SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Notas de la Materia Métodos Numéricos

Julio 2016

Clave ELC-1019

Lino Coria Cisneros

Ingeniería Eléctrica

Instituto Tecnológico de Morelia

INDICE

INTRODUCCIÓN A LOS MÉTODOS NUMÉRICOS

INTRODUCCIÓN. Aunque las bases de esta disciplina son antiguas, como lo atestiguan los nombres de muchos de los métodos, cobró enorme importancia hacia la segunda mitad del siglo XX. Existe referencia a matemáticos tan lejanos como el siglo XV, sin embargo los recursos para aplicar de manera eficiente se hizo posible con la aparición de las computadoras digitales electrónicas comerciales en los años 50's del siglo pasado. El nombre de análisis numérico se asoció por primera vez con el Instituto de Análisis Numérico (Institute of Numerical Analysis), fundado en 1947 en la Universidad de California en Los Angeles.

Es difícil encontrar una definición que satisfaga a todo mundo, como ocurre en muchos casos; sin embargo, para una posible definición podría la que propone P. Henrici [PH], y que se proporciona a continuación.

ANÁLISIS NUMÉRICO: Teoría de métodos constructivos en análisis matemático.

Por *métodos constructivos* se entiende un procedimiento que nos permite obtener la solución de un problema matemático con una precisión arbitraria, en un número finito de pasos que pueden realizarse racionalmente. El número de pasos depende de la exactitud deseada.

Un método constructivo usualmente consiste de un conjunto de instrucciones para el funcionamiento de ciertas operaciones aritméticas y lógicas, en un orden predeterminado. Un conjunto de instrucciones que conducen a la solución de un problema dado se denomina ALGORITMO. El análisis numérico está asociado con el diseño y análisis de algoritmos para resolver problemas matemáticos que surgen en ciencia e ingeniería. Por esta razón, cada vez es más común encontrar el término *Cálculo Científico* (scientific computing) alternativo al concepto de métodos numéricos. La mayoría de los problemas de la matemática continua no pueden resolverse en un número de pasos finito, por lo que se recurre a resolverlos por un proceso iterativo (teóricamente infinito) que converge a la solución.

En la búsqueda de la solución a un problema, una estrategia básica general, es reemplazar un problema complicado con otro más fácil que tenga una solución igual o

muy cercana a la solución de dicho problema. Como información al respecto, enumeramos las estrategias más comunes:

- Reemplazar un procesos infinitos, con procesos finitos; como reemplazar integrales o sumas infinitas con series finitas o derivadas con diferencias finitas
- Reemplazar matrices generales con matrices con formas más simples
- Reemplazar funciones complicadas con funciones simples como polinomios
- Reemplazar problemas no lineales con problemas lineales
- Reemplazar ecuaciones diferenciales con ecuaciones algebraicas
- Reemplazar sistemas de alto orden con sistemas de bajo orden

Una característica importante del cálculo científico es que sus resultados son aproximados, en oposición a la percepción de precisión que nos produce el aprendizaje de las matemáticas en la escuela. En la realidad, esto es en la aplicación del cálculo científico a ingeniería y ciencia, los resultados son aproximados, pero debemos tener conciencia de dicha aproximación y, en la medida de, lo posible, control del error involucrado en el cálculo. Realmente en el mundo real, muy pocos de los modelos matemáticos usados, no tienen solución analítica o de forma cerrada y por tanto la única opción que nos permite efectuar las cuantificaciones es el uso de los métodos numéricos. Este hecho deja clara la trascendencia de esta disciplina en la actividad científica e ingenieril. El efecto de las principales fuentes de inexactitud empieza antes del proceso de cálculo y se pueden resumir en los siguientes conceptos [MTH].

MODELADO. Es el proceso que, a partir de las leyes físicas que rigen el fenómeno sujeto del estudio, nos permite obtener una o un conjunto de ecuaciones matemáticas que replican el comportamiento del fenómeno (o dispositivo en el caso de ingeniería). En ocasiones se requiere simplificar u omitir algunas características del problema o estudio sujeto a estudio, cuyo resultado será una imprecisión en dicho modelo.

MEDICIONES EMPÍRICAS. La precisión de los instrumentos de medición, tienen una precisión limitada. Aunado a esto, la precisión puede verse afectada por un tamaño pequeño de la muestra utilizada, o bien por el efecto de ruido aleatorio a las que pueden ser sujetas las lecturas. Ejemplos de esto aparecen en todos los campos de estudio, en ingeniería y ciencias.

CÁLCULOS PREVIOS. Los datos de entrada usados en las simulaciones computacionales pueden estar basados en resultados previamente cuyos resultados son solo aproximaciones.

Las aproximaciones o inexactitudes mencionadas están fuera de nuestro control, pero juegan un papel muy importante en la exactitud de nuestros futuros cálculos. Por esta razón, en esta disciplina se da mucha importancia a las aproximaciones sobre las que si podemos tener influencia. Dichas aproximaciones ocurren en el transcurso del proceso numérico y se definen a continuación.

ERROR DE TRUNCAMIENTO O DISCRETIZACIÓN.

Este error es denominado por algunos autores atinadamente con el nombre de error del método. Se refiere a la aproximación asociada al método numérico; los ejemplos pueden ser muy variados y es inherente a los métodos numéricos y producto de su carácter finito. Por ejemplo el uso ineludible del número finito de términos de un serie infinita, la aproximación de una derivada como una diferencia finita.

ERROR REDONDEO O DE REPRESENTACIÓN.

El instrumento de cálculo empleado, la computadora, dispone obviamente de registros finitos para representar los números reales y por tanto los resultados de las operaciones aritméticas basadas en dichas cantidades son inexactos.

La exactitud de los resultados finales de un cálculo puede reflejar el efecto de la combinación de todas las aproximaciones mencionadas y los efectos asociados pueden resultar en perturbaciones que se amplifiquen debido a la naturaleza del problema o el algoritmo empleado. El estudio del efecto de dichas aproximaciones en la exactitud y estabilidad de los algoritmos numéricos se conocen como **análisis del error**.

Nota.-Es importante comentar que en la literatura de métodos numéricos y áreas relacionadas de los años 60's y 70's del siglo pasado, se pueden encontrar definiciones de diferentes y hasta contrapuestas a las expuestas aquí. Si se hace uso de literatura de esa época, de la cual hay ejemplos excelentes, hay que tomar en cuenta este hecho.

La definición básica del error, aunque muy obvia, es el inicio para analizar este concepto

Valor Verdadero = Aproximación + Error

De donde

E_t = Valor Verdadero – Aproximación (+/-)

El subíndice **t** significa verdadero, por lo que el término de la izquierda es el error verdadero, es decir, basado en el conocimiento del valor verdadero.

A partir de lo anterior, definimos el concepto del *error verdadero fraccional relativo*, que es un valor normalizado del error

$$\text{Error verdadero fraccional relativo} = \frac{\text{error verdaero}}{\text{valor verdadero}}$$

Una alternativa muy común, es expresar en forma porcentual este concepto

$$\text{Error verdadero relativo porcentual, } \varepsilon_t = \frac{\text{error verdadero}}{\text{valor verdadero}} \times 100\%$$

Aunque básicas y ciertas, las aplicaciones de las definiciones anteriores tienen el inconveniente de que el valor verdadero de una variable nunca se conoce, salvo en casos muy hipotéticos que se usan para ejemplificar algún procedimiento en los textos. Por lo que la aplicación de las últimas definiciones del error debe adecuarse para utilizarlas en los casos reales. Esto nos conduce a definiciones alternativas basadas en las cantidades de que realmente disponemos, por lo que tendríamos

$$\varepsilon_a = \frac{\text{Error Aproximado}}{\text{Aproximacion}} \times 100\%$$

La definición del *error aproximado porcentual*.

Muchos, si no es que la mayoría, de los métodos numéricos constituyen procesos iterativos, es decir, procedimientos que se repiten de manera sistemática para obtener valores a través de aproximaciones sucesivas. En este caso, el concepto

del error adecuado a este procedimiento estará en función de valores aproximados consecutivos de la variable de interés

$$\varepsilon_{\text{a}} = \frac{\text{Aproximacion Actual - Aproximacion Previa}}{\text{Aproximacion Actual}} \times 100\%$$

Es evidente que ante la imposibilidad de obtener un valor cero del error en el proceso iterativo, buscamos encontrar un valor lo más aproximado posible al valor ideal del error, es decir cero; lo anterior conduce a definir algún criterio que nos indique en qué punto del proceso (en qué iteración) debemos adoptar el valor de la variable como el adecuado, de acuerdo a nuestras necesidades de precisión. Lo anterior constituye lo que se denomina como *criterio de paro* y consiste en definir nosotros una cantidad de referencia, contra la cual comprara los sucesivos valores de la variable en cada iteración. Dicho criterio se define como

$$\left|\varepsilon_a\right| \leq \varepsilon_s$$

En este caso ε_s se denomina *tolerancia.*

Seleccionar el valor de la tolerancia usada en el proceso iterativo depende comúnmente de aspectos concretos que tienen que ver con la naturaleza física de la variable. Así por ejemplo, en el proceso iterativo asociado con la aplicación de flujos de potencia en un sistema de potencia, aunque siempre se calcula en base a cantidades normalizadas, usar el valor de la variable directamente involucrada, que es el voltaje nodal, no es recomendable; por lo que se recomienda usar un valor que más confiable para el criterio de convergencia, que depende de la variable, en este caso la potencia inyectada. Quién aplica estos métodos en el cálculo en ciencia o ingeniería, no tiene inconvenientes en seleccionar un valor adecuado. Sin embargo siempre queda la necesidad de definir este criterio en términos que no dependan de la naturaleza de la variable y que, sin embargo, sea confiable. Scarborough [JBS] propuso la siguiente definición, que cumple con este fin

$$\varepsilon_{\text{s}} = (0.5 \times 10^{(2\text{-n})})\%$$

La definición anterior implica que el resultado será correcto en al menos ***n*** cifras significativas.

NÚMEROS DE PUNTO FLOTANTE.

Los *errores de truncamiento*, es decir, errores del método se analizarán al final de cada uno de los métodos del curso. Éstos dependen de aspectos matemáticos que tienen que ver con el método y requieren de análisis para cada método. Los *errores de redondeo o de representación* son independientes del método empleado y tienen que ver esencialmente, como su nombre indica, con las limitantes que impone el tamaño de los registros disponibles en la computadora para almacenar los datos numéricos. Para entender este tipo de errores, es clave el análisis de la representación de los números reales en el formato de **punto flotante**; este es el formato empleado desde tiempos remotos en la historia de las computadoras. El material cubierto en este punto no es exhaustivo; lo que se discute a continuación es suficiente para entender este tipo de errores y su efecto en el proceso de cómputo numérico.

Un sistema de números de punto flotante se define como

$$x = \pm .b_1 b_2 b_3 \cdots b_t \times \beta^e$$

La cifra numérica $.b_1 b_2 b_3 \cdots b_t$ se denomina mantisa. . La longitud de la mantisa, o el número de términos que la conforman, es ***t***. El término es la base **β** elevada a un exponente ***e***. El exponente satisface $l \le e \le u$. b_i, *i=1,2⋯t* son los dígitos de la base β y satisfacen $0 \le b_i \le \beta - 1$.

Los números de punto flotante distintos de cero están normalizados; por tanto el ***cero*** es un número de punto flotante cuya mantisa y exponente son cero.

Supongamos un sistema de punto flotante definido como

$$(\beta, t, l, u) = (10, 2, -1, 2)$$

Existen 90 mantisas normalizadas positivas: **(.10,.11,.12,⋯,.98,.99)**, así como 4 posibles exponentes: **(-1,0,1,2).** Si consideramos los números negativos y el cero, el total de números de punto flotante será: **2x90x4+1=721**.

Además el número positivo más pequeño será: **m = 0.1 x 10^{-1} = 0.01** y el número más grande: **M = .99 x 10^2 = 99**. El conjunto de números de punto flotante es *finito* y su espaciamiento *no es uniforme*.

Si ***x*** es un número real, denotamos como ***fl(x)*** al número de punto flotante, es decir, a la versión almacenada de ***x***. El error absoluto relativo es

$$\frac{\left| fl(x) - x \right|}{\left| x \right|}$$

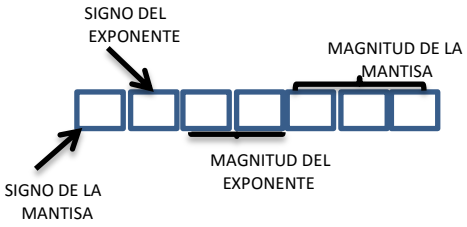
Para un sistema de números de punto flotante $F(\beta, t, l, u)$ donde $x \in \mathbb{R}$ y $\beta^l \langle |x| \langle \beta^u$, dicho error está acotado por

$$\frac{\left| fl(x) - x \right|}{\left| x \right|} \leq \frac{1}{2} \beta^{1-t}$$

A la cantidad de la derecha, $\frac{1}{2}\beta^{1-t}$, se le denomina el **ε** (épsilon) de la máquina, traducción del anglicismo **ε**-machine.

La discusión sobre estos temas puede parecer muy críptica inicialmente, por lo que puede ayudar a entender estas ideas de manera más amigable un ejemplo. Para no confundir con demasiadas cifras y su manejo, utilizamos un sistema numérico de punto flotante hipotético que se describe a continuación.

Suponemos que nuestra computadora hipotética dispone de un registro de 7 bits para representar números de punto flotante [ChC]. El primer dígito (a la izquierda) se emplea para el signo del números representado; el segundo dígito se utiliza para el signo del exponente; los siguientes dos dígitos, es decir, el 3º y 4º dígitos, se emplean para representar el exponente; los 3 dígitos restantes, 5º 6º y 7º dígito, representan el valor de la mantisa.



Presentamos la secuencia de números positivos representable en el caso hipotético del ejemplo. Es evidente que existe las misma cantidad de cifras representadas en la parte negativa de la recta numérica y sería idéntica a la mostrada a continuación en todo, salvo en el primer dígito del registro que, en este caso, sería un **1** en lugar del **0** mostrado, en cuyo caso representa el signo positivo.

EL primer bloque de números de punto flotante empieza con el número menor en la mantisa, asociado al menor valor del exponente, -3. No se considera la mantisa $\mathbf{(000)_2}$ debido a la normalización, razón por la cual empezamos con $\mathbf{(100)_2}$ que es igual a $\mathbf{(0.5)_{10}}$

$$0 \quad 111 \quad 100 = \left(1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{-3} = (0.062500)_{10}$$
$$0 \quad 111 \quad 101 = \left(1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{-3} = (0.078125)_{10}$$
$$0 \quad 111 \quad 110 = \left(1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{-3} = (0.093750)_{10}$$
$$0 \quad 111 \quad 111 = \left(1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{-3} = (0.0109375)_{10}$$

El segundo bloque sigue la misma secuencia en la mantisa por supuesto, pero ahora incrementamos el exponente de $(111)_2$ a $(110)_2$, es decir a, de $(2^{-3})_{10}$ a $(2^{-2})_{10}$

$$0 \quad 110 \quad 100 = \left(1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{-2} = (0.125000)_{10}$$
$$0 \quad 110 \quad 101 = \left(1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{-2} = (0.156250)_{10}$$
$$0 \quad 110 \quad 110 = \left(1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{-2} = (0.187500)_{10}$$
$$0 \quad 110 \quad 111 = \left(1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{-2} = (0.218750)_{10}$$

Procediendo de la misma manera, ahora incrementamos el exponente de $\mathbf{(110)_2}$ a $\mathbf{(101)_2}$, es decir, de $\mathbf{(2^{-2})_{10}}$ a $\mathbf{(2^{-1})_{10}}$

$$0 \quad 101 \quad 100 = \left(1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{-1} = (0.250000)_{10}$$
$$0 \quad 101 \quad 101 = \left(1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{-1} = (0.312500)_{10}$$
$$0 \quad 101 \quad 110 = \left(1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{-1} = (0.375000)_{10}$$
$$0 \quad 101 \quad 111 = \left(1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{-1} = (0.437500)_{10}$$

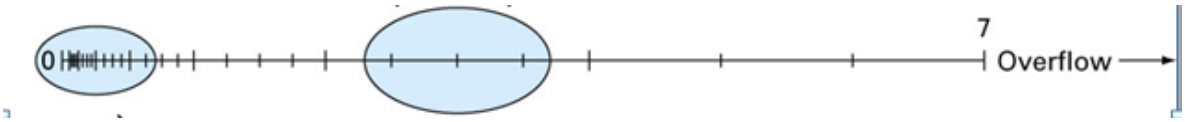
Aumentando el exponente con la misma variación de la mantisa podemos obtener todos los números representables por este registro hipotético. El último bloque, correspondiente al exponente mayor, es decir $\mathbf{(2^3)_{10}}$ o $\mathbf{(011)_2}$ se muestran a continuación

$$0 \quad 011 \quad 100 = \left(1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{3} = (4.0)_{10}$$
$$0 \quad 011 \quad 101 = \left(1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{3} = (5.0)_{10}$$
$$0 \quad 011 \quad 110 = \left(1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3}\right)\times 2^{3} = (6.0)_{10}$$
$$0 \quad 011 \quad 111 = \left(1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3}\right)\times 2^{3} = (7.0)_{10}$$

Lo primero que salta a la vista es que no podemos representar todas las cifras numéricas existentes, es decir, que hay un ***rango limitado*** de cantidades representables. Por lo que esta limitación explica por sí misma la importancia del concepto del ***error por redondeo o de representación numérica***.

Es importante notar en el patrón de representatividad numérica, que el espacio numérico entre cifras representables no es constante en toda la recta numérica representada. Así por ejemplo podemos observar que el espacio entre cada par de cifras numéricas del primer bloque es de $\boldsymbol{(0.015625)_{10}}$, mientras que en el segundo resulta $\boldsymbol{(0.031250)_{10}}$, y en el tercero $\boldsymbol{(0.062500)_{10}}$; finalmente, en el último bloque la diferencia entre los números es de $\boldsymbol{(1.0)_{10}}$. Con esto salta a la vista que además de no ser igual ese espacio entre números, se incrementa con la magnitud del número de punto flotante que queremos representar.

Gráficamente la distribución de las cantidades en el rango positivo de la recta numérica es como se muestra [ChC]



En forma resumida podemos decir que la finitud de ***e*** (el exponente) es una limitación en ***rango***; mientras que la finitud de ***t*** constituye una limitación en ***precisión***.

Lo anterior explica por qué se recomienda, y más aún, es indispensable, la normalización en los procesos numéricos y en general en los procesos de cálculo. Cuando normalizamos referimos las cifras numéricas a una base adecuada, que generalmente se recomienda sea una cantidad que tenga un valor alrededor de la media numérica de las cifras, con el fin de reducir el espectro numérico al que pertenecen las cifras usadas en el cálculo y con esto evitar, en lo posible, el manejo de números demasiado grandes los cuales tiene implícito un error por redondeo proporcional a su valor, es decir, muy grande también.

ERRORES DE TRUNCAMIENTO.

Los errores de truncamiento también se conocen como errores del método. El origen de estos proviene del diseño del algoritmo, que en la mayoría de los casos (quizás exceptuando los métodos directos para la solución de sistemas de ecuaciones linéales) éstos incluyen aproximaciones por truncamiento de series infinitas. Por esta razón, de manera particular, en cada método numérico se analiza dicho tipo de error.

Una de las formas más usadas para analizar el error de truncamiento usa la serie de Taylor; por lo tanto básicamente en esta sección, tratamos de dicha serie y su residuo de manera resumida, dejando para el apéndice de series infinitas algunos aspectos trascendentales a mayor detalle.

SERIE DE TAYLOR.

La serie de Taylor es un caso muy importante, y muy utilizado en el análisis de error por truncamiento, de las denominadas series de potencias [MLB]. En el apéndice de *series infinitas*, se discute en más detalle este tema. La serie de Taylor nos permite calcular, de forma aproximada, el valor de una función en un punto dado, basado en el conocimiento del valor de dicha función y sus derivadas en otro punto.

Enunciado en forma compacta el teorema de la serie de Taylor establece: si una función ***f*** y sus ***n+1*** derivadas son continuas en una intervalo que contiene a los puntos ***x*** y ***a***, entonces el valor de la función en el punto ***x*** estará dado por

$$f(x) = f(a) + f'(a)(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

$$|R_n| \le \left| M \frac{x^n}{n!} \right| = \left| M \frac{1}{n!} \right| \le 0.000001$$

$$\max f^{(n)}(\xi) \quad \text{y} \quad [x_0, x] = [0, 1]$$

$$f(x) = e^x, \quad f^{(n)} = e^x$$

$$e = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \cdots + \frac{1}{(n+1)!} + R_n$$

$$\left| f^{(n)}(\xi) \right| \le M = 3$$

$$\left| M \frac{1}{n!} \right| = 3\frac{1}{n!} \le 0.000001$$

$$n! \ge 3 \times 10^6$$

cuyo residuo R_n se define

$$R_n = \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt$$

Al valor de $x = a$ se le denomina frecuentemente como *punto base*.

Es importante familiarizarnos con el formato utilizado para la fórmula de Taylor, mostrada arriba, en el ámbito de los métodos numéricos. En los métodos numéricos generalmente aplicamos dicha fórmula en una función discretizada, es decir, una función definida en puntos concretos con cierta de frecuencia de muestreo, que está relacionada con el espacio entre puntos de la muestra.
Para esto, tomamos en general el intervalo alrededor del punto $\boldsymbol{X_i}$ y la fórmula de Taylor para la aproximación de la función en el punto $\boldsymbol{X_{i+1}}$ estará definida por

$$f(x_{i+1}) = f(\mathrm{x}_i) + f'(\mathrm{x}_i)\mathrm{h} + \frac{f^{(2)}(\mathrm{x}_i)}{2!}h^2 + \frac{f^{(3)}(\mathrm{x}_i)}{3!}h^3$$
$$+ \cdots + \frac{f^{(n)}(\mathrm{x}_i)}{n!}h^n + R_n$$

En cuyo caso el residuo se define

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x_{i+1} - \mathrm{x}_i)^{(n+1)} \qquad \xi \in [\mathrm{x}_i, x_{i+1}]$$

También se encontrará muy a menudo la expresión anterior modificada por la definición, $x_{i+1} - \mathrm{x}_i = h$

$$f(x_{i+1}) = f(\mathrm{x}_i) + f'(\mathrm{x}_i)\mathrm{h} + \frac{f^{(2)}(\mathrm{x}_i)}{2!}h^2 + \frac{f^{(3)}(\mathrm{x}_i)}{3!}h^3$$
$$+ \cdots + \frac{f^{(n)}(\mathrm{x}_i)}{n!}h^n + R_n$$

.

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{(n+1)} \qquad \xi \in [\mathrm{x}_i, x_{i+1}]$$

Un ejemplo puede ayudar a familiarizarnos con el uso de esta fórmula y su residuo.

Supongamos que queremos obtener el valor de ***e*** con un error menor o igual a 10^{-6}

Expandemos la serie de Taylor alrededor del punto ***x = 0***, la cual se denomina serie de Maclaurin, con el resultado

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Dado que nos interesa calcular el valor de ***e***, ***x = 1*** y entonces tenemos

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

Lo cual nos conduce a

$$e = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \cdots + \frac{1}{(n+1)!} + R_n$$

La condición del error requerido se reduce

$$\left|R_n\right| \le \left|M\,\frac{x^n}{n!}\right| = \left|M\,\frac{1}{n!}\right| \le 0.000001$$

Dado que M es $\max f^{(n)}(\xi)$ y $[x_0, x] = [0, 1]$.

Además $f(x) = e^x$, $f^{(n)} = e^x$ y en el intervalo $[x_0, x] = [0, 1]$, ***M = e*** (valor máximo de la derivada ***n-ésima***). Como el valor requerido es ***e*** y es igual con $\max f^{(n)}(\xi)$, asignamos un valor aproximado para el valor de ***e*** y suponiendo que $e < 3$, obtendremos

$\left|f^{(n)}(\xi)\right| \le M = 3.$

Regresando a la condición descrita arriba y tomando en cuenta lo anterior, obtenemos

$\left| M \frac{1}{n!} \right| = 3\frac{1}{n!} \leq 0.000001$. Lo anterior implica que $n! \geq 3 \times 10^6$, cumple los requisitos planteados.

Por intento-y-error podemos fácilmente encontrar: ***8!=40320 , 9!=362880, 10!=3628800***. Por lo que podemos concluir que requerimos ***n = 10*** términos de la serie para cumplir con el requisito.

Es importante tomar en cuenta que el único criterio que priva en este ejemplo es matemático, porque es muy importante tener en cuenta que este resultado, en la realidad, se vería afectado por el efecto del error por redondeo. Cuando hacemos cálculos numéricos debemos tener en cuenta el efecto de ambos errores. Hay en la literatura muy buenas referencias a este hecho. Por citar una, se sugiere consultar [BT].

SOLUCIÓN

DE ECUACIONES

NO LINEALES

DE UNA VARIABLE

El material del capítulo que actual, se refiere a la solución por métodos iterativos de ecuaciones no lineales. Estos métodos se basan de soluciones de raíces reales, que en el caso de los polinomios, limita su uso. Esto no quiere decir que no se puedan resolver las raíces de polinomios, pues en cuanto a sus raíces reales es perfectamente posible hacerlo con estos procedimientos, sin embargo sabemos que, en general, los polinomios con coeficientes reales pueden dar lugar a raíces complejas; los métodos que vamos a ver, salvo el caso del método de Newton, no permiten obtener dichas raíces. El método de Newton o Newton-Raphson, como también se le conoce, se puede adaptar para obtener raíces complejas [GW], aunque no se discute ese material en este curso. Para resolver las raíces de polinomios, existen métodos apropiados que tampoco se cubren en este curso, pero que abundan en la literatura sobre el tema.

Los métodos para resolver ecuaciones no lineales se dividen en dos grupos: ***métodos de intervalo o métodos cerrados*** y ***métodos abiertos***. Los nombres pueden variar de un libro a otro, pero su esencia no. Como el curso es un curso introductorio sobre el tema, el material dista mucho de ser exhaustivo, sin embargo trata de ser representativo en el sentido de que cubre el material básico, con el cual se puede profundizar en el tema. Por esta razón, se cubren dos métodos, los más representativos sobre el tema, en cada uno de los grupos mencionado.

MÉTODOS DE INTERVALO.

Estos métodos se caracterizan por trabajar con una "ventana", o sea, con un intervalo, que contiene la raíz de interés. De manera sistemática estos métodos redicen el intervalo o ventana que contiene la raíz, hasta que dicho intervalo se reduzca la cantidad suficiente para considerar que el resultado es preciso, de acuerdo a un criterio de convergencia. La diferencia entre éstos métodos consiste en la forma de reducir dicho intervalo.

Método de Bisección.

Este método es el más simple y consiste simplemente en que, una vez seleccionado el intervalo que contiene a la raíz, el intervalo se reduce dividiendo el intervalo a la mitad, de ahí el nombre del método, hasta que se cumpla el criterio de convergencia.

El intervalo de trabajo se delimita con dos valores que denominamos $\mathbf{X_l}$ y $\mathbf{X_u}$. En cada paso del proceso iterativo, requerimos determinar, además de que el intervalo completo contenga la raíz, una vez divido el intervalo en otros dos intervalos, cual sub-intervalo contiene la raíz, con el fin de desechar el que no la contenga y con ello reducir, en el caso de este método, a la mitad el intervalo original. El criterio empleado para encontrar si un intervalo contiene la raíz, es evaluar la función en los puntos extremos de dicho intervalo y verificar que hay un cambio de signo, que indica el cruce por cero de dicha función, esto es que $f(x_l) * f(x_u) < 0$.

Algoritmo del Método de Bisección.

Paso 1. Asegurarse de que la función contiene la raíz en el intervalo inicial, es decir que

$$f(x_l) * f(x_u) < 0$$

Paso 2. Dividimos el intervalo por mitad $x_m = \frac{(x_l + x_u)}{2}$.

Paso 3. Encuentre el sub-intervalo que contiene la raíz:

- Si $f(x_l) \cdot f(x_m) \langle 0$, entonces la raíz está en el intervalo $[x_l, x_m]$. Desechamos el intervalo $[x_m, x_u]$ asignando $x_u \Leftarrow x_m$. Probar convergencia: ¿converge? Ir al paso terminar, la raíz es x_m. ¿No converge? Ir al paso 2.
- Si $f(x_l) \cdot f(x_m) \rangle 0$, entonces la raíz está en el intervalo $[x_m, x_u]$.
- Desechamos el intervalo $[x_l, x_m]$ asignando $x_l \Leftarrow x_m$. Probar convergencia: ¿converge? Ir al paso terminar, la raíz es x_m. ¿No converge? Ir al paso 2.

- Si $f(x_l) \cdot f(x_m) = 0$ (caso por demás improbable), entonces la raíz está en x_m.

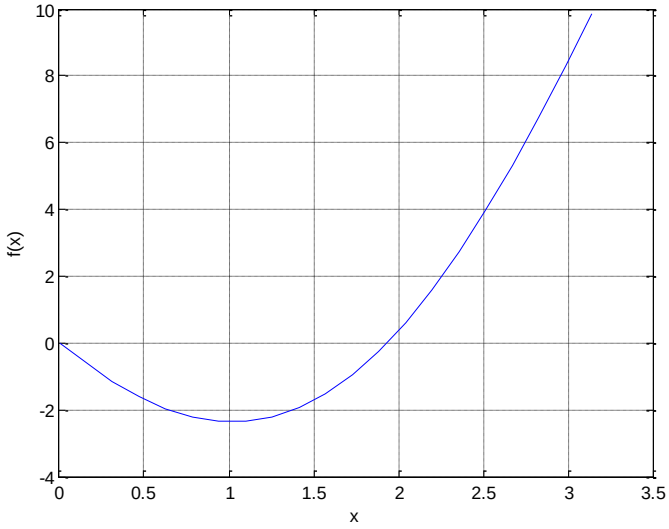
La prueba de convergencia puede hacerse de varias formas. La primera podría ser una prueba directa, que consiste en evaluar la función y comprar con el criterio de convergencia ε_s, que será seleccionado de acuerdo a los criterios mencionados en el capítulo anterior. En caso de que $\varepsilon_s \leq f(x_m)$, hemos encontrado la solución. Otra opción, que es la más utilizada en la literatura, no por ello la más adecuada en mi opinión, es hacer una prueba indirecta, calculando el error aproximado relativo porcentual ε_a que en este caso se calcularía como $\frac{|x_l - x_m|}{|x_l - x_u|} \times 100\%$ o bien $\frac{|x_u - x_m|}{|x_l - x_u|} \times 100\%$, según sea el caso.

Al reducir el intervalo a través de bisección, este método tiene la característica de que se puede anticipar el número de iteraciones que tomaría encontrar la solución [ChC].

Inicialmente, el error aproximado es $E_a^0 = x_u^0 - x_l^0 = \Delta x^0$; después de la primera iteración, el error se ha reducido $E_a^1 = \frac{\Delta x^0}{2}$. Dado que cada iteración divide exactamente a la mitad el intervalo, podemos deducir que después de ***n*** iteraciones, el error es $E_a^n = \frac{\Delta x^0}{2^n}$. Si denominamos como $E_{a,d}$ al error deseado, entonces tendremos

$$n = \frac{\log\left(\Delta x^0 / E_{a,d}\right)}{\log 2} = \log_2\left(\frac{\Delta x^0}{E_{a,d}}\right).$$

Complementando con un ejemplo, consideremos la ecuación $f(x) = x^2 - 4\,sen(x) = 0$ y supongamos un intervalo de inicio $[x_l, x_u] = [1, 3]$.



GRÁFICA DE LA FUNCIÓN $f(x) = x^2 - 4\,sen(x) = 0$

EL proceso iterativo inicia, una vez que estamos seguros de que el intervalo de inicio contiene la raíz, con el cálculo del valor medio $x_m^0 = \frac{x_l + x_u}{2} = \frac{1+3}{2} = 2$. Indagamos en cuál de los sub-intervalos está la raíz, con el objeto de desechar el que no la contiene:

$$f(x_l) * f(x_m^0) = f(1) * f(2) = (-2.36588)(0.36281) < 0$$

Lo anterior implica que en el intervalo $\left[x_l, x_m^0\right]$ está contenida la raíz, por tanto $x_u \Leftarrow x_m^0$, con lo que el nuevo intervalo será $\left[x_l, x_u\right]^1 = [1, 2]$. Se revisa si el resultado es satisfactorio, como se mencionó anteriormente y en caso positivo el valor medio del intervalo de esta iteración sería el resultado. Caso contrario la determinación de $\left[x_l, x_u\right]^1$ da inicio a la segunda iteración, como se muestra a continuación.

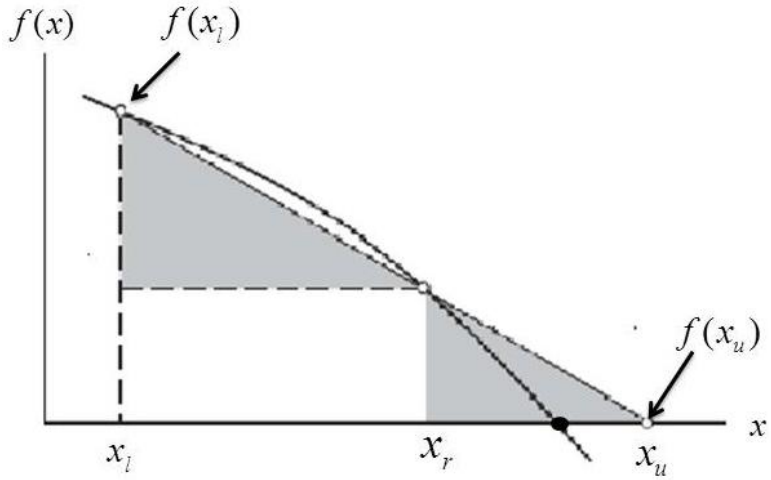
Iniciamos determinando el valor medio del intervalo actual $x_m^1 = \frac{x_l + x_u}{2} = \frac{1+2}{2} = 1.5$ y

de nueva cuenta indagamos cual sub-intervalo contiene la raíz:

$f(x_l) * f(x_m^1) = f(1) * f(1.5) = (-2.36588)(-1.73998) \rangle 0$ lo cual implica que la raíz está contenida en $\left[x_m^1, x_u \right]$ y por tanto el nuevo intervalo es $\left[x_l, x_u \right]^2 = [1.5, 2]$. Se prueba si ya se obtuvo convergencia y se concluye como se hizo antes. El resultado final corrido en MATLAB® con formato *long*, resulta, en *22 iteraciones* un valor de la raíz de ***1.933753490447998***, con valor fe la función igual a ***-1.440250095186002e-006***

MÉTODO DE LA FALSA POSICIÓN.

El método de la *falsa posición* o también conocido como *regula falsi*, es una alternativa para reducir más rápido el intervalo durante el proceso iterativo de los métodos abiertos o de intervalo. En este método se aprovecha la geometría de la función para acelerar dicha reducción, al encontrar el cruce por cero de una cuerda que una los puntos $\mathbf{X_l}$ y $\mathbf{X_u}$. Es como si aproximáramos la función f(x) en dicho intervalo por la recta que une los puntos mencionados y determinamos el punto que cruza el eje ***x*** dicha recta [ChC].



Representación geométrica del método de la falsa posición

Nuestro objetivo es encontrar una expresión para determinar el punto $\mathbf{X_r}$ lo cual se puede obtener de varias maneras; una es utilizando la ecuación de la recta que pasa por dos puntos, $\mathbf{X_l}$ y $\mathbf{X_u}$, y la otra por triángulos semejantes. Ésta última es la más fácil.

En función de los triángulos que se muestran sombreados o achurados, podemos escribir

$$\frac{f(x_l)}{x_r - x_l} = \frac{f(x_u)}{x_r - x_u}$$

De la cual despejamos el punto buscado, que con un poco de algebra extra [ChC], nos conduce a la ecuación que usará en la determinación del punto que dividirá el intervalo $\left[x_l, x_u\right]$ en los sub-intervalos $\left[x_l, x_r\right]$ y $\left[x_r, x_u\right]$

$$x_r = x_u - \frac{f(x_u)(x_l - x_u)}{f(x_l) - f(x_u)}.$$

Salvo casos muy especiales, éste método converge más rápido que el de bisección. Un análisis más detallado en cuanto análisis matemático de la convergencia se puede encontrar en la bibliografía, [GW], [BT] por ejemplo.

EL mismo ejemplo de la sección anterior, en el que se quiere determinar el cero correspondiente a la función $f(x) = \mathrm{x}^2 - 4\,sen(x) = 0$ $\qquad f(x_l)\cdot f(x_u)\langle 0$ confirma que la raíz está en el intervalo inicial.

La primera iteración inicia calculando el punto que divide el intervalo

$$x_r^0 = x_u - \frac{f(x_u)(x_l - x_u)}{f(x_l) - f(x_u)} = 3 - \left(8.435520\right)\frac{1-3}{\left(-2.365884 - 8.435520\right)} = 1.41885$$

Evaluamos la función en este punto $f(x_r^0) = \left(1.41885\right)^2 - 4\,sen(1.41885) = -1.94078$

Por lo que $f(x_l)\cdot f(x_r^0)\rangle 0$ sabemos que el nuevo intervalo es $\left[x_l, x_u\right]^1 = \left[1.41885, 3\right]$.

La segunda iteración es necesaria dado que no se cumple el criterio de convergencia, por lo que calculamos el nuevo punto para el intervalo $\left[x_l, x_u\right]^1 = \left[1.41885, 3\right]$

$$x_r^1 = x_u - \frac{f(x_u)(x_l - x_u)}{f(x_l) - f(x_u)} = 3 - \left(8.435520\right)\frac{1.41885 - 3}{\left(-1.94078 - 8.435520\right)} = 1.71434$$

Dado que $f(x_r^1) = -1.0199$; $f(x_l) \cdot f(x_r^1) \rangle 0$ por lo que $\left[x_l, x_u\right]^2 = \left[1.41885, 3\right]$.

La solución numérica completa por medio de la computadora resulta: en ***14 iteraciones*** un valor de la raíz de ***1.933753331679808***, con valor fe la función igual a

- 4.673004150301807e-007.

MÉTODOS ABIERTOS.

La característica de estos métodos es que, a diferencia de los anteriores, no se requiere un intervalo, por lo que el inicio del proceso de búsqueda de la raíz se hace con un punto. En estas notas se describen tres métodos, dos de los cuales son los más representativos de esta familia de métodos.

MÉTODO DE ITERATIVO PUNTO FIJO.

Iniciamos con las definiciones esenciales asociadas con este método.

Definición: Sea $g(x)$ una función en los reales y x un valor tal que

$$x = g(x)$$

Entonces x se denomina un **punto fijo** de la función g, dado que x permanece sin cambio cuando g se aplica a dicho valor.

Los problemas de punto fijo tiene valor en matemáticas en sí mismos, pero en nuestro caso lo importante consiste en que nos permiten replantear las funciones no lineales como un problema de punto fijo, con el fin de resolver dichas ecuaciones no lineales.

Muchos algoritmos iterativos para resolver funciones no lineales se basan en formulaciones del tipo

$$x_{k+1} = g(x_k)$$

En este caso se selecciona la función g de manera que sus puntos fijos sean soluciones de la función $f(x)=0$. Dicho algoritmo se denomina ***iteración de punto fijo*** o ***iteración funcional***, dado que la función g se aplica de manera repetida a partir de un valor inicial X_0.

Una función no lineal $f(x)=0$ puede plantearse de varias formas como un problema de punto fijo $x= g(x)$, pero no todas las formulaciones resultantes son igualmente adecuadas; de hecho el problema puede ir desde unas opciones con mayor rapidez de convergencia que otras y en muchos casos inclusive formulaciones que resulten ***divergentes*** [MTH].

