SECRETARIA DE RECURSOS HIDRICOS

INSTITUTO NACIONAL DE CIENCIA Y TECNICA HIDRICAS

LABORATORIO DE HIDRAULICA APLICADA

ERRORES EN LA SOLUCION NUMERICA DE ECUACIONES DIFERENCIALES

por

Dr. Angel N. MENENDEZ

Jefe del Dto. de Hidráulica Computacional

Informe LHA 064 - 004 - 88

EZEIZA, mayo de 1987

RESUMEN

En este apunte se desarrollan los conceptos fundamentales sobre el análisis de los errores presentes en las soluciones numéricas de ecuaciones diferenciales. Se describen las principales fuentes de errores y se discute sobre la consistencia de las aproximaciones en diferencias y la convergencia, estabilidad y precisión de las soluciones numéricas

DESCRIPTORES TEMATICOS:

Simulación numérica, modelación matemática, diferencias finitas, errores.

PROLOGO

Este apunte es una continuación del titulado "Introducción a la simulación numérica de problemas hidráulicos"(Informe LHA 064-003-87), en el cual se introdujeron nociones elementales sobre los métodos de simulación numérica. Aquí se desarrolla un tópico fundamental del análisis numérico: la estimación y control de los errores en los resultados del proceso de cálculo. En particular, se estudia este tema en relación a la resolución numérica de ecuaciones diferenciales.

El profesional que trabaja en simulación numérica tiene una percepción dramática de este problema cuando se topa con casos de inestabilidad numérica que se manifiestan "explosivamente", produciendo resultados intermedios o finales que superan el rango de punto flotante de la computadora. Sin embargo, esas inestabilidades no siempre se hacen evidentes: a veces, es necesario desentrañarlas. En rigor, lo que debe comprenderse es que un resultado no es un valor (o una serie de valores) sino un intervalo que hay que estimar y controlar. En el apunte se tratan de exponer las nociones fundamentales para lograr ese cometido.

Ezeiza, mayo de 1988

INDICE

CAPITULO 1

ERRORES EN EL CALCULO NUMERICO

1.1. Tipos de errores

En un proceso de cálculo el error de los resultados proviene de una diversidad de fuentes. Para estimarlo es necesario analizar su "composición" remitiéndolo a esas fuentes. De esta manera pueden distinguirse distintos tipos de errores, para cuya estimación se requieren, en general, tratamientos diferentes.

Los tres tipos principales de error son: los inherentes, los de truncamiento y los de redondeo. Los errores inherentes son los asociados a los datos de entrada del proceso de cálculo. Puede tratarse de errores de medición. Tambien se refiere a los que resultan de representar un número por una secuencia finita de dígitos (como hace necesariamente una computadora digital), aunque aquél sea conocido, teóricamente, con una precisión infinita. Además, tambien deben ser considerados como inherentes los errores en los resultados de un cálculo, cuando éstos constituyen datos de entrada de un nuevo proceso de cálculo.

Los errores de truncamiento son los que se producen por truncar un proceso matemáticamente infinito (es decir, un proceso que involucra, en algún sentido, un paso al límite). Tal es lo que sucede al evaluar series por medio de sumas finitas, integrales por fórmulas de cuadratura, derivadas por aproximaciones en diferencias finitas, etc.

Los errores de redondeo, finalmente, son los que se producen por conservar solo un número finito de dígitos durante las operaciones aritméticas (lo cual está ligado, nuevamente, a la esencia de la computadora digital).

En la Fig. 1.1 se representa esquemáticamente como interviene cada uno de estos tres tipos de errores en la afectación de los resultados de un proceso de cálculo.

1.2. Propagación de errores inherentes y de redondeo

Cuando la influencia de los errores de redondeo es despreciable, la propagación de errores inherentes (es decir, la estimación de los errores en los resultados como fruto de los errores en los datos de entrada) se lleva a cabo por la conocida fórmula general de propagación (1). Esta establece que si la variable y es función de las variables x_1, x_2, ...x_n, de las cuales se conocen estimaciones (en rigor, cotas superiores) de sus errores inherentes Δx_1, Δx_2, ...Δx_n, una estimación del error en y puede obtenerse como

$$(1.1) \qquad \Delta y = \sum_{i=1}^{n} \left| \frac{\partial y}{\partial x_i} \right| \Delta x_i$$

A partir de la Ec. (1.1) es fácil deducir las conocidas reglas para propagación del error en las cuatro operaciones elementales (suma, resta, multiplicación y división).

Cuando la influencia de los errores de redondeo es significativa (lo cual sucede, por ejemplo, cuando la cantidad de operaciones involucradas en el proceso de cálculo es muy grande), el análisis de la propagación se hace más complicado. En este caso, una posibilidad consiste en hacer un seguimiento del detalle del proceso de cálculo, desglosándolo en pequeños sub-procesos que involucren solo a las operaciones elementales (por ejemplo, la función seno debería representarse por su desarrollo en serie de Taylor). El error de los datos de entrada de cada sub-proceso podría propagarse de acuerdo a la fórmula, pero antes de utilizar el resultado como dato de entrada de un nuevo sub-proceso habría que adicionarle el error de redondeo. Una forma sistemática, aunque tediosa, de llevar a cabo este procedimiento se describe en la Ref.(2) bajo la denominación de gráfica de procesos. Alternativamente, puede recurrirse a la técnica de "análisis retrospectivo de errores" (1), que consiste en representar el error de redondeo cometido en cada sub-proceso como un error inherente extra en los datos de entrada de ese sub-proceso, si éste fuera realizado con precisión infinita. Procediendo sucesivamente de atrás hacia adelante, el problema original se reduce a otro con los datos de entrada afectados de nuevos errores inherentes, por lo cual el error en los resultados puede estimarse utilizando la fórmula general de propagación, Ec. (1.1).

Estos tratamientos de propagación de errores adolecen de una limitación básica, a saber, solo son útiles para acotar el error. Si bien ésto puede ser adecuado cuando el error predominante es de tipo sistemático, resulta en gruesas sobrestimaciones cuando la naturaleza de los errores es estocástica (la referencia, en este caso, es a los errores inherentes, ya que los de redondeo son siempre de tipo aleatorio). En esta última situación es conveniente introducir un tratamiento estadístico, y caracterizar el error, por ejemplo, por medio de la desviación estandar (1).

Todos los procedimientos de propagación de errores inherentes y de redondeo mencionados hasta aquí descansan sobre el supuesto de que el proceso de cálculo, o algoritmo, puede ser analizado completamente. En algoritmos complejos, y ésto hoy en día se refiere a la mayoría de los involucrados en estudios de ingeniería, es prácticamente imposible llevar a cabo un tal análisis. En consecuencia, se recurre a una técnica de perturbaciones experimentales, consistente en efectuar varios cálculos variando los valores de los datos de entrada, y estudiando la sensibilidad de los resultados a esas variaciones (1). Para sopesar la influencia de los errores de redondeo, puede repetirse un cálculo con los mismos valores de los datos de entrada pero variando la precisión de la máquina (por ejemplo, de simple a doble).

1.3. Estimación de los errores de truncamiento

Para estimar los errores en los resultados de un proceso de cálculo debidos al truncamiento, es necesario disponer de una mejor estimación de los resultados del proceso infinito del cual el algoritmo en cuestión es una aproximación. Cuando los errores de truncamiento son dominantes, esta estimación puede llevarse a cabo recalculando los resultados con el truncamiento efectuado en un nivel superior (por ejemplo, incluyendo más términos en la evaluación de una serie). En ciertos casos pueden desarrollarse fórmulas que den cuenta, al menos, de la dependencia de los errores de truncamiento respecto de los parámetros del problema (por ejemplo, el número de términos utilizado en la evaluación de una serie).

Cuando el número de operaciones involucradas en el algoritmo es muy grande (o cuando el proceso es inestable), entran a tallar los errores de redondeo. Un ejemplo simple, pero muy ilustrativo, se presenta en la Fig. 1.2, tomada de la Ref.(2). Allí se muestra, para el problema de la evaluación de una integral mediante fórmulas de cuadratura, cómo el error de los resultados disminuye, al aumentar el número de intervalos, mientras domina el error de truncamiento. En cambio, comienza a aumentar en cuanto los errores de redondeo se hacen dominantes.

1.4. Estabilidad de algoritmos

El problema de la estabilidad de un algoritmo consiste en analizar la sensibilidad de los resultados a pequeñas variaciones en los datos de entrada. En términos laxos, se denomina estable a aquél cuyos resultados muestran poca sensibilidad a esas variaciones, y viceversa. Obviamente, todas estas consideraciones requieren ser expresadas con mayor rigor matemático, lo cual se hará en lo que sigue.

En primer lugar, es obvio que se necesita alguna unidad de medida para cuantificar las variaciones. La unidad elemental u es la denominada unidad de máquina o de redondeo, que es el máximo error relativo con que la máquina puede representar a cualquier número real, dentro de su rango de punto flotante (para redondeo simétrico $u=0,5*B^{-t+1}$, donde B es la base y t la cantidad de dígitos de la mantisa; para redondeo truncado u vale el doble que en el caso anterior).

En segundo lugar, hay que determinar la causa de la inestabilidad, cuando ésta se presenta. Existen dos posibilidades. Una es que esa sea una caracaterística del problema (puede reflejar, por ejemplo, una inestabilidad física), en cuyo caso se trata de un problema mal condicionado o matemáticamente inestable. La otra posibilidad, más común que la anterior, es que el algoritmo esté pobremente construido; se habla, entonces, de un algoritmo mal condicionado o numéricamente inestable. Hay que tener en cuenta, de todos modos, que la inestabilidad no se manifiesta necesariamente para cualquier juego de valores de los datos de entrada.

Se plantea, entonces, la cuestión de qué hacer en el caso de tener un algoritmo inestable. La respuesta obvia es que éste debe ser reemplazado por un algoritmo "mejor". En general, existe más de una manera de plantear un proceso de cálculo.

Esto conduce al concepto de algoritmos matemáticamente equivalentes, que son aquellos que darían resultados idénticos, a partir de los mismos datos de entrada, si las operaciones se efectuaran con precisión infinita (es decir, sin errores de redondeo). En la mayoría de los casos, un algoritmo puede ser derivado a partir de otro matemáticamente equivalente mediante manipulaciones algebraicas. Ahora bien, debido a la introducción de errores de redondeo durante los cálculos, algoritmos matemáticamente equivalentes no dan, en general, los mismos resultados. Si las diferencias entre resultados son pequeñas, se trata de algoritmos numéricamente equivalentes. Más precisamente, se dice que dos algoritmos son numéricamente equivalentes cuando sus resultados, usando los mismos datos de entrada, no difieren entre sí en más de lo que los resultados exactos (es decir, calculados con precisión infinita) lo harían entre sí, para dos cálculos efectuados con los datos originales uno y con los datos perturbados en unos pocos u el otro (esta última diferencia debería ser evaluada, en principio, por medio de la fórmula general de propagación de errores, Ec. (1.1)). A menudo se da el caso de que dos algoritmos matemáticamente equivalentes no son numéricamente equivalentes. Precisamente, en el caso de tener un algoritmo numéricamente inestable (para un problema bien condicionado), éste debería ser reemplazado por un algoritmo matemáticamente equivalente, pero estable.

En algoritmos iterativos la presencia de inestabilidades se hace rápidamente evidente, ya que, en general, los resultados tienden a crecer en valor absoluto (ya sea en forma monótona u oscilatoria) con una tendencia exponencial, llegando a superar el rango de la máquina. Esto no sucede, en general, con los algoritmos directos.

CAPITULO 2

CONSISTENCIA DE UNA APROXIMACION EN DIFERENCIAS

2.1.- Error de discretización

Al aproximar una ecuación diferencial por medio de una ecuación en diferencias (utilizando cualquier método de discretización) se comete un cierto error de truncamiento. Se denomina error de discretización al residuo que resulta al reemplazar en la ecuación en diferencias la solución de la ecuación diferencial (que se denominará, en adelante, solución analítica). Este error indica, obviamente, en qué medida la solución analítica satisface la ecuación en diferencias. La relación entre el error de discretización y el error de truncamiento de la solución de la ecuación en diferencias (que se denominará solución numérica) se explicará más abajo.

