,2) = \frac{1}{3}\begin{bmatrix} -1 & -2 & 0 & -2 \\ -2 & 2 & 0 & -1 \\ 0 & 0 & 3 & 0 \\ -2 & -1 & 0 & 2 \end{bmatrix},$$

   y con esta calculamos la matriz $A^{(2)}$ siguiente:

$$A^{(2)} = Q^{(2)}A^{(1)} = \frac{1}{3}\begin{bmatrix} -1 & -2 & 0 & -2 \\ -2 & 2 & 0 & -1 \\ 0 & 0 & 3 & 0 \\ -2 & -1 & 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & -2 & 1 \\ 2 & 5 & 1 \\ 0 & 5 & -1 \\ 2 & 3 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -\frac{14}{3} & -\frac{5}{3} \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 4 & -1 \\ 0 & \frac{5}{3} & -\frac{1}{3} \end{bmatrix}.$$

2. De la segunda columna de la matriz $A^{(2)}$, elegimos elvector $\overrightarrow{v} = \begin{bmatrix} 0 \\ \frac{11}{3} \\ 4 \\ \frac{5}{3} \end{bmatrix}$ y calculamos su norma, tenemos $\|\overrightarrow{v}\| = \frac{\sqrt{290}}{3}$, con lo que $\alpha = -\frac{\sqrt{290}}{3}$, pués $v_2 = \frac{11}{3}$. Luego

$$\frac{1}{r} = \frac{1}{\alpha\,(\alpha - v_2)} = \frac{1}{-\frac{\sqrt{290}}{3}\left(-\frac{\sqrt{290}}{3} - \frac{11}{3}\right)} = \frac{9}{\sqrt{290}\left(11 + \sqrt{290}\right)} \simeq 0{,}018855.$$

   Se define $\overrightarrow{w} = \begin{bmatrix} 0. \\ 9{,}34313 \\ 4. \\ 1{,}6667 \end{bmatrix}$. La matriz de Housholder está definida como $Q^{(2)} = I - \frac{1}{r}\overrightarrow{w}\overrightarrow{w}^T$, esto

es,

$$Q^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - 0{,}018855 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 81{,}2941 & 37{,}37252 & 15{,}57191 \\ 0 & 37{,}37252 & 16{,}0 & 6{,}666667 \\ 0 & 15{,}57191 & 6{,}666667 & 2{,}777778 \end{bmatrix}$$

$$= \begin{bmatrix} 1. & 0. & 0. & 0. \\ 0. & -0{,}64593 & -0{,}70466 & -0{,}29361 \\ 0. & -0{,}70466 & 0{,}69832 & -0{,}12570 \\ 0. & -0{,}29361 & -0{,}12573 & 0{,}94762 \end{bmatrix}$$

Se define $A^{(3)} = Q^{(2)}A^{(2)}$. Resulta

$$A^3 = \begin{bmatrix} 1. & 0. & 0. & 0. \\ 0. & -0{,}64593 & -0{,}70466 & -0{,}29361 \\ 0. & -0{,}70466 & 0{,}69832 & -0{,}12570 \\ 0. & -0{,}29361 & -0{,}12573 & 0{,}94762 \end{bmatrix} \begin{bmatrix} -3 & -\frac{13}{3} & -\frac{5}{3} \\ 0. & \frac{11}{3} & -\frac{1}{3} \\ 0. & 4. & -1 \\ 0. & \frac{5}{3} & -\frac{1}{3} \end{bmatrix}$$

$$= \begin{bmatrix} -3 & -4{,}66667 & -1{,}66667 \\ 0. & -5{,}67639 & 1{,}01784 \\ 0. & 0. & -0{,}42153 \\ 0. & 0. & -0{,}09231 \end{bmatrix}.$$

Por lo tanto la matriz R está definida como

$$R = \begin{bmatrix} -3 & -4{,}6667 & -1{,}66667 \\ 0. & -5{,}67639 & 1{,}01784 \\ 0. & 0. & -0{,}42153 \end{bmatrix}$$

y su inversa

$$R^{-1} = \begin{bmatrix} -0{,}33333 & 0{,}2740 & 1{,}9796 \\ 0. & -0{,}1761 & -0{,}4254 \\ 0. & 0. & -2{,}3723 \end{bmatrix}.$$

Se define $\overrightarrow{h} = Q\overrightarrow{b} = Q^{(2)}Q^{(1)}\overrightarrow{b}$. Se tiene

$$\overrightarrow{h} = \frac{1}{3} \begin{bmatrix} 1. & 0. & 0. & 0. \\ 0. & -0{,}64593 & -0{,}70466 & -0{,}29361 \\ 0. & -0{,}70466 & 0{,}69832 & -0{,}12570 \\ 0. & -0{,}29361 & -1{,}12570 & 0{,}94762 \end{bmatrix} \begin{bmatrix} -1 & -2 & 0 & -2 \\ -2 & 2 & 0 & -1 \\ 0 & 0 & 3 & 0 \\ -2 & -1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} -4{,}3333 \\ -2{,}0748 \\ 2{,}3971 \\ 1{,}0821 \end{bmatrix} = \begin{bmatrix} \overrightarrow{h}_1 \\ \overrightarrow{h}_2 \end{bmatrix},$$

de donde $\overrightarrow{h}_1 = \begin{bmatrix} -4{,}3333 \\ -2{,}0748 \\ 2{,}3971 \end{bmatrix}$

3. La solución en mínimos cuadrado está definida como

$$\overrightarrow{x} = R^{-1}\overrightarrow{h}_1 = \begin{bmatrix} -0{,}3333 & 0{,}2740 & 1{,}9796 \\ 0. & -0{,}1761 & -0{,}4254 \\ 0. & 0. & -2{,}3723 \end{bmatrix} \begin{bmatrix} -4{,}3333 \\ -2{,}0748 \\ 2{,}3971 \end{bmatrix} = \begin{bmatrix} 5{,}6213 \\ -0{,}6544 \\ -5{,}6867 \end{bmatrix}.$$

Note que en realidad no se requiere del cálculo de la matriz R^{-1}. Se resuelve directamente el sistema de ecuaciones lineales triangular superior $R\overrightarrow{x} = \overrightarrow{h}_1$.

9.4. Ajuste de datos polinomial

Para simplificar la escritura y que a su vez no pierda de generalidad, hemos seleccionado como problema de ajuste polinomial con un polinomio de tercer grado. El procedimiento que a continuación se describe, se aplica directamente al ajuste polinomial con polinomios de grados uno, dos, etc.

Supongamos que se dispone de un conjunto de n pares de datos experimentales $S = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, ..., n\}$. Se desea encontrar un polinomio P de grado 3: $P(x) = a + bx + cx^2 + dx^3 \quad x \in \mathbb{R}$, de modo que P se ajuste de la mejor manera al conjunto de datos S. El polinomio P queda perfectamente bien definido si se conocen todos sus coeficientes a, b, c, d. Estos coeficientes son calculados mediante el denominado método de mínimos cuadrados discreto que describimos a continuación.

Denotemos con r_i el residuo en cada medición, esto es,

$$y_i = P(x_i) + r_i = a + bx_i + cx_i^2 + dx_i^3 + r_i \quad i = 1, ..., n.$$

En forma matricial, el conjunto de ecuaciones precedente, se escribe como el siguiente sistema de ecuaciones lineales:

$$\begin{bmatrix} y_1 \\ \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} r_1 \\ \vdots \\ \vdots \\ r_n \end{bmatrix}.$$

El residuo en cada medición depende de los coeficientes a, b, c, d del polinomio P.

Definimos los vectores $\overrightarrow{x}$, $\overrightarrow{y}$ y el residuo $\overrightarrow{r}(\overrightarrow{x})$ con como sigue:

$$\overrightarrow{x} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \quad \overrightarrow{y} = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, \quad \overrightarrow{r}(\overrightarrow{x}) = \begin{bmatrix} r_1(\overrightarrow{x}) \\ \vdots \\ \vdots \\ r_n(\overrightarrow{x}) \end{bmatrix}.$$

Definimos la matriz A siguiente: $A = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$.

El sistema de ecuaciones lineales arriba propuesto se transforma en el siguiente: $\overrightarrow{y} = A\overrightarrow{x} + \overrightarrow{r}(\overrightarrow{x})$, de donde el residuo $\overrightarrow{r}(\overrightarrow{x}) = \overrightarrow{y} - A\overrightarrow{x}$ con $\overrightarrow{x} \in \mathbb{R}^4$. El problema de hallar el "mejor polinomio" que se ajusta al conjunto de datos S se expresa como sigue:

$$\text{hallar } \widehat{x}^T = (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d}) \in \mathbb{R}^4 \text{ tal que } \|\overrightarrow{r}(\widehat{x})\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^4}{\text{Min}} \|\overrightarrow{r}(\widehat{x})\|^2,$$

o de modo equivalente

$$\|\overrightarrow{y} - A\widehat{x}\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^4}{\text{Min}} \|\overrightarrow{y} - A\widehat{x}\|^2.$$

Este problema, como ya hemos señalado, se conoce como método de mínimos cuadrados y se ha demostrado que conduce a resolver el sistema de ecuaciones $A^T A\overrightarrow{x} = A^T\overrightarrow{y}$, donde A^T denota la matriz transpuesta de A.

Veamos la generalización de este problema. Sea $C = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots, n\}$ un conjunto de datos experimentales y p un polinomio de grado $\leq m$. Ponemos

$$p(x) = a_0 + a_1 x_1 + \ldots + a_m x^m \quad x \in \mathbb{R},$$

donde $a_0, \ldots, a_m \in \mathbb{R}$ se determinan mediante el método de mínimos cuadrados. Supongamos

$$\begin{cases} y_1 = a_0 + a_1 x_1 + \ldots + a_m x_1^m + r_1 \\ \quad \vdots \\ y_n = a_0 + a_1 x_n + \ldots + a_m x_n^m + r_n, \end{cases}$$

con $r_1, \ldots, r_n$ los errores cometidos en la medición. En forma matricial, el sistema de ecuaciones precedente, se expresa como $\overrightarrow{y} = A\overrightarrow{x} + \overrightarrow{r}$, donde $\overrightarrow{y} = (y_1, \ldots, y_n)^T$, $\overrightarrow{r} = (r_1, \ldots, r_n)^T$, $\overrightarrow{x} = (a_0, \ldots, a_m)^T$,

$$A = (a_{ij}) = \begin{bmatrix} 1 & x_1 & \ldots & x_1^m \\ & \vdots & & \\ 1 & x_n & \ldots & x_n^m \end{bmatrix} = (A_0, \ldots, A_m),$$

con $A_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \ldots, A_0 = \begin{bmatrix} x_1^m \\ \vdots \\ x_n^m \end{bmatrix}$ las columnas de A.

Note que si $\overrightarrow{x} = (a_0, \ldots, a_m)^T \in \mathbb{R}^{m+1}$, se tiene $A\overrightarrow{x} = a_0 A_0 + \ldots + a_m A_m$. Sea $W = \left\{ A\overrightarrow{x} \mid \overrightarrow{x} \in \mathbb{R}^{m+1} \right\}$. Resulta que W es un subespacio de $\mathbb{R}^n$ con $\dim W = m < n$. Se define $\overrightarrow{r}(\overrightarrow{x}) = \overrightarrow{y} - A\overrightarrow{x} \quad \overrightarrow{x} \in \mathbb{R}^{m+1}$. El problema de mínimos cuadrados consiste en determinar $\hat{x} \in \mathbb{R}^{m+1}$ tal que

$$\|\overrightarrow{r}(\hat{x})\|^2 \leq \|\overrightarrow{r}(\overrightarrow{x})\|^2 \quad \forall \overrightarrow{x} \in \mathbb{R}^{m+1} \Longleftrightarrow \|\overrightarrow{y} - A\hat{x}\|^2 = \underset{\overrightarrow{x} \in \mathbb{R}^{m+1}}{\text{Min}} \|\overrightarrow{y} - A\overrightarrow{x}\|^2.$$

Note que, en general, $\overrightarrow{y} \notin W$. Por el teorema precedente, existe $\hat{y} \in W$ tal que

$$\|\overrightarrow{y} - \hat{y}\| \leq \|\overrightarrow{y} - \overrightarrow{z}\| \quad \forall \overrightarrow{z} \in W,$$
$$\langle \overrightarrow{y} - \hat{y},\ \overrightarrow{w} \rangle = 0 \quad \forall \overrightarrow{w} \in W.$$

Como $\hat{y} \in W$ si y solo si $\overrightarrow{x} \in \mathbb{R}^{m+1}$ tal que $\hat{y} = A\hat{x}$, $\overrightarrow{z} \in W$ si y solo si existe $\overrightarrow{x} \in \mathbb{R}^{m+1}$ tal que $\overrightarrow{z} = A\overrightarrow{x}$.

Así,

$$\|\overrightarrow{y} - A\hat{x}\|^2 \leq \|\overrightarrow{y} - A\overrightarrow{x}\|^2 \quad \forall \overrightarrow{x} \in \mathbb{R}^{m+1},$$
$$\langle \overrightarrow{y} - A\hat{x},\ \overrightarrow{w} \rangle = 0 \quad \forall \overrightarrow{w} \in W.$$

Para $\overrightarrow{w} = A\overrightarrow{x} \quad \overrightarrow{x} \in \mathbb{R}^{m+1}$, se obtiene

$$0 = \langle \overrightarrow{y} - A\hat{x},\ Ax \rangle = (A\overrightarrow{x})^T (\overrightarrow{y} - A\hat{x}) = \overrightarrow{x}^T A^T \overrightarrow{y}^T - \overrightarrow{x}^T A^T A\hat{x} = \overrightarrow{x}^T \left(A^T \overrightarrow{y} - A^T A\hat{x} \right).$$

Luego

$$\overrightarrow{x}^T \left(A^T \overrightarrow{y} - A^T A\hat{x} \right) = 0 \quad \forall \overrightarrow{x} \in \mathbb{R}^{m+1} \Longleftrightarrow A^T A\hat{x} = A^T \overrightarrow{y},$$

que se conoce como ecuación normal. O sea $\hat{x}$ es solución del sistema de ecuaciones normal.

9.4.1. Ajuste de datos con polinomios de grado 1.

Consideramos el conjunto de datos $S = \left\{ (x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots, n \right\}$ y supongamos que dicho conjunto de puntos tiene una tendencia como la de un polinomio de grado 1, esto es, $p(x) = a + bx \quad x \in \mathbb{R}$, donde $a, b \in \mathbb{R}$ se determinan como soluciones en mínimos cuadrados. Tenemos

$$y_i = a + bx_i + r_i \quad i = 1, \ldots, n.$$

Ponemos $\overrightarrow{x}^T = (a, b) \in \mathbb{R}^2$, $\overrightarrow{y}^T = (y_1, \ldots, y_n)$, $A = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, el residuo definido como $\overrightarrow{r}(\overrightarrow{x})^T = (r_1(\overrightarrow{x}), \ldots, r_n(\overrightarrow{x}))$. El sistema de ecuaciones precedente se escribe como $\overrightarrow{y} = A\overrightarrow{x} + \overrightarrow{r}(\overrightarrow{x}) \quad \overrightarrow{x} \in \mathbb{R}^2$. Luego, el problema en mínimos cuadrados está definido como

$$E(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \|\overrightarrow{r}(\overrightarrow{x})\|^2 = \|\overrightarrow{y} - A\overrightarrow{x}\|^2 \quad \overrightarrow{x} \in \mathbb{R}^2.$$

La solución en mínimos cuadrados se expresa como

$$\widehat{x} = \left(A^T A\right)^{-1} A^T \overrightarrow{y}.$$

Puesto que

$$\begin{aligned} A^T A &= \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}, \\ A^T \overrightarrow{y} &= \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}. \end{aligned}$$

Es este caso resulta fácil el cálculo de $\left(A^T A\right)^{-1}$, tenemos

$$\left(A^T A\right)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix},$$

luego

$$\begin{aligned} \widehat{x} &= \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \left(A^T A\right)^{-1} A^T \overrightarrow{y} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) & -\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right) \\ -\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) & n\sum_{i=1}^{n} x_i y_i \end{bmatrix} \end{aligned}$$

y de esta igualdad obtenemos los coeficientes $\widehat{a}$, $\widehat{b}$:

$$\begin{aligned} \widehat{a} &= \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}, \\ \widehat{b} &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}. \end{aligned}$$

Si aplicamos las condiciones necesarias de extremo a la función $E(a,b)$, tenemos

$$\begin{cases} \dfrac{\partial E}{\partial a}(a,b) = 0 \\ \dfrac{\partial E}{\partial b}(a,b) = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \end{cases} \Leftrightarrow \begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i, \end{cases}$$

cuya solución es

$$a = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

que coincide exactamente con $\widehat{a},\ \widehat{b}$.

Note que hemos supuesto que $\left(\sum_{i=1}^{n} x_i\right)^2 \neq n\sum_{i=1}^{n} x_i^2$.

Para calcular las constantes $\widehat{a},\ \widehat{b}$ mediante el método de mínimos cuadrados se requiere de la siguiente información: el número de puntos $n \geq 2$ y el conjunto de datos $S = \{(x_i, y_i) \mid i = 1, \ldots, n\}$; y, se deben calcular las sigientes sumas: $S_1 = \sum_{i=1}^{n} x_i$, $S_2 = \sum_{i=1}^{n} x_i^2$, $R_1 = \sum_{i=1}^{n} y_i$, $R_2 = \sum_{i=1}^{n} x_i y_i$. Con estos resultados se pasa al cálculo de $\widehat{a}$ y $\widehat{b}$. Se propone el siguiente algoritmo de cálculo de $\widehat{a}$ y $\widehat{b}$.

Algoritmo

Datos de entrada: $n \in \mathbb{Z}^+,\ (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$.

Datos de salida: Mensaje1, Mensaje 2, $a,\ b$.

1. Si $n < 2$, continuar en 14)
2. $S_1 = 0$.
3. $S_2 = 0$.
4. $R_1 = 0$.
5. $R_2 = 0$.
6. Para $i = 1, \ldots, n$

   $S_1 = S_1 + x_i$

   $S_2 = S_2 + x_i^2$

   $R_1 = R_1 + y_i$

   $R_2 = R_2 + x_i y_i$

   Fin de bucle i.
7. $z = nS_2 - S_1^2$
8. Si $z = 0$, continuar en 13)
9. $a = \dfrac{S_2 R_1 - S_1 R_2}{z}$,
10. $b = \dfrac{nR_2 - S_1 R_1}{z}$
11. Imprimir $a,\ b$, continuar en 14)
12. Mensaje 1: $n \geq 2$ continuar en 14)

13. Mensaje 2: el sistema no tiene solución o tiene infinitas soluciones.

14. Fin

En la gráfica siguiente se ilustra un conjunto de puntos $S = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots, n\}$ que siguen una tendencia de una recta, y, la gráfica de la recta solución en mínimos cuadrados, de ecuación $y = a + bx$ $x \in \mathbb{R}$.

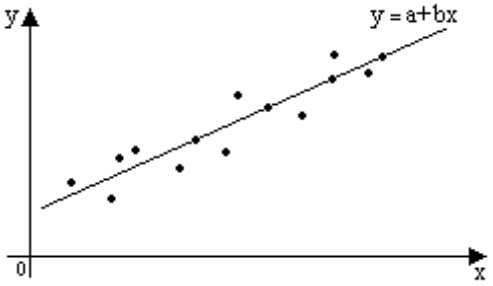


Figura 84

Ejemplo

Apliquemos el algoritmo al siguiente conjunto de datos:

$$S = \{(1,\ 3{,}6), (1{,}5,\ 4{,}35), (2{,}1,\ 5{,}25), (2{,}9,\ 6{,}45), (3{,}2,\ 6{,}9)\}.$$

Buscamos una función p de la forma $p(x) = a + bx \quad x \in \mathbb{R}$, con $a,\ b$ constantes calculadas con el método de mínimos cuadrados

$$\begin{aligned} S_1 &= \sum_{i=1}^{5} x_i = 1 + 1{,}5 + 2{,}1 + 2{,}9 + 3{,}2 = 10{,}7, \\ S_2 &= \sum_{i=1}^{n} x_i^2 = 1 + 2{,}25 + 4{,}41 + 8{,}41 + 10{,}24 = 26{,}31, \\ R_1 &= \sum_{i=1}^{n} y_i = 3{,}6 + 4{,}35 + 5{,}25 + 6{,}45 + 6{,}9 = 26{,}55, \\ R_2 &= \sum_{i=1}^{n} x_i y_i = 3{,}6 + 6{,}525 + 11{,}025 + 18{,}705 + 22{,}08 = 61{,}935. \\ z &= nS_2 - S_1^2 = 5 \times 26{,}31 - (10{,}7)^2 = 17{,}06. \end{aligned}$$

La solución es

$$\begin{aligned} a &= \frac{S_2 \times R_1 - S_1 R_2}{z} = \frac{26{,}55 \times 26{,}31 - 10{,}7 \times 61{,}935}{17{,}06} = 2{,}1 \\ b &= \frac{nR_2 - S_1 \times R_1}{z} = \frac{5 \times 61{,}935 - 10{,}07 \times 26{,}55}{17{,}06} = 1{,}5. \end{aligned}$$

El polinomio de grado 1 buscado es $p(x) = 2{,}1 + 1{,}5x \quad x \in \mathbb{R}$. Note que $p(x)$ es la ecuación cartesiana de la recta a la que se le denomina recta de mejor ajuste en mínimos cuadrados.

9.4.2. Ajuste polinomial con polinomios de grado 2.

Dado el conjunto de datos $S = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots, n\}$ que tiene una tendencia de una función cuadrática del tipo $p(x) = a + bx + cx^2 \quad x \in \mathbb{R}$, donde $a,\ b,\ c$ son constantes reales que se determinan como soluciones del método de mínimos cuadrados. tenemos

$$y_i = a + bx_i + cx_i^2 + r_i \quad i = 1, \ldots, n.$$

Siguiendo el mismo procedimiento descrito anteriormente, definimos los vectores de $\mathbb{R}^n$ siguientes: $\overrightarrow{y}^T = (y_1, \ldots, y_n)$, $\overrightarrow{r}\left(\overrightarrow{x}\right)^T = \left(r_1\left(\overrightarrow{x}\right), \ldots, r_n\left(\overrightarrow{x}\right)\right)$ con $\overrightarrow{x}^T = (a, b, c) \in \mathbb{R}^3$; y, se define la matriz $A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$. Entonces

$$\overrightarrow{y} = A\overrightarrow{x} + \overrightarrow{r}\left(\overrightarrow{x}\right) \quad \overrightarrow{x} \in \mathbb{R}^3.$$

Se define la función E de $\mathbb{R}^3$ en $[0, \infty[$ como sigue

$$E(a, b, c) = \|\overrightarrow{r}(a, b, c)\|^2 = \|\overrightarrow{y} - A\overrightarrow{x}\|^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right)^2 \quad (a, b, c) \in \mathbb{R}^3.$$

El sistema de ecuaciones normal está definido como $A^T A\overrightarrow{x} = A^T\overrightarrow{y}$. Calculemos $A^T A$ y $A^T\overrightarrow{y}$. Tenemos

$$A^T A = \begin{bmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ x_1^2 & \ldots & x_n^2 \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{bmatrix},$$

$$A^T\overrightarrow{y} = \begin{bmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ x_1^2 & \ldots & x_n^2 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix}.$$

La solución $\widehat{x}^T = \left(\widehat{a}, \widehat{b}, \widehat{c}\right)$ se obtiene del sistema de ecuaciones lineales

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix}.$$

Si aplicamos las condiciones necesarias de extremo a la función $E(a, b, c)$, tenemos

$$\begin{cases} \frac{\partial E}{\partial a}(a, b, c) = 0 \\ \frac{\partial E}{\partial b}(a, b, c) = 0 \\ \frac{\partial E}{\partial c}(a, b, c) = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right) = 0 \\ \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right) x_i = 0 \\ \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right) x_i^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i, \end{cases}$$

que es exactamente el sistema de ecuaciones normal.

Primeramente, para resolver el sistema de ecuaciones, hemos de calcular cada uno de los sumatorios. Ponemos

$$S_1 = \sum_{i=1}^{n} x_i,\ S_2 = \sum_{i=1}^{n} x_i^2,\ S_3 = \sum_{i=1}^{n} x_i^3,\ S_4 = \sum_{i=1}^{n} x_i^4,$$

$$R_1 = \sum_{i=1}^{n} y_i,\ R_2 = \sum_{i=1}^{n} x_i y_i,\ R_3 = \sum_{i=1}^{n} x_i^2 y_i,$$

con lo que el sistema de ecuaciones lineales se escribe como

$$\begin{bmatrix} n & S_1 & S_2 \\ S_1 & S_2 & S_3 \\ S_2 & S_3 & S_4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix}.$$

Este sistema de ecuaciones se resuelve con el método de eliminación gaussiana.

Proponemos como ejercicio la elaboración de un algoritmo para el cálculo de $\widehat{x}^T = \left(\widehat{a}, \widehat{b}, \widehat{c}\right)$.

Ejemplo

Consideremos el conjunto de datos S siguiente: $S = \{(0,3), (1,2), (2,3), (3,6), (4,11)\}$. Determinemos un polinomio $p(x) = a_1 + a_2 x + a_3 x^2 \quad x \in \mathbb{R}$ en mínimos cuadrados.

De los resultados arriba establecidos, se tienen $\overrightarrow{x}^T$, $\overrightarrow{y}^T$ y A definidos como sigue: $\overrightarrow{x} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$, $\overrightarrow{y}^T = (3,2,3,6,11)$, $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}$. Luego

$$\begin{aligned} A^T A &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} = \begin{bmatrix} 5 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}, \\ A^T \overrightarrow{y} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 4 & 9 & 16 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 3 \\ 6 \\ 11 \end{bmatrix} = \begin{bmatrix} 25 \\ 70 \\ 244 \end{bmatrix}. \end{aligned}$$

El sistema normal de ecuaciones lineales es

$$\begin{bmatrix} 5 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 25 \\ 70 \\ 244 \end{bmatrix}.$$

Apliquemos el método de eliminación gaussiana con pivoting total. Para el efecto, ponemos

$$\widetilde{A} = \left[\begin{array}{ccc|c} 354. & 100. & 30. & 244. \\ 100. & 30. & 10. & 70. \\ 30. & 10. & 5. & 25. \end{array}\right]$$

entonces

$$\begin{aligned} \widetilde{A}^{(1)} &= \left[\begin{array}{ccc|c} 354. & 100. & 30. & 244. \\ 0. & 1{,}75141243 & 1{,}525423729 & 1{,}073446329 \\ 0. & 1{,}525423729 & 2{,}457627119 & 4{,}322033899 \end{array}\right], \\ \widetilde{A}^{(2)} &= \left[\begin{array}{ccc|c} 354. & 100. & 30. & 244. \\ 0. & 1{,}75141243 & 1{,}525423729 & 1{,}073446329 \\ 0. & 0. & 1{,}129032259 & 3{,}387096774 \end{array}\right], \end{aligned}$$

y de este sistema de ecuaciones triangular superior, se obtiene

$$\widehat{a}_1 = 2{,}999999997 \simeq 3{,}0, \quad \widehat{a}_2 = -1{,}999999996 \simeq -2{,}0, \quad \widehat{a}_3 = 1.$$

El polinomio buscado es $p(x) = 3 - 2x + x^2 \quad x \in \mathbb{R}$.

9.5. Ajuste de datos con funciones afines de n variables

Supongamos que se dispone de un conjunto de datos

$$S=\left\{\left(x_1^{(k)},\ldots,x_n^{(k)},z_k\right)\in\mathbb{R}^{n+1}\mid k=1,\ldots,m\right\},$$

y se desea hallar los coeficientes de la función f definida como

$$z=f\left(x_1,\ldots,x_n\right)=a_0+a_1x+\cdots+a_nx_n\quad\left(x_1,\ldots,x_n\right)\in\mathbb{R}^n.$$

Se supone que $m\geq n+1$.

Si $n=1$, la función f se escribe como $f\left(x\right)=a+bx\quad x\in\mathbb{R}$ y coincide con la de un polinomio de grado 1 que ya hemos arriba tratado. Del punto de vista geométrico, la gráfica de f representa una recta del plano, por este motivo llamamos recta de mejor ajuste.

Si $n=2$, se tiene $z=f\left(x,y\right)=a_0+a_1x+a_2y\quad\left(x,y\right)\in\mathbb{R}^2$, que del punto de vista geométrico, la gráfica de f se identifica con un plano.

Si $n\geq 3$, a la función f lo identificamos con la ecuación cartesiana de un hiperplano.

Tal como en el caso del ajuste de datos polinomial, se tiene el siguiente sistema de ecuaciones:

$$z_k=a_0+a_1x_1^{(k)}+\cdots+a_nx_n^{(k)}+r_k\quad k=1,\ldots,m,$$

donde r_k denota la perturbación del dato experimental $\left(x_1^{(k)},\ldots,x_n^{(k)},z_k\right)$ y que depende de $a_0,a_1,\ldots,a_n$.

Se define $\overrightarrow{x}^T=\left(a_0,a_1,\ldots,a_n\right)\in\mathbb{R}^{n+1}$, $\overrightarrow{z}^T=\left(z_1,\ldots z_m\right)$, $\overrightarrow{r}\left(\overrightarrow{x}\right)^T=\left(r_1\left(\overrightarrow{x}\right),\ldots,r_m\left(\overrightarrow{x}\right)\right)$ y $A=\begin{bmatrix}1 & x_1^{(1)} & \ldots & x_n^{(1)}\\ \vdots & \vdots & & \vdots\\ 1 & x_n^{(1)} & \ldots & x_n^{(m)}\end{bmatrix}$.

El sistema de ecuaciones precedente se escribe en forma matricial como $\overrightarrow{z}=A\overrightarrow{x}+\overrightarrow{r}\left(\overrightarrow{x}\right)$ o bien $r\left(\overrightarrow{x}\right)=\overrightarrow{z}-A\overrightarrow{x}\quad\overrightarrow{x}\in\mathbb{R}^{n+1}$.

Se define la función E de $\mathbb{R}^{n+1}$ en $[0,\infty[$ como sigue:

$$E\left(a_0,a_1,\ldots,a_n\right)=\|\overrightarrow{r}\left(\overrightarrow{x}\right)\|^2=\|\overrightarrow{z}-A\overrightarrow{x}\|^2=\sum_{k=1}^{m}\left(z_k-a_0-a_1x_1^{(k)}-\cdots-a_nx_n^{(k)}\right)^2$$

y consideramos el problema de minimización siguiente:

$$\text{hallar }\widehat{x}^T=\left(\widehat{a}_0,\ldots,\widehat{a}_n\right)\in\mathbb{R}^{n+1}\text{ tal que }E\left(\widehat{a}_0,\ldots,\widehat{a}_n\right)=\min_{\left(a_0,\ldots,a_n\right)\in\mathbb{R}^{n+1}}E\left(a_0,\ldots,a_n\right).$$

Las condiciones necesarias de extremo implican

$$\left\{\begin{array}{c}\dfrac{\partial E}{\partial a_0}\left(a_0,\ldots,a_n\right)=0\\ \vdots\\ \dfrac{\partial E}{\partial a_n}\left(a_0,\ldots,a_n\right)=0\end{array}\right.\Leftrightarrow\left\{\begin{array}{c}\displaystyle\sum_{k=1}^{m}\left(z_k-a_0-a_1x^{(k)}-\cdots-a_nx_n^{(k)}\right)=0,\\ \vdots\\ \displaystyle\sum_{k=1}^{m}\left(z_k-a_0-a_1x_1^{(k)}-\cdots-a_nx_n^{(k)}\right)x_n^{(k)}=0,\end{array}\right.$$

que se expresa en forma matricial como

$$\left(A^TA\right)\overrightarrow{x}=A^T\overrightarrow{z}\quad x\in\mathbb{R}^{n+1},$$

que es el sistema normal de ecuaciones lineales.

Este sistema se obtuvo como solución en mínimos cuadrados del sistema de ecuaciones lineales $A\overrightarrow{x}=\overrightarrow{z}$ $\overrightarrow{x}\in\mathbb{R}^{n+1}$.

Ejemplos

1. Consideremos el conjunto de datos S siguiente:

$$S = \{(1,0{,}5,9),(2,2{,}5,16),(3,3,20),(4,3{,}5,24),(5,5,32)\}.$$

Determinemos una función f del tipo $f(x,y,z) = ax + by + c \quad (x,y,z) \in \mathbb{R}^3$ y $a,\ b,\ c$ constantes determinadas con el método de mínimos cuadrados.

Definimos la matriz A y los vectores $\overrightarrow{x},\ \overrightarrow{y}$ como sigue:

$$A = \begin{bmatrix} 1 & 0{,}5 & 1 \\ 2 & 2{,}5 & 1 \\ 3 & 3 & 1 \\ 4 & 3{,}5 & 1 \\ 5 & 6 & 1 \end{bmatrix}, \quad \overrightarrow{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \in \mathbb{R}^3, \quad \overrightarrow{y} = \begin{bmatrix} 9 \\ 16 \\ 20 \\ 24 \\ 32 \end{bmatrix}$$

El problema en mínimos cuadrados es el siguiente:

$$\text{hallar } \widehat{x} \in \mathbb{R}^3 \text{ solución de } \|A\widehat{x} - \widehat{y}\|^2 = \min_{\overrightarrow{x} \in \mathbb{R}^3} \|A\overrightarrow{x} - \overrightarrow{y}\|^2.$$

La solución en mínimos cuadrados satisface la ecuación normal $A^T A\overrightarrow{x} = A^T\overrightarrow{y}$. Note que el rango de la matriz es 3, luego $A^T A$ tiene también rango 3 y en consecuencia $A^T A$ es invertible. Se tiene

$$A^T A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0{,}5 & 2{,}5 & 3 & 3{,}5 & 6 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0{,}5 & 1 \\ 2 & 2{,}5 & 1 \\ 3 & 3 & 1 \\ 4 & 3{,}5 & 1 \\ 5 & 6 & 1 \end{bmatrix} = \begin{bmatrix} 55 & 58{,}5 & 15 \\ 58{,}5 & 63{,}75 & 15{,}5 \\ 15 & 15{,}5 & 5 \end{bmatrix}$$

$$A^T\overrightarrow{y} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0{,}5 & 2{,}5 & 3 & 3{,}5 & 6 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 \\ 16 \\ 20 \\ 24 \\ 32 \end{bmatrix} = \begin{bmatrix} 357 \\ 380{,}5 \\ 101 \end{bmatrix}.$$

El sistema normal de ecuaciones lineales se escribe:

$$\begin{bmatrix} 55 & 58{,}5 & 15 \\ 58{,}5 & 63{,}75 & 15{,}5 \\ 15 & 15{,}5 & 5 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 357 \\ 380{,}5 \\ 101 \end{bmatrix}.$$

Para hallar la solución de este sistema de ecuacones lineales aplicamos el método de eliminación gaussiana con pivoting parcial. Tenemos la matriz ampliada en la que se intercambiaron la primera y segunda filas

$$\widetilde{A} = \left[\begin{array}{ccc|c} 58{,}5 & 63{,}75 & 15{,}5 & 380{,}5 \\ 55 & 58{,}5 & 15 & 357 \\ 15 & 15{,}5 & 5 & 101 \end{array}\right],$$

$$\widetilde{A}^{(1)} = \left[\begin{array}{ccc|c} 58{,}5 & 63{,}75 & 15{,}5 & 380{,}5 \\ 0. & -1{,}43589744 & 0{,}42735043 & -0{,}7350427 \\ 0. & -0{,}84615385 & 1{,}025641026 & 3{,}43589744 \end{array}\right],$$

$$\widetilde{A}^{(2)} = \left[\begin{array}{ccc|c} 58{,}5 & 63{,}75 & 15{,}5 & 380{,}5 \\ 0. & -1{,}43589744 & 0{,}42735043 & -0{,}7350427 \\ 0. & 0. & 0{,}7738095222 & 3{,}869047603 \end{array}\right].$$

La solución del sistema de ecuaciones triangular superior nos da los resultados siguientes:

$$\widehat{c} = 4{,}99999999 \simeq 5{,}0, \quad \widehat{b} = 1{,}9999999976 \simeq 2{,}0 \quad , \widehat{a} \simeq 3{,}0$$

La función buscada está definda como

$$f(x,y) = 3x + 2y + 5 \quad (x,y) \in \mathbb{R}^2.$$

2. Consideremos el conjunto de datos S siguiente:

$$S = \{(5,1,2{,}5)\,,\ (15,3,3{,}5)\,,\ (20,4,4)\,,\ (40,8,6)\}\,.$$

Determinemos, siempre que sea posible, una función del tipo $z = f(x,y) = a+bx+cy \quad (x,y) \in \mathbb{R}^2$ y $a,\ b,\ c$ constantes que se determinen con el método de mínimos cuadrados.

Tal como en el ejemplo anterior, definimos los vectores $\overrightarrow{x},\ \overrightarrow{z}$ y la matriz A como sigue:

$$\overrightarrow{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \in \mathbb{R}^3, \quad \overrightarrow{z} = \begin{bmatrix} 2{,}5 \\ 3{,}5 \\ 4{,}0 \\ 6{,}0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 5 & 1 \\ 1 & 15 & 3 \\ 1 & 29 & 4 \\ 1 & 40 & 8 \end{bmatrix}.$$

Entonces

$$A^TA = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 15 & 20 & 40 \\ 1 & 3 & 4 & 8 \end{bmatrix} \begin{bmatrix} 1 & 5 & 1 \\ 1 & 15 & 3 \\ 1 & 29 & 4 \\ 1 & 40 & 8 \end{bmatrix}. = \begin{bmatrix} 4 & 80 & 16 \\ 80 & 2250 & 450 \\ 16 & 450 & 90 \end{bmatrix},$$

$$A^T\overrightarrow{z} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 15 & 20 & 40 \\ 1 & 3 & 4 & 8 \end{bmatrix} \begin{bmatrix} 2{,}5 \\ 3{,}5 \\ 4{,}0 \\ 6{,}0 \end{bmatrix} = \begin{bmatrix} 16 \\ 385 \\ 77 \end{bmatrix}.$$

El sistema normal de ecuaciones lineales $A^TA\overrightarrow{x} = A^T\overrightarrow{z}$ se expresa como

$$\begin{bmatrix} 4 & 80 & 16 \\ 80 & 2250 & 450 \\ 16 & 450 & 90 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 16 \\ 385 \\ 77 \end{bmatrix}.$$

Debemos notar que la matriz A tiene rango 2, por lo tanto la matriz A^TA no es invertible. En este caso el sistema normal de ecuaciones lineales debe ser resuelto con el método QR de Householder.

3. Considerar el conjunto de datos dados en la tabla siguiente

i	x_i	y_i	z_i	w_i
1	0	0,25	4.	14,5
2	1	0,5	6.	21.
3	2	1.	10.	33.
4	3	1,5	20.	63.
5	4	2.	25.	78.

Sea f la función real definida por

$$w = f(x,y,z) = a + bx + cy + dz, \quad (x,y,z) \in \mathbb{R}^3,$$

donde $a,\ b,\ c,\ d$ son constantes reales que deben determinarse utilizando el método de mínimos cuadrados.

i) Halle el sistema normal de ecuaciones.

ii) Aplique el método de Householder al sistema de ecuaciones normales para hallar una solución (si existe) del sistema de ecuaciones normales

Solución

Se busca una función real f definida como

$$f(x,y,z) = a + bx + cy + dz, \quad (x,y,z) \in \mathbb{R}^3,$$

donde $a,\ b,\ c,\ d$ son constantes reales que deben determinarse utilizando el método de mínimos cuadrados y usando la información suministrada en la tabla, esto es

$$w_i = f(x_i, y_i, z_i) = a + bx_i + cy_i + dz_i + r_i, \quad i = 1, \ldots, n \tag{1}$$

con $r_i \in \mathbb{R}$ el error cometido en cada observación.

Ponemos $\overrightarrow{x}^T = (a, b, c, d)$, $\overrightarrow{w}^T = (w_1, w_2, w_3, w_4, w_5)$, $\overrightarrow{r}^T = (r_1, r_2, r_3, r_4, r_5)$ y $A = (a_{ij}) \in M_{4\times 5}[\mathbb{R}]$ dada por

$$A = \begin{bmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ 1 & x_3 & y_3 & z_3 \\ 1 & x_4 & y_4 & z_4 \\ 1 & x_5 & y_5 & z_5 \end{bmatrix}.$$

El valor $\overrightarrow{r}$ es función de $\overrightarrow{x}$, escribiremos $\overrightarrow{r}(\overrightarrow{x})$.

El sistema de ecuaciones (1) se escribe en forma matricial como el siguiente:

$$A\overrightarrow{x} + \overrightarrow{r}(\overrightarrow{x}) = \overrightarrow{w}. \tag{2}$$

De (2) se obtiene

$$\overrightarrow{r}(\overrightarrow{x}) = \overrightarrow{w} - A\overrightarrow{x}.$$

El problema consiste en determinar $\widehat{x} \in \mathbb{R}^4$ solución de

$$\min_{\overrightarrow{x}\in\mathbb{R}^4} \|\overrightarrow{r}(\overrightarrow{x})\|^2$$

o lo que es lo mismo

$$\|\overrightarrow{w} - A\widehat{x}\|^2 = \min_{\overrightarrow{x}\in\mathbb{R}^4} \|\overrightarrow{w} - A\overrightarrow{x}\|^2. \tag{3}$$

i) Denotamos con $R(A)$ el rango de la matriz A. Se ha demostrado que si $R(A) = 4$, existe un único $\widehat{x} \in \mathbb{R}^4$ solución de

$$A^T A\overrightarrow{x} = A^T\overrightarrow{w}, \tag{4}$$

y que minimiza el funcional $J(\overrightarrow{x}) = \|\overrightarrow{w} - A\overrightarrow{x}\|^2$, $\overrightarrow{x} \in \mathbb{R}^4$.

Observación. Si $A = (a_{ij}) \in M_{m\times n}[\mathbb{R}]$ y $R(A) = n, \exists\widehat{x} \in \mathbb{R}^n$ tal que $A^T A\widehat{x} = A^T\overrightarrow{b}$. En nuestro caso $n = 4$.

Con la información suministrada en la tabla, se tiene

$$A = \begin{bmatrix} 1 & 0 & 0{,}25 & 4 \\ 1 & 1 & 0{,}5 & 6 \\ 1 & 2 & 1. & 10 \\ 1 & 3 & 1{,}5 & 20 \\ 1 & 4 & 2. & 25 \end{bmatrix}, \quad \overrightarrow{w} = \begin{bmatrix} 14{,}5 \\ 21 \\ 33 \\ 63 \\ 78 \end{bmatrix}.$$

Ponemos $B = A^T A$. Entonces

$$B = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0{,}25 & 0{,}5 & 1 & 1{,}5 & 2 \\ 4 & 6 & 10 & 20 & 25 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0{,}25 & 4 \\ 1 & 1 & 0{,}5 & 6 \\ 1 & 2 & 1. & 10 \\ 1 & 3 & 1{,}5 & 20 \\ 1 & 4 & 2. & 25 \end{bmatrix} = \begin{bmatrix} 5 & 10 & 5{,}25 & 65 \\ 10 & 30 & 15 & 186 \\ 5{,}25 & 15 & 7{,}5625 & 94 \\ 65 & 186 & 94 & 1177 \end{bmatrix}$$

Sea $\overrightarrow{b} = A^T\overrightarrow{w}$. Se tiene

$$\overrightarrow{b} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0{,}25 & 0{,}5 & 1 & 1{,}5 & 2 \\ 4 & 6 & 10 & 20 & 25 \end{bmatrix} \begin{bmatrix} 14{,}5 \\ 21 \\ 33 \\ 63 \\ 78 \end{bmatrix} = \begin{bmatrix} 209{,}5 \\ 588 \\ 297{,}625 \\ 3724 \end{bmatrix}$$

El sistema normal de ecuaciones es $B\overrightarrow{x} = \overrightarrow{b}$ o en forma explícita:

$$\begin{bmatrix} 5 & 10 & 5{,}25 & 65 \\ 10 & 30 & 15 & 186 \\ 5{,}25 & 15 & 7{,}5625 & 94 \\ 65 & 186 & 94 & 1177 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 209{,}5 \\ 588 \\ 297{,}625 \\ 3724 \end{bmatrix}$$

ii) Apliquemos ahora el método de Householder.

Recordemos que la matriz de Householder tiene la forma $H = I - 2\overrightarrow{u}\,\overrightarrow{u}^T$, con $\overrightarrow{u} \in \mathbb{R}^n$ y $\|\overrightarrow{u}\| = 1$. Esta matriz es simétrica y ortogonal. Por otro lado, dado $\overrightarrow{v} \in \mathbb{R}^n$ con $\overrightarrow{v} \neq 0$, existen $\alpha \in \mathbb{R}$ y H tales que $H\overrightarrow{v} = \alpha\overrightarrow{e}_1$.

Consideramos nuevamente el sistema normal de ecuaciones arriba propuesto:

$$\begin{bmatrix} 5 & 10 & 5{,}25 & 65 \\ 10 & 30 & 15 & 186 \\ 5{,}25 & 15 & 7{,}5625 & 94 \\ 65 & 186 & 94 & 1177 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 209{,}5 \\ 588 \\ 297{,}625 \\ 3724 \end{bmatrix}.$$

Etapa 1

Sea $\overrightarrow{v}^T = (5,\ 10,\ 5{,}25,\ 65)$. Entonces $\|\overrightarrow{v}\| = 66{,}16315062$, $\alpha = -66{,}16315062$,

$r = \alpha(\alpha - v_1) = -66{,}16315062\,(-66{,}16315062 - 5) = 4708{,}378253$.

Luego

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_1 = \begin{bmatrix} 5 \\ 10 \\ 2{,}25 \\ 65 \end{bmatrix} + 66{,}16315062 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 71{,}16315062 \\ 10 \\ 5{,}25 \\ 65 \end{bmatrix},$$

$$\begin{aligned} H^{(1)} &= I - \frac{1}{r}\overrightarrow{w}\,\overrightarrow{w}^T \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - 0{,}0002123873543 \begin{bmatrix} 71{,}16315062 \\ 10 \\ 5{,}25 \\ 65 \end{bmatrix} (71{,}16315062, 10, 5{,}25, 65) \\ &= \begin{bmatrix} -0{,}075570787 & -0{,}1511415328 & -0{,}07934930475 & -0{,}9824199634 \\ -0{,}1511415328 & 0{,}9787612646 & -0{,}0111503361 & -0{,}1380517803 \\ -0{,}07934930475 & -0{,}0111503361 & 0{,}9941460736 & -0{,}07247718465 \\ -0{,}9824199634 & -0{,}1380517803 & -0{,}07247718465 & 0{,}1026634281 \end{bmatrix} \end{aligned}$$

Se pone $B^{(1)} = H^{(1)}B$. Resulta

$$B^{(1)} = \begin{bmatrix} -66{,}16315063 & -189{,}2103064 & -95{,}6114252 & -1196{,}791557 \\ 0 & 2{,}006536434 & 0{,}8267341518 & 8{,}690318587 \\ 0 & 0{,}3034316251 & 0{,}1215354329 & 0{,}912417267 \\ 0 & 4{,}042300817 & 1{,}87388202 & 24{,}48707075 \end{bmatrix}$$

Además, $\overrightarrow{b}^{(1)} = H^{(1)}\overrightarrow{b}$, se tiene

$$\overrightarrow{b}^{(1)} = \begin{bmatrix} -3786{,}851578 \\ 26{,}42402386 \\ 2{,}797612564 \\ 73{,}75615504 \end{bmatrix}.$$

Etapa 2

Sea $\overrightarrow{v}^T = (0, 2{,}006536434, 0{,}3034316251, 4{,}042300817)$. Obtenemos $\|\overrightarrow{v}\| = 4{,}523102377$,

$\alpha = -4{,}523102377$, $r = \alpha(\alpha - v_1) = -4{,}523102377\,(-4{,}523102377 - 2{,}006536434) = 29{,}53422483$,

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_2 = \begin{bmatrix} 0 \\ 2{,}006536434 \\ 0{,}3034316251 \\ 4{,}042300817 \end{bmatrix} + 4{,}523102377 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 6{,}529638811 \\ 0{,}3034316251 \\ 4{,}042300817 \end{bmatrix}$$

Se pone $H^{(2)} = I - \dfrac{1}{r}\overrightarrow{w}\overrightarrow{w}^T$. Entonces

$$\begin{aligned} H^{(2)} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - 0{,}03385902307 \begin{bmatrix} 0 \\ 6{,}529638811 \\ 0{,}3034316251 \\ 4{,}042300817 \end{bmatrix} \\ &\qquad (0, 6{,}529638811, 0{,}3034316251, 4{,}042300817) \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0{,}4436195038 & -0{,}06708484571 & -0{,}8937009334 \\ 0 & -0{,}06708484571 & 0{,}9968825743 & -0{,}04153018787 \\ 0 & -0{,}8937009334 & -0{,}04153018787 & 0{,}4467369301 \end{bmatrix}. \end{aligned}$$

Se define $B^{(2)} = H^{(2)}B^{(1)}$. Se tiene

$$B^{(2)} = \begin{bmatrix} -66{,}16315063 & -189{,}2103064 & -95{,}6114252 & -1196{,}791557 \\ 0 & -4{,}523102374 & -2{,}049500383 & -25{,}80052218 \\ 0 & 0 & -0{,}01212288182 & -0{,}6903684562 \\ 0 & 0 & 0{,}09318268742 & 3{,}13484012 \end{bmatrix}$$

Además, $\overrightarrow{b}^{(2)} = H^{(2)}\overrightarrow{b}^{(1)}$. Obtenemos

$$\overrightarrow{b}^{(2)} = \begin{bmatrix} -3786{,}851578 \\ -77{,}82583436 \\ -2{,}046867324 \\ 9{,}218238109 \end{bmatrix}.$$

Etapa 3

Sea $\overrightarrow{v} = \begin{bmatrix} 0 \\ 0 \\ -0{,}01212288182 \\ -0{,}09318268742 \end{bmatrix}$, $\|\overrightarrow{v}\| = 0{,}09396795996$, $\alpha = 0{,}09396795996$,

$r = \alpha(\alpha - v_1) = 0{,}09396795996\,(0{,}09396795996 + 0{,}01212288182) = 0{,}009969139979$.

$$\overrightarrow{w} = \overrightarrow{v} - \alpha\overrightarrow{e}_3 = \begin{bmatrix} 0 \\ 0 \\ -0{,}01212288182 \\ 0{,}09318268742 \end{bmatrix} - 0{,}09396795996 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -0{,}1060908418 \\ 0{,}09318268742 \end{bmatrix}$$

Se define $H^{(3)} = I - \dfrac{1}{r}\overrightarrow{w}\overrightarrow{w}^T$. Luego

$$H^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0{,}1290108007 & 0{,}991643188 \\ 0 & 0 & 0{,}991643188 & 0{,}129010802 \end{bmatrix}.$$

Además, $B^{(3)} = H^{(3)}B^{(2)}$, con lo que

$$B^{(3)} = \begin{bmatrix} -66{,}16315063 & -189{,}2103064 & -95{,}6114252 & -1196{,}791557 \\ 0 & -4{,}523102374 & -2{,}049500383 & -25{,}80052218 \\ 0 & 0 & 0{,}09396795991 & 3{,}197707838 \\ 0 & 0 & 0 & -0{,}2801709388 \end{bmatrix},$$

$$\overrightarrow{b}^{(3)} = H^{(3)}\overrightarrow{b}^{(2)} = \begin{bmatrix} -3786{,}851578 \\ -77{,}82583436 \\ 9{,}405271019 \\ -0{,}8405097475 \end{bmatrix}$$

Resolución del sistema triangular superior:

$$\begin{aligned} d &= \frac{-0{,}8405097475}{-0{,}2801709388} = 2{,}999989046, \\ c &= \frac{9{,}405271019 - 3{,}197707838 \times d}{0{,}09396795991} = -1{,}998739434, \\ b &= \frac{-77{,}82583436 + 2{,}049500383 \times c + 25{,}80052218}{-4{,}523102374} = 0{,}9995280443, \\ a &= \frac{-3786{,}851578 + 189{,}2103064 \times b + 95{,}6114252 \times c + 1196{,}791557}{-66{,}16315063} = 2{,}999726189 \end{aligned}$$

La solución es:

$$\overrightarrow{x}^T = (2{,}999726189, 0{,}9995280443, -1{,}998739434, 2{,}999989046)\,.$$

9.6. Ajuste de datos con funciones dependientes de un parámetro

Sea $S = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots n\}$ un conjunto de datos experimentales. Supongamos que se busca una función real f dependiente de la variable $x \in \mathbb{R}$ y del parámetro $a \in \mathbb{R}$ a determinar según el conjunto de datos S, esto es,

$$y_i = f(x_i, a) + r_i \quad i = 1, \ldots, n,$$

donde r_i es una perturbación de y_i que depende del parámetro a, escribiremos $r_i(a)$. En forma matricial este sistema de ecuaciones se escribe como

$$\overrightarrow{y} = \overrightarrow{b}(a) + \overrightarrow{r}(a),$$

donde $\overrightarrow{y}, \overrightarrow{b}(a), \overrightarrow{r}(a) \in \mathbb{R}^n$ definimos a continuación

$$\overrightarrow{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \overrightarrow{b}(a) = \begin{bmatrix} f(x_1, a) \\ \vdots \\ f(x_n, a) \end{bmatrix}, \quad \overrightarrow{r}(a) = \begin{bmatrix} r_1(a) \\ \vdots \\ r_n(a) \end{bmatrix}.$$

Resulta

$$\overrightarrow{r}(a) = \overrightarrow{y} - \overrightarrow{b}(a) = \begin{bmatrix} y_1 - f(x_1, a) \\ \vdots \\ y_n - f(x_n, a) \end{bmatrix}.$$

La determinación del parámetro a y del vector $\overrightarrow{r}(a)$, en general es complicada. recurrimos entonces a aproximar el parámetro a mediante el método de mínimos cuadrados. Para el efecto definimos una función real E como sigue:

$$E(a) = \|\overrightarrow{r}(a)\|^2 = \left\|\overrightarrow{y} - \overrightarrow{b}(a)\right\|^2 = \sum_{i=1}^{n} (y_i - f(x_i, a))^2 \quad a \in \mathbb{R}$$

y consideramos el problema siguiente:

$$\text{hallar } \widehat{a} \in \mathbb{R} \text{ solución de } \min_{a \in \mathbb{R}} E(a),$$

que a su vez es equivalente al siguiente:

$$\text{hallar } \widehat{a} \in \mathbb{R} \text{ tal que } E(\widehat{a}) \leq E(a) \quad \forall a \in \mathbb{R}.$$

Escribiremos $E(\widehat{a}) = \min_{a\in\mathbb{R}} E(a)$.

Supondremos que la función f es diferenciable y que $\dfrac{\partial f}{\partial a}(x,a)$ es continua. De las condiciones necesarias de extremo se tiene que $E'(a) = 0$. Luego

$$E'(a) = \frac{d}{da}\sum_{i=1}^{n}(y_i - f(x_i,a))^2 = \sum_{i=1}^{n} -2(y_i - f(x_i,a))\frac{\partial f}{\partial a}(x_i,a)$$

con lo que

$$E'(a) = 0 \Leftrightarrow \sum_{i=1}^{n}(y_i - f(x_i,a))\frac{\partial f}{\partial a}(x_i,a) = 0.$$

Note que la función f depende de x y de a.

Se define $\varphi(a) = \sum_{i=1}^{n}(y_i - f(x_i,a))\dfrac{\partial f}{\partial a}(x_i,a)$ y consideramos la ecuación:

$$\text{hallar } \widehat{a} \in \mathbb{R} \text{ tal que } \varphi(\widehat{a}) = 0.$$

La ecuación precedente, en general, es no lineal. Su resolución es a menudo muy complicada y se recurre a la aplicación de métodos iterativos de resolución de ecuaciones no lineales que han sido estudiados anteriormente, estre ellos, el método de bisección, punto fijo modificado y método de Newton son los más utilizados y que están relacionados con la regularidad de la función f.

A continuación presentamos algunas funciones f que son muy utilizadas:

1. $f(x,a) = \lambda\exp(ax) \quad x \in \mathbb{R}$,
2. $f(x,a) = \lambda\,\text{sen}(ax) \quad x \in \mathbb{R}$,
3. $f(x,a) = \lambda\cos(ax) \quad x \in \mathbb{R}$,
4. $f(x,a) = \lambda + x^a \quad x \in \mathbb{R}$,
5. $f(x,a) = \lambda\arctan(ax) \quad x \in \mathbb{R}$,

donde $\lambda \in \mathbb{R}$ es una constante conocida y $a \in \mathbb{R}$ es el parámetro a determinar.

Para calcular una aproximación del parámetro a, para algunas funciones f como las indicadas, recurriremos a proponer un problema alternativo en el que se ha linealizado mediante algún procedimiento particular. Aclararemos esta situación con ejemplos.

Ejemplos

1. Con el propósito de realizar comparaciones consideramos la función g definida como $g(x) = 5{,}2\exp(-0{,}282x) \quad x \geq 0$ y con esta función g obtenemos el conjunto de datos S siguiente:

   $$S = \{(0,5{,}2),(1{,}5,3{,}406),(4,1{,}683),(8{,}5,0{,}473),(10,0{,}31)\}.$$

   Buscamos una función f definida como $f(x,a) = 5{,}2\exp(ax) \quad x \geq 0$, con a un parámetro a determinar como solución de la ecuación

   $$\begin{aligned}
   \sum_{i=1}^{n}(y_i - f(x_i,a))\frac{\partial f}{\partial a}(x_i,a) &= 0 \Longleftrightarrow \\
   \sum_{i=1}^{n}(y_i - 5{,}2\exp(ax_i))(5{,}2x_i\exp(ax_i)) &= 0 \Longleftrightarrow \\
   \sum_{i=1}^{n}(y_i - 5{,}2\exp(ax_i))x_i\exp(ax_i) &= 0.
   \end{aligned}$$

Este tipo de ecuaciones son muy laboriosas. Otra alternatica ampliamente utilizada es la siguiente. Ponemos $y = 5{,}2\exp(ax)$ entonces $\ln(y) = ax + \ln(5{,}2)$. Denotamos con $u = \ln(y)$ y $b = \ln(5{,}2) \simeq 1{,}648658626$, $u_i = \ln(y_i) \quad i = 1, \ldots, n$.

