Métodos iterativos para la solución de problemas lineales y no-lineales. Notas del curso

Mario Storti, Rodrigo Paz
Centro Internacional de Métodos Computacionales en Ingeniería
http://www.cimec.org.ar,
mario.storti@gmail.com,rodrigop@intec.unl.edu.ar

28 de abril de 2011

Índice general

Parte I

Métodos iterativos para la resolución de ecuaciones lineales

Capítulo 1

Conceptos básicos de métodos iterativos estacionarios

1.1. Notación y repaso

Denotamos a un sistema lineal como

$$Ax = b \tag{1.1}$$

con A no-singular de $N \times N$ y $b \in R^N$. La solución del sistema la denotamos como $x^* = A^{-1}b \in R^N$, mientras que x representa una solución potencial.

Los métodos iterativos se basan en encontrar una secuencia $\{x_k\}_{k=0}^{\infty}$ tal que

$$\lim_{k\to\infty} x_k = x^* \tag{1.2}$$

Debemos asegurar la convergencia del método iterativo y si es posible determinar la *tasa de convergencia*, es decir como se comporta el error $\|x - x_k\|$ para $k \to \infty$.

1.1.1. Normas inducidas

Dada una norma para vectores $\| \cdot \|$, denotamos por $\|A\|$ la norma de A inducida por $\| \cdot \|$, definida por

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \tag{1.3}$$

Las normas inducidas tienen la importante propiedad de que

$$\|Ax\| \leq \|A\| \, \|x\| \tag{1.4}$$

y también:

$$\|AB\| \leq \|A\| \, \|B\| \tag{1.5}$$

ya que

$$\begin{aligned} \|AB\| &= \max_{\|x\|=1} \|ABx\| \\ &\leq \max_{\|x\|=1} \|A\| \, \|Bx\| \\ &= \|A\| \max_{\|x\|=1} \|Bx\| \\ &= \|A\| \, \|B\| \end{aligned} \tag{1.6}$$

Obviamente $\|I\| = 1$ y $\|0\| = 0$.

A continuación se describen diferentes normas inducidas que serán utilizadas,

- $\|A\|_2 = \sqrt{\text{maximo autovalor de}(A^T A)}$
- $\|A\|_\infty = \max_i \left(\sum_j |a_{ij}|\right)$
- $\|A\|_1 = \max_j \left(\sum_i |a_{ij}|\right)$

Norma inducida L_2 de una matriz. Sea $B = A^T A$ entonces B es simétrica y semi-definida positiva. Sean $\{v_j, \lambda_j\}_{j=1}^N$ los autovalores de B, $v_j^T v_k = \delta_{jk}$, entonces

$$\|A\|_2^2 = \max_{x \neq 0} \frac{x^T B x}{x^T x} \tag{1.7}$$

Si $x = \sum_j \alpha_j v_j$ entonces

$$Bx = \sum_j \alpha_j \lambda_j v_j \tag{1.8}$$

y

$$\begin{aligned} x^T B x &= (\sum \alpha_k v_k^T)(\sum \alpha_j \lambda_j v_j) \\ &= \sum_{jk} \alpha_k \, \alpha_j \, \lambda_j \, v_k^T \, v_j \\ &= \sum_j \alpha_j^2 \lambda_j \end{aligned} \tag{1.9}$$

Por otra parte,

$$x^T x = \sum_j \alpha_j^2 \tag{1.10}$$

y

$$\|A\|_2^2 = \max_{\alpha \in \mathbb{R}^N} \frac{\sum_j \alpha_j^2 \lambda_j}{\sum_j \alpha_j^2} = \max_j \lambda_j \tag{1.11}$$

Además, si A es simétrica, entonces es diagonalizable, con autovalores λ'_j, entonces $A^T A = A^2$ y

$$\text{autovalores de } A^2 = (\lambda'_j)^2 \tag{1.12}$$

de manera que

$$\|A\|_2 = \max |\text{autovalores de } A| \tag{1.13}$$

Norma infinito inducida de una matriz.

Por definición

$$\|x\|_\infty = \max |(x)_i| \tag{1.14}$$

y entonces, la norma inducida es

$$\begin{aligned} \|Ax\|_\infty &= \max_i |\sum_j a_{ij} x_j| \leq \max_i \sum_j |a_{ij}| \, |x_j| \\ &\leq \max_i \left\{ \sum_j |a_{ij}| \, \max_k |x_k| \right\} \\ &= \max_i \left(\sum_j |a_{ij}| \right) \, \|x\|_\infty \end{aligned} \tag{1.15}$$

por lo tanto

$$\|A\|_\infty \leq \max_i \left(\sum_j |a_{ij}| \right). \tag{1.16}$$

Tomemos v tal que

$$(v)_j = \operatorname{sign} a_{ij}, \quad \text{con} \quad i = \operatorname*{argmax}_k \sum_j |a_{kj}| \tag{1.17}$$

entonces, es obvio que

$$\|v\|_\infty = 1 \tag{1.18}$$

y

$$\begin{aligned} |(Av)_k| &= \left| \sum a_{kj}\, v_j \right| \leq \sum |a_{kj}|\, |v_j| \\ &\leq \sum_j |a_{kj}| < \sum_j |a_{ij}| \end{aligned} \tag{1.19}$$

Para $k = i$ se satisface que

$$\begin{aligned} |(Av)_i| &= \left| \sum a_{ij}\, v_j \right| \\ &= \left| \sum a_{ij}\ \operatorname{sign} a_{ij} \right| \\ &= \sum |a_{ij}|. \end{aligned} \tag{1.20}$$

Por lo tanto,

$$\|Av\|_\infty = \frac{\|Av\|_\infty}{\|v\|_\infty} = \max_i \sum_j |a_{ij}| \tag{1.21}$$

Norma inducida L_1 de una matriz.
Por definición,

$$\|x\|_1 = \sum_i |(x)_i| \tag{1.22}$$

y entonces,

$$\begin{aligned} \|Ax\|_1 = \sum_i |(Ax)_i| &= \sum_i |\sum_j a_{ij}\, x_j| \\ &\leq \sum_i \sum_j |a_{ij}|\, |x_j| = \sum_j \left(\sum_i |a_{ij}| \right)\ |x_j| \\ &\leq \sum_j \left[\max_k \sum_i |a_{ik}| \right]\ |x_j| \\ &= \left(\max_k \sum_i |a_{ik}| \right)\ \|x\|_1 \end{aligned} \tag{1.23}$$

y entonces

$$\|A\|_1 \leq \max_k \sum_i |a_{ik}| \tag{1.24}$$

Sea v tal que $(v)_i = \delta_{ij}$ con $j = \operatorname{argmax}_k \sum_i |a_{ik}|$, entonces

$$\|v\|_1 = \sum_i \delta_{ij} = 1 \tag{1.25}$$

y como

$$(Av)_i = a_{ij} \tag{1.26}$$

entonces

$$\|Av\|_1 = \sum_i |a_{ij}| = \max_k \sum_i |a_{ik}| \tag{1.27}$$

y por definición de norma inducida

$$\|A\|_1 \geq \frac{\|Av\|_1}{\|v\|_1} = \max_k \sum_i |a_{ik}| \tag{1.28}$$

y entonces, de (1.24) y (1.28) se deduce que

$$\|A\|_1 = \max_k \sum_i |a_{ik}| \tag{1.29}$$

Normas no inducidas. Es fácil demostrar que existen normas que no son inducidas, es decir que satisfacen

$$\begin{aligned} &\|A\| \neq 0, \quad \text{si } A \neq 0 \\ &\|\alpha A\| = |\alpha| \, \|A\| \\ &\|A + B\| \leq \|A\| + \|B\| \end{aligned} \tag{1.30}$$

pero no provienen de ninguna norma para vectores. Por ejemplo, si definimos la norma $\| \, . \, \|_*$ como

$$\|A\|_* = c \, \|A\|_2 \tag{1.31}$$

con $c > 0, c \neq 1$, entonces es claro que es una norma pero $\|I\|_* = c \neq 1$ y por lo tanto no es inducida.

1.1.2. Número de condición

El número de condición de una matriz no-singular en la norma $\| \, . \, \|$ es

$$\kappa(A) = \|A\| \left\|A^{-1}\right\| \tag{1.32}$$

Podemos ver que

$$1 = \|I\| = \left\|AA^{-1}\right\| \leq \|A\| \left\|A^{-1}\right\| = \kappa(A) \tag{1.33}$$

Si A es singular tomamos como que $\kappa(A) = \infty$.

1.1.3. Criterios de convergencia de los métodos iterativos

Desaríamos *"detener"* un método iterativo cuando $\|e_k\| = \|x^* - x_k\| < \text{tol}$, pero como no conocemos x^* esto es imposible en los casos prácticos. Pero sí podemos calcular el residuo de la ecuación para la iteración k como

$$r_k = b - Ax_k \tag{1.34}$$

Si $\|r_k\|$ es suficientemente pequeño esperamos que $\|e_k\|$ sea pequeño y podemos poner como criterio de detención

$$\frac{\|r_k\|}{\|r_0\|} \leq \text{tol} \tag{1.35}$$

El siguiente lema nos permite relacionar ambos criterios de detención

Lema 1.1.1. Dados $b, x, x_0 \in \mathbb{R}^N$, A no-singular y $x^* = A^{-1}b$, entonces

$$\frac{\|e\|}{\|e_0\|} \leq \kappa(A)\frac{\|r\|}{\|r_0\|} \tag{1.36}$$

Demostración.

$$r = b - Ax = Ax^* - Ax = A(x - x^*) = -Ae \tag{1.37}$$

entonces

$$\|e\| = \left\|A^{-1}Ae\right\| \leq \left\|A^{-1}\right\| \|Ae\| = \left\|A^{-1}\right\| \|r\| \tag{1.38}$$

y

$$\|r_0\| \leq \|A\| \|e_0\| \tag{1.39}$$

Por lo tanto

$$\frac{\|e\|}{\|e_0\|} \leq \|A\| \left\|A^{-1}\right\| \|r\| \ \|r_0\| = \kappa(A) \frac{\|r\|}{\|r_0\|} \ \square. \tag{1.40}$$

La división por $\|e_0\|$ y $\|r_0\|$ es para *adimensionalizar* el problema. Por ejemplo, si consideramos un problema térmico, entonces x puede representar temperaturas nodales y por lo tanto tendrá dimensiones de temperatura ($^\circ$C), el miembro derecho q tendrá dimensiones de potencia (Watts) y los coeficientes $[A_{ij}] = \mathrm{W}/^\circ\mathrm{C}$, pero lo importante es que el residuo r tendrá las mismas dimensiones que r_0 es decir potencia (Watts). Ahora si hacemos un cambio de unidades tomando como unidad de potencia a cal/s entonces el criterio de parada $\|r\| < \mathrm{tol}$ es completamente diferente, mientras que $\|r\| / \|r_0\|$ es el mismo en los dos sistemas de unidades. Por otra parte, este criterio depende de la solución inicial x_0 lo cual puede llevar a iterar demasiado en el caso en que partimos de una buena solución ($\|r_0\|$ muy pequeño). Otra posibilidad es

$$\frac{\|r_k\|}{\|b\|} < \mathrm{tol} \tag{1.41}$$

Ambos coinciden si $x_0 = 0$.

1.2. El lema de Banach

La forma más directa de obtener (y posteriormente analizar) un método iterativo es reescribir (1.1) como un problema de iteración de punto fijo. Una forma de hacer esto es reescribir la ecuación como

$$x = (I - A)x + b \tag{1.42}$$

lo cual induce el siguiente método iterativo de Richardson

$$x_{k+1} = (I - A)x_k + b \tag{1.43}$$

El análisis de tales secuencias recursivas es más general y vamos a estudiar la convergencia de relaciones recursivas generales de la forma

$$x_{k+1} = Mx_k + c \tag{1.44}$$

donde M es la llamada *matriz de iteración* o *matriz de amplificación*. Estos métodos son llamados métodos iterativos estacionarios porque la transición de x_k a x_{k+1} no depende de la historia anterior:

Métodos estacionarios: $x_{k+1} = f(x_k)$

Métodos no estacionarios: $x_{k+1} = f(x_k, x_{k-1}, x_{k-2}, \ldots)$

Los métodos de Krylov que discutiremos más adelante *no son métodos estacionarios*. Es interesante también ver que el esquema de Richardson puede ponerse de la forma

$$x_{k+1} = x_k + (b - Ax_k) = x_k + r_k \tag{1.45}$$

Nótese que r_k actúa como una pequeña corrección a la iteración k para obtener la siguiente $k+1$. Si estamos suficientemente cerca de la solución x^* entonces r_k será pequeño y la corrección será pequeña.

Aplicando recursivamente (1.44) obtenemos una expresión general para la iteración x_k. Asumamos por simplicidad que $x_0 = 0$, entonces

$$\begin{aligned} x_1 &= c \\ x_2 &= Mc + c = (M+I)\,c \\ x_3 &= M(M+I)\,c + c = (M^2 + M + I)\,c \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ x_k &= \left(\sum_{j=0}^{k-1} M^j\right) c \end{aligned} \tag{1.46}$$

Es claro que la convergencia del método está ligada a la convergencia de la suma de M^j.

Lemma 1.2.1. Si $M \in \mathbb{R}^{N\times N}$ satisface $\|M\| < 1$ entonces $I - M$ es no-singular y

$$\left\|(I-M)^{-1}\right\| \leq \frac{1}{1-\|M\|} \tag{1.47}$$

Demostración. Veremos que la serie converge a $(I-M)^{-1}$. Sea $S_k = \sum_{l=0}^{k} M^l$. Mostraremos que es una secuencia de Cauchy, sea $k < m$

$$\begin{aligned} \|S_m - S_k\| &= \left\| \sum_{l=k+1}^{m} M^l \right\| \\ &\leq \sum_{l=k+1}^{m} \left\| M^l \right\| \\ &\leq \sum_{l=k+1}^{m} \|M\|^l \\ &= \|M\|^{k+1} \left(\frac{1 - \|M\|^{m-k}}{1-\|M\|} \right) \\ &\to 0 \end{aligned} \tag{1.48}$$

para $m, k \to \infty$. Entonces $S_k \to S$ para algún S. Pero entonces tomando el límite en la relación de recurrencia

$$MS_k + I = S_{k+1} \tag{1.49}$$

obtenemos

$$MS + I = S \tag{1.50}$$

Por lo tanto

$$(I-M)S = I \tag{1.51}$$

de donde $I - M$ es no-singular y $S = (I - M)^{-1}$. Por otra parte

$$\left\|(I-M)^{-1}\right\| \leq \sum_{l=0}^{\infty} \|M\|^l = (1 - \|M\|)^{-1} \ \square. \tag{1.52}$$

Matrices diagonalizables bajo la norma 2. En el caso en que M es simétrica y consideramos la norma $\|M\|_2$ es más fácil de visualizar las implicancias del lema. Como M es simétrica, por lo tanto es diagonalizable. Sean S no-singular y Λ diagonal las matrices que dan la descomposición diagonal de M

$$\begin{aligned} M &= S^{-1}\Lambda S \\ (I-M) &= S^{-1}(I-\Lambda)S \end{aligned} \tag{1.53}$$

Como $\|M\| = \max_j |\lambda_j| < 1$ esto quiere decir que todos los autovalores λ_j de M, que son reales ya que M es simétrica, están estrictamente contenidos en el intervalo $-1 < \lambda_j < 1$. Por otra parte

$$\begin{aligned} \left\|(I-M)^{-1}\right\|_2 &= \max_j \left|\frac{1}{1-\lambda_j}\right| \\ &= \frac{1}{\min|1-\lambda_j|} \\ &= \frac{1}{1-\max\lambda_j} \\ &\leq \frac{1}{1-\max|\lambda_j|} = \frac{1}{1-\|M\|_2} \end{aligned} \tag{1.54}$$

Corolario 1.2.1. Si $\|M\| < 1$ entonces la iteración (1.44) converge a $x = (I - M)^{-1} c$ para cualquier punto inicial x_0.

Demostración. Si $x_0 = 0$ ya está demostrado. Si $x_0 \neq 0$ haciendo el cambio de variables $x'_k = x_k - x_0$ llegamos al esquema recursivo

$$\begin{aligned} x_{k+1} - x_0 &= M(x_k - x_0) + c - (I-M)x_0 \\ x'_{k+1} &= Mx'_k + c' \end{aligned} \tag{1.55}$$

El cual converge y converge a $(I - M)^{-1} c'$, por lo tanto x_k converge a

$$\begin{aligned} (I-M)^{-1}c' + x_0 &= (I-M)^{-1}[c - (I-M)x_0] + x_0 \\ &= (I-M)^{-1} c \ \square. \end{aligned} \tag{1.56}$$

Una consecuencia del corolario es que la iteración de Richardson (1.6) converge si $\|I - A\| < 1$. A veces podemos precondicionar el sistema de ecuaciones multiplicando ambos miembros de (1.1) por una matriz B

$$BA\,x = Bb \tag{1.57}$$

de manera que la convergencia del método iterativo es mejorada. En el contexto de la iteración de Richardson las matrices B tales que permiten aplicar el Lema de Banach y su corolario se llaman *inversas aproximadas*

Definición 1.2.1. B es una inversa aproximada de A si $\|I - BA\| < 1$.

El siguiente teorema es llamado comúnmente *Lema de Banach*.

Teorema 1.2.1. Si A y $B \in \mathbb{R}^{N\times N}$ y B es una inversa aproximada de A, entonces A y B son no singulares y

$$\left\|A^{-1}\right\| \leq \frac{\|B\|}{1-\|I-BA\|}, \quad \left\|B^{-1}\right\| \leq \frac{\|A\|}{1-\|I-BA\|}, \tag{1.58}$$

y

$$\left\|A^{-1}-B\right\| \le \frac{\|B\|\,\|I-BA\|}{1-\|I-BA\|}, \quad \left\|A-B^{-1}\right\| \le \frac{\|A\|\,\|I-BA\|}{1-\|I-BA\|}, \tag{1.59}$$

Demostración. Sea $M = I - BA$, entonces $I - M = BA$ es no singular y por lo tanto B y A son no-singulares y

$$\left\|(BA)^{-1}\right\| = \left\|A^{-1}B^{-1}\right\| \le \frac{1}{1-\|I-BA\|} \tag{1.60}$$

y

$$\left\|A^{-1}\right\| \le \left\|A^{-1}B^{-1}\right\|\|B\| \le \frac{\|B\|}{1-\|I-BA\|}. \tag{1.61}$$

Por otra parte,

$$\left\|A^{-1}-B\right\| = \left\|(I-BA)A^{-1}\right\| \le \frac{\|B\|\,\|I-BA\|}{1-\|I-BA\|} \tag{1.62}$$

La demostración de las otras desigualdades es similar. □

Notar que hay una cierta simetría en el rol jugado por A y B. De hecho deberíamos definir inversa aproximada *por derecha* y *por izquierda* y B sería la inversa aproximada por derecha de A.

La iteración de Richardson, precondicionada aproximadamente, tiene la forma

$$x_{k+1} = (I - BA)\,x_k + Bb \tag{1.63}$$

Si $\|I - BA\| \ll 1$ las iteraciones convergen rápido y además, (por el Lema 1.2.1) las decisiones basadas en el residuo precondicionado $\|B(b - A\,x)\|$ reflejarán muy bien el error cometido.

1.3. Radio espectral

El análisis de §1.2 relacionó la convergencia del esquema iterativo (1.44) a la norma de la matriz M. Sin embargo la norma de M puede ser pequeña en alguna norma y grande en otras. De aquí que la performance de la iteración no sea completamente descripta por $\|M\|$. El concepto de *radio espectral* da una descripción completa. Sea $\sigma(A)$ el conjunto de autovalores de A.

Definición 1.3.1. El radio espectral de A es

$$\rho(A) = \max_{\lambda\in\sigma(A)} |\lambda| = \lim_{n\to\infty} \|A^n\|^{1/n} \tag{1.64}$$

El radio espectral de A es independiente de la norma particular de M. De hecho

$$\rho(A) \le \|A\| \tag{1.65}$$

ya que, si $Av = \lambda_{\max} v$, entonces

$$\|A\| \ge \frac{\|Av\|}{\|v\|} = |\lambda_{\max}| = \rho(A) \tag{1.66}$$

Siempre que $\|\,.\,\|$ sea una norma inducida. En realidad puede demostrarse algo así como que $\rho(A)$ es el ínfimo de todas las normas de A. Esto es el enunciado del siguiente teorema (que no demostraremos aquí).

Teorema 1.3.1. Sea $A \in \mathbb{R}^{N\times N}$. Para cada $\epsilon > 0$ existe una norma inducida $\|\,.\,\|$ tal que $\rho(A) > \|A\| - \epsilon$.

Puede verse que, si $\rho(M) \geq 1$ entonces existen x_0 y c tales que (1.44) diverge. Efectivamente, sea v el autovalor tal que $Mv = \lambda v$ y $|\lambda| = \rho(M) \geq 1$. Tomemos $x_0 = v$, $c = 0$, entonces

$$x_k = M^k x_0 = \lambda^k x_0 \tag{1.67}$$

es claro que no converge.

Teorema 1.3.2. Sea $M \in \mathbb{R}^{N\times N}$. La iteración (1.44) converge para todos $x_0, c \in \mathbb{R}^{N\times N}$ si y solo si $\rho(M) < 1$.

Demostración. Si $\rho(M) > 1$ entonces en cualquier norma tal que $\|M\| > 1$ la iteración no converge por el párrafo anterior. Por otra parte, si $\rho(M) < 1$ entonces, tomando $\epsilon = (1 - \rho(M))/2$, existe una norma tal que

$$\|M\| < \rho(M) + \epsilon = \frac{1}{2}(1 + \rho(M)) < 1 \tag{1.68}$$

y por lo tanto converge. □

Tasa de convergencia de métodos estacionarios. Para un esquema convergente de la forma

$$x_{k+1} = (I - BA)\, x_k + Bb \tag{1.69}$$

podemos estimar la tasa de convergencia como,

$$\begin{aligned} x_k - x^* &= (I - BA)\, x_{k-1} + BA\, x^* - x^* \\ &= (I - BA)\, (x_{k-1} - x^*) \end{aligned} \tag{1.70}$$

y por lo tanto

$$\begin{aligned} \|x_k - x^*\| &\leq \|I - BA\|\ \|x_{k-1} - x^*\| \\ &\leq \|I - BA\|^k\ \|x_0 - x^*\| \end{aligned} \tag{1.71}$$

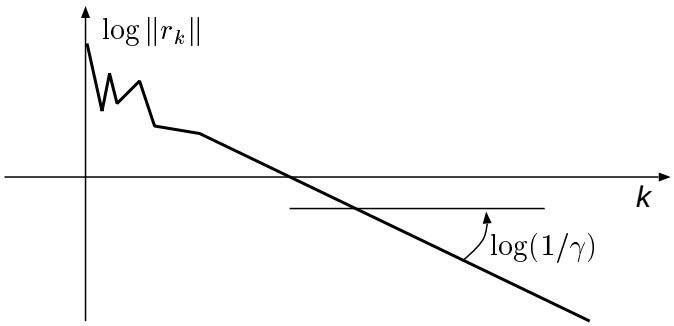

Figura 1.1: Historia de convergencia típica

Es usual visualizar la convergencia del método graficando $\|r_k\|$ versus k. Como $\|r_k\|_2$ puede reducirse en varios órdenes de magnitud durante la iteración, es usual usar ejes logarítmicos para $\|r_k\|_2$. Es fácil demostrar que una relación muy similar a (1.36), pero intercambiando r con x, es decir

$$\frac{\|r\|}{\|r_0\|} \leq \kappa(A) \frac{\|e\|}{\|e_0\|} \tag{1.72}$$

esto se debe a que (1.36) proviene del hecho de que $r = Ae$, pero también tenemos que $e = A^{-1}r$, de donde se puede deducir fácilmente que

$$\frac{\|r\|}{\|r_0\|} \leq \kappa(A^{-1}) \frac{\|e\|}{\|e_0\|} \tag{1.73}$$

pero $\kappa(A^{-1}) = \kappa(A)$. Entonces

$$\frac{\|r_k\|}{\|r_0\|} \leq \kappa(A)\frac{\|e_k\|}{\|e_0\|} \leq \kappa(A)\,\|I - BA\|^k \tag{1.74}$$

Como podemos tomar arbitrariamente cualquier norma de A para el número de condición, podemos usar el radio espectral, o sea que

$$\begin{aligned} \frac{\|r_k\|}{\|r_0\|} &\leq \kappa(A)\gamma^k, \\ \gamma &= \rho(I - BA). \end{aligned} \tag{1.75}$$

Por otra parte (1.75) no sólo es una estimación de la tasa de convergencia sino que, de hecho, muchas veces el residuo de la ecuación termina comportándose de esta forma, después de un cierto transitorio inicial (ver figura) es decir

$$\|r_k\| \sim \gamma^k\,\|r_0\| \tag{1.76}$$

Este tipo de comportamiento se refleja en una recta de pendiente $\log\gamma$ en el gráfico. Un índice de la velocidad de convergencia es el número de iteraciones n necesario para bajar el residuo un factor 10,

$$\begin{aligned} &\|r_{k+n}\| = 1/10\;\|r_k\| \\ &\gamma^{k+n}\,\|r_0\| = 1/10\,\gamma^k\,\|r_0\| \\ &\log 10 = -n\log\gamma, \quad n = \frac{\log 10}{\log(1/\gamma)} \end{aligned} \tag{1.77}$$

para problemas muy mal condicionados

$$\gamma = 1 - \epsilon, \quad \epsilon \ll 1 \tag{1.78}$$

y entonces,

$$\log(1/\gamma) \sim \epsilon, \quad n = \frac{\log 10}{\epsilon} \tag{1.79}$$

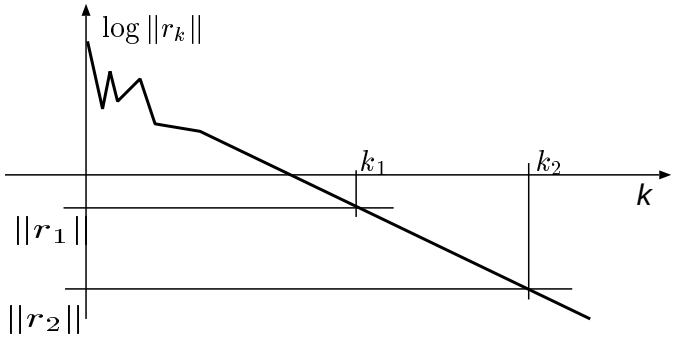

Figura 1.2: Estimación de la tasa de convergencia

A veces podemos calcular la tasa de convergencia "experimentalmente" conociendo el valor del residuo $\|r_k\|$ para dos iteraciones k_1, k_2. Asumiendo que en el intervalo $k_1 \leq k \leq k_2$ la tasa de convergencia es constante como en (1.76) (ver figura 1.2), entonces

$$\|r_2\| = \gamma^{k_2 - k_1}\,\|r_1\| \tag{1.80}$$

de donde

$$\log \gamma = \frac{\log(\|r_2\| \,/\, \|r_1\|)}{k_2 - k_1} \tag{1.81}$$

y entonces,

$$n = \frac{(\log 10)\,(k_2 - k_1)}{\log(\|r_1\| \,/\, \|r_2\|)} \tag{1.82}$$

1.4. Saturación del error debido a los errores de redondeo.

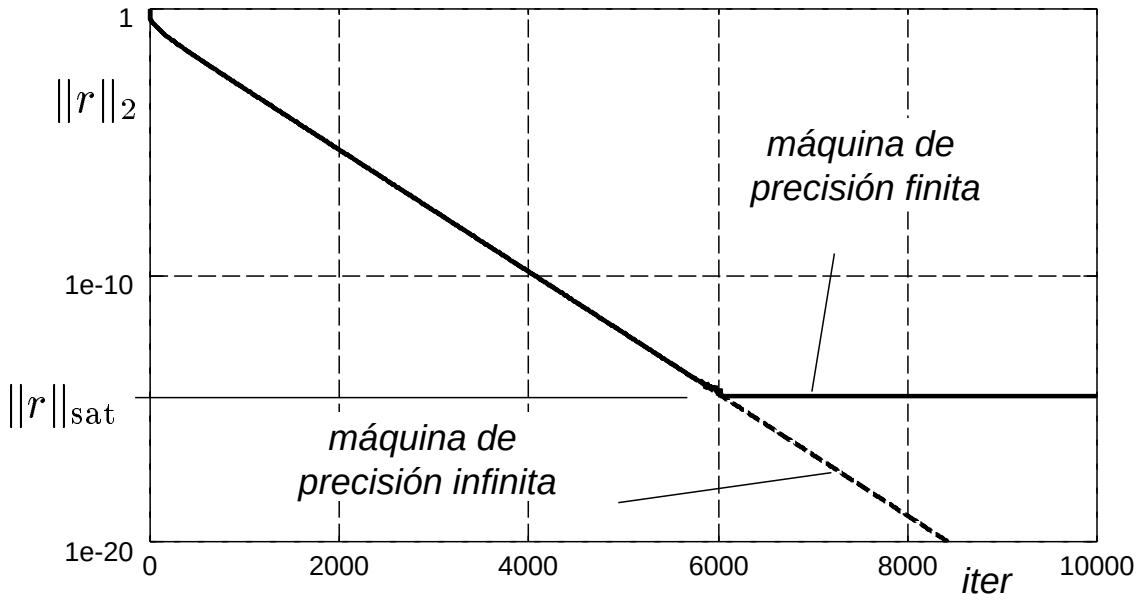

Figura 1.3: Saturación del error por errores de redondeo.

A medida que vamos iterando el residuo va bajando en magnitud. Cuando el valor del residuo pasa por debajo de cierto valor umbral dependiendo de la precisión de la máquina ($\sim 10^{-15}$ en Octave, Fortran (`real *8`) o C (tipo `double`)) se produce, debido a errores de redondeo, un efecto de *saturación* (ver figura 1.3). Es decir, al momento de calcular (1.34), es claro que incluso reemplazando x_k por la solución exacta x^* el residuo (calculado en una máquina de precisión finita) dara un valor no nulo del orden de la precisión de la máquina. Por supuesto este umbral de saturación es relativo a la norma de cada uno de los términos intervinientes y además para sistemas mal condicionados, el umbral se alcanza antes por un factor $\kappa(A)$, es decir que

$$\|r\|_{\text{sat}} \approx 10^{-15} \times \kappa(A) \times \|b\| \tag{1.83}$$

1.5. Métodos iterativos estacionarios clásicos

Hay otras formas alternativas a (1.42) de llevar $Ax = b$ a un problema de punto fijo. Los métodos como Jacobi, Gauss-Seidel y relajaciones sucesivas se basan en descomposiciones de A (*"splittings"*) de la forma,

$$A = A_1 + A_2 \tag{1.84}$$

con A_1 no-singular y fácil de factorizar. El nuevo problema de punto fijo es

$$x = A_1^{-1}\left(b - A_2 x\right) \tag{1.85}$$

El análisis del método se basa en estimar el radio espectral de $M = -A_1^{-1}A_2$.

Iteración de Jacobi. Corresponde a tomar

$$\begin{aligned} A_1 &= D = \operatorname{diag}(A) \\ A_2 &= L + U = A - D \end{aligned} \tag{1.86}$$

donde L, U, D son las partes triangular inferior, superior y diagonal de $A = L + U + D$. El esquema iterativo es

$$(x_{k+1})_i = a_{ii}^{-1}\left(b_i - \sum_{j\neq i} a_{ij}(x_k)_j\right) \tag{1.87}$$

Notar que A_1 es diagonal y, por lo tanto, trivial de invertir. La matriz de iteración correspondiente es

$$M_{\text{Jac}} = -D^{-1}\left(L + U\right) \tag{1.88}$$

Teorema 1.4.1. Sea $A \in \mathbb{R}^{N\times N}$ y asumamos que para cualquier $1 \leq i \leq N$

$$0 \leq \sum_{j\neq i} |a_{ij}| < |a_{ii}| \tag{1.89}$$

entonces A es no-singular y Jacobi converge.

Demostración. Veamos que

$$\sum_{j=1}^{N} |m_{ij}| = \frac{\sum_{j\neq i}|a_{ij}|}{|a_{ii}|} < 1 \tag{1.90}$$

Entonces,

$$\|M_{\text{Jac}}\|_\infty = \max_i \sum_j |m_{ij}| < 1 \tag{1.91}$$

Por lo tanto la iteración converge a la solución de

$$\begin{aligned} x &= Mx + D^{-1}b \\ x &= (I - D^{-1}A)x + D^{-1}b \end{aligned} \tag{1.92}$$

entonces $Ax = b$ y $I - M = D^{-1}A$ es no-singular y A es no-singular.

Iteración de Gauss-Seidel. En este esquema se reemplaza la solución aproximada con el nuevo valor tan pronto como éste es calculado,

$$(x_{k+1})_i = a_{ii}^{-1}\left(b_i - \sum_{j<i} a_{ij}\left(x_{k+1}\right)_j - \sum_{j>i} a_{ij}\left(x_k\right)_j\right) \tag{1.93}$$

que puede escribirse como

$$\left(D + L\right) x_{k+1} = b - U\, x_k \tag{1.94}$$

El split correspondiente es

$$A_1 = D + L, \quad A_2 = U \tag{1.95}$$

y la matriz de iteración

$$M_{\mathrm{GS}} = -(D+L)^{-1}U \tag{1.96}$$

A_1 es triangular inferior y por lo tanto A_1^{-1} es fácil de calcular. Notar también que a diferencia de Jacobi, depende del ordenamiento de las incógnitas. También podemos hacer un "backward Gauss-Seidel" con el splitting

$$A_1 = D+U, A_2 = L \tag{1.97}$$

La iteración es

$$(D+U)\,x_{k+1} = (b - Lx_k) \tag{1.98}$$

y la matriz de iteración

$$M_{\mathrm{BGS}} = -(D+U)^{-1}L \tag{1.99}$$

Gauss-Seidel simétrico. Podemos combinar alternando una iteración de forward GS con una de backward GS poniendo

$$\begin{aligned}(D+L)\,x_{k+1/2} &= b - Ux_k \\ (D+U)\,x_{k+1} &= b - Lx_{k+1/2}\end{aligned} \tag{1.100}$$

La matriz de iteración es

$$M_{\mathrm{SGS}} = M_{\mathrm{BGS}}\,M_{\mathrm{GS}} = (D+U)^{-1}\,L\,(D+L)^{-1}\,U \tag{1.101}$$

Si A es simétrica, entonces $U = L^T$ y

$$M_{\mathrm{SGS}} = (D+L^T)^{-1}\,L\,(D+L)^{-1}\,L^T \tag{1.102}$$

Queremos escribir estos esquemas como una iteración de Richardson precondicionada, es decir que queremos encontrar B tal que $M = I - BA$ y usar B como inversa aproximada. Para la iteración de Jacobi,

$$B_{\mathrm{Jac}} = D^{-1} \tag{1.103}$$

y para Gauss-Seidel simétrico

$$B_{\mathrm{SGS}} = (D+L^T)^{-1}\,D\,(D+L)^{-1} \tag{1.104}$$

Verificación. Debemos demostrar que $M_{\mathrm{SGS}} = I - B_{\mathrm{SGS}}\,A$. Efectivamente,

$$\begin{aligned}I - B_{\mathrm{SGS}}\,A &= I - (D+L^T)^{-1}\,D\,(D+L)^{-1}\,(D+L+L^T) \\ &= I - (D+L^T)^{-1}\,D\,(I + (D+L)^{-1}L^T) \\ &= (D+L^T)^{-1}\,[D + L^T - D - D(D+L)^{-1}L^T] \\ &= (D+L^T)^{-1}\,[I - D(D+L)^{-1}]\,L^T \\ &= (D+L^T)^{-1}\,[(D+L) - D]\,(D+L)^{-1}\,L^T \\ &= (D+L^T)^{-1}\,L\,(D+L)^{-1}\,L^T \\ &= M_{\mathrm{SGS}}\end{aligned} \tag{1.105}$$

Sobrerelajación. Consideremos iteración simple de Richardson

$$x_{k+1} = x_k + (b - Ax_k) \tag{1.106}$$

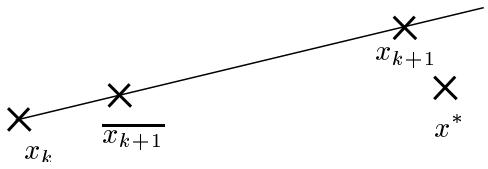

Figura 1.4: Aceleración de la convergencia por sobre-relajación

Si vemos que x_k se acerca monótonamente y lentamente a x^* podemos pensar en "acelerarlo" (ver figura 1.4) diciendo que el método iterativo en realidad nos predice un estado intermedio $\overline{x_{k+1}}$, y buscamos el vector de iteración x_{k+1} sobre la recta que une x_k con $\overline{x_{k+1}}$

$$\begin{aligned} \overline{x_{k+1}} &= x_k + (b - Ax_k) \\ x_{k+1} &= x_k + \omega(\overline{x_{k+1}} - x_k) \end{aligned} \tag{1.107}$$

Veremos más adelante que el valor óptimo de ω está relacionado con la distribución de autovalores de la matriz de iteración en el plano complejo. Mientras tanto podemos ver intuitivamente que

- $\omega = 1$ deja el esquema inalterado
- $\omega > 1$ tiende a acelerar la convergencia si el esquema converge lenta y monótonamente
- $\omega < 1$ tiende a desacelerar la convergencia si el esquema se hace inestable.

Esquema iterativo de Richardson con relajación para matrices spd. Aplicando el método de relajación a la iteración básica de Richardson obtenemos el esquema

$$x_{k+1} = x_k + \omega r_k \tag{1.108}$$

que puede reescribirse como

$$x_{k+1} = (I - \omega A)\, x_k + \omega b \tag{1.109}$$

de manera que la matriz de iteración es

$$M_{\mathrm{SR}} = I - \omega A \tag{1.110}$$

Asumiendo que A es simétrica y definida positiva o *"spd"* (por *symmetric positive definite*) , el espectro de M_{SR} esta dado por

$$\sigma(M_{\mathrm{SR}}) = 1 - \omega\sigma(A) \tag{1.111}$$

pero como A es spd, los autovalores $\lambda \in \sigma(A)$ son reales y positivos. Asumamos que están ordenados $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. También denotaremos $\lambda_{\min} = \lambda_N$, $\lambda_{\max} = \lambda_1$. Los autovalores de M_{SR} se comportan en función de ω como se observa en la figura 1.5. Para un dado ω todos los autovalores de M_{SR} están comprendidos entre los autovalores correspondientes a $\lambda_{\min}$ y $\lambda_{\max}$ (la región rayada de la figura). Para ω suficientemente chico todos los autovalores se concentran cerca de la unidad. Para un cierto valor ω_{crit} el autovalor correspondiente a $\lambda_{\max}$ se pasa de -1 con lo cual la iteración no converge. El valor de ω_{crit} está dado entonces por

$$1 - \omega_{\mathrm{crit}}\lambda_{\max} = -1, \quad \omega_{\mathrm{crit}} = \frac{2}{\lambda_{\max}} \tag{1.112}$$

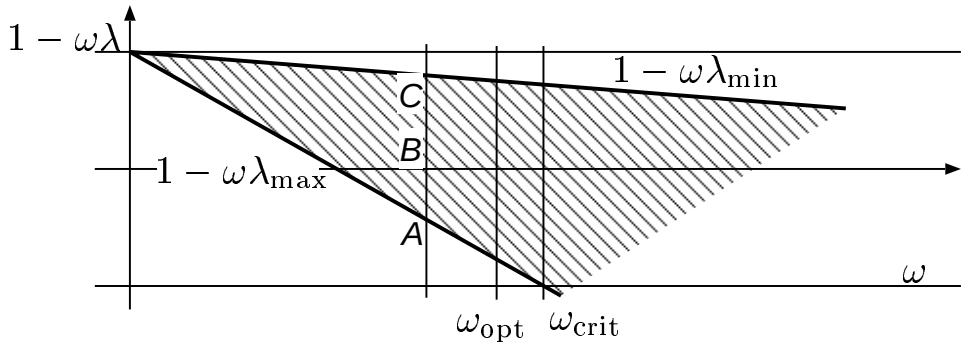

Figura 1.5:

La tasa de convergencia está dada por

$$\gamma = \rho(M) = \max_j |1 - \omega\lambda_j|, \tag{1.113}$$

y queremos buscar el ω que da la mejor tasa de convergencia, es decir el mínimo γ. Como los $1-\omega\lambda_j$ están comprendidos en el intervalo $1-\omega\lambda_{\min}, 1-\omega\lambda_{\max}$, se cumple que

$$\gamma = \max(|1 - \omega\lambda_{\min}|, |1 - \omega\lambda_{\max}|) \tag{1.114}$$

Ahora bien, para un cierto intervalo de valores $0 < \omega < \omega_{\text{opt}}$ el máximo corresponde a $1-\omega\lambda_{\min}$, mientras que para $\omega_{\text{opt}} < \omega < \omega_{\text{crit}}$ el máximo está dado por $-1+\omega\lambda_{\max}$. Para $\omega = \omega_{\text{opt}}$ ambos valores coinciden y entonces

$$1 - \omega_{\text{opt}}\lambda_{\min} = -1 + \omega_{\text{opt}}\lambda_{\max} \tag{1.115}$$

de donde

$$\omega_{\text{opt}} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \tag{1.116}$$

Además es fácil ver que γ es mínimo para $\omega = \omega_{\text{opt}}$, de ahí el nombre de *coeficiente de relajación óptimo*.

Podemos ver también que, para números de condición muy altos el valor óptimo se encuentra muy cerca del valor crítico,

$$\omega_{\text{opt}} = \frac{\omega_{\text{crit}}}{1 + \kappa(A)^{-1}} \sim \omega_{\text{crit}} \tag{1.117}$$

La tasa de convergencia que se obtiene para ω_{opt} es de

$$\begin{aligned} n_{\text{opt}} &= \frac{\log 10}{\log(1/\gamma_{\text{opt}})} \\ &= \frac{\log 10}{\log[1 - 2\lambda_{\min}/(\lambda_{\max} + \lambda_{\min})]} \\ &= \frac{\log 10}{\log[(\kappa+1)/(\kappa-1)]} \end{aligned} \tag{1.118}$$

que para sistemas mal condicionados ($\kappa \gg 1$) es

$$n_{\text{opt}} = \frac{\kappa \log 10}{2} \sim 1.15\,\kappa \tag{1.119}$$

Método de relajaciones sucesivas. La combinación de Gauss-Seidel puede mejorarse dramáticamente con sobre-relajación para una cierta elección apropiada del parámetro de relajación. Este método es muy popular y se llama SSOR por *"Successive Standard Over-Relaxation"*. Partiendo de (1.94) y reescribiéndolo como

$$D\,\overline{x_{k+1}} + L\,x_{k+1} = b - U\,x_k \tag{1.120}$$

y combinando con la sobre-relajación estándar (1.107)

$$\overline{x_{k+1}} = \omega^{-1}(x_{k+1} - (1-\omega)\,x_k) \tag{1.121}$$

llegamos a

$$D[x_{k+1} - (1-\omega)\,x_k] + \omega L x_{k+1} = \omega\,(b - U\,x_k) \tag{1.122}$$

$$(D + \omega\,L)x_{k+1} = [(1-\omega)\,D - \omega U]x_k + \omega b \tag{1.123}$$

de manera que la matriz de iteración es

$$M_{\text{SOR}} = (D + \omega\,L)^{-1}\,[(1-\omega)\,D - \omega\,U] \tag{1.124}$$

1.6. Guía Nro 1. Métodos Iterativos Estacionarios

1. Generar aleatoriamente una matriz simétrica y positiva definida A y un miembro derecho b con
```
n=5; c=1;
a=rand(n);
a=(a+a');
a=expm(c*a);
b=rand(n,1);
```
Porqué es esta matriz spd?
2. Determinar para qué valores del parámetro de relajación ω el esquema de Richardson es convergente. Determinar la tasa de convergencia.

$$x_{k+1} = x_k + \omega\,(b - Ax_k) \tag{1.125}$$

3. Observar el comportamiento de una componente dada en función de las iteraciones (Graficar curvas de $(xk)_i$ en función de k. Que se observa para $\omega = \omega_{\text{opt}}$? Explicar.
4. **Ecuación de Poisson 1D:** Consideremos la ecuación de Poisson 1-dimensional en un dominio $0 < x < 1$ con condiciones Dirichlet en $x = 0, 1$:

$$\phi'' = -f \tag{1.126}$$

Consideramos una discretización por diferencias finitas de paso constante $h = 1/N$. Los nodos son $x_j = jh$, $j = 0, \ldots, N$ y las ecuaciones para los nodos interiores $1 < j < N - 1$ son

$$h^{-2}(-\phi_{j+1} + 2\phi_j - \phi_{j-1}) = f_j \tag{1.127}$$

ϕ_0 y ϕ_N son datos dados por las condiciones de contorno Dirichlet. El sistema puede ponerse de la forma $Ax = b$ con

$$A = h^{-2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ 0 & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} \tag{1.128}$$

   a) Verificar que A es simétrica y definida positiva.
   b) Calcular los autovalores de A y el número de condición de A, para $N = 10, 30, 100$.
   c) Verificar que el número de condición de A se comporta de la forma $\text{cond}(A) \sim N^2 = 1/h^2$, al refinar.
   d) Graficar las autofunciones de A para los autovalores más bajos y para los más altos.
   e) Verificar cuán buena es la aproximación de $\|A\|$ al radio espectral de A.
   f) Efectuar experimentos numéricos con $\omega = \omega_{\text{opt}}, 0.7\omega_{\text{opt}}$, etc... como antes.
   g) Evaluar las tasas de convergencia en forma experimental y teórica.