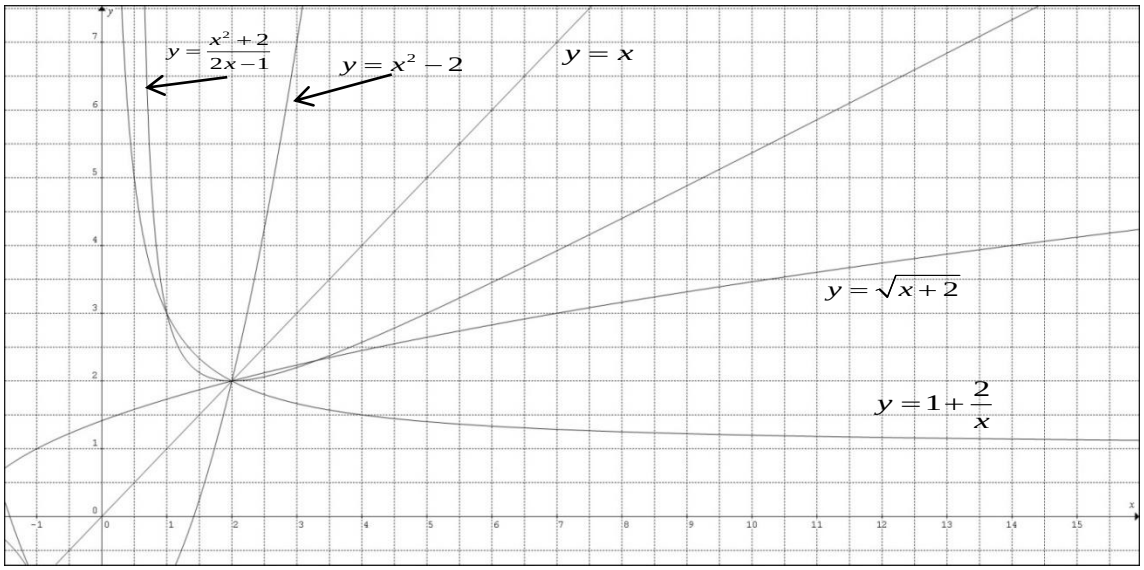
Lo más adecuado es ilustrar lo anterior con un ejemplo [MTH]. Supongamos una función no lineal, simple para ilustrar con facilidad lo que se quiere, definida por $f(x) = x^2 - x - 2 = 0$. Esta ecuación tiene dos raíces reales en $x = -1 \quad x = 2$.

Debemos obtener una función de punto fijo de la forma mencionada anteriormente; en este punto vemos que puede haber varias formas de obtener dicha función de punto fijo, es decir, $g(x)$ puede tener varias formas. De hecho, vamos a analizar este caso con cuatro formas de obtener la función de punto fijo, las cuales se enumeran

$$g_1(x) = x^2 - 2 \qquad g_2(x) = \sqrt{x+2}$$

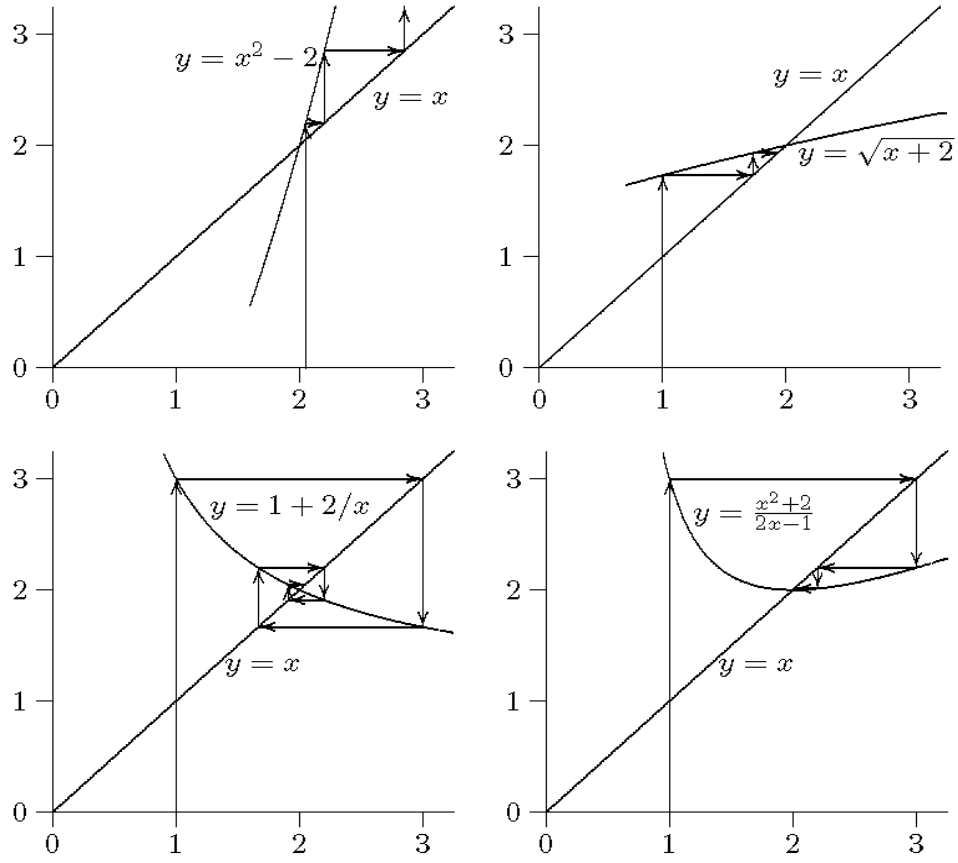
$$g_3(x) = 1 + \frac{2}{x} \qquad g_4(x) = \frac{x^2+2}{2x-1}$$

La gráfica mostrada a continuación, muestras estas cuatro funciones, junto con la función de punto fio complementaria del algoritmo $f(x) = x$.



Resulta que no todas estas funciones de punto fijo se comportan de la misma manera; no solamente presentan rapidez de convergencia distinta en general, sino que incluso puede esperar que exista divergencia en alguna de ellas. Este el caso como vamos a mostrar a continuación.

La gráfica que se muestra a continuación, contiene los cuatro el resultado, en forma geométrica, de la simulación de las cuatro funciones de punto fijo descritas por las ecuaciones anteriores y en el mismo orden de éstas, concordante con el subíndice de dichas ecuaciones.



1er CASO	2º CASO	3er CASO	4º CASO
se exedió el máximo número de iteraciones	**k =12.0000e+000**	**k =23.0000e+000**	**k = 7.0000e+000**
k = 100.0000e+000	**p = 2.0000e+000**	**p = 2.0000e+000**	**p = 2.0000e+000**
p = Inf	**err = 784.3656e-009**	**err =1.0729e-006**	**err =698.4919e-012**
err = NaN	P =1.0000e+000	P =1.0000e+000	P =1.0000e+000
P = 2.1000e+000	1.7321e+000	3.0000e+000	3.0000e+000
2.4100e+000	1.9319e+000	1.6667e+000	2.2000e+000
3.8081e+000	1.9829e+000	2.2000e+000	2.0118e+000
12.5016e+000	1.9957e+000	1.9091e+000	2.0000e+000
154.2906e+000	1.9989e+000	2.0476e+000	2.0000e+000
23.8036e+003	1.9997e+000	1.9767e+000	2.0000e+000
566.6115e+006	1.9999e+000	2.0118e+000	
321.0486e+015	2.0000e+000	1.9942e+000	
103.0722e+033	2.0000e+000	2.0029e+000	
10.6239e+069	2.0000e+000	1.9985e+000	
112.8668e+138	2.0000e+000	2.0007e+000	
12.7389e+279		1.9996e+000	
Inf		2.0002e+000	
Inf		1.9999e+000	
Inf		2.0000e+000	
Inf		2.0000e+000	
Inf		2.0000e+000	
		2.0000e+000	
		2.0000e+000	
		2.0000e+000	
		2.0000e+000	
		2.0000e+000	

En las páginas anteriores se muestran las cuatro simulaciones correspondientes a las funciones punto fijo, en el mismo orden de las gráficas mostradas. En estos textos, los el parámetro **k**: es el número de iteraciones, **p**: es el valor de la raíz, **err**: es el error (se usó un criterio de convergencia de 10^{-6}) y finalmente **P** muestra la secuencia de valores del proceso iterativo. Hay que recordar que el valor ***inf*** en MATLAB®, así como ***NAN***, significan respectivamente *infinito* y *Not A Number*.

¿Qué es lo que caracteriza a los casos divergentes?. EL procedimiento siguiente demuestra la condición de convergencia [ChC], que probablemente al observar las gráficas ya haya descubierto.

Iniciamos con la fórmula recursiva del método de punto fijo

$$x_{i+1} = g(x_i)$$

Suponemos que la solución verdadera está dado por

$$x_r = g(x_r)$$

Restamos estas dos ecuaciones y resulta

$$x_r - x_{i+1} = g(x_r) - g(x_i) \qquad (*)$$

TEOREMA DEL VALOR MEDIO

Si una función $g(x)$ y su primera derivada
son continuas en el intervalo $a \le x \le b$
entonces existe un valor de $x = \xi$ en dicho intervalo,
tal que

$$g'(\xi) = \frac{g(b) - g(a)}{b - a}$$

Lo anterior nos dice simplemente que si tenemos una curva delimitada por el intervalo mencionado, en algún punto dentro del intervalo, identificado como $x=\xi$, $\xi \in [a,b]$, la pendiente de la recta tangente a la curva en dicho punto es paralela a la cuerda que une los puntos $(a, f(a))$ y $(\mathrm{b}, f(\mathrm{b}))$.

Si aplicamos este teorema al caso que nos interesa tendremos

$$g(x_r) - \mathrm{g}(x_i) = (x_r - x_i)\, g'(\xi)$$

con $\xi \in [x_i \; x_r]$.

Aplicado esto al caso nuestro tenemos y tomando en cuenta la ecuación (*), obtenemos

$$x_r - x_{i+1} = (x_r - x_i)\, g'(\xi)$$

Siendo x_r la solución verdadera de la función, la diferencia de ésta con cualquier valor de la variable en el proceso iterativo, es igual al error verdadero en la iteración correspondiente, por lo que de acuerdo con esto

$$x_r - x_i = E_{t,i} \qquad x_r - x_{i+1} = E_{t,i+1}$$

Y sustituyendo en la ecuación anterior obtenemos

$$E_{t,i+1} = g'(\xi)\, E_{t,i}$$

- Si $|g'(\xi)| < 1$, los errores decrecen
- Si $|g'(\xi)| > 1$, los errores se incrementan

Por otro lado:

- Si la derivada es positiva, los errores serán positivos y la solución es monotónica
- Si la derivada es negativa, los errores oscilarán

La conclusión de la última ecuación, sintetizada en el cuadro, es que para que exista convergencia, se deben obtener valores en cada iteración más pequeños que el error de la iteración anterior, es decir que

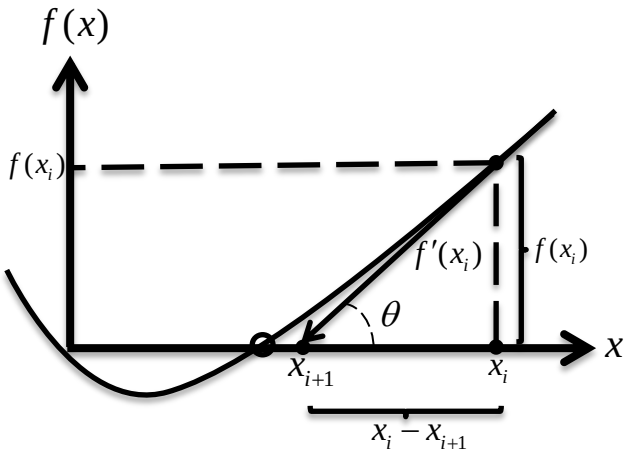
$$\frac{E_{t,i+1}}{E_{t,i}} = g'(\xi)\langle 1$$

Que es lo mismo que afirmar que la condición de convergencia requiere que la pendiente de la función de punto fijo debe ser menor de 1.

MÉTODO DE NEWTON.

Desde el punto de vista geométrico, en lugar de usar funciones de punto fijo, un método iterativo más robusto y estable consiste en utilizar la recta tangente a la función no lineal, para encontrar puntos de cruce de dicha línea recta, aproximarnos a la raíz buscada. Esta es la esencia del método de Newton o Newton-Raphson. Es un método sencillo y de convergencia más rápida que el método de punto fijo; la ´desventaja más común que le atribuyen es que requiere de la evaluación de la derivada en cada iteración, lo cual puede constituir un proceso costoso computacionalmente. En el vecindario de la solución, este método es muy rápido, pues lo caracteriza una convergencia cuadrática. Sin embargo, globalmente puede ser cuestionada esta ventaja, por lo que debemos iniciar el proceso iterativo con suficiente cercanía a la raíz deseada.

Geométricamente podemos visualizar de forma sencilla este método.



En base a la figura anterior, el punto de la iteración x_{i+1} se calcula de forma gráfica a partir de

$$f'(x_i) = \tan\theta = \frac{f(x_i)}{x_i - x_{i+1}}$$

De donde obtenemos

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Una vez calculado un nuevo punto, tomamos ese punto como una actualización del proceso iterativo y en caso de no estar suficientemente cerca, de acuerdo con un criterio de convergencia, en base a ese punto calculamos el valor de la función y su derivada y la fórmula anterior nos permitirá encontrar un nuevo punto, con el que procederemos de nueva cuenta a evaluar si cumple con el criterio de convergencia y, en caso negativo, a proceder con el proceso iterativo.

Un ejemplo ilustra la forma de proceder descrita arriba. Supongamos la función no lineal $f(x) = x^2 - 4sen(x) = 0$, usada anteriormente. Seleccionamos el punto inicial $x_0 = 1.5$ y un criterio de convergencia $\varepsilon = 10^{-6}$.

La derivada de la función, necesaria en este método, es $f'(x) = 2\mathrm{x} - 4\cos(\mathrm{x})$. Con esto la primera iteración inicia evaluando la función y su derivada en el punto inicial $f(1.5) = -1.73998 \quad y \quad f'(1.5) = 2.71705$. Con esto recurrimos al cálculo de actualización de la variable en la 1ª iteración

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1.5 - \frac{-1.73998}{2.71705} = 2.14039.$$

En este punto probamos convergencia; existen dos opciones: la primera, y muy utilizada en muchos textos [ChC], es calcular el error relativo porcentual

$$\varepsilon_a = \left|\frac{2.14039 - 1.5}{2.14039}\right| \cdot 100 = 29.92\%$$

La 2ª opción, que en mi opinión es más adecuada, ya que requiere calcular la función en el valor actualizado de la variable, valor necesario para efectuar la siguiente iteración y que nos proporciona una prueba directa de convergencia, es decir

$$f(x_1) = f(2.14039) \leq \varepsilon_S \, ?$$

Dado que $f(2.14039) = 1.21279 \gg \varepsilon_S (= 10^{-3})$, requerimos seguir el proceso iterativo.

Aunque la 2ª iteración se puede decir que empezó en párrafo anterior, realmente en este caso ese sería el final de la iteración anterior, pues si se cumple con el criterio de convergencia el método termina en ese punto. En este caso debemos evaluar la derivada de la función para proceder a actualizar la variable calculando x_2, para cuyo efecto evaluamos primero

$$f'(x_1) = 2 * 2.14039 - 4 * \cos(2.14039) = 6.43794$$

Con lo cual evaluamos

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 2.14039 - \frac{1.21279}{6.43794} = 1.95201$$

Y después de calcular y probar criterio de convergencia:

$$f(x_2) = f(1.95201) = 0.097488 \leq \varepsilon_S \, ?$$

Iniciamos la 3ª iteración, calculando la derivada

$$f'(x_2) = 2 * 1.95201 - 4 * \cos(1.95201) = 5.39221$$

Y con esto calculamos

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 1.95201 - \frac{0.097488}{5.39221} = 1.93393$$

Y dado que

$$f(x_3) = f(1.93393) = 0.000932 \leq \varepsilon_S (= 10^{-3})$$

La raíz será $x_r = x_3 = 1.93393$.

ERROR DEL MÉTODO DE NEWTON.

Para obtener el error de truncamiento o error del método en este caso, es muy importante que entendamos que aunque desarrollamos la fórmula que caracteriza a este método con un enfoque geométrico, en realidad un enfoque más analítico requiere considerar que dicho método se obtiene a partir de una aproximación de primer orden de la serie de Taylor; en efecto si nosotros partimos de dicha serie

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \Re_1(x_{i+1})$$

Y aproximamos hasta el término de la primera derivada, es decir, eliminamos el término del residuo, obtenemos

$$f(x_{i+1}) \simeq f(x_i) + f'(x_i)(x_{i+1} - x_i)$$

Dado que en el cruce por cero del eje-x es $f(x_{i+1}) = 0$

$$0 = f(x_i) + f'(x_i)(x_{i+1} - x_i)$$

Despejamos el valor correspondiente de la variable y obtenemos la consabida fórmula

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Como ya vimos antes, el valor verdadero de la raíz de la función se denomina X_0, por lo que

$$f(x_r) = 0 = f(x_i) + f'(x_i)(x_r - x_i) + \frac{f''(\xi)}{2!}(x_r - x_i)^2$$

Restando de esta ecuación la aproximación propuesta de primer orden, de la cual se eliminó el residuo, obtenemos

$$0 = f'(x_i)(x_r - x_{i+1}) + \frac{f''(\xi)}{2!}(x_r - x_i)^2$$

Recordamos que la definición de error verdadero es la diferencia del valor verdadero menos en valor aproximado, por lo que es evidente que

$$E_{t,i+1} = x_r - x_{i+1} \qquad y \qquad E_{t,i} = x_r - x_i$$

Por lo que la ecuación anterior se convierte en

$$0 = f'(x_i)\, E_{t,i+1} + \frac{f''(\xi)}{2!} E_{t,i}^2$$

Y finalmente

$$E_{t,i+1} = \frac{-f''(x_r)}{2 f'(x_r)} E_{t,i}^2$$

Esto significa:

- En una iteración dada, el error en este método será proporcional al cuadrado del error de la iteración anterior
- Lo anterior implica que el número de cifras decimales correctas, se duplica en cada iteración
- Esta característica se conoce como *convergencia cuadrática*.

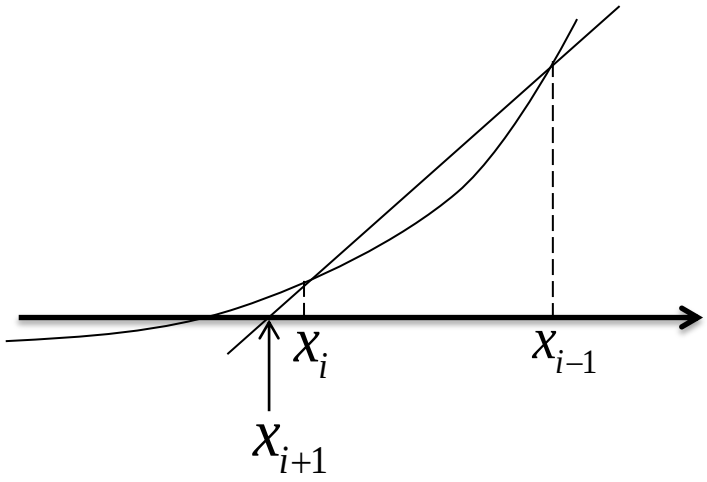
MÉTODO DE LA SECANTE.

Muchos autores ven en el cálculo de la derivada en cada iteración un inconveniente en el método de Newton; esto puede tener un costo computacional importante si la función es trigonométrica o trascendental, pues su evaluación requiere del uso de series.

Una solución es calcular la derivada numéricamente a través de cocientes de diferencias finitas, como se verá en la siguiente unidad. Sin embargo esto requiere también calcular la función correspondiente en cada derivada. Con el fin de evitar dicha evaluación, se formuló un método que hace uso de los valores de las funciones evaluadas en iteraciones sucesivas, en cuyo caso la función utilizada debe evaluarse de todas maneras, con el fin usarlas para aproximar la mencionada diferencia finita. Esta propuesta conduce al *método de la secante*, cuyo algoritmo está definido por la ecuación

$$x_{i+1} = x_i - f(x_i)\frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})}$$

En este método la función se aproxima por medio de la recta secante y el cruce por cero de dicha línea recta constituye una aproximación de la raíz. La gráfica siguiente ilustra lo anterior



Es interesante observar que, a diferencia del método de Newton que usa un punto de arranque, este método requiere de un intervalo para generar los dos puntos requeridos correspondientes a iteraciones sucesivas.

Un ejemplo nos ayuda a precisar lo que se ha comentado en los párrafos anteriores.

Resolvemos la misma función que usamos en el ejemplo del método anterior, es decir,

$$f(x) = x^2 - 4sen(x) = 0 \quad con\ \varepsilon = 10^{-6}$$

En este caso arrancamos con dos puntos inicialmente, los cuales serán

$$x_0 = 1.5 \quad y \quad x_1^i = 2.0$$

El superíndice del segundo número se usa con el fin de diferenciarlo de los siguientes valores, los cuales provienen del cálculo de cada iteración.

Iniciamos evaluando las funciones de los puntos de arranque que serán utilizadas en la fórmula de la secante

$$f(x_0) = f(1.5) = 1.73998 \quad y \quad f(x_1^i) = f(2.0) = 0.36281$$

Con esto calculamos

$$x_1 = x_0 - f(x_0)\frac{x_0 - x_1^i}{f(x_0) - f(x_1^i)} = 1.5 - \frac{1.5 - 2.0}{-1.73998 - 0.36281} = 1.91373$$

Evaluamos probamos criterio de convergencia

$$f(x_1) = f(1.91373) = -0.104727 \leq \varepsilon_S\ ?$$

En cuyo caso, requerimos al menos otra iteración. Iniciamos evaluando el valor de la variable correspondiente a esta iteración

$$x_2 = x_1 - f(x_1)\frac{x_1 - x_0}{f(x_1) - f(x_0)} = 1.91373 - (-0.104727)\frac{1.91373 - 1.5}{-0.104727 - (-1.73998)} = 1.94023$$

Y procediendo de la manera mencionada

$$f(\mathrm{x}_2) = 0.034364 \rangle \varepsilon_S (= 10^{-3})$$

Por lo que

$$x_3 = x_2 - f(x_2)\frac{x_2 - x_1}{f(x_2) - f(x_1)} = 1.94023 - (0.034364)\frac{1.94023 - 1.91373}{0.034364 - (-0.104727)} = 1.93368$$

Y dado que

$$f(x_3) = -0.00039 \leq \varepsilon_S (= 10^{-3})$$

Concluimos que este último valor es la raíz buscada

$$x_r = x_3 = 1.93368$$

INTERPOLACIÓN

INTRODUCCIÓN.

El principio de la interpolación consiste en obtener, a partir de una colección de puntos o pares ordenados $\mathbf{x_i,f(x_i)}$, un polinomio, denominado **polinomio interpolante** o simplemente **interpolante**, que pase por todos esos puntos y que aproxime a dichos puntos en el intervalo definido por la colección de puntos; en dichos puntos, llamados *nodos*, el valor del interpolante es el mismo que el de los datos, siendo el valor del polinomio interpolante una aproximación en el resto del intervalo. Si $x \in [x_1, x_2, \cdots, x_{n+1}]$ hablamos de **interpolación**; si por el contrario $x \notin [x_1, x_2, \cdots, x_{n+1}]$, entonces hablamos de **extrapolación**. En el enunciado anterior **x** es el valor cuya ordenada o valor de la función **f(x)** queremos aproximar.

Lo anterior establece la definición de interpolación en su sentido más general. La utilidad de la interpolación, sin embargo, va más allá del problema estipulado arriba. Sus usos son muy variados y, en el caso nuestro, es muy importante también porque una familia de métodos de integración numérica, fórmulas de Newton-Cotes, se basa precisamente en calcular la aproximación de una integral usando interpolantes en lugar de la función original, como se discutirá en la siguiente unidad.

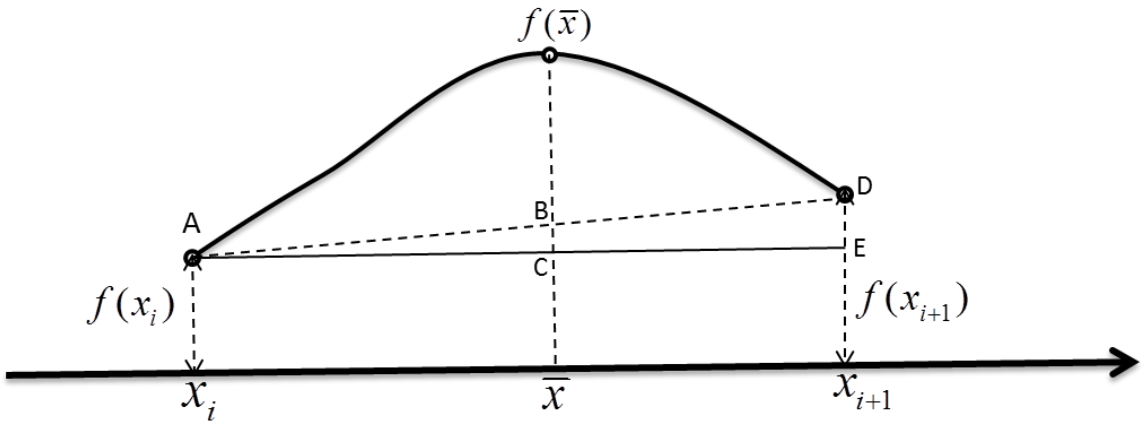
INTERPOLACIÓN LINEAL.

Las fórmulas de la interpolación lineal comúnmente se llegan a utilizar en cursos de física y matemáticas a nivel básico, sin que se conozca el concepto de manera precisa. En esta sección discutimos la interpolación lineal o de primer orden con detalle, como base para generalizar más adelante el manejo de interpolantes de cualquier orden.

Suponemos una lista de *n* puntos $(x_1, f(x_1)), (x_2, f(x_2)), (x_3, f(x_3)) \cdots (x_i, f(x_i)), (x_{i+1}, f(x_{i+1})) \cdots (x_n, f(x_n))$. Requerimos calcular el valor de la ordenada para un punto cuyo valor **x** se conoce $x = \overline{x}$; sabemos además que $\overline{x} \in [x_i, x_{i+1}]$.

La siguiente gráfica define el problema .

Conociendo $\overline{x}$ requerimos obtener el valor de $f(\overline{x})$.



El valor verdadero de $f(\overline{x})$ se aproxima como la ordenada a la recta $\overline{AD}$, el cual denominamos $f_{\text{int}}(\overline{x})$. De dicha figura vemos que

$$\frac{BC}{AC} = \frac{DE}{AE}$$

De donde obtenemos

$$BC = \frac{AC}{AE} DE = \frac{\overline{x} - x_i}{x_{i+1} - x_{i+1}} \left[f(x_{i+1}) - f(x_i) \right]$$

Además

$$f_{\text{int}}(\overline{x}) = f(x_i) + BC$$

$$f_{\text{int}}(\overline{x}) = f(x_i) + \frac{\overline{x} - x_i}{x_{i+1} - x_{i+1}} \left[f(x_{i+1}) - f(x_i) \right]$$

POLINOMIO INTERPOLANTE.

Como se discutió en la sección anterior, dos puntos definen una ***única*** recta que pasa por éstos; tres puntos definen una parábola única y en general ***(n+1)*** puntos pueden definir un polinomio que pase por ***m*** de ese conjunto de puntos. Dicho polinomio sería de grado ***(m-1)*** y tendría la forma

$$p_{m-1}(x) = a_{m-1}x^{m-1} + a_{m-2}x^{m-2} + \cdots + a_2x^2 + a_1x + a_0$$

La forma más directa y visible de obtener dicho polinomio, es decir, de encontrar sus coeficientes, consistiría en avaluar el polinomio en cada uno de los puntos disponibles como datos y resolver el conjunto de ecuaciones resultante.

Para mostrar lo anterior, definamos un polinomio de grado ***n***, en cuyo caso disponemos de los ***n+1*** puntos $(x_0, f(x_0)), (x_1, f(x_1)) \cdots (x_n, f(x_n))$ como

$$p_n(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0$$

Evaluamos el polinomio anterior para cada punto del conjunto de datos mencionado

$$p_n(x_0) = a_nx_0^n + a_{n-1}x_0^{n-1} + \cdots + a_2x_0^2 + a_1x_0 + a_0 = f(x_0)$$

$$p_n(x_1) = a_nx_1^n + a_{n-1}x_1^{n-1} + \cdots + a_2x_1^2 + a_1x_1 + a_0 = f(x_1)$$

.

.

.

$$p_n(x_n) = a_nx_n^n + a_{n-1}x_n^{n-1} + \cdots + a_2x_n^2 + a_1x_n + a_0 = f(x_n)$$

Es evidente que las incógnitas de este sistema de ecuaciones son los coeficientes del polinomio; el resto son valores conocidos.

El sistema matricial asociado con este planteamiento es

$$\begin{bmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \cdots & x_0^2 & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \cdots & x_1^2 & x^1 & 1 \\ & & & \ddots & & & \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \cdots & x_n^2 & x_n & 1 \end{bmatrix} \begin{bmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$$

Dado que los ***n+1*** puntos disponibles en los datos son distintos y determinan además a un polinomio de grado ***n único***, el sistema de ecuaciones resultante tiene una solución *única*.

De hecho el determinante de este sistema de ecuaciones

$$\begin{vmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \cdots & x_0^2 & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \cdots & x_1^2 & x^1 & 1 \\ & & & \ddots & & & \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \cdots & x_n^2 & x_n & 1 \end{vmatrix}$$

Es de un tipo conocido como *determinante de Vandermonde* y se sabe que su valor *no es cero*, por lo que el sistema de ecuaciones lineales es un *sistema no singular*.

Supongamos que disponemos de un muestreo de la función ***f(x) = sen (x)***, en el rango que se muestra en la tabla siguiente

i	x_i	$f(x_i)$
0	π/6	0.5
1	π/4	0.707107
2	π/3	0.866025
3	5π/12	0.965926

Como puede observarse el intervalo de muestreo es de 15^o = π/12 radianes.

Basados en la formulación vista antes, desarrollamos la ecuación lineal correspondiente para obtener el polinomio interpolante correspondiente.

$$\begin{bmatrix} x_0^3 & x_0^2 & x_0 & 1 \\ x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_1^2 & x_1 & 1 \\ x_3^3 & x_3^2 & x_3 & 1 \end{bmatrix} \begin{bmatrix} a_3 \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix}$$

El sistema con los datos del problema, se convierte el sistema de orden 3 mostrado

$$\begin{bmatrix} 0.143540 & 0.274156 & 0.523599 & 1 \\ 0.484473 & 0.616850 & 0785398 & 1 \\ 1.148380 & 1.096620 & 1.04720 & 1 \\ 2.242930 & 1.713470 & 1.30900 & 1 \end{bmatrix} \begin{bmatrix} a_3 \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.707107 \\ 0.866107 \\ 0.965926 \end{bmatrix}$$

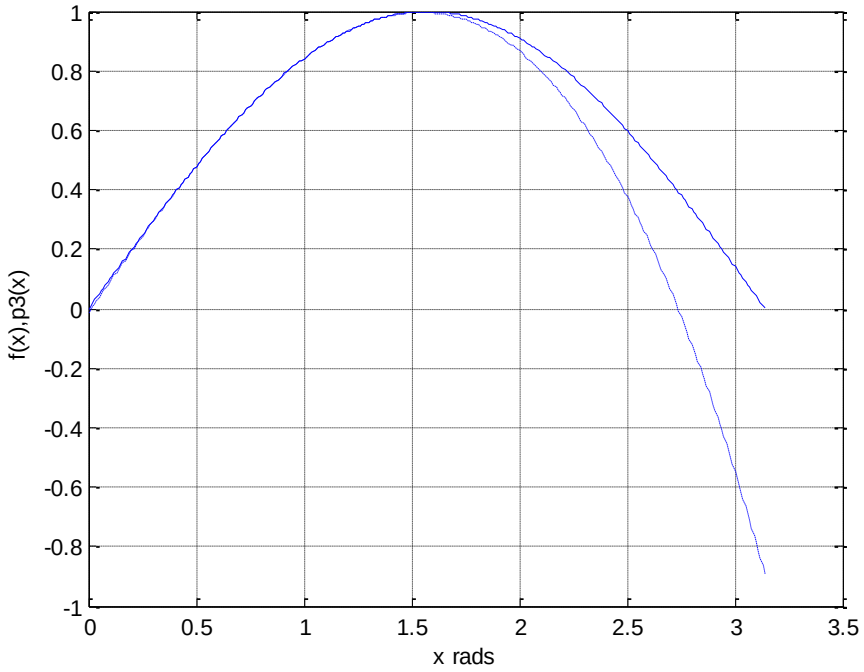
El resultado es

$$\begin{bmatrix} a_3 \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \text{-100.5433e-003} \\ \text{-114.6573e-003} \\ \text{1.0721e+000} \\ \text{-15.4885e-003} \end{bmatrix}$$

Y el polinomio de 3er grado

$$p_3(x) = \text{-100.5433e-003}\,x^3 - \text{114.6573e-003}\,x^2 + 1.0721x - \text{15.4885e-003}$$

La siguiente gráfica muestra la relación entre la función original y el interpolante.



La curva puntada corresponde a $p_3(x)$. Es evidente en la gráfica que a medida que nos alejamos del rango de muestreo el error se incremente. A simple vista, se puede ver que en el rango de 0 a π/2 la aproximación es aceptable; pero después del extremo derecho de dicho rango el error crece notoriamente. Es importante que estas observaciones son cualitativas, por lo que solamente con los valores numéricos correspondiente podríamos hacer observaciones más objetivas.

POLINOMIO INTERPOLANTE DE LAGRANGE.

Es método discutido anteriormente tiene es muy importante en la discusión del tema de interpolación, sin embargo no es computacionalmente adecuado, pues el costo de resolver sistemas de ecuaciones lineales es alto; además se requiere de métodos más eficientes. La primera propuesta en este sentido la constituye el método del polinomio de Lagrange que discutimos a continuación.

Iniciamos con el caso lineal

$$p_1(x) = a_1 x + a_0$$

Que como sabemos

$$p_1(x_0) = f(x_0) \quad y \quad p_1(x_1) = f(x_1)$$

De lo anterior tenemos las siguientes ecuaciones

$$\begin{aligned} p_1(x) - a_1 x - a_0 &= 0 \\ f(x_0) - a_1 x_0 - a_0 &= 0 \\ f(x_1) - a_1 x_1 - a_0 &= 0 \end{aligned}$$

En el curso de Algebra Lineal, se discute que la condición para que un sistema de ecuaciones lineal homogéneo tenga una solución distinta de la trivial, se requiere que su determinan te sea cero

$$\begin{vmatrix} p_1(x) & -x & -1 \\ f(x_0) & -x_0 & -1 \\ f(x_1) & -x_1 & -1 \end{vmatrix} = 0$$

Expandimos el determinante por su primera columna y obtenemos

$$(-1)^2 p_1(x) \begin{vmatrix} -x_0 & -1 \\ -x_1 & -1 \end{vmatrix} + (-1)^3 f(x_0) \begin{vmatrix} -x & -1 \\ -x_1 & -1 \end{vmatrix} + (-1)^4 f(x_1) \begin{vmatrix} -x & -1 \\ -x_0 & -1 \end{vmatrix} = 0$$

De donde obtenemos

$$p_1(x)(x_0 - x_1) - f(x_0)(x - x_1) + f(x_1)(x - x_0) = 0$$

Finalmente

$$p_1(x) = f(x_0)\frac{(x - x_1)}{(x_0 - x_1)} + f(x_1)\frac{(x - x_0)}{(x_1 - x_0)}$$

Esta última ecuación representa el polinomio de Lagrange de primer orden, que en forma compacta definimos

$$p_1(x) = f(x_0)L_0(x) + f(x_1)L_1(x)$$

Donde

$$L_0 = \frac{(x - x_1)}{(x_0 - x_1)} \qquad L_1(x) = \frac{(x - x_0)}{(x_1 - x_0)}$$

En general, para un polinomio de Lagrange de grado ***n***:

$$p_n(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + \cdots + f(x_n)L_n(x)$$

Tomando en cuenta que

$$p_n(x_i) = f(x_i) \qquad \mathrm{i} = 0,1,\cdots,n$$

$$p_n(x_1) = f(x_0)\cdot 0 + f(x_1)\cdot 1 + \cdots + f(x_n)\cdot 0 = f(x_1)$$

Con lo anterior vemos que los coeficientes $L_i(x)$ deben tener una forma tal que para un valor de **x**, digamos x_i, todos los coeficientes deben hacerse cero, menos el coeficiente $L_i(x)$.

La estructura requerida deberá por tanto tener la forma

$$\mathrm{L}_i(x) = \frac{(x - x_0)(x - x_1)(x - x_2)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)}{(x_i - x_0)(x_i - x_1)(x_i - x_2)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}$$

Para cualquier valor $x_k \neq x_i$, $L_i(x) = 0$, mientras que para $x = x_i$, $L_i(x) = 1$.

El numerador de $L_i(x)$ es un polinomio de grado ***n*** en ***x***, mientras que el denominador , que no depende de ***x***, en constante y por tanto $\mathbf{p_n(x)}$ es una suma de polinomios de grado ***n*** y, por tanto, un polinomio en grado ***n*** el mismo.

En forma compacta los coeficientes de Lagrange se definen

$$L_j(x) = \frac{\prod_{\substack{j=0 \\ j\neq i}}^{n} \left(x - x_j\right)}{\prod_{\substack{k=0 \\ k\neq i}}^{n} \left(x_i - x_k\right)}$$

Su pongamos que queremos calcular por medio de un polinomio cuadrático la función seno del ejemplo previo para un valor de x = 0.75 rads. Repetimos la tabla de datos por comodidad

i	x_i	$f(x_i)$
0	π/6	0.5
1	π/4	0.707107
2	π/3	0.866025
3	5π/12	0.965926

El polinomio interpolante y los coeficientes de Lagrange se muestran a continuación

$$p_2(x) = L_0(x)\,f(x_0) + L_1(x)\,f(x_1) + L_2(x)\,f(x_2)$$

$$L_0(0.75) = \frac{(0.75-\pi/4)(0.75-\pi/3)}{(\pi/6-\pi/4)(\pi/6-\pi/3)} = 0.076747 \qquad L_1(0.75) = \frac{(0.75-\pi/6)(0.75-\pi/3)}{(\pi/4-\pi/6)(\pi/4-\pi/3)} = 0.981718$$

$$L_2(0.75) = \frac{(0.75-\pi/6)(0.75-\pi/4)}{(\pi/3-\pi/6)(\pi/3-\pi/4)} = -0.058465$$

De donde obtenemos

$$p_2(0.75) = (0.076747)(0.5) + (0.707107)(0.981718) + (0.866025)(-0.058465)$$

Finalmente $p_2(0.75) = 0.681921$.

DERIVADAS NUMÉRICAS.

Antes de entrar en el siguiente tema, es indispensable dar un vistazo a la versión numérica de las derivadas; sin pretender ser exhaustivo en el tema, vemos la parte esencial que es indispensable para el análisis del siguiente tema, el polinomio de Newton.

Existen varios enfoques en el desarrollo de estos conceptos, el más natural [ChC] es que tratamos de exponer. Partimos de la expansión en series de Taylor de primer orden

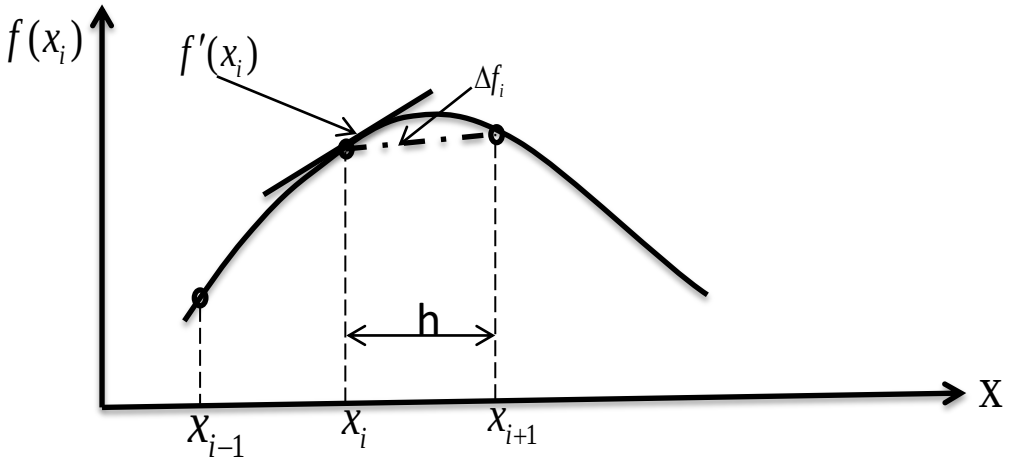
$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \Re_1(x_i)$$

Despejando la primera derivada de la expresión anterior obtenemos

$$f'(x_i) \simeq \frac{\Delta f_i}{h} + O(x_{i+1} - x_i) \qquad donde: \Delta f_i = f(x_{i+1}) - f(x_i)\ ,\ h = x_{i+1} - x_i$$

Al término Δf_i se le denomina ***primera diferencia finita dividida hacia adelante***. El adjetivo hacia adelante, se refiere al orden mostrado en la diferencia y se hará más evidente más adelante cuando veamos otras definiciones de diferencias. El término "dividida" se refiere al denominador, ***h***, de la definición.