El problema alternatico a resolver es el siguiente. Se define

$$E_1(a) = \sum_{i=1}^{n} (u_i - ax_i - b)^2$$

y de las condiciones de extremo se tiene

$$\begin{aligned} E_1'(a) &= 0 \Leftrightarrow \sum_{i=1}^{n} (u_i - ax_i - b)\, x_i = 0 \Leftrightarrow \sum_{i=1}^{n} u_i x_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ &\Leftrightarrow a = \frac{\sum_{i=1}^{n} y_i x_i - b\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}. \end{aligned}$$

Calculamos $u_1 = \ln(5{,}2)$, $u_2 = \ln(3{,}406)$, $u_3 = \ln(1{,}683)$, $u_4 = \ln(0{,}473)$ y $u_5 = \ln(0{,}31)$. Además $\sum_{i=1}^{n} x_i = 24$, $\sum_{i=1}^{n} x_i^2 = 190{,}5$, $\sum_{i=1}^{n} u_i x_i = -14{,}15481935$, resulta

$$a = \frac{\sum_{i=1}^{n} y_i x_i - b\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = \frac{-14{,}15481935 - 1{,}648658626 \times 24}{190{,}5} = -0{,}2820085373$$

que redondeando se tiene $a = -0{,}282$.

La función buscada es por lo tanto $f(x, a) = 5{,}2\exp(-0{,}282) = g(x) \quad x \geq 0$. En la figura siguiente se muestran los puntos del conjunto S y la gráfica de la función f.

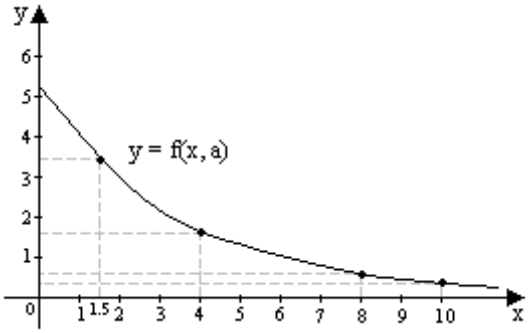


Figura 85

2. Nuevamente, con el propósito de realizar comparaciones, consideramos la función g definida como $g(x) = 3\,\mathrm{sen}(2{,}5x) \quad x \in [0, \pi]$ y S el conjunto de datos siguiente:

$$S = \{(0,0), (0{,}52, 2{,}9), (0{,}78, 2{,}79), (1, 1{,}8), (1{,}57, -2{,}12), (2{,}1, -2{,}57), (2{,}62, 0{,}8), (\pi, 0)\}.$$

Buscamos una función f de la forma $f(x, a) = 3\,\mathrm{sen}(ax) \quad x \in [0, \pi]$. De acuerdo a los resultados obtenidos anteriormente, para determinar el parámetro $a \in \mathbb{R}$ con el método de mínimos cuadrados debemos resolver la ecuación

$$\sum_{i=1}^{n} (y_i - f(x_i, a)) \frac{\partial f}{\partial a}(x_i, a) = 0.$$

Como $f(x,a) = 3\,\mathrm{sen}\,(ax)$ se sigue que $\dfrac{\partial f}{\partial a}(x,a) = 3x\cos(ax)$, con lo que la ecuación a resolver.se escribe como

$$\sum_{i=1}^{n} (y_i - f(x_i,a))\, x_i \cos(ax_i) = 0,$$

ecuación que es muy laboriosa de resolver.

Por otro lado, fijado $a \in \mathbb{R}$, la función f es solución de la ecuación diferencial $u''(x) + a^2 u(x) = 0$ $\forall x \in \mathbb{R}$ pues $\dfrac{\partial^2 f}{\partial x^2}(x,a) = -a^2 f(x,a)$. Es esta ecuación diferencial aparece axplícitamente el parámetro $a \in \mathbb{R}$ y no figuran las funciones $\mathrm{sen}\,(ax)$ y $\cos(ax)$. Por lo tanto el problema alternatico a resolver se construye como sigue:

$$y_i'' - a^2 y_i = r_i \quad i = 2, \ldots, n-1$$

donde y_i'' es la aproximación de la derivada segunda de f respecto de x, aproximación que se le calcula con diferencias finitas centrales estudiadas en el capítulo II.

Se define

$$\overrightarrow{w} = \begin{bmatrix} y_2'' \\ \vdots \\ y_{n-1}'' \end{bmatrix}, \quad \overrightarrow{y} = \begin{bmatrix} y_2 \\ \vdots \\ y_{n-1} \end{bmatrix}, \quad \overrightarrow{r}(a) = \begin{bmatrix} r_2(a) \\ \vdots \\ r_{n-1}(a) \end{bmatrix},$$

con lo que la relación precedente se escribe como

$$\overrightarrow{w} - a^2 \overrightarrow{y} = r(\overrightarrow{a}),$$

y definimos la función $E_1(a)$ como

$$E_1(a) = \|\overrightarrow{r}(a)\|^2 = \left\|\overrightarrow{w} - a^2\overrightarrow{y}\right\|^2 = \sum_{i=2}^{n-1} \left(y_i'' - a^2 y_i\right)^2,$$

donde $\|\cdot\|$ es la norma euclídea en $\mathbb{R}^{n-3}$.

El problema alternativo que se propone es el siguiente:

$$\text{hallar } \widehat{a} \in \mathbb{R} \text{ tal que } E_1(\widehat{a}) = \min_{a\in\mathbb{R}} E_1(a),$$

que por las condiciones necesarias de extremo conduce a resolver la ecuación $E_1'(a) = 0$. Así,

$$E_1'(a) = \sum_{i=3}^{n-2} \left(y_i'' - a^2 y_i\right) a y_i = 0 \Leftrightarrow \sum_{i=3}^{n-2} y_i'' y_i - a^2 y_i^2 = 0 \Leftrightarrow a^2 = \frac{\sum_{i=3}^{n-2} y_i'' y_i}{\sum_{i=3}^{n-2} y_i^2}.$$

Queda calcular $y_i'' \quad i = 2, \ldots, n-1$, $\sum_{i=3}^{n-2} y_i'' y_i$, $\sum_{i=3}^{n-2} y_i^2$.

Recordemos que la derivada primera se aproxima con diferencias finitas centrales mediante el cociente

$$y_i' = \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \quad i = 2, \ldots, n-1,$$

y la derivada segunda se aproxima con diferencias finitas centrales mediante el cociente

$$y_i'' = \frac{y_{i+1}' - y_{i-1}'}{x_{i+1} - x_{i-1}} \quad i = 3, \ldots, n-2..$$

Se propone como ejercicio realizar los cálculos para obtener el valor de a.

9.7. Mínimos cuadrados continuos.

Sean $V = C([a,b])$ el espacio de funciones continuas provisto del producto escalar

$$\langle f, g\rangle = \int_a^b f(x)g(x)dx \quad \forall f, g \in C([a,b]).$$

Sea $\mathbb{K}_m[\mathbb{R}]$ el espacio de polinomios de grado $\leq m$. Se tiene $\dim \mathbb{K}_m[\mathbb{R}] = m+1$ y $\mathbb{K}_m[\mathbb{R}]$ es un subespacio cerrado de $C([a,b])$. Sea $\{\varphi_0, \ldots, \varphi_m\}$ una base de $\mathbb{K}_m[\mathbb{R}]$. Entonces

$$\mathbb{K}_m[\mathbb{R}] = \left\{ \sum_{i=o}^{m} \alpha_i \varphi_i \mid \alpha_i \in \mathbb{R},\ i = 0, \ldots, m \right\}.$$

Sea $f \in C([a,b])$. Existe $\hat{g} \in \mathbb{K}_m[\mathbb{R}]$ tal que

$$\|f - \hat{g}\| = \operatorname*{Min}_{g \in \mathbb{K}_m[\mathbb{R}]} \|f - g\|, \qquad \langle f - \hat{g}, w\rangle = 0 \quad \forall w \in \mathbb{K}_m[\mathbb{R}].$$

Note que $\hat{g} \in \mathbb{K}_m[\mathbb{R}] \Longleftrightarrow \exists \hat{a}_0, \ldots, \hat{a}_m \in \mathbb{R}$ tales que $\hat{g} = \sum_{i=o}^{m} \hat{a}_i \varphi_i$. Además,

$$\|f - \hat{g}\|^2 = \operatorname*{Min}_{g \in \mathbb{K}_m[\mathbb{R}]} \|f - g\|^2,$$

$$\langle f - \hat{g}, w\rangle = 0 \quad \forall w \in \mathbb{K}_m[\mathbb{R}] \Longleftrightarrow \int_a^b \left(f(x) - \sum_{i=o}^{m} \hat{a}_i \varphi_i(x) \right) w(x)dx = 0 \quad \forall a_0, \ldots, a_m \in \mathbb{R}.$$

Poniendo sucesivamente $w = \varphi_i \quad i = 1, \cdots, n$ se obtiene el sistema de ecuaciones siguiente:

$$\int_a^b f(x)\varphi_j(x)dx - \sum_{i=o}^{m} \hat{a}_i \int_a^b \varphi_i(x)\varphi_j(x)dx = 0 \quad j = 0, \ldots, m,$$

o en forma explícita dicho sistema tiene la forma

$$\begin{cases} \sum_{i=o}^{m} \hat{a}_i \int_a^b \varphi_i(x)\varphi_0(x)dx = \int_a^b f(x)\varphi_0(x)dx, \\ \vdots \\ \sum_{i=o}^{m} \hat{a}_i \int_a^b \varphi_i(x)\varphi_m(x)dx = \int_a^b f(x)\varphi_m(x)dx, \end{cases}$$

que en forma matricial se expresa como $A\overrightarrow{x} = \overrightarrow{b}$, donde $A = (a_{ij})$ es la matriz definida como $a_{ij} = \int_a^b \varphi_i(x)\varphi_j(x)dx$, $\overrightarrow{b} = (b_0, \ldots, b_m)$ con $b_j = \int_a^b f(x)\varphi_j(x)dx \quad j = 0, \ldots, m$, $\overrightarrow{x}^T = (\hat{a}_0, \ldots, \hat{a}_m)$ es el vector de las incógnitas..La matriz A es simétrica, definida positiva por lo tanto invertible.

9.8. Aproximación numérica de series de Fourier

9.8.1. Preliminares

Sean $L > 0$, a_0, a_k, $b_k \in \mathbb{R} \quad$ con $k = 1, 2, \cdots$. Las series de funciones de la forma

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \right)$$

se llaman series de Fourier y aparecen en algunas aplicaciones de la matemática tales como el procesamiento de la señal y de imágenes, en la resolución de algunas clases de ecuaciones en derivadas

parciales. Nos interesamos en la aproximación numérica de esta clase de series de funciones. Para ello primeramente introducimos algunas notaciones.

Sean $L > 0$. Se denota con $C([-L, L])$ al espacio vectorial de las funciones continuas en $[-L, L]$. Proveemos a $C([-L, L])$ del producto escalar notado $\langle \cdot, \cdot \rangle$ y definido como (véase en el apéndice los espacios con producto interior)

$$\langle u, v \rangle = \int_{-L}^{L} u(x) v(x) \, dx \quad \forall u, v \in C([-L, L]),$$

y la norma asociada a este producto escalar se nota $\| \cdot \|$ y se define como

$$\| u \| = \left(\int_{-L}^{L} | u(x) |^2 \, dx \right)^{\frac{1}{2}} \quad \forall u \in C([-L, L]).$$

Definición 1 *Sean $L > 0$ y u una función real definida en todo $\mathbb{R}$. Se dice que u es periódica de período $2L$ si y solo si se verifica*

$$u(x + 2L) = u(x) \quad \forall x \in \mathbb{R}.$$

En la figura siguiente se muestra la gráfica de una función u periódica de período $2L$.

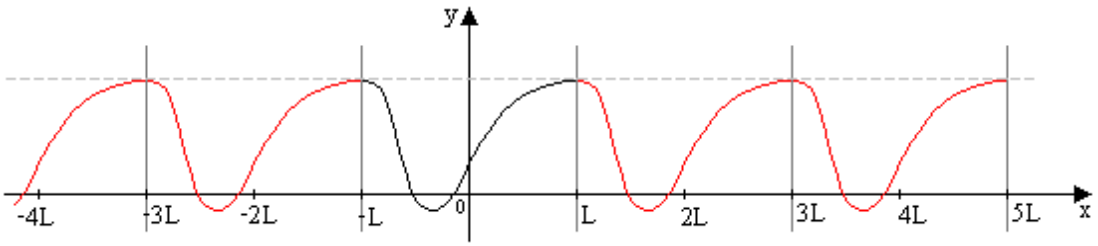


Figura 86

Observe que para $x = -L$ se tiene $u(L) = u(-L)$.

Sea $u \in C([-L, L])$. A la función u lo extendemos a todo $\mathbb{R}$ por periodicidad y por abuso de lenguaje lo notamos aún con u, es decir que

$$u(x + 2L) = u(x) \quad \forall x \in \mathbb{R}.$$

Definición 2 *Sea $u \in C([-L, L])$.*

i) Se dice que u es par si y solo si u verifica la propiedad $u(-x) = u(x) \quad \forall x \in [-L, L]$.

ii) Se dice que u es impar si y solo si verifica la propiedad $u(-x) = -u(x) \quad \forall x \in [-L, L]$.

En la izquierda de la figura siguiente se muestra la gráfica de una función par; y, a la derecha está

representada la gráfica de una función impar.

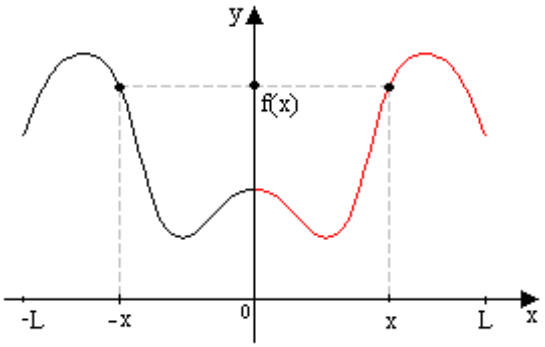


Figura 87

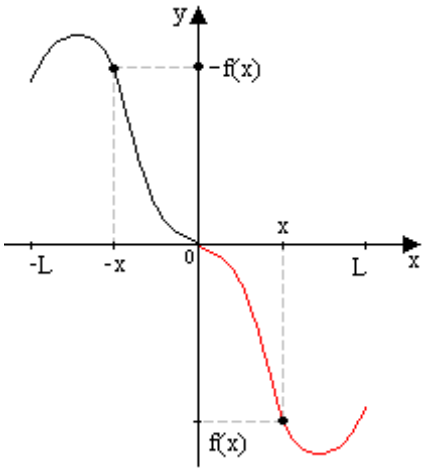


Figura 88

Sea $u \in C([-L, L])$. Se verifica inmediatamente las siguientes propiedades.

i) Si u es impar, $\displaystyle\int_{-L}^{L} u(x)\,dx = 0$.

ii) Si u es par, $\displaystyle\int_{-L}^{L} u(x)\,dx = 2\int_{0}^{L} u(x)\,dx$.

iii) La función f definida como $f(x) = \dfrac{1}{2}(u(x) + u(-x)) \quad x \in [-L, L]$ es par.

iv) La función g definida como $g(x) = \dfrac{1}{2}(u(x) - u(-x)) \quad x \in [-L, L]$ es impar.

Además $u(x) = f(x) + g(x) \quad \forall x \in [-L, L]$.

Definición 3

i) Sean $u, v \in C([-L, L])$. Se dice que u es ortogonal o perpendicular a v, que se nota $u \perp v$, si y solo si $\langle u, v\rangle = 0$.

ii) Sean $u \in C([-L, L])$, $M \subset C([-L, L])$ con $M \neq \phi$. Se dice que u es ortogonal a M, que se escribe $u \perp M$, si y solo si $\langle u, v\rangle = 0 \quad \forall v \in M$.

iii) Sean M, N dos subconjuntos no vacíos de $C([-L, L])$. Se dice que M es ortogonal a N, que se nota $M \perp N$, si y solo si $\langle u, v\rangle = 0 \quad \forall u \in M$, $\forall v \in N$.

iv) Sea $M \subset C([-L, L])$ con $M \neq \emptyset$. Se dice que M es ortogonal si y solo si $\langle u, v\rangle = 0 \quad \forall u, v \in M$, $u \neq v$.

v) Se dice que M es ortonomal si y solo si M es ortogonal, y $\forall u \in M$, $\|u\| = 1$.

Los siguientes conjuntos de funciones son muy importantes en el desarrollo en series de Fourier de funciones reales periódicas de período $2L$ y continuas a trozos en el intervalo $[-L, L]$. Sean M, N los subconjuntos de $C([-L, L])$ definidos como $M = \{\varphi_k \mid k \in \mathbb{N}\}$, $N = \{\psi_k \mid k \in \mathbb{Z}^+\}$, donde

$$\begin{aligned}
\varphi_0(x) &= 1 \quad x \in [-L, L], \\
\varphi_k(x) &= \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \quad k = 1, 2, \ldots, \\
\psi_k(x) &= \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \quad k = 1, 2, \ldots
\end{aligned}$$

Se tiene

i) M es un conjunto ortogonal.

ii) N es un conjunto ortogonal.

iii) $M \perp N$.

Sea $f \in C([-L, L])$. Supongamos que f se prolonga por periodicidad a todo $\mathbb{R}$ con periódo $2L$, esto es,

$$f(x + 2L) = f(x) \quad \forall x \in \mathbb{R}.$$

Admitimos que la función f se representa como una serie de Fourier, es decir que se verifica

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)\right) \quad x \in [-L, L],$$

donde a_0, a_k, $b_k \in \mathbb{R}$ $k = 1, 2, \ldots$, son los coeficientes de Fourier definidos a continuación:

$$\begin{aligned} a_0 &= \frac{1}{L}\int_{-L}^{L} f(x)\, dx, \\ a_k &= \frac{1}{L}\int_{-L}^{L} f(x) \cos\left(\frac{k\pi x}{L}\right) dx \quad k = 1, 2, \ldots, \\ b_k &= \frac{1}{L}\int_{-L}^{L} f(x) \operatorname{sen}\left(\frac{k\pi x}{L}\right) dx \quad k = 1, 2, \ldots. \end{aligned}$$

Si la función f es par, se tiene que $f(x) \operatorname{sen}\left(\frac{k\pi x}{L}\right)$ $x \in [-L, L]$ es impar y en consecuencia $b_k = 0$ $k = 1, 2, \ldots$. Resulta que la serie de Fourier de f se escribe como

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L],$$

y

$$a_k = \frac{2}{L}\int_{0}^{L} f(x) \cos\left(\frac{k\pi x}{L}\right) dx \quad k = 0, 1, \ldots.$$

Si la función f es impar, se tiene $f(x) \cos\left(\frac{k\pi x}{L}\right) dx$ $x \in [-L, L]$ impar. Luego $a_k = 0$ $k = 0, 1, \ldots,$ con lo que la serie de Fourier de f se escribe como sigue:

$$f(x) = \sum_{k=1}^{\infty} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L],$$

con

$$b_k = \frac{2}{L}\int_{0}^{L} f(x) \operatorname{sen}\left(\frac{k\pi x}{L}\right) dx \quad k = 1, 2, \ldots.$$

Definición 4 *Sean $m \in \mathbb{Z}^+$. La función $Q_m \in C([-L, L])$ definida como*

$$Q_m(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \left(a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)\right) \quad x \in [-L, L]$$

se llama polinomio trigonométrico, donde $a_0, a_k, b_k \in \mathbb{R}$ $k = 1, \ldots, m$.

Sean $u_0(x) = \frac{a_0}{2}$, $u_k(x) = a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)$ $x \in [-L, L]$. A las funciones u_0, u_k $k = 1, \ldots, m$ se les denomina armónicos de Q_m.

En el caso en que a_0, a_k, b_k son los coeficientes de Fourier, el polinomio trogonométrico Q_m se le denomina polinomio trogonométrico de Fourier.

En lo que sigue, nos interesamos en los polinomios trigonométricos de Fourier siguientes:

i) Si $f \in C([-L,L])$ es par, $Q_m(x) = \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L,L]$,

donde $a_k = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{k\pi x}{L}\right)dx \quad k = 0,1,\ldots,m.$

ii) Si $f \in C([-L,L])$ es impar, $Q_m(x) = \sum_{k=1}^{m} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L,L]$,

donde $b_k = \frac{2}{L}\int_0^L f(x)\operatorname{sen}\left(\frac{k\pi x}{L}\right)dx \quad k = 1,\ldots,m.$

iii) Si $f \in ([-L,L])$ es arbitraria

$$Q_m(x) = \frac{a_0}{2} + \sum_{k=1}^{m}\left(a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)\right) \quad x \in [-L,L]$$

y a_0, a_k, b_k $k = 1,\ldots,m$ los coeficientes de Fourier arriba definidos.

Teorema 3 *Sean $f \in C([-L,L])$, $m \in \mathbb{Z}^+$. Se definen*

$$\begin{aligned} Q_m(x) &= \sum_{k=0}^{m} a_k \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L,L], \\ E(a_0,\ldots,a_m) &= \|f - Q_m\|^2 = \int_{-L}^{L}\left[f(x) - \left(a_0 + a_1\cos\left(\frac{\pi x}{L}\right) + \cdots + a_m\cos\left(\frac{m\pi x}{L}\right)\right)\right]^2 dx. \end{aligned}$$

Existen $\widehat{a}_0,\ldots,\widehat{a}_m \in \mathbb{R}$ tales que $E(\widehat{a}_0,\ldots,\widehat{a}_m) = \min_{(a_0,\ldots,a_m)\in\mathbb{R}^{m+1}} E(a_0,\ldots,a_m)$, y $\widehat{a}_j \quad j = 1,\ldots,m$ los coeficientes de Fourier definidos como $a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\,dx$ y $a_j = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{j\pi x}{L}\right)dx$ $j = 1,\ldots,m$.

Demostración. De las condiciones necesarias de extremo, se tiene

$$\bigtriangledown E(a_0,\ldots,a_m) = \overrightarrow{0} \Leftrightarrow \frac{\partial E}{\partial a_j}(a_0,\ldots,a_m) = 0 \quad j = 0,1,\ldots,m.$$

Además, de la definición del funcional E, se tiene

$$\begin{aligned} \frac{\partial E}{\partial a_j}(a_0,\ldots,a_m) &= \frac{\partial}{\partial a_j}\int_{-L}^{L}\left[f(x) - \left(a_0 + a_1\cos\left(\frac{\pi x}{L}\right) + \cdots + a_m\cos\left(\frac{m\pi x}{L}\right)\right)\right]^2 dx \\ &= \int_{-L}^{L} 2\left[f(x) - \left(a_0 + a_1\cos\left(\frac{\pi x}{L}\right) + \cdots + a_m\cos\left(\frac{m\pi x}{L}\right)\right)\left(-\cos\left(\frac{j\pi x}{L}\right)\right)\right]dx \\ &= -2\int_{-L}^{L}\left[f(x)\cos\left(\frac{j\pi x}{L}\right) - \left(a_0\cos\left(\frac{j\pi x}{L}\right) + \cdots + a_m\cos\left(\frac{m\pi x}{L}\right)\cos\left(\frac{j\pi x}{L}\right)\right)\right]dx \end{aligned}$$

Luego,

$$\begin{aligned} \frac{\partial E}{\partial a_j}(a_0,\ldots,a_m) &= 0 \Leftrightarrow a_0\int_{-L}^{L}\cos\left(\frac{j\pi x}{L}\right)dx + \cdots + a_m\int_{-L}^{L}\cos\left(\frac{m\pi x}{L}\right)\cos\left(\frac{j\pi x}{L}\right)dx = \\ &= \int_{-L}^{L} f(x)\cos\left(\frac{j\pi x}{L}\right)dx \quad j = 0,1,\ldots,m. \end{aligned}$$

Por otro lado, como el conjunto de funciones $\{\varphi_j \mid j = 0,1,\ldots,m\}$ con $\varphi_0(x) = 1$, $\varphi_j(x) = \cos\left(\frac{j\pi x}{L}\right)$ $x \in [-L,L]$, $\quad j = 1,\ldots,m$, es un conjunto ortogonal, entonces $\langle\varphi_k,\varphi_j\rangle = 0$ si $j \neq k$ y $\langle\varphi_k,\varphi_j\rangle = \|\varphi_k\|^2$

si $j = k$. Además $\|\varphi_0\|^2 = 2L$, y $\left\|\varphi_j\right\|^2 = L \quad j = 1, \ldots, m$. Por lo tanto, el sistema de ecuaciones precedente se reduce al siguiente:

$$\begin{aligned} \text{para } j &= 0, \quad 2La_0 = \int_{-L}^{L} f(x)\,dx \Rightarrow a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\,dx, \\ \text{para } j &= 1, \quad a_1 L = \int_{-L}^{L} f(x)\cos\left(\frac{\pi x}{L}\right)dx \Rightarrow a_1 = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{\pi x}{L}\right)dx, \end{aligned}$$

así sucesivamente, para $j = m$ obtenemos $a_m = \dfrac{1}{L}\displaystyle\int_{-L}^{L} f(x)\cos\left(\frac{m\pi x}{L}\right)dx$. ■

Claramente, los $a_k \quad k = 0, 1, \ldots, m$ son los coeficientes de Fourier, por lo tanto Q_m es el polinomio trigonométrico de Fourier de cosenos.

Sea $f \in C([-L, L])$, m$\in \mathbb{Z}^+$ y consideramos el conjunto de funciones $M = \{\psi_k \mid k = 1, \ldots, m\}$ con $\psi_k(x) = \operatorname{sen}\left(\dfrac{k\pi x}{L}\right) \quad x \in [-L, L], \quad k = 1, \ldots, m$. Este conjunto M es ortogonal. Se definen

$$\begin{aligned} Q_m(x) &= \sum_{k=1}^{m} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \\ E(b_1, \ldots, b_m) &= \|f - Q_m\|^2 = \int_{-L}^{L} \left(f(x) - Q_m(x)\right)^2 dx = \int_{-L}^{L} \left[f(x) - \sum_{k=1}^{m} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)\right]^2 dx \end{aligned}$$

Procediendo en forma similar a la prueba del teorema precedente, se obtienen los coeficientes de Fourier $\widehat{b}_j \quad j = 1, \ldots, m$ definidos como $\widehat{b}_j = \dfrac{1}{L}\displaystyle\int_{-L}^{L} f(x)\operatorname{sen}\left(\frac{j\pi x}{L}\right)dx$, y

$$E\left(\widehat{b}_1, \ldots, \widehat{b}_m\right) = \min_{(b_1,\ldots,b_m)\in\mathbb{R}^m} E(b_1, \ldots, b_m).$$

Más aún, si combinamos los dos resultados precedentes con $f \in C([-L, L])$ y se definen

$$\begin{aligned} Q_m(x) &= \sum_{k=0}^{m} a_k \cos\left(\frac{k\pi x}{L}\right) + \sum_{k=1}^{m} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \\ E(a_0, \ldots, a_m, b_1, \ldots, b_m) &= \|f - Q_m\|^2 = \int_{-L}^{L} \left(f(x) - Q_m(x)\right)^2 dx \end{aligned}$$

resulta que los coeficientes de Fourier

$$a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\,dx,\ \widehat{a}_k = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{k\pi x}{L}\right)dx,\ \widehat{b}_k = \frac{1}{L}\int_{-L}^{L} f(x)\operatorname{sen}\left(\frac{k\pi x}{L}\right)dx \quad k = 1, 2, \ldots, m$$

minimizan el funcional $E(a_0, \ldots, a_m, b_1, \ldots, b_m)$.

Teorema 4 (Igualdad de Bessel) *Sean* $M = \{\varphi_k \mid k = 0, 1, \ldots, m\}$, *donde* $\varphi_k(x) = \cos\left(\dfrac{k\pi x}{L}\right)$ $x \in [-L, L], \quad k = 0, 1, \ldots, m$, *y* $f \in C([-L, L])$. *Entonces, para cada* $m \in \mathbb{Z}^+$ *se tiene*

$$\left\| f - \sum_{k=0}^{m} \frac{\langle f, \varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k \right\|^2 = \|f\|^2 - \sum_{k=0}^{m} \frac{|\langle f, \varphi_k\rangle|^2}{\|\varphi_k\|^2}.$$

Demostración. El conjunto M es ortogonal y de la definición de la norma asociada al producto escalar, se tiene

$$\left\| f - \sum_{k=0}^{m} \frac{\langle f, \varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k \right\|^2 = \left\langle f - \sum_{k=0}^{m} \frac{\langle f, \varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k, f - \sum_{k=0}^{m} \frac{\langle f, \varphi_j\rangle}{\left\|\varphi_j\right\|^2}\varphi_j \right\rangle$$

y de la linealidad del producto escalar respecto de cada variable (véase el apéndice, espacios con producto interior) resulta

$$\begin{aligned}\left\|f-\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2 &= \langle f,f\rangle-2\left\langle f,\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\rangle+\left\langle\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k,\sum_{j=1}^{m}\frac{\langle f,\varphi_j\rangle}{\|\varphi_j\|^2}\varphi_j\right\rangle\\ &= \|f\|^2-2\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\langle f,\varphi_k\rangle+\sum_{k=0}^{m}\sum_{j=1}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\frac{\langle f,\varphi_j\rangle}{\|\varphi_j\|^2}\langle\varphi_k,\varphi_j\rangle.\end{aligned}$$

Por hipótesis $\langle\varphi_k,\varphi_j\rangle=0$ si $k\neq j$ y $\langle\varphi_k,\varphi_k\rangle=\|\varphi_k\|^2$, entonces

$$\sum_{k=0}^{m}\sum_{j=1}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\frac{\langle f,\varphi_j\rangle}{\|\varphi_j\|^2}\langle\varphi_k,\varphi_j\rangle=\sum_{k=0}^{m}\left(\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\right)^2\|\varphi_k\|^2=\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}.$$

Por lo tanto,

$$\begin{aligned}\left\|f-\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2 &= \|f\|^2-2\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}+\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}\\ &= \|f\|^2-\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}.\end{aligned}$$

■

Observaciones

1. Puesto que $\left\|f-\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2\geq 0$ se sigue que $\|f\|^2-\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}\geq 0\Leftrightarrow\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}\leq\|f\|^2$ (desigualdad de Bessel).

   Por el teorema de Pitágoras, se tiene

$$\left\|\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2=\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^4}\|\varphi_k\|^2=\sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2}.$$

   Por lo tanto, la igualdad y desigualdad de Bessel se escriben como sigue:

$$\begin{aligned}\left\|f-\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2 &= \|f\|^2-\left\|\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2,\\ \sum_{k=0}^{m}\frac{|\langle f,\varphi_k\rangle|^2}{\|\varphi_k\|^2} &= \left\|\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k\right\|^2\leq\|f\|^2.\end{aligned}$$

   Note que

$$Q_m(x)=\sum_{k=0}^{m}\widehat{a}_k\cos\left(\frac{k\pi x}{L}\right)dx=\sum_{k=0}^{m}\frac{\langle f,\varphi_k\rangle}{\|\varphi_k\|^2}\varphi_k(x)\quad x\in[-L,L],$$

   el polinomio trigonométrico de Fourier de senos, con $\widehat{a}_k\quad k=1,\ldots,m$ los coeficientes de Fourier.

2. En forma similar a la precedente se obtiene con el conjunto ortogonal $N=\{\psi_k\mid k=1,\ldots,m\}$ con $\psi_k(x)=\operatorname{sen}\left(\frac{k\pi x}{L}\right)\quad x\in[-L,L],\quad k=1,\ldots,m$. Se tiene

$$\left\|f-\sum_{k=1}^{m}\frac{\langle f,\psi_k\rangle}{\|\psi_k\|^2}\psi_k\right\|^2=\|f\|^2-\sum_{k=1}^{m}\frac{|\langle f,\psi_k\rangle|^2}{\|\psi_k\|^2}.$$

   En este caso

$$\sum_{k=1}^{m}\frac{\langle f,\psi_k\rangle}{\|\psi_k\|^2}\psi_k=\sum_{k=1}^{m}\widehat{b}_k\operatorname{sen}\left(\frac{k\pi x}{L}\right)=Q_m(x),$$

   el polinomio trigonométrico de Fourier de senos, con $\widehat{b}_k\quad k=1,\ldots,m$ los coeficientes de Fourier.

9.8.2. Aproximación numérica

Para calcular valores aproximados de los coeficientes de Fourier, aplicamos el método de los trapecios.

Sea $n \in \mathbb{Z}^+$ y $\tau(n)$ una partición uniforme del intervalo $[0, L]$, esto es $h = \dfrac{L}{h}$ y $x_j = jh \quad j = 0, 1, \ldots, n.$

i) Supongamos que la función $f \in C([-L, L])$ es par, entonces el polinomio trigonométrico de Fourier Q_m está definido como

$$Q_m(x) = \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L],$$

donde

$$a_k = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{k\pi x}{L}\right)dx = \frac{2}{L}\int_{0}^{L} f(x)\cos\left(\frac{k\pi x}{L}\right)dx \quad k = 0, 1, \ldots, m,$$

que se aproxima con la fórmula de los trapecios. Tenemos

$$\begin{aligned} a_k &= \frac{2}{L}\int_{0}^{L} f(x)\cos\left(\frac{k\pi x}{L}\right)dx \\ &\simeq \frac{2}{L}\left[\frac{h}{2}\left(f(0)\cos(0) + f(L)\cos(k\pi)\right) + h\sum_{j=1}^{n-1} f(x_j)\cos\left(\frac{k\pi x_j}{L}\right)\right] \\ &= \frac{1}{n}\left(f(0) + f(L)\cos(k\pi)\right) + \frac{2}{n}\sum_{j=1}^{n-1} f(x_j)\cos\left(\frac{k\pi j}{n}\right). \end{aligned}$$

Ponemos $y_j = f(x_j) \quad j = 0, 1, \ldots, n$. Entonces

$$\widehat{a}_k = \frac{1}{n}\left(y_0 + y_n\cos(k\pi)\right) + \frac{2}{n}\sum_{j=1}^{n-1} y_j\cos\left(\frac{k\pi j}{L}\right) \quad k = 0, 1, \ldots, m,$$

y se define $\widehat{Q}_m(x)$ como sigue:

$$\widehat{Q}_m(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \widehat{a}_k\cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L],$$

$\widehat{Q}_m(x)$ es una aproximación de $Q_m(x)$ en $x \in [-L, L]$.

ii) Supongamos que la función $f \in C([-L, L])$ es impar, el polinomio trogonométrico de Fourier está definido como

$$Q_m(x) = \sum_{k=1}^{m} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L],$$

con

$$b_k = \frac{1}{L}\int_{-L}^{L} f(x)\operatorname{sen}\left(\frac{k\pi x}{L}\right)dx = \frac{2}{L}\int_{0}^{L} f(x)\operatorname{sen}\left(\frac{k\pi x}{L}\right)dx \quad k = 1, 2, \ldots, m,$$

los mismos que se aproximan con la fórmula de los trapecios.

Resulta

$$\begin{aligned} b_k &= \frac{2}{L}\int_{0}^{L} f(x)\operatorname{sen}\left(\frac{k\pi x}{L}\right)dx \\ &\simeq \frac{2}{L}\left[\frac{h}{2}\left(f(0)\operatorname{sen}(0) + f(L)\operatorname{sen}(k\pi)\right) + h\sum_{j=0}^{n-1} f(x_j)\operatorname{sen}\left(\frac{k\pi x_j}{L}\right)\right]. \end{aligned}$$

Puesto que $\operatorname{sen}(k\pi) = 0 \quad k = 1, 2, \ldots, m$, se sigue que

$$b_k \simeq \frac{2}{n}\sum_{j=1}^{n-1} f(x_j)\operatorname{sen}\left(\frac{k\pi x_j}{L}\right) = \frac{2}{n}\sum_{j=1}^{n-1} f(x_j)\operatorname{sen}\left(\frac{k\pi j}{n}\right).$$

Nuevamente, se define $y_j = f(x_j) \quad j = 1, \ldots, n-1$, $b_k = \frac{2}{n} \sum_{j=1}^{n-1} y_j \operatorname{sen}\left(\frac{k\pi j}{n}\right)$ y con estos coeficientes, el polinomio trigonométrico $\widehat{Q}_m(x) = \sum_{k=1}^{m} \widehat{b}_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L]$.

Entonces $\widehat{Q}_m(x)$ es una aproximación de $Q_m(x)$

iii) Sea $f \in C([-L, L])$. Se define las funciones $u, v \in C([-L, L])$ como sigue:

$$\begin{cases} u(x) = \frac{1}{2}(f(x) + f(-x)) \\ v(x) = \frac{1}{2}(f(x) - f(-x)) \end{cases} \qquad x \in [-L, L],$$

entonces u es par y v es impar. Se tiene $f = u + v$.

El polinomio trogonométrico de Fourier

$$\begin{aligned} Q_m(x) &= \frac{a_0}{2} + \sum_{k=1}^{m} \left(a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right)\right) \\ &= \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos\left(\frac{k\pi x}{L}\right) + \sum_{k=1}^{m} b_k \operatorname{sen}\left(\frac{k\pi x}{L}\right) \\ &= P_m(x) + R_m(x) \quad x \in [-L, L] \end{aligned}$$

donde $P_m(x) = \frac{1}{2}(Q_m(x) + Q_m(-x))$, $R_m(x) = \frac{1}{2}(Q_m(x) - Q_m(-x))$ con P_m par y R_m impar que se aproximan como en i) y ii) precedentes.

Para calcular $\widehat{Q}_m(x)$ se requiere de la siguiente información: número de términos del polinomio trigonométrico $Q_m(x)$, extremo derecho $L > 0$ del intervalo $[-L, L]$ par. Para calcular aproximaciones de los coeficientes mediante el método de los trapecios se requiere conocer el número de puntos n del intervalos $[0, L]$. Adicionalmente requerimos de una aproximación de π. Con esta información, proponemos el siguiente algoritmo de cálculo de $Q_m(x)$

Algoritmo

Datos de entrada: $m, n \in \mathbb{Z}^+$, L, $x \in \mathbb{R}$, función f.

Datos de salida: $\widehat{Q}_m(x)$.

1. $pi = 3{,}1415926536$.
2. $h = \frac{L}{n}$.
3. $y_0 = f(0)$.
4. $y_1 = f(L)$.
5. Para $k = 0, 1, \ldots, m$

   $S = 0$.

   Si $k = 0$ entonces

   para $j = 1, \ldots, n-1$

   $x_j = jh$

   $S = S + f(x_j)$

   fin bucle j

$a_0 = \frac{1}{n}(y_0 + y_1) + \frac{2}{n} S$

$S = 0$

Si $a < k \leq n$ entonces

para $j = 1, \ldots, n-1$

$$S = s + f(x_j) * \cos\left(\frac{k\pi j}{n}\right)$$

fin de bucle j

$$\widehat{a}_k = \frac{1}{n}\left(y_0 + y_1 \cos(k\pi)\right) + \frac{2}{n}S$$

Fin de bucle k

6. $S = 0$
7. Para $k = 1, \ldots, m$

$$S = S + \widehat{a}_k * \cos\left(\frac{k\pi x}{L}\right)$$

Fin de bucle k

8. $Q_m(x) = \frac{1}{2}\widehat{a}_0 + S$
9. Imprimir $x,\ Q_m(x)$
10. Fin

Para los otros casos se elaboran algoritmos muy similares, por lo que se propone como ejercicio

9.9. Ejercicios

1. Para la función f definida en $[0,1]$ que en cada item se propone, hallar el polinomio P que mejor se aproxima en mínimos cuadrados a la función f.

   a) $f(x) = e^x$, y $P(x) = a + bx \quad x \in [0,1]$ **b)** $f(x) = e^{-x}$, y $P(x) = a + bx + cx^2 \quad x \in [0,1]$

   c) $f(x) = \operatorname{sen} x$, y $P(x) = a + bx \quad x \in [0,1]$. **d)** $f(x) = \sqrt{x+1}$, y $P(x) = a + bx + cx^2 \quad x \in [0,1]$.

2. Supongamos que se dispone del conjunto de n pares de datos experimentales $S = \{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, ..., n\}$ que en cada item se indica. Encontrar un polinomio P de grado n que se indica y que mejor se ajusta al conjunto de datos.

   a) $S = \{(0,1), (1,-1), (2,1), (3,3)\}$ y $P(x) = a + bx \quad x \in \mathbb{R}$.

   b) $S = \{(0,0), (1,-1), (2,0), (3,3), (4,8)\}$ y $P(x) = a + bx + cx^2 \quad x \in \mathbb{R}$.

   c) $S = \{(0,0), (1,-1), (2,0), (3,3), (4,8)\}$ y $P(x) = a + bx + cx^2 \quad x \in \mathbb{R}$.

   d) $S = \{(0,1), (1,4), (2,9), (3,16), (4,25)\}$ y $P(x) = a + bx + cx^2 \quad x \in \mathbb{R}$.

3. Sea f una función real dependiente de un parámetro c. Escribiremos $t = f(x,c)$. Suponga que $\dfrac{\partial f}{\partial c}$ es continua.

   Se dispone de un conjunto de datos experimentales

   $$S = \left\{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots, n\right\}$$

   y se asume que cada $y_i = f(x_i, c) + r_i(c)$, donde $r_i(c)$ denota el error en la observación y_i, $i = 1, \ldots, n$.

En el método de mínimos cuadrados se considera el problema siguiente:

$$\min_{c\in\mathbb{R}} \sum_{i=1}^{n} r_i^2(c).$$

Se define $E(c) = \sum_{i=1}^{n} r_i^2(c) = \sum_{i=1}^{n} (y_i - f(x_i, c))^2$.

a) Elaborar un algoritmo para aproximar $\widehat{c} \in \mathbb{R}$ tal que

$$E(\widehat{c}) = \min_{c\in\mathbb{R}} E(c).$$

b) Se considera la siguiente información experimental:

$$S = \{(1, 1{,}35), (1{,}5, 0{,}498), (2, 0{,}183), (2{,}2, 0{,}123)\}$$

Aplique el método de mínimos cuadrados para calcular la constante $\widehat{c} > 0$ tal que $f(t) = 10e^{-\widehat{c}t}$.

c) Se considera el siguiente conjunto de datos

$$S = \{(0{,}26, 5), (0{,}785, 5), (0{,}5, 8{,}7), (1{,}05, 8{,}7)\}$$

Aplique el método de mínimos cuadrados para calcular la constante $\widehat{c}$ tal que $f(t) = 10\sin(ct)$.

4. Sea F una función real dependiente de dos parámetros a y b.

Escribiremos $y = f(x, a, b)$. Se supone que $\dfrac{\partial^2 f}{\partial a^2}, \dfrac{\partial^2 f}{\partial b^2}, \dfrac{\partial^2 f}{\partial a \partial b}$ son continuas.

Se dispone de un conjunto de datos experimentales

$$S = \left\{(x_i, y_i) \in \mathbb{R}^2 \mid i = 1, \ldots, n\right\}, \quad n \geq 3.$$

Y se asume que cada $y_i = f(x_i, a, b) + r_i(a, b)$, donde $r_i(a, b)$ denota el error en la observación $y_i, i = 1, \ldots, n$.

En el método de mínimos cuadrados se considera el problema

$$\min_{(c)a,b\in\mathbb{R}^2} \sum_{i=1}^{n} r_i^2(a, b). \tag{P}$$

A fin de calcular los parámetros a y b usando la información experimental, definimos

$$E(a, b) = \sum_{i=1}^{n} r_i^2(a, b) = \sum_{i=1}^{n} (y_i - f(x_i, a, b))^2.$$

a) Utilice el método de Newton y elabore un algoritmo para aproximar $\widehat{a}, \widehat{b} \in \mathbb{R}$ tales que

$$E\left(\widehat{a}, \widehat{b}\right) = \min_{(a,b)\in\mathbb{R}} E(a, b).$$

b) Considere la siguiente información experimental

$$S = \{(2, 20), (10, 20{,}2), (50, 21{,}03), (100, 22{,}1), (500, 33)\}$$

Aplique el método de mínimos cuadrados para calcular los parámetros $\widehat{a}, \widehat{b}$ tales que $f(t) = \widehat{a}e^{\widehat{b}t}$ $t \geq 0$.

c) Se dispone del conjunto de datos

$$S = \{(0{,}1, 50), (2, 3{,}85), (4, 1{,}02), (5, 0{,}66)\}$$

Aplique el método de mínimos cuadrados para calcular los parámetros $\widehat{a}, \widehat{b}$ tales que $f(x) = \frac{a}{1+bx^2}$ $x \geq 0$.

9.10. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Análisis Matemático, Segunda Edición, Editorial Reverté, Barcelona, 1982.
2. N. Bakhvalov, Metodos Numéricos, Editorial Paraninfo, Madrid, 1980.
3. Åke Björck, Numerical Methods for Least Squares Problems, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1996.
4. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.
5. John P. Boyd, Chebyshev and Fourier Spectral Methods, Second Edition (Revised), Editorial Dover Publications, Inc.,Mineola, 2001.
6. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
7. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
8. P. G. Ciarlet, Introduction á L' Analyse Numérique Matricielle et á L' Optimisation, Editorial Masson, París, 1990.
9. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
10. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.
11. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.
12. J. E. Dennis, Jr., Robert B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1996.
13. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volumi I, II Pitagora Editrice Bologna, 1983.
14. John E. Freund, Ronald E. Walpole, Estadística Matemática con Aplicaciones, Cuarta Edición, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1990.
15. Clude Gasquet, Patrick Witomski, Analyse de Fourier et Applications: Filtrage, Calcul Numérique et Ondeletles, Editorial-Dunod, París, 2000.
16. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.
17. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.
18. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.
19. Franz E. Hohn, Algebra de Matrices, Editorial Trillas, México, 1979.
20. Robert W. Hornbeck, Numerical Methods, Quantum Publishers, Inc., New York, 1975.
21. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.
22. Erwin Kreyszig, Introducción a la Estadística Matemática, Editorial Limusa, México, 1981.

23. Charles L. Lawson, Richard J. Hanson, Solving Least Squares Problems, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1995.

24. L. Lebart, A. Morineau, J.-P. Fénelon, Tratamiento Estadístico de Datos, Editorial Marcombo Boixareu Editores, Barcelona, 1985.

25. Thomas M. Little, F. Jackson Hills, Métodos Estdísticos para la Investigación en la Agricultura, Editorial Trillas, México, 2002.

26. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

27. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

28. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

29. Fazlollah Reza, Los Espacios Lineales en la Ingeniería, Editorial Reverté, S. A., Barcelona, 1977.

30. Francis Scheid, Theory and Problems of Numerical Analysis, Schaum's Outline Series, Editorial McGraw-Hill, New York, 1968.

31. M. Sibony, J. Cl. Mardon, Analyse Numérique I, Sustèmes Linéaires et non Linéaires, Editorial Hermann, París, 1984.

32. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

33. Gilbert Strang, Algebra Lineal y sus Aplicaciones, editorial Fondo Educativo Interamericano, México, 1982.

34. V. Voïévodine, Principes Numériques D' Algèbre Linéaire, Editions Mir, Moscú, 1976.

Capítulo 10

Splines

Resumen

La teoría de los splines tiene aplicaciones en dos direcciones importantes de la matemática: la una en los métodos de resolución de ecuaciones diferenciales ordinarias, particularmente los problemas de valor inicial y los problemas de valores en la frontera; las ecuaciones en derivadas parciales y ecuaciones integrales; la otra dirección lo constituye la computación gráfica, particularmente los modelos geométricos con splines, y el objetivo de este capítulo es dar una introducción a esta teoría. Se abordan dos clases de splines: los de interpolación y se da más énfasis a los splines cúbicos; los B-splines y particularmente los de interpolación cúbicos . En este capítulo se ha limitado los ejercicios. Al final del capítulo se incluye una amplia bibliografía.

10.1. Introducción

Una spline es una función definida a trozos sobre intervalos de $\mathbb{R}$ que se unen entre si obedeciendo a ciertas condiciones de regularidad. La terminología fue introducida por I. J. Schoenberg (1946).

El nombre de spline proviene del nombre del intrumento mecánico del mismo nombre que consiste en un alambre flexible que puede ser utilizado para dibujar curvas suaves a través de puntos asignados. Esta clase de instrumentos fueron utilizados para dibujo técnico en las industrias aeronáuticas, automotriz, naval, etc.

Como aplicaciones simples de splines podemos citar el método de Euler para construir una aproximación polinomial a trozos para la solución de problemas de valor inicial de las ecuaciones diferenciales ordinarias. Este tipo de aproximación es a menudo utilizada para establecer el teorema de Peano para la existencia de soluciones de tales problemas. Con este punto de vista podemos citar también los artículos de C. Runge (1901), W. Quade y L. Collatz (1938), J. Favord (1940), R. Curant (1943). Entre los textos sobre splines, publicados recientemente, podemos citar: C. de Boor (1978); A Practical Guide to splines; L. L. Schumaker (1981): Spline Functions: Basic Theory.

En la actualidad, las funciones splines se aplican fundamentalmente en grafismo en las industrias automotriz, aeronáutica, naval; en diseño y arquitetura; en métodos numéricos para la solución numérica de ecuaciones diferenciales ordinarias o en derivadas parciales con valores iniciales y/o valores ligados a los métodos de Rayleigh-Ritz-Galerkin y Petrov-Galerkin; y se cuentan miles de artículos de splines y de sus aplicaciones.

10.2. Espacio de funciones splines

Sean $n \in \mathbb{N}$. Se llama conjunto de nodos un conjunto de puntos $\tau(n) = \{x_j\}_{j=0,\ldots,n}$, donde $a = x_0 < x_1 < \cdots < x_n = b$. Estos nodos forman una partición del intervalo $[a, b] \subset \mathbb{R}$ en subintervalos $[x_{j-1}, x_j]$,

$j = 1, \ldots, n$. Los puntos $x_1, \ldots, x_{n-1}$ se llaman nodos interiores, y lospuntos $x_0 = a$ y $x_n = b$ se llaman nodos frontera.

Un conjunto de puntos $S_n = \{(x_i, y_i) \mid x_i \in \tau(n),\ y_i \in \mathbb{R},\ i = 0, 1, \ldots, n\}$, se llama conjunto de puntos de base.

Notamos con P_m el espacio de polinomios de grado $\leq m$. Se designa con $C_{-1}([a,b])$ el espacio de funciones continuas a trozos en $[a,b]$. Se denota con $C_{k-1}([a,b])$, $a \leq k \leq m$, el espacio de funciones que poseen derivadas continuas hasta el orden $k-1$ en $[a,b]$ ($C_0([a,b]) = C([a,b])$ es el espacio de las funciones en $[a,b]$).

Definición 1 *Sea $m \in \mathbb{N}$. Una función $S : [a,b] \to \mathbb{R}$ se llama función spline polinomial de grado m si ella posee las propiedades siguientes:*

i) $S \in C_{m-1}([a,b])$;

ii) $S \in P_m$ *para* $x \in [x_{j-1}, x_j[$, $\quad j = 1, \ldots, n$.

Denotamos con $S_m(\tau(n))$ el conjunto de todas las funciones splines polinomiales de grado m asociadas a la subdivisión $\tau(n)$ de $[a,b]$.

En lo sucesivo nos limitaremos a los splines polinomiales y nos referiremos a ellas simplemente como splines.

Ejemplos

1. En la figura siguiente se ilustra la gráfica de una función spline de grado 0.

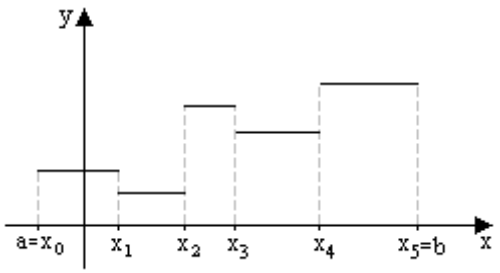


Figura 89

2. Sean S_n un conjunto de puntos de base. La línea poligonal que consiste en los segmentos de recta que une puntos sucesivos de S_n es un ejemplo de función spline de grado 1. En la figura que a continuación se indica se traza una spline de grado 1.

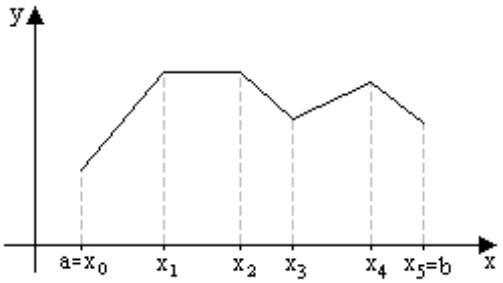


Figura 90

3. Sean $\tau(n)$ una subdivisión del intervalo $[a,b]$ y $m \in \mathbb{N}$. La familia de funciones $\{q_{m,j} \mid j = 0,\ldots,n-1\}$ definidas como sigue

$$q_{m,j}(x) = \begin{cases} (x-x_j)^m, & \text{si } x \in [x_j, b], \\ 0, & \text{si } x \in [a, x_j[, \end{cases}$$

son splines de grado m asociadas a la subdivisión $\tau(n)$. En la figura siguiente se ilustran estas funciones.

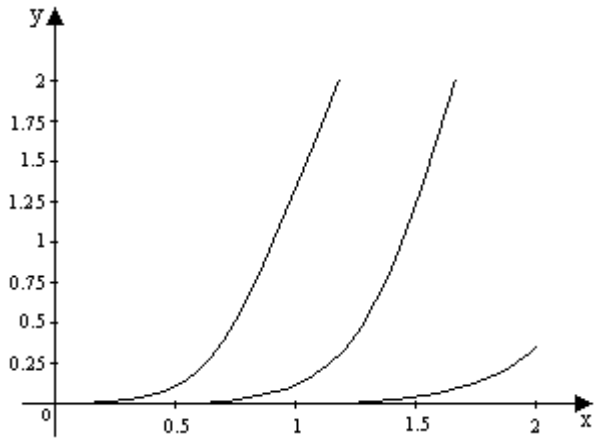


Figura 91

Las funciones $q_{m,j}, \quad j = 1,\ldots,n$, se llaman splines de un solo lado.

Base de $S_m(\tau(n))$

El conjunto $S_m(\tau(n))$ provisto de las operaciones habituales entre funciones (adición y producto de un número real por una función) es un espacio vectorial real de dimensión $m+n$ y una base de dicho espacio es el conjunto de funciones

$$\{p_0, p_1, \ldots, p_m, q_{m,1}, \ldots, q_{m,n-1}\},$$

donde

$$p_i(x) = x^i, \qquad i = 0,1,\ldots,m,$$

y las funciones $q_{m,j}$ están definidas en el ejercicio 3).

Se puede probar que toda función $S \in S_m(\tau(n))$ se escribe de manera única en la forma

$$S(x) = a_0 + \sum_{i=1}^{m} a_i x^j + \sum_{j=1}^{n-1} b_j q_{m,j}(x) \qquad x \in [a,b],$$

donde $a_0, a_i, b_j \in \mathbb{R}, \quad i = 1,\ldots,m; \;\; j = 2,\ldots,m.$

10.3. Interpolación mediante splines

Centraremos nuestra atención en la interpolación mediante splines de grado uno y tres que son las más utilizadas en las aplicaciones.

La ventaja del método de interpolación mediante splines es el uso de polinomios de grado bajo para producir globalmente interpolantes suaves, al tiempo que evita la desventaja del uso de polinomios de interpolación de grado alto.

Splines de interpolación de grado 1

Sea $f : [a,b] \longrightarrow \mathbb{R}$ una función continua definida en $[a,b]$ y $\tau(n)$ una subdivisión de $[a,b]$.

Sea $S_n = \{(x_i, f(x_i)) \mid x_i \in \tau(n),\ \ i = 0, 1, \ldots, n\}$ un conjunto de puntos de base. El ejemplo más simple de splines de grado 1 es el spline lineal que consiste en segmentos de recta que unen puntos sucesivos de S_n; de este modo se obtiene una línea poligonal.

La spline interpolante de grado 1 está definida de manera única en cada subintervalo $[x_{j-1}, x_j] \quad j = 1, \ldots, n$, como una función afín definida en dicho subintervalo y que globalmente es la única línea poligonal obtenida al juntar todos los segmentos de recta.

Buscamos una función real S definida en $[a, b]$ que verifique las propiedades siguientes:

i) $S \in C([a, b])$;

ii) $S \in P_1$ para $x \in [x_{j-1}, x_j[\,, \quad j = 1, \ldots, n$;

iii) $S(x_j) = f(x_j), \quad j = 0, 1, \ldots, n.$

Para construir una tal función S, determinemos αj, β_j constantes tales que

$$\begin{aligned} S(x) &= \alpha_j + \beta_j x, \quad x \in [x_{j-1}, x_j],\ \ j = j = 1, \ldots, n, \\ S(x_{j-1}) &= f(x_{j-1}), \\ S(x_j) &= f(x_j). \end{aligned}$$

Entonces

$$\begin{aligned} S(x_{j-1}) &= \alpha_j + \beta_j x_{j-1} = f(x_{j-1}), \\ S(x_j) &= \alpha_j + \beta_j x_j = f(x_j). \end{aligned}$$

Resolviendo el sistema de ecuaciones, obtenemos

$$\begin{aligned} \alpha_j &= f(x_{j-1}) - \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}} x_{j-1}, \\ \beta_j &= \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}}. \end{aligned}$$

Por lo tanto la función S se escribe

$$S(x) = f(x_{j-1}) + \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}}(x - x_{j-1}), \quad x \in [x_{j-1}, x_j] \quad j = 1, \ldots, n,$$

que es la ecuación de la recta que pasa por los puntos $(x_{j-1}, f(x_{j-1}))$ y $(x_j, f(x_j))$ restringida al intervalo $[x_{j-1}, x_j]$, la misma que se escribe

$$S(x) = \frac{x - x_j}{x_{j-1} - x_j} f(x_j) + \frac{x - x_{j-1}}{x_j - x_{j-1}} f(x_{j-1}), \quad j = 1, \ldots, n.$$

Estimación del error

Definimos

$$p_{j,1}(x) = \frac{x - x_j}{x_{j-1} - x_j}, \quad p_{j,2} = \frac{x - x_{j-1}}{x_j - x_{j-1}} \quad \forall x \in [x_{j-1}, x_j].$$

Se tiene entonces que

$$\begin{aligned} p_{j,1}(x) &\geq 0, \quad p_{j,1}(x) \geq 0, \quad \forall x \in [x_{j-1}, x_j], \\ p_{j,1}(x) + p_{j,2}(x) &= 1, \quad \forall x \in [a, b], \quad j = 1, \ldots, n, \end{aligned}$$

de donde $f(x) = p_{j,1}(x) f(x) + p_{j,2}(x) f(x)$ y

$$|f(x) - S(x)| \leq \text{Max}\{|f(x) - f(x_{j-1})|,\ |f(x) - f(x_j)|\}.$$

El módulo de continuidad de f relativo al intervalo $[x_{j-1}, x_j]$ se define por

$$w_f(\delta) = \operatorname*{Sup}_{\substack{|t_1-t_2|\leq\delta \\ t_1,t_2\in[x_{j-1},x_j]}} |f(t_1) - f(t_2)|,$$

donde $\delta > 0$.

El módulo de continuidad verifica las propiedades siguientes:

i) $w_f(\delta_1) \leq w_f(\delta_2)$ para $0 < \delta_1 \leq \delta_2$,

ii) $w_f(\delta) \underset{\delta\to 0}{\longrightarrow} 0$.