La estimación del error de discretización se lleva a cabo mediante un procedimiento que se introducirá por medio de un ejemplo. Sea el siguiente problema de valores iniciales, que es el problema más simple de difusión

$$(2.1)\quad \frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} \qquad , t \geq 0 \ , \quad -\infty < x < \infty$$

donde u(x,t) es la función incógnita de las variables independientes x y t, que se supondrá representan el espacio y el tiempo, respectivamente, y ν es una constante positiva (que representa una viscosidad cinemática). Una de las aproximaciones en diferencias más simples de la Ec. (2.1) consiste en utilizar el siguiente esquema explícito centrado:

$$(2.2)\quad \frac{u_j^{n+1} - u_j^n}{\Delta t} = \nu \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

donde el subíndice j identifica el nodo espacial, el supraíndice n indica el paso de cálculo temporal, y Δt y Δx son los intervalos temporal y espacial de discretización, respectivamente. Reemplazando en la Ec. (2.2) la solución numérica u_j^n por la analítica $u(x_j, t^n)$, donde x_j es la coordenada x del nodo j y t^n el tiempo correspondiente al nivel n, se obtiene el error de discretización:

$$(2.3)\quad \epsilon_j^n = \frac{u(x_j,t^{n+1}) - u(x_j,t^n)}{\Delta t} - \nu \frac{u(x_{j+1},t^n) - 2u(x_j,t^n) + u(x_{j-1},t^n)}{\Delta x^2}$$

Ahora, como los intervalos de discretización Δt y Δx son "pequeños", es posible efectuar desarrollos en serie de Taylor de la solución analítica (suponiéndola diferenciable hasta el orden que se necesite) alrededor del punto (x_j, t^n). De esta manera se tiene que

$$(2.4)\quad u(x_j,t^{n+1}) = u(x_j,t^n) + \left.\frac{\partial u}{\partial t}\right|_j^n \Delta t + \left.\frac{\partial^2 u}{\partial t^2}\right|_j^n \frac{\Delta t^2}{2} + O(\Delta t^3)$$

$$(2.5)\quad u(x_{j\pm1},t^n) = u(x_j,t^n) \pm \left.\frac{\partial u}{\partial x}\right|_j^n \Delta x + \left.\frac{\partial^2 u}{\partial x^2}\right|_j^n \frac{\Delta x^2}{2} \pm$$

$$\pm \left.\frac{\partial^3 u}{\partial x^3}\right|_j^n \frac{\Delta x^3}{6} + \left.\frac{\partial^4 u}{\partial x^4}\right|_j^n \frac{\Delta x^4}{24} + O(\Delta x^5)$$

donde el símbolo $O(k^m)$ significa términos de orden m (o superior) en el incremento k. Reemplazando las Ecs. (2.4) y (2.5) en la Ec. (2.3) se obtiene

$$(2.6)\quad \epsilon_j^n = \left.\frac{\partial^2 u}{\partial t^2}\right|_j^n \frac{\Delta t}{2} - \nu \left.\frac{\partial^4 u}{\partial x^4}\right|_j^n \frac{\Delta x^2}{12} + O(\Delta t^2, \Delta x^4)$$

Una estimación del error de discretización estará dada, en general, por los dos primeros términos de la Ec. (2.6).

Una primera interpretación del error de discretización como error de truncamiento de la solución numérica puede hacerse a partir de la introducción del concepto de solución analítica local. Esta se define, para cada nivel n, como la solución de la ecuación diferencial para $t \geq t^n$, cuando se toma como condición inicial el valor de la solución numérica en $t=t^n$. Es decir, denominando $\tilde{u}^n(x,t)$ a la solución analítica local, la condición inicial es

$$(2.7)\quad \tilde{u}^n(x_j,t^n) = u_j^n$$

para todos los nodos j (nótese que al no estar definidas las condiciones iniciales fuera de los puntos nodales, la solución $\tilde{u}^n$ no es única). Por su parte, la solución numérica local exacta $(\tilde{u}^n)_j^m$ para cada nivel n es la solución exacta (es decir, con precisión infinita) de la ecuación en diferencias cuando se parte de las mismas condiciones iniciales, es decir

$$(2.8)\quad (\tilde{u}^n)_j^n = u_j^n$$

La diferencia entre la solución analítica local y la solución numérica local exacta en el nivel de tiempos n+1 es el error de truncamiento local:

$$e_j^n = \tilde{u}^n(x_j, t^{n+1}) - (\tilde{u}^n)_j^{n+1} \tag{2.9}$$

que se representa esquemáticamente en la Fig. 2.1. La solución analítica puede ser desarrollada en serie de Taylor:

$$\tilde{u}^n(x_j, t^{n+1}) = u_j^n + \left.\frac{\partial \tilde{u}^n}{\partial t}\right|_j^n \Delta t + \left.\frac{\partial^2 \tilde{u}^n}{\partial t^2}\right|_j^n \frac{\Delta t^2}{2} + O(\Delta t^3) \tag{2.10}$$

donde se ha utilizado la Ec. (2.7). Por su parte, la solución numérica satisface la Ec. (2.2), es decir

$$(\tilde{u}^n)_j^{n+1} = (\tilde{u}^n)_j^n + \frac{\nu \Delta t}{\Delta x^2}\left[(\tilde{u}^n)_{j+1}^n - 2(\tilde{u}^n)_j^n + (\tilde{u}^n)_{j-1}^n\right] \tag{2.11}$$

o, de acuerdo a la Ec. (2.8),

$$(\tilde{u}^n)_j^{n+1} = u_j^n + \frac{\nu \Delta t}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) \tag{2.12}$$

Pero, atendiendo a la Ec. (2.7), los valores numéricos en el nivel n pueden ser reemplazados por la solución analítica local. Entonces, desarrollando ésta en serie de Taylor, la Ec. (2.12) conduce a

$$\begin{aligned}(\tilde{u}^n)_j^{n+1} = u_j^n + \frac{\nu \Delta t}{\Delta x^2}\Bigg\{&\left[u_j^n + \left.\frac{\partial \tilde{u}^n}{\partial x}\right|_j^n \Delta x + \left.\frac{\partial^2 \tilde{u}^n}{\partial x^2}\right|_j^n \frac{\Delta x^2}{2} + \right.\\ &+ \left.\left.\frac{\partial^3 \tilde{u}^n}{\partial x^3}\right|_j^n \frac{\Delta x^3}{6} + \left.\frac{\partial^4 \tilde{u}^n}{\partial x^4}\right|_j^n \frac{\Delta x^4}{24} + \left.\frac{\partial^5 \tilde{u}^n}{\partial x^5}\right|_j^n \frac{\Delta x^5}{120} + O(\Delta x^6)\right] -\\ &- 2u_j^n + \left[u_j^n - \left.\frac{\partial \tilde{u}^n}{\partial x}\right|_j^n \Delta x + \left.\frac{\partial^2 \tilde{u}^n}{\partial x^2}\right|_j^n \frac{\Delta x^2}{2} - \right.\\ &- \left.\left.\frac{\partial^3 \tilde{u}^n}{\partial x^3}\right|_j^n \frac{\Delta x^3}{6} + \left.\frac{\partial^4 \tilde{u}^n}{\partial x^4}\right|_j^n \frac{\Delta x^4}{24} - \left.\frac{\partial^5 \tilde{u}^n}{\partial x^5}\right|_j^n \frac{\Delta x^5}{120} + O(\Delta x^6)\right] =\\ &= u_j^n + \nu \Delta t\left[\left.\frac{\partial^2 \tilde{u}^n}{\partial x^2}\right|_j^n + \frac{\Delta x^2}{12}\left.\frac{\partial^4 \tilde{u}^n}{\partial x^4}\right|_j^n + O(\Delta x^4)\right]\end{aligned} \tag{2.13}$$

Introduciendo las Ecs. (2.10) y (2.13) en la Ec. (2.9), y teniendo en cuenta que $\tilde{u}^n$ satisface la Ec. (2.1), se obtiene

$$e_j^n = \Delta t\left[\left.\frac{\partial^2 \tilde{u}^n}{\partial t^2}\right|_j^n \frac{\Delta t}{2} - \nu \left.\frac{\partial^4 \tilde{u}^n}{\partial x^4}\right|_j^n \frac{\Delta x^2}{12} + O(\Delta t^2, \Delta x^4)\right] \tag{2.14}$$

Finalmente, admitiendo que la diferencia entre la solución analítica global u(x,t) y la local $\tilde{u}^n(x,t)$ es de primer orden, y comparando las Ecs. (2.6) y (2.14), se tiene que

$$e_j^n \approx \Delta t\, \epsilon_j^n \tag{2.15}$$

es decir que el error de discretización y el error de truncamiento local difieren, esencialmente, en un factor Δt. En otras palabras, el primero es una medida del segundo.

Una segunda interpretación del error de discretización lo liga al error de truncamiento global (ver Fig. 2.1), definido como

$$E_j^n = u(x_j,t^n) - \hat{u}_j^n \tag{2.16}$$

donde $\hat{u}_j^n$ es la solución numérica exacta, es decir, la que satisface la Ec. (2.2) si se calcula con precisión infinita. Restando la Ec. (2.2) a la (2.3), y utilizando la definición (2.16), se obtiene

$$E_j^{n+1} = (1-2r)E_j^n + r(E_{j+1}^n + E_{j-1}^n) + \Delta t\, \epsilon_j^n \tag{2.17}$$

donde $r=\nu\,\Delta t/\Delta x^2$. De la Ec. (2.17) surge que

$$|E_j^{n+1}| \leq (|1-2r|+2r) \max_{\forall k} |E_k^n| + \Delta t \max_{\forall k} |\epsilon_k^n| \tag{2.18}$$

Entonces, si $r \leq 1/2$, la Ec. (2.18) conduce a

$$|E_j^{n+1}| \leq \max_{\forall k} |E_k^n| + \Delta t \max_{\forall k} |\epsilon_k^n| \tag{2.19}$$

La Ec. (2.19) puede utilizarse recurrentemente hacia atrás en el tiempo, obteniéndose

$$|E_j^{n+1}| \leq \Delta t \sum_{m=0}^{n} \max_{\forall k} |\epsilon_k^m| \tag{2.20}$$

donde se ha tenido en cuenta que $E_j^0 = 0$. De la Ec. (2.20) surge que

$$|E_j^{n+1}| \leq (n+1)\Delta t \max_{\substack{\forall k \\ 0\leq m\leq n}} |\epsilon_k^m| = t^{n+1} \max_{\substack{\forall k \\ 0\leq m\leq n}} |\epsilon_k^m| \tag{2.21}$$

La Ec. (2.21) muestra que el error de discretización (en rigor, su magnitud máxima) es una medida del error de truncamiento global, con un factor de proporcionalidad dado por el tiempo transcurrido. Si bien, como se vio más arriba, esta relación está sujeta a la condición $r \leq 1/2$, se mostrará más adelante que esta restricción es, de todos modos, necesaria para que el cálculo permanezca estable.

2.2. Definición de consistencia

Se dice que un esquema numérico es consistente con una ecuación diferencial, si el error de discretización tiende a anularse cuando disminuyen continuamente los intervalos de discretización, es decir,

$$\epsilon_j^n \xrightarrow[\substack{\Delta t \to 0 \\ \Delta x \to 0}]{} 0 \qquad (2.22)$$

De acuerdo a la interpretación dada (en la sección anterior) al error de discretización en términos del error de truncamiento global, la condición de consistencia es necesaria (aunque no suficiente) para que la solución numérica converja a la solución analítica cuando se afina la malla de cálculo.

Para el ejemplo desarrollado en la sección anterior, la Ec. (2.6) muestra claramente que la ecuación en diferencias (2.2) es consistente con la ecuación diferencial (2.1). Pero no siempre el tema es así de sencillo. Si a la Ec. (2.1) se la aproxima por el esquema de DuFort-Frankel

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = \nu \frac{u_{j+1}^n - u_j^{n+1} - u_j^{n-1} + u_{j-1}^n}{\Delta x^2} \qquad (2.23)$$

el error de discretización es (3)

$$\epsilon_j^n = \nu \left(\frac{\Delta t}{\Delta x}\right)^2 \frac{\partial^2 u}{\partial t^2}\bigg|_j^n + O\left[\Delta t^2, \Delta x^2, \left(\frac{\Delta t}{\Delta x}\right)^2 \Delta t^2\right] \qquad (2.24)$$

La Ec. (2.24) muestra que hay consistencia solo si

$$\left(\frac{\Delta t}{\Delta x}\right) \xrightarrow[\substack{\Delta t \to 0 \\ \Delta x \to 0}]{} 0 \qquad (2.25)$$

Es decir que, por ejemplo, se puede tomar $\Delta t \sim \Delta x^2$. En cambio, si el límite se realiza manteniendo el cociente $\beta = \Delta t/\Delta x$ constante, el esquema (2.23) resulta consistente con la ecuación diferencial

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} + \nu\beta \frac{\partial^2 u}{\partial t^2} = 0 \qquad (2.26)$$

2.3. Orden de precisión de un esquema numérico

Se denomina orden de precisión de un esquema numérico en cada una de las variables independientes, a las potencias dominantes de los intervalos de discretización correspondientes a cada variable, tal cual aparecen en el error de discretización. El orden de precisión mide con qué ritmo se anula el error de discretización cuando se afina la malla.

Por ejemplo, el esquema (2.2) es, de acuerdo a la Ec. (2.6), de orden de precisión 1 en el tiempo y 2 en el espacio.

Nótese que, de acuerdo a la Ec. (2.21), el orden de precisión del esquema es tambien el orden de precisión de la solución numérica respecto de la analítica.

En el caso del esquema de DuFort-Frankel, Ec. (2.23), y tomando $\Delta t \sim \Delta x^2$ para que haya consistencia, el orden de precisión resulta, de acuerdo a la Ec. (2.24), 2 en el espacio y en el tiempo. Esto significa, obviamente, que este esquema debería resultar más preciso que el anterior.

Es interesante mostrar que podría realizarse un subterfugio para hacer más preciso al esquema (2.2). En efecto, diferenciando la Ec. (2.1) pueden obtenerse las siguientes relaciones

$$\frac{\partial^2 u}{\partial t^2} = \nu \frac{\partial^3 u}{\partial x^2 \partial t} \tag{2.27}$$

$$\frac{\partial^3 u}{\partial x^2 \partial t} = \nu \frac{\partial^4 u}{\partial x^4} \tag{2.28}$$

Combinando las Ecs. (2.27) y (2.28) resulta que

$$\frac{\partial^2 u}{\partial t^2} = \nu^2 \frac{\partial^4 u}{\partial x^4} \tag{2.29}$$

Introduciendo la Ec. (2.29) en la Ec. (2.6) se obtiene

$$\epsilon_j^n = \frac{\nu}{2} \frac{\partial^4 u}{\partial x^4}\Big|_j^n \left(\nu \Delta t - \frac{\Delta x^2}{6}\right) + O(\Delta t^2, \Delta x^4) \tag{2.30}$$

Entonces, eligiendo $\Delta t = \nu \Delta x^2 / 6$, la Ec. (2.30) muestra que los términos dominantes del error de discretización se anulan, quedando un esquema de orden 2 en el tiempo y 4 en el espacio.

