Appunti del corso: Analisi numerica Prof. Dario Bini

Stefano Maggiolo
http://poisson.phc.unipi.it/~maggiolo/
maggiolo@mail.dm.unipi.it

2007–2008

Indice

1 Errori **3**
1.1 Rappresentazione in base di numeri reali 3
1.2 Analisi degli errori 4

2 Algebra lineare numerica **8**
2.1 Primo e secondo teorema di Gerschgorin 8
2.2 Riducibilità e terzo teorema di Gerschgorin 11
2.3 Forma canonica di Schur 13
2.3.1 Forma di Schur delle matrici normali 14

3 Norme **15**
3.1 Nozioni generali 15
3.2 Norme matriciali 17
3.3 Norme matriciali e raggio spettrale 18

4 Sistemi lineari: metodi diretti **19**
4.1 Condizionamento 20
4.2 Metodi diretti 21
4.3 Fattorizzazioni LU e QR 21
4.4 Matrici elementari 21
4.5 Decomposizioni con matrici elementari 22
4.5.1 Decomposizione LU con matrici di Gauss 22
4.5.2 Decomposizione QR con matrici di Householder 23
4.6 Costo del calcolo della fattorizzazione LU 23
4.7 Stabilità numerica del calcolo della fattorizzazione LU 24
4.8 Pivoting della fattorizzazione LU 24
4.9 Stabilità numerica del calcolo della fattorizzazione QR 26

5 Sistemi lineari: metodi iterativi **27**
5.1 Convergenza del metodo iterativo 27
5.2 Riduzione asintotica dell'errore 27
5.3 Condizioni di arresto dell'iterazione 28
5.3.1 Valutazione del residuo 28
5.3.2 Valutazione del passo 28
5.4 Metodo di Jacobi 29
5.5 Metodo di Gauss-Seidel 29
5.6 Teoremi di convergenza 30

6 Problemi di punto fisso **31**
6.1 Convergenza 31
6.2 Calcolo degli zeri di una funzione continua 33
6.3 Tipi di convergenza 33
6.4 Metodo di Newton 36
6.4.1 Comportamento vicino a zeri multipli 37
6.5 Velocità di convergenza 38

7 Interpolazione **39**
7.1 Interpolazione polinomiale 39
7.1.1 Resto dell'interpolazione 40
7.1.2 Interpolazione polinomiale di Lagrange 41
7.2 Trasformata discreta di Fourier (DFT) 41
7.2.1 Algoritmi per il calcolo della DFT 42
7.3 Applicazioni della DFT 43
7.3.1 Moltiplicazione di polinomi 43
7.3.2 Moltiplicazione di numeri 44
7.3.3 Interpolazione trigonometrica 44

1 Errori

26/09/2007

Esempio 1.1. Si può scrivere $e^x = \sum_{i=0}^{+\infty} {}^{x^i}/_{i!}$; con questa formula si approssima l'esponenziale troncando la serie a un numero finito di termini (per esempio, quando il termine che si deve aggiungere è più piccolo della precisione di macchina, cioè quando per il calcolatore il risultato non cambia). Si può calcolare l'esponenziale anche in un altro modo:

$$e^x = \frac{1}{e^{-x}} = \frac{1}{\sum_{i=0}^{+\infty} \frac{(-x)^i}{i!}};$$

matematicamente sono espressioni identiche, tuttavia calcolando e^{13}, la prima dà un'approssimazione accurata, mentre la seconda dà addirittura un risultato negativo.

1.1 Rappresentazione in base di numeri reali

Sia β un intero maggiore o uguale a 2.

Teorema 1.2. *Dato $x \in \mathbb{R} \setminus \{0\}$ qualsiasi, esiste unico $p \in \mathbb{Z}$ ed esiste unica una successione $(d_i)_{i \in \mathbb{N} \setminus \{0\}}$ tale che*

$$x = \operatorname{sgn}(x) \left(\sum_{i=1}^{+\infty} d_i \beta^{-i} \right) \beta^p,$$

dove $0 \leq d_i < \beta$, $d_1 \neq 0$ e d_i non sono definitivamente uguali a $\beta - 1$.

In altri termini, d_i sono le *cifre*, p è l'*esponente* e l'espressione tra parentesi è la *mantissa*. Questa rappresentazione è detta *rappresentazione in base normalizzata* o *floating point* (*virgola mobile*).

Esempio 1.3. Se $\beta = 10$, π andrebbe scritto come $+(0.31415\ldots)10^1$. Si richiede che il numero venga scritto con uno 0 prima del punto decimale per garantire l'unicità; per lo stesso motivo si eliminano le espressioni del tipo $0.1999\ldots$ che possono essere scritte anche come 0.2. La richiesta che il primo numero dopo il punto sia diverso da 0, oltre che per l'unicità, è utile per accorciare la lunghezza di una rappresentazione. Un esempio di ciò si ha per la rappresentazione di ${}^1/_{3000}$: si potrebbe scrivere come $0.000333\ldots$, ma la stessa informazione che danno le tre cifre nulle dopo il punto decimale si ha con un'unica cifra all'esponente.