Utilizando el módulo de continuidad, tenemos

$$\operatorname*{Max}_{x\in[x_{j-1},x_j]} |f(x) - S(x)| \leq w_f(|x_j - x_{j-1}|), \quad j = 1, \ldots, n.$$

Sea $h = \operatorname*{máx}_{j=1,\ldots,n} |x_j - x_{j-1}|$. Se tiene entonces la siguiente estimación del error:

$$\|f - S\|_{L^\infty(a,b)} = \operatorname*{Max}_{x\in[a,b]} |f(x) - S(x)| \leq w_f(h).$$

De la definición de $w_f(h)$, se sigue que

$$\|f - S\|_{L^\infty(a,b)} \underset{h\to 0}{\longrightarrow} 0,$$

esto es, las splines de interpolación lineales convergen uniformemente a $f \in C([a,b])$ cuando $h \longrightarrow 0$.

Si $f \in C_1([a,b])$, se tiene la siguiente estimación de error:

$$\|f - S\|_{L^\infty(a,b)} \leq \frac{h}{4} w_{f'}(h).$$

Si $f \in C_2([a,b])$, entonces

$$\|f - S\|_{L^\infty(a,b)} \leq \frac{h^2}{8} \|f'\|_{L^\infty(a,b)}.$$

10.3.1. Splines cúbicas de interpolación

Consideremos $f \in C_2([a,b])$, $\tau(n)$ una subdivisión de $[a,b]$ y

$$S_n = \{(x_i, f(x_i)) \mid x_i \in \tau(n), \quad i = 0, \ldots, n\}$$

un conjunto de puntos de base.

Puesto que $\dim S_3(\tau(n)) = n + 3$, si se requiere interpolar en cada uno de los $n + 1$ nodos $x_0, \ldots, x_n$, entonces quedan 2 parámetros libres que pueden ser utilizados en los tipos de splines siguientes:

a) Interpolación con condiciones de frontera de Hermite

Hallar $S \in S_3(\tau(n))$ tal que

i) $S(x_j) = f(x_j), \quad j = 0, 1, \ldots, n,$

ii) $S'(a) = f'(a)$;

iii) $S'(b) = f'(b)$.

b) Intepolación con condiciones de frontera naturales

Suponemos que $n \geq 2$.

Hallar $S \in S_3(\tau(n))$ tal que

i) $S(x_j) = f(x_j), \quad j = 0, 1, \ldots, n;$
ii) $S''(a) = S''(b) = 0.$

c) Interpolación con condiciones de frontera periódicas ($f(a) = f(b)$ y $f'(a) = f'(b)$)

Hallar $S \in S_3(\tau(n))$ tal que

i) $S(x_j) = f(x_j), \quad j = 0, 1, \ldots, n;$
ii) $S'(a) = S'(b);$
iii) $S''(a) = S''(b).$

Con el propósito de mostrar que los problemas a), b) y c) tienen solución única, enunciamos la propiedad siguiente de las splines cúbicas, conocida como relación integral.

Relación integral

Sea $f \in C_2([a,b])$ y $S \in S_3(\tau(n))$ una función spline de interpolación de f tal que la diferencia $E(x) = f(x) - S(x) \;\; x \in [a,b]$, satisface la condición de frontera

$$S''(a)\,E'(a) = S''(b)\,E'(b).$$

Entonces

$$\int_a^b \left[f''(x)\right]^2 dx = \int_a^b \left[f''(x) - S''(x)\right]^2 dx + \int_a^b \left[S''(x)\right]^2 dx.$$

De esta relación, vemos que

1. Si $E'(a) = E'(b) = 0$, entonces se tiene las splines del tipo a).
2. Si $S''(a) = S''(b) = 0$, entonces se tiene las splines del tipo b).
3. Si $S''(a) = S''(b)$ y $E'(a) = E'(b)$, corresponden entonces a las splines del tipo c).

Usando la relación integral se prueba que los problemas de interpolación a), b) y c) tienen siempre una única solción $S \in S_3(\tau(n))$.

Construcción

Dada una función $f \in C_2([a,b])$, para construir la función spline de interpolación S, aplicamos las condiciones de las funciones splines y de las splines cúbicas al polinomio cúbico siguiente:

$$\begin{aligned} S_j(x) &= a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3, \quad x \in [x_j, x_{j+1}], \quad j = 0, 1, \ldots, n-1, \\ S(x) &= S_j(x), \quad x\,[\times_j, x_{j+1}], \quad j = 0, 1, \ldots, n-1. \end{aligned}$$

Por i) de los tipos de splines a), b) y c), se tiene

$$S_j(x_j) = a_j = f(x_j), \quad j = 0, 1, \ldots, n-1,$$

y ponemos $a_n = f(x_n)$.

De la definición de función spline (continuidad en cada nodo), se obtiene

$$\begin{aligned} a_{j+1} &= S_{j+1}(x_{j+1}) = S_j(x_{j+1}) \\ &= a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3, \end{aligned}$$

para $j = 0, 1, \ldots, n-1$.

Notamos con $h_j = x_{j+1} - x_j$, $j = 0, 1, \ldots, n-1$. Entonces la relación precedente se escribe

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3.$$

La derivada de $S_j(x)$ es la función

$$S_j'(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2, \quad x \in [x_j, x_{j+1}], \quad j = 0, 1, \ldots, n-1,$$

de donde

$$S_j'(x_j) = b_j, \quad j = 0, 1, \ldots, n-1.$$

Definimos $b_n = S'(x_n)$.

Por la cotinuidad de S_j' en cada nodo x_j, tenemos

$$b_{j+1} = S_{j+1}'(x_{j+1}) = S_j'(x_{j+1}) = b_j + 2c_j h_j + 3d_j h_j^2, \quad j = 0, 1, \ldots, n-1.$$

La derivada segunda de $S_j(x)$ está dada por

$$S_j''(x) = 2c_j + 6d_j(x - x_j), \quad x \in [x_j, x_{j+1}], \quad j = 0, 1, \ldots, n-1,$$

de donde

$$S_j''(x_j) = 2c_j, \quad j = 0, 1, \ldots, n-1,$$

y definimos $c_n = \dfrac{1}{2}S''(x_n)$,

Nuevamente, utilizando la continuidad de $S_j''(x)$ en cada nodo x_j, tenemos:

$$C_{j+1} = \frac{1}{2}S_{j+1}''(x_{j+1}) = \frac{1}{2}S_j''(x_{j+1}) = \frac{1}{2}(2c_j + 6d_j h_j) = c_j + 3d_j h_j, \quad j = 0, 1, \ldots, n.$$

Obtengamos relaciones que liguen los coeficientes b_j, c_j, d_j en términos de los datos $a_j = f(x_j)$, $j = 0, 1, \ldots, n$.

Resulta

$$d_j = \frac{c_{j+1} - c_j}{3h_j},$$

con lo cual

$$\begin{aligned} a_{j+1} &= a_j + b_j h_j + c_j h_j^2 + \frac{c_{j+1} - c_j}{3h_j} h_j^3 = a_j + b_j h_j + \frac{1}{3}(2c_j + c_{j+1}) h_j^2, \\ b_{j+1} &= b_j + 2c_j h_j + 3\frac{c_{j+1} - c_j}{3h_j} h_j^2 = b_j + (cj + c_{j+1}) h_j, \quad j = 0, 1, \ldots, n-1. \end{aligned}$$

Para obtener la relación final entre los coeficientes, de la igualdad

$$a_{j+1} = a_j + b_j h_j + \frac{1}{3}(2c_j + c_{j+1}) h_j^2,$$

obtenemos

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \quad j = 0, 1, \ldots, n-1,$$

y para $j = 0, 1, \ldots, n$:

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_j - 1) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j),$$

con lo cual la relación

$$b_{j+1} = b_j + (c_j + c_{j+1}) h_j$$

se expresa en la forma

$$b_j = b_{j-1} + (c_{j-1} + c_j) h_{j-1}, \quad j = 1, \ldots, n,$$

y

$$\frac{1}{h_j}\left(a_{j+1}-a_j\right)-\frac{h_j}{3}\left(2c_j+c_{j+1}\right)=\frac{1}{h_{j-1}}\left(a_j-a_{j-1}\right)-\frac{h_{j-1}}{3}\left(2c_{j-1}+c_j\right)+\left(c_{j-1}+c_j\right)h_{j-1},$$

de donde

$$\begin{aligned}\frac{1}{h_j}\left(a_{j+1}-a_j\right)-\frac{1}{h_{j-1}}\left(a_j-a_{j-1}\right) &= \frac{h_j}{3}\left(2c_j+c_{j+1}\right)-\frac{h_{j-1}}{3}\left(2c_{j-1}+c_j\right)+\left(c_{j-1}+c_j\right)h_{j-1}\\ &= \frac{1}{3}h_{j-1}c_{j-1}+\frac{2}{3}\left(h_{j-1}+h_j\right)c_j+\frac{1}{3}h_jc_{j+1},\end{aligned}$$

o bien

$$\left(h_{j-1},2\left(h_{j-1}+h_j\right),h_j\right)\begin{pmatrix}c_{j-1}\\c_j\\c_{j+1}\end{pmatrix}=\frac{3}{h_j}\left(a_{j+1}-a_j\right)-\frac{3}{h_{j-1}}\left(a_j-a_{j-1}\right),\quad j=1,\ldots,n-1.$$

Ponemos $\overrightarrow{c}^t=\left(c_0,c_1,\ldots,c_n\right)$. El sistema de ecuaciones precedente involucra únicamente el vector $\overrightarrow{c}$, las longitudes de los subintervalos $[x_{j-1},x_j]$, $j=1,\ldots,n$ y los valores de f en los puntos $\tau(n)=\{x_j\}_{j=1,\ldots,n}$ de la subdivisión de $[a,b]$.

10.3.2. Interpolación con condiciones de frontera de Hermite

Sea $f\in C^{(4)}([a,b])$, f tiene una única spline cúbica de interolaciíon $S\in S_3(\tau(n))$ que satisface las condiciones de frontera de hermite $S'(a)=f(a)$ y $S'(b)=f'(b)$.

En efecto,

$$S'(a)=S'(x_0)=b_0=f'(a),$$

y para $j=0$, b_0 está dado por

$$b_0=\frac{1}{h_0}\left(a_1-a_0\right)-\frac{h_0}{3}\left(2c_0+c_1\right),$$

resulta que

$$2h_0c_0+h_0c_1=-3f'(a)+\frac{3}{h_0}\left(a_1-a_0\right).$$

De manera similar, tenemos

$$S'(b)=S'(x_n)=b_n=f'(b).$$

Como

$$b_n=b_{n-1}+h_{n-1}\left(c_{n-1}+c_n\right),$$

y

$$b_{n-1}=\frac{1}{h_{n-1}}\left(a_n-a_{n-1}\right)-\frac{h_{n-1}}{3}\left(2c_{n-1}+c_n\right),$$

se tiene entonces que

$$\begin{aligned}f'(b) &= \frac{1}{h_{n-1}}\left(a_n-a_{n-1}\right)-\frac{h_{n-1}}{3}\left(2c_{n-1}+c_n\right)+h_{n-1}\left(c_{n-1}+c_n\right)\\ &= \frac{1}{h_{n-1}}\left(a_n-a_{n-1}\right)+\frac{h_{n-1}}{3}\left(c_{n-1}+2c_n\right),\end{aligned}$$

con lo cual

$$h_{n-1}c_{n-1}+2h_{n-1}c_n=3f'(b)-\frac{3}{h_{n-1}}\left(a_n-a_{n-1}\right).$$

En resumen, tenemos que

$$\begin{aligned}2h_0c_0+h_0c_1 &= -3f'(a)+\frac{3}{h_0}\left(a_1-a_0\right),\\ \left(h_{j-1},2\left(h_{j-1}+h_j\right),h_j\right)\begin{pmatrix}c_{j-1}\\c_j\\c_{j+1}\end{pmatrix} &= \frac{3}{h_j}\left(a_{j+1}-a_j\right)-\frac{3}{h_{j+1}}\left(a_j-a_{j-1}\right),\quad j=1,\ldots,n-1,\\ h_{n-1}c_{n-1}+2h_{n-1}c_n &= 3f'(b)-\frac{3}{h_{n-1}}\left(a_n-a_{n-1}\right),\end{aligned}$$

que puede expresarse en forma compacta como un sistema de ecuaciones

$$A\overrightarrow{C} = \overrightarrow{b},$$

donde

$$A = \begin{pmatrix} 2h_0 & h_0 & 0 & 0 & \dots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \dots & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & h_{n-1} \\ 0 & \dots & \dots & \dots & h_{n-1} & 2h_{n-1} \end{pmatrix}.$$

$$\overrightarrow{b} = \begin{pmatrix} -3f'(a) + \dfrac{3}{h_0}(a_1 - a_0) \\ \dfrac{3}{h_1}(a_2 - a_1) - \dfrac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \dfrac{3}{h_{n-1}}(a_n - a_{n-1}) - \dfrac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \dfrac{3}{h_{n-1}}(a_n - a_{n-1}) \end{pmatrix}$$

La matriz A es simétrica, estrictamente diagonal dominante, por lo tanto el sistema de ecuaciones precedente tiene solución única.

El método de resolución numérica que puede utilizarse es el de factorización LU de Crout o de Doolitle.

Una vez calculados los coeficientes $c_0, c_1, \dots, c_n$, los coeficientes b_j se calculan usando la relación

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \quad j = 1, \dots, n-1,$$

y los coeficientes d_j por

$$d_j = \frac{c_{j+1} + c_j}{3h_j}, \quad j = 0, 1, \dots, n-1.$$

Finalmente, se define $S(x) = S_j(x), \quad x \in [x_{j-1}, x_j], \quad j = 1, \dots, n.$

El error de interpolación en la norma $L^\infty(a,b)$ satisface la desigualdad siguiente;

$$\|f - S\|_{L^\infty(a,b)} \leq \frac{5}{384} M h^4,$$

donde $M = \left\|f^{(4)}\right\|_{L^\infty(a,b)}, h = \underset{j=0,1,\dots,n}{\text{Max}} h_j.$

Es claro que $\|f - S\|_{L^\infty(a,b)} \underset{h\to 0}{\longrightarrow} 0$; es decir que S converge uniformemente a f cuando $h \longrightarrow 0$.

10.3.3. Interpolación con condiciones de frontera naturales

Sea $f \in C^{(4)}([a,b])$, f tiene una única splice cúbica de interpolación $S \in S_3(\tau(n))$ que satisface las condiciones de frontera naturales $S''(a) = S''(b) = 0$. Efectivamente,

$$0 = S''9a = S''(x_0) = 2c_0 + 6d_0(x_0 - x_0),$$

de donde $c_0 = 0$,

$$c_n = \frac{S''(x_n)}{2} = \frac{S''(b)}{2} = 0.$$

Así, $c_0 = 0,\ c_n = 0$. Para $j = 1, \dots, n-1$, tenemos

$$(h_{j-1}, 2(h_{j-1} + h_j), h_j) \begin{pmatrix} c_{j-1} \\ c_j \\ c_{j+1} \end{pmatrix} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}),$$

que podemos escribir como un sistema de ecuaciones lineales

$$A\overrightarrow{C} = \overrightarrow{b},$$

donde

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \dots & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

$$\overrightarrow{b} = \begin{pmatrix} 0 \\ \frac{3}{h_1}(a_2-a_1) - \frac{3}{h_0}(a_1-a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n-a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1}-a_{n-2}) \\ 0 \end{pmatrix}$$

La matriz A es estrictamente diagonalmente dominante. Esto implica que el siste,a de ecuaciones precedente tiene solución única. El método numérico de resolución de tal sistema es el de factorización deCrout o de Doolitle.

Sea $\overrightarrow{C}^t = (c_0, c_1, \dots, c_n)$ la solución del sistema de ecuaciones $A\overrightarrow{C} = \overrightarrow{b}$. Los coeficientes b_j y d_j se calculan usando las fórmulas siguientes:

$$\begin{aligned} b_j &= \frac{a_{j+1}-a_j}{h_j} - \frac{h_j}{3}(c_{j+1}+2c_j) \quad j = 0, 1, \dots, n-1, \\ d_j &= \frac{c_{j+1}-c_j}{3h_j} \quad j = 0, 1, \dots, n-1. \end{aligned}$$

Note que $b_n = S'(b)$ y $d_n = 0$.

Definimos $S(x) = S_j(x) \quad x \in [x_{j-1}, x_j], \quad j = 1, \dots, n.$

Se tiene entonces la siguiente estimación de error

$$\|f - S\|_{L^\infty(a,b)} \leq C \left\| f^{(4)} \right\|_{L^\infty(a,b)} h^4,$$

donde $C > 0$ es una constante independiente de n y $h = \underset{j=1,\dots,n}{\text{Max}} h_j$.

10.3.4. Interpolación con condiciones de frontera periódicas

Sea $f \in C^4([a,b])$. f tiene una única spline cúbica de interpolación $S \in S_3(\tau(n))$ que satisface las condiciones de frontera $S'(a) = S'(b)$, $S''(a) = S''(b)$.

Mediante un razonamiento similar a los dos casos a) y b), se obtiene el sistema de ecuaciones lineales siguiente:

$$A\overrightarrow{C} = \overrightarrow{b},$$

donde

$$A = \begin{pmatrix} 2(h_0-h_{n-1}) & h_0 & 0 & 0 & \dots & -h_{n-1} & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \dots & 0 & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & h_0 & \dots & \dots & 0 & -h_{n-1} & 2(h_0-h_{n-1}) \end{pmatrix}$$

$$\overrightarrow{b} = \begin{pmatrix} \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_0}(a_1 - a_0) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_0}(a_1 - a_0) \end{pmatrix}$$

El valor de $S(x)$ para $x \in [a,b]$ se obtiene de manera análoga a los casos a) y b).

El error de interpolación es idéntico al caso b)

10.4. Splines cuadráticas

El espacio $S_2(\tau(n))$ de las splines cuadráticas correspondientes a la subdivisión $\tau(n) = \{x_j\}_{j=0,\ldots,n}$ tiene dimensión $n+2$. Si deseamos construir una función spline S de interpolación en cada nodo, nos queda entonces exactamente un parámetro llibre, y por lo tanto es imposible imponer condiciones de frontera simétricas como en el caso de splines de grado impar discutidas en la seccion precedente.

A continuación proponemos dos problemas de interpolación que conducen a definir de manera única splines cuadráticas y que tienen condiciones simétricas de frontera. Para lograr esto, introducimos las subdivisiones de $[a,b]$ siguientes:

$$\tau_1(n-1) = \{y_j\}_{j=0,1,\ldots,n-1}, \quad \tau(n) = \{x_j\}_{j=0,1,\ldots,n}$$

tales que

$$a = x_0 = y_0 < x_1 < y_1 < x_2 < \cdots < x_{n-1} < y_{n-1} = k_n = b.$$

Entonces el espacio $S_2(\tau_1(n-1))$ tiene dimensión $n+1$, mientras que $S_2(\tau(n))$ tiene dimensión $n+2$.

a) Hallar $S \in S_3(\tau_1(n-1))$ tal que

$$S(x_j) = f(x_j), \quad j = 0,1,\ldots,n,$$

donde f es una función dada definida en $[a,b]$.

b) Sea $f \in C_1([a,b])$. Hallar $S \in S_2(\tau(n))$ tal que

$$\begin{aligned} S(y_j) &= f(y_j), \quad j = 0,1,\ldots,n-1, \\ S'(y_0) &= f'(a), \\ S'(y_{n-1}) &= f'(b). \end{aligned}$$

Utilizando el teorema de Rolle se demuestra que los problemas a) y b) tienen solución única.

Interpolación cuadrática para el problema a)

Sea f una función definida en $[a,b]$. Consideremos las subdivisiones $\tau_1(n-1)$ y $\tau(n)$ de $[a,b]$ arriba definidas. Buscamos una función $S \in S_2(\tau_1(n-1))$ tal que

$$S(x_j) = f(x_j), \quad j = 0,1,\ldots,n.$$

De la definición de función spline, S es un polinomio de grado 2 en cada subintervalo $[y_{j-1}, y_j]$, $j = 1,\ldots,n-1$.

Definimos

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2, \quad j = 0,1,\ldots,n-1,$$

y determinemos las constantes a_j, b_j y c_j. Tenemos

$$S_j(x_j) = a_j = f(x_j), \quad j = 0, 1, \ldots, n.$$

De la continuidad de S en y_j, tenemos

$$S_{j+1}(y_{j+1}) = S_j(y_{j+1}), \quad j = 1, \ldots, n-2,$$

de donde

$$a_{j+1} + b_{j+1}(y_{j+1} - x_{j+1}) + c_{j+1}(y_{j+1} - x_{j+1})^2 = a_j + b_j(y_{j+1} - x_j) + c_j(y_{j+1} - x_j)^2.$$

La derivada de S_j es la función definida por

$$S_j'(x) = b_j + 2c_j(x - x_j), \quad j = 0, 1, \ldots, n-1.$$

La continuidad de S' en y_j nos permite obtener la relación siguiente:

$$b_{j+1} + 2c_{j+1}(y_{j+1} - x_{j+1}) = b_j + 2c_j(y_{j+1} - x_j), \quad j = 1, \ldots, n-2.$$

Para determinar los coeficientes b_j y c_j, imponemos la condición $S'(x_j) = 0$. Entonces

$$0 = S'(x_j) = S_j'(x_j) = b_j, \quad j = 0, 1, \ldots, n.$$

Luego

$$c_{j+1} = c_j \frac{y_{j+1} - x_j}{y_{j+1} - x_{j+1}}, \quad j = 1, \ldots, n-2.$$

Puesto que

$$a_{j+1} + c_{j+1}(y_{j+1} - x_{j+1})^2 = a_j + c_j(y_{j-1} - x_j)^2, \quad j = 1, \ldots, n-2,$$

pues $b_j = 0, \quad \forall j = 1, \ldots, n-2$. Resulta que

$$a_{j+1} + cj\frac{y_{j+1} - x_j}{y_{j+1} - x_{j+1}}(y_{j+1} - x_j)^2 = a_j + c_j(y_{j+1} - x_j)^2,$$

de donde

$$c_j = \frac{a_{j+1} - a_j}{(y_{j+1} - x_j)(x_{j+1} - x_j)} = \frac{f(x_{j+1}) - f(x_j)}{(y_{j+1} - x_j)(x_{j+1} - x_j)}.$$

Así,

$$S_j(x) = f(x_j) + \frac{f(x_{j+1}) - f(x_j)}{(y_{j+1} - x_j)(x_{j+1} - x_j)}(x - x_j)^2, \quad j = 0, 1, \ldots, n-1.$$

Definimos

$$S(x) = S_j(x), \quad x \in [y_{j-1}, y_j], \quad j = 1, \ldots, n-1.$$

Con frecuencia la subdivisión $\tau_1(n-1)$ se selecciona de la manera siguiente:

$$y_0 = a, \;\; y_{n-1} = b \;\; \text{y} \;\; y_j = x_j + t_j(x_{j+1} - x_j), \quad j = 1, \ldots, n-2,$$

donde $t_j \in]0, 1[$.

Si $t_j = \dfrac{1}{2}$, entonces $y_j = \dfrac{1}{2}(x_j + x_{j+1}), \quad j = 1, \ldots, n-2$ que corresponden a los puntos medios de los subintervalos $[x_j, x_{j+1}], \quad j = 1, \ldots, n-2$.

Interpolación cuadrática para el problema b)

Definimos

$$S_j(x) = a_j + b_j(x - y_j) + c_j(x - y_j)^2, \quad x \in [x_{j-1}, x_j], \quad j = 0, 1, \ldots, n-1.$$

Entonces

$$S_j(y_j) = a_j + f(y_j), \quad j = 0, 1, \ldots, n-1.$$

La continuidad de S en cada nodo x_j nos conduce a la siguiente relación:

$$a_{j+1}+b_{j+1}\left(x_{j+1}-y_j\right)+c_{j+1}\left(x_{j+1}-y_j\right)^2=a_j+b_j\left(x_{j+1}-y_j\right)+c_j\left(x_{j+1}-y_j\right)^2,\quad j=0,1,\ldots,n-2,$$

y la continuidad de S' en cada nodo x_j, nos da la igualdad siguiente:

$$b_{j+1}+2c_{j+1}\left(x_{j+1}-y_j\right)=b_j+2c_j\left(x_{j+1}-y_j\right),\quad j=0,1,\ldots,n-2.$$

Por otro lado, $S_0'(x)=b_0+2c_0(x-a)$, entonces

$$S_0'(a)=b_0=f'(a).$$

Además

$$S_{n-1}'(x)=b_{n-1}+sc_{n-1}\left(x-y_{n-1}\right),$$

con lo cual

$$S_{n-1}'(b)=b_{n-1}=f'(b).$$

Combinando las relaciones anteriores, obtenemos

$$\begin{aligned} c_j &= \frac{a_j-a_{j+1}}{\left(x_{j+1}-y_j\right)^2}=\frac{f(y_j)-f(y_{j+1})}{\left(x_{j+1}-y_j\right)^2},\quad j=0,1,\ldots,n-2,\\ b_{j+1} &= b_j+2\left(x_{j+1}-y_j\right)\left(c_j-c_{j+1}\right),\quad j=0,1,\ldots,n-3. \end{aligned}$$

Note que

$$\begin{aligned} S_0(x) &= f(a)+f'(a)(x-a)+\frac{f(a)-f(y_1)}{(x_1-a)^2}(x-a)^2,\\ S_{n-1}(x) &= f(b)+f'(b)(x-b)+\frac{f(y_{n-2})-f(b)}{(b-y_{n-2})^2}(x-b)^2. \end{aligned}$$

Finalmente

$$S_j(x)=f(y_j)+b_j(x-y_j)+c_j(x-y_j)^2,\quad x\in[x_{j-1},x_j],\quad j=1,\ldots,n-2.$$

10.5. B - Splines

En las secciones precedentes construimos los espacios de splines $S_m(\tau(n))$ para una subdivisión dada $\tau(n)=\{x_j\}_{j=0,\ldots,n}$. Estos espacios tienen dimensión $m+n$ y una base de $S_m(\tau(n))$ en la familia de funciones

$$\{p_0,p_1,\ldots,p_m,q_{m,1},\ldots,q_{m,n}\}.$$

En esta sección discutiremos bases alternativas para espacios de splines mejor adaptadas a los aspectos numéricos. Estas funciones fueron introducidas por Schoenberg y las denominó "Curvas básicas de Splines" que en la actualidad se conocen simplemente como B - Splines.

Notamos con $\tau_\infty=\{x_j\}_{j\in\mathbb{Z}}$ una subdivisión de $\mathbb{R}$ tal que $x_j \underset{j\to-\infty}{\longrightarrow} -\infty$, $x_j \underset{j\to+\infty}{\longrightarrow} +\infty$, y $x_j<x_{j+1}$, $\forall j\in\mathbb{Z}$.

Definición 2 *Sea τ_∞ una subdivisión de $\mathbb{R}$. Se nota con $B_{m,j}$ la función de $\mathbb{R}$ en $\mathbb{R}$ tal que*

i) $B_{m,j}(x)=0$ *si* $x\in\mathbb{R}-[x_j,x_{j+m+1}[$, $j\in\mathbb{Z}$;

ii) $B_{m,j}\in P_m$ *sobre cada subintervalo* $[x_i,x_{j+1}]$, $i=j,\ldots,j+m+1$;

iii) $\displaystyle\int_{-\infty}^{+\infty}B_{m,j}(x)\,dx=\int_{x_j}^{x_{j+m+1}}B_{m,j}(x)\,dx=1$

Las funciones $B_{m,j}$ se llaman B - Splines. La condición iii) se conoce con el nombre de condición de normalización. Se puede probar que existe una única función $B_{m,j}$ que verifica i), ii) y iii).

Las funciones $\{B_{m,j} \mid j \in \mathbb{Z}\}$ forman una base del espacio de splines $S_m(\tau_\infty)$.

Ejemplos

1. B - spline de grado 0.

   Se nota con $B_{0,j}$ a las B-splines de grado cero. En la figura siguiente se muestra la gráfica de $B_{0,j}$.

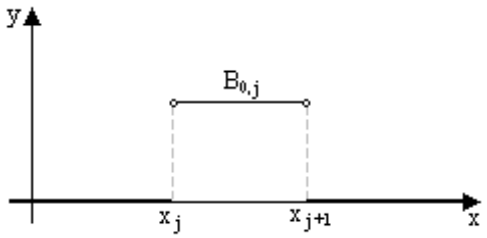


   Figura 92

   De la definición de B-splines, se tiene que

$$B_{0,j}(x) = \begin{cases} 0, & \text{si } x \in \mathbb{R} - [x_i, x_{j+1}[\,, \\ \dfrac{1}{x_{j+1} - x_j}, & \text{si } x \in [x_i, x_{j+1}[\,. \end{cases}$$

   Note que $B_{0,j} \in P_0$ sobre $[x_i, x_{j+1}[$, e $\displaystyle\int_{-\infty}^{+\infty} B_{0,j}(x)\,dx = 1$.

   Se tiene las siguientes propiedades:

   i) $Sop(B_{0,j}) = [x_i, x_{j+1}]$, $\quad \forall j \in \mathbb{Z}$.

   ii) $B_{0,j}(x) \geq 0$, $\quad \forall x \in \mathbb{R}$, $\quad \forall j \in \mathbb{Z}$.

   iii) $B_{0,j}$ es continua por la derecha en todo $\mathbb{R}$.

   iv) $\displaystyle\sum_{j=-\infty}^{+\infty} B_{0,j}(x) = 1$, $\quad \forall x \in \mathbb{R}$.

   v) La familia $\{B_{0,j} \mid j \in \mathbb{Z}\}$ es una base de $S_0(\tau_\infty)$, pues si $S \in S_0(\tau_\infty)$, entonces $S(x) = c_j$ si $x \in [x_j, x_{j+1}[$, $\quad j \in \mathbb{Z}$. Resulta que $S(x) = \displaystyle\sum_{j=-\infty}^{+\infty} B_{0,j}(x)$.

2. B-spline de grado 1.

   Notamos con $B_{1,j}$ a las B-splines de grado uno y que se definen como sigue:

$$B_{1,j}(x) = \begin{cases} 0, & \text{si } x \in \mathbb{R} - [x_j, x_{j+2}]\,, \\ \dfrac{x - x_j}{x_{j+1} - x_j}, & \text{si } x \in [x_j, x_{j+1}[\,, \\ \dfrac{x_{j+2} - x}{x_{j+2} - x_{j+1}}, & \text{si } x \in [x_{j+1}, x_{j+2}[\,, \end{cases} \qquad j \in \mathbb{Z}.$$

En la figura siguiente se ilustra la gráfica de esta función.

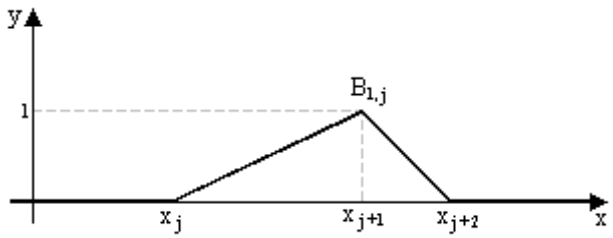


Figura 93

A estas funciones se les denomina también funciones techo. Se tienen las propiedades siguientes:

i) $Sop\left(B_{1,j}\right)=\left[x_j, x_{j+2}\right], \quad \forall j \in \mathbb{Z}.$

ii) $B_{1,j}(x) \geq 0, \quad \forall x \in \mathbb{R}, \quad \forall j \in \mathbb{Z}.$

iii) $B_{1,j} \in C_1(\mathbb{R})$.

iv) $\sum_{j=-\infty}^{+\infty} B_{1,j}(x)=1, \quad \forall x \in \mathbb{R}.$

v) $\sum_{j=-\infty}^{+\infty} c_j B_{1,j}=\sum_{j=-\infty}^{+\infty}\left[c_j v_{1,j}+c_{j-1}\left(1-v_{1,j}\right)\right] B_{0,j}$, donde $\left\{v_{1,j} \mid j \in \mathbb{Z}\right\}$ es la familia de funciones definidas por

$$v_{1,j}(x)=\frac{x-x_j}{x_{j+1}-x_j}, \quad j \in \mathbb{Z}.$$

Note que $B_{1,j}=v_{1,j} B_{0,j}+\left(1-v_{1,j+1}\right) B_{0,j+1}$.

vi) Para cada $S \in S_1(\tau(n))$ en el intervalo $\left[x_0, x_n\right]$, se tiene una única representación en términos de B-splines:

$$S(x)=\sum_{i=-1}^{n-1} \alpha_1 B_{1,i}(x), \quad \alpha_i \in \mathbb{R}.$$

B - splines cuadráticos.

3. Sean $h>0$, $x_0 \in \mathbb{R}$ y $x_j=x_0+jh$, $j \in \mathbb{Z}$.

Consideremos ahora splines en puntos igualmente espaciados $\tau_\infty=\left\{x_j\right\}_{j \in \mathbb{Z}}$.

Una B-spline cuadrática de grado 2 con respecto de τ_∞ se nota con $B_{2,j}$ y se define por

$$B_{2,j}(x)=\begin{cases} \dfrac{1}{2h^2}\left(x-x_j\right)^2, & \text{si } \ x \in\left[x_j, x_{j+1}\right[, \\ \dfrac{1}{2h^2}\left[h^2+2h\left(x-x_{j+1}\right)-2\left(x-x_{j+1}\right)^2\right], & \text{si } \ x \in\left[x_{j+1}, x_{j+2}\right[, \\ \dfrac{1}{2h^2}\left(x_{j+1}-x\right)^2, & \text{si } \ x \in\left[x_{j+2}, x_{j+3}\right[, \\ 0, & \text{si } \ x \in \mathbb{R}-\left[x_j, x_{j+3}\right], \end{cases} \quad j \in \mathbb{Z}.$$

4. **B - splines cúbicas**

Tal como en el caso de B-splines cuadráticas, consideramos $\tau_\infty=\left\{x_j\right\}_{j \in \mathbb{Z}}$ una subdivisión en puntos igualmente espaciados.

Una B-spline cúbica de grado 3 se define por

$$B_{3,j}(x)=\frac{1}{6h^3}\begin{cases} \left(x-x_j\right)^3, & \text{si } \ x \in\left[x_j, x_{j+1}\right[, \\ h^3+3h^2\left(x-x_{j+1}\right)+3h\left(x-x_{j+1}\right)^2-3\left(x-x_{j+1}\right)^3, & \text{si } \ x \in\left[x_{j+1}, x_{j+2}\right[, \\ h^3+3h^2\left(x_{j+3}-x\right)+3h\left(x_{j+3}-x\right)^2-3\left(x_{j+3}-x\right)^2, & \text{si } \ x \in\left[x_{j+2}, x_{j+3}\right[, \\ \left(x_{j+4}\right)^3, & \text{si } \ x \in\left[x_{j+3}, x_{j+4}\right[, \\ 0, & \text{si } \ x \in \mathbb{R}-\left[x_j, x_{j+4}\right], \end{cases}$$

Las funciones B-splines descritas en los ejemplos 1) a 4) son ampliamente utilizadas en el método de elementos finitos para la resolución numérica de problemas de valores de frontera y/o de condiciones iniciales.

10.5.1. Interpolaciones mediante B-splines cúbicas

Sea $f : [a, b] \to \mathbb{R}$ una función definida en $[a, b]$. Buscamos una función $S \in S_3(\tau(n))$ tal que

$$S(x_j) = f(x_j), \quad j = 0, 1, \ldots, n,$$

donde $\tau(n)$ es una subdivisión en puntos igualmente espaciados.

Para lograrlo, necesitamos los valores de los B-splines $B_{3,j}, \quad j = -3, \ldots, n-1$ en los nodos $x_0 = a, x_1, \ldots, x_n = b$, así como los valores de las derivadas $B'_{3,j}$ o $B''_{3,j}$ en $x_0 = a$ para $j = -3, -2, -1$ y en $x_n = b$ para $j = n-3, n-2, n-1$.

En la tabla siguiente se ilustran estos valores:

	x_j	x_{j+1}	x_{j+2}	x_{j+3}	x_{j+4}
$B_{3,j}(x)$	0	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	0
$B'_{3,j}(x)$	0	$\frac{1}{2h}$	0	$-\frac{1}{2h}$	0
$B''_{3,j}(x)$	0	$\frac{1}{h^2}$	$-\frac{2}{h^2}$	$\frac{1}{h^2}$	0

Sea $S \in S_3(\tau(n))$. Supongamos que

$$S(x) = \sum_{j=-3}^{n-1} \alpha_j B_{3,j}(x), \quad x \in [a, b].$$

Los problemas a), b) y c) discutidos en la sección de interpolación mediante splines cúbicas se escriben como sigue:

$$\sum_{j=-3}^{n-1} \alpha_J B_{3,j}(x_k) = f(x_k), \quad k = 0, 1, \ldots, n.$$

Condiciones de frontera:

a) $\sum_{j=k-3}^{k-1} \alpha_j B'_{3,j}(x_k) = f'(x_k), \quad k = 0, n;$

b) $\sum_{j=-3}^{-1} \alpha_j B''_{3,j}(x_k) = 0, \quad k = 0, n;$

c) $\sum_{j=3}^{-1} \alpha_j B'_{3,j}(a) = \sum_{j=n-3}^{n-1} \alpha_j B'_{3,j}(b);$

$\sum_{j=-3}^{-1} \alpha_j B''_{3,j}(a) = \sum_{j=n-3}^{n-1} \alpha_j B''_{3,j}(b).$

El sistema de ecuaciones resultante

$$A\overrightarrow{C} = \overrightarrow{b},$$

para los tres problemas, tiene las formas siguientes:

a) **Condiciones de frontera de Hermite**

$$A = \frac{1}{6}\begin{pmatrix} -\frac{3}{h} & 0 & \frac{3}{4} & 0 & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & 1 & 4 & 1 \\ 0 & \cdots & \cdots & -\frac{3}{h} & 0 & \frac{3}{h} \end{pmatrix},$$

$\overrightarrow{b}^t = (f'(a), f(x_0), \ldots, f(x_n), f'(b))$.

b) **Splines naturales**

$$A = \frac{1}{6}\begin{pmatrix} \frac{6}{h^2} & -\frac{12}{h^2} & \frac{6}{h^2} & 0 & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & 1 & 4 & 1 \\ 0 & \cdots & \cdots & \frac{6}{h^2} & -\frac{12}{h^2} & \frac{6}{h^2} \end{pmatrix},$$

$\overrightarrow{b}^t = (0, f(x_0), \ldots, f(x_n), 0)$.

c) **Splines periódicas**

$$A = \frac{1}{6}\begin{pmatrix} -\frac{3}{h} & 0 & \frac{3}{h} & 0 & \cdots & 0 & \frac{3}{h} & 0 & -\frac{3}{h} \\ \frac{6}{h^2} & -\frac{12}{h^2} & \frac{6}{h^2} & 0 & \cdots & 0 & -\frac{6}{h^2} & \frac{12}{h^2} & -\frac{6}{h^2} \\ 1 & 4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 1 & 4 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 4 & 1 \end{pmatrix},$$

$\overrightarrow{b}^t = (0, 0, f(x_0), \ldots, f(x_{n-1}), f(a))$.

10.6. Ejercicios

1. Construir la gráfica de $B_{2,j}$ y construir sus propiedades.

2. Proponemos como ejercicios construir la gráfica de B_3, j así como enunciar sus propiedades.

10.7. Lecturas complementarias y bibliografía

1. Richard H. Bartels, John C. Beatty, Brian A. Barsky, An Introduction to Splines for use in Computer Graphics and Geometric Medeling, Editorial Morgan Kaufmann Publishers, Inc., San Mateo, California, 1987.

2. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.

3. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.

4. Elaine Cohen, Richard F. Riesenfeld, Gershon Elber, Geometric Modeling with Splines, Editorial A. K. Peters, Natick, Massachusetts, 2001.

5. Gerald Farin, Curves and Surfaces for CAGD, Fifth Edition, Editorial Morgan Kaufmann Publishing, San Francisco, 2002.

6. James D. Foley, Andries van Dam, Steven K. Feiner, John F. Hughes, Computer Graphics, Principles and Practice, Second Edition in C, Editorial Addison-Ewsley, Boston , 1997.

7. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

8. Günther Hämmerlin, Karl-Heinz Hoffmann, Numerical Mathematics, Editorial Springer-Verlag, New York, 1991.

9. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

10. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

11. H. Prautzsch, W. Boehm, M Paluszny, Bézier and B-Spline Techniques, Editorial Springer-Verlag, Berlín, 2000.

12. Helmuth Späth, One Dimensional Spline Interpolation algorithms, Editorial A. K. Peters, Wellesley, Massachusetts, 1995.

13. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

14. Grace Wahba, Spline Models for Observational Data, Editorial Society for Industrial and Applied Mathematics (SIAM); Philadelphia, 1990.

Capítulo 11

Métodos numéricos de resolución de ecuaciones diferenciales ordinarias

Resumen

En este capítulo se tratan discretizaciones de dos tipos de problemas de ecuaciones diferenciales ordinarias: el problema de valor inicial de Cauchy, y, los problemas de valores en la frontera 1d. Se tratan los métodos clásicos de discretizaciones de problemas de valor inicial de Cauchy como son: la familia de los métodos de Runge-Kutta y el método θ que conduce a los métodos de Euler explícito e implícito y al método implícito de Crank-Nicolson. Por otro lado, se presenta el método de discretización del tipo Petrov-Galerkin que conduce a métodos implícitos del tipo Crank-Nicolson. La discretización de los problemas de valores en la frontera 1d se limita a las ecuaciones diferenciales de segundo orden y se aplica el método de diferencias finitas. Se estudia la consistencia, la estabilidad y la convergencia del método. Se concluye con la discretización en mallas no uniformes de ecuaciones diferenciales de segundo orden con valores en la frontera. Al final del capítulo se incluye una amplia bibliografía.

11.1. Introducción

En ingeniería y las ciencias físicas y químicas, las ciencias biológicas, la economía y sociales surgen modelos gobernados por ecuaciones diferenciales ordinarias, es decir, ecuaciones de la forma:

$$\begin{aligned} u'(t) &= f(t,u(t)) \quad t \in]0,T[, \\ u(0) &= u_0, \end{aligned}$$

donde $T > 0$, f es una función real definida en $[0,T] \times \mathbb{R}$, que en lo sucesivo supondremos al menos continua, $u_0 \in \mathbb{R}$ se denomina condición inicial y u es una función real definida en el intervalo $[0,T]$, u es la función incógnita. El problema de hallar una función u solución de la ecuación diferencial y que satisfaga la condición inicial, se conoce con el nombre de problema de Cauchy de valor inicial. Con más generalidad, se considera el problema de Cauchy de valor inicial siguiente:

$$\begin{aligned} \overrightarrow{u}'(t) &= \overrightarrow{f}(t,\overrightarrow{u}(t)) \quad t \in]0,T[, \\ \overrightarrow{u}(0) &= \overrightarrow{u}_0, \end{aligned}$$

donde $\overrightarrow{f}$ es una función $[0,T] \times \mathbb{R}^n$ en $\mathbb{R}^n$, $\overrightarrow{u}_0 \in \mathbb{R}^n$.

Cuando $\overrightarrow{f}(t,\overrightarrow{u}(t)) = A\overrightarrow{u}(t) \quad t \in [0,T]$, se tiene un sistema lineal de ecuaciones diferenciales ordinarias de primer orden, donde $A = (a_{ij})$ es una matriz real de $n \times n$. En forma explícita se escribe como sigue:

$$\begin{cases} u_1'(t) = a_{11}u_1(t) + \cdots + a_{1n}u_n(t) & \\ \qquad\qquad \vdots & t \in]0,T[, \\ u_n'(t) = a_{n1}u_1(t) + \cdots + a_{nn}u_n(t) & \end{cases}$$

que se le conoce como sistema de ecuaciones diferencial lineal autónomo.

Una ecuación diferencial lineal homogénea de orden n con coeficientes constantes es una ecuación de la forma

$$u^{(n)}(t) + a_1 u^{(n-1)}(t) + \cdots + a_n u(t) = 0 \quad t \in [0,T],$$

donde $a_1, \cdots, a_n \in \mathbb{R}$. Este tipo de ecuaciones diferenciales se transforma en un sistema de ecuaciones diferenciales lineales como el precedente mediante la siguiente transformación: $\overrightarrow{u}(t) = \begin{bmatrix} u(t) \\ u'(t) \\ \vdots \\ u^{(n-1)}(t) \end{bmatrix}$, resulta

$$\begin{aligned} \overrightarrow{u}'(t) &= \begin{bmatrix} u'(t) \\ u''(t) \\ \vdots \\ u^{(n)}(t) \end{bmatrix} = \begin{bmatrix} u'(t) \\ u''(t) \\ \vdots \\ -a_1 u^{(n-1)}(t) - \cdots - a_n u(t) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & \cdots & 0 \\ & \vdots & & \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} u(t) \\ u'(t) \\ \vdots \\ u^{(n-1)}(t) \end{bmatrix}. \end{aligned}$$

Ponemos $\overrightarrow{u}(t) = \begin{bmatrix} u(t) \\ u'(t) \\ \vdots \\ u^{(n-1)}(t) \end{bmatrix} \in \mathbb{R}^n$, $A = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ & \vdots & & \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{bmatrix} \in M_{n\times n}[\mathbb{R}]$, resulta , $\overrightarrow{u}'(t)) = A\overrightarrow{u}(t) \quad t \in [0,T]$. Con mayor generalidad, una ecuación diferencial lineal no homogénea de orden n con coeficientes constantes es una ecuación de la forma

$$u^{(n)}(t) + a_1 u^{(n-1)}(t) + \cdots + a_n u(t) = f(t) \quad t \in [0,T],$$

donde $a_1, \cdots, a_n \in \mathbb{R}$, f es una función real definida en $[0,T]$ y que se supondrá allí continua . Este tipo de ecuaciones diferenciales se transforma en un sistema de ecuaciones diferenciales lineales como el precedente si se define $\overrightarrow{u}(t) = \begin{bmatrix} u(t) \\ u'(t) \\ \vdots \\ u^{(n-1)}(t) \end{bmatrix}$, entonces

$$\begin{aligned} \overrightarrow{u}'(t) &= \begin{bmatrix} u'(t) \\ u''(t) \\ \vdots \\ u^{(n)}(t) \end{bmatrix} = \begin{bmatrix} u'(t) \\ u''(t) \\ \vdots \\ f(t) - a_1 u^{(n-1)}(t) - \cdots - a_n u(t) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & \cdots & 0 \\ & \vdots & & \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} u(t) \\ u'(t) \\ \vdots \\ u^{(n-1)}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ f(t) \end{bmatrix}. \end{aligned}$$

Se pone $\overrightarrow{u}(t) = \begin{bmatrix} u(t) \\ u'(t) \\ \vdots \\ u^{(n-1)}(t) \end{bmatrix} \in \mathbb{R}^n$, $\overrightarrow{b}(t) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ f(t) \end{bmatrix} \in \mathbb{R}^n$, $t \in [0,T]$ y A la matriz arriba definida.
Resulta $\overrightarrow{u}'(t) = A\overrightarrow{u}(t) + \overrightarrow{b}(t) \quad t \in [0,T]$. En este caso, $\overrightarrow{f}(t, \overrightarrow{u}(t)) = A\overrightarrow{u}(t) + \overrightarrow{b}(t)$.

Si $\overrightarrow{f}(t,\overrightarrow{u}(t))= A(t)\overrightarrow{u}(t)$, donde para cada $t \in [0,T]$, $A(t) = (a_{ij}(t))$ es una matriz real de $n \times n$ dependiente de t, el sistema de ecuaciones diferenciales $\overrightarrow{u}'(t) = A(t)\overrightarrow{u}(t)$ se conoce como sistema lineal no autónomo.

Si $\overrightarrow{u}'(t) = \overrightarrow{f}(\overrightarrow{u}(t))$ con $\overrightarrow{f}$ una función de $\mathbb{R}^n$ en $\mathbb{R}^n$ que no se se expresa como $A\overrightarrow{u}$, se llama sistema de ecuaciones diferenciales no lineal autónomo.

En lo que sigue, supondremos que $\overrightarrow{f}$ es una función $[0,T]\times\mathbb{R}^n$ en $\mathbb{R}^n$ lipchisiana respecto de la segunda variable, es decir, existe $k > 0$ tal que $\parallel \overrightarrow{f}(t,\overrightarrow{u}_1) - \overrightarrow{f}(t,\overrightarrow{u}_2) \parallel \leq k \parallel \overrightarrow{u}_1 - \overrightarrow{u}_2 \parallel \quad \forall \overrightarrow{u}_1, \overrightarrow{u}_2 \in \mathbb{R}^n$, $\forall t \in [0,T]$. Esta hipótesis garantiza la existencia de una sola solución $\overrightarrow{u} \in C^1([0,T])^n$.

11.2. El método θ

Sean $T > 0$, $u_0 \in \mathbb{R}$ y f una función real definida en $[0,T]\times\mathbb{R}$ que suponemos lipschisiana; esto es, existe $L > 0$ tal que para todo $y_1, y_2 \in \mathbb{R}$, se verifica

$$|f(t,y_1) - f(t,y_2)| \leq L|y_1 - y_2| \quad \forall t \in [0,T].$$

Consideramos el problema de valor inicial de Cauchy:

$$\begin{cases} u'(t) = f(t,u(t)) & t \in]0,T[, \\ u(0) = u_0. \end{cases}$$

De la hipótesis sobre f, esta ecuación tiene una única solución $u \in C^1([0,T])$.

En muy pocos casos se puede resolver esta ecuación directamente y obtener la solución exacta u. En la generalidad de los casos, la solución u no puede obtenerse directamente y debe recurrirse a los métodos numéricos, esto significa que podemos calcular soluciones aproximadas de u.

Sean $n \in \mathbb{Z}^+$, $\tau(n) = \{t_j \mid l = 0,1,\ldots,n\}$ una partición del intervalo $[0,T]$, esto es, $t_0 = 0$, $t_{j-1} < t_j \quad j = 1,\ldots,n$, $t_n = T$. Ponemos $h_j = t_j - t_{j-1} \quad j = 1,\ldots,n$ y $\widehat{h} = \max\limits_{j=1,\ldots,n} h_j$. En el caso de la partición uniforme, tenemos $h = \dfrac{T}{n}$, $t_j = jh \quad j = 0,1,\ldots,n$ con lo que $\tau(n) = \{jh \mid j = 0,1,\ldots,n\}$ y $\widehat{h} = h$.

Denotamos con u_j una aproximación de $u(t_j)$. El valor de $f(t_j,u(t_j))$ se aproxima como $f(t_j,u_j)$.

La derivada $u'(t_j)$ lo aproximamos mediante una diferencia finita progresiva de primer orden, esto es,

$$u'(t_j) \simeq \frac{u(t_{j+1}) - u(t_j)}{h_{j+1}} \quad j = 0,1,\ldots,n-1,$$

entonces $u'(t_j)$ se aproxima como

$$u'(t_j) \simeq \frac{u_{j+1} - u_j}{j_{j+1}} \quad j = 0,1,\ldots,n-1.$$

Sea $\theta \in [0,1]$. El método θ consiste en discretizar la ecuación diferencial mediante el esquema numérico siguiente:

$$\frac{u_{j+1} - u_j}{h_{j+1}} = \theta f(t_j,u_j) + (1-\theta) f(t_{j+1},u_{j+1}) \quad j = 0,1,\ldots,n-1,$$

cuyos datos son los pasos temporales h_j, $j = 1,\ldots,n$, los tiempos $t_j \quad j = 0,1,\ldots,n$. De este esquema numérico, se tiene interés en tres métodos numéricos conocidos como Euler explícito, Euler implícito y Crank-Nicolson.

1. Método de Euler explícito

Para $\theta = 1$ se obtiene el siguiente esquema numérico

$$\frac{u_{j+1} - u_j}{h_{j+1}} = f(t_j, u_j) \quad j = 0, 1, \ldots, n-1$$

y de este resultado

$$u_{j+1} = u_j + h_{j+1} f(t_j, u_j) \quad j = 0, 1, \ldots, n-1,$$

que se conoce como esquema numérico de Euler explícito.

La razón de ser método explícito se explica a continuación. Para $j = 0$, el esquema numérico precedente se expresa como

$$u_1 = u_0 + h_1 f(t_0, u_0).$$

Note que a partir de los datos conocidos: condición inicial $u(0) = u_0$, paso temporal h_1, tiempo t_0, podemos calcular directamente u_1 al instante t_1. Con u_1 calculado y los datos: paso temporal h_2, tiempo t_1 pasamos a calcular u_2 al instante t_2 mediante el esquema numérico que se obtiene haciendo $j = 1$, esto es,

$$u_2 = u_1 + h_2 f(t_1, u_1).$$

Así sucesivamente.

En el caso particular de una partición uniforme, el método de Euler explícito se escribe como

$$u_{j+1} = u_j + h\, f(t_j, u_j) \quad j = 0, 1, \ldots, n-1.$$

Más adelante estudiamos la convergencia del método de Euler explícito.

2. Método de Euler implícito

En el método θ hacemos $\theta = 0$, obtenemos

$$\frac{u_{j+1} - u_j}{h_{j+1}} = f(t_{j+1}, u_{j+1}) \quad j = 0, 1, \ldots, n-1,$$

y de esta igualdad resulta

$$u_{j+1} - h_{j+1} f(t_{j+1}, u_{j+1}) = u_j \quad j = 0, 1, \ldots, n-1,$$

que se conoce como esquema numérico de Euler implícito.

Para $j = 0$ se tiene la siguiente ecuación

$$u_1 - h_1 f(t_1, u_1) = u_0$$

cuya incógnita es u_1 al instante t_1. Solo en muy pocos casos se puede resolver esta ecuación directamente y calcularse u_1. En la generalidad de los casos, u_1 se calcula en forma aproximada como solución de dicha ecuación. Con u_1 calculado al instante t_1 se pasa inmediatamente a calcular u_2 como solución aproximada de la ecuación que se obtiene con $j = 1$, así

$$u_2 - h_2 f(t_2, u_2) = u_1.$$

El proceso continua hasta calcular $u_3, \ldots, u_{n-1}$ en los instantes $t_3, \ldots, t_n = T$.

Para una partición uniforme, el esquema numérico de Euler implícito se escribe como

$$u_{j+1} - h f(t_{j+1}, u_{j+1}) = u_j \quad j = 0, 1, \ldots, n-1.$$

Para calcular $u_j \quad j = 1, \ldots, n$, definimos la función real G como

$$G(x) = x - h_{j+1} f(t_{j+1}, x) - u_j \quad x \in \mathbb{R}.$$

Como $u_{j+1} - h_{j+1} f(t_{j+1}, u_{j+1}) - u_j = 0$, se sigue que $G(u_{j+1}) = 0$, es decir que $\widehat{x} = u_{j+1}$ es raíz de la ecuación $G(x) = 0$. Esta raíz $\widehat{x}$ es aproximada aplicando cualquiera de los métodos numéricos de resolución de ecuaciones no lineales que por supuesto dependen de la regularidad de la función f.

El análisis de la convergencia del método de Euler implícito lo haremos más adelante.

3. Método de Crank-Nicolson

Este método fue propuesto por J. Crank y P. Nicolson en 1947. Para $\theta = \frac{1}{2}$ se tiene el siguiente esquema numérico

$$\frac{u_{j+1} - u_j}{h_{j+1}} = \frac{1}{2} f(t_j, u_j) + \frac{1}{2} f(t_{j+1}, u_{j+1}) \quad j = 0, 1, \ldots, n-1$$

y de esta igualdad obtenemos

$$u_{j+1} - \frac{h_{j+1}}{2} f(t_{j+1}, u_{j+1}) = u_j + \frac{h_{j+1}}{2} f(t_j, u_j) \quad j = 0, 1, \ldots, n-1$$

que se conoce como esquema numérico de Crank-Nicolson. Este es un esquema numérico implícito.

Al igual que en el método de Euler implícito, para $j = 0$ y con los datos: paso temporal h_1, tiempo t_0, condición inicial $u(0) = u_0$ se tiene la ecuación

$$u_1 - \frac{h_1}{2} f(t_1, u_1) = u_0 + \frac{h_1}{2} f(t_0, u_0)$$

cuya incógnita es u_1 la misma que se aproxima como solución numérica de dicha ecuación. Calculado u_1 al instante t_1, con datos el paso temporal h_2, los tiempos t_1 y t_2, podemos calcular el valor aproximado de u_2 como solución numérica de la ecuación

$$u_2 - \frac{h_2}{2} f(t_2, u_2) = u_1 - \frac{h_1}{2} f(t_1, u_1).$$

Así sucesivamente.

En el caso de una partición uniforme, el esquema numérico de Crank-Nicolson se escribe como

$$u_{j+1} - \frac{h}{2} f(t_{j+1}, u_{j+1}) = u_j + \frac{h}{2} f(t_j, u_j) \quad j = 0, 1, \ldots, n-1.$$

De manera similar que el método de Euler implícito, definimos la función real G como

$$G(x) = x - \frac{h}{2} f(t_{j+1}, x) - u_j - \frac{h}{2} f(t_j, u_j) \quad x \in \mathbb{R}.$$

Resulta que $\widehat{x} = u_{j+1}$ es raíz de la ecuación $G(x) = 0$. El método de resolución numérica de la ecuación no lineal $G(x) = 0$ está relacionado con la regularidad de la función f.

La convergencia de este método será tratado más adelante.

Los esquemas numéricos de Euler implícito, explícito, y de Crank - Nicolson obtenidos para una sola ecuación diferencial pueden extenderse inmediatamente a los sistemas de ecuaciones diferenciales:

$$\begin{cases} \overrightarrow{u}'(t) = \overrightarrow{F}(t, \overrightarrow{u}(t)) & t \in]0, T[, \\ \overrightarrow{u}(0) = \overrightarrow{u}_0, \end{cases}$$

donde $T > 0$, $\overrightarrow{u}_0^T = \left(u_1^{(0)}, \ldots, u_m^{(0)}\right) \in \mathbb{R}^m$, $\overrightarrow{F}^T = (f_1, \ldots, f_m)$ una función vectorial de $[0, T] \times \mathbb{R}^m$ en $\mathbb{R}^m$ que suponemos lipschisiana. Este sistema de ecuaciones diferenciales se expresa como

$$\begin{cases} u_1'(t) = f_1(t, \overrightarrow{u}(t)) \\ \qquad \vdots \qquad\qquad\qquad t \in]0, T[. \\ u_m'(t) = f_m(t, \overrightarrow{u}(t)), \end{cases}$$

Sea $\tau(n) = \{t_j \mid j = 0, 1, \ldots, n\}$ una partición del intervalo $[0, T]$. Denotamos con $\overrightarrow{u}^T = \left(u_1^{(j)}, \ldots, u_m^{(j)}\right)$ una aproximación de $\overrightarrow{u}^T(t_j) \quad j = 1, \ldots, n$. Los esquemas numéricos anteriores se expresan como sigue.