CAPITULO 3

CONVERGENCIA DE UNA SOLUCION NUMERICA

3.1. Definición de convergencia

Se dice que una solución numérica converge a la solución analítica cuando, para cada punto (en el espacio de las variables independientes), la primera tiende a la segunda al refinar la grilla de cálculo, es decir

$$(3.1)\quad u_j^n \xrightarrow[\substack{\Delta t\to 0\\ \Delta x\to 0\\ x_j\ fijo\\ t^n\ fijo}] \quad u(x_j,t^n)$$

Nótese que al efectuar el paso al límite indicado en la Ec. (3.1) los índices j y n deben variar (específicamente, $j,n\to\infty$) para que x_j y t^n permanezcan fijos (por ejemplo, $x_j = j\,\Delta x$ y $t^n = n\,\Delta t$).

Para ilustrar el concepto se estudiará el problema (2.1). Su solución analítica puede ser expresada en términos de una serie de Fourier (3)

$$(3.2)\quad u(x,t) = \sum_{m=-\infty}^{\infty} A_m\, e^{-m^2\nu t}\, e^{imx}$$

donde los coeficientes A_m pueden calcularse a partir de las condiciones iniciales. Ahora, para la ecuación en diferencias (2.2) puede proponerse un desarrollo similar

$$(3.3)\quad u_j^n = \sum_{m=-\infty}^{\infty} A_m\, \xi_m^n\, e^{imj\Delta x}$$

(los coeficientes A_m son los mismos que antes pues deben satisfacerse las mismas condiciones iniciales). Introduciendo la Ec. (3.3) en la (2.2) se obtiene que (3)

$$(3.4)\quad \xi_m = 1 - 2r\,(1-\cos m\Delta x)$$

donde, como antes, $r=\nu\Delta t/\Delta x^2$. Comparando las Ecs. (3.2) y (3.3), y teniendo en cuenta (3.1), se concluye que la solución numérica convergerá a la analítica solo si

$$(3.5)\quad \xi_m^n \xrightarrow[\substack{\Delta t\to 0\\ \Delta x\to 0\\ j\Delta x\ fijo\\ t=n\Delta t\ fijo}] \quad e^{-m^2\nu t}$$

Una restricción para que se verifique la relación (3.5) surge de reconocer que el miembro de la derecha disminuye cuando t aumenta, cualquiera sea el valor de m, mientras que el de la izquierda puede crecer indefinidamente con n a menos que

$$(3.6) \qquad |\xi_m| \leq 1$$

para todo valor de m. La Ec. (3.4) muestra que $\xi_m \leq 1$ siempre, de modo que la relación (3.6) solo requiere que $\xi_m > -1$, lo cual conduce a

$$(3.7) \qquad r \leq \frac{1}{1 - \cos m\Delta x}$$

Como la relación (3.7) debe verificarse para todos los valores posibles de m, se concluye que r debe ser menor o igual que el mínimo valor que puede tomar el miembro de la derecha, es decir,

$$(3.8) \qquad r \leq \frac{1}{2}$$

La restricción (3.8) ya había sido hallada por otro camino en la sección 2.1.

Ahora puede demostrarse que la relación (3.5) efectivamente se verifica. Teniendo en cuenta que $n=t/\Delta t=\nu t/r\,\Delta x^2$ se tiene

$$(3.9) \qquad \ln \xi_m^n = \frac{\nu t}{r} \frac{\ln [1 - 2r(1-\cos m\Delta x)]}{\Delta x^2}$$

Si en la Ec. (3.9) se toma el límite para $\Delta x \to 0$ usando la regla de L'Hospital, y se considera que r permanece constante, se obtiene

$$(3.10) \qquad \ln \xi_m^n \to -m\nu t \frac{\operatorname{sen} m\Delta x}{\Delta x} \frac{1}{[1-2r(1-\cos m\Delta x)]} \to -m^2\nu t$$

que es lo que se quería demostrar.

3.2. Análisis de la convergencia

Para datos de entrada determinados (es decir, haciendo abstracción de errores inherentes), la solución numérica de un problema (calculada con la computadora) diferirá de la solución analítica debido a dos causas: el error de truncamiento y los errores de redondeo. Que la solución numérica sea convergente significa que, al afinar la malla de cálculo, el error de truncamiento disminuye, mientras los errores de redondeo permanecen bajo control.

Está claro que el procedimiento utilizado en la sección anterior para demostrar la convergencia no es aplicable en general, ya que ello demandaría conocer las soluciones analítica y numérica exacta (lo cual no solo no es posible en general, sino que tornaría ocioso hallar la solución numérica). En consecuencia, es necesario disponer de otros medios de análisis de la convergencia.

La consistencia del esquema numérico es necesaria para que el error de truncamiento efectivamente disminuya al refinar la malla. Si bien como criterio resulta, entonces, incompleto, tiene la ventaja de que su determinación es automática. Para complementar el análisis, es necesario demostrar que el problema en diferencias permanece estable. Estos conceptos se expresan en el teorema de equivalencia de Lax, válido para ecuaciones diferenciales lineales con coeficientes constantes. El teorema expresa que (3):

> Dado un problema de valores iniciales bien planteado y una aproximación en diferencias finitas que satisface la condición de consistencia, la estabilidad es una condición necesaria y suficiente para la convergencia.

Se necesita, entonces, desarrollar métodos de análisis de la estabilidad de ecuaciones en diferencias, lo cual constituye el tópico central de los próximos dos capítulos. Allí se verá que la restricción (3.8) es, efectivamente, una condición para la estabilidad del esquema numérico en cuestión.

"...Lo que había en la mente de Arquimedes era diferente de lo que había en la de Newton, y ésto, a su vez, diferia de lo que había en la de Gauss. No es solo una cuestión de 'más', es decir que Gauss sabia más matemáticas que Newton quien, a su vez, sabia más que Arquimedes. Tambien es una cuestión de 'diferente'. El estado actual de conocimientos está entrelazado en una red de motivaciones y aspiraciones diferentes y de interpretaciones y potencialidades tambien diferentes..."

P.J. Davis, R. Hersh
"The Mathematical Experience"
Birkhäuser Boston, 1981

CAPITULO 4

ESTABILIDAD DE PROBLEMAS DE VALORES INICIALES

4.1. Definición de estabilidad

El concepto de estabilidad se discutió, en términos generales, en el capitulo 1. Aqui se presentarán conceptos y metodologias para esudiar la estabilidad del problema en diferencias (la estabilidad de algoritmos particulares solo se discutirá donde se considere pertinente). En el caso de problemas diferenciales, es necesario distinguir entre la estabilidad de la solución analitica, la estabilidad de la solución numérica y la estabilidad de la solución numérica respecto de la analitica. Obviamente, cuando el problema diferencial es estable, el esquema numérico que lo aproxima tambien debe serlo, en cuyo caso la estabilidad del segundo respecto del primero está asegurada. Sin embargo, aunque en la práctica es menos común, puede plantearse la cuestión de simular numéricamente problemas inestables. En este caso lo relevante es asegurar la estabilidad de la solución numérica respecto de la analitica, que resulta ser, entonces, el criterio más general.

Hablar de la estabilidad de la solución numérica respecto de la analitica, significa acotar de alguna manera la diferencia $|u(x_j, t^n) - u_j^n|$. En problemas de valores iniciales (que resultan en algoritmos iterativos) puede adoptarse la siguiente definición (3) : es estable si

$$|u(x_j,t^n) - u_j^n| \underset{\substack{n\to\infty \\ \Delta x,\Delta t \text{ fijos}}}{<} \infty \qquad (4.1)$$

es decir, la diferencia entre la solución analitica y numérica (para cada nodo del eje espacial) permanece acotada cuando se avanza indefinidamente el cálculo (en la dirección temporal), manteniendo fijos los intervalos de discretización. Una restricción de este tipo solo puede proveer, en principio, condiciones necesarias para la estabilidad.

En el caso de estar considerando un problema con solución (analitica) acotada, es decir

$$|u(x,t)| \underset{\substack{t\to\infty \\ x \text{ fijo}}}{<} \infty \qquad (4.2)$$

el criterio (4.1) se reduce a

$$(4.2a) \quad |u_j^n| \underset{\substack{n\to\infty \\ \Delta x, \Delta t \text{ fijos}}}{<} \infty$$

es decir que la solución numérica tambien debe permanecer acotada. Es fácil ver que, en el caso de ecuaciones lineales y homogéneas, las condiciones (4.2) y (4.2a) son necesarias para la estabilidad de las soluciones analítica y numérica, respectivamente. En efecto, si u'(x,t) es una solución perturbada respecto de u(x,t) (por ejemplo debido a un pequeño cambio en las condiciones iniciales), es necesario que

$$(4.2b) \quad |u'(x,t) - u(x,t)| \underset{\substack{t\to\infty \\ x \text{ fijo}}}{<} \infty$$

para que u sea estable. Pero u'-u satisface la misma ecuación diferencial que u (por ser lineal y homogénea), por lo cual ella misma debe verificar esa condición, es decir, debe cumplir con (4.2). La situación es idéntica con la solución numérica.

En la práctica, las inestabilidades se manifiestan, en general, en forma "explosiva", en el sentido de que la solución crece rápidamente en valor absoluto (ya sea en forma monótona u oscilatoria) hasta superar el rango de punto flotante de la computadora (produciendo un "overflow").

Obviamente, es vital disponer de métodos de análisis de la estabilidad de esquemas numéricos, de modo de poder detectar las eventuales fuentes de inestabilidades y, de ser posible, producir correcciones.

4.2. Método de von Neumann

En el caso de ecuaciones en diferencias lineales y homogéneas, es siempre posible plantear soluciones en términos de su serie de Fourier, tal como la de la Ec. (3.3). En este caso, para que se verifique la condición (4.2a) es necesario que ninguno de los armónicos crezca en forma no acotada con n. Esto se verifica siempre que

$$(4.3) \quad |\xi_m| \leq 1$$

para todo valor de m. La condición (4.3) coincide con la (3.6) y conduce, en el caso de la Ec. (2.2), a la restricción (3.8), que se repite a continuación

$$(4.4) \quad r = \frac{\nu \Delta t}{\Delta x^2} \leq \frac{1}{2}$$

La condición (4.4) se interpreta, en general, como una restricción sobre el valor máximo permisible de Δt, para un dado Δx, para que el cálculo permanezca estable. La Fig. 4.1, tomada de la Referencia (3), muestra como se manifiesta la inestabilidad cuando se toma $r > 1/2$, en el caso del problema (2.1) pero sometido a condiciones de contorno de Dirichlet homogéneas en los bordes, y con una forma inicial triangular.

Este método de análisis fue introducido por von Neumann, y puede ser extendido a ecuaciones no lineales e inhomogéneas (ver próximo capitulo).

Es interesante estudiar un problema algo más general que el (2.1), a saber

$$(4.5) \quad \frac{\partial u}{\partial t} + U \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \quad , \quad t \geq 0 \; , \; -\infty < x < \infty$$

donde U es una constante positiva (que representa una velocidad de convección). Se discretizará la Ec. (4.5) mediante el siguiente esquema explícito de segundo orden en el espacio:

$$(4.6) \quad \frac{u_j^{n+1} - u_j^n}{\Delta t} + U \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = \nu \frac{(u_{j+1}^n - 2u_j^n + u_{j-1}^n)}{\Delta x^2}$$

Tratándose, nuevamente, de una ecuación lineal y homogénea, puede plantearse una solución armónica, aunque más general que la Ec. (3.3):

$$(4.7) \quad u_j^n = \sum_{m=-\infty}^{\infty} A_m \, \xi_m^n \, e^{i k_m j \Delta x}$$

donde k_m es el número de ondas (no necesariamente entero, como en el caso del problema (2.1)). Introduciendo la Ec. (4.7) en la (4.6) se obtiene

$$(4.8) \quad \xi_m = 1 - 2r(1 - \cos k_m \Delta x) - i p \, \mathrm{sen}(k_m \Delta x)$$

donde $p = U \Delta t / \Delta x$. Nótese que el "factor de amplificación" ξ_m resulta complejo. Su módulo está dado por

$$(4.9) \quad |\xi_m|^2 = 1 - 4r(1 - \cos k_m \Delta x)[1 - r(1 - \cos k_m \Delta x)] + p^2 \mathrm{sen}^2(k_m \Delta x)$$

La condición (4.3) impone que

$$(4.10) \quad p^2 \mathrm{sen}^2(k_m \Delta x) \leq 4r(1 - \cos k_m \Delta x)[1 - r(1 - \cos k_m \Delta x)]$$

Una primera restricción surge a partir de (4.10) notando que el miembro de la izquierda no puede ser negativo. Entonces se debe cumplir que el factor entre corchetes del miembro de la derecha no debe ser negativo, lo cual conduce (como antes) a

(4.11) $$r \leq \frac{1}{2}$$

Ahora, retornando a la Ec. (4.10), y definiendo la variable

(4.12) $$\chi \equiv 1 - \cos k_m \Delta x ,$$

ésta puede reescribirse como

(4.13) $$p^2 \leq 4r f(\chi)$$

donde

(4.14) $$f(\chi) = \frac{1 - r\chi}{2 - \chi}$$

Como

(4.15) $$f'(\chi) = \frac{1 - 2r}{(2-\chi)^2}$$

la función f es monótonamente creciente, de acuerdo a la Ec. (4.11). Entonces, su valor mínimo lo toma en $\chi=0$, y la Ec. (4.13) implica que

(4.16) $$p^2 \leq 2r$$

o

(4.17) $$\Delta t \leq \frac{2\nu}{u^2}$$

que es una nueva restricción sobre el paso temporal para que el esquema permanezca estable.