Definizione 1.4. L'insieme $\mathscr{F}(\beta, t, m, M)$ è l'*insieme dei numeri di macchina* o *insieme dei numeri floating point*, cioè l'insieme dei numeri rappresentabili in base β, con t cifre nella mantissa e esponente compreso tra $-m$ e M, unito allo 0:

$$\mathscr{F}(\beta, t, n, m) \coloneqq \{0\} \cup \left\{ \pm\beta^p \sum_{i=1}^{t} d_i \beta^{-i} \mid 0 \leq d - i < \beta, d_1 \neq 0, -m \leq p \leq M \right\}.$$

Dato $x \in \mathbb{R}$, si scrive innanzitutto $x = \operatorname{sgn}(x)\beta^p \sum_{i=1}^{+\infty} d_i \beta^{-i}$. Se $-m \leq p \leq M$, si considera $\tilde{x} \coloneqq \operatorname{sgn}(x)\beta^p \sum_{i=1}^{t} d_i \beta^{-i}$ come numero di macchina corrispondente a x; allora l'*errore relativo*, cioè $\left|\frac{\tilde{x}-x}{x}\right|$, è minore di β^{1-t}. Si osserva che l'errore relativo è maggiorato da una costante indipendente dal numero che si

	byte	β	t	m	M	u
single	4	2	24	125	128	$6 \cdot 10^{-8}$
double	8	2	53	1021	1024	$1.1 \cdot 10^{-16}$
extended	10	2	64	16381	16384	$5 \cdot 10^{-20}$

Tabella 1: Rappresentazioni dello standard IEEE.

sta rappresentando; questa costante si denota con u e si chiama *precisione di macchina*. Se $p > M$, non si può rappresentare x come numero di macchina e questa situazione si chiama *overflow*; se $p < -m$, la situazione è detta *underflow*; in entrambi i casi il processo di calcolo viene arrestato, oppure viene data la possibilità, nel secondo caso, di sostituire il numero con 0. Questa scelta però non garantisce la maggiorazione uniforme dell'errore relativo con β^{1-t}, dato che in questo caso l'errore relativo è uguale a 1.

Esempio 1.5. Si possono usare 4 byte di un calcolatore per rappresentare i numeri di macchina: nel primo byte viene memorizzato il numero $p + m$ (per esempio scegliendo $m = 127$ e, di conseguenza, $M = 128$); nel bit seguente si inserisce il segno (con una qualche convenzione) e successivamente la mantissa, partendo da d_2 fino a d_{24}. Si omette la prima cifra che essendo diversa da 0 in base 2, è necessariamente 1. Quindi con 4 byte si rappresentano i numeri di $\mathscr{F}(2, 24, 127, 128)$, a meno dello 0. Per convenzione, 0 si rappresenta come il più piccolo numero positivo, cioè dato da tutti 0. La precisione con questo modello è circa di $6 \cdot 10^{-8}$.

I modelli per le rappresentazioni al calcolatore sono definiti da uno standard IEEE, come in tabella 1.1. Oltre ai numeri floating point, questi standard possono rappresentare dei valori particolari: NaN (*not a number*), + Infty e − Infty.

1.2 Analisi degli errori

27/09/2007

Definizione 1.6. L'*errore di rappresentazione* di un numero x è $\varepsilon_x := \left|\frac{\tilde{x}-x}{x}\right|$.

Si può scrivere anche $\tilde{x} = x(1 + \varepsilon_x)$; se si calcola l'errore relativo a $\tilde{x}$, invece che a x, si ottiene un numero η_x, minore o uguale alla precisione di macchina u. Si può anche scrivere $\tilde{x} = x(1 + \eta_x)^{-1}$.

Si può definire su $\mathscr{F}$ un'aritmetica: se $\tilde{x}, \tilde{y} \in \mathscr{F}$, e $*$ è un'operazione sui numeri reali, si denota con $\circledast$ il troncamento di $\tilde{x} * \tilde{y}$. Per quest'aritmetica troncata si ha $\tilde{x} \circledast \tilde{y} = (\tilde{x} * \tilde{y})(1 + \delta)$, con $|\delta| \leq u$. L'errore δ è detto *errore locale dell'operazione*.

Finora si hanno quindi due errori: il primo è quello per rappresentare un numero reale qualsiasi come numero di macchina, il secondo è quello causato da ogni operazione aritmetica tra numeri di macchina.

Sia $f(x)$ una funzione che è possibile calcolare con un numero finito di operazioni (il che implica necessariamente che f sia razionale). Innanzitutto, anche se si desidera calcolare $f(x)$, è possibile calcolare solo $f(\tilde{x})$. L'errore

$$\varepsilon_{\text{in}} := \frac{f(\tilde{x}) - f(x)}{f(x)}$$

è detto *errore inerente* ed è intrinseco al problema. Successivamente, si ha l'errore locale di ogni operazione eseguita per calcolare $f(\tilde{x})$; sia φ la funzione che

si ottiene da f sostituendo le operazioni troncate a quelle normali: l'errore

$$\varepsilon_{\text{alg}} \coloneqq \frac{\varphi(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}$$

è detto invece *errore algoritmico*. Infine, l'*errore totale* è semplicemente

$$\varepsilon_{\text{tot}} \coloneqq \frac{\varphi(\tilde{x}) - f(x)}{f(x)}.$$

Con un semplice calcolo, si ha che $\varepsilon_{\text{tot}} = \varepsilon_{\text{in}} + \varepsilon_{\text{alg}} + \varepsilon_{\text{in}}\varepsilon_{\text{alg}} \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{alg}}$, dove l'approssimazione è giustificata dal fatto che l'errore totale è un polinomio in u senza termine noto, e generalmente si può considerare solo la parte lineare rispetto a u. Grazie a questa suddivisione dell'errore totale, si possono trattare separatamente i due tipi di errore, inerente e algoritmico.