5. **Ecuación de Laplace 2-dimensional:** Lo mismo que en el ejercicio anterior pero en 2D en el dominio $0 < x, y < 1$ con condiciones dirichlet homogéneas en todo el contorno y $f = 1$. Una ventaja de los métodos iterativos es que no es necesario armar la matriz del sistema. En efecto, sólo necesitamos poder calcular el residuo r_k a partir del vector de estado x_k. Estudiar

la implementación que se provee a través del script `poisson.m` que llama sucesivamente a la función `laplacian.m`. En `poisson.m` el vector de estado es de tamaño $(N-1)^2$ donde hemos puesto todos los valores de ϕ encolumnados. La función `laplacian(phi)` calcula el laplaciano del vector de iteración ϕ. El laplaciano es calculado convirtiendo primero el vector de entrada a una matriz cuadrada de $(N-1)\times(N-1)$ y despues se evalúa la aproximación estándar de 5 puntos al laplaciano

$$(\Delta\phi)_{ij} = h^{-2}(\phi_{i,j+1} + \phi_{i,j-1} + \phi_{i+1,j} + \phi_{i-1,j} - 4\phi_{ij}) \tag{1.129}$$

a) Estimar el autovalor máximo con la norma 1 de A.

b) Efectuar experimentos numéricos con varios valores de ω. Evaluar tasas de convergencia en forma experimental.

6. **Analogía entre el método de Richardson relajado y la solución seudo-temporal.** Consideremos el sistema ODE's

$$\frac{dx}{dt} = -Ax + b \tag{1.130}$$

Entonces si A tiene autovalores con parte real positiva, la solución $x(t)$ tiende a la la solución de $Ax = b$ para $t \to \infty$. Entonces podemos generar métodos iterativos integrando este sistema en forma numérica. *Consigna:* Demostrar que aplicar el método de forward Euler a esta ecuación ($dx/dt \approx (x_{k+1} - x_k)/\Delta t$) equivale al método de Richardson, donde el paso de tiempo equivale al factor de relajación ω.

Capítulo 2

Método de Gradientes Conjugados

2.1. Métodos de Krylov y propiedad de minimización

Los métodos de Krylov (a diferencia de los métodos estacionarios que vimos hasta ahora), no tienen una matriz de iteración. Los que describiremos más en profundidad son Gradientes Conjugados y GMRES. Ambos se basan en que la k-ésima iteración minimiza alguna medida del error en el espacio afín

$$x_0 + \mathcal{K}_k \tag{2.1}$$

donde x_0 es la iteración inicial y el subespacio de Krylov $\mathcal{K}_k$ es

$$\mathcal{K}_k = \text{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{k-1}r_0\}, \quad k \geq 1 \tag{2.2}$$

El residuo es $r = b - Ax$, de manera que $\{r_k\}$ para $k \geq 0$ denota la secuencia de residuos $r_k = b - Ax_k$. Como antes, $x^* = A^{-1}b$, es la solución del sistema. El método de GC es en realidad un método directo, en el sentido de que llega a la solución en un número finito de pasos, de hecho veremos más adelante que converge en N (o menos) iteraciones, es decir que

$$x_k = x^*, \quad \text{para un cierto } k \text{ con } k \leq N \tag{2.3}$$

GC es un método que sirve en principio sólo para *matrices simétricas y definidas positivas* (spd). Recordemos que A es simétrica si $A = A^T$ y definida positiva si

$$x^T Ax > 0, \quad \text{para todo } x \neq 0 \tag{2.4}$$

Luego veremos que podemos extender cualquier sistema (no simétrico ni definido positivo) a un sistema spd. Como A es spd. podemos definir una norma como

$$\|x\|_A = \sqrt{x^T Ax} \tag{2.5}$$

es la llamada *"norma A"* o *"norma energía"* ya que en muchos problemas prácticos el escalar resultante ($\|x\|_A^2$) representa la energía contenida en el campo representado por el vector x.

El esquema a seguir será

- Descripción formal del método y consecuencias de la propiedad de minimización
- Criterio de terminación, performance, precondicionamiento
- Implementación

La k-ésima iteración de GC x_k minimiza el funcional cuadrático

$$\phi(x) = (1/2)x^T Ax - x^T b \tag{2.6}$$

en $x_0 + \mathcal{K}_k$.

Notemos que si hacemos el mínimo sobre todo $\mathbb{R}^N$, entonces si

$$\tilde{x} = \underset{x\in\mathbb{R}^N}{\operatorname{argmin}}\, \phi(x), \tag{2.7}$$

vale que

$$\nabla\phi(\tilde{x}) = A\tilde{x} - b = 0, \quad \text{implica} \;\; \tilde{x} = x^* \tag{2.8}$$

Lema 2.1.1. Sea $S \subset \mathbb{R}^N$, si x_k minimiza ϕ sobre S, entonces también minimiza $\|x^* - x\|_A = \|r\|_{A^{-1}}$ sobre S.

Demostración. Notemos que

$$\begin{aligned} \|x - x^*\|_A^2 &= (x - x^*)^T\, A\, (x - x^*) \\ &= x^T\, A\, x - x^{*T}\, A\, x - x^T\, A\, x^* + x^{*T}\, A\, x^* \end{aligned} \tag{2.9}$$

pero $A = A^T$, entonces $x^{*T}\, A\, x = x^T\, A^T\, x^* = x^T\, A\, x^*$ y $Ax^* = b$, de manera que

$$\begin{aligned} \|x - x^*\|_A^2 &= x^T\, Ax - 2x^T\, b + x^{*T}\, A\, x^* \\ &= 2\phi(x) + x^{*T}\, A\, x^* \end{aligned} \tag{2.10}$$

Pero $x^{*T}\, A\, x^*$=cte de manera que x_k minimiza $\|x - x^*\|_A$. Sea $e = x - x^*$, entonces,

$$\begin{aligned} \|e\|_A^2 &= e^T\, A\, e \\ &= [A(x - x^*)]^T\, A^{-1}\, [A(x - x^*)] \\ &= (b - Ax)^T\, A^{-1}\, (b - Ax) \\ &= \|b - Ax\|_{A^{-1}}^2 \;\; \square. \end{aligned} \tag{2.11}$$

Usaremos este lema para el caso $S = x_0 + \mathcal{K}_k$.

2.2. Consecuencias de la propiedad de minimización.

El lema 2.1.1 implica que, como x_k minimiza ϕ sobre $x_0 + \mathcal{K}_k$,

$$\|x^* - x_k\|_A \leq \|x^* - w\|_A \tag{2.12}$$

para todo $w \in x_0 + \mathcal{K}_k$. Como $w \in x_0 + \mathcal{K}_k$ puede ser escrito como

$$w = \sum_{j=0}^{k-1} \gamma_j\, A^j\, r_0 + x_0 \tag{2.13}$$

para ciertos coeficientes $\{\gamma_j\}$, podemos expresar $x^* - w$ como

$$x^* - w = x^* - x_0 - \sum_{j=0}^{k-1} \gamma_j\, A^j\, r_0 \tag{2.14}$$

Pero

$$r_0 = b - Ax_0 = A(x^* - x_0) \tag{2.15}$$

entonces

$$\begin{aligned} x^* - w &= (x^* - x_0) - \sum_{j=0}^{k-1} \gamma_j \, A^{j+1} \, (x^* - x_0) \\ &= p(A) \, (x^* - x_0) \end{aligned} \tag{2.16}$$

donde el polinomio

$$p(z) = 1 - \sum_{j=0}^{k-1} \gamma_j \, z^{j+1} \tag{2.17}$$

tiene grado k y satisface $p(0) = 1$. Entonces,

$$\|x^* - x_k\|_A = \min_{p \in \mathcal{P}_k, p(0)=1} \|p(A) \, (x^* - x_0)\|_A \tag{2.18}$$

$\mathcal{P}_k$ denota el conjunto de los polinomios de grado k. El teorema espectral para matrices spd afirma que existe una base de autovectores $\{u_i\}_{i=1}^N$ con autovalores $\{\lambda_i\}$

$$A \, u_i = \lambda_i \, u_i \tag{2.19}$$

con u_i, λ_i reales, $\lambda_i > 0$ y los u_i ortogonales entre sí, es decir que

$$u_i^T u_j = \delta_{ij} \tag{2.20}$$

Además formamos las matrices

$$\begin{aligned} U &= \left[\begin{array}{cccc} u_1 & u_2 & \dots & u_N \end{array} \right] \\ \Lambda &= \mathrm{diag}\{\lambda_1, \lambda_2, \dots, \lambda_N\} \end{aligned} \tag{2.21}$$

La descomposición diagonal de A es

$$A = U \, \Lambda \, U^T \tag{2.22}$$

Además

$$\begin{aligned} A^j &= (U \Lambda U^T) \, (U \Lambda U^T) \, \dots \, (U \Lambda U^T) \\ &= U \Lambda^j U^T \end{aligned} \tag{2.23}$$

y

$$p(A) = U \, p(\Lambda) \, U^T \tag{2.24}$$

Definamos

$$A^{1/2} = U \, \Lambda^{1/2} \, U^T \tag{2.25}$$

y notemos que

$$\|x\|_A^2 = x^T \, A \, x = \left\| A^{1/2} \, x \right\|_2^2 \tag{2.26}$$

Entonces, para todo $x \in \mathbb{R}^N$

$$\begin{aligned} \|p(A) \, x\|_A &= \left\| A^{1/2} \, p(A) \, x \right\|_2 \\ &= \left\| p(A) \, A^{1/2} \, x \right\|_2 \\ &\leq \|p(A)\|_2 \, \left\| A^{1/2} \, x \right\|_2 \\ &= \|p(A)\|_2 \, \|x\|_A \end{aligned} \tag{2.27}$$

Volviendo a (2.18)

$$\|x_k - x^*\|_A \leq \|x_0 - x^*\|_A \min_{p\in\mathcal{P}_k, p(0)=1} \left\{ \max_{z\in\sigma(A)} |p(z)| \right\} \tag{2.28}$$

donde $\sigma(A)$ es el conjunto de autovalores de A.

Notar que en (2.27) podríamos haber usado directamente la norma energía, entonces hubiéramos llegado inmediatamente a

$$\|p(A)\, x\|_A = \|p(A)\|_A \, \|x\|_A \tag{2.29}$$

la diferencia es que en está expresión la norma de $p(A)$ está expresada en la norma energía y en (2.27) está en la norma 2. Pero resultan ser lo mismo, ya que

$$\begin{aligned}
\|p(A)\|_A^2 &= \max_{x\neq 0} \frac{\|p(A)x\|_A^2}{\|x\|_A^2}, \\
&= \max_{x\neq 0} \frac{x^T\, p(A)\, A\, p(A)\, x}{x^T A x}, \\
&= \max_{x\neq 0} \frac{x^T\, p(A)\, A^{1/2}\, A^{1/2}\, p(A)\, x}{x^T A x}, \\
&= \max_{x\neq 0} \frac{x^T\, A^{1/2}\, p(A)\, p(A)\, A^{1/2}\, x}{x^T\, A^{1/2}\, A^{1/2}\, x}, \\
&= \max_{y\neq 0} \frac{y^T\, p(A)\, p(A)\, y}{y^T y}, \\
&= \max_{y\neq 0} \frac{\|\, p(A)\, y\|_2^2}{\|y\|_2^2}, \\
&= \|p(A)\|_2 .
\end{aligned} \tag{2.30}$$

Aquí hemos usado que $A^{1/2} p(A) = p(A)\, A^{1/2}$ (que es fácil de demostrar) y hemos hecho el cambio de variable $y = A^{1/2} x$. **Corolario 2.2.1.** Sea A spd y $\{x_k\}$ las iteraciones de GC. Sea $\bar{p}_k$ cualquier polinomio de grado k tal que $\bar{p}_k(0) = 1$, entonces

$$\frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \leq \max_{z\in\sigma(A)} |\bar{p}_k(z)| \tag{2.31}$$

Los polinomios que satisfacen $\bar{p}_k(0) = 1$ se llaman *polinomios residuales*.

Definición 2.2.1. El conjunto de polinomios residuales de orden k es

$$\mathcal{P}_k = \{p \;/\; p \text{ es un polinomio de grado } k \text{ y } p(0) = 1\} \tag{2.32}$$

La forma de estimar la convergencia de GC es construir secuencias de polinomios residuales basados en la información de como están distribuidos los autovalores de A (es decir de $\sigma(A)$). Un primer resultado es ver que GC es un método directo

Teorema 2.2.1. Sea A spd. Entonces GC converge antes de las N iteraciones.

Demostración. Sea $\{\lambda_i\}_{i=1}^N$ los autovalores de A. Tomemos como polinomio residual

$$\bar{p}(z) = \prod_{i=1}^{N} (\lambda_i - z)/\lambda_i \tag{2.33}$$

Pero $\bar{p} \in \mathcal{P}_N$ ya que tiene grado N y $\bar{p}(0) = 1$. Entonces, de la estimación de error (2.31)

$$\|x_N - x^*\|_A \leq \|x_0 - x^*\|_A \max_{z\in\sigma(A)} |\bar{p}(z)| = 0 \tag{2.34}$$

ya que, por construcción, $\bar{p}(z) = 0$ para todos los $z \in \sigma(A)$. □.

Sin embargo, desde el punto de vista práctico esto no es tan bueno como suena. Veremos que, bajo ciertas condiciones, la convergencia puede ser muy lenta y desde el punto de vista práctica haya que esperar hasta N iteraciones para que el residuo baje de la tolerancia aceptable.

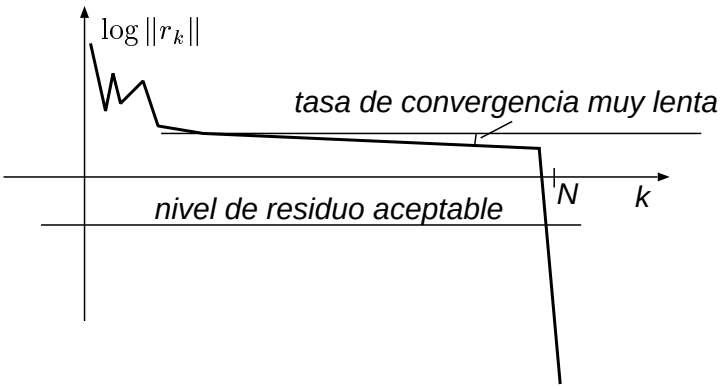

Figura 2.1: GC con tasa de convergencia muy lenta.

Es decir, si evaluamos a GC como un método iterativo, entonces debemos poder evaluar la tasa de convergencia para $k < N$ (la pendiente de la curva de convergencia).

Teorema 2.2.2. Sea A spd con autovalores $\{u_i\}_{i=1}^{N}$. Sea b una combinación lineal de k de los autovectores de A. Por simplicidad asumiremos que los autovectores de A están ordenados de manera que estos autovectores son los primeros k

$$b = \sum_{l=1}^{k} \gamma_l \, u_l \tag{2.35}$$

Entonces la iteración de GC para $Ax = b$ con $x_0 = 0$ termina en, a lo sumo, k iteraciones

Demostración. Por el teorema espectral

$$x^* = \sum_{l=1}^{k} (\gamma_l/\lambda_l) \, u_l \tag{2.36}$$

ya que

$$\begin{aligned} Ax^* &= \sum_{l=1}^{k} (\gamma_l/\lambda_l) A u_l \\ &= \sum_{l=1}^{k} (\gamma_l/\lambda_l) \lambda_l \, u_l \\ &= b \end{aligned} \tag{2.37}$$

Usamos el polinomio residual

$$\bar{p}(z) = \prod_{l=1}^{k} (\lambda_l - z)/\lambda_l \tag{2.38}$$

y puede verse que $\bar{p} \in \mathcal{P}_k$ y $\bar{p}(\lambda_l) = 0$ para $1 \leq l \leq k$ y

$$\bar{p}(A)\, x^* = \sum_{l=1}^{k} \bar{p}(\lambda_l)\, (\gamma_l/\lambda_l)\, u_l = 0 \tag{2.39}$$

De manera que, por (2.18) y $x_0 = 0$ se deduce que

$$\|x^* - x_k\|_A \leq \|\bar{p}(A)\, x^*\|_A = 0 \ \square. \tag{2.40}$$

Teorema 2.2.3. Sea A spd y supongamos que hay exactamente $k \leq N$ autovalores distintos de A. Entonces GC termina en, a lo sumo, k iteraciones

Demostración. Usar el polinomio residual

$$\bar{p}_k(z) = \prod_{l=1}^{k} (\lambda_l - z)/\lambda_l \ \square. \tag{2.41}$$

2.3. Criterio de detención del proceso iterativo.

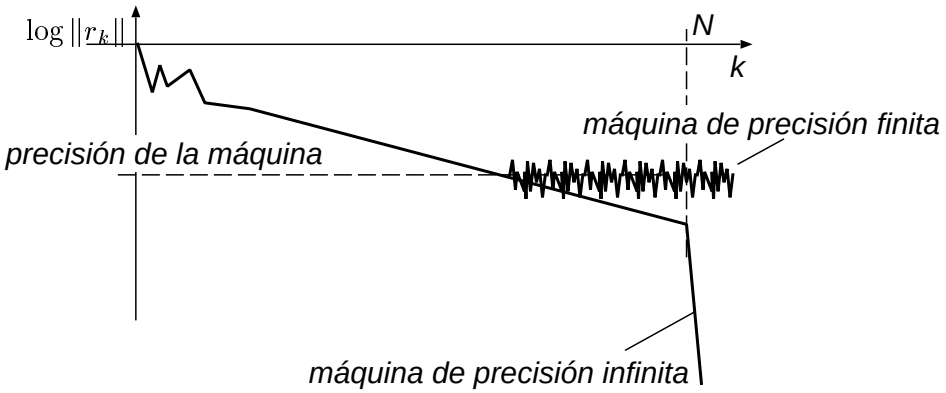

Figura 2.2: Comportamiento del residuo en máquinas de precisión finita.

En una máquina de precisión infinita debemos ver que el residuo desciende bruscamente a cero en la iteración N (puede ser antes bajo ciertas condiciones, como hemos visto en algunos casos especiales), pero en la práctica no se itera GC hasta encontrar la solución exacta sino hasta que cierto criterio sobre el residuo (por ejemplo $\|r_k\| < 10^{-6}$) es alcanzado. De hecho, debido a errores de redondeo, ocurre que no se puede bajar de un cierto nivel de residuo relacionado con la precisión de la máquina (10^{-15} en Fortran doble precisión y también en Octave) y el número de operaciones necesarias para calcular el residuo. Entonces, si ese nivel de residuo está por encima del valor del residuo que obtendríamos en la iteración $N - 1$ (ver figura 2.2), no veremos el efecto de la convergencia en N iteraciones, sino que GC se comporta como un método iterativo, de manera que lo importante es la tasa de convergencia media del método.

En la práctica se usa usualmente como criterio de detención del proceso iterativo

$$\|b - Ax_k\|_2 \leq \eta\, \|b\|_2 \tag{2.42}$$

Sin embargo, las estimaciones de error basadas en la propiedad de minimización están expresadas en la norma energía del error

$$\frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \leq \max_{z \in \sigma(A)} |\bar{p}_k(z)| \tag{2.43}$$

El siguiente lema relaciona la norma euclídea del error con la norma energía del error

Lema 2.3.1. Sea A spd, con autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. Entonces para todo $z \in \mathbb{R}^N$

$$\left\| A^{1/2} z \right\|_2 = \|z\|_A \tag{2.44}$$

y

$$\lambda_N^{1/2} \, \|z\|_A \leq \|Az\|_2 \leq \lambda_1^{1/2} \, \|z\|_A \tag{2.45}$$

Demostración.

$$\|z\|_A^2 = z^T A z = (A^{1/2} z)^T (A^{1/2} z) = \left\| A^{1/2} z \right\|_2^2 \tag{2.46}$$

Sean $\{u_i, \lambda_i\}$ los autovectores, autovalores de A, con $u_i^T u_j = \delta_{ij}$. Entonces,

$$\begin{aligned} z &= \sum_{i=1}^{N} (u_i^T z) \, u_i \\ Az &= \sum_{i=1}^{N} \lambda_i \, (u_i^T z) \, u_i \end{aligned} \tag{2.47}$$

Entonces,

$$\begin{aligned} \lambda_N \left\| A^{1/2} \, z \right\|_2^2 &= \lambda_N \sum_{i=1}^{N} \lambda_i \, (u_i^T z)^2 \\ &\leq \sum_{i=1}^{N} \lambda_i^2 \, (u_i^T z)^2 = \|Az\|_2^2 \\ &\leq \lambda_1 \sum_{i=1}^{N} \lambda_i \, (u_i^T z)^2 = \lambda_1 \left\| A^{1/2} \, z \right\|_2^2 \ \square. \end{aligned} \tag{2.48}$$

Lema 2.3.2.

$$\frac{\|b - Ax_k\|_2}{\|b\|_2} \leq \frac{\sqrt{\kappa_2(A)} \, \|r_0\|_2}{\|b\|_2} \frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \tag{2.49}$$

Recordemos que en la norma L_2 para matrices spd. vale que

$$\begin{aligned} \|A\|_2 &= \text{máx autovalor de } A = \lambda_1 \\ \left\|A^{-1}\right\|_2 &= \frac{1}{\text{mín autovalor de } A} = \frac{1}{\lambda_N} \\ \kappa_2(A) &= \lambda_1 / \lambda_N \end{aligned} \tag{2.50}$$

Usando (2.44) y (2.45)

$$\frac{\|b - Ax_k\|_2}{\|b - Ax_0\|_2} = \frac{\|A\,(x^* - x_k)\|_2}{\|A\,(x^* - x_0)\|_2} \leq \frac{\lambda_1^{1/2} \, \|x^* - x_k\|_A}{\lambda_N^{1/2} \, \|x^* - x_0\|_A} \ \square. \tag{2.51}$$

Consideremos ahora un ejemplo simple

$$x_0 = 0, \quad \lambda_1 = 11, \quad \lambda_N = 9, \tag{2.52}$$

Por lo tanto $k_2 = 1.22$ (relativamente pequeño). Tomemos el polinomio (ver figura 2.3)

$$\bar{p}(z) = (10 - z)^k / 10^k \in \mathcal{P}_k \tag{2.53}$$

Como vemos en la figura todos los polinomios se anulan en $z = 10$ la mitad del intervalo donde se encuentran los autovalores. Esta región está sombreada en la figura. El máximo valorde $\bar{p}_k(z)$ sobre el intervalo se alcanza en los extremos del intervalo

$$\max_{9 \le z \le 11} |\bar{p}_k(z)| = |\bar{p}_k(9)| = |\bar{p}_k(11)| = 10^{-k} \tag{2.54}$$

lo cual implica que la tasa de convergencia es $\gamma = 1/10$ y el número de iteraciones por orden de magnitud es

$$n = \frac{\log(10)}{\log(1/\gamma)} = 1 \tag{2.55}$$

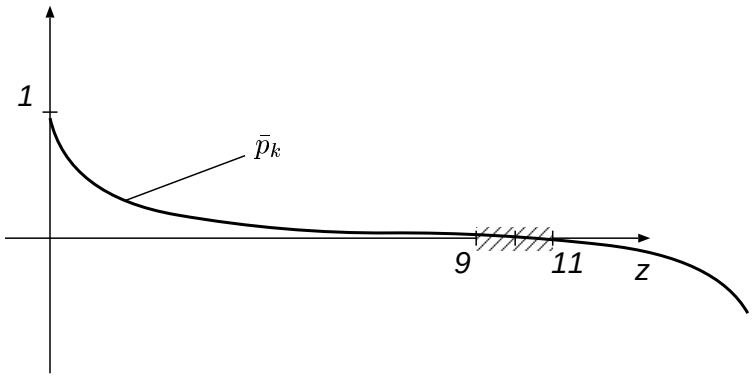

Figura 2.3: Polinomio residual apropiado para el caso (2.52)

Mientras tanto, para los residuos

$$\begin{aligned} \frac{\|Ax_k - b\|_2}{\|b\|_2} &\le \sqrt{1.22} \times 10^{-k} \\ &= 1.10 \times 10^{-k} \end{aligned} \tag{2.56}$$

Para obtener la mejor estimación basada solamente en la información del autovalor máximo y mínimo debemos construir un polinomio de tal forma que tome los valores más bajos posibles en el intervalo $\lambda_1 \le z \le \lambda_n$ (ver figura 2.4). Estos polinomios pueden construirse en base a los *polinomios de Tchebyshev*. Efectivamente, hagamos el cambio de variable

$$(z - \lambda_N)/(\lambda_1 - \lambda_N) = (1 - \cos\theta)/2 = (1 - x)/2 \tag{2.57}$$

de manera que $\theta = 0$ en $z = \lambda_N$ y $\theta = \pi$ en $z = \lambda_1$. Los polinomios de Tchebyshev $T_n(x)$ se definen como

$$T_n(x) = \cos n\theta \tag{2.58}$$

Por ejemplo

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \end{aligned} \tag{2.59}$$

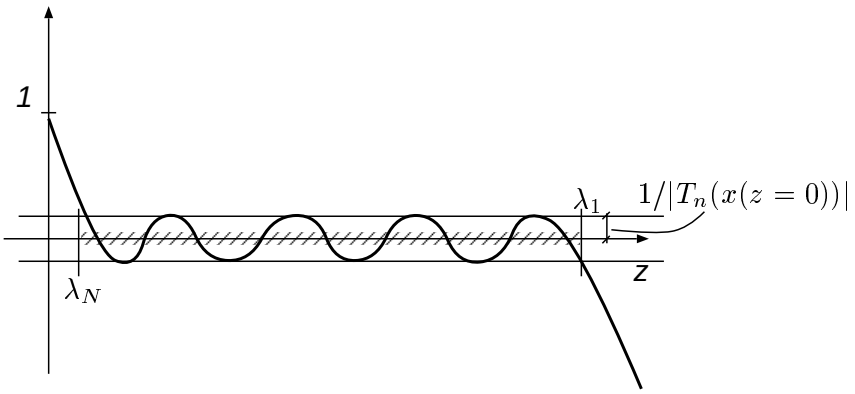

Figura 2.4: Polinomio residual basado en los polinomios de Tchebyshev

Por construcción vemos que $|T_n| \leq 1$ en $|x| \leq 1$, o sea $\lambda_n \leq z \leq \lambda_1$. Entonces tomamos $p_k(z) = T_n(x(z))/T_n(x(z=0))$. Como $x(z=0)$ cae fuera del intervalo $|x| \leq 1$ y T_n crece fuertemente (como x^n) fuera del mismo, es de esperar que $T_n(x(z=0)) \gg 1$ y entonces $|p_k(z)| \ll 1$ en $\lambda_n \leq z \leq \lambda_1$. Mediante estimaciones apropiadas para el valor de $T_n(x(z=0))$ puede demostrarse que

$$\|x_k - x^*\|_A \leq 2 \|x_0 - x^*\|_A \left(\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1}\right)^k \tag{2.60}$$

Podemos ver que, si $\kappa \gg 1$, entonces podemos aproximar

$$\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \approx 1 - \frac{2}{\sqrt{\kappa}} \tag{2.61}$$

y entonces

$$n \approx \frac{\log(10)}{2} \sqrt{\kappa} = 1.15\sqrt{\kappa} \tag{2.62}$$

que debe ser comparada con $n \approx 1.15\kappa$ para Richardson con $\omega = \omega_{\text{opt}}$. La ventaja es clara, teniendo en cuenta que (como veremos después) el costo (número de operaciones) de una iteración de gradientes conjugados es similar al de una iteración de Richardson.

Sin embargo, la convergencia puede ser mucho mejor que la dada por la estimación anterior, dependiendo de la distribución de los autovalores. Por ejemplo, asumamos que los autovalores están distribuidos en (ver figura 2.5)

$$1 \leq \lambda \leq 1.5 \text{ ó } 399 \leq \lambda \leq 400 \tag{2.63}$$

Entonces $\kappa = 400$ y la estimación anterior (basada sólo en el número de condición) da

$$n = 1.15\sqrt{400} = 23 \tag{2.64}$$

mientras que si tomamos el polinomio residual de la forma

$$\bar{p}_{3k} = \frac{(1.25 - z)^k (400 - z)^{2k}}{1.25^k \, 400^{2k}} \tag{2.65}$$

Entonces debemos estimar

$$\max_{1 \leq z \leq 1.5} |p_{3k}(z)| = (0.25/1.25)^k = 0.2^k \tag{2.66}$$

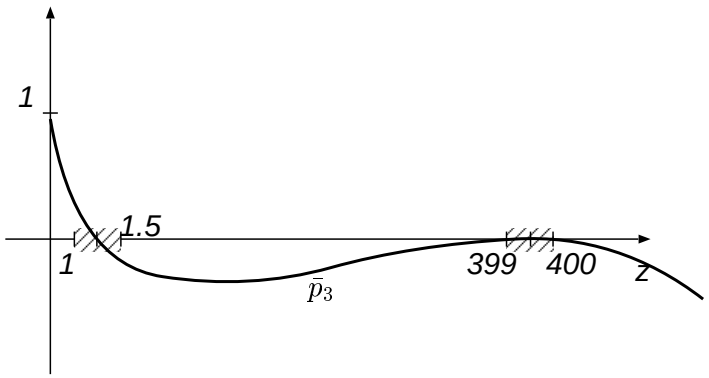

Figura 2.5: Polinomio residual apropiado para un espectro con dos clusters de autovalores

y

$$\begin{aligned} \max_{399\le z\le 400} |p_{3k}(z)| &\le (400/1.25)^k \, (1/400)^{2k} \\ &= 1/(1.25\times 400)^k \end{aligned} \tag{2.67}$$

Por lo tanto,

$$\max_{z\in\sigma(A)} |\bar{p}_{3k}(z)| \le 0.2^k \tag{2.68}$$

y $\gamma = 0.2^{1/3}$,

$$n = \frac{\log 10}{1/3\log(1/0.2)} = 4.29 \tag{2.69}$$

que representa una tasa de convergencia 5 veces más rapida que la predicha sólo en base al número de condición de A.

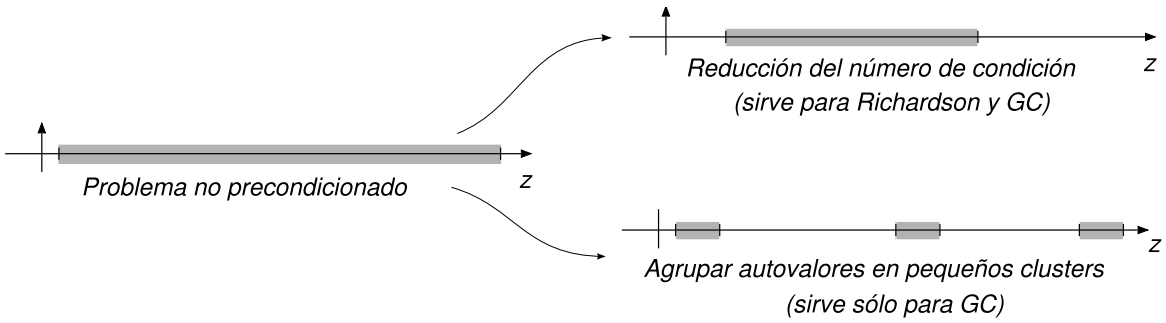

Figura 2.6: Posibles estrategias de precondicionamiento. El eje z son los autovalores de A.

En el contexto de GC la convergencia se puede mejorar tratando de encontrar precondicionamientos que disminuyan el número de condición o bien que los autovalores están agrupados en pequeños *"clusters"* (ver figura 2.6).

2.4. Implementación de gradientes conjugados

La implementación de GC se basa en que conociendo x_k pueden ocurrir dos situaciones. O bien $x_{k+1} = x_k = x^*$, o bien

$$x_{k+1} = x_k + \alpha_{k+1}\, p_{k+1} \tag{2.70}$$

donde $p_{k+1} \neq 0$ es una *dirección de búsqueda* que puede obtenerse en forma sencilla y α_{k+1} es un escalar que se obtiene minimizando el funcional de GC sobre la recta,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\,\phi(x_k + \alpha p_{k+1}) = 0, \text{ en } \alpha = \alpha_{k+1} \tag{2.71}$$

Recordar que el funcional estaba definido como

$$\phi(x) = \frac{1}{2}\, x^T A x - x^T b \tag{2.72}$$

Entonces

$$\begin{aligned} \frac{\mathrm{d}\phi}{\mathrm{d}\alpha} &= p_{k+1}^T\, \nabla\phi \\ &= p_{k+1}^T\, [A(x_k + \alpha_{k+1}\, p_{k+1}) - b] = 0 \end{aligned} \tag{2.73}$$

de donde

$$\alpha_{k+1} = \frac{p_{k+1}^T\, (b - A x_k)}{p_{k+1}^T\, A\, p_{k+1}} \tag{2.74}$$

Si $x_{k+1} = x_k$ entonces $\alpha = 0$, pero esto sólo ocurre si x_k es la solución.

Lema. $r_l \in \mathcal{K}_k$ para todo $l < k$

Demostración. Lo haremos por inducción en k. Para $k = 1$ $\mathcal{K}_k = \mathrm{span}\{r_0\}$ y obviamente se verifica que $r_0 \in \mathcal{K}_k$. Ahora asumiendo que es válido para k lo demostraremos para $k+1$. Para $l < k$, se cumple que $r_l \in \mathcal{K}_k \subset \mathcal{K}_{k+1}$, por lo tanto sólo resta demostrarlo para r_k. Pero

$$x_k = x_0 + \sum_{j=0}^{k-1} \alpha_j\, A^j\, r_0 \tag{2.75}$$

y entonces,

$$\begin{aligned} r_k &= b - A x_k \\ &= r_0 - \sum_{j=0}^{k-1} \alpha_j\, A^{j+1}\, r_0 \in \mathcal{K}_{k+1}\ \square. \end{aligned} \tag{2.76}$$

Lema 2.4.1. Sea A spd. y $\{x_k\}$ las iteraciones de GC, entonces

$$r_k^T r_l = 0, \quad \text{para todo } 0 \leq l < k \tag{2.77}$$

Demostración. Como x_k minimiza ϕ en $x_0 + \mathcal{K}_k$, entonces para todo $\xi \in \mathcal{K}_k$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(x_k + t\xi) = \nabla\phi(x_k + t\xi)^T\, \xi = 0, \quad \text{en } t = 0 \tag{2.78}$$

pero $\nabla\phi = -r$, entonces

$$-\, r_k^T \xi = 0, \quad \text{para todo } \xi \in \mathcal{K}_k\ \square. \tag{2.79}$$

Ahora bien, si $x_{k+1} = x_k$ entonces $r_k = r_{k+1}$ y

$$\|r_k\|_2^2 = r_k^T r_k = r_k^T r_{k+1} = 0 \tag{2.80}$$

por lo tanto $x_k = x^*$. De paso esto permite ver también que GC converge en N iteraciones (cosa que ya hemos demostrado basándonos en la propiedad de minimización.) Efectivamente, supongamos que $r_k \neq 0$ entonces

$$\dim \mathcal{K}_{k+1} = \dim \mathcal{K}_k + 1 \tag{2.81}$$

Pero para $k = N$ $\dim \mathcal{K}_k = N$ entonces $r_{k+1} = 0$.

Lema 2.4.2. Si $x_k \neq x^*$ entonces $x_{k+1} = x_k + \alpha_{k+1}\, p_{k+1}$ donde p_{k+1} está determinado (a menos de una constante multiplicatva) por

$$p_{k+1} \quad \in \mathcal{K}_{k+1} \tag{2.82}$$

$$p_{k+1}^T A\xi \quad = 0, \quad \text{para todo } \xi \in \mathcal{K}_k \tag{2.83}$$

Demostración. Como $\mathcal{K}_k \subset \mathcal{K}_{k+1}$

$$\nabla\phi(x_{k+1})^T \xi = 0, \quad \text{para todo } \xi \in \mathcal{K}_k \tag{2.84}$$

$$\left[Ax_k + \alpha_{k+1} A p_{k+1} - b\right]^T \xi = 0 \tag{2.85}$$

pero $Ax_k - b = r_k \perp \xi$ para todo $\xi \in \mathcal{K}_k$ de manera que

$$p_{k+1}^T A\xi = 0 \tag{2.86}$$

y como $\dim \mathcal{K}_{k+1} = \dim \mathcal{K}_k + 1$, esto define únicamente a p_{k+1}.

Esta propiedad se llama que p_{k+1} es *"conjugado"* a $\mathcal{K}_k$. Ahora bien, tomando r_k y ortogonalizándolo con respecto a $\mathcal{K}_k$ podemos obtener p_{k+1}. Entonces,

$$p_{k+1} = r_k + w_k, \quad \text{con } w_k \in \mathcal{K}_k \tag{2.87}$$

Teorema 2.4.1. Sea A spd. y asumamos que $r_k \neq 0$. Sea $p_0 = 0$. Entonces $p_{k+1} = r_k + \beta_{k+1} p_k$ para algún escalar β_{k+1} y $k \geq 0$.

Demostración. ver libro.

Lema 2.4.3. Las siguientes fórmulas son también válidas para obtener los α_k y β_k.

$$\alpha_{k+1} = \frac{\|r_k\|_2^2}{p_{k+1}^T\, Ap_{k+1}}, \quad \beta_{k+1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} \tag{2.88}$$

Demostración. Ver libro.

Algoritmo 2.4.1. `cg`$(x, b, A, \epsilon, k_{\max})$

1. $r = b - Ax$, $\rho_0 = \|r\|_2^2$, $k = 1$
2. Do while $\sqrt{\rho_{k-1}} > \epsilon \|b\|_2$ y $k < K_{\max}$
   a. If $k = 1$ then
      $p = r$
   else
      $\beta = \rho_{k-1}/\rho_{k-2}$
      $p = r + \beta p$
   endif
   b. $w = Ap$
   c. $\alpha = \rho_{k-1}/(p^T w)$
   d. $x = x + \alpha p$
   e. $r = r - \alpha w$
   f. $\rho_k = \|r_k\|_2^2$
   g. $k = k + 1$

Notar que no es necesario formar explítamente la matriz A (ni siquiera en forma "sparse", es decir, la lista de $\{i, j, a_{ij}\}$), sino que solo es necesario definir una rutina que, dado x calcule Ax. Esto se llama una *operación matriz-vector*. Debido a esta propiedad de no necesitar tener la matriz almacenada, GC es llamado un método *mátrix free*.

Costo de GC. Las variables que se deben mantener almacenadas durante la iteración son x, w, p, r o sea $4N$ elementos. En cuanto al número de operaciones,

- 1 producto matriz vector (paso 2.b)
- 2 productos escalares (pasos 2.c y 2.f)
- 3 "daxpys" (pasos 2.a, 2.d y 2.e).

Una operación "daxpy" es aquella de la forma $y \leftarrow \alpha x + y$, es decir adicionar a un vector y un múltiplo escalar de otro vector x. (El nombre proviene de que la rutina de la librería BLAS que realiza esta operación y que a su vez es una descripción mnemotécnica de la operación que se realiza. La "d" inicial corresponde a la versión doble precisión). De todos, el que requiere más operaciones es generalmente el producto matriz vector, sobre todo en la versión *matrix free* y sobre todo cuanto más compleja es la física del problema. Por ejemplo, si x representa el vector de potenciales nodales guardados por columnas para una malla homogénea de paso h (como en la función `laplacian.m` usada en la Guía 1), entonces la implementación de Ax es

```
phi=reshape(phi,n-1,n-1);

% range of indices of internal nodes
II=(2:n);
JJ=II;

% Include the boundary nodes
Phi=zeros(n+1);
Phi((2:n),(2:n))=phi;

% Use the standard 5 point formula
lap_phi=((-Phi(II+1,JJ)...
         -Phi(II-1,JJ)...
         -Phi(II,JJ+1)...
         -Phi(II,JJ-1)...
         +4*Phi(II,JJ))/h^2);

lap_phi=lap_phi(:);
```

donde vemos que básicamente el número de operaciones necesario es $O(5N)$, ya que 5 es el número de puntos involucrados en el stencil de Poisson. Sin embargo, si la complejidad de la representación aumenta el cómputo se mantiene en general $O(N)$ pero con cada vez más operaciones por nodo. Por ejemplo, si la malla está refinada hacia los lados, es decir si el h no es constante, entonces hay que multiplicar cada fila por un h distinto. Si además permitimos que las líneas de la malla no sean paralelas a los ejes, entonces habrá que calcular los *coeficientes métricos* de la malla. Si consideramos el uso de mallas no estructuradas con el método de los elementos finitos el cálculo de las matrices de cada elemento aumentará todavía más el número de operaciones por nodo.

Costo de GC como método directo. Si bien el método de GC se usa raramente como método directo, por razones que veremos más adelante, vamos a estimar el número de operaciones necesarios y compararlo con eliminación de Gauss. Considerando un cuadrado de $N = n \times n$ nodos en 2D y un cubo de $N = n \times n \times n$ en 3D, tenemos que,

- Ancho de banda de la matriz: $m = n$ en 2D, $m = n^2$ en 3D.
- Número total de incógnitas: $N = n^2$ en 2D, $N = n^3$ en 3D.
- Número de op. Gauss: N^2 en 2D, $N^{2.33}$ en 3D
- Número de op. GC: N^2 en 2D, N^2 en 3D

Hemos hecho uso de que el número de operaciones para eliminación de Gauss es Nm^2. Para GC hemos usado

$$\begin{aligned} \text{número de op.} &= O(\text{número de iter.} \times \text{nro. de op. por iter.}) && (2.89)\\ &= O(N^2) && (2.90) \end{aligned}$$

Vemos que, incluso como método directo, GC llega a ser más competitivo en 3D. De todas formas, como ya hemos mencionado, por el problema de precisión finita (ver figura 2.2) en general para problemas grandes se llega a la precisión de la máquina antes de las N iteraciones.

En cuanto a la capacidad de almacenamiento, GC obviamente requiere mucho menos memoria ($O(N)$ para GC contra $O(N^{1.5})$ en 2D y $O(N^{1.66})$ en 3D para Gauss).

Comparación como método iterativo con Richardson. Tanto para Richardson como para GC el costo para bajar un orden de magnitud el residuo es, básicamente

$$\begin{aligned} &\text{Nro. de opr. para bajar el residuo un orden de magnitud} = && (2.91)\\ &= n \times \text{nro. de oper. por iteración} \end{aligned}$$

Donde n es el número de iteraciones necesario para bajar el error un orden de magnitud. Asumiendo que el costo de las iteraciones es prácticamene el mismo para los dos métodos (lo cual es válido si hacemos una implementación *matrix free* con una formulación relativamente compleja, por ejemplo FEM), entonces los costos relativos están dados por las tasas de convergencia y, usando que en 3D $\kappa \sim n^2 = N^{2/3}$.

$$\begin{aligned} n(\text{Richardson con } \omega = \omega_{\text{opt}}) &\sim \kappa = N^{2/3} && (2.92)\\ n(\text{GC }) &\sim \sqrt{\kappa} = N^{1/3} && (2.93) \end{aligned}$$

con lo cual la ganancia es evidente.