Es interesante hacer un comentario sobre el esquema (4.6). En problemas donde la convección es dominante ($\nu \to 0$), la restricción (4.17) puede tornar impráctico el cálculo, ya que las escalas de movimiento temporales pueden ser órdenes de magnitud más grandes que el paso temporal requerido. Más aún, en ausencia de difusión ($\nu=0$), la condición (4.17) no se puede cumplir, es decir, el esquema se torna incondicionalmente inestable. En la próxima sección se verá una manera de corregir esta deficiencia.

4.3. Método de acotamiento

En el caso de métodos explícitos, se puede intentar obtener condiciones para acotar el crecimiento de la solución, lo cual siempre es posible en ecuaciones lineales y homogéneas. El procedimiento es análogo al utilizado en la sección 2.1. Por ejemplo, la Ec. (4.6) puede reescribirse como

$$u_j^{n+1} = (1-2r)u_j^n + (r+\tfrac{p}{2})u_{j-1}^n + (r-\tfrac{p}{2})u_{j+1}^n \tag{4.18}$$

De la Ec. (4.18) surge que

$$|u_j^{n+1}| \leq |1-2r|\,|u_j^n| + (r+\tfrac{p}{2})|u_{j-1}^n| + |r-\tfrac{p}{2}|\,|u_{j+1}^n| \leq$$
$$\leq (|1-2r| + (r+\tfrac{p}{2}) + |r-\tfrac{p}{2}|) \max_{\forall k} |u_k^n| \tag{4.19}$$

para todo valor de j. Entonces, si

$$1-2r \geq 0 \tag{4.20}$$

y

$$r-\frac{p}{2} \geq 0 \tag{4.21}$$

la Ec. (4.19) se reduce a

$$|u_j^{n+1}| \leq \max_{\forall k} |u_k^n| \tag{4.22}$$

es decir, la solución se "achata" a medida que avanza el cálculo, por lo cual permanece acotada. La condición (4.20) es equivalente a la (4.11). En cambio, la (4.21) conduce a

$$\Delta x \leq \frac{2\nu}{U} \tag{4.23}$$

Para compararla con la (4.17), la (4.23) puede reescribirse como

$$\Delta t \leq \frac{4r\nu}{U^2} \tag{4.24}$$

que, salvo para r=1/2 (en que coincide con la anterior), resulta algo más restrictiva que la (4.17). Si bien, entonces, este método puede resultar en condiciones más restrictivas que el de von Neumann, es mucho más simple de aplicar.

En la sección anterior se discutió sobre una deficiencia básica del esquema (4.6) para problemas en los cuales la convección es dominante. El remedio más simple para corregirla consiste en descentrar el término convectivo, tomando para $\partial u/\partial x$

positivo o negativo, respectivamente. Como ésto significa "traer" el valor de la derivada desde la zona de donde apunta la velocidad, esta forma de discretizar se denomina "de aguas arriba" ("upwinding"). Nótese que el precio que se paga es el de una disminución del order de precisión espacial del esquema (de 2 a 1). Se mostrará que esta técnica corrige, efectivamente, la deficiencia apuntada, utilizando el método de acotamiento introducido más arriba.

Admitiendo que U es positivo, un esquema de aguas arriba para el problema (4.5) es

$$(4.25) \quad \frac{u_j^{n+1} - u_j^n}{\Delta t} + U \frac{u_j^n - u_{j-1}^n}{\Delta x} = \nu \frac{(u_{j+1}^n - 2u_j^n + u_{j-1}^n)}{\Delta x^2}$$

que puede reescribirse como

$$(4.26) \quad u_j^{n+1} = (1 - p - 2r) u_j^n + (p + r) u_{j-1}^n + r u_{j+1}^n$$

De la Ec. (4.26) surge que

$$(4.27) \quad |u_j^{n+1}| \leq (|1 - p - 2r| + p + 2r) \max_{\forall k} |u_k^n|$$

para todo valor de j. Entonces, si $1-p-2r \geq 0$, la Ec. (4.27) se reduce a la (4.22), y se verifica el criterio de estabilidad. La única restricción es

$$(4.28) \quad \Delta t \leq \frac{\Delta x}{U + \frac{2\nu}{\Delta x}}$$

que no limita inaceptablemente el valor de Δt cuando $\nu \to 0$, ya que, en este caso, la Ec. (4.28) se reduce a

$$(4.29) \quad \Delta t \leq \frac{\Delta x}{U}$$

que es la denominada condición de Courant para problemas hiperbólicos. Esta condición puede interpretarse en términos de un criterio general, que se denominará criterio de Courant, y que establece que el dominio de influencia (respecto de un punto) de la solución analítica debe de estar contenido en el de la solución numérica, tal cual se ilustra en la Fig. 4.2. La violación de este criterio produce inestabilidades de tipo oscilatorio.

4.4. Método de Hirt

La idea básica del método de Hirt (4) parte de reconocer que, en el caso de un problema diferencial matemáticamente estable, la inestabilidad del problema en diferencias se debe a la influencia del error de truncamiento. En consecuencia, ésta puede explicarse analizando ese error. El detalle del procedimiento se introducirá con un ejemplo

Sea el problema (4.5), y su aproximación mediante el esquema (4.6). Para determinar cuál es la "verdadera" ecuación diferencial que satisface la solución numérica, puede hacerse un desarrollo formal en serie de Taylor, con el resultado

$$(4.30)\quad \frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} + U\frac{\Delta x^2}{3}\frac{\partial^3 u}{\partial x^3} - \nu\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4} + O(\Delta t^2, \Delta x^4) = 0$$

Nótese que los términos extra, que cuantifican el error de truncamiento, son formalmente análogos a los del error de discretización. Ahora bien, a primer orden (es decir, despreciando términos de orden mayor o igual a 2) la Ec. (4.30) es hiperbólica. Sus dos familias de curvas características tienen pendientes dadas por (5)

$$(4.31)\quad \frac{dx}{dt} = \pm\left(\frac{2\nu}{\Delta t}\right)^{1/2}$$

Dado que el dominio de influencia de la solución numérica es el segmento comprendido entre las rectas $\pm\Delta x/\Delta t$, la aplicación del criterio de Courant conduce a la condición

$$(4.32)\quad \left(\frac{2\nu}{\Delta t}\right)^{1/2} \leq \frac{\Delta x}{\Delta t}$$

que es equivalente a la (4.11). Su violación produce, entonces, inestabilidades de tipo oscilatorio.

Más información sobre el efecto del error de truncamiento puede obtenerse expresándolo solo en términos de derivadas espaciales, las cuales pueden interpretarse como mecanismos de atenuación o amplificación. Este tipo de técnica ya se utilizó en la sección 2.3. De la Ec. (4.30) surge que:

$$(4.33)\quad \frac{\partial^2 u}{\partial t^2} + U\frac{\partial^2 u}{\partial x \partial t} - \nu\frac{\partial^3 u}{\partial x^2 \partial t} + O(\Delta t, \Delta x^2) = 0$$

$$(4.34)\quad \frac{\partial^2 u}{\partial x \partial t} + U\frac{\partial^2 u}{\partial x^2} - \nu\frac{\partial^3 u}{\partial x^3} + O(\Delta t, \Delta x^2) = 0$$

$$(4.35)\quad \frac{\partial^3 u}{\partial x^2 \partial t} + U\frac{\partial^3 u}{\partial x^3} - \nu\frac{\partial^4 u}{\partial x^4} + O(\Delta t, \Delta x^2) = 0$$

Combinando las Ecs. (4.33) a (4.35) se llega a

$$(4.36)\quad \frac{\partial^2 u}{\partial t^2} = U^2\frac{\partial^2 u}{\partial x^2} - 2\nu U\frac{\partial^3 u}{\partial x^3} + \nu^2\frac{\partial^4 u}{\partial x^4} + O(\Delta t, \Delta x^2)$$

Introduciendo la Ec. (4.36) en la (4.30) se obtiene

$$(4.37)\quad \frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} = \left(\nu - \frac{U^2\Delta t}{2}\right)\frac{\partial^2 u}{\partial x^2} + \nu U\Delta t\,\frac{\partial^3 u}{\partial x^3} - \frac{\nu^2\Delta t}{2}\,\frac{\partial^4 u}{\partial x^4} + O(\Delta t^2, \Delta x^2)$$

que se denomina ecuación modificada (6). El primer término del segundo miembro de la Ec. (4.37) representa un mecanismo de difusión. Para que el coeficiente sea positivo debe verificarse la condición (4.17). En caso contrario se producirá un crecimiento monótono de la solución.

Se observa, entonces, que con el método de Hirt pudieron obtenerse las mismas condiciones de estabilidad que con el método de von Neumann. Sin embargo, no siempre es así. Sucede a menudo que el primero da menos condiciones que el segundo, ya que el método de Hirt, al considerar solo las derivadas de orden menor, limita su búsqueda a las mayores escalas espaciales y temporales. Como compensación, el método de Hirt puede detectar inestabilidades provenientes de mecanismos no lineales, lo cual es imposible por principio por el método de von Neumann (ver próximo capítulo).

4.5. Estabilidad de esquemas implícitos

Hasta ahora, todas las ejemplificaciones se hicieron sobre esquemas numéricos explícitos. Estos, en general, requieren menos operaciones por paso de tiempo que los implícitos. Por su parte, los implícitos suelen estar libres de algunos tipos de inestabilidades que afectan a los explícitos.

Sea el problema (2.1) y el siguiente método numérico implícito ponderado:

$$(4.38)\quad \frac{u_j^{n+1} - u_j^n}{\Delta t} = \nu\left[\theta\,\frac{(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1})}{\Delta x^2} + (1-\theta)\,\frac{(u_{j+1}^{n} - 2u_j^{n} + u_{j-1}^{n})}{\Delta x^2}\right]$$

donde θ es un coeficiente tal que $0 \leqslant \theta \leqslant 1$. Utilizando el método de von Neumann se obtiene el siguiente factor de amplificación (3):

$$(4.39)\quad \xi_m = \frac{1 - 2r(1-\theta)(1-\cos m\Delta x)}{1 + 2r\theta(1-\cos m\Delta x)}$$

que, obviamente, se reduce a (3.4) cuando $\theta=0$. La Fig. 4.3 muestra la variación de ξ_m (que es un número real) con la variable $y \equiv 2r(1-\cos m\Delta x)$, para distintos valores de θ. Se observa que $|\xi_m| \leqslant 1$, para todo valor de m, si $\theta \geqslant 1/2$. Es decir que el esquema resulta incondicionalmente estable para $\theta \geqslant 1/2$. En caso contrario, es decir para $0 \leqslant \theta < 1/2$, siempre decrece por abajo del valor -1 para valores grandes de y. Esto conduce a la restricción

$$(4.40) \qquad r \leq \frac{1}{1-2\theta}$$

que, obviamente, se reduce a (4.11) para $\theta=0$.

Debe tenerse en cuenta que los análisis efectuados se dirigen a determinar la estabilidad del problema en diferencias. A ello debe agregarse, aún, el tema de la estabilidad del algoritmo particular mediante el cual se implementa la resolución de ese problema. Esta cuestión es especialmente relevante en el caso de esquemas numéricos implícitos, que conducen a sistemas de ecuaciones acopladas. En general, en estos casos resultan matrices de coeficientes del tipo banda, que pueden ser resueltas eficientemente por el método de eliminación de Gauss. Las condiciones para que este algoritmo permanezca estable puede consultarse en la bibliografía (1).

El ejemplo anterior muestra que pueden construirse esquemas implícitos incondicionalmente estables. Si bien ésto puede resultar "cómodo" desde el punto de vista práctico, al no tener que limitar el valor de Δt, es necesario reconocer que condiciones del tipo de la (4.40) tambien resultan necesarias para garantizar la precisión de la solución numérica. En efecto, en el problema de difusión que se está considerando, los efectos difusivos se "sienten" (es decir, se hacen significativos) sobre una distancia Δx en tiempos del orden de $\Delta x^2/\nu$, lo cual significa que r debe ser del orden de 1 para seguir con precisión la marcha de la solución. Esto significa que no puede elegirse un paso temporal Δt demasiado distinto al que resultaría para un esquema explícito, lo cual cuestiona la aparente ventaja del método implícito, ya que éste requiere, en general, muchas más operaciones por paso de tiempo.

La ventaja manifiesta de los esquemas implícitos sobre los explícitos es en problemas que involucran escalas de tiempo disímiles, cuando están activadas solo las escalas mayores. En efecto, en esta situación mantener la precisión requiere calcular con un paso temporal pequeño frente a las escalas activas. En cambio, para mantener la estabilidad de un esquema explícito se necesita restringir el paso en relación a las escalas menores, aunque no estén activas. Un ejemplo clásico es el problema de traslación de ondas en ríos, que se describe por medio de las Ecuaciones de Saint Venant (7,8). En ese problema conviven escalas rápidas, caracterizadas por la celeridad de Lagrange $(gh)^{1/2}$ (g es la gravedad y h la profundidad) con escalas lentas, caracterizadas por la velocidad del agua u. Su relación $F=u/(gh)^{1/2}$ es el número de Froude del escurrimiento. Las ondas de inundación provocadas por crecidas naturales están asociadas a escalas lentas. Cuando $F \ll 1$, como sucede en ríos de llanura, las escalas son disímiles, y conviene utilizar un método implícito par describirlas, de modo de no estar limitados por las escalas rápidas.