Errore inerente. È ragionevole assumere che tra x e $\tilde{x}$ non ci sia uno zero del denominatore della funzione razionale; se è così, si può sviluppare f in serie di potenze con centro x e applicare la formula $\tilde{x} = x(1 + \varepsilon_x)$, ottenendo:

$$\varepsilon_{\text{in}} = \frac{\left(\cancel{f(x)} + (\tilde{x} - x)f'(x) + \frac{(\tilde{x}-x)^2}{2}f''(\xi)\right) - \cancel{f(x)}}{f(x)} \doteq \varepsilon_x \frac{xf'(x)}{f(x)}.$$

Il coefficiente di ε_x in questa espressione, ${xf'(x)}/{f(x)}$, è detto *coefficiente di amplificazione* di f in x. Ci sono problemi in analisi numerica, anche innocui all'apparenza, che hanno un coefficiente di amplificazione che cresce in modo esponenziale con la dimensione (per esempio, come si vedrà in seguito, la risoluzione del sistema lineare relativo alla matrice di Hilbert). Nel caso di una funzione razionale in più variabili, con un calcolo analogo si ottiene la formula

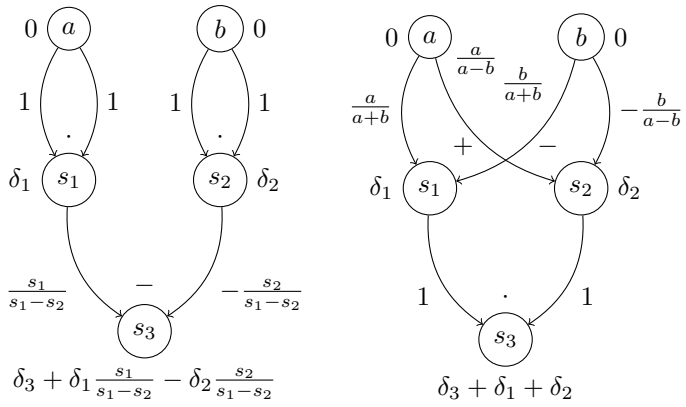
$$\varepsilon_{\text{in}} = \frac{f(\tilde{x}_1, \ldots, \tilde{x}_n) - f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)} \doteq \sum_{i=1}^{n} \varepsilon_{x_i} \frac{x_i \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)};$$

in questo caso per ogni variabile si ha un relativo coefficiente di amplificazione.

Esempio 1.7. Sia $f(x_1, x_2) \coloneqq x_1 x_2$. Il coefficiente di amplificazione rispetto a x_1 è dato da $\frac{x_1 x_2}{x_1 x_2} = 1$, e così per x_2; la moltiplicazione quindi non amplifica l'errore (con un'analisi al primo ordine). Anche per $f(x_1, x_2) \coloneqq {x_1}/{x_2}$ il coefficiente di amplificazione è limitato per entrambe le variabili (1 rispetto a x_1, -1 rispetto a x_2). Invece, per $f(x_1, x_2) \coloneqq x_1 + x_2$, il coefficiente di amplificazione rispetto a x_i è $\frac{x_i}{x_1+x_2}$. Se x_1 e x_2 sono concordi entrambi i coefficienti sono minori di 1, ma se sono discordi uno dei due coefficienti diventa maggiore di 1 e in particolare se i numeri da sommare sono vicini in modulo l'errore può amplificarsi anche di molto. Di conseguenza, si deve evitare di addizionare due numeri discordi (è il motivo per la differenza di precisione delle due formule nell'esempio 1.1).

Esempio 1.8. Sia $at^2 + bt + c = 0$ un'equazione di secondo grado; se $b \neq 0$ e l'equazione ha due soluzioni distinte, per trovare le due soluzioni si devono effettuare una somma tra numeri concordi in un caso e tra numeri discordi nel caso (poiché la radice è sempre positiva). Per calcolare la soluzione che presenta problemi, si può usare il fatto che il prodotto delle soluzioni è $-ac$, il che permette di esprimere una radice in funzione dell'altra usando solo moltiplicazioni e divisioni.

Figura 1: Calcolo dell'errore algoritmico tramite grafo.



(a) Primo algoritmo. (b) Secondo algoritmo.

Definizione 1.9. Il fenomeno che si presenta nella somma di due numeri di segno opposto e valore assoluto simile è detto *cancellazione*.

Errore algoritmico.

Esempio 1.10. Si vuole calcolare la funzione $a^2 - b^2 = (a + b)(a - b)$. Si può calcolare con questi due algoritmi:

$$\begin{aligned} s_1 &= a \cdot a & \qquad s_1 &= a + b \\ s_2 &= b \cdot b & \qquad s_2 &= a - b \\ s_3 &= s_1 - s_2 & \qquad s_3 &= s_1 s_2. \end{aligned}$$

Il secondo algoritmo è preferibile dal punto di vista del costo computazionale (generalmente, la moltiplicazione ha un costo maggiore rispetto alla somma), ma potrebbe essere conveniente usare il secondo se il suo errore algoritmico è minore. Per calcolarlo, si costruisce un grafo (figura 1) che rappresenta l'algoritmo, che ha come nodi ogni quantità in ingresso o calcolata (a, b, s_1, s_2, s_3) e un arco da ogni operando a ogni risultato, pesato con il coefficiente di amplificazione relativo. Successivamente, si associa a ogni nodo, partendo dalle sorgenti, l'errore algoritmico corrispondente, che è un certo δ sommato agli errori degli operandi moltiplicati per i relativi coefficienti di amplificazione (poiché si sta calcolando l'errore algoritmico, a e b si suppongono essere già numeri di macchina, cioè privi di errore). Infine si può maggiorare in valore assoluto l'errore algoritmico grazie al fatto che $\delta_i \leq u$.

Nel primo algoritmo, si ha la limitazione

$$|\varepsilon_{\text{alg}}| \dot{\leq} u \left(1 + \left|\frac{a^2}{a^2 - b^2}\right| + \left|\frac{b^2}{a^2 - b^2}\right|\right),$$

perciò l'errore può essere molto grande se a^2 e b^2 sono vicini. Al contrario, per il secondo algoritmo si ha $|\varepsilon_{\text{alg}}| \dot{\leq} 3u$, che è indipendente da a e b e quindi a maggior ragione limitato.