2.5. Los "verdaderos residuos".

El algoritmo `cg(...)` (ver pág. 33) descripto más arriba tiene la particularidad de que los residuos no son calculados directamente usando (1.34) sino que son calculados en el paso (2e). Si bien las expresiones coinciden en una máquina de precisión infinita, debido a errores de redondeo los valores numéricos obtenidos con uno u otro método pueden diferir en una máquina de precisión finita. En la figura 2.7 vemos los residuos calculados en un experimento calculados de las dos formas. El comportamiento de los verdaderos residuos es similar al observado para Richardson en §1.4. Sin embargo, es todavía peor para GC. El residuo no sólo no baja de el umbral $\|r\|_{\text{sat}}$ de saturación sino que empieza

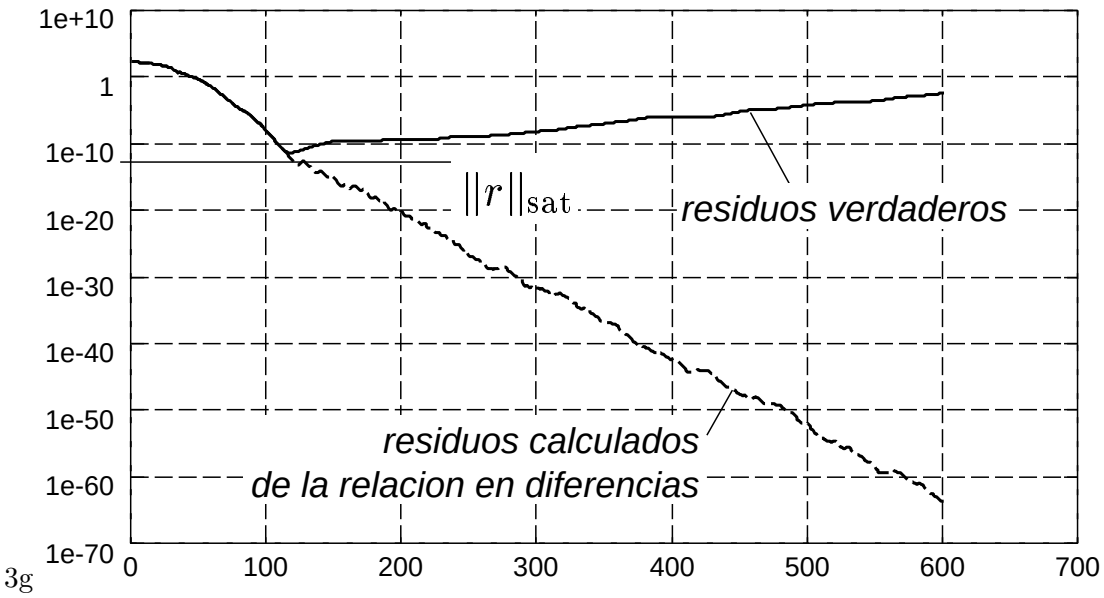

Figura 2.7: Gradientes Conjugados y los residuos verdaderos.

a crecer lentamente. Esto es muy peligroso desde el punta de vista práctico, ya que en el caso de Richardson el efecto de saturación sólo ocasiona un gasto de tiempo de cálculo innecesario, pero en el caso de GC puede ocasionar una pérdida de precisión. Podemos decir que Richardson es un método *estable* ante errores de redondeo, mientras que GC es *inestable*. Esta inestabilidad de GC se debe al hecho de que asume que los residuos van formando una base ortogonal y las direcciones de búsqueda van siendo conjugadas (ec. (2.83)), y estas propiedades se van perdiendo por errores de redondeo. Esta inestabilidad es compartida por todos los métodos que se basan en estas propiedades de conjugación, por ejemplo GMRES y los métodos que veremos después. Para peor los residuos calculados por la expresión recursiva (2e) tienden a seguir descendiendo, de manera que si el usuario no verifica el verdadero valor del residuo, puede creer que realmente el error ha bajado (después de 600 iteraciones) hasta 2×10^{-64}, mientras que el error verdadero está en el orden de 3×10^{-3}.

Cuando calculamos los residuos directamente a partir de (1.34) decimos que se trata de los *verdaderos residuos*. El porqué los residuos calculados según el algoritmo `cg(...)` (ver pág. 33) tienden a bajar, en vez de *rebotar* como lo hacen los verdeaderos residuos, puede entenderse más fácilmente en el algoritmo de Richardson. Efectivamente, puede demostrarse simplemente que los residuos también satisfacen una relación como la (1.70) a saber

$$r_{k+1} = (I - BA)\, r_k \tag{2.94}$$

De manera que también podríamos calcular los residuos utilizando recursivamente esta relación. Pero esta relación no involucra diferencias entre magnitudes grandes como en (1.34) y por lo tanto $\|r_k\|$ calculado por esta expresión sigue descendiendo, independientemente de la precisión de la máquina.

Una forma de corregir el el algoritmo `cg(...)` (ver pág. 33) es agregar una línea en cada iteración que calcule el verdadero residuo, *sólo a los efectos de chequear convergencia*, sin embargo esto involucra un producto matriz vector adicional por iteración, es decir prácticamente duplica el costo del método.

Precondicionamiento. Asumamos que tenemos una matriz M fácil de invertir y tal que $\kappa(MA) \ll \kappa(A)$ o tal que los autovalores están agrupados en clusters. Entonces podemos tratar de resolver

$$(MA)\, x = (Mb) \tag{2.95}$$

en vez del sistema original, ya que este está bien condicionado. Sin embargo, incluso si M es spd. el producto de dos matrices spd. no es necesariamente spd. Basta con ver el ejemplo,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \tag{2.96}$$

$$M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{2.97}$$

A y M son spd. pero

$$MA = \begin{bmatrix} 2 & 2 \\ 1 & 4 \end{bmatrix} \tag{2.98}$$

no es simétrica. Una posibilidad es buscar $M = S^2$ y entonces redefinir

$$(SAS)\, y = (Sb) \tag{2.99}$$

y una vez obtenido y obtener $x = Sy$. Como SAS es equivalente a $S^2A = MA$ tienen la misma distribución de autovalores y SAS sí es spd. Pero falta resolver el problema de hallar S, lo cual puede ser más complicado que hallar M y por otra parte el cálculo de SAS por un vector involucra el cálculo de dos productos matriz-vector adicionales, contra uno sólo si precondicionamos de la forma (2.95). La solución pasa por reproducir el algoritmo de Gradiente Conjugado pero reemplazando el producto escalar por $x^T Mx$ y finalmente resulta en el siguiente algoritmo de GC precondicionado,

Algoritmo 2.4.1. `pcg`$(x, b, A, M, \epsilon, k_{\max})$

1. $r = b - Ax$, $\tau_0 = r^T Mr$, $\rho_0 = \|r\|_2$, $k = 1$
2. Do while $\sqrt{\rho_{k-1}} > \epsilon \|b\|_2$ y $k < K_{\max}$
   a. If $k = 1$ then
      $z = Mr$
   else
      $\beta = \tau_{k-1}/\tau_{k-2}$
      $p = Mr + \beta p$
   endif
   b. $w = Ap$
   c. $\alpha = \tau_{k-1}/(p^T w)$
   d. $x = x + \alpha p$
   e. $r = r - \alpha w$
   f. $\rho_k = \|r_k\|_2^2$,$\tau_k = r_k^T Mr$
   g. $k = k + 1$

La modificación principal del algoritmo pasa por reemplazar p por Mp en las definiciones de los p y reemplazar los productos escalares $x^T y$ por la forma cuadrática $x^T My$. Además, ahora calculamos escalares τ_k para el cálculo de las direcciones p_k y escalares ρ_k para chequear la convergencia. (Mientras que en la versión no precondicionada, $\rho_k = \tau_k$).

El costo adicional más importante para el esquema precondicionado es un producto matriz-vector adicional con la matriz precondicionada. El costo del producto matriz-vector para el precondicionamiento puede variar mucho y depende de la complejidad del precondicionamiento. Por ejemplo, si tomamos M como la inversa de la parte diagonal de A (precondicionamiento Jacobi), entonces el costo es ínfimo, pero en el otro extremo podemos tomar $M = A^{-1}$. Entonces GC converge en una iteración pero el costo de calcular Mx es el costo de invertir A!!! Tenemos entonces que

$$\text{Costo total} = n \times \text{nro de oper. por iteración} \tag{2.100}$$

Cuanto más complejo es el precondicionador el n tiende a disminuir pero el número de operaciones tiende a aumentar debido al costo del cálculo del precondicionamiento. Lo importante entonces, al evaluar la efectividad de un precondicionador es no sólo evaluar cómo aumenta la tasa de convergencia sino *también tener en cuenta el costo de evaluar* Mx.

Precondicionadores. Existen una variedad de precondicionadores propuestos para diferentes problemas. Algunos son muy específicos para ciertas aplicaciones otros son puramente algebraicos, es decir su aplicación es general para toda clase de problema (tal vez no su efectividad!!). En el contexto de la resolución de sistemas lineales provenientes de la discretización de ecuaciones en derivadas parciales por diferencias finitas o volúmenes finitos podemos mencionar los siguientes

- Intrínsecos al problema
  - Resolvedores rápidos de Poisson
  - Multigrilla
  - Descomposición de dominios
  - ADI
- De tipo general
  - Jacobi
  - Factorización incompleta
  - Precondicionamiento polinomial

A continuación haremos una breve descripción de los mismos

Resolvedores rápidos de Poisson. Si consideramos la ecuación de Poisson 1D con condiciones de contorno periódicas la matriz resulta ser

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 & \dots & -1 \\ -1 & 2 & -1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 2 & -1 & \dots & 0 \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ -1 & 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix} \tag{2.101}$$

Sea x de la forma

$$(x^k)_i = \mathrm{e}^{\mathrm{i}2\pi ki/N} \tag{2.102}$$

donde i es la unidad imaginaria. Puede verse que $(x)_i$ es un autovector de A. Esto vale, no sólo para la matriz del laplaciano 1D, sino también para cualquier operador homogéneo (que no depende de x) discretizado sobre una malla homogénea. En este caso las filas de la matriz A son las mismas, sólo que se van corriendo una posición hacia la derecha a medida que bajamos una fila, es decir:

$$A_{ij} = \hat{A}_{i-j} \tag{2.103}$$

donde $\hat{A}_p$ es cíclico en p de período N, es decir $\hat{A}_{p+N} = \hat{A}_p$. Pero los x de la forma (2.102) son la base que induce la transformada de Fourier, de manera que si F es la matriz que tiene como columnas estos autovectores, vale que

$$Fz = \mathrm{fft}(z), \qquad \forall z \tag{2.104}$$

donde fft(z) indica el operador de transformada de Fourier tal como está definido en Matlab. Como las columnas de F son autovectores de A, entonces $F^{-1}AF$ es diagonal y podemos tomar como precondicionamiento

$$M = \text{diag}(\text{diag}(F^{-1}\,AF\,)^{-1}) \tag{2.105}$$

por lo visto, para matrices periódicas, $M = A^{-1}$ y entonces GC convergerá en una iteración. La idea es que (2.105) puede ser un buen precondicionamiento incluso en condiciones más generales, por ejemplo cuando la matriz no es periódica o cuando la malla no es homogénea. En cuanto al costo del precondicionamiento, multiplicar por F y F^{-1} es equivalente a aplicar transformadas y antitransformadas de Fourier (aplicar las operaciones `fft( )` y `ifft( )` de Matlab. Estas requieren $O(N\log_2(N))$ operaciones y la inversión de la parte diagonal de A sólo requiere $O(N)$ operaciones. La idea puede extenderse fácilmente a 2D y 3D.

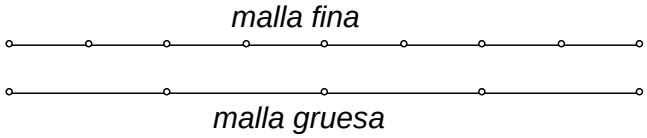

Figura 2.8: Mallas fina y gruesa para las técnicas de multigrilla.

Multigrilla. Si hacemos una descomposición modal del error, puede puede verse que el error en las componentes correspondientes a los autovalores más pequeños converge en general mucho más lentamente que para los autovalores más altos. Esto es más evidente para el método de Richardson, donde los factores de amplificación $1-\omega\lambda_i$ son muy cercanos a 1 para los autovalores más pequeños. A su vez, como ya hemos visto en los ejemplos, los autovalores altos están asociados a frecuencias altas (funciones que son muy oscilatorias espacialmente) mientras que los autovalores bajos están asociados a autofunciones suaves. Esto significa que después de un cierto número de iteraciones el error para las frecuencias altas se debe haber reducido mucho mientras que el grueso del error está en las componentes de baja frecuencia. Como las funciones suaves pueden ser restringidas a una malla más gruesa sin perder información, esto sugiere la idea de proyectar el problema a una malla más gruesa (digamos con $h' = 2h$ y por lo tanto con la mitad de nodos, ver figura 2.8) y realizar una serie de iteraciones sobre esta malla para obtener una corrección. Una vez obtenida ésta se interpola sobre la malla original y se agrega al vector de iteración. La idea es que la corrección va a ser tan buena como si hubiéramos iterado sobre la malla original pero a un costo igual a la mitad ($1/4$ en $2D$ y $1/8$ en 3D) ya que la malla gruesa tiene la mitad de nodos. Esta idea ser puede aplicar recursivamente, ya que al iterar en la malla gruesa va a volver a ocurrir que despues de una serie de iteraciones va a quedar una fuerte componente del error en las frecuencias bajas y estas las podemos corregir iterando sobre una malla $h'' = 2h' = 4h$ y así siguiendo. De esta manera, multigrilla se basa en resolver el problema en una serie de mallas con paso de malla $h, 2h, 4h, \ldots, 2^m h$, haciendo una serie de iteraciones en la malla fina seguida de iteraciones sobre la malla más gruesa y así siguiendo.

Descomposición de dominios. Ver §4.1

Precondicionamiento Jacobi. Consiste en simplemente tomar $M = \text{diag}\,A^{-1}$. Como la matriz a invertir es diagonal el costo de invertirla es muy bajo ($O(N)$ operaciones). Este precondicionamiento no aporta nada en situaciones donde los elementos de la diagonal son aproximadamente constantes (precondicionar con un múltiplo escalar de la identidad no puede nunca mejorar el número de condición). Esto ocurre por ejemplo para el Laplaciano (tanto en 1D como en más dimensiones) cuando el paso de la malla es constante. Cuando la malla no es homogénea (ver figura- 2.9) entonces el número

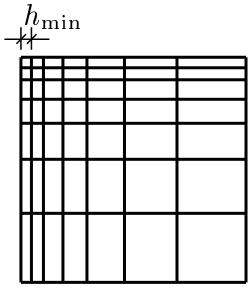

Figura 2.9: Malla refinada hacia una esquina.

de condición es

$$\kappa(A) = O\left(\left(\frac{L}{h_{\min}}\right)^2\right) \tag{2.106}$$

donde $h_{\min}$ es el mínimo h sobre toda la malla y entonces puede ser bastante peor que $O(n^2)$ donde n es el número de elementos en una dirección característica. Los elementos diagonales de A son $\propto h_x^{-2} + h_y^{-2} = O(\min(h_x, h_y)^{-2})$ y puede verse que este precondicionamiento corrige el mal condicionamiento producido por el refinamiento de manera que

$$\kappa(\mathrm{diag}(A)^{-1}A) = O(n^2) \tag{2.107}$$

Factorización incompleta. Consideremos la factorización Cholesky A, $A = BB^T$. Entonces este precondicionamiento consiste en descartar elementos de B de manera de reducir su ancho de banda. En la implementación práctica el descarte se va haciendo a medida de que se va factorizando la matriz, basándose en el valor absoluto del elemento B_{ij} que se está evaluando y su distancia a la diagonal $|i - j|$. Por supuesto se trata de descartar aquellos elementos que tienen un menor valor absoluto y que se encuentran más alejados de la diagonal.

Precondicionamiento polinomial. Consiste en encontrar un polinomio $p(z)$ tal que $M = p(A) \approx A^{-1}$. El criterio para construir el polinomio en cuestión es similar al utilizado para construir los polinomios residuales que permiten obtener la estimación de convergencia (2.60) en base a los polinomios de Tchebyshev. El costo de un tal precondicionmiento es evaluar $Mx = p(A)x$ que involucra m productos matriz-vector, donde m es el orden del polinomio de Tchebyshev. Como es usual en la evaluación de polinomios, se utiliza la forma anidada, también llamada de Hörner, como en el siguiente script

```
y=0;
for k=0:m
  y=p(k)*x+A*y;
end
```

donde (usando la notación de Octave) $p(x) = p_1\, x^m + p_2\, x^{m-1} + \ldots + p_{m+1}$ y los coeficientes p_i están almacenados en un vector de longitud $m + 1$. En la versión *matrix free*, el producto Ax está definido por medio de una rutina. Sea `w=prodvec(x)` la rutina que retorna $w = Ax$ cuando se le pasa x entonces el algoritmo se escribe como

```
y=0;
for k=0:m
  y=p(k)*x+prodvec(y);
end
```

2.6. Métodos CGNR y CGNE

Si A es nosimétrica o no definida positiva, entonces podemos considerar resolver

$$(A^T A)\, x = (A^T b) \tag{2.108}$$

en el cual aparece la matriz $A^T A$ que es simétrica y definida positiva si A es no singular. Notar que en cada iteración de GC sobre este sistema estamos minimizando

$$\begin{aligned} \|x^* - x\|_{A^T A} &= (x^* - x)^T A^T A\, (x^* - x) && (2.109)\\ &= (b - Ax)^T (b - Ax) = \|r\|_2^2 && (2.110) \end{aligned}$$

de ahí el nombre de *"Conjugate Gradient on the Normal Equations to minimize the Residual"* (CGNR). Otra posibilidad es hacer el cambio de variable $x = A^T y$ y resolver

$$(AA^T) y = b \tag{2.111}$$

para y. Una vez encontrado y, x se obtiene con una operación matriz vector adicional $x = A^T y$. La norma que se minimiza es en este caso

$$\begin{aligned} \|y^* - y\|_{AA^T} &= (y^* - y)^T AA^T\, (y^* - y) && (2.112)\\ &= (A^T y^* - A^T y)^T\, (A^T y^* - A^T y) && (2.113)\\ &= (x^* - x)^T\, (x^* - x) = \|x^* - x\|_2^2 && (2.114) \end{aligned}$$

de ahí el nombre de *"Conjugate Gradient on the Normal Equations to minimize the Error"* (CGNE).
Observaciones

- En general ocurre que $\kappa(A^T A) \sim \kappa(A)^2$, lo cual augura muy bajas tasas de convergencia si el $\kappa(A)$ es grande.
- Se necesitan 2 productos matriz vector por iteración
- En la versión *matrix free* hay que programar no sólo una rutina que calcule Ax sino también una que calcule $A^T x$. Para problemas provenientes de discretizaciones por FEM o FDM de PDE's, esto puede resultar bastante complicado.

2.7. Guía Nro 2. Conjugate Gradients

1. Si f(x) es una forma cuadrática $f(x) = \frac{1}{2}x^T Ax - b^T x + c$, donde A es una matriz en $\mathbb{R}^{n\times n}$, x y b son vectores en $\mathbb{R}^n$ y c es un escalar constate. Calcular $f'(x)$. Mostrar que si A es *spd*, $f(x)$ es minimizada por la solución de $Ax = b$.

2. Para el Método de Steepest Descent demostrar que si el error e_i en la iteración i es un autovalor de la matriz A cuyo autovalor es λ_e, se convergerá a la solución exacta en la próxima iteración, i.e., $i + 1$.

3. Dar una interpretación geométrica del ejercicio anterior y decir cuanto tiene que valer α (la long del paso en la dirección de búsqueda) para obtener convergencia inmediata.

4. **GC como método directo**.

   Resolver la ecuación de Poisson

   $$\Delta\phi = -f, \quad \text{en} \quad \Omega = \{x, y \ / \ 0 \leq x, y \leq 1\} \tag{2.115}$$
   $$\phi = 0, \quad \text{en} \quad \partial\Omega \tag{2.116}$$

   con una grilla de diferencias finitas de $(N + 1) \times (N + 1)$ puntos. Usar $N = 4, 6, 8$ y 10 y ver en cuantas iteraciones converge a precisión de máquina. Calcular los autovalores de A y ver cuantos distintos hay. Inicializar con `x0=rand(n,1)` y ver en cuantas iteraciones converge. Porqué?. Puede usar las rutinas de Octave provistas por la cátedra.

5. **GC como método iterativo**.

   Idem que el ej. anterior pero ahora para $N = 100$. (No tratar de construir la matriz!! La matriz llena ocupa 800Mbytes y la banda 8Mbytes). Graficar la curva de convergencia y comparar el n experimental con el teórico (basado en una estimación del número de condición de la matriz). Comparar la convergencia con el método de Richardson (con $\omega = \omega_{\text{ópt}}$), en números de iteraciones y en tiempo de CPU. Puede usar las rutinas de Octave provistas por la cátedra.

6. Resolver el punto anterior en el cluster en forma secuencial usando las rutinas de PETSc provistas, con y sin precondicionamiento de Jacobi (*point Jacobi*). Sacar conclusiones acerca de los resultados obtenidos en relación a los resultados de los puntos anteriores.

7. Resolver los puntos anteriores en el cluster usando 4 procesadores, con y sin precondicionamiento de Jacobi (*point Jacobi*). Sacar conclusiones acerca de los resultados obtenidos en relación a los resultados de los puntos anteriores.

8. Verificar la escalabilidad del Método CG (con prec. point Jacobi) para el problema de Poisson usando una grilla de $100 \cdot \sqrt{n_{\text{proc}}} \times 100 \cdot \sqrt{n_{\text{proc}}}$. Es decir el comportamiento de la cantidad de iteraciones de CG para bajar el residuo un cierto orden de magnitud (por ejemplo 5 órdenes) en función de la cantidad de procesadores cuando el problema crece en la misma proporción al número de procesadores n_{proc}. Sacar conclusiones acerca de la escalabilidad del algoritmo.

Capítulo 3

El método GMRES

3.1. La propiedad de minimización para GMRES y consecuencias

GMRES (por *"Generalized Minimum RESidual"* fue popuesto en 1986 por Y. Saad y m. Schulz como un método iterativo por subespacios de Krylov para sistemas no-simétricos y no necesariamente definidos positivos. En contraposición con CGNR o CGNE *no* requiere el cálculo de productos de A^T con un vector, lo cual es una gran ventaja en muchos casos, pero es necesario guardar una base de $\mathcal{K}_k$ lo cual requiere un costo de almacenamiento adicional a medidad que la iteración progresa.

La iteración k-ésima ($k \geq 1$) se realiza de tal manera que

$$x_k = \underset{x \in x_0 + \mathcal{K}_k}{\operatorname{argmin}} \|b - A\,x\|_2 \tag{3.1}$$

Nótese que esta propiedad de minimización es muy similar con la de Gradientes Conjugados, con la diferencia que en GC se aplica al funcional (2.6). Como aquí estamos contemplando la posibilidad que A no sea simétrica o definida positiva, entonces no podemos tomar este funcional. Si consideramos que minimizar $\|b - A\,x\|_2$ es equivalente a minimizar $\|b - A\,x\|_2^2$ el cual contiene en la forma cuadrática $A^T\,A$, entonces vemos que GMRES es equivalente a CGNR con la salvedad de que con GMRES no debemos calcular productos A^T-vector pero en contraparte debemos mantener una base del espacio de Krylov.

Como $x \in x_0 + \mathcal{K}_k$ puede ponerse de la forma

$$x = x_0 + \sum_{j=0}^{k-1} \gamma_j\, A^j\, r_0 \tag{3.2}$$

entonces

$$b - A\,x \quad = b - A\,x_0 - \sum_{j=0}^{k-1} \gamma_j\, A^{j+1}\, r_0 \tag{3.3}$$

$$= r_0 - \sum_{j=1}^{k} \gamma_{j-1}\, A^j\, r_0 \tag{3.4}$$

$$= \bar{p}(A)\, r_0 \tag{3.5}$$

donde $\bar{p} \in \mathcal{P}_k$ es un polinomio residual.

Teorema 3.1.1. Sea A no-singular y x_k la k-ésima iteración de GMRES. Entonces para todo $\bar{p} \in \mathcal{P}_k$

$$\|r_k\|_2 = \min_{\bar{p} \in \mathcal{P}_k} \|\bar{p}(A)\, r_0\|_2 \leq \|\bar{p}(A)\, r_0\|_2 \tag{3.6}$$

Corolario 3.1.1. Sea A no-singular, entonces

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \|\bar{p}_k(A)\|_2 \tag{3.7}$$

Teorema 3.1.2. Sea A no-singular. Entonces GMRES encuentra la solución dentro de las N iteraciones

Demostración. Usar $\bar{p}(z) = p(z)/p(0)$, donde $p(z) = \det(A - zI)$ es el polinomio característico. □

Para GC hemos usado el teorema espectral para encontrar estimaciones de la tasa de convergencia. Esto no puede asumirse en general si A no es simétrica. Sin embargo podemos restringir el análisis al caso de que A sea diagonalizable aunque los autovalores y autovectores pueden ser ahora complejos.

Nota sobre la condición de diagonalizabilidad de matrices. Recordemos que

- A es diagonalizable si $A = V\Lambda V^{-1}$ con Λ diagonal. Cuando A es real y simétrica Λ es real y podemos elegir a V como real. Cuando A es no-simétrica tando Λ como V pueden ser complejos.
- En álgebra compleja muchos conceptos pueden extenderse fácilmente si reemplazamos la "tranpuesta" de A por la "transpuesta conjugada" de A que denotaremos como A^H.
- El producto escalar de dos vectores pertenecientes a C^N se define como $x^H y = \sum_{k=1}^{n} \overline{(x)_k}(y)_k$.
- V es *unitaria* si $V^H V = I$ (es la extensión del concepto de matriz ortogonal (ortonormal) a complejos).
- A es *normal* si es diagonalizable y si la matriz de cambio de base correspondiente V es unitaria.
- Puede verse que si A conmuta con su transpuesta conjugada (es decir $A^H A = A A^H$) entonces A es normal. Si $A^H = A$ entonces es Hermitiana.
- Es obvio que si A es hermitiana (simétrica en el caso real) entonces A es normal

Teorema 3.1.3. Sea $A == V\Lambda V^{-1}$ una matriz no singular diagonalizable y x_k la solución k=th de GMRes, entonces para todo $\bar{p}_k \in \mathcal{P}_k$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \kappa_2(V) \max_{z \in \sigma(A)} |\bar{p}_k(Z)| \tag{3.8}$$

Demostración. Basta con ver que

$$\begin{aligned} \|\bar{p}_k(A)\|_2 &\leq \|V\|_2 \left\|V^{-1}\right\|_2 \|\bar{p}_k(\Lambda)\|_2 && (3.9)\\ &= \kappa_2(V) \max_{z \in \sigma(A)} |\bar{p}_k(Z)|\square. && (3.10) \end{aligned}$$

No está claro como puede estimarse el $\kappa_2(V)$, si existe. Si A es normal, entonces V es unitaria, preserva la norma y entonces $\|V\|_2 = \left\|V^{-1}\right\|_2 = \kappa_2(V) = 1$

$$\begin{aligned} \|V\|_2 &= \max_{x \neq 0} \frac{\|V\,x\|_2}{\|x\|_2} && (3.11)\\ &= \max_{x \neq 0} \frac{\sqrt{(V\,x)^H\,(V\,x)}}{\sqrt{x^H\,x}} && (3.12)\\ &= \max_{x \neq 0} \frac{\sqrt{x^H\,(V^H V)\,x}}{\sqrt{x^H\,x}} && (3.13)\\ &= 1 && (3.14) \end{aligned}$$

y similarmente para V^{-1}. Por otra parte, si A se aproxima a una matriz no diagonalizable, entonces $\kappa_A() \to \infty$. Esto puede verse con un ejemplo simple. La matriz

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{3.15}$$

es un *bloque de Jordan* y es claramente no diagonalizable. Perturbando ligeramente uno de los elementos diagonales de la forma

$$A = \begin{bmatrix} 0 & 1 \\ 0 & \epsilon \end{bmatrix} \tag{3.16}$$

con ϵ muy pequeño la matriz pasa a ser diagonalizable, ya que tiene dos autovalores distintos ($\lambda = 0, \epsilon$). Sin embargo podemos ver que la matriz de cambio de base V correspondiente tiene un número de condición que crece como $2/\epsilon$:

```
octave> a=[0 1;0 1e-5]
a =
  0.00000  1.00000
  0.00000  0.00001
octave> [v,d]=eig(a)
v =
  1.00000  1.00000
  0.00000  0.00001
d =
   0.0000e+00   0.0000e+00
   0.0000e+00   1.0000e-05
octave> cond(v)
ans =  2.0000e+05
octave>
```

Ahora consideremos que ocurre si utilizamos una rutina numérica de diagonalización (como el `eig( )` de Octave) sobre una matriz que no es diagonalizable. Lo deseable sería que la rutina numérica nos mostrara un mensaje de eror diciendo que la matriz no es diagonalizable. Sin embargo, debido a los errores de redondeo, la matriz aparece ser como diagonalizable desde el punto de vista numérico y entonces la forma efectiva de verificar *"cuán diagonalizable es una matriz"* es chequear $\kappa_2(V)$. Cuanto más grande es este número *"menos diagonalizable"* es la matriz.

```
octave>a=[0 1;0 0]
a =

  0  1
  0  0

octave> [v,d]=eig(a)
v =

   1.00000  -1.00000
   0.00000   0.00000

d =
```

```
  0   0
  0   0

octave> cond(v)
ans =  1.9958e+292
```

Esto es similar a cuando nos preguntamos si una matriz es inversible o no. Si calculamos el determinante de la matriz, por errores de redondeo este número siempre resulta ser diferente de cero. Además, la misma matriz multiplicada por una factor $a < 1$ tiene un determinante que es un factor a^N veces menor. Para matrices grandes (digamos $N = 100$) un pequeño factor 0.1 puede representar un cambio en la magnitud del determinante por un factor 10^{-100}. Todo esto indica que el determinante no es un buen indicador de *"cuan singular es una matriz"* y un análisis más detallado muestra que el indicador correcto es el número de condición de la matriz: cuanto más alto es el número de condición *"más singular es la matriz"*.

Los siguientes Teoremas reproducen los resultados ya obtenidos para GC. Su demostración se basa nuevamente en la construcción de polinomios residuales apropiados que se anulan en los autovalores de la matriz.

Teorema 3.1.4. Si A tienen autovalores distintos entonces GMRES converge en k iteraciones.

Teorema 3.1.5. Si r_0 es una combinación lineal de k autovectores de A entonces GMRES converge dentro de las k iteraciones.

3.2. Criterio de detención:

Pensando a GMRES como un método iterativo las estimaciones de la tasa de convergencia son similares a las de GC. Como en GC el criterio de detención es

$$\|r_k\|_2 \leq \texttt{tol}\ \|b\|_2$$

Una estimación bastante cruda de la tasa de convergencia se puede obtener asumiendo que existe un ω tal que $\|I - \omega A\|_2 = \rho < 1$. Tomando el polinomio de $\bar{p}_k(z) = (1 - \omega z)^k$ podemos obtener la estimación de convergencia

$$\|r_k\|_2 \leq \rho^k \|r_0\|_2 \tag{3.17}$$

Esta estimación de convergencia tiene algunos puntos en común con las estimaciones de convergencia que hemos obtenido para el método de Richardson. Notemos que bajo estas mismas condiciones el método de Richardson predice la misma tasa de convergencia. Sin embargo la diferencia fundamental está en que con GMRES no es necesario conocer el valor de ω, de hecho, como (3.17) es válido para cualquier ω es válido para aquel que minimize $\|I - \omega A\|_2$, es decir que su convergencia es mejor que la de Richardson sobre-relajado con el mejor valor del parámetro de relajación que pudiéramos obtener, sin necesidad de conocer ningúna norma ni estimación de los autovalores de A. Por otra parte, GMRES tiene en su contra que debe guardar la base del espacio de Krylov. Pero si hacemos la estrategia del "restart" que veremos más adelante, según la cual se realizan un cierto número m (bajo) de iteraciones de GMRES y obtenido el x_m se vuelve a iterar GMRES de cero desde x_m, entonces este GMRES con restart tendría una tasa de convergencia mejor que la mejor que podríamos obtener con el mejor parámetro de relajación ω, a un costo similar por iteración y con un requerimiento de capacidad de almacenamiento no demasiado alto.

Como esta estimación no depende de una diagonalización de A podríamos esperar que nos de alguna estimación de la convergencia en el caso en que A no es diagonalizable. Desafortunadamente,

puede verificarse que para un bloque de Jordan de la forma

$$A = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & 0 & \dots \\ \vdots & 0 & \lambda & 1 & \dots \\ 0 & \ddots & \ddots & \ddots & 1 \\ 0 & 0 & \dots & 0 & \lambda \end{bmatrix} \tag{3.18}$$

vale que $\|I - \omega A\|_2 > 1$ para todo ω, es decir que no hay ω que haga convergente a Richardson y, a la vez, nos permita obtener una estimación de la tasa de convergencia para GMRES. Todo esto haría pensar que si la matriz no es diagonalizable (o *"casi no diagonalizable'*) entonces GMRES no convergerá. Pero si la forma de Jordan de una Matriz incluye un pequeño bloque de Jordan de dimension k_J y el resto es diagonalizable, entonces basta con tomar polinomios residuales de la forma

$$\bar{p}_k(z) = \left[\frac{(z - \lambda_J)}{\lambda_J}\right]^{k_J} q_{k-k_J}(z) \tag{3.19}$$

para $k > k_J$, donde λ_J es el autovalor correspondiente al bloque de Jordan y q es un polinomio apropiado para estimar una buena convergencia sobre el espectro de autovalores restantes. Por ejemplo, si el resto de los autovalores es real y positivo, entonces podríamos usar los polinomios de Tchebyshev usados para estimar la tasa de convergencia de GC.

3.3. Precondicionamiento

La forma de implementar el precondicionamiento en GMRES difiere con GC en cuanto a que para GMRES no es necesario que el sistema precondicionado

$$(MA)\, x = (Mb) \tag{3.20}$$

sea ni simétrico ni definido positivo. Entonces basta con encontrar un precondicionamiento M tal que $\|I - MA\|_2 < 1$ y se resuelve directamente el sistema precondicionado. Debido a esto, la rutina de GMRES no incluye nada en especial en cuanto al precondicionamiento, sino que directamente se pasa Mb como miembro derecho, y la rutina que calcula el producto matriz vector retorna $(MA)\, x$ en vez de $A\, x$.

Por otra parte se pueden usar *precondicionadores por derecha* y *por izquierda*. El precondicionamiento por izquierda es como fue expuesto en el párrafo anterior mientras que el precondicionamiento po derecha consiste en encontrar un M tal que $\|I - AM\|_2 < 1$ y entonces hacer el cambio de variable $x = My$, resolver con GMRES el sistema

$$(AM)\, y = b \tag{3.21}$$

y finalmente, una vez encontrado y obtener x de $x = My$.

En cuanto a las ideas para desarrollar precondicionadores son similares a las utilizadas con GC.

3.4. Implementación básica de GMRES

Recordemos que x_k se obtiene del problema de minimización (3.1). Sea V_k una matriz que contiene los vectores que expanden $\mathcal{K}_k$ es decir

$$V_k = \begin{bmatrix} r_0 & Ar_0 & \dots & A^{k-1}r_0 \end{bmatrix} \tag{3.22}$$

Entonces $(x - x_0)$ es una combinación lineal de las columnas de V_k, es decir

$$x - x_0 = V_k y, \quad y \in \mathbb{R}^k \tag{3.23}$$

Entonces y debe ser tal que minimize

$$\|b - A(x_0 + V_k y)\|_2 = \|r_0 - AV_k y\|_2 \tag{3.24}$$

Sea ahora

$$B_k = AV_k = \left[Ar_0 \; A^2 r_0 \; \ldots \; A^k r_0\right] \tag{3.25}$$

entonces el cuadrado de la norma se escribe como

$$\|r_0 - B_k y\|_2^2 = (r_0 - B_k y)^T (r_0 - B_k y) \tag{3.26}$$

$$= r_0^T r_0 - 2 r_0^T B_k y + y^T \left(B_k^T B_k\right) y \tag{3.27}$$

que alcanza su mínimo cuando

$$- B_k^T r_0 + (B_k^T B_k) y = 0 \tag{3.28}$$

lo cual ocurre cuando

$$y = \left(B_k^T B_k\right)^{-1} B_k^T r_0 \tag{3.29}$$

Esto puede implementarse en Octave con el comando

```
y=(B'*B)\ B'*r0
```

pero de una manera aún más simple

```
y=B\r0
```

hace lo mismo ya que para matrices rectangulares el operador \ se interpreta como resolver el sistema de mínimos cuadrados asociado.

En cuanto a la implementación práctica, notemos que el vector

$$q_k = B_k^T r_0 \tag{3.30}$$

$$= \begin{bmatrix} r_0 A\, r_0 \\ r_0 A^2\, r_0 \\ \vdots \\ r_0 A^k\, r_0 \end{bmatrix} = \begin{bmatrix} q_{k-1} \\ r_0 A^k\, r_0 \end{bmatrix} \tag{3.31}$$

de donde vemos que en cada iteración sólo necesitamos calcular el último elemento del vector, lo cual involucra $O(N)$ operaciones. Algo similar ocurre con el cálculo de la matriz $H_k = B_k^T B_k$ a invertir.

$$H_k = B_k^T\, B_k \tag{3.32}$$

$$= \begin{bmatrix} B_{k-1}^T \\ (A^k r_0)^T \end{bmatrix} \left[B_{k-1} \; A^k r_0\right] \tag{3.33}$$

$$= \begin{bmatrix} H_{k-1} & B_{k-1}^T A^k r_0 \\ (B_{k-1}^T A^k r_0)^T & r_0^T (A^k)^T A^k r_0 \end{bmatrix} \tag{3.34}$$

con lo cual sólo es necesario calcular la última columna y el elemento diagonal. La última fila es igual a la transpuesta de la última columna ya que H_k es simétrica y definida positiva por construcción. El cálculo de esta última columna requiere de $O(kN)$ operaciones. Finalmente la inversión del sistema cuya matriz es H_k requiere $O(k^3)$ operaciones. Mientras mantengamos $k \ll N$ (más estrictamente es $k^2 \ll N$) este costo es despreciable contra las k^N operaciones.

3.5. Implementación en una base ortogonal

En la práctica es muy común ir construyendo una base ortonormal de $\mathcal{K}_k$ mediante el *proceso de ortogonalización de Gram-Schmidt*. El algoritmo es el siguiente

1. $r_0 = b - Ax_0$, $v_1 = r_0 / \|r_0\|_2$
2. Para $i = 1, \ldots, k-1$
   $w = Av_i - \sum_{j=1}^{i}((Av_i)^T v_j)\, v_j$
   $v_{i+1} = w_i / \|w_i\|_2$

Si en algún i sucede que $w = 0$ entonces decimos que el algoritmo *falla* o *colapsa* (*"breakdown"*). Asumiendo que los $r_0, Ar_0, \ldots, A^{k-1}r_0$ son linealmente independientes (o sea que: $\dim \mathcal{K}_k = k$), entonces podemos que

- El algoritmo no falla.
- Los $\{v_i\}_{i=1}^{k}$ así generados forman una base ortonormal de $\mathcal{K}_k$.
- $v_k = \alpha_k\, A^{k-1}r_0 + w_k$ con $w_k \in \mathcal{K}_{k-1}$ y $\alpha_k \neq 0$.

Esto se puede demostrar por inducción en i. Para $i = 1$ es obvio, ya que v_1 es igual a r_0 normalizado. Para demostrarlo para $i+1$, observemos que w es ortogonal a todos los v_j para $j \leq i$

$$v_j^T\, A\, v_i - v_j^T \sum_{l=1}^{i}((Av_i)^T v_l)\, v_l) \qquad = v_j^T\, A\, v_i - \sum_{l=1}^{i}(Av_i)^T v_l)\, \delta_{jl} \tag{3.35}$$

$$= v_j^T\, A\, v_i - (Av_i)^T v_j) \tag{3.36}$$

$$= 0 \tag{3.37}$$

Ahora bien Av_i pertenece a $\mathcal{K}_{i+1}$ y los v_l para $l = 1, \ldots, i$ pertenecen a $\mathcal{K}_i$. Si $w \neq 0$ entonces podemos normalizar w para obtener v_{i+1} y $\{v_l\}_{l=1}^{i+1}$ es un conjunto de vectores ortonormales en $\mathcal{K}_{i+1}$. Como $\mathcal{K}_{i=1}$ es de dimensión a lo sumo $i+1$, v_{i+1} y $\{v_l\}_{l=1}^{i+1}$ es una base ortogonal. Sólo falta demostrar entonces que bajo las hipótesis asumida el algoritmo no falla.

Ahora bien, por inducción podemos asumir que

$$Av_i = \alpha_i\, A^i r_0 + Aw_i \tag{3.38}$$

Pero si el algoritmo falla, entonces quiere decir que Av_i es una combinación lineal de los $\{v_l\}_{l=1}^{i}$ y eso implicaría que $A^i r_0$ es una combinación lineal de los mismos e indicaría que los $\{A^{j-1}r_0\}_{j=1}^{i}$ son linealmente dependientes.

3.5.1. Colapso de GMRES (Breakdown)

Puede ocurrir que $w = 0$ para algún i, en cuyo caso (por lo visto en el párrafo anterior) indicaría que $\{A^{j-1}r_0\}_{j=1}^{i}$ son linealmetne dependientes. Podemos ver que esto ocurre si $x^* - x_0 \in \mathcal{K}_k$.

Lemma 3.4.1. Si $w_{i+1} = 0$ entonces $x^* = A^{-1}b \in \mathcal{K}_i$

Demostración. Si $w = 0$ para algún i entonces eso implica

$$Av_i = \sum_{i=1}^{j} \alpha_j\, v_j \in \mathcal{K}_i \tag{3.39}$$

Por construcción $Av_j \in \mathcal{K}_i$ para $j < i$, con lo que resulta que $A\mathcal{K}_i \subset \mathcal{K}_i$. Pero como

$$V_i = [v_1 \; v_2 \; \ldots \; v_i] \tag{3.40}$$

es una base de $\mathcal{K}_i$ entonces

$$AV_i = V_i H \tag{3.41}$$

para una cierta matriz H de $i \times i$. La columna j-ésima de H son los coeficientes de la expansión de Av_j en término de los $\{v_l\}_{l=1}^{j+1}$. H es no singular ya que A no lo es. Efectivamente si $z \neq= 0$, $z \in \mathbb{R}^i$ es tal que $Hz = 0$, entonces $V_i z$ es un vector no nulo y

$$A\,(V_i z) = V_i\, H\, z = 0 \tag{3.42}$$

Consideremos ahora el residuo en la i-ésima iteración

$$r_i = b - Ax_i = r_0 - A(x_i - x_0) = V_i y \tag{3.43}$$

con $y \in \mathbb{R}^i$ ya que $x_i - x_0 \in \mathcal{K}_i$. Además r_0 por construcción es proporcional a v_1 la primera columna de V_i lo cual puede escribirse como

$$r_0 = \beta V_i e_1 \tag{3.44}$$

con $e_1^T = [1 \; 0 \; \ldots \; 0]$. Como V_i es ortogonal, preserva la norma

$$\begin{aligned} \|r_i\|_2^2 &= \|V_i\,(\beta e_1 - Hy)\|_2^2 && (3.45)\\ &= (\beta e_1 - Hy)^T\, V_i^T\, V_i(\beta e_1 - Hy) && (3.46)\\ &= \|(\beta e_1 - Hy)\|_2^2 && (3.47) \end{aligned}$$

Pero basta con tomar $y = \beta H^{-1} e_1$ para el cual $\|r_i\|_2 = 0$ y GMRES ha convergido.