1. Euler explícito

$$\overrightarrow{u}_{j+1} = \overrightarrow{u}_j + h_{j+1}\overrightarrow{F}\left(t_j, \overrightarrow{u}_j\right) \quad j = 0, 1, \ldots, n-1.$$

2. Euler implícito

$$\overrightarrow{u}_{j+1} - h_{j+1}\overrightarrow{F}\left(t_{j+1}, \overrightarrow{u}_{j+1}\right) = \overrightarrow{u}_j \quad j = 0, 1, \ldots, n-1.$$

3. Crank - Nicolson

$$\overrightarrow{u}_{j+1} - \frac{h_{j+1}}{2}\overrightarrow{F}\left(t_{j+1}, \overrightarrow{u}_{j+1}\right) = \overrightarrow{u}_j + \frac{h_{j+1}}{2}\overrightarrow{F}\left(t_j, \overrightarrow{u}_j\right) \quad j = 0, 1, \ldots, n-1.$$

Más particularmente en el caso de las funciones vectoriales del tipo

$$\overrightarrow{F}\left(t, \overrightarrow{y}\right) = A\left(t\right)\overrightarrow{y} + \overrightarrow{g}\left(t\right) \quad t \in [0, T],$$

donde para cada $t \in [0, T]$, $A\left(t\right) = \left(a_{ij}\left(t\right)\right)$ es una matriz de $m \times m$ no nula, $\overrightarrow{g}^T\left(t\right) = \left(g_1\left(t\right), \ldots, g_m\left(t\right)\right)$ con $g_j \in C\left([0, T]\right) \quad j = 1, \ldots, m$.

El método de Euler explícito se escribe como sigue:

$$\begin{aligned} \overrightarrow{u}_{j+1} &= \overrightarrow{u}_j + h_{j+1}\left[A\left(t_j\right)\overrightarrow{u}_j + \overrightarrow{g}\left(t_j\right)\right] \\ &= \left[I + h_{j+1}A\left(t_j\right)\right]\overrightarrow{u}_j + h_{j+1}\overrightarrow{g}\left(t_j\right) \quad j = 0, 1, \ldots, n-1, \end{aligned}$$

donde I denota la matriz identidad de $m \times m$.

El método de Euler implícito se escribe como

$$\overrightarrow{u}_{j+1} - h_{j+1}\left[A\left(t_{j+1}\right)\overrightarrow{u}_{j+1} + \overrightarrow{g}\left(t_{j+1}\right)\right] = \overrightarrow{u}_j \quad j = 0, 1, \ldots, n-1,$$

que a su vez podemos expresarlo como el siguiente

$$\left(I - h_{j+1}A\left(t_{j+1}\right)\right)\overrightarrow{u}_{j+1} = \overrightarrow{u}_j + h_{j+1}\overrightarrow{g}\left(t_{j+1}\right) \quad j = 0, 1, \ldots, n-1.$$

El método de Crank - Nicolson se expresa de la manera siguiente:

$$\overrightarrow{u}_{j+1} - \frac{h_{j+1}}{2}\left[A\left(t_{j+1}\right)\overrightarrow{u}_{j+1} + \overrightarrow{g}\left(t_{j+1}\right)\right] = \overrightarrow{u}_j + \frac{h_{j+1}}{2}\left(A\left(t_j\right)\overrightarrow{u}_j + \overrightarrow{g}\left(t_j\right)\right) \quad j = 0, 1, \ldots, n-1,$$

Luego

$$\left(I - \frac{h_{j+1}}{2}A\left(t_{j+1}\right)\right)\overrightarrow{u}_{j+1} = \left(I + \frac{h_j}{2}A\left(t_j\right)\right)\overrightarrow{u}_j + \frac{1}{2}h_{j+1}\left(\overrightarrow{g}\left(t_j\right) + \overrightarrow{g}\left(t_{j+1}\right)\right) \quad j = 0, 1, \ldots, n-1.$$

Se observa que tanto en el método de Euler implícito como en el de Crank-Nicolson, que para calcular $\overrightarrow{u}_{j+1}$ se requieren que las matrices $I - h_{j+1}A\left(t_{j+1}\right)$ e $I - \frac{h_{j+1}}{2}A\left(t_{j+1}\right)$ sean invertibles.

11.3. Método de Petrov-Galerkin.

Sean $T > 0$, $u_0 \in \mathbb{R}$, f una función real definida en $[0, T] \times \mathbb{R}$ que suponemos lipschisiana. Consideramos el problema de valor inicial de Cauchy siguiente: hallar una función u definida en $[0, T]$ solución de

$$\begin{cases} u'\left(t\right) = f\left(t, u\left(t\right)\right) & t \in]0, T[, \\ u\left(0\right) = u_0. \end{cases}$$

Antes de describir el método de Petrov-Galerkin para resolver el problema de Cauchy precedente, requerimos introducir algunas notaciones y algunos espacios de funciones.

Recordemos que una función s se dice escalonada en $[0, T]$ si y solo si existe una partición $\tau\left(n\right) = \{t_j \mid j = 0, 1, \ldots, n\}$ del intervalo $[0, T]$ tal que $s\left(t\right) = s_i \quad t \in]t_{j-1}, t_j[\quad j = 1, \ldots, n$, donde $s_i \in \mathbb{R}$. En

los puntos $t_j \quad j = 0, 1, \ldots, n$ la función s debe estar definida de cualquier modo. En la figura siguiente se muestra una función escalonada en $[0, T]$.

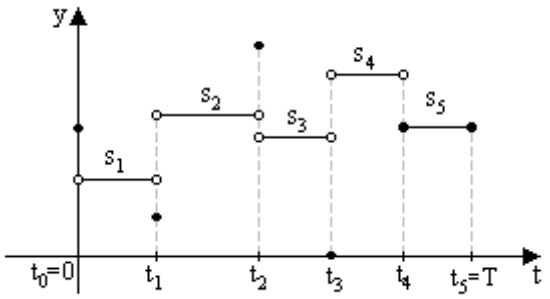


Figura 94

Definición 1 *Diremos que una función real f definida en $[0, T]$ es discontinua en $a \in [0, T]$ con salto de primera especie si y solo si f es discontinua en a, y $|f(a^+) - f(a^-)| < \infty$, donde $f(a^+) = \lim\limits_{\substack{h>0 \\ h\to 0}} f(a+h)$, $f(a^-) = \lim\limits_{\substack{h<0 \\ h\to 0}} f(a+h)$, son los límites por derecha e izquierda respectivamente.*

Si $a = 0$ se tiene únicamente el límite por la derecha finito, esto es $|f(0^+)| < \infty$ y si $a = T$, se tiene el límite por la izquierda $|f(T^-)| < \infty$.

Definición 2 *Sea f una función real definida en $[0, T]$. Decimos que f es acotada y continua a trozos en $[0, T]$ si y solo si existen $a_k \in [0, T] \quad k = 1, \ldots, m$ tales que $0 \leq a_1 < a_2 < \cdots < a_m \leq T$, f no es continua en a_k, y $\left|f\left(a_k^+\right) - f\left(a_k^-\right)\right| < \infty$.*

Se tiene que f es continua en el conjunto $[0, T] \setminus \{a_k \mid k = 1, \ldots, m\}$.

En el caso en que $\tau(m) = \{t_j \mid j = 0, 1, \ldots, m\}$ es una partición de $[0, T]$ y f es una función acotada y continua a trozos, entonces f es continua en cada subintervalo $]t_{j-1}, t_j[$ y $\left|f\left(t_j^+\right) - f\left(t_j^-\right)\right| < \infty$ $j = 1, \ldots, m$. Una función escalonada en $[0, T]$ es una función continua a trozos.

Es claro que f es continua en $[0, T]$ si y solo si para cada $t \in [0, T]$, $f\left(t_j^+\right) = f\left(t_j^-\right) = f(t)$.

Denotamos con $C_d([0, T])$ el conjunto de funciones reales acotadas y continuas a trozos. Con las operaciones habituales de funciones: adición " + " y producto de números reales por funciones " $\bullet$ ", $C_d([0, T])$ es un espacio vectorial real. Se tiene que $C([0, T]) \subset C_d([0, T])$.

En el espacio $C_d([0, T])$ se define el producto escalar siguiente:

$$\langle u, v \rangle = \int_0^T u(t)\, v(t)\, dt \quad \forall u, v \in C_d([0, T]).$$

Se denota con $C_d^1([0, T])$ el espacio de funciones reales u cuya derivada u' pertenece a $C_d([0, T])$, esto es,

$$C_d^1([0, T]) = \left\{u \in C_d([0, T]) \mid u' \in C_d([0, T])\right\}.$$

En la figura siguiente se muestra una función $u \in C_d^1([0,T])$.

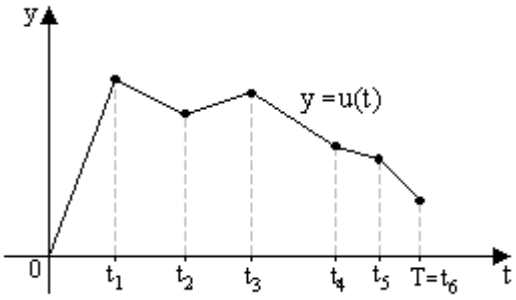


Figura 95

En la siguiente figura se muestra la función $u'(t)$

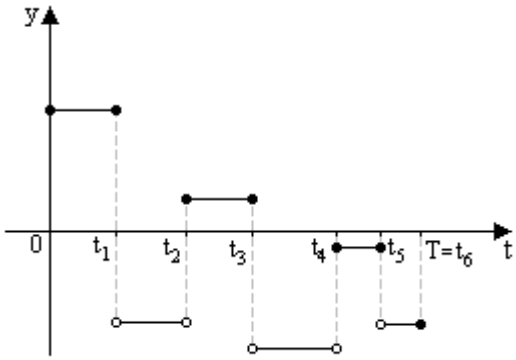


Figura 96

Introducimos el subespacio $C_*^1([0,T])$ de $C_d^1([0,T])$ siguiente:

$$C_*^1([0,T]) = \left\{ v \in C_d^1([0,T]) \mid v(T) = 0 \right\}.$$

Pasemos a describir el método de Petrov-Galerkin. Sea $u \in C^1([0,T])$ la solución del problema de Cauchy arriba propuesto.

Sea $v \in C_*^1([0,T])$. Multiplicamos a la ecuación diferencial por v e integramos sobre el intervalo $[0,T]$, esto es,

$$\int_0^T u'(t)\, v(t)\, dt = \int_0^T f(t, u(t))\, v(t)\, dt.$$

Apliquemos el método de integración por partes al primer miembro de la igualdad precedente, tenemos

$$\int_0^T u'(t)\, v(t)\, dt = u(t)\, v(t)\Big|_0^T - \int_0^T u(t)\, v'(t)\, dt = u(T)\, v(T) - u(0)\, v(0) - \int_0^T u(t)\, v'(t)\, dt$$

Puesto que $v \in C_*^1([0,T])$ entonces $v(T) = 0$ y como u es solución del problema de Cauchy de valor inicial, se tiene $u(0) = u_0$. La igualdad precedente se reduce a la siguiente

$$\int_0^T u'(t)\, v(t) = -u_0 v(0) - \int_0^T u(t)\, v'(t)\, dt$$

y en consecuencia

$$-u_0 v(0) - \int_0^T u(t)\, v'(t)\, dt = \int_0^T f(t, u(t))\, v(t)\, dt$$

o lo que es lo mismo

$$-\int_0^T u(t)\,v'(t)\,dt = u_0 v(0) + \int_0^T f(t,u(t))\,v(t)\,dt.$$

Así, el problema de Cauchy de valor inicial propuesto es equivalente a la ecuación precedente, la misma que es una ecuación integral. Esta ecuación integral es la que da lugar al método de Petrov-Galerkin.

Note que en esta ecuación el orden de derivación ha disminuido en 1 y la función incógnita u figura bajo el signo de integración. Ahora buscamos una función u que sea continua en $[0,T]$ y que verifique la ecuación integral, lo que significa que se ha bajado la regularidad de la función u (antes se buscaba u de modo que la sea derivable). Esto permite amplificar el campo de acción de la solución de ecuaciones diferenciales en las que la función f cumpla condiciones de regularidad más débiles. Es precisamente esta situación la de mayor interés.

Definición 3 *Diremos que $u \in C_d^1([0,T])$ es solución de la ecuación integral si y solo si satisface la ecuación*

$$-\int_0^T u(t)\,v'(t)\,dt = u_0 v(0) + \int_0^T f(t,u(t))\,v\,|t|\,dt \quad \forall v \in C_*^1([0,T]).$$

La formulación dada en la definición precedente es la conocida como método del tipo Petrov-Galerkin. note que la función incógnita u pertenece al espacio $C_d([0,T])$ mientras que las denominadas funciones de prueba o funciones test pertenecen al espacio $C_*^1([0,T])$.

Pasemos a la discretización de la formulación del método del tipo Petrov-Galerkin arriba enunciado.

Sea $n \in \mathbb{Z}^+$, $\tau(n) = \{t_j \mid j = 0,1,\ldots,n\}$ una partición del intervalo $[0,T]$. Ponemos $h_j = t_j - h_{j-1}$ $j = 1,\ldots,n$, $\widehat{h} = \max\limits_{j=1,\ldots,n} h_j$. En el caso de una partición uniforme del intervalo $[0,T]$, ponemos $h = \dfrac{T}{n}$, $\tau(n) = \{jh \mid j = 0,1,\ldots,n\}$ y $\widehat{h} = h$.

Definimos el subespacio U_h de $C_d([0,T])$ como sigue:

$$U_h = \left\{ u_h \in C_d([0,T]) \mid u_{h/_{]t_{j-1},t_j[}} = cte \quad j = 1,\ldots,n \right\}$$

donde $u_{h/_{]t_{j-1},t_j[}} = cte$ denota la restricción de la función u_h al intervalo abierto $]t_{j-1},t_j[$ y que en dicho intervalo la función es constante.

Una base del espacio U_h es la familia de funciones $\{\chi_1,\ldots,\chi_n\}$ definidas como se indica:

$$\chi_j(t) = \begin{cases} 1, \text{ si } t \in]t_{j-1},t_j[, \\ 0, \text{ si } t \in [0,T] \setminus]t_{j-1},t_j[, \end{cases} \quad j = 1,\ldots,n.$$

Se advierte inmediatamente que cada función χ_j es la función indicatriz del intervalo $]t_{j-1},t_j[$. En la figura siguiente se muestra esta función

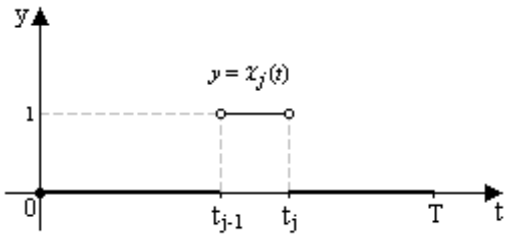


Figura 97

De la definición de espacio U_h y de la base $\{\chi_1,\ldots,\chi_n\}$ de U_h se tiene que $u_h \in U_h$ si y solo si existe $u_1,\ldots,u_n \in \mathbb{R}$ tales que $u_h = \sum\limits_{j=1}^{n} u_j \chi_j$.

Introducimos el subespacio V_h de $C_*^1([0,T])$:

$$V_h = \left\{ v_h \in C_*^1([0,T]) \mid v_{h/}{}_{]t_{j-1},t_j[} \in \mathcal{P}_1, \quad j = 1,\ldots,n \right\},$$

donde $\mathcal{P}_1$ denota el espacio de polinomios reales de grado ≤ 1, y $v_{h/}{}_{]t_{j-1},t_j[} \in \mathcal{P}_1$ designa la restricción de la función v_h al intervalo cerrado $[t_{j-1},t_j]$ en el que v_h es un polinomio de grado 1, o dicho de otro modo, v_h restringido al intervalo $[t_{j-1},t_j]$ es una función afín de la forma $v_h(t) = a_j + b_j t \quad t \in [t_{j-1},t_j]$ y $a_j, b_j \in \mathbb{R}$ son constantes escogidas apropiadamente, $j = 1,\ldots,n$. Una base del espacio V_h es la familia de funciones $\{\varphi_0,\ldots,\varphi_{n-1}\}$ definidas como sigue:

$$\varphi_0(t) = \begin{cases} -\dfrac{t-t_i}{h_i}, \text{ si } t \in [0,t_i], \\ 0, \text{ si } t \in [0,T] \setminus [0,t_1], \end{cases}$$

y para $j = 1,\ldots,n-1$ se tiene

$$\varphi_j(t) = \begin{cases} \dfrac{t-t_{j-1}}{h_j}, \text{ si } t \in [t_{j-1},t_j], \\ -\dfrac{t-t_{j+1}}{h_{j+1}}, \text{ si } t \in]t_j,t_{j+1}], \\ 0, \text{ si } t \in [0,T] \setminus [t_{j-1},t_{j+1}]. \end{cases}$$

En la figura de la izquierda se muestra la gráfica de la función φ_0 y en la derecha se muestra la gráfica de la función φ_j.

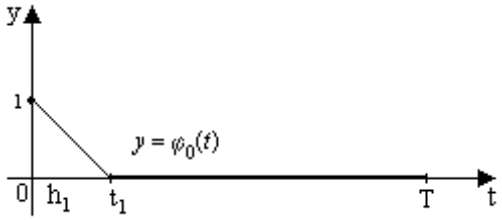


Figura 98

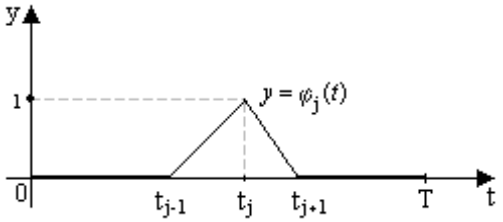


Figura 99

Estas funciones φ_j se las denomina funciones techo.

De la definición del espacio V_h y de la base $\{\varphi_0,\ldots,\varphi_{n-1}\}$ resulta que $v_h \in V_h$ si y solo si existen $v_0,\ldots,v_{n-1} \in \mathbb{R}$ tales que $v_h = \sum_{i=0}^{n-1} v_i\varphi_i$. Note que $v_h(T) = 0$ pués $\varphi_i(T) = 0 \quad \forall i = 0,\cdots,n-1$.

Definición 4 *Diremos que $u_h \in U_h$ es solución aproximada del problema de Cauchy de valor inicial si y solo si satisface la ecuación integral*

$$-\int_0^T u_h(t)\, v_h'(t)\, dt = u_0 v_h(0) + \int_0^T f(t,u_h(t))\, v_h(t)\, dt \quad \forall v_h \in V_h.$$

Esta es la formulación discreta de la formulación del tipo Petrov-Galerkin arriba definida, a la que nos referimos como formulación discreta del tipo Petrov-Galerkin.

Observe que la función incógnita u_h se busca en el espacio U_h mientras que las funciones v_h, denominadas funciones test, están en el espacio V_h.

Aplicando la formulación discreta del tipo Petrov-Galerkin se construye a continuación un esquema numérico.

Sea $u_h \in V_h$ la solución. Existe $u_1, \ldots, u_n \in \mathbb{R}$ tales que $u_h = \sum_{j=1}^{n} u_j \chi_j$. Remplazando en la formulación discreta del tipo Petrov-Galerkin, tenemos

$$-\int_0^T \left(\sum_{j=1}^{n} u_j \chi_j (t) \right) v_h' (t)\, dt = u_0 v_h (0) + \int_0^T f \left(t, \sum_{j=1}^{n} u_j \chi_j (t) \right) v_h (t)\, dt \quad \forall v_h \in V_h,$$

y tomando en consideración la linealidad de la integral en el primer miembro, esta ecuación se expresa como

$$-\sum_{j=1}^{n} u_j \int_0^T \chi_j (t)\, v_h' (t)\, dt = u_0 v_h (0) + \int_0^T f \left(t, \sum_{j=1}^{n} u_j \chi_j (t) \right) v_h (t)\, dt.$$

Debemos calcular $u_1, \ldots, u_n$ y tenemos una ecuación válida para todo $v_h \in V_h$, en particular lo es para los elementos de la base $\{\varphi_0, \ldots, \varphi_{n-1}\}$ de V_h lo que nos permite obtener las n ecuaciones requeridas. En efecto, hacemos $v_h{=}\varphi_i \in V_h$. Tenemos

$$-\sum_{j=1}^{n} u_j \int_0^T \chi_j (t)\, \varphi_i' (t)\, dt = u_0 \varphi_i (0) + \int_0^T f \left(t, \sum_{j=1}^{n} u_j \chi_j (t) \right) \varphi_i (t)\, dt \qquad i = 0, \ldots, n-1,$$

lo que da lugar a las n ecuaciones. Determinemos estas ecuaciones.

Para $i = 0$ se tiene $\varphi_0 (0) = 1$, $\varphi_0' (t) = \begin{cases} -\dfrac{1}{h_1}, \text{ si } t \in]0, t_1[\,, \\ 0, \text{ en otro caso,} \end{cases}$ además en el intervalo $]0, t_1[$ intervienen u_1 y $\chi_1 (t) = 1 \quad t \in]0, t_1[$ con lo que se obtiene

$$-u_1 \int_0^{t_1} -\frac{1}{h_1} dt = u_0 + \int_0^{t_1} f (t, u_1)\, \varphi_0 (t)\, dt.$$

Como $h_1 = t_1$ e $\int_0^{t_1} -\dfrac{1}{h_1} dt = -1$, entonces

$$u_1 = u_0 + \int_0^{t_1} f (t, u_1)\, \varphi_0 (t)\, dt.$$

Para $i = 2, \ldots, n$ se tiene $\varphi_i (0) = 0$, más aún $\varphi_i (t_j) = \begin{cases} 1, \text{ si } i = j \\ 0, \text{ si } i \neq j, \end{cases}$ $\varphi_i' (t) = \begin{cases} \dfrac{1}{h_i}, \text{ si } t \in]t_{i-1}, t_i[\,, \\ -\dfrac{1}{h_{i+1}}, \text{ si } t \in]t_i, t_{i+1}[\,, \\ 0, \text{ en otro caso.} \end{cases}$ Además en los intervalos $]t_{i-1}, t_i[$ y $]t_i, t_{i+1}[$ intervenen $u_i,\ u_{i+1},\ \chi_i (t) = 1 \quad t \in]t_{i-1}, t_i[\,,\ \chi_{i+1} (t) = 1 \quad t \in]t_i, t_{i+1}[$. Entonces

$$-\left(u_i \int_{t_{i-1}}^{t_i} \frac{1}{h_i} dt + u_{i+1} \int_{t_i}^{t_{i+1}} -\frac{1}{h_{i+1}} dt \right) = \int_{t_{i-1}}^{t_i} f (t, u_i)\, \varphi_i (t)\, dt + \int_{t_i}^{t_{i+1}} f (t, u_{i+1})\, \varphi_i (t)\, dt.$$

Puesto que $\int_{t_{i-1}}^{t_i} \dfrac{1}{h_i} dt = 1, \quad \int_{t_i}^{t_{i+1}} \dfrac{1}{h_{i+1}} dt = 1$, se sigue que

$$-(u_i - u_{i+1}) = \int_{t_{i-1}}^{t_i} f (t, u_i)\, \varphi_i (t)\, dt + \int_{t_i}^{t_{i+1}} f (t, u_{i+1})\, \varphi_i (t)\, dt$$

que a su vez se expresa como

$$u_{i+1} - \int_{t_i}^{t_{i+1}} f (t, u_{i+1})\, \varphi_i (t)\, dt = u_i + \int_{t_{i-1}}^{t_i} f (t, u_i)\, \varphi_i (t)\, dt \quad i = 1, \ldots, n-1.$$

En resumen, el esquema numérico que se obtiene es el siguiente:

$$\begin{cases} u_i - \int_0^{t_i} f(t,u_i)\,\varphi_0 dt = u_0, \\ \qquad\qquad \vdots \\ u_{i+1} - \int_{t_i}^{t_{i+1}} f(t,u_{i+1})\,\varphi_i(t)\,dt = u_i + \int_{t_{i-1}}^{t_i} f(t,u_i)\,\varphi_i(t)\,dt \quad i=1,\ldots,n-1. \end{cases}$$

Lastimosamente en este esquema numérico se debe aún calcular las integrales. Para ello aplicamos el método de los trapecios; esto es, si $g \in C([a,b])$ entonces $\int_a^b g(t)\,dt \simeq \dfrac{b-a}{2}\left(g(a)+g(b)\right)$. Entonces

$$\int_0^{t_1} f(t,u_1)\,\varphi_0(t)\,dt \simeq \frac{t_1}{2}\left[f(0,u_1)\,\varphi_0(0) + f(t_1,u_1)\,\varphi_0(t_1)\right]$$

y como $h_1 = t_1$, $\varphi_0 = 1$, $\varphi_0(t_1) = 0$ resulta

$$\int_0^{t_1} f(t,u_1)\,\varphi_0(t)\,dt \simeq \frac{h_1}{2} f(0,u_i).$$

De manera similar

$$\begin{aligned} \int_{t_{i-1}}^{t_i} f(t,u_i)\,\varphi_i(t)\,dt &\simeq \frac{t_i - t_{i-1}}{2}\left[f(t_{i-1},u_i)\,\varphi_i(t_{i-1}) + f(t_i,u_i)\,\varphi_i(t_i)\right], \\ \int_{t_i}^{t_i+1} f(t,u_{i+1})\,\varphi_i(t)\,dt &\simeq \frac{t_{i+1} - t_i}{2}\left[f(t_i,u_{i+1})\,\varphi_i(t_i) + f(t_{i+1},u_i)\,\varphi_i(t_{i+1})\right], \end{aligned}$$

y por la definición de h_i, h_{i+1}, y de las funciones $\varphi_1,\ldots,\varphi_{n-1}$, tenemos $\varphi_i(t_{i-1}) = \varphi_i(t_{i+1}) = 0$, $\varphi_i(t_i) = 0$.

Entonces

$$\begin{aligned} \int_{t_{i-1}}^{t_i} f(t,u_i)\,\varphi_i(t)\,dt &\simeq \frac{h_i}{2} f(t_i,u_i), \\ \int_{t_i}^{t_{i+1}} f(t,u_{i+1})\,\varphi_i(t)\,dt &\simeq \frac{h_{i+1}}{2} f(t_i,u_{i+1}). \end{aligned}$$

Por abuso de lenguaje designamos nuevamente con $u_1,\ldots,u_n$ a las incógnitas que satisfacen el esquema numérico siguiente:

$$\begin{cases} u_1 - \dfrac{h_1}{2} f(0,u_1) = u_0 \\ \qquad\qquad \vdots \\ u_{i+1} - \dfrac{h_{i+1}}{2} f(t_i,u_{i+1}) = u_i + \dfrac{h_i}{2} f(t_i,u_i) \quad i=1,\ldots,n-1. \end{cases}$$

Se observa que este esquema numérico es similar al de Crank-Nicolson al que nos referimos como esquema numérico del tipo Crank-Nicolson, el mismo que fue propuesto por HB-PB .

Otra forma de obtener este esquema numérico es la siguiente. Definimos la función φ_n como sigue:

$$\varphi_n(t) = \begin{cases} 0, \text{ si } t \in [0, t_{n-1}[\,, \\ \dfrac{t - t_{n-1}}{h_n}, \text{ si } t \in [t_{n-1}, T]\,. \end{cases}$$

En la figura siguiente se muestra la gráfica de la función φ_n.

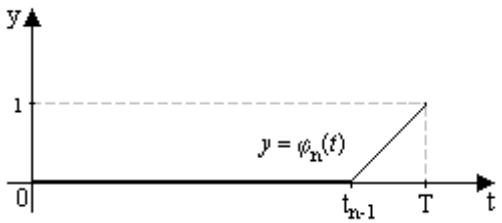


Figura 100

Introducimos los subespacios U_h y V_h de $C_d^1([0,T])$ y $C_*^1([0,T])$ respectivamente como sigue:

$$\begin{aligned} U_h &= \left\{ \sum_{i=0}^{n} u_i \varphi_i \mid u_i \in \mathbb{R} \quad i = 0, \ldots, n \right\}, \\ V_h &= \left\{ \sum_{i=0}^{n-1} \alpha_i \varphi_i \mid \alpha_i \in \mathbb{R} \quad i = 0, 1, \ldots, n-1 \right\}. \end{aligned}$$

Se tiene $\dim U_h = n+1$, $\dim V_h = n$. una base de U_h es la familia de funciones $\{\varphi_0, \ldots, \varphi_n\}$ y una de V_h es $\{\varphi_0, \ldots, \varphi_{n-1}\}$. La formulación discreta de la formulación del tipo Petrov-Galerkin se expresa como sigue: hallar una función $u_h \in U_h$ solución de

$$-\int_0^T u_h(t) v_h'(t)\, dt = u_0 v_h(0) + \int_0^T f(t, u_h(t)) v_h(t)\, dt \quad \forall v_h \in V_h.$$

A esta acción nos referimos como formulación discreta del tipo Petrov-Galerkin.

La función $u_h \in U_h$ se escribe como $u_h = \sum_{i=0}^{n} u_i \varphi_i = u_0 \varphi_0 + \sum_{i=1}^{n} u_i \varphi_i$, con $u(0) = u_0$ la condición inicial. Luego

$$-\int_0^T \left(u_0 \varphi_0(t) + \sum_{i=1}^{n} u_i \varphi_i(t) \right) v_h'(t)\, dt = u_0 v_h(0) + \int_0^T f\left(t, u_0 \varphi_0(t) + + \sum_{i=1}^{n} u_i \varphi_i \right) v_h(t)\, dt \quad \forall v_h \in V_h.$$

Se observa en esta ecuación que las incógnitas son $u_1, \ldots, u_n \in \mathbb{R}$. Para poder calcular estas incógnitas requerimos generar n ecuaciones, para ello remplazamos sucesivamente v_h por los elementos de la base $\{\varphi_0, \ldots, \varphi_{n-1}\}$ de V_h, es decir $v_h = \varphi_j \quad j = 0, 1, \ldots, n-1$. Así,

$$\begin{aligned} -u_0 \int_0^T \varphi_0(t) \varphi_j'(t)\, dt - \sum_{i=1}^{n-1} u_i \int_0^T \varphi_i(t) \varphi_j'(t)\, dt &= u_0 \varphi_j(0) + \\ + \int_0^T f\left(t, u_0 \varphi_0(t) + \sum_{i=1}^{n} u_i \varphi_i(t) \right) \varphi_j(t)\, dt \qquad j &= 0, 1, \ldots, n-1. \end{aligned}$$

Para $j = 0$, de la definición de φ_0 se tiene $\varphi_0(0) = 1$, $\varphi_0(t) = 0$ si $t \in [t_1, T]$, $\varphi_0'(t) = \begin{cases} -\dfrac{1}{h_1}, \text{ si } o < t < t_1, \\ 0, \text{ si } t_1 < t < T. \end{cases}$ Tomando en consideración esta información, la ecuación precedente se reduce a la siguiente

$$-u_0 \int_0^{t_1} \varphi_0(t) \left(-\frac{1}{h_1} \right) dt - u_i \int_0^{t_1} \varphi_1(t) \left(-\frac{1}{h_1} \right) dt = u_0 + \int_0^{t_1} f(t, u_0 \varphi_0(t) + u_1 \varphi_1(t)) \varphi_0(t)\, dt$$

e integrando, obtenemos

$$\frac{1}{2} u_0 + \frac{1}{2} u_1 = u_0 + \int_0^{t_1} f(t, u_0 \varphi_0(t) + u_1 \varphi_1(t)) \varphi_0(t)\, dt.$$

Para $j = 1$, de la definición de φ_1 se tiene $\varphi_1(t_1) = 1, \varphi_1(0) = 0, \varphi_1(t) = 0$ si $t \in [t_2, T]$ y $t = 0$, la derivada de φ_1 está definida como sigue:$\varphi_1'(t) = \begin{cases} \dfrac{1}{h_1}, \text{ si } 0 < t < t_1, \\ -\dfrac{1}{h_2}, \text{ si } t_1 < t < t_2, \\ 0, \text{ si } t_2 < t < T. \end{cases}$ Entonces $u_h = u_0 \varphi_0 + u_1 \varphi_1 + u_2 \varphi_2$ sobre $[0, t_2]$ en consecuencia, por la aditividad respecto del dominio de integración, tenemos

$$\begin{aligned} -u_0 \int_0^{t_1} \varphi_0(t) \varphi_1'(t)\, dt - u_1 \int_0^{t_1} \varphi_1(t) \varphi_1'(t)\, dt - u_1 \int_{t_1}^{t_2} \varphi_1(t) \varphi_1'(t)\, dt - u_2 \int_{t_1}^{t_2} \varphi_2(t) \varphi_1'(t)\, dt \\ = u_0 \varphi_1(0) + \int_0^{t_2} f(t, u_0 \varphi_0(t) + u_1 \varphi_1(t) + u_2 \varphi_2(t)) \varphi_1(t)\, dt, \end{aligned}$$

y de esta , resulta

$$\begin{aligned} & -u_0\int_0^{t_1}\varphi_0(t)\left(\frac{1}{h_1}\right)dt-u_1\int_0^{t_1}\varphi_1(t)\frac{1}{h_1}dt-u_1\int_{t_1}^{t_2}\varphi_1(t)\left(-\frac{1}{h_2}\right)dt-u_2\int_{t_1}^{t_2}\varphi_2(t)\left(-\frac{1}{h_2}\right)dt \\ = \;& u_0\varphi_1(0)+\int_0^{t_1}f(t,u_0\varphi_0(t)+u_1\varphi_1(t)+u_2\varphi_2(t))\varphi_1(t)\,dt \\ & +\int_{t_1}^{t_2}f(t,u_0\varphi_0(t)+u_1\varphi_1(t)+u_2\varphi_2(t))\varphi_1(t)\,dt \end{aligned}$$

Integrando los tres primeros términos obtenemos el siguiente resultado

$$\begin{aligned} -\frac{u_0}{2}+\frac{u_2}{2} &= \int_0^{t_1}f(t,u_0\varphi_0(t)+u_1\varphi_1(t)+u_2\varphi_2(t))\varphi_1(t)\,dt \\ &\quad +\int_{t_1}^{t_2}f(t,u_0\varphi_0(t)+u_1\varphi_1(t)+u_2\varphi_2(t))\varphi_1(t)\,dt. \end{aligned}$$

Continuando con este proceso, obtenemos para $1\leq j\leq n-1$,

$$\begin{aligned} -\frac{1}{2}u_{j-1}+\frac{1}{2}u_{j+1} &= \int_{t_{j-1}}^{t_j}f\left(t,u_{j-1}\varphi_{j-1}(t)+u_j\varphi_j(t)+u_{j+1}\varphi_{j+1}(t)\right)\varphi_j(t)\,dt \\ &\quad +\int_{t_j}^{t_{j+1}}f\left(t,u_{j-1}\varphi_{j-1}(t)+u_j\varphi_j(t)+u_{j+1}\varphi_{j+1}(t)\right)\varphi_j(t)\,dt. \end{aligned}$$

Para el cálculo de las integrales aplicamos la fórmula de integración del punto medio, esto es, si $g\in C([a,b])$,

$$\int_a^b g(t)\,dt\simeq(b-a)\,g\left(\frac{a+b}{2}\right).$$

Entonces, de la definición de las funciones φ_0, φ_1 se tiene

$$\int_0^{t_1}f(t,u_0,\varphi_0(t)+u_1\varphi_1(t))u_0(t)\,dt\simeq\frac{h_1}{2}f\left(\widetilde{t}_0,\frac{1}{2}(u_0+u_1)\right),$$

donde $\widetilde{t}_0=\frac{1}{2}(t_0+t_1)$. Note que $\varphi_0\left(\widetilde{t}_0\right)=\frac{1}{2}$.

De manera similar, de la definición de las funciones φ_0, φ_1, φ_2 se tiene

$$\begin{aligned} \int_0^{t_1}f(t,u_0\varphi_0(t)+u_1\varphi_1(t)+u_2\varphi_2(t))\varphi_1(t)\,dt &\simeq \frac{h_1}{2}f\left(\widetilde{t}_0,\frac{1}{2}(u_0+u_1)\right), \\ \int_{t_1}^{t_2}f(t,u_0\varphi_0(t)+u_1\varphi_1(t)+u_2\varphi_2(t))\varphi_2(t)\,dt &\simeq \frac{h_2}{2}f\left(\widetilde{t}_1,\frac{1}{2}(u1+u_2)\right), \end{aligned}$$

donde $\widetilde{t}_1=\frac{1}{2}(t_1+t_2)$, $\varphi_1\left(\widetilde{t}_0\right)=\frac{1}{2}$, $\varphi_1\left(\widetilde{t}_1\right)=\frac{1}{2}$.

De manera general, obtenemos

$$\begin{aligned} \int_{t_{j-1}}^{t_j}f\left(t,u_{j-1}\varphi_{j-1}(t)+u_j\varphi_j(t)+u_{j+1}\varphi_{j+1}(t)\right)\varphi_j(t)\,dt &= \frac{h_j}{2}f\left(\widetilde{t}_{j-1},\frac{1}{2}(u_{j-1}+u_j)\right) \\ \int_{t_j}^{t_{j+1}}f\left(t,u_{j-1}\varphi_{j-1}(t)+u_j\varphi_j(t)+u_{j+1}\varphi_{j+1}(t)\right)\varphi_j(t)\,dt &\simeq \frac{h_{j+1}}{2}f\left(\widetilde{t}_j,\frac{1}{2}(u_j+u_{j+1})\right) \end{aligned}$$

con $\widetilde{t}_{j-1}=\frac{1}{2}(t_{j-1}+t_j)$, $\widetilde{t}_j=\frac{1}{2}(t_j+t_{j+1})$.

Resulta que para $j=0$, $j=1,\ldots,n-1$ el esquema numérico se expresa como

$$\begin{cases} \frac{1}{2}(u_0+u_1) \simeq u_0 + \frac{h_1}{2} f\left(\widetilde{t}_0, \frac{1}{2}(u_0+u_1)\right), \\ -\frac{1}{2}u_0 + \frac{1}{2}u_2 \simeq \frac{h_1}{2} f\left(\widetilde{t}_0, \frac{1}{2}(u_0+u_1)\right) + \frac{h_2}{2} f\left(\widetilde{t}_1, \frac{1}{2}(u_1+u_2)\right), \\ \qquad\qquad \vdots \\ -\frac{1}{2}u_{j-1} + \frac{1}{2}u_{j+1} \simeq \frac{h_j}{2} f\left(\widetilde{t}_{j-1}, \frac{1}{2}(u_{j-1}+u_j)\right) + \frac{h_{j+1}}{2} f\left(\widetilde{t}_j, \frac{1}{2}(u_j+u_{j+1})\right). \end{cases}$$

Sumando y restando $\frac{1}{2}u_1$ y de manera general $\frac{1}{2}u_j$, se tiene

$$\begin{cases} \frac{1}{2}(u_0+u_1) - \frac{h_1}{2} f\left(\widetilde{t}_0, \frac{1}{2}(u_0+u_1)\right) \simeq u_0, \\ \frac{1}{2}u_0 + \frac{1}{2}u_2 - \frac{h_2}{2} f\left(\widetilde{t}_1, \frac{1}{2}(u_1+u_2)\right) \simeq \frac{1}{2}(u_0+u_1) + \frac{h_1}{2} f\left(\widetilde{t}_0, \frac{1}{2}(u_0+u_1)\right), \\ \qquad\qquad \vdots \\ \frac{1}{2}(u_j+u_{j+1}) - \frac{h_{j+1}}{2} f\left(\widetilde{t}_j, \frac{1}{2}(u_j+u_{j+1})\right) \simeq \frac{1}{2}(u_{j-1}+u_j) + \frac{h_j}{2} f\left(\widetilde{t}_{j-1}, \frac{1}{2}(u_{j-1}+u_j)\right). \end{cases}$$

En base a este resultado obtenemos el siguiente esquema numérico: buscamos $\widetilde{u}_1,\ldots,\widetilde{u}_n \in \mathbb{R}$ solución del sistema de ecuaciones siguiente:

$$\begin{cases} \widetilde{u}_1 - \frac{h_1}{2} f\left(\widetilde{t}_0, \widetilde{u}_1\right) = u_0, \\ \widetilde{u}_2 - \frac{h_2}{2} f\left(\widetilde{t}_1, \widetilde{u}_2\right) = \widetilde{u}_1 + \frac{h_1}{2} f\left(\widetilde{t}_0, \widetilde{u}_1\right), \\ \qquad\qquad \vdots \\ \widetilde{u}_{j+1} - \frac{h_{j+1}}{2} f\left(\widetilde{t}_j, \widetilde{u}_{j+1}\right) = \widetilde{u}_j + \frac{h_j}{2} f\left(\widetilde{t}_{j-1}, \widetilde{u}_j\right) \quad j=1,2,\ldots,n-1. \end{cases}$$

que es el esquema numérico del tipo Crank-Nicolson que hemos obtenido anteriormente.

El esquema numérico obtenido para el caso escalar de una ecuación diferencial se extiende inmediatamente a sistemas de ecuaciones diferenciales. Así, consideremos el sistema de ecuaciones diferenciales siguiente:

$$\begin{cases} \overrightarrow{u}'(t) = \overrightarrow{F}\left(t, \overrightarrow{u}(t)\right) \quad t \in]0,T[, \\ \overrightarrow{u}(0) = \overrightarrow{u}_0, \end{cases}$$

donde $T>0$, $\overrightarrow{u}_0^T = \left(u_1^{(0)},\ldots,u_m^{(0)}\right) \in \mathbb{R}^m$, $\overrightarrow{F}$ es una función vectorial de $[0,T]\times\mathbb{R}^m$ en $\mathbb{R}^m$ que suponemos lipchisiana, esto es, existe $L>0$ tal que $\forall \overrightarrow{y}_1, \overrightarrow{y}_2 \in \mathbb{R}^m$,

$$\left\| \overrightarrow{F}\left(t, \overrightarrow{y}_1\right) - \overrightarrow{F}\left(t, \overrightarrow{y}_2\right) \right\| \leq L \left\| \overrightarrow{y}_1 - \overrightarrow{y}_2 \right\| \quad \forall t \in [0,T].$$

Entonces

$$\begin{cases} \overrightarrow{u}_1 - \frac{h_1}{2} \overrightarrow{F}\left(\widetilde{t}_0, \overrightarrow{u}_1\right) = \overrightarrow{u}_0, \\ \overrightarrow{u}_{j+1} - \frac{h_{j+1}}{2} \overrightarrow{F}\left(\widetilde{t}_j, \overrightarrow{u}_{j+1}\right) = \overrightarrow{u}_j + \frac{h_j}{2} \overrightarrow{F}\left(t_{j-1}, \widetilde{u}_j\right) \end{cases}$$

donde $\widetilde{t}_j = \frac{1}{2}(t_{j-1}+t_j)$ los puntos medios de los intervalos $[t_{j-1},t_j]$ de la partición $\tau(n)$ de $[0,T]$.

Más particularmente, si la función $\overrightarrow{F}$ tiene la forma

$$\overrightarrow{F}\left(t, \overrightarrow{y}\right) = A(t)\overrightarrow{y} + \overrightarrow{b}(t) \quad t \in [0,T],$$

con $A(t) = (a_{ij}(t))$ una matriz de $m\times m$ no nula y cada $a_{ij}(t)$ función continua, $\overrightarrow{b}$ una función vectorial continua.

El esquema numérico se expresa como sigue;

$$\begin{cases} \left[I - \frac{h_1}{2} A\left(\widetilde{t}_0\right)\right] \overrightarrow{u}_1 = \overrightarrow{u}_0 + \frac{h_1}{2} \overrightarrow{b}\left(t_0\right) \\ \qquad \vdots \\ \left[I - \frac{h_{j+1}}{2} A\left(\widetilde{t}_j\right)\right] \overrightarrow{u}_{j+1} = \left(I - \frac{h_j}{2} A\left(\widetilde{t}_{j-1}\right)\right) \overrightarrow{u} j + \frac{1}{2} h_j \overrightarrow{b}\left(\widetilde{t}_{j-1}\right) + \frac{1}{2} h_{j+1} \overrightarrow{b}\left(\widetilde{t}_j\right) \quad j = 1, \ldots, n-1. \end{cases}$$

11.4. Método de diferencias finitas para problemas de valores en la frontera 1d.

El método de diferencias finitas (MDF) es uno de los primeros métodos que fueron implementados en la resolución numérica tanto de ecuaciones diferenciales ordinarias como en derivadas parciales para problemas uni, bi y tridimensionales. Su popularidad radica en el hecho de la simplicidad con la que se discretizan tales ecuaciones mediante el uso de aproximaciones de las derivadas por medio de cocientes incrementales.

En este capítulo iniciaremos con los operadores en diferencias finitas que luego serán aplicados a una clase de problemas con valores en la frontera unidimensionales siguientes:

$$-\frac{d}{dx}\left(p\frac{du}{dx}\right) + r\frac{du}{dx} + qu = f \quad \text{sobre} \quad]0, L],$$

de donde $L > 0$, p, q, r, $f \in C^0\left([0, L]\right)$ tales que

i. $p\left(x\right) \geq \alpha > 0 \quad \forall x \in [0, L]$,

ii. $q\left(x\right) \geq 0 \quad \forall x \in [0, L]$.

Para el problema propuesto consideramos cuatro condiciones de frontera que precisamos a continuación.

1. Condiciones de frontera de Dirichlet: $u\left(0\right) = a_0, \quad u\left(L\right) = a_1$.
2. Condiciones de frontera de Neumann: $u'\left(0\right) = a, \quad u'\left(L\right) = b$.
3. Condiciones de frontera mixtas: $u'\left(0\right) + \alpha u\left(0\right) = a, \quad u'\left(L\right) + \beta u\left(L\right) = b$.
4. Condiciones de frontera periódicas: $u\left(0\right) = u\left(L\right), \quad u'\left(0\right) = u'\left(L\right)$. En este último problema debemos suponer que las funciones p, q, r, f se extienden por periodicidad a todo $\mathbb{R}$ y conservan la continuidad. Nótese que

$$\begin{aligned} u\left(x + L\right) &= u\left(x\right) \quad \forall x \in \mathbb{R}, \\ u'\left(x + L\right) &= u'\left(x\right) \quad \forall x \in \mathbb{R}. \end{aligned}$$

Suponemos que para el problema propuesto, se conocen resultados de existencia, unicidad, regularidad de la solución.

Con el propósito de introducir las nociones de consistencia, estabilidad y convergencia que serán abordados más adelante, consideramos el problema modelo siguiente:

$$\text{hallar } u \in C^2\left([0, L]\right) \text{ solución de } \begin{cases} -u'' + qu = f \quad \text{sobre }]0, L[, \\ u\left(0\right) = u\left(L\right) = 0, \end{cases} \tag{P}$$

donde f, $q \in C^0\left([0, L]\right)$ con $q\left(x\right) \geq 0 \quad \forall x \in [0, L]$.

Este problema es una simplificación del problema planteado en la introducción. Con este problema abordaremos otros desde el punto de vista informático que consiste en la puesta en marcha del método de diferencias finitas.

Sea $n \in \mathbb{Z}^+$, $h = \dfrac{L}{n}$ y $x_j = jh$, $j = 0, 1, ..., n$. Ponemos

$$\tau(n) = \{x_j = jh \mid j = 0, 1, ..., n\}.$$

El conjunto $\tau(n)$ se llama discretización del intervalo $[0, L]$ o también malla de $[0, L]$.

Puesto que

$$u''(x_j) = \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + r_h(x_j), \quad j = 1, ..., n-1,$$

(véase el capítulo 2, diferencias finitas centrales) se sigue que (P) se discretiza del modo siguiente:

$$\begin{cases} -\dfrac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + q(x_j)\, u(x_j) + r_h(x_j) = f(x_j) \quad j = 1, ..., n-1, \\ u(0) = 0, \quad u(L) = 0 \end{cases} \tag{P$_1$}$$

En el esquema numérico precedente (P$_1$) se desconocen $u(x_j)$ y $r_h(x_j)$, $j = 1, ..., n-1$. Deseamos que $r_h(x_j) \to_{h\to 0} 0$ $j = 1, ..., n-1$.

Denotamos con u_j una aproximación de $u(x_j)$. Asumimos que $r_h(x_j) \simeq 0$ y en consecuencia, se tiene el esquema numérico siguiente.

$$\begin{cases} u_0 = 0, \quad u_n = 0, \\ -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + q(x_j)\, u_j = f(x_j) \quad j = 1, ..., n-1. \end{cases}$$

Para $j = 1$, se tiene

$$\left[\frac{2}{h^2} + q(x_1)\right] u_1 - \frac{1}{h^2} u_2 = f(x_1).$$

Para $1 < j < n-1$,

$$-\frac{1}{h^2} u_{j-1} + \left[\frac{2}{h^2} + q(x_j)\right] u_j - \frac{1}{h^2} u_{j+1} = f(x_j).$$

Para $j = n-1$,

$$-\frac{1}{h^2} u_{n-2} + \left[\frac{2}{h^2} + q(x_{n-1})\right] u_{n-1} = f(x_{n-1}).$$

El conjunto de ecuaciones precedente, en forma matricial se escribe en la siguiente forma:

$$\begin{bmatrix} \frac{2}{h^2} + q(x_1) & -\frac{1}{h^2} & 0 & \cdots & 0 \\ -\frac{1}{h^2} & \frac{2}{h^2} + q(x_2) & -\frac{1}{h^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -\frac{1}{h^2} \\ 0 & \cdots & \cdots & -\frac{1}{h^2} & \frac{2}{h^2} + q(x_{n-1}) \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_{n-2}) \\ f(x_{n-1}) \end{bmatrix} \tag{1}$$

Ponemos $\overrightarrow{u}_h = (u_1, ..., u_{n-1})^T$, $\overrightarrow{b} = (f(x_1), ..., f(x_{n-1}))^T$, $A_h = (a_{ij}(h))$ con

$$\begin{aligned} a_{ii}(h) &= \frac{2}{h^2} + q(x_i) \quad i = 1, ..., n-1, \\ a_{i-1i}(h) &= a_{ii-1}(h) = -\frac{1}{h^2}, \quad i = 2, ..., n-1. \end{aligned}$$

El sistema de ecuaciones lineales (1) se escribe en forma compacta como

$$A_h \overrightarrow{u}_h = \overrightarrow{b}. \tag{2}$$

La matriz A_h es de $(n-1) \times (n-1)$, tridiagonal, simétrica y se demostrará más adelante que es definida positiva, por lo que el sistema de ecuaciones (2) tiene una única solución $\overrightarrow{u}_n \in \mathbb{R}^{n-1}$.

11.4.1. Aspectos informáticos del método de diferencias finitas

El problema propuesto en la sección precedente sugiere proponer la siguiente metodología para su resolución.

1. **Lectura de datos de entrada.**
   Leer $L > 0,\ \ n \in \mathbb{Z}^+$, funciones q y f.

2. **Preparación de los datos de entrada.**

   i. Generación de la malla $\tau(n) = \{x_j \mid j = 0, 1, ..., n\}$.
   a. Generación manual.
   b. Generación automática.

   ii. Construcción de los vectores $\overrightarrow{q}$ y $\overrightarrow{b}$.

$$\begin{aligned} \overrightarrow{q} &= (q(x_1), ..., q(x_{n-1})), \\ \overrightarrow{b} &= (f(x_1), ..., f(x_{n-1})). \end{aligned}$$

   iii. Condiciones de frontera: $u(0),\ \ u(L)$.

$$\begin{aligned} u(0) &= 0, \\ u(L) &= 0. \end{aligned}$$

   iv. Construcción de la matriz $A_h = (a_{ij}(h))$.

$$\begin{aligned} a_{ii}(h) &= \frac{2}{h^2} + q(x_i) \quad i = 1, ..., n-1, \\ a_{i-1i}(h) &= a_{ii-1}(h) = -\frac{1}{h^2} \quad i = 2, ..., n-1. \end{aligned}$$

3. **Ejecución del algoritmo.**
   Resolución del sistema de ecuaciones

$$A_h \overrightarrow{u}_h = \overrightarrow{b}. \qquad (*)$$

   a) Método directo: factorización LU.
   b) Método iterativo: $S.O.R$.

4. **Preparación de los datos de salida.**

   i. Construir el archivo que contiene n, $\tau(n)$ y $\overrightarrow{u}_h$; por ejemplo u_h.dat contiene

$$\begin{array}{cc} & n \\ x_i \quad u_{h,i}, & i = 0, ..., n. \end{array}$$

   ii. Con el propósito de efectuar pruebas, es recomendable conocer la solución exacta u. Con esta solución se debe construir el vector $\overrightarrow{u} = (u(0), ..., u(L))$ y en consecuencia generar el archivo u. dat que contiene

$$\begin{array}{ccc} & & n \\ x_i & u_i & i = 0, ..., n. \end{array}$$

   iii. Supongamos que $u \in V$, donde V es un espacio de funciones provisto de la norma $\|\cdot\|_V$. Calcular el error $r(n) = \|u - u_h\|_V$ para diferentes discretizaciones. Generar un archivo que contiene n y $e(n)$, por ejemplo: e_h.dat:

$$n \quad e(n) \qquad n = n_1,\ 2n_1,\ 4n_1,\ 8n_1, ..., 2^{n_0} n_1,$$

   donde $n_0 \in \mathbb{Z}^+$ (por ejemplo: $n_1 = 20,\ \ n_0 = 5$).

5. **Presentación de resultados.**

   i. Gráficas de u y u_h.

   ii. Curva de errores.

Nota: Se conocen las soluciones u y u_h en los puntos $x_i \in \tau(n)$; esto es, se tienen los conjuntos de puntos:

$$\begin{aligned} G_1 &= \{(x_i, u(x_i)) \mid i = 0, 1, ..., n\}, \\ G_2 &= \{(x_i, u_{h,i}) \mid i = 0, 1, ..., n\}. \end{aligned}$$

Al representar G_1 y G_2 en el sistema de coordenadas rectangulares XY, se tienen puntos en el plano. Estos puntos pueden ser unidos con segmentos de recta, polinomios de grado 2, polinomios cúbicos, etc. En definitiva, se utilizarán B-splines de orden 1, 2, 3, etc. Por sencillez se utilizarán los B-splines de orden 1. Para ello definimos

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i}, & x \in [x_{i-1}, x_i], \\ -\dfrac{x - x_{i+1}}{h_{i+1}}, & x \in]x_i, x_{i+1}], \\ 0, & x \in [0, L] \setminus [x_{i-1}, x_{i+1}] \end{cases} \quad i = 1, ..., n-1.$$

Note que $\varphi_i(x_j) = \delta_{ij} \quad i, j = 1, ..., n-1$; además $h_i = h, \ \ i = 1, ..., n$.

Ponemos

$$\begin{aligned} \widetilde{u}(x) &= \sum_{i=1}^{n-1} u(x_i)\, \varphi_i(x), \\ u_h(x) &= \sum_{i=1}^{n-1} u_{n,i} \varphi_i(x) \\ e(n) &= \sum_{k=1}^{n_0} e(n_j)\, \psi_k(n) \end{aligned}$$

con $\psi_k(n)$ definida de manera análoga a $\varphi_i(x)$.

11.4.2. Consistencia, estabilidad, convergencia

1. Sea $L > 0$. Consideramos el problema siguiente:

$$\text{Hallar } u \in C^2([0, L]) \text{ solución de } \begin{cases} -u'' + qu = f \quad \text{sobre }]0, L[, \\ u(0) = u(L) = 0, \end{cases} \tag{1}$$

donde $q,\ f \in C^0([0, L])$ con $q(x) \geq 0 \quad \forall x \in [0, L]$.

El esquema numérico que aproxima la solución u es el siguiente:

$$\begin{cases} u_0 = u_n = 0 \\ -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + q(x_j)\, u_j = f(x_j), \quad j = 1, ..., n-1, \end{cases} \tag{2}$$

donde $\tau(n) = \{x_j = jh \mid j = 0, 1, ..., n\}$ es el conjunto de nodos de $[0, L]$, $h = \frac{L}{n}$ y $n \in \mathbb{Z}^+$.

El esquema numérico (2), en forma matricial, se escribe:

$$A_h \overrightarrow{u}_h = \overrightarrow{b}, \tag{3}$$

con $\overrightarrow{u}_h = (u_1, ..., u_{n-1})^T$, $\overrightarrow{b} = (f(x_1), ..., f(x_{n-1}))$, $A_h = (a_{ij}(h))$ la matriz definida por:

$$\begin{aligned} a_{ii}(h) &= \frac{2}{h^2} + q(x_i) \quad i = 1, ..., n-1, \\ a_{ii-1}(h) &= -\frac{2}{h^2} = a_{i-1,i}(h) \quad i = 2, ..., n-1. \end{aligned}$$

Definición 5 *Toda función* $g:\begin{cases} \tau(n) & \to & \mathbb{R} \\ x_j & \to & g(x_j)=y_j \end{cases}$ *se llama función reticular que se escribirá* $\overrightarrow{g}=(g(x_0),...,g(x_n))$ *o bien* $\overrightarrow{g}=(y_0,...,y_n)$.

Se denota con V_h al conjunto de todas las funciones reticulares definidas en $\tau(n)$. Con las operaciones habituales de funciones: adición y producto por escalares, V_h es un espacio vectorial de dimensión $n+1$.

Se denota con $V_0=\{g\in V_h \mid g(0)=g(L)=0\}$.

Teorema 1 *Las siguientes son normas en* V_0.

i) $\|g\|_\infty=\underset{i=1,...,n-1}{\text{Max}}|g(x_i)|$

ii) $\|g\|_2=\left[\sum_{i=1}^{n-1}hg^2(x)\right]^{\frac{1}{2}}$.

iii) $\|g\|_{1,2}=\left[\sum_{i=0}^{n-1}h\left(\frac{g(x_{i+1})-g(x_i)}{h}\right)^2\right]^{\frac{1}{2}}$.