Esempio 1.11. Se si vuole calcolare $\sum_{i=1}^{n} x_i$ (dove $x_i > 0$ per ogni i), seguendo l'algoritmo sequenziale, cioè sommando da sinistra a destra, l'errore algoritmico è maggiorato da $(n-1)u$; viceversa, se si effettua l'algoritmo parallelo, cioè si somma prima a coppie e così via, l'errore è maggiorato da $u(\lceil \log_2 n \rceil + 1)$.

Definizione 1.12. Un algoritmo è detto *numericamente stabile* se l'errore algoritmico generato è "piccolo" (in un qualche senso da specificare); se questo non accade, si dice *numericamente instabile*. Un problema si dice *ben condizionato* se piccole perturbazioni dei dati provocano piccole perturbazioni dei risultati (cioè se il coefficiente di amplificazione è piccolo); se accade il contrario, si *mal condizionato*.

Nel caso di problemi mal condizionati, non è sufficiente cambiare algoritmo per migliorare la situazione. Ci si può appoggiare a pacchetti specifici che permettono di lavorare in precisione arbitraria, oppure si può trattare il problema in modo simbolico, usando solo operazioni che non introducono errori.

28/09/2007

Analisi all'indietro. Il tipo di analisi visto finora è detto *analisi in avanti*; non è l'unico: esiste anche l'*analisi all'indietro*. Sia come prima $\varphi(x)$ il valore effettivamente calcolato cercando di calcolare la funzione f in x; l'analisi all'indietro si concentra sul dominio e non sul codominio, cioè cerca un valore y tale che $f(y) = \varphi(x)$. L'errore algoritmico assume in questo caso la forma $\frac{f(y)-f(x)}{f(x)}$, cioè si presenta formalmente come un errore inerente. A questo punto basta studiare i coefficienti di amplificazione di f per calcolare l'errore algoritmico.

Esempio 1.13. Sia $f(x,y) \coloneqq x+y$; allora $\varphi(x,y) = (x+y)(1+\delta)$, con $|\delta| \leq u$. Ma allora $\varphi(x,y) = x(1+\delta) + y(1+\delta) = f(x(1+\delta), y(1+\delta)) = f(\hat{x}, \hat{y})$.

Esempio 1.14. A volte, l'analisi in avanti è lunga, mentre l'analisi all'indietro è più vantaggiosa. Sia $p(t) \coloneqq at^2 + bt + c$; fissato un punto, per valutarvi il polinomio, si possono calcolare tutte le potenze del punto e poi moltiplicare per i coefficienti, o più efficientemente si può usare l'algoritmo di Horner-Ruffini: $p(t) = (at+b) * t + c$, che usa n moltiplicazioni e n addizioni, con $n = \deg p$. Si calcola l'errore algoritmico di questo modo con l'analisi all'indietro:

$$\begin{aligned}\varphi(a,b,c) &= \Big(\big((at)(1+\delta_1)+b\big)(1+\delta_2)t(1+\delta_3)+c\Big)(1+\delta_4) = \\ &= c(1+\delta_4) + bt\prod_{i=4}^{2}(1+\delta_i) + at^2\prod_{i=4}^{1}(1+\delta_i) = \\ &= \hat{c} + \hat{b}t + \hat{a}t^2 = f(\hat{a},\hat{b},\hat{c}).\end{aligned}$$

Si osserva che calcolando si ottiene il valore esatto di un polinomio coi coefficienti perturbati: c può essere perturbato di una quantità limitata da u, b da $3u$, a da $4u$.

Se si scopre la stabilità di un algoritmo con l'analisi all'indietro, si dice che l'algoritmo è *stabile all'indietro*.

Errore analitico Per calcolare una funzione non razionale f (per esempio, e^x o $\cos(x)$), si può solo sostituire a f una funzione razionale h che bene la

approssima, che spesso è lo sviluppo in serie di f troncato. L'errore che segue dall'approssimazione è detto *errore analitico*:

$$\varepsilon_{\text{an}} \coloneqq \frac{h(x) - f(x)}{f(x)}.$$

Se si considera l'errore totale (rispetto a f) si ottiene facilmente che

$$\varepsilon_{\text{tot}} \doteq \varepsilon_{\text{an}} + \varepsilon_{\text{in}} + \varepsilon_{\text{alg}} = \frac{h(x) - f(x)}{f(x)} + \frac{h(\tilde{x}) - h(x)}{h(x)} + \frac{\varphi(\tilde{x}) - h(\tilde{x})}{h(\tilde{x})}.$$

Si può calcolare l'errore inerente sulla funzione f anziché su h, poiché spesso il calcolo risulta molto più semplice.

Esempio 1.15. Si vuole calcolare $e^{x+y} = e^x e^y$; per capire quale algoritmo è migliore, si suppone di avere una funzione razionale h approssimante l'esponenziale, tale che $h(x) = e^x(1+\vartheta)$ (dove ϑ non è altro che l'errore analitico sommato all'errore algoritmico dell'operazione). Il coefficiente di amplificazione per l'operazione di elevamento a potenza è l'esponente stesso. Applicando il metodo dei grafi, l'errore totale per il primo algoritmo risulta

$$\varepsilon_{\text{tot}\,1} = \vartheta + (x+y)\left(\delta + \frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y\right) = \vartheta + x\varepsilon_x + y\varepsilon_y + (x+y)\delta,$$

da cui togliendo l'errore inerente, si ha che il modulo dell'errore algoritmico è $\left|\varepsilon_{\text{alg}_1}\right| = |\vartheta + (x+y)\delta| \leq (1 + |x+y|)u$. Per il secondo algoritmo si ottiene

$$\varepsilon_{\text{tot}\,2} = \delta + \vartheta_1 + x\varepsilon_x + \vartheta_2 + y\varepsilon_y,$$

da cui il modulo dell'errore algoritmico è $\left|\varepsilon_{\text{alg}_2}\right| = |\delta + \vartheta_1 + \vartheta_2| \leq 3u$. Perciò seconda del valore di x e di y è cambia l'algoritmo migliore.