4.6. Estabilidad de sistemas acoplados

Las técnicas de análisis de la estabilidad que se introdujeron en este capítulo pueden ser extendidas sin dificultad a sistemas de ecuaciones. Por ejemplo, sea el caso del siguiente sistema de ecuaciones diferenciales de tipo

hiperbólico, que representa ondas de gravedad de pequeña amplitud:

$$\frac{\partial h}{\partial t} + H \frac{\partial u}{\partial x} = 0 \qquad (4.41)$$

$$\frac{\partial u}{\partial t} + g \frac{\partial h}{\partial x} = 0 \qquad (4.42)$$

donde h(x,t) y u(x,t) son las funciones incógnita, y g y H constantes positivas. Su discretización por medio de un esquema explícito centrado da

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + H \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \qquad (4.43)$$

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + g \frac{h_{j+1}^n - h_{j-1}^n}{2\Delta x} = 0 \qquad (4.44)$$

Para aplicar el método de von Neumann se toma

$$\begin{bmatrix} h \\ u \end{bmatrix}_j^n = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}^n e^{ikj\Delta x} \qquad (4.45)$$

Introduciendo las Ecs. (4.45) en las Ecs. (4.43)-(4.44) se obtiene

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}^{n+1} = \begin{bmatrix} 1 & -isH\,\mathrm{sen}\,k\Delta x \\ -isg\,\mathrm{sen}\,k\Delta x & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}^n \qquad (4.46)$$

donde $s \equiv \Delta t/\Delta x$. La matriz de coeficientes de las Ecs. (4.46) se denomina "matriz de amplificación" y se la denotará como G. Para poder extender el criterio (4.3) al caso de un sistema, como el presente, es necesario diagonalizar G, de modo de que ese criterio pueda ser aplicado a cada elemento de la diagonal, es decir, de cada autovalor. Los autovalores λ surgen de la ecuación

$$\begin{vmatrix} 1-\lambda & -isH\,\mathrm{sen}\,k\Delta x \\ -isg\,\mathrm{sen}\,k\Delta x & 1-\lambda \end{vmatrix} = 0 \qquad (4.47)$$

que conduce a los dos autovalores

$$\lambda_{1,2} = 1 \pm is(gH)^{1/2}\,\mathrm{sen}\,k\Delta x \qquad (4.48)$$

La Ec. (4.48) muestra que $|\lambda_1| = |\lambda_2| > 1$, de modo que el sistema (4.43)-(4.44) es incondicionalmente inestable.

Ante esta dificultad, Lax introdujo un esquema alternativo, que lleva su nombre, y que consiste en efectuar los siguientes reemplazos en las Ecs. (4.43)-(4.44): $h_j^n \longrightarrow 1/2\ (h_{j-1}^n + h_{j+1}^n)$, $u_j^n \longrightarrow 1/2(u_{j-1}^n + u_{j+1}^n)$. De esta manera se conserva el orden de precisión del esquema, resultando la siguiente matriz de amplificación:

$$(4.49) \quad G = \begin{bmatrix} \cos k\Delta x & -is\sqrt{H}\,\mathrm{sen}\, k\Delta x \\ -is\sqrt{g}\,\mathrm{sen}\, k\Delta x & \cos k\Delta x \end{bmatrix}$$

cuyos autovalores son

$$(4.50) \quad \lambda_{1,2} = \cos k\Delta x \pm is(gH)^{1/2}\,\mathrm{sen}\, k\Delta x$$

De esta forma, se tiene que

$$(4.51) \quad |\lambda_1|^2 = |\lambda_2|^2 = 1 - (1 - s^2 gH)\,\mathrm{sen}^2(k\Delta x)$$

que satisface el criterio de estabilidad solo si

$$(4.52) \quad \Delta t \leq \frac{\Delta x}{(gH)^{1/2}}$$

La Ec. (4.52), formalmente análoga a la (4.29), es la condición de Courant para este problema.

Una metodología similar puede utilizarse para analizar esquemas numéricos que involucran más de dos niveles de tiempo. Por ejemplo, el esquema de Richardson para la Ec. (2.1) es

$$(4.53) \quad \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = \nu\ \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

que es explícito y tiene un orden de precisión 2 en el espacio y en el tiempo (el esquema (2.2) solo tiene orden de precisión 1 en el tiempo). La Ec. (4.53) es una ecuación en diferencias de orden 2 en el tiempo, ya que involucra 3 niveles de tiempo. Mediante el artilugio de introducir la variable $v_j^n = u_j^{n-1}$, puede convertirse en un sistema de dos ecuaciones en diferencias de primer orden:

$$(4.54) \quad u_j^{n+1} = v_j^n + 2r(u_{j+1}^n - 2u_j^n + u_{j-1}^n)$$

$$(4.55) \quad v_j^{n+1} = u_j^n$$

A partir de este punto el análisis es como en el caso anterior. La matriz de amplificación es

$$(4.56) \qquad G = \begin{bmatrix} -4r\,\mathrm{sen}^2(k\Delta x/2) & 1 \\ 1 & 0 \end{bmatrix}$$

cuyos autovalores son

$$(4.57) \qquad \lambda_{1,2} = -2r\,\mathrm{sen}^2\frac{k\Delta x}{2} \pm \left(1 + 4r^2\,\mathrm{sen}^4\frac{k\Delta x}{2}\right)^{1/2}$$

Entonces

$$(4.58) \qquad |\lambda_{1,2}|^2 = 1 + 8r^2\,\mathrm{sen}^4\frac{k\Delta x}{2} \mp 4r\,\mathrm{sen}^2\frac{k\Delta x}{2}\left(1 + 4r^2\,\mathrm{sen}^4\frac{k\Delta x}{2}\right)^{1/2}$$

La Ec. (4.58) muestra que $|\lambda_2| \geq 1$, de modo que el esquema de Richardson es incondicionalmente inestable. Esto suele ocurrir cuando se utiliza para aproximar un esquema de orden mayor que el de la ecuación diferencial (en este ejemplo, la ecuación es de orden 1 en el tiempo, y el esquema es de orden 2). Lo que sucede es que aparecen componentes espurias (en este ejemplo, una) que, si son inestables, terminan dominando el comportamiento de la solución.

Debido a esta dificultad es que se desarrolló el esquema de DuFort-Frankel, Ec. (2.23), que tiene el mismo orden de precisión que el de Richardson. Su matriz de amplificación es

$$(4.59) \qquad G = \begin{bmatrix} \dfrac{4r\cos k\Delta x}{1+2r} & \dfrac{1-2r}{1+2r} \\ 1 & 0 \end{bmatrix}$$

que tienen autovalores

$$(4.60) \qquad \lambda_{1,2} = \frac{2r\cos k\Delta x \mp (1 - 4r^2\,\mathrm{sen}^2 k\Delta x)^{1/2}}{1+2r}$$

No es difícil demostrar que $|\lambda_{1,2}| \leq 1$, de modo que el esquema de DuFort-Frankel es incondicionalmente estable, a pesar de ser explícito.

4.7. Aproximación de problemas inestables

Cuando el problema diferencial es inestable, el esquema numérico que lo aproxima tambien debe serlo. En este caso el criterio (4.3) no resulta adecuado, sino que hay que relajarlo para permitir el crecimiento de la solución numérica. Si se pide que (3)

$$(4.61) \qquad |\xi_m| \leq 1 + O(\Delta t)$$

se esta admitiendo la posibilidad de un crecimiento exponencial legítimo. En efecto,

$$(4.62)\quad \ln\left[\,|\xi_m|^n\right] \xrightarrow[\substack{n\to\infty \\ \Delta t \text{ fijo}}]{} n \ln\left[1+O(\Delta t)\right] = O(n\Delta t)$$

es decir que

$$(4.63)\quad |\xi_m|^n \xrightarrow[\substack{n\to\infty \\ \Delta t \text{ fijo}}]{} e^{\alpha n \Delta t}$$

donde α es una constante.

Por ejemplo, para la ecuación

$$(4.64)\quad \frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} + bu$$

donde ν y b son constantes positivas, el esquema numérico

$$(4.65)\quad \frac{u_j^{n+1} - u_j^n}{\Delta t} = \nu \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + b u_j^n$$

tiene un factor de amplificación

$$(4.66)\quad \xi_m = 1 - 4 \frac{\nu \Delta t}{\Delta x^2} \operatorname{sen}^2 \frac{k_m \Delta x}{2} + b\Delta t$$

que satisface la condición (4.61) y, en consecuencia, provee una solución numérica estable respecto de la analítica.

4.8. Condiciones suficientes de estabilidad

Tal cual se ha hecho notar anteriormente, con las metodologías empleadas se obtienen condiciones necesarias de estabilidad. Sin embargo, en el caso de ecuaciones lineales y homogéneas, las condiciones obtenidas con el criterio de von Neumann (4.61) pueden ser tambien suficientes. Dos de esas situaciones son las siguientes:

i) cuando la ecuación en diferencias es de dos niveles (orden 1) y tiene solo una variable dependiente.

ii) cuando los elementos de la matriz de amplificación G están acotados, y todos sus autovalores caen en un círculo interior al círculo unidad ($|\lambda_i| < 1$), con la posible excepción de uno que satisface (4.61).

CAPITULO 5

ESTABILIDAD DE PROBLEMAS MAS COMPLEJOS

5.1. Problemas lineales de coeficientes variables

Si bien en el capítulo anterior se introdujeron muchas nociones generales sobre la estabilidad numérica de problemas de valores iniciales, se analizaron solo problemas lineales y homogéneos con coeficientes constantes.

En el caso en que los coeficientes son variables, es decir, dependen de las variables independientes (x,t), los métodos de análisis se aplican sin modificaciones, excepto que debe tenerse en cuenta explícitamente esa dependencia. Entonces, resultan condiciones de estabilidad locales, es decir, dependientes del punto particular de la grilla de cálculo. Por ejemplo, si en el problema (2.1) se supone que $\nu = \nu(x)$ y se aplica el esquema (2.2), con el reemplazo $\nu \to \nu_j$, la condición de estabilidad (4.4) sigue vigente, es decir

$$\Delta t \leq \frac{\Delta x^2}{2\nu_j} \tag{5.1}$$

para todo valor de j. Entonces, la condición más restrictiva es

$$\Delta t \leq \frac{\Delta x^2}{2 \max_{\forall k} (\nu_k)} \tag{5.2}$$

La presencia de coeficientes variables puede introducir, a veces, términos extras que imponen nuevas condiciones de estabilidad. Por ejemplo, una formulación más general del problema (2.1) es

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\nu(x)\frac{\partial u}{\partial x}\right], \quad t \geq 0, \quad -\infty < x < \infty \tag{5.3}$$

que puede aproximarse por el siguiente esquema explícito centrado

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{1}{\Delta x^2}\left[\nu_{j+1/2}(u_{j+1}^n - u_j^n) - \nu_{j-1/2}(u_j^n - u_{j-1}^n)\right] \tag{5.4}$$

La Ec. (5.4) puede reescribirse como

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \left(\frac{\nu_{j+1/2} - \nu_{j-1/2}}{\Delta x}\right)\left(\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}\right) = \left(\frac{\nu_{j+1/2} + \nu_{j-1/2}}{2}\right)\left(\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}\right) \tag{5.5}$$

que no es más que una aproximación centrada para

$$(5.6) \quad \frac{\partial u}{\partial t} - \frac{\partial \nu}{\partial x}\frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}$$

resultante de desarrollar la Ec. (5.3). Comparando las Ecs. (5.5) y (4.6) se concluye que, además de la condición (5.2), tambien debe verificarse una restricción del tipo (4.17), es decir

$$(5.7) \quad \Delta t \leq \frac{(\nu_{j+1/2} + \nu_{j-1/2})}{(\nu_{j+1/2} - \nu_{j-1/2})^2} \Delta x^2 \approx \left[\frac{2\nu}{(\partial \nu/\partial x)^2} \right]_j$$

5.2. Problemas no lineales

La no linealidad de un problema complica fuertemente el tema del análisis de la estabilidad, ya que introduce nuevas fuentes que no es sencillo detectar. El método de von Neumann todavia puede aplicarse, pero el análisis debe hacerse no sobre la ecuación en diferencias original, sino sobre la ecuación para una pequeña perturbación introducida en la solución numérica. Esta ecuación surge de la primera mediante un proceso de linealización.

Por ejemplo, una reformulación no lineal clásica del problema (4.5) es

$$(5.8) \quad \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \quad , \quad t \geq 0 \quad , \quad -\infty < x < \infty$$

que puede aproximarse con el esquema (4.25), si se hace el reemplazo $U \rightarrow u_j^n$ (y se supone que $u_j^n \geq 0$). Para linealizar esta ecuación en diferencias se hace el reemplazo

$$(5.9) \quad u_j^n \longrightarrow u_j^n + \epsilon_j^n$$

y se supone que ϵ_j^n es una corrección de primer orden a u_j^n. Entonces, utilizando el hecho de que u_j^n es solución de la ecuación en diferencias y conservando solo términos de orden 1 en la perturbación, se obtiene

$$(5.10) \quad \epsilon_j^{n+1} - \epsilon_j^n + s\left[u_j^n(\epsilon_j^n - \epsilon_{j-1}^n) + (u_j^n - u_{j-1}^n)\epsilon_j^n\right] = r(\epsilon_{j+1}^n - 2\epsilon_j^n + \epsilon_{j-1}^n)$$

donde $s = \Delta t/\Delta x$. Nótese que la Ec. (5.10) es lineal y homogénea en ϵ_j^n, por lo cual admite los mismos tipos de tratamientos vistos en el capitulo anterior. Si se admite que u_j^n varia en forma lo suficientemente suave como para que $u_j^n - u_{j-1}^n = O(\Delta x)$, puede despreciarse, en primera aproximación, el segundo término del corchete de la Ec. (5.10), por lo cual ésta queda reducida a una

forma análoga a la de la Ec. (4.26). Por lo tanto, la condición de estabilidad es formalmente análoga a la (4.28), a saber

$$(5.11) \qquad \Delta t \leq \frac{\Delta x}{u_j^n + \frac{2\nu}{\Delta x}}$$

para todo valor de j y n. La condición (5.11) es de tipo local, pero, a diferencia de la (5.1), depende de la solución numérica. Esto torna menos controlable su verificación. En el mejor de los casos, que es cuando de alguna manera puede acotarse el valor de u_j^n, puede resultar globalmente demasiado restrictiva. Esto revaloriza la utilidad de los métodos implícitos, los cuales, aunque siga vigente una condición similar a la (5.11) por razones de precisión, no se tornan inestables.