Demostración. Son inmediatas (véase el apéndice, espacios normados). ■

Nota: para todo $f,\ g\in V_0$,

$$\begin{aligned}\langle f,g\rangle &= \sum_{i=1}^{n-1}hf(x_i)g(x_i),\\ \langle f,g\rangle_{1,2} &= \sum_{i=0}^{n-1}\frac{[f(x_{i+1})-f(x_i)][g(x_{i+1})-g(x_i)]}{h}\end{aligned}$$

son productos escalares en V_0. Además,

$$\begin{aligned}\|f\|_2 &= \langle f,f\rangle^{\frac{1}{2}}.\\ \|f\|_{1,2} &= \langle f,f\rangle_{1,2}^{\frac{1}{2}}.\\ |\langle f,g\rangle| &\le \|f\|_2\|g\|_2 \quad \forall f,\ g\in V_0,\\ |\langle f,g\rangle| &\le \|f\|_{1,2}\|g\|_{1,2} \quad \forall f,\ g\in V_0\end{aligned}$$

Teorema 2 *Para todo* $f\in V_0$, *se verifica que*

$$L^{-\frac{1}{2}}\|f\|_2\le\|f\|_\infty\le L^{\frac{1}{2}}\|f\|_{1,2}.$$

Demostración.

i. Probemos primeramente que $\|f\|_\infty\le L^{\frac{1}{2}}\|f\|_{1,2}\quad\forall f\in V_0$. En efecto, $y_i=\sum_{j=0}^{i-1}(y_{j+1}-y_j)$ ya que $y_0=0,\ \ i=1,...,n$. Luego, aplicando la desigualdad de Cauchy-Schwarz, se tiene

$$\begin{aligned}|y_i| &= \sum_{j=0}^{i-1}|(y_{j+1}-y_j)|\le\left(\sum_{j=0}^{i-1}1^2\right)^{\frac{1}{2}}\left(\sum_{j=0}^{i-1}(y_{j+1}-y_j)^2\right)^{\frac{1}{2}}\\ &\le n^{\frac{1}{2}}\left(\sum_{j=0}^{n-1}(y_{j+1}-y_j)^2\right)^{\frac{1}{2}}=\left[\sum_{j=0}^{n-1}n(y_{j+1}-y_j)^2\right]^{\frac{1}{2}}.\end{aligned}$$

Como $h=\frac{L}{h}\Rightarrow n=\frac{L}{h}$. Obtenemos

$$|y_i|\le\left(\sum_{j=0}^{n-1}\frac{L}{h}(y_{j+1}-y_j)^2\right)^{\frac{1}{2}}=L^{\frac{1}{2}}\left(\sum_{j=0}^{n-1}h\left(\frac{y_{j+1}-y_j}{h}\right)^2\right)^{\frac{1}{2}}=L^{\frac{1}{2}}\|f\|_{1,2}\quad i=1,...,n-1.$$

En consecuencia $\|f\|_\infty \leq L^{\frac{1}{2}} \|f\|_{1,2}$.

ii. Probemos que $\|f\|_2 \leq L^{\frac{1}{2}} \|f\|_\infty$.

Puesto que

$$\|f\|_2^2 = \sum_{i=1}^{n-1} h y_i^2 = h \sum_{i=1}^{n-1} y_i^2 \leq h \sum_{i=1}^{n-1} \|f\|_\infty^2 = hn \|f\|_\infty^2 = L \|f\|_\infty^2 .$$

De la no negatividad de la norma, se sigue que $\|f\|_2 \leq L^{\frac{1}{2}} \|f\|_\infty$.

De i) y ii) se deduce

$$L^{-\frac{1}{2}} \|f\|_2 \leq \|f\|_\infty \leq L^{\frac{1}{2}} \|f\|_{1,2} \quad \forall f \in V_0.$$

■

Observación. Puesto que $|y_i| \leq L^{\frac{1}{2}} \|f\|_{1,2} \quad i = 1, ..., n$, se sigue que

$$\|f\|_2^2 = \sum_{i=1}^{n-1} h y_i^2 \leq \sum_{i=1}^{n-1} hL \|f\|_{1,2}^2 = nhL \|f\|_{1,2}^2 = L^2 \|f\|_{1,2}^2 ,$$

así,

$$\|f\|_2^2 \leq L^2 \|f\|_{1,2}^2$$

de donde

$$\|f\|_2 \leq L \|f\|_{1,2} .$$

que es la análoga a la desigualdad de Poincaré: $f \in H_0^1(o, L)$, $\exists c > 0$ tal que $\|f\|_{L^2(o,L)} \leq c \|f'\|_{L^2(0,L)}$.

Definición de consistencia

Sea $\overrightarrow{U}_h = (u(x_0), ..., u(x_n))^T$ el vector constituido por la solución exacta en los puntos $x_i, \quad i = 0, ..., n$ de la malla $\tau(n)$. Definimos

$$\begin{aligned} \overrightarrow{e}_h &= \overrightarrow{U}_h - \overrightarrow{u}_h \quad \text{el error sobre la solución numérica,} \\ \overrightarrow{r}_h &= A_h \overrightarrow{U}_h - \overrightarrow{b} \quad \text{error de consistencia.} \end{aligned}$$

Definición 6 *1. Diremos que (1) y(3) son consistentes, para una norma $\|\cdot\|$ de V_0, si $\lim_{h \to 0} \|\overrightarrow{r}_h\| = 0$.*

Definición 7 *2. Diremos que (1) y (3) tienen una consistencia de orden $m > 0$, si existe una constante $c_1 > 0$ independiente de h, tal que*

$$\|\overrightarrow{r}_h\| \leq c_1 h^m \quad \forall h > 0.$$

La consistencia es necesaria, pero no suficiente para que el sistema discreto (esquema numérico (2)) sea convergente.

Sea $\mathcal{L} : C^2([0, L]) \to C^0([0, L])$ el operador diferencial definido por $\mathcal{L}u = -u'' + qu$.

Para $x \in [0, L]$, escribiremos $\mathcal{L}u(x) = -u''(x) + q(x) u(x)$.

Denotamos con $L_\pi : C^0(]0, L[) \to \mathbb{R}$ el operador definido por:

$$L_\pi u(x) = -\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + q(x) u(x) ,$$

donde $h > 0, \quad x - h,\ x,\ x + h \in [0, L]$.

Ponemos $L_\pi u_i = f(x_i), \quad i = 1, ..., n-1$. Observe que

$$L_\pi u(x_i) = -\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i+1})}{h^2} + q(x_i)\, u(x_i) \quad i = 1, ..., n-1.$$

La consistencia establece que el operador L_π aproxima al operador diferencial $\mathcal{L}$ cuando $h \to 0$. Más precisamente, tenemos el siguiente teorema.

Teorema 3 *Supongamos que u, solución de (1), es de clase* $C^4([0, L])$. *Entonces*

$$\begin{aligned} \|\overrightarrow{r}_h\|_\infty &\leq \frac{h^2}{12} \left\|u^{(iv)}\right\|_{L^\infty(0,L)}, \\ \|L_\pi u_h - \mathcal{L}u\|_\infty &\leq \frac{h^2}{12} \left\|u^{(iv)}\right\|_{L^\infty(0,L)}. \end{aligned}$$

Demostración. Esto es, el esquema numérico (2) es consistente.

De la definición de $\overrightarrow{r}_h$, se sigue que

$$\begin{aligned} \overrightarrow{r}_h &= A_h \overrightarrow{U}_h - \overrightarrow{b} \Leftrightarrow \\ r_{ih} &= -\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} + q(x_i)\, u(x_i) - f(x_i). \end{aligned}$$

Por otro lado,

$$-u''(x) + q(x)\, u(x) = f(x) \qquad x \in]0, L[,$$

entonces

$$0 = -u''(x_i) + q(x_i)\, u(x_i) - f(x_i), \quad i = 1, ..., n-1,$$

de donde

$$r_{ih} = u''(x_i) - \frac{1}{h^2}\left(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\right).$$

Además,

$$\begin{aligned} u(x_{i+1}) &= u(x_i) + hu'(x_i) + \frac{h^2}{2!}u''(x_i) + \frac{h^3}{3!}u'''(x_i) + \frac{h^4}{4!}u^{iv}(\Im_1), \\ u(x_{i-1}) &= u(x_i) - hu'(x_i) + \frac{h^2}{2!}u''(x_i) - \frac{h^3}{3!}u'''(x_i) + \frac{h^4}{4!}u^{iv}(\Im_2), \end{aligned}$$

con $\Im_1 \in [x_i, x_{i+1}]$, $\Im_2 \in [x_{i-1}, x_i]$, $i = 1, ..., n-1$.

Entonces

$$\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = u''(x_i) + \frac{h^2}{4!}\left(u^{iv}(\Im_1) + u^{iv}(\Im_2)\right),$$

con lo cual

$$\begin{aligned} r_{ih} &= -\frac{h^2}{4!}\left(u^{iv}(\xi_1) + u^{iv}(\xi_2)\right) \\ |r_{ih}| &\leq \frac{h^2}{12}\left\|u^{iv}\right\|_{L^\infty(0,L)}, \quad i = 1, ..., n-1 \end{aligned}$$

de donde

$$\|\overrightarrow{r}_h\|_\infty = \max_{i=1,...,n-1} |r_{ih}| \leq \frac{h^2}{12}\left\|u^{iv}\right\|_{L^\infty(0,L)}.$$

Puesto que $\mathcal{L}u = f$ y $L_\pi u(x_i) = f(x_i) \quad i = 1, ..., n-1$, se sigue que

$$L_\pi u(x_i) - \mathcal{L}u(x_i) = u''(x_i) - \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = r_{ih} \quad i = 1, ..., n-1.$$

Luego

$$\left\|L_\pi \overrightarrow{U}_h - \mathcal{L}u\right\|_\infty \underset{h \to 0}{\longrightarrow} 0.$$

■

Nótese que $\|\overrightarrow{r}_h\|_\infty = \left\|L_\pi \overrightarrow{U}_h - \mathcal{L}u\right\|_\infty \leq \frac{h^2}{12} \underset{x\in[0,L]}{\text{Max}} \left|u^{iv}(x)\right|$, que muestra que el esquema numérico es de orden 2.

Definición 8 *Sean $\|\cdot\|_{h,1}$, $\|\cdot\|_{h,2}$ dos normas en V_0. Diremos que el esquema numérico (2) es estable con respecto de las normas $\|\cdot\|_{h,1}$, $\|\cdot\|_{h,2}$, si existe una constante $C_2 > 0$ independiente de h, tal que*

$$\left\|A_h^{-1}\overrightarrow{u}\right\|_{h,1} \leq C_2 \|\overrightarrow{u}\|_{h,2} \quad \forall \overrightarrow{u} \in V_0, \ \forall h > 0.$$

Definición 9 *Sea $\|\cdot\|$ una norma en V_0.*

i. Se dice que el esquema numérico (2) es convergente con respecto de $\|\cdot\|$ si $\lim_{h\to 0} \|\overrightarrow{e}_h\| = 0$.

ii. Se dice que el esquema numérico tiene un orden de convergencia $p > 0$, si existe una constante $C_3 > 0$ independiente de h tal que $\|\overrightarrow{e}_h\| \leq C_3 h^p \quad \forall h > 0$.

Teorema 4 *El sistema de ecuaciones $A_h \overrightarrow{u}_h = \overrightarrow{b}$ tiene una única solución u_h.*

Demostración. Probemos que A_h es invertible. Para el efecto, mostremos que A_h es definida positiva, esto es,

$$\overrightarrow{u}^T A_h \overrightarrow{u} > 0 \quad \forall \overrightarrow{u} \in \mathbb{R}^{n-1} \quad \text{con} \quad \overrightarrow{u} \neq 0.$$

Sea $\overrightarrow{u} = (u_1, ..., u_n) \in \mathbb{R}^{n-1}$, $u_0 = u_n = 0$. Entonces

$$\begin{aligned}
\overrightarrow{u}^T A_h \overrightarrow{u} &= (u_1, ..., u_n) \begin{bmatrix} -\frac{u_0}{h^2} + \left(\frac{2}{h^2} + q(x_1)\right) u_1 - \frac{u_2}{h^2} \\ \vdots \\ -\frac{u_{i-1}}{h^2} \left(\frac{2}{h^2} + q(x_i)\right) u_i - \frac{u_{i+1}}{h^2} \\ \vdots \\ -\frac{u_{n-2}}{h^2} + \left(\frac{2}{h^2} + q(x_{n-1})\right) u_{n-1} - \frac{u_n}{h^2} \end{bmatrix} \\
&= \sum_{i=1}^{n-1} u_i \left(-\frac{u_{i-1}}{h^2}\left(\frac{2}{h^2} + q(x_i)\right) u_i - \frac{u_{i+1}}{h^2}\right) \\
&= \sum_{i=1}^{n-1} \frac{1}{h^2} u_i \left(-u_{i-1} + 2u_i - u_{i+1}\right) + \sum_{i=1}^{n-1} q(x_i) u_i^2 \\
&= \frac{1}{h_2}\left(\sum_{i=1}^{n-1} u_i (u_i - u_{i-1}) - \sum_{i=1}^{n-1} u_i (u_{i+1} - u_i)\right) + \sum_{i=1}^{n-1} q(x_i) u_i^2 \\
&= \frac{1}{h^2}\left(\sum_{i=0}^{n-2} u_{i+1} (u_{i+1} - u_i) - \sum_{i=1}^{n-1} u_i (u_{i+1} - u_i)\right) + \sum_{i=1}^{n-1} q(x_i) u_i^2.
\end{aligned}$$

Tomando en cuenta que $u_0 = 0$ u $u_n = 0$, se sigue que

$$\begin{aligned}
\overrightarrow{u}^T A_h \overrightarrow{u} &= \frac{1}{h^2}\left(\sum_{i=0}^{n-1} u_{i+1} (u_{i+1} - u_i) - \sum_{i=0}^{n-1} u_i (u_{i+1} - u_i)\right) + \sum_{i=1}^{n-1} q(x_i) u_i^2 \\
&= \frac{1}{h^2} \sum_{i=0}^{n-1} (u_{i+1} - u_i)^2 + \sum_{i=1}^{n-1} q(x_i) u_i^2 \\
&= \frac{1}{h} \sum_{i=0}^{n-1} h \left(\frac{u_{i+1} - u_i}{h}\right)^2 + \sum_{i=1}^{n-1} q(x_i) u_i^2 \\
&\geq \frac{1}{h} \|\overrightarrow{u}\|_{1,2}^2.
\end{aligned}$$

Luego $\overrightarrow{u}^T A_h \overrightarrow{u} \geq \frac{1}{h} \|\overrightarrow{u}\|_{1,2}^2 \quad \forall \overrightarrow{u} \in \mathbb{R}^n$ con $\overrightarrow{u} \neq 0$

Si $\ker(A_h) \neq 0\{0\}$, existe $\widetilde{u} \in \mathbb{R}^{n-1}$ con $u_0 = u_n = 0$ tal que $\widetilde{u} \neq 0$ y $A_h \widetilde{u} = 0$. Resulta que

$$0 \geq \frac{1}{h} \|\widetilde{u}\|_{1,2}^2 \Rightarrow \widetilde{u} = 0$$

en otra contradicción con lo supuesto. En consecuencia $\ker(A_h) = \{0\}$, que muestra que A_h es invertible. Por lo tanto, $\forall \overrightarrow{b} \in \mathbb{R}^{n-1}$, el sistema de ecuaciones $A_h \overrightarrow{u} = \overrightarrow{b}$ tiene solución única. ■

Observación: La estabilidad del esquema numérico (2) significa que pequeños errores en los datos de entrada producen pequeños errores en los datos de salida, esto a su vez que existe una constante $c > 0$ tal que $\left\|A_h^{-1}\right\| \leq c \quad \forall h > 0$. Note que $\lim_{h \to 0} \left\|A_h^{-1}\right\| \leq c$.

La consistencia y la estabilidad son dos nociones independientes.

Teorema 5 *Para todo $\overrightarrow{u} \in V_0$, $\left\|A_h^{-1}\overrightarrow{u}\right\|_{1,2} \leq L \|\overrightarrow{u}\|_2$; esto es, se tiene estabilidad con respecto de las normas $\|\cdot\|_{1,2}$ y $\|\cdot\|_2$.*

Demostración. Probaremos el teorema en dos etapas.

i. Probemos que para todo $\overrightarrow{u} \in V_0$,

$$h \overrightarrow{u}^T B_h \overrightarrow{u} = \|\overrightarrow{u}\|_{1,2}^2,$$

donde $B_h = (b_{ij}(h))$ denota la matriz de $n-1 \times n-1$ definida por

$$\begin{aligned} b_{ii}(h) &= \frac{2}{h^2} \quad i = 1, ..., n-1, \\ b_{ii-1}(h) &= -\frac{1}{h^2} = b_{i-1i}(h) \quad i = 2, ..., n-1. \end{aligned}$$

Procediendo de manera similar al teorema precedente, tenemos

$$h \overrightarrow{u}^T B_h \overrightarrow{u} = \frac{1}{h} \sum_{i=1}^{n-1} u_i (-u_{i-1} + 2u_i - u_{i+1}) = \|\overrightarrow{u}\|_{1,2}^2.$$

ii. Probemos que $\left\|A_h^{-1}\overrightarrow{u}\right\|_{1,2} \leq c \|\overrightarrow{u}\|_{1,2}$.

Sea $\overrightarrow{v} = A_h^{-1}\overrightarrow{u}$ entonces $\overrightarrow{u} = A_h \overrightarrow{v}$. Ponemos $A_h = B_h + Q_h$ con B_h definida en i) precedente y $Q_h = diag(q(x_1), ..., q(x_{n-1}))$.

Multiplicando por $\overrightarrow{v}^T$, se tiene

$$\overrightarrow{v}^T \overrightarrow{u} = \overrightarrow{v}^T A_h \overrightarrow{v} = \overrightarrow{v}^T (B_h + Q_h) \overrightarrow{v} = \overrightarrow{v}^T B_h \overrightarrow{v} + \overrightarrow{v}^T Q_h \overrightarrow{v}.$$

Luego

$$h \overrightarrow{v}^T \overrightarrow{u} = h \overrightarrow{v}^T B_h \overrightarrow{v} + h \overrightarrow{v}^T + Q_h \overrightarrow{v} = \|\overrightarrow{v}\|_{1,2}^2 + \sum_{i=1}^{n-1} hq(x_i) v_i^2 \geq 0$$

de donde

$$0 \leq \|\overrightarrow{v}\|_{1,2}^2 + \sum_{i=1}^{n-1} hq(x_i) v_i^2 = h \overrightarrow{v}^T \overrightarrow{u} \leq \|\overrightarrow{v}\|_2 \|\overrightarrow{u}\|_2.$$

Además,

$$\|\overrightarrow{v}\|_{1,2}^2 \leq \|\overrightarrow{v}\|_{1,2}^2 + \sum_{i=1}^{n-1} hq(x_i) v_i^2 \leq \|\overrightarrow{v}\|_2 \|\overrightarrow{u}\|_2.$$

Puesto que $\|\overrightarrow{v}\|_2 \leq L \|\overrightarrow{v}\|_{1,2}$, entonces

$$\|\overrightarrow{v}\|_{1,2}^2 \leq L \|\overrightarrow{v}\|_{1,2} \|\overrightarrow{u}\|_2,$$

de donde

$$\|\overrightarrow{v}\|_{1,2} \leq L \|\overrightarrow{u}\|_2 ,$$

o bien

$$\left\|A_h^{-1}\overrightarrow{u}\right\|_{1,2} \leq L \|\overrightarrow{u}\|_2 .$$

Estabilidad + consistencia $\Rightarrow$ convergencia. ■

Teorema 6 *El esquema numérico (2) es convergente para las normas* $\|\cdot\|_2$, $\|\cdot\|_\infty$ *y* $\|\cdot\|_{1,2}$.

Demostración. Puesto que $U_h = (u(x_1), ..., u(x_{n-1}))$, $u(0) = u(L) = 0$,y

$$\overrightarrow{r}_h = A_h\overrightarrow{U}_h - \overrightarrow{b}$$

con lo cual

$$A_h\overrightarrow{U}_h = \overrightarrow{r}_h + \overrightarrow{b}.$$

Además,

$$A_h\overrightarrow{u}_h = \overrightarrow{b}.$$

Luego, el error $\overrightarrow{e}_h = \overrightarrow{U}_h - \overrightarrow{u}_h$ satisface la ecuación

$$A_h\overrightarrow{e}_h = \overrightarrow{r}_h,$$

pues

$$A_h\left(\overrightarrow{U}_h - \overrightarrow{u}_h\right) = \overrightarrow{r}_h.$$

Se tiene $\overrightarrow{e}_h \in V_0$ y $\overrightarrow{e}_h = A_h^{-1}\overrightarrow{r}_h$.

Por el teorema precedente (estabilidad para las normas $\|\cdot\|_{1,2}$ y $\|\cdot\|_2$) y por el teorema relativo a la consistencia, se tiene

$$\|\overrightarrow{e}_h\|_{1,2} = \left\|A_h^{-1}\overrightarrow{r}_h\right\|_{1,2} \leq L \|\overrightarrow{r}_h\|_2 .$$

Puesto que

$$L^{-\frac{1}{2}} \|\overrightarrow{r}_h\|_2 \leq \|\overrightarrow{r}_h\|_\infty \leq L^{\frac{1}{2}} \|\overrightarrow{r}_h\|_{1,2} ,$$

resulta que

$$\|\overrightarrow{r}_h\|_2 \leq L^{\frac{1}{2}} \|\overrightarrow{r}_h\|_\infty \leq \frac{L^{\frac{1}{2}}}{12}h^2 \operatorname*{Max}_{x\in(0,L)} \left|u^{iv}(x)\right|$$

y en consecuencia

$$\|\overrightarrow{e}_h\|_{1,2} \leq L \|\overrightarrow{r}_h\|_2 \leq \frac{L^{\frac{3}{2}}}{12}h^2 \operatorname*{Max}_{x\in[0,L]} \left|u^{iv}(x)\right| . \qquad (*)$$

Por lo tanto,

$$\|\overrightarrow{e}_h\|_{1,2} \to_{h\to 0} 0.$$

Por otro lado,

$$L^{-\frac{1}{2}} \|\overrightarrow{e}_h\|_\infty \leq \|\overrightarrow{e}_h\|_{1,2} \leq \frac{L^{\frac{3}{2}}}{12}h^2 \operatorname*{Max}_{x\in[0,L]} \left|u^{iv}(x)\right| , \qquad (**)$$

$$\|\overrightarrow{e}_h\|_\infty \leq \frac{L^2}{12}h^2 \operatorname*{Max}_{x\in[0,L]} \left|u^{iv}(x)\right| \to_{h\to 0} 0. \qquad (11.1)$$

Finalmente,

$$L^{-1} \|\overrightarrow{e}_h\|_2 \leq \|\overrightarrow{e}_h\|_{1,2} \leq \frac{L^{\frac{3}{2}}}{12}h^2 \operatorname*{Max}_{x\in[0,L]} \left|u^{iv}(x)\right| ,$$

$$\|\overrightarrow{e}_h\|_2 \leq \frac{L^{\frac{5}{2}}}{12}h^2 \operatorname*{Max}_{x\in[0,L]} \left|u^{iv}(x)\right| \to_{h\to 0} 0. \qquad (***)$$

■

Observe que

$$L^{-\frac{1}{2}} \left\| \overrightarrow{e}_h \right\|_2 \leq \left\| \overrightarrow{e}_h \right\|_\infty \leq L^{-\frac{1}{2}} \left\| \overrightarrow{e}_h \right\|_{1,2} \leq ch^2,$$

con $c = \frac{c'}{12} \underset{x\in[0,L]}{\text{Max}} \left|u^{iv}(x)\right|$ y $c' = L^{\frac{3}{2}}$, L^2, $L^{\frac{5}{2}}$ de acuerdo a (*), (**) y (***) respectivamente.

Adicionalmente, el teorema muestra que el esquema numérico (2) tiene un orden de convergencia 2 para las normas $\|\cdot\|_\infty$, $\|\cdot\|_2$ y $\|\cdot\|_{1,2}$.

11.4.3. Orden de convergencia

1. **Ecuación diferencial con condiciones de frontera de Dirichlet no homogéneas.**

   Consideramos el problema siguiente:

$$\text{hallar } u \in C^2([0,L]) \text{ solución de } \begin{cases} -u'' + qu = f \text{ sobre }]0,L[, \\ u(0) = a, \quad u(L) = b. \end{cases} \tag{P.}$$

   Sean $n \in \mathbb{Z}^+$, $h = \frac{L}{n}$, $\tau(h) = \{x_j = jh \mid j = 0,1,...,n\}$ la malla de $[0,L]$.

   El esquema numérico del problema (P) es el siguiente:

$$\begin{cases} u_0 = a, \quad u_n = b, \\ -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + q(x_j)\, u_j = f(x_j) \quad j = 1,...,n-1. \end{cases} \tag{P$_1$.}$$

   El esquema numérico (P_1) en forma explícita se escribe: para $j = 1$, $u_0 = a$, y

$$-\frac{u_2 - 2u_1 + a}{h^2} + q(x_1)\, u_1 = f(x_1)$$

   con lo cual

$$\left[\frac{2}{h^2} + q(x_1)\right] u_1 - \frac{1}{h^2} u_2 = f(x_1) + \frac{a}{h^2}.$$

   Para $j = n-1$, $u_n = b$, luego

$$-\frac{b - 2u_{n-1} + u_{n-2}}{h^2} + q(x_{n-1})\, u_{n-1} = f(x_{n-1})$$

   de donde

$$-\frac{1}{h^2} u_{n-2} + \left[\frac{2}{h^2} + q(x_{n-1})\right] u_{n-1} = f(x_{n-1}) + \frac{b}{h^2}.$$

   El esquema numérico (P_1) es:

$$\begin{cases} \left[\frac{2}{h^2} + q(x_1)\right] u_1 - \frac{1}{h^2} u_2 = f(x_1) + \frac{a}{h^2} \\ -\frac{1}{h^2} u_{j-1} + \left[\frac{2}{h^2} + q(x_j)\right] u_j - \frac{1}{h^2} u_{j+1} = f(x_j) \quad j = 2,...,n-2 \\ -\frac{1}{h^2} u_{n-2} + \left[\frac{2}{h^2} + q(x_{n-1})\right] u_{n-1} = f(x_{n-1}) + \frac{b}{h^2}. \end{cases} \tag{P$_2$.}$$

   Sea $w : [0,L] \Rightarrow \mathbb{R}$ la función definida por

$$w(x) = \frac{x}{L} b + \left(1 - \frac{x}{L}\right) a.$$

   Entonces $w(0) = a$, $w(L) = b$. Se define $\widetilde{u} = u - w$. Se tiene

$$\begin{cases} \widetilde{u}(0) = u(0) - w(0) = 0, \\ \widetilde{u}(L) = u(L) - w(L) = 0. \end{cases}$$

   La función $\widetilde{u}$ satisface las condiciones de frontera de Dirichlet homogénea. La ecuación lineal del problema (P) se escribe

$$-\widetilde{u}'' + q(\widetilde{u} + w) = f$$

de donde

$$-\widetilde{u}'' + q\widetilde{u} = f - qw.$$

En consecuencia

$$\left\{ \begin{array}{c} -\widetilde{u}'' + q\widetilde{u} = g \quad \text{sobre }]0, L[\,, \\ \widetilde{u}(0) = \widetilde{u}(L) = 0 \end{array} \right. \tag{$\widetilde{\mathrm{P}}$.}$$

con $g = f - qw$.

Note que $g(x_j) = f(x_j) - (qw)(x_j) = f(x_j) - q(x_j)\left[\frac{x_j}{L}b + \left(1 - \frac{x_j}{L}\right)a\right]$.

Desde el punto de vista informático, si se utiliza $\left(\widetilde{\mathrm{P}}\right)$, el esquema numérico es idéntico al establecido en la sección 3, el sistema de ecuaciones $A_h \overrightarrow{u}_h = \overrightarrow{b}$ es similar al establecido en esa sección a condición de remplazar $\overrightarrow{b} = (f(x_1), ..., f(x_n))^T$ por el vector $\overrightarrow{b} = (g(x_1), ..., g(x_n))^T$.

Si se utiliza el esquema numérico (P_2), la matriz A_h es la misma que la de la sección 3, el vector $\overrightarrow{b}$ se remplaza por el vector $\left(f(x_1) + \frac{a}{h^2}, f(x_2), ..., f(x_{n-2}), f(x_1) + \frac{b}{h^2}\right)$.

La estabilidad, consistencia y convergencia para el problema $\left(\widetilde{\mathrm{P}}\right)$ ha sido discutida en la sección 5. Por lo tanto, el esquema numérico (P_2) es convergente, con orden de convergencia igual a 2.

2. **Ecuación diferencial con condiciones de frontera de Neumann.**

Consideramos el problema siguiente:

$$\text{hallar } u \in C^2([0, L]) \text{ solución de } \left\{ \begin{array}{c} -u'' + qu = f \quad \text{sobre }]0, L[\,, \\ u'(0) = a, \quad u'(L) = b. \end{array} \right. \tag{P_n.}$$

Sea w la función real definifa por:

$$w(x) = \frac{x^2}{2L}b - \frac{(L-x)^2}{2L}a, \quad x \in [0, L].$$

Entonces

$$\begin{array}{rcl} w'(x) & = & \dfrac{x}{L}b + \left(1 - \dfrac{x}{L}\right)a, \\ w''(x) & = & \dfrac{b}{L} - \dfrac{a}{L} = \dfrac{b-a}{L}, \\ w'(0) & = & a, \quad w'(L) = b. \end{array}$$

Se define $\widetilde{u} = u - w$. Entonces

$$\begin{array}{rcl} \widetilde{u}'(x) & = & u'(x) - w'(x) \quad x \in [0, L], \\ \widetilde{u}'(0) & = & 0, \quad \widetilde{u}'(L) = 0, \\ \widetilde{u}''(x) & = & u''(x) - w''(x) = u''(x) - \dfrac{b-a}{L}. \end{array}$$

Consecuentemente, el problema (P_n) se escribe

$$-\left(\widetilde{u}''(x) + \frac{b-a}{L}\right) + q(x)\left(\widetilde{u}(x) + w(x) = f(x)\right) \quad x \in]0, L[\,,$$

con lo cual

$$\left\{ \begin{array}{c} -\widetilde{u}'' + q\widetilde{u} = f - \frac{b-a}{L} - qw \quad \text{sobre }]0, L[\,, \\ \widetilde{u}'(0) = \widetilde{u}'(L) = 0. \end{array} \right. \tag{$\widetilde{\mathrm{P}}_n$}$$

El resultado precedente muestra que debemos estudiar la ecuación diferencial con las condiciones de Neumann homogéneas; eso es, consideramos el problema (P_0) siguiente:

$$\left\{ \begin{array}{c} -u'' + qu = f \quad \text{sobre }]0, L[\,, \\ u'(0) = u'(L) = 0. \end{array} \right. \tag{P_0}$$

Para construir un esquema numérico que aproxime la solución de (P_0), comenzamos con la aproximación de $u'(0)$ y $u'(L)$.

Aproximación de $u'(0)$ y $u''(0)$

Sea $n \in \mathbb{Z}^+$, $h = \frac{L}{h}$ y $\tau(n) = \{x_j = jh \mid j = 0, 1, ..., n\}$. Entonces

$$u''(0) \simeq \frac{u'(\widetilde{x}_1) - u'(0)}{\frac{h}{2}} = \frac{2}{h}\left(u'(\widetilde{x}_1) - a\right),$$

donde $\widetilde{x}_1$ es el punto medio del intervalo $[0, x_1]$. Para la aproximación de $u'(\widetilde{x}_1)$ utilizamos las diferencias finitas centrales. Tenemos

$$u'(\widetilde{x}_1) \simeq \frac{u(x_1) - u(0)}{h},$$

luego

$$u''(\widetilde{x}_1) \simeq \frac{2}{h}\left[\frac{u(x_1) - u(0)}{h} - a\right] = \frac{2}{h^2}\left[u(x_1) - u(0) - ah\right].$$

Utilizando el polinomio de Taylor, tenemos

$$u(x_1) = u(0) + hu'(0) + \frac{h^2}{2!}u''(0) + \frac{h^3}{3!}u'''(\xi) \text{ con } \xi \in [0, x_1],$$

con lo cual

$$\frac{2}{h^2}\left[u(x_1) - u(0) - ah\right] = u''(0) + \frac{h}{3}u'''(\xi).$$

Aproximación de $u'(L)$ y $u''(L)$.

Sea $\widetilde{x}_n$ el punto medio del intervalo $[x_{n-1}, L]$. Procediendo de modo similar al caso $u'(L)$, se deduce que

$$\begin{aligned} u''(L) &\simeq \frac{u'(L) - u'(\widetilde{x}_{n-1})}{\frac{h}{2}} = \frac{2}{h}\left[b - u'(\widetilde{x}_{n-1})\right] \\ &\simeq \frac{2}{h}\left[b - \frac{u(L) - u(x_{n-1})}{h}\right] = \frac{2}{h^2}\left(-u(L) + u(x_{n-1}) + bh\right). \end{aligned}$$

Por el desarrollo de Taylor, se tiene

$$u(x_{n-1}) = u(L) - hu'(L) + \frac{h^2}{2!}u''(L) - \frac{h^3}{3!}u'''(\xi_n) \quad con \;\; \xi_n \in [x_{n-1}, L],$$

entonces

$$\frac{2}{h^2}\left(-u(L) + u(x_{n-1}) + bh\right) = u''(L) - \frac{h}{3}u'''(\xi_n).$$

Definimos $\mathcal{L}u = -u'' + qu$ y suponemos $u \in C^4([0, L])$. Entonces

$$\begin{aligned} \mathcal{L}u(x) &= -u''(x) + q(x)u(x) \quad x \in [0, L], \\ \mathcal{L}_\pi u(0) &= -\frac{2}{h^2}\left(u(x_1) - u(0) - ah\right) + q(0)u(0), \\ \mathcal{L}_\pi u(L) &= -\frac{2}{h^2}\left(-u(L) + u(x_{n-1}) + bh\right) + q(L)u(L). \end{aligned}$$

Se tiene

$$\begin{aligned} |\mathcal{L}u(0) - \mathcal{L}_\pi(0)| &= \left|-u''(0) + q(0)u(0) + \frac{2}{h^2}\left(u(x_1) - u(0) - ah\right) - q(0)u(0)\right| \\ &= \frac{h}{3}\left|u'''(\xi_1)\right| \quad \text{con} \quad \xi_1 \in [0, x_1], \\ |\mathcal{L}u(L) - \mathcal{L}_\pi u(L)| &= \left|-\frac{h}{3}u'''(\xi_n)\right| = \frac{h}{3}\left|u'''(\xi_n)\right| \quad \text{con} \quad \xi_n \in [x_{n-1}, L]. \end{aligned}$$

Para $j = 0, ..., n$, el esquema numérico tiene la forma siguiente:

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + q(x_j)\, u_j = f(x_j)$$

Luego

$$\left\{ \begin{array}{c} \left(\frac{2}{h^2} + q(0)\right) u(0) - \frac{2}{h^2} u_1 = f(0) - \frac{2a}{h}, \\ -\frac{1}{h^2} u_{j-1} + \left(\frac{2}{h^2} + q(x_j)\right) u_j - \frac{1}{h} u_{j+1} = f(x_j), \\ -\frac{2}{h^2} u_{n-1} + \left(\frac{2}{h^2} + q(L)\right) u_n = f(L) + \frac{2b}{h}. \end{array} \right. \tag{P$_1$}$$

Poniendo $\mathcal{L}_\pi u(x_j) = -\dfrac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + q(x_j)\, u(x_j)$, se tiene

$$|\mathcal{L}u(x_j) - \mathcal{L}_\pi u(x_j)| \leq \frac{h^2}{12} \underset{x \in [0,L]}{\text{Max}} \left| u^{(iv)}(x) \right| \quad j = 1, ..., n-1.$$

Como

$$\begin{aligned} |\mathcal{L}u(0) - \mathcal{L}_\pi u(0)| &\leq \frac{h}{3} \underset{x \in [0,L]}{\text{Max}} \left| u'''(x) \right|, \\ |\mathcal{L}u(L) - \mathcal{L}_\pi u(L)| &\leq \frac{h}{3} \underset{x [0,L]}{\text{Max}} \left| u'''(x) \right|, \end{aligned}$$

se sigue

$$\|\overrightarrow{r}_h\| \leq \frac{h}{3} \text{Max} \left\{ \left\| u''' \right\|_{L^\infty(o,L)} \left\| u^{(iv)} \right\|_{L^\infty(0,L)} \right\},$$

que muestra que el esquema numérico es de orden 1.

Se puede observar que la matriz $A_h = (a_{ij}(h))$ del esquema numérico (P$_1$) está definida por

$$\begin{aligned} a_{ii}(h) &= -\frac{2}{h^2} + q(x_i), \quad i = 0, 1, ..., n, \\ a_{12}(h) &= -\frac{2}{h^2}, \;\; a_{nn-1}(h) = -\frac{2}{h^2}, \\ a_{i-1i}(h) &= a_{ii-1}(h) = -\frac{1}{h^2} \quad i = 2, ..., n-1. \end{aligned}$$

La matriz A_h no es simétrica. Además,

$$\left\{ \begin{array}{c} a_{ii}(h) \geq -a_{i-1i}(h) - a_{ii-1}(h), \quad i = 1, ..., n-1, \\ a_{11}(h) \geq -a_{12}(h), \\ a_{nn}(h) \geq -a_{nn-1}(h). \end{array} \right.$$

Si $q \neq 0$, existe $j = 0, 1, ..., n$ tal que $q(x_j) > 0$ con lo cual

$$a_{jj}(h) > -a_{j-1j}(h) - a_{jj-1}(h)$$

que prueba que la matriz A_h es diagonalmente dominante, y en consecuencia A_h es invertible.

Conclusión: El esquema numérico (P$_1$) es convergente con orden de convergencia igual a 1.

11.4.4. Método de diferencias finitas en mallas no uniformes

Posición del problema

Sea $L > 0$. Consideramos el problema (P) siguiente:

$$\text{hallar } u : [0, L] \to \mathbb{R} \text{ solución de } \left\{ \begin{array}{c} -\frac{d}{dx}\left(k(x) \frac{du}{dx}(x)\right) + q(x)\, u(x) = f(x) \quad x \in]0, L[, \\ u(0) = 0, \quad u(L) = 0, \end{array} \right. \tag{P}$$

donde$k,\ q,\ f \in C^0([0,L])$ tales que

$$\begin{aligned} k(x) &\geq \alpha > 0 \quad \forall x\,[0,L], \\ q(x) &\geq 0 \quad \forall x \in [0,L]. \end{aligned}$$

El término $-k\frac{du}{dx}$ se interpreta como el flujo.

Sean $n \in \mathbb{Z}^+$ y $\tau(n) = \{x_0 = 0 < x_1 < ... < x_{n-1} < x_n = L\}$ la malla de $[0,L]$ no necesariamente uniforme. Ponemos $h_j = x_j - x_{j-1} \quad j = 1,...,n$; y $h = \underset{j=1,...,n}{\text{Max}}\ h_j$.

Deseamos aproximar la solución de u de (P) en los nodos $x_1,...,x_{n-1}$.

Discretizando de (P)

Sean $a,\ b \in [0,L]$ tales que $a < b$. Entonces

$$\begin{aligned} -\int_a^b \frac{d}{dx}\left(k(x)\frac{du}{dx}\right)dx + \int_a^b q(x)\,u(x)\,dx &= \int_a^b f(x)\,dx \\ -k(x)\left.\frac{du}{dx}\right|_a^b + \int_a^b q(x)\,u(x)\,dx &= \int_a^b f(x)\,dx. \end{aligned}$$

Para $a = x_i - \dfrac{h_i}{2},\ \ b = x_i + \dfrac{h_{i+1}}{2}$ (a y b son los puntos medios de los intervalos $[x_{i-1},x_i]$ y $[x_i,x_{i+1}]$ respectivamente) se tiene

$$\begin{aligned} &-\left[k\left(x_i+\frac{h_{i+1}}{2}\right)\frac{du}{dx}\left(x_i+\frac{h_{i+1}}{2}\right) - k\left(x_i-\frac{h_i}{2}\right)\frac{du}{dx}\left(x_i-\frac{h_i}{2}\right)\right] + \int_{x_i-\frac{h_i}{2}}^{x_i+\frac{h_{i+1}}{2}} q(x)\,u(x)\,dx \\ = &\int_{x_i-\frac{h_i}{2}}^{x_i+\frac{h_{i+1}}{2}} f(x)\,dx \end{aligned}$$

Las derivadas $\frac{du}{dx}\left(x_i+\frac{h_{i+1}}{2}\right)$ y $\frac{du}{dx}\left(x_i-\frac{h_i}{2}\right)$ se aproximan mediante diferencias finitas centrales

$$\begin{aligned} \frac{du}{dx}\left(x_i-\frac{h_i}{2}\right) &\simeq \frac{u(x_i)-u(x_i-h_i)}{h_i} = \frac{u(x_i)-u(x_{i-1})}{h_i}, \\ \frac{du}{dx}\left(x_i+\frac{h_{i+1}}{2}\right) &\simeq \frac{u(x_i+h_{i+1})-u(x_i)}{h_{i+1}} = \frac{u(x_{i+1})-u(x_i)}{h_{i+1}}. \end{aligned}$$

Las integrales del modo siguiente

$$\int_{x_i-\frac{h_i}{2}}^{x_i+\frac{h_{i+1}}{2}} q(x)\,u(x)\,dx \simeq \frac{h_i+h_{i+1}}{2} q(x_i)\,u(x_i), \quad i = 1,...,n-1, \tag{1}$$

$$\int_{x_i-\frac{h_i}{2}}^{x_i+\frac{h_{i+1}}{2}} f(x)\,dx \simeq \frac{h_i+h_{i+1}}{2} f(x_i) \quad i = 1,...,n-1. \tag{2}$$

Las fórmulas (1) y (2) son una variante de la fórmula del punto medio siguiente: si $g \in C^0([\alpha,\beta])$,

$$\int_\alpha^\beta g(x)\,dx \simeq (\beta-\alpha)\,g\left(\frac{\alpha+\beta}{2}\right).$$

Denotamos con u_i una aproximación de $u(x_i)$. Se establece el esquema numérico siguiente:

$$\begin{aligned} -k\left(x_i+\frac{h_{i+1}}{2}\right)\frac{u_{i+1}-u_i}{h_{i+1}} + k\left(x_i-\frac{h_i}{2}\right)\frac{u_i-u_{i-1}}{h_i} + \frac{h_i+h_{i+1}}{2} q(x_i)\,u_i &= \\ \frac{h_i+h_{i+1}}{2} f(x) \quad i &= 1,...,n-1. \end{aligned}$$

de donde

$$-\frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i}u_{i-1}+\left[\frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i}+\frac{k\left(x_i+\frac{h_{i+1}}{2}\right)}{h_{i+1}}+\frac{h_i+h_{i+1}}{2}q\left(x_i\right)\right]u_i-\frac{k\left(x_i+\frac{h_{i+1}}{2}\right)}{h_{i+1}}u_{i+1}$$
$$= \frac{h_i+h_{i+1}}{2}f\left(x_i\right) \quad i=1,...,n-1.$$

Tomando en consideración las condiciones de frontera $u\left(0\right)=u_0=0$ y $u\left(L\right)=u_n=0$, resulta:

$$\begin{cases} \left[\frac{k\left(x_1-\frac{h_1}{2}\right)}{h_1}+\frac{k\left(x_1+\frac{h_2}{2}\right)}{h_2}+\frac{1}{2}\left(h_1+h_2\right)q\left(x_1\right)\right]u_1-\frac{k\left(x_1+\frac{h_2}{2}\right)}{h_2}u_2=\frac{1}{2}\left(h_1+h_2\right)f\left(x_1\right), \\ -\frac{k\left(x_i-\frac{h_1}{2}\right)}{h_i}u_{i-1}+\left[\frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i}+\frac{k\left(x_i+\frac{h_{i+1}}{2}\right)}{h_{i+1}}+\frac{h_i+h_{i+1}}{2}q\left(x_i\right)\right]u\left(x_i\right)-\frac{k\left(x_i+\frac{h_{i+1}}{2}\right)}{h_{i+1}}u_{i+1}=\frac{1}{2}\left(h_i+h_{i+1}\right)f\left(x_i\right), \\ -\frac{k\left(x_{n-1}-\frac{h_{n-1}}{2}\right)}{h_{n-1}}u_{n-1}+\left[\frac{k\left(x_{n-1}-\frac{h_{n-1}}{2}\right)}{h_{n-1}}+\frac{k\left(x_{n-1}+\frac{h_n}{2}\right)}{h_n}+\frac{h_{n-1}+h_n}{2}q\left(x_{n-1}\right)\right]u_{n-1}=\frac{1}{2}\left(h_{n-1}+h_n\right)f\left(x_n\right) \end{cases} \tag{3}$$

Ponemos

$$\begin{aligned} a_i &= -\frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i} \quad i=2,...,n-1, \quad a_1=0, \\ b_i &= \frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i}+\frac{k\left(x_i+\frac{h_{i+1}}{2}\right)}{h_{i+1}}\frac{1}{2}\left(h_i+h_{i+1}\right)q\left(x_i\right) \quad i=1,...,n-1, \\ c_i &= -\frac{k\left(x_i+\frac{h_{i+1}}{2}\right)}{h_{i+1}}, \quad i=1,...,n-2, \quad c_{n-1}=0, \\ f_i &= \frac{1}{2}\left(h_i+h_{i+1}\right)f\left(x_i\right) \quad i=1,...,n-1 \end{aligned}$$

El esquema numérico (3) se escribe

$$a_iu_{i-1}+b_iu_i+c_iu_{i+1}=f_i \quad i=1,...,n-1. \tag{4}$$

Por otro lado, se pone $\overrightarrow{u}=\left(u_1,...,u_{n-1}\right)^T$, y se define $A_h=\left(a_{ij}\left(h\right)\right)$ con

$$\begin{aligned} a_{ii}\left(h\right) &= \frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i}+\frac{k\left(x_i+h_{i+1}\right)}{h_{i+1}}+\frac{h_i+h_{i+1}}{2}q\left(x_i\right) \quad i=1,...,n-1 \\ a_{i-i}i\left(h\right) &= a_{ii-1}\left(h\right)=-\frac{k\left(x_i-\frac{h_i}{2}\right)}{h_i} \quad i=2,...,n-1, \\ \overrightarrow{v} &= \left(b_1,...,b_{n-1}\right)^T \text{ con } b_i=\frac{h_i+h_{i+1}}{2}f\left(x_i\right). \end{aligned}$$

El esquema numérico (3) se escribe en la forma

$$A_h\overrightarrow{u}=\overrightarrow{b}. \tag{5}$$

Se verifican las condiciones siguientes:

i. $A_h=A_h^T$, es decir que A_h es simétrica.

ii. $b_i>0 \quad i=1,...,n-1.$

iii. $c_1<0, \;\; a_0<0, \;\; c_i<0 \quad i=2,...,n-2, \;\; a_{n-1}<0.$

iv. $\left\{ \begin{array}{l} b_i > -a_i - c_i \quad i = 2, ..., n-2 \\ b_1 > -c_1, \quad b_{n-1} > -a_{nn-1}. \end{array} \right.$

Se demuestra que A_h es positiva, esto es, $A_h^{-1} > 0$.

Observación

Si se considera el problema (P) siguiente:

$$\left\{ \begin{array}{c} -\frac{d}{dx}\left(k\left(x\right)\frac{du(x)}{dx}\right) + q\left(x\right)u\left(x\right) = f\left(x\right) \quad x \in \left]0, L\right[, \\ u\left(0\right) = a, \quad u\left(L\right) = b \end{array} \right. \tag{P}$$

donde a, $b \in \mathbb{R}$, k, q, f funciones que satisfacen las condiciones citadas precedentemente.

Se define $w\left(x\right) = \frac{x}{L}b + \frac{(L-x)}{L}a \quad x \in [0, L]$. Se tiene

$$w\left(0\right) = a,\, w\left(L\right) = b,\, w'\left(x\right) = \frac{b-a}{L} \quad \forall x \in [0, L]\,.$$

Se define $\widetilde{u} = u - w$. Entonces

$$\left\{ \begin{array}{c} -\frac{d}{dx}\left(k\frac{d\widetilde{u}}{dx}\right) + q\widetilde{u} = f - qw + \frac{b-a}{L}\frac{dk}{dx}. \\ \widetilde{u}\left(0\right) = 0, \quad \widetilde{u}\left(L\right) = 0 \end{array} \right. \tag{$\widetilde{\text{P}}$}$$

Note que en este caso $k \in C^1\left([0, L]\right)$.

El problema $\left(\widetilde{\text{P}}\right)$ tiene condiciones de frontera de Dirichlet homogéneas.

Condiciones de frontera mixtas

Consideramos ahora el caso en el que las condiciones de frontera son las siguientes:

$$\left\{ \begin{array}{l} \alpha_1 u'\left(0\right) + \alpha_2 u\left(0\right) = g_1 \\ \beta_1 u'\left(L\right) + \beta_2 u\left(L\right) = g_2 \end{array} \right.$$

con α_1, α_2, β_1, β_2, g_1, $g_2 \in \mathbb{R}$.

Condiciones de frontera en $x = 0$.

Tomando $a = 0$, $b = \frac{h_1}{2}$ se tiene

$$\begin{array}{rcl} -\int_0^{\frac{h_1}{2}} \frac{d}{dx}\left(k\left(x\right)\frac{du}{dx}\right)dx + \int_0^{\frac{h_1}{2}} q\left(x\right)u\left(x\right)dx & = & \int_0^{\frac{h_1}{2}} f\left(x\right)dx \\ -k\left(x\right)\frac{du}{dx}\Big|_0^{\frac{h_1}{2}} + \int_0^{\frac{h_1}{2}} q\left(x\right)u\left(x\right)dx & = & \int_0^{\frac{h_1}{2}} f\left(x\right)dx. \end{array}$$

Consideramos el término $-k\left(x\right)\frac{du}{dx}\Big|_0^{\frac{h_1}{2}}$. Tenemos

$$-k\left(x\right)\frac{du}{dx}\Big|_0^{\frac{h_1}{2}} = -k\left(\frac{h_1}{2}\right)\frac{du}{dx}\left(\frac{h_1}{2}\right) + k\left(0\right)\frac{du}{dx}\left(0\right).$$

Puesto que $\alpha_1 u'\left(0\right) + \alpha_2 u\left(0\right) = g_1$, suponemos que $\alpha_1 \neq 0$, con lo cual $u'\left(0\right) = \frac{1}{\alpha_1}\left(g_1 - \alpha_2 u\left(0\right)\right)$.

Por otro lado, la derivada $\frac{du}{dx}\left(\frac{h_1}{2}\right)$ se aproxima mediante diferencias finitas centrales. Se tiene

$$\frac{du}{dx}\left(\frac{h_1}{2}\right) \simeq \frac{u\left(x_1\right) - u\left(0\right)}{h_1},$$

en consecuencia $-k(x)\left.\frac{du}{dx}\right|_0^{\frac{h_1}{2}}$ se aproxima como

$$-k(x)\left.\frac{du}{dx}\right|_0^{\frac{h_1}{2}} \simeq -k\left(\frac{h_1}{2}\right)\frac{u(x_1)-u(0)}{h_1}+k(0)\frac{1}{\alpha_1}(g_1-\alpha_2 u(0)).$$

Además,

$$\begin{aligned}\int_0^{\frac{h_1}{2}} q(x)\,u(x)\,dx &\simeq \frac{h_1}{2}q(0)\,u(0),\\ \int_0^{\frac{h_1}{2}} f(x)\,dx &\simeq \frac{h_1}{2}f(0).\end{aligned}$$

Resulta

$$-k\left(\frac{h_1}{2}\right)\frac{u_1-u_0}{h_1}+k(0)\frac{1}{\alpha_1}(g_1-\alpha_2 u_0)+\frac{h_1}{2}q(0)\,u_0=\frac{h_1}{2}f(0),$$

o bien

$$\left[\frac{k\left(\frac{h_1}{2}\right)}{h_1}-\frac{\alpha_2}{\alpha_1}k(0)+\frac{h_1}{2}q(0)\right]u_0-\frac{k\left(\frac{h_1}{2}\right)}{h_1}u_1=\frac{h_1}{2}f(0).$$

En la práctica $\alpha_1=-1,\ \alpha_2\geq 0$ con lo cual

$$\left[\frac{k\left(\frac{h_1}{2}\right)}{h_1}+\alpha_2 k(0)+\frac{h_1}{2}q(0)\right]u_0-\frac{k\left(\frac{h_1}{2}\right)}{h_1}u_1=\frac{h_1}{2}f(0)+k(0)\,g_1.$$

Condición de frontera en $x=L$.

Tomamos $a=L-\frac{h_n}{2},\ \ b=L$. Entonces

$$-k(x)\left.\frac{du}{dx}\right|_{L-\frac{h_n}{2}}^{L}+\int_{L-\frac{h_n}{2}}^{L} q(x)\,u(x)\,dx=\int_{L-\frac{h_n}{2}}^{L} f(x)\,dx,$$

con lo cual

$$-k(x)\left.\frac{du}{dx}\right|_{L-\frac{h_n}{2}}^{L}=-k(L)\frac{du}{dx}(L)+k\left(L-\frac{h_n}{2}\right)\frac{du}{dx}\left(L-\frac{h_n}{2}\right).$$

La derivada $\frac{du}{dx}\left(L-\frac{h_n}{2}\right)$ se aproxima mediante diferencias finitas centrales, esto es,

$$\frac{du}{dx}\left(L-\frac{h_n}{2}\right)\simeq\frac{u(L)-u(x_{n-1})}{h_n}.$$

Por otro lado, $\beta_1 u'(L)+\beta_2 u(L)=g_2\Rightarrow u'(L)=\frac{1}{\beta_1}(g_2-\beta_2 u(L))$ con $\beta_1\neq 0$. Luego,

$$-k(x)\left.\frac{du}{dx}\right|_{L-\frac{h_n}{2}}^{L}=-\frac{k(L)}{\beta_1}(g_2-\beta_2 u(L))+k\left(L-\frac{h_n}{2}\right)\frac{u(L)-u(x_{n-1})}{h_n}.$$

Por otro lado,

$$\begin{aligned}\int_{L-\frac{h_n}{2}}^{L} q(x)\,u(x)\,dx &\simeq \frac{h_n}{2}q(L)\,u(L),\\ \int_{L-\frac{h_n}{2}}^{L} f(x)\,dx &\simeq \frac{h_n}{2}f(L).\end{aligned}$$

Entonces

$$-\frac{k(L)}{\beta_1}(g_2-\beta_2 u(L))+k\left(L-\frac{h_n}{2}\right)\frac{u(L)-u(x_{n-1})}{h_n}+\frac{h_n}{2}q(L)\,u(L)\simeq\frac{h_n}{2}f(L).$$

de dond

$$\frac{-k\left(L-\frac{h_n}{2}\right)}{h_n}u_{n-1}+\left[\frac{k\left(L-\frac{h_n}{2}\right)}{h_n}+\frac{k\left(L\right)\beta_2}{\beta_1}+\frac{h_n}{2}q\left(L\right)\right]u_n=\frac{h_n}{2}f\left(L\right)+\frac{k\left(L\right)g_2}{\beta_1}.$$

En la práctica $\beta_1=1$. Así

$$\frac{-k\left(L-\frac{h_n}{2}\right)}{h_n}u_{n-1}+\left[\frac{k\left(L-\frac{h_n}{2}\right)}{h_n}+k\left(L\right)\beta_2+\frac{h_n}{2}q\left(L\right)\right]u_n=\frac{h_n}{2}f\left(L\right)+k\left(L\right)g_2.$$

La matriz $A_h=\left(a_{ij}\left(h\right)\right)\in M_{(n+1)\times(n+1)}\left[\mathbb{R}\right]$ satisface las siguientes propiedades:

i. $A_h=A_h^T$.

ii. $a_{ii}\left(h\right)>0\quad i=0,1,...,n,$

iii. $a_{i-1i}\left(h\right)=a_{ii-1}\left(h\right)<0\quad i=2,...,n,$

iv. $\begin{cases} a_{ii}\left(h\right)\geq-a_{i-1i}\left(h\right)-a_{ii-1}\left(h\right)\quad i=2,...,n-1, \\ a_{11}\left(h\right)>-a_{21}\left(h\right),\ \ a_{nn}\left(h\right)>-a_{nn-1}\left(h\right). \end{cases}$

La matriz A_h es positiva.

11.5. Ejercicios resueltos

1. Considerar el problema de valores de frontera:

$$\begin{cases} -u''\left(x\right)+x^2u\left(x\right)=1+x\quad x\in\left]0,1\right[, \\ u\left(0\right)=0,\ \ u\left(1\right)=1. \end{cases}$$

Aplicar el método de diferencias finitas para aproximar la solución con $n=5$. La matriz debe factorarse con el método de Choleski.

Solución

Utilizando diferencias finitas centrales, se tiene

$$u''\left(x_j\right)\simeq\frac{u\left(x_{j+1}\right)-2u\left(x_j\right)+u\left(x_{j-1}\right)}{h_x^2},$$

donde $h_x=\dfrac{1}{5}=0{,}2$ y $x_j=jh_x\quad j=0,1,\ldots,5.$

Sea u_j una aproximación de $u\left(x_j\right)$. Entonces el problema de valores de frontera:

$$\begin{cases} -u''\left(x\right)+x^2u\left(x\right)=1+x\quad x\in\left]0,1\right[, \\ u\left(0\right)=0,\ \ u\left(1\right)=1. \end{cases}$$

se discretiza del modo siguiente:

$$\begin{cases} u_0=0,\ \ u\left(1\right)=u_5=1, \\ -\dfrac{u\left(x_{j+1}\right)-2u\left(x_j\right)+u\left(x_{j-1}\right)}{h_x^2}+x_j^2u_j=1+x_j \end{cases}$$

o de manera explícita:

$$\begin{aligned} j &= 1, \quad -\frac{u_2-2u_1}{h_x^2}+x_1^2u_1=1+x_1 \\ j &= 2, \quad -\frac{u_3-2u_2+u_1}{h_x^2}+x_2^2u_2=1+x_2 \\ j &= 3, \quad -\frac{u_4-2u_3+u_2}{h_x^2}+x_3^2u_3=1+x_3 \\ j &= 4, \quad -\frac{1-2u_4+u_3}{h_x^2}+x_4^2u_4=1+x_4 \end{aligned}$$

que expresado en forma matricial, se tiene

$$\begin{bmatrix} \frac{2}{h_x^2}+x_1^2 & -\frac{1}{h_x^2} & 0 & 0 \\ -\frac{1}{h_x^2} & \frac{2}{h_x^2}+x_2^2 & -\frac{1}{h_x^2} & 0 \\ 0 & -\frac{1}{h_x^2} & \frac{2}{h_x^2}+x_3^2 & -\frac{1}{h_x^2} \\ 0 & 0 & -\frac{1}{h_x^2} & \frac{2}{h_x^2}+x_4^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 1+x_1 \\ 1+x_2 \\ 1+x_3 \\ 1+x_4+\frac{1}{h_x^2} \end{bmatrix}.$$

Remplazando $h_x = 0{,}2$ y $x_j = jh_x \quad j = 1,2,3,4$ se obtiene

$$A = \begin{bmatrix} 50{,}04 & -25 & 0 & 0 \\ -25 & 50{,}16 & -25 & 0 \\ 0 & -25 & 50{,}36 & -25 \\ 0 & 0 & -25 & 50{,}64 \end{bmatrix}, \quad \overrightarrow{b} = \begin{bmatrix} 1{,}2 \\ 1{,}4 \\ 1{,}6 \\ 26{,}8 \end{bmatrix}$$

Ponemos $\overrightarrow{u} = (u_1, u_2, u_3, u_4)^T$. Para resolver el sistema de ecuaciones $A\overrightarrow{u} = \overrightarrow{b}$ aplicamos el método de factorización LU.