3.6. El algoritmo de Gram-Schmidt modificado

Uno de los principales problemas de GMRES es que por errores de redondeo se va perdiendo la ortogonalidad de los vectores v_i, por lo cual debe prestarse especial atención al proceso de ortogonalización. Una primera observación es que la siguiente versión modificada del proceso de ortogonalización de Gram-Schmidt resulta ser mucho más estable ante errores de redondeo

$v_{k+1} = Av_k$
for $j = 1, \ldots, k$
    $v_{k+1} = v_{k+1} - (v_{k+1}^T v_j) v_j$

3.7. Implementación eficiente

La matriz H que contiene los coeficientes que expanden Av_i en término de los v_l tiene una estructura particular que permite una implementación más eficiente del algoritmo. Consideremos la expresión

$$v_{i+1} = \|w_{i+1}\|_2^{-1}\; (Av_i - \sum_{j=1}^{i} \alpha_j\, v_j) \tag{3.48}$$

Esto quiere decir que Av_i es una combinación lineal de los $\{v_j\}_{j=1}^{i+1}$. Como la columna j-ésima de H_k son los coeficientes de la expansión de Av_j en término de los $\{v_l\}_{l=1}^{j+1}$

$$A\,V_k = V_{k+1}\,H_k \tag{3.49}$$

vemos los h_{lj} deben ser de la forma $h_{lj} = 0$ para $l > j + 1$. Es decir

$$H_k = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \dots \\ h_{21} & h_{22} & h_{23} & \dots \\ 0 & h_{32} & h_{33} & \dots \\ 0 & 0 & h_{43} & \dots \\ \ddots & \ddots & \ddots & \ddots \end{bmatrix} \tag{3.50}$$

Este tipo de matriz se llama *Hessenberg superior* (*upper Hessenberg*).

El residuo se puede escribir como

$$r_k = b - Ax_k = r_0 - A(x_k - x_0) \tag{3.51}$$
$$= V_{k+1}(\beta e_1 - H_k y^k) \tag{3.52}$$

y como V_k es ortogonal

$$\|r_k\|_2 = \left\|\beta e_1 - H_k y^k\right\|_2 \tag{3.53}$$

Para resolver este problema se recurre a una estrategia similar a la anterior. Es decir, en Octave `y=beta^{-1}*H\e1` si usamos la solución por mínimos cuadrados interna de Octave, o `y=beta^{-1}*(H'*H)\H*e1` si lo resolvemos explícitamente. Finalmente, existen algoritmos especiales que permiten factorizar la matriz con un algoritmo QR en forma más eficiente usando la estructura Hessenberg superior de H_k.

Independientemente de la implementación de la resolución del problema de cuadrados mínimos, el algoritmo de GMRES es

Algoritmo `gmresb`$(x, b, A, \epsilon, k_{\max}, \rho)$

1. $r = b - Ax$, $v_1 = r/\|r\|_2$, $\rho = \|r\|_2$, $\beta = \rho$, $k = 0$
2. While $\rho > \epsilon \|b\|_2$ y $k < k_{\max}$
   - (a) $k = k + 1$
   - (b) $v_{k+1} = Av_k$. Para $j = 1, \dots, k$
     - (i) $h_{jk} = v_{k+1}^T v_j$
     - (ii) $v_{k+1} = v_{k+1} - h_{jk}\, v_j$
   - (c) $h_{k+1,k} = \|v_{k+1}\|_2$
   - (d) $v_{k+1} = v_{k+1}/\|v_{k+1}\|_2$
   - (e) $e_1 = [1\ 0\ 0\ \dots\ 0]^T$
     $y_k = \operatorname{argmin}_{y \in \mathbb{R}^k} \left\|\beta e_1 - H_k y^k\right\|_2$
   - (f) $\rho = \left\|\beta e_1 - H_k\, y^k\right\|_2$
3. $x_k = x_0 + V_k\, y^k$

3.8. Estrategias de reortogonalización

Se puede perder ortogonalidad por errores de redondeo. Una solución es hacer un segundo paso de ortogonalización sea entodas las iteraciones, sea cuando algún indicador de pérdida de ortogonalidad se activa.

3.9. Restart

Como debemos almacenar una base para $\mathcal{K}_k$ esto requiere kN reales lo cual va creciendo con k y eventualmente excede la memoria de la máquina. Entonces la idea es iterar GMRES m iteraciones y

empezar de nuevo a partir de la última iteración, es decir aplicar GMRES inicializando con $x_0 \leftarrow x_m$. El siguiente algoritmo `gmresm` refleja esto,

Algoritmo `gmresm`$(x, b, A, \epsilon, k_{\max}, m, \rho)$
1. `gmres`$(x, b, A, \epsilon, m, \rho)$
2. $k = m$
3. While $\rho > \epsilon \|b\|_2$ y $k < k_{\max}$
   (a) `gmres`$(x, b, A, \epsilon, m, \rho)$
   (b) $k = k + m$

3.10. Otros métodos para matrices no-simétricas

GMRES, CGNE y CGNR comparten las siguientes (buenas) características

- Son fáciles de implementar
- Se analizan con residuos polinomiales

Por otra parte los CGN* tienen la desventaja de que necesitan un producto $A^T x$ y su tasa de convergencia está regida por $\kappa(A^T A) \sim \kappa(A)^2$, mientras que GMRES sólo necesita calcular $A x$ y la convergencia está basada en $\kappa(A)$, pero es necesario guardar una base de $\mathcal{K}_k$, lo cual representa una capacidad de almacenamiento que va creciendo con las iteraciones. Como esto es virtualmente imposible para grandes sistemas desde el punto de vista práctico se utiliza el GMRESm, lo cual puede involucrar un serio deterioro de la convergencia.

El método ideal debería ser como CG:

- Sólo necesitar calcular $A x$ (no $A^T x$)
- Estar basado en una propiedad de minimización o conjugación
- Requerir baja capacidad de almacenamiento. (Sobre todo que no crezca con las iteraciones).
- Converger en N iteraciones.

En esta sección describiremos algunos métodos que tratan de aproximarse a estos requerimientos con menor o mayor éxito.

Bi-CG: (por *Biconjugate Gradient*) No se basa en una propiedad de minimización sino de ortogonalidad

$$r_k^T w = 0, \quad \text{para todo } w \in \overline{\mathcal{K}_k} \tag{3.54}$$

donde $\overline{\mathcal{K}_k}$

$$\overline{\mathcal{K}_k} = \text{span}\{\hat{r}_0, A^T \hat{r}_0, \ldots, (A^T)^{k-1} \hat{r}_0\} \tag{3.55}$$

es el *espacio de Krylov conjugado*. $\hat{r}_0$ es un vector que debe ser provisto por el usuario, pero lo más usual es tomar $\hat{r}_0 = r_0$. El algoritmo genera secuencias de residuos y direcciones de búsqueda $\{r_k, p_k\}$ y sus correspondientes conjugados $\{\hat{r}_k, \hat{p}_k\}$ tales que hay biortogonalidad entre los residuos

$$\hat{r}_k^T r_l = 0, \quad k \neq l \tag{3.56}$$

y de bi-conjugación entre las direcciones de búsqueda

$$\hat{p}_k^T A p_l = 0, \quad k \neq l \tag{3.57}$$

Si A es simétrica y definida positiva y $\hat{r}_0 = r_0$, entonces Bi-CG es equivalente a GC (pero computa todo el doble, es decir que tiene el doble de costo por iteración). Bi-CG necesita un cálculo $A^T x$, pero la ventaja con respecto a los CGN* es que su tasa de convergencia está basada en $\kappa(A)$ no en $\kappa(A^T A)$. Es muy útil si A es aproximadamente spd, es decir si los autovalores de A están cerca de la recta real.

Podemos comparar Bi-CG con GMRES en cuanto a la tasa de convergencia si consideramos que resulta ser que

$$r_k = \bar{p}_k(A)\, r_0 \tag{3.58}$$

con $\bar{p}_k$ un polinomio residual, y entonces por la propiedad de minimización de GMRES

$$\left\|(r_k)^{\mathrm{GMRES}}\right\|_2 \leq \left\|(r_k)^{\mathrm{Bi-CG}}\right\|_2 \tag{3.59}$$

pero debe recordarse que Bi-CG comparado con GMRES no necesita un espacio de almacenamiento creciente. Además una iteración de Bi-CG requiere dos productos matriz vector, pero el costo de GMRES también crece con las iteraciones.

CGS (por *Conjugate Gradient Squared*). Este método trata de ser una extensión de Bi-CG pero tal que no necesita calcular $A^T x$. Puede verse que en la iteración k

$$r_k = \bar{p}_k(A)\, r_0, \quad \hat{r}_k = \bar{p}_k(A^T)\, \hat{r}_0 \tag{3.60}$$

entonces el factor $r_k^T\, \hat{r}_k$ (que juega el papel de ρ_k en GC, (ver el algoritmo `cg(...)` en pág. 33), puede reescribirse de la forma

$$\begin{aligned} r_k^T\, \hat{r}_k &= (\bar{p}_k(A)\, r_0)^T\, (\bar{p}_k(A^T)\, \hat{r}_0) && (3.61)\\ &= ([\bar{p}_k(A)]^2\, r_0)^T\, \hat{r}_0 && (3.62) \end{aligned}$$

en el cual no es necesario calcular $A^T\, x$. Con manipulaciones similares se puede llegar a transformar todos las expresiones donde aparece A^T, pero el algoritmo resulta modificado, es decir las iteraciones de Bi-CG *no son las mismas* que para CGS.

Bi-CGstab (por Biconjugate Gradient stabilized). Trata de mejorar CGS reemplazando

$$r_k = q_k(A)\, \bar{p}_k(A)\, r_0 \tag{3.63}$$

donde

$$q_k(z) = \prod_{i=1}^{k} (1 - \omega_i z) \tag{3.64}$$

donde ω_i se selecciona para minimizar

$$\|r_i\|_2 = \|q_i(A)\, \bar{p}_i(A)\, r_0\|_2 \tag{3.65}$$

como función de ω_i, esto es usualmente conocido como *line-searching*. Sea

$$\begin{aligned} r_i &= (1 - \omega_i A) \left[\prod_{i=1}^{k} (1 - \omega_i A)\, \bar{p}_i(A)\, r_0 \right] && (3.66)\\ &= (1 - \omega_i A)\, w && (3.67)\\ &= w - \omega_i A\, w && (3.68)\\ & && (3.69) \end{aligned}$$

de manera que

$$\begin{aligned} \|r_i\|_2^2 &= (w - \omega_i A\, w)^T\, (w - \omega_i A\, w) && (3.70)\\ &= \|w\|_2^2 - 2\omega_i\, w^T\, Aw + \omega_i^2 (Aw)^T\, Aw && (3.71) \end{aligned}$$

y el valor que minimiza es

$$\omega_i = \frac{w^T(Aw)}{\|Aw\|_2^2} \tag{3.72}$$

Bi-CGstab entonces no necesita del cálculo de $A^T x$ como CGS pero tiende a tener una tasa de convergencia mejor. El costo computacional involucrado por iteración es

- Almacenamiento para 7 vectores
- 4 productos escalares
- 2 productos matriz-vector (Ax)

3.11. Guía Nro 3. GMRES

1. Usando el algoritmos de Arnoldi, encontrar una matriz ortogonal $\mathbf{Q}$ tal que $\mathbf{Q^T A Q} = \mathbf{H}$ es Hessemberg superior y $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$

2. Demostrar los Teoremas T3.1.4 y T3.1.5 del apunte.

3. Consideremos la ec. de *advección-difusión* lineal escalar

$$k\phi'' - a\phi' = 0 \tag{3.73}$$
$$\phi(0) = 0 \tag{3.74}$$
$$\phi(1) = 1 \tag{3.75}$$

donde

- ϕ = temperatura del fluido
- k = conductividad térmica del fluido
- a = velocidad del fluido (puede ser positiva o negativa)

La solución exacta es

$$\phi(x) = \frac{\mathrm{e}^{2\mathrm{Pe}x} - 1}{\mathrm{e}^{2\mathrm{Pe}} - 1} \tag{3.76}$$

donde

$$\mathrm{Pe} = \frac{aL}{2k} \tag{3.77}$$

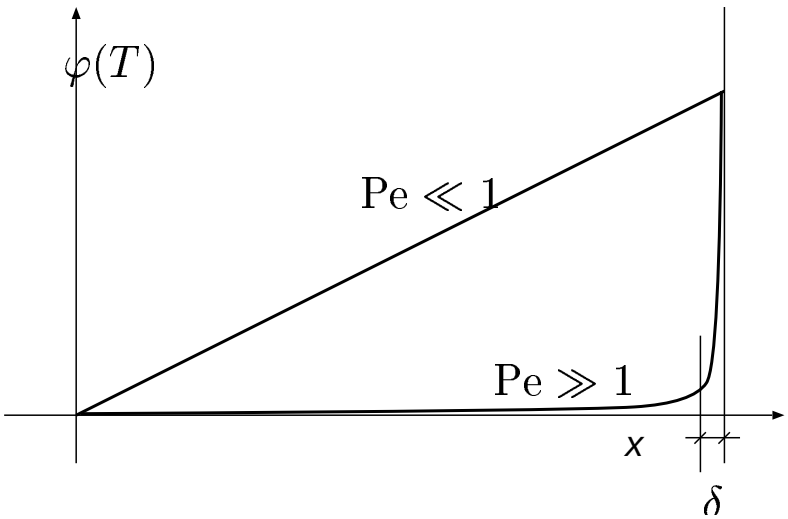

Figura 3.1: Solución para el problema de advección difusión, en los límites dominado por difusión y por advección

siendo $L = 1$. En la figura 3.1 vemos Para $a \to 0$ el problema se acerca a la ec. de Laplace y la solución tiende a ser una recta que une los valores de contorno, mientras que para Pe grandes y positivos el fluido va en la dirección $+x$ y arrastra el valor de la condición *aguas arriba* una

cierta longitud hasta que a una distancia pequeña δ sube rápidamente para tomar el valor dado por la condición en $x = 1$.

La discretización numérica pasa por reemplazar las derivadas por diferencias numéricas

$$\phi''_j \approx (\phi_{j+1} - 2\phi_j + \phi_{j-1})/h^2 \tag{3.78}$$

y

$$\phi'_j \approx= (\phi_{j+1} - \phi_{j-1})/(2h) \tag{3.79}$$

Lamentablemente, esta discretización falla cuando el Pe $\gg 1$, en el sentido de que produce fuertes oscilaciones numéricas. La solución usual es agregar una cierta cantidad de *difusión numérica*, es decir que se modifica la ecuación original a

$$(k + k_{\text{num}})\,\phi'' - a\phi' = 0 \tag{3.80}$$

donde

$$k_{\text{num}} = |a|h/2 = \text{Pe}_h k, \qquad \text{siendo} \tag{3.81}$$

$$\text{Pe}_h = \frac{ah}{2k} \tag{3.82}$$

Realizar las siguientes tareas:

a) Calcular la matriz para $\text{Pe}_h = 0.01$, 1 y 100

b) Calcular los autovalores. Ver la distribución en el plano complejo.

c) Verificar que la matriz no es simétrica. Ver que la parte correspondiente a ϕ'' es simétrica mientras que la que corresponde a ϕ' es antisimétrica.

d) Ver a que tiende la matriz para $k \to 0$. Es diagonalizable?

e) Resolver con los métodos de GMRes y CGNE modificando las rutinas utilizadas para la GTP de CG (se provee un script de octave con la implementación de GMRes). Realizar los cambios necesarios en `gmres.m` para utilizar precondicionamiento point Jacobi. Tener en cuenta que en PETSc se reportan los residuos del problema precondicionado (si es que se usó).

f) Al igual que para el algoritmo CG, estudiar la escalabilidad de GMRes usando una malla de $2000 \cdot n_{\text{proc}}$.

g) Pensar y explicar como se podría hacer para calcular $A^T x$ sin construir A.

Capítulo 4

Descomposición de dominios.

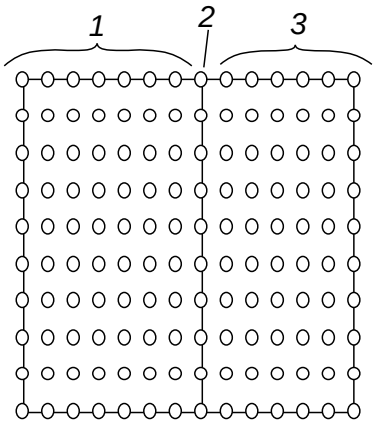

Figura 4.1: Descomposición de dominios

Consideremos la solución de una ecuación en derivadas parciales en un dominio como muestra la figura 4.1. Descomponemos el problema en dos dominios de manera que podemos separar las incógnitas en x en tres grupos, aquellos que están estrictamente contenidos en el dominio de la izquierda x_1, los estricatamente contenidos en el dominio de la derecha x_3 y los que están sobre la interfase x_2. La ecuación se descompone en bloques de la siguiente forma

$$\begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{4.1}$$

La descomposición en bloques refleja el hecho de que como los grados de libertad contenidos en x_1 y x_3 no están compartidos por ningún elemento, los bloques correspondientes A_{13} y A_{31} son nulos. Podemos eliminar x_1 y x_3 de la primera y última línea de ecuaciones y obtener una ecuación para los grados de libertad de interfase x_2

$$(-A_{23}A_{33}^{-1}A_{32} + A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 = b_2' \tag{4.2}$$

$$S\,x_2 = b_2' \tag{4.3}$$

Ahora consideremos resolver este sistema por GC. En cada iteración debemos calcular el producto de la matriz entre paréntesis por un vector x_2. Para los términos primero y tercero de la matriz es

necesario factorizar y resolver un sistema lineal con las matrices A_{33} y A_{11}. Esto corresponde a resolver problemas independientes sobre cada uno de los dominios con condiciones Dirichlet sobre la interfase, por lo tanto estos problemas pueden ser resueltos en procesadores independientes lo cual implica un alto grado de paralelización del problema. Esto se puede extender a más procesadores, y en el límite podemos lograr que en cada procesador se resuelva un sistema lineal lo suficientemente pequeño como para ser resuelto en forma directa. La eficiencia del método puede ser mejorada aún más encontrando un buen precondicionamiento.

Si separamos la matriz S que aparece en el sistema (4.3) en la contribución por el dominio de la izquierda S_L y de la derecha S_R

$$S = S_R + S_L \tag{4.4}$$
$$S_L = (1/2)A_{22} - A_{21}A_{11}^{-1}A_{12} \tag{4.5}$$
$$S_R = -A_{23}A_{33}^{-1}A_{32} + (1/2)A_{22} \tag{4.6}$$

Entonces podemos poner como precondicionamiento

$$M = (1/4)\,(S_L^{-1} + S_R^{-1}) \tag{4.7}$$

Es M un buen precondicionamiento? O mejor dicho: Porqué habría M de parecerse a S^{-1}? Bien, si el operador es simétrico (el laplaciano lo es) y los dominios son iguales y la malla es simétrica, entonces puede verse que $S_L = S_R$ y entonces M y S^{-1} coinciden

$$M = S^{-1} = (1/2)S_L^{-1} \tag{4.8}$$

Por otra parte resolver $x = My$ es equivalente a resolver

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & (1/2)\,A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_{2L} \end{bmatrix} = \begin{bmatrix} 0 \\ (1/2)\,y \end{bmatrix} \tag{4.9}$$

$$\begin{bmatrix} (1/2)\,A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_{2R} \\ x_3 \end{bmatrix} = \begin{bmatrix} (1/2)\,y \\ 0 \end{bmatrix} \tag{4.10}$$

y después

$$x = (1/2)(x_{2L} + x_{2R}) \tag{4.11}$$

y el punto es que ambos sistemas (4.9) y (4.10) equivale a resolver problemas *independientes* con condiciones tipo Neumann sobre la interfase. De nuevo, el hecho de que ambos problemas sobre cada dominio sean independientes favorece la paralelización del algoritmo.

4.1. Condicionamiento del problema de interfase. Análisis de Fourier.

Obviamente la aplicabilidad de la descomposición de dominios descripta depende del número de iteraciones necesarias para resolver el problema de interfase y por lo tanto del número de condición de la matriz complemento de Schur S o de su precondicionada MS. Para hacer una estimación de estos números de condición nos basaremos en un análisis de Fourier del problema del continuo para la ecuación de Laplace. Las ecuaciones de gobierno son

$$\Delta\phi = -f \text{ en } \Omega \tag{4.12}$$
$$\phi = \bar{\phi} \text{ en } \Gamma_1 \tag{4.13}$$
$$\tag{4.14}$$

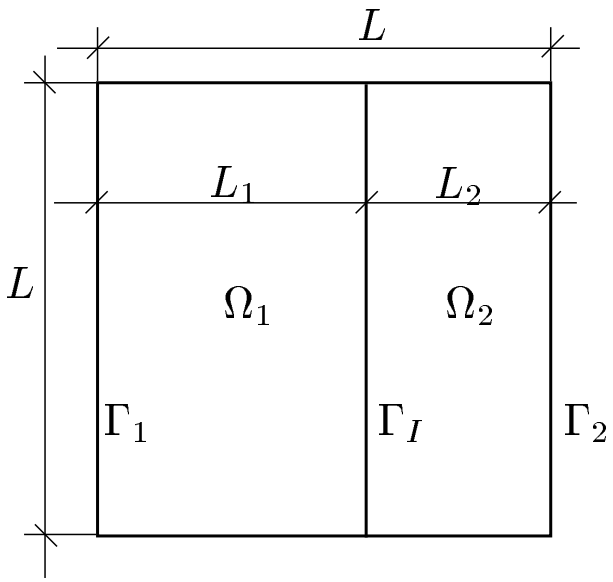

Figura 4.2: Descomposición de dominios. Problema del continuo.

Consideremos la solución en dos dominios Ω_1, Ω_2. La restricción de ϕ a cada uno de los dominios debe satisfacer

$$\Delta\phi_1 = -f \text{ en } \Omega_1 \tag{4.15}$$
$$\phi_1 = \bar{\phi}_1 \text{ en } \Gamma_1 \tag{4.16}$$
$$\tag{4.17}$$

y

$$\Delta\phi_2 = -f \text{ en } \Omega_2 \tag{4.18}$$
$$\phi_2 = \bar{\phi}_2 \text{ en } \Gamma_2 \tag{4.19}$$
$$\tag{4.20}$$

y la continuidad de ϕ y su derivada normal a través de Γ_I

$$(\phi_1)_{\Gamma_I} = (\phi_2)_{\Gamma_I} \tag{4.21}$$
$$\left(\frac{\partial\phi_1}{\partial x}\right)_{\Gamma_I} = \left(\frac{\partial\phi_2}{\partial x}\right)_{\Gamma_I} \tag{4.22}$$

Ahora consideremos una descomposición $\phi = \psi + \tilde{\phi}$, de manera que $\psi = 0$ en Γ_I, es decir

$$\Delta\psi_1 = -f \text{ en } \Omega_1 \tag{4.23}$$
$$\psi_1 = \bar{\psi}_1 \text{ en } \Gamma_1 \tag{4.24}$$
$$\psi_1 = 0 \text{ en } \Gamma_I \tag{4.25}$$

y

$$\Delta\psi_2 = -f \text{ en } \Omega_2 \tag{4.26}$$
$$\psi_2 = \bar{\phi}_2 \text{ en } \Gamma_2 \tag{4.27}$$
$$\psi_2 = 0 \text{ en } \Gamma_I \tag{4.28}$$

y por lo tanto falta hallar $\tilde{\phi}$ definido por

$$\begin{aligned} \Delta\tilde{\phi}_1 &= 0 \text{ en } \Omega_1 \\ \tilde{\phi}_1 &= 0 \text{ en } \Gamma_1 \\ \tilde{\phi}_1 &= u \text{ en } \Gamma_I \end{aligned} \tag{4.29}$$

y

$$\begin{aligned} \Delta\tilde{\phi}_2 &= 0 \text{ en } \Omega_2 \\ \tilde{\phi}_2 &= 0 \text{ en } \Gamma_2 \\ \tilde{\phi}_2 &= u \text{ en } \Gamma_I \end{aligned} \tag{4.30}$$

donde u es una incógnita del problema y debe ser tal que

$$\left(\frac{\partial\tilde{\phi}_1}{\partial x}\right)_{\Gamma_I} - \left(\frac{\partial\tilde{\phi}_2}{\partial x}\right)_{\Gamma_I} = \tilde{b} \tag{4.31}$$

donde

$$\tilde{b} = -\left\{\left(\frac{\partial\psi_1}{\partial x}\right)_{\Gamma_I} - \left(\frac{\partial\psi_2}{\partial x}\right)_{\Gamma_I}\right\} \tag{4.32}$$

Definimos ahora el operador de *Stekhlov-Poincaré* $\mathcal{S}_1$ del dominio Ω_1 como

$$\mathcal{S}_1\{u\} = \frac{\partial\phi_1}{\partial n} = \frac{\partial\phi_1}{\partial x} \tag{4.33}$$

donde $\frac{\partial\phi_1}{\partial n}$ denota la derivada normal tomando la normal exterior al dominio en cuestión, que para el dominio Ω_1 coincide con la dirección de $+x$, y ϕ_1 es la solución de (4.29), para el correspondiente u. Análogamente se puede definir $\mathcal{S}_2$ por

$$\mathcal{S}_2\{u\} = \frac{\partial\phi_2}{\partial n} = -\frac{\partial\phi_2}{\partial x} \tag{4.34}$$

donde ϕ_2 está definido en función de u por (4.30) y el signo $-$ viene de que la normal exterior a Ω_2 sobre Γ_I va ahora según la dirección $-x$. Entonces

$$\mathcal{S}\{u\} = g \tag{4.35}$$

donde $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$.

Ec. (4.35) es la versión del continuo de (4.3). Calcularemos el número de condición de S a partir del cálculo de autovalores de $\mathcal{S}$.

Cada uno de los operadores de Stekhlov-Poincaré actúa sobre el espacio de funciones definidas sobre Γ_I. Podemos mostrar que las funciones de la forma

$$u_m = \sin(k_m y), \quad k_m = m\pi/L, \quad m = 1, 2, \ldots, \infty \tag{4.36}$$

son autofunciones de estos operadores. La solución $\tilde{\phi}_1$ que es solución de (4.29) correspondiente a $u = u_m$ es de la forma

$$\tilde{\phi}_1 = \hat{\phi}(x)\, \sin(k_m y), \tag{4.37}$$

Reemplazando en las ecuaciones que definen (4.29), la primera de las ecuaciones resulta ser

$$\hat{\phi}'' - k_m^2\hat{\phi} = 0 \tag{4.38}$$

de donde

$$\hat{\phi}(x) = a\, e^{k_m y} + b\, e^{-k_m y} \tag{4.39}$$

y de las condiciones de contorno en $x = 0$, L_1 resulta ser

$$\hat{\phi}(x) = \frac{\sinh k_m y}{\sinh k_m L_1} \tag{4.40}$$

la derivada normal en Γ_I es entonces

$$\frac{\partial \tilde{\phi}_1}{\partial n} = k_m \frac{\cosh k_m L_1}{\sinh k_m L_1} = \frac{k_m}{\tanh k_m L_1} \tag{4.41}$$

de manera que

$$\mathcal{S}_1\{u_m\} = \left(\frac{k_m}{\tanh k_m L_1}\right) u_m \tag{4.42}$$

Vemos entonces, que u_m es una autofunción de $\mathcal{S}_1$ con autovalor

$$\lambda_m^1 = \frac{k_m}{\tanh k_m L_1} \tag{4.43}$$

Análogamente, el operador $\mathcal{S}_2$ tiene autovalores

$$\lambda_m^2 = \frac{k_m}{\tanh k_m L_2} \tag{4.44}$$

El operador $\mathcal{S}$ tiene autovalores

$$\lambda_m = k_m \left(\frac{1}{\tanh k_m L_1} + \frac{1}{\tanh k_m L_2}\right) \tag{4.45}$$

Ahora bien, dijimos que estimaríamos el número de condición de la matriz S a partir de los autovalores de $\mathcal{S}$. En el continuo $\mathcal{S}$ tiene infinitos autovalores y los λ_m van a infinito para $m \to \infty$, de manera que el número de condición de $\mathcal{S}$ va a infinito. Obtendremos una estimación para el número de condición de S asumiendo que los autovalores de S se aproximan a los primeros N_I autovalores de $\mathcal{S}$, donde N_I es la dimensión de S. El autovalor más bajo es

$$\lambda_{\min} = \lambda_1 = \frac{\pi}{L}\left(\frac{1}{\tanh(L_1/L)} + \frac{1}{\tanh(L_2/L)}\right) \tag{4.46}$$

si $L_1 = L_2 = L/2$, entonces

$$\lambda_{\min} = \frac{\pi}{L}\frac{2}{\tanh(1/2)} = \frac{13.6}{L} \tag{4.47}$$

El autovalor máximo se obtiene para $m = N_I$ y asumiendo que $N_I \gg 1$ y $L_1/L, L_2/L$ no son demasiado pequeños, entonces $k_m L_1 = N_I \pi L_1/L \gg 1$ y $\tanh k_m L_1 \approx 1$ y similarmente $\tanh k_m L_2 \approx 1$ y por lo tanto

$$\lambda_{\max} = 2k_m = 2N_I\pi/L \tag{4.48}$$

y el número de condición es

$$\kappa(S) \approx \tanh(1/2)\, N_I = 0.46\, N_I \tag{4.49}$$

Notar que $\kappa(S)$ va como $1/h$ al refinar, mientras que recordemos que el número de condición de A (la matriz del sistema) va como $1/h^2$ (Guía de ejercicios Nro. 1).

Otro punto interesante a notar es que, para $m \to \infty$ los autovalores tienden a hacerse independientes de las longitudes de los dominios en la dirección normal a la interfase. Por ejemplo, para los autovalores λ_m^1 de $\mathcal{S}_1$, L_1 aparece en el argumento de la tangente hiperbólica y esta tiende a 1 para $m \to \infty$, independientemente del valor de L_1. Incluso puede verse que para $m \to \infty$ los autovalores se hacen independientes del tipo de condición de contorno sobre el borde opuesto (es decir sobre $x = 0$). Esta observación se basa en el hecho de que el operador de Laplace es muy *local*. Si u es de la forma $\sin m\pi y/L$ sobre Γ_I, entonces la longitud de onda es $\beta = 2L/m$ la cual se va achicando a medida que $m \to \infty$. Las perturbaciones inducidas decaen como $e^{-n/\beta}$ donde n es una coordenada en la dirección normal a Γ_I y si β es muy pequeño la perturbación no llega al contorno opuesto.

Ahora consideremos los autovalores del problema precondicionado. Como todos las u_m de la forma (4.36) son autofunciones de los operadores $\mathcal{S}_1$, $\mathcal{S}_2$ y $\mathcal{S}$, entonces los autovalores de MS son aproximadamente los primeros N_I autovalores de $(1/4)(\mathcal{S}_1^{-1} + \mathcal{S}_2^{-1})(\mathcal{S}_1 + \mathcal{S}_2)$, es decir

$$\lambda_m^{\text{prec}} = (1/4)\ [(\lambda_m^1)^{-1} + (\lambda_m^2)^{-1}]\,(\lambda_m^1 + \lambda_m^2) \tag{4.50}$$

Como mencionamos previamente (ver pág. 58), si el problema es simétrico (en el caso del continuo basta con $L_1 = L_2$, en el caso del discreto además la malla también tiene que ser simétrica alrededor de $x = 1/2$), entonces $M = S^{-1}$. En el caso del continuo ocurre lo mismo ya que si los dos dominios son iguales, entonces

$$\lambda_m^1 = \lambda_m^2 \tag{4.51}$$

y por lo tanto $\lambda_m^{\text{prec}} = 1$ para todo m. Si $L_1 \neq L_2$, entonces puede verse que para m grandes $\lambda_1 \approx \lambda_2$ ya que ambos se hacen independientes de $L_{1,2}$ y de la condición de contorno en el contorno opuesto y por lo tanto

$$\lambda_m^{\text{prec}} \to 1 \text{ para } m \to \infty \tag{4.52}$$

Pero entonces esto quiere decir que $\kappa(MS)$ se hace independiente de N_I para m suficientemente grandes. Este resultado es muy importante desde el punto de vista práctico, *el número de condición del problema precondicionado no se deteriora bajo refinamiento para el precondicionamiento Dirichlet-to-Neumann.*

Parte II

Métodos iterativos para la resolución de ecuaciones no-lineales

Capítulo 5

Conceptos básicos e iteración de punto fijo

Queremos resolver un sistema no-lineal de la forma

$$F(x) = 0, \quad F : \mathbb{R}^N \to \mathbb{R}^N \tag{5.1}$$

Denotaremos con f_i la i-ésima componente de F. Si bien muchos de los métodos son aplicables a sistemas generales de ecuaciones no-lineales, en general tendremos en mente sistemas que provienen de la discretización por FDM o FEM de PDE's. Por ejemplo:

Ec. de Burgers: Consideremos la ecuación de *advección difusión* que ya viéramos como un sistema que genera ecuaciones no-simétricas (Guía 3) y representa el transporte unidimensional de una cantidad escalar u por un fluido con velocidad a y difusividad k. La ecuación de advección difusión es de la forma

$$a\frac{\partial u}{\partial x} - k\frac{\partial^2 u}{\partial x^2} = 0 \tag{5.2}$$

con condiciones de contorno Dirichlet $u(0) = 0$, $u(1) = 1$. La ecuación de Burgers representa una complejidad adicional en la cual la velocidad de propagación a depende de la misma cantidad u, es decir

$$a(u)\frac{\partial u}{\partial x} - k\frac{\partial^2 u}{\partial x^2} = 0 \tag{5.3}$$

Si $a(u)$ no es constante, entonces la ecuación pasa a ser no lineal y dependiendo de la forma de $a(u)$ la ecuación puede desarrollar *"ondas de choque"* (*"shock waves"* por lo cual esta ecuación es un modelo muy simple para problemas mucho más complicados de mecánica de fluidos como las ecuaciones de fluidos compresible o la de propagación de ondas de gravedad en canales (Ecs. de Saint-Venant). El caso más simple que desarrolla ondas de choque es tomar $a(u) = u$ y puede ser puesto de la forma

$$\frac{1}{2}\frac{\partial u^2}{\partial x} - k\frac{\partial^2 u}{\partial x^2} = 0 \tag{5.4}$$

Notemos que si $k = 0$ entonces las soluciones son de la forma $u^2 =$cte, o sea $u = \pm u_0$. Los puntos de discontinuidad donde u pasa de u_0 a $-u_0$ son las ondas de choque y pueden ser *"de compresión"* si las características de un lado y del otro de la discontinuidad fluyen hacia la misma o de *"descompresión"* en caso contrario. En el caso de la ec. de Burgers discutida aquí, la velocidad de las características es $a(u)$ de manera que es positiva si $u > 0$ y viceversa, por lo tanto las ondas de choque donde u pasa de $+u_0$ a $-u_0$ en el sentido de x positivo son de

compresión y viceversa. Un resultado interesante es que al agregar una pequeña difusividad o un término temporal las ondas de choque de compresión tienden a permanecer estable, mientras que las de decompresión se desarman en *"abanicos de expansión"*. En el caso de la propagación de ondas de gravedad en canales, las discontinuidades se llaman *"resaltos hidráulicos"*.

Ec. de advección difusión con cambio de fase: (También llamado "Problema de Stefan") Consideremos ahora la ecuación de advección-difusión no-lineal para el balance de energía

$$a\rho C_p \frac{\partial T}{\partial x} - \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) = 0 \tag{5.5}$$

donde a es la velocidad, ρ la densidad, C_p el calor específico y k la conductividad. Asumimos condiciones Dirichlet de la forma $T(0) = T_0$, $T(L) = T_L$. C_p y k son propiedades del material y, en general, pueden depender de la temperatura, en cuyo caso las ecuaciones resultantes son nolineales. Podemos definir la *entalpía* como

$$h(T) = \int_{T_0}^{T} C_p(T')\ \mathrm{d}T' \tag{5.6}$$

$h(T)$ representa el calor necesario para llevar al material desde una temperatura de referencia T_0 a otra T. La ecuación (5.5) queda entonces como

$$a\rho \frac{\partial}{\partial x} h(T(x)) - \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) = 0 \tag{5.7}$$

En el caso de haber un cambio de fase a una temperatura T_m dada, se necesita una cantidad finita de calor L llamada *"calor latente"* para llevar al material desde $T_\mathrm{m} - \epsilon$ hasta $T_\mathrm{m} + \epsilon$, es decir que la curva de entalpía $h(T)$ tiene una discontinuidad en T_m. Notemos que entonces $C_p(T)$ de (5.6) tiene un comportamiento tipo $\delta(T - T_\mathrm{m})$ donde δ es la distribución delta de Dirac. La ecuación del calor con cambio de fase sigue siendo válida *en el sentido de las distribuciones* debido a la derivada de la entalpía con respecto a x. Integrando entre dos puntos x_1 y x_2 se llega a un "balance de energía" de la forma

$$q_2 - q_1 = k\frac{\partial T}{\partial x}\bigg|_{x_2} - k\frac{\partial T}{\partial x}\bigg|_{x_1} = a\rho[h(T_2) - h(T_1)] = a\Delta h \tag{5.8}$$

donde $q = k\,(\partial T/\partial x)$ es el flujo de calor y Δh es el incremento de entalpía de T_1 a T_2. Haciendo tender $x_{1,2}$ por derecha y por izquierda al punto donde se produce el cambio de fase $x = s$, es decir tal que $T(s) = T_\mathrm{m}$ tenemos el balance de energía en la interfase

$$k\frac{\partial T}{\partial x}\bigg|_{s^+} - k\frac{\partial T}{\partial x}\bigg|_{s^-} = a\rho L \tag{5.9}$$

Entonces, el problema puede ponerse como un problema de frontera libre de la siguiente manera

$$a\rho C_p \frac{\partial T}{\partial x} - \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) = 0, \quad \text{en } 0 < x < s, \quad s < x, L \tag{5.10}$$

con condiciones de contorno $T(0) = T_0$, $T(L) = T_L$, y condiciones de empalme en la interfase,

$$T(s^+) = T(s^-), \qquad \text{continuidad de la temperatura} \tag{5.11}$$

$$k\frac{\partial T}{\partial x}\bigg|_{s^+} - k\frac{\partial T}{\partial x}\bigg|_{s^-} = a\rho L, \qquad \text{balance de energía en la interfase} \tag{5.12}$$

$$T(s) = T_\mathrm{m}, \qquad \text{posición de la interfase} \tag{5.13}$$

Este problema es no-lineal debido a la determinación de la posición de la interfase s.

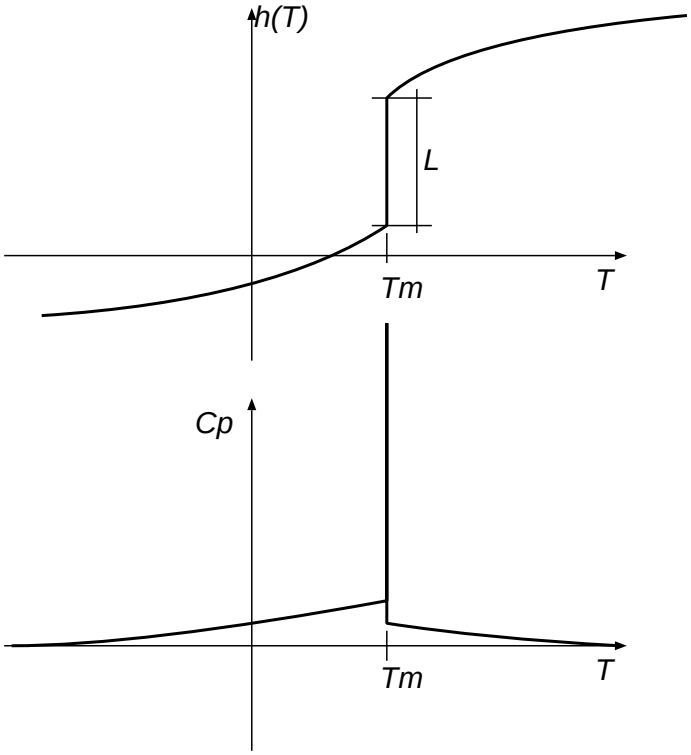

Figura 5.1: Curva de entalpía para un material con cambio de fase

La base de muchos métodos es hacer un desarrollo de Taylor alrededor de la iteración x^n de la forma

$$F(x) \approx F(x^n) + F'(x)\,(x - x^n) \tag{5.14}$$

donde

$$\{F'(x)\}_{ij} = \frac{\partial f_i}{\partial x_j}(x) \tag{5.15}$$

El *Teorema Fundamental del Cálculo* dice que

Teorema 4.0.1: Sea F diferenciable en un conjunto abierto $\omega \subset \mathbb{R}^N$ y sea $x^* \in \Omega$. Entonces para todo $x \in \Omega$ suficientemente cerca de x^* tenemos que

$$F(x) - F(x^*) = \int_0^1 F'(x^* + t(x - x^*))\,(x - x^*)\,\mathrm{d}t \tag{5.16}$$

5.1. Tipos de convergencia

Los métodos iterativos se clasifican de acuerdo a como convergen a la solución

Definición 4.1.1. Sea $\{x_n\} \subset \mathbb{R}^N$ y $x^* \in \mathbb{R}^N$. Entonces

- $x_n \to x^*$ *q-cuadráticamente* si $x_n \to x^*$ y existe una constante $K > 0$ tal que

$$\|x_{n+1} - x^*\| \leq K\,\|x_n - x^*\|^2 \tag{5.17}$$

  para n suficientemente grande.

- $x_n \to x^*$ *q-superlinealmente* con q-orden $\alpha > 1$ si $x_n \to x^*$ y existe una constante $K > 0$ tal que

$$\|x_{n+1} - x^*\| \leq K\,\|x_n - x^*\|^\alpha \tag{5.18}$$

- $x_n \to x^*$ *q-superlinealmente*
  si
  $$\lim_{n\to\infty} \frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|} = 0 \tag{5.19}$$
- $x_n \to x^*$ *q-linealmente* con q-factor $0 < \sigma < 1$ si
  $$\|x_{n+1} - x^*\| \leq \sigma \, \|x_n - x^*\| \tag{5.20}$$
  para n suficientemente grande.

Notar que para la convergencia lineal no es necesario el requisito de que la serie converge a x^* ya que esto es una consecuencia directa de (5.20), mientras que en el caso cuadrático y superlineal esto solo ocurre si alguno de los x_n es tal que el producto $K\,\|x_n - x^*\| < 1$ (más sobre esto después).

Definición 4.1.2.: Un método iterativo se dice que converge localmente q-cuadráticamente, superlinealmente, linealmente, etc... si sus iteraciones convergen q-cuadráticamente, etc... asumiendo que el valor inicial x_0 está suficientemente cerca de la solución x^*.

Obviamente la convergencia cuadrática es una convergencia superlineal de orden 2.

Notar que la noción de convergencia local no está directamente relacionada con la global. La convergencia global es mucho más difícil de determinar que en sistemas lineales pero sí se puede decir mucho sobre la convergencia local, basándose en un análisis linealizado.

Acá también vale la contraposición entre "precisión" (aquí rapidez de convergencia) y "estabilidad". Aquellos algoritmos que exhiben muy buenas tasas de convergencia (como Newton) tienden a ser los más inestables si se miran desde el punto de vista global. Además, cuando se comparan diferentes métodos debe además evaluarse el costo computacional de la iteración.

5.2. Iteración de punto fijo

Muchos sistemas de ecuaciones no-lineales se pueden poner naturalmente de la forma

$$x = K(x) \tag{5.21}$$

donde K es un mapeo no-lineal. Una solución x^* de (5.21) se llama un *"punto fijo"* del mapeo K. La iteración de punto fijo es

$$x_{n+1} = K(x_n) \tag{5.22}$$

Esto se conoce también como *iteración no-lineal de Richardson*, *iteración de Picard* o *método de sustituciones sucesivas*.

Recordemos que dado $\Omega \subset \mathbb{R}^N$ y $G : \Omega \to \mathbb{R}^N$, G es continua Lipschitz con constante de Lipschitz γ si $\|G(x) - G(y)\| \leq \gamma \, \|x - y\|$, para todo $x, y \in \Omega$.