Altri tipi di approccio sono quelli di tipo probabilistico, dove gli errori sono considerati come variabili aleatorie, il che però appesantisce notevolmente l'analisi; oppure si può usare la cosiddetta aritmetica degli intervalli, cioè non si lavora su numeri ma su intervalli che contengono il numero corretto.

2 Algebra lineare numerica

01/10/2007

L'algebra lineare numerica (o NLA, da *numerical linear algebra*), per quanto si vedrà, servirà essenzialmente per risolvere sistemi lineari. Nel proseguo, le matrici saranno indicate con lettere maiuscole; le entrate della matrice A con $a_{i,j}$; lo spazio delle matrici con n righe e m colonne su un campo K si denoterà con $K^{n\times m}$. Gli elementi del tipo $a_{i,i}$ sono gli elementi della *diagonale principale*, e vengono detti elementi *diagonali* o *principali*.

2.1 Primo e secondo teorema di Gerschgorin

Data $A \in \mathbb{C}^{n\times n}$, si denota con K_i il cerchio chiuso

$$\left\{ z \in \mathbb{C} \mid |z - a_{i,i}| \leq \sum_{j\neq i} |a_{i,j}| \right\};$$

K_i è detto *i-esimo cerchio di Gerschgorin*. Per esempio, sia A la matrice

$$\begin{pmatrix} 4 & -1 & 1 \\ 0 & 3 & -2 \\ -1 & 4 & 6 \end{pmatrix};$$

allora il primo centro di Gerschgorin ha centro 4 e raggio 2 (la somma degli elementi della prima riga tranne quello diagonale); il secondo cerchio ha centro 3 e raggio 2; il terzo cerchio ha centro 6 e raggio 5.

Teorema 2.1 (primo teorema di Gerschgorin)**.** *Gli autovalori di $A \in \mathbb{C}^{n\times n}$ appartengono a $\bigcup_{i=1}^{n} K_i$.*

Dimostrazione. Sia λ un autovalore per A e x un autovettore corrispondente; allora vale $Ax = \lambda x$, che sulla i-esima riga diventa $\sum_{j=1}^{n} a_{i,j}x_j = \lambda x_i$. Sia ora x_k una delle componenti di x di massimo modulo (in particolare x_k è non nullo); si può scrivere $(a_{k,k} - \lambda)x_k = -\sum_{j\neq k} a_{k,j}x_j$, da cui, dividendo per x_k, si ottiene

$$a_{k,k} - \lambda = -\sum_{j\neq k} a_{k,j}\frac{x_j}{x_k}.$$

Ma $|x_j/x_k| \leq 1$, perciò, grazie alla disuguaglianza triangolare, $|a_{k,k} - \lambda| \leq \sum_{j\neq k} |a_{k,j}|$, cioè $\lambda \in K_k$. □

Esempio 2.2. Sia $P(x) \coloneqq a_n x^n + \cdots + a_0$; grazie al primo teorema di Gerschgorin si può dare una stima sulle radici: innanzitutto si suppone che P sia monico, cioè che $a_n = 1$; successivamente si può costruire la matrice

$$C \coloneqq \begin{pmatrix} 0 & \cdots & 0 & -a_0 \\ & & & -a_1 \\ & I_{n-1} & & \vdots \\ & & & -a_{n-1} \end{pmatrix},$$

detta *matrice di Frobenius* del polinomio P. Ora, il polinomio caratteristico di C si dimostra, per induzione, essere $(-1)^n P(\lambda)$. Di conseguenza gli autovalori sono le radici di $P(x)$ e grazie al teorema si ha la stima sulle radici.

Teorema 2.3 (secondo teorema di Gerschgorin)**.** *Sia $A \in \mathbb{C}^{n\times n}$ una matrice tale che $\bigcup K_i = M_1 \cup M_2$, dove M_1 e M_2 sono unioni di cerchi di Gerschgorin, con $M_1 \cap M_2 = \varnothing$. Allora M_i contiene tanti autovalori quanti sono i cerchi che la costituiscono.*

Dimostrazione. Sia $D = \operatorname{diag}(a_{1,1}, \ldots, a_{n,n})$ la matrice diagonale con la stessa diagonale di A; si definisce $A_t \coloneqq D + t(A - D)$: al variare di t tra 0 e 1, A_t traccia un segmento nello spazio delle matrici (in particolare, è una funzione continua). Allora anche gli elementi di A_t variano in modo continuo con t e così anche i suoi autovalori: infatti si può dimostrare che gli zeri di un polinomi variano con continuità al variare dei coefficienti, che a loro volta sono espressi in termini delle entrate secondo le funzioni simmetriche elementari, che sono ancora continue (sono somme di prodotti).

Si osserva che il segmento così definito va da $A_0 = D$ a $A_1 = A$ e che i cerchi di Gerschgorin di A_t hanno lo stesso centro di quelli di A e raggio minore o

uguale a quelli di A (in particolare, i cerchi che compongono M_1 sono sempre disgiunti da quelli che compongono M_2). Per $t = 0$, i cerchi degenerano in un punto, che è l'autovalore, perciò per A_0 il teorema vale. All'aumentare di t, il numero di autovalori nella prima componente non può cambiare, altrimenti un autovalore dovrebbe spostarsi da una componente all'altra, ma si è osservato che gli spostamenti degli autovalori sono continui, perciò questo è impossibile. □

Esempio 2.4. Si può usare il secondo teorema di Gerschgorin per capire se una matrice a entrate reali ha autovalori reali, sapendo che in questo caso gli autovalori non reali si presentano in coppie coniugate. Infatti, se esiste un sottoinsieme dei cerchi la cui unione è disgiunta dagli altri cerchi e la cui cardinalità è dispari, necessariamente ci deve essere un autovalore reale, dato che quelli complessi si presentano in coppie di coniugati e che i cerchi di Gerschgorin hanno tutti centro sull'asse reale (dunque se contengono un punto contengono anche il suo coniugato).