Está claro que el método de von Neumann no puede detectar todas las potenciales fuentes de inestabilidades. Las que sí identifica se denominan inestabilidades "de tipo lineal". Pero en problemas no lineales una fuente importante radica en la transferencia de "energía" de unas escalas a otra de movimiento. En particular, cuando esta transferencia se da desde las escalas mayores a las menores, y no existe un mecanismo eficaz de disipación a este nivel, se acumula energía en estas últimas (es decir, en las de longitud de onda $2\Delta x$). Esto otorga una aspecto "ruidoso" a la solución numérica, que puede estar preanunciando la eclosión de un proceso inestable. Esto suele suceder en problemas convectivos como el (5.8). En efecto, nótese que, en este caso, un armónico de número de ondas k produce, a través del término convectivo (no lineal), una contribución de la forma

$$(5.12) \qquad ik\, e^{2ikx}$$

que es un término fuente para el armónico 2k. De todos modos, en el problema (5.8) existe un mecanismo de disipación particularmente efectivo para escalas cortas, que es el término de difusión. Claro que, en problemas con un "número de Reynolds de grilla" alto, es decir, $u_j^n \Delta x/\nu \gg 1$, ese mecanismo puede resultar insuficiente. En esta situación puede ser necesario agregar "viscosidad artificial" (por ejemplo, un término difusivo extra con un coeficiente ν lo suficientemente grande como para disipar eficientemente la energía de las escalas más cortas, es decir, $\nu \sim u_j^n \Delta x$) o utilizar esquemas numéricos "disipativos", es decir, aquellos que introducen una difusión numérica significativa (ver próximo capítulo). Por supuesto, debe actuarse cuidadosamente de modo que la estabilidad no se logre a costa de una inaceptable pérdida de precisión.

El método de Hirt constituye una herramienta interesante para detectar las fuentes de inestabilidades de tipo no lineal, ya que no requiere una linealización previa. Sea, por ejemplo, el sistema de ecuaciones

$$(5.13) \qquad \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0$$

$$(5.14) \qquad \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + g\frac{\partial h}{\partial x} = 0$$

que constituyen las ecuaciones para aguas poco profundas en una dimensión espacial (con simetria en la segunda dirección espacial). Una posible discretización de la Ec. (5.13), suponiendo que u siempre es positivo, es el siguiente esquema de aguas arriba:

$$(5.15)\quad \frac{h_j^{n+1} - h_j^n}{\Delta t} + \frac{h_j^n u_{j+1/2}^n - h_{j-1}^n u_{j-1/2}^n}{\Delta x} = 0$$

Efectuando desarrollos en serie de Taylor alrededor del nodo (j,n) se obtiene

$$(5.16)\quad \frac{\partial h}{\partial t} + \frac{\Delta t}{2}\frac{\partial^2 h}{\partial t^2} + O(\Delta t^2) + \frac{1}{\Delta x}\left\{ h\left[u + \frac{\Delta x}{2}\frac{\partial u}{\partial x} + \frac{\Delta x^2}{8}\frac{\partial^2 u}{\partial x^2} + \right.\right.$$
$$\left. + \frac{\Delta x^3}{48}\frac{\partial^3 u}{\partial x^3} + O(\Delta x^4)\right] - \left[h - \Delta x\frac{\partial h}{\partial x} + \frac{\Delta x^2}{2}\frac{\partial^2 h}{\partial x^2} + O(\Delta x^3)\right] \times$$
$$\left. \times \left[u - \frac{\Delta x}{2}\frac{\partial u}{\partial x} + \frac{\Delta x^2}{8}\frac{\partial^2 u}{\partial x^2} - \frac{\Delta x^3}{48}\frac{\partial^3 u}{\partial x^3} + O(\Delta x^4)\right]\right\} = 0$$

Ahora, diferenciando las Ecs. (5.13) y (5.14) surgen las siguientes relaciones:

$$(5.17)\quad \frac{\partial^2 h}{\partial t^2} + u\frac{\partial^2 h}{\partial x \partial t} + \frac{\partial u}{\partial t}\frac{\partial h}{\partial x} + \frac{\partial h}{\partial t}\frac{\partial u}{\partial x} + h\frac{\partial^2 u}{\partial x \partial t} = 0$$

$$(5.18)\quad \frac{\partial^2 h}{\partial x \partial t} + u\frac{\partial^2 h}{\partial x^2} + 2\frac{\partial u}{\partial x}\frac{\partial h}{\partial x} + h\frac{\partial^2 u}{\partial x^2} = 0$$

$$(5.19)\quad \frac{\partial^2 u}{\partial x \partial t} + u\frac{\partial^2 u}{\partial x^2} + \left(\frac{\partial u}{\partial x}\right)^2 + g\frac{\partial^2 h}{\partial x^2} = 0$$

de donde se concluye que

$$(5.20)\quad \frac{\partial^2 h}{\partial t^2} = (u^2 + c^2)\frac{\partial^2 h}{\partial x^2} + \ldots$$

donde $c=(gh)^{1/2}$ y, en el miembro derecho, solo se han tenido en cuenta los términos de tipo difusivo. Operando en la Ec. (5.16), utilizando la Ec. (5.20) y colectando solo los términos difusivos, se obtiene el siguiente coeficiente de viscosidad numérica

$$(5.21)\quad \frac{\Delta x}{2}u - \frac{\Delta t}{2}(u^2 + c^2) - \frac{\Delta x^2}{4}\frac{\partial u}{\partial x} + \frac{\Delta x^3}{16}\frac{\partial^2 u}{\partial x^2} - \frac{\Delta x^4}{96}\frac{\partial^3 u}{\partial x^3}$$

La Ec. (5.21) indica toda una variedad de posibilidades que pueden generar una inestabilidad no lineal: un fuerte gradiente (positivo) de velocidades ($\partial u/\partial x$), una fuerte convexidad ($\partial^2 u/\partial x^2 < 0$) del perfil de velocidades, etc. Tambien muestra, a través del primer término, la influencia estabilizadora de la descentralización de la derivada. Hirt (4) presenta resultados del cálculo para un problema muy similar al (5.13)-(5.14), donde se ponen en evidencia, prácticamente, cada una de las inestabilidades posibles.

5.3. Problemas mixtos de valores iniciales y de contorno

Hasta ahora se han tratado problemas puros de valores iniciales, es decir, no se ha analizado la influencia que, sobre la estabilidad, pueden tener las condiciones de contorno. Estas, de hecho, suelen constituirse en nuevas fuentes de inestabilidades.

Desde el punto de vista del análisis armónico, las condiciones de contorno pueden generar nuevos modos de oscilación (en adición a los estudiados por el método de von Neumann), que decaen al alejarse del contorno. La técnica de análisis más sistemática, aunque engorrosa, de problemas mixtos es el método de la energía, en cuyos detalles no se entrará, pudiéndose consultar la bibliografía (3).

Resta mencionar que, al resolver el problema en diferencias que aproxima un dado problema diferencial, suele ocurrir que deben especificarse más condiciones de contorno que las que requiere el propio problema diferencial, para que la formulación cierre (es decir, para que haya el mismo número de ecuaciones que de incógnitas). Esas condiciones extra deben ser compatibles con el resto de la formulación o, al menos, producir una influencia débil sobre la solución numérica. En caso contrario, pueden disparar nuevas inestabilidades.

5.4. Problemas de valores de contorno

Los problemas puros de valores de contorno están asociados a ecuaciones diferenciales elípticas. Su discretización directa conduce a un sistema de ecuaciones algebraicas, eventualmente no lineales, fuertemente acoplado. Supuesto el problema estable desde el punto de vista analítico, la determinación de la estabilidad numérica está asociada al algoritmo utilizado para resolver el sistema de ecuaciones algebraicas.

Cuando el sistema es lineal puede resolverse, en principio, por métodos directos. El método directo más eficiente es la clásica eliminación de Gauss, con sus simplificaciones en el caso de tratarse de una matriz de coeficientes de estructura banda. Utilizando la técnica de refinamiento de la solución puede estimarse la estabilidad del proceso (1). La alternativa es recurrir a los métodos iterativos. El más utilizado es el método de Gauss-Seidel, eventualmente con sobre-relajación (SOR) para aumentar su velocidad de convergencia. La estabilidad se manifiesta a través de la convergencia del proceso iterativo.

Cuando el sistema de ecuaciones es no lineal debe recurrirse, necesariamente, a técnicas iterativas. Estas consisten, en general, en plantear en cada iteración un nuevo sistema lineal, el cual debe ser resuelto de acuerdo a lo discutido más arriba. La técnica más difundida es el método de Newton-Raphson. La estabilidad se refleja, entonces, en la convergencia del proceso iterativo.

Un planteo alternativo interesante para problemas de valores de contorno es transformarlo en un problema mixto, mediante el agregado de algún término con derivada temporal. Se define así un problema seudo-evolucionario que puede resolverse de acuerdo a las técnicas desarrolladas para problemas de valores iniciales, y cuya estabilidad puede analizarse por los métodos discutidos. No estando interesados en los resultados intermedios sino en la solución asintótica luego de muchos pasos de cálculo, el valor de Δt solo necesita ser restringido por razones de estabilidad (y no de precisión). En rigor, este método representa una manera de plantear técnicas iterativas, con una estimación inicial interpretada como condición inicial del problema seudo-evolucionario. Solo que, al interpretar el proceso iterativo como un proceso físico real, puede aprovecharse el conocimiento físico del problema para plantear procedimientos de cálculo más eficientes.

CAPITULO 6

PRECISION DE UNA SOLUCION NUMERICA

6.1. Definición de precisión

La precisión de una solución numérica u_j^n está medida por su diferencia con la solución analítica $u(x_j, t^n)$ para cada nodo (j,n) de la grilla de discretización. Está claro que esta diferencia incluye tanto el error de truncamiento, Ec. (2.16), como los de redondeo. Como se vio en el capítulo 2, la consistencia y estabilidad de un esquema numérico garantizan que esos errores están controlados. Sin embargo, controlar la precisión significa avanzar un paso más y cuantificar esa diferencia.

En rigor, en un problema estable estimar la precisión significa cuantificar el error de truncamiento, ya que los errores de redondeo no solo dan, en general, una contribución mucho menor, sino tambien que esa contribución tiene un carácter estocástico.

6.2. Difusión y dispersión numéricas

Dado que la precisión está íntimamente relacionada con el error de truncamiento, se puede inferir bastante acerca de aquella si se analiza el error de discretización. Para problemas de valores iniciales, este error puede expresarse solo en términos de derivadas espaciales, tal cual se vio en las secciones 2.3 y 4.4. Entonces, la ecuación modificada que satisface la solución numérica es de la forma

$$(6.1) \quad \mathcal{L}(u) + \sum_{m=1}^{\infty} \mu_m \frac{\partial^m u}{\partial x^m} = 0$$

donde $\mathcal{L}(u)$ simboliza la ecuación diferencial original, y los coeficientes μ_m dependen, al menos, de los pasos de discretización. La serie de la Ec. (6.1) (formalmente análoga al error de discretización) puede descomponerse en una serie para las derivadas de orden par y otra para las de orden impar. En problemas de valores iniciales de orden 1 en el tiempo, es decir, aquellos para los cuales el operador $\mathcal{L}$ solo contiene como derivada temporal a $\partial/\partial t$, la serie par representa términos de difusión, mientras que la impar está ligada a fenómenos de dispersión. Como el origen de estos téminos es puramente numérico, se denominan difusión y dispersión numéricas.

En términos físicos, el fenómeno de difusión es un proceso particular de atenuación, tanto más intenso cuanto menor es la escala de movimiento. La dispersión, por su parte, significa que distintas escalas de movimiento están asociadas a velocidades de traslación diferentes. Para poner en evidencia estos efectos se los analizará separadamente del problema original, por lo cual se tomará $\mathcal{L}(u) = \partial u/\partial t$. Además, se considerará la siguiente oscilación armónica:

$$u(x,t) = e^{i(kx-\beta t)} \tag{6.2}$$

donde $\beta = \omega + i\,A$ depende de k, siendo ω la frecuencia angular y A un coeficiente de atenuación. Entonces, se tiene que

$$\frac{\partial u}{\partial t} = -i\beta u \tag{6.3}$$

$$\sum_{p=1}^{\infty} \mu_{2p} \frac{\partial^{2p} u}{\partial x^{2p}} = \left[\sum_{p=1}^{\infty} (-1)^p k^{2p} \mu_{2p} \right] u \tag{6.4}$$

$$\sum_{p=0}^{\infty} \mu_{2p+1} \frac{\partial^{2p+1} u}{\partial x^{2p+1}} = i \left[\sum_{p=0}^{\infty} (-1)^p k^{2p+1} \mu_{2p+1} \right] u \tag{6.5}$$

Introduciendo las Ecs. (6.3)-(6.5) en la Ec. (6.1) se obtiene

$$A = \sum_{p=1}^{\infty} (-1)^{p-1} k^{2p} \mu_{2p} \tag{6.6}$$

$$c \equiv \frac{\omega}{k} = \sum_{p=0}^{\infty} (-1)^p k^{2p} \mu_{2p+1} \tag{6.7}$$

La Ec. (6.6) da directamente el coeficiente de atenuación numérica para cada número de ondas k. Si $A(k) < 0$ se dice que el esquema es disipativo para ese número de ondas. En cambio, si $A(k) > 0$ para algún valor de k, pueden dispararse inestabilidades. El caso de transición $A(k)=0$ corresponde a un esquema neutro o conservativo. Si bien esta última situación puede ser deseable desde el punto de vista de la precisión, no siempre es adecuada por ser solo marginalmente estable. Tal cual se discutió anteriormente, ésto es particularmente crítico en problemas no lineales, para los cuales la difusión numérica puede constituir un mecanismo eficaz de disipación de la energía de las escalas de movimiento menores resolubles.