La siguiente figura ilustra el concepto de la aproximación de la derivada a través de Δf_i



El intervalo de la definición anterior es $[x_i, x_{i+1}]$, con x_i como punto base. La siguiente definición que vamos a ver, está basada en el intervalo $[x_{i-1}, x_i]$ y parte del desarrollo de la serie de Taylor de primer orden en dicho intervalo y con el mismo punto base, como se muestra

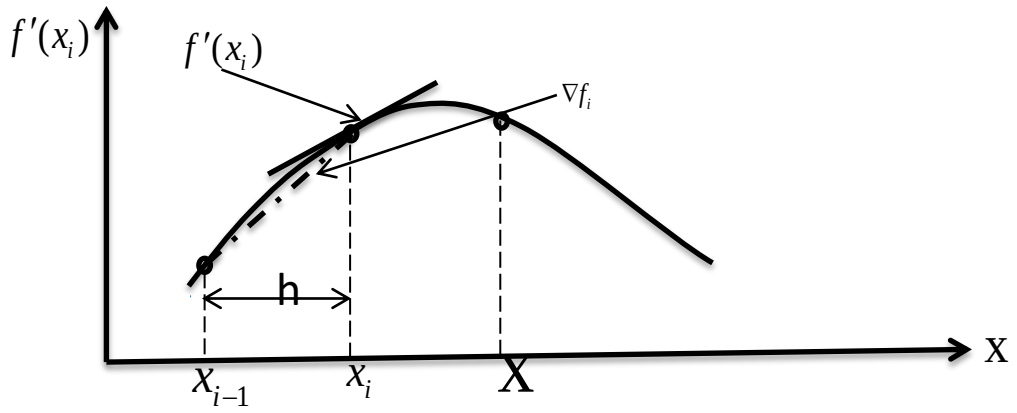
$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \Re_1(x_i)$$

De nueva cuenta, despejamos la primera derivada para obtener

$$f'(x_i) \simeq \frac{f(x_i) - f(x_{i-1})}{h} = \frac{\nabla f_i}{h} \qquad donde: \nabla f_i = f(x_i) - f(x_{i-1})$$

Observe la secuencia de la diferencia en la definición de la primera derivada numérica en este caso. Al término ∇f_i se le conoce como la ***primera diferencia dividida hacia atrás***. Observando el orden de la resta en la división, comparado con la definición previa, se comprenderá sin más explicación el adjetivo "hacia atrás".

La siguiente figura ilustra el concepto de la aproximación de la derivada a través de ∇f_i



Otra opción para definir la aproximación numérica de la primera derivada es a través de las denominadas diferencias centrales.

Para su obtención usamos las expansiones de Taylor que se muestran a continuación

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{f''(x_i)}{2!}h^2 - \cdots$$

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \cdots$$

Restando las dos ecuaciones anteriores obtenemos

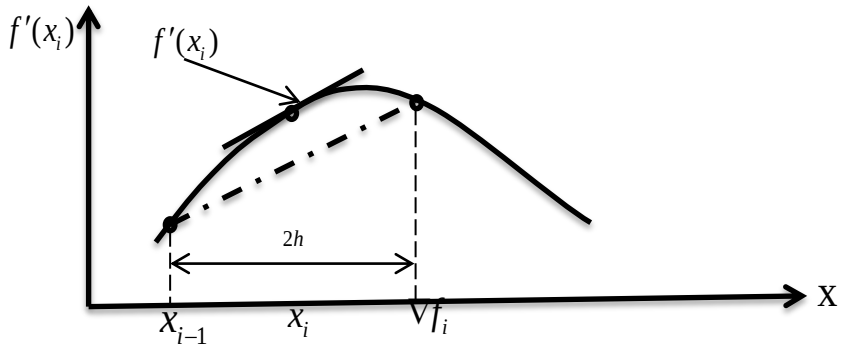
$$f(x_{i-1}) - f(x_{i+1}) = -2hf'(x_i)$$

De donde obtenemos

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_{i-1})}{2h}$$

Al numerador del cociente mostrado se le conoce como diferencia central y al cociente completo se le conoce como ***diferencia central dividida***.

La figura siguiente muestra el concepto geométrico de este resultado



La obtención de derivadas de orden superior aes un poco más complicada, sin embargo con un poco de algebra extra, podemos obtener la segunda derivada como ejemplo y con el fin de ver la tendencia de las diferencias finitas, que son la base del siguiente método de interpolación que se discutirá.

Partimos de dos series de Taylor des arrolladas con el punto base que las anteriores y de hecho la primera se repite por comodidad, pues ya se usó anteriormente

$$f(x_{i+2}) = f(x_i) + f'(x_i)(2h) + \frac{f''(x_i)}{2!}(2h)^2 + \cdots$$

A la serie anterior se le resta 2 veces la serie que se muestra a continuación, es decir se resta la expresión que sigue a continuación

$$2\times\left\{ f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \cdots \right\}$$

El resultado de la resta mencionada es

$$f(x_{i+2}) - 2f(x_{i+1}) = -f(x_i) + f''(x_i)h^2 + \cdots$$

Truncando la serie a partir del término de tercer orden y resolviendo para la segunda derivada, obtenemos la expresión para la segunda derivada

$$f''(x_i) \simeq \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{h^2}$$

Esta expresión puede reescribirse en términos de primeras diferencias como se muestra

$$f''(x_i) \simeq \frac{\frac{f(x_{i+1}) - f(x_i)}{h} - \frac{f(x_i) - f(x_{i-1})}{h}}{h}$$

El resultado mostrado es muy ilustrativo, pues vemos que la segunda derivada se define en términos de un cociente que incluye la ***diferencia de primeras diferencias***, lo cual constituye una ***segunda diferencia***; en este caso se trata de primeras (diferencias de primer orden) diferencias hacia adelante que dan lugar auna diferencia de segundo orden hacia adelante. Es decir, la segunda diferencia finita es la aproximación numérica de la segunda derivada. En general podemos obtener diferencias de orden superior en términos de diferencias de diferencias finitas del orden inmediato inferior, y estas diferencias de orden superior serán las aproximaciones numéricas de las correspondientes derivadas.

POLINOMIO INTERPOLANTE DE NEWTON.

Anteriormente habíamos llegado a la expresión para la interpolación lineal, cuyo resultado repetimos en este punto

$$f_{\text{int}}(\overline{x}) = f(x_i) + \frac{\overline{x} - x_i}{x_{i+1} - x_{i+1}}\left[f(x_{i+1}) - f(x_i)\right]$$

La expresión mostrada representa el polinomio interpolante de primer grado (lineal) de Newton, que presentamos de una forma alternativa como

$$p_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

El formato presentado es más adecuado para el desarrollo general del polinomio de Newton. Algunos autores usan otra nomenclatura [ChC]. Como autor de estas notas utilizaré una nomenclatura adoptada de algunos autores, pero con variantes. No se debe a un eclecticismo caprichoso, sino lo que dicta la experiencia en la enseñanza. Posiblemente a algunos no les parezca buena idea, pero lo hago con la idea de evitar confusión en donde, según mi criterio, puede presentarse.

En la expresión anterior se ha utilizado ***x*** como el valor de la abscisa para la cual se quiere obtener una aproximación por interpolación, en lugar de $\overline{x}$. Se puede observar que el factor del segundo término que consiste de un cociente , representa una ***primera diferencia hacia adelante dividida***, y se basa en la información de dos datos (interpolante de primer orden) $\left(x_0, f(x_0)\right), \left(x_1, f(x_1)\right)$.

Para encontrar el polinomio de Newton de segundo grado, utilizaremos un formato diferente al comúnmente utilizado en la representación de los polinomios; este formato se expresa adecuadamente en forma de productos como se muestra a continuación

$$p_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

Es conveniente diferenciar los coeficientes en este formato para no confundirnos con los coeficientes de la forma común usada anteriormente; por esta razón se usan coeficientes ***b*** en este caso. Es importante recordar que para este caso, disponemos de tres puntos conocidos como dato $\left(x_0, f(x_0)\right), \left(x_1, f(x_1)\right), \left(x_2, f(x_2)\right)$. Nuestro objetivo es determinar los coeficientes del polinomio en función de estos datos.

Evaluamos el polinomio para $x = x_0$ en cuyo caso se cancelan el segundo y tercer términos de la derecha, por lo que obtenemos entonces que $b_0 = f(x_0)$.

Procediendo de manera semejante, para $x = x_1$, $p_2(x)$ resulta

$$p_2(x_1) = b_0 + b_1(x - x_0) = f(x_1)$$

Recordemos que en los nodos (los puntos proporcionados como datos), el interpolante y la función discretizada por medio de los datos **coinciden** exactamente, de ahí la igualdada mostrada.

Sustituyendo el valor previamente obtenido para b_0 obtenemos

$$f(x_0) + b_1(x - x_0) = f(x_1)$$

De donde resulta

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Finalmente evaluamos la expresión del polinomio para $x = x_2$, sustituimos los coeficientes previamente determinados, b_0 y b_1 para obtener el coeficiente b_2

$$p_2(x_2) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_2 - x_0) + b_2(x_2 - x_0)(x_2 - x_1) = f(x_2)$$

De donde

$$b_2(x_2 - x_0)(x_2 - x_1) = f(x_2) - f(x_0) - \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_2 - x_0)$$

Después de un poco de algebra finalmente podemos obtener

$$b_2 = \frac{\dfrac{f(x_2) - f(x_1)}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

Aquí es indispensable hacer varias acotaciones. Seguramente el lector notó que el cociente que define el coeficiente b_1 es una *diferencia finita dividida de primer orden*, que constituye una derivada numérica de primer orden; mientras que el cociente que define b_2 define una *diferencia finita dividida de segundo orden*, que representa la aproximación numérica de la segunda derivada.

Usamos la siguiente nomenclatura para representar dichas diferencias divididas [ChC]

$$f[x_1, x_0] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \qquad f[x_2, x_1, x_0] = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

Procediendo de manera similar, podemos generalizar los resultados anteriores para polinomios de mayor grado y generalizar sus características, como enumeramos a continuación.

Las definiciones de los coeficientes y de las diferencias divididas finitas asociadas con dichos coeficientes polinomiales son

$$b_0 = f(x_0)$$
$$b_1 = f[x_1, x_0]$$
$$b_2 = f[x_2, x_1, x_0]$$
$$\vdots$$
$$b_n = f[x_n, x_{n-1}, \cdots, x_1, x_0]$$
$$f[x_i, x_j] = \frac{f(x_i) - f(x_j)}{x_i - x_j}$$
$$f[x_i, x_j, x_k] = \frac{f[x_i, x_j] - f[x_j, x_k]}{x_i - x_k}$$
$$\vdots$$
$$f[x_n, x_{n-1}, \ldots, x_1, x_0] = \frac{f[x_n, x_{n-1}, \ldots, x_1] - f[x_{n-1}, x_{n-2}, \ldots, x_0]}{x_n - x_0}$$

Lo anterior muestra que una forma simple de obtener las diferencias requeridas es a través de una tabla de diferencias, y con esto de manera ordenada, a partir de cada conjunto de diferencias de cierto orden, nos permiten obtener las diferencia de orden inmediatamente superior; es decir, las diferencias finitas de segundo orden se obtienen como diferencias de las correspondientes diferencias finitas de primer orden; las diferencias de tercer orden se obtiene como diferencias de las diferencia finitas de segundo orden y así sucesivamente, hasta llegar a las diferencias finitas de orden ***n*** que se obtendrán como obtendrá como diferencias de las correspondientes diferencias finitas de orden ***n-1***.

De acuerdo a lo anterior, el polinomio general interpolante de Newton de orden ***n*** será

$$p_n(x) = f(x_0) + (x - x_0) f[x_1, x_0] + (x - x_0)(x - x_1) f[x_2, x_1, x_0] + \cdots + (x - x_0)(x - x_1) \cdots (x - x_{n-1}) f[x_n, x_{n-1}, \cdots, x_0]$$

La tabla siguiente muestra la forma tabular de obtener las diferencias finitas, para el caso consistente en cinco pares de puntos; es simple ver que en general si tenemos ***n*** puntos podemos obtener diferencias finitas hasta un orden ***n-1***, lo cual es consistente con lo que habíamos anticipado en un principio.

0	x_0	$f(x_0)$	$f[x_1, x_0]$	$f[x_2, x_1, x_0]$	$f[x_3, x_2, x_1, x_0]$	$f[x_4, x_3, x_2, x_1, x_0]$
1	x_1	$f(x_1)$	$f[x_2, x_1]$	$f[x_3, x_2, x_1]$	$f[x_4, x_3, x_2, x_1]$	
2	x_2	$f(x_2)$	$f[x_3, x_2]$	$f[x_4, x_3, x_2]$		
3	x_3	$f(x_3)$				
4	x_4	$f(x_4)$				

Si recurrimos al ejemplo de la interpolación de la función seno usado en la sección anterior, cuyos datos repetimos por comodidad

i	x_i	$f(x_i)$
0	π/6	0.5
1	π/4	0.707107
2	π/3	0.866025
3	5π/12	0.965926

La primera diferencia finita de primer orden es

$$f[x_1, x_0] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{0.707107 - 0.5}{\pi/4 - \pi/6} = 0.79107$$

La siguiente sería

$$f[x_2, x_1] = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{0.866025 - 0.707107}{\pi/3 - \pi/4} = 0.607022$$

En términos de estas diferencias finitas de primer orden podemos obtener las diferencias finitas de segundo orden (observe que se omite el término dividida para no hacer tan largo el nombre, pero es evidente que son diferencias divididas)

$$f[x_2, x_1, x_0] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} = \frac{0.607022 - 0.79107}{\pi/3 - \pi/6} = -0.351506$$

El cambio de signo no es casual, es consistente con el hecho de que la segunda diferencia es una aproximación numérica de la segunda derivada, que en el caso de la función seno es negativa.

Los resultados de la tabla de diferencias en notación científica (*short eng* en MATLAB®) resultan como se muestra

i	x_i	$f(x_i)$	$f[x_1, x_0]$	$f[x_2, x_1, x_0]$	$f[x_3, x_2, x_1, x_0]$
0	π/6	0.5	***0.0000e+000***	***0.0000e+000***	***0.0000e+000***
1	π/4	0.707107	***791.0916e-003***	***0.0000e+000***	***0.0000e+000***
2	π/3	0.866025	***607.0160e-003***	***-351.5571e-003***	***0.0000e+000***
3	5π/12	0.965926	***381.5928e-003***	***-430.5239e-003***	***-100.5433e-003***

Como ejemplo, supongamos que queremos encontrar el valor de la función para un

x = 0.75 rads (42.97° aprox) a través de una interpolación cuadrática.

Calculamos

$$p_2(0.75) = f\left(\frac{\pi}{4}\right) + \left(0.75 - \frac{\pi}{4}\right) \cdot f\left[x_1, x_0\right] + \left(0.75 - \frac{\pi}{4}\right)\left(0.75 - \frac{\pi}{3}\right) \cdot f\left[x_2, x_1, x_0\right]$$
$$= 0.707107 + (-0.035398) \cdot (0.7910916) + (-0.035398) \cdot (-0.297198) \cdot (-0.351557)$$
$$= 0.707107 - 0.028003 - 0.003698 = 0.675406$$

El valor verdadero es ***sen(0.75) = 0.681639***, por lo que el error verdadero relativo porcentual será

$$\varepsilon_v = \frac{0.681639 - 0.675406}{0.681639} \times 100 = 0.91\%$$

ERROR DEL POLINOMIO DE NEWTON.

Anteriormente hicimos notar que si observamos el polinomio de Newton de orden ***n***, encontramos una similitud muy evidente con la serie de Taylor; de hecho, corriendo el riesgo de que algún matemático se escandalice, podemos decir que el polinomio interpolante de Newton es la versión discreta de la serie de Taylor. Vimos que las diferencias finitas de orden ***n*** en general, corresponden a las diferencias finitas del orden correspondiente. Tomando esto en cuenta, podemos utilizar el concepto del residuo de la serie de Taylor con el fin de escribir la versión finita, que corresponderá al residuo del polinomio de Newton.

La versión finita de un polinomio de Newton de orden ***n*** será entonces

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n) \qquad \xi \in \left[x, x_i\right]$$

En este caso es importante definir los elementos del rango de pertenencia de ξ. La ***x*** representa la incógnita (el valor para el cual se quiere interpolar) y $\mathbf{x}_i$ representa el conjunto de datos.

Existe otra limitante y es que se supone que en general no disponemos del valor de la derivada de orden ***n*** de la expresión anterior, por lo que debemos adecuadarla al entorno matemático en que trabajamos, es decir, al hecho de que trabajamos con datos finitos.

Dicho lo anterior sustituimos el operador de dicha derivada con la correspondiente diferencia finita, con lo que la expresión final será

$$R_n \cong f[x_{n+1}, x_n, x_{n-1}, \ldots, x_0](x-x_0)(x-x_1)\cdots(x-x_n)$$

Observamos que si queremos evaluar el error del método en una interpolación, debemos contar con al menos la diferencia finita de orden superior al grado del polinomio usado.

Para el caso del ejemplo, el valor del residuo de $p_2(0.75)$ es

$$\begin{aligned} R_2(0.75) &\cong f[x_3, x_2, x_1, x_0](0.75-x_0)(0.75-x_1)(0.75-x_2)(00.75-x_3) \\ &= (-0.1005433)\left(0.75-\frac{\pi}{6}\right)\left(0.75-\frac{\pi}{4}\right)\left(0.75-\frac{\pi}{3}\right)\left(0.75-\frac{5\pi}{12}\right) \\ &= (-0.1005433)(0.226401)(-0.035398)(-0.297198)(-0.558997) = 0.000134 \end{aligned}$$

INTEGRACIÓN

NUMÉRICA

INTRODUCCIÓN.

La ***integración numérica***, llamada también CUADRATURA, es el cálculo aproximado de las integrales del tipo $\int_a^b f(x)dx$.

Estos métodos se requieren en dos casos generales: cuando no tenemos la función del integrando en forma explícita y en su lugar tenemos una colección de pares ordenados que representan a dicha función; el otro caso es que, aun cuando se tenga la expresión funcional del integrando, sea difícil integrarlo por los métodos que se discuten en los cursos de cálculo o que simplemente la primitiva de la función del integrando no se puede obtener. Ejemplos de lo anterior serían integrales de la forma $\int_0^1 e^{-x^2} dx$ o bien $\int_0^{\frac{\pi}{2}} \sqrt{1+\cos^2 x}\ dx$.

En la realidad, es decir, en las aplicaciones a las ciencias y la ingeniería, existen muchos ejemplos del segundo caso; se podría decir que en una mayoría de los casos, en estas áreas las integrales resultantes no tienen primitiva o es demasiado costoso el cálculo analítico. Ejemplos hay muchos, simplemente para mencionar uno muy antiguo en la ingeniería eléctrica, mencionamos el caso del modelado de la tierra como conductor en los problemas de transmisión. Este problema se ha estudiado desde el siglo XIX y de manera muy intensa en el siglo pasado. La solución obtenida parte de desarrollar series infinitas para aproximar el integrando; dichas series son más factibles de evaluar y proporcionan una solución aproximada. Existen dos formulaciones que datan de la segunda mitad de 1920's; una conduce a las famosas series de Carson y la otra a las series de Pollaczek. En la actualidad se sigue trabajando en métodos computacionales para resolver dicho problema de forma más eficiente. Los ejemplos podrían seguir en todas las áreas de las ciencias y la ingeniería.

En términos generales construimos métodos numéricos de integración, aproximando la función integrando por funciones fáciles de integrar: polinomios.

Hay dos formas de obtener una buena aproximación:

- Aproximar ***f*** por un único polinomio interpolante de alto grado;
- Aproximar ***f*** por diferentes polinomios interpolante de bajo grado en pequeños intervalos de integración.

Existen dos grupos de métodos de integración numérica:

- FÓRMULAS DE INTEGRACIÓN DE NEWTON-COTES
- CUADRATURA GAUSSIANA.

FÓRMULAS DE INTEGRACIÓN DE NEWTON-COTES.

La idea esencial en estos métodos consiste en integrar un polinomio como sustituto de la función original

$$I = \int_a^b f(x)dx \cong \int_a^b p_n(x)dx$$

Con

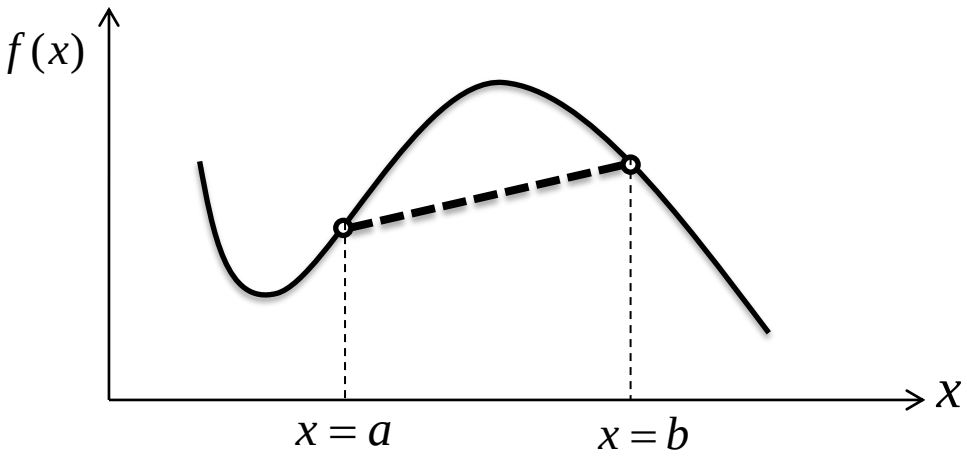
$$p_n(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + a_n x^n$$

$p_n(x)$ es un polinomio interpolante que aproxima a la función del integrando $f(x)$.

Dependiendo del grado del polinomio, será el método concreto referido. Si usamos un polinomio de grado uno $p_1(x)$, entonces estamos usando una recta que pasa por los puntos definidos por los límites de integración $(a, f(a))$ y $(\mathrm{b}, f(\mathrm{b}))$

$$I = \int_a^b f(x)dx \cong \int_a^b p_1(x)dx$$

Cuya representación gráfica se muestra enseguida.



En este caso, el área comprendida entre las rectas *x = a* y *x = b*, así como el eje de las *x* y la recta punteada en la figura, será la aproximación del área bajo la curva *f(x)*. Es obvio que estamos aproximando dicha área por el área de un trapecio, como se puede observar en la figura. Lo anterior nos conduce a obtener de manera geométrica la primera fórmula de integración numérica del grupo de métodos de Newton-Cotes, que se denomina por obvias razones REGLA TRAPECIAL.

Definimos el rango de integración en *el eje-x*, como paso de integración $h = b - a$.

Dado lo anterior y haciendo uso de las matemáticas elementales, recordamos que la fórmula del área del trapecio es $Area = \left(base\,menor + base\,mayor\right) \times \frac{altura}{2}$

Usando la nomenclatura apropiada para este caso la fórmula anterior se convierte en la fórmula de integración del trapecio

$$I \simeq (b-a)\frac{f(a)+f(b)}{2} .$$

El desarrollo anterior está basado en consideraciones geométricas, lo cual por supuesto es válido; sin embargo con el fin de obtener el error del método, es necesario usar un desarrollo más analítico.

DERIVACIÓN DE LA FÓRMULA TRAPECIAL. La obtención de la fórmula trapecial de forma analítica inicia usando la definición que mencionamos al principio de esta sección, es decir, a partir de la definición $I = \int_a^b f(x)dx \cong \int_a^b p_n(x)dx$

Con el interpolante de primer orden $p_1(x)$ dado por

$$p_1(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

Si desarrollamos la expresión anterior tendremos

$$p_1(x) = \frac{f(b) - f(a)}{b - a}x + f(a) - \frac{a\,f(b) - a\,f(a)}{b - a}$$

Si usamos el concepto antes mencionado: $I = \int_a^b f(x)dx \cong \int_a^b p_1(x)dx$ obtenemos

$$I = \frac{f(b) - f(a)}{b - a}\frac{x^2}{2} + \frac{b\,f(a) - a\,f(b)}{b - a}x\Bigg]_a^b$$

Y una vez evaluados los límites

$$I = \frac{f(b) - f(a)}{b - a}\frac{\left(b^2 - a^2\right)}{2} + \frac{b\,f(a) - a\,f(b)}{b - a}(b - a)\Bigg]$$

Tomando en cuenta $(b^2 - a^2) = (b - a)(b + a)$, finalmente obtenemos

$$I = \left[f(b) - f(a)\right]\frac{b + a}{2} + b\,f(a) - a\,f(b)$$

Finalmente factorizando obtenemos el resultado buscado

$$I = (b - a)\frac{f(a) + f(b)}{2}.$$

El cálculo el error del método, está obviamente relacionado con el residuo de $p_1(x)$.

A continuación se muestra la forma de obtenerlo [ChC]. Se podría utilizar cualquier forma del interpolante, sin embargo la más ilustrativa en mi opinión requiere de escribir en polinomio interpolante en un formato especial denominado polinomio de *Newton-Gregory* y que se aplica para el caso especial en que el paso de integración es constante.

Los puntos correspondientes a los datos discretos se pueden escribir como se muestra a continuación, dada la condición del paso de integración constante

$$x_1 = x_0 + h$$
$$x_2 = x_1 + h = x_0 + 2h$$
.
.
$$x_n = x_0 + nh$$

Definiendo $\alpha = \frac{x - x_0}{h}$, podemos escribir las expresiones anteriores como

$$x - x_0 = \alpha h$$
$$x - x_0 - h = \alpha h - h = h(\alpha - 1)$$
.
.
$$x - x_0 - (n-1)h = \alpha h - (n-1)h = h(\alpha - n + 1)$$

Por lo que usando estas expresiones, podemos escribir el polinomio como

$$p_n(x) = f(x_o) + \frac{\Delta f(x_0)}{h}(x - x_0) + \frac{\Delta^2 f(x_0)}{2!h^2}(x - x_0)(x - x_0 - h) + \cdots$$
$$\cdots + \frac{\Delta^n f(x_0)}{n!h^n}(x - x_0)(x - x_0 - h)\cdots\left[x - x_0 - (n-1)h\right] + \Re_n$$

Escribiendo la fórmula anterior en términos de α

$$p_n(x) = f(x_0) + \Delta f(x_0)\alpha + \frac{\Delta^2 f(x_0)}{2!}\alpha(\alpha-1) + \cdots$$

$$\cdots + \frac{\Delta^n f(x_0)}{n!}\alpha(\alpha-1)\cdots(\alpha-n+1) + \Re_n$$

Con

$$\Re_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}\alpha(\alpha-1)\cdots(\alpha-n)$$

Esta última fórmula es la que se utiliza en la obtención analítica de las reglas de integración numérica de los métodos de Newton-Cotes.

Para nuestro caso, la regla trapecial, integramos nuevamente el polinomio interpolante correspondiente

El interpolante de Newton-Gregory de primer está dado por

$$f(a) + \Delta f(a)\alpha + \frac{f''(\xi)}{2}\alpha(\alpha-1)h^2$$

Por lo que

$$I \simeq \int_a^b \left[f(a) + \Delta f(a)\alpha + \frac{f''(\xi)}{2}\alpha(\alpha-1)h^2 \right] dx$$

Tomando en cuenta que $\alpha = \frac{(x-a)}{h}$ y $dx = h\,d\alpha$, podemos encontrar los nuevos límites de integración, en los que ***a*** y ***b*** corresponden a ***0*** y ***1***, respectivamente por lo que

$$I \simeq h\int_0^1 \left[f(a) + \Delta f(a)\alpha + \frac{f''(\xi)}{2}\alpha(\alpha-1)h^2 \right] d\alpha$$

$$I \simeq h\left[\alpha f(a) + \frac{\alpha^2}{2}\Delta f(a) + \left(\frac{\alpha^3}{6} - \frac{\alpha^2}{4}\right) f''(\xi)h^2 \right]_0^1$$

Cuyo resultado es

$$I \simeq h\left[f(a) + \frac{\Delta f(a)}{2} \right] - \frac{1}{12} f''(\xi) h^3$$

Recordamos que $\Delta f(a) = f(b) - f(a)$, por lo que finalmente

$$I \simeq h \frac{f(a) + f(b)}{2} - \frac{1}{12} f''(\xi) h^3$$

Identificamos el primer término de la derecha como la fórmula de la regla trapecial, por lo que el segundo corresponde al *error del método* o *error de truncamiento*.

Dado que lo que sabemos es que $\xi \in [a,b]$, no conocemos el valor exacto del residuo, por lo que lo que podemos concluir es que el error es proporcional al cubo del paso de integración.

A la fórmula de la regla trapecial que se discutió hasta ahora, la podemos llamar aplicación simple de la fórmula de integración. Es decir, dado que el resultado del análisis del error nos dice evidentemente, que mientras más pequeño es el paso de integración ***h***, pues el error se reduce en forma proporcional al cubo, utilizar la el rango de integración como paso de integración no es adecuado. La solución es dividir el rango de integración en segmentos más pequeños y aplicar la regla trapecial a cada uno de estos segmentos, por lo que la suma de las áreas de los trapecios individuales, se aproximará con menor error al valor de la integral buscada. A lo anterior se le conoce como ***aplicación múltiple de la regla trapecial***.

Definimos el número de segmentos, ***n***, que vamos a emplear para la aplicación múltiple y determinamos el tamaño del paso correspondiente; definimos

$$h = \frac{b-a}{n} \qquad a = x_0 \quad b = x_n$$

Por lo que $I = \int_a^b f(x)\,dx$ se transforma en

$$I = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)dx$$

De donde aplicando la regla trapecial a cada integral obtenemos

$$I \simeq h\frac{f(x_0)+f(x_1)}{2}+h\frac{f(x_1)+f(x_2)}{2}+\cdots+h\frac{f(x_{n-1})+f(x_n)}{2}$$

Que una vez factorizada podemos expresar

$$I \simeq \frac{h}{2}\left[\, f(x_0)+2\sum_{i=1}^{n-1} f(x_i)+f(x_n)\right]$$

Algunos autores [ChC] prefieren escribir la fórmula compuesta en función de los límites originales, por lo que definen

$$I \simeq (b-a)\,\frac{f(x_0)+2\sum_{i=1}^{n-1} f(x_i)+f(x_n)}{2n}$$

En función de la definición anterior, la expresión del error se escribe como sigue

$$E_t = -\frac{(b-a)^3}{12n^3}\sum_{i=1}^{n} f''(\xi_i)$$

Es importante observar que $\xi_i \in [x_i, x_{i+1}]$, es decir, como era de esperarse el error depende de la segunda derivada de la función en cada uno de los intervalos de las integrales del método compuesto, por lo que para hacer útil la expresión, se define un valor de la derivada de para el rango completo, como promedio de la segunda derivada en cada segmento, es decir

$$\overline{f}'' \cong \frac{\sum_{i=1}^{n} f''(\xi_i)}{n}$$

Por tanto $\sum f''(\xi_i) \cong n\,\overline{f}''$, de donde obtenemos finalmente

$$E_a = -\frac{(b-a)^3}{12n^2}\,\overline{f}''$$

El ejemplo siguiente ayudará a comprender los detalles de lo que se ha discutido.

Queremos calcular numéricamente $\int_0^1 \frac{1}{1+x}dx = \ln(2)$ cuyo valor analítico se puede determinar fácilmente y nos sirve para ver e efecto del paso de integración de la regla trapecial. Vamos (a) integrar con un solo intervalo, ***h = 1***; (b) con ***h = 0.25***; (c) con *h = 0.125*.

(a) $I \simeq \frac{1-0}{2}\left(f(1)-f(0)\right) = \frac{1}{2}\left(\frac{1}{2}+1\right) = 0.750000$

Dado que, sin considerar el error de truncamiento en la evaluación de ésta función, $\ln(2) = 0.693147$; esto constituye lo que usamos como valor verdadero

De acuerdo con esto, el error verdadero será

$$\varepsilon_v = \left|0.693147 - 0.75\right| = 0.056853$$

(b) Con ***h = 0.25***, calculamos el número de segmentos $n = \frac{1-0}{0.25} = 4$, por lo que calculamos la tabla de valores requerida

i	x_i	$f(x_i)$
0	0	1.0
1	0.25	0.80
2	0.50	0.666667
3	0.75	0.571429
4	1.0	0.5

Con lo anterior tenemos

$$I \simeq \frac{0.25}{2}\left[1.0 + (2)\left(0.8 + 0.666667 + 0.571429\right) + 0.5\right] = \left(\frac{0.25}{2}\right)\left(5.576192\right) = 0.697024$$

Y con esto el error verdadero

$$\varepsilon_v = \left|0.693147 - 0.697024\right| = 0.026784$$

(c) con *h = 0.125*, el número de segmentos es $n = \frac{1-0}{0.125} = 8$, por lo que

i	x_i	$f(x_i)$
0	0.0	1.0
1	0.125	0.888889
2	0.250	0.80
3	0.375	0.727273
4	0.50	0.666667
5	0.625	0.615385
6	0.75	0.571429
7	0.875	0.533333
8	1.0	0.50

La integral aproximada de la regla trapecial resulta

$$I \simeq \frac{0.125}{2}\left[1.0 + (2)\left(0.8 + 0.88889 + 0.8 + 0.727273 + 0.666667 + 0.615385 + 0.571429 + 0.533333\right) + 0.5\right] =$$

$$= \left(\frac{0.125}{2}\right)\left(1.0 + 9.605952 + 0.5\right) = 0.694122$$

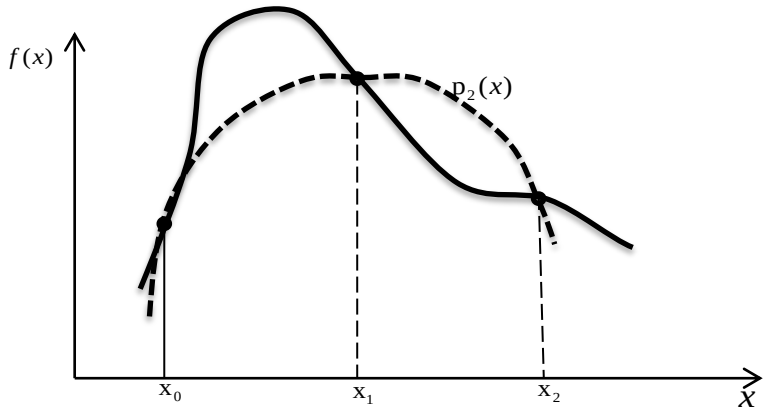
Con lo que $\varepsilon_v = \left|0.693147 - 0.694122\right| = 0.000975$.

REGLA DE SIMPSON.

Seguramente a estas alturas se habrá observado que el principio usado para obtener la regla trapecial de integración numérica, se debe poder extender al uso de polinomios de mayor grado que el usado esta regla. En efecto, siguiendo la misma idea uno puede imaginar que no habiendo problema para generar más puntos dentro del intervalo de integración, se podrían proponer polinomios de orden superior a uno y obtener reglas de integración numéricas alternativas. Esto es exactamente lo que conduce a las denominadas reglas de Simpson, que son el siguiente tema en estas notas.

Tomando como base puntos equidistantes, podemos obtener un punto intermedio en el rango de integración y con ello integrar un polinomio de segundo grado.

Gráficamente la idea se muestra a continuación, referida por supuesto referida a la aplicación de la regla en forma simple, en contrapartida de la aplicación múltiple de la regla, que se comentará más adelante.



Al igual que antes, utilizamos el formato del polinomio de Newton-Gregory, en este caso de segundo grado, el cual integramos en los límites integración $x = a = x_0 \quad y \quad x = b = x_2$

$$I \simeq \int_{x_0}^{x_2} \left[f(x_0) + \Delta f(x_0)\alpha + \frac{\Delta^2 f(x_0)}{2}\alpha(\alpha-1) + \frac{\Delta^3 f(x_0)}{6}\alpha(\alpha-1)(\alpha-2) + \frac{f^{(4)}(\xi)}{24}\alpha(\alpha-1)(\alpha-2)(\alpha-3)h^4 \right] dx$$

EL polinomio utilizado es de tercer grado, aparentemente contradiciendo la idea expuesta; sin embargo en el desarrollo, que se muestra enseguida, se verá la razón de esto. Debemos configurar la integral en función del parámetro α, para lo cual debemos obtener su diferencial y los límites correspondientes; los límites de x_0, x_1 corresponden a $\alpha = 0, 2$, respectivamente. De esta manera la integral finalmente toma la forma

$$I = h\int_0^2 \left[f(x_0) + \Delta f(x_0)\alpha + \frac{\Delta^2 f(x_0)}{2}\alpha(\alpha-1) + \frac{\Delta^3 f(x_0)}{6}\alpha(\alpha-1)(\alpha-2) + \frac{f^{(4)}(\xi)}{24}\alpha(\alpha-1)(\alpha-2)(\alpha-3)h^4 \right] d\alpha$$

Para evitar confusiones, es pertinente recordar que como se definió anteriormente $\alpha = \frac{x - x_0}{h}$, tenemos que $dx = h \cdot d\alpha$.

Efectuando la integración mostrada obtenemos

$$I \simeq h\left[\alpha f(x_0) + \frac{\alpha^2}{2}\Delta f(x_0) + \left(\frac{\alpha^3}{6} - \frac{\alpha^2}{4}\right)\Delta^2 f(x_0) + \left(\frac{\alpha^4}{24} - \frac{\alpha^3}{6} + \frac{\alpha^2}{6}\right)\Delta^3 f(x_0) + \left(\frac{\alpha^5}{120} - \frac{\alpha^4}{16} + \frac{11\alpha^3}{72} - \frac{\alpha^2}{8}\right) f^{(4)}(\xi) h^4 \right]_0^2$$

Evaluada la expresión anterior en los límites obtenemos

$$I = h\left[2\, f(x_0) + 2\Delta f(x_0) + \frac{\Delta^2 f(x_0)}{3} + (0)\Delta^3 f(x_0) - \frac{1}{90} f^{(4)}(\xi) h^4 \right]$$

Observe que el coeficiente de la diferencia finita de tercer orden es cero y esto está relacionado con la observación hecha antes.

Escribiendo las diferencia finitas en términos de las ordenadas de los datos, recordemos que $\Delta f(x_0) = f(x_1) - f(x_0)$ y $\Delta^2 f(x_0) = f(x_2) - 2f(x_1) + f(x_0)$, por lo que sustituidas en la última ecuación obtenemos finalmente

$$I = \frac{h}{3}\left[f(x_0) + 4\, f(x_1) + f(x_2) \right] - \frac{1}{90} f^{(4)}(\xi) h^5 .$$

Esta es la ***fórmula de Simpson 1/3***, denominada así para distinguirla de la regla se Simpson que se mencionará más adelante. El segundo término de la derecha es evidente que constituye el término del residuo y por tanto relacionado con el error de truncamiento del método de Simpson 1/3.

APLICACIÓN MÚLTIPLE DE LA REGLA DE SIMPSON 1/3.