Poiché gli autovalori di una matrice non cambiano se la matrice viene trasposta, gli autovalori sono contenuti anche nei cerchi

$$H_i \coloneqq \left\{ z \in \mathbb{C} \mid |z - a_{i,i}| \leq \sum_{j \neq i} |a_{j,i}| \right\} :$$

sono cerchi con gli stessi centri dei K_i ma con raggi calcolati sommando i termini sulla colonna e non sulla riga. Questi cerchi possono dare una stima diversa da quella ottenuta con i K_i.

Esempio 2.5. Una classe importante di matrici sono le matrici *tridiagonali*: sono matrici nulle ovunque tranne che sulla diagonale principale e su quelle direttamente sopra e sotto. Si considera la matrice

$$A \coloneqq \begin{pmatrix} 100 & 1 & 0 & 0 & 0 \\ 2 & 1 & 2 & 0 & 0 \\ 0 & -1 & 95 & 4 & 0 \\ 0 & 0 & 1 & 10 & 1 \\ 0 & 0 & 0 & -1 & 100 \end{pmatrix};$$

non si hanno garanzie sulla non singolarità di A, dato che K_2 contiene l'origine. Tuttavia si può considerare una congruenza che moltiplica la seconda riga di un fattore ε (e di conseguenza divide per ε la seconda colonna). Scegliendo opportunamente ε, si può restringere il cerchio K_2, al prezzo di allargare i cerchi K_1 e K_3; se si riesce a farlo abbastanza da escludere l'origine, senza includerla in uno degli altri due cerchi, si dimostra che la matrice originale è non singolare.

Definizione 2.6. Una matrice A si dice *dominante diagonale in senso stretto* o *fortemente dominante diagonale* se per ogni i si ha

$$|a_{i,i}| > \sum_{j \neq i} |a_{i,j}| .$$

Una matrice si dice *dominante diagonale* se la disuguaglianza vale per almeno una riga e se per tutte le righe vale con il segno rilassato.

In particolare, una matrice dominante diagonale in senso stretto non può essere singolare per il primo teorema di Gerschgorin. Se una matrice fortemente dominante diagonale è a coefficienti reali, il numero di autovalori con parte reale positiva (negativa) corrisponde al numero di elementi principali positivi (negativi).

2.2 Riducibilità e terzo teorema di Gerschgorin

Definizione 2.7. Una matrice P è *di permutazione* se $p_{i,j} = \delta_{\sigma i,j}$ con σ una permutazione.

Si dice matrice di permutazione perché se A è una qualsiasi matrice di n righe (o un vettore), PA è la matrice A con le righe permutate, mentre PAP^t è una permutazione sia delle righe che delle colonne di A; inoltre $PP^t = I = P^tP$, cioè P è una matrice ortogonale.

Definizione 2.8. Una matrice $A \in \mathbb{C}^{n\times n}$ si dice *riducibile* se esiste una matrice di permutazione $P \in \mathbb{C}^{n\times n}$ per cui

$$PAP^t = \begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{pmatrix},\ A_{1,1} \in \mathbb{C}^{m\times m}.$$

Visto nell'ottica dei sistemi lineari, la riducibilità permette di scomporre il sistema in due parti. Poiché il costo di risolvere un sistema è più che lineare, risolvere due sistemi di dimensione minore è più vantaggioso di risolverne uno grande.

02/10/2007

Data una matrice A, si associa a essa un grafo che ha tanti nodi quant'è la dimensione della matrice e che ha un arco dal nodo i al nodo j se e solo se $a_{i,j} \neq 0$. Si dimostrerà che la forte connessione (ovvero la possibilità di trovare un percorso tra qualunque coppia ordinata di vertici) del grafo associato è equivalente alla irriducibilità della matrice.

Esempio 2.9. Non è intuitivo capire se le matrici

$$A = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ & & & 0 \\ & I_{n-1} & & \vdots \\ & & & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

siano riducibili, anche se sono costituite per lo più da entrate nulle. Con il metodo dei grafi però si scopre che sono irriducibili, dato che il grafo associato a A è costituito da un unico ciclo, mentre quello associato a B è un segmento con archi diretti in entrambi i sensi.

Teorema 2.10. *Una matrice è riducibile se e solo se il grafo corrispondente non è fortemente connesso.*

Dimostrazione. Si inizia con l'osservare che i grafi associati a A e a $B := PAP^t$ differiscono solo per una permutazione dei nodi; in particolare A è fortemente connessa se e solo se lo è B.

($\Rightarrow$) Se A è riducibile, si può supporre che sia già in una forma in cui $A_{2,1} = 0$ e si osserva facilmente che non esiste nessun cammino tra un nodo di indice maggiore di m a un nodo di indice minore o uguale a m.

($\Leftarrow$) Se il grafo di A non è fortemente connesso, esiste una coppia di nodi (p, q) per cui non esiste un cammino da p a q. Siano Q l'insieme (non vuoto) dei nodi non raggiungibili da p e P l'insieme di quelli raggiungibili; chiaramente da un nodo di P non si può raggiungere un nodo di Q. Se si permutano le righe e le colonne in modo che gli indici di Q siano messi in testa e quelli di P in coda, si ha la riducibilità di A. □

Esercizio 2.11. Una matrice tridiagonale con entrate tutte diverse da 0 sulle tre diagonali è irriducibile.