La Ec. (6.7) expresa una velocidad de fase, que debe interpretarse como la perturbación que el esquema numérico introduce en la velocidad de fase correspondiente a la solución analítica. Entonces, si $c(k) > 0$ la perturbación apuntará en la dirección de las x positivas, y viceversa, lo cual puede producir un atraso o un adelanto relativo de la solución numérica, dependiendo de hacia donde se mueve la oscilación correspondiente a la solución analítica.

Para estimar los valores de A(k) y c(k), y así cuantificar la precisión, debe tenerse en cuenta que, en general, interesan más los movimienots de mayor escala (k pequeños). Entonces, pueden considerarse solo los primeros términos no nulos de las series de las Ecs. (6.6) y (6.7). Si éstos son directamente los primeros de cada serie, se tiene que

$$A \approx k^2 \mu_2 \tag{6.8}$$

(6.9) $$c \approx k\mu_1$$

Nótese que A y c representan, respectivamente, la atenuación y el desplazamiento relativos de la solución numérica respecto de la analítica, por unidad de tiempo. Entonces, el grado de precisión no solo dependerá de las características propias del esquema numérico, sino tambien del intervalo de tiempo sobre el cual quiere calcularse la solución.

6.3. Factor de propagación

Una alternativa a analizar el error de discretización consiste en comparar las soluciones analítica y numérica exacta para un caso simplificado. Por ejemplo, sea la ecuación hiperbólica

(6.10) $$\frac{\partial u}{\partial t} + U \frac{\partial u}{\partial x} = 0$$

aproximada por el siguiente esquema explícito de aguas arriba:

(6.11) $$\frac{u_j^{n+1} - u_j^n}{\Delta t} + U \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0$$

La oscilación (6.2) es solución de la Ec. (6.10) siempre que

(6.12) $$\beta = kU$$

o, dado que $\beta = \omega + i\,A$,

(6.13) $$\frac{\omega}{k} = U$$

(6.14) $$A = 0$$

La Ec. (6.13) expresa que la velocidad de fase de la oscilación coincide con la velocidad de convección U, mientras que la Ec. (6.14) muestra que la oscilación es neutra. Esta solución se denomina onda analítica.

Ahora, si se impone que una oscilación similar, pero eventualmente con distinto valor de β (β'), es solución exacta de la Ec. (6.11), se obtiene

(6.15) $$e^{-i\beta'\Delta t} = 1 - p(1-\cos k\Delta x) - ip \operatorname{sen} k\Delta x$$

donde $p = U\Delta t/\Delta x$. De la Ec. (6.15) surge que

(6.16) $$\omega' = \frac{1}{\Delta t} \operatorname{arctg}\left[\frac{p \operatorname{sen} k\Delta x}{1 - p(1-\cos k\Delta x)}\right] = \frac{1}{\Delta t} \operatorname{arctg}\left[\frac{\operatorname{tg} k\Delta x}{1 + \frac{1-p}{1}}\right]$$

$$(6.17) \quad A' = \frac{1}{\Delta t} \ln \left[1 + 2p(p-1)(1-\cos k\Delta x) \right]$$

Esta oscilación se denomina onda numérica. Nótese que si p=1, las Ecs. (6.16) y (6.17) se reducen, respectivamente, a $\omega' = \omega$ y $A'=0=A$, es decir que las ondas analitica y numérica coinciden (en efecto, la Ec. (6.11) daría, en este caso, $u_j^{n+1} = u_{j-1}^n$, lo cual es exacto ya que expresa la conservación del valor de u a lo largo de las curvas características). En cambio, si $p < 1$ (consistente con la condición de estabilidad $p \leq 1$) la Ec. (6.17) muestra que $A'<0$, es decir, la onda numérica se atenúa, mientras que la Ec. (6.15) indica que $\omega' < \omega$, es decir, la onda numérica se retrasa.

Una forma práctica de expresar estas relaciones consiste en utilizar el factor de propagación T (9), definido como la relación entre las ondas numérica y analitica luego de un intervalo de tiempo igual al período de la onda analitica, es decir

$$(6.18) \quad T \equiv \left[\frac{e^{i(kx-\beta' t)}}{e^{i(kx-\beta t)}} \right]_{t=\frac{2\pi}{\omega}} = e^{-i\frac{2\pi}{\omega}(\beta'-\beta)}$$

El módulo del factor de propagación da una medida del decaimiento de la amplitud de la onda numérica debido exclusivamente al esquema numérico (atenuación o difusión numérica), mientras que el argumento del factor de propagación es una medida del desfasaje que provoca el esquema numérico (dispersión numérica).

Es común representar el módulo y el argumento de T no directamemnte en función de k, sino en términos del número N de nodos computacionales necesario para representar la longitud de onda asociada ($2\pi/k$), es decir, $N=2\pi/k\,\Delta x$. La Fig. 6.1 muestra una presentación típica para el caso del esquema de Preissman aplicado a una forma simplificada de las Ecuaciones de Saint Venant (10). Allí, además del argumento de T, figura un eje con la relación entre las velocidades de fase de las ondas numérica (c_n) y analítica (c_a), dada por

$$(6.19) \quad \frac{c_n}{c_a} = 1 + \frac{arg(T)}{2\pi\, signo(\omega)}$$

Los parámetros Cr, θ, Fr y σ caracterizan el problema (así como p es el parámetro del ejemplo de más arriba).

Obviamente, la evaluación del factor de propagación puede complicarse fuertemente en problemas complejos, por lo cual su utilidad es, en este sentido, limitada.

REFERENCIAS

1.-DAHLQUIST, G., BJORCK, A., Numerical Methods, Prentice Hall, 1974 (tambien edición castellana en El Ateneo).

2.-McCRACKEN, D.D., DORN, W.S., Métodos Numéricos y Programación FORTRAN, Limusa, 1982.

3.-RICHTMYER, R., MORTON, K.W., Difference Methods for Initial-Value Problems, Wiley, 1967.

4.-HIRT, C.W., Heuristic Stability Theory for Finite-Difference Equation, J. Comp. Phys., 2, 339-355, 1968.

5.-MENENDEZ, A.N., Introducción a la simulación numérica de problemas hidráulicos, Informe LHA-INCYTH 064-003-87, setiembre de 1987.

6.-WARMING, R.F., HYETT, B.J., The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods, J. Comp. Phys., 14, 159-179, 1974.

7.-HENDERSON, F.M., Open Channel Flow, MacMillan, 1966.

8.-PUJOL, A., MENENDEZ, A.N., Análisis unidimensional de escurrimiento en canales, EUDEBA, 1987.

9.-LEENDERSTSE, J.J., Aspects of a Computational Model for Long-Period Water-Wave Propagation, RAND Memorandum RH-5299-PR, 1967.

10.-CARRERAS, P.E., MENENDEZ, A.N., Un método numérico para simular ondas de inundación con frentes abruptos en escurrimientos con cambio de régimen, Informe LHA-INCYTH S5-034-87, enero de 1987.

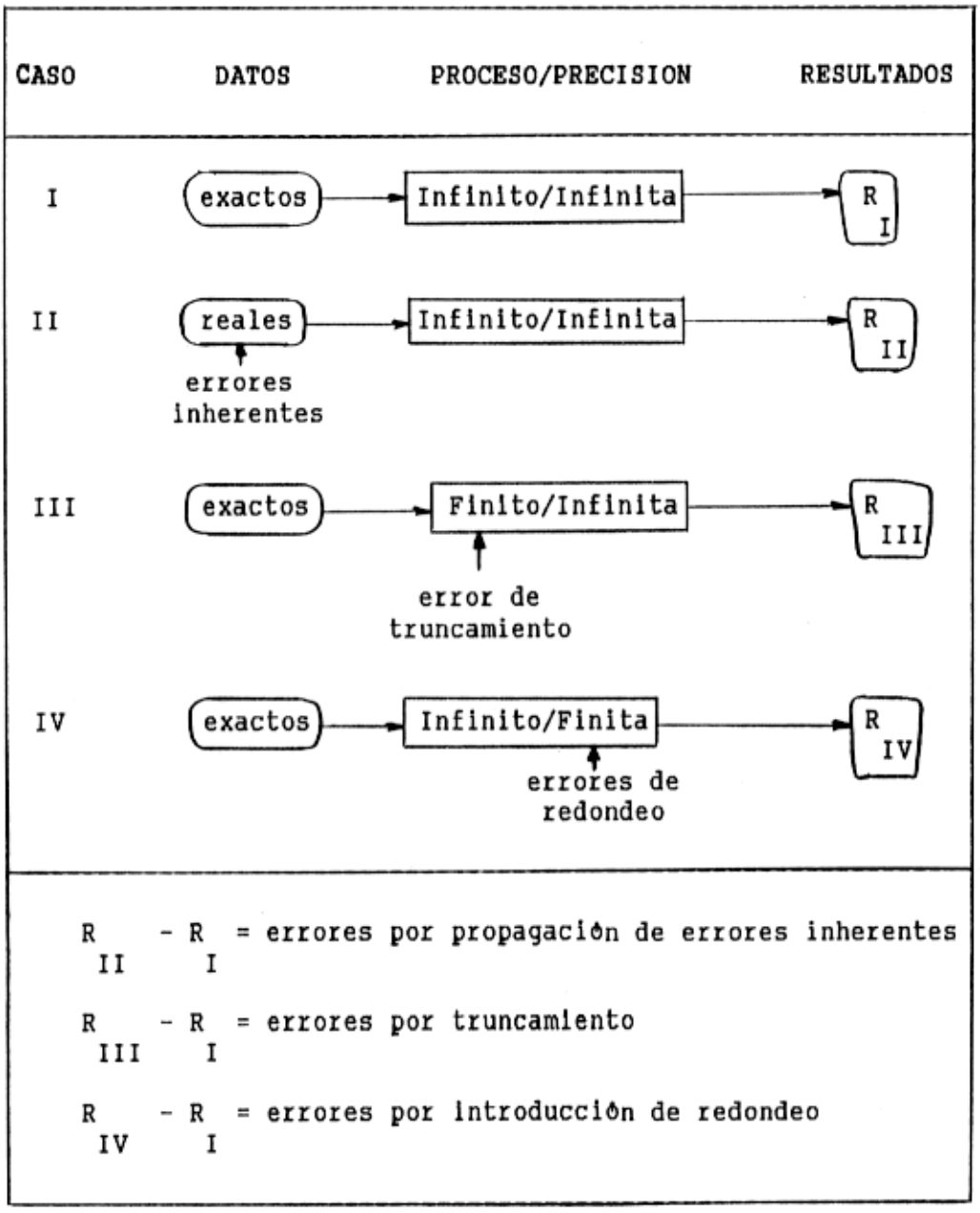


Figura 1.1. Errores presentes en un proceso de cálculo.