Sean $L = \begin{bmatrix} L_{11} & 0 & 0 & 0 \\ L_{21} & L_{22} & 0 & 0 \\ 0 & L_{32} & L_{33} & 0 \\ 0 & 0 & L_{43} & L_{44} \end{bmatrix}$, $U = \begin{bmatrix} 1 & u_{12} & 0 & 0 \\ 0 & 1 & u_{23} & 0 \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Entonces

$$A = LU = \begin{bmatrix} L_{11} & L_{22}u_{12} & 0 & 0 \\ L_{21} & L_{21}u_{12}+L_{22} & L_{22}u_{23} & 0 \\ 0 & L_{32} & L_{32}u_{23}+L_{33} & L_{33}u_{34} \\ 0 & 0 & L_{43} & L_{43}u_{34}+L_{44} \end{bmatrix}$$

$$\begin{aligned} L_{11} &= 50{,}04, \quad u_{12} = -\frac{25}{50{,}04} = -0{,}4996003197 \\ L_{21} &= -25, \quad L_{22} = 50{,}16 - (-25)\left(-\frac{25}{50{,}04}\right) = 37{,}66999201, \\ & \qquad u_{23} = -\frac{25}{37{,}66999201} = -0{,}663658224 \\ L_{32} &= -25, \quad L_{33} = 50{,}36 - (-25)(-0{,}663658224) = 33{,}7685444, \\ & \qquad u_{34} = -\frac{25}{33{,}7685444} = -0{,}740333954 \\ L_{43} &= -25, \quad L_{44} = 50{,}64 - (-25)(-0{,}740333954) = 32{,}13165115 \end{aligned}$$

$$\begin{aligned} L &= \begin{bmatrix} 50{,}04 & 0 & 0 & 0 \\ -25 & 37{,}66999201 & 0 & 0 \\ 0 & -25 & 33{,}7685444 & 0 \\ 0 & 0 & -25 & 32{,}13165115 \end{bmatrix} \\ U &= \begin{bmatrix} 1 & -0{,}4996003197 & 0 & 0 \\ 0 & 1 & -0{,}663658224 & 0 \\ 0 & 0 & 1 & -0{,}740333954 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

El sistema de ecuaciones $A\overrightarrow{x} = \overrightarrow{b}$ se transforma en los siguientes: $LU\overrightarrow{u} = \overrightarrow{b} \Leftrightarrow \begin{cases} L\overrightarrow{y} = \overrightarrow{b} \\ U\overrightarrow{u} = \overrightarrow{y} \end{cases}$.

Solución del sistema lineal de ecuaciones $L\overrightarrow{y} = \overrightarrow{b}$. En forma explícita, este sistema de ecuaciones lineales se expresa como sigue:

$$\begin{bmatrix} 50{,}04 & 0 & 0 & 0 \\ -25 & 37{,}66999201 & 0 & 0 \\ 0 & -25 & 33{,}7685444 & 0 \\ 0 & 0 & -25 & 32{,}13165115 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1{,}2 \\ 1{,}4 \\ 1{,}6 \\ 26{,}8 \end{bmatrix},$$

luego

$$\begin{aligned} y_1 &= \frac{1{,}2}{50{,}04} = 0{,}023980815, \\ y_2 &= \frac{1{,}4 + 25 \times 0{,}023980815}{37{,}66999201} = 0{,}05307992589, \\ y_3 &= \frac{1{,}6 + 25 \times 0{,}05307992589}{33{,}7685444} = 0{,}0866782444, \\ y_4 &= \frac{26{,}8 + 25 \times 0{,}0866782444}{32{,}13165115} = 0{,}9015084838, \end{aligned}$$

Solución del sistema lineal de ecuaciones $U\overrightarrow{u} = \overrightarrow{y}$. Este sistema en forma explícita se escribe como sigue:

$$\begin{bmatrix} 1 & -0{,}4996003197 & 0 & 0 \\ 0 & 1 & -0{,}663658224 & 0 \\ 0 & 0 & 1 & -0{,}740333954 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0{,}023980815 \\ 0{,}05307992589 \\ 0{,}0866782444 \\ 0{,}9015084838 \end{bmatrix},$$

$$\begin{aligned} u_4 &= 0{,}9015084838, \\ u_3 &= 0{,}0866782444 + 0{,}740333954 \times 0{,}9015084838 = 0{,}750955848, \\ u_2 &= 0{,}05307992589 + 0{,}663658224 \times 0{,}750955848 = 0{,}5535416624, \\ u_1 &= 0{,}023980815 + 0{,}4996003197 \times 0{,}5535416624 = 0{,}3005304065, \end{aligned}$$

con 3 cifras $\overrightarrow{u}^T = (0{,}3,\ 0{,}554,\ 0{,}751,\ 0{,}902)$, $u_0 = 0$ y $u_5 = 1$.

2. Considerar el problema de valores de frontera:

$$\begin{cases} -u''(x) + (1+x)\,u(x) = x^2 \quad x \in]0,1[, \\ u(0) = 1, \ \ u(1) = 0. \end{cases}$$

Aplicar el método de diferencias finitas para aproximar la solución con $n = 5$. La matriz debe factorarse con el método LU.

Solución

Sean $n = 5$, $h_x = \dfrac{1}{5} = 0{,}2$, $x_j = 0{,}2j \quad j = 0, 1, \ldots, 5$. Utilizando diferencias finitas centrales, se tiene

$$u''(x_j) \simeq \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h_x^2}.$$

Sea u_j una aproximación de $u(x_j)$. Entonces el problema de valores de frontera:

$$\begin{cases} -u''(x) + (1+x)\,u(x) = x^2 \quad x \in]0,1[, \\ u(0) = 1, \ \ u(1) = 0. \end{cases}$$

se aproxima mediante el esquema siguiente:

$$\begin{cases} -\dfrac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h_x^2} + (1+x_j)\,u_j = x_j^2, \quad j = 1, 2, 3, 4, \\ u(0) = u_0 = 1, \ \ u(1) = u_5 = 0. \end{cases}$$

En forma explícita, este conjunto de ecuaciones se escribe:

$$\begin{aligned}
j &= 1, \quad -\frac{u_2 - 2u_1 + 1}{h_x^2} + (1 + x_1)\, u_1 = x_1^2, \quad u_0 = 1, \\
j &= 2, \quad -\frac{u_3 - 2u_2 + u_1}{h_x^2} + (1 + x_2)\, u_2 = x_2^2, \\
j &= 3, \quad -\frac{u_4 - 2u_3 + u_2}{h_x^2} + (1 + x_3)\, u_3 = x_3^2, \\
j &= 4, \quad -\frac{-2u_4 + u_3}{h_x^2} + (1 + x_4)\, u_4 = x_4^2, \quad u_5 = 0,
\end{aligned}$$

que en forma matricial, se escribe como sigue:

$$\begin{bmatrix}
\frac{2}{h_x^2} + 1 + x_1 & -\frac{1}{h_x^2} & 0 & 0 \\
-\frac{1}{h_x^2} & \frac{2}{h_x^2} + 1 + x_2 & -\frac{1}{h_x^2} & 0 \\
0 & -\frac{1}{h_x^2} & \frac{2}{h_x^2} + 1 + x_3 & -\frac{1}{h_x^2} \\
0 & 0 & -\frac{1}{h_x^2} & \frac{2}{h_x^2} 1 + + x_4
\end{bmatrix}
\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} =
\begin{bmatrix} x_1^2 + \frac{1}{h_x^2} \\ x_2^2 \\ x_3^2 \\ x_4^2 \end{bmatrix}.$$

Poniendo $h_x = 0{,}2$ y $x_1 = 0{,}2,\ x_2 = 0{,}4,\ x_3 = 0{,}6,\ x_4 = 0{,}8,$ se tiene

$$\begin{bmatrix}
51{,}20 & -25 & 0 & 0 \\
-25 & 51{,}4 & -25 & 0 \\
0 & -25 & 51{,}6 & -25 \\
0 & 0 & -25 & 51{,}8
\end{bmatrix}
\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} =
\begin{bmatrix} 25{,}04 \\ 0{,}16 \\ 0{,}36 \\ 0{,}64 \end{bmatrix}.$$

Para hallar la solución de este sistema de ecuaciones lineales, apliquemos el método de factorización LU. Para el efecto, factoramos $A = LU$, donde

$$L = \begin{bmatrix}
L_{11} & 0 & 0 & 0 \\
L_{21} & L_{22} & 0 & 0 \\
0 & L_{32} & L_{33} & 0 \\
0 & 0 & L_{43} & L_{44}
\end{bmatrix}, \quad
U = \begin{bmatrix}
1 & u_{12} & 0 & 0 \\
0 & 1 & u_{23} & 0 \\
0 & 0 & 1 & u_{34} \\
0 & 0 & 0 & 1
\end{bmatrix}.$$

Puesto que

$$LU = \begin{bmatrix}
L_{11} & L_{22}u_{12} & 0 & 0 \\
L_{21} & L_{21}u_{12} + L_{22} & L_{22}u_{23} & 0 \\
0 & L_{32} & L_{32}u_{23} + L_{33} & L_{33}u_{34} \\
0 & 0 & L_{43} & L_{43}u_{34} + L_{44}
\end{bmatrix}$$

y de la igualdad $A = LU$ se obtienen las matrices LU como siguen:

$$\begin{aligned}
L_{11} &= 51{,}20, \quad u_{12} = \frac{a_{12}}{L_{11}} = -\frac{25}{51{,}20} = -0{,}48828125, \\
L_{21} &= -25, \quad L_{22} = a_{22} - L_{21}u_{12} = 51{,}4 - (-25)\,(-0{,}48828125) = 39{,}19296875, \\
&\qquad u_{23} = \frac{a_{23}}{L_{22}} - \frac{25}{39{,}19296875} = -0{,}6378695158, \\
L_{32} &= -25, \quad L_{33} = a_{33} - L_{32}u_{23} = 51{,}66 - (-25)\,(-0{,}6378695158) = 35{,}65326211, \\
&\qquad u_{34} = \frac{a_{34}}{L_{33}} = -\frac{25}{35{,}65326211} = -0{,}7011981098, \\
L_{43} &= -25, \quad L_{44} = a_{44} - L_{43}u_{34} = 51{,}8 - (-25)\,(-0{,}7011981098) = 34{,}27004726,
\end{aligned}$$

Ponemos $\overrightarrow{u}^T = (u_1, u_2, u_3, u_4)$. El sistema de ecuaciones $A\overrightarrow{u} = \overrightarrow{b}$ es equivalente al siguiente:

$$LU\overrightarrow{u} = \overrightarrow{b} \Leftrightarrow \begin{cases} L\overrightarrow{y} = \overrightarrow{b}, \\ U\overrightarrow{u} = \overrightarrow{y}. \end{cases}$$

Hallemos la solución del sistema de ecuaciones $L\overrightarrow{y} = \overrightarrow{b}$. Tenemos

$$\begin{bmatrix} 51{,}20 & 0 & 0 & 0 \\ -25 & 39{,}19296875 & 0 & 0 \\ 0 & -25 & 35{,}65326211 & 0 \\ 0 & 0 & -25 & 34{,}27004726 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 25{,}04 \\ 0{,}16 \\ 0{,}36 \\ 0{,}64 \end{bmatrix},$$

luego

$$\begin{aligned} y_1 &= \frac{25{,}04}{51{,}20} = 0{,}4890625, \\ y_2 &= \frac{0{,}16 + 25 \times 0{,}4890625}{39{,}19296875} = 0{,}316040425, \\ y_3 &= \frac{1{,}6 + 25 \times 0{,}316040425}{35{,}65326211} = 0{,}2317042008, \\ y_4 &= \frac{26{,}8 + 25 \times 0{,}2317042008}{34{,}27004726} = 0{,}1877034184, \end{aligned}$$

Hallemos la solución del sistema de ecuaciones $U\overrightarrow{u} = \overrightarrow{y}$ que en forma explícita se expresa como sigue:

$$\begin{bmatrix} 1 & -0{,}48828125 & 0 & 0 \\ 0 & 1 & -0{,}6378695158 & 0 \\ 0 & 0 & 1 & -0{,}7011981098 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0{,}4890625 \\ 0{,}316040425 \\ 0{,}2317042008 \\ 0{,}1877034184 \end{bmatrix},$$

de donde

$$\begin{aligned} u_4 &= 0{,}1877034184, \\ u_3 &= 0{,}2317042008 + 0{,}7011981098 \times 0{,}1877034184 = 0{,}363321483, \\ u_2 &= 0{,}316040425 + 0{,}6378695158 \times 0{,}363321483 = 0{,}5477921234, \\ u_1 &= 0{,}4890625 + 0{,}48828125 \times 0{,}5477921234 = 0{,}7565391228, \end{aligned}$$

La aproximación de la ecuación diferencial en cuatro nodos internos con una precisión de 3 cifras es

$$\overrightarrow{u}^T = (0{,}757,\ 0{,}548,\ 0{,}363,\ 0{,}188)\,.$$

3. Considerar el problema de valores de frontera no lineal siguiente:

$$\begin{cases} -u'' + u^3 = f \quad \text{sobre }]0, 12[\,, \\ u\,(0) = u\,(12) = 0, \end{cases}$$

donde $f\,(x) = \begin{cases} 0, \text{ si } x \in [0, 5]\,, \\ 1, \text{ si } x \in]5, 12[\,, \end{cases}$ y el esquema numérico siguiente:

$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + u_j^3 = f\,(x_j)\,, \quad j = 1, \ldots, n-1 \\ u_0 = 0, \quad u_n = 0, \end{cases}$$

con $h = \dfrac{12}{n}$, $n \in \mathbb{Z}^+$, $x_j = jh$, $j = 0, 1, \ldots, n$ y u_j una aproximación de $u\,(x_j)$.

i) Para $n = 6$, construir el sistema no lineal correspondiente al esquema numérico propuesto.

ii) Considere una aproximación inicial $\overrightarrow{u}^{(0)} = (0, 1, 1, 1, 1, 1, 0)$. Aplique el método de Newton y dos iteraciones.

Solución

Puesto que se requiere generar un algoritmo general para aproximar la solución del problema de valores de frontera, hemos de proceder en ese sentido para luego particularizar al caso $n = 6$.

Sean $L > 0$. Suponemos que f es una función definida en $[0, L]$. Sea $n \in \mathbb{Z}^+$. Ponemos $h = \dfrac{L}{n}$, $x_j = jh$, $j = 0, 1, \ldots, n$, $f_j = f(x_j)$, $j = 0, 1, \ldots, n$. Del esquema numérico

$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + u_j^3 = f(x_j), \quad j = 1, \ldots, n-1, \\ u_0 = 0, \quad u_n = 0, \end{cases}$$

se obtiene las siguientes ecuaciones.

Para $j = 1$,

$$-\frac{u_2 - 2u_1 + u_0}{h^2} + u_1^3 = f_1,$$

o bien

$$\frac{2}{h^2}u_1 - \frac{1}{h^2}u_2 + u_1^3 = f_1.$$

Para $1 < j < n - 1$,

$$-\frac{1}{h^2}u_{j-1} + \frac{2}{h^2}u_j - \frac{1}{h^2}u_{j+1} + u_j^3 = f_j.$$

Para $j = n - 1$

$$-\frac{u_n - 2u_{n-1} + u_{n-2}}{h^2} + u_{n-1}^3 = f_{n-1,}$$

de donde

$$-\frac{1}{h^2}u_{n-2} + \frac{2}{h^2}u_{n-1} + u_{n-1}^3 = f_{n-1}.$$

Ponemos $\overrightarrow{u}^T = (u_1, \ldots, u_{n-1})$, $u_0 = 0$, $u_n = 0$. Definimos la matriz $A = (a_{ij}(h))$ y el vector $B(\overrightarrow{u}) \in \mathbb{R}^{n-1}$ como sigue

$$a_{ij}(h) = \begin{cases} 0, & \text{si} \quad |i-j| > 1, \\ \dfrac{2}{h^2}, & \text{si} \quad i = j, \\ -\dfrac{1}{h^2}, & \text{si} \quad |i-j| = 1 \end{cases} \qquad i, j = 1, \ldots, n-1, \quad b_j(\overrightarrow{u}) = u_j^3, \quad j = 1, \ldots, n-1.$$

Se define $\overrightarrow{f}^T = (f(x_1), \ldots, f(x_{n-1})) = (f_1, \ldots, f_{n-1})$.

Entonces, el esquema numérico propuesto, discretización del problema de valores de frontera

$$\begin{cases} -u'' + u^3 = f \quad \text{sobre }]0, L[\\ u(0) = 0 = u(L), \end{cases}$$

se transforma en el siguiente sistema de ecuaciones no lineales:

$$A\overrightarrow{u} + B(\overrightarrow{u}) = \overrightarrow{f}.$$

Se define $F : \mathbb{R}^{n-1} \longrightarrow \mathbb{R}^{n-1}$ como $F(\overrightarrow{u}) = A\overrightarrow{u} + B(\overrightarrow{u}) - \overrightarrow{f} \quad \overrightarrow{u} \in \mathbb{R}^4$. Para aplicar el método de Newton, requerimos de la matriz jacobiana $DF(\overrightarrow{u})$. Para el efecto, hallemos la derivada de Gâteau (derivada direccional de F) $D\overrightarrow{y}F(\overrightarrow{u})$ según la dirección $\overrightarrow{y} \in \mathbb{R}^{n-1}$ en $\overrightarrow{u} \in \mathbb{R}^{n-1}$. Por definición

$$D\overrightarrow{y}F(\overrightarrow{u}) = \lim_{t \to 0} \frac{F(\overrightarrow{u} + t\overrightarrow{y}) - F(\overrightarrow{u})}{t},$$

siempre que el límite exista.

Sea $t \neq 0$. Entonces

$$\begin{aligned} F(\overrightarrow{u} + t\overrightarrow{y}) - F(\overrightarrow{u}) &= A(\overrightarrow{u} + t\overrightarrow{y}) + B(\overrightarrow{u} + t\overrightarrow{y}) - \overrightarrow{f} - \left(A\overrightarrow{u} + B(\overrightarrow{u}) - \overrightarrow{f}\right) \\ &= tA\overrightarrow{y} + B(\overrightarrow{u} + t\overrightarrow{y}) - B(\overrightarrow{u}). \end{aligned}$$

El j-ésimo componente de $B(\overrightarrow{u} + t\overrightarrow{y}) - B(\overrightarrow{u})$ es

$$\begin{aligned} (u_j + ty_j)^3 - u_j^3 &= u_j^3 + 3tu_j^2 y_j + 3t^2 u_j y_j^2 + t^3 y_j^3 - u_j^3 \\ &= ty_j\left(3u_j^2 + 3tu_j y_j + t^2 y_j^2\right), \end{aligned}$$

de donde

$$\lim_{t\to 0}\frac{(u_j+ty_j)^3-u_j^3}{t}=3y_ju_j^2,\quad j=1,\ldots,n-1,$$

luego,

$$\begin{aligned}D\overrightarrow{y}F(\overrightarrow{u}) &= \lim_{t\to 0}\frac{F(\overrightarrow{u}+t\overrightarrow{y})-F(\overrightarrow{u})}{t}=\lim_{t\to 0}\left(A\overrightarrow{y}+\frac{B(\overrightarrow{u}+t\overrightarrow{y})-B(\overrightarrow{u})}{t}\right)\\ &= A\overrightarrow{y}+D\overrightarrow{y}B(\overrightarrow{u}),\end{aligned}$$

donde

$$D\overrightarrow{y}B(\overrightarrow{u})=3\begin{bmatrix}u_1^2 & 0 & \ldots & 0\\ 0 & u_2^2 & \ldots & 0\\ \vdots & & \ddots & \vdots\\ 0 & \ldots & \ldots & u_{n-1}^2\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_{n-1}\end{bmatrix}=3c(\overrightarrow{u})\,\overrightarrow{y},$$

con $c(\overrightarrow{u})=(c_{ij}(\overrightarrow{u}))$ la matriz definida como

$$c_{ij}(\overrightarrow{u})=\begin{cases}u_i^3, & \text{si } i=j\\ 0, & \text{si } i\neq j\end{cases},\quad i,j=1,\ldots,n-1.$$

Así,

$$D\overrightarrow{y}F(\overrightarrow{u})=[A+3c(\overrightarrow{u})]\,\overrightarrow{y}\quad\forall\overrightarrow{y}\in\mathbb{R}^{n-1},$$

con lo cual, la matriz jacobiana está definida como

$$DF(\overrightarrow{u})=A+3c(\overrightarrow{u}),\quad\forall\overrightarrow{u}\in\mathbb{R}^{n-1}.$$

El método de Newton está definido como sigue:

$$\begin{cases}\widetilde{u}^{(0)} \text{ aproximación inicial},\\ DF\left(\widetilde{u}^{(k)}\right)\widetilde{w}=-F\left(\widetilde{u}^{(k)}\right), & k=0,1,\ldots,N_{\text{máx}},\\ \widetilde{u}^{(k+1)}=\widetilde{u}^{(k)}+\widetilde{w}.\end{cases}$$

i) Para $n=6$ se tiene $h=\dfrac{L}{n}=\dfrac{12}{6}=2$, $x_j=jh$, $j=0,1,\ldots,n$, la partición del intervalo [0,12] está constituida por los siguientes nodos: $x_0=0$, $x_1=2$, $x_2=4$, $x_3=6$, $x_4=8$, $x_5=10$, $x_6=12$, y la función f en dichos nodos tiene los valores siguientes: $f(0)=0$, $f(2)=0$, $f(4)=0$, $f(6)=1$, $f(8)=1$, $f(10)=1$, $f(12)=1$.

Luego, $\widetilde{f}=(0,0,1,1,1)$,

$$A=\begin{bmatrix}\frac{2}{h^2} & -\frac{1}{h^2} & 0 & 0 & 0\\ -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} & 0 & 0\\ 0 & -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} & 0\\ 0 & 0 & -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2}\\ 0 & 0 & 0 & -\frac{1}{h^2} & \frac{2}{h^2}\end{bmatrix}=\begin{bmatrix}\frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0\\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0\\ 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0\\ 0 & 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4}\\ 0 & 0 & 0 & -\frac{1}{4} & \frac{1}{2}\end{bmatrix},$$

$$B(\overrightarrow{u})=\begin{bmatrix}u_1^3\\ u_2^3\\ u_3^3\\ u_4^3\\ u_5^3\end{bmatrix},\quad c(\overrightarrow{u})=\begin{bmatrix}u_1^2 & 0 & 0 & 0 & 0\\ 0 & u_2^2 & 0 & 0 & 0\\ 0 & 0 & u_3^2 & 0 & 0\\ 0 & 0 & 0 & u_4^2 & 0\\ 0 & 0 & 0 & 0 & u_5^2\end{bmatrix}.$$

El sistema no lineal de ecuaciones es

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 \\ 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ 0 & 0 & 0 & -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} u_1^3 \\ u_2^3 \\ u_3^3 \\ u_4^3 \\ u_5^3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

por lo tanto

$$F\left(\overrightarrow{u}\right) = A\overrightarrow{u} + B\left(\overrightarrow{u}\right) - \overrightarrow{f} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 \\ 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ 0 & 0 & 0 & -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} u_1^3 \\ u_2^3 \\ u_3^3 \\ u_4^3 \\ u_5^3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

ii) Sea $\left(\overrightarrow{u}^{(0)}\right)^T = (0,1,1,1,1,1,0)$ una aproximación inicial. Ponemos $\left(\overrightarrow{u}^{(0)}\right)^T = (1,1,1,1,1)$. Entonces

$$B\left(\widetilde{u}^{(0)}\right) = \begin{bmatrix} 1^3 \\ 1^3 \\ 1^3 \\ 1^3 \\ 1^3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad A\widetilde{u}^{(0)} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 \\ 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ 0 & 0 & 0 & -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ 0 \\ 0 \\ 0 \\ \frac{1}{4} \end{bmatrix},$$

$$F\left(\widetilde{u}^{(0)}\right) = A\widetilde{u}^{(0)} + B\left(\widetilde{u}^{(0)}\right) - \overrightarrow{f} = \begin{bmatrix} \frac{1}{4} \\ 0 \\ 0 \\ 0 \\ \frac{1}{4} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{4} \\ 1 \\ 0 \\ 0 \\ \frac{1}{4} \end{bmatrix},$$

$$DF\left(\widetilde{u}^{(0)}\right) = A + 3c\left(\widetilde{u}^{(0)}\right) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 \\ 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ 0 & 0 & 0 & -\frac{1}{4} & \frac{1}{2} \end{bmatrix} + 3 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{7}{2} & -\frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{7}{2} & -\frac{1}{4} & 0 & 0 \\ 0 & -\frac{1}{4} & \frac{7}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{7}{2} & -\frac{1}{4} \\ 0 & 0 & 0 & -\frac{1}{4} & \frac{7}{2} \end{bmatrix}$$

El sistema de ecuaciones lineales es:

$$\begin{bmatrix} \frac{7}{2} & -\frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{7}{2} & -\frac{1}{4} & 0 & 0 \\ 0 & -\frac{1}{4} & \frac{7}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{7}{2} & -\frac{1}{4} \\ 0 & 0 & 0 & -\frac{1}{4} & \frac{7}{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix} = - \begin{bmatrix} \frac{5}{4} \\ 1 \\ 0 \\ 0 \\ \frac{1}{4} \end{bmatrix}.$$

Método de Crout. Sea D la matriz del sistema de ecuaciones precedente. Entonces, $D = LU$ y $D\overrightarrow{w} = \overrightarrow{b} \Leftrightarrow LU\widetilde{w} = \widetilde{b} \Leftrightarrow \begin{cases} L\widetilde{y} = \widetilde{b} \\ U\widetilde{w} = \widetilde{y}. \end{cases}$

Comencemos con la factorización

$$D = LU = \begin{bmatrix} L_{11} & 0 & 0 & 0 & 0 \\ L_{21} & L_{22} & 0 & 0 & 0 \\ 0 & L_{32} & L_{33} & 0 & 0 \\ 0 & 0 & L_{43} & L_{44} & 0 \\ 0 & 0 & 0 & L_{54} & L_{55} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & 0 & 0 & 0 \\ 0 & 1 & u_{23} & 0 & 0 \\ 0 & 0 & 1 & u_{34} & 0 \\ 0 & 0 & 0 & 1 & u_{45} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} L_{11} & L_{11}u_{12} & 0 & 0 & 0 \\ L_{21} & L_{21}u_{12} + L_{22} & L_{22}u_{23} & 0 & 0 \\ 0 & L_{32} & L_{32}u_{23} + L_{33} & L_{33}u_{34} & 0 \\ 0 & 0 & L_{43} & L_{43}u_{34} + L_{44} & L_{44}u_{45} \\ 0 & 0 & 0 & L_{54} & L_{44}u_{45} + L_{55} \end{bmatrix}$$

Obtenemos

$$\begin{aligned}
L_{11} &= \frac{7}{2},\\
u_{12} &= \frac{a_{12}}{L_{11}} = \frac{-\frac{1}{4}}{\frac{7}{2}} = -\frac{1}{14},\\
L_{21} &= -\frac{1}{4},\\
L_{22} &= a_{22} - L_{21}u_{12} = \frac{7}{2} - \left(-\frac{1}{4}\right)\left(-\frac{1}{14}\right) = \frac{195}{56},\\
u_{23} &= \frac{a_{23}}{L_{22}} = \frac{-\frac{1}{4}}{\frac{195}{56}} = -\frac{14}{195},\\
L_{32} &= -\frac{1}{4},\\
L_{33} &= a_{33} - L_{32}u_{23} = \frac{7}{2} - \left(-\frac{1}{4}\right)\left(-\frac{14}{195}\right) = \frac{1358}{390} = \frac{679}{195},\\
u_{34} &= \frac{a_{34}}{L_{33}} = \frac{-\frac{1}{4}}{\frac{1358}{390}} = -\frac{195}{2716},\\
L_{43} &= -\frac{1}{4},\\
L_{44} &= a_{44} - L_{43}u_{34} = \frac{7}{2} - \left(-\frac{1}{4}\right)\left(-\frac{195}{2716}\right) = \frac{37829}{10864},\\
u_{45} &= \frac{a_{45}}{L_{44}} = \frac{-\frac{1}{4}}{\frac{37829}{10864}} = -\frac{2716}{37829},\\
L_{54} &= -\frac{1}{4},\\
L_{55} &= a_{55} - L_{54}u_{45} = \frac{7}{2} - \left(-\frac{1}{4}\right)\left(-\frac{2716}{37829}\right) = \frac{263445}{75658}.
\end{aligned}$$

Resolución del sistema triangular inferior $L\overrightarrow{y} = \overrightarrow{b}$, esto es,

$$\begin{bmatrix}
\frac{7}{2} & 0 & 0 & 0 & 0\\
-\frac{1}{4} & \frac{195}{56} & 0 & 0 & 0\\
0 & -\frac{1}{4} & \frac{679}{195} & 0 & 0\\
0 & 0 & -\frac{1}{4} & \frac{37829}{10864} & 0\\
0 & 0 & 0 & -\frac{1}{4} & \frac{263445}{75658}
\end{bmatrix}
\begin{bmatrix} y_1\\ y_2\\ y_3\\ y_4\\ y_5 \end{bmatrix} =
\begin{bmatrix} -\frac{5}{4}\\ -1\\ 0\\ 0\\ -\frac{1}{4} \end{bmatrix}.$$

Se tiene

$$\begin{aligned}
y_1 &= \frac{-\frac{5}{4}}{\frac{7}{2}} = -\frac{5}{14} = -0{,}3571428571,\\
-\frac{1}{4}y_1 + \frac{195}{56}y_2 &= -1 \Rightarrow y_2 = \frac{56}{195}\left(-1 + \frac{1}{4}y_1\right) = -0{,}3128205128,\\
-\frac{1}{4}y_2 + \frac{679}{195}y_3 &= 0 \Rightarrow y_3 = \frac{195}{679} \times \frac{1}{4}y_2 = -0{,}02245941926,\\
-\frac{1}{4}y_3 + \frac{37829}{10864}y_4 &= 0 \Rightarrow y_4 = \frac{10864}{37829} \times \frac{1}{4}y_3 = -0{,}0004031298739,\\
-\frac{1}{4}y_4 + \frac{263445}{75658}y_5 &= -\frac{1}{4} \Rightarrow y_5 = \frac{75658}{263445}\left(-\frac{1}{4} + \frac{1}{4}y_4\right) = -0{,}07182571318,
\end{aligned}$$

Resolución del sistema triangular superior $U\overrightarrow{w} = \overrightarrow{y}$:

$$\begin{bmatrix}
1 & -\frac{1}{14} & 0 & 0 & 0\\
0 & 1 & -\frac{14}{195} & 0 & 0\\
0 & 0 & 1 & -\frac{195}{2716} & 0\\
0 & 0 & 0 & 1 & -\frac{2716}{37829}\\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix} w_1\\ w_2\\ w_3\\ w_4\\ w_5 \end{bmatrix} =
\begin{bmatrix} -0{,}3571428571\\ -0{,}3128205128\\ -0{,}02245941926\\ -0{,}0004031298739\\ -0{,}07182571318 \end{bmatrix}$$

Obtenemos

$$\begin{aligned}
w_5 &= -0{,}07182571318,\\
w_4 - \frac{2716}{37829}w_5 &= -0{,}0004031298739,\\
w_4 &= \frac{2716}{37829}w_5 - 0{,}0004031298739 = -0{,}00555998406,\\
w_3 - \frac{195}{2716}w_4 &= -0{,}02245949926,\\
w_3 &= \frac{195}{2716}w_4 - 0{,}02245949926 = -0{,}228586881,\\
w_2 - \frac{14}{195}w_3 &= -0{,}3128205128,\\
w_2 &= \frac{14}{195}w_3 - 0{,}3128205128 = -0{,}3144616494,\\
w_1 - \frac{1}{14}w_2 &= -0{,}3571428571,\\
w_1 &= \frac{1}{14}w_2 - 0{,}3571428571 = -0{,}3796044035,
\end{aligned}$$

Luego

$$\widetilde{u}^{(1)} = \widetilde{u}^{(0)} + \widetilde{w} = \begin{bmatrix} 1\\1\\1\\1\\1 \end{bmatrix} + \begin{bmatrix} -0{,}3796044035\\ -0{,}3144616494\\ -0{,}228586881\\ -0{,}00555998406\\ -0{,}07182571318 \end{bmatrix} = \begin{bmatrix} 0{,}6203955965\\ 0{,}6855383506\\ 0{,}9771413119\\ 0{,}9944400159\\ 0{,}9281742868 \end{bmatrix}.$$

Segunda iteración

La matriz A y el vector $\widetilde{f}$ no cambian. Calculemos $B\left(\widetilde{u}^{(1)}\right)$. Tenemos

$$B\left(\widetilde{u}^{(1)}\right) = \begin{bmatrix} \left(u_1^{(1)}\right)^3 \\ \left(u_2^{(1)}\right)^3 \\ \left(u_3^{(1)}\right)^3 \\ \left(u_4^{(1)}\right)^3 \\ \left(u_5^{(1)}\right)^3 \end{bmatrix} = \begin{bmatrix} 0{,}238784493 \\ 0{,}3221775434 \\ 0{,}9341234603 \\ 0{,}9987910978 \\ 0{,}7996291156 \end{bmatrix}.$$

Calculemos $F\left(\widetilde{u}^{(1)}\right) = A\widetilde{u}^{(1)} + B\left(\widetilde{u}^{(1)}\right) - \widetilde{f}$. Tenemos

$$\begin{bmatrix} 0{,}5 & -0{,}25 & 0 & 0 & 0 \\ -0{,}25 & 0{,}5 & -0{,}25 & 0 & 0 \\ 0 & -0{,}25 & 0{,}5 & -0{,}25 & 0 \\ 0 & 0 & -0{,}25 & 0{,}5 & -0{,}25 \\ 0 & 0 & 0 & -0{,}25 & 0{,}5 \end{bmatrix} \begin{bmatrix} 0{,}6203955965 \\ 0{,}6855383506 \\ 0{,}9771413119 \\ 0{,}9944400159 \\ 0{,}9281742868 \end{bmatrix} + \begin{bmatrix} 0{,}238784493 \\ 0{,}3221775434 \\ 0{,}9341234603 \\ 0{,}9987910978 \\ 0{,}7996291156 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0{,}3775977034 \\ 0{,}2654626945 \\ 0{,}00160990548 \\ 0{,}022160836 \\ 0{,}0138170415 \end{bmatrix}$$

Calculemos $DF\left(\widetilde{u}^{(1)}\right) = A + 3c\left(\widetilde{u}^{(1)}\right)$.

Se tiene

$$A + 3c\left(\widetilde{u}^{(1)}\right) = \begin{bmatrix} 0{,}5 & -0{,}25 & 0 & 0 & 0 \\ -0{,}25 & 0{,}5 & -0{,}25 & 0 & 0 \\ 0 & -0{,}25 & 0{,}5 & -0{,}25 & 0 \\ 0 & 0 & -0{,}25 & 0{,}5 & -0{,}25 \\ 0 & 0 & 0 & -0{,}25 & 0{,}5 \end{bmatrix}$$

$$+3 \begin{bmatrix} \left(u_1^{(1)}\right)^2 & 0 & 0 & 0 & 0 \\ 0 & \left(u_2^{(1)}\right)^2 & 0 & 0 & 0 \\ 0 & 0 & \left(u_3^{(1)}\right)^2 & 0 & 0 \\ 0 & 0 & 0 & \left(u_4^{(1)}\right)^2 & 0 \\ 0 & 0 & 0 & 0 & \left(u_5^{(1)}\right)^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1{,}654672089 & -0{,}25 & 0 & 0 & 0 \\ -0{,}25 & 1{,}90988849 & -0{,}25 & 0 & 0 \\ 0 & -0{,}25 & 3{,}366756292 & -0{,}25 & 0 \\ 0 & 0 & -0{,}25 & 3{,}497581708 & -0{,}25 \\ 0 & 0 & 0 & -0{,}25 & 3{,}08452252 \end{bmatrix}$$

El sistema de ecuaciones lineales correspondiente $DF\left(\widetilde{u}^{(1)}\right)\widetilde{w} = -F\left(\widetilde{u}^{(1)}\right)$ es el siguiente:

$$\begin{bmatrix} 1{,}654672089 & -0{,}25 & 0 & 0 & 0 \\ -0{,}25 & 1{,}90988849 & -0{,}25 & 0 & 0 \\ 0 & -0{,}25 & 3{,}366756292 & -0{,}25 & 0 \\ 0 & 0 & -0{,}25 & 3{,}497581708 & -0{,}25 \\ 0 & 0 & 0 & -0{,}25 & 3{,}08452252 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix}$$

$$= \begin{bmatrix} -0{,}3775977034 \\ -0{,}2654626945 \\ -0{,}00160990548 \\ -0{,}02217505055 \\ -0{,}0138170415 \end{bmatrix}$$

Resolvemos este sistema utilizando el método de Crout. Para ello la matriz del sistema lo factoramos en la forma LU como se procedió en la primera iteración. Note que la matriz del sistema es simétrica, estrictamente diagonalmente dominante. Se obtienen los siguientes resultados:

$$\begin{aligned}
L_{11} &= 1{,}654672089, \\
u_{12} &= \frac{-0{,}25}{1{,}654672089} = -0{,}1510873373, \\
L_{21} &= -0{,}25, \\
L_{22} &= a_{22} - L_{21}u_{12} = 1{,}872116656, \\
u_{23} &= \frac{a_{23}}{L_{22}} = \frac{-0{,}25}{0{,}872116656} = -0{,}1335386869, \\
L_{32} &= -0{,}25, \\
L_{33} &= a_{33} - L_{32}u_{23} = 3{,}33337162, \\
u_{34} &= \frac{a_{34}}{L_{33}} = -0{,}07499913855, \\
L_{43} &= -0{,}25, \\
L_{44} &= a_{44} - L_{43}u_{34} = 3{,}478831923, \\
u_{45} &= \frac{a_{45}}{L_{44}} = 0{,}7186320165, \\
L_{54} &= -0{,}25, \\
L_{55} &= a_{55} - L_{54}u_{45} = 3{,}06655672.
\end{aligned}$$

El sistema triangular inferior $L\overrightarrow{y} = -F\left(\widetilde{u}^{(1)}\right)$ es el siguiente:

$$\begin{bmatrix}
1{,}654672089 & 0 & 0 & 0 & 0 \\
-0{,}25 & 1{,}872116656 & 0 & 0 & 0 \\
0 & -0{,}25 & 3{,}33337162 & 0 & 0 \\
0 & 0 & -0{,}25 & 3{,}478831923 & 0 \\
0 & 0 & 0 & -0{,}25 & 3{,}06655682
\end{bmatrix}
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$
$$= \begin{bmatrix}
-0{,}3775977034 \\
-0{,}2654626945 \\
-0{,}00160990548 \\
-0{,}02217505055 \\
-0{,}0138170415
\end{bmatrix}$$

cuya solución es

$$\begin{aligned}
y_1 &= \frac{-0{,}3775977034}{1{,}654672089} = -0{,}2282009263, \\
y_2 &= \frac{0{,}25y_1 - 0{,}2654626945}{1{,}872116656} = -0{,}1722718107, \\
y_3 &= \frac{0{,}25y_2 - 0{,}00160990548}{3{,}33337162} = -0{,}0134032035, \\
y_4 &= \frac{0{,}25y_3 - 0{,}02217505055}{3{,}478831923} = -0{,}007337477635, \\
y_5 &= \frac{0{,}25y_4 - 0{,}0138170415}{3{,}06655682} = -0{,}005103903935.
\end{aligned}$$

Resolución del sistema triangular superior $U\widetilde{w} = \overrightarrow{y}$:

$$\begin{bmatrix} 1 & -0{,}1510873383 & 0 & 0 & 0 \\ 0 & 1 & -0{,}1335386869 & 0 & 0 \\ 0 & 0 & 1 & -0{,}07499913855 & 0 \\ 0 & 0 & 0 & 1 & -0{,}7186320165 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix}$$

$$= \begin{bmatrix} -0{,}2282009263 \\ -0{,}1722718107 \\ -0{,}0134032035 \\ -0{,}007337477635 \\ -0{,}005103903935 \end{bmatrix}$$

Obtenemos

$$\begin{aligned} w_5 &= -0{,}005103903935, \\ w_4 &= -0{,}01100530641, \\ w_3 &= -0{,}014228592, \\ w_2 &= -0{,}1741718782, \\ w_1 &= -0{,}2545160916, \end{aligned}$$

En consecuencia:

$$\widetilde{u}^{(2)} = \widetilde{u}^{(1)} + \overrightarrow{w} = \begin{bmatrix} 0{,}6203955965 \\ 0{,}6855383506 \\ 0{,}9775405007 \\ 0{,}9995968701 \\ 0{,}9281742868 \end{bmatrix} + \begin{bmatrix} -0{,}2545160916 \\ -0{,}1741718782 \\ -0{,}014228592 \\ -0{,}01100530641 \\ -0{,}005103903935 \end{bmatrix} = \begin{bmatrix} 0{,}3658795049 \\ 0{,}5113664724 \\ 0{,}9633119087 \\ 0{,}9885915637 \\ 0{,}9230703829 \end{bmatrix},$$

y con $u_0^{(2)} = 0,\ u_6^{(2)} = 0$ se obtiene $\overrightarrow{u}^{(2)}$.

11.6. Lecturas complementarias y bibliografía

1. Tom M. Apostol, Calculus, Volumen 1, Segunda Edición, Editorial Reverté, Barcelona, 1977.
2. Uri M. Ascher, Robert M. M. Mattheij, Robert D. Russell, Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1995.
3. N. Bakhvalov, Metodos Numéricos, Editorial Paraninfo, Madrid, 1980.
4. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.
5. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
6. Steven C. Chapra, Raymond P. Canale, Numerical Methods for Engineers, Third Edition, Editorial McGraw-Hill, Boston, 1998.
7. S. D. Conte, Carl de Boor, Análisis Numérico, Segunda Edición, Editorial Mc Graw-Hill, México, 1981.
8. M. Crouzeix, A. L. Mignot, Analyse Numérique des Equations Différentielles, Seconde Edition, Editorial Masson, París, 1989.
9. Jean-Pierre Demailly, Analyse Numérique et Equations differentielles, Presses Universitaires de Grenoble, Grenoble, 1991.

10. Peter Deuflhard, Folkmar Bornemann, Scientific Computing with Ordinary Differential Equations, Editorial Springer-Verlag, New York, 2002.

11. B. P. Demidovich, I. A. Maron, E. Cálculo Numérico Fundamental, Editorial Paraninfo, Madrid, 1977.

12. B. P. Demidovich, I. A. Maron, E. S. Schuwalowa, Métodos Numéricos de Análisis, Editorial Paraninfo, Madrid, 1980.

13. James W. Demmel, Applied Numerical Linear Algebra, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.

14. C. H. Edwards, Jr., David E. Penney, Ecuaciones Diferenciales Elementales y Problemas con Condiciones en la Frontera, Tercera Edición, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1993.

15. Ferruccio Fontanella, Aldo Pasquali, Calcolo Numerico. Metodi e Algoritmi, Volume II Pitagora Editrice Bologna, 1983.

16. M. K. Gavurin, Conferencias sobre los Métodos de Cálculo, Editorial Mir, Moscú, 1973.

17. Curtis F. Gerald, Patrick O. Wheatley, Análisis Numérico con Aplicaciones, Sexta Edición, Editorial Pearson Educación de México, México, 2000.

18. E. Hairer, S. P. Norsett, G. Wanner, Solving Ordinary Differential Equations I, Second Revised Edition, Editorial Springer-Verlag, Berlín, 2000.

19. R. Kent Nagle, Edward B. Saff, Arthur David Snider, Ecuaciones Diferenciales y Problemas con Valores en la Frontera, Tercera Edición, editorial Pearson Educación, México, 2001.

20. David Kincaid, Ward Cheney, Análisis Numérico, Editorial Addison-Wesley Iberoamericana, Wilmington, 1994.

21. Melvin J. Maron, Robert J. López, Análisis Numérico, Tercera Edición, Compañía Editorial Continental, México, 1995.

22. R. M. M. Mattheij, J. Molenaar, Ordinary Differential Equations in Theory and Practice, Editorial John Wiley & Sons, New York, 1996.

23. Shoichiro Nakamura, Métodos Numérico Aplicados con Software, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1992.

24. Antonio Nieves, Federico C. Dominguez, Métodos Numéricos Aplicados a la Ingeniería, Tercera Reimpresión, Compañía Editorial Continental, S. A. De C. V., México, 1998.

25. S. Nikolski, Fórmulas de Cuadratura, Editorial Mir, Moscú, 1990.

26. J. M. Ortega, W. C. Rheinbolodt, Iterative Solution of Nonlinear Equatios in Several Variables, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.

27. Anthony Ralston, Introducción al Análisis Numérico, Editorial Limusa, México, 1978.

28. Werner C. Rheinboldt, Methods for Solving Systems of Nonlinear Equations, Second Edition, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

29. A. A. Samarski, Introducción a los Métodos Numéricos, Editorial Mir, Moscú, 1986.

30. M. Sibony, J. Cl. Mardon, Analyse Numérique II, Approximations et Equations Différentielles, Editorial Hermann, París, 1988.

31. J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Editorial Springer-Verlag, 1980.

Capítulo 12

Apendice

Resumen

Este apéndice tiene como objetivo refrescar algunos resultados de los espacios vectoriales, los espacios normados y los espacios con producto interior. Al final se provee de una amplia bibliografía sobre estos tópicos.

12.1. Espacios vectoriales reales.

12.1.1. Definición de espacio vectorial. Ejemplos.

Se denota con $\mathbb{R}$ al cuerpo de los números reales. Nos limitamos en definir los espacios vectoriales reales.

Definición 1 *Un espacio vectorial V sobre $\mathbb{R}$ consiste en un conjunto no vacío V en el que se ha definido dos operaciones: adición "+" en V que a cada par de elementos x, y de V le asocia un único elemento $x+y$ de V, y, producto de números reales por elementos de V dicha también producto por escalares que a cada $\alpha \in \mathbb{R}$ y $x \in V$ le asocia un único elemento αx de V; y, estas operaciones satisfacen las propiedades siguientes:*

i. Conmutativa: para todo $x, y \in V$, $x+y=y+x$.

ii. Asociativa: para todo $x, y, z \in V$, $(x+y)+z=x+(y+z)$.

iii. Existencia de elemento neutro: existe $0 \in V$ tal que para todo $x \in V$, $x+0=0+x=x$.

iv. Existencia de opuestos aditivos: para cada $x \in V$, existe $y \in V$ tal que $x+y=0$.

v. Para todo $\alpha \in \mathbb{R}$, $x, y \in \mathbb{R}$, $\alpha(x+y)=\alpha x+\alpha y$.

vi. Para todo $x \in V$, $\alpha, \beta \in \mathbb{R}$, $(\alpha+\beta)x=\alpha x+\beta x$.

vii. Para todo $x \in V$, $\alpha, \beta \in \mathbb{R}$, $\alpha(\beta x)=(\alpha\beta)x$.

viii. Para todo $x \in V$, $1 \cdot x = x$.

Los elementos de V se llaman vectores y los elementos de $\mathbb{R}$ se llaman escalares. El espacio vectorial V sobre $\mathbb{R}$ se dirá simplemente espacio vectorial real. El conjunto V con la operación adición "+" que satisface las propiedades i) a iv) se dice grupo conmutativo que se nota $(V,+)$.

El elemento $0 \in V$ de iii) es único y se denomina elemento nulo. El elemento y de iv) se escribe $-x$, además es único. La propiedad iv) se expresa como sigue: $\forall x \in V,\ \exists -x \in V$ tal que $x+(-x)=0$.

Para todo $x, y, z \in V$, se escribe $x+y+z$ en vez de $(x+y)+z$ o de $x+(y+z)$.

En todo espacio vectorial real se verifican las propiedades siguientes cuyas demostraciones son inmediatas y se dejan como ejercicio.

i. Para todo $x \in V$, $0x = 0$.

ii. Para todo $\alpha \in \mathbb{R}$, $\alpha 0 = 0$.

iii. Para todo $\alpha \in \mathbb{R}$, $x \in V$, $(-\alpha)\, x = -\alpha x$.

iv. $\alpha x = 0 \Leftrightarrow \alpha = 0$ o $x = 0$.

Ejemplos

1. El espacio vectorial $\mathbb{R}^n$. Sea $n \in \mathbb{Z}^+$. Se denota con $\mathbb{R}^n$ al conjunto $\{(x_1, ..., x_n) \mid x_i \in \mathbb{R},\ i = 1, ..., n\}$, esto es $\mathbb{R}^n = \{(x_1, ..., x_n) \mid x_i \in \mathbb{R},\ i = 1, ..., n\}$. A los elementos de $\mathbb{R}^n$ los notamos como $\overrightarrow{x}$, $\overrightarrow{y}$, etc. También escribiremos $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ y los denominaremos vectores de $\mathbb{R}^n$. El elemento nulo de $\mathbb{R}^n$ se escribe $\overrightarrow{0} = (0, ..., 0)$.

   En $\mathbb{R}^n$ se define la igualdad, adición y producto de escalares por elementos de $\mathbb{R}^n$ como sigue. Sean $\overrightarrow{x} = (x_1, ..., x_n)$, $\overrightarrow{y} = (y_1, ..., y_n)$ dos elementos de $\mathbb{R}^n$, $\alpha \in \mathbb{R}$.

   Igualdad: diremos $\overrightarrow{x} = \overrightarrow{y}$ si y solo si $x_i = y_i$, $i = 1, ..., n$.

   Adición: $\overrightarrow{x} + \overrightarrow{y} = (x_1, ..., x_n) + (y_1, ..., y_n) = (x_1 + y_1, ..., x_n + y_n)$.

   Producto por escalares: $\alpha \overrightarrow{x} = \alpha (x_1, ..., x_n) = (\alpha x_1, ..., \alpha x_n)$.

   De la definición de adición en $\mathbb{R}^n$, se tiene $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n \Rightarrow \overrightarrow{x} + \overrightarrow{y} \in \mathbb{R}^n$, y, de la definición de producto por escalares $\alpha \in \mathbb{R}$, $\overrightarrow{x} \in \mathbb{R}^n \Rightarrow \alpha \overrightarrow{x} \in \mathbb{R}^n$. Mas aún, se prueba fácilmente que $\mathbb{R}^n$ es un espacio vectorial real.

   El opuesto aditivo de $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ es $-\overrightarrow{x} = (-x_1, ..., -x_n) \in \mathbb{R}^n$.

   Para $n = 1$, se tiene que $\mathbb{R}$ es un espacio vectorial sobre si mismo.

   Para $n = 2$, $\mathbb{R}^2 = \{(x, y) \mid x,\ y \in \mathbb{R}\}$. Note que si $\overrightarrow{x} = (a, b)$, $\overrightarrow{y} = (c, d) \in \mathbb{R}^2$, $\alpha \in \mathbb{R}$ se tiene

   Igualdad: $\overrightarrow{x} = \overrightarrow{y} \Leftrightarrow a = c$ y $b = d$,

   Adición: $\overrightarrow{x} + \overrightarrow{y} = (a, b) + (c, d) = (a + c, b + d)$.

   Producto por escalares: $\alpha \overrightarrow{x} = \alpha (a, b) = (\alpha a, \alpha b)$.

   Para $n = 3$, $\mathbb{R}^3 = \{(x, y, z) \mid x,\ y,\ z \in \mathbb{R}\}$. Sean $\overrightarrow{x} = (x_1, y_1, z_1)$, $\overrightarrow{y} = (x_2, y_2, z_2) \in \mathbb{R}^3$, $\alpha \in \mathbb{R}$. Se tiene

$$\begin{aligned} \overrightarrow{x} &= \overrightarrow{y} \Leftrightarrow x_1 = x_2,\ y_1 = y_2,\ z_1 = z_2, \\ \overrightarrow{x} + \overrightarrow{y} &= (x_1, y_1, z_1) + (x_2, y_2, z_2) = (x_1 + x_2, y_1 + y_2, z_1 + z_2), \\ \alpha \overrightarrow{x} &= \alpha (x_1, y_1, z_1) = (\alpha x_1, \alpha y_1, \alpha z_1). \end{aligned}$$

   Los elementos de $\mathbb{R}^n$ se escribirán también como vectores columna, así: $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$.

2. Espacio de matrices $M_{m\times n}[\mathbb{R}]$. Sean m, $n \in \mathbb{Z}^+$. Una matriz de $m \times n$ con valores en $\mathbb{R}$ es un arreglo rectangular de la forma:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

   donde $a_{i,j} \in \mathbb{R}$, $i = 1, ..., m$, $j = 1, ..., n$.

Los números naturales $i = 1, ..., m, \quad j = 1, ..., n$ se llaman índices. Cuando i es fijo, los elementos $a_{i1}, a_{i2}, ..., a_{in}$ forman el i-ésimo renglón de la matriz y se puede considerar como un vector de $\mathbb{R}^n$, esto es, $(a_{i1}, a_{i2}, ..., a_{in}) \in \mathbb{R}^n$. Para j fijo, los elementos forman la j-ésima columna de la matriz. Esta puede considerarse como un vector columna de $\mathbb{R}^m$, es decir $\begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix} \in \mathbb{R}^m$.

A una matriz de $m \times n$ la representaremos abreviadamente como $(a_{ij})_{m\times n}$. También escribiremos $A = (a_{ij})_{m\times n}$ y simplemente (a_{ij}) si no hay peligro de confusión. Se nota con $M_{m\times n}[\mathbb{R}]$ al conjunto de todas la matrices de $m \times n$ con valores en $\mathbb{R}$.

Cuando $m = n$, los elementos de $M_{n\times n}[\mathbb{R}]$ los denominamos matrices cuadradas. Si $A = (a_{ij})_{n\times n}$ es una matriz cuadrada, si no hay peligro de confusión, escribiremos simplemente $A = (a_{ij})$.

Sean $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{m\times n} \in M_{m\times n}[\mathbb{R}]$ y $\alpha \in \mathbb{R}$. Definimos la igualdad, adición de matrices y producto de escalares por matrices, como sigue:

Igualdad: $A = B \Leftrightarrow a_{ij} = b_{ij}, \quad i = 1, ..., m, \quad j = 1, ..., n.$

Adición: $A + B = (a_{ij})_{m\times n} + (b_{ij})_{m\times n} = (a_{ij} + b_{ij})_{m\times n}$.

Producto por escalares: $\alpha A = \alpha (a_{ij})_{m\times n} = (\alpha a_{ij})_{m\times n}$.

Por la definición de adición en $M_{m\times n}[\mathbb{R}]$, tenemos $A, B \in M_{m\times n}[\mathbb{R}] \Rightarrow A + B \in M_{m\times n}[\mathbb{R}]$, y por la definición de producto de escalares por matrices tenemos $\alpha \in \mathbb{R}, \quad A \in M_{m\times n}[\mathbb{R}] \Rightarrow \alpha A \in M_{m\times n}[\mathbb{R}]$.

Se demuestra fácilmente que $M_{m\times n}[\mathbb{R}]$ es un espacio vectorial real.

El elemento neutro de $M_{m\times n}[\mathbb{R}]$ es la matriz nula o matriz cero y la representamos con $0 = (0)_{m\times n}$, es decir que la matriz nula 0 es aquella que sus elementos son $0 \in \mathbb{R}$. El opuesto aditivo de $A = (a_{ij})_{m\times n}$ es la matriz notada $-A = (-a_{ij})_{m\times n}$.

Por otro lado, si $n = 1$ las matrices de $M_{m\times 1}[\mathbb{R}]$ coinciden con los vectores columna de $\mathbb{R}^m$ y si $m = 1$, las matrices de $M_{1\times n}[\mathbb{R}]$ coinciden con los vectores fila de $\mathbb{R}^n$.

3. Los espacios $C([a, b])$ y $C^1([a, b])$.

Revisemos brevemente algunos conceptos sobre funciones reales, operaciones con funciones reales, límites, continuidad, derivación e integración que son tratados en el curso de Análisis Matemático.

Sea $A \subset \mathbb{R}$, $A \neq \emptyset$. Una función real f definida en el conjunto A es un subconjunto F del producto cartesiano $A \times \mathbb{R}$ que satisface con las dos propiedades siguientes:

i. Para cada $x \in A$, existe un único $y \in \mathbb{R}$ tal que $y = f(x)$, o bien $(x, y) \in F$.

ii. Si $(x, y_1), \ (x, y_2) \in F$, entonces $y_1 = y_2$.

Sea f una función real definida en A. Escribiremos $f : \begin{cases} A & \to & \mathbb{R} \\ x & \to & f(x), \end{cases}$ que se lee f es la función de A en $\mathbb{R}$ que a cada $x \in A$ le asocia un único elemento $f(x)$ en $\mathbb{R}$. Se dirá también f es la función real de A en $\mathbb{R}$ que a cada $x \in A$ le asocia o le corresponde $f(x) \in \mathbb{R}$. El conjunto A se llama dominio de f y se designa con $\text{Dom}(f)$. El conjunto $\mathbb{R}$ se llama conjunto de llegada de f y el conjunto $\text{Rec}(f) = \{f(x) \mid x \in A\}$ se llama recorrido de f. Claramente $\text{Rec}(f) \subset \mathbb{R}$ y $\text{Rec}(f) \neq \emptyset$. Se designa con $\mathcal{F}(A)$ al conjunto de todas las funciones definidas en A.

La función nula $0 \in \mathcal{F}(A)$ se define como $0(x) = 0 \quad \forall x \in A$, y la función unidad $\mathbf{1} \in \mathcal{F}(A)$ está definida como $\mathbf{1}(x) = 1 \quad \forall x \in A$.

En $\mathcal{F}(A)$ se define la igualdad, adición y producto por escalares o producto de números reales por funciones como sigue:

Igualdad: Sean $f, g \in \mathcal{F}(A), \ f = g$ si y solo si $f(x) = g(x) \quad \forall x \in A$.

Adición: Sean $f, g \in \mathcal{F}(A)$. Se define $f + g \in \mathcal{F}(A)$ como $(f + g)(x) = f(x) + g(x) \quad \forall x \in A$.

Producto por escalares: Sean $\alpha \in \mathbb{R}, \quad f \in \mathcal{F}(A)$. Se define $\alpha f \in \mathcal{F}(A)$ como $(\alpha f)(x) = \alpha f(x) \quad \forall x \in A$.

Se demuestra fácilmente que $\mathcal{F}(A)$ es un espacio vectorial real denominado espacio de funciones definidas en A.

Funciones continuas

Sea A un intervalo de $\mathbb{R}$, $f \in \mathcal{F}(A)$, $x_0 \in A$ y $L \in \mathbb{R}$. Se dice que $f(x)$ tiende a L cuando x tiende a x_0 que se escribe $f(x) \underset{x \to x_0}{\to} L$, si y solo si se satisface la siguiente condición:

$$\forall \varepsilon > 0, \ \exists \delta > 0 \text{ tal que } \forall x \in A \ \text{ con } \ 0 < |x - x_0| < \delta \Rightarrow |f(x) - L| < \varepsilon.$$

Escribiremos también $\lim_{x \to x_0} f(x) = L$ que se lee límite de $f(x)$ cuando x tiende a x_0 es igual a L.