Al igual que comentamos la conveniencia de aplicar la fórmula trapecial varias veces, con el fin de reducir el paso de integración y con esto, de acuerdo a la fórmula del error, reducir el error en el cálculo numérico de la integral, lo más adecuado también en el presente caso, es aplicar de forma múltiple la fórmula de Simpson 1/3.

En general, esta regla de integración supone mejor precisión que la regla trapecial; sin embargo la aplicación múltiple está condicionada a discretizar el rango de integración en un número par de segmentos, asociado por tanto a un número impar de puntos, de otra forma no se podría llevar a cabo dicha operación.

En el ejemplo siguiente ilustramos, tanto la aplicación simple de la fórmula, como su aplicación múltiple de la misma, para el mismo problema.

Un ejemplo de la aplicación de este método ayuda. Resolvemos el mismo problema del caso de la regla trapecial, es decir,

$$\int_0^1 \frac{1}{1+x}\,dx = \ln(2)$$

Primero hacemos uso de la regla Simpson 1/3 simple, por lo que se requiere generan un punto extra ***h = 0.5***

i	x_i	$f(x_i)$
0	0	1.0
1	0.5	0.666667
2	1.0	0.5

La aplicación de la regla resulta

$$I_{S_{1/3}} = (1/4)(1/3)\left[1+(4)\cdot(0.666667)+0.5\right] =$$
$$= (1/6)\cdot(4.166667) = 0.694445$$

Lo cual resulta en un error absoluto verdadero de

$$\varepsilon_v = \left|0.93147 - 0.694445\right| = 0.001298$$

Aplicamos ahora la regla múltiple a través de dividir en dos intervalos de integración con tres puntos en cada intervalo y dos segmentos por intervalo, lo que significan ***n = 4*** segmentos y por tanto *5 puntos de muestreo* , por lo cual ***h =1/4=0.25***. Esto se muestra a continuación

i	x_i	$f(x_i)$
0	0	1.0
1	0.25	0.80
2	0.50	0.666667
3	0.75	0.571429
4	1.0	0.5

De acuerdo con el planteamiento, la fórmula múltiple es

$$I = \frac{h}{3}\left[f(x_0) + 4\cdot\left(f(x_1) + f(x_3)\right) + 2\cdot f(x_2) + f(x_4)\right]$$

Observe que el punto $(x_2, f(x_2))$ es común a los dos intervalos de integración, razón por la cual aparece $f(x_2)$ con un factor 2 ntepuesto en la anterio expresión.

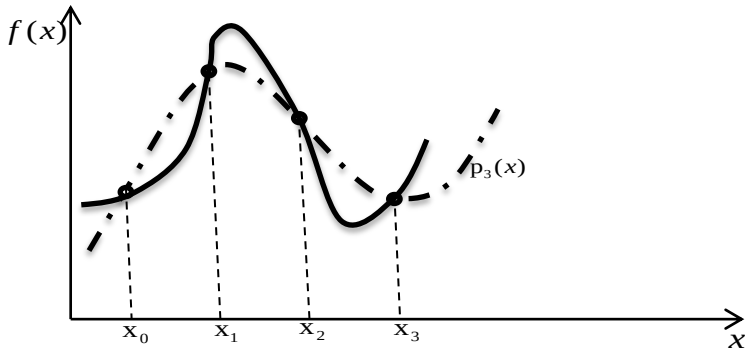
Con esto

$$\begin{aligned} I &= \frac{h}{3}\left[f(x_0) + 4\cdot\left(f(x_1) + f(x_3)\right) + 2\cdot f(x_2) + f(x_4)\right] \\ &= (1/4)(1/3)\left[1 + (4)\cdot(0.8 + 0.571429) + (2)\cdot(0.666667) + 0.5\right] \\ &= (1/12)(8.31905) = 0.693254 \end{aligned}$$

Con lo cual

$$\varepsilon_v = \left|0.93147 - 0.693254\right| = 0.000107$$

La obtención de la segunda fórmula de Simpson, denominada Simpson 3/8, es más involucrada en cuanto a su desarrollo, por lo que no lo haremos. Se anima al lector a intentarlo, siguiendo para ello el mismo procedimiento. La gráfica siguiente es útil para entender los conceptos concernientes al método.



En este caso se utiliza un polinomio de 3er grado ***p3(x)*** como interpolante, por lo que se requieren ***4 puntos*** y ***3 segmentos*** en este caso y la fórmula del método resulta

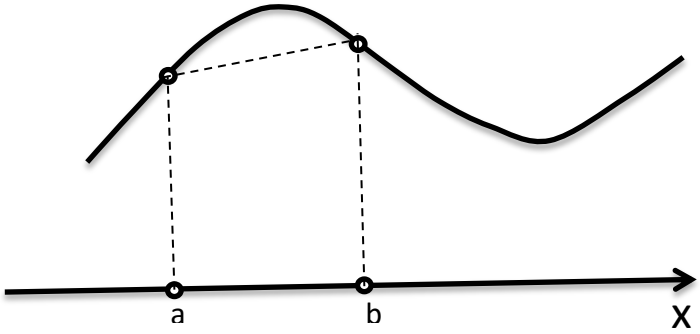
$$I \cong \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right]$$

$$h = \frac{(b-a)}{3} \qquad\qquad E_t = -\frac{(b-a)5}{6480} f^{(4)}(\xi)$$

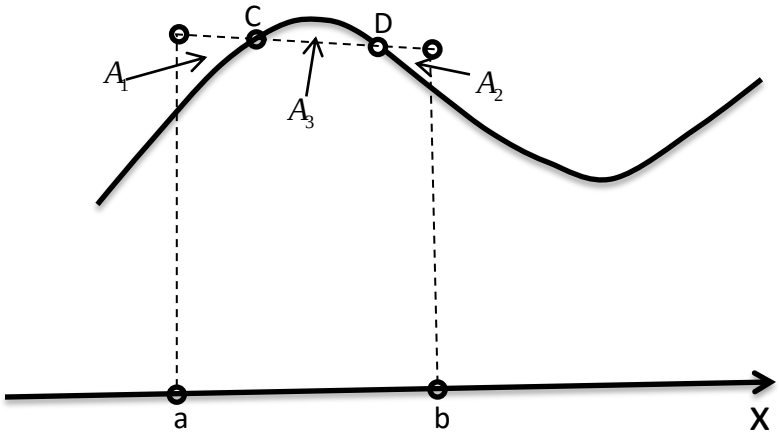
CUADRATURA DE GAUSS.

Una de las características de los métodos de Newton-Cotes es que utilizan segmentos del mismo ancho o equi-esapaciados, como los llaman algunos autores. La pregunta que se planteó Gauss fue si esta característica limita la exactitud. Su apreciación es que efectivamente lo anterior ocurría, por lo que planteo una forma alternativa de resolver la cuadratura. Primero trataremos de explicar este planteamiento de una manera informal, a través de un par de gráficas, y después recurriremos a la formulación a través del método de coeficientes indeterminados.

La siguiente figura muestra la interpretación gráfica de la regla trapecial, en la cual se usan los puntos que definen el rango de integración de la función del integrando, para aproximar por medio del trapecio correspondiente, el área correspondiente a la integral, que es el área bajo la curva y delimitada por las rectas ***x = a*** y ***x = b***.



Una alternativa se muestra en la siguiente figura donde Gauss plantea la idea de no limitar a la recta que une los puntos correspondientes al rango de integración, sino buscar una recta (interpolante lineal) cuya posición reduzca el error al mínimo



Se puede observar que el error representado por las tres áreas que no contienen exactamente el área bajo la curva, tienen signos diferentes, es decir, las pequeñas áreas de las esquinas superiores, A_1 y A_2, toman porciones de área extra, que no corresponden al área de la integral, en cuyo caso tiene un error de signo diferente a la pequeña área comprendida por la curva entre los puntos C y D y el segmento punteado que se muestra, es decir el área A_3. Es decir, gráficamente vemos que él es error correspondiente

$$\varepsilon = A_1 + A_2 - A_3$$

El planteamiento en conclusión sería minimizar el error mencionado antes, determinando la recta cuya posición produzca el resultado deseado. Esto conlleva a tener que determinar las abscisas de los puntos C y D sobre la curva.

La formulación comienza con recordar que en los métodos anteriores, y en general en los métodos de integración numérica, la fórmula correspondiente es

$$I \simeq w_1 \cdot f(x_1) + w_2 \cdot f(x_2) + \cdots w_n \cdot f(x_n)$$

Para el caso trapecial la fórmula es, correspondiente a una interpolación lineal asociada,

$$I_{TRAP} \simeq w_1 \cdot f(x_1) + w_2 \cdot f(x_2)$$

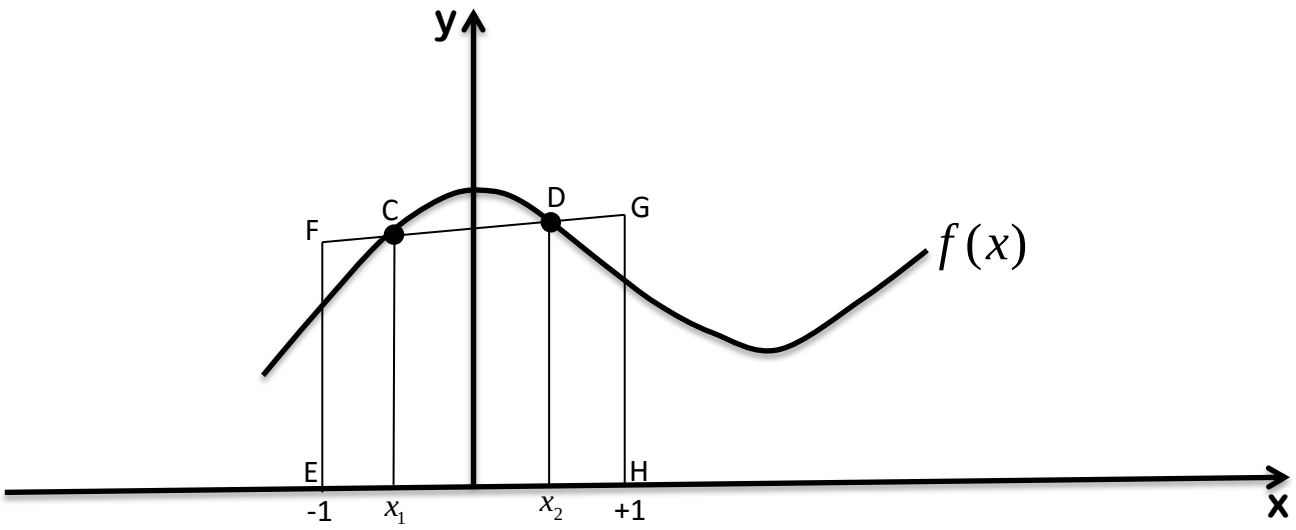
Mientras que para la fórmula se Simpson 1/3

$$I_{S\,IMP1/3} \simeq w_1 \cdot f(x_1) + w_2 \cdot f(x_2) + w_3 \cdot f(x_3)$$

A los coeficientes ***w*** se les conoce como *pesos* o *coeficientes de ponderación*.

Vamos a desarrollar el caso más simple (de menor orden) de la cuadratura de Gauss, que corresponde al ***método de dos puntos***. (En el apéndice correspondiente se muestra, para quién se interese en profundizar, el desarrollo más general).

Para concluir la formulación, la siguiente figura nos ayuda precisar lo discutido arriba. Es muy común en matemáticas, formular ecuaciones usando un rango base alrededor del origen del sistema cartesiano, lo cual implica que se usarán en la formulación como límites de integración ***x = +1*** y ***x = -1***. Esto se muestra en la figura. El cambio de límites se hará posteriormente, como se discute más adelante.



En el desarrollo de la obtención de la fórmula de dos puntos de la cuadratura de Gauss, utilizaremos el método de coeficientes indeterminados, que vamos a ejemplificar con la obtención de la fórmula de Simpson 1/3 como ejemplo.

El método de coeficientes indeterminados es una alternativa para derivar las fórmulas de Newton-Cotes que antes derivamos. En éstas notas se optó por un método que está más a tono con el concepto expuesto en dichos métodos, sin embargo en la derivación de la cuadratura gaussiana se utilizará el de coeficientes indeterminados, razón por la cual se ilustra el uso de este método para desarrollar la fórmula del método de Simpson 1/3 .

Partimos de la forma general de las reglas de integración numéricas que se mencionó anteriormente; para el caso del Simpson 1/3, se requieren tres puntos y por tanto la fórmula general a resolver es

$$I \simeq w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3)$$

Los puntos igualmente espaciados son

$$x_1 = a \quad x_2 = (a+b)/2 \quad x_3 = b \quad \in [a,b]$$

Existen tres curvas que se interpolan exactamente con tres puntos: una función constante ***f(x)=1*** por ejemplo; otra recta ***f(x)=x***; y una parábola ***f(x)=x^2***. A estas funciones se les denomina *funciones monomiales* y serán las mismas que usaremos en la derivación de la cuadratura Gaussiana de tres puntos.

A partir de estos datos y la ecuación expuesta arriba, podemos obtener 3 ecuaciones para resolver las 3 incógnitas del caso, que para el Simpson 1/3 son los coeficientes de ponderación. Dichas ecuaciones son

$$w_1 \cdot 1 + w_2 \cdot 1 + w_3 \cdot 1 = \int_a^b 1\,dx = x\Big|_a^b = b - a$$

$$w_1 \cdot a + w_2 \cdot (a+b)/2 + w_3 \cdot b = \int_a^b x\,dx = \left(x^2/2\right)\Big|_a^b = \left(b^2 - a^2\right)/2$$

$$w_1 \cdot a^2 + w_2 \cdot \left((a+b)/2\right)^2 + w_3 \cdot b^2 = \int_a^b x^2\,dx = \left(x^3/3\right)\Big|_a^b = \left(b^3 - a^3\right)/3$$

En forma matricial

$$\begin{bmatrix} 1 & 1 & 1 \\ a & (a+b)/2 & b \\ a^2 & \left((a+b)/2\right)^2 & b^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} b-a \\ \left(b^2 - a^2\right)/2 \\ \left(b^3 - a^3\right)/3 \end{bmatrix}$$

Resolviendo este sistema de ecuaciones (sistema Vandermonde), encontramos que los coeficientes de ponderación o pesos son

$$w_1 = \left(b-a\right)/6 \qquad w_2 = 2(b-a)/3 \qquad w_3 = (b-a)/6$$

En los cuales se reconoce la regla de Simpson.

Regresando a nuestro problema, desarrollar la fórmula de la cuadratura gaussiana de 3 puntos, recordemos que tenemos 4 incógnitas, pues además de los coeficientes de ponderación, también desconocemos los puntos x_1 y x_2.

Dado lo mencionado, requerimos 4 ecuaciones, por lo que las funciones monomiales son

$$y=1 \quad y=x \quad y=x^2 \quad y=x^3$$

Y el conjunto de ecuaciones

$$w_1 \cdot 1 + w_2 \cdot 1 = \int_{-1}^{+1} 1\,dx = x\Big|_{-1}^{+1} = 2$$

$$w_1 \cdot x_1 + w_2 \cdot x_2 = \int_{-1}^{+1} x\,dx = \left(x^2/2\right)\Big|_{-1}^{+1} = 0$$

$$w_1 \cdot x_1^2 + w_2 \cdot x_2^2 = \int_{-1}^{+1} x^2\,dx = \left(x^3/3\right)\Big|_{-1}^{+1} = \frac{2}{3}$$

$$w_1 \cdot x_1^3 + w_2 \cdot x_2^3 = \int_{-1}^{+1} x^3\,dx = \left(x^4/4\right)\Big|_{-1}^{+1} = 0$$

De la primera ecuación vemos que

$$w_1 + w_2 = 2$$

Además si observamos

$$w_1 = w_2 \quad \text{y} \quad x_1 = -x_2$$

Satisfacen la 3ª y 4ª ecuaciones. De aquí podemos escoger

$$w_1 = w_2 = 1 \quad \text{y} \quad x_1 = -x_2$$

Y sustituyendo en la 3ª ecuación

$$w_1 \cdot x_1^2 + w_2 \cdot x_2^2 = \frac{2}{3}$$

$$x_1^2 + (-x_1)^2 = \frac{2}{3}$$

$$2 \cdot x_1^2 = \frac{2}{3}$$

Por lo que finalmente

$$x_1 = \pm \frac{1}{\sqrt{3}} = \pm 0.57735.....$$

$$\int_{-1}^{+1} f(x)\,dx = w_1 \cdot f(x_1) + w_2 \cdot f(x_2) = f\left(\frac{+1}{\sqrt{3}}\right) + f\left(\frac{-1}{\sqrt{3}}\right)$$

La última ecuación se denomina ***fórmula de Gauss-Legendre de dos puntos***.

Es interesante mencionar que los valores de x_1 y x_2 corresponden a las raíces del polinomio de Legendre de grado 2, de ahí el nombre del método. En el apéndice correspondiente a este tema de la cuadratura Gaussiana, se desarrolla la teoría más general y se desarrollan las fórmulas para cualquier grado, así como una discusión resumida de los polinomios de Legendre.

En este punto, lo que falta es resolver la cuestión relativa al cambio de los límites de integración, ya que en el desarrollo de la fórmula usaron los límites ***-1*** y ***+1***, por lo que hay que hacer el cambio de límites correspondiente.

La derivación de la fórmula del error en este caso es más complicada que en los casos de las fórmulas de Newton-Cotes. Hay literatura disponible por supuesto para quién tiene interés en dicho procedimiento [IK],[RR].

Sin embargo, se demuestra que el error en los métodos de Gauss-Legendre, está dado por [ChC]

$$E_t = \frac{2^{2n+3}\left[(n+1)!\right]^4}{(2n+3)\left[(2n+2)!\right]^3} f^{(2n+2)}(\xi)$$

De lo anterior podemos concluir que siempre que las derivadas de alto orden no se incrementen con el correspondiente incremento del número de puntos (orden del método), los métodos de Gauss-Legendre son superiores a los métodos de Newton-Cotes. En [ChC] se propone un problema al final del capítulo muy ilustrativo sobre este particular.

Antes de dejar el tema, debemos tratar acerca de la adaptación de los límites reales de la aplicación de este método en un problema particular a los límites usados en la derivación del mismo, que es en los que están basados los resultados obtenidos arriba.

Se usaron los límites ***-1*** y ***+1*** en dicho desarrollo y suponemos que los del problema concreto los límites de la integral son ***x=a*** (límite inferior) y ***x=b*** (límite superior).

La relación entre dichos límites es lineal, por lo que hacemos un cambio de variables como se muestra.

Sea

$$x = a_0 + a_1 t$$

En este caso ***t*** es la variable correspondiente a los límites de la derivación del método, es decir *límite inferior = -1* , *límite superior = +1*.

De lo anterior vemos que

$$\text{si } x = a \quad \text{corresponde a} \quad t = -1 \Rightarrow a = a_0 + a_1 \cdot (-1)$$

$$\text{si } x = b \quad \text{corresponde a} \quad t = +1 \Rightarrow a = a_0 + a_1 \cdot (+1)$$

Esto conduce a un par de ecuaciones lineales simultáneas

$$a_0 - a_1 = a$$
$$a_0 + a_1 = b$$

Matricialmente

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a_a \\ a_1 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$$

Cuya solución es

$$a_0 = \frac{a+b}{2} \qquad a_1 = \frac{b-a}{2}$$

Y nos conduce finalmente a

$$x = \frac{(a+b)+(b-a)t}{2}$$
$$dx = \frac{b-a}{2}\,dt$$

Un ejemplo de este método podemos aplicarlo a la siguiente integral, cuya solución analítica es fácil de obtener

$$\int_0^1 \left(6 - 6x^5\right) dx = 5$$

En este caso, con ***a=0*** y ***b=1*** obtenemos

$$x = \frac{1+t}{2} \qquad \text{y} \qquad dt = \frac{1}{2}\,dt$$

La función modificada será

$$f(t) = \frac{1}{2}\left(6 - 6 \cdot \left(\frac{1+t}{2}\right)^5\right)$$

Y evaluamos

$$f\left(1/\sqrt{3}\right)=2.998736 \qquad f\left(-1/\sqrt{3}\right)=2.084598$$

De donde obtenemos

$$I \simeq f\left(1/\sqrt{3}\right)+f\left(-1/\sqrt{3}\right)=5.083333$$

EL problema que trabajamos en los métodos de Newton_Cotes, aplicando la cuadratura se resolvería como se muestra enseguida. Recordemos la integral que se quiere resolver

$$\int_0^1 \frac{1}{1+x}\,dx$$

Si efectuamos el cambio de variable, tendríamos

$$x=\frac{(1+0)+(1-0)t}{2}=\frac{1+t}{2} \qquad dx=\frac{(1-0)}{2}\,dt=\frac{1}{2}\,dt$$

Dado lo anterior la integral correspondiente es

$$\int_{-1}^{+1} \frac{1}{3+t}\,dt$$

Por lo que la nueva función del integrando es

$$f(t)=\frac{1}{3+t} \qquad f\left(\frac{1}{\sqrt{3}}\right)=0.279537 \qquad f\left(\frac{-1}{\sqrt{3}}\right)=0.412771$$

Finalmente

$$I \simeq f\left(\frac{1}{\sqrt{3}}\right)+f\left(\frac{-1}{\sqrt{3}}\right)=0.279537+0.412771=0.692308$$

$$\varepsilon_v = \left|0.693147 - 0.692308\right| = 0.000839$$

El método presentado aquí es el de más bajo orden entre los métodos de Gauss-Legendre. En cualquier libro se pueden encontrar tablas que contienen los coeficientes de ponderación y los puntos correspondientes, para utilizarse en métodos de mayor orden, y por tanto más precisos [ChC], [MTH], [GW], [BT], [CLW], por citar algunos.

Por último, en [ChC] aparece un problema muy interesante; en el capítulo 22, problema 8 de la 4ª edición, se propone resolver un problema típico donde este tipo de métodos tiene problema para converger a un valor adecuado. La razón es que si grafican la segunda derivada, asociado con el método particular de dos puntos, verán que su valor se llega a ser muy grande. Esto hace que en este caso, el método proporcione resultados muy pobres. El número del problema puede cambiar en cada edición de dicha referencia, por lo que se proporciona la integral a resolver

$$\int_{-3}^{+3} \frac{2}{1+2x^2}\,dx$$

SOLUCIÓN NUMERICA

DE

SISTEMAS

DE

ECUACIONES LINEALES

INTRODUCCIÓN.

Un sistema de ecuaciones lineales de orden *n*

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
&a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
&\cdot \\
&\cdot \\
&\cdot \\
&a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n
\end{aligned}
$$

En el cual suponemos que todos los coeficientes son reales y que en forma compacta se puede expresar

$$A\underset{\sim}{x} = \underset{\sim}{b}$$

Donde

$$
A = \begin{pmatrix}
a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\
a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\
\cdot & & & & & \\
\cdot & & & & & \\
\cdot & & & & & \\
a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn}
\end{pmatrix}
$$

$$\underset{\sim}{b} = \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{pmatrix}$$

En ocasiones es útil describir *A* como *n* vectores columna:

$$A = \begin{bmatrix} a_{\bullet 1} & a_{\bullet 2} & \cdot & \cdot & \cdot & a_{\bullet n} \end{bmatrix}$$

Donde definimos

$$a_{\bullet j} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ . \\ . \\ . \\ a_{nj} \end{bmatrix}$$

Por lo que podemos escribir el sistema de ecuaciones de la forma:

$$x_1 a_{\bullet 1} + x_2 a_{\bullet 2} + \cdots + x_n a_{\bullet n} = \underset{\sim}{b} .$$

- De acuerdo a lo anterior el problema de resolver el sistema $A\underset{\sim}{x} = \underset{\sim}{b}$ es equivalente a representar el vector $\underset{\sim}{b}$, como una combinación lineal de los vectores $a_{\bullet 1}, a_{\bullet 2}, \cdots, a_{\bullet n}$.
- El sistema tiene solución para cada $\underset{\sim}{b}$, si y solo si, los vectores columna constituyen una ***base*** en $\mathbb{R}^n$, o sea, si los vectores columna son linealmente independientes.
- Una matriz *A* cuyos vectores columna son linealmente independientes se denomina ***no singular***.
- Si *A* es no singular existe una solución única para cada vector $\underset{\sim}{b}$.

El método básico para resolver el sistema de ecuaciones $A\underset{\sim}{x} = \underset{\sim}{b}$ es la eliminación gaussiana. En este método se suman múltiplos de las ecuaciones de manera sistemática, para convertir el sistema de ecuaciones en un sistema triangular:

$$\left[A \vdots \underset{\sim}{b}\right] \Rightarrow \left[U \vdots \underset{\sim}{c}\right]$$

donde *U* es una matriz triangular superior.

La solución de $U\underset{\sim}{x} = \underset{\sim}{c}$ se obtiene por sustitución regresiva o hacia atrás.

Una matriz $U = \left(u_{ij}\right)$ de orden $n \times n$ se denomina *triangular superior* si $u_{ij} = 0$ para todo $i \rangle j$, mientras que una matriz $L = \left(l_{ij}\right)$ se denomina *triangular inferior* si $l_{ij} = 0$ para todo $i \langle j$.

Suponiendo que todos los elementos diagonales de la matriz *U* son distintos de cero, en cuyo caso U es *no singular*, el sistema

$$\begin{bmatrix} u_{11} & u_{12} & \cdot & \cdot & \cdot & u_{1n} \\ & u_{22} & \cdot & \cdot & \cdot & u_{2n} \\ & & & & & \cdot \\ & & & & & \cdot \\ & & & & & \cdot \\ & & & & & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{bmatrix}$$

puede resolverse por *sustitución regresiva*:

$$x_n = c_n / u_{nn}$$

$$x_i = \frac{\left(c_i - \sum_{j(i+1)}^{n} u_{ij} x_j \right)}{u_{ii}} \qquad i = n-1, n-2, \cdots, 1.$$

Describiremos ahora el proceso de eliminación gaussiana para un sistema de *n* ecuaciones lineales

$$\left[A : \underset{\sim}{b} \right] = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} & \vdots & b_1 \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & \cdot & \vdots & b_2 \\ \cdot & & & & & & \vdots & \cdot \\ \cdot & & & & & & \vdots & \cdot \\ \cdot & & & & & & \vdots & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn} & \vdots & b_n \end{bmatrix}$$

En el primer paso de eliminación hacemos *cero* los elementos de la primera columna, excluyendo el elemento diagonal a_{11}; suponiendo que $a_{11} \neq 0$, restamos múltiplos adecuados del primer renglón a cada renglón, del segundo en adelante, con lo cual obtenemos

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} & \vdots & b_1 \\ 0 & a'_{22} & \cdot & \cdot & \cdot & a'_{2n} & \vdots & b'_2 \\ \cdot & & & & & & \vdots & \cdot \\ \cdot & & & & & & \vdots & \cdot \\ \cdot & & & & & & \vdots & \cdot \\ 0 & a'_{n2} & \cdot & \cdot & \cdot & a'_{nn} & \vdots & b'_n \end{bmatrix}$$

El apóstrofe significa que los elementos correspondientes, han sido modificados en el proceso. Los detalles del proceso de eliminación se detallan más adelante.

En el segundo paso del proceso de eliminación, eliminamos (hacemos cero) los elementos de la segunda columna, a partir del siguiente elemento debajo de a'_{22}, el cual suponemos es distinto de cero, obteniendo como resultado

$$\left[\begin{array}{ccccccc|c} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} & \vdots & b_1 \\ 0 & a'_{22} & \cdot & \cdot & \cdot & a'_{2n} & \vdots & b'_2 \\ 0 & 0 & \cdot & \cdot & \cdot & a'_{3n} & \vdots & b'_3 \\ \cdot & & \cdot & & & & \vdots & \cdot \\ \cdot & & & \cdot & & & \vdots & \cdot \\ \cdot & & & & \cdot & & \vdots & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & a'_{nn} & \vdots & b'_n \end{array}\right]$$

Después de *(k-1)* pasos del proceso de eliminación la matriz se ha transformado como se muestra

$$\left[\begin{array}{ccccccccccc|c} a_{11} & a_{12} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & a_{1n} & \vdots & b_1 \\ & a_{22} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & a_{2n} & \vdots & b_2 \\ & & \cdot & & & & & & & & & \vdots & \cdot \\ & & & \cdot & & & & & & & & \vdots & \cdot \\ & & & & \cdot & & & & & & & \vdots & \cdot \\ & & & & & a_{kk} & a_{k,k+1} & \cdot & \cdot & \cdot & a_{kn} & \vdots & b_k \\ & & & & & \cdot & \cdot & & & & & \vdots & \cdot \\ & & & & & \cdot & \cdot & & & & & \vdots & \cdot \\ & & & & & \cdot & \cdot & & & & & \vdots & \cdot \\ & & & & & a_{ik} & a_{i,k+1} & \cdot & \cdot & \cdot & a_{in} & \vdots & b_i \\ & & & & & \cdot & \cdot & & & & & \vdots & \cdot \\ & & & & & \cdot & \cdot & & & & & \vdots & \cdot \\ & & & & & \cdot & \cdot & & & & & \vdots & \cdot \\ & & & & & a_{nk} & a_{n,k+1} & \cdot & \cdot & \cdot & a_{nn} & \vdots & b_n \end{array}\right]$$

El siguiente paso de eliminación, *(k),* nos conduce a la matriz mostrada, suponiendo de nuevo que $a_{kk} \neq 0$

$$\left[\begin{array}{ccccccccccc|c}
a_{11} & a_{12} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & a_{1n} & b_1 \\
 & a_{22} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & a_{2n} & b_2 \\
 & & \cdot & & & & & & & & & \cdot \\
 & & & \cdot & & & & & & & & \cdot \\
 & & & & \cdot & & & & & & & \cdot \\
 & & & & & a_{kk} & a_{k,k+1} & \cdot & \cdot & \cdot & a_{kn} & b_k \\
 & & & & & \cdot & \cdot & & & & \cdot & \cdot \\
 & & & & & \cdot & \cdot & & & & \cdot & \cdot \\
 & & & & & \cdot & \cdot & & & & \cdot & \cdot \\
 & & & & & 0 & a'_{i,k+1} & \cdot & \cdot & \cdot & a'_{in} & b'_i \\
 & & & & & \cdot & \cdot & & & & \cdot & \cdot \\
 & & & & & \cdot & \cdot & & & & \cdot & \cdot \\
 & & & & & \cdot & \cdot & & & & \cdot & \cdot \\
 & & & & & 0 & a'_{n,k+1} & \cdot & \cdot & \cdot & a'_{nn} & b'_n
\end{array}\right]$$

Los elementos transformados son obtenidos a través de las siguientes expresiones recursivas:

$$a'_{ij} = a_{ij} - m_{ik} a_{kj} \qquad j = k+1, \ldots, n \quad i = k+1, \ldots, n$$

$$b'_i = b_i - m_{ik} b_k \qquad i = k+1, \ldots, n \qquad \text{con} \quad m_{ik} = \frac{a_{ik}}{a_{kk}} \qquad i = k+1, \ldots, n\,.$$

El renglón *k-ésimo* que es usado para hacer cero los elementos debajo de él, en la misma columna *k-ésima*, se denomina *renglón pivote*; mientras que al elemento a_{kk} se le llama *elemento pivote*.

Si aplicamos las ecuaciones recursivas al renglón *i-ésimo*, obtenemos para la columna *k*:

$a'_{ik} = a_{ik} - m_{ik} a_{kk} = 0$, dado que $m_{ik} = \frac{a_{ik}}{a_{kk}}$.

El seudo-código relacionado con este caso

for *i = k+1* **to** *n* **do** $m_{ik} = \frac{a_{ik}}{a_{kk}}$

for all (i,j), $k+1 \le i, j \le n$ **do**

$a_{ij} = a_{ij} - m_{ik} * a_{kj}$;

for *i = k+1* **to** *n* **do** $b_i = b_i - m_{ik} * b_k$ **;**

El seudo-código mostrado está formado por instrucciones parecidas a las que ocurren en todos los superlenguajes más comúnmente usados.

Si repetimos esta operación recursiva para *k = 1,…,n-1* el seudo-código final será como se muestra a continuación

for *k = 1* **to** *n-1* **do**

for *i = k+1* **to** *n* **do** $m_{ik} = \frac{a_{ik}}{a_{kk}}$

for all (i,j), $k+1 \le i, j \le n$ **do**

$a_{ij} = a_{ij} - m_{ik} * a_{kj}$;

for *i = k+1* **to** *n* **do** $b_i = b_i - m_{ik} * b_k$ **;**

PIVOTEO.

Iniciamos el tema mostrando los efectos de valores muy pequeños en la diagonal, usando para este propósito un sistema de ecuaciones de *2x2*.

Consideremos un sistema de ecuaciones como el que se muestra

$$\varepsilon x + By = C$$
$$Dx + Ey = F$$

en el cual ε es un número muy pequeño.

La segunda ecuación después de efectuar la eliminación gaussiana resulta

$$\left(E - \frac{DB}{\varepsilon}\right) y = \left(F - \frac{DC}{\varepsilon}\right).$$

Resolviendo para *y* tenemos

$$y = \frac{F - \dfrac{DC}{\varepsilon}}{E - \dfrac{DB}{\varepsilon}} = \frac{F - \dfrac{C}{\varepsilon}}{E - \dfrac{B}{\varepsilon}} \approx \frac{C}{B}$$

De donde obtenemos

$$x \approx \frac{C - B\dfrac{C}{B}}{\varepsilon} \approx \frac{C - C}{\varepsilon} = 0\,.$$

Lo anterior implica que *x = 0* para cualquier valor de *C* y *F*.

El problema mostrado se resuelve usando *pivoteo*, como se muestra a continuación.

Consideremos el mismo sistema usado arriba, pero con las ecuaciones intercambiadas

$$Dx + Ey = F$$
$$\varepsilon x + By = C$$

Supongamos, sin perder generalidad, que

$$F = D + E$$
$$C = B + \varepsilon$$

Por lo que entonces tendremos

$$Dx + Ey = \left(D + E\right)$$
$$\varepsilon x + By = \left(B + \varepsilon\right)$$

Es obvio que la solución verdadera es *x = 1, y = 1*, como podemos probar por sustitución.

Por eliminación gaussiana sin embargo obtenemos

$$y = \frac{C - \left(\frac{F\varepsilon}{D}\right)}{B - \left(\frac{\varepsilon E}{D}\right)} \approx \frac{C}{B} = \frac{B + \varepsilon}{B} = 1 + \frac{\varepsilon}{B} = 1$$

Lo anterior se obtiene tomando en cuenta que ε es número muy pequeño.

Si sustituimos en la primera ecuación completamos la solución, como se muestra

$$x = \frac{\left(D + E\right) - Ey}{D} = \frac{\left(D + E\right) - E \bullet (1)}{D} = 1.$$

Al existir valores muy pequeños en la diagonal, los factores m_{ik} son muy grandes. Lo anterior conduce a que la información de los coeficientes originales se pierde en le producto cruzado $a_{ij}^{'} = a_{ij} - m_{ik}a_{kj}$.

El procedimiento apropiado para resolver este problema consiste en intercambiar renglones que contengan el valor más grande en la columna, correspondiente a la columna cuyos elementos se harán cero, con el renglón pivote. A este procedimiento se le conoce como *pivoteo*.

Para entender el procedimiento mencionado, suponemos que el *k-ésimo* paso de la eliminación está en curso

$$\begin{bmatrix}
a_{11} & a_{12} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & a_{1n} \\
 & a_{22} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & a_{2n} \\
 & & \cdot & & & & & & & & \\
 & & & \cdot & & & & & & & \\
 & & & & \cdot & & & & & & \\
 & & & & & a_{kk} & a_{k,k+1} & \cdot & \cdot & \cdot & a_{kn} \\
 & & & & & \cdot & & & & & \\
 & & & & & \cdot & & & & & \\
 & & & & & \cdot & & & & & \\
 & & & & & a_{ik} & a_{i,k+1} & \cdot & \cdot & \cdot & a_{in} \\
 & & & & & \cdot & & & & & \\
 & & & & & \cdot & & & & & \\
 & & & & & \cdot & & & & & \\
 & & & & & a_{nk} & a_{n,k+1} & \cdot & \cdot & \cdot & a_{nn}
\end{bmatrix}$$

A continuación se lleva a cabo una búsqueda del elemento de mayor valor absoluto en la columna *k*, a partir del renglón *k-ésimo* en adelante. En otras palabras buscamos el índice de renglón *v* tal que

$$\left|a_{vk}\right| = \max_{k \le i \le n} \left|a_{ik}\right|.$$

Efectuada la búsqueda, se intercambiarán los renglones *k* y *v*, procediendo a continuación a llevar a cabo la eliminación de manera ordinaria. Este procedimiento se conoce como *pivoteo parcial*.

El término parcial se refiere al hecho de que el *pivoteo total* , la búsqueda se extiende no nada más a la columna *k-ésima*, sino a la submatriz completa

$$\begin{bmatrix} a_{kk} & \cdot & \cdot & \cdot & a_{kn} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ a_{nk} & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}$$

y una vez encontrado, se intercambian tanto columnas como renglones. Este procedimiento ya casi no se utiliza.

Para modificar el seudo-código de la eliminación, agregamos las líneas de código mostradas abajo, al seudo-código de la eliminación mencionado anteriormente.

Definimos una función asociada con el índice del renglón al cual pertenece el elemento de mayor valor absoluto: *v=indmax(a,k,n)*, de tal manera que $|a_{vk}| = \max_{k \le i \le n} |a_{ik}|$. Enseguida efectuamos el intercambio de renglones, en cuyo caso usaremos la instrucción: *swap(a,b,k,v,n)*, que indicará el intercambio de los renglones *k* y *v*, incluyendo los elementos correspondientes al vector ***b*** por supuesto.

SEUDO-CODIGO DE ELIMINACION GAUSSIANA CON PIVOTEO PARCIAL

for *k =1* **to** *n-1* **do**

v=indmax(a,k,n);

swap(a,b,k,v,n);

for *k = 1* **to** *n-1* **do**

for *i = k+1* **to** *n* **do** $m_{ik} = \frac{a_{ik}}{a_{kk}}$

for all (i,j), $k+1 \le i, j \le n$ **do**

$$a_{ij} = a_{ij} - m_{ik} * a_{kj} ;$$

for *i = k+1* **to** *n* **do** $b_i = b_i - m_{ik} * b_k$ **;**

Recordemos que el objetivo del procedimiento de pívoteo, es evitar elementos matriciales muy grandes durante la eliminación, y con esto la pérdida de precisión que ocurre al efectuar operaciones con cantidades de gran magnitud.

Es importante notar que en el procedimiento de pivoteo parcial, todos los multiplicadores m_{ik} son menores a 1 en magnitud, es decir, $|m_{ik}| = \left|\frac{a_{ik}}{a_{kk}}\right| \le 1$.

En todos los casos de solución de *Ax = b* se debe efectuar el procedimiento de pivoteo parcial con el objetivo de obtener una precisión adecuada.

Existen dos casos especiales en los cuales el procedimiento de pivoteo parcial no es necesario. Estos dos casos corresponden a casos en los que la matriz *A* es:

- Simétrica y positiva definida, y
- Diagonalmente dominante.

Una matriz *A* es *positiva definida* si $\underset{\sim}{x}^T A \underset{\sim}{x} \rangle 0 \quad \forall \underset{\sim}{x} \neq 0$

Una matriz *A* es *diagonalmente dominante* si $\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \leq |a_{ii}|, \qquad i = 1, 2, \ldots, n$

con al menos un *i* que cumpla la desigualdad estrictamente.

Como se estableció anteriormente, un sistema de ecuaciones lineales tiene una solución única si el conjunto de vectores columna de *A* forma una base en $\mathbb{R}^n$. La pregunta en este punto es si la misma condición se cumple en le caso de la eliminación gaussiana con pivoteo parcial.

Empezamos estableciendo que le algoritmo falla si en el *k-ésimo* paso no podemos encontrar un elemento pivote *no cero*, esto es, si $a_{ik} = 0 \quad i = k, k+1, \ldots, n$.

Lo anterior ocurre solamente si la columna *k* es una combinación lineal de las columnas *1,2,…,k-1*.