Teorema 2.12 (terzo teorema di Gerschgorin)**.** *Se $A \in \mathbb{C}^{n\times n}$ è una matrice irriducibile e λ un suo autovalore per cui vale la proprietà $\lambda \in K_i \Rightarrow \lambda \in \partial K_i$, allora λ appartiene a tutti i K_i (e quindi a tutte le frontiere).*

Dimostrazione. Ripercorrendo la dimostrazione del primo teorema di Gerschgorin, dato che $\lambda \in \partial K_k$, le maggiorazioni devono essere uguaglianze:

$$|a_{k,k} - \lambda| = \sum_{j\neq k} |a_{k,j}| \left|\frac{x_j}{x_k}\right| = \sum_{j\neq k} |a_{k,j}| \,;$$

di conseguenza, se $a_{k,j} \neq 0$, x_k e x_j devono essere uguali in modulo. Poiché la matrice è irriducibile, esiste k_1 tale che $a_{k,k_1} \neq 0$, e quindi $|x_k| = |x_{k_1}|$. Di conseguenza, si può scegliere come indice di componente di modulo massimo k_1; iterando il procedimento, per l'irriducibilità, si può trovare una successione k_i che copre tutti gli indici. Dunque, sempre riprendendo la dimostrazione del primo teorema di Gerschgorin, ciascuna componente può essere presa come componente di modulo massimo, ossia λ appartiene a tutti i cerchi di Gerschgorin. □

Esempio 2.13. Nei problemi di deformazione unidimensionale, interviene la matrice tridiagonale con 2 sulla diagonale e -1 sulla sopradiagonale e sottodiagonale. I cerchi di Gerschgorin hanno tutti centro 2 e raggio 1 o 2. Per il terzo teorema di Gerschgorin, la matrice non può essere singolare, perché se 0 fosse un autovalore, soddisferebbe le ipotesi del teorema, ma 0 non sta sulla frontiera di tutti i cerchi.

03/10/2007

Esercizio 2.14. Si definisce l'*ovale di Cassini* come

$$C_{i,j} \coloneqq \{\, z \in \mathbb{C} \mid \; |z - a_{i,i}|\,|z - a_{j,j}| \leq S_i S_j \,\}\,,$$

dove S_i e S_j sono i raggi dei cerchi di Gerschgorin K_i e K_j. In realtà, la forma di un ovale di Cassini può essere un ovale schiacciato al centro o due cerchi. Si dimostri che gli autovalori appartengono all'unione degli ovali di Cassini, che a loro volta sono contenuti nei cerchi di Gerschgorin. Quindi si ha una stima migliore degli autovalori, ma meno pratica perché la forma dell'insieme è meno facile da comprendere.

Se una matrice è dominante diagonale (anche non strettamente) e irriducibile, per il terzo teorema di Gerschgorin è non singolare: infatti, 0 sta (al più) sulla frontiera dell'unione dei cerchi, quindi soddisfa le ipotesi del teorema. Allora, se 0 fosse un autovalore, dovrebbe appartenere a tutti i cerchi, ma per definizione ce n'è almeno uno che non contiene l'origine.

2.3 Forma canonica di Schur

Oltre alla trasposizione di una matrice, si introduce un'altra operazione: se $a+ib$ è un numero complesso, il suo *coniugato* è $\overline{a+ib} \coloneqq a-ib$; per le matrici, si indica con A^h la trasposta coniugata di A, cioè la matrice B con $b_{i,j} \coloneqq \overline{a_{j,i}}$. La lettera "$h$" si usa in onore di Charles Hermite.

Se $x, y \in \mathbb{R}^n$ sono vettori colonna, la quantità reale $\sum_{i=1}^{n} x_i y_i$ può essere indicata con $x^t y$ (il prodotto matriciale tra x trasposto e y) o con $\langle x, y \rangle$, dato che è un prodotto scalare euclideo; in particolare, la lunghezza di un vettore secondo questo prodotto scalare è data da $\sqrt{x^t x}$. Nel campo complesso, se $x, y \in \mathbb{C}^n$, si definisce il *prodotto scalare hermitiano* $\langle x, y \rangle \coloneqq x^h y$. Grazie alla disuguaglianza di Cauchy-Schwartz, si ha una nozione di angolo tra due vettori, che permette di interpretare geometricamente il prodotto scalare: si dice che x e y sono *ortogonali* se $\langle x, y \rangle = 0$.

Definizione 2.15. Una matrice A si dice:

- *simmetrica*, se $A = A^t$;
- *hermitiana*, se $A = A^h$;
- *antisimmetrica*, se $A = -A^t$;
- *antihermitiana*, se $A = -A^h$;
- *ortogonale*, se è quadrata, reale e $AA^t = I$;
- *unitaria*, se è quadrata, complessa e $AA^h = I$.

In particolare, se A è ortogonale o unitaria e $x = Ay$, la lunghezza (con il prodotto scalare corrispondente) di x e y è uguale. Ancora, se $x_i = Ay_i$ per $i \in \{1, 2\}$, l'angolo tra y_1 e y_2 è uguale all'angolo tra x_1 e x_2. In realtà, si dimostra che una matrice è ortogonale (unitaria) se e solo se conserva il prodotto scalare euclideo (hermitiano). Un'altra proprietà delle matrici ortogonali o unitarie è che le colonne (e le righe) della matrice sono ortonormali, grazie alla relazione $AA^t = I$.

Se si considera una matrice A, si può mettere in forma canonica di Jordan, cioè si può trovare una matrice invertibile S tale che $A = SJS^{-1}$ e J è una matrice diagonale a blocchi, dove i blocchi sono nulli se non sulla diagonale, dove risiede un autovalore, e sulla sopradiagonale, dove valgono 1. In questa forma canonica, l'informazione sulle proprietà della matrice è concentrata nella matrice J. Esistono anche altre forme canoniche: per esempio, è importante la *forma canonica di Schur*, che scrive una matrice A come QTQ^h, dove Q è unitaria e T è triangolare superiore. In questo caso, nella matrice T sono codificati solo gli autovalori di A, mentre le informazioni sulle molteplicità si perdono; tuttavia si ha la condizione molto forte che Q sia unitaria.