Sea $A \subset \mathbb{R}$, $A \neq \emptyset$ y $f \in \mathcal{F}(A)$. Se dice que f es continua en $x_0 \in A$ si y solo si se satisfacen las dos condiciones siguientes:

i. $f(x_0)$ está bién definido.

ii. $\lim_{x \to x_0} f(x) = f(x_0)$.

Se dice f continua en A si y solo si f es continua en todo punto $x_0 \in A$.

Se designa con $C(A)$ al conjunto de todas las funciones continuas en A. En el curso de análisis matemático se prueba que la suma de dos funciones continuas es continua, y que el producto de un número real λ por una función continua f es también una función continua, esto es,

$$f, g \in C(A) \Rightarrow f + g \in C(A), \quad \lambda \in \mathbb{R}, \ f \in C(A) \Rightarrow \lambda f \in C(A).$$

Se prueba además que $C(A)$ es un espacio vectorial real.

En particular, si $A = [a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$ es un intervalo cerrado y acotado de $\mathbb{R}$, el conjunto $C(A)$ se denota $C([a, b])$ y se le denomina espacio de funciones continuas en $[a, b]$.

Funciones derivables

Sean $A \subset \mathbb{R}$, $A \neq \emptyset$, f una función real definida en A y $x_0 \in A$. Si $\lim_{h \to 0} \dfrac{f(x_0 + h) - f(x_0)}{h}$ existe, este se denomina derivada de f en x_0 que se escribe $f'(x_0)$ o también $\dfrac{df}{dx}(x_0)$; esto es,

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Se dice que f es derivable en A si $f'(x_0)$ existe en todo punto $x_0 \in A$ y se define una nueva función f' llamada función derivada de f.

Se designa con $C^1(A)$ al conjunto de todas las funciones f tales que f' es continua en A, y diremos que f es de clase C^1 en A. Particularmente si $A = [a, b]$, escribiremos $C^1([a, b])$ y diremos espacio de funciones de clase C^1en $[a, b]$.

Funciones integrables

En esta parte proponemos algunos resultados importantes de la teoría de la integración de funciones reales acotadas.

Definición 2 *Sean $a, b \in \mathbb{R}$ con $a < b$, $n \in \mathbb{Z}^+$. Una subdivisión o partición del intervalo $[a, b]$ se nota con $\Pi(n)$ y se define como el conjunto $\{x_0, x_1, ..., x_n\}$, donde $x_0 = a$, $x_n = b$, $x_i < x_{i+1}$, $i = 0, ..., n-1$.*

Si $\Pi(n)$ es una subdivisión de $[a, b]$, $[x_{i-1}, x_i]$, $i = 1, ..., n$ designa el i-ésimo subintervalo de $[a, b]$. Se pone $h_i = x_i - x_{i-1}$ la longitud del intervalo $[x_{i-1}, x_i]$, $i = 1, ..., n$, y $h = \max_{i=0,1,...,n-1} h_i$.

Definición 3 *Sean $m, n \in \mathbb{Z}^+$. Se dice que $\Pi(m)$ es una subdivisión mas fina que $\Pi(n)$ si se verifica que $\Pi(n) \subset \Pi(m)$.*

Particularmente, si $\Pi(m)$, $\Pi(n)$ son dos subdivisiones de $[a,b]$, $\Pi(m)\cup\Pi(n)$ es una subdivisión mas fina que $\Pi(m)$ y $\Pi(n)$.

Definición 4 *Sea f una función real definida en $[a,b]$. Se dice que f es acotada en $[a,b]$ si y solo si $f([a,b])=\{f(x)\mid x\in[a,b]\}$ es acotado, es decir, existe $\beta>0$ tal que $|f(x)|\leq\beta \quad \forall x\in[a,b]$.*

Sean $n\in\mathbb{Z}^+$, $\Pi(n)$ una subdivisión de $[a,b]$ y f una función acotada en $[a,b]$. Se pone

$$\begin{aligned}\alpha_i &= \operatorname*{Inf}_{x\in[x_{i-1},x_i]} f(x), \quad \beta_i = \operatorname*{Sup}_{x\in[x_{i-1},x_i]} f(x), \quad i=1,...,n,\\ \alpha &= \operatorname*{Min}_{i=1,...,n}\alpha_i, \quad \beta = \operatorname*{Max}_{i=1,...,n}\beta_i.\end{aligned}$$

Definición 5 *La oscilación de f en $[x_{i-1},x_i]$, $i=1,...,n$ se define como*

$$\omega_i = \operatorname*{Sup}_{x\in[x_{i-1},x_i]} f(x) - \operatorname*{Inf}_{x\in[x_{i-1},x_i]} f(x), \quad i=1,...,n.$$

Definición 6 *Sea f una función real definida en $[a,b]$. Se dice que f es una función escalonada si y solo si existe una subdivisión $\Pi(n)=\{x_0=a,x_1,...,x_n=b\}$ y $c_1,...,c_n\in\mathbb{R}$ tales que*

$$f(x)=c_i, \quad x\in\,]x_{i-1},x_i[\,, \quad i=1,...,n.$$

La subdivisión $\Pi(n)$ se dice asociada a f.

Note que f es una función definida en todo $[a,b]$ y en cada subintervalo abierto $]x_{i-1},x_i[$, $i=1,...,n$, f es constante.

Definición 7 *Sea s una función escalonada en $[a,b]$ con $\Pi(n)=\{a=x_0,x_1,...,x_n=b\}$ la partición asociada a s y $s(x)=c_i$, $x\in\,]x_{i-1},x_i[$, $i=1,...,n$. La integral de s sobre el intervalo $[a,b]$ se nota $\int_a^b s(x)\,dx$ y se define como $\int_a^b s(x)\,dx=\sum_{i=1}^{n} c_i h_i$, donde $h_i=x_i-x_{i-1}$, $i=1,...,n$.*
La notación $\int_a^b s(x)\,dx$ se lee integral de la función s con respecto de x en el intervalo $[a,b]$. El número real a es extremo inferior de integración, y el número real b el extremo superior de integración.

Sean s, t dos funciones escalonadas en $[a,b]$ y f una función acotada en $[a,b]$ tales que

$$s(x)\leq f(x)\leq t(x) \quad \forall x\in[a,b],$$

la función s se llama función escalonada inferior a f, y t se llama función escalonada superior a f. Particularmente, sea $n\in\mathbb{Z}^+$ y $\Pi(n)$ una subdivisión de $[a,b]$. Se definen las funciones escalonadas s_n y t_n como sigue:

$$\begin{aligned}s_n(x) &= \alpha_i = \operatorname*{Inf}_{x\in[x_{i-1},x_i]} f(x), \quad x\in\,]x_{i-1},x_i[\,, \quad i=1,...,n,\\ t_n(x) &= \beta_i = \operatorname*{Sup}_{x\in[x_{i-1},x_i]} f(x), \quad x\in\,]x_{i-1},x_i[\,, \quad i=1,...,n.\end{aligned}$$

Estas funciones satisfacen la siguiente desigualdad:

$$s_n(x)\leq f(x)\leq t_n(x) \quad \forall x\in\,]x_{i-1},x_i[\,, \quad i=1,...,n.$$

Se tiene que s_n es una función escalonada inferior a f, t_n es una función escalonada superior a f. Las integrales de estas funciones se definen como:

$$S_n(f)=\int_a^b s_n(x)\,dx=\sum_{i=1}^{n}\alpha_i h_i, \quad T_n(f)=\int_a^b t_n(x)\,dx=\sum_{i=1}^{n}\beta_i h_i,$$

donde $h_i = x_i - x_{i-1}, \quad i = 1, ..., n$. Se verifica

$$\alpha (b-a) \leq S_n (f) \leq T_n (f) \leq \beta (b-a)$$

y

$$0 \leq T_n (f) - S_n (f) \leq \sum_{i=1}^{n} \omega_i h_i \leq (\beta - \alpha)(b-a),$$

donde ω_i es la oscilación de f, $\alpha = \underset{x\in[a,b]}{\text{Inf}} f(x), \quad \beta = \underset{x\in[a,b]}{\text{Sup}} f(x)$.

Definición 8 *Sea f una función real, acotada en $[a,b]$.*

i. La integral inferior de f se designa con $\underline{I}(f)$ y se define como $\underline{I}(f) = \underset{n\in\mathbb{Z}^+}{\text{Sup}} S_n(f) = \underset{n\in\mathbb{Z}^+}{\text{Sup}} \int_a^b s_n(x)\,dx$, donde s_n es escalonada inferior a f.

ii. La integral superior de f se designa con $\overline{I}(f)$ y se define como $\overline{I}(f) = \underset{n\in\mathbb{Z}^+}{\text{Inf}} T_n(f) = \underset{n\in\mathbb{Z}^+}{\text{Inf}} \int_a^b t_n(x)\,dx$, donde t_n es escalonada superior a f.

Se verifica inmediatamente que si $s_n \leq f \leq t_n, \quad \int_a^b s_n(x)\,dx \leq \underline{I}(f) \leq \overline{I}(f) \leq \int_a^b t_n(x)\,dx$.

Definición 9 *Sea f una función real, acotada en $[a,b]$. Se dice que f es integrable en $[a,b]$ si y solo si $\underline{I}(f) = \overline{I}(f)$. En tal caso, escribimos $I(f) = \int_a^b f(x)\,dx$ y al número real $I(f)$ lo denominamos la integral de la función f en el intervalo $[a,b]$.*

Las funciones monótonas en $[a,b]$ (crecientes, decrecientes), las funciones continuas en $[a,b]$ son ejemplos de funciones integrables en $[a,b]$.

Se denota con $\mathcal{I}([a,b])$ al conjunto de todas las funciones integrables en $[a,b]$. Con las operaciones habituales de adición " + " de funciones y producto de escalares por funciones, $\mathcal{I}([a,b])$ es un espacio vectorial real denominado espacio de funciones integrables en $[a,b]$. Se tiene

i. $f, g \in \mathcal{I}([a,b]) \Rightarrow f+g \in \mathcal{I}([a,b])$, e $\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$,

ii. $\lambda \in \mathbb{R}$, $f \in \mathcal{I}([a,b]) \Rightarrow \lambda f \in \mathcal{I}([a,b])$, e $\int_a^b \lambda f(x)\,dx = \lambda \int_a^b f(x)\,dx$.
Si $f \in \mathcal{I}([a,b])$ y $\alpha \in [a,b]$ se define $\int_\alpha^\alpha f(x)\,dx = 0$.

Sea $f \in \mathcal{I}([a,b])$. Se verifican las siguientes propiedades:

i. Si $c \in [a,b], \quad \int_a^b f(x)\,dx = \int_a^c f(x) + \int_c^b f(x)\,dx$.

ii. $\int_a^b f(x)\,dx = \int_{a+c}^{b+c} f(x-c)\,dx \quad \forall c \in \mathbb{R}$.

iii. Para $\alpha \neq 0, \quad \int_a^b f(x)\,dx = \frac{1}{\alpha}\int_{\alpha a}^{\alpha b} f\left(\frac{x}{\alpha}\right)dx$.

iv. Si $f(x) \geq 0 \quad \forall x \in [a,b], \quad \int_a^b f(x)\,dx \geq 0$.

v. $\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx$.

vi. Si $[c,d] \subset [a,b]$ y $f(x) \geq 0 \quad \forall x \in [a,b], \ \int_c^d f(x)\,dx \leq \int_a^b f(x)\,dx$.

Por otro lado, si $f(x) \geq 0 \quad \forall x \in [a,b]$, geométricamente $\int_a^b f(x)\,dx$ se interpreta como el área de la región

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \mid 0 \leq y \leq f(x),\ \ x \in [a,b]\right\}.$$

Ejercicios

1. Demuestre que el conjunto $\mathbb{R}^n$ en el que se ha definido la igualdad y las operaciones de adición y producto por escalares, es un espacio vectorial real.
2. Demuestre que el conjunto de matrices $M_{m\times n}[\mathbb{R}]$ en el que se ha definido la igualdad y las operaciones de adición y producto de números reales por matrices, es un espacio vectorial real.
3. Demuestre que el conjunto $\mathcal{F}(A)$ de funciones reales definidas en A en el que se ha definido la igualdad de funciones, y las operaciones de adición y producto de números reales por funciones, es un espacio vectorial real.
4. Pruebe que con las operaciones de adición de funciones y producto de números reales por funciones definidas en $\mathcal{F}(A)$, los siguientes conjuntos son espacios vectoriales reales.
   i. Conjunto $\mathcal{I}([a,b])$ de funciones integrables en $[a,b]$.
   ii. Conjunto $C([a,b])$ de funciones continuas en $[a,b]$.
   iii. Conjunto $C^1([a,b])$ de funciones derivables con derivada continua en $[a,b]$.

12.1.2. Subespacios vectoriales. Ejemplos.

Definición 10 *Sea V un espacio vectorial sobre $\mathbb{R}$ y W un subconjunto no vacío de V. Se dice que W es un subespacio de V si W es un espacio vectorial real con las mismas operaciones definidas en V.*

El cualquier espacio vectorial V, $W = V$ y $W = \{0\}$ con $0 \in V$ son subespacios de V llamados subespacios triviales de V.

Teorema 1 *Un subconjunto no vacío W de un espacio vectorial V es un subespacio de V si y solo si se satisfacen las tres condiciones siguientes:*

i. $0 \in V \Rightarrow 0 \in W$.

ii. $x,\ y \in W \Rightarrow x + y \in W$.

iii. $\alpha \in \mathbb{R},\ \ x \in W \Rightarrow \alpha x \in W$.

Si $W_1,\ W_2$ son dos subespacios de V, entonces $W_1 \cap W_2$ es también en subespacio de V.

Definición 11 *Sean V un espacio vectorial real, $S_1,\ S_2$ dos subconjuntos no vacíos de V. Se define $S_1 + S_2$ como sigue*

$$S_1 + S_2 = \{x + y \mid x \in S_1\ \ y \in S_2\}$$

y se denomina subconjunto suma de S_1 con S_2.

Sea $x_0 \in V$ y W un subespacio de V. El subconjunto

$$x_0 + W = \{x_0 + x \mid x \in W\}$$

se llama trasladado del subespacio W o subconjunto afín.

Teorema 2 *Si W_1, W_2 son subespacios de un espacio vectorial V, $W_1 + W_2$ es un subespacio de V. El subespacio $W_1 + W_2$ se llama suma de los subespacios W_1 y W_2.*

Definición 12 *Sean W_1, W_2 dos subespacios de V. Se dice que V es suma directa de W_1 y W_2 que se escribe $V = W_1 \oplus W_2$ si y solo si se satisfacen las dos condiciones siguientes:*

i. $V = W_1 + W_2$.

ii. $W_1 \cap W_2 = \{0\}$.

Teorema 3 *Sean W_1, W_2 dos subespacios de V. Entonces, $V = W_1 \oplus W_2$ si y solo si cada $x \in V$ se escribe de manera única en la forma $x = x_1 + x_2$, donde $x_1 \in W_1$, $x_2 \in W_2$.*

Ejemplos

1. El conjunto $W = \{(x_1, ..., x_{n-1}, 0) \mid x_i \in \mathbb{R},\ i = 1, ..., n-1\}$ es un subespacio de $\mathbb{R}^n$.

2. Sea $n = 3$, $W_1 = \{(x, y, 0) \mid x,\ y \in \mathbb{R}\}$, $W_2 = \{(2t, 3t, t) \mid t \in \mathbb{R}\}$ son subespacios de $\mathbb{R}^3$. Se verifica que $W_1 \cap W_2 = \{0\}$ y que $\mathbb{R}^3 = W_1 \oplus W_2$.

   Note que si $\overrightarrow{x} = (x, y, z) \in \mathbb{R}^3$, existen $\overrightarrow{x}_1 = (x - 2z, y - 3z, 0) \in W_1$ y $\overrightarrow{x}_2 = (2z, 3z, z) \in W_2$ tales que $\overrightarrow{x} = \overrightarrow{x}_1 + \overrightarrow{x}_2$. El vector $\overrightarrow{x} = \overrightarrow{x}_1 + \overrightarrow{x}_2$ se escribe de esta manera en forma única como $\overrightarrow{x}_1 \in W_1$, $\overrightarrow{x}_2 \in W_2$.

3. El espacio $C([a, b])$ de funciones continuas en $[a, b]$ es un subespacio de $\mathcal{I}([a, b])$.

4. Un polinomio$P$ de grado $\leq n$ con coeficientes reales se define como

$$P(x) = a_0 + a_1 x + ... + a_n x^n = \sum_{k=0}^{n} a_k x^k, \qquad x \in \mathbb{R},\ a_i \in \mathbb{R},\ i = 0, 1, ..., n.$$

   El polinomio nulo se define como $P(x) = 0 \quad \forall x \in \mathbb{R}$.

   Se designa con $\mathbb{K}_n[\mathbb{R}]$ el conjunto de todos los polinomios de grado $\leq n$. Se define la igualdad de polinomios, adición y producto de números reales por polinomios como sigue:

   Igualdad: Sean $P, Q \in \mathbb{K}_n[\mathbb{R}]$ con $P(x) = \sum_{k=0}^{n} a_k x^k$, $Q(x) = \sum_{k=0}^{n} b_k x^k$, $x \in \mathbb{R}$.

$$P(x) = Q(x) \Leftrightarrow a_k = b_k, \quad k = 0, 1, ..., n.$$

   Adición: Sean $P, Q \in \mathbb{K}_n[\mathbb{R}]$ con $P(x) = \sum_{k=0}^{n} a_k x^k$, $Q(x) = \sum_{k=0}^{n} b_k x^k$, $x \in \mathbb{R}$. Se define $P + Q \in \mathbb{K}_n[\mathbb{R}]$ como $(P + Q)(x) = \sum_{k=0}^{n} (a_k + b_k) x^k \quad x \in \mathbb{R}$.

   Producto por escalares: Sean $\lambda \in \mathbb{R}$, $P \in \mathbb{K}_n[\mathbb{R}]$ con $P(x) = \sum_{k=0}^{n} a_k x^k$, $x \in \mathbb{R}$. Se define $\lambda P \in \mathbb{K}_n[\mathbb{R}]$ como $(\lambda P)(x) = \lambda P(x) = \sum_{k=0}^{n} \lambda a_k x^k, \quad x \in \mathbb{R}$.

   Se demuestra que $\mathbb{K}_n[\mathbb{R}]$ es un espacio vectorial real denominado espacio de polinomios de grado $\leq n$.

   Sea $V = C([a, b])$. El conjunto de todos los polinomios de grado $\leq n$ restringidos a $[a, b]$ con las operaciones de adición y producto por escalares, es un subespacio de $C([a, b])$. A este subespacio lo notaremos con $\mathbb{K}_n([a, b])$.

Bases de V

Definición 13 *Sea A un subconjunto no vacío de un espacio vectorial V. Se dice que $x \in V$ es una combinación lineal de elementos de A si existe un número finito $x_1, ..., x_n \in A$ y $\alpha_1, ..., \alpha_n \in \mathbb{R}$ tales que $x = \sum_{i=1}^{n} \alpha_i x_i$.*

Particularmente, si $A = \{x_1, ..., x_n\} \subset V$, $x \in V$ es combinación lineal de elementos de A si existen $\alpha_1, ..., \alpha_n \in \mathbb{R}$ tales que $x = \sum_{i=1}^{n} \alpha_i x_i$.

Teorema 4 *Sea A un subconjunto no vacío de V. El subconjunto W de V constituído por todas las combinaciones lineales de elementos de A es un subespacio de V. Este subconjunto W de V se denomina subespacio generado por A. Escribiremos $W = L(A)$.*

Definición 14 *Sea A un subconjunto no vacío de un espacio vectorial V. Si $L(A) = V$ diremos que A genera a V o que A es un conjunto generador de V.*

Si $A \subset V$ y $L(A) = V$, entonces $x \in V$ si y solo si existen $\alpha_1, ..., \alpha_n \in \mathbb{R}$, $x_1, ..., x_n \in A$ tales que $x = \sum_{i=1}^{n} \alpha_i x_i$. Particularmente, si $A = \{x_1, ..., x_n\} \subset V$, escribiremos explícitamente al subespacio generado por A como

$$W = L(x_1, ..., x_n) = \left\{ \sum_{i=1}^{n} \alpha_i x_i \,|\alpha_i \in \mathbb{R}, \quad i = 1, ..., n \right\}.$$

Si $W = V$, escribiremos $V = L(x_1, ..., x_n)$ y diremos que $\{x_1, ..., x_n\}$ es un conjunto generador de V.

Definición 15 *Sea A un subconjunto no vacío de un espacio vectorial V.*

i. Se dice que A es linealmente independiente si para todo $x_1, .., x_n \in A$, $\sum_{i=1}^{n} \alpha_i x_i = 0 \Rightarrow \alpha_i = 0, \quad i = 1, ..., n$.

ii. Se dice que A es linealmente dependiente si A no es linealmente independiente.

De la definición de dependencia lineal se sigue que A es linealmente dependiente si existen $x_1, ..., x_n \in A$ y $\alpha_1, ..., \alpha_n \in \mathbb{R}$ no todos nulos tales que $\sum_{i=1}^{n} \alpha_i x_i = 0$.

Sean A, B dos subconjuntos no vacíos de un espacio vectorial V tales que $A \subset B$. Entonces, si A es linealmente dependiente, B también lo es; y , si B es linealemente independiente, A también lo es.

Definición 16 *Se dice que un subconjunto $\mathcal{B}$ de un espacio vectorial V es una base de V si y solo si se satisfacen las dos condiciones siguientes:*

i. $\mathcal{B}$ es linealmente independiente.

ii. $\mathcal{B}$ genera a V.

Definición 17

i. Un espacio vectorial real V es de dimensión finita n si toda base $\mathcal{B}$ de V está constituida por exactamente n elementos. Al único número natural n se le llama dimensión de V y se le denota $\dim V$, esto es $\dim V = n$.

ii. Se dice que un espacio vectorial real V es de dimensión infinita si cualquier base $\mathcal{B}$ de V tiene un número infinito o numerable de elementos.

Ejemplos

1. El espacio $\mathbb{R}^n$ es un espacio vectorial de dimensión finita n. La base $\mathcal{B} = \{\overrightarrow{e}_1, ..., \overrightarrow{e}_n\}$ se conoce como base canónica de $\mathbb{R}^n$, donde

$$\overrightarrow{e}_1 = (1, 0, ..., 0), \cdots, \overrightarrow{e}_n = (0, ..., 0, 1).$$

Sea $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Se tiene $\overrightarrow{x} = \sum_{i=1}^{n} x_i \overrightarrow{e}_i = x_1 \overrightarrow{e}_1 + ... + x_n \overrightarrow{e}_n$.

Para $n = 2$, el conjunto $\mathcal{B} = \left\{ \overrightarrow{i}, \overrightarrow{j} \right\}$ con $\overrightarrow{i} = (1, 0)$, $\overrightarrow{j} = (0, 1)$, es la base canónica de $\mathbb{R}^2$. Note que $\overrightarrow{x} = (a, b) \in \mathbb{R}^2$ se escribe en la forma $\overrightarrow{x} = a\overrightarrow{i} + b\overrightarrow{j}$.

Para $n = 3$, el conjunto $\mathcal{B} = \left\{ \overrightarrow{i}, \overrightarrow{j}, \overrightarrow{k} \right\}$ con $\overrightarrow{i} = (1, 0, 0)$, $\overrightarrow{j} = (0, 1, 0)$, $\overrightarrow{k} = (0, 0, 1)$, es la base canónica de $\mathbb{R}^3$. Si $\overrightarrow{x} = (a, b, c) \in \mathbb{R}^3$ entonces $\overrightarrow{x} = a\overrightarrow{i} + b\overrightarrow{j} + c\overrightarrow{k}$.

2. Sea $V = C([0, 2\pi])$. Consideremos las funciones $\varphi_0, \varphi_1, ..., \varphi_n$ definidas en $[0, 2\pi]$ como sigue:

$$\varphi_0(x) = 1, \ \ \varphi_1(x) = \operatorname{sen} x, \ \ \varphi_2(x) = \operatorname{sen}(2x), ..., \varphi_n(x) = \operatorname{sen}(nx).$$

El conjunto $\mathcal{B} = \{\varphi_0, \varphi_1, ..., \varphi_n\}$ es linealmente independiente y genera un espacio W constituido por todas las combinaciones lineales de $\varphi_0, \varphi_1, .., \varphi_n$, esto es

$$\begin{aligned} W &= \left\{ \sum_{i=1}^{n} \alpha_i \varphi_i \,|\alpha_i \in \mathbb{R}, \ \ i = 0, 1, ..., n \right\} \\ f &\in W \Leftrightarrow f(x) = \alpha_0 + \alpha_1 \operatorname{sen}(x) + ... + \alpha_n \operatorname{sen}(nx), \quad x \in [0, 2\pi] \end{aligned}$$

donde $\alpha_0, ..., \alpha_n \in \mathbb{R}$ son elegidos apropiadamente.

3. El espacio vectorial de matrices de $M_{m\times n}[\mathbb{R}]$ es de dimensión finita $m \times n$. La base canónica $\mathcal{B}$ de $M_{m\times n}[\mathbb{R}]$ está formada por las matrices $A_1 = \left(a_{ij}^{(1)}\right), ..., A_{m\times n} = \left(a_{ij}^{(m\times n)}\right)$ donde

$$a_{ij}^{(1,1)} = \begin{cases} 1, & \text{si } \ i = 1, \, j = 1, \\ 0, & \text{si } 1 < i \leq m, \ \ 1 < j \leq n, \end{cases} \qquad \cdots \qquad a_{ij}^{(m\times n)} = \begin{cases} 1, & \text{si } \ i = m, \, j = n, \\ 0, & \text{si } 1 \leq i < m, \ \ 1 \leq j < n. \end{cases}$$

Por ejemplo si $m = 2$, $n = 3$, la base canónica del espacio vectorial de matrices $M_{2\times 3}[\mathbb{R}]$ está formada por las siguientes matrices:

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad A_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad A_6 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Sea $A = (a_{ij}) \in M_{2\times 3}[\mathbb{R}]$, entonces $A = a_{11}A_1 + ... + a_{23}A_6$.

4. El espacio vectorial $C([a, b])$ de funciones continuas en $[a, b]$ es de dimensión infinita.

5. El espacio vectorial $\mathcal{I}([a, b])$ de funciones integrables en $[a, b]$ es de dimensión infinita.

6. El espacio $\mathbb{K}_n([a, b])$ es de dimensión finita $n + 1$. La base canónica de $\mathbb{K}_n([a, b])$ está constituido por el conjunto de funciones $\{P_0, P_1, ..., P_n\}$ con $P_0(x) = 1, \ \ P_j(x) = x^j \quad x \in [a, b], \ \ j = 1, ..., n$.

12.2. Definición de espacio normado.

Definición 18 *Sea V un espacio vectorial sobre $\mathbb{R}$. Una norma en V es una función N de V en $\mathbb{R}$ que satisface las siguientes propiedades:*

i. $N(x) \geq 0 \quad \forall x \in V$,

ii. $N(x) = 0 \Leftrightarrow x = 0$,

iii. $N(\lambda x) = |\lambda| N(x) \quad \forall \lambda \in \mathbb{R}, \forall x \in V$,

iv. $N(x + y) \leq N(x) + N(y) \quad \forall x, y \in V$ *(desigualdad triangular).*

El número real no negativo $N(x)$ se llama norma de x. El par (V, N) se llama espacio normado.

Observación

Si la función N de V en $\mathbb{R}$ verifica las propiedades i), iii) y iv) de la definición de norma, pero no se verifica ii), la función N se dice seminorma en V.

Note en iv) que $x + y$ es la suma de los elementos x, y de V, mientras que $N(x) + N(y)$ es la suma de los números reales no negativos $N(x)$ y $N(y)$. En iii), λx es el producto del escalar λ (número real) por el elemento x de V y $|\lambda| N(x)$ es el producto de los números reales no negativos $|\lambda|$ y $N(x)$.

En ii), $x = 0$ denota el elemento neutro o nulo de V y $N(x) = 0$ es el elemento neutro o nulo de $\mathbb{R}$.

Notación

Si N es una norma en V, es usual escribir esta función con el símbolo $\|\cdot\|$ en vez de N y el espacio normado se escribirá $(V, \|\cdot\|)$ o se dirá V espacio normado provisto de la norma $\|\cdot\|$. Para $x \in V$, la norma de x se escribirá $\|x\|$.

Si en V se han definido varias normas, es preciso señalar que norma se está utilizando.

Proposición 5 *Sea V un espacio normado con $\|\cdot\|$ su norma. Se verifican las siguientes propiedades:*

i. $\|x - y\| = \|y - x\| \quad \forall x, y \in V.$

ii. $|\,\|x\| - \|y\|\,| \leq \|x - y\| \quad \forall x, y \in V.$

Demostración.

i. Sean $x, y \in V$. Entonces $\|x - y\| = \|(-1)(x - y)\| = |-1| \|y - x\| = \|y - x\|$.

ii. Sean $x, y \in V$. De la desigualdad triangular, se tiene $\|x\| = \|(x - y) + y\| \leq \|x - y\| + \|y\|$, de donde $\|x\| - \|y\| \leq \|x - y\|$. Además,

$$\|y\| = \|(y - x) + x\| \leq \|y - x\| + \|x\| = \|x - y\| + \|x\|$$

y de esta desigualdad se obtiene la siguiente: $\|y\| - \|x\| \leq \|x - y\|$ que multiplicándola por -1 resulta $-\|x - y\| \leq \|x\| - \|y\|$.

Por lo tanto, $-\|x - y\| \leq \|x\| - \|y\| \leq \|x - y\|$, que es equivalente a la desigualdad $|\,\|x\| - \|y\|\,| \leq \|x - y\|$.

Nota: Recuerde que si $a \geq 0$, $|t| \leq a \Leftrightarrow -a \leq t \leq a$.

■

12.3. Ejemplos de espacios normados.

Comenzamos esta sección considerando el ejemplo más simple de espacio normado: el espacio vectorial $\mathbb{R}$ provisto de la función valor absoluto $|\cdot|$.

Sea $V = \mathbb{R}$. Se define la función $\|\cdot\|$ de $\mathbb{R}$ en $\mathbb{R}$ como sigue: $\|x\| = |x| \quad \forall x \in \mathbb{R}$. Entonces, la función $\|\cdot\|$ definida en $\mathbb{R}$ es una norma en $\mathbb{R}$. La verificación de las propiedades i) a iv) siguen inmediatamente de las propiedades del valor absoluto siguientes:

i. $|x| \geq 0 \quad \forall x \in \mathbb{R}$.

ii. $|x| = 0 \Leftrightarrow x = 0$.

iii. $|\lambda x| = |\lambda| |x| \quad \forall \lambda, x \in \mathbb{R}$.

iv. $|x + y| \leq |x| + |y| \quad \forall x, y \in \mathbb{R}$, (desigualdad triangular).

12.3.1. Normas en $\mathbb{R}^n$.

En el espacio vectorial real $\mathbb{R}^n$ se consideran dos normas importantes: la del máximo que se denota $\|\cdot\|_\infty$ y las hölderianas $\|\cdot\|_p$ con $p \in [1,\ \infty[$.

Norma $\|\cdot\|_\infty$.

Sea $V = \mathbb{R}^n$. Se define la función $\|\cdot\|_\infty$ de $\mathbb{R}^n$ en $\mathbb{R}$ como $\|\overrightarrow{x}\|_\infty = \underset{i=1,...,n}{\text{Max}} |x_i| \quad \forall \overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Se verifica que $\|\cdot\|_\infty$ es una norma en $\mathbb{R}^n$.

i) Es claro que para todo $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, $\|\overrightarrow{x}\|_\infty \geq 0$.

ii) Si $\overrightarrow{x} = \overrightarrow{0} = (0, ..., 0) \in \mathbb{R}^n$, se tiene $\|\overrightarrow{x}\|_\infty = 0$. Recíprocamente, si $\|\overrightarrow{x}\|_\infty = 0$ se sigue que $\underset{i=1,...,n}{\text{Max}} |x_i| = 0$ y como $0 \leq |x_i| \leq \underset{i=1,...,n}{\text{Max}} |x_i| = 0 \quad i = 1, ..., n$, resulta que $x_i = 0,\ \ i = 1, ..., n$, esto es, $\overrightarrow{x} = \overrightarrow{0}$. Así, $\|\overrightarrow{x}\|_\infty = 0 \Leftrightarrow \overrightarrow{x} = \overrightarrow{0}$.

iii) Sean $\lambda \in \mathbb{R}$, $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Entonces $\lambda\overrightarrow{x} = (\lambda x_1, ..., \lambda x_n)$, y

$$\|\lambda\overrightarrow{x}\|_\infty = \underset{i=1,...,n}{\text{Max}} |\lambda x_i| = \underset{i=1,...,n}{\text{Max}} |\lambda|\,|x_i| = |\lambda| \underset{i=1,...,n}{\text{Max}} |x_i| = |\lambda|\,\|\overrightarrow{x}\|_\infty.$$

Luego $\|\lambda\overrightarrow{x}\|_\infty = |\lambda|\,\|\overrightarrow{x}\|_\infty$.

iv) Sean $\overrightarrow{x} = (x_1, ..., x_n), \quad \overrightarrow{y} = (y_1, ..., y_n) \in \mathbb{R}^n$. Puesto que

$$\overrightarrow{x} + \overrightarrow{y} = (x_1 + y_1, ..., x_n + y_n), \qquad |x_i + y_i| \leq |x_i| + |y_i| \quad i = 1, ..., n,$$

y de la definición de la función $\|\cdot\|_\infty$, se sigue

$$\begin{aligned} \|\overrightarrow{x} + \overrightarrow{y}\|_\infty &= \underset{i=1,...,n}{\text{Max}} |x_i + y_i| \leq \underset{i=1,...,n}{\text{Max}} \{|x_i| + |y_i|\} \\ &= \underset{i=1,...,n}{\text{Max}} |x_i| + \underset{i=1,...,n}{\text{Max}} |y_i| = \|\overrightarrow{x}\|_\infty + \|\overrightarrow{y}\|_\infty. \end{aligned}$$

Luego, $\|x + y\|_\infty \leq \|x\|_\infty + \|y\|_\infty$.

Conclusión: $\|\cdot\|_\infty$ es una norma sobre $\mathbb{R}^n$.

Sean $n = 2$, $\overrightarrow{x} = (x, y) \in \mathbb{R}^2$, la norma $\|\cdot\|_\infty$ en $\mathbb{R}^2$ está definida como $\|\overrightarrow{x}\|_\infty = \text{Max}\{|x|, |y|\}$. Note que $|x| \leq \|\overrightarrow{x}\|_\infty$ y $|y| \leq \|\overrightarrow{x}\|_\infty$.

Sean $n = 3$, $\overrightarrow{x} = (x, y, z) \in \mathbb{R}^3$, la norma $\|\cdot\|_\infty$ en $\mathbb{R}^3$ está definida como $\|\overrightarrow{x}\|_\infty = \text{Max}\{|x|, |y|, |z|\}$. Además, se verifican las siguientes desigualdades: $|x| \leq \|\overrightarrow{x}\|_\infty, \quad |y| \leq \|\overrightarrow{x}\|_\infty, \quad |z| \leq \|\overrightarrow{x}\|_\infty$.

Si $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Se tiene $|x_i| \leq \|\overrightarrow{x}\|_\infty, \quad i = 1, ..., n$.

Normas hölderianas en el espacio $\mathbb{R}^n$.

Sea $p \in [1, \infty[$. Se define la función $\|\cdot\|_p$ de $\mathbb{R}^n$ en $\mathbb{R}$ como sigue:

$$\|\overrightarrow{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad \forall \overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n.$$

Entonces $\|\cdot\|_p$ es una norma en $\mathbb{R}^n$, llamada norma de Hölder o norma hölderiana. Verifiquemos las propiedades i) a iv) de la definición de norma.

i)Sea $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Puesto que $|x_i| \geq 0 \quad i = 1, ..., n$, de la definición de $\|\cdot\|_p$, se sigue $\|x\|_p \geq 0$.

ii) Si $\overrightarrow{x} = \overrightarrow{0} = (0, ..., 0)$ se tiene $\|\overrightarrow{x}\|_p = 0$. Supongamos que $\|\overrightarrow{x}\|_p = 0$ entonces $\sum_{i=1}^{n} |x_i|^p = 0$. Se tiene la siguiente desigualdad: $0 \leq |x_i|^p \leq \sum_{i=1}^{n} |x_i|^p = 0 \quad i = 1, ..., n$, consecuentemente $x_i = 0,\ \ i = 1, ..., n$, o sea $\overrightarrow{x} = (0, ..., 0)$. Luego, $\|\overrightarrow{x}\|_p = 0 \Leftrightarrow \overrightarrow{x} = \overrightarrow{0}$.

iii) Sean $\lambda \in \mathbb{R}$, $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Entonces $\lambda \overrightarrow{x} = (\lambda x_1, ..., \lambda x_n)$ y por la definición de $\|\cdot\|_p$, se tiene

$$\begin{aligned} \|\lambda \overrightarrow{x}\|_p &= \left(\sum_{i=1}^{n} |\lambda x_i|^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} |\lambda|^p |x_i|^p \right)^{\frac{1}{p}} = \left(|\lambda|^p \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \\ &= |\lambda| \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} = |\lambda| \, \|\overrightarrow{x}\|_p . \end{aligned}$$

Por lo tanto, $\|\lambda \overrightarrow{x}\|_p = |\lambda| \, \|\overrightarrow{x}\|_p$.

iv) Para probar la desigualdad triangular, se requieren de dos resultados preliminares: la desigualdad de Young y la desigualdad de Hölder.

Desigualdad de Young.

Esta se establece en los siguientes términos: sean $p, q \in]1, \infty[$ tales que $\frac{1}{p} + \frac{1}{q} = 1$, $\alpha \geq 0$, $\beta \geq 0$. Entonces

$$\alpha^{\frac{1}{p}} \beta^{\frac{1}{q}} \leq \frac{1}{p}\alpha + \frac{1}{q}\beta.$$

Para probar esta desigualdad, estudiemos la función siguiente:

$$f : \begin{cases} [0, \infty[& \to \mathbb{R} \\ x & \to \quad f(x) = \frac{1}{q} + \frac{1}{p}x - x^{\frac{1}{p}}. \end{cases}$$

Para $x > 0$ calculemos la derivada $f'(x)$ y determinemos los puntos críticos y los intervalos donde $f'(x) > 0$, $f'(x) < 0$, tenemos

$$\begin{aligned} f'(x) &= \frac{1}{p} - \frac{1}{p} x^{\frac{1}{p}-1} = \frac{1}{p} - \frac{1}{p} x^{-\frac{1}{q}} = \frac{1}{p}\left(1 - x^{-\frac{1}{q}}\right), \\ f'(x) &= 0 \Leftrightarrow x^{-\frac{1}{q}} = 1 \Leftrightarrow x = 1, \\ f'(x) &> 0 \Leftrightarrow 1 - x^{-\frac{1}{q}} > 0 \Leftrightarrow x > 1, \\ f'(x) &< 0 \Rightarrow 0 < x < 1. \end{aligned}$$

La función f es decreciente sobre $]0, 1[$ y creciente si $x \geq 1$. Puesto que $f(1) = 0$, la función f tiene un mínimo local en $x = 1$. Además, $f(0) = \frac{1}{q} > 0$ y $f(x) \underset{x \to \infty}{\to} +\infty$. Luego $f(1) = 0$ es un mínimo global. En consecuencia, para todo $x \geq 0$, $f(x) \geq 0$. En particular, para $x \geq 1$ y siendo f creciente, se tiene

$$0 = f(1) \leq f(x) = \frac{1}{q} + \frac{1}{p}x - x^{\frac{1}{p}}$$

y de esta desigualdad se obtiene $x^{\frac{1}{p}} \leq \frac{1}{q} + \frac{1}{p}x$.

Para $\alpha = 0$ o $\beta = 0$, la desigualdad de Young se verifica trivialmente. Supongamos que $\alpha > 0$, $\beta > 0$ y $x = \frac{\alpha}{\beta} \geq 1$ si $\alpha \geq \beta$ (en el caso contrario ponemos $x = \dfrac{\beta}{\alpha} \geq 1$). Remplazando $x = \frac{\alpha}{\beta}$ en la desigualdad precedente, resulta $\left(\frac{\alpha}{\beta}\right)^{\frac{1}{p}} \leq \frac{1}{q} + \frac{1}{p}\frac{\alpha}{\beta}$, luego $\alpha^{\frac{1}{p}} \beta^{-\frac{1}{p}+1} \leq \frac{1}{p}\alpha + \frac{1}{q}\beta$. Como $\frac{1}{p} + \frac{1}{q} = 1$, se tiene $\frac{1}{q} = 1 - \frac{1}{p}$, con lo que $\alpha^{\frac{1}{p}} \beta^{\frac{1}{q}} \leq \frac{1}{p}\alpha + \frac{1}{q}\beta$.

Desigualdad de Hölder.

Esta desigualdad se expresa como a continuación se indica:

$$\sum_{i=1}^{n} |x_i y_i| \leq \|\overrightarrow{x}\|_p \, \|\overrightarrow{y}\|_q \qquad \forall \overrightarrow{x} = (x_1, ..., x_n), \overrightarrow{y} = (y_1, ..., y_n) \in \mathbb{R}^n,$$

donde $p,\, q \in]1,\infty[$ tales que $\dfrac{1}{p}+\dfrac{1}{q}=1$.

Si $\overrightarrow{x}=0$ o $\overrightarrow{y}=0$, la desigualdad de Hölder se verifica trivialmente. Supongamos $\overrightarrow{x}\neq 0,\ \ \overrightarrow{y}\neq 0$.

Sean $\alpha=\dfrac{|x_i|^p}{\|\overrightarrow{x}\|_p^p},\quad \beta=\dfrac{|y_i|^q}{\|\overrightarrow{y}\|_q^q}$. Apliquemos la desigualdad de Young. Resulta

$$\left(\frac{|x_i|^p}{\|\overrightarrow{x}\|_p^p}\right)^{\frac{1}{p}}\left(\frac{|y_i|^q}{\|\overrightarrow{y}\|_q^q}\right)^{\frac{1}{q}}\leq\frac{1}{p}\frac{|x_i|^p}{\|\overrightarrow{x}\|_p^p}+\frac{1}{q}\frac{|y_i|^q}{\|\overrightarrow{y}\|_q^q}\qquad i=1,...,n,$$

con lo cual

$$\frac{|x_i|}{\|\overrightarrow{x}\|_p}\frac{|y_i|}{\|\overrightarrow{y}\|_q}\leq\frac{1}{p}\frac{|x_i|^p}{\|\overrightarrow{x}\|_p^p}+\frac{1}{q}\frac{|y_i|^q}{\|\overrightarrow{y}\|_q^q}\qquad i=1,...,n.$$

Sumando de 1 a n en cada miembro de la última desigualdad, obtenemos

$$\frac{1}{\|\overrightarrow{x}\|_p}\frac{1}{\|\overrightarrow{y}\|_q}\sum_{1=n}^{n}|x_i|\,|y_i|\leq\frac{1}{p\,\|\overrightarrow{x}\|_p^p}\sum_{i=1}^{n}|x_i|^p+\frac{1}{q\,\|\overrightarrow{y}\|_q^q}\sum_{i=1}^{n}|y_i|^q\,.$$

Puesto que $\|\overrightarrow{x}\|_p^p=\sum\limits_{i=1}^{n}|x_i|^p\,,\quad \|\overrightarrow{y}\|_q^q=\sum\limits_{i=1}^{n}|y_i|^q$, entonces $\frac{1}{\|\overrightarrow{x}\|_p}\frac{1}{\|\overrightarrow{y}\|_q}\sum_{1=n}^{n}|x_i|\,|y_i|\leq\frac{1}{p}+\frac{1}{q}=1$, y de esta desigualdad se deduce la lesigualdad de Hölder: $\sum_{1=n}^{n}|x_iy_i|\leq\|\overrightarrow{x}\|_p\,\|\overrightarrow{y}\|_q\,.$

Desigualdad triangular**:** $\|\overrightarrow{x}+\overrightarrow{y}\|_p\leq\|\overrightarrow{x}\|_p+\|\overrightarrow{y}\|_p\quad\forall\overrightarrow{x},\,\overrightarrow{y}\in\mathbb{R}^n$.

De la definición de $\|\cdot\|_p$, obtenemos

$$\begin{aligned}\|\overrightarrow{x}+\overrightarrow{y}\|_p^p &= \sum_{i=1}^{n}|x_i+y_i|^p=\sum_{i=1}^{n}|x_i+y_i|^{p-1}\,|x_i+y_i|\leq\sum_{i=1}^{n}|x_i+y_i|^{p-1}\,(|x_i|+|y_i|)\\ &= \sum_{i=1}^{n}|x_i+y_i|^{p-1}\,|x_i|+\sum_{i=1}^{n}|x_i+y_i|^{p-1}\,|y_i|\,.\end{aligned}$$

Apliquemos la desigualdad de Hölder a cada sumando del lado derecho. Ya que $\frac{1}{p}+\frac{1}{q}=1$, se obtiene $p+q=pq$ con lo que $p=q\,(p-1)$. Luego

$$\begin{aligned}\sum_{i=q}^{n}|x_i+y_i|^{p-1}\,|x_i| &\leq \left(\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}}\left(\sum_{i=1}^{n}|x_i+y_i|^{(p-1)q}\right)^{\frac{1}{q}}=\|\overrightarrow{x}\|_p\left(\sum_{i=1}^{n}|x_i+y_i|^p\right)^{\frac{1}{q}}\\ &= \|\overrightarrow{x}\|_p\left(\|\overrightarrow{x}+\overrightarrow{y}\|_p^p\right)^{\frac{1}{q}},\end{aligned}$$

Análogamente, $\sum_{i=1}^{n}|x_i+y_i|^{p-1}\,|y_i|\leq\|\overrightarrow{y}\|_p\left(\|\overrightarrow{x}+\overrightarrow{y}\|_p^p\right)^{\frac{1}{q}}$. Por lo tanto

$$\begin{aligned}\|\overrightarrow{x}+\overrightarrow{y}\|_p^p &\leq \sum_{i=1}^{n}|x_i+y_i|^{p-1}\,|x_i|+\sum_{i=1}^{n}|x_i+y_i|^{p-1}\,|y_i|\\ &\leq \|\overrightarrow{x}\|_p\left(\|\overrightarrow{x}+\overrightarrow{y}\|_p^p\right)^{\frac{1}{q}}+\|\overrightarrow{y}\|_p\left(\|\overrightarrow{x}+\overrightarrow{y}\|_p^p\right)^{\frac{1}{q}}\\ &= \left(\|\overrightarrow{x}\|_p+\|\overrightarrow{y}\|_p\right)\|\overrightarrow{x}+\overrightarrow{y}\|_p^{\frac{p}{q}},\end{aligned}$$

y de esta desigualdad, se deduce $\|\overrightarrow{x}+\overrightarrow{y}\|_p^{p-\frac{p}{q}}\leq\|\overrightarrow{x}\|_p+\|\overrightarrow{y}\|_p$. Nuevamente $\frac{1}{p}+\frac{1}{q}=1$ entonces $1=p-\frac{p}{q}$, con lo que se obtiene la desigualdad triangular: $\|\overrightarrow{x}+\overrightarrow{y}\|_p\leq\|\overrightarrow{x}\|_p+\|\overrightarrow{y}\|_p\,.$

Para $n=2,\ \ \overrightarrow{x}=(x,y)\in\mathbb{R}^2$ la norma de Hölder está definida como sigue:

$$\|\overrightarrow{x}\|_p=(|x|^p+|y|^p)^{\frac{1}{p}}\quad\text{con}\quad p\in[1,\infty[\,.$$

Así, para $p=1$, la norma $\|\cdot\|_1$ está definida como $\|\overrightarrow{x}\|_1 = |x|+|y|$. Para $p=2$, la norma $\|\cdot\|_2$ se escribe como $\|\overrightarrow{x}\|_2 = \left(x^2+y^2\right)^{\frac{1}{2}}$. Esta se conoce como norma euclídea. Para $p=3$, la norma $\|\cdot\|_3$ se escribe como $\|\overrightarrow{x}\|_3 = \left(|x|^3+|y|^3\right)^{\frac{1}{3}}$.

Para $n=3$, $\overrightarrow{x}=(x,y,z)\in\mathbb{R}^3$ la norma de Hölder $\|\cdot\|_p$ está definida como

$$\|\overrightarrow{x}\|_p = \left(|x|^p+|y|^p+|z|^p\right)^{\frac{1}{p}} \text{ con } p\in[1,\infty[\,.$$

Particularmente, para $p=1$, $\|\overrightarrow{x}\|_1 = |x|+|y|+|z|$. Para $p=2$, $\|\overrightarrow{x}\|_2 = \left(x^2+y^2+z^2\right)^{\frac{1}{2}}$. Esta es la norma euclídea. Para $p=3{,}5=\frac{7}{2}$, $\|\overrightarrow{x}\|_{\frac{7}{2}} = \left(|x|^{\frac{7}{2}}+|y|^{\frac{7}{2}}+|z|^{\frac{7}{2}}\right)^{\frac{2}{7}}$.

Consecuencias

1. Para $p=1$ la norma $\|\cdot\|_1$ está definida como $\|\overrightarrow{x}\|_1 = \sum_{i=1}^n |x_i|$ $\overrightarrow{x}=(x_1,...,x_n)\in\mathbb{R}^n$. Además, para $\overrightarrow{y}=(y_1,...,y_n)\in\mathbb{R}^n$, se verifica la desigualdad de Hölder para $p=1$ y $q=\infty$:

$$\sum_{i=1}^n |x_iy_i| \le \|\overrightarrow{x}\|_1 \|\overrightarrow{y}\|_\infty .$$

En efecto, de la desigualdad $|y_i| \le \underset{i=1,n}{\text{Max}}|y_i| = \|\overrightarrow{y}\|_\infty$ $i=1,\dots,n$, se sigue que

$$\sum_{i=1}^n |x_iy_i| = \sum_{i=1}^n |x_i|\,|y_i| \le \sum_{i=1}^n \left(|x_i|\,\|\overrightarrow{y}\|_\infty\right) \le \|\overrightarrow{y}\|_\infty \sum_{i=1}^n |x_i| = \|\overrightarrow{y}\|_\infty \|\overrightarrow{x}\|_1 .$$

Así, la desigualdad de Hölder es válida para $p=1$ y $q=\infty$.

2. Sean $p\in\,]1,\infty[$ y $\overrightarrow{x}=(x_1,...,x_n)\in\mathbb{R}^n$ con $\overrightarrow{x}\neq 0$. Mostremos que $\lim\limits_{p\to\infty}\|\overrightarrow{x}\|_p = \|\overrightarrow{x}\|_\infty$. Primeramente $\|\overrightarrow{x}\|_\infty \le \|\overrightarrow{x}\|_p$ $\forall\overrightarrow{x}\in\mathbb{R}^n$. En efecto,

$$\|\overrightarrow{x}\|_\infty = \underset{i=1,\dots,n}{\text{Max}}|x_i| \le \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} = \|\overrightarrow{x}\|_p .$$

Por otro lado,

$$\begin{aligned}\|\overrightarrow{x}\|_p &= \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} \le \left(\sum_{i=1}^n \left(\underset{i=1,\dots,n}{\text{Max}}|x_i|\right)^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^n \|\overrightarrow{x}\|_\infty^p\right)^{\frac{1}{p}} \\ &= \|\overrightarrow{x}\|_\infty \left(\sum_{i=1}^n 1\right)^{\frac{1}{p}} = n^{\frac{1}{p}}\|\overrightarrow{x}\|_\infty .\end{aligned}$$

De los dos resultados previos, se deduce la siguiente desigualdad: $\|\overrightarrow{x}\|_\infty \le \|\overrightarrow{x}\|_p \le n^{\frac{1}{p}}\|\overrightarrow{x}\|_\infty$. Tomando límite cuando $p\to\infty$, se obtiene

$$\lim_{p\to\infty}\|\overrightarrow{x}\|_\infty \le \lim_{p\to\infty}\|\overrightarrow{x}\|_p \le \lim_{p\to\infty} n^{\frac{1}{p}}\|\overrightarrow{x}\|_\infty ,$$

y como $\lim\limits_{p\to\infty} n^{\frac{1}{p}} = 1$, se deduce $\|\overrightarrow{x}\|_\infty \le \lim_{p\to\infty}\|\overrightarrow{x}\|_p \le \|\overrightarrow{x}\|_\infty$. Así, $\lim_{p\to\infty}\|\overrightarrow{x}\|_p = \|\overrightarrow{x}\|_\infty$.

3. Relación entre $\|\cdot\|_p$ y $\|\cdot\|_q$ con $1 \leq p < q$. Mostremos que $\|\overrightarrow{x}\|_p \leq n^{\frac{q-p}{pq}} \|\overrightarrow{x}\|_q \quad \forall \overrightarrow{x} \in \mathbb{R}^n$ con $q > p \geq 1$.

   Sean $p, q \in [1, \infty[$ con $q > p$. Sea $r = \dfrac{q}{p} > 1$ y $s \in]1, \infty[$ tal que $\dfrac{1}{r} + \dfrac{1}{s} = 1$. Por la desigualdad de Hölder, para cada $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ se tiene

$$\begin{aligned} \|\overrightarrow{x}\|_p^p &= \sum_{i=1}^n |x_i|^p \leq \left(\sum_{i=1}^n 1^s\right)^{\frac{1}{s}} \left(\sum_{i=1}^n (|x_i|^p)^r\right)^{\frac{1}{r}} = n^{\frac{1}{s}} \left(\sum_{i=1}^n |x_i|^{pr}\right)^{\frac{1}{r}} = n^{\frac{1}{s}} \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{1}{r}} \\ &= n^{\frac{1}{s}} \left(\|\overrightarrow{x}\|_q^q\right)^{\frac{1}{r}} = n^{\frac{1}{s}} \|\overrightarrow{x}\|_q^{\frac{q}{r}} = n^{\frac{1}{s}} \|\overrightarrow{x}\|_q^p . \end{aligned}$$

   Note que $r = \frac{q}{p}$ entonces $pr = q$ y $\frac{q}{r} = p$. Se tiene $\|\overrightarrow{x}\|_p^p \leq n^{\frac{1}{s}} \|\overrightarrow{x}\|_q^p$ y tomando la raíz p-ésima se deduce $\|\overrightarrow{x}\|_p \leq n^{\frac{1}{sp}} \|\overrightarrow{x}\|_q$. Como $\dfrac{1}{r} + \dfrac{1}{s} = 1$ y $r = \dfrac{q}{p}$ se sigue que $s = \frac{q}{q-p}$ y en consecuencia $\|\overrightarrow{x}\|_p \leq n^{\frac{q-p}{pq}} \|\overrightarrow{x}\|_q$.

   Así, si $p, q \in [1, \infty[$ tales que $q > p \geq 1$, $\|\overrightarrow{x}\|_p \leq n^{\frac{q-p}{pq}} \|\overrightarrow{x}\|_q \quad \forall \overrightarrow{x} \in \mathbb{R}^n$.

4. Si $p = 2$ y $\overrightarrow{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, la norma $\|\cdot\|_2$ viene dada por $\|\overrightarrow{x}\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{1}{2}}$ que se conoce con el nombre de norma euclídea. Además, para $p = q = 2$, $\overrightarrow{x} = (x_1, ..., x_n)$, $\overrightarrow{y} = (y_1, ..., y_n) \in \mathbb{R}^2$, la desigualdad de Hölder se escribe $\sum_{i=1}^n |x_i y_i| \leq \|\overrightarrow{x}\|_2 \|\overrightarrow{y}\|_2$, que coincide con la conocida desigualdad de Cauchy-Schwarz que se verá más adelante.

5. Sean $n \in \mathbb{Z}^+$, $\alpha_j \in \mathbb{R}^+$, $j = 1, ..., n$, las siguientes son normas en $\mathbb{R}^n$:

   a) $\|\overrightarrow{v}\| = \underset{j=1,...,n}{\text{Max}} \{\alpha_j |v_j|\} \quad \overrightarrow{v} = (v_1, ..., v_n) \in \mathbb{R}^n$.

   b) $\|\overrightarrow{v}\| = \sum_{j=1}^n \alpha_j |v_j| \quad \overrightarrow{v} = (v_1, ..., v_n) \in \mathbb{R}^n$.

   c) $\|\overrightarrow{v}\| = \left(\sum_{j=1}^n \alpha_j |v_j|^2\right)^{\frac{1}{2}} \quad \overrightarrow{v} = (v_1, ..., v_n) \in \mathbb{R}^n$.

   d) $\|\overrightarrow{v}\| = \underset{j=1,...,n}{\text{Max}} \left\{\sum_{i=1}^j \alpha_j |v_j|\right\} \quad \overrightarrow{v} = (v_1, ..., v_n) \in \mathbb{R}^n$.

12.3.2. Normas geométricas de matrices.

Sean $V = \mathbb{R}^n$, $W = \mathbb{R}^m$, se designa con $\mathcal{B}_V = \{\overrightarrow{e}_1, ..., \overrightarrow{e}_n\}$ y $\mathcal{B}_W = \left\{\overrightarrow{f}_1, ..., \overrightarrow{f}_m\right\}$ las bases canónicas de $\mathbb{R}^n$ y $\mathbb{R}^m$, $p, q \in [1, \infty]$ y $\|\cdot\|_p$, $\|\cdot\|_q$ normas en $\mathbb{R}^n$ y $\mathbb{R}^m$, respectivamente. Se denota con $M_{m\times n}[\mathbb{R}]$ el espacio vectorial de las matrices reales de $m \times n$ y $A \in M_{m\times n}[\mathbb{R}]$. Se define la aplicación lineal T de $\mathbb{R}^n$ en $\mathbb{R}^m$ como

$$T(\overrightarrow{x}) = A\overrightarrow{x} \quad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

Entonces T es continua en todo punto $\overrightarrow{x_o} \in \mathbb{R}^n$. Más aún, debido a la linealidad de T, se tiene $T(\overrightarrow{x} - \overrightarrow{x}_o) = T(\overrightarrow{x}) - T(\overrightarrow{x}_o) \quad \forall \overrightarrow{x}, \overrightarrow{x}_o \in \mathbb{R}^n$, por lo que la continuidad de T en $\overrightarrow{x_o}$ es equivalente a la continuidad de T en el origen. Por lo tanto, dado $\epsilon > 0$, $\exists\ \delta > 0$ tal que $\forall \overrightarrow{x} \in \mathbb{R}^n$ con $\|\overrightarrow{x}\|_p < \delta \Longrightarrow \|T(\overrightarrow{x})\|_q < \epsilon$.