Consideremos las primeras *k* columnas de la matriz

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1,k-1} & a_{1k} \\ & a_{22} & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & \cdot & & & \cdot & \cdot \\ & & & \cdot & & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot \\ & & & & & a_{k-1,k-1} & a_{k-1,k} \\ & & & & & 0 & 0 \\ & & & & & \cdot & \cdot \\ & & & & & \cdot & \cdot \\ & & & & & \cdot & \cdot \\ & & & & & 0 & 0 \end{bmatrix}$$

La determinación de la columna *k-ésima* como una combinación lineal de las otras *(k-1)* columnas, equivale a resolver el sistema triangular

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1,k-1} \\ & a_{22} & \cdot & \cdot & \cdot & a_{2,k-1} \\ & & \cdot & & & \cdot \\ & & & \cdot & & \cdot \\ & & & & \cdot & \cdot \\ & & & & & a_{k-1,k-1} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_{k-1} \end{bmatrix} = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \cdot \\ \cdot \\ \cdot \\ a_{k-1,k} \end{bmatrix}$$

Podemos ver que la solución de este sistema existe, debido a que los elementos diagonales de la matriz, $a_{ii} \quad i = 1, 2, \ldots, k-1$, son *no cero*.

Concluimos que bajo la suposición de que las columnas de *A* son linealmente independientes, es decir, que *A* es *no singular*, la eliminación gaussiana con pivoteo parcial nos conduce a una solución correcta.

MATRICES DE PERMUTACION Y ELIMINACION GAUSSIANA.

Una propiedad importante que conduce a los métodos de solución por factorización triangular de matrices, es que la solución gaussiana es equivalente a la factorización de la matriz.

Para mostrar lo anterior recurrimos a matrices de transformación elementales.

El pivoteo por renglón es equivalente a la multiplicación por la izquierda (premultiplicación) por una matriz de permutación. La matriz simple de permutación, P_{ij}, se construye intercambiando los renglones *i* y *j* en la matriz identidad. Por ejemplo, para orden 4, P_{13} será

$$P_{13} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

La matriz anterior se denomina *matriz de permutación simple ó matriz de transposición*.

Suponga que multiplicamos un vector ***x*** de orden *n* por P_{ij}. El resultado será un vector cuyos renglones *i* y *j* se intercambian

$$P_{ij}\underset{\sim}{x} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ x_{i-1} \\ x_j \\ x_{i+1} \\ \cdot \\ \cdot \\ \cdot \\ x_{j-1} \\ x_i \\ x_{j+1} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

De manera similar, si multiplicamos por la izquierda la matriz *A*, *nxn*, por P_{ij}, obtendremos

$$P_{ij}A = \begin{bmatrix} & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ a_{j1} & a_{j2} & \cdot\ \cdot\ \cdot & & a_{jn} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ a_{i1} & a_{i2} & \cdot\ \cdot\ \cdot & & a_{in} \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \end{bmatrix}$$

Un análisis sencillo nos demuestra que $P_{ij}^{-1} = P_{ij}$.

Lo anterior es fácil de ver pues $P_{ij}^{-1} * P = I$ significa que si permuto dos veces los mismos renglones, el resultado será la matriz de partida, es decir, la identidad; o bien, para obtener la identidad debo efectuar la misma permutación dos veces y por tanto $P_{ij}^{-1} = P_{ij}$.

Una propiedad importante establece que un producto de matrices de permutación es otra matriz de permutación

$$P_{i1j1}P_{i2j2}\ldots P_{ikjk} = P.$$

Si se quiere efectuar permutación de columnas, el mecanismo es el mismo, solo que el producto no es por la izquierda, sino por la derecha usando P^T, es decir, AP^T.

La transformación gaussiana puede definirse como una matriz triangular inferior L_j, obtenida a partir de la matriz identidad y diferenciándose de ésta solamente en los elementos debajo de la diagonal, como se muestra

$$\begin{bmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & \ddots & & & & \\ & & & & 1 & & & \\ & & & & m_{j+1,j} & 1 & & \\ & & & & m_{j+2,j} & & 1 & \\ & & & & \vdots & & & \ddots \\ & & & & m_{n,j} & & & & 1 \end{bmatrix}$$

Lo anterior equivale a sumar elementos $m_{\nu j}x_j$ a los elementos x_ν, $\nu = j+1, j+2, \ldots, n$.

A través de un razonamiento simple, podemos ver que

$$L_j^{-1} = \begin{bmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & \ddots & & & & & & \\ & & & \ddots & & & & & \\ & & & & 1 & & & & \\ & & & & -m_{j+1,j} & 1 & & & \\ & & & & -m_{j+2,j} & & 1 & & \\ & & & & \vdots & & & \ddots & \\ & & & & \vdots & & & & \ddots \\ & & & & -m_{j+n,j} & & & & 1 \end{bmatrix}$$

Un resultado importante a tener en cuenta es que si $k \rangle j$, entonces

$$L_j L_k = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & \ddots & & & & & & \\ & & & \ddots & & & & & \\ & & & & 1 & & & & \\ & & & & m_{j+1,j} & 1 & & & \\ & & & & m_{j+2,j} & & \ddots & & \\ & & & & \vdots & & & \ddots & \\ & & & & \vdots & & & & \ddots \\ & & & & \vdots & & & & 1 \\ & & & & \vdots & & & m_{k+1,k} & 1 \\ & & & & & & & \vdots & \ddots \\ & & & & & & & \vdots & \ddots \\ & & & & & & & \vdots & \ddots \\ & & & & m_{nj} & & & m_{nk} & 1 \end{bmatrix}$$

Lo cual muestra que el producto de matrices triangular inferior en secuencia, genera una matriz triangular inferior, cuyas características se muestran.

Por otro lado, dado un vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

en el cual $x_1 \neq 0$, y sea

$$L_1 = \begin{bmatrix} 1 & & & & & \\ m_{21} & 1 & & & & \\ m_{31} & 0 & 1 & & & \\ \cdot & \cdot & 0 & \cdot & & \\ \cdot & \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdot & & & \cdot \\ m_{n1} & 0 & 0 & & & 1 \end{bmatrix}$$

con $m_{i1} = \dfrac{x_i}{x_1}$ $\quad i = 1, 2, \ldots, n$, entonces

$$L_1^{-1} x = \begin{bmatrix} x_1 \\ x_2 - m_{21} x_1 \\ x_3 - m_{31} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n - m_{n1} x_1 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$, de acuerdo con la definición de m_{i1}.

Además si $x_j \neq 0$ y

$$L_j = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & m_{j+1,j} & 1 & \\ & & & \vdots & & \ddots \\ & & & m_{nj} & & 1 \end{bmatrix}, \quad m_{ij} = \frac{x_i}{x_j}, \quad i = j+1, j+2, \ldots, n$$

Entonces

$$L_j^{-1} x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Antes de proceder con le método LU, revisamos algunas propiedades de la partición de matrices.

Consideremos

$$A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \text{ y } B=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Suponemos que la partición de B_{ij} tiene la misma dimensión de A_{ij} , de otra manera el producto no sería conformable.

Las matrices se pueden partir de diferentes formas, pero es conveniente que los bloques diagonales (A_{11}, A_{22}) sean cuadrados. En muchos casos la multiplicación de dos matrices partidas (o particionadas) se puede interpretar como si los bloques fueran escalares, lo cual requiere que dichos bloques sean de las dimensiones apropiadas.

Lo anterior conduce a

$$AB=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}=\begin{bmatrix} (A_{11}+B_{11}+A_{12}B_{21}) & (A_{11}B_{12}+A_{12}B_{22}) \\ (A_{21}B_{11}+A_{22}B_{21}) & (A_{21}B_{12}+A_{22}B_{22}) \end{bmatrix}$$

Por ejemplo

$$A=\begin{bmatrix} 5 & 1 & 0 \\ 2 & 4 & 1 \\ -1 & 0 & 3 \end{bmatrix}=\begin{bmatrix} \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \begin{bmatrix} -1 & 0 \end{bmatrix} & \begin{bmatrix} 3 \end{bmatrix} \end{bmatrix}=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \text{ y}$$

$$B=\begin{bmatrix} 3 & -1 & 0 \\ 3 & 6 & 2 \\ 2 & 3 & 6 \end{bmatrix}=\begin{bmatrix} \begin{bmatrix} 3 & -1 \\ 3 & 6 \end{bmatrix} & \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\ \begin{bmatrix} 2 & 3 \end{bmatrix} & \begin{bmatrix} 6 \end{bmatrix} \end{bmatrix}=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

Entonces tenemos

$$A_{11}B_{11}=\begin{bmatrix}5 & 1\\ 2 & 4\end{bmatrix}\begin{bmatrix}3 & -1\\ 3 & 6\end{bmatrix}=\begin{bmatrix}18 & 1\\ 18 & 22\end{bmatrix};\ A_{12}B_{21}=\begin{bmatrix}0\\ 1\end{bmatrix}\begin{bmatrix}2 & 3\end{bmatrix}=\begin{bmatrix}0 & 0\\ 2 & 3\end{bmatrix}$$

$$A_{11}B_{12}=\begin{bmatrix}5 & 1\\ 2 & 4\end{bmatrix}\begin{bmatrix}0\\ 2\end{bmatrix}=\begin{bmatrix}2\\ 8\end{bmatrix};\ A_{12}B_{22}=\begin{bmatrix}0\\ 1\end{bmatrix}\begin{bmatrix}6\end{bmatrix}=\begin{bmatrix}0\\ 6\end{bmatrix};\ A_{21}B_{11}=\begin{bmatrix}-1 & 0\end{bmatrix}\begin{bmatrix}3 & -1\\ 3 & 6\end{bmatrix}=\begin{bmatrix}-3 & 1\end{bmatrix}$$

$$A_{22}B_{21}=\begin{bmatrix}3\end{bmatrix}\begin{bmatrix}2 & 3\end{bmatrix}=\begin{bmatrix}6 & 9\end{bmatrix};\ A_{21}B_{12}=\begin{bmatrix}-1 & 0\end{bmatrix}\begin{bmatrix}0\\ 2\end{bmatrix}=\begin{bmatrix}0\end{bmatrix};\ A_{22}B_{22}=\begin{bmatrix}3\end{bmatrix}\begin{bmatrix}6\end{bmatrix}=\begin{bmatrix}18\end{bmatrix}$$

$$AB=\begin{bmatrix}\begin{bmatrix}18 & 1\\ 20 & 25\end{bmatrix} & \begin{bmatrix}2\\ 9\end{bmatrix}\\ \begin{bmatrix}3 & 10\end{bmatrix} & \begin{bmatrix}18\end{bmatrix}\end{bmatrix}.$$

En general si dos matrices se parten (o particionan) confortablemente, esto es, que sus dimensiones hagan conformables los productos, entonces si

$$A=\begin{bmatrix}A_{11} & A_{12} & \cdot & \cdot & \cdot & A_{1N}\\ A_{21} & A_{22} & \cdot & \cdot & \cdot & A_{2N}\\ \cdot & \cdot & & & & \cdot\\ \cdot & \cdot & & & & \cdot\\ \cdot & \cdot & & & & \cdot\\ A_{N1} & A_{N2} & \cdot & \cdot & \cdot & A_{NN}\end{bmatrix},\ B=\begin{bmatrix}B_{11} & B_{12} & \cdot & \cdot & \cdot & B_{1N}\\ B_{21} & B_{22} & \cdot & \cdot & \cdot & B_{2N}\\ \cdot & \cdot & & & & \cdot\\ \cdot & \cdot & & & & \cdot\\ \cdot & \cdot & & & & \cdot\\ B_{N1} & B_{N2} & \cdot & \cdot & \cdot & B_{NN}\end{bmatrix}$$

El producto ***C=AB*** se puede particionar de la misma forma, y los bloques en ***C*** serán calculados a partir de

$$C_{ij}=\sum_{k=1}^{N}A_{ik}B_{kj}\ .$$

De lo discutido podemos ver que si $L_1 = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{m} & [I] \end{bmatrix}$,

donde $\underset{\sim}{m} = \begin{bmatrix} m_{21} \\ m_{31} \\ . \\ . \\ . \\ m_{n1} \end{bmatrix}$ e [***I***] es la matriz identidad.

Sea

$$P_n = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & [P_{n-1}] \end{bmatrix},$$

donde [P_{n-1}] es una matriz de permutación de orden ***n-1***, entonces

$$P_n L_1 P_n^{-1} = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & [P_{n-1}] \end{bmatrix} \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{m} & [I] \end{bmatrix} \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & [P_{n-1}] \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \underset{\sim}{0} \\ [P_{n-1}\underset{\sim}{m}] & [P_{n-1}] \end{bmatrix} \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & [P_{n-1}^{-1}] \end{bmatrix} = \begin{bmatrix} 1 & \underset{\sim}{0} \\ [P_{n-1}\underset{\sim}{m}] & [I] \end{bmatrix}.$$

FACTORIZACION ***LU***.

En las páginas anteriores se mostró que la eliminación gaussiana con pivoteo parcial es equivalente a la factorización ***PA=LU***, donde ***P*** es una matriz de permutación, ***L*** es una matriz triangular inferior y ***U*** es una matriz triangular superior.

Nos referimos aun ejemplo simple antes de efectuar una prueba constructiva del teorema para matrices 3x3.

Supongamos una matriz

$$A=\begin{bmatrix} 0.6 & 1.52 & 3.5 \\ 2 & 4 & 1 \\ 1 & 2.8 & 1 \end{bmatrix}.$$

En ***U*** se almacenará la matiz triangular superior que resulta de la eliminación gaussiana con pivoteo parcial, mientras que ***L*** se forma con los factores m_{ik} y con 1 en la diagonal.

Primero intercambiamos los renglones 1 y 2:

$$A=\begin{bmatrix} 2 & 4 & 1 \\ 0.6 & 1.52 & 3.5 \\ 1 & 2.8 & 1 \end{bmatrix},$$

donde $m_{21}=\frac{0.6}{2}$, $m_{31}=\frac{1}{2}=0.5$, por lo que

$$A'=\begin{bmatrix} 0.32 & 3.2 \\ 0.8 & 0.5 \end{bmatrix}, \quad L=\begin{bmatrix} 1 & \\ 0.3 & \cdot \\ 0.5 & \cdot \end{bmatrix}, \quad U=\begin{bmatrix} 2 & 4 & 1 \\ & \cdot & \\ & & \cdot \end{bmatrix}.$$

Intercambiamos los renglones de la matriz A'; notar que la columna 1 no se muestra debido a que los elementos de interés (renglones 2 y 3 en A') son cero. Lo anterior implica el intercambio de dichos renglones en *L* también. El resultado de esto se muestra a continuación

$$A'' = \begin{bmatrix} 0.8 & 0.5 \\ 0.32 & 3.2 \end{bmatrix} \quad m_{32} = \frac{0.32}{0.8} = 0.4$$

Las matrices correspondientes de *L* y *U* serán

$$L = \begin{bmatrix} 1 & & \\ 0.5 & 1 & \\ 0.3 & 0.4 & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 2 & 4 & 1 \\ & 0.8 & 0.5 \\ & & \cdot \end{bmatrix}$$

Además tenemos que $a''_{33} = 3.2 - (0.4)(0.5) = 3.0$, por lo que $A''' = [3.0]$, por lo que las matrices factor resultan

$$L = \begin{bmatrix} 1 & & \\ 0.5 & 1 & \\ 0.3 & 0.4 & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 2 & 4 & 1 \\ & 0.8 & 0.5 \\ & & 3.0 \end{bmatrix}.$$

Efectuando el producto *LU*, podemos comprobar que el resultado es la matriz original con los renglones intercambiados, de acuerdo al proceso de pivoteo parcial:

$$LU = \begin{bmatrix} 2 & 4 & 1 \\ 1 & 2.8 & 1 \\ 0.6 & 1.52 & 3.5 \end{bmatrix}.$$

La matriz de permutación asociada es

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Mostramos ahora el proceso general del método, en una matriz de 3x3. Sea una matriz de 3x3, no singular

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{23} & a_{33} \end{bmatrix}$$

El primer paso del proceso de eliminación se obtiene primeramente multiplicando, por la izquierda, la matriz *A* por una matriz de permutación P_1, debido a que se puede requerir un intercambio de renglones. El resultado es

$$A'=P_1A=\begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{bmatrix}$$

En este caso los apóstrofes en los elementos de la matriz implican un intercambio de renglones, efectuado por P_1.

Definimos ahora

$$L_1=\begin{bmatrix} 1 & & \\ m_{21} & 1 & \\ m_{31} & 0 & 1 \end{bmatrix}, \text{ donde } m_{21}\frac{a'_{21}}{a'_{11}}, \quad m_{31}\frac{a'_{31}}{a'_{11}}.$$

Multiplicar *A* por la izquierda por L'_1, equivale a restar $(m_{21}\times \mathrm{Re}\,ng\,1)$ del (*Reng 2*); al mismo tiempo se efectúa también la operación $(\mathrm{Re}\,ng\,3)-(m_{31}\times \mathrm{Re}\,ng\,1)$.

El resultado de lo anterior es

$$A^{(1)}=L'_1A'=\begin{bmatrix} 1 & & \\ -m_{21} & 1 & \\ -m_{31} & 0 & 1 \end{bmatrix}\begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{bmatrix}=\begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ 0 & a^{(1)}_{22} & a^{(1)}_{23} \\ 0 & a^{(1)}_{32} & a^{(1)}_{33} \end{bmatrix}$$

Donde $a^{(1)}_{ij}=a'_{ij}-m_{i1}a'_{1j} \qquad i=1,2,3\,;\; j=2,3.$

En el siguiente paso, posiblemente, efectuamos pivoteo para intercambiar los renglones 2 y 3, llevando a cabo las operaciones para hacer cero los elementos correspondientes de la columna 2:

$$A'' = P_2 A^{(1)} = \begin{bmatrix} a'_{11} & a'_{21} & a'_{31} \\ 0 & a'_{22} & a'_{23} \\ 0 & a'_{32} & a'_{33} \end{bmatrix}$$

Por lo que

$$U = A^{(2)} = L_2^{-1} A'' = \begin{bmatrix} 1 & & \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{bmatrix} \begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ 0 & a''_{22} & a''_{23} \\ 0 & a''_{32} & a''_{33} \end{bmatrix} = \begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ 0 & a''_{22} & a''_{23} \\ 0 & 0 & a_{33}^{(2)} \end{bmatrix}$$

Donde $m_{32} = \dfrac{a''_{32}}{a''_{22}}$, $\quad a_{33}^{(2)} = a''_{33} - m_{32} a''_{23}$.

Sustituyendo las operaciones asociadas con el paso de eliminación anterior obtenemos $U = L_2^{-1} P_2 L_1^{-1} P_1 A$, lo cual conduce a $A = P_1 L_1 P_2 L_2 U$, dado que, recordemos, $P_{ij}^{-1} = P_{ij}$.

Definimos $P = P_2 P_1$ y efectuamos el producto $PA = P\left(P_1 L_1 P_2 L_2 U\right) = P_2 L_1 P_2 L_2 U$.

Recordemos que P_2 es una matriz de conmutación que involucra únicamente los renglones 2 y 3, entonces la estructura de P_2 es $\begin{bmatrix} 1 & 0 \\ 0 & [Q] \end{bmatrix}$. En este caso la matriz [*Q*] es de 2x2, por lo que definimos

$$\bar{L}_1 = P_2 L_1 P_2 = \begin{bmatrix} 1 & & \\ m'_{21} & 1 & \\ m'_{31} & 0 & 1 \end{bmatrix}.$$

Los apóstrofes en los factores, indican que pudo haber intercambio de renglones en los elementos de la columna 1.

Tomando en cuenta todo lo anterior tendremos

$$PA = \bar{L}_1 L_2 U = LU = \begin{bmatrix} 1 & & \\ m'_{21} & 1 & \\ m'_{31} & m_{32} & 1 \end{bmatrix} U .$$

Lo anterior muestra que para el caso *m = 3*, la eliminación gaussiana con pivoteo parcial es equivalente a la descomposición ó factorización *LU* de la matriz original.

La prueba de los resultados anteriores se pueden generalizar. En consecuencia se puede enunciar el resultado anterior a través de un

TEOREMA DE FACTORIZACION *LU*. Cualquier matriz *A no singular nxn*, puede factorizarse como

PA = LU

P es una matriz de permutación, *L* es una matriz factor triangular inferior con diagonal unitaria (1 en todos lo elementos de la diagonal principal) y *U* es la matriz factor triangular superior.

Podemos agregar que si para una secuencia de pivoteo dada, es decir, que *P* está definida, entonces los factores *L* y *U* son determinados de manera única.

La gran ventaja de factorizar la matriz de coeficientes *A*, consiste en que la solución de sistemas de ecuaciones lineales con diferentes vectores de términos independientes $\underset{\sim}{b}$ y

en los cuales la matriz de coeficientes no cambia, es altamente eficiente, como se muestra a continuación.

Sea

$$A\underset{\sim}{x} = \underset{\sim}{b}$$

Con

$$LU = A, \text{ entonces}$$

$$LU\underset{\sim}{x} = \underset{\sim}{b}$$

Definimos

$$U\underset{\sim}{x} = \underset{\sim}{y}, \text{ un vector}$$

y por tanto

$$L\underset{\sim}{y} = \underset{\sim}{b}.$$

La solución de $A\underset{\sim}{x} = \underset{\sim}{b}$ se llevará a cabo en dos etapas, asociadas con las dos últimas ecuaciones como sigue.

Primera etapa: resolvemos por sustitución hacia delante

$$L\underset{\sim}{y} = \underset{\sim}{b}, \text{ y}$$

Segunda etapa: resolvemos por sustitución hacia atrás

$$U\underset{\sim}{x} = \underset{\sim}{y}$$

Debemos ahora modificar el código de la eliminación gaussiana, para incluir la información concerniente a este algoritmo de solución de sistemas de ecuaciones lineales por factorización *LU*.

Se debe guardar en memoria *L* y generar un vector entero $\underset{\sim}{p}$, donde se almacene la información del pivoteo, el cual se inicializa como $\underset{\sim}{p}_i = i$, donde $i = 1, 2, \ldots, n$.

Recordemos que *U* queda arriba de la diagonal principal de *A* y por tanto ya está guardada su información en el algoritmo de eliminación gaussiana.

El nuevo código ya no opera sobre el vector $\underset{\sim}{b}$.

El seudo-código resultante es

for *k=1* **to** *n-1* **do**

ν =indmax(a,k,n);

$p_\nu = p_k;\ p_k = \nu;$

$swap(a, k, \nu, n);$

for *i = k+1 to n* **do**

for all *(i,j),* $k+1 \le i,\ j \le n$ **do**

$a_{ij} = a_{ij} - a_{ik} * a_{kj};$

Cuando se resuelve $L\underset{\sim}{y} = P\underset{\sim}{b}$, requerimos de la información de $\underset{\sim}{p}$, el cual contiene la secuencia de pivoteo. El código correspondiente es como sigue:

for i = 1 **to** *n* **do**

$s = bp_i$ **;**

for *j = 1* **to** *i-1* **do**

$s = s - a_{ij} * y_j;$

$y_i = s;$

Factorización matricial para matrices positivas definidas.

Si *A* es una matriz simétrica y positiva definida, el elemento de mayor magnitud es positivo y está en la diagonal principal.

Lo anterior implica que todos los elementos que aparecen el a factorización *LU*, de *A*, sin usar pivoteo parcial, son más pequeños en magnitud ó iguales al elemento más grande en *A*.

Mostramos a continuación que para este caso, es posible obtener una factorización simétrica de una matriz simétrica y positiva definida.

Consideremos una descomposición *LU*:

$$A = LU$$

Con

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdot & \cdot & \cdot & u_{1n} \\ & u_{22} & \cdot & \cdot & \cdot & u_{2n} \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & u_{nn} \end{bmatrix}$$

Definimos ahora una matriz diagonal

$$D = \begin{bmatrix} u_{11} & & & & & \\ & u_{22} & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & u_{nn} \end{bmatrix}$$

El producto $D^{-1}U$ es equivalente a dividir los elementos de cada renglón en U por le elemento diagonal correspondiente:

$$D^{-1}U = \begin{bmatrix} u_{11}^{-1} & & & & & \\ & u_{22}^{-1} & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & u_{nn}^{-1} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdot & \cdot & \cdot & u_{1n} \\ & u_{22} & \cdot & \cdot & \cdot & u_{2n} \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & u_{nn} \end{bmatrix} = U' = \begin{bmatrix} 1 & u_{12}' & \cdot & \cdot & \cdot & u_{1n}' \\ & & \cdot & \cdot & \cdot & u_{2n}' \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & 1 \end{bmatrix}$$

Donde $u_{ij}' = \dfrac{u_{ij}}{u_{ii}} \qquad i = 1, 2, \ldots, n \qquad j = i, i+1, \ldots, n$.

Con lo anterior vemos que: $A = LU = LDD^{-1}U = LDU'$. Además con *A* simétrica tenemos

$$LU = A = A^T = \left(LDU'\right)^T = \left(U'\right)^T D^T L^T .$$

De acuerdo con lo mostrado vemos que $\left(U'\right)^T$ es una matriz triangular inferior con diagonal principal formada por 1; también podemos ver fácilmente que DL^T es una matriz triangular superior.

Dado que la factorización *LU* de una matriz es única, concluimos que $\left(U'\right)^T = L$.

Por lo que la factorización puede escribirse como $A = LDL^T$. Esta factorización se conoce como factorización LDL^T.

Un ejemplo complementa lo anterior. Sea una matriz simétrica positiva definida

$$A = \begin{bmatrix} 8 & 4 & 2 \\ 4 & 6 & 0 \\ 2 & 0 & 3 \end{bmatrix}$$

la cual tiene las matrices factor

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & -0.25 & 1 \end{bmatrix} \begin{bmatrix} 8 & 4 & 2 \\ 0 & 4 & -1 \\ 0 & 0 & 2.25 \end{bmatrix}$$

y la factorización LDL^T, con

$$D = \begin{bmatrix} 8 & & \\ & 4 & \\ & & 2.25 \end{bmatrix}.$$

En la última de estas matrices factor, no es necesario efectuar primero la factorización *LU*; en su lugar se obtienen los elementos de *L* y *D* directamente, como se indica a continuación.

Consideremos $A = LDL^T$

$$A = \begin{bmatrix} 1 & & & & & & & & & \\ l_{21} & 1 & & & & & & & & \\ \cdot & \cdot & & & & & & & & \\ \cdot & & \cdot & & & & & & & \\ \cdot & & & \cdot & & & & & & \\ l_{i1} & \cdot & \cdot & \cdot & l_{i,i-1} & 1 & & & & \\ \cdot & & & & & & \cdot & & & \\ \cdot & & & & & & & \cdot & & \\ \cdot & & & & & & & & \cdot & \\ l_{n1} & \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & 1 \end{bmatrix} \begin{bmatrix} d_1 & d_1 l_{21} & \cdot & \cdot & \cdot & d_1 l_{j1} & \cdot & \cdot & \cdot & d_1 l_{n1} \\ 0 & d_2 & \cdot & \cdot & \cdot & d_2 l_{j2} & \cdot & \cdot & \cdot & d_2 l_{n2} \\ 0 & 0 & \cdot & & & \cdot & & & & \\ & & & \cdot & & \cdot & & & & \\ \cdot & \cdot & & & \cdot & \cdot & & & & \\ \cdot & \cdot & & & & d_j & & & & \\ \cdot & \cdot & & & & 0 & \cdot & & & \\ & & & & & \cdot & & \cdot & & \\ & & & & & \cdot & & & \cdot & \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & & 0 & d_n \end{bmatrix}$$

Para $i \rangle j$ obtenemos

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik} d_k l_{jk} \; + d_j l_{ij}$$

Mientras que para $i = j$

$$a_{jj} = \sum_{k=1}^{j-1} l_{jk}^2 d_k \; + d_j \, .$$

Supongamos que ya hemos calculado $d_1, d_{2,\ldots,} d_{j-1}$ y los elementos en las columnas *1* a *j-1* en *L.*

De la última ecuación entonces $d_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k$ y por tanto, los elementos en la columna *j* a partir de la penúltima ecuación serán:

$$l_{ij} = \frac{\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_k l_{jk} \right)}{d_j} \qquad i = j+1, j+2, \ldots, n \, .$$

De las últimas ecuaciones recursivas, podemos ver que los factores $d_k l_{jk}$ en las sumas no dependen de *i*; por lo que podemos reorganizar el proceso recursivo como sigue:

$$r_k = d_k l_{jk} \qquad k = 1, 2, \ldots, j-1$$

$$d_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk} r_k$$

$$l_{jj} = \frac{\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} r_k \right)}{d_j} \qquad i = j+1, j+2, \ldots, n .$$

El proceso consiste en calcular los elementos de *L* por columna, de izquierda a derecha, así como los elementos de *D* al mismo tiempo.

Debido a que, como se mencionó antes, los elementos en *D* son positivos, una variante del método LDL^T será:

$$A = LDL^T = \left(LD^{\frac{1}{2}} \right)\left(D^{\frac{1}{2}} \right) L^T = U^T U$$

donde

$$D^{\frac{1}{2}} = \begin{bmatrix} \sqrt{d_1} & & & & \\ & \sqrt{d_2} & & & \\ & & . & & \\ & & & . & \\ & & & & . \\ & & & & \sqrt{d_n} \end{bmatrix}$$

y *U* es una matriz triangular superior.

Esta variante se conoce como el *método de factorización de cholesky*.

SISTEMAS DE ECUACIONES MAL CONDICIONADOS.

Con el fin de establecer algunos conceptos relativos a sistemas de ecuaciones mal condicionados (ill-conditioned), se presenta un resumen de conceptos de álgebra lineal necesarios para establecer dichos conceptos.

Una definición esencial es la norma vectorial y matrcicial.

Sea *A* una matriz de orden *n*, la **norma** de *A*, denotada $\|A\|$, debe cumplir:

1. $\|A\| \geq 0$ y $\|A\| = 0$, si y solo si $A = 0$

2. $\|kA\| = |k|\|A\|$

3. $\|A + B\| \leq \|A\| + \|B\|$ (desigualdad del triángulo)

4. $\|AB\| \leq \|A\|\|B\|$.

Para vectores, definimos la *norma euclidiana* (que corresponde a la longitud de un vector en el espacio de 2 y 3 dimensiones). Esta norma se define como

$$\|x\|_e = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}.$$

Existen otras formas de definir la norma. Una de ellas es mediante la suma de los valores absolutos de los x_i; también el valor máximo de las magnitudes de los x_i se puede usar.

En general podemos definir la *norma-p* como:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

De donde podemos definir:

$\|x\|_1 = \sum_{i=1}^{n} |x_i|$ suma de magnitudes

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}$$ norma euclidiana

$$\|x\|_\infty = \max_{1 \le i \le n} |x|_i$$ norma de la máxima magnitud.

Para ejemplificar los conceptos anteriores definimos un vector

$$x = \begin{bmatrix} 1.25 & 0.02 & -5.15 & 0 \end{bmatrix}.$$

Entonces las normas definidas resultan:

$$\|x\|_1 = |1.25| + |0.02| + |-5.15| + |0| = 6.42$$
$$\|x\|_2 = \left((1.25)^2 + (0.02)^2 + (-5.15)^2 + 0^2 \right)^{\frac{1}{2}} = 5.2996$$
$$\|x\|_\infty = |-5.15| = 5.15$$

La *norma de una matriz* de una matriz, se puede obtener por correspondencia a las anteriores como sigue

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|$$ Máxima suma de columnas

$$\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$$ Máxima suma de renglón.

La norma $\|A\|_2$ de una matriz se define en términos de los *valores característicos* (eigenvalores) de la matriz $A^T * A$.

Supongamos que *r* es el valor característico más grande de $A^T * A$, entonces a $\|A\|_2 = r^{1/2}$ se le denomina la *norma espectral de A* y siempre es menor o igual que $\|A\|_1$ y $\|A\|_\infty$.

Ejemplificamos lo anterior. Sea $A = \begin{bmatrix} 5 & 9 \\ -2 & 1 \end{bmatrix}$, entonces tendremos $\|A\|_1 = 10$ y $\|A\|_\infty = 14$.

Por otro lado, para obtener la norma $\|A\|_2$, debemos obtener los valores característicos de $A^T * A$. Dichos valores característicos resultan 4.9901 y 106.0099, por lo que $r = 106.0099$ y por tanto

$$\|A\|_2 = \sqrt{r} = \sqrt{106.0099} = 10.2961.$$

El mal condicionamiento, ill-conditioned, de los sistemas de ecuaciones es sinónimo de inestabilidad numérica.

Estos sistemas mal condicionados se caracterizan porque su solución es muy sensible a pequeños cambios (los cuales pueden ser errores) en la matriz ó en el vector de términos independientes.

Para aclarar estos conceptos recurrimos a un ejemplo sencillo y muy ilustrativo. Consideremos el sistema

$$\begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.00 \\ 2.00 \end{bmatrix}$$

Aquí es obvio que la solución es $x = 1 \quad y = 1$.

Consideremos ahora una pequeña modificación en el vector de términos independientes $\underset{\sim}{b}$:

$$\begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.02 \\ 1.98 \end{bmatrix}.$$

Podemos ver fácilmente que la solución es $x = 2 \quad y = 0$.

Sin embargo consideremos ahora otra pequeña variación en le mismo vector $\underset{\sim}{b}$, como se muestra:

$$\begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.98 \\ 2.02 \end{bmatrix},$$

lo cual conduce a la solución $x = 0 \quad y = 2$!.

Resumiendo, tres vectores $\underset{\sim}{b}$ muy parecidos conducen a soluciones muy distintas:

$$\underset{\sim}{b}_1 = \begin{bmatrix} 2.00 \\ 2.00 \end{bmatrix} \Rightarrow \underset{\sim}{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$\underset{\sim}{b}_2 = \begin{bmatrix} 2.02 \\ 1.98 \end{bmatrix} \Rightarrow \underset{\sim}{x}_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$
$$\underset{\sim}{b}_3 = \begin{bmatrix} 1.98 \\ 2.02 \end{bmatrix} \Rightarrow \underset{\sim}{x}_3 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

NUMERO DE CONDICION.

El grado de mal condicionamiento de una matriz se puede medir a través del denominado *número de condición* de una matriz. Este importante concepto se define como el producto de dos normas matriciales:

$$cond\,(A) = \|A\|\|A^{-1}\|.$$

Si el *cond(A)* es muy grande, estamos en presencia de un sistema mal condicionado.

Ejemplo, si

$$A = \begin{bmatrix} 3.02 & -1.05 & 2.53 \\ 4.33 & 0.56 & -1.78 \\ -0.83 & -0.54 & 1.47 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 5.661 & -7.273 & -18.55 \\ 200.5 & -268.3 & -669.9 \\ 76.85 & -102.6 & -255.9 \end{bmatrix}$$

Usando *normas*-∞, suma máxima de renglón, *cond(A)* será:

$$\|A\|\|A^{-1}\| = (6.67)(1138.7) = 7595.$$

Los elementos de A^{-1} serán muy grandes comparados con los de *A* cuando el sistema está mal condicionado.

El *cond(·)* permite relacionar la magnitud del error de la solución obtenida con la magnitud del residuo.

Sea $e = x - \bar{x}$, donde *x* es la solución exacta de $A\underset{\sim}{x} = \underset{\sim}{b}$, y $\bar{x}$ es una solución aproximada.

Definimos $\underset{\sim}{r} = \underset{\sim}{b} - A\bar{x}$, el residuo. De aquí tendremos

$$\underset{\sim}{r} = \underset{\sim}{b} - A\bar{x} = A\underset{\sim}{x} - A\bar{x} = A\left(\underset{\sim}{x} - \bar{x}\right) = A\underset{\sim}{e},$$

Por lo que

$$\underset{\sim}{e} = A^{-1}\underset{\sim}{r}\,.$$

Tomando en cuenta la norma de esta ecuación y recordando la propiedad: $\|AB\| \le \|A\|\,\|B\|$, tenemos

$$\|\underset{\sim}{e}\| \le \|A^{-1}\|\,\|\underset{\sim}{r}\|$$

Además como $\underset{\sim}{r} = A\underset{\sim}{e}$, tenemos

$$\|\underset{\sim}{r}\| \le \|A\|\,\|\underset{\sim}{e}\| \Rightarrow \frac{\|\underset{\sim}{r}\|}{\|A\|} \le \|\underset{\sim}{e}\|$$

Por lo que sustituyendo en la ecuación anterior obtenemos:

$$\frac{\|\underset{\sim}{r}\|}{\|A\|} \le \|\underset{\sim}{e}\| \le \|A^{-1}\|\,\|\underset{\sim}{r}\| \qquad (1)$$

Aplicamos el mismo proceso para $A\underset{\sim}{x} = \underset{\sim}{b}$, así como para $\underset{\sim}{x} = A^{-1}\underset{\sim}{b}$, obteniendo:

$$\frac{\|\underset{\sim}{b}\|}{\|A\|} \le \|\underset{\sim}{x}\| \le \|A^{-1}\|\,\|\underset{\sim}{b}\| \qquad (2).$$

Si combinamos (1) y (2):

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|\underset{\sim}{r}\|}{\|\underset{\sim}{b}\|} \le \frac{\|\underset{\sim}{e}\|}{\|\underset{\sim}{x}\|} \le \|A\|\|A^{-1}\|\frac{\|\underset{\sim}{r}\|}{\|\underset{\sim}{b}\|} \qquad (3)$$

Lo anterior conduce a

$$\frac{1}{cond\,(A)}\frac{\|\underset{\sim}{r}\|}{\|\underset{\sim}{b}\|} \le \frac{\|\underset{\sim}{e}\|}{\|\underset{\sim}{x}\|} \le cond\,(A)\frac{\|\underset{\sim}{r}\|}{\|\underset{\sim}{b}\|} \qquad (4).$$

Definimos: $\frac{\|\underset{\sim}{e}\|}{\|\underset{\sim}{x}\|} = \frac{\|\underset{\sim}{x} - \overline{x}\|}{\|\underset{\sim}{x}\|}$ como el *error relativo del vector solución calculado* $\overline{x}$, y $\frac{\|\underset{\sim}{r}\|}{\|\underset{\sim}{b}\|}$ como el *residuo relativo*, por lo que vemos que la expresión (4) nos indica que:

1. El error relativo puede ser tan grande como el residuo relativo multiplicado por el número de condición, *cond(·)*.
2. El error relativo puede ser tan pequeño como el residuo relativo dividido por el número de condición, *cond(·)*.

Lo anterior implica que cuando ***cond(·)* es muy grande**, el residuo da poca información acerca de la precisión de $\overline{x}$. Por otro lado, cuando ***cond(·)* es muy cercano a la unidad** el residuo relativo es una buena medida del error relativo de $\overline{x}$.

SOLUCIÓN

DE

SISTEMAS

DE

ECUACIONES NO LINEALES

Nuestro objetivo es resolver un sistema de ecuaciones cuya forma general es

$$\begin{aligned} &f_1\left(x_{1,}x_{2,}...,x_n\right)=0\\ &f_2\left(x_{1,}x_{2,}...,x_n\right)=0\\ &\bullet\\ &\bullet\\ &\bullet\\ &f_n\left(x_{1,}x_{2,}...,x_n\right)=0 \end{aligned} \qquad (1)$$

Un caso especial al de arriba lo representa el sistema de ecuaciones lineales

$$\begin{aligned} &a_{11}x_1+a_{12}x_2+...+a_{1n}x_n=b_1\\ &a_{21}x_1+a_{22}x_2+...+a_{2n}x_n=b_2\\ &\bullet\\ &\bullet\\ &\bullet\\ &a_{n1}x_1+a_{n2}x_2+...+a_{nn}x_n=b_n \end{aligned} \qquad (2)$$

que en forma compacta se escribe $A\underset{\sim}{x}=\underset{\sim}{b}$, donde *A* es la matriz de coeficientes [a_{ij}]

$\underset{\sim}{x}=\left[x_1,x_2,...,x_n\right]^T$ y $\underset{\sim}{b}=\left[b_1,b_2,...,b_n\right]^T$.