Definición 4.2.2.: K es un mapeo de contracción si K es Lipschitz continua con constante de Lipschitz $\gamma < 1$

Teorema 4.2.1. del mapeo de contracción. Sea Ω un subconjunto cerrado de $\mathbb{R}^N$ y K un mapeo de contracción tal que $K(x) \in \Omega$ para todo $x \in \Omega$, entonces existe un único punto fijo x^* de K y la iteración de Richardson (5.22) converge linealmente a x^* con factor γ para cualquier x_0 inicial con $x_0 \in \Omega$.

Demostración: Sea $x_0 \in \Omega$, entonces es fácil ver que $x_n \in \Omega$ para todo $n \geq 1$. La secuencia $\{x_n\}$ es acotada ya que

$$\begin{aligned} \|x_{i+1} - x_i\| &= \|K(x_i) - K(x_{i-1})\| && (5.23)\\ &\leq \gamma \|x_i - x_{i-1}\| && (5.24)\\ &\leq \gamma^i \|x_1 - x_0\| && (5.25)\\ & && (5.26) \end{aligned}$$

y entonces

$$\begin{aligned} \|x_n - x_0\| &= \left\| \sum_{i=1}^{n-1} x_{i+1} - x_i \right\| && (5.27)\\ &\leq \sum_{i=1}^{n-1} \|x_{i+1} - x_i\| && (5.28)\\ &\leq \|x_1 - x_0\| \sum_{i=1}^{n-1} \gamma^i && (5.29)\\ &\leq \frac{1}{1-\gamma} \|x_1 - x_0\| && (5.30) \end{aligned}$$

Análogamente, puede mostrarse fácilmente que

$$\|x_{n+k} - x_n\| \leq \frac{\gamma^n}{1-\gamma} \|x_1 - x_0\| \tag{5.31}$$

de manera que $\{x_n\}$ es una secuencia de Cauchy y tiene un límite x^*. Si K tiene dos puntos fijos x^*, y^* entonces

$$\|x^* - y^*\| = \|K(x^*) - K(y^*)\| \leq \gamma \|x^* - y^*\| \tag{5.32}$$

Pero esto es una contradicción ya que asumimos $\gamma < 1$. □

Un caso típico de problema que lleva a un punto fijo no-lineal es la resolución de ODE's (*"Ordinary Differential Equations"*) no-lineales. Consideremos un tal sistema de la forma

$$y' = f(y), \quad y(t_0) = y_0 \tag{5.33}$$

Discretizando por el método de Backward Euler, llegamos a

$$y^{n+1} - y^n = h f(y^{n+1}) \tag{5.34}$$

donde

$$y^k \approx y(t_0 + hk) \tag{5.35}$$

y h es el paso de tiempo. Ec. (5.34) es un sistema de ecuaciones no-lineal en y^{n+1} de la forma

$$y = K(y) = y^n + h\, f(y) \tag{5.36}$$

Asumiendo que f es acotada y Lipschitz continua

$$\|f(y)\| \leq m \quad \text{y} \quad \|f(x) - f(y)\| \leq M \|x - y\| \tag{5.37}$$

para todos x, y, entonces si elegimos h suficientemente chico tal que $hM < 1$, podemos ver que

$$\|K(x) - K(y)\| = h \|f(x) - f(y)\| \leq hM \|x - y\| \tag{5.38}$$

Con lo cual podemos ver que para h suficientemente pequeño el esquema resulta ser convergente para cada paso de tiempo.

La convergencia local se puede estudiar, viendo que el hecho de que K sea un mapeo de contracción implica que $\|K'\| < 1$. En el caso de una sola dimensión

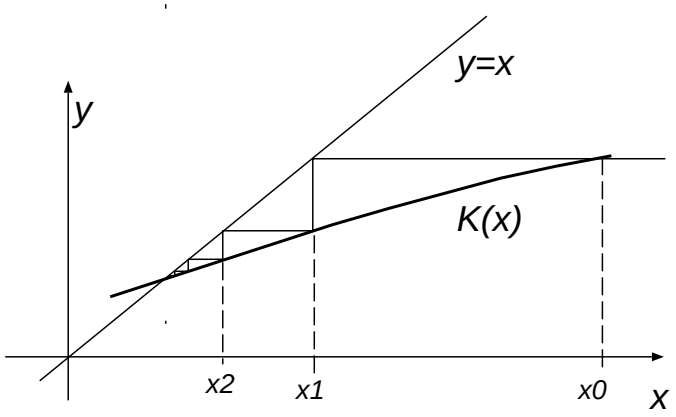

Figura 5.2: Convergencia monótona en el caso $0 < K'(x^*) < 1$

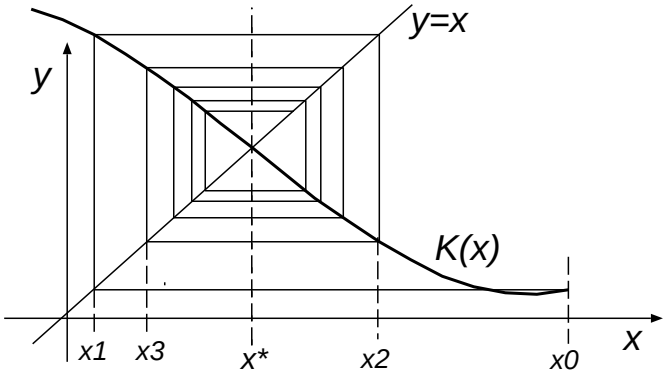

Figura 5.3: Convergencia oscilatoria en el caso $-1 < K'(x^*) < 0$

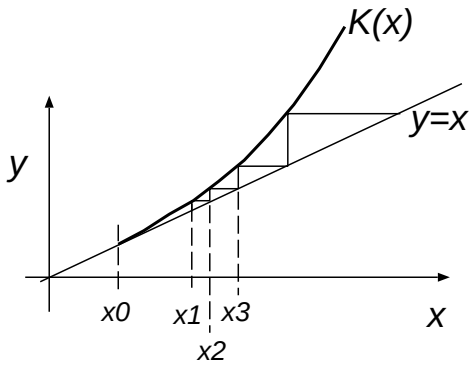

Figura 5.4: Divergencia monótona en el caso $K'(x^*) > 1$

5.3. Suposiciones estándar

En el marco de métodos no-lineales aparecen frecuentemente suposiciones sobre la continuidad de la función residuo $F(x)$. Llamaremos a estas suposiciones *"estándar"* y haremos frecuente mención a las mismas. Las usaremos inmediatamente para demostrar la convergencia del Método de Newton.

Suposición 4.3.1.

1. La ecuación (5.21) tiene una única solución x^*
2. $F' < \Omega \to \mathbb{R}^{N\times N}$ es Lipschitz continua con constante de Lipschitz γ
3. $F'(x^*)$ es no-singular.

Denotaremos por $\mathcal{B}(r)$ la bola de radio r alrededor de x^*.

Lema 4.3.1. Asumiendo las suposiciones 4.3.1., entonces existe $\delta > 0$ tal que para todo $x \in \mathcal{B}(\delta)$

$$\left\|F'(x)\right\| \leq 2\left\|F'(x^*)\right\| \tag{5.39}$$

$$\left\|F'(x)^{-1}\right\| \leq 2\left\|F'(x^*)^{-1}\right\| \tag{5.40}$$

$$\tfrac{1}{2}\left\|F'(x)^{-1}\right\|^{-1}\|e\| \leq \|F(x)\| \leq 2\left\|F'(x)\right\|\|e\| \tag{5.41}$$

donde $e = x - x^*$ es el error.

Capítulo 6

Método de Newton

Supongamos que tenemos x_n y queremos obtener x_{n+1}. Haciendo un desarrollo de Taylor a primer orden en $s = x_{n+1} - x_n$ e igualando este desarrollo a cero obtenemos una ecuación lineal para x_{n+1}

$$F(x_{n+1}) \approx F(x_n) + F'(x_n)\,(x_{n+1} - x_n) = 0 \tag{6.1}$$

De donde sale

$$x_{n+1} = x_n - F'(x_n)^{-1}\,F(x_n) \tag{6.2}$$

Podemos verlo como una iteración de punto fijo para

$$x = K(x) = x - F'(x)^{-1}\,F(x) \tag{6.3}$$

En adelante denotaremos también

$$x_n \text{ como } x_c \text{ por “current” (actual)} \tag{6.4}$$
$$x_{n+1} \text{ como } x_+ \text{ el estado “avanzado”} \tag{6.5}$$
$$\tag{6.6}$$

Teorema 5.1.1.: Asumiendo las suposiciones estándar 3.4.1., entonces existe $K > 0$ y $\delta > 0$ tales que si $x_c \in \mathcal{B}(\delta)$, entonces x_c y x_+ relacionados por la iteración de Newton

$$x_+ = x_c - F'(x_c)^{-1}\,F(x_c) \tag{6.7}$$

satisfacen

$$\|e_+\| \leq K\,\|e_c\|^2 \tag{6.8}$$

Demostración: Sea δ suficientemente pequeño tal que valen las conclusiones del Lema 4.3.1. Por el Teorema 4.0.1 tenemos que

$$\begin{aligned} e_+ &= e_c - F'(x_c)^{-1}F(x_c) \\ &= e_c - F'(x_c)^{-1}\left\{\int_0^1 F'(x^* + te_c)\,e_c\,\mathrm{d}t\right\} \\ &= F'(x_c)^{-1}\int_0^1 \left\{F'(x_c) - F'(x^* + te_c)\right\}\,e_c\,\mathrm{d}t \end{aligned} \tag{6.9}$$

De manera que,

$$\begin{aligned}
\|e_+\| &\leq \left\|F'(x_c)^{-1}\right\| \int_0^1 \left\|F'(x_c) - F'(x^* + te_c)\right\| \|e_c\| \, \mathrm{d}t && (6.10)\\
&\leq \left\|F'(x_c)^{-1}\right\| \gamma \int_0^1 (1-t)\, \mathrm{d}t \, \|e_c\|^2 && (6.11)\\
&\leq \frac{\gamma}{2} \left\|F'(x_c)^{-1}\right\| \|e_c\|^2 && (6.12)\\
&\leq \gamma \left\|F'(x^*)^{-1}\right\| \|e_c\|^2 && (6.13)\\
&\leq K \|e_c\|^2 && (6.14)\\
& && (6.15)
\end{aligned}$$

usando (5.41) y tomando $K = \gamma \left\|F'(x^*)^{-1}\right\|$. □

Sin embargo, notar que la relación (6.8) no implica automáticamente la convergencia ya que, por ejemplo la serie $x_n = n$ la satisface y sin embargo diverge. Para que esta relación implique convergencia hay que imponer que la iteración inicial esté suficientemente cerca de la solución, de manera que $K \|x_0 - x^*\| < 1$.

Teorema 5.1.2.: Existe δ tal que si $x_0 \in \mathcal{B}(\delta)$, entonces Newton converge cuadráticamente.

Demostración: Sea δ suficientemente pequeño tal que vale el Teorema 5.1.1. y además $K\delta = \eta < 1$, entonces

$$\|e_{n+1}\| \leq K \|e_n\|^2 \leq \eta \|e_n\| \tag{6.16}$$

de manera que

$$\|x_{n+1}\| \in \mathcal{B}(\eta\delta) \subset \mathcal{B}(\delta) \tag{6.17}$$

y entonces

$$\|e_{n+1}\| \leq \eta^n \|e_0\| \tag{6.18}$$

y $x^n \to x^*$ converge q-quadráticamente. □

Es útil la convergencia local? La suposición de que $x_0 \in \mathcal{B}(\delta)$, es decir que x_0 este suficientemente cerca de x^* puede parecer un poco artificial, pero hay situaciones en las que esto ocurre naturalmente.

Integración implícita de ODE's: Sea una sistema de ODE's de la forma

$$y' = f(y) \tag{6.19}$$

Haciendo una discretización con el método de Backward Euler tenemos

$$\frac{y^{n+1} - y^n}{\Delta t} = f(y^{n+1}) \tag{6.20}$$

Lo cual representa un sistema de ecuaciones no-lineal en y^{n+1} para cada paso de tiempo. Ahora si consideramos y^{n+1} como función de Δt, vemos que se acerca a y^n para $\Delta t \to 0$, de manera que tomando un paso de tiempo suficientemente pequeño garantizamos la convergencia de la resolución del sistema no-lineal en cada paso de tiempo

Refinamiento de soluciones numéricas: Supongamos que estamos resolviendo un problema de mecánica del continuo por FDM o FEM. Es usual ir refinando la malla hasta obtener una solución aceptable. Si el problema a resolver es no-lineal, es usual inicializar el esquema iterativo en la malla más fina $h = h_2$ (h= paso de la malla) con una interpolación de la solución en la malla más gruesa $h = h_1$. Es de esperar que si el esquema de discretización es convergente las soluciones en las dos mallas esten muy próximas entre sí.

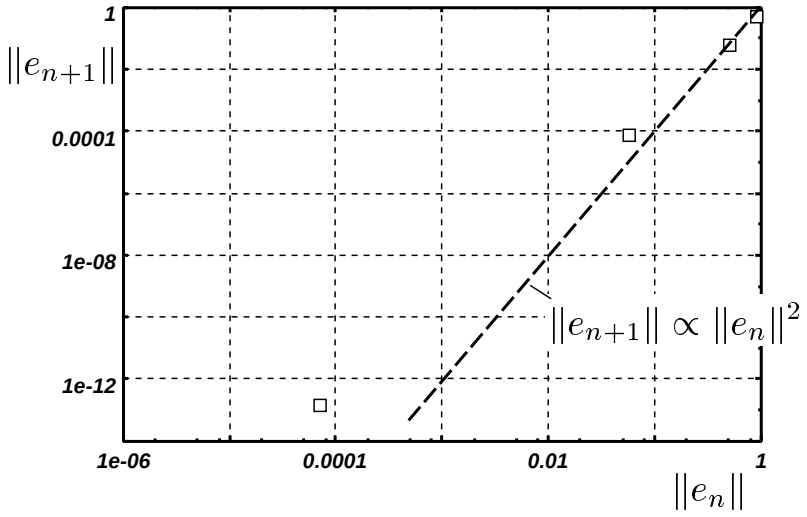

Figura 6.1: Determinación del orden de convergencia. (En este caso cuadrática.)

Método de continuación: Muchas veces tenemos un problema no-lineal dependiente de un parámetro λ

$$F(x, \lambda) = 0 \tag{6.21}$$

λ puede ser un parámetro físico (el número de Reynolds en un problema de mecánica de fluidos, o la carga aplicada en un problema de elasticidad). En cierto rango de λ el problema es lineal (pequeños números de Reynolds, pequeñas cargas) y se hace más no-lineal en otro rango. La idea es que tal vez si podemos resolver el problema para un cierto valor de λ, entonces tomando esta solución x_λ como inicialización para $\lambda + \Delta\lambda$ para $\Delta\lambda$ pequeño, tengamos muchas más probabilidades de converger. Notar que haciendo $\Delta\lambda$ arbitrariamente pequeño, las soluciones también tienden a acercarse tan cerca como uno quiera. Además, muchas veces el contar con la solución para todo el rango de valores de λ tiene su propio interés (esto es cierto en los ejemplos mencionados).

Por otra parte la convergencia local describe siempre como será la convergencia en la etapa final, independientemente del proceso inicial.

Cómo se determina experimentalmente el orden de convergencia. Como siempre, lo más usual es graficar el error versus el número de iteración. La convergencia lineal da una recta, pero las convergencias de más alto orden son un poco más difícil de determinar. En ese caso, La mejor forma de determinar el orden de convergencia de un algoritmo es graficar log-log $\|e_{n+1}\|$ versus $\|e_n\|$ y estimar la pendiente de la curva, que determina el orden convergencia. Por ejemplo en la figura 6.1 vemos el gráfico correspondiente al método de Newton y, por lo tanto, se espera una convergencia cuadrática. Vemos que hay una cierta proximidad a la relación cuadrática. Hay que tener en cuenta que, debido a la rapidez de la convergencia, usualmente se observan pocos puntos en la curva. Notar que, cuando se entra en la zona de convergencia cuadrática las razones entre dos residuos sucesivos se comportan cuadráticamente. Es decir, asumiendo que la igualdad vale en (6.8),

$$\frac{\|e_{n+1}\|}{\|e_n\|} = \frac{K \|e_n\|^2}{K \|e_{n-1}\|^2} = \left(\frac{\|e_n\|}{\|e_{n-1}\|}\right)^2 \tag{6.22}$$

Digamos por ejemplo que si el residuo baja dos órdenes de magnitud en una iteración, en la siguiente bajará 4. Esto es fácil de verificar mediante una simple cuenta.

6.1. Criterios de detención de la iteración

A partir de los resultados del Lema 4.3.1. podemos demostrar que la norma del residuo es equivalente a la del error. Tomando la desigualdad izquierda de (5.41) en x y la derecha en x_0 tenemos que

$$\|F(x)\| \geq \left\|F'(x^*)^{-1}\right\|^{-1} \|e\| /2 \tag{6.23}$$
$$\|F(x_0)\| \leq 2 \left\|F'(x^*)\right\| \|e_0\| \tag{6.24}$$

Dividiendo miembro a miembro,

$$\frac{\|F(x)\|}{\|F(x_0)\|} \geq \frac{\|e\|}{4 \|e_0\| \kappa(F'(x^*))} \tag{6.25}$$

Análogamente puede demostrarse una cota inferior para $\|e\|$ y queda

$$\frac{1}{4\kappa} \frac{\|e\|}{\|e_0\|} \leq \frac{\|F(x)\|}{\|F(x_0)\|} \leq 4\kappa \frac{\|e\|}{\|e_0\|} \tag{6.26}$$

donde $\kappa = \kappa(F'(x^*))$.

Recordemos que para sistemas lineales teníamos la opción de detener el proceso en base a $\|r\| / \|r_0\|$ o $\|r\| / \|b\|$. Aquí ya no existe naturalmente el concepto de miembro derecho b, pero muchas veces los sistemas no-lineales provenientes de sistemas físicos consisten en una serie de términos. Puede tomarse entonces como referencia aquel término que sea relativamente independiente de la solución y también que sea un término relativamente importante dentro del balance.

En la versión de los macros de Kelley tenemos dos parámetros que controlan la detención del esquema, a saber una tolerancia absoluta τ_a y una relativa τ_r. El esquema se detiene cuando

$$\|F(x)\| \leq \tau_r \|F(x_0)\| + \tau_a \tag{6.27}$$

Poniendo τ_r a cero el esquema se detiene cuando se obtiene el criterio absoluto y viceversa. Si ambos no son nulos, resulta en un criterio mixto.

Otra forma de detener el algoritmo es en base al tamaño del paso s

$$s = x_+ - x_c = -F'(x_c)^{-1} F(x_c) \tag{6.28}$$

y detener cuando $\|s\|$ sea chico. Este criterio tiene su justificación en la convergencia cuadrática ya que

$$\begin{aligned} e_c &= -\|s\| + e_+ \\ \|e_c\| &= \|s\| + O(\|e_c\|^2) \end{aligned} \tag{6.29}$$

Lo cual indica que si estamos suficientemente cerca de la solución, entonces prácticamente el paso s es el error.

Esto no es cierto para otros métodos si la convergencia no es tan buena. Por ejemplo consideremos convergencia lineal

$$\|e_+\| \leq \sigma \|e_c\| \tag{6.30}$$

Entonces,

$$\begin{aligned} \|e_c\| &\leq \|e_+\| + \|s\| \\ &\leq \sigma \|e_c\| + \|s\| \end{aligned} \tag{6.31}$$

de donde

$$\|e_c\| \leq \frac{1}{1-\sigma} \|s\| \tag{6.32}$$

De manera que en este caso sólo es válido asumir $\|s\| \approx \|e_c\|$ si $\sigma \ll 1$, es decir si el algoritmo converge muy rápidamente.

6.2. Implementación de Newton

En cuanto al conteo de las operaciones involucradas, asumiremos que se usa un método directo para resolver el paso de Newton (podríamos usar un método iterativo.) También asumiremos que el jacobiano es denso. Básicamente el método se basa en

1. Cálculo del residuo $F(x_c)$
2. Cálculo del jacobiano $F'(x_c)$
3. Factorización del jacobiano
4. Resolución del sistema lineal para el paso s
5. Update (actualización) del vector de iteración $x_+ = x_c + s$.

Normalmente los pasos más costosos (en tiempo de máquina) son los 2 y 3 (para grandes problemas predomina (3)). En el caso de matrices densas la factorización requiere $O(N^3)$ operaciones de punto flotante. Si evaluamos F' por diferencias finitas, entonces la columna j se puede aproximar por diferencias centradas como

$$[F'(x_c)\, e_j] \approx \frac{F(x_c + he_j) - F(x_c - he_j)}{2h} \tag{6.33}$$

Esto requiere 2 evaluaciones del residuo. Asumiremos que cada evaluación del residuo requiere $O(N)$ operaciones. Como hay que evaluar N columnas, el costo total de la evaluación del jacobiano es de $O(2N^2)$ ops. Si usamos diferencias avanzadas

$$[F'(x_c)\, e_j] \approx \frac{F(x_c + he_j) - F(x_c)}{h} \tag{6.34}$$

entonces el $F(x_c)$ se puede calcular una sola vez y el costo baja a $O((N+1)N)$ ops. con un sacrificio en la precisión.

En muchos casos el jacobiano puede ser evaluado directamente, con la consiguiente ganancia en rapidez y precisión, pero en otros casos el cálculo del jacobiano es una tarea muy laboriosa (en cuanto a la programación).

Interfase de los scripts dentro del paquete de Kelley: Los argumentos de entrada son el estado inicial x, el mapeo F definido como el nombre de una función, y un vector de tolerancias $\tau = [\tau_a \tau_r]$.

Algoritmo `newton(`x, F, τ`)`

1. $r_0 = \|F(x)\|$
2. Do while $\|F(x)\| > \tau_r r_0 + \tau_a$
   a. Calcule $F'(x)$
   b. Factorice $F'(x) = LU$
   c. Resolver $LUs = -F(x)$
   d. $x = x + s$
   e. Evaluar $F(x)$

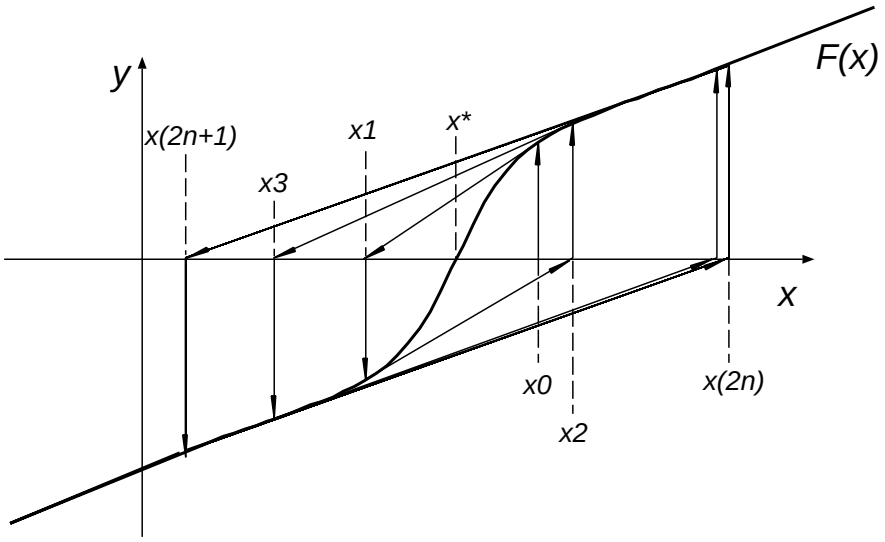

Figura 6.2: Newton no converge globalmente en ciertos casos.

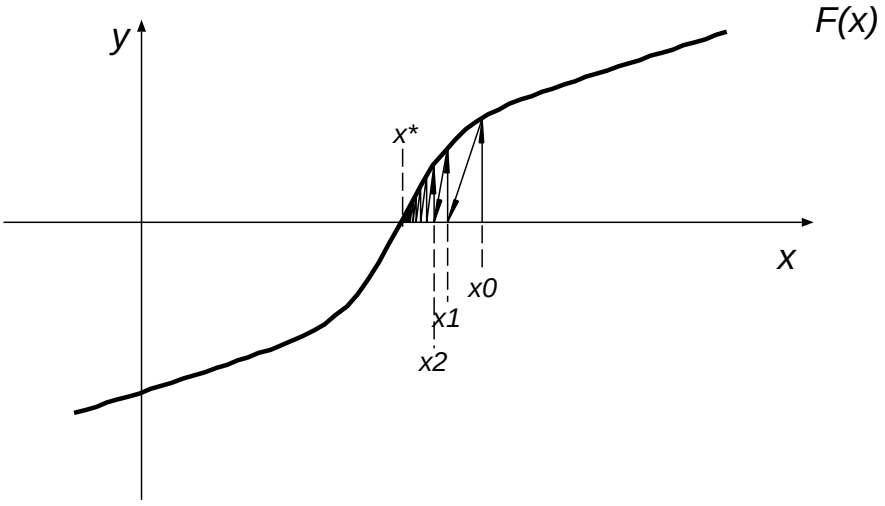

Figura 6.3: Newton subrelajado con $\omega = 0.4$ converge globalmente.

6.3. Sub-relajación de Newton

Muchas veces Newton puede no converger globalmente. Un caso típico es cuando la curva tiene una forma tipo "S" como en la figura. El algoritmo puede caer en un lazo, del cual no puede salir. Una forma de evitar esto es agregar *"sub-relajación"*. Básicamente es escalear el incremento s por una

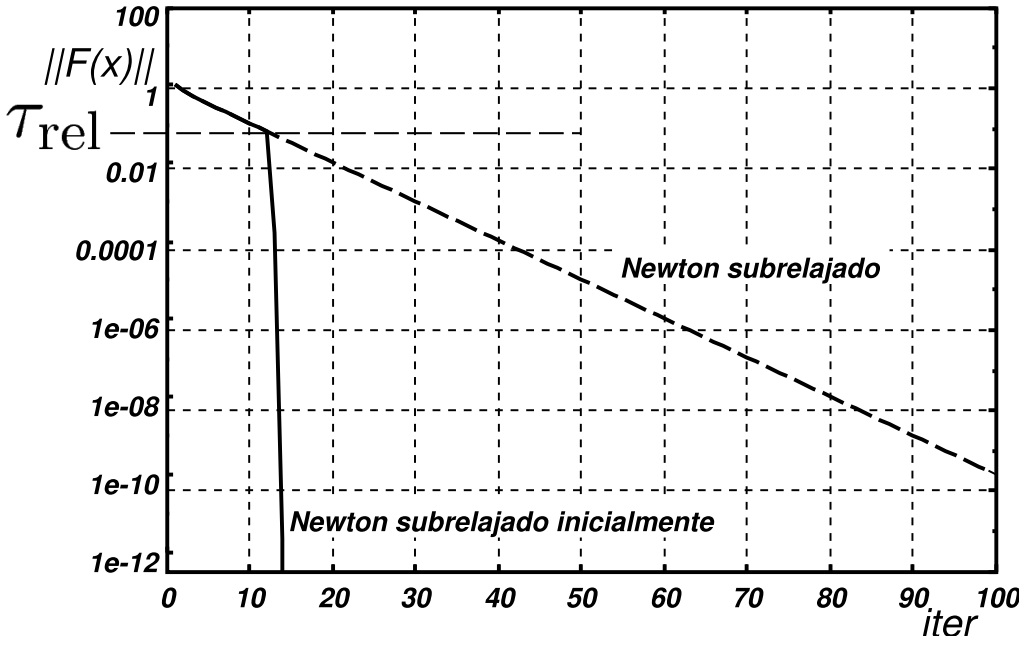

Figura 6.4: Convergencia del método de Newton subrelajado y subrelajado sólo inicialmente.

constante de relajación ω a determinar

$$x_{n+1} = x_n - \omega F'(x_n)^{-1} F(x_n) \tag{6.35}$$

Poniendo $\omega = 1$ recuperamos Newton, con ω pequeños tendemos a frenarlo y por lo tanto hacerlo más estable. Por ejemplo, con $\omega = 0.4$ el mismo caso se vuelve convergente, ver figura 6.3. En el caso de converger, la convergencia con subrelajación (en general con $\omega \neq 1$) será sólo lineal, ver figura 6.4. Una posibilidad de recuperar la convergencia cuadrática es volver el parámetro de relajación a $\omega = 1$ una vez que el residuo bajo suficientemente. A partir de ahí, la convergencia será cuadrática, si el algoritmo converge. Esto está reflejado en el siguiente algoritmo `nwtnsubr`.

Algoritmo `nwtnsubr`$(x, F, \omega_{\text{ini}}, \tau)$

1. $r_0 = \|F(x)\|$
2. Do while $\|F(x)\| > \tau_r r_0 + \tau_a$
   a. Calcule $F'(x)$
   b. Factorice $F'(x) = LU$
   c. Resolver $LUs = -F(x)$
   d. if $\|F(x)\| \leq \tau_{\text{rel}}$ then
      $\omega = \omega_{\text{ini}}$
      else
      $\omega = 1$
      endif
   e. $x = x + \omega s$
   f. Evaluar $F(x)$

El algoritmo tiene dos parámetros adicionales a saber el factor de relajación a ser usado en las iteraciones iniciales ω_{ini} y la tolerancia τ_{rel} para el residuo a partir del cual el factor de relajación es

reseteado a 1, esta forma parte del vector τ. En la figura 6.4 vemos las historias de convergencia para Newton subrelajado y subrelajado sólo inicialmente. Vemos que las dos curvas inicialmente son iguales hasta que se llega a la tolerancia τ_{rel}. A partir de allí el algoritmo converge cuadráticamente cuando se deja de subrelajar, mientras que converge linealmente cuando se continua subrelajando. Por supuesto, la elección de los parámetros τ_{rel} y ω_{ini} es fundamental y usualmente se basa en la experiencia del usuario. Un ω_{ini} o τ_{rel} demasiado altos puede hacer que se pierda la convergencia, mientras que tomar los valores demasiado bajos redunda en una pérdida de eficiencia.

6.4. Update condicional del jacobiano. Métodos de la cuerda y Shamanskii

Suponiendo que el jacobiano no varía mucho, podríamos reducir mucho el costo si sacamos (2a) y (2b) del lazo. Es decir, reemplazar la iteración por

$$x_+ = x_c - F'(x_0)^{-1} F(x_c) \tag{6.36}$$

Este es el método de la cuerda,

Algoritmo `chord`(x, F, τ)

1. $r_0 = \|F(x)\|$
2. Calcule $F'(x)$
3. Factorice $F'(x) = LU$
4. Do while $\|F(x)\| > \tau_r r_0 + \tau_a$
   a. Resolver $LUs = -F(x)$
   b. $x = x + s$
   c. Evaluar $F(x)$

La única diferencia es que F' es calculado y factorizado antes de empezar el lazo. El método de la cuerda puede ser localmente no-convergente.

6.5. El método de Shamanskii

Se basa en alternar una serie de m iteraciones tipo cuerda con una tipo Newton. El paso de x_c a x_+ es entonces

$$y_1 = x_c - F'(x_c)^{-1} F(x_c) \tag{6.37}$$
$$y_{j+1} = y_j - F'(x_c)^{-1} F(y_j) \quad 1 \le j \le m-1 \tag{6.38}$$
$$x_+ = y_m \tag{6.39}$$

Notar que para $m = 1$ coincide con Newton y para $m = \infty$ coincide con el método de la cuerda, donde $\{y_j\}$ son las iteraciones. Otra forma de pensarlo es una modificación al algoritmo de Newton en el que el jacobiano es calculado y factorizado sólo cada m iteraciones.

Algoritmo `sham`(x, F, τ, m)

1. $r_0 = \|F(x)\|$
2. Do while $\|F(x)\| > \tau_r r_0 + \tau_a$
   a. Calcule $F'(x)$
   b. Factorice $F'(x) = LU$

c. Resolver $LUs = -F(x)$
   i. Para $j = 1 \ldots, m$
   ii. $x = x + s$
   iii. Evaluar $F(x)$
   iv. Si $\|F(x)\| \leq \tau_r r_0 + \tau_a$ salir

El método de Shamanskii converge localmente, como Newton. Sin embargo también puede requerir de subrelajación para mejorar las propiedades de convergencia global.

6.6. Error en la función y las derivadas

Como ya mencionamos, el costo fundamental del método de Newton es el cálculo y la factorización del jacobiano $F'(x_c)$. Entonces, la mayoría de los métodos alternativos se basan en la aproximación del jacobiano, por ejemplo el método de la cuerda se basa en calcular el jacobiano una vez, factorizarlo y no modificarlo en lo sucesivo. Otra posibilidad es que reemplazar $F(x_c)^{-1}$ por la inversa de otra matriz que sea más fácil de factorizar, o por una aproximación a la inversa de $F(x_c)^{-1}$ (factorización incompleta o aproximada). Veamos ahora como influye el uso de una aproximación así como errores (probablemente de redondeo) en la evaluación del residuo. Sea entonces la iteración de la forma

$$x_+ = x_c - \left(F'(x_c) + \Delta(x_c)\right)^{-1} \left(F(x_c) + \epsilon(x_c)\right) \tag{6.40}$$

Podemos demostrar entonces el siguiente

Teorema 5.4.1.: Asumamos válidas las suposiciones estándar §5.3, entonces existe $\bar{K} > 0, \delta, \delta_1 > 0$ tal que si $x_c \in \mathcal{B}(\delta)$ y $\|\Delta(x_c)\| < \delta_1$ entonces x_+ está definida y satisface

$$\|e_+\| \leq \bar{K} \left(\|e_c\|^2 + \|\Delta(x_c)\| \, \|e_c\| + \|\epsilon(x_c)\|\right) \tag{6.41}$$

Demostración: Con "x_+ está definida" queremos básicamente decir que $F'(x_c) + \Delta(x_c)$ es no-singular. Sea δ suficientemente pequeño tal que vale el lema 4.3.1. (sección §5.3). Sea

$$x_+^N = x_c - F'(x_c)^{-1} F(x_c) \tag{6.42}$$

la iteración de Newton y notemos que

$$\begin{aligned} x_+ &= x_+^N + \left[F'(x_c)^{-1} - \left(F'(x_c) + \Delta(x_c)\right)^{-1}\right] F(x_c) && (6.43) \\ &\quad -\left(F'(x_c) + \Delta(x_c)\right)^{-1} \epsilon(x_c) && (6.44) \end{aligned}$$

Restando miembro a miembro la solución x^* obtenemos una expresión para los errores, y tomando normas

$$\begin{aligned} \|e_+\| &\leq \left\|e_+^N\right\| + \left\|F'(x_c)^{-1} - \left(F'(x_c) + \Delta(x_c)\right)^{-1}\right\| \, \|F(x_c)\| && (6.45) \\ &\quad + \left\|\left(F'(x_c) + \Delta(x_c)\right)^{-1} \epsilon(x_c)\right\| && (6.46) \end{aligned}$$

El primer término es el error que cometeríamos si usáramos el método de Newton y por lo tanto está acotado por el Teorema 5.1.1. (ecuación (6.8)). Con respecto al segundo término (el error cometido por la aproximación en el jacobiano), el factor $F(x_c)$ está acotado por el Lema 4.3.1. (sección §5.3) por

$$\|F(x_c)\| \leq 2 \left\|F'(x*)\right\| \|e_c\| \tag{6.47}$$

El primer factor es más complicado. Asumamos que tomamos $\Delta(x_c)$ tal que

$$\|\Delta(x_c)\| \leq \frac{1}{4} \left\|F'(x^*)^{-1}\right\|^{-1} \tag{6.48}$$

Entonces, por el Lema 4.3.1. (pág. 5.3)

$$\|\Delta(x_c)\| \leq \frac{1}{2}\left\|F'(x_c)^{-1}\right\|^{-1} \tag{6.49}$$

y por el Lema de Banach (pág. 10) tenemos que $F'(x_c)+\Delta(x_c)$ es no singular. Efectivamente, tomando

$$\begin{aligned} A &= F'(x_c)+\Delta(x_c) && (6.50)\\ B &= F'(x_c)^{-1} && (6.51)\end{aligned}$$

tenemos que

$$\begin{aligned} \|I-BA\| &= \left\|I-F'(x_c)^{-1}\left(F'(x_c)+\Delta(x_c)\right)\right\| && (6.52)\\ &= \left\|F'(x_c)^{-1}\,\Delta(x_c)\right\| && (6.53)\\ &\leq \left\|F'(x_c)^{-1}\right\|\,\|\Delta(x_c)\| && (6.54)\\ &\leq \left\|F'(x_c)^{-1}\right\|\left(\frac{1}{2}\left\|F'(x_c)^{-1}\right\|^{-1}\right) \quad \text{por (6.49)} && (6.55)\\ &\leq \frac{1}{2} && (6.56)\end{aligned}$$

Entonces vale el Lema de Banach, $A = F'(x_c)+\Delta(x_c)$ y por (1.58)

$$\begin{aligned} \left\|A^{-1}\right\| &= \left\|\left(F'(x_c)+\Delta(x_c)\right)^{-1}\right\| && (6.57)\\ &\leq \frac{\|B\|}{1-\|I-BA\|} && (6.58)\\ &\leq \frac{\left\|F'(x_c)^{-1}\right\|}{1-\frac{1}{2}} && (6.59)\\ &\leq 2\left\|F'(x_c)^{-1}\right\| \leq 4\left\|F'(x^*)^{-1}\right\| \quad \text{por Lema 4.3.1.} && (6.60)\end{aligned}$$

de manera que

$$\begin{aligned} \|F'(x_c)^{-1} &- \left(F'(x_c)+\Delta(x_c)\right)\| \leq && (6.61)\\ &\leq \left\|F'(x_c)^{-1}\left(F'(x_c)+\Delta(x_c)-F'(x_c)\right)\left(F'(x_c)+\Delta(x_c)\right)^{-1}\right\| && (6.62)\\ &\leq \left\|F'(x_c)^{-1}\right\|\,\|\Delta(x_c)\|\,\left\|\left(F'(x_c)+\Delta(x_c)\right)^{-1}\right\| && (6.63)\\ &\leq \left\|F'(x_c)^{-1}\right\|\,\|\Delta(x_c)\|\,\left(2\left\|\left(F'(x_c)^{-1}\right\|\right) \quad \text{por (6.60)} && (6.64)\\ &\leq 8\left\|F'(x^*)^{-1}\right\|^2\,\|\Delta(x_c)\| \quad \text{por el Lema 4.3.1.} && (6.65)\end{aligned}$$

Volviendo entonces a (6.46)

$$\begin{aligned} \|e_+\| \leq\ & K\|e_c\|^2 + 16\left\|F'(x^*)^{-1}\right\|^2\left\|F'(x^*)\right\|\,\|\Delta(x_c)\|\,\|e_c\| \\ & +4\left\|F'(x^*)^{-1}\right\|\,\|\epsilon(x_c)\| \end{aligned} \tag{6.66}$$

Poniendo

$$\bar{K} = K + 16\left\|F'(x^*)^{-1}\right\|^2\left\|F'(x^*)\right\| + 4\left\|F'(x^*)^{-1}\right\| \tag{6.67}$$

se llega a (6.41). □

Ahora aplicaremos este resultado al método de la cuerda.

6.7. Estimación de convergencia para el método de la cuerda

Recordemos que el método de la cuerda es

$$x_+ = x_c - F'(x_0)^{-1} F(x_c) \tag{6.68}$$

En el lenguaje del teorema 5.4.1. (pág. 6.6) tenemos

$$\begin{aligned} \epsilon(x_c) &= 0 && (6.69)\\ \Delta(x_c) &= F'(x_0) - F'(x_c) && (6.70) \end{aligned}$$

Si x_c, $x_0 \in \mathcal{B}(\delta) \subset \Omega$ entonces

$$\begin{aligned} \|\Delta(x_c)\| &\leq \gamma \|x_0 - x_c\| && (6.71)\\ &= \gamma \|(x_0 - x^*) - (x_c - x^*)\| && (6.72)\\ &\leq \gamma(\|e_0\| + \|e_c\|) && (6.73) \end{aligned}$$

Podemos estimar una estimación de la convergencia aplicando el teorema 5.4.1. (pág. 6.6).

Teorema 5.4.2." Asumamos las suposiciones estándar (pág. 5.3), entonces existe $K_c > 0$ y $\delta > 0$ tales que si $x_0 \in \mathcal{B}(\delta)$, la iteración de la cuerda (6.68) converge q-linealmente a x^* y

$$\|e_{n+1}\| \leq K_c \, \|e_0\| \, \|e_n\| \tag{6.74}$$

Demostración': Aplicando el Teorema 5.4.1 (pág. 6.6) con los valores (6.69,6.70)tenemos que

$$\begin{aligned} \|e_{n+1}\| &\leq \bar{K} \left[\|e_n\|^2 + \gamma(\|e_0\| + \|e_n\|) \, \|e_n\| \right] && (6.75)\\ &\leq \bar{K} \left[\|e_n\| \, (1+\gamma) + \gamma \, \|e_0\| \right] \|e_n\| && (6.76)\\ &\leq \bar{K} \, (1+2\gamma) \, \delta \, \|e_n\| && (6.77) \end{aligned}$$

Tomando δ suficientemente pequeño tal que

$$\bar{K} \, (1+2\gamma) \, \delta = \eta < 1 \tag{6.78}$$

entonces vemos que converge linealmente, por lo tanto $\|e_n\| \leq \|e_0\|$ y entonces

$$\|e_{n+1}\| \leq \bar{K} \, (1+2\gamma) \, \|e_0\| \, \|e_n\| \tag{6.79}$$

con lo cual el teorema queda demostrado poniendo $K_c = \bar{K} \, (1+2\gamma)$. □

Igualmente, pueden establecerse la convergencia para el método de Shamanskii y otros. Esto está discutido en más detalle en el libro de Kelley.

6.8. Aproximación por diferencias del Jacobiano

Un problema práctico que se encuentra en la práctica es como escoger la magnitud de la perturbación al hacer aproximaciones por diferencias finitas para el cálculo del jacobiano. Nosotros queremos aproximar el producto del jacobiano por una dirección arbitraria w como

$$F'(x) \, w \approx \frac{F(x + hw) + \epsilon(x + hw) - F(x) - \epsilon(x)}{h} \tag{6.80}$$

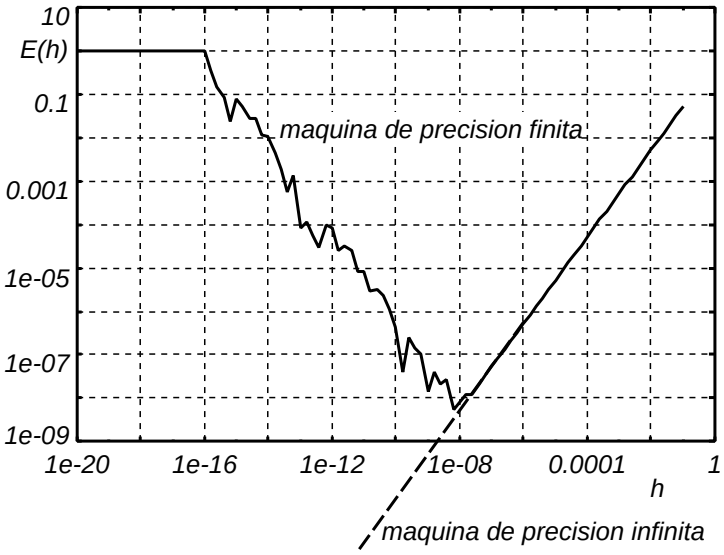

Figura 6.5: Error por aproximación por diferencias.