Sea $\overrightarrow{v} \in \mathbb{R}^n$ tal que $\|\overrightarrow{v}\|_p = 1$ y sea $\overrightarrow{x} = \frac{\delta}{2}\overrightarrow{v}$. Entonces,

$$\|\overrightarrow{x}\|_p = \left\|\frac{\delta}{2}\overrightarrow{v}\right\|_p = \frac{\delta}{2}\|\overrightarrow{v}\|_p = \frac{\delta}{2} < \delta,$$

y por la linealidad de T, se tiene $T(\overrightarrow{x}) = T(\frac{\delta}{2}\overrightarrow{v}) = \frac{\delta}{2}T(\overrightarrow{v})$, en consecuencia

$$\|T(\overrightarrow{x})\|_q = \frac{\delta}{2}\|T(\overrightarrow{v})\|_q < \epsilon,$$

de donde $\|T(\overrightarrow{v})\|_q < \frac{2\epsilon}{\delta} = M$.

Así, $\|T(\overrightarrow{v})\|_q \leq M \quad \forall \overrightarrow{v} \in \mathbb{R}^n$ con $\|\overrightarrow{v}\|_p = 1$, y de la definición de T se sigue que el conjunto

$\left\{\|A\overrightarrow{x}\|_q \mid \overrightarrow{x} \in \mathbb{R}^n \text{ con } \|\overrightarrow{x}\|_p = 1\right\}$ es acotado superiormente. Este resultado nos permite definir las normas geométricas de matrices que a continuación se propone.

Definición 19 *Sea $A = (a_{ij}) \in M_{m\times n}[\mathbb{R}]$. La norma geométrica de la matriz A se denota con $|||A|||$ y se define como $|||A||| = \sup_{\|\overrightarrow{x}\|_p \leq 1} \|A\overrightarrow{x}\|_q$.*

Se verifica inmediatamente que $|||\cdot|||$ es una norma en $M_{m\times n}[\mathbb{R}]$. La prueba se deja como ejercicio. Además, de la definición de norma geométrica de una matriz se sigue inmediatamente que para toda matriz $A \in M_{m\times n}[\mathbb{R}]$ se verifica la desigualdad siguiente:

$$\|A\overrightarrow{x}\|_q \leq |||A|||\,\|\overrightarrow{x}\|_p \quad \forall \overrightarrow{x} \in \mathbb{R}^n.$$

Teorema 6 *Sea $A = (a_{ij}) \in M_{m\times n}[\mathbb{R}]$. Entonces,*

i. $|||A|||_1 = \sup_{\|\overrightarrow{x}\|_1 \leq 1} \|A\overrightarrow{x}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$ *(máximo por columnas de A).*

ii. $|||A|||_\infty = \sup_{\|\overrightarrow{x}\|_\infty \leq 1} \|A\overrightarrow{x}\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$ *(máximo por filas de A).*

iii. $|||A|||_2 = \sup_{\|\overrightarrow{x}\|_2 \leq 1} \|A\overrightarrow{x}\|_2 = \left(\max_{i=1,\ldots,n} |\lambda_i|\right)^{\frac{1}{2}}$, *donde $\lambda_1, \ldots \lambda_n$ son los valores propios de A^TA y A^T denota la matriz transpuesta de A.*

Demostración.

i. a) Probemos que $|||A|||_1 \leq \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$. Sea $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$ tal que $\|\overrightarrow{x}\|_1 = \sum_{j=1}^{n} |x_j| = 1$. Entonces,

$$A\overrightarrow{x} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{bmatrix},$$

y por la definición de la norma $\|\cdot\|_1$ en $\mathbb{R}^m$, se sigue que

$$\begin{aligned} \|A\overrightarrow{x}\|_1 &= \sum_{i=1}^{m} \left(\left| \sum_{j=1}^{n} a_{ij}x_j \right| \right) \leq \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|\,|x_j| = \sum_{j=1}^{n} \left(|x_j| \sum_{i=1}^{m} |a_{ij}| \right) \\ &\leq \sum_{j=1}^{n} \left(|x_j| \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \right) = \left(\max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \right) \sum_{j=1}^{n} |x_j| = \left(\max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \right). \end{aligned}$$

Por lo tanto, de la definición de la norma geométrica $|||\cdot|||_1$, se obtiene la desigualdad siguiente:

$$|||A|||_1 = \sup_{\|\overrightarrow{x}\|_1 \leq 1} \|A\overrightarrow{x}\|_1 \leq \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|.$$

b) Probemos que $\max\limits_{j=1,...,n}\sum\limits_{i=1}^{m}|a_{ij}| \leq |||A|||_1$. Para el efecto, sea k la columna para la cual se verifica la igualdad $\max\limits_{j=1,...,n}\sum\limits_{i=1}^{m}|a_{ij}| = \sum\limits_{i=1}^{m}|a_{ik}|$. Para $\overrightarrow{x} = \overrightarrow{e_k}$, el k-ésimo vector de la base canónica de $\mathbb{R}^n$, se tiene $A\overrightarrow{e_k} = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{mk} \end{bmatrix}$, y en consecuencia $\|A\overrightarrow{e}_k\|_1 = \sum_{i=1}^{m}|a_{ik}| = \max_{j=1,...,n}\sum_{i=1}^{m}|a_{ik}|$.

Nuevamente, de la definición de la norma geométrica $|||\cdot|||_1$, se obtiene la desigualdad siguiente:

$$\max_{j=1,...,n}\sum_{i=1}^{n}|a_{ij}| = \|A\overrightarrow{e}_k\|_1 \leq \sup_{\|\overrightarrow{x}\|_1\leq 1}\|A\overrightarrow{x}\|_1 = |||A|||_1 .$$

De las desigualdades obtenidas en las partes a) y b) se deduce finalmente el resultado buscado:

$$|||A|||_1 = \sup_{\|\overrightarrow{x}\|\leq 1}\|A\overrightarrow{x}\|_1 = \max_{j=1,...,n}\sum_{i=1}^{n}|a_{ij}| .$$

ii. Primeramente, obtenemos la desigualdad $|||A|||_\infty \leq \max\limits_{i=1,...,m}\sum\limits_{j=1}^{n}|a_{ij}|$. Sea $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$ tal que $\|\overrightarrow{x}\|_\infty = \max\limits_{j=1,...,n}|x_j| = 1$. Entonces, $|x_j| \leq \|\overrightarrow{x}\|_\infty$, $j = 1, ..., n$, y

$$\|A\overrightarrow{x}\|_\infty = \max_{i=1,...,m}\left|\sum_{j=1}^{n}a_{ij}x_j\right| \leq \max_{i=1,...,m}\sum_{j=1}^{n}|a_{ij}|\,|x_j| \leq \max_{i=1,...,m}\sum_{j=1}^{n}|a_{ij}| ,$$

consecuentemente

$$\|A\overrightarrow{x}\|_\infty \leq \sup_{\|\overrightarrow{x}\|_\infty\leq 1}\|A\overrightarrow{x}\|_\infty = |||A|||_\infty \leq \max_{i=1,...,m}\sum_{j=1}^{n}|a_{ij}| .$$

Mostremos a continuación la desigualdad $\max\limits_{i=1,...,m}\sum\limits_{j=1}^{n}|a_{ij}| \leq |||A|||_\infty$. Para el efecto, sea k el índice para el cual la fila k-ésima de A es tal que $\max_{i=1,...,m}\sum_{j=1}^{n}|a_{ij}| = \sum_{j=1}^{n}|a_{kj}|$. Sea $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$ definido como sigue: $x_j = \begin{cases} \dfrac{|\,a_{kj}\,|}{a_{kj}}, & \text{si } a_{kj} \neq 0, \\ 0, & \text{si } a_{kj} = 0. \end{cases}$ Se tiene $\|\overrightarrow{x}\|_\infty = 1$, y

$$\|A\overrightarrow{x}\|_\infty = \max_{i=1,...,m}\left|\sum_{j=1}^{n}a_{ij}x_j\right| = \sum_{j=1}^{n}|a_{kj}| ,$$

de donde

$$\sum_{j=1}^{n}|a_{kj}| = \|A\overrightarrow{x}\|_\infty \leq \sup_{\|\overrightarrow{x}\|_\infty\leq 1}\|A\overrightarrow{x}\|_\infty = |||A|||_\infty .$$

Por lo tanto, $|||A|||_\infty = \max_{i=1,...,m}\sum_{j=1}^{n}|a_{ij}|$.

iii. Las normas euclídeas en $\mathbb{R}^n$ y $\mathbb{R}^m$ están definidas como

$$\begin{aligned} \|\overrightarrow{x}\|_2 &= \left(\overrightarrow{x}^T\ \overrightarrow{x}\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n}x_i^2\right)^{\frac{1}{2}} \quad \forall\overrightarrow{x} \in \mathbb{R}^n, \\ \|A\overrightarrow{x}\|_2^2 &= \left(A\overrightarrow{x}\right)^T\ A\overrightarrow{x} = \overrightarrow{x}^T A^T A\overrightarrow{x} \quad \forall\overrightarrow{x} \in \mathbb{R}^n. \end{aligned}$$

Definimos la función $\overrightarrow{g}$ de $\mathbb{R}^n$ en $\mathbb{R}$ como sigue $g(x) = \overrightarrow{x}^T A^T A\overrightarrow{x} \quad \overrightarrow{x} \in \mathbb{R}^n$. Esta función es diferenciable y el gradiente de g está definido como $\nabla g(\overrightarrow{x}) = 2A^T A\overrightarrow{x} \quad \forall\overrightarrow{x} \in \mathbb{R}^n$.

Sea $S = \left\{ \overrightarrow{x} \in \mathbb{R}^n \mid \overrightarrow{x}^T \overrightarrow{x} = 1 \right\}$, y consideramos el problema siguiente: $\text{máx}_{\overrightarrow{x} \in S} g(x)$. Note que $g(\overrightarrow{x}) = \|A\overrightarrow{x}\|_2^2$, de modo que

$$|||A|||_2 = \sup_{\|\overrightarrow{x}\|_2 \leq 1} \|A\overrightarrow{x}\|_2 = \sup_{\|\overrightarrow{x}\|_2 \leq 1} [g(\overrightarrow{x})]^{\frac{1}{2}}.$$

Apliquemos el método de los multiplicadores de Lagrange. Definimos

$$\Phi(\overrightarrow{x}, \lambda) = g(\overrightarrow{x}) + \lambda\left(1 - \overrightarrow{x}^T \overrightarrow{x}\right) \quad \overrightarrow{x} \in \mathbb{R}^n,$$

y λ es el multiplicador de Lagrange. Por las condiciones necesarias de extremo, tenemos el par de ecuaciones siguiente:

$$\begin{aligned} \nabla_{\overrightarrow{x}} \Phi(\overrightarrow{x}, \lambda) &= \nabla g(\overrightarrow{x}) - 2\lambda\overrightarrow{x} = 2A^T A\overrightarrow{x} - 2\lambda\overrightarrow{x} = 0, \\ \nabla_{\lambda} \Phi(\overrightarrow{x}, \lambda) &= 1 - \overrightarrow{x}^T \overrightarrow{x} = 0. \end{aligned}$$

Se obtiene el siguiente sistema de ecuaciones: $\overrightarrow{x} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$, $\begin{cases} \left(A^T A - \lambda I\right)\overrightarrow{x} = 0, \\ \overrightarrow{x}^T \overrightarrow{x} = 1. \end{cases}$ Así, la determinación de los puntos críticos de $\Phi(\overrightarrow{x}, \lambda)$ se transforma en el clásico problema de valores propios: $A^T A\overrightarrow{x} = \lambda\overrightarrow{x}$.

Puesto que $A^T A$ es simétrica, se sabe que los valores propios son reales. Sean $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ tales valores propios, y $\overrightarrow{x}_1, \ldots, \overrightarrow{x}_n \in \mathbb{R}^n$ los respectivos vectores propios tales que $\|\overrightarrow{x}_i\|_2 = 1$ $i = 1, \ldots, n$; esto es,

$$\begin{cases} A^T A\overrightarrow{x}_i = \lambda_i x_i \quad i = 1, \ldots, n \\ \|\overrightarrow{x}_i\|_2^2 = 1. \end{cases}$$

De la definición de la función g se deduce

$$0 \leq g(\overrightarrow{x}_i) = \overrightarrow{x}_i^T A^T A\overrightarrow{x}_i = \overrightarrow{x}_i^T \lambda_i \overrightarrow{x}_i = \lambda_i \quad i = 1, \ldots, n.$$

Por lo tanto $|||A|||_2 = \sup_{\|\overrightarrow{x}\|_2 \leq 1} \|A\overrightarrow{x}\|_2 = \left(\max_{i=1,\ldots,n} |\lambda_i|\right)^{\frac{1}{2}}$.

■

Observación

1. Sea $\overrightarrow{A} = (a_1, ..., a_n) \in \mathbb{R}^n$. Se tiene $A = (a_1, ..., a_n) \in M_{1\times n}[\mathbb{R}]$. Entonces $|||A|||_1 = \left\|\overrightarrow{A}\right\|_\infty$, y $|||A|||_\infty = \left\|\overrightarrow{A}\right\|_1$.

   Las normas geométricas de matrices son submultiplicativas, como se muestra en el siguiente teorema.

Teorema 7 *Sean A, $B \in M_{n\times n}[\mathbb{R}]$. Entonces*

i. $|||AB|||_1 \leq |||A|||_1 |||B|||_1$,

ii. $|||AB|||_\infty \leq |||A|||_\infty |||B|||_\infty$.

iii. $|||AB|||_2 \leq |||A|||_2 |||B|||_2$.

Demostración.

i. Sea $\overrightarrow{x} \in \mathbb{R}^n$ con $\overrightarrow{x} \neq 0$ tal que $\overrightarrow{y} = B\overrightarrow{x} \neq 0$. Entonces $\frac{\|AB\overrightarrow{x}\|_1}{\|\overrightarrow{x}\|_1} = \frac{\|A\overrightarrow{y}\|_1}{\|\overrightarrow{y}\|_1} \cdot \frac{\|\overrightarrow{y}\|_1}{\|\overrightarrow{x}\|_1}$. Luego

$$\begin{aligned} |||AB|||_1 &= \sup_{\|\overrightarrow{x}\|_1 \leq 1} \frac{\|AB\overrightarrow{x}\|_1}{\|\overrightarrow{x}\|_1} = \sup_{\|\overrightarrow{x}\|_1 \leq 1} \frac{\|A\overrightarrow{y}\|_1}{\|\overrightarrow{y}\|_1} \cdot \frac{\|\overrightarrow{y}\|_1}{\|\overrightarrow{x}\|_1} \\ &\leq \sup_{\|\overrightarrow{y}\|_1 \leq 1} \frac{\|A\overrightarrow{y}\|_1}{\|\overrightarrow{y}\|_1} \cdot \sup_{\|\overrightarrow{x}\|_1 \leq 1} \frac{\|B\overrightarrow{y}\|_1}{\|\overrightarrow{x}\|_1} = |||A|||_1 |||B|||_1. \end{aligned}$$

Otra forma de obtener este resultado se muestra a continuación:

$$\|AB\overrightarrow{x}\|_1 = \|AB\overrightarrow{x}\|_1 \leq |||A|||_1 \, |||B\overrightarrow{x}|||_1 \leq |||A|||_1 \, |||B|||_1 \, \|\overrightarrow{x}\|_1 ,$$

de donde

$$|||A|||_1 = \max_{\|\overrightarrow{x}\|_1 \leq 1} \|AB\overrightarrow{x}\|_1 \leq |||A|||_1 \, |||B|||_1 .$$

ii. Sea $\overrightarrow{x} \in \mathbb{R}^n$. Entonces,

$$\|AB\overrightarrow{x}\|_\infty = \|A\,(B\overrightarrow{x})\|_\infty \leq |||A|||_\infty \, |||B\overrightarrow{x}|||_\infty \leq |||A|||_\infty \, |||B|||_\infty \, \|\overrightarrow{x}\|_\infty ,$$

y para $\|\overrightarrow{x}\|_\infty \leq 1$, se obtiene $|||AB|||_\infty \leq |||A|||_\infty \, |||B|||_\infty$.

iii. Es inmediata.

■

Otra clase de normas en $M_{m\times n}[\mathbb{R}]$ se definen a continuación, donde $A = (a_{ij})_{m\times n} \in M_{m\times n}[\mathbb{R}]$.

a) $N(A) = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|$.

b) $N(A) = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}$.

c) $N(A) = \operatorname{Max}_{i=1,\ldots,m} \operatorname{Max}_{j=1,\ldots,n} |a_{ij}|$.

d) $N(A) = \operatorname{Max}_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$.

e) $N(A) = \operatorname{Max}_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$.

f) Sean $p \in [1,\infty[$ y $A = (a_{ij})_{m\times n} \in M_{m\times n} \in [\mathbb{R}]$. Se define $\|A\|_p = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^p\right)^{\frac{1}{p}}$, entonces $\|\cdot\|_p$ es una norma de Hölder sobre $M_{m\times n}[\mathbb{R}]$.

12.3.3. Normas en el espacio de funciones continuas $C([a,b])$.

Sean $a, b \in \mathbb{R}$ tales que $a < b$. Se denota con $C([a,b])$ al espacio de todas las funciones continuas en $[a,b]$. En este espacio se van a definir dos normas: la norma de Chebyshev notada $\|\cdot\|_\infty$ y la norma de Hölder que se denota con $\|\cdot\|_p$, donde $p \in [1,\infty[$.

Norma de Chebyshev.

Se define la función $\|\cdot\|_\infty$ de $C([a,b])$ en $\mathbb{R}$ como se indica a continuación:

$$\|f\|_\infty = \operatorname{Max}_{x\in[a,b]} |f(x)| \quad \forall f \in C([a,b]) .$$

Esta se conoce como norma de Chebyshev. Probemos que $\|\cdot\|_\infty$ es una norma sobre $C([a,b])$. En efecto,

i) De la definición de $\|\cdot\|_\infty$, se tiene $\|f\|_\infty \geq 0$.

ii) Si $f = 0$, esto es, $f(x) = 0 \quad \forall x \in [a,b]$, $\|f\|_\infty = 0$. Recíprocamente, si $\|f\|_\infty = 0 = \operatorname{Max}_{x\in[a,b]} |f(x)|$, entonces $f(x) = 0 \ \forall x \in [a,b]$, es decir que $f = 0$.

iii) Sean $\lambda \in \mathbb{R}$, $f \in C([a,b])$. Entonces $\|\lambda f\|_\infty = \operatorname{Max}_{x\in[a,b]} |\lambda f(x)| = |\lambda| \, \|f\|_\infty$.

iv) Sean $f,\, g \in C([a,b])$. Entonces $|f(x)+g(x)| \leq |f(x)| + |g(x)| \quad \forall x \in [a,b]$. Resulta,

$$\|f+g\|_\infty = \underset{x\in[a,b]}{\text{Max}} |f(x)+g(x)| \leq \underset{x\in[a,b]}{\text{Max}} |f(x)| + \underset{x\in[a,b]}{\text{Max}} |g(x)| \leq \|f\|_\infty + \|g\|_\infty .$$

Conclusión**:** $\|\cdot\|_\infty$ es una norma en $C([a,b])$.

Notación**:** El espacio $C([a,b])$ provisto de la norma $\|\cdot\|_\infty$ se le nota $\mathcal{L}^\infty([a,b])$.

Normas hölderianas en el espacio de funciones continuas $C([a,b])$.

Sea $p \in [1,\infty[$. Se define la función $\|\cdot\|_p$ de $C([a,b])$ en $\mathbb{R}$ como sigue:

$$\|f\|_p = \left(\int_a^b |f(x)|^p\, dx\right)^{\frac{1}{p}} \qquad \forall f \in C([a,b]).$$

Probemos que $\|\cdot\|_p$ es una norma sobre $C([a,b])$ denominada norma hölderiana. Para el efecto mostremos que $\|\cdot\|_p$ satisface las cuatro propiedades de la definición de norma.

i) Es claro que $\|f\|_p \geq 0 \quad \forall f \in C([a,b])$.

ii) Si $f = 0$ se tiene $\|f\|_p = 0$. Supongamos que $\|f\|_p = 0$ y probemos que $f = 0$, o lo que es equivalente a probar que $f \neq 0 \Rightarrow \|f\|_p \neq 0$. Recuerde que si u, v son proposiciones, se tiene la siguiente tautología: $(u \Longrightarrow v) \Longleftrightarrow [(\sim v) \Longrightarrow (\sim u)]$. Efectivamente, si $f \neq 0$, existe $x_0 \in [a,b]$ tal que $f(x_0) \neq 0$ y como f es continua, existe $[\alpha,\beta] \subset [a,b]$ tal que $f(x) \neq 0 \quad \forall x \in [\alpha,\beta]$. Luego

$$0 < \int_\alpha^\beta |f(x)|^p\, dx \leq \int_a^b |f(x)|^p\, dx = \|f\|_p^p ,$$

es decir que $\|f\|_p > 0$. Así, $f \neq 0 \Rightarrow \|f\|_p > 0$, o lo que es lo mismo $\|f\|_p = 0 \Rightarrow f = 0$.

iii) Sean $\lambda \in \mathbb{R}$, $f \in C([a,b])$. Se verifica inmediatamente que $\|\lambda f\|_p = |\lambda|\, \|f\|_p$.

iv) Mediante un procedimiento análogo al de la demostración de la desigualdad triangular para la norma hölderiana en $\mathbb{R}^n$, se obtiene de la desigualdad de Young, la desigualdad de Hölder que para funciones continuas se establece del modo siguiente: para todo $p,\, q \in [1,\infty[$ tales que $\dfrac{1}{p} + \dfrac{1}{q} = 1$, $f, g \in \mathcal{C}([a,b])$, entonces $fg \in \mathcal{C}([a,b])$ y

$$\|fg\|_1 = \int_a^b |f(x)\, g(x)|\, dx \leq \|f\|_p\, \|g\|_q ,$$

o lo que es lo mismo

$$\int_a^b |f(x)\, g(x)|\, dx \leq \left(\int_a^b |f(x)|^p\, dx\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q\, dx\right)^{\frac{1}{q}} .$$

Para probar esta desigualdad se pone $\alpha = \frac{|f(x)|^p}{\|f\|_p^p}$, $\beta = \frac{|g(x)|^q}{\|g\|_q^q}$ con $f \neq 0$, $g \neq 0$ en la desigualdad de Young que ha sido establecida anteriormente. Se obtiene

$$\frac{|f(x)|}{\|f\|_p}\frac{|g(x)|}{\|g\|_q} \leq \frac{1}{p}\frac{|f(x)|^p}{\|f\|_p^p} + \frac{1}{q}\frac{|g(x)|^q}{\|g\|_q^q} \qquad x \in [a,b],$$

y luego se integra sobre el intervalo $[a,b]$, esto es

$$\frac{1}{\|f\|_p\, \|g\|_q}\int_a^b |f(x)\, g(x)|\, dx \leq \frac{1}{p}\frac{1}{\|f\|_p^p}\int_a^b |f(x)|^p\, dx + \frac{1}{q}\frac{1}{\|g\|_q^q}\int_a^b |g(x)|^q\, dx = 1,$$

y de esta desigualad se obtiene la desigualdad de Hölder.

Sean $f,\, g \in C([a,b])$. Utilizando la desigualdad de Hölder se obtiene la desigualdad triangular siguiente:

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p .$$

Conclusión: $\|\cdot\|_p$ es una norma sobre $C([a,b])$, $p \in [1,\infty[$.

Notación: El espacio $C([a,b])$ provisto de la norma $\|\cdot\|_p$ se le nota $\mathcal{L}^p([a,b])$.

Sea $f \in C([a,b])$. Para $p = 1$, la norma $\|\cdot\|_1$ está definida como $\|f\|_1 = \int_a^b |f(x)|\,dx$. Note que $f \in \mathcal{L}^1([a,b])$. Para $p = 2$, $\|f\|_2 = \left(\int_a^b |f(x)|^2\,dx\right)^{\frac{1}{2}}$ es la norma euclídea. Se tiene $f \in \mathcal{L}^2([a,b])$. Si $p = 4$, la función $\|\cdot\|_4$ está definida como $\|f\|_4 = \left(\int_a^b |f(x)|^4\,dx\right)^{\frac{1}{4}}$. Note que $f \in \mathcal{L}^4([a,b])$.

Para $p = q = 2$, la desigualdad de Hölder se expresa como sigue:

$$\|fg\|_1 = \int_a^b |f(x)\,g(x)|\,dx \leq \|f\|_2 \|g\|_2 \quad \forall f, g \in C([a,b]),$$

que coincide con la conocida desigualdad de Cauchy-Schwarz.

Es importante observar el significado de los espacios $\mathcal{L}^p([a,b])$ que aquí hemos dado: simplemente es el espacio $C([a,b])$ provisto de la norma $\|\cdot\|_p$. Estos espacios $\mathcal{L}^p([a,b])$ son diferentes de los espacios $L^p(a,b)$ que designan a los espacios de Lebesgue que se tratan en los cursos de Análisis Funcional, Teoría de Integración, etc. Para información $\mathcal{L}^p([a,b]) \subset L^p(a,b)$. Similarmente $\mathcal{L}^\infty([a,b]) \subset L^\infty(a,b)$, donde $L^\infty(a,b)$ pertenece a la clase de los espacios de Lebesgue.

Sean $\omega \in C([a,b])$ tal que $\omega(x) > 0 \quad \forall x \in [a,b]$, las siguientes son normas en $C([a,b])$:

a) $\|f\| = \int_a^b \omega(x)\,|f(x)|\,dx \quad f \in C([a,b])$.

b) $\|f\| = \left(\int_a^b \omega(x)\,f^2(x)\,dx\right)^{\frac{1}{2}} \quad f \in C([a,b])$.

Otras normas en $C^1([0,10])$ se definen a continuación:

a) $N(u) = \underset{x\in[0,10]}{\text{Max}} \{|u(x)|, |u'(x)|\}$.

b) $M(u) = \int_0^{10} (|u(x)| + |u'(x)|)\,dx$.

c) $R(u) = \left(\int_0^{10} (|u(x)|^p + |u'(x)|^p)\,dx\right)^{\frac{1}{p}}$ para $p \in]1,\infty[$.

12.4. Espacios con producto interno.

En esta sección revisamos brevemente una clase de espacios vectoriales reales V en los que se define un producto escalar (dicho también producto interno o producto punto) que los denominaremos espacios con producto escalar o espacios con producto punto, o espacios euclídeos. En esta clase de espacios se introducirán las nociones geométricas de ángulo, de perpendicularidad u ortogonalidad, la conocida ley del paralelogramo y el teorema de Pitágoras.

Enfatizaremos en dos clases de espacios $\mathbb{R}^n$ y $C([a,b])$.

Definición 20 *Sea V un espacio vectorial real. Un producto interno o producto escalar en V es una función denotada $\langle\cdot,\cdot\rangle$ de $V \times V$ en $\mathbb{R}$ que satisface las siguientes propiedades:*

i. $\langle x, y\rangle = \langle y, x\rangle \quad \forall x, y \in V$,

ii. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle \quad \forall x, y, z \in V$,

iii. $\langle \lambda x, y\rangle = \lambda \langle x, y\rangle \quad \forall \lambda \in \mathbb{R}, \quad \forall x, y \in V$,

iv. $\langle x, x\rangle = 0 \Leftrightarrow x = 0$,
$\langle x, x\rangle > 0 \Leftrightarrow x \neq 0 \quad x \in V$.
Para $x, y \in V$, el número real $\langle x, y\rangle$ se llama producto escalar o producto interno de x con y.

Definición 21 *Un espacio vectorial V en el que se ha definido un producto escalar $\langle \cdot, \cdot \rangle$ se denomina espacio con producto interior, espacio con producto escalar o espacio prehilbertiano.*

Observación

Si V es un espacio vectorial complejo y $\langle \cdot, \cdot \rangle$ denota un producto escalar en V, la propiedad i) se escribe como $\langle x, y \rangle = \overline{\langle y, x \rangle} \quad \forall x, y \in V,$ donde el lado derecho de la igualdad designa el número complejo conjugado de $\langle y, x \rangle$. Las propiedades ii), iii) y iv) de la definición de producto escalar permanecen invariables.

Ejemplos

1. Sea $V = \mathbb{R}^n$. Un producto escalar $\langle \cdot, \cdot \rangle$ en $\mathbb{R}^n$ se define como $\langle \overrightarrow{x}, \overrightarrow{y} \rangle = \sum_{i=1}^{n} x_i y_i$, donde $\overrightarrow{x} = (x_1, ..., x_n), \quad \overrightarrow{y} = (y_1, ..., y_n) \in \mathbb{R}^n$. En notación matricial, esto es, los elementos de $\mathbb{R}^n$ se escriben como vectores columna, el producto escalar $\langle \cdot, \cdot \rangle$ en $\mathbb{R}^n$ se escribe como

$$\langle \overrightarrow{x}, \overrightarrow{y} \rangle = \overrightarrow{x}^T \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i,$$

donde $\overrightarrow{x}^T = (x_1, ..., x_n)$, $\overrightarrow{y}^T = (y_1, ..., y_n) \in \mathbb{R}^n$ denotan los vectores transpuestos de los vectores columna $\overrightarrow{x}$ e $\overrightarrow{y}$. Las propiedades i) a iv) de la definición de producto escalar se verifican fácilmente y se dejan como ejercicio.

Para $n = 2$, el producto $\langle \cdot, \cdot \rangle$ en $\mathbb{R}^2$ está dado como sigue: si $\overrightarrow{x} = (a_1, b_1)$, $\overrightarrow{y} = (a_2, b_2) \in \mathbb{R}^2$, entonces $\langle \overrightarrow{x}, \overrightarrow{y} \rangle = a_1 a_2 + b_1 b_2$.

Para $n = 3$, el producto $\langle \cdot, \cdot \rangle$ en $\mathbb{R}^3$ está dado como $\langle \overrightarrow{x}, \overrightarrow{y} \rangle = a_1 a_2 + b_1 b_2 + c_1 c_2$ con $\overrightarrow{x} = (a_1, b_1, c_1)$, $\overrightarrow{y} = (a_2, b_2, c_2) \in \mathbb{R}^3$.

2. Sea $V = C([a, b])$. Un producto escalar $\langle \cdot, \cdot \rangle$ en $C([a, b])$ se define como

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx \quad \forall f, g \in C([a, b]).$$

Por ejemplo, si $f, g \in C([-1, 1])$ están dadas como $f(x) = x^3$, $g(x) = x^2 + 1 \quad x \in [-1, 1]$. Entonces

$$\langle f, g \rangle = \int_{-1}^{1} f(x) g(x) \, dx = \int_{-1}^{1} x^3 \left(x^2 + 1\right) dx = 0.$$

Probemos que la función $\langle \cdot, \cdot \rangle$ de $C([a, b]) \times C([a, b])$ en $\mathbb{R}$ satisface las cuatro propiedades de la definición de producto escalar. Para ello utilicemos algunas propiedades de las funciones integrables y de las funciones continuas. Sean f, g, $h \in C([a, b])$ y $\lambda \in \mathbb{R}$. Entonces

i. $\langle f, g \rangle = \int_a^b f(x) g(x) \, dx = \int_a^b g(x) f(x) \, dx = \langle g, f \rangle$.

ii. $\langle f + g, h \rangle = \int_a^b [f(x) + g(x)] h(x) \, dx = \int_a^b f(x) h(x) \, dx + \int_a^b g(x) h(x) \, dx = \langle f, h \rangle + \langle g, h \rangle$.

iii. $\langle \lambda f, g \rangle = \int_a^b \lambda f(x) g(x) \, dx = \lambda \int_a^b f(x) g(x) \, dx = \lambda \langle f, g \rangle$.

iv. Si $f = 0$ es claro que $\langle f, f \rangle = 0$. Mostremos que si $\langle f, f \rangle = 0$ entonces $f = 0$. Para ello, haciendo uso de la tautología $(p \Rightarrow q) \Leftrightarrow [(\sim q) \Rightarrow (\sim p)]$ en la que p, q son las proposiciones siguientes $p: \ f = 0, \ q: \langle f, f \rangle = 0$; tenemos la proposición siguiente: $f \neq 0 \Rightarrow \langle f, f \rangle > 0$.
Si $f \neq 0$, existe $x_0 \in [a, b]$ tal que $f(x_0) \neq 0$. Por hipótesis f es continua, por lo tanto es continua en x_0 y siendo $f(x_0) \neq 0$, existe un intervalo $[\alpha, \beta] \subset [a, b]$ tal que $f(x) \neq 0$ $\forall x \in [\alpha, \beta]$. Luego

$$0 < \int_\alpha^\beta f^2(x) \, dx \leq \int_a^b f^2(x) \, dx = \langle f, f \rangle.$$

Así, $f \neq 0 \Rightarrow \langle f, f \rangle > 0$ y por la tautología antes citada se deduce $\langle f, f \rangle = 0 \Rightarrow f = 0$. Consecuentemente, $\langle f, f \rangle = 0 \Leftrightarrow f = 0$. Además, del resultado precedente, es claro que $\langle f, f \rangle > 0 \quad \forall f \in C([a, b])$ con $f \neq 0$.

3. Sean $V = M_{n\times n}[\mathbb{R}]$, $A = (a_{ij}) \in M_{n\times n}[\mathbb{R}]$. La traza de la matriz A se nota $tr(A)$ y se define como $tr(A) = \sum_{i=1}^{n} a_{ii}$. Una función $\langle\cdot,\cdot\rangle$ de $M_{n\times n}[\mathbb{R}] \times M_{n\times n}[\mathbb{R}]$ en $\mathbb{R}$ definida como

$$\langle A, B\rangle = tr\left(B^T A\right) \quad \forall A,\, B \in M_{n\times n}[\mathbb{R}]$$

es un producto escalar en $M_{n\times n}[\mathbb{R}]$.

Propiedades adicionales del producto escalar.

En un espacio prehilbertiano real V se verifican las propiedades siguientes:

i. $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle \quad \forall x,\, y,\, z \in V,$

ii. $\langle x, \lambda y\rangle = \lambda \langle x, y\rangle \quad \forall \lambda \in \mathbb{R},\ \forall x,\, y \in V,$

iii. $\langle x - y, z\rangle = \langle x, z\rangle - \langle y, z\rangle \quad \forall x,\, y,\, z \in V,$
$\langle x, y - z\rangle = \langle x, y\rangle - \langle x, z\rangle \quad \forall x,\, y,\, z \in V,$

iv. $\langle 0, x\rangle = \langle x, 0\rangle = 0 \quad \forall x \in V.$

v. Sean $x,\, y \in V$, si para todo $z \in V$, $\langle x, z\rangle = \langle y, z\rangle$, entonces $x = y$.

vi. Sean $x_1, ..., x_n,\ y \in V$, $\alpha_1, ..., \alpha_n \in \mathbb{R}$. Entonces

$$\left\langle \sum_{i=1}^{n} \alpha_i x_i,\ y \right\rangle = \sum_{i=1}^{n} \alpha_i \langle x_i,\ y\rangle, \quad \text{y,} \quad \left\langle y, \sum_{i=1}^{n} \alpha_i x_i \right\rangle = \sum_{i=1}^{n} \alpha_i \langle x_i,\ y\rangle.$$

En un espacio vectorial real V se pueden definir una infinidad de productos escalares. En los ejercicios se exhiben algunos productos escalares definidos en $\mathbb{R}^2$ y en $C([0,1])$.

Longitud o norma de un vector

Definición 22 *Sea V un espacio vectorial real provisto de un producto escalar $\langle\cdot,\cdot\rangle$. La longitud o norma de $x \in V$ se nota $\|x\|$ y se define como* $\|x\| = (\langle x, x\rangle)^{\frac{1}{2}}$.

Esta norma $\|\cdot\|$ se dice asociada al producto escalar $\langle\cdot,\cdot\rangle$ y se le denomina norma euclídea.

Ejemplos

1. En el caso en que $V = \mathbb{R}^n$, la norma del vector $\overrightarrow{x}^T = (x_1, ..., x_n) \in \mathbb{R}^n$ asociada al producto escalar definido como $\overrightarrow{x}^T \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$, se escribe $\|\overrightarrow{x}\|_2 = \left(\overrightarrow{x}^T \overrightarrow{x}\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$. Esta norma coincide con la norma hölderiana en $\mathbb{R}^n$ con $p = 2$.

2. Sea $V = C([-1,1])$. La norma de $f \in C([-1,1])$ asociada al producto escalar $\langle\cdot,\cdot\rangle$ antes definido, está definida como $\|f\|_2 = (\langle f, f\rangle)^{\frac{1}{2}} = \left(\int_{-1}^{1} f^2(x)\, dx\right)^{\frac{1}{2}}$. Esta norma coincide con la norma hölderiana en $C([a,b])$ con $p = 2$.

3. Se denota $\mathbb{K}_n([a,b])$ al espacio vectorial de los polinomios reales de grado $\leq n$ restringidos al intervalo $[a,b] \subset \mathbb{R}$ El espacio $\mathbb{K}_n([a,b])$ es un subespacio de $C([a,b])$ de dimensión $n+1$. Definido un producto escalar $\langle\cdot,\cdot\rangle$ en $C([a,b])$, este es un producto escalar en $\mathbb{K}_n([a,b])$ y la norma asociada se escribe

$$\|p\|_2 = (\langle p, p\rangle)^{\frac{1}{2}} = \left(\int_a^b p^2(t)\, dt\right)^{\frac{1}{2}} \qquad \forall p \in \mathbb{K}_n([a,b]).$$

En tal caso diremos que $\mathbb{K}_n([a,b])$ es un espacio con producto intermo inducido por el de $C([a,b])$ y que la norma $\|\cdot\|_2$ en $\mathbb{K}_n([a,b])$ es la inducida por la norma $\|\cdot\|_2$ en $C([a,b])$.

4. Sea $V = C^1([-1,1])$. Un producto escalar $\langle \cdot, \cdot \rangle$ en $C^1([-1,1])$ se define como

$$\langle f, g \rangle = \int_{-1}^{1} \left[f(x)\, g(x) + \frac{df}{dx}(x)\, \frac{dg}{dx}(x) \right] dx$$

y la norma de $f \in C^1([-1,1])$ asociada a este producto escalar se define como

$$\|f\|_{1,2} = \left(\int_{-1}^{1} \left(|f(x)|^2 + \left| \frac{df}{dx}(x) \right|^2 \right) dx \right)^{\frac{1}{2}}.$$

5. Sean $\Omega = [-1,1] \times [-1,1] \subset \mathbb{R}^2$. Se denota con $C(\Omega)$ al espacio vectorial de funciones continuas en Ω. Un producto escalar $\langle \cdot, \cdot \rangle$ en $C(\Omega)$ se define como

$$\langle f, g \rangle = \int_{-1}^{1} \int_{-1}^{1} f(x,y)\, g(x,y)\, dxdy \quad \forall f, g \in C(\Omega).$$

Se propone como ejercicio probar que efectivamente la función $\langle \cdot, \cdot \rangle$ definida en $C(\Omega) \times C(\Omega)$ es un producto escalar. La norma de $f \in C(\Omega)$ asociada a este producto escalar se define como

$$\|f\| = \left(\int_{\Omega} |f(x,y)|^2\, dxdy \right)^{\frac{1}{2}}.$$

6. Considerar el espacio de funciones $C^1([-1,1])$ que poseen derivada continua en $[-1,1]$. Se define la función $\langle \cdot, \cdot \rangle_1$ de $C^1([-1,1]) \times C^1([-1,1])$ en $\mathbb{R}$ como sigue:

$$\langle u, v \rangle_1 = \int_{-1}^{1} \left[u(x)\, v(x) + u'(x)\, v'(x) \right] dx \quad \forall u,\, v \in C^1([-1,1]).$$

entonces $\langle \cdot, \cdot \rangle$ es un producto escalar en $C^1([-1,1])$.

7. Sean $\Omega = [-1,1] \times [-1,1] \subset \mathbb{R}^2$. Se denota con $C(\Omega)$ al espacio vectorial de funciones continuas en Ω. Un producto escalar $\langle \cdot, \cdot \rangle$ en $C(\Omega)$ se define como

$$\langle f, g \rangle = \int_{-1}^{1} \int_{-1}^{1} f(x,y)\, g(x,y)\, dxdy \quad \forall f, g \in C(\Omega).$$

La norma de $f \in C(\Omega)$ asociada a este producto escalar se define como $\|f\| = \left(\int_{\Omega} |f(x,y)|^2\, dxdy \right)^{\frac{1}{2}}$.

8. Sea $\Omega = [-1,1] \times [-1,1] \subset \mathbb{R}^2$. Se designa con $C^1(\Omega)$ al espacio de funciones reales que poseen derivadas parciales primeras continuas en Ω. En $C^1(\Omega)$ se define la función real $\langle \cdot, \cdot \rangle_{1,2}$ como a continuación se indica:

$$\langle f, g \rangle_{1,2} = \int_{\Omega} \left(f(x,y) g(x,y) + \frac{\partial f}{\partial x}(x,y) \frac{\partial g}{\partial x}(x,y) + \frac{\partial f}{\partial y}(x,y) \frac{\partial g}{\partial y}(x,y) \right) dxdy \quad \forall f, g \in C^1(\Omega).$$

entonces $\langle \cdot, \cdot \rangle_{1,2}$ es un producto escalar en $C^1(\Omega)$. Este producto escalar se escribe en forma abreviada como $\langle f, g \rangle_{1,2} = \int_{\Omega} (fg + \nabla f \cdot \nabla g) \quad \forall f, g \in C^1(\Omega)$ y la norma asociada a este producto escalar se escribe como $\|f\|_{1,2} = \left(\int_{\Omega} \left(f^2 + |\nabla f|^2 \right) \right)^{\frac{1}{2}} \quad \forall f \in C^1(\Omega)$.

Teorema 8 *Sea V un espacio vectorial real provisto de un producto escalar $\langle\cdot,\cdot\rangle$. La longitud o norma $\|\cdot\|$ en V satisface las siguientes propiedades:*

i. $\|x\| \geq 0 \quad \forall x \in V.$

ii. $\|x\| = 0 \Leftrightarrow x = 0.$

iii. $\|\lambda x\| = |\lambda| \, \|x\| \quad \forall \lambda \in \mathbb{R}, \ x \in V.$

iv. $|\langle x, y\rangle| \leq \|x\| \, \|y\| \quad \forall x, y \in V$ *(desigualdad de Cauchy-Schwarz).*

v. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in V$ *(desigualdad triangular).*

vi. $|\|x\| - \|y\|| \leq \|x - y\| \quad \forall x, y \in V.$

vii. $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \forall x, y \in V$ *(ley del paralelogramo).*

viii. $\langle x, y\rangle = \frac{1}{4}\left(\|x + y\|^2 - \|x - y\|^2\right) \quad \forall x, y \in V$ *(identidad de polarización).*

Demostración.

i. Puesto que la función $\langle\cdot,\cdot\rangle$ de $V \times V$ en $\mathbb{R}$ es un producto escalar, esta tiene la propiedad siguiente: para $x \in V$, $\langle x, x\rangle = 0 \Leftrightarrow x = 0, \quad \langle x, x\rangle > 0 \Leftrightarrow x \neq 0$, y de la definición de norma, se tiene $\|x\| \geq 0 \quad \forall x \in V.$

ii. Como $\langle x, x\rangle = 0 \Leftrightarrow x = 0$, resulta $\|x\| = (\langle x, x\rangle)^{\frac{1}{2}} = 0 \Leftrightarrow x = 0.$

iii. Sean $\lambda \in \mathbb{R}$, $x \in V$. Entonces

$$\|\lambda x\| = (\langle \lambda x, \lambda x\rangle)^{\frac{1}{2}} = \left(\lambda^2 \langle x, x\rangle\right)^{\frac{1}{2}} = |\lambda| \, (\langle x, x\rangle)^{\frac{1}{2}} = |\lambda| \, \|x\|.$$

Así, $\|\lambda x\| = |\lambda| \, \|x\| \quad \forall \lambda \in \mathbb{R}, \quad \forall x \in V.$

iv. Sean $x, y \in V$. Por la parte i) de este teorema se tiene $\|x + \lambda y\|^2 \geq 0 \quad \forall \lambda \in \mathbb{R}$. Luego, por la definición de norma y las propiedades del producto escalar, se tiene

$$\begin{aligned} 0 &\leq \|x + \lambda y\|^2 = \langle x + \lambda y, x + \lambda y\rangle = \langle x, x\rangle + \langle x, \lambda y\rangle + \langle \lambda y, x\rangle + \langle \lambda y, \lambda y\rangle \\ &= \langle x, x\rangle + \lambda \langle x, y\rangle + \lambda \langle y, x\rangle + \lambda^2 \langle y, y\rangle = \|x\|^2 + 2\lambda \langle x, y\rangle + \lambda^2 \|y\|^2. \end{aligned}$$

Sea P el polinomio de grado ≤ 2 definido por $P(\lambda) = \|x\|^2 + 2\langle x, y\rangle \lambda + \|y\|^2 \lambda^2, \quad \lambda \in \mathbb{R}.$ Por la parte precedente se verifica $P(\lambda) \geq 0 \quad \forall \lambda \in \mathbb{R}$, con lo que el discriminante $d = (2\langle x, y\rangle)^2 - 4\|x\|^2 \|y\|^2 \leq 0$ de donde $4(\langle x, y\rangle)^2 \leq 4\|x\|^2 \|y\|^2$. Tomando la raíz cuadrada y considerando que la norma $\|\cdot\|$ es no negativa, se deduce la desigualdad de Cauchy-Schwarz: $|\langle x, y\rangle| \leq \|x\| \, \|y\|.$

v. Sean $x, y \in V$. Entonces

$$\|x + y\|^2 = \langle x + y, x + y\rangle = \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2.$$

Se tiene $\langle x, y\rangle \leq |\langle x, y\rangle|$ y por la desigualdad de Cauchy-Schwarz resulta

$$\langle x, y\rangle \leq |\langle x, y\rangle| \leq \|x\| \, \|y\| \quad \forall x, y \in V.$$

Luego,

$$\|x + y\|^2 = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2 \leq \|x\|^2 + 2\|x\| \, \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

Tomando la raíz cuadrada y considerando que la norma es no negativa, se obtiene la desigualdad triangular

$$\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in V.$$

vi. La desigualdad $|\|x\| - \|y\|| \leq \|x - y\| \quad \forall x, y \in V$ ya fue probada en la sección de los espacios normados.

vii. De la definición de norma y de las propiedades del producto escalar, se tiene

$$\begin{aligned} \|x+y\|^2 + \|x-y\|^2 &= \langle x+y, x+y\rangle + \langle x-y, x-y\rangle \\ &= \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle + \langle x,x\rangle - 2\langle x,y\rangle + \langle y,y\rangle \\ &= 2\left(\|x\|^2 + \|y\|^2\right). \end{aligned}$$

viii. Se propone como ejercicio.

■

Definición 23 *Sean V un espacio prehilbertiano, $x, y \in V$. La distancia de x a y se denota y se define como $d(x,y) = \|x - y\|$.*

Si $\langle \cdot, \cdot \rangle$ es un producto escalar definido en V y $\|\cdot\|$ la norma asociada, se tiene

$$d(x,y) = \|x-y\| = (\langle x-y, x-y\rangle)^{\frac{1}{2}} \quad \forall x, y \in V.$$

Teorema 9 *Sea V un espacio prehilbertiano.*
La función d de $V \times V$ en $\mathbb{R}$ definida como $d(x,y) = \|x-y\| \quad \forall x, y \in V$ es una métrica en V, es decir que satisface las propiedades siguientes:

i. $d(x,y) \geq 0 \quad \forall x, y \in V.$

ii. $d(x,y) = 0 \Leftrightarrow x = y, \;\; x, y \in V.$

iii. $d(x,y) = d(y,x) \quad \forall x, y \in V.$

iv. $d(x,z) \leq d(x,y) + d(y,z) \quad \forall x, y, z \in V$ *(desigualdad triangular).*

Demostración. La prueba es inmediata y se propone como ejercicio. ■

Si V es un espacio vectorial real provisto de un producto interior o producto escalar $\langle \cdot, \cdot \rangle$, la norma asociada a este producto escalar está dado por $\|x\| = (\langle x, x\rangle)^{\frac{1}{2}} \quad x \in V$, y la métrica asociada a la norma $\|\cdot\|$ está definida como $d(x,y) = \|x-y\| \quad x, y \in V$, con lo cual V es un espacio métrico que escribimos (V, d). En la siguiente sección trataremos más en detalle las métricas sobre un conjunto no vacío E.

12.4.1. Ortogonalidad o perpendicularidad.

Definición 24 *Sea V un espacio vectorial provisto del porducto escalar $\langle \cdot, \cdot \rangle$.*

i) *Sean $x, y \in V$. Se dice que x es ortogonal o perpendicular a y, que se nota $x \perp y$, si y solo si $\langle x, y\rangle = 0$.*

ii) *Sean $x \in V$, $M \subset V$ con $M \neq \phi$. Se dice que x es ortogonal a M, que se escribe $x \perp M$, si y solo si $\langle x, y\rangle = 0 \;\; \forall y \in M$.*

iii) *Sean M, N dos subconjuntos no vacíos de V. Se dice que M es ortogonal a N, que se nota $M \perp N$, si y solo si $\langle x, y\rangle = 0 \;\; \forall x \in M, \forall y \in N$.*

iv) *Sea $M \subset V$ con $M \neq 0$. Se dice que M es ortogonal si y solo si $\langle x, y\rangle = 0 \;\; \forall x, y \in M, \; x \neq y$.*

v) *Se dice que M es ortonomal si y solo si M es ortogonal, y $\forall x \in M, \; \|x\| = 1$.*

Ejemplos

1. Sea $V = \mathbb{R}^n$. El conjunto $M = \{\overrightarrow{e}_1, \ldots, \overrightarrow{e}_n\}$, donde $\overrightarrow{e}_1^T = (1, 0, \ldots, 0), \ldots, \overrightarrow{e}_n^T = (0, \ldots, 0, 1)$ son los vectores de la base canónica de $\mathbb{R}^n$, es un conjuto ortogonal, pués

$$\overrightarrow{e}_j^T \overrightarrow{e}_k = 0 \text{ si } j \neq k, \quad \text{y,} \quad \|\overrightarrow{e}_j\| = 1 \quad j = 1, \ldots, n.$$

2. Sean $L > 0$. Se denota con $C([-L, L])$ al espacio vectorial de las funciones continuas en $[-L, L]$. Proveemos a $C([-L, L])$ del producto escalar $\langle \cdot, \cdot \rangle$ definido por

$$\langle u, v \rangle = \int_{-L}^{L} u(x) v(x) \, dx \quad \forall u, v \in C([-L, L]).$$

Los siguientes conjuntos de funciones son muy importantes en el desarrollo en series de Fourier de funciones reales periódicas de período $2L$ y continuas a trozos en el intervalo $[-L, L]$. Sean M, N los subconjuntos de $C([-L, L])$ definidos como $M = \{\varphi_k \mid k \in \mathbb{N}\}$, $N = \{\psi_k \mid k \in \mathbb{Z}^+\}$, donde

$$\begin{aligned} \varphi_0(x) &= 1 \quad x \in [-L, L], \quad \varphi_k(x) = \cos\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \quad k = 1, 2, \ldots, \\ \psi_k(x) &= \operatorname{sen}\left(\frac{k\pi x}{L}\right) \quad x \in [-L, L], \quad k = 1, 2, \ldots \end{aligned}$$

Se tiene

i) M es un conjunto ortogonal.

ii) N es un conjunto ortogonal.

iii) $M \perp N$.

i) Probemos que M es ortogonal, esto es,

$$\begin{aligned} \langle \varphi_0, \varphi_k \rangle &= 0 \quad \forall k = 1, 2, \ldots \\ \langle \varphi_j, \varphi_k \rangle &= 0 \quad \forall j, k \in \mathbb{Z} \text{ con } j \neq k. \end{aligned}$$

En efecto, de la definición de φ_0 y φ_k, se tiene

$$\langle \varphi_0, \varphi_k \rangle = \int_{-L}^{L} \varphi_0(x) \varphi_k(x) \, dx = \int_{-L}^{L} \cos\left(\frac{k\pi x}{L}\right) dx = \frac{L}{k\pi} \operatorname{sen}\left(\frac{k\pi x}{L}\right)\Big|_{-L}^{L} = 0 \quad k = 1, 2, \ldots$$

Por otro lado,

$$\langle \varphi_j, \varphi_k \rangle = \int_{-L}^{L} \varphi_j(x) \varphi_k(x) \, dx = \int_{-l}^{L} \cos\left(\frac{j\pi x}{L}\right) \operatorname{sen}\left(\frac{k\pi x}{L}\right) dx \quad \forall x \in [-L, L].$$

Como $\cos\left(-\frac{j\pi x}{L}\right) = \cos\left(\frac{j\pi x}{L}\right)$, $\operatorname{sen}\left(-\frac{k\pi x}{L}\right) = -\operatorname{sen}\left(\frac{k\pi x}{L}\right)$ entonces la función $\varphi_j \varphi_k$ es impar para $j \neq k$. Luego $\langle \varphi_j, \varphi_k \rangle = 0$.

ii) Pasemos a probar que N es ortogonal, es decir,

$$\langle \psi_j, \psi_k \rangle = \int_{-l}^{L} \operatorname{sen}\left(\frac{j\pi x}{L}\right) \operatorname{sen}\left(\frac{k\pi x}{L}\right) dx = 0 \quad \forall j, k \in \mathbb{Z} \text{ con } j \neq k.$$

Sean $a, b \in \mathbb{R}$, de las identidades trigonométricas

$$\cos(a + b) = \cos a \cos b - \operatorname{sen} a \operatorname{sen} b, \qquad \cos(a - b) = \cos a \cos b + \operatorname{sen} a \operatorname{sen} b,$$

se obtiene

$$\operatorname{sen} a \operatorname{sen} b = \frac{1}{2}\left[\cos(a - b) - \cos(a + b)\right],$$

y poniendo $a = \frac{j\pi x}{L}$, $b = \frac{k\pi x}{L}$, resulta

$$\begin{aligned} \langle \psi_j, \psi_k \rangle &= \int_{-L}^{L} \frac{1}{2}\left[\cos\left(\frac{\pi(j-k)x}{L}\right) - \cos\left(\frac{\pi(j+k)x}{L}\right)\right] dx \\ &= \frac{1}{2}\left(\frac{L}{\pi(j-k)} \operatorname{sen}\left(\frac{\pi(j-k)x}{L}\right) - \frac{L}{\pi(j+k)} \operatorname{sen}\left(\frac{\pi(j+k)x}{L}\right)\right)\Bigg|_{-L}^{L} \\ &= 0 \quad \text{si } j \neq k. \end{aligned}$$

iii) Para mostrar que $M \perp N$, se debe probar que $\langle \varphi_0, \psi_k \rangle = 0 \quad k = 1, 2, \ldots$ y $\langle \varphi_j, \psi_k \rangle = 0 \quad j, k \in \mathbb{Z}^+$. La verificación se propone como ejercicio.

Definición 25 *Sea V un espacio con producto interior, $x, y \in V$. El ángulo $\theta \in [0, \pi]$ que forman los vectores x e y se define como* $\cos\theta = \frac{\langle x,y \rangle}{\|x\|\|y\|} \quad x \neq 0,\ y \neq 0.$

Teorema 10 (de Pitágoras) *Sea V un espacio con producto interior, $x, y \in V$. Si $x \perp y$ se tiene*

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 .$$

De manera mas general, sea $\{x_1, \cdots, x_n\}$ un conjunto ortogonal de V, entonces

$$\left\| \sum_{i=1}^{n} x_i \right\|^2 = \sum_{i=1}^{n} \| x_i \|^2 .$$

Demostración. De la definición de la norma $\|.\|$, se tiene

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + 2 \langle x, y \rangle + \langle y, y \rangle .$$

Por hipótesis $x \perp y$, luego $\langle x, y \rangle = 0$, y como $\langle x, x \rangle = \|x\|^2$, $\langle y, y \rangle = \|y\|^2$, se concluye

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 .$$

La prueba de la generalización del teorema de Pitágoras a una familia ortogonal finita se realiza por inducción y se propone como ejercicio. ■

En el caso de espacios vectoriales reales, se tiene que: si $x, y \in V$ tales que $\|x + y\|^2 = \|x\|^2 + \|y\|^2$ entonces $x \perp y$. Efectivamente, si $x, y \in V$ entonces $\|x + y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2$. Por hipótesis, $\|x + y\|^2 = \|x\|^2 + \|y\|^2$, y de la igualdad $\|x\|^2 + \|y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2$ de donde $\langle x, y \rangle = 0$, o sea $x \perp y$.

12.5. Lecturas complementarias y bibliografía

1. Owe Axelsson, Iterative Solution Methods, Editorial Cambridge University Press, Cambridge, 1996.
2. E. K. Blum, Numerical Analysis and Computation. Theory and Practice, Editorial Addison-Wesley Publishing Company, Reading, Massachusetts, 1972.
3. Richard L. Burden, J. Douglas Faires, Análisis Numérico, Séptima Edición, International Thomson Editores, S. A., México,2002.
4. P. G. Ciarlet, Introduction á l'Analyse Numérique Matricielle et á l'Optimisation, Editorial Masson, París, 1990.
5. James W. Demmel, Applied Numerical Linear Algebra, Editorial Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.
6. V. N. Faddeva, Métodos de Cálculo de Algebra Lineal, Editorial Paraninfo, Madrid, 1967.
7. Francis G. Florey, Fundamentos de Algebra Lineal y Aplicaciones, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1980.
8. Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence, Algebra Lineal, Editorial Publicaciones Cultural, S. A., México, 1982.
9. Noel Gastinel, Análisis Numérico Lineal, Editorial Reverté, S. A., Barcelona, 1975.

10. Gene H. Golub, Charles F. Van Loan, Matrix Computations, Second Edition, The Johns Hopkins University Press, Baltimore, 1989.

11. Kenneth Hoffman, Ray Kunze, Algebra Lineal, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1987.

12. Franz E. Hohn, Algebra de Matrices, Editorial Trillas, México, 1979.

13. Roger A. Horn, Charles R. Johnson, Matrix Analysis, Editorial Cambridge University Press, Cambridge, 1999.

14. A. N. Kolmogórov, S. V. Fomín, Elementos de la Teoría de Funciones y del Análisis Funcional. Editorial Mir, Moscú, 1972.

15. Peter Linz, Theoretical Numerical Analysis, Editorial Dover Publications, Inc., New York, 2001.

16. Anthony N. Michel, Charles J. Herget, Applied Algebra and Functional Analysis, Editorial Dover Publications, Inc., New York, 1981.

17. Ben Noble, James W. Daniel, Algebra Lineal Aplicada, Editorial Prentice-Hall Hispanoamericana, S. A., México, 1989.

18. Fazlollah Reza, Los Espacios Lineales en la Ingeniería, Editorial Reverté, S. A., Barcelona, 1977.

19. Gilbert Strang, Algebra Lineal y sus Aplicaciones, Editorial Fondo Educativo Interamericano, México, 1982.

20. Arthur Wouk, A Course of Applied Functional Analysis, Editorial John Wiley&Sons, New York, 1979.