El sistema de ecuaciones (1) se resuelve, invariablemente, usando *técnicas numéricas iterativas*. El sistema de ecuaciones (2) se resuelven mediante el empleo de *métodos directos,* o bien mediante métodos iterativos, que en algunos casos pueden ser ventajosos en la soluciones de grandes sistemas de ecuaciones lineales y dispersos. Dado que en diferentes materias se cubre el material correspondiente a la solución de (2), mediante métodos directos, y que el uso de métodos iterativos requiere del conocimiento de material que nos se cubre en un curso introductorio de sistemas eléctricos de potencia, nos limitaremos a exponer un repaso del material correspondiente

a métodos iterativos para resolver sistemas de ecuaciones no lineales. Se expondrán dos métodos: el Gauss-Seidel y el Newton-Raphson.

METODO DE GAUSS-SEIDEL.

Expresamos (1) en la forma

$$\begin{aligned} x_1 &= \Phi_1\left(x_1, x_2, ..., x_n\right) \\ x_2 &= \Phi_2\left(x_1, x_2, ..., x_n\right) \\ x_3 &= \Phi_3\left(x_1, x_2, ..., x_n\right) \\ &\bullet \\ &\bullet \\ &\bullet \\ x_n &= \Phi_n\left(x_1, x_2, ..., x_n\right) \end{aligned} \qquad (3).$$

De manera compacta

$$x_i = \Phi_i\left(\underset{\sim}{x}\right) \qquad i = 1, 2, ..., n$$

Suponiendo un vector solución inicial $\underset{\sim}{x}_i^{(0)}$, las estimaciones del nuevo vector $x_i^{(k+1)} = \left[x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right]^T$, pueden ser obtenidas mediante: el método de Jacobi (llamado también método iterativo de Gauss), y el método de Gauss-Seidel.

Los métodos iterativos inician con un vector solución propuesto ***x***, el cual se mejora sistemáticamente a través de un esquema.

Los métodos iterativos tienen ventajas que los hacen atractivos

en algunos casos:

- Es posible almacenar y operar únicamente los elementos *no cero* de la matriz de coeficientes. Esto los hace ventajosos en el caso de sistemas de gran escala y dispersos (con un enorme porcentaje de ceros en la matriz de coeficientes). De hecho generalmente no se requiere ni almacenar una matriz de coeficientes.

- Los métodos iterativos son auto-correctivos, lo cual significa que los errores de redondeo (incluso errores aritméticos) en un ciclo iterativo, se corrigen en los ciclos subsecuentes.

Como desventaja importante podemos citar que estos métodos no siempre convergen a una solución.

La convergencia está garantizada si la matriz de coeficientes es diagonal dominante, es decir, si

$$\left|a_{ii}\right| \geq \sum_{j \neq i} \left|a_{ij}\right|$$

Método de Jacobi.

En el método de Jacobi, las iteraciones se definen por

$$x_i^{(k+1)} = \Phi_i\left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right) \qquad i = 1, 2, ..., n\,.$$

Método de Gauss-Seidel.

En el método de Gauss-Seidel, los valores recientemente calculados se usan en las ecuaciones, es decir, en la evaluación de las ecuaciones se utilizan los valores más actualizados de que disponemos, es decir

$$x_i^{(k+1)} = \Phi_i\left(x_1^{(k+1)}, x_2^{(k+1)}, ..., x_{i-1}^{(k+1)}, x_{i+1,...,}^{(k)} x_n^{(k)}\right) \qquad i = 1, 2, ..., n$$

En forma compacta

$$x_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k-1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}\right), \quad i = 1, 2, \ldots, n$$

Las iteraciones se continúan hasta que la máxima diferencia entre valores consecutivos de $x_i \left(i = 1, 2, ..., n\right)$, es menor que un valor predeterminado ε, esto es,

$$\underset{i}{Max}\left|x_i^{(k+1)} - x_i^{(k)}\right| \leq \varepsilon .$$

La convergencia de este método se puede mejorar mediante una técnica conocida como *relajación*. Esta técnica consiste en extrapolar el valor de la variable, basada en el valor de la iteración actual y el valor de la iteración anterior

$$x_{i,rel}^{(k)} = \alpha\, x_i^{(k)} + \left(1 - \alpha\right) x_i^{(k-1)}$$

El factor α se denomina *factor de relajación* o *factor de aceleración*.

Se puede observar que:

- $\alpha = 0$, no existe relajación o aceleración
- $\alpha < 1$, la ecuación anterior representa una interpolación (subrelajación)
- $\alpha > 1$, la ecuación anterior representa una extrapolación (sobreralajación)

El valor óptimo del factor de relajación no se puede estimar previamente a la simulación. Sin embargo se puede obtener un estimado del factor de relajación óptimo durante la simulación, como se muestra enseguida.

Definimos la magnitud del cambio en ***x*** durante la *k-ésima* iteración (llevada a cabo sin relajación), como

$$\Delta x^{(k)} = \left|x^{(k-1)} - x^{(k)}\right|$$

Si *k* es suficientemente grande, digamos $k \geq 5$, se puede demostrar que una estimación del valor óptimo será (*p* es un entero positivo)

$$\alpha_{opt} \approx \frac{2}{1 + \sqrt{1 - \left(\dfrac{\Delta x^{(k+p)}}{\Delta x^{(k)}}\right)^{1/p}}}$$

Los puntos esenciales del algoritmo de Gauss-Seidel con relajación se muestran a continuación:

1. Efectuar *k* iteraciones con ***α = 1*** (*k = 10* es razonable). Después de la *k-ésima* iteración guardar el valor de ***Δx(k).***
2. Realice ***p*** iteraciones adicionales y guarde ***Δx(k+p)*** para la última iteración.
3. Realice todas las subsecuentes iteraciones con $\boldsymbol{\alpha = \alpha_{opt}}$, donde $\boldsymbol{\alpha_{opt}}$ es calculada por medio de la ecuación antes mostrada.

METODO DE NEWTON-RAPHSON.

El método de Newton-Raphson es aplicado directamente al sistema de ecuaciones (1). Constituye una extensión del caso de 1er orden , por lo cual es conveniente recordarlo brevemente .

Consideremos la ecuación no lineal $f(x)=0$. Suponiendo un valor de arranque $x^{(0)}$, expandamos en serie de Taylor $f(x)$ alrededor de $x^{(0)}$, o sea, tomando como punto base $x^{(0)}$. La ecuación resulta entonces

$$f\left(x^{(0)}\right)+\left(x-x(0)\right)f'\left(x^{(0)}\right)+\tfrac{1}{2!}\left(x-x(0)\right)^2 f''\left(\ddot{x}^{(0)}\right)+....=0.$$

Despreciando los términos de segundo orden y orden superior, obtenemos

$$f\left(x^{(0)}\right)+\left(x-x^{(0)}\right)f'\left(x^{(0)}\right)=0.$$

De esta última ecuación despejamos *x*, con el fin de obtener un estimado más cercano a la solución

$$x^{(1)} = x^{(0)} - \frac{f\left(x^{(0)}\right)}{f'\left(x^{(0)}\right)}$$

en donde a la *x* la hemos denominado $x^{(1)}$ en la última ecuación.

La ecuación anterior puede aplicarse de manera iterativa, hasta alcanzar el valor deseado, mediante la ecuación general

$$x^{(k+1)} = x^{(k)} - \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}.$$

La convergencia puede probarse mediante el criterio $|f| \leq \varepsilon$. De hecho si el método iterativo converge, $f \to 0$. Es importante recordar que la ecuación $f(x) = 0$ puede tener varias soluciones, por lo que en caso de converger, el método probablemente lo hará al valor más cercano al valor de arranque.

Sistema de ecuaciones no lineales. Consideramos el sistema de ecuaciones (1), que repetimos por comodidad

$$\begin{aligned}
&f_1(x_1, x_2, ..., x_n) = y_1 \\
&f_2(x_1, x_2, ..., x_n) = y_2 \\
&\bullet \\
&\bullet \\
&\bullet \\
&f_n(x_1, x_2, ..., x_n) = y_n
\end{aligned}$$

En este sistema mostrado hay una diferencia con respecto al descrito en (1) sin embargo, y es que en lugar de estar igualadas a cero las ecuaciones, estas están igualadas a un valor constante, $y_1, y_2,\ldots,y_n$. Lo anterior no debe representar ningún problema, puesto que es obvio que se trata del mismo sistema de ecuaciones, solamente que la forma del expuesto arriba es más apropiada para la formulación del problema de flujos, como se verá más adelante.

Siguiendo el esquema del caso de 1er orden, efectuamos la expansión en serie de Taylor para cada una de las funciones que constituyen el sistema de ecuaciones no lineales. Si denominamos al vector $\underset{\sim}{x}^{(0)} = \left[x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)} \right]^T$, vector de arranque, y suponemos que $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$, son las correcciones requeridas para que el vector $\underset{\sim}{x}^{(0)}$ sea la solución, tendremos que al sustituir en la ecuación anterior

$$\begin{aligned}
&f_1\left(x_1^{(0)} + \Delta x_1, x_2^{(0)} + \Delta x_2, \ldots, x_n^{(0)} + \Delta x_n\right) = y_1 \\
&f_2\left(x_1^{(0)} + \Delta x_1, x_2^{(0)} + \Delta x_2, \ldots, x_n^{(0)} + \Delta x_n\right) = y_2 \\
&\bullet \\
&\bullet \\
&\bullet \\
&f_n\left(x_1^{(0)} + \Delta x_1, x_2^{(0)} + \Delta x_2, \ldots, x_n^{(0)} + \Delta x_n\right) = y_n
\end{aligned} \qquad (4)$$

Aplicamos el teorema de Taylor a cada una de las ecuaciones del conjunto (4).

Para la primera ecuación obtenemos

$$f_1\left(x_1^{(0)} + \Delta x_1, x_2^{(0)} + \Delta x_2, \ldots, x_n^{(0)} + \Delta x_n\right) = f_1\left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right) + \Delta x_1 \left.\frac{\partial f_1}{\partial x_1}\right|_0 + \Delta x_2 \left.\frac{\partial f_1}{\partial x_2}\right|_0 + \ldots + \Delta x_n \left.\frac{\partial f_1}{\partial x_n}\right|_0 + \Phi_1$$

en este caso Φ_1 es una función de potencias de Δx_1 , $\Delta x_2, \ldots, \Delta x_n$, de grado mayor a 1, así como de derivadas de alto orden de f_1. Si los estimados iniciales (vector de arranque)

están cerca de la solución, los valores de $\Delta x_1, \Delta x_2, ..., \Delta x_n$ serán muy pequeños y por tanto se podrán despreciar los términos con potencias de grado superior.

De acuerdo a lo anterior, el sistema de ecuaciones tendrá la forma

$$f_1\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) + \Delta x_1 \left.\frac{\partial f_1}{\partial x_1}\right|_0 + \Delta x_2 \left.\frac{\partial f_1}{\partial x_2}\right|_0 + ... + \Delta x_n \left.\frac{\partial f_1}{\partial x_n}\right|_0 = y_1$$

$$f_2\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) + \Delta x_1 \left.\frac{\partial f_1}{\partial x_1}\right|_0 + \Delta x_2 \left.\frac{\partial f_1}{\partial x_2}\right|_0 + ... + \Delta x_n \left.\frac{\partial f_1}{\partial x_n}\right|_0 = y_2$$

•

•

•

$$f_n\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) + \Delta x_1 \left.\frac{\partial f_1}{\partial x_1}\right|_0 + \Delta x_2 \left.\frac{\partial f_1}{\partial x_2}\right|_0 + ... + \Delta x_n \left.\frac{\partial f_1}{\partial x_n}\right|_0 = y_n$$

Por lo que despejando los primeros términos, y usando notación matricial, tendremos

$$\begin{bmatrix} y_1 - f_1\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) \\ y_2 - f_2\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) \\ ..... \\ y_n - f_n\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) \end{bmatrix} = \begin{bmatrix} \left.\frac{\partial f_1}{\partial x_1}\right|_0 & \left.\frac{\partial f_1}{\partial x_2}\right|_0 & ... & \left.\frac{\partial f_1}{\partial x_n}\right|_0 \\ \left.\frac{\partial f_2}{\partial x_1}\right|_0 & \left.\frac{\partial f_2}{\partial x_2}\right|_0 & ... & \left.\frac{\partial f_2}{\partial x_n}\right|_0 \\ . & . & . & . \\ \left.\frac{\partial f_n}{\partial x_1}\right|_0 & \left.\frac{\partial f_n}{\partial x_2}\right|_0 & ... & \left.\frac{\partial f_n}{\partial x_n}\right|_0 \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ ... \\ \Delta x_n \end{bmatrix} \qquad (5)$$

El proceso se trabaja en forma iterativa, en cuyo caso el sistema general sería como

$$\begin{bmatrix} y_1 - f_1\left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right) \\ y_2 - f_2\left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right) \\ ..... \\ y_n - f_n\left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right) \end{bmatrix} = \begin{bmatrix} \left.\frac{\partial f_1}{\partial x_1}\right|_k & \left.\frac{\partial f_1}{\partial x_2}\right|_k & ... & \left.\frac{\partial f_1}{\partial x_n}\right|_k \\ \left.\frac{\partial f_2}{\partial x_1}\right|_k & \left.\frac{\partial f_2}{\partial x_2}\right|_k & ... & \left.\frac{\partial f_2}{\partial x_n}\right|_k \\ . & . & . & . \\ \left.\frac{\partial f_n}{\partial x_1}\right|_k & \left.\frac{\partial f_n}{\partial x_2}\right|_k & ... & \left.\frac{\partial f_n}{\partial x_n}\right|_k \end{bmatrix} \begin{bmatrix} \Delta x_1^k \\ \Delta x_2^k \\ ... \\ \Delta x_n^k \end{bmatrix} \qquad (6)$$

En forma compacta

$$\left[J\right]\underset{\sim}{C} = \underset{\sim}{D} \qquad (7)$$

donde $\underset{\sim}{C}$ es el vector de correcciones, mientras $\underset{\sim}{D}$ es el vector de desajustes, o sea de diferencias de los valores constantes y las funciones evaluadas en el vector obtenido en la iteración correspondiente.

El vector izquierdo contiene las diferencias de los términos conocidos menos las funciones evaluadas con los vectores obtenidos en cada iteración. Lo denominamos vector de diferencias. La matriz de primeras derivadas parciales se conoce como matriz Jacobiana, y sus elementos son valores numéricos obtenidos al evaluar las expresiones obtenidas al evaluar las derivadas indicadas con los vectores obtenidos en cada iteración. Finalmente el vector de la derecha es el vector de correcciones, pues como se indicó anteriormente, representa el vector requerido para corregir el vector solución de la iteración anterior, rumbo a la solución.

Con el objeto de entender el algoritmo, se muestra un ejemplo con un sistema de orden 2.

El objetivo es encontrar la solución del sistema de ecuaciones

$$f_1\left(x_1, x_2\right) = x_1^2 + 3x_1x_2 - 4$$
$$f_2\left(x_1, x_2\right) = x_1x_2 - 2x_2^2 + 5$$

Arrancamos el proceso iterativo con

$$\underset{\sim}{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Evaluamos las expresiones de la matriz Jacobiana:

$$[J] = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix}$$

donde las derivadas parciales estarán dadas por las siguientes expresiones

$$\frac{\partial f_1}{\partial x_1} = 2x_1 + 3x_2 \qquad \frac{\partial f_1}{\partial x_2} = 3x_1$$

$$\frac{\partial f_2}{\partial x_1} = x_2 \qquad \frac{\partial f_2}{\partial x_2} = x_1 - 4x_2$$

Observamos que $y_1 = 4$, $y_2 = -5$, por lo que $\underset{\sim}{f}\left(\underset{\sim}{x}^{(0)}\right) = \begin{bmatrix} 4-7 \\ -5-(-6) \end{bmatrix} = -\begin{bmatrix} 3 \\ -1 \end{bmatrix}$.

Si calculamos la matriz Jacobiana y la invertimos obtendremos

$$[J]^{-1} = \begin{bmatrix} 0.1129 & 0.04839 \\ 0.03226 & -0.12903 \end{bmatrix}$$

por lo que $\underset{\sim}{x}^{(1)} = \underset{\sim}{x}^{(0)} + \left[J\right]^{-1} \underset{\sim}{f}\left(\underset{\sim}{x}^{(0)}\right)$, resulta en

$$\underset{\sim}{x}^{(1)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0.1129 & 0.04839 \\ 0.03226 & -0.12903 \end{bmatrix} \left\{ - \begin{bmatrix} 3 \\ -1 \end{bmatrix} \right\} = \begin{bmatrix} 0.70968 \\ 1.77419 \end{bmatrix}.$$

Si efectuamos las iteraciones subsecuentes, siguiendo el mismo procedimiento, obtendremos

$$\underset{\sim}{x}^{(2)} = \underset{\sim}{x}^{(1)} + \left[J\right]^{-1} \underset{\sim}{f}\left(\underset{\sim}{x}^{(1)}\right) = \begin{bmatrix} 0.67302 \\ 1.75831 \end{bmatrix}$$

para la segunda iteración.

Antes de seguir con los resultados de las siguientes iteraciones, es importante mencionar que el criterio de convergencia se aplica al vector de diferencias, dado que cuando este vector sea cero, entonces el vector empleado para evaluar las funciones que conforman dicho vector es la solución del problema, de acuerdo a (6).

Para la tercera iteración tenemos que la inversa de la matriz Jacobiana es

$$\left[J\right]^{-1} = \begin{bmatrix} 0.13929 & 0.044220 \\ 0.03851 & -0.14500 \end{bmatrix}$$

de donde obtenemos

$$\underset{\sim}{x}^{(3)} = \underset{\sim}{x}^{(2)} + \left[J\right]^{-1} \underset{\sim}{f}\left(\underset{\sim}{x}^{(2)}\right) = \begin{bmatrix} 0.67259 \\ 1.75820 \end{bmatrix}.$$

Podemos verificar fácilmente que $\max_i \left(\underset{\sim}{x}^{(3)} \right) < \varepsilon$. Por tanto, el vector $\underset{\sim}{x}^{(3)}$ es la solución.

SOLUCIÓN NUMÉRICA

DE

ECUACIONES

DIFERENCIALES

ORDINARIAS

INTRODUCCIÓN. MÉTODOS DE UN SOLO PASO.

El objetivo de estas notas complementarias al tema de solución numérica de ecuaciones diferenciales ordinarias es dar una introducción simple al tema, basada en principios de cálculo. Antes de entrar al tema en términos más generales, este enfoque permite establecer los métodos más simples del tipo de *un paso* o de *paso simple*. Posteriormente la desarrollar los métodos Runge-Kutta, será indispensable otro enfoque del tema, basado en la serie de Taylor.

El objetivo en este tema es resolver la ecuación diferencial ordinaria de primer orden

$$\frac{dy}{dt} = f\left(y,t\right) \qquad (1)$$

Sujeta a la condición inicial $y(0) = Y_0$. El problema de condiciones de frontera no será tratado en este curso de introducción al tema, dado que su carácter es más complejo, pero de cualquier manera los conceptos aprendidos en este curso de introducción serán la base de estudios más avanzados donde se cubren dichos temas.

Por otro lado es importante recalcar que aunque el planteamiento se refiere a ecuaciones diferenciales ordinarias de primer orden, se puede resolver cualquier ecuación de diferencial ordinaria de más alto orden convirtiéndola en un conjunto de ecuaciones diferenciales ordinarias de primer orden, en número igual al orden de la ecuación diferencial ordinaria original.

Regresando al planteamiento de la solución numérica de la ecuación diferencial ordinaria (EDO), es importante comentar que dicha solución, obviamente, consistirá de una colección de puntos, que representarán aproximaciones de la solución real o verdadera, la cual no conocemos por supuesto. Esto significa que lo que obtendremos es una representación finita y aproximada de la curva solución verdadera *y(t)*. El punto de inicio será por supuesto la condición inicial $y(0) = Y_0$.

Tomemos la ecuación (1) y empecemos por escribirla de la forma

$$dy = f\left(y,t\right)dt$$

Si integramos esta última ecuación entre los valores de t_k y t_{k+1}, partiendo de que estamos en el paso $k+1$ del proceso recursivo, y por tanto conocemos el valor de $y(t_k)$ que denotaremos por simplicidad como y_k, entonces obtenemos el valor de $y\left(k+1\right)$ que denotaremos como y_{k+1}, como se muestra continuación.

$$\int_{y(k)}^{y(k+1)} dy = \int_{t_k}^{t_{k+1}} f\left(y,t\right)dt$$

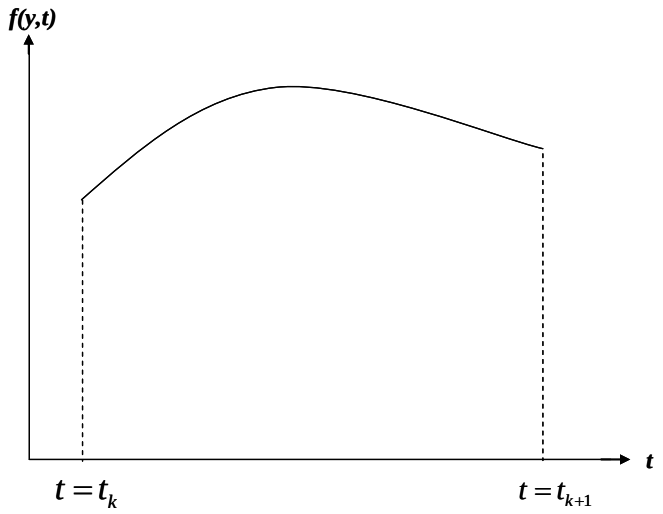
Lo cual resulta

$$y_{k+1} - y_k = \int_{t_k}^{t_{k+1}} f\left(y,t\right)dt$$

De donde obtenemos finalmente

$$y_{k+1} = y_k + \int_{t_k}^{t_{k+1}} f\left(y,t\right)dt \qquad (2)$$

Esta última ecuación es la base para obtener los métodos de paso simple que obtendremos. Su obtención depende de la solución numérica, por supuesto, de la integral del lado derecho de (2).



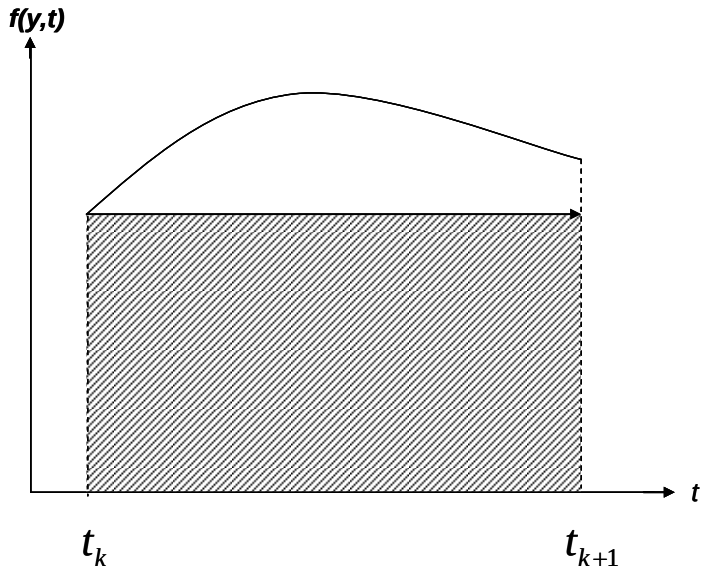
Integral de la función f(y,t) en intervalo

Podemos aproximar dicha integral a través de la aproximación lineal de la curva *f(y,t)*, para lo cual existen tres opciones.

1. Aproximación por recta constante e igual a la ordenada en el punto t_k, es decir $f(y_k, t_k) = f_k$, por lo que la integral $\int_{t_k}^{t_{k+1}} f(y,t)\, dt$, que representa como sabemos el área bajo la curva *f(y,t)*, comprendida entre las rectas t_k y t_{k+1}, se aproxima por el área del rectángulo de área igual a

$$f_k \left(t_{k+1} - t_k \right)$$

Lo anterior se muestra en la figura siguiente.



Aproximación numérica de la integral.

Método de Euler explícito

Si definimos la diferencia $\left(t_{k+1}-t_k\right)$ como *h*, que es el paso de integración, entonces obtenemos la ecuación recursiva del método presente como

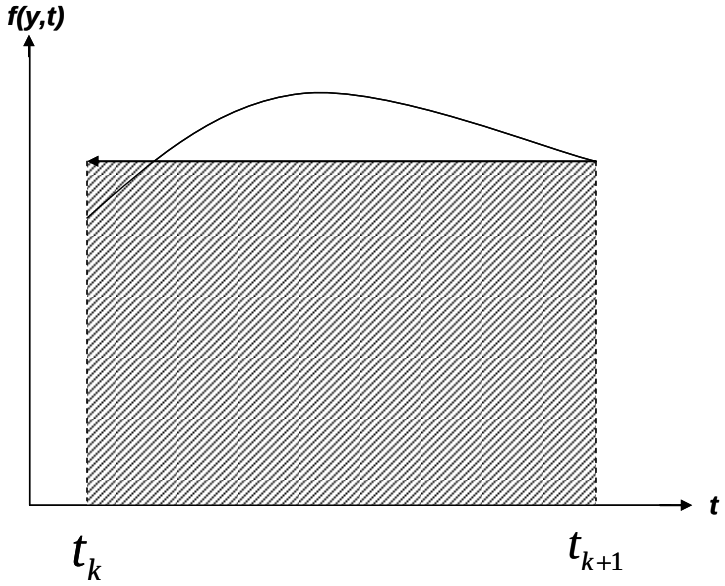
$$y_{k+1}=y_k+hf_k \qquad (3)$$

La ecuación anterior es la fórmula recursiva del método denominado de *Euler hacia adelante* o *Euler explícito*.

2. Aproximación por recta constante e igual a la ordenada en le punto t_{k+1}, es decir, $f(y_{k+1}, t_{k+1}) = f_{k+1}$, por lo que la integral $\int_{t_k}^{t_{k+1}} f(y,t)\, dt$, que representa como sabemos el área bajo la curva *f(y,t)*, comprendida entre las rectas t_k y $t_{k+1,}$ se aproxima por el área del rectángulo de área igual a

$$f_{k+1}\left(t_{k+1} - t_k\right).$$

Esto se muestra en la figura a continuación.



Aproximación numérica de la integral.

Método de Euler implícito.

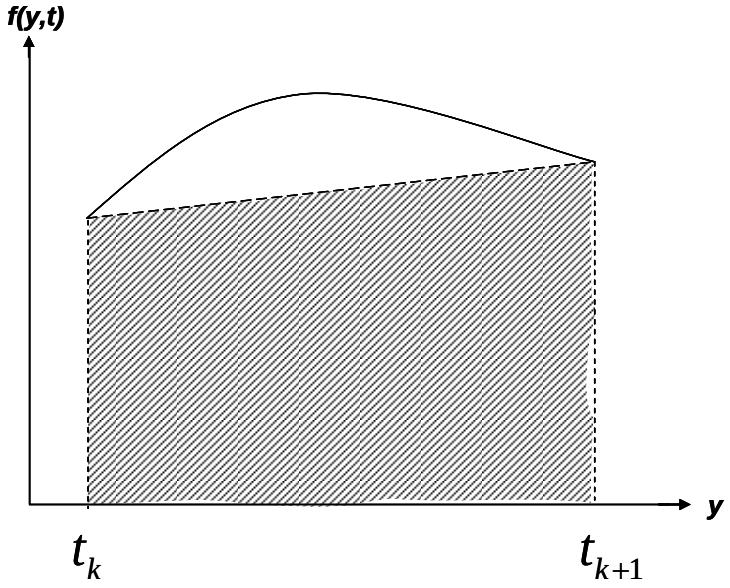
Por lo que sustituyendo en la ecuación (2) obtenemos la fórmula recursiva

$$y_{k+1} = y_k + h\, f_{k+1} \qquad (4)$$

La fórmula anterior se conoce como la fórmula recursiva del método de *Euler hacia atrás o Euler implícito*.

El nombre de los dos últimos métodos hace evidente que las ecuaciones recursivas que los definen, son del tipo explícitas, en el primer caso, e implícitas en el segundo. Lo anterior indica que en el primero y_{k+1} está en función de t_k y y_k, mientras que en el segundo está en función de t_{k+1} y y_{k+1}. Si la función del integrando es lineal, la ecuación recursiva del *Euler implícito* se puede escribir de forma explícita y por tanto el método será de paso simple, de lo contrario habría que desarrollar otro procedimiento del tipo denominado *predictor-corrector*.

3. El tercer caso corresponde a la aproximación numérica de la integral, por medio de la regla trapecial, es decir, aproximando la curva *f(y,t)* por una recta que une los puntos correspondientes a las coordenadas (t_k, f_k) y (t_{k+1}, f_{k+1}), como se muestra en la figura siguiente.



Aproximación numérica de la integral.

Método de la regla trapecial.

El área que representa la aproximación numérica en este caso será igual a

$$\frac{(t_{k+1}-t_k)}{2}(f_{k+1}+f_k)$$

O bien

$$\frac{h}{2}(f_{k+1}+f_k)$$

De donde sustituimos en la ecuación (2) para obtener así el tercer método de donde resulta la fórmula recursiva del método trapecial

$$y_{k+1} = y_k + \frac{h}{2}(f_{k+1} + f_k) \qquad (5).$$

Al igual que en el caso anterior, este método produce una fórmula recursiva implícita, por lo que es un método implícito, en general. Sin embargo en ambos casos, si la función *f(y,t)* es lineal, se podrán hacer las factorizaciones apropiadas para escribir la fórmula en forma explícita. Lo anterior no se podría efectuar en el caso de que la función mencionada fuera no lineal, en cuyo caso habría que diseñar un método iterativo para resolver el problema concreto por estos métodos implícitos.

ERROR DE TRUNCAMIENTO DEL EULER EXPLÍCITO.

Utilizaremos el método de Euler explícito, denominado también Euler hacia adelante, para desarrollar una fórmula que nos permita cuantificar el error de truncamiento de los métodos para resolver ecuaciones diferenciales ordinarias numéricamente.

Para tal efecto, usamos la serie de Taylor, suponiendo que le punto base está definido por (x_i,t_i), el cual es el punto en que nos basamos para evaluar la ordenada del siguiente punto que estamos calculando, x_{i+1}

$$y_{i+1} = y_i + y_i^{'} h + \frac{y_i^{''}}{2!} h^2 + \cdots + \frac{y_i^{(n)}}{n!} h^n + \Re_n$$

$$\text{donde: } \Re_n = \frac{y_i^{(n+1)}(\xi)}{(n+1)!} h^{n+1} \qquad h = t_{i+1} - t_i$$

Dado que $y_i^{'} = f(t_i, y_i)$ obtenemos, a partir de la expresión anterior

$$y_{i+1} = y_i + f(t_i, y_i)h + \frac{f'(t_i, y_i)}{2!} h^2 + \cdots + \frac{f^{(n+1)}(t_i, y_i)}{n!} h^n + O(h^{n+1})$$

El término $O(h^{n+1})$ representa el resto de la serie y especifica que el error de truncamiento es función del paso de integración elevado a la potencia ***n+1***.

Comparando esta última expresión con la fórmula del método de Euler explícito, vemos que dicha fórmula está contenida en los dos primeros términos de la derecha de la expresión, lo que significa que el resto constituye el error de truncamiento, es decir

$$E_t = \frac{f'(t_i, y_i)}{2!} h^2 + \cdots + O(h^{n+1})$$

Debido a que la serie es infinita, la mejor aproximación representa la mejor estimación del error de truncamiento. Los valores de los términos en esta serie decrecen de forma monótona evidentemente, por lo que podemos afirmar que el mayor valor lo contiene el primer término de la serie de $\boldsymbol{E_t}$.

Todos los métodos de un solo paso requieren el mismo tratamiento en la evaluación del erro de truncamiento. Hemos usado como ejemplo el presente método, dado que resulta en el análisis más simple; sin embargo el procedimiento en los demás casos es igual, tomando en cuenta que debido a las características de estos, el procedimiento es un poco más complicado [RR].

ESTBILIDAD DE LA SOLUCION NUMERICA.

Definiciones.

La estabilidad es una de las propiedades más críticas de los métodos numéricos para resolver ecuaciones diferenciales. En esta sección aprovechamos la introducción al tema, a través del desarrollo de las fórmulas recursivas presentadas, con el fin de discutir este complejo tema, de manera introductoria. El tema es complejo y existe literatura que lo trata de manera exclusiva.

Es posible que la solución numérica d una ecuación diferencial crezca sin límite, aún cuando la *solución exacta* (solución analítica, no conocida por lo general) permanezca acotada. Por supuesto también existirán casos en los que la solución exacta crezca indefinidamente.

En nuestro caso, nos limitaremos a la discusión de estabilidad de ecuaciones diferenciales para las cuales la solución exacta está acotada.

Comenzamos considerando la ecuación diferencial ordinaria (1) y un método numérico. En el análisis de estabilidad buscamos las condiciones y parámetros del método numérico para los cuales la solución numérica permanece acotada. El parámetro más importante es el paso de integración *h*.

Tenemos tres clases de métodos numéricos:

- Esquema numérico estable: Su solución numérica está acotado, es decir, no crece sin control con cualquier selección de los parámetros, principalmente del paso de integración *h*. Su robustez puede tener alto costo computacional.

- Esquema numérico inestable: Su solución numérica crece sin límite, sin importar el valor seleccionado de los parámetros. Estos es que más carecen de utilidad, aun cuando fueran precisos.

- Esquema condicionalmente estable: La solución permanece acotada solamente con ciertos valores de parámetros.

Estabilidad de los métodos.

La estabilidad de los métodos se estudia por medio d una ecuación diferencial especial, denominada *problema modelo*:

$$y' = \lambda y \qquad (6)$$

Cuya solución exacta es

$y = y(0)e^{\lambda t}$, donde λ puede ser real ó complejo.

Euler explícito:

La fórmula de este método, $y_{k+1} = y_k + h\, f(y_k, t_k)$, nos conduce a

$$x_1 = x_0 + h(\lambda x_0) = x_0(1+\lambda h)$$
$$x_2 = x_1 + h(\lambda x_1) = x_1(1+\lambda h) = x_0(1+\lambda h)^2$$
•
•
•
$$x_n = x_{n-1} + h(\lambda x_{n-1}) = x_{n-1}(1+\lambda h) = x_0(1+\lambda h)^n$$

A medida que crece *n* a ∞, un resultado finito para una ecuación diferencial estable $(\Re e\{\lambda\} \langle 0)$ requiere

$$|1+\lambda h| \leq 1 \qquad (7).$$

La desigualdad anterior es la condición de estabilidad para el método de Euler implícito.

λ puede ser compleja, aunque *h* sea real. Por lo que definimos

$$h\lambda = q = u + jv \qquad (8).$$

Sustituimos en (7) para obtener

$$\left|1 + u + jv\right| \leq 1$$

O bien

$$\left(1+u\right)^2 + v^2 \leq 1 \qquad (9).$$

El lugar geométrico de la expresión anterior es un círculo con centro en (-1,0) y radio unitario, el cual pasa por el origen. La región asociada con la (8) incluye el interior de dicho círculo.

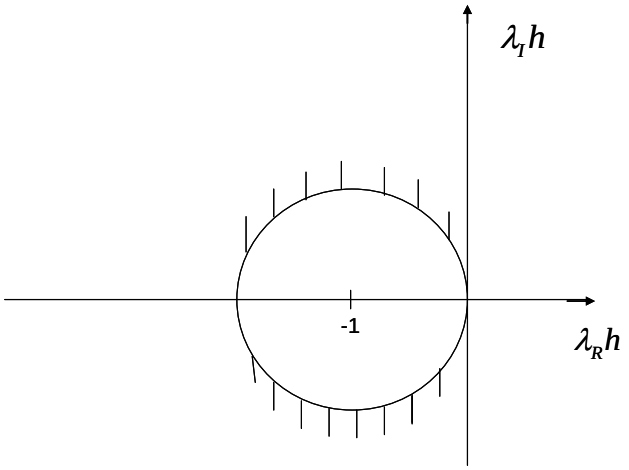


Diagrama de estabilidad para el método Euler explícito

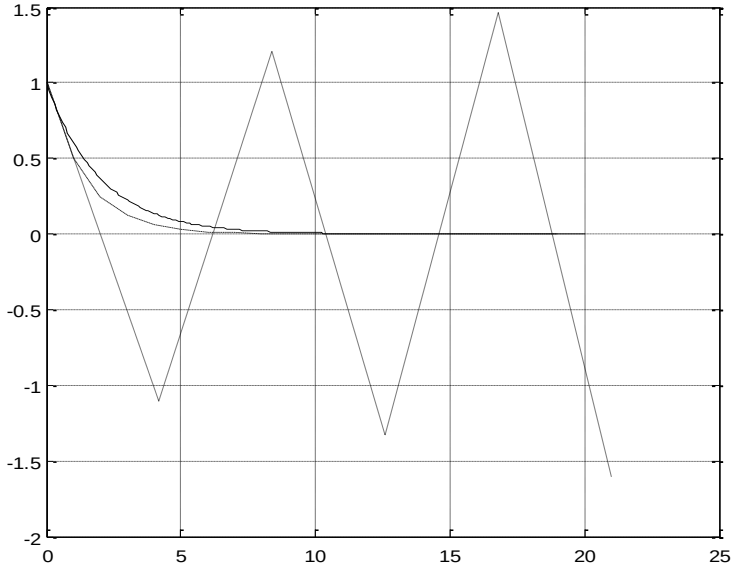
Lo anterior implica que si suponemos $\Re e\{\lambda\}\langle 0$ (solución estable), el valor de *h* debe ser tal que el producto $q = \lambda h$ representa un punto dentro círculo.

Un ejemplo ilustrativo de anterior lo ilustra la solución numérica de la sencilla ecuación diferencial

$$y' + 0.5y = 0$$
$$y(0) = 1 \qquad 0 \le t \le 20$$

Usamos dos valores del paso de integración. El primero $h = 1$ y el segundo $h = 4.2$. De la ecuación (7) vemos que la desigualdad se cumple en le primer caso, es decir, para $h = 1$, mientras que para el segundo, $h = 4.2$, dicha desigualdad no se cumple.

La gráfica siguiente muestra los trazos correspondientes a la solución, en trazo continuo, el primer caso, $h = 1$ en trazo con .- **(estable)**, y el segundo caso (**inestable**, oscilatorio y creciente) en línea punteada - -.



Solución Numérica de la EDO por el método de Euler explícito.

<u>Euler implícito</u>:

La fórmula recursiva del método es

$$y_{k+1} = y_k + h\, f\left(y_{k+1}, t_{k+1}\right).$$

Aplicamos esta fórmula recursiva a la ecuación (6) para obtener

$$y_1 = y_0 + h\left(\lambda y_1\right)$$

De donde obtenemos

$$y_1 = \frac{y_0}{1-\lambda h} = \frac{y_0}{1-q}$$

Si procedemos con el método en forma recursiva, en el paso *n-ésimo* tendremos

$$y_n = \left[\frac{1}{1-q}\right]^n y_0 \qquad (10)$$

La condición asociada con un método estable requiere que a medida que $n \to \infty$,

$$\left|\frac{1}{1-q}\right| \leq 1$$

Por lo que tendremos

$$1 \leq (1-u)^2 + v^2 \qquad (11).$$

La igualdad de la ecuación anterior representa un círculo con centro en (1,0) que pasa por el origen. La desigualdad de (11) se satisface fuera del círculo.

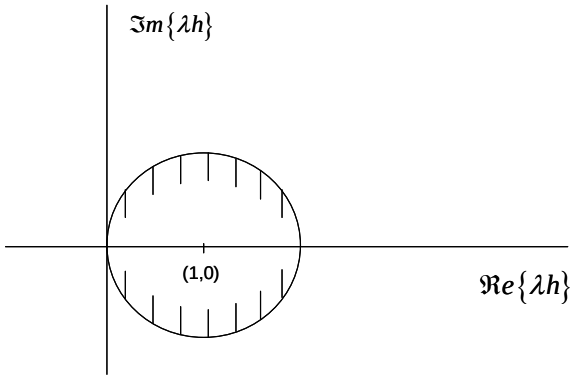


Diagrama de estabilidad para el método Euler implícito.