Como siempre, $\epsilon(x)$ representa aquí el error de redondeo. Es obvio que para h suficientemente pequeño la representación binaria de $x + hw$ será igual a la de x, de manera que la derivada aproximada numéricamente dará cero. Por otra parte, para h muy grandes tendremos el error típico $O(h)$ (para derivadas laterales) de la discretización numérica. Es de esperar entonces que haya un h óptimo para el cuál el error sea mínimo. Efectivamente, llamando $E(h)$ a la diferencia entre $F'(x)\,w$ y la discretización numérica, y haciendo una expansión de Taylor de orden 2 para $F(x + hw)$ en h obtenemos

$$E(h) = \frac{F(x + hw) + \epsilon(x + hw) - F(x) - \epsilon(x)}{h} - F'(x)\,w \tag{6.81}$$

$$\sim O(h + \frac{\bar{\epsilon}}{h}) \tag{6.82}$$

Notar que no podemos expandir $\epsilon(x + hw)$ debido a que, como es un error de redondeo, no podemos esperar que sea una función diferenciable de x. $\bar{\epsilon}$ representa un valor característico de $\epsilon(x)$. El mínimo de $E(h)$ se obtiene cuando

$$-1 + \frac{\bar{\epsilon}}{h^2} = 0 \tag{6.83}$$

o sea cuando $h = \sqrt{\bar{\epsilon}}$. En la figura 6.5 vemos un ejemplo práctico donde estamos calculando la derivada de e^x en $x = x_0 = 0.032$. Graficamos el error $E(h)$ versus h para h decrecientes desde 0.1 hasta 10^{-20}. Inicialmente (para $10^{-8} < h < 0.1$) el error decrece monótonamente y $\propto h$, como es de esperar por el análisis del error de truncamiento. Esto continuaría así para valores menores de h, de no ser por la aparición del error de truncamiento, debido al cual el error empieza a crecer desde allí, hasta saturarse para $h \approx 10^{-16}$. Efectivamente, para valores de $h < 10^{-16}$, la representación interna de $x + h$ es la misma que la de x y entonces la aproximación a la derivada es nula para esos valores de h, por lo tanto el error tiende a F' en una máquina de precisión finita (Matlab con doble precisión).

6.9. Guía 1. Método de Newton e iteración de punto fijo. Ejemplo sencillo 1D.

Consideremos la siguiente ecuación de una sola incógnita

$$F(x) = cx + \tanh(x) = 0 \tag{6.84}$$

La única solución es obviamente $x^* = 0$. Analizaremos las propiedades de convergencia local y global de los diferentes métodos partiendo de $x_0 \neq 0$.

a. Reescribir como iteración de punto fijo y ver si la iteración de Richardson no-lineal converge localmente (partir de x_0 suficientemente pequeño) y para que valores de c.

b. Aplicar un factor de relajación ω para hacer converger.

c. Aplicar Newton y comprobar convergencia local y cuadrática para diferentes valores de c.

d. Para $c = 0.2$ ver desde donde converge globalmente Newton.

e. Probar con Newton subrelajado. Se mantiene la convergencia cuadrática?

f. Agregar un test para que vuelva a $\omega = 1$ si $\|r\| < \tau_{\text{rel}}$. Hay convergencia global? Cómo es la convergencia local?

g. Aplicar `chord` y `sham`. Hace falta subrelajar?

Capítulo 7

Aplicación de resolución de sistemas no-lineales a problemas de PDE en 1D

7.1. Modelo simple de combustión

Consideremos una cierta región del espacio donde hay una cierta concentración de combustible y de oxígeno (O_2), por ejemplo la salida de un típico mechero Bunsen. (Ver figura 7.1). Veremos la posibilidad de que haciendo un aporte de energía inicial, el combustible entre en combustión y ésta se mantenga en forma estable. Empezaremos por el modelo más simple posible: el balance de energía para un punto representativo de la mezcla

$$\text{tasa de acumulación} = \text{generación} - \text{pérdida al medio ambiente} \tag{7.1}$$

$$\rho C_p \frac{\mathrm{d}T}{\mathrm{d}t} = q(T) - h\,(T - T_\infty) \tag{7.2}$$

donde ρ, C_p y T son la densidad, calor específico y temperatura de la mezcla, $q(T)$ el calor generado, T_∞ la temperatura del medio ambiente, y h un coeficiente que engloba las pérdidas de calor hacia el mismo. El calor de reacción $q(T)$ es positivo, ya que se trata de una reacción *"exotérmica"* y en

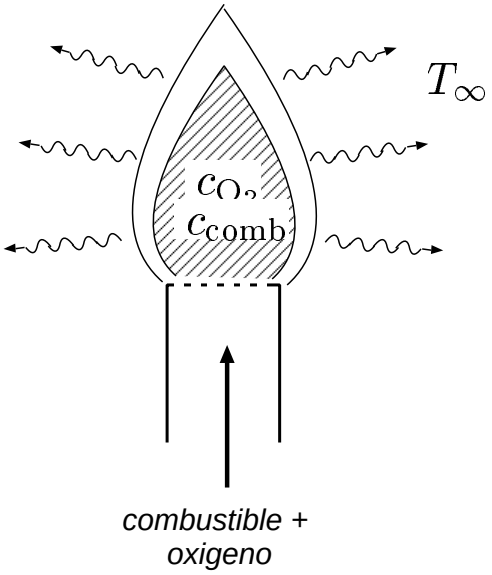

Figura 7.1: Modelo simple de combustión

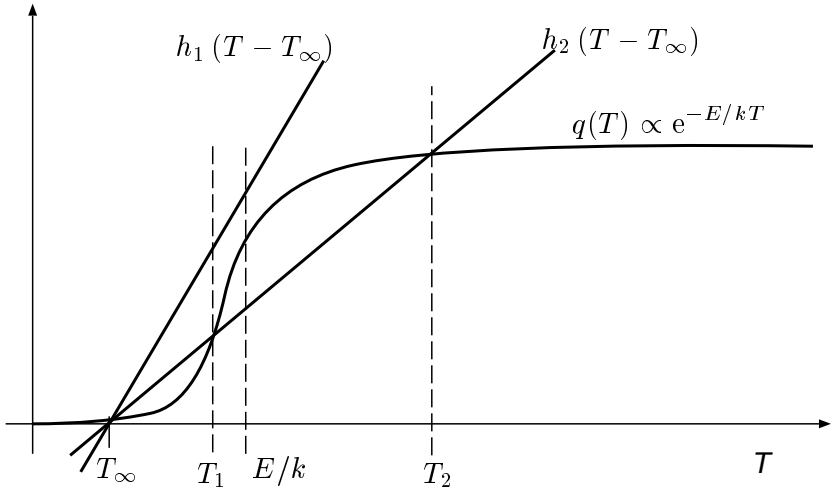

Figura 7.2: Factor exponencial que aparece en la ecuación de combustión

general es de la forma

$$q(T) = c_{\mathrm{O}_2} c_{\mathrm{comb}} A \exp(-E/kT) \tag{7.3}$$

es decir, es proporcional a la concentración de las sustancias que intervienen en la combustión (O_2 y el combustible), y un factor dependiente de la temperatura. E es una energía de activación, T es la temperatura absoluta (es decir en *"grados Kelvin"* $^\circ$K, $T[^\circ\mathrm{K}] = T[^\circ\mathrm{C}] + 273.16^\circ\mathrm{K}$). Como asumimos que la concentración de O_2 y el combustible es constante el único factor no constante es la exponencial que incluye la temperatura. Para temperaturas bajas $T \to 0$, tiende a cero mientras que a altas temperaturas tiende a una constante, resultando en una curva tipo "S". El cambio se produce alrededor de la temperatura característica de la reacción E/k. Ahora consideremos la suma de los dos términos del miembro derecho. Si esta suma se anula para una dada temperatura T^* entonces $T \equiv T^*$ =cte es una solución *"estacionaria"* de la ecuación diferencial. Para valores pequeños de h, como el h_1 en la figura 7.2, las curvas se intersectan en un sólo punto, cerca de T_∞. Llamaremos a esta la solución *"apagada"*. Para valores mayores de h, como h_2 en la figura, existen tres soluciones que llamaremos $T_{0,1,2}$, estando T_0 muy cerca de T_∞ y $T_{1,2}$ por encima de la la temperatura de activación E/k.

Ahora, llamemos $-c\varphi(T) = q(T) - h\,(T - T_\infty)$ donde $c > 0$ es una constante que usaremos después para hacer al análisis dimensional del problema. Asumiremos que h es tal que hay tres soluciones estacionarias y entonces $\varphi(T)$ tiene la forma general que se ve en la figura 7.3. de Haremos un *"análisis de estabilidad"* cerca de las soluciones estacionarias $T_{0,1,2}$. Consideremos, por ejemplo, que la condición inicial es $T(t = 0) = T_0 + \Delta t$. Hagamos un desarrollo en serie de Taylor de $\varphi(T)$ alrededor de T_0, entonces

$$\frac{\mathrm{d}T}{\mathrm{d}t} \approx \varphi(T_0) + \varphi'(T_0)(T - T_0) \tag{7.4}$$

cuya solución es

$$T(t) = T_0 + \Delta T \mathrm{e}^{-\varphi'(T_0)t} \tag{7.5}$$

pero $\varphi'(T_0) > 0$ de manera que la solución *se aproxima exponencialmente* a T_0. Llamaremos a este tipo de solución estacionaria *"estable"*. Básicamente queremos decir que si perturbamos un poco la solución de la estacionaria T_0, la *"fuerza restitutiva"* φ la tiende a hacer volver a T_0. La analogía mecánica es una bolita en un valle. El mismo razonamiento nos indica que para $\varphi' > 0$ la solución es *"inestable"*, es decir si perturbamos un poco la solución desde T_1 entonces la fuerza restitutiva la tiende a hacer apartar aún más de T_1 y diverge exponencialmente de la forma

$$T(t) = T_1 + \Delta T \mathrm{e}^{+|\varphi'(T_1)|\,t} \tag{7.6}$$

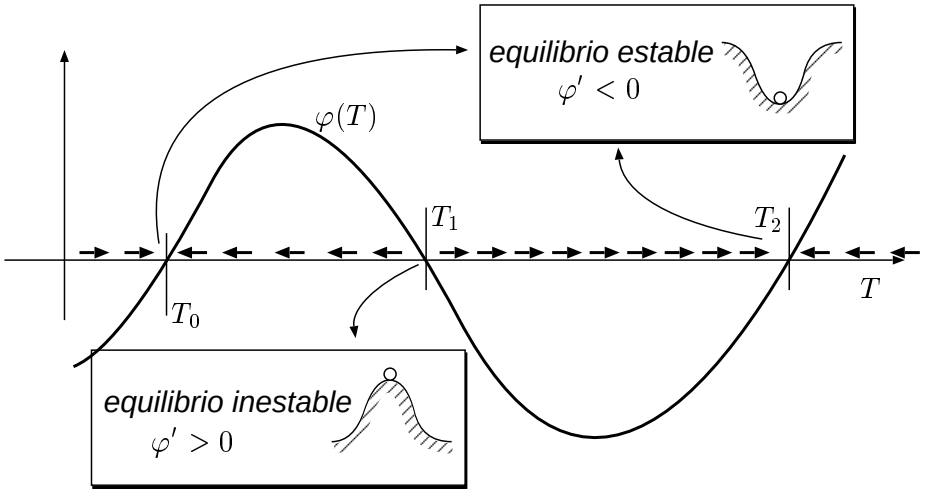

Figura 7.3: Función de combustión

Esta aproximación no es válida cuando T se aparta demasiado de T_1, pero de alguna forma T se aleja de T_1 y termina aproximándose exponencialmente a T_0 o T_2.

Ahora consideremos la resolución de la ecuación $\varphi(T) = 0$ para encontrar las soluciones estacionarias. El algoritmo de Newton converge localmente, por supuesto, independientemente del signo de la derivada φ', es decir, independientemente de si la solución es físicamente estable o no. Por otra parte podemos generar un algoritmo de punto fijo, integrando la ODE (7.2) en el tiempo con un método de Euler hacia adelante

$$\frac{T^{n+1} - T^n}{\Delta t} - c\,\varphi(T^n) \tag{7.7}$$

entonces

$$T^{n+1} = T^n - c\Delta t\,\varphi(T^n) = K(T^n) \tag{7.8}$$

Consideremos el criterio de convergencia del método de punto fijo, que indica que K debe ser un mapeo de contracción, en particular debe ser

$$|K'| = |1 - c\Delta t\,\varphi'(T^*)| < 1 \tag{7.9}$$

de manera que es claro que el esquema no será nunca convergente si $\varphi' < 0$, es decir si la solución estacionaria es inestable. Este método entonces, permite capturar *"solamente aquellas soluciones que son físicamente estables"*. Esta propiedad puede ser muy importante en problemas donde hay muchas soluciones y necesitamos descartar aquellas que son físicamente inaceptables.

Volvamos ahora a nuestro modelo de combustión, consideramos ahora que la combustión se produce en una región de una cierta extensión, por simplicidad en una sola dimensión $0 < x < L$ y que existe difusión del calor a través de la mezcla, con una difusividad k. Además, consideramos el problema estacionario, de manera que

$$-k\Delta T + \varphi(T) = 0 \tag{7.10}$$

con condiciones de contorno Dirichlet generales $T(x = 0) = T_0$, $T(x = L) = T_L$, aunque lo más usual es que tomaremos $T_{0,L} = T_\infty$. Si consideramos $k \to 0$, el problema deja de estar bien planteado, pero podemos pensar que en cada punto (es decir para cada x) tenemos una solución como la unidimensional

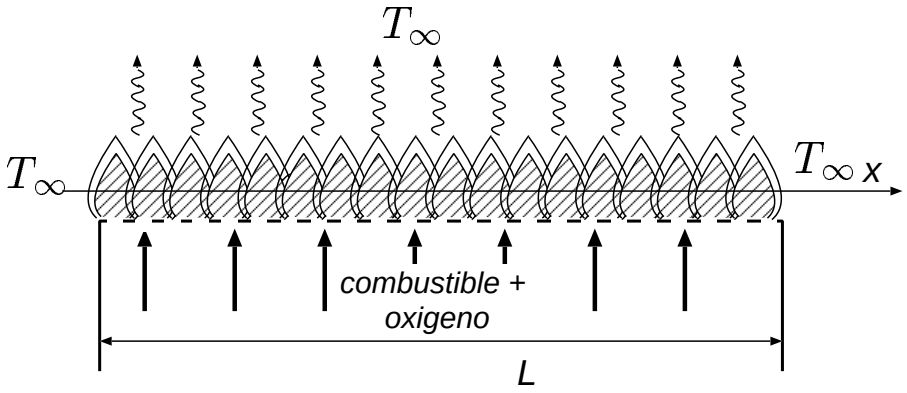

Figura 7.4: Problema de combustión unidimensional

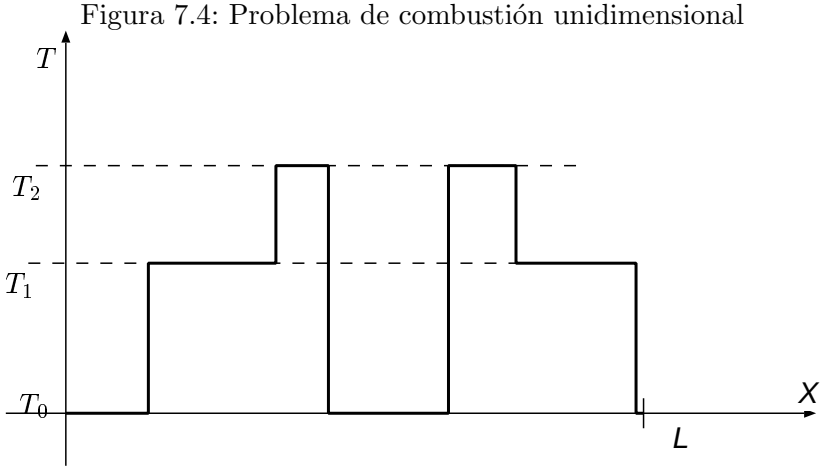

Figura 7.5: Soluciones para $k \to 0$.

anterior y entonces la temperatura puede tomar localmente cualquiera de los valores $T_{0,1,2}$, es decir, hay infinitas soluciones, cada una de ellas es constante de a trozos tomando los valores $T_{0,1,2}$. Ahora bien, el agregado de una cierta cantidad de difusividad $k > 0$ tiende a hacer que los saltos discontinuos del perfile $T(x)$ se redondeen, pero además, tiene un efecto similar al del término temporal, en cuanto a que los segmentos con $T = T_1$ se vuelven inestables, de manera que la solución para $k > 0$ será de la forma mostrada en la figura 7.6. A medida que k aumenta los saltos se vuelven más redondeados. Además a partir de un cierto k no es posible encontrar una solución *"prendida"*, es decir que sólo existe la solución *"apagada"* $T = T_0$.

Discretización: Asumiendo una grilla con paso constante $h = L/N$,

$$x_j = (j-1)h, \quad j = 1, \ldots, N+1 \tag{7.11}$$

Usando la discretización del operador de Laplace con el stencil de tres puntos, como ya hemos visto varias veces, la aproximación más sencilla posible para (7.10) es

$$\frac{-T_{i+1} + 2T_i - T_{i-1}}{h^2} + c\varphi(T_i) = 0, \quad i = 2, \ldots, N \tag{7.12}$$

puesto de forma matricial

$$AT + F(T) = 0 \tag{7.13}$$

donde A es la matriz del operador de Laplace con condiciones Dirichlet, como descripto en la Guía 1 y F es un operador no-lineal, pero "diagonal" en el sentido que F_i sólo depende de T_i.

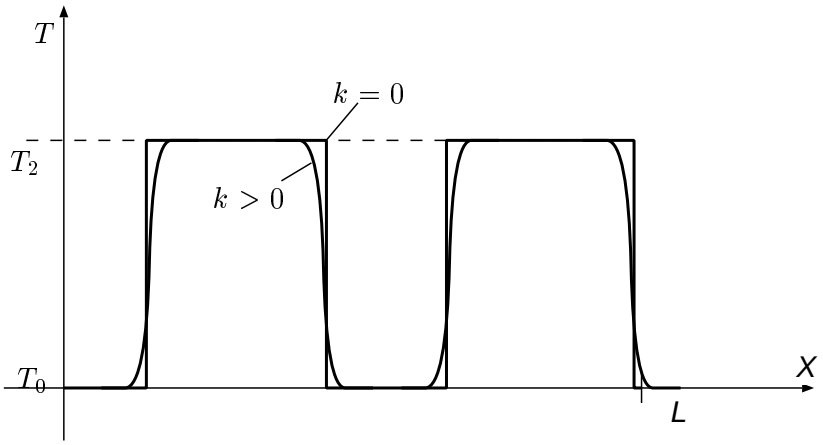

Figura 7.6: Soluciones para $k > 0$.

7.2. El problema de Stefan.

Consideremos el problema de un fluido moviéndose con velocidad U. La ecuación del balance de energía es

$$U\rho C_p \frac{\partial T}{\partial x} = k\Delta T \tag{7.14}$$

En el caso en que U, ρC_p y k son constantes, esto es que no dependen de la temperatura, este problema es lineal y ya lo hemos visto como la ecuación de *"advección-difusión"*. Ahora consideremos el caso cuando ρC_p depende de la temperatura, el problema se transforma en no-lineal. Definamos la cantidad

$$h(T) = \int_{T_{\text{ref}}}^{T} \rho C_p(T')\,\mathrm{d}T' \tag{7.15}$$

conocida como *"entalpía"* o *"contenido de calor"*. T_{ref} es una temperatura de referencia arbitraria. $h(T_2) - h(T_1)$ representa la cantidad de calor necesario para llevar al material desde la temperatura T_1 a la temperatura T_2. La única restricción física sobre $h(T)$ es que debe ser monótona creciente, ya que ante un agregado de calor un cuerpo no puede decrecer su temperatura. Notemos que el término no-lineal de advección (el miembro derecho de (7.14)) puede ponerse como $(\partial h/\partial x)$ de manera que la ecuación puede ponerse como

$$\frac{\partial h}{\partial x} = k\Delta T \tag{7.16}$$

Ahora consideremos que ocurre cuando hay un cambio de fase. Supongamos que el material está por debajo de la temperatura de cambio de fase T_{m} y vamos agregando calor de a poco, lo cual hace incrementar su temperatura. Al llegar a la temperatura de cambio de fase, necesitamos una cantidad de calor finita L llamado *"calor latente"* para llevar al cuerpo de $T_{\text{m}} - \epsilon$ a $T_{\text{m}} + \epsilon$. Esto nos indica que la curva de entalpía tiene una discontinuidad en $T = T_{\text{m}}$ (ver figura 5.1). Ahora descompongamos h en una componente regular y absorbamos la discontinuidad en un término proporcional a la función *"escalón de Heaviside"*

$$h(T) = h_{\text{reg}}(T) + L\,H(T - T_{\text{m}}) \tag{7.17}$$

donde

$$H(x) = \begin{cases} 0 & ;\ \text{si } x < 0 \\ 1 & ;\ \text{si } x \geq 0 \end{cases} \tag{7.18}$$

Ahora bien, si insertamos esto directamente en (7.16) llegamos a una ecuación de la forma

$$U(\rho C_p)_{\text{reg}} \frac{\partial T}{\partial x} = k\Delta T - UL\,\delta(x - s) \tag{7.19}$$

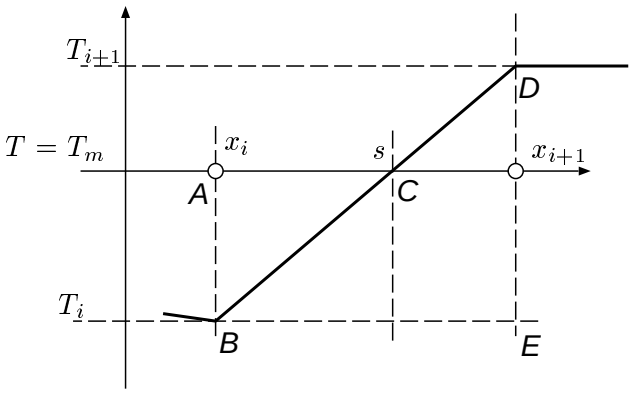

Figura 7.7: Posición de la interfase en el elemento.

donde

$$(\rho C_p)_{\text{reg}} = \frac{\partial}{\partial T}(\rho C_p) \tag{7.20}$$

y δ representa la distribución *"Delta de Dirac"*. s es el punto donde se produce el cambio de fase, es decir $T(s) = T_{\text{m}}$, es decir que s depende del campo de temperaturas, lo cual hace no-lineal el problema, incluso en el caso de que $(\rho C_p)_{\text{reg}}$ fuera constante. Notar que, si s fuera un punto fijo, entonces el problema previo representa un problema de transmisión del calor con una fuente de calor puntual.

Discretización: Recordemos que el problema de advección difusión tenía un problema adicional, que no tiene nada que ver con el cambio de fase, relacionado con la estabilidad a altos números de Peclet. Para evitar esto y poder concentrarnos en el problema del cambio de fase, asumiremos que en todos casos estamos muy por debajo del límite de estabilidad $\text{Pe}_h \ll 1$, de manera que no hace falta agregar los términos difusividad numérica o *"upwind"*. Entonces la discretización de (7.19) es

$$k\frac{T_{i+1} - 2T_i + T_{i-1}}{h^2} - (\rho C_p)_{\text{reg}} U \frac{T_{i+1} - T_{i-1}}{2h} - F(\tau) = 0 \tag{7.21}$$

donde F representa el término generado por la fuente puntual, τ el vector de temperaturas nodales. A continuación describiremos la discretización para F. Primero, debemos ubicar el punto en el cual se produce el cambio de fase. En el espíritu del método de elementos finitos, asumimos que las temperaturas son interpoladas linealmente entre los nodos. De manera que el cambio de fase se produce entre x_i y x_{i+1} si

$$(T_{i+1} - T_{\text{m}})(T_i - T_{\text{m}}) < 0 \tag{7.22}$$

Por similaridad de los triángulos ABC y BDE

$$\frac{s - x_i}{T_{\text{m}} - T_i} = \frac{h}{T_{i+1} - T_i} \tag{7.23}$$

de donde

$$\beta = \frac{s - x_i}{h} = \frac{T_{\text{m}} - T_i}{T_{i+1} - T_i} \tag{7.24}$$

donde $0 < \beta < 1$. Obviamente, si la interface cae en un nodo $s = x_i$ ponemos

$$F_j = \begin{cases} UL; & j = i \\ 0; & j \neq i \end{cases} \tag{7.25}$$

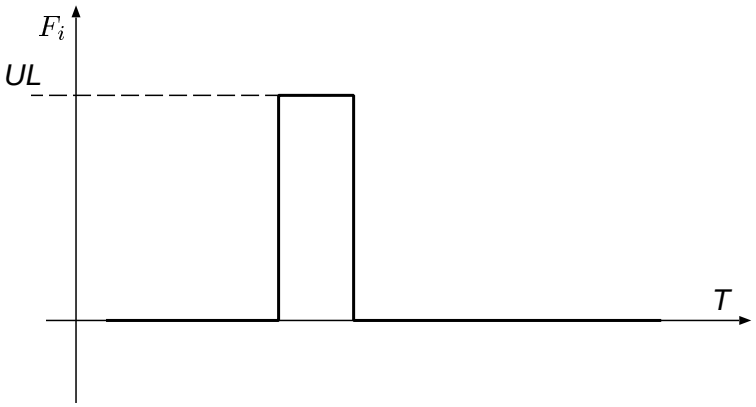

Figura 7.8: Contribución discontinua al residuo por el cambio de fase.

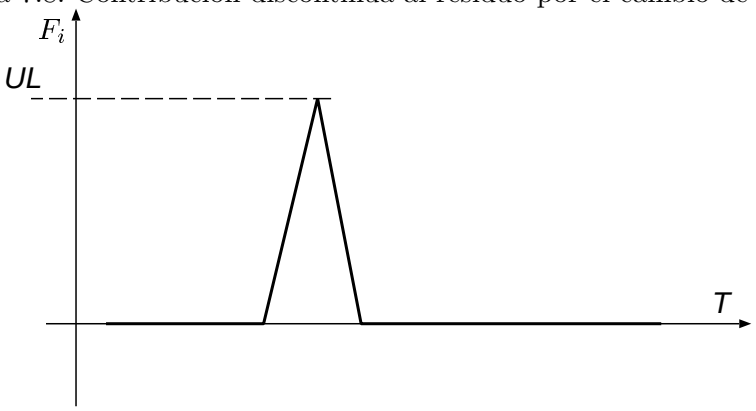

Figura 7.9: Contribución continua al residuo por el cambio de fase.

Si la interfase no coincide exactamente en un nodo es más complicado. Podríamos definir que agregamos todo el calor liberado por el cambio de fase UL en el nodos i o el $i+1$ dependiendo de si la interfase está más cerca de uno u otro nodo

$$F_i = UL, \quad F_{i+1} = 0 \qquad \text{si} \qquad \beta < 1/2 \tag{7.26}$$

$$F_i = 0, \quad F_{i+1} = UL \qquad \text{si} \qquad \beta > 1/2 \tag{7.27}$$

Ahora bien, la contribución al residuo por el cambio de fase $F(\tau)_i$ será de la forma mostrada en la figura 7.8. Esta contribución es discontinua y por lo tanto la función residuo total seguramente no será continua Lipschitz. De hecho, existe muy pocas probabilidades de que algún método converja para tal discretización. Otra posibilidad es distribuir el calor liberado UL linealmente entre los dos nodos

$$F_i = (1-\beta)UL, \quad F_{i+1} = \beta UL \tag{7.28}$$

De esta forma obtenemos una contribución al residuo continua (ver figura 7.9). En los scripts de Matlab distribuidos esta implementada esta última opción.

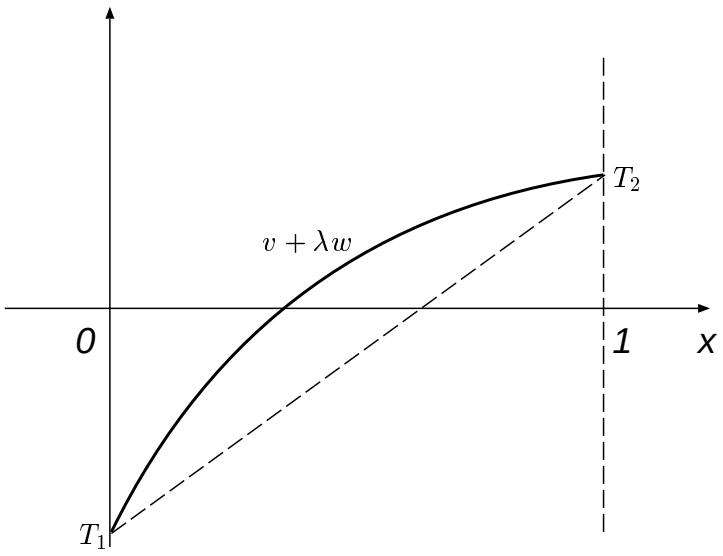

Figura 7.10: Campo de temperaturas dependiente de un parámetro.

7.3. Guía 3. Aplicación de resolución de sistemas no-lineales a problemas de PDE en 1D

- Hallar las soluciones y mostrar convergencias para diferentes valores de L.
- Cambiar la forma de repartir el calor generado por el cambio de fase como se indica en (7.27). Converge? Porqué? Graficar el residuo para algún nodo i sobre la recta que pasa por el punto v en $\mathbb{R}^N$ y dirección w, es decir graficar $F_i(v + \lambda w)$. Donde v y w están dados por

$$\begin{aligned} v_i &= T_1 + (T_2 - T_1)x_i \\ w_i &= x_i(1 - x_i) \end{aligned} \tag{7.29}$$

Capítulo 8

Newton inexacto.

Como mencionamos, el principal costo computacional del método de Newton es la resolución del sistema lineal que debe resolverse en cada iteración para obtener el paso s

$$F'(x_c)s = -F(x_c) \tag{8.1}$$

Entonces, es tentador de usar un método iterativo para resolver este sistema lineal. La pregunta es entonces, con que precisión es necesario resolverlo. En el lenguaje de los métodos iterativos para sistemas lineales que hemos usado en la primera parte del curso, tenemos

$$A \rightarrow F'(x_c) \tag{8.2}$$
$$x \rightarrow s \tag{8.3}$$
$$b \rightarrow F(x_c) \tag{8.4}$$

Como fue discutido en su momento (sección 1.1.3, pág. 7). Lo más común es tomar como criterio de detención para el esquema iterativo lineal una tolerancia en el residuo relativa sea a $\|b\|$ o al residuo en el vector de inicialización $\|r_0\|$. En este caso asumiremos que partimos como vector inicial para la iteración sobre s de $s = 0$ con lo cual ambos criterios coinciden y el criterio de detención es

$$\left\|F'(x_c)s + F(x_c)\right\| \leq \eta_c \left\|F(x_c\right\| \tag{8.5}$$

Llamaremos a $r = F'(x_c)s + F(x_c)$ el *"residuo lineal"*, o *"residuo del lazo interior"*. También es común llamar a η_c como el *"término forzante"*.

8.1. Estimaciones básicas. Análisis directo.

Teorema 6.1.1. Asumamos las suposiciones estándar (sección 5.3, pág. 70), entonces existe δ y K_I tales que si $x_c \in \mathcal{B}(\delta)$ y s satisface (8.5), $x_+ = x_c + s$, entonces

$$\|e_+\| \leq K_I(\|e_c\| + \eta_c)\ \|e_c\| \tag{8.6}$$

Demostración: Sea δ suficientemente pequeño tal que vale el Lema 4.3.1. (sección 5.3, pág. 70). Sea $r = -F'(x_c)s - F(x_c)$ el residuo lineal, entonces

$$s + F'(x_c)\, F(x_c) = -F(x_c)^{-1}\, r \tag{8.7}$$

y

$$e_+ = e_c + s \tag{8.8}$$
$$= e_c - F'(x_c)^{-1}\, F(x_c) - F'(x_C)^{-1}\, r \tag{8.9}$$

donde el segundo término del miembro derecho lo podemos identificar como el paso s si usáramos Newton y el tercer término como el error en el paso introducido al no resolver en forma exacta el problema lineal. La suma de los dos primeros términos podemos acotarla como el error del método de Newton (6.8)

$$\left\| e_c - F'(x_c)^{-1} F(x_c) \right\| \leq K \|e_c\|^2 \tag{8.10}$$

Mientras que el término restante lo podemos acotar como

$$\begin{aligned} \left\| F(x_c)^{-1} r \right\| &\leq \left\| F(x_c)^{-1} \right\| \|r\| && (8.11)\\ &\leq \left\| F'(x_c)^{-1} \right\| \eta_c \|F(x_c)\| && (8.12)\\ &\leq (2 \left\| F'(x^*)^{-1} \right\|) \eta_c (2 \|F(x^*)\|) \quad \text{por Lema 4.3.1.} && (8.13)\\ &= 4\kappa(F'(x^*)) \eta_c \|e_c\| && (8.14) \end{aligned}$$

donde $\kappa(F'(x^*))$ es el número de condición del jacobiano en la solución. Volviendo a (8.9) llegamos a que

$$\|e_+\| \leq K \|e_c\|^2 + 4\kappa(F'(x^*)) \eta_c \|e_c\| \tag{8.15}$$

De donde sale la demostración poniendo

$$K_I = K + 4\kappa(F'(x^*)) \tag{8.16}$$

□.

El resultado es bastante razonable ya que nos dice que si resolvemos el lazo interior en forma exacta ($\eta_c = 0$) entonces recuperamos una convergencia cuadrática como en Newton, y si no, tenemos una convergencia lineal, siendo el factor de convergencia proporcional a η_c.

Teorema 6.1.2.: Asumamos las suposiciones estándar (sección 5.3, pág. 70). Entonces existen δ y $\bar{\eta}$ tales que si $x_0 \in \mathcal{B}(\delta)$, y $\{\eta_n\} \subset [0, \bar{\eta}]$, entonces el Newton inexacto dado por

$$x_{n+1} = x_n + s_n \tag{8.17}$$

$$\left\| F'(x_n) s_n + F(x_n) \right\| \leq \eta_n \|F(x_n)\| \tag{8.18}$$

converge q-linealmente a x^*. Además

- Si $\eta_n \to 0$ la convergencia es q-superlineal
- Si $\eta_n \leq K_\eta \|F(x_n)\|^p$, $0 < p < 1$ para algún $K_\eta > 0$, entonces la convergencia es q-superlineal con q-orden $1 + p$.

Demostración: Sea δ suficientemente pequeño tal que vale el Teorema 6.1.1. (sección 8.1, pág. 92). Tomamos δ y $\bar{\eta}$ suficientemente pequeños tales que

$$K_I(\delta + \bar{\eta}) < 1 \tag{8.19}$$

Entonces, para $n \geq 0$ y $x_n \in \mathcal{B}(\delta)$ tenemos que

$$\begin{aligned} \|e_{n+1}\| &\leq K_I(\|e_n\| + \eta_n) \|e_n\| && (8.20)\\ &\leq K_I(\delta + \bar{\eta}) \|e_n\| && (8.21)\\ &\leq \|e_n\| && (8.22) \end{aligned}$$

lo cual asegura convergencia q-lineal con q-factor $K_I(\delta + \bar{\eta})$. Entonces tenemos convergencia superlineal ya que

$$\frac{\|e_{n+1}\|}{\|e_n\|} \leq K_I(\|e_n\| + \eta_n) \tag{8.23}$$

y el miembro derecho tiende a 0 para $n \to \infty$ ya que $\|e_n\|$ tiende a cero por la convergencia lineal ya demostrada y η_n tiende a cero por las hipótesis del teorema.

Por otra parte, si $\eta_n \leq K_\eta \|F(x_n)\|^p$ entonces

$$\begin{aligned} \|e_{n+1}\| &\leq K_I(\|e_n\|^2 + K_\eta \|F(x_n)\|^p \|e_n\|) && (8.24)\\ &\leq K_I(\|e_n\|^2 + K_\eta (2\|F(x_*)\|)^p \|e_n\|^{1+p}) \quad \text{por Lema 4.3.1} && (8.25)\\ &\leq K_I(\|e_n\|^{1-p} + K_\eta (2\|F(x_*)\|)^p) \|e_n\|^{1+p} && (8.26) \end{aligned}$$

Obviamente, usar $p > 1$ no puede mejorar la convergencia más allá de la cuadrática dada por el método de newton exacto.

8.2. Análisis en normas pesadas

Como $K_I = O(\kappa(F'(x^*))$ se podría esperar que si $F'(x^*)$ está mal condicionada entonces deberíamos permitir sólo muy pequeños términos forzantes. Veremos que éste no es el caso. Básicamente veremos que la única restricción es mantener $\{\eta_n\} \leq \eta < 1$. El análisis se basa en tomar normas pesadas con el jacobiano $F'(x*)$ es decir en la norma $\|\cdot\|_* = \|F'(x^*)\cdot\|$.

Teorema 6.1.3.: Asumamos las suposiciones estándar (sección 5.3, pág. 70). Entonces existe δ tal que si $x_c \in \mathcal{B}(\delta)$, y s satisface

$$\begin{aligned} \|F'(x_c)\, s + F(x_c)\| &\leq \eta_c \|F(x_c)\| && (8.27)\\ x_+ = x_c + s, \quad &\text{y } \eta_c \leq \eta < 1 && (8.28) \end{aligned}$$

entonces

$$\|F'(x^*)\, e_+\| \leq \bar{\eta} \|F'(x^*)\, e_c\| \tag{8.29}$$

Demostración: Por el Teorema fundamental del cálculo 4.0.1 (sección 5, pág. 66)

$$F(x) - F(x^*) = \int_0^1 F'(x^* + te_c)\, e_c \, \mathrm{d}t \tag{8.30}$$

Pero $F(x^*) = 0$ y sumando y restando en el integrando y tomando normas

$$\begin{aligned} \|F(x_c)\| &= \int_0^1 \|F'(x^* + t\, e_c)\, e_c - F'(x^*)\, e_c + F'(x^*)\, e_c\| \, \mathrm{d}t \\ &\leq \int_0^1 \left[\|F'(x^* + t\, e_c)\, e_c - F'(x^*)\, e_c\| + \|F'(x^*)\, e_c\|\right] \mathrm{d}t \end{aligned} \tag{8.31}$$

pero

$$\begin{aligned} \|F'(x^* + t\, e_c)\, e_c - F'(x^*)\, e_c\| &\leq \gamma \|te_c\| \|e_c\| \\ &\text{(cont. Lipschitz de } F') \\ &= \gamma t \|e_c\|^2 \end{aligned} \tag{8.32}$$

Volviendo a (8.31)

$$\|F(x_c\| \leq \|F'(x^*)\| \|e_c\| + \frac{1}{2}\gamma \|e_c\|^2 \tag{8.33}$$

Ahora el objetivo es acotar los errores en la norma $\|F'(x^*)\cdot\|$. Entonces, hacemos uso de que

$$\begin{aligned} \|e_c\| &= \|F'(x^*)^{-1} F'(x^*)\, e_c\| \\ &\leq \|F'(x^*)^{-1}\| \|F'(x^*)\, e_c\| \end{aligned} \tag{8.34}$$

y reemplazando en (8.33)

$$\begin{aligned}\|F(x_c)\| &\leq (1+\frac{1}{2}\gamma \left\|F'(x^*)\right\| \|e_c\|) \left\|F'(x^*)\, e_c\right\| \\ &= (1+M_0\delta) \left\|F'(x^*)\, e_c\right\|\end{aligned} \tag{8.35}$$

donde $M_0 = \frac{1}{2}\gamma \|F'(x^*)\|$.

Por otra parte, restando x^* de ambos miembros de (8.28)

$$e_+ = e_c + s \tag{8.36}$$

y aplicando $F'(x^*)$ a ambos miembros y tomando normas

$$\begin{aligned}\|F(x^*)e_+\| &= \left\|F(x^*)\,(e_c - F'(x_c)^{-1}\,F(x_c) - F'(x_c)^{-1}r)\right\| \\ &\leq \left\|F(x^*)\,[e_c - F'(x_c)^{-1}\,F(x_c)]\right\| + \left\|F(x^*)\,F'(x_c)^{-1}r)\right\|\end{aligned} \tag{8.37}$$

En el primer término aparece el error de la iteración de Newton, de manera que podemos usar (6.8) para acotarlo

$$\left\|F(x^*)\,[e_c - F'(x_c)^{-1}\,F(x_c)]\right\| \leq K\,\|F(x^*)\|\,\|e_c\|^2 \tag{8.38}$$

Por otra parte,

$$\begin{aligned}\left\|F(x^*)\,F'(x_c)^{-1}r)\right\| &\leq \|r\| + \left\|(F(x^*) - F'(x_c))\,F'(x_c)\,r\right\| \\ &\leq \|r\| + (2\gamma\,\|e_c\|) \left\|F(x^*)^{-1}\right\| \|r\| \quad \text{(Lema 4.3.1. y Lips.)}\end{aligned} \tag{8.39}$$

Poniendo,

$$M_1 = 2\gamma \left\|F(x^*)^{-1}\right\|, \quad M_2 = K\,\|F(x^*)\| \tag{8.40}$$

Llegamos a que

$$\begin{aligned}\|F(x^*)\,e_+\| &\leq (1+M_1\delta)\,\|r\| + M_2\,\delta\,\|e_c\| \\ &\leq (1+M_1\delta)\,\eta_c\,(1+M_0\delta)\,\|F(x^*)\,e_c\| \\ &\quad + M_2\,\delta \left\|F(x^*)^{-1}\right\| \|F(x^*)e_c\| \quad \text{por (8.27,8.35)} \\ &\leq \left[(1+M_1\delta)\,\eta_c\,(1+M_0\delta) + M_2\,\delta \left\|F(x^*)^{-1}\right\|\right] \|F(x^*)e_c\|\end{aligned} \tag{8.41}$$

Ahora bien, si $\eta_c \leq \eta$, entonces para cualquier $\eta < \bar{\eta} < 1$ existe un δ tal que el término entre corchetes en (8.41) es menor que $\bar{\eta}$ de manera que se cumple (8.29). □

8.3. Guía 4. Newton Inexacto

Considerar el siguiente problema unidimensional

$$-k\Delta T + c\phi(T) = q \tag{8.42}$$

La ecuación es similar al caso de combustión (sección 7.1, pág. 84), pero hemos agregado un término fuente q y modificaremos la función $\phi(T)$ con respecto a la estudiada en esa sección. El sistema discreto es (7.13) que lo denotaremos como

$$F(T) = AT + G(T) = 0 \tag{8.43}$$

(notar que F denota aquí el residuo y $G(T)$ el término de reacción y pérdidas) y el jacobiano correspondiente es

$$F' = A + G' \tag{8.44}$$

Ahora bien, A es la matriz correspondiente a $-\Delta$, de manera que es simétrica y definida positiva. Por otra parte G' es una matriz diagonal con elementos diagonales

$$G'_{ii} = \phi'(T_i) \tag{8.45}$$

de manera que el jacobiano total es simétrico. Para simplificar el análisis del Newton inexacto, trataremos de aplicar Gradientes Conjugados, por lo cual necesitamos poder asegurar que el jacobiano sea simétrico y definido positivo. Para que esto último sea verdadero basta con asegurar que $\phi' > 0$ es decir que ϕ sea monótona creciente. Esto es ciertamente no válido para el caso de combustión estudiado, por lo que modificaremos la función $\phi(T)$ apropiadamente. Tomaremos

$$\phi(T) = c \operatorname{atanh}\left(\frac{T}{T_{\max}}\right) \tag{8.46}$$

con constantes c, $T_{\max}$ a determinar. Podemos pensar a $\phi(T)$ como una ley de enfriamiento no-lineal. Para T acercándose a $T_{\max}$ el ϕ va a infinito, de manera que es de esperar que T se mantenga acotada por $|T| < T_{\max}$. Sin embargo, es de notar que el operador así definido no es Lipschitz. Para evitar este problema la "regularizamos" modificamos la ϕ de manera que a partir de cierta temperatura T_{cut} cerca de $T_{\max}$ reemplazamos la ϕ por su tangente (ver figura 8.1).

Concretamente, la función de enfriamiento utilizada será

$$\tilde{\phi}(T) = \begin{cases} \phi(T) & ; \text{ si } |T| < T_{\text{cut}} \\ \phi(T_{\text{cut}}) + \phi'(T_{\text{cut}})(T - T_{\text{cut}}) & ; \text{ si } |T| \geq T_{\text{cut}} \end{cases} \tag{8.47}$$

Asumiremos entonces que la función está regularizada y dejaremos de lado la tilde al referirnos a la misma.