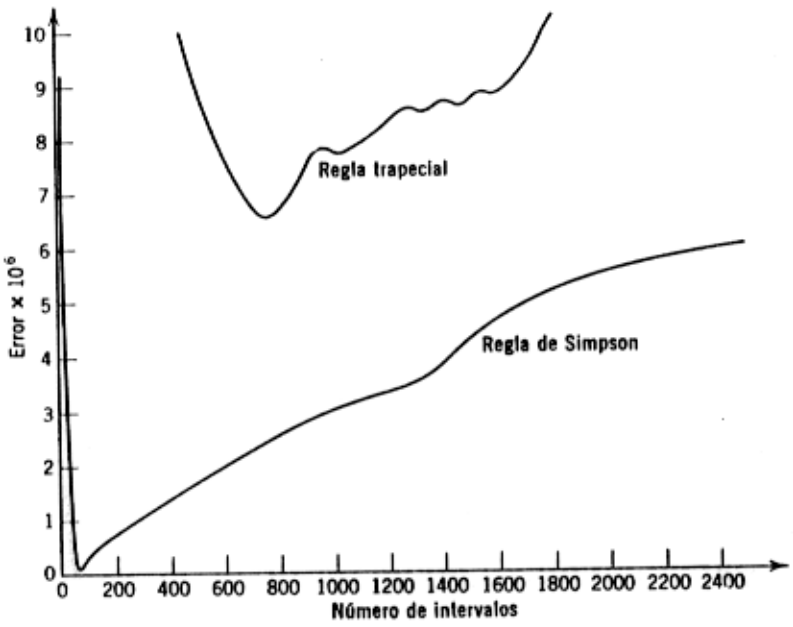
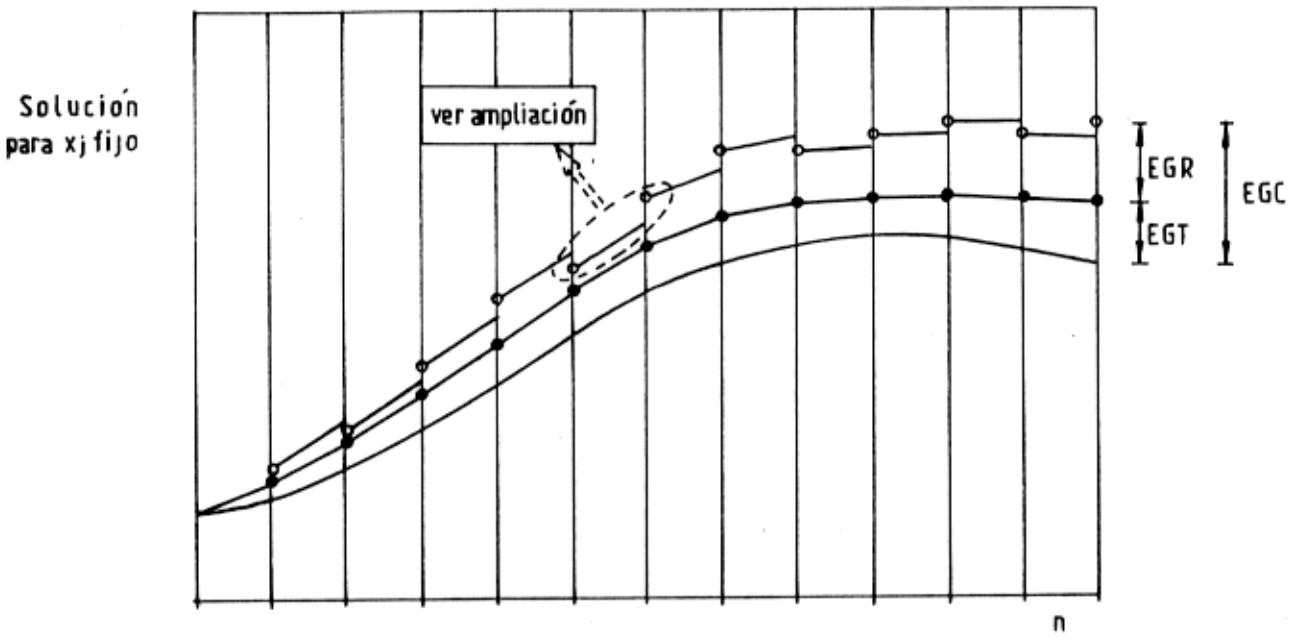


Fig. 1.2. Gráficas del error total (por truncamiento y por redondeo) al integrar sen x en el intervalo de 0 a π mediante la regla trapecial y mediante la regla de Simpson.



——— Solución analítica $u(x_j, t^n)$

●——— Solución numérica exacta $\hat{u}_j^n$

o——— Solución numérica computada u_j^n

EGT : Error global de truncamiento E_j^n

EGR : Error global de redondeo

EGC : Error global del cómputo

AMPLIACION

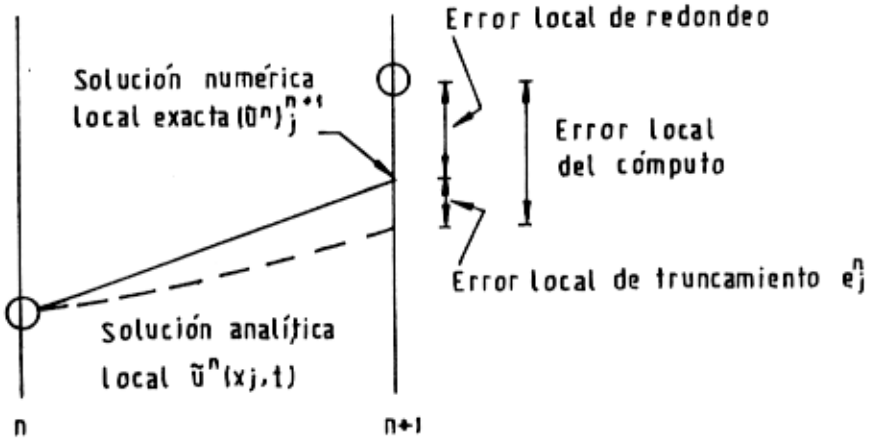


FIGURA 2.1 Representación de las soluciones analítica y numérica.

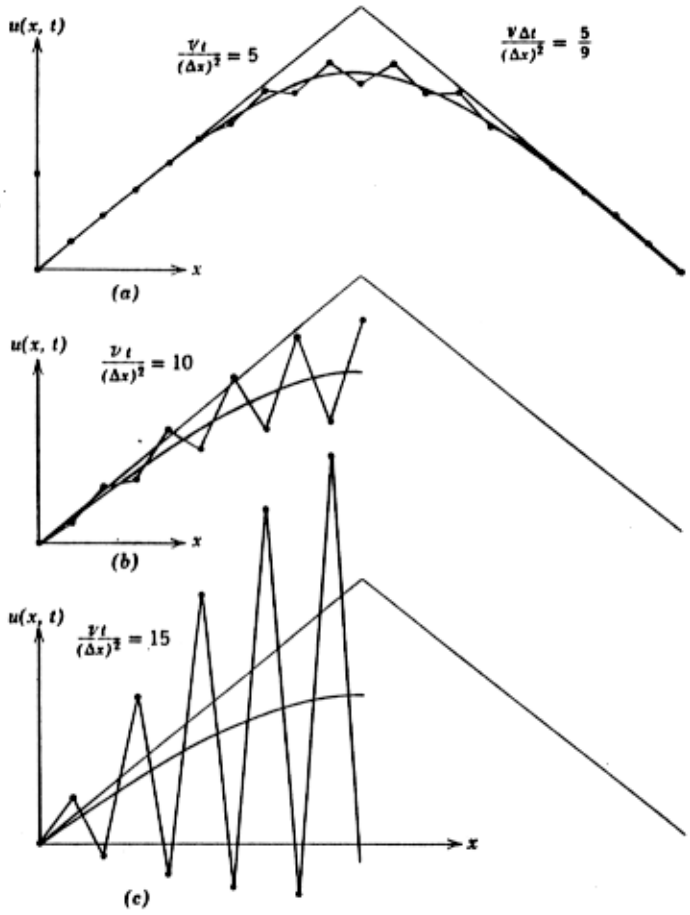
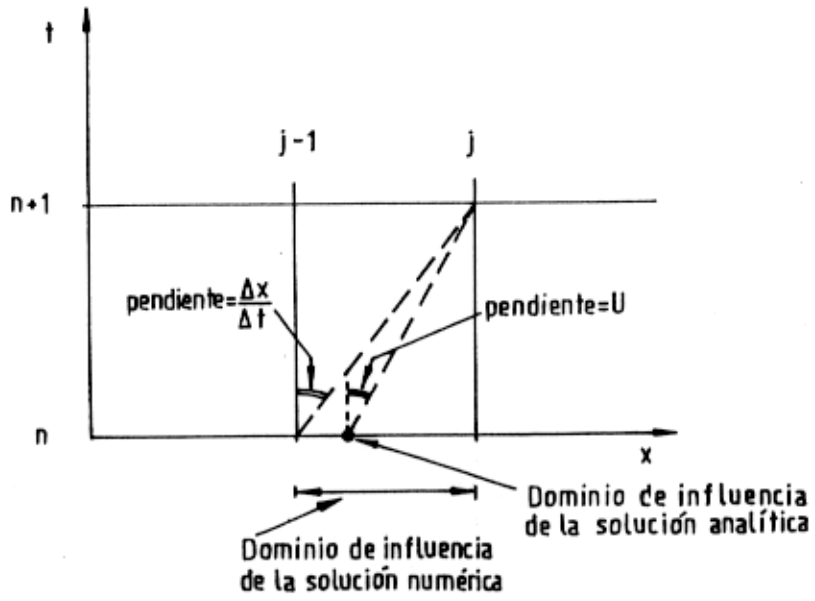


Figura 4.1. Manifestación de la inestabilidad numérica para el problema (2.1) cuando r=5/9 ($>$1/2).



Ecuación diferencial : $\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} = 0$

Ecuación en diferencias: $u_j^{n+1} = (1-p)\,uj^n + p\,u_{j-1}^n$

Criterio de Courant: $U \leq \Delta x/\Delta t$

FIGURA 4.2 Criterio de Courant para la estabilidad de una ecuación en diferencias.

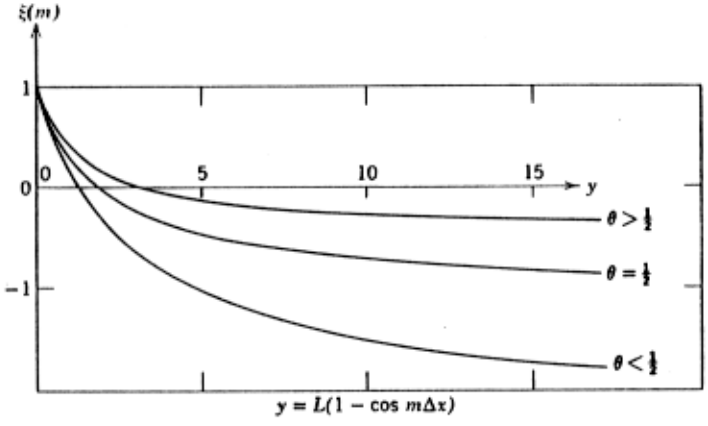


FIGURA 4.3 Factor de amplificación para el esquema implícito (4.38). L=2r.

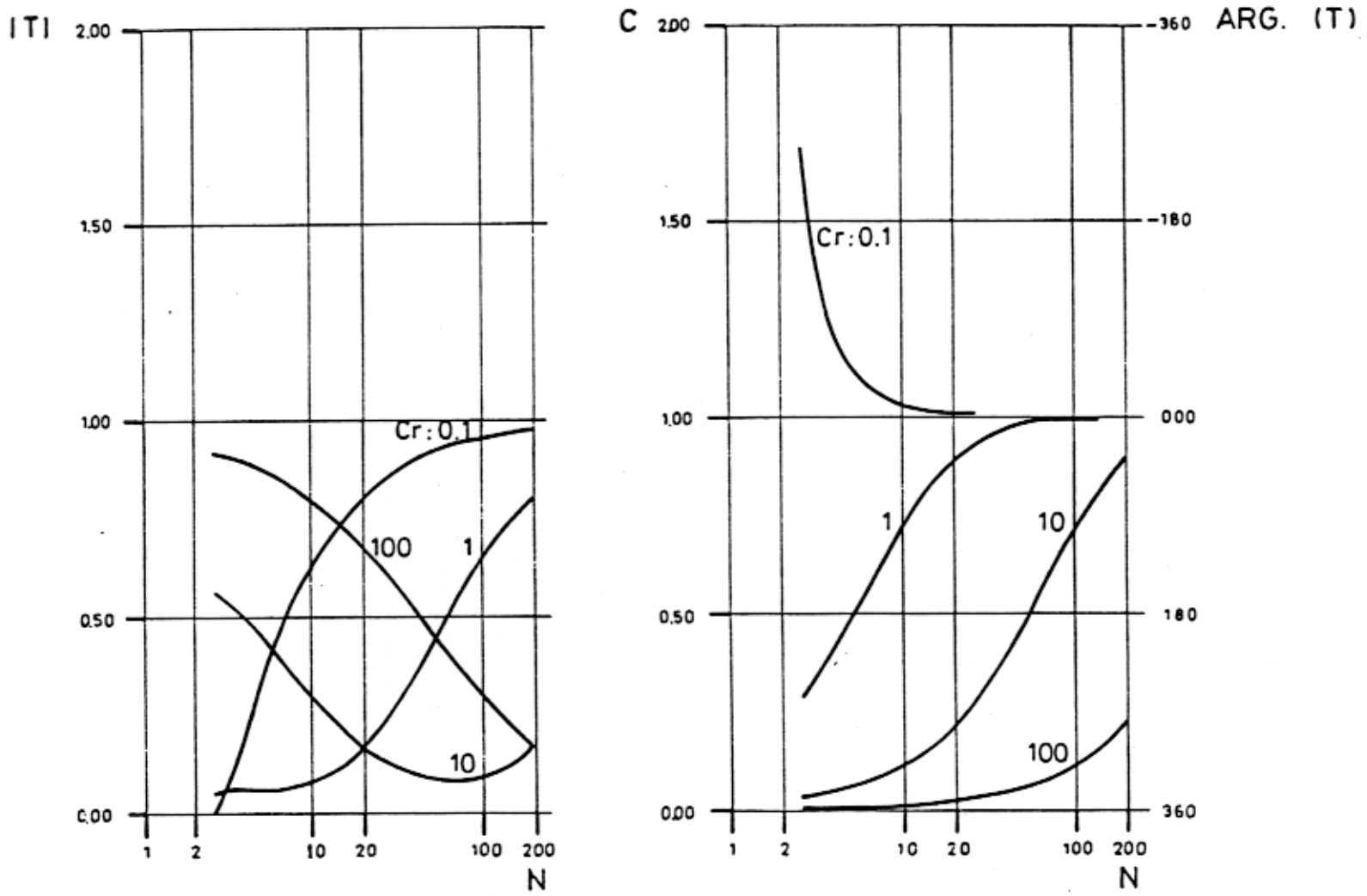


Figura 6.1. Factor de propagación para el esquema de Preissman (onda de avance, $\theta = 1$, $Fr = 1{,}1$ y $\varsigma \to \infty$).