Lo anterior implica que el método es estable para todo valor de λ en el semiplano izquierdo. Si λ está en el semiplano derecho, el método muestra inestabilidad solamente en el caso de que λ esté dentro del círculo. Si λ se encuentra fuera del círculo unitario mencionado, la fórmula provee una secuencia convergente, aunque la respuesta real crece sin límite.

Método Trapecial

La fórmula recursiva para este método es:

$$y_{k+1} = y_k + \frac{h}{2}\left(f_{k+1} + f_k \right).$$

Aplicada a la ecuación modelo (6) obtenemos para el primer valor del proceso recursivo

$$y_1 = y_0 + \frac{2h}{2}\left(y_1 + y_0 \right)$$

Por lo que

$$y_1 = y_0 \left(\frac{1+\frac{\lambda h}{2}}{1-\frac{\lambda h}{2}} \right)$$

Para el paso *n-ésimo* obtendremos

$$y_n = \left(\frac{2+q}{2-q} \right)^n y_0$$

El requisito de estabilidad de este método para cuando $n \to \infty$ es

$$\left|\frac{2+q}{2-q}\right| \leq 1$$

Lo cual conduce a la desigualdad siguiente

$$\left|\frac{2+u+jv}{2+u-jv}\right| \leq 1 .$$

Simplificando lo anterior obtenemos finalmente

$$4u \leq 0$$

Lo anterior implica una región de estabilidad consistente en el semiplano izquierdo, cuya frontera es el eje imaginario. La fórmula será estable para cualquier valor de λ con $\Re e\{\lambda\} \langle 0$.

El resultado de lo anterior implica que se obtendrán respuestas estables para funciones inestables. Lo anterior no implica un resultado correcto, solamente que cualquier error en el cálculo no crecerá en los pasos subsecuentes.

MÉTODOS DE RUNGE-KUTTA.

Los problemas expuestos para el método de Euler explícito se pueden mejorar ostensiblemente, usando la familia de métodos denominados Runge-Kutta.

Estos métodos obtienen la precisión de una serie de Taylor, pero sin el requerimiento de calcular derivadas de alto orden.

Derivamos las ecuaciones recursivas correspondientes a uno de los métodos más simples de esta familia, pero nos permite entender las bases de los demás. La obtención de los métodos de Runge-Kutta de 2º orden se muestra a continuación. En un apéndice, se muestra la derivación del método de Runge-Kutta de 4º orden, para quién esté interesado, la cual por razones obvias es más complicada que la que se muestra a continuación.

La formulación de esta familia de métodos inicia con la idea de que los métodos que hemos visto hasta ahora, tiene la forma general:

$$y_{i+1} = y_i + \phi(t_i, y_i, h)\, h$$

A la función $\phi(t_i, y_i, h)$ se le conoce como ***función incremento*** y se puede interpretar como la pendiente sobre el intervalo. En general dicha función se puede definir como $\phi = a_1 k_1 + a_2 k_2 + \cdots + a_n k_n$. En este caso los coeficientes a_1, a_2,..., a_n son constantes y los coeficientes ***k*** se definen como sigue:

$$\mathrm{k}_1 = f(t_i, y_i)$$
$$\mathrm{k}_2 = f(t_i + p_1 h, y_i + q_{11} k_1 h)$$
$$\mathrm{k}_3 = f(t_i + p_2 h, y_i + q_{21} k_1 h + q_{22} k_2 h)$$
.
.
.
$$\mathrm{k}_n = f(t_i + p_{n-1} h, y_i + q_{n-1,1} k_1 h + q_{n-1,2} k_2 h + \cdots + q_{n-1,n-1} k_{n-1} h)$$

Los coeficientes ***p*** y ***q*** son constantes.

DERIVACIÓN DE LOS MÉTODOS DE RUNGE-KUTTA DE 2º ORDEN.

Para este caso, la ecuación asociada de acuerdo con la definición general está dada por [ChC]:

$$\mathrm{y}_{\mathrm{i+1}} = y_i + \left(a_1 k_1 + a_2 k_2 \right) h$$

Con

$$k_1 = f(t_i, y_i) \qquad (1)$$

$$k_2 = f(t_i + p_1 h, y_i + q_{11} k_1 h)$$

El problema consiste en determinar los valores de a_1, a_2, p_1 y q_{11}.

Para esto, requerimos comparar la ecuación recursiva general, para el caso de segundo orden, con la serie de Taylor para y_{i+1} alrededor del punto base (t_i, y_i).

La serie de Taylor está dada por

$$y_{i+1} = y_i + f(t_i, y_i)h + \frac{f'(t_i, y_i)}{2!} h^2 \qquad (2)$$

Recordemos que para ahorrar en escritura hemos simplificado la notación en la representación de la función $f(t_i, y_i)$, la cual en estricto sentido tiene la forma $f(t_i, y(t_i))$.

Esto es importante a la hora de evaluar las derivadas de la serie de Taylor; de acuerdo a la regla de la cadena, dicha derivada es por tanto

$$f'(t_i y_i) = \frac{\partial f(t_i, y_i)}{\partial t} + \frac{\partial f(t_i, y_i)}{\partial y} \frac{dy}{dt}$$

De acuerdo con lo anterior y sustituyendo en la ecuación (2) de la serie de Taylor obtenemos

$$y_{i+1} = y_i + f(t_i, y_i)h + \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} \frac{dy}{dt} \right) \frac{h^2}{2!} \qquad (3).$$

La estrategia de los métodos de Runge-Kutta consiste en hacer equivalente la ecuación inicial de este desarrollo con la última ecuación mostrada arriba.

Para llevar a cabo lo anterior, requerimos obtener la expansión en serie de Taylor la función

$$f(t_i + p_1 h, y_i + q_{11} k_1 h)$$

Para llevar a cabo dicha expansión en serie de Taylor, debemos tomar en cuenta la expansión para una función compuesta tiene la forma

$$g(t+r, y+s) = g(t,y) + r\frac{\partial g}{\partial t} + s\frac{\partial g}{\partial y} + \cdots$$

Por lo que para el caso presente tenemos

$$f(t_i + p_1 h, y_i + q_{11} k_1 h) = f(t_i, y_i) + p_1 h \frac{\partial f}{\partial t} + q_{11} k_1 h \frac{\partial f}{\partial y} + O(h^2)$$

Sustituyendo en la ecuación (*) tenemos

$$\mathrm{y}_{i+1} = y_i + a_1 h\, f(t_i, y_i) + a_2 h f(t_i, y_i) + a_2 p_1 h^2 \frac{\partial f}{\partial t} + a_2 q_{11} h^2 f(t_i, y_i)\frac{\partial f}{\partial y} + O(h^3)$$

La cual conduce factorizando términos de *h* a la expresión

$$y_{i+1} = y_i + \left[a_1 f(t_i, y_i) + a_2 f(t_i, y_i)\right] h + \left[a_2 p_1 \frac{\partial f}{\partial t} + a_2 q_{11} f(t_i, y_i)\frac{\partial f}{\partial y}\right] h^2 + O(h^3)$$

Finalmente comparamos la ecuación (3) con esta última para obtener, sin olvidar que $\frac{dy}{dt} = f(\mathrm{t}, \mathrm{y})$:

$$a_1 + a_2 = 1$$

$$a_2 p_1 = \frac{1}{2}$$

$$a_2 q_{11} = \frac{1}{2}$$

Lo anterior constituye un sistema de 3 ecuaciones lineales en 4 incógnitas, a_1, a_2, p_1 y q_{11}, por lo que dicho sistema *no tiene solución única*. Existe por tanto una familia de métodos de Runge-Kutta de 2º orden, cuyos casos más conocidos se muestran a continuación [ChC].

Método de Heun con un solo corrector $a_2=1/2$

Definimos $a_2 = \frac{1}{2}$, por lo que resolviendo las ecuaciones obtenemos

$$a_1 = \frac{1}{2} \qquad p_1 = q_{11} = 1$$

De donde sustituyendo en la definición general, ecuación (1), obtenemos

$$y_{1+1} = y_i + \left(\frac{1}{2} k_1 + \frac{1}{2} k_2 \right) h$$

$$k_1 = f(t_i, y_i)$$

$$k_2 = f(t_i + h, y_i + k_1 h)$$

Método del punto medio: $a_2=1$

En este caso fácilmente que la suposición conduce a $a_1 = 0$ y además $p_1 = q_{11} = \frac{1}{2}$, por lo que la ecuación recursiva resulta

$$y_{i+1} = y_i + k_2 h$$

$$k_1 = f(t_i, y_i)$$

$$k_2 = f(t_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h).$$

Método de Ralston: $a_2=2/3$

Con $a_2 = \frac{2}{3}$ de la primera de las ecuaciones encontramos que $a_1 = 1 - a_2 = \frac{1}{3}$ y con esto obtenemos $p_1 = q_{11} = \frac{1}{2} \cdot \frac{1}{a_2} = \frac{3}{4}$, por lo que la ecuación recursiva para este caso resulta

$$y_{i+1} = y_1 + \left(\frac{1}{3}k_1 + \frac{2}{3}k_2 \right) h$$

$$k_1 = f(t_i, y_i)$$

$$k_2 = f\left(t_i + \frac{3}{4}h, y_i + \frac{3}{4}k_1 h \right)$$

Existen formulaciones para métodos de Runge-Kutta de diversos órdenes. Los más conocidos, a juzgar por la difusión evidente en la literatura al respecto, son los métodos de Runge-Kutta de 4º orden. En esta familia hay varias versiones, a saber, con coeficientes de Runge, con coeficientes de Kutta y con coeficientes de Gill; al menos de acuerdo al conocimiento del autor de estas notas.

Se anexan las ecuaciones correspondientes al más conocido, correspondiente al RK-4º orden, cuya formulación completa está contenida en uno de los apéndices.

MÉTODO DE RUNGE-KUTTA DE 4º ORDEN.

$$\mathrm{y}_{\mathrm{i+1}} = y_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)h$$

$$k_1 = f(t_i, y_i)$$

$$k_2 = f\left(t_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right)$$

$$k_3 = f\left(t_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2 h\right)$$

$$k_4 = f\left(t_i + h, y_i + k_3 h\right)$$

Todos los métodos anteriores se denominan de un solo paso. La característica de estos métodos es que, como puede observarse, la aproximación de la ordenada de un punto $\mathrm{y}_{\mathrm{i+1}}$, se obtiene en base a un solo punto, el anterior, (t_i, y_i). Todos estos casos están asociados con aproximaciones lineales, desde el caso simple de los métodos de Euler y trapecial, discutidos al inicio de este capítulo, hasta método más sofisticados, como la familia de métodos de Runge-Kutta, que consisten en extrapolaciones lineales a través de una recta que resulta del promedio geométrico ponderado de varias rectas; dos en el caso de los métodos de 2º orden, cuatro en el caso de los de 4º orden [ChC].

Existen los métodos multipaso como alternativa, en los cuales se obtiene el valor aproximado de $\mathrm{y}_{\mathrm{i+1}}$ a través de la extrapolación de interpolantes de grado mayor que uno. Para esto se requieren más de un punto conocido para obtener dicha aproximación. En este caso el inicio de estos métodos requiere el uso de un método de un solo paso para encontrar los puntos necesarios para iniciar el proceso recursivo del método multipaso, a diferencia de los que hemos discutido aquí, en los cuales se inicia con las condiciones iniciales.

Además de esta alternativa, existen esquemas alternativos más sofisticados, que utilizan algoritmos que obtienen las ordenadas y_{i+1} a través de un proceso de dos etapas, una de las cuales consiste de un proceso iterativo. La primera etapa obtiene un estimado del valor de la ordenada y_{i+1}^{p} que se denomina ***predictor***; la segunda etapa obtiene, iniciando con el valor anterior y mediante un proceso iterativo valores sucesivos de la variable, y_{i+1}^{c}, denominados ***correrctor***, caracterizados con una precisión definida por el usuario. A estos métodos se les conoce como ***métodos predictor-corrector***.

Se puede observar que el método asociado con la etapa de la obtención del predictor debe ser un *método explícito*, mientras el de la etapa del corrector debe ser una *fórmula implícita*.

La extensión de las notas presentes no hace posible cubrir estos importantes métodos, pero al final se proporciona una muestra bibliográfica, que aunque no es obviamente exhaustiva, constituye una fuente modesta de información.

APENDICES

SERIES INFINITAS.

Definiciones y notación.

A la suma de una sucesión de términos se denomina SERIE y el valor de dicha suma, si es que tiene alguno, se define como

$$S = \lim_{n\to\infty} S_n \, .$$

Un ejemplo de serie infinita, denominada así debido a que dicha sucesión es infinita, es la denominada serie geométrica, la cual se obtiene a partir de un térmno inicial multiplicado por una cantidad constante, p. ej. $a + ar + ar^2 + ar^3 + \cdots + ar^{n-1} + \cdots$. En este caso la cantidad inicial *a* es multiplicada por la cantidad constante *r* para obtener dicha serie infinita.

En general una serie infinita significa una expresión de la forma

$$a_1 + a_2 + a_3 + \cdots + a_n + \cdots,$$

donde las a_n son números o funciones dadas por alguna regla o fórmula. Los tres puntos significan que la serie nunca termina. Si se tiene duda de cómo es la regla usada en la formación e la serie, el término general o término *n*-ésimo deberá expresarse, p. ej.

$$1^2 + 2^2 + \cdots + n^2 + \cdots$$

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

También usaremos formas abreviadas para denotar las series, p. ej. las series anteriores, la forma abreviada son

$$\sum_{n=1}^{\infty} n^2$$

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1} x^n}{(n-1)!} \, .$$

Las aplicaciones de las series infinitas son muchas, pero mencionamos como lo más importante para nosotros en este momentos, su uso en la solución de problemas matemáticos que no pueden resolverse en términos de funciones elementales (potencias, raíces, funciones trigonométricas y sus inversas, logaritmos y exponenciales y combinaciones de estos), o en caso de que puedan resolverse, es muy complicado trabajar con ellos. En estos casos encontramos una respuesta en función de una serie y usamos los términos requeridos de acuerdo a la presición deseada. Las ecuaciones diferenciales son resueltas en muchas ocasiones en función de series infinitas. Una integral definida, por ejemplo, $\int_0^{0.1} e^{-x^2}\,dx$, para la cual no hay solución en términos de funciones elementales, se puede resolver su expandiendo su integrando en una serie e integrando término a término dicha serie.

SERIES CONVERGENTES Y DIVERGENTES.

Existen series caracterizadas por tener una *suma finita*. Pero también existen series cuya suma no es finita. Si la serie tiene una suma finita, se denomina *serie convergente*, mientras que en caso contrario se denomina *serie divergente*.

Es muy importante saber si una serie es o no convergente. Pueden ocurrir cosas raras si tratamos de aplicar algebra ordinaria a una serie divergente. Supongamos la siguiente serie: $S = 1+2+4+8+16+\cdots$. Entonces $2S = 2+4+16+\cdots = S-1$, y de aquí podríamos concluir que $S=-1$, lo cual obviamente no tiene sentido.

En este punto podríamos reconocer que la serie

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots$$

es divergente

Mientras la serie

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\cdots,$$

es convergente como se muestra, pero se puede tener la suma que se quiera reacomodando el orden de los términos!, como se podrá ver más adelante. Lo anterior

nos muestra la importancia de trabajar con series que sean convergentes. Esto implica que nuestro interés se centrará en series que cumplan la condición $S = \lim_{n\to\infty} S_n$.

De aquí podemos decir que si la suma parcial S_n (la suma de los *n* primeros términos) tiende a un límite, entonces la serie es *convergente*. En caso contrario se dice que la serie es *divergente*. Al valor límite de la serie *S* se denomina *suma de la serie*. Por otro lado a la diferencia $R_n = S - S_n$ se le denomina *residuo*. De la definición mostrada antes tendremos

$$\lim_{n\to\infty} R_n = \lim_{n\to\infty}(S - S_n) = S - S = 0.$$

PRUEBAS DE CONVERGENCIA.

Primero discutimos la denominada *prueba preliminar*. En muchos casos debemos intentar aplicar esta simple prueba antes de aplicar un método más complicado, aunque no en todos los casos es útil.

Prueba preliminar. Si los términos de una serie infinita no tienden a cero, esto es si $\lim_{n\leftarrow\infty} a_n \neq 0$, la serie diverge. Si $\lim_{n\leftarrow\infty} a_n = 0$, debemos de probar por otro método más avanzado. Es importante hacer notar que esta prueba preliminar resulta útil para eliminar pruebas complicadas en series notoriamente divergentes, pero también hay que notar que esta misma prueba *nunca nos dice que la serie converge*, es decir, *no nos* dice que la serie converge si $a_n \to 0$, y de hecho a menudo es el caso. Un ejemplo de lo anterior lo constituye la denominada serie armónica $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots$, en la que el *n*-ésimo término tiende a cero, pero se puede demostrar que la serie $\sum_{n=1}^{\infty} 1/n$, es divergente. Por otro lado en la serie $\frac{1}{2}+\frac{2}{3}+\frac{3}{4}+\frac{4}{5}+\cdots$, los términos tienden a 1, y de acuerdo a la prueba preliminar la serie diverge y no hay caso hacer más pruebas.

PRUEBAS PARA CONVERGENCIA DE SERIES DE TERMINOS POSISTIVOS.

CONVERGENCIA ABSOLUTA.

Consideramos ahora cuatro pruebas útiles para probar la convergencia de series que contienen únicamente términos positivos. Si la serie contiene términos negativos, aun consideraremos la serie que resulta de todos los términos positivos , es decir, la serie cuyos términos son los valores absolutos de la serie original. Si la nueva serie converge, llamamos a la serie original *absolutamente convergente*. Se puede probar que si una serie converge absolutamente, entonces es convergente. Lo anterior significa que si la serie de valores absolutos converge, la serie aun es convergente cuando ponemos los signos negativos en los términos que lo son, aunque el valor de la suma sea diferente.

PRUEBA DE COMPARACION.

Esta prueba consiste de dos partes, que llamaremos (a) y (b).

(a). Sea $m_1 + m_2 + m_3 + m_4 + \cdots$, una serie de términos positivos, la cual sabemos que converge. Entonces la serie que queremos probar: $a_1 + a_2 + a_3 + a_4 + \cdots$, es absolutamente convergente si $|a_n| \leq m_n$, para todo *n* a partir de un punto en adelante (ya sea a partir del 2^0 o el millonésimo término), esto es, si el valor absoluto de cada término de la serie *a* no es mayor que el correspondiente término de la serie *m*.

(b). Sea $d_1 + d_2 + d_3 + d_4 + \cdots$, una serie de términos positivos que sabemos que diverge. Entonces la serie $|a_1| + |a_2| + |a_3| + |a_4| + \cdots$ diverge si $|a_n| \geq d_n$, para todo *n* a partir de un punto en adelante.

Es importante hacer notar que ni $|a_n| \geq m_n$, ni $|a_n| \leq d_n$ nos dice nada. Es decir, que si una serie tiene términos mayores que aquellos de una serie convergente, puede aún converger o también podría divergir; debemos hacer más pruebas. Por otro lado, si una serie tiene términos más pequeños que los de una serie divergente, aun puede ser divergente o convergente.

Ejemplo. Probemos si la serie $\sum_{n=1}^{\infty}\frac{1}{n!}=1+\frac{1}{2}+\frac{1}{6}+\frac{1}{24}+\cdots$ converge. Como comparación usaremos la serie geométrica $\sum_{n=1}^{\infty}\frac{1}{2^n}=\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots$, que sabemos converge.

Notar que no nos importan los primeros términos en una serie (de hecho, cualquier número finito de términos), dado que estos pueden afectar el valor de la suma, pero no su convergencia. Cuando preguntamos si una serie converge o no, estamos preguntando qué ocurre cuando agregamos más y más términos para *n* más y más grandes. ¿La suma se incrementa indefinidamente o se aproxima a un límite?. Que los primeros cinco, cien o un millón de términos no tiene efecto sobre si, eventualmente la suma se incrementa indefinidamente o se aproxima a un límite. Consecuentemente, a menudo ignoramos los primeros términos en la prueba de convergencia de una serie.

En el ejemplo presente, los términos de $\sum_{n=1}^{\infty}\frac{1}{n!}$ son más pequeños que los correspondientes de $\sum_{n=1}^{\infty}\frac{1}{2^n}$, para todo $n\rangle 3$, y como sabemos que la serie geométrica converge, entonces concluimos que $\sum_{n=1}^{\infty}\frac{1}{n!}$, converge también.

PRUEBA POR INTEGRACION.

Podemos usar esta prueba cuando los términos de la serie son positivos y no se incrementan, esto es, cuando $a_{n+1}\le a_n$. Recordemos que podemos ignorar un número finito de términos de la serie; aun así el resto aún puede usarse, aun cuando la condición $a_{n+1}\le a_n$ no se cumpla para un número finito de términos. Para aplicar esta prueba, pensamos en a_n como una función de la variable *n*, olvidándonos del significado atribuido a *n;* de esta forma, permitimos que tome todos los valores y no nada más valores enteros. La prueba puede enunciarse como sigue:

Si $0\le a_{n+1}\le a_n$ *para* $n\rangle N$, *entonces* $\sum^{\infty} a_n$ *converge si* $\int^{\infty} a_n\, dn$ *es finita y diverge si la integral es infinita.* Es importante notar que la integral se evaluará *solamente* en le límite superior; no se requiere límite inferior.

Para entender esta prueba, imaginemos una gráfica de a_n como función de *n*. Supongamos por ejemplo la serie armónica $\sum_{n=1}^{\infty} 1/n$; consideramos la gráfica de la función $y = 1/n$, similar a la que se muestra en las figura 1 y 2, donde suponemos que *n* toma todos los valores, no únicamente valores enteros. Entonces los valores de *y* en la gráfica en $n = 1,2,3,\cdots$, son términos de las series. En las figuras 1 y 2, las áreas de los rectángulos son simplemente los términos de la serie. Note que en la figura 1, la parte superior del rectángulo está por encima de la curva, de tal manera que el área del los rectángulos es mayor que el área debajo de la curva. Por otro lado, en la figura 2 los rectángulos están por debajo de la curva, por lo que su área es menor que el área debajo de la curva. El área de los rectángulos son los términos de la serie simplemente, mientras que el área bajo la curva es una integral de $y\,dn$ o $a_n\,dn$. El límite superior de la integral es ∞ y el límite inferior puede hacerse que corresponda a cualquier término de la serie con que se quiera arrancar.

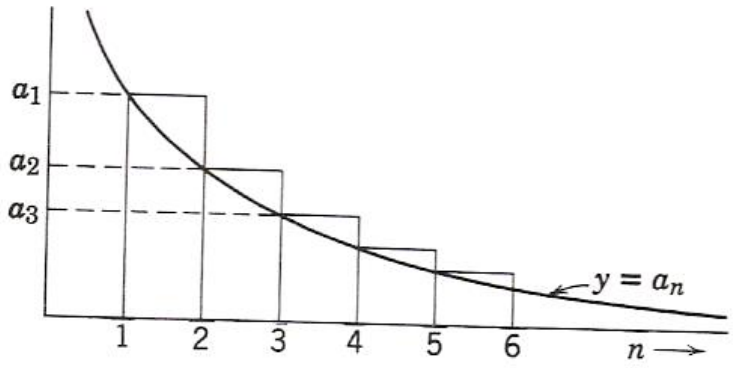


Figura 1. Prueba de convergencia por integración.

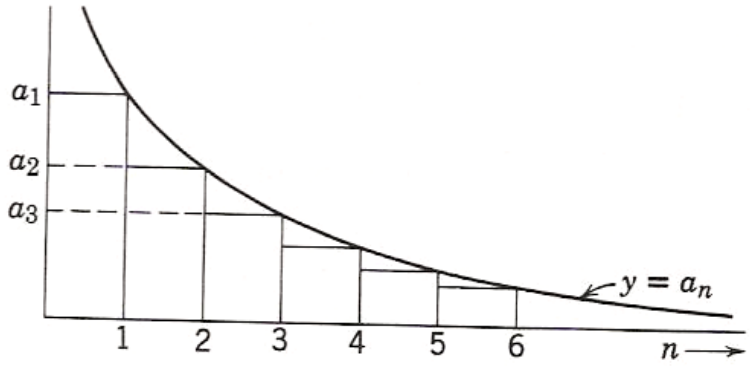


Figura 2. Prueba de convergencia por integración.

Por ejemplo, de la figura 1, $\int_3^\infty a_n\, dn$ es menor que la suma de la serie de a_3 en adelante, pero (figura 2) mayor que la suma de la serie de a_4 en adelante. Si la integral es finita, entonces la suma de la serie de a_4 en adelante es finita, esto es, la serie converge. Note nuevamente que los términos del inicio de la serie no influyen en la convergencia. Por el otro lado, si la integral es infinita, entonces la suma de la serie de a_3 en adelante es infinita y la serie diverge. Dado que los términos iniciales no son d interés, entonces no hace falta el límite inferior de la integral y evaluamos simplemente $\int^\infty a_n\, dn$.

Probemos la serie armónica: $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots$

Usando la prueba de la integración evaluamos

$$\int^\infty \frac{1}{n}\, dn = \ln n\Big|^\infty = \infty .$$

Dado que la integral es infinita, la serie es divergente.

PRUEBA DEL COCIENTE.

La integración de $a_n\ dn$ no siempre es fácil, por lo que podemos considerar otra prueba que puede resolver muchos casos que no pueden resolverse por la prueba de integración.

Empezamos por definir los siguientes términos:

$$\rho_n = \left| \frac{a_{n+1}}{a_n} \right| ,$$

$$\rho = \lim_{n\to\infty} \rho_n .$$

Entonces la prueba del cociente se puede enunciar como sigue:

Si:

$\rho < 1,$ ***entonces la serie converge***

$\rho = 1,$ ***usar otra prueba (esta no es concluyente)***

$\rho > 1,$ ***la serie diverge***.

Tomemos como ejemplo la serie

$$1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots$$

Usando las definiciones anteriores tenemos

$$\rho_n = \left| \frac{1}{(n+1)!} \div \frac{1}{n!} \right|$$

$$= \frac{n!}{(n+1)!} = \frac{n(n-1)\cdots 3\cdot 2\cdot 1}{(n+1)(n)(n-1)\cdots 3\cdot 2\cdot 1} = \frac{1}{n+1}$$

De donde tenemos

$$\rho = \lim_{n\to\infty} \rho_n = \lim_{n\to\infty} \frac{1}{n+1} = 0\,.$$

Dado que $\rho < 1$, la serie converge.

Otro ejemplo muy ilustrativo de la aplicación de este método de prueba ocurre con la serie armónica

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots$$

En este caso encontramos que

$$\rho_n = \left| \frac{1}{n+1} \div \frac{1}{n} \right| = \frac{n}{n+1},$$

$$\rho = \lim_{n\to\infty} \frac{n}{n+1} = \lim_{n\to\infty} \frac{1}{1+\frac{1}{n}} = 1.$$

De acuerdo al enunciado del método, éste no nos dice nada y debemos usar una prueba diferente. Es importante alertar que $\rho_n = \frac{n}{n+1}$ es siempre menor que 1. Se debe tener cuidado de no confundir este cociente con ρ y concluir incorrectamente que la serie converge. De hecho, como mostramos en la prueba de la integración, la serie es divergente.

PRUEBA ESPECIAL DE COMPARACION.

Esta prueba tiene dos partes: (a) prueba de convergencia, y (b) prueba de divergencia.

(a) Si $\sum_{n=1}^{\infty} b_n$ es una serie convergente de términos positivos y $a_n \geq 0$ y a_n/b_n tiende a un límite (finito), entonces $\sum_{n=1}^{\infty} a_n$ converge.

(b) Si $\sum_{n=1}^{\infty} d_n$ es una serie divergente de términos positivos y además $a_n \geq 0$ y a_n/d_n tiende a un límite mayor que 0 (o tiende a +∞), entonces $\sum_{n=1}^{\infty} a_n$ diverge.

Tomemos la serie

$$\sum_{n=2}^{\infty} \frac{3^n - n^3}{n^5 - 5n^2} \quad , \quad \text{para ejemplificar este método.}$$

Primero debemos decidir que término es más importante a medida que *n*→∞; ¿es 3^n ó bien n^3?. Podemos comparar sus logaritmos para indagar la respuesta, dado que *ln* N y N crecen o decrecen juntos. Ahora sabemos que $\ln 3^n = n \cdot \ln 3$, y $\ln n^3 = 3 \cdot \ln n$, pero *ln n* es más pequeño que *n*, por lo que para *n* grande tenemos $n \cdot \ln 3 \rangle 3$ y $3^n \rangle n^3$. (Puede calcular p. ej. $100^3 = 10^6$, y $3^{100} \rangle 5 \cdot 10^{47}$. El denominador de la serie entonces será aproximadamente n^5. De lo anterior vemos que la serie para comparación será $\sum_{n=2}^{\infty} \frac{3^n}{n^5}$.

Si hacemos uso de la prueba del cociente, vemos que esta serie es divergente, como se muestra a continuación:

$$\rho_n = \left| \frac{3^{n+1}}{(n+1)!} \div \frac{3^n}{n^5} \right| = \left| \frac{3^{n+1}/3^n}{(n+1)^5/n^5} \right| = \left| \frac{3}{(1+1/n)^5 \cdot n^5/n^5} \right| = \left| \frac{3}{(1+1/n)^5} \right|$$

De donde

$\rho = \lim_{n\to\infty} \frac{3}{\left(1+\frac{1}{n}\right)^5} = 3$, por lo que vemos que la serie es divergente.

Ahora por la prueba (b)

$$\lim_{n\to\infty} \frac{a_n}{d_n} = \lim_{n\to\infty} \left(\frac{3^n - n^3}{n^5 - 5n^2} \div \frac{3^n}{n^5} \right) = \lim_{n\to\infty} \frac{1-\frac{n^3}{3^n}}{1-\frac{5}{n^3}} = 1$$

La cual es mayor que cero, por lo que concluimos que la serie diverge.

SERIES ALTERNANTES.

Hasta ahora hemos considerado series de términos positivos y consideramos ahora un caso importante de series cuyos términos tiene signos mixtos; una serie alternante e una serie cuyos términos son alternativamente positivos y negativos. Por ejemplo la serie

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots + \frac{(-1)^{n+1}}{n} + \cdots$$

es una serie alternante. Dos preguntas esenciales en el caso de series alternantes son: ¿converge la serie?, ¿converge absolutamente (esto es, cuando hacemos todos los signos positivos)?. Consideremos la segunda pregunta primero. Para el ejemplo anterior, la serie de valores absolutos es

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} - \cdots + \frac{1}{n} + \cdots$$

Esta es la serie armónica, la cual sabemos que diverge. Entonces decimos que la serie alternante no es absolutamente convergente. La siguiente pregunta es si esta serie converge tal como está; si hubiera convergido absolutamente, entonces no habrá

necesidad de hacernos la pregunta anterior, dado que, se puede demostrar, una serie absolutamente convergente, es convergente también. Sin embargo, una serie absolutamente divergente, puede o no ser convergente; deberemos probar por otros métodos entonces. Para una serie alternante la prueba es muy simple y se enuncia como sigue:

Prueba para una Serie Alternante. Una serie alternante converge si el valor absoluto de los términos decrece monótonamente (es decir, de manera permanente) a cero, esto es, si $|a_{n+1}| \leq |a_n|$ y además $\lim_{n\to\infty} a_n = 0$.

En nuestro ejemplo $\frac{1}{n+1} < \frac{1}{n}$, y $\lim_{n\leftarrow\infty} \frac{1}{n} = 0$, por lo tanto la serie converge.

ALGUNAS PROPIEDADES UTILES DE LAS SERIES.

Enumeramos algunas propiedades de las series:

1. La convergencia o divergencia de una serie no se afecta al multiplicar cada término de la serie por la misma constante. Tampoco se afecta al cambiar un número finito de términos (por ejemplo, omitiendo pocos términos del inicio).

2. Dos series convergentes $\sum_{n=1}^{\infty} a_n$ y $\sum_{n=1}^{\infty} b_n$ pueden sumarse (o restarse) término a término. (Sumando "término a término" significa que el *n*-ésimo término de la suma es $a_n + b_n$). La serie resultante es convergente, y su suma es obtenida sumando (o restando) las sumas de las series dadas.

3. Los términos de una *serie absolutamente convergente* puede reacomodarse en cualquier orden sin afectar la convergencia o la suma de la serie. Esto sin embargo *no es cierto* en el caso de *series condicionalmente convergentes*.

SERIES DE POTENCIAS.

Existen series cuyos términos no son constantes, sino funciones de *x*; hay muchos ejemplos de dichas series, pero nosotros únicamente consideraremos aquellas series en las que el *n*-ésimo término es igual al producto de una constante por x^n o bien, una constante multiplicando a $(x-a)$. donde *a* es una constante. Dichas series se conocen como *series de potencias*, porque sus términos son múltiplos de *x* ó $(x-a)$. Existen otras series cuyos términos pueden ser *senos* y *cosenos* (series de Fourier) ó bien polinomios u otro tipo de funciones (Legendre, Bessel, p ej.) y que son sumamente útiles en una gran cantidad de aplicaciones. Aquí estudiaremos únicamente las series de potencias y en particular la serie de Taylor.

Por definición, una serie de potencia de potencias tiene la forma

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$$

O bien

$$\sum_{n=0}^{\infty} a_n (x-a)^n = a_0 + a_1 (x-a) + a_2 (x-a)^2 + a_3 (x-a)^3 + \cdots$$

Los coeficientes a_n son constantes.

A continuación unos ejemplos:

(a) $1-\frac{x}{2}+\frac{x^2}{4}-\frac{x^3}{8}+\cdots+\frac{(-x)^n}{2^n}+\cdots,$

(b) $x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\cdots+\frac{(-1)^{n+1}x^n}{n}+\cdots,$

(c) $x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots+\frac{(-1)^{n+1}x^{2n-1}}{(2n-1)!}+\cdots$

(d) $1+\frac{(x+2)}{\sqrt{2}}+\frac{(x+2)^2}{\sqrt{3}}+\cdots+\frac{(x+2)^n}{\sqrt{n+1}}+\cdots.$

La convergencia de estas series depende de los valores que se consideren para la variable *x*. A menudo se usa la prueba del cociente para encontrar los valores de *x* para los cuales converge la serie. Probemos las series ejemplificadas arriba.

1. Para la serie mostrada en (a), tenemos

$$\rho_n=\left|\frac{(-x)^{n+1}}{2^{n+1}}\div\frac{(-x)^n}{2^n}\right|=\left|\frac{x}{2}\right|,$$

de donde obtenemos

$$\rho=\left|\frac{x}{2}\right|.$$

La serie converge para $\rho\langle 1$, esto es, $\left|\frac{x}{2}\right|\langle 1$ ó bien $|x|\langle 2$, y diverge para $\rho\rangle 1$ por lo que podemos mostrar fácilmente que esto implica $|x|\rangle 2$. Esto significa que para cualquier valor de *x* comprendido entre -2 y 2, la serie converge; lo anterior excluye los extremos de la recta numérica comprendida entre estos valores, por lo que debemos indagar que ocurre con la convergencia de la serie cuando *x* toma los valores extremos, es decir -2 y 2. Si $x=2$ vemos que la serie es: $1-1+1-1+\cdots$, la cual es divergente. Cuando $x=-2$ la serie es: $1+1+1+1+\cdots$, por lo que la serie es divergente. De acuerdo a estos resultados concluimos que el intervalo de convergencia se establece como: $-2\langle x\langle 2$.

2. Para la serie del caso (b) encontramos

$$\rho_n = \left| \frac{x^{n+1}}{n+1} \div \frac{x^n}{n} \right| = \left| \frac{nx}{n+1} \right|,$$

$$\rho = \lim_{n\to\infty} \left| \frac{nx}{n+1} \right| = |x| .$$

La serie converge para $|x| < 1$. De nuevo debemos explorar los puntos extremos del intervalo de convergencia, $x = 1$ y $x = -1$. Para $x = 1$ la serie es: $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$; esta es una serie armónica alternante y se puede demostrar que es convergente.

Para $x = -1$ la serie es: $-1 - \frac{1}{2} - \frac{1}{3} - \frac{1}{4} - \cdots$; observamos que esta es la serie armónica multiplicada por -1 y es divergente. Por lo anterior vemos que el intervalo de convergencia es $-1 < x \le 1$.

3. Observamos que para la serie del caso (c), el valor absoluto del *n*-ésimo término es $\left| \frac{x^{2n-1}}{(2n-1)!} \right|$. De acuerdo con esto, el término $n+1$ se obtiene sustituyendo *n* por $n+1$ y el valor absoluto del término $n+1$ es $\left| \frac{x^{2n+1}}{(2n+1)!} \right|$.

Por lo anterior tenemos

$$\rho_n = \left| \frac{x^{2n+1}}{(2n+1)!} \div \frac{x^{2n-1}}{(2n-1)!} \right| = \left| \frac{x^2}{(2n+1)(2n)} \right|,$$

$$\rho = \lim_{n\to\infty} \left| \frac{x^2}{(2n+1)(2n)} \right| = 0 .$$

Dado que $\rho < 1$, para todos los valores de *x*, esta serie converge para todos los valores de *x*.

4. Finalmente para el caso (d) tenemos

$$\rho_n = \left| \frac{(x+2)^{n+1}}{\sqrt{n+2}} \div \frac{(x+2)^n}{\sqrt{n+1}} \right|,$$

$$\rho = \lim_{n\to\infty} \left| (x+2) \frac{\sqrt{n+1}}{\sqrt{n+2}} \right| = |x+2| .$$

Esta serie converge para $|x+2| < 1$; esto es, $-1 < (x+2) < 1$, ó bien $-3 < x < -1$.

Para $x = -3$, la serie es $1 - \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} - \frac{1}{\sqrt{4}} + \cdots$, que es convergente por la prueba de las series alternantes. Para $x = -1$, la serie es $1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots = \sum_{n=0}^{\infty} \frac{1}{\sqrt{n+1}}$ la cual es divergente por la misma prueba de las series alternantes.

De lo anterior concluimos que la serie converge para $-3 \leq x < 1$.

TEOREMAS ACERCA DE SERIES DE POTENCIAS.

El valor de la suma de una serie depende del valor que tome la variable *x* por lo que denotamos por *S(x)* al valor de dicha suma. Por esto las series de potencias definen una función de *x*, que llamaremos *S(x)*. En este sentido decimos que la serie converge a la función *S(x)*. Aquí la idea es obtener la función a partir de una serie dada. Estaremos interesados en obtener una serie que converja a la función dada. Las series de potencias son muy útiles porque las podemos manejar muy parecido a como manejamos los polinomios. A continuación listamos cuatro teoremas de mucha utilidad en la obtención y aplicación de dichas series.

1. Una serie de potencias puede diferenciarse o integrarse término a término; la serie resultante converge a la derivada o la integral de la función representada por la serie original dentro del mismo intervalo de convergencia de la serie original (esto es, no necesariamente en los puntos extremos del intervalo).
2. Dos series de potencias pueden sumarse, restarse o multiplicarse; la serie resultante converge al menos en el intervalo común de convergencia. Se pueden dividir dos series si el denominador de la serie no es cero en $x = 0$, o si siendo cero se puede cancelar por el numerador (como por ejemplo en $\frac{sen\, x}{x}$). La serie resultante tendrá algún intervalo de convergencia.
3. Una serie puede ser sustituida en otra serie, siempre y cuando los valores de la serie sustituida están en le intervalo de convergencia de la otra serie.
4. La serie de potencias de una función es *única*, esto es, existe únicamente una serie de la forma $\sum_{n=0}^{\infty} a_n x^n$ que converge a la función dada.

EXPANSION DE FUNCIONES. SERIE DE TAYLOR.