- Resolver el problema de Usar Newton y ver convergencia para diferentes valores de los parámetros.
- Modificar el script `nsol.m` de forma de que use CG para resolver el subproblema lineal.
- Usar como criterio del lazo interior $\|r\| \leq \eta \|b\|$ para una serie de valores η distribuidos logarítmicamente entre 10^{-8} y 1. Graficar
  - Nro. de iteraciones en el lazo exterior

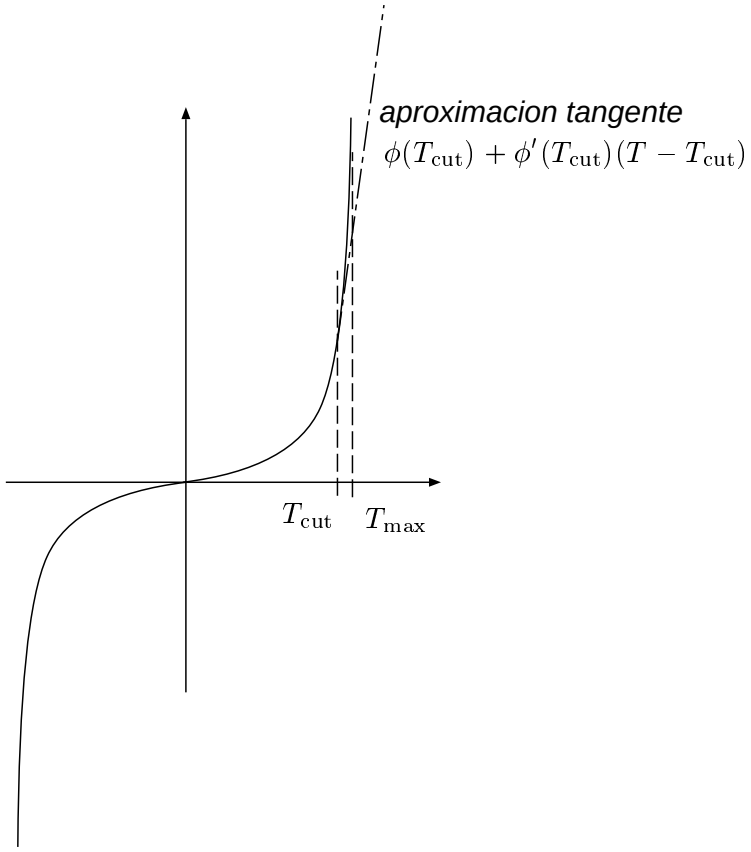

Figura 8.1: Función de enfriamiento regularizada.

- Nro. de iteraciones en el lazo interior (promedio y total).

en función de η. Estimar (si existe) el η_{opt}.

- Reemplazar el cálculo de $F'w$ por una aproximación en diferencias. Verificar el valor de h óptimo que minimiza el error.

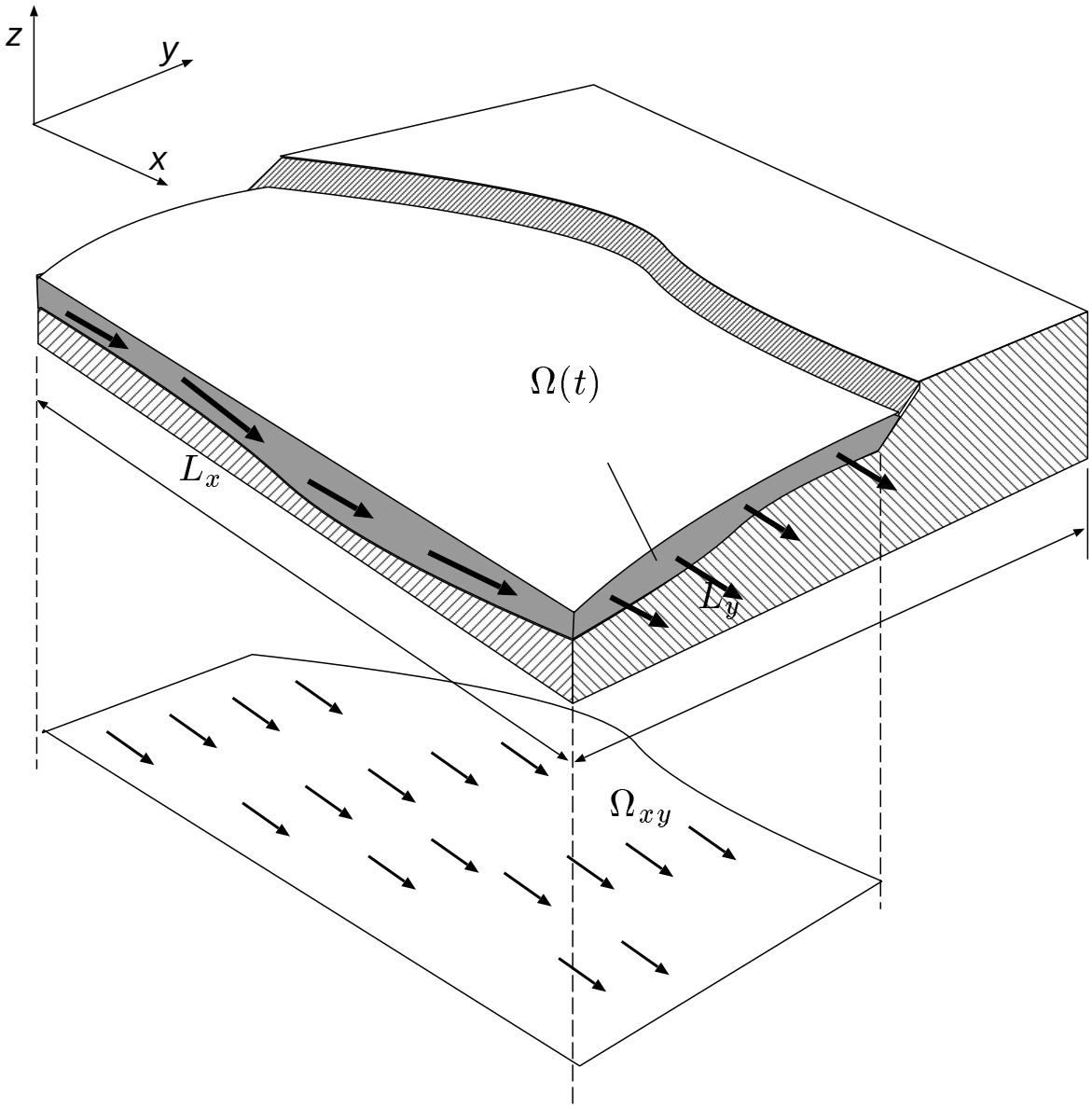

Figura 9.1: Escurrimiento sobre un fondo variable con aguas poco profundas. (3D)

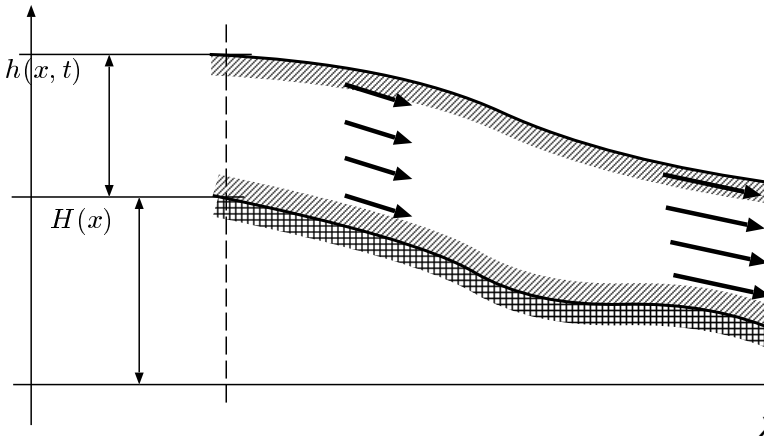

Figura 9.2: Escurrimiento sobre un fondo variable con aguas poco profundas. Versión 1D.

Capítulo 9

Las ecuaciones de shallow-water

Consideremos el escurrimiento de cursos de agua (ríos o lagunas) sobre un fondo variable (ver figuras 9.1, 9.2). La forma del fondo está dada por la altura del fondo $H(x, y)$ con respecto a un nivel de referencia fijo con respecto al nivel del mar y es un dato del problema. Se asume que el flujo es incompresible y las incógnitas son

- La posición de la superficie del agua $h(x, y, t)$.
- El campo de velocidades $\mathbf{u}(x, y, z, t)$, y presiones $p(x, y, z, t)$ en el dominio $\Omega(t)$ ocupado por el agua, es decir

$$\Omega(t) = \{(x, y, z) \quad \text{tales que } (x, y) \in \Omega_{xy} \text{ y } H(x, y) < z < H(x, y) + h(x, y, t)\} \tag{9.1}$$

El modelo más completo consiste en resolver las ecuaciones de Navier-Stokes incompresible en el dominio ocupado por el fluido. Como la posición de la superficie es desconocida a priori, se impone una ecuación adicional a saber la igualdad de presiones de ambos lados de la superficie libre. La presión del lado del aire la podemos asumir constante e igual a la presión atmosférica, que es un dato. Así puesto, el problema se vuelve intratable incluso para geometrías simples, debido a la gran cantidad de incógnitas a resolver y a la complejidad de las ecuaciones de Navier-Stokes.

El problema puede simplificarse notablemente si asumimos que la profundidad del agua h es mucho menor que las dimensiones del problema y que la longitud característica de las variaciones del fondo, es

decir que las pendientes son suaves. Asumiendo un perfil de velocidades conocido en la dirección vertical e integrando las ecuaciones en esa misma dirección, se obtienen las *"ecuaciones para flujo en aguas poco profundas"* (*"shallow water equations"*. Al integrar la ecuación en la dirección z esta desaparece y sólo quedan x, y como variables independientes. Las incógnitas son el campo de velocidades $\bar{\mathbf{u}}(x, y, t)$, donde ahora $\bar{\mathbf{u}} = (\bar{u}, \bar{v})$, que de alguna forma representa el promedio del campo de velocidades tridimensional $\mathbf{u}(x, y, z, t)$ para un instante t a lo largo de la recta vertical que pasa por un punto x, y. La variable presión desaparece y pasamos de un dominio variable $\Omega(t)$ en 3D a un dominio fijo Ω_{xy} en 2D. Estas son (en la versión simplificada 1D, es decir cuando no hay variación según la dirección transversal y) las ecuaciones de flujo en aguas poco profundas

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0 \quad \text{ec. de continuidad} \tag{9.2}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -g\frac{\partial}{\partial x}(h + H) \quad \text{ec. de balance de momento} \tag{9.3}$$

$g = 9.81\text{m/seg}^2$ es la aceleración de la gravedad. (Hemos dejado de lado también la barra en las componentes de velocidad, es decir $\bar{u} \to u$.) La ecuación de continuidad (también llamada ecuación de balance de masa) expresa la conservación de la cantidad total de agua. En el caso estacionario, implica

$$Q = hu = \text{cte} \tag{9.4}$$

donde Q es el caudal total de agua. Por otra parte, en el caso no estacionario consideremos un volumen de control comprendido entre $x - \Delta x$ y $x + \Delta x$. Reemplazando la derivada con respecto a x por una expresión en diferencias

$$2\Delta x\frac{\partial h}{\partial t} = Q_{x-\Delta x} - Q_{x+\Delta x} \tag{9.5}$$

$$\left(\begin{array}{c}\text{tasa de incremento del}\\ \text{contenido de agua}\end{array}\right) = (\text{ flujo neto de agua entrante }) \tag{9.6}$$

Con respecto a la ecuación de momento, en el estado estacionario esta expresa simplemente que

$$\frac{1}{2}u^2 + g(h + H) = E + gH = \text{cte} \tag{9.7}$$

donde E es el *"flujo de energía"* que posee el fluido. El primer término es la *"energía cinética"* y el segundo la *"energía potencial"*. La conservación de la energía proviene de no haber tenido en cuenta el rozamiento con el fondo, ni la disipación de energía por la viscosidad o turbulencia. En el caso de tener en cuenta esos efectos, deberíamos tener $(\partial E/\partial x) < 0$, asumiendo que el fluido va en la dirección de x positivo, es decir que $u > 0$.

Notar que el sistema de ecuaciones involucra sólo derivadas espaciales de primer orden de las incógnitas. Esto es una característica de los *"sistemas hiperbólicos"* (Si bien no es la única condición requerida para decir que un sistema es hiperbólico. Más sobre esto después... (sección 9.1, pág. 106)).

Si no hay rozamiento, la conservación de la energía implica que si $h + H$, es decir, la posición de la superficie libre disminuye, entonces la velocidad aumenta. Por otra parte, como $hu = Q = \text{cte}$ tenemos que

$$\frac{1}{2}u^2 + g\frac{Q}{u} + gH = \text{cte} \tag{9.8}$$

Consideremos el caso de fondo plano ($H = \text{cte}$). Sea u_1 y h_1 las variables del fluido en una cierta sección x_1. Debemos tener entonces que

$$\frac{1}{2}u^2 + g\frac{Q}{u} = \frac{1}{2}u_1^2 + g\frac{Q}{u_1} \tag{9.9}$$

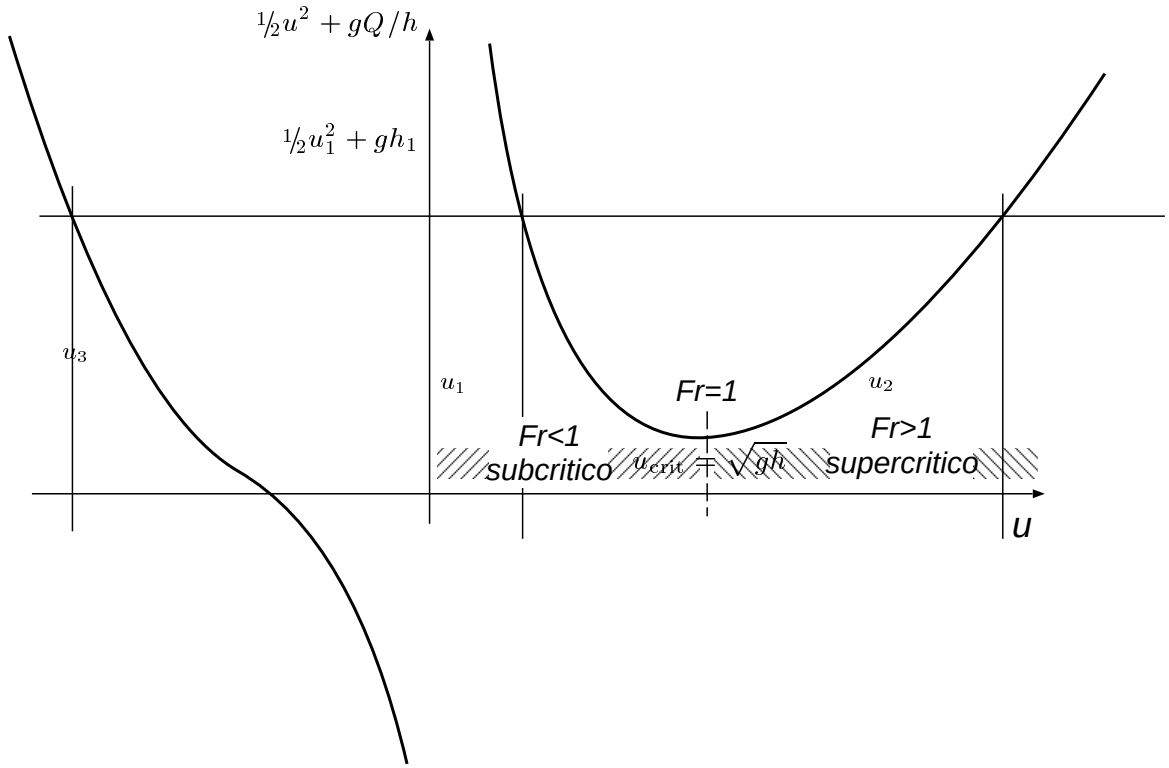

Figura 9.3: Curva de energía y diferentes tipos de flujo.

Asumiendo que $u_1 > 0$, el aspecto del miembro izquierdo de (9.8) es como se ve en la figura 9.3. Además de u_1 hay dos valores u_2 y u_3 que satisfacen (9.9). Estos valores pueden encontrarse fácilmente resolviendo una ecuación cúbica para u. Los valores correspondientes de altura h_2 y h_3 pueden sacarse del caudal $h_j = Q/u_j$. Podemos formar soluciones estacionarias tomando (h, u) constante de a trozos, tomando los valores (h_1, u_1), (h_2, u_2) y (h_3, u_3). La solución negativa podemos descartarla ya que como $Q = h_1 u_1 > 0$ eso implicaría que $h_3 = Q/u_3 < 0$ lo cual no tiene sentido físico. Podemos formar entonces soluciones constantes de a trozo con los valores (h_1, u_1), (h_2, u_2). En la figura, hemos asumido que u_1 está antes del mínimo de la curva. El mínimo de la curva se produce cuando

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{1}{2}u^2 + g\frac{Q}{u}\right) = u - g\frac{Q}{u^2} = 0 \tag{9.10}$$

o sea

$$u^2 = gQ/u = gh \tag{9.11}$$

La cantidad adimensional

$$\mathrm{Fr} = \frac{|u|}{\sqrt{gh}} \tag{9.12}$$

se llama *"número de Froude"* y por lo visto anteriormente se cumple que $\mathrm{Fr} = 1$ en el mínimo de la curva. Además

$$\mathrm{Fr} = \frac{u^{3/2}}{\sqrt{gQ}} \tag{9.13}$$

de manera que a caudal constante el Froude crece monótonamente con la velocidad y decrece monótonamente con la altura h. Por lo tanto, $\mathrm{Fr} < 1$ para velocidades menores (alturas mayores) que las críticas y $\mathrm{Fr} > 1$ para velocidades mayores (alturas menores). A estas dos condiciones de flujo se le

	$(\mathrm{d}H/\mathrm{d}x) > 0$ (desfavorable.)		$(\mathrm{d}H/\mathrm{d}x) < 0$ (favorable.)	
Cantidad	Fr < 1 (subcr.)	Fr > 1 (supercr.)	Fr < 1 (subcr.)	Fr > 1 (supercr.)
H	aumenta $\uparrow$	aumenta $\uparrow$	disminuye $\downarrow$	disminuye $\downarrow$
$E = \frac{1}{2}u^2 + gh$	disminuye $\downarrow$	disminuye $\downarrow$	aumenta $\uparrow$	aumenta $\uparrow$
u	aumenta $\uparrow$	disminuye $\downarrow$	disminuye $\downarrow$	aumenta $\uparrow$
h	disminuye $\downarrow$	aumenta $\uparrow$	aumenta $\uparrow$	disminuye $\downarrow$
$h + H$	disminuye $\downarrow$	aumenta $\uparrow$	aumenta $\uparrow$	disminuye $\downarrow$
Fr	aumenta $\uparrow$	disminuye $\downarrow$	disminuye $\downarrow$	aumenta $\uparrow$

Cuadro 9.1: Comportamiento de diferentes cantidades en pendiente favorable/desfavorable, flujo subcrítico/supercrítico

llama flujo *"subcrítico"* y *"supercrítico"*, respectivamente. Ahora bien, cuando la pendiente del fondo es creciente en flujo subcrítico se produce un decrecimiento de $E = \frac{1}{2}u^2 + gh$ por la conservación de la energía (9.7). Pero entonces, pasando a la figura 9.3, vemos que en flujo subcrítico esto implica que u debe aumentar y por lo tanto h debe disminuir por conservación de masa (ver ec. (9.4)) y $h + H$, la posición de la superficie libre, debe disminuir por (9.7). Para flujo, supercrítico, en cambio una disminución de E trae aparejado una disminución de u. Esto está reflejado en la Tabla 9. En la figura 9.4 vemos un ejemplo práctico de una presa. Antes de entrar a la presa el agua se encuentra en un estado de muy baja velocidad y gran profundidad, por lo tanto es altamente probable que se encuentre en flujo subcrítico. Al acercarse a la presa el gradiente del fondo se vuelve adverso ($(\mathrm{d}H/\mathrm{d}x) > 0$ y $u > 0$) y por lo tanto el Fr tiende a aumentar. La posición $H + h$ de la superficie libre decrece. Si en la cresta de la presa el Fr llega a un valor $\mathrm{Fr}_{\max} < 1$ entonces pasando la cima de la presa el gradiente se hace favorable ($(\mathrm{d}H/\mathrm{d}x) < 0$, $u > 0$) y el Fr vuelve a bajar, de manera que el flujo es subcrítico en todo el recorrido. Notar que, si el perfil de la presa es simétrico, entonces el perfil de todas las cantidades es simétrico, ya que sólo dependen de la profundidad. Por otra parte, si el Fr llega a uno en la cresta, entonces pasando la misma pueden ocurrir dos situaciones. Si el flujo se vuelve subcrítico, entonces estamos en una situación similar a la anterior. Si el flujo se hacer supercrítico, entonces al disminuir el H el Fr aumentará, adelgazándose la profundidad del agua (h disminuye) como puede verse en la figura 9.4. Finalmente, el flujo pasa de supercrítico a subcrítico a través de un resalto hidráulico. Si bien, esto no es contemplado por la teoría, tal como lo hemos descripto aquí, los resaltos hidráulicos van acompañados de una gran disipación de energía en forma de turbulencia. Cual de los dos casos ocurre en realidad depende de las condiciones de contorno del problema. Las condiciones de contorno en sistemas hiperbólicos es un tema delicado, por ahora sólo podemos decir que como es un sistema de primer orden en las derivadas espaciales, sólo podemos imponer un número de m condiciones de contorno, donde m es el número de campos incógnita (en este caso $m = 2$ ya que las incógnitas son h y u). Por ahora pensamos a las condiciones de contorno como una especie de *"exclusas"* que están ubicadas aguas arriba y aguas abajo de la presa. Dejamos la exclusa de aguas arriba abierta y vamos abriendo la exclusa de aguas abajo. Inicialmente al estar la exclusa cerrada el flujo es en general lento y subcrítico en todo el dominio, como en la situación de la figura 9.4, a medida que vamos abriendo la exclusa de aguas abajo el flujo se empieza acelerar hasta llegar a la situación de la figura 9.5. Esto lo veremos en los ejemplos numéricos.

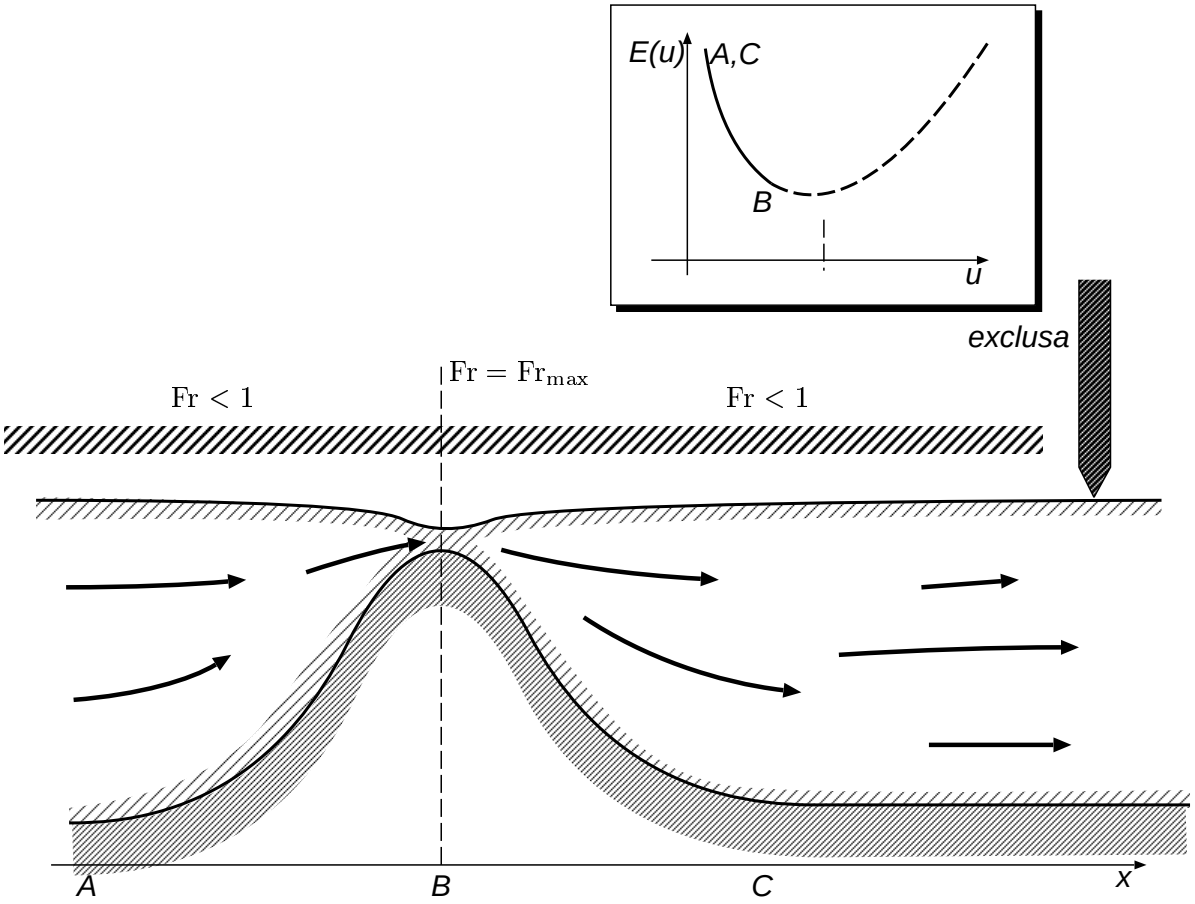

Figura 9.4: Comportamiento de diferentes cantidades en pendiente favorable/desfavorable para un obstáculo en el fondo, en flujo subcrítico.

9.1. Análisis temporal

Hasta ahora hemos hecho un análisis puramente estacionario del problema, pero para entenderlo mejor es imprescindible hacer un análisis temporal. El sistema (9.2,9.3) se puede poner en la forma general de un sistema hiperbólico

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x} F(U) = S_{\mathrm{w}} \tag{9.14}$$

donde

$$U = \begin{bmatrix} h \\ u \end{bmatrix} \tag{9.15}$$

$$F(U) = \begin{bmatrix} hu \\ \frac{1}{2}u^2 + gh \end{bmatrix} \tag{9.16}$$

$$S = \begin{bmatrix} 0 \\ -g\,(\mathrm{d}H/\mathrm{d}x) \end{bmatrix} \tag{9.17}$$

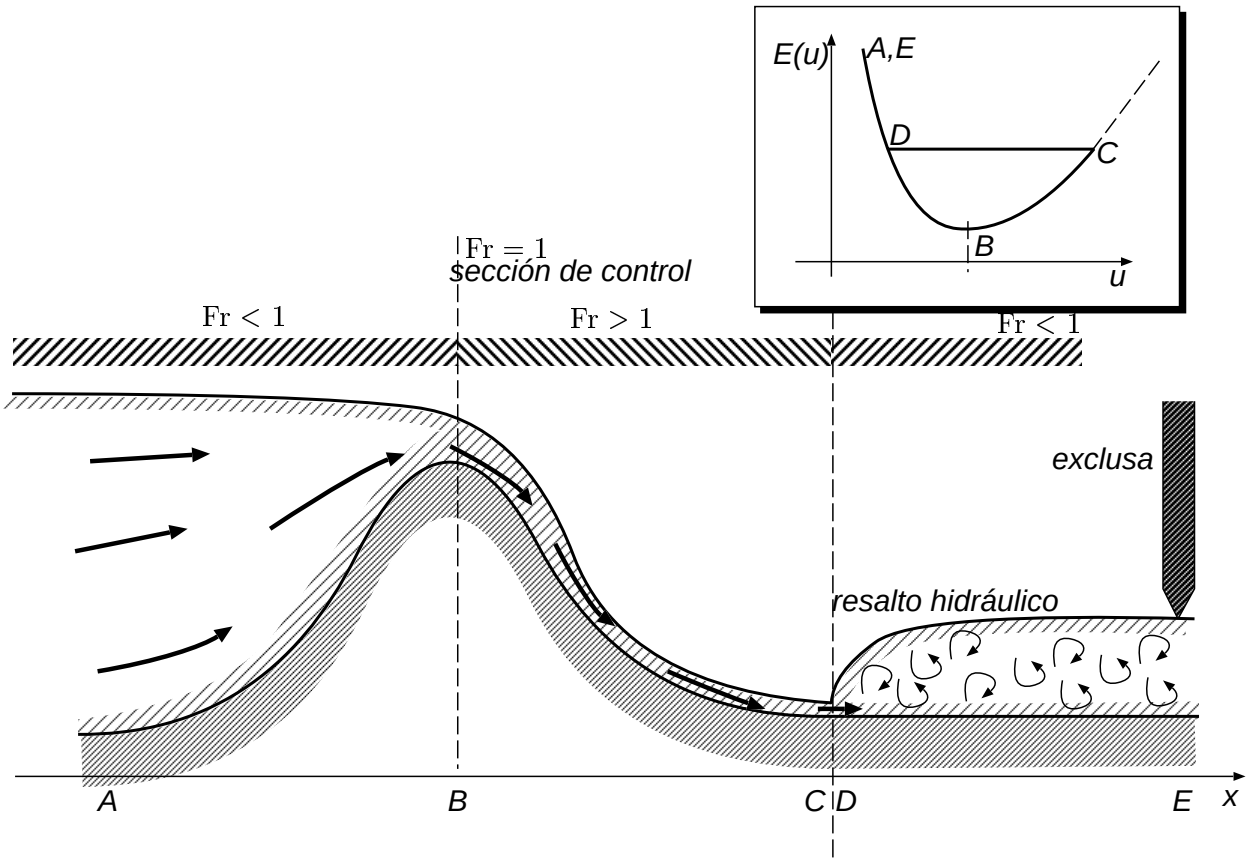

Figura 9.5: Comportamiento de diferentes cantidades en pendiente favorable/desfavorable, flujo subcrítico/supercrítico, para una presa.

son los vectores *"de estado"*, *"de flujos"* y *"término fuente"*. Si hacemos una discretización espacial y temporal (Euler explícita) del problema y

$$\mathbf{U}^n = \begin{bmatrix} h_1^n \\ u_1^n \\ h_2^n \\ u_2^n \\ \vdots \\ h_N^n \\ u_N^n \end{bmatrix} \tag{9.18}$$

es el vector de estado *"global"* al tiempo n, h_j^n es la aproximación al valor de $h(x_j, t^n)$ y similar para u_j^n. Entonces tendremos una serie de ecuaciones

$$\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} = \mathbf{G}(\mathbf{U}^n) \tag{9.19}$$

Despejando U^{n+1} obtenemos un método de punto fijo o de Richardson no-lineal para encontrar $\mathbf{U}$ solución del problema estacionario. Lo bueno de este método es que, de alguna forma es más probable que de convergencia global, ya que (al menos en el límite $\Delta t \to 0$) sigue la evolución real del sistema, el

cual, a menos que el sistema sea inestable o no disipativo, debería converger a cierto estado estacionario. (Ya veremos más adelante como combinar este esquema con Newton.) Bajo esta óptica, la integración temporal es un esquema iterativo más para resolver el sistema de ecuaciones que lleva al estado estacionario.

Primero estudiaremos las propiedades de los sistemas hiperbólicos en el dominio temporal. Además, asumiremos una linealización del problema. Aplicando la regla de la cadena al término advectivo en (9.14), obtenemos

$$\frac{\partial U}{\partial t} + A(U)\frac{\partial U}{\partial x} = S_{\mathrm{w}} \tag{9.20}$$

donde

$$A(U) = \frac{\partial F}{\partial U} \tag{9.21}$$

es el *"jacobiano de los flujo advectivos"*. La hipótesis de linealidad consiste en asumir que el flujo consiste de pequeñas perturbaciones a un flujo constante $U = U_0 = \mathrm{cte} \neq U(x)$, y entonces podemos reemplazar los jacobianos por $A(U_0)$. Para simplificar aún más, asumiremos que tenemos un sólo campo incógnita, que es decir $m = 1$, y que no hay término fuente ($S_{\mathrm{w}} \equiv 0$), en ese caso la ecuación típica es

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{9.22}$$

con $a = \mathrm{cte}$. Las soluciones a (9.22) son de la forma

$$u(x,t) = f(x - at) \tag{9.23}$$

lo cual quiere decir que los valores de u son constantes sobre las *"rectas características"* $x - at = \mathrm{cte}$. a tiene dimensiones de velocidad y representa entonces la velocidad con la cual se propaga u.

En el caso de más campos incógnita $m > 1$, supongamos que el jacobiano A puede diagonalizarse, es decir

$$Av_j = \lambda_j v_j \tag{9.24}$$

para un conjunto de m autovectores linealmente independientes v_j. Entonces formando la matriz de cambio de base

$$S = \begin{bmatrix} v_1 & v_2 & \dots & v_m \end{bmatrix} \tag{9.25}$$

y definiendo

$$\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \dots, \lambda_m\} \tag{9.26}$$

entonces (9.24) es equivalente a

$$AS = S\Lambda \tag{9.27}$$

y haciendo el cambio de variables $U \to V$ definido por

$$U = SV \tag{9.28}$$

la ecuación linealizada, multidimensional, sin término fuente

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0 \tag{9.29}$$

queda como

$$S\frac{\partial U}{\partial t} + AS\frac{\partial U}{\partial x} = 0 \tag{9.30}$$

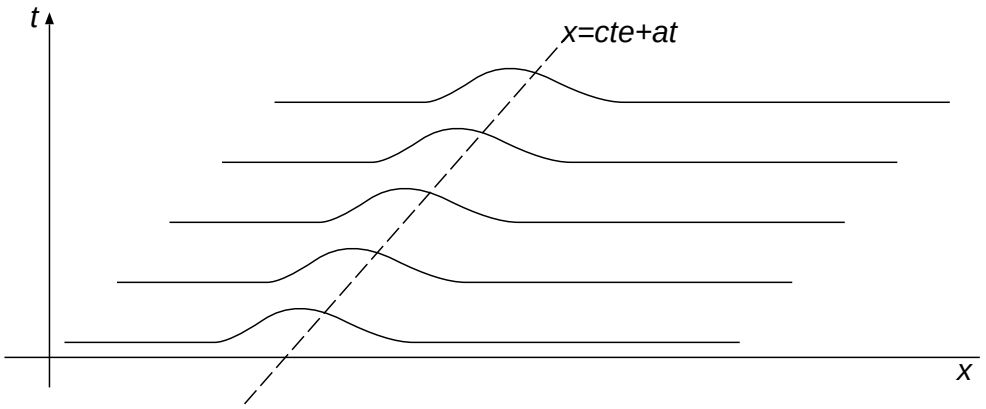

Figura 9.6: Propagación de una onda para la ec. de advección escalar.

y multiplicando la ecuación por S^{-1}

$$\frac{\partial U}{\partial t} + (S^{-1}AS)\frac{\partial U}{\partial x} = 0 \tag{9.31}$$

$$\implies \quad \frac{\partial U}{\partial t} + \Lambda\frac{\partial U}{\partial x} = 0 \tag{9.32}$$

pero como Λ es diagonal, esto se desacopla en una serie de m ecuaciones de advección unidimensionales iguales a (9.22)

$$\frac{\partial v_j}{\partial t} + \lambda_j\frac{\partial v_j}{\partial x} = 0 \tag{9.33}$$

donde v_j es la componente j de V. Como vimos anteriormente, esto significa que

$$v_j = f_j(x - \lambda_j t) \tag{9.34}$$

y por lo tanto, los λ_j son las velocidades características del sistema. Nótese que todo esto es válido si A es diagonalizable y si los autovalores de A son reales. Justamente esta es la definición de sistema hiperbólico, *que los jacobianos advectivos sean diagonalizables y con autovalores reales para cualquier estado U admisible.*

El jacobiano para las ecuaciones de shallow water es

$$A = \begin{bmatrix} (\partial F_1/\partial h) & (\partial F_1/\partial u) \\ (\partial F_2/\partial h) & (\partial F_2/\partial u) \end{bmatrix} = \begin{bmatrix} u & h \\ g & u \end{bmatrix} \tag{9.35}$$

Calculando los autovalores

$$\det(A - \lambda I) = \det \begin{bmatrix} u-\lambda & h \\ g & u-\lambda \end{bmatrix} = (u-\lambda)^2 - gh = 0 \tag{9.36}$$

de donde

$$\lambda_\pm = u \pm \sqrt{gh} \tag{9.37}$$

De manera que las ecuaciones de shallow water son un sistema hiperbólico.

En agua en reposo ($u, \mathrm{Fr} = 0$), las dos son iguales y de sentido contrario $\pm\sqrt{gh}$. Por lo tanto si hacemos una perturbación en altura, veremos dos ondas propagándose hacia x positivos y negativos con velocidad $\pm\sqrt{gh}$ (ver figura 9.7). Manteniendo la profundidad constante y aumentando la velocidad

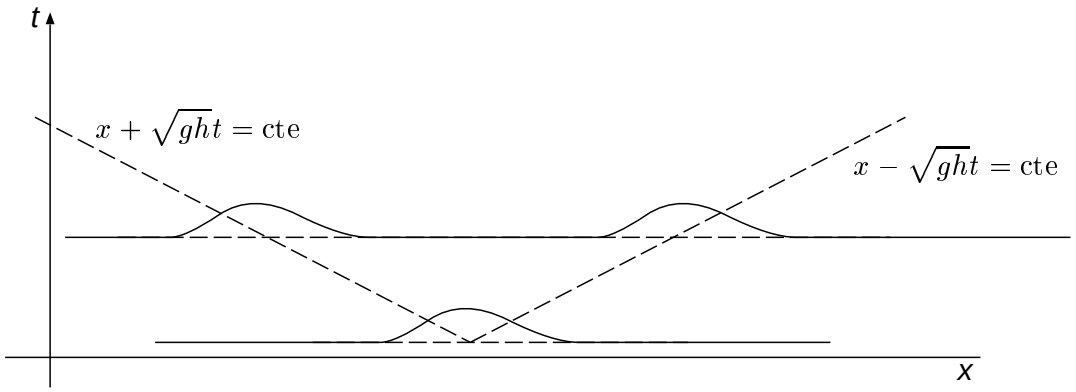

Figura 9.7: Propagación de ondas de altura para la ec. de advección para las ec. shallow water para agua en reposo $u = 0$.

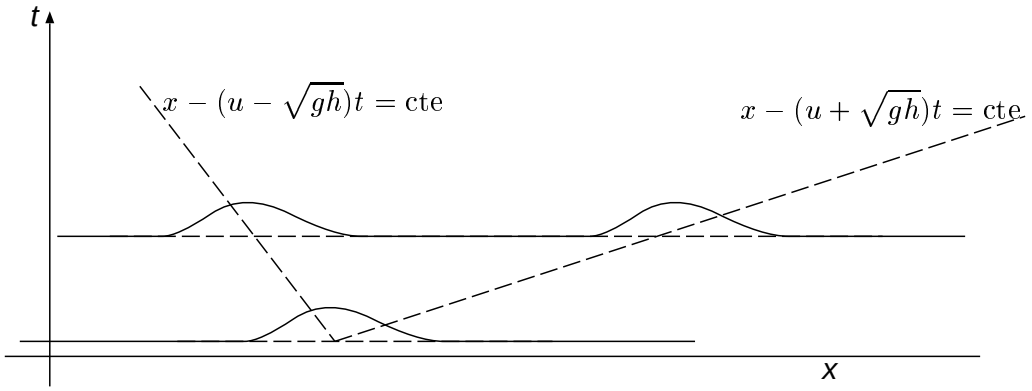

Figura 9.8: Propagación de ondas de altura para la ec. de advección para las ec. shallow water para agua en movimiento, flujo subcrítico Fr < 1.

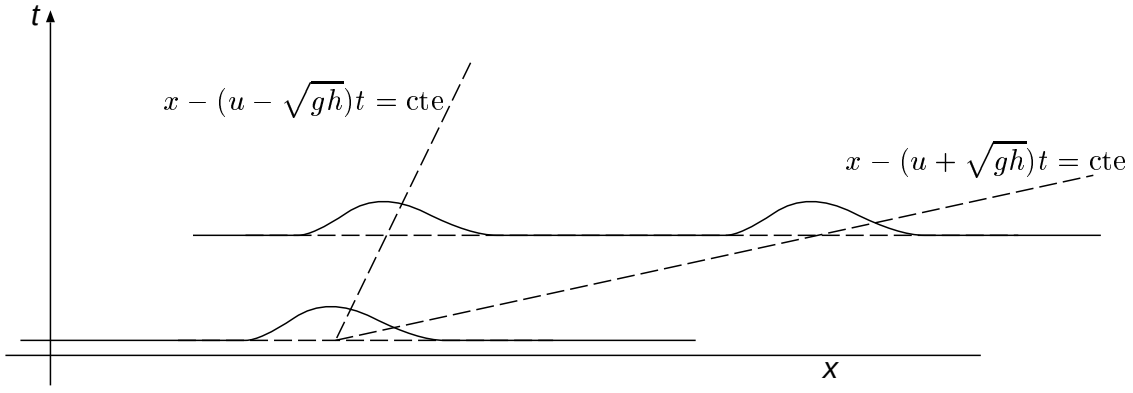

Figura 9.9: Propagación de ondas de altura para la ec. de advección para las ec. shallow water para agua en movimiento, flujo supercrítico Fr > 1.

a $u > 0$ es equivalente a dejar el agua quieta y describir el fenómeno desde un sistema que se mueve con respecto al agua con velocidad $-u < 0$. Por lo tanto vemos a las ondas propagarse también hacia x positivos y negativos con velocidad $u \pm \sqrt{gh}$ (ver figura 9.8). Cuando la velocidad del fluido llega a la crítica $u = \sqrt{gh}$ la ola que va hacia x negativos se mantiene estacionaria con respecto al observador. Para velocidades mayores las dos velocidades son positivas y por lo tanto ambas ondas se propagan en la misma dirección. Esto es muy importante ya que implica que el flujo supercrítico no es afectado aguas arriba por una perturbación producida aguas abajo.