Una de las aplicaciones más importantes de las series de potencias es la representación de funciones. Para ilustrar este punto usaremos la función *sen x*, para la cual suponemos que existe una serie de potencias. Lo anterior implica que buscamos los coeficientes *a*'s de la serie:

$$sen\, x = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots ,$$

que haga que la identidad anterior sea correcta. Dado que el intervalo de una serie de potencias contiene el origen, debe cumplirse en $x=0$. Si sustituimos $x=0$ en la expresión anterior obtenemos que $a_0 = 0$ debido a que todos los términos, excepto el primero, se hacen cero. De la misma forma, si evaluamos la primera derivada de la serie, $\cos x = a_1 + 2a_2 x + 3a_3 x^2 + \cdots$, obtenemos que $a_1 = 1$. Si volvemos a diferenciar obtenemos la expresión: $-sen\, x = 2a_2 + 3 \cdot 2a_3 x + 4 \cdot 3a_4 x^2 + \cdots$, la cual evaluada a su vez en $x = 0$, nos conduce a la igualdad $0 = 2a_2$. Continuando el proceso, obteniendo la siguiente derivada y evaluando la expresión resultante en $x = 0$, obtenemos

$-\cos x = 3\cdot 2a_3 + 4\cdot 3\cdot 2a_4 x + \cdots$, que evaluada en $x = 0$ nos proporciona el valor de $a_3 = -\frac{1}{3!}$. De la misma manera, derivando de nuevo obtenemos

$$sen\, x = 4\cdot 3\cdot 2a_4 + 5\cdot 4\cdot 3\cdot 2a_5 x + \cdots ,$$

que evaluada en $x = 0$ nos conduce a $0 = 5!a_5$. Cuando sustituimos los valores obtenidos, nos resulta la conocida serie de la función *seno*:

$$sen\, x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

Las series obtenidas de esta forma se denominan series de Maclaurin o series de Taylor alrededor del origen. Una serie de Taylor en general significa una serie de potencias de $(x-c)$, donde *c* es alguna constante y se encuentra usando $(x-c)$ en lugar de *x*. La obtención de los coeficientes *a* de la serie, se lleva a cabo por el mismo procedimiento mostrado en la obtención de la función *seno*, sustituyendo $x = c$ en la función y sus derivadas, en lugar de $x = 0$.

Efectuemos este procedimiento para una función general $f(x)$, suponiendo que existe la serie de Taylor de dicha función:

$$f(x) = a_0 + a_1(x-c) + a_2(x-c)^2 + a_3(x-c)^3 + a_4(x-c)^4 + \cdots + a_n(x-c)^n + \cdots$$

$$f'(x) = a_1 + 2a_2(x-c) + 3a_3(x-c)^2 + 4a_4(x-c)^3 + \cdots + na_n(x-c)^{n-1} + \cdots$$

$$f''(x) = 2a_2 + 3\cdot 2a_3(x-c) + 4\cdot 3a_4(x-c)^2 + \cdots + n(n-1)a_n(x-c)^{n-2} + \cdots$$

$$f'''(x) = 3!a_3 + 4\cdot 3\cdot 2a_4(x-c) + \cdots + n(n-1)(n-2)a_n(x-c)^{n-3} + \cdots$$

.

$$f^{(n)}(x) = n(n-1)(n-2)\cdots 1a_n + \Re_n ,$$

Donde el término $\Re_n$ representa los términos de derivada superior que no se muestran, y está relacionado con un elemento de la serie que definimos más adelante. Ahora evaluamos las expresiones anteriores en $x = c$ y obtenemos:

$$f(c) = a_0 \quad , f'(c) = a_1 \quad , f''(c) = 2a_2$$

$$f'''(c) = 3!a_{3,} \qquad \cdots \quad f^{(n)}(c) = n!a_n \quad ,$$

Por lo que podemos escribir entonces la serie de Taylor para $f(x)$ alrededor de $x = c$:

$$f(x) = f(c) + (x-c)f'(c) + \frac{1}{2!}(x-c)^2 f''(c) + \cdots + \frac{1}{n!}(x-c)^n f^{(n)}(c) + \cdots$$

La serie de Maclaurin para $f(x)$ es la serie de Taylor alrededor del origen. Haciendo $x = 0$ en la expresión anterior obtenemos la serie de Maclaurin para $f(x)$:

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

En general, y de manera no formal, podemos decir que una función $f(x)$ puede expanderse alrededor de un punto *c*, que denominaremos punto base, en una serie de Taylor, siempre que exista dicha serie, como

$$f(x) = f(c) + (x-c)f'(c) + \frac{1}{2!}(x-c)^2 f''(c) + \cdots + \frac{1}{n!}(x-c)^n f^{(n)}(c) + \Re_n(x)$$

En la que el término

$$\Re_n(x) = \frac{(x-c)^{n+1} f^{(n+1)}(\xi)}{(n+1)\ !} \qquad \xi \in [x, c],$$

se denomina el residuo de orden *n* de la serie, y la expresión presentada se denomina la forma de Lagrange del residuo.

La serie de Taylor es de suma importancia en muchas aplicaciones, sin embargo esto implica el uso de derivadas de alto orden, dependiendo del número de términos que se desee, y estas derivadas no siempre son simples de encontrar; considere por ejemplo el caso de la función $e^{\tan x}$, por citar un ejemplo, y se verá que encontrar sus derivadas de alto orden no es trivial. En estos casos hay una serie de métodos que resultan de suma utilidad para obtener la serie de Taylor, a partir de la combinación de series de funciones sencillas, que nos permiten, al combinarlas, encontrar series de funciones más complejas.

Ilustraremos con ejemplos una variedad de métodos para obtener series de funciones complicadas. Para esto establecemos las series correspondientes a funciones sencillas que se muestran a continuación.

1. $sen\ x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad \forall\ x$

2. $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \qquad \forall\ x$

3. $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \qquad \forall\ x$

4. $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad -1 < x \le 1$

5. $(1+x)^p = 1 + px + \frac{p(p-1)}{2!}x^2 + \frac{p(p-1)(p-2)}{3!}x^3 + \cdots \qquad |x| < 1.$

Esta última es la serie binomial; *p* es cualquier número real, positivo ó negativo.

Ejemplo 1. Encontrar la serie de la función $(x+1)\,sen\,x$. En este caso usamos la serie para la función *seno* y la multiplicamos por el 1er término:

$$(x+1)\,sen\,x = (x+1)\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots\right) = x + x^2 - \frac{x^3}{3!} - \frac{x^4}{4!} + \cdots$$

Ejemplo 2. Encontrar la serie de $e^x \cos x$, combinando las series 2 y 3 mostradas arriba:

$$e^x \cos x = \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots\right)\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right)$$

$$= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$
$$- \frac{x^2}{2!} - \frac{x^3}{2!} - \frac{x^4}{2!\,2!} \cdots$$
$$+ \frac{x^4}{4!} \cdots$$

$$= 1 + x + 0x^2 - \frac{x^3}{3} - \frac{x^4}{6} \cdots = 1 + x - \frac{x^3}{3} - \frac{x^4}{6} \cdots$$

Ejemplo 3. Encontramos la serie que resulta de un cociente de funciones, $\frac{1}{x}\ln(1+x)$.

Usando la 4ª serie mostrada en la lista obtenemos

$$\frac{1}{x}\ln(1+x) = \frac{1}{x}\left(x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots\right)$$
$$= 1 - \frac{x}{2} + \frac{x^2}{3} - \frac{x^3}{4} + \cdots$$

Ejemplo 4. Usamos la expresión 5 como una generalización del teorema del binomio $(a+b)^n$, con $a = 1, \quad b = x, \quad n = p$;la diferencia es que en este caso *p* puede ser negativo o fraccional, y en estos casos, la expansión es una serie infinita. Esta serie converge para $|x| < 1$, como se puede probar usando la prueba del cociente. Usaremos lo anterior para obtener

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

Otros casos interesantes lo constituyen la sustitución de polinomios o series, por la variable de otra serie. El siguiente ejemplo ilustra lo anterior.

Ejemplo 5. Encontrar la serie de la función e^{-x^2}. Usando la serie de e^x obtenemos

$$e^{-x^2} = 1 - x^2 + \frac{\left(-x^2\right)^2}{2!} + \frac{\left(-x^2\right)^3}{3!} + \cdots$$
$$= 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \cdots$$

CASOS AVANZADOS

DE

INTEGRACIÓN NUMÉRICA

DERIVACIÓN DE LOS MÉTODOS DE RUNGE-KUTTA 4º ORDEN

En este apéndice se muestran para el interesado, algunos aspectos más avanzados de los métodos de integración numérica que se discuten en el curso. El primer caso consiste en la obtención del método de Runge-Kutta de 4o orden. Se trata de las versiones más conocidas, denominadas coeficientes de Runge. Existe al menos otra versión, obtenida en fecha posterior a las versiones mostradas en este apéndice, y por ende en el curso, que se denomina coeficientes de Gill. El interesado puede encontrar literatura al respecto, aunque no sea tema tan difundido como el que se muestra.

Los métodos de RK se derivan a partir de la serie de Taylor. La forma general de la ecuación usada para formular el método de RK es

$$y_{n+1} = y_n + \Delta y_n \qquad , \quad \Delta y_n = \phi(t_n, y_n)h$$

Δy_n es la función incremental que puede interpretarse como la pendiente representativa del intervalo.

En general

$$\phi = a_1 k_1 + a_2 k_2 + ... + a_n k_n$$

Las *a*'s son constantes y las *k*'s se definen como:

$$k_1 = f(t_n, y_n)$$
$$k_2 = f(t_n + p_1 h, y_n + q_{11} k_1 h)$$
$$k_3 = f(t_n + p_2 h, y_n + q_{21} k_1 h + q_{22} k_2 h)$$
$$.$$
$$.$$
$$k_n = f(t_n + p_{n-1} h, y_n + q_{n-1,1} k_1 h + q_{n-2} k_2 h + ... + q_{n-1,n-1} k_{n-1} h)$$

Para derivar los valores de las constantes *a* y *k* en el método de RK de 2^0, escribimos

$$y_{n+1} = y_n + (a_1 k_1 + a_2 k_2)h$$

donde

$$k_1 = f(t_n, y_n)$$
$$k_2 = f(t_n + p_1 h, y_n + q_{11} k_1 h)$$

Desarrollamos y_{n+1} en serie de Taylor de 2^0 orden, tomando como punto base y_n

$$y_{n+1} = y_n + f(t_n, y_n)\, h + \frac{f'(t_n, y_n)}{2!} h^2$$

la derivada de f (t_n , y_n) se desarrolla por medio de la regla de la cadena

$$f'(t_n, y_n) = \frac{\partial f(t, y)}{\partial t} + \frac{\partial f(t, y)}{\partial y} \frac{dy}{dt}.$$

Sustituyendo esta expresión obtenemos

$$y_{n+1} = y_n + f(t_n, y_n)h + \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} \frac{dy}{dt} \right) \frac{h^2}{2!} \qquad (1)$$

Además la expansión de $f(t_n + p_1 h, y_n + q_{11} k_1 h)$ en serie de Taylor, toma la forma dada por:

$$g(x+r, y+s) = g(x,y) + r\frac{\partial g}{\partial x} + s\frac{\partial g}{\partial y} + \ldots$$

entonces, la serie de Taylor de la función mencionada resulta:

$$f(t_n + p_1 h, y_n + q_{11} k_1 h) = f(t_n, y_n) + p_1 h\frac{\partial f}{\partial t} + q_{11} k_1 h\frac{\partial f}{\partial y} + O(h^2)$$

de donde tendremos:

$$y_{n+1} = y_n + (a_1 k_1 + a_2 k_2)h = y_n + a_1 h\ f(t_n, y_n) + a_2 h\ f(t_n + p_1 h, y_n + q_{11} k_1 h)$$

$$= y_n + a_1 h\ f(t_n, y_n) + a_2 h\left[f(t_n, y_n) + p_1 h\frac{\partial f}{\partial x} + q_{11} k_1 h\frac{\partial f}{\partial y} + O(h^2)\right]$$

$$= y_n + a_1 h\ f(t_n, y_n) + a_2 h\ f(t_n, y_n) + a_2 p_1 h^2\frac{\partial f}{\partial x} + a_2 q_{11} k_1 h^2\frac{\partial f}{\partial y} + O(h^3)$$

$$y_{n+1} = y_n + [a_1 f(t_n, y_n) + a_2 f(t_n, y_n)]h + \left[a_2 p_1\frac{\partial f}{\partial x} + a_2 q_{11} f(t_n, y_n)\frac{\partial f}{\partial y}\right]h^2 + O(h^3)$$

Recordando que $k = f(t_n, y_n)$.

Comparando esta última ecuación con (1) arriba

$$a_1 + a_2 = 1$$
$$a_2 p_1 = \tfrac{1}{2}$$
$$a_2 q_{11} = \tfrac{1}{2}$$

que representan 3 ecuaciones en 4 incógnitas. Existen una familia de métodos de RK de 2^0 orden, uno de ellos es el definido por

$$y_{n+1} = y_n + (\tfrac{1}{2} k_1 + \tfrac{1}{2} k_2)h$$

METODO DE RUNGE-KUTTA (CON COEFICIENTES DE RUNGE).

La fórmula del método RK de 4^0 orden es:

$$y_{n+1} = y_n + \Delta y_n \quad , \quad \Delta y_n = \frac{\Delta t}{6}\left(k_0 + 2k_1 + 2k_2 + k_3\right)$$

Donde

$$k_0 = f\left(t_n, y_n\right)$$
$$k_1 = f\left(t_n + \frac{\Delta t}{2}, y_n + \frac{k_0}{2}\Delta t\right)$$
$$k_2 = f\left(t_n + \frac{\Delta t}{2}, y_n + \frac{k_1}{2}\Delta t\right)$$
$$k_3 = f\left(t_n + \Delta t, y_n + k_2\Delta t\right)$$

Las cantidades k representan las pendientes en varios puntos:

- K_0 es la pendiente en el punto inicial del intervalo
- K_3 es la pendiente en el punto final del intervalo
- K_2 es una de las pendientes a mitad del intervalo con ordenada $y_n + \frac{1}{2}k_1\Delta t$
- K_1 es la 2ª pendiente a mitad del intervalo con ordenada $y_n + \frac{1}{2}k_0\Delta t$

Mientras que el método de Euler utiliza una pendiente, el método RK usa un promedio ponderado de pendientes.

DERIVACIÓN DE LA FORMULA DE RK 4^0 ORDEN CON COEFICIENTES DE RUNGE.

Las fórmulas de los métodos de RK se desarrollan a partir de las serie de Taylor:

$$y_{n+1} = y_n + y'_n(t_{n+1} - t_n) + \frac{y''_n}{2!}(t_{n+1} - t_n)^2 + \frac{y'''_n}{3!}(t_{n+1} - t_n)^3 + \frac{y^{iv}_n}{4!}(t_{n+1} - t_n)^4 + ...$$

como $t_{n+1} = t_n + \Delta t$,

$$y_{n+1} = y_n + y'_n \Delta t + \frac{y''_n}{2!}\Delta t^2 + \frac{y'''_n}{3!}\Delta t^3 + \frac{y^{iv}_n}{4!}\Delta t^4 + ...$$

definamos $y_{n+1} - y_n = = \Delta y_n$, entonces tendremos

$$\Delta y_n = y'_n \Delta t + \frac{y''_n}{2!}\Delta t^2 + \frac{y'''_n}{3!}\Delta t^3 + \frac{y^{iv}_n}{4!}\Delta t^4 + ... \qquad (A)$$

Denotamos $y' = f(t, y)$, entonces:

$$y'' = f' = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\frac{dy}{dt} = f_t + f_y f$$

$$y''' = f'' = \frac{\partial f'}{\partial t} + \frac{\partial f'}{\partial y} f =$$
$$= \left[f_{tt} + \left(f_{yt} f + f_y f_t \right) \right] + \left[f_{ty} + \left(f_{yy} f + f_y^2 \right) \right] f$$

$$y^{iv} = f_{ttt} + ...$$

Sustituyendo lo anterior en la ecuación de Δy_n resulta

$$\Delta y_n = f_n \Delta t + \left(\tfrac{1}{2!}\right)\left(f_t + f_y f \right)_n (\Delta t)^2 +$$
$$\left(\tfrac{1}{3!}\right)\left[f_{tt} + 2 f_{ty} f + f_{yy} f^2 + \left(f_t + f_y f \right) f_y \right]_n (\Delta t)^3 + \left(\tfrac{1}{4!}\right)\left[f_{ttt} + ... \right](\Delta t)^4 \qquad \text{(B)}$$

La evaluación de tanta derivada es poco práctico y para evitar esta dificultad tomamos arbitrariamente

$$\Delta y_n = \left(\mu_0 z_0 + \mu_1 z_1 + \mu_2 z_2 + ... + \mu_m z_m \right)$$

$$z_0 = f\left(t_n, y_n\right)\Delta t$$
$$z_1 = f\left(t_n + \alpha_1 \Delta t, y_n + \beta_{10} z_0\right)\Delta t$$
$$z_2 = f\left(t_n + \alpha_2 \Delta t, y_n + \beta_{20} z_0 + \beta_{21} z_1\right)\Delta t$$
.
.
$$z_m = f\left(t_n + \alpha_n \Delta t, y_n + \beta_{m0} z_0 + \beta_{m1} z_1 + ...\right)\Delta t$$

El objetivo es determinar los tres conjuntos de constantes μ , α , β. El subíndice *m* indica que las expresiones para Δy_n debe coincidir hasta el término que contiene a $(\Delta t)^{m-1}$.

Con *m = 3* :

$$\Delta y_n = \mu_0 z_0 + \mu_1 z_1 + \mu_2 z_2 + \mu_3 z_3 \qquad \text{(C)}$$

$$\begin{aligned} z_0 &= f(t_n, y_n)\Delta t \\ z_1 &= f(t_n + \alpha_1 \Delta t, y_n + \beta_{10} z_0)\Delta t \\ z_2 &= f(t_n + \alpha_2 \Delta t, y_n + \beta_{20} z_0 + \beta_{21} z_1)\Delta t \\ z_3 &= f(t_n + \alpha_3 \Delta t, y_n + \beta_{30} z_0 + \beta_{31} z_1 + \beta_{32} z_2)\Delta t \end{aligned}$$

Ahora buscamos la forma de determinar las 13 constantes μ_0 , μ_1 , μ_2 , μ_3 , α_1 , α_2 , α_3 , α_4 , β_{10} , β_{20}, β_{30}, β_{21}, β_{31}, β_{32}. Para esto, partimos de la serie de Taylor para dos variables independientes alrededor del punto (a , b):

$$f(a+h, b+k) = f(a,b) + f_x(a,b) + f_y(a,b)k + \frac{1}{2!}\left[f_{xx}(a,b)h^2 + 2 f_{xy}(a,b)hk + f_{yy}(a,b)k^2\right] + \ldots.$$

La serie anterior a menudo se escribe simbólicamente como sigue:

$$f(a+h, b+k) = f(a,b) + \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) f(a,b) + \\ + \frac{1}{2!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 f(a,b) + \\ + \frac{1}{3!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^3 f(a,b) + \ldots$$

Ahora sustituimos z_0 , z_1 , z_2 y z_3 en la ultima ecuación para obtener:

$$z_0 = f_n \Delta t$$

$$z_1 = \left[f_n + \left(\alpha_1 \Delta t \frac{\partial}{\partial t} + \beta_{10} z_0 \frac{\partial}{\partial y} \right) f_n + \frac{1}{2!} \left(\alpha_1 \Delta t \frac{\partial}{\partial t} + \beta_{10} z_0 \frac{\partial}{\partial y} \right)^2 f_n + \frac{1}{3!} \left(\alpha_1 \Delta t \frac{\partial}{\partial t} + \beta_{10} z_0 \frac{\partial}{\partial y} \right)^3 f_n + \ldots \right] \Delta t$$

$$z_2 = \left[\begin{array}{l} f_n + \left(\alpha_2 \Delta t \frac{\partial}{\partial t} + \left(\beta_{20} z_0 + \beta_{21} z_1 \right) \frac{\partial}{\partial y} \right) f_n + \frac{1}{2!} \left(\alpha_2 \Delta t \frac{\partial}{\partial t} + \left(\beta_{20} z_0 + \beta_{21} z_1 \right) \frac{\partial}{\partial y} \right)^2 f_n \\ + \frac{1}{3!} \left(\alpha_2 \Delta t \frac{\partial}{\partial t} + \left(\beta_{20} z_0 + \beta_{21} z_1 \right) \frac{\partial}{\partial y} \right)^3 f_n + \ldots \end{array} \right] \Delta t$$

$$z_3 = \left[\begin{array}{l} f_n + \left(\alpha_3 \Delta t \frac{\partial}{\partial t} + \left(\beta_{30} z_0 + \beta_{31} z_1 + \beta_{32} z_2 \right) \frac{\partial}{\partial y} \right) f_n + \frac{1}{2!} \left(\alpha_2 \Delta t \frac{\partial}{\partial t} + \left(\beta_{30} z_0 + \beta_{31} z_1 + \beta_{32} z_2 \right) \frac{\partial}{\partial y} \right)^2 f_n + \\ + \frac{1}{3!} \left(\alpha_2 \Delta t \frac{\partial}{\partial t} + \left(\beta_{30} z_0 + \beta_{31} z_1 + \beta_{32} z_2 \right) \frac{\partial}{\partial y} \right)^3 f_n + \ldots \end{array} \right] \Delta t$$

Si igualamos los coeficientes de (B) con (C) y estas ultimas relaciones de las *z*'s, obtenemos 11 ecuaciones en 13 incógnitas :

$$\alpha_1 = \beta_{10}$$
$$\alpha_2 = \beta_{20} + \beta_{21}$$
$$\alpha_3 = \beta_{30} + \beta_{31} + \beta_{32}$$
$$\mu_0 + \mu_1 + \mu_2 + \mu_3 = 1$$
$$\mu_1 \alpha_1 + \mu_2 \alpha_2 + \mu_3 \alpha_3 = \tfrac{1}{2}$$
$$\mu_1 \alpha_1^2 + \mu_2 \alpha_2^2 + \mu_3 \alpha_3^2 = \tfrac{1}{3}$$
$$\mu_1 \alpha_1^3 + \mu_2 \alpha_2^3 + \mu_3 \alpha_3^3 = \tfrac{1}{4}$$

$$\mu_2 \alpha_1 \beta_{21} + \mu_3 \left(\alpha_1 \beta_{31} + \alpha_2 \beta_{32} \right) = \tfrac{1}{6}$$
$$\mu_2 \alpha_1^2 \beta_{21} + \mu_3 \left(\alpha_1^2 \beta_{31} + \alpha_2^2 \beta_{32} \right) = \tfrac{1}{12}$$
$$\mu_2 \alpha_1 \alpha_2 \beta_{21} + \mu_3 \left(\alpha_1 \beta_{31} + \alpha_2 \beta_{32} \right) \alpha_3 = \tfrac{1}{8}$$
$$\mu_3 \alpha_1 \beta_{21} \beta_{32} = \tfrac{1}{24}$$

Con 11 ecuaciones y 13 incógnitas debemos escoger 2 de las incógnitas y resolver el sistema de ecuaciones.

Para $\mu_1 = \mu_2 = \frac{1}{3}$ obtenemos:

$\mu_0 = 1/6$, $\mu_1 = 1/3$, $\mu_2 = 1/3$, $\mu_3 = 1/6$;

$\alpha_1 = ½$, $\alpha_2 = ½$, $\alpha_3 = 1$;

$\beta_{10} = 1/2$, $\beta_{20} = 0$, $\beta_{30} = 0$, $\beta_{21} = 1/2$, $\beta_{31} = 0$, $\beta_{32} = 1$.

Los coeficientes anteriores se denominan *coeficientes de Runge* .

MÉTODO DE CUADDRATURA DE GAUSS-LEGENDRE

Este método de basa en muestrear el integrando de la función cuya integral se desea encontrar, a valores que representan raíces de *polinomios ortogonales*. Los más populares de éstos son los *polinomios de Legendre*.

En general un conjunto de funciones $\phi_0(x), \phi_1(x), \cdots, \phi_n(x)$ se conocen como **ortogonales** en un intervalo $a \le x \le b$, si

$$\int_a^b w(x)\,\phi_m(x)\,\phi_n(x)\,dx = 0, \quad m \neq n \tag{1}$$

Donde $w(x)$ es una función de ponderación no negativa en $[a \quad b]$.

Si las funciones $\phi_m(x)$ son polinomios, estos se designan como **polinomios ortogonales**.

POLINOMIOS DE LEGENDRE.

Los primeros cinco polinomios de **Legendre** son:

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \tfrac{1}{2}(3x^2 - 1) \\ P_3(x) &= \tfrac{1}{2}(5x^3 - 3x) \\ P_4(x) &= \tfrac{1}{8}(35x^4 - 30x^2 + 3) \end{aligned} \tag{2}$$

El polinomio de Legendre de grado *n* se puede obtener por medio d la fórmula de Rodrigues

$$P_n(x) = \frac{1}{2^n\, n!}\,\frac{d^n}{dx^n}(x^2 - 1)^n$$

O bien a partir de la fórmula recursiva:

$$(n+1)\cdot P_{n+1}(x)-(2n+1)\cdot x\cdot P_n(x)+n\cdot P_{n-1}(x)=0$$

Las relaciones de ortogonalidad y normalización, con las funciones de ponderación (peso) igual a 1, son:

$$\int_{-1}^{1} P_n(x)\,P_m(x)\,dx = \begin{cases} 0 & m \neq n \\ \dfrac{2}{2n+1} & m = n \end{cases} \qquad (3)$$

Todas las raíces de cada $P_n(x)=0$ son reales y distintas, además están contenidas en el intervalo $[-1 \quad 1]$.

CUADRATURA GAUSSIANA.

El propósito es discutir la fórmula de integración Gaussiana que aproxima

$$\int_{-1}^{1} f(x)\,dx \qquad (4)$$

y mostrar que con un simple cambio de variable se pueden extender los límites de integración a valores distintos a $[-1 \quad 1]$.

La aproximación d la integral definida se puede definir como

$$\int_{-1}^{1} f(x) = w_0\ f(x_0) + w_1\ f(x_1) + w_2\ f(x_2) + \cdots + w_n\ f(x_n) = \sum_{k=0}^{n} w_k\ f(x_k) \qquad (5)$$

$w_0, w_1, \cdots, w_n$ son los coeficientes ponderados ó pesos.

El problema consiste en encontrar las $(2n+2)$ constantes $(w_i, f(x_i))$. Para encontrar las mencionadas constantes, partimos de la suposición básica de que la fórmula (2) representa sin aproximación, es decir, exactamente un polinomio de orden $2n+1$ ó menor.

Primero mostramos que los puntos x_k $(k = 0, \cdots, n)$, son iguales a las raíces del polinomio de Legendre $P_{n+1}(x)$.

Tomemos un polinomio arbitrario $g_n(x)$ de grado *n*. En términos de polinomios de Legendre $g_n(x)$ puede expresarse como

$$g_n(x) = \beta_0\, P_0(x) + \beta_1\, P_1(x) + \cdots + \beta_n\, P_n(x) \qquad (6)$$

Como ejemplo supongamos

$$g_2(x) = 1 + 2x + x^2\,.$$

De la ecuación (6) y (2) obtendremos:

$$g_2(x) = \beta_0 + \beta_1 x + \frac{\beta_2}{2}(3x^2 - 1) = \left(\beta_0 - \frac{\beta_2}{2}\right) + \beta_1 x + \tfrac{3}{2}\beta_2 x^2$$

Comparando esta última expresión con la $g_2(x)$ inicial obtenemos:

$$\beta_0 - \frac{\beta_2}{2} = 1, \quad \beta_1 = 2, \quad \tfrac{3}{2}\beta_2 = 1,$$

De donde obtenemos finalmente: $\beta_0 = \frac{4}{3}, \quad \beta_1 = 2, \quad \beta_2 = \frac{2}{3}$.

Sustituyendo esto en (6), obtenemos

$$g_2(x) = \tfrac{4}{3}P_0(x) + 2P_1(x) + \tfrac{2}{3}P_2(x) .$$

Este simple ejemplo muestra que cualquier polinomio $g_n(x)$ se puede escribir en términos de polinomios de Legendre.

A partir de la definición de ortogonalidad expresada en (3):

$$\int_{-1}^{1} g_n(x)P_{n+1}(x)\,dx =$$
$$\int_{-1}^{1} \beta_0 P_0(x)P_{n+1}(x) + \int_{-1}^{1} \beta_1 P_1(x)P_{n+1}(x) + \cdots + \int_{-1}^{1} \beta_n P_n(x)P_{n+1}(x) = 0 \qquad (7)$$

Observamos que $g_n(x)P_{n+1}(x)$, es un polinomio de grado $2n+1$, y por tanto representa *exactamente* polinomios de grado $2n+1$ ó menos, lo cual constituye el requisito básico mencionado antes, en la definición de la ecuación (5), para la selección de w_k y $x_k \quad (k=0,\cdots,n)$.

Comparando (7) con (5) obtenemos:

$$w_0 g_n(x_0)P_{n+1}(x_0) + w_1 g_n(x_1)P_{n+1}(x_1) + \cdots + w_n g_n(x_n)P_{n+1}(x_n) = 0 \qquad (8)$$

Como $g_n(x)$ es un polinomio arbitrario, $g_n(x_k) \quad (k=0,\cdots,n)$ no es cero en general. Así mismo las $n+1$funciones de ponderación ó pesos $w_k \quad (k=0,\cdots,n)$ no pueden ser todos cero, de lo contrario la ecuación (5) será igual a cero, lo cual constituye el caso trivial.

Dado lo anterior la única condición para la ecuación (8) será:

$$P_{n+1}(x_0) = 0$$
$$P_{n+1}(x_1) = 0$$
$$.$$
$$.$$
$$.$$
$$P_{n+1}(x_n) = 0$$

Lo anterior implica que $x_0, x_1, \cdots, x_n$ son las *raíces del polinomio de Legendre* $P_{n+1}(x) = 0$.

Para $P_{n+1}(x) \in [-1 \quad +1]$ existen *n+1* raíces distintas.

Como ejemplo, para *n=1*,

$$P_{n+1}(x) = P_2(x) = \tfrac{1}{2}\left(3x^2 - 1\right) = 0$$

por lo que las raíces son $x = \pm 1/\sqrt{3}$.

Mientras que para el caso *n=2*,

$$P_3(x) = \tfrac{1}{2}\left(5x^3 - 3x\right) = \tfrac{1}{2}\,x\left(5x^2 - 3x\right) = 0\,,$$

por lo que las raíces son $x = 0, \quad x = \pm\sqrt{3/5}$.

Para la determinación de los coeficientes $w_k \ \ (k = 0, \cdots, n)$ de nuevo tomamos en consideración el requisito establecido en (5), esto es, que si el integrando *f(x)* es un polinomio de grado *n+1* ó menos, dicha ecuación no involucra una aproximación.

Por definición, el polinomio de Lagrange para aproximar cualquier polinomio $h_n(x)$ de grado *n*, que pasa por *n+1* puntos $x_k \ \ (k = 0, \cdots, n)$ se puede expresar como

$$h_n(x) = \sum_{k=0}^{n} h(x_k) L_k(x)$$

Por lo que

$$\int_{-1}^{+1} h_n(x)\,dx = \int_{-1}^{+1} \sum_{k=0}^{n} h(x_k) L_k(x)\,.$$

Dado que $h(x_k)$ es una constante

$$\int_{-1}^{+1} h_n(x)\,dx = \sum_{k=0}^{n} h(x_k) \int_{-1}^{+1} L_k(x) \qquad (9)$$

Comparando (5) con (9) tenemos

$$w_k = \int_{-1}^{+1} L_k(x) \qquad k = 0, \cdots, n \qquad (10).$$

Es común encontrar la definición de L_k y por tanto de w_k en términos de polinomios de Legendre. Esto se obtiene como sigue.

El polinomio $\frac{P_{n+1}(x)}{x - x_k}$ es igual a cero para todo $x = x_j$, $j = 0, \cdots, n$, pero $j \neq k$.

De acuerdo a la regla de L'Hopital

$$\lim_{x \to x_k} \frac{P_{n+1}(x)}{x - x_k} = \left[\frac{\dfrac{dP_{k+1}(x)}{dx}}{\dfrac{d\left(x - x_k\right)}{dx}} \right]_{x = x_k} = \frac{dP_{k+1}(x_k)}{dx} = P'_{n+1}(x_k)$$

(Dado que la derivada del denominador es igual a 1), donde x_k es una de las raíces del polinomio de Legendre $P_{n+1}(x)=0$.

De acuerdo a lo anterior, el polinomio de Lagrange puede expresarse como

$$L_k = \frac{1}{P'_{n+1}(x_k)} \frac{P_{n+1}(x)}{\left(x - x_k\right)}$$

por tanto las funciones de ponderación (pesos) se definen alternativamente como

$$w_k = \frac{1}{P'_{n+1}(x_k)} \int_{-1}^{+1} \frac{P_{n+1}(x)}{\left(x - x_k\right)}\, dx \qquad (11).$$

Para ejemplificar consideremos *n=1*, $P_{n+1}(x) = P_2(x) = \frac{1}{2}\left(3x^2 - 1\right)$ cuyas raíces son $x_0 = 1/\sqrt{3}$, $x_1 = -1/\sqrt{3}$ y su derivada $P'_2(x) = \frac{1}{2}\left(6x\right) = 3x$. De aquí entonces

$$w_0 = \frac{1}{3\left(1/\sqrt{3}\right)} \int_{-1}^{+1} \frac{\frac{1}{2}\left(3x^2 - 1\right)}{x + \frac{1}{\sqrt{3}}}\, dx$$

$$w_1 = \frac{1}{3\left(-1/\sqrt{3}\right)} \int_{-1}^{+1} \frac{\frac{1}{2}\left(3x^2 - 1\right)}{x + \frac{1}{\sqrt{3}}}\, dx$$

Para *n=2*, $P_{n+1}(x) = P_3(x) = \frac{1}{2}\left(5x^3 - 3x\right)$. Las raíces de $P_3(x)$ se determinaron previamente y resultaron $x_0 = -\sqrt{\frac{3}{5}}, \quad x_1 = 0, \quad x_2 = \sqrt{\frac{3}{5}}$ y la derivada de $P_3(x)$, $P_3^{'}(x) = \frac{3}{2}\left(5x^2 - 1\right)$, por lo que obtenemos

$$w_0 = \frac{1}{\frac{3}{2}\left(5\cdot\frac{3}{5} - 1\right)} \int_{-1}^{+1} \frac{\frac{1}{2}\left(5x^3 - 3x\right)}{x + \sqrt{\frac{3}{5}}}\, dx = \frac{2}{3(3-1)} \int_{-1}^{+1} \frac{1}{2}\left(x^2 - \sqrt{\frac{3}{5}}x \right) dx = \frac{5}{9}$$

$$w_1 = \frac{1}{\frac{3}{2}\left(5\cdot 0 - 1\right)} \int_{-1}^{+1} \frac{\frac{1}{2}\left(5x^3 - 3x\right)}{x - 0}\, dx = \frac{8}{9}$$

$$w_2 = \frac{1}{\frac{3}{2}\left(5\cdot\frac{3}{5} - 1\right)} \int_{-1}^{+1} \frac{\frac{1}{2}\left(5x^3 - 3x\right)}{x - \sqrt{\frac{3}{5}}}\, dx = \frac{2}{3(3-1)} \int_{-1}^{+1} \frac{1}{2}\left(x^2 + \sqrt{\frac{3}{5}}x \right) dx = \frac{5}{9}$$

El procedimiento descrito arriba puede extenderse para diferentes valores de *n*, es decir, para tres puntos, cuatro puntos, cinco puntos, etcétera. La siguiente tabla muestra algunos de estos casos, y en [1] se pueden encontrar una lista más grande.

Raíces de los polinomios de Legendre $P_{n+1}(z)$ y sus factores de ponderación para la cuadratura de Gauss-Legendre.

Raíces (z_i)	$\int_{-1}^{+1} F(z)\,dz = \sum_{i=0}^{n} w_i\,F(z_i)$	Factores de ponderación (peso)
±0.57735 02691 89626	n = 1 fórmula de dos puntos	1.00000 00000 00000
0.00000 000000 ±0.77459 66692 41483	n = 2 fórmula de tres puntos	0.88888 88888 88889
±0.33998 10435 84856 ±0.86113 63115 94053	n = 3 fórmula de cuatro puntos	0.65214 51548 62546 0.34785 48451 37454
0.00000 00000 000000 ±0.53846 93101 05683 ±0.90617 98459 38664	n = 4 fórmula de cinco puntos	0.56888 88888 88889 0.47862 86704 99366 0.23692 68850 56189

Límites de Integración.

Dado que los límites de integración asociados con es te desarrollo son -1 y +1, en un problema de aplicación habrá que ajustar el procedimiento de la cuadratura Gaussiana a los límites de la aplicación particular. Lo anterior se logra mediante un simple cambio de variable.

Definimos una relación lineal con la nueva variable

$$x = \frac{(b-a)\,t + (b+a)}{2} \qquad dx = \frac{b-a}{2}\,dt$$

En este caso $\int_a^b f(x)\,dx$ se convertirá en

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{+1} f\left(\frac{(b-a)\,t_k + (b+a)}{2}\right)dt$$

Dado que la cuadratura de Gauss-Legendre se define

$$\int_{-1}^{+1} f(x)\,dx = \sum_{k=0}^{n} w_k \;\; f(x_k)$$

La integral anterior se puede aproximar como

$$\int_a^b f(x)\,dx = \frac{(b-a)}{2}\sum_{k=0}^{n} w_k \;\; f\left(\frac{(b-a)\,t_k + (b+a)}{2}\right)$$

Esta formulación es la apropiada para usarse en la programación de este método en computadora, en lugar de usar una transformación simbólica de *f (x)*. En este caso los puntos base t_k se transforman y los factores de ponderación w_k se modifican al multiplicarse por la constante $\left(\frac{b-a}{2}\right)$.

Por ejemplo, usamos la fórmula de cuadratura de Gauss-Legendre de dos puntos para calcular

$$\int_2^4 \left(x^2 - 2x + 1 \right) dx = \frac{26}{3}$$

La fórmula de cuadratura Gauss-Legendre será (para el método de dos puntos)

$$\int_2^4 \left(x^2 - 2x + 1 \right) dx =$$

$$= \frac{(4-2)}{2}\left[(1.0) * f\left(\frac{-0.577350269189626*(4-2)+4+2}{2} \right) + (1.0) * f\left(\frac{0.577350269189626*(4-2)+4+2}{2} \right) \right]$$

$$= 2.0239322565749 + 6.6427344100918$$

$$= 8.666666666667 .$$

BIBLIOGRAFÍA

- [PH] Elements of Numerical Analysis.

  Peter Henrrici

  John Wiley & Sons 1964

- [MTH] Scientific Computing. An Introductory Survey.

  Michael T. Heath

  McGraw Hill 4th edition 1997

- [ChC] Métodos Numéricos para Ingenieros.

  S. C. Chapra, R. P. Canale

  McGraw Hill 4ª edición 2002

- [PAS] Introduction to Numerical Methods

  Peter A. Stark

  MacMillan Publishing Co. 1970

- [CM] Numerical Computing with MATLAB©

  Cleve Moler (Gratuito en la página web de Cleve moler en http://www.mathworks.com/)

  SIAM (Society forIndustrial and Applied Mathematics)

- [JBS] Numerical Mathematical Analysis. Sixth edition

  James B. Scarborough

  The Johns Hopkins Press 1966

- [GW] Applied Numerical Analysis. 6th edition

  Curtis F. Gerald, Patrick O. Wheatley

  Addison Wesley 1997

- [BT] Numerical Methods and Analysis

  James I. Buchanan, Peter R. Turner

  Mc Graw Hill 1992

- [MLB] Mathematical Methods in the Physical Sciences 3rd edition

  Mary L. Boas

  John Wiley & Sons 2006

- [NJH] Accuracy and Stability of Numerical Algorithms

  Nicholas J. Higham

  SIAM 1996

- [IK] Analysis of Numerical Methods

  Eugene Issacson, Herbert B. Keller

  John Wiley & Sons. 1966

  Dover Publicatioons 2012

- [RR] A First Course in Numerical Analysis 2nd edition

  Anthony Ralston, Philip Rabinowitz

  Dover Publications 2012

- [CLW] Applied Numerical Methods

  Bruce Carnahan, H. A. Luther, James O. Wilkes

  John Wiley & Sons 1969