9.2. Detalles de discretización

Consideremos la discretización en una grilla de paso constante

$$x_i = i\Delta x, \quad i = 0, N \tag{9.38}$$

con $\Delta x = L/N$ el paso de la malla. Recordemos (ver sección §3.11) que para la ecuación de advección (3.73) debe agregarse una difusión numérica k_{num}. En el caso nuestro no tenemos difusión física, es decir $k = 0$ y por lo tanto $\text{Pe}_{\Delta x} \to \pm\infty$ dependiendo del signo de a. Puede verificarse que el límite corresponde a

$$k_{\text{num}} \to \frac{|a|\Delta x}{2}, \quad \text{para } |\text{Pe}_{\Delta x}| \to \infty \tag{9.39}$$

De manera que el esquema apropiado para (9.14) (consideramos término fuente nulo) es

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{F_{j+1}^n - F_{j-1}^n}{2\Delta x} = \left(\frac{|A|\Delta x}{2}\right) \frac{(U_{j+1} - 2U_j + U_{j-1})^n}{\Delta x^2} \tag{9.40}$$

Primero que todo, debemos especificar que aquí $|A|$ debe interpretarse como valor absoluto de A en el sentido matricial, es decir si A se descompone como en (9.27) entonces

$$|A| = S\,\text{diag}\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_m|\}S^{-1} \tag{9.41}$$

Asumiendo por ahora que A es constante, (9.40) puede ponerse en la forma

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{(F^*)_{j+1/2}^n - (F^*)_{j-1/2}^n}{\Delta x} = 0 \tag{9.42}$$

donde

$$(F^*)_{j+1/2}^n = \frac{F(U_{j+1}^n) + F(U_j^n)}{2} - |A|\,\frac{U_{j+1} - U_j}{2} \tag{9.43}$$

Lo importante de la formulación (9.42) es que, en el estado estacionario se cumple

$$(F^*)_{j+1/2} = (F^*)_{j-1/2} = (F^*)_{j-3/2} = \ldots \tag{9.44}$$

(dejamos de lado el supraíndice n ya que estamos en el estado estacionario) y esto fuerza la conservación de F a través de una variación abrupta como es un resalto hidráulico, como veremos a continuación. Tomemos $(F^*)_{j+1/2}^n$ de tal forma que dependa sólo de los valores $U_{j,j+1}^n$, por ejemplo

$$(F^*)_{j+1/2}^n = \frac{F(U_{j+1}^n) + F(U_j^n)}{2} - |A(U_{j+1/2}^n)|\,\frac{U_{j+1} - U_j}{2} \tag{9.45}$$

donde $U_{j+1/2}^n = 1/2(U_j^n + U_{j+1}^n)$. Entonces si, en el estado estacionario, hay una cierta variación grande de U en una región $|x| < \delta$ y después

$$U_j \to U_{L,R}, \quad \text{para} j \to \pm\infty \tag{9.46}$$

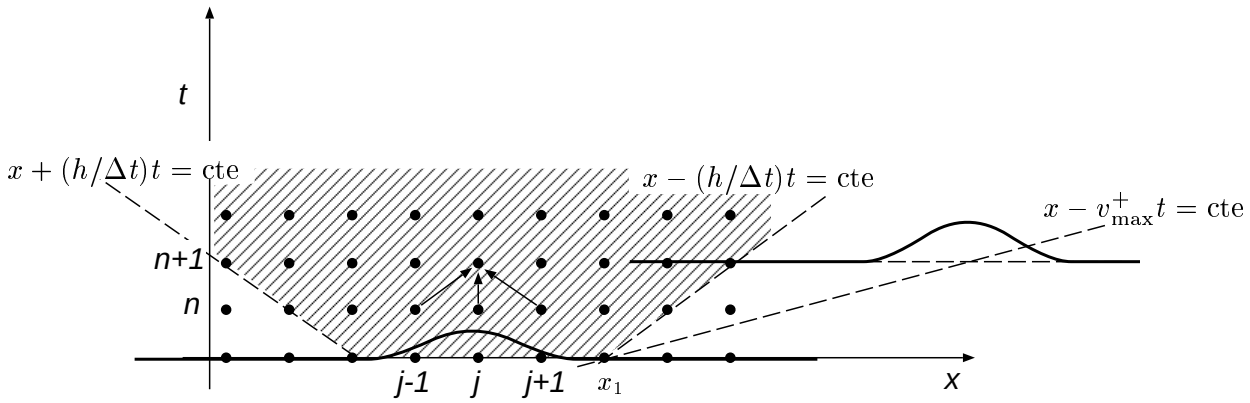

Figura 9.10: Dependencia de los datos en el esquema temporal explícito

entonces como los flujos F^* se conservan, tendremos que $F^*_L = F^*_R$. Pero como para $j \to \infty$ tenemos $U_j \to U_R$, tenemos que, por (9.45), el término difusivo (el que contiene a $|A|$) desaparece y por lo tanto $F_j \to F^*_j \to F^*_R$. Análogamente $F_j \to F^*_j \to F^*_L$ para $j \to -\infty$ y por lo tanto F se conserva a través del resalto. Esto quiere decir que los flujos F se conservan, por eso se llaman *"esquemas conservativos"*. Básicamente, estos esquemas aseguran que no habrá pérdida de masa ni momento ni otras cantidades físicas, incluso en el caso de que haya grandes variaciones de las variables, como ocurre en el resalto hidráulico.

9.3. Integración temporal. Paso de tiempo crítico

Consideremos ahora el esquema de integración temporal (9.19). Poniéndolo de la forma

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t \mathbf{G}(\mathbf{U}^n) \tag{9.47}$$

vemos que, como ya mencionáramos, es equivalente a un Richardson no-lineal, donde Δt juega el papel de una parámetro de relajación. Igual que para los sistemas lineales, cuanto más alto podamos poner el parámetro de relajación más rápido convergeremos al estacionario, pero subiéndolo demasiado corremos el riesgo de pasarnos del límite de estabilidad y que el esquema diverja. Necesitamos entonces, una estimación del *"paso de tiempo crítico"* Δt_{crit} para el cual el esquema se vuelve inestable. Para sistemas hiperbólicos, el Δt_{crit} puede estimarse considerando la velocidad de propagación física de las ondas y la velocidad de propagación *"inherente"* al esquema.

Efectivamente, consideremos la ecuación de evolución discreta para U^{n+1}_j (9.42). Esta ecuación nos dice que U^{n+1}_j depende sólo de U^n_j y de los flujos $(F^*)^n_{j\pm1/2}$. A su vez, estos dos últimos dependen sólo de U^n_k para $k = j-1, j, j+1$, (ver figura 9.10). Decimos que *el esquema explícito avanza un paso de malla Δx por paso de tiempo Δt*. Esto quiere decir que si a t^n tenemos una cierta perturbación en una región $x_0 < x < x_1$, por esta dependencia de los datos del esquema numérico, la perturbación puede haber llegado, en el paso de tiempo t^m a lo sumo a la región comprendida entre

$$x_0 - (m-n)\Delta x < x < x_1 + (m-n)\Delta x \tag{9.48}$$

Eso quiere decir que la parte más avanzada/retrasada de la onda se mueve sobre una recta $x \pm (\Delta x/\Delta t)t = \text{cte}$. Por otra parte, la perturbación inicialmente contenida entre $x_0 < x < x_1$ se debería (desde el punto de vista de la física del problema, es decir de la mecánica del continuo) descomponer

en una serie de componentes modales que se propagan cada una con la velocidad λ_j (ver ec. (9.34)). O sea que la perturbación debería ocupar la región

$$x_0 - v^-_{\max}(m-n)\Delta t < x < x_1 + v^+_{\max}(m-n)\Delta t \tag{9.49}$$

donde

$$v^-_{\max} = \max_{\substack{j=1,\ldots,m \\ \lambda_j<0}} |\lambda_j| \tag{9.50}$$

$$v^+_{\max} = \max_{\substack{j=1,\ldots,m \\ \lambda_j>0}} |\lambda_j| \tag{9.51}$$

son la máxima velocidad hacia x positivos y negativos, respectivamente. Si $v^+_{\max} > \Delta x/\Delta t$ (como se ve en la figura) entonces después de transcurrido un cierto tiempo la posición donde debería estar la onda queda fuera de la región donde la información se puede haber propagado por las restricciones del esquema. Obviamente bajo estas condiciones, el esquema numérico no puede seguir la física del problema y lo que ocurre en la práctica es que el esquema numérico se hace inestable. Tenemos entonces que la restricción para que esto no ocurra es $v^+_{\max} < \Delta x/\Delta t$. Lo mismo ocurre para las ondas que se propagan hacia la derecha, de manera que la restricción combinada es la llamada *"condición de Courant-Friedrichs-Lewy"* o también *"condición CFL"* que dice que para que el esquema sea estable debe cumplirse que

$$\text{Co} = \frac{v_{\max}\Delta t}{\Delta x} < 1 \tag{9.52}$$

donde

$$v_{\max} = \max\{v^-_{\max}, v^+_{\max}\} = \max_{j=1,\ldots,m} |\lambda_j| \tag{9.53}$$

y Co es el *"número de Courant"*. En el caso de las ecuaciones de shallow water las velocidades de propagación están dadas por (9.37), y

$$v_{\max} = |u| + \sqrt{gh} \tag{9.54}$$

9.4. Guía Nro. 5. Ecuaciones de shallow water.

Ejer. 1: Propagación de ondas alrededor de flujo homogéneo ($H \equiv \text{cte}$).

a.) Considerando condiciones iniciales

$$\begin{cases} h &= h_0 + 0.1\, \mathrm{e}^{-(x-x_0)^2/\sigma^2} \\ u &= u_0 \end{cases} \tag{9.55}$$

Con

$$L = 5 \tag{9.56}$$
$$\sigma = 0.7 \tag{9.57}$$
$$x_0 = L/2 \tag{9.58}$$
$$h_0 = 1 \tag{9.59}$$
$$u_0 = 0.2 \tag{9.60}$$

Ver como se propagan las ondas hasta llegar al estacionario. Evaluar la velocidad de propagación de cada modo y comparar con la velocidad teórica. Ídem para $u_0 = 1.5$.

b.) Condiciones absorbentes: Considerar la condición de contorno

$$\begin{cases} u(x=0,t) &= u_0 = 0.2 \\ h(x=L,t) &= h_0 \end{cases} \tag{9.61}$$

Hay reflexión de las ondas en los contornos? Comparar con la condición absorbente utilizada. Como afecta la convergencia hacia el estado estacionario? Repetir lo mismo para

$$\begin{cases} u(x=0,t) &= u_0 \\ h(x=0,t) &= h_0 \\ u(x=L,t) &= 0.3 \\ h(x=L,t) &= h_0 \end{cases} \tag{9.62}$$

Ejer. 2. Flujo subcrítico con fondo variable Tomar $H = 0.2\, \mathrm{e}^{-(x-x_0)^2/\sigma^2}$. En este y siguientes ejercicios inicializaremos un valor constante U_0 en la mitad izquierda y U_L en la derecha, es decir con

$$U(x,0) = \begin{cases} U_0 & 0 < x < L/2 \\ U_L & L/2 < x < L \end{cases} \tag{9.63}$$

Tomar $U_0 = U_L = \begin{bmatrix} 1 \\ 0.3 \end{bmatrix}$. Se forma el perfil subcrítico? Es simétrico? Como varía $\mathrm{Fr}(x)$?

Repetir el experimento para $U_L = \begin{bmatrix} h^* \\ 0.3 \end{bmatrix}$ con $h^* = 0.9, 0.8, 0.7 \ldots$. Ir graficando $\mathrm{Fr}(x)$. Se forma un resalto? Donde se produce el cambio de flujo subcrítico a supercrítico y viceversa?

Ejer. 3. Ondas de choque y abanicos de expansión. Consideremos fondo plano. Sea $U^- = \begin{bmatrix}1 & 0.3\end{bmatrix}'$ y U^+ el estado supercrítico correspondiente, es decir tal que $F(U^+) = F(U^-)$ (usar la función `supercr.m`). Probar a inicializar con

- Subcrítico/supercrítico: $U_0 = U^-$, $U_L = U^+$
- Supercrítico/subcrítico: $U_0 = U^+$, $U_L = U^-$
- Subcrítico/supercrítico con transición suave:

$$U(x) = \frac{U^- + U^+}{2} - \frac{U^- - U^+}{2} \tanh \frac{x - x_0}{\sigma} \tag{9.64}$$

- Supercrítico/subcrítico con transición suave:

$$U(x) = \frac{U^- + U^+}{2} + \frac{U^- - U^+}{2} \tanh \frac{x - x_0}{\sigma} \tag{9.65}$$

Se forman discontinuidades? Depende esto de si el perfil inicial es suave o abrupto?

Ejer. 4. Resaltos no estacionarios Sean $h^\pm, u^\pm$ las componentes de $U^\pm$.

Inicializar con $U_0 = \begin{bmatrix}u^+ + 0.1, & h^+\end{bmatrix}'$, $U_L = \begin{bmatrix}u^- + 0.1, & h^-\end{bmatrix}'$. Se forma un resalto no-estacionario? Porqué?

Probar con $U_0 = \begin{bmatrix}u^+ - 0.1, & h^+\end{bmatrix}'$, $U_L = \begin{bmatrix}u^- - 0.1, & h^-\end{bmatrix}'$.

Capítulo 10

Las ecuaciones de shallow-water 2D

La extensión de las ecuaciones (9.2-9.3) a 2D es la siguiente

$$\frac{\partial h}{\partial t} + \nabla(h\mathbf{u}) = 0 \tag{10.1}$$

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot)\mathbf{u} = -g\nabla(h+H) \tag{10.2}$$

donde $\mathbf{u} = (u, v)$ es el vector velocidad ∇ el operador vector gradiente y $\nabla\cdot$ el operador divergencia. Si bien esta notación es muy compacta, a veces puede ser un poco confusa en cuanto a cuales índices están contraídos. La siguiente es la *"forma tensorial"*

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x_j}(hu_j) = 0 \tag{10.3}$$

$$\frac{\partial u_i}{\partial t} + u_j\frac{\partial u_i}{\partial x_j} = -g\frac{\partial}{\partial x_i}(h+H) \tag{10.4}$$

donde hemos usado la *"convención de Einstein"* de suma implícita sobre índices repetidos. Hemos denotado $(u_1, u_2) = (u, v)$.

10.1. Forma conservativa

Para llevar estas ecuaciones a la forma de conservación debemos proceder en forma un tanto diferente a la usada en el caso 1D. Multiplicando (10.3) por u_i, (10.4) por h y sumando ambas ecuaciones llegamos a

$$\frac{\partial hu_i}{\partial t} + \frac{\partial}{\partial x_j}(hu_iu_j) + g\frac{1}{2}\frac{\partial h^2}{\partial x_i} = -gh\frac{\partial H}{\partial x_i} \tag{10.5}$$

que puede ponerse como

$$\frac{\partial hu_i}{\partial t} + \frac{\partial}{\partial x_j}(hu_iu_j + g\delta_{ij}\frac{1}{2}h^2) = -gh\frac{\partial H}{\partial x_i} \tag{10.6}$$

donde la *"delta de Kronecker"* está definida como

$$\delta_{ij} = \begin{cases} 1 & ;\ \text{si } i = j \\ 0 & ;\ \text{si } i \neq j \end{cases} \tag{10.7}$$

Podemos poner las ecuaciones en forma conservativa (es decir como en (9.14) poniendo

$$U = \begin{bmatrix} h \\ hu \\ hv \end{bmatrix} = \begin{bmatrix} h \\ h\mathbf{u} \end{bmatrix}, \quad F_x = \begin{bmatrix} hu \\ hu^2 + gh^2/2 \\ huv \end{bmatrix}, \quad F_y = \begin{bmatrix} hv \\ huv \\ hv^2 + gh^2/2 \end{bmatrix}, \tag{10.8}$$

$$S = \begin{bmatrix} 0 \\ -gh\,(\partial H/\partial x) \\ -gh\,(\partial H/\partial y) \end{bmatrix} = \begin{bmatrix} 0 \\ -gh\nabla H \end{bmatrix} \tag{10.9}$$

y entonces

$$\frac{\partial U}{\partial t} + \frac{\partial F_j}{\partial x_j} = S \tag{10.10}$$

Los flujos advectivos F_i pueden ponerse en forma más compacta como

$$a_j F_j = \begin{bmatrix} h\,u_j a_j \\ h\,(a_j u_j)\,\mathbf{u} + (gh^2/2)\,\mathbf{a} \end{bmatrix} \tag{10.11}$$

donde $\mathbf{a}$ es cualquier vector de $\mathbb{R}^2$.

10.2. Linealización de las ecuaciones

Como en el caso 1D (9.20) estudiaremos la forma en que se propagan perturbaciones a un flujo homogéneo y sin término fuente, es decir una linealización del problema. La ec. (10.10) puede ponerse como

$$\frac{\partial U}{\partial t} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} = 0 \tag{10.12}$$

$$\frac{\partial U}{\partial t} + A_x \frac{\partial U}{\partial x} + A_y \frac{\partial U}{\partial y} = 0 \tag{10.13}$$

O en forma tensorial

$$\frac{\partial U}{\partial t} + A_j \frac{\partial U}{\partial x_j} = 0 \tag{10.14}$$

Asumiremos que los A_j son constantes y que la U es en realidad una pequeña perturbación con respecto al flujo medio.

10.3. Velocidad de propagación. Proyección unidimensional.

Cuál es ahora la velocidad de propagación de las perturbaciones? Primeramente podemos ver que depende de la dirección. Propongamos soluciones de la forma

$$U(\mathbf{x}, t) = \bar{U}(\mathbf{x} \cdot \hat{\mathbf{n}}, t) = \mathbf{U}(\xi, t) \tag{10.15}$$

es decir, donde U sólo varía en una dirección dada por el versor unitario $\hat{\mathbf{n}}$. ξ es una coordenada según esa dirección. Reemplazando esta forma para U en (10.14) llegamos a

$$\frac{\partial \bar{U}}{\partial t} + A_x \frac{\partial \bar{U}}{\partial \xi} n_x + A_y \frac{\partial \bar{U}}{\partial \xi} n_y = 0 \tag{10.16}$$

$$\frac{\partial \bar{U}}{\partial t} + (n_x A_x + n_y A_y) \frac{\partial \bar{U}}{\partial \xi} = 0 \tag{10.17}$$

$$\tag{10.18}$$

Es decir que el sistema se comporta como un sistema advectivo 1D de la forma (9.20) donde el jacobiano está dado por la proyección del vector de jacobianos A_j según la dirección $\hat{\mathbf{n}}$

$$A_n = n_j A_j \tag{10.19}$$

Como habíamos visto, la velocidad de propagación de las ondas estaba dada por los autovalores de este jacobiano. Primero calculemos explícitamente la proyección del jacobiano

$$A_n = n_j A_j = n_j \frac{\partial F_j}{\partial U} = \frac{\partial}{\partial U}(n_j F_j) = \frac{\partial F_n}{\partial U} \tag{10.20}$$

donde F_n sale de (10.11). Pero para calcular los jacobianos debemos poner F_n en términos de las componentes de U. Llamemos $w_j = hu_j$, entonces (10.11) puede ponerse como

$$F_n = \begin{bmatrix} (w_j n_j) \\ (w_j n_j)\, w_k/h + gh^2/2\, n_k \end{bmatrix} \tag{10.21}$$

Aquí en la segunda fila k es un índice que vale $k = 1$ para calcular la componente 2 de F_n y $k = 2$ para la tercera componente. Es decir

$$F_n = \begin{bmatrix} (w_j n_j) \\ (w_j n_j)\, w_1/h + gh^2/2\, n_1 \\ (w_j n_j)\, w_2/h + gh^2/2\, n_2 \end{bmatrix} \tag{10.22}$$

Esto hace hincapié en el carácter vectorial de las dos últimas componentes de F_n.

Ahora bien,

$$A_n = \begin{bmatrix} (\partial F_n/\partial h) & (\partial F_n/\partial w_l) \end{bmatrix} \tag{10.23}$$

Aquí l juega un rol similar al de la ecuación previa. Realizando las derivadas, obtenemos que

$$A_n = \begin{bmatrix} 0 & n_l \\ -(n_j w_j)\, w_k/h^2 + ghn_k & [(n_j w_j)\delta_{kl} + w_k n_l]/h \end{bmatrix} \tag{10.24}$$

$$= \begin{bmatrix} 0 & n_l \\ -(n_j u_j)\, u_k + ghn_k & (n_j u_j)\delta_{kl} + u_k n_l \end{bmatrix} \tag{10.25}$$

Poniendo $\hat{\mathbf{n}} = (1, 0)$ podemos obtener el jacobiano A_x y para $\hat{\mathbf{n}} = (0, 1)$ obtenemos A_y

$$A_x = \begin{bmatrix} 0 & 1 & 0 \\ -u^2 + gh & 2u & 0 \\ -uv & v & u \end{bmatrix}, \quad A_y = \begin{bmatrix} 0 & 0 & 1 \\ -vu & v & u \\ -v^2 + gh & 0 & 2v \end{bmatrix} \tag{10.26}$$

Calculemos sus autovalores

$$\det(A_x - \lambda I) = \det \begin{bmatrix} -\lambda & 1 & 0 \\ -u^2 + gh & 2u - \lambda & 0 \\ -uv & v & u - \lambda \end{bmatrix} \tag{10.27}$$

$$= (u - \lambda)\left[-\lambda(2u - \lambda) - (-u^2 + gh)\right] \tag{10.28}$$

$$= (u - \lambda)\,(\lambda^2 - 2u\lambda + u^2 - gh) \tag{10.29}$$

$$= (u - \lambda)\left[(\lambda - u)^2 - gh\right] \tag{10.30}$$

de donde sale

$$\lambda_1 = u \tag{10.31}$$

$$\lambda_{2,3} = u \pm \sqrt{gh} \tag{10.32}$$

Notar que los dos últimos autovalores $\lambda_{2,3}$ coinciden con los del problema puramente unidimensional (9.37). El primer autovalor corresponde a perturbaciones de la componente y de velocidad, siempre según la dirección x.

Análogamente, para si asumimos sólo variación según la dirección y,

$$\begin{aligned}
\det(A_x - \lambda I) &= \det \begin{bmatrix} -\lambda & 0 & 1 \\ -vu & v-\lambda & u \\ -v^2+gh & 0 & 2v-\lambda \end{bmatrix} && (10.33)\\
&= (v-\lambda)\left[-\lambda(2v-\lambda) - (-v^2+gh)\right] && (10.34)\\
&= (v-\lambda)\left(\lambda^2 - 2v\lambda + v^2 - gh\right) && (10.35)\\
&= (v-\lambda)\left[(\lambda - v)^2 - gh\right] && (10.36)
\end{aligned}$$

O sea que las soluciones son

$$\begin{aligned}
\lambda_1 &= v && (10.37)\\
\lambda_{2,3} &= v \pm \sqrt{gh} && (10.38)
\end{aligned}$$

En ambos casos la solución puede ponerse como

$$\begin{aligned}
\lambda_1 &= u_j n_j && (10.39)\\
\lambda_{2,3} &= u_j n_j \pm \sqrt{gh} && (10.40)
\end{aligned}$$

y efectivamente puede verse que esto vale para cualquier normal n_j.

10.4. Velocidad de fase

Un caso particular de campo de velocidades que varía según una sola dirección como en (10.15) es el caso de las *"ondas planas"*

$$U(\mathbf{x},t) = \begin{bmatrix} \bar{U}_1 \cos(k_j x_j - \omega t + \varphi_1) \\ \bar{U}_2 \cos(k_j x_j - \omega t + \varphi_2) \\ \bar{U}_3 \cos(k_j x_j - \omega t + \varphi_3) \end{bmatrix} \tag{10.41}$$

donde cada componente varía temporalmente según una sinusoide con una frecuencia $\omega = 2\pi/T$, donde T es el período. Si consideramos, como antes, una coordenada $\xi = (k_j x_j)/k$, donde $k = \sqrt{k_j k_j}$ es el módulo del vector k_j, entonces la sinusoide tiene una *"longitud de onda"*

$$\lambda = \frac{2\pi}{k} \tag{10.42}$$

A k_j se le llama el *"vector número de onda"* ya que para cualquier longitud L, kL es 2π veces el número de ondas que entran en L. Notar que si bien todas las amplitudes tienen la misma frecuencia temporal y vector número de onda, cada una puede tener su amplitud U_j y fase φ_j. Ec. (10.41) puede ponerse en forma más compacta como

$$U(\mathbf{x},t) = \mathrm{Re}\left\{\tilde{U}\, \mathrm{e}^{\mathrm{i}(k_j x_j - \omega t)}\right\} \tag{10.43}$$

donde $\mathrm{Re}\{X\}$ denota la parte real del complejo X, y los $\tilde{U}_j$ son complejos dados por

$$\tilde{U}_j = \bar{U}_j\, \mathrm{e}^{\mathrm{i}\varphi_j} \tag{10.44}$$

Reemplazando en (10.14) y operando llegamos a

$$\mathrm{Re}\left\{\left[(-\mathrm{i}\omega I + \mathrm{i}A_j k_j)\,\tilde{U}\right]\, \mathrm{e}^{\mathrm{i}(k_j x_j - \omega t)}\right\} = 0 \tag{10.45}$$

Ahora notemos que todo el término entre corchetes es una constante compleja que no depende ni espacialmente ni del tiempo y la ecuación debe ser válida para todo $\mathbf{x}$ y t. Entonces, tomando $x_j = 0$ y $t = 0$ llegamos a que

$$\mathrm{e}^{\mathrm{i}(k_j x_j - \omega t)} = 1 \tag{10.46}$$

y entonces

$$\mathrm{Re}\left\{\left[(-\mathrm{i}\omega I + \mathrm{i}A_j k_j)\,\tilde{U}\right]\right\} = 0 \tag{10.47}$$

Ahora, tomando $x_j = 0$ y $t = \pi/(2\omega) = T/4$ la exponencial compleja se hace

$$\mathrm{e}^{\mathrm{i}(k_j x_j - \omega t)} = \mathrm{e}^{-\mathrm{i}\pi/2} = -\mathrm{i} \tag{10.48}$$

de manera que

$$\mathrm{Re}\left\{-\mathrm{i}\left[(-\mathrm{i}\omega I + \mathrm{i}A_j k_j)\,\tilde{U}\right]\right\} = \mathrm{Im}\left\{\left[(-\mathrm{i}\omega I + \mathrm{i}A_j k_j)\,\tilde{U}\right]\right\} = 0 \tag{10.49}$$

Juntando (10.49) y (10.47) llegamos a que todo el complejo entre corchetes debe ser nulo y entonces

$$(-\omega I + A_j k_j)\,\tilde{U} = 0 \tag{10.50}$$

es decir que ω y $\tilde{U}$ deben ser autovalor y autovector de la proyección del jacobiano según la dirección dada por el vector número de onda. Entonces, las frecuencias son por (10.39,10.40)

$$\omega_1 = u_j k_j \tag{10.51}$$

$$\omega_{2,3} = u_j k_j \pm \sqrt{gh}k \tag{10.52}$$

Notar que en el segundo término de (10.52) debimos agregar el factor k ya que en (10.39,10.40) se ha hecho uso de que n_j es un vector de valor absoluto unitario. La *"velocidad de fase"* v_ϕ es aquella con la que se mueven los planos de fase constante. Estos son aquellos determinados por

$$k_j x_j - \omega t = \mathrm{cte} \tag{10.53}$$

de donde

$$v_\phi = \frac{\omega}{k} \tag{10.54}$$

Esta velocidad está dirigida según la normal a los planos de fase constante, es decir que en el sentido vectorial

$$v_{\phi,j} = \left(\frac{\omega}{k}\right)\left(\frac{k_j}{k}\right) = \frac{\omega k_j}{k^2} \tag{10.55}$$

10.5. Velocidad de grupo

En realidad la idea de una onda perfectamente monocromática como (10.43). Veamos que sucede con una *"paquete de ondas"*, es decir con una onda monocromática modulada por una función suave, para lo cual elegiremos una Gaussiana (ver figura 10.1). Tomamos como estado inicial (a $t = 0$) una onda monocromática modulada por una Gaussiana

$$U(\mathbf{x}, t = 0) = \tilde{U}\mathrm{e}^{|x - x_0|^2/\sigma^2}\,\mathrm{e}^{\mathrm{i}k_j x_j} \tag{10.56}$$

En un abuso de notación hemos omitido el símbolo "parte real". Queremos ver como evoluciona en el tiempo para $t > 0$. Haciendo una transformada de Fourier en el espacio podemos descomponer a este paquete de ondas en ondas monocromáticas

$$U(\mathbf{x}, t = 0) = \sum_{\mathbf{k}' \approx \mathbf{k}} \hat{U}(\mathbf{k})\,\mathrm{e}^{\mathrm{i}k'_j x_j} \tag{10.57}$$

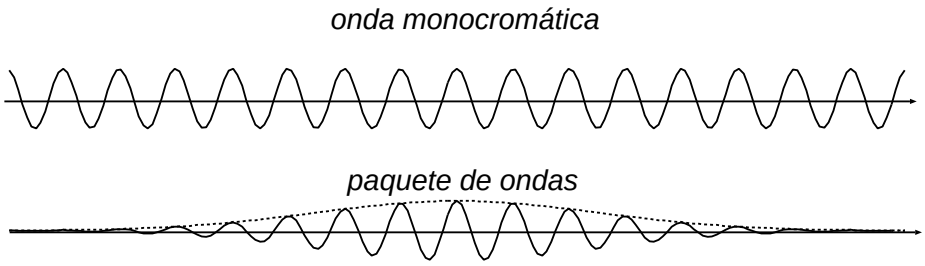

Figura 10.1:

donde como la envolvente es suave las componentes involucradas van a a ser aquellas que están cerca de k. Para cada componente monocromática la evolución temporal viene dada por el término temporal $-\mathrm{i}\omega t$ en la exponencial compleja

$$U(\mathbf{x},t) = \sum_{\mathbf{k}'\approx\mathbf{k},\mu} \hat{U}_\mu(\mathbf{k})\, \mathrm{e}^{\mathrm{i}(k'_j x_j - \omega_\mu t)} \tag{10.58}$$

μ va de 1 a 3 y es un índice sobre los autovalores, para cada vector número de onda $\mathbf{k}$. Como en la suma es sólo sobre vectores de onda $\mathbf{k}'$ cercanos a $\mathbf{k}$ entonces podemos hacer un desarrollo en serie de ω_μ alrededor de $\mathbf{k}$

$$\omega_\mu(\mathbf{k}') = \omega_\mu(\mathbf{k}) + \frac{\partial \omega_\mu}{\partial \mathbf{k}}\,(\mathbf{k}'-\mathbf{k}) + O(|\mathbf{k}'-\mathbf{k}|^2) \tag{10.59}$$

La cantidad $(\partial\omega_\mu/\partial\mathbf{k})$ es un vector con dimensiones de velocidad y se llama la *"velocidad de grupo"* y la denotaremos por como

$$v_{G\mu,j} = \frac{\partial \omega_\mu}{\partial k_j} \tag{10.60}$$

Puede verse (no lo mostraremos aquí) que la envolvente de la onda se propaga con la velocidad de grupo mientras que las crestas de las ondas lo hacen con la de fase (ver figura 10.2). Notar que esto implica que hay un cierto "deslizamiento" de las crestas de las ondas con respecto a la envolvente. Así, la cresta que está en el centro de la envolvente a tiempo t_0 está ubicado 4 longitudes de onda adelante del centro de la envolvente a $t = t_1$.

El concepto de velocidad de grupo ha demostrado ser muy importante en todas las ramas de la física ondulatoria y, en general, puede decirse que tanto la energía como la *"información"* se propagan con la velocidad de grupo.

Calculemos ahora las velocidades de grupo a partir de las leyes de dispersión (10.51,10.52). Tenemos

$$v_{G1,j} = \frac{\partial}{\partial k_j}(k_l u_l) = u_j \tag{10.61}$$

$$v_{G23,j} = \frac{\partial}{\partial k_j}(k_l u_l \pm \sqrt{gh}k) = u_j \pm \sqrt{gh}\frac{k_j}{k} \tag{10.62}$$

Notar que para el primer modo la velocidad de grupo es constante (no depende del vector número de onda), mientras que para los modos 2 y 3 depende de la dirección de $\mathbf{k}$ pero no de su valor absoluto. Ahora supongamos que a $t = 0$ provocamos una perturbación en $\mathbf{x} = 0$. Podemos descomponer esta perturbación como una integral de Fourier, es decir en paquetes de ondas planas de número de onda $\mathbf{k}$ para los diferentes modos $\mu = 1$ a 3. A un cierto tiempo $t^* > 0$ el centro de cada uno de estos paquetes se va a encontrar en $\mathbf{x} = \mathbf{v}_{G\mu} t^*$. Por ejemplo si la velocidad del agua es $\mathbf{u} = 0$, entonces los paquetes

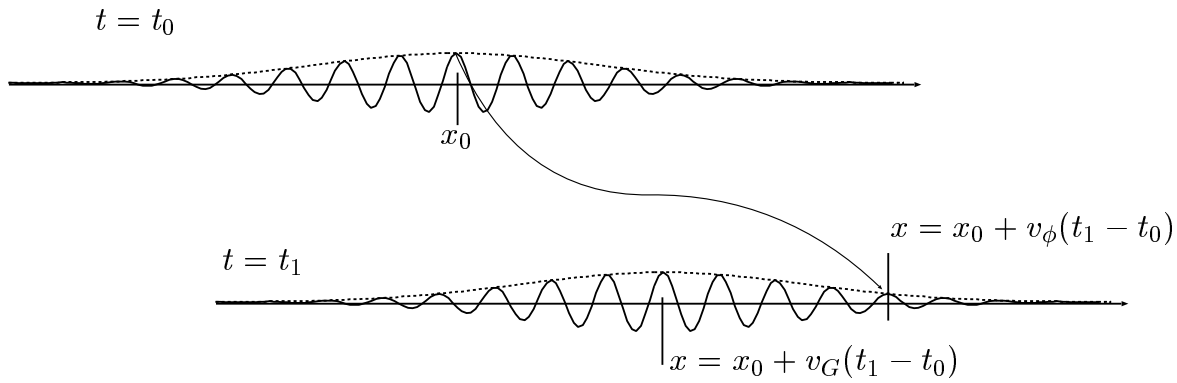

Figura 10.2: La envolvente de una paquete de ondas se propaga con velocidad de grupo, mientras que las crestas de las ondas lo hacen con la velocidad de fase.

que corresponden al primer modo van a formar un punto ubicado en $\mathbf{x} = 0$ mientras que los otros dos van a formar dos círculos superpuestos de radio $\sqrt{gh}t^*$ (ver figura 10.3) . Ahora supongamos que esta perturbación la producimos a intervalos periódicos Δt, es decir a $t^n = n\Delta t$, entonces a $t = t^*$ la posición del frente de onda S^n para la perturbación producida a $t = t^n$ es un círculo con centro en $\mathbf{x} = 0$ de radio $\sqrt{gh}\,(t^* - t^n)$. Esto forma un sistema de círculos concéntricos. Ahora consideremos el caso en que el agua se mueve con velocidad $\mathbf{u}$, entonces la posición del frente de onda S^n es un círculo de radio $\sqrt{gh}\,(t^* - t^n)$ pero ahora centrado en $\mathbf{x} = \mathbf{u}\,(t^* - t^n)$. Asumamos por ahora que el flujo es subcrítico ($u < \sqrt{gh}$) y sin pérdida de generalidad que $\mathbf{u}$ es paralelo al eje x. Entonces los frentes de onda son como se muestra en la figura. La posición de la intersección de los frentes con el eje $x > 0$ es $(u + \sqrt{gh})\,(t^* - t^n)$ mientras que la de las intersecciones con el eje $x < 0$ son $-(\sqrt{gh} - u)\,(t^* - t^n)$. Notar que el espaciamiento aguas abajo $((u + \sqrt{gh})\,\Delta t)$ es mayor que el espaciamiento aguas arriba $((u - \sqrt{gh})\,\Delta t)$. Para velocidades supercríticas los frentes de onda son arrastrados por el fluido con mayor velocidad de la que pueden propagarse hacia aguas arriba de manera que tienden a formar un cono hacia aguas abajo de la fuente de perturbación. Todos los frentes de onda están confinados dentro de un sector angular (un cono para flujo compresible en 3D) llamado *"cono de Mach"*. El ángulo interior θ de este cono puede calcularse (ver figura 10.4)

$$\sin\theta = \frac{\sqrt{gh}\,(t^* - t^n)}{u\,(t^* - t^n)} = \frac{1}{\mathrm{Fr}} \tag{10.63}$$

10.6. Detalles de discretización

Consideremos la resolución de las ecuaciones de shallow-water 2D en un dominio rectangular $0 < x < L_x$, $0 < y < L_y$. Necesitamos ahora una malla 2D de nodos equiespaciados con N_x, N_y intervalos de tal forma que hay $(N_x + 1)(N_y + 1)$ nodos $\mathbf{x}_{ij} = (x_i, y_j)$, con

$$x_i = (i-1)\frac{L_x}{N_x}, \quad i = 1, \ldots, N_x + 1 \tag{10.64}$$

$$y_j = (j-1)\frac{L_y}{N_y}, \quad j = 1, \ldots, N_y + 1 \tag{10.65}$$

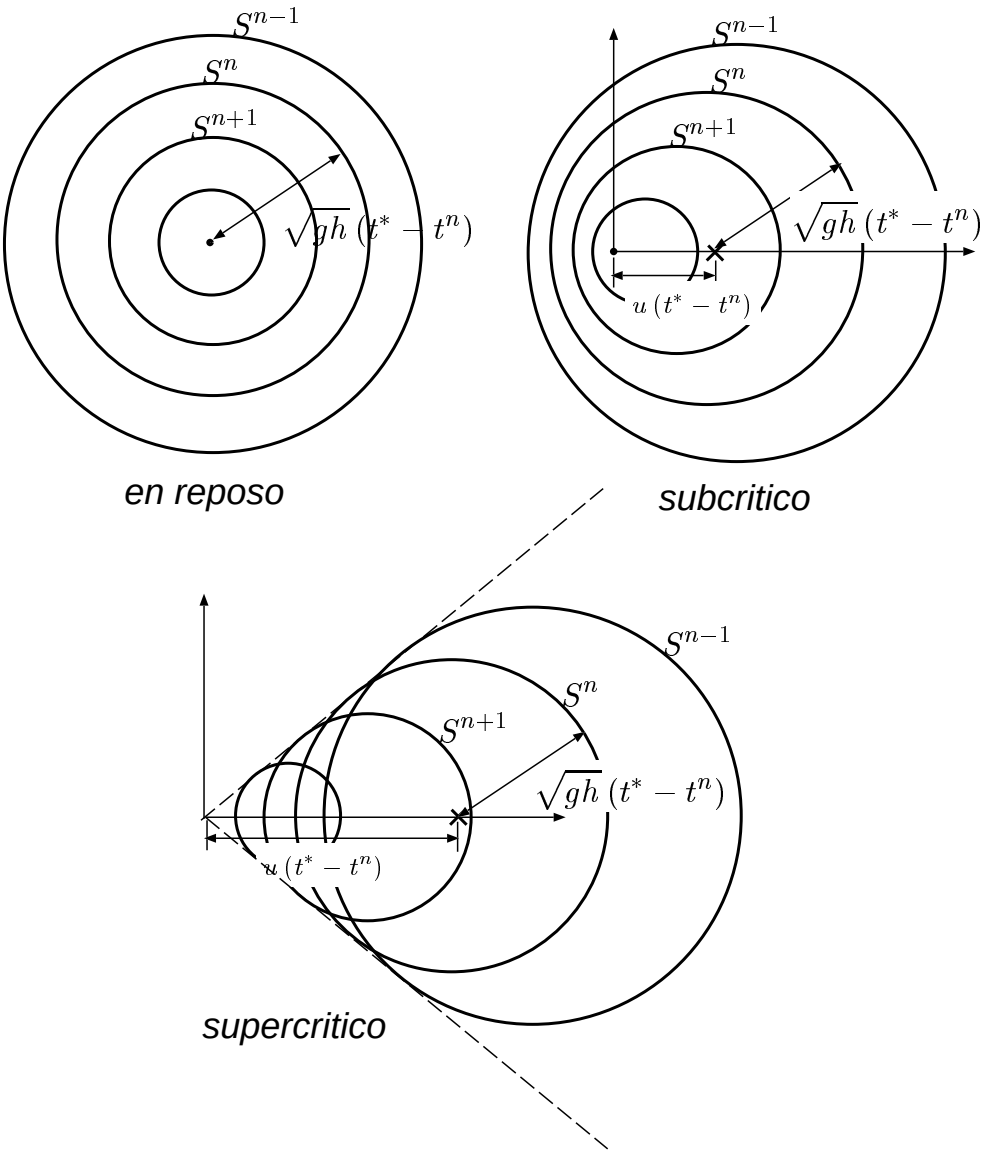

Figura 10.3: Perturbaciones periódicas en diferentes regímenes de flujo.

La ecuación correspondiente al nodo i, j es una aproximación por diferencias a (10.10) a saber

$$\frac{U_{ij}^{n+1} - U_{ij}^{n}}{\Delta t} + \frac{F_{x,i+1/2,j}^{*n} - F_{x,i-1/2,j}^{*n}}{\Delta x} + \frac{F_{y,i,j+1/2}^{*n} - F_{y,i,j-1/2}^{*n}}{\Delta y} = S_{ij}^{n} \tag{10.66}$$

Análogamente al caso 1D los flujos F^* contienen un término de difusión numérica y pueden ponerse de la forma

$$F_{x,i+1/2,j}^{*} = Fx, i + 1/2, j + |A_{x,i+1/2,j}|\,(U_{i+1,j} - U_{i-1,j}) \tag{10.67}$$

donde los flujos centrados F son

$$F_{x,i+1/2,j} = F_(U_{i+1/2,j}) = F\left(\frac{U_{i,j} + U_{i+1,j}}{2}\right) \tag{10.68}$$

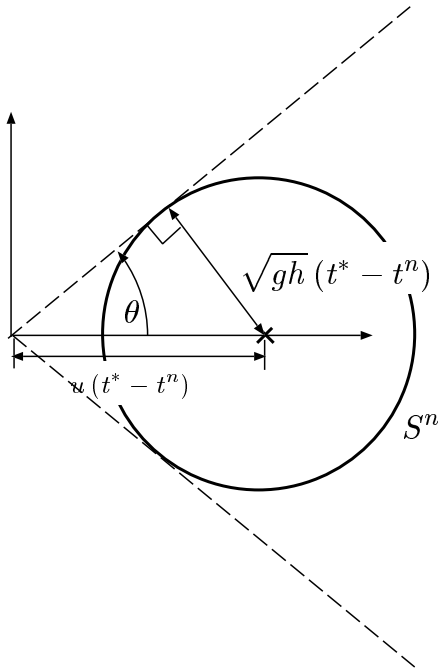

Figura 10.4: Ángulo formado por el cono de Mach.

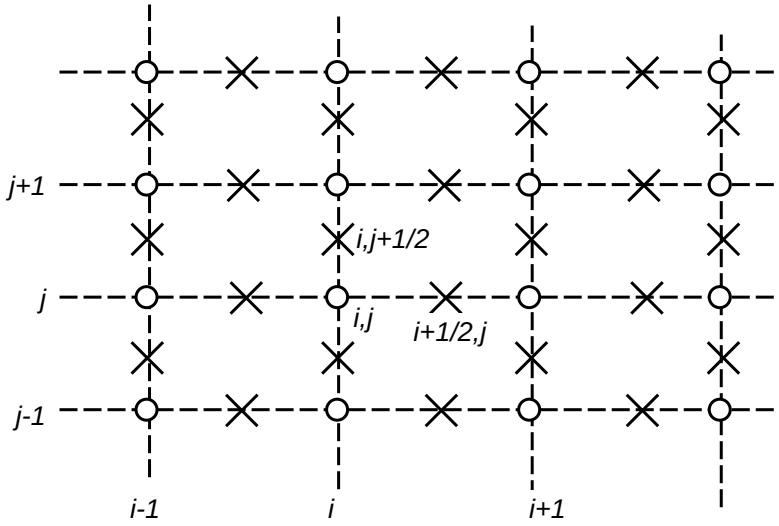

Figura 10.5: Malla 2d para las ecuaciones de shallow-water

10.7. Guía 6. Ec. de shallow water 2D

a. Considerando las siguientes condiciones

$$L_x = 8, \qquad N_x = 80 \tag{10.69}$$
$$L_y = 3, \qquad N_x = 30 \tag{10.70}$$
$$u_0 = u_L = \begin{bmatrix} 1 & 0.4 & 0 \end{bmatrix} \tag{10.71}$$
$$H = H_{\max} \mathrm{e}^{-[(x-x_c)^2+y^2]/\sigma^2} \cos \omega t \qquad \sigma = 0.5 \tag{10.72}$$
$$H_{\max} = 0.2 \tag{10.73}$$

Donde hemos introducido un fondo variable en el tiempo para perturbar el sistema y ver como se propagan las perturbaciones.

a) Llegar al estado periódico y ver la zona de influencia de la perturbación.

b) Ídem en régimen supercrítico

$$u_0 = u_L = \begin{bmatrix} 1 & 1.4 & 0 \end{bmatrix} \tag{10.74}$$

b. Ídem ejercicio previo (en régimen subcrítico y supercrítico) pero con un fondo constante en el tiempo, es decir

$$H = H_{\max} \mathrm{e}^{-[(x-x_c)^2+y^2]/\sigma^2} \tag{10.75}$$

Verificar la ley que relaciona la abertura del cono con el Fr.

c. Buscar la solución estacionaria para el fondo (10.75) con

$$\begin{aligned} \sigma &= 1.5 \\ H_{\max} &= 0.2 \\ x_c &= L_x/4 \end{aligned}$$

con la siguiente inicialización

$$U = \begin{cases} [1, 0.4, 0] & ; \text{ para } x < x_c \\ [h_L, 0.4, 0] & ; \text{ para } x > x_c \end{cases} \tag{10.76}$$

y probar con h_L =1., 0.9, 0.8, ... Se forma un resalto hidráulico? Se mantiene estacionario. Como es la convergencia en cada caso.