Per dimostrare l'esistenza della forma di Schur, si può partire dalla forma di Jordan e ortonormalizzare la matrice S con un processo di Gram-Schmidt. Tuttavia, dal punto di vista numerico questo non è produttivo, perché calcolare la forma di Jordan è numericamente impossibile (per esempio, se si considera una matrice costituita da un unico blocco di Jordan di autovalore 0 e molteplicità algebrica n, e si perturba mettendo un ε in posizione $(n, 1)$, gli autovalori cambiano radicalmente, diventano le radici n-esime di ε).

Teorema 2.16. *Ogni matrice ammette una scrittura in forma di Schur.*

Dimostrazione. Si dimostrerà per induzione sulla dimensione di A. Se $n = 1$, si scrive semplicemente $A = 1A1$. Ora si suppone che il risultato valga per matrici di dimensione $n-1$ e si considera una matrice A di dimensione n. Siano λ un autovalore di A e x un autovettore corrispondente, normalizzato in modo che $x^h x = 1$; si costruisce una matrice U che ha x sulla prima colonna, mentre le altre colonne si scelgono nell'ortogonale di x e in modo che siano ortonormali tra loro (questa scelta è possibile, anche computazionalmente, per esempio utilizzando l'algoritmo di Gram-Schmidt). Di conseguenza, $U^h U = I$; ora si vuole capire com'è fatta la prima colonna di $U^h AU$, cosa che si può fare moltiplicando tutto per il vettore colonna e_1:

$$U^h AU e_1 = U^h Ax = U^h \lambda x = \lambda U^h x = \lambda e_1.$$

Perciò, $U^h AU = \begin{pmatrix} \lambda & b^t \\ 0 & A_{n-1} \end{pmatrix}$; per ipotesi induttiva, $A_{n-1} = Q_{n-1} T Q_{n-1}^h$ ed è sufficiente considerare $Q \coloneqq UQ_n$, dove $Q_n = \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix}$. □

2.3.1 Forma di Schur delle matrici normali

Data una matrice hermitiana A, calcolando la sua forma di Schur si ottiene $A = QTQ^h$, da cui $QT^h Q^h = A^h = A$, cioè $T = T^h$, che significa che T è diagonale e reale. In particolare, gli autovalori di A sono reali e in loro corrispondenza si può trovare una base ortonormale. Se A è antihermitiana, con lo stesso procedimento si verifica che T è diagonale con elementi immaginari puri. Se A è antisimmetrica, in particolare è antihermitiana; se ha dimensione dispari, c'è almeno un autovalore reale, che deve essere anche immaginario puro, quindi è 0 e la matrice è singolare.

Definizione 2.17. Una matrice A si dice *normale* se $A^h A = AA^h$.

Le matrici normali sono importanti perché ammettono buone proprietà di stabilità; si vedrà che questo è così perché gli autovettori di una matrice normale sono ortogonali, mentre le matrici più difficili da trattare sono quelle che hanno autovettori molto vicini. Tra le classi di matrici già viste, le matrici hermitiane, antihermitiane, unitarie sono normali. Ci si può chiedere com'è fatta la forma di Schur di una matrice normale.

Teorema 2.18. *Una matrice A è normale se e solo se esistono Q unitaria e D diagonale tali che $A = QDQ^h$ (o equivalentemente, se e solo se ha autovettori ortogonali).*

Dimostrazione.

($\Leftarrow$) Se $A = QDQ^h$, per la commutatività delle matrici diagonali, si ottiene

$$\begin{aligned} A^h A &= QD^h Q^h QDQ^h = QD^h DQ^h = \\ &= QDD^h Q^h = QDQ^h QD^h Q^h = AA^h. \end{aligned}$$

($\Rightarrow$) Se A è normale, si considera la forma di Schur $A = QTQ^h$; procedendo come prima, si ottiene $QTT^h Q^h = QT^h TQ^h$, da cui si ha che T è normale. Essendo anche triangolare, si può dimostrare per induzione che questo

implica che T è diagonale: se $n = 1$, non c'è nulla da dimostrare; se è vero per matrici di dimensione $n-1$ e T ha dimensione n, si controlla l'entrata $(1,1)$ di $T^hT = TT^h$: a sinistra si ottiene $\overline{t_{1,1}}t_{1,1} = |t_{1,1}|^2$, mentre a destra $\sum_{j=1}^{n} |t_{1,j}|^2$, perciò $t_{1,j} = 0$ per $1 < j \leq n$. Allora la sottomatrice costituita dalle ultime $n-1$ righe e $n-1$ colonne è normale e si può usare l'ipotesi induttiva. □

Nel caso delle matrici unitarie, questo significa che la matrice ha autovettori ortonormali che corrispondono ad autovalori di modulo 1.

Esercizio 2.19. Data una matrice normale A e un polinomio $P(A) \coloneqq a_0 I + a_1 A + \cdots + a_n A^n$, capire se $P(A)$ è normale.

3 Norme

09/10/2007

3.1 Nozioni generali

Definizione 3.1. Una *norma vettoriale* è un'applicazione $\|\bullet\| : \mathbb{C}^n \to \mathbb{R}$ tale che:

1. $\|x\| \geq 0$ per ogni $x \in \mathbb{C}^n$;
2. $\|x\| = 0$ se e solo se $x = 0$;
3. $\|\alpha x\| = |\alpha| \, \|x\|$ per ogni $\alpha \in \mathbb{C}$, $x \in \mathbb{C}^n$;
4. $\|x + y\| \leq \|x\| + \|y\|$ per ogni $x, y \in \mathbb{C}^n$.

La norma relativa alla lunghezza euclidea si denota con $\|\bullet\|_2$ ed è definita da $\|x\|_2 \coloneqq \sqrt{\sum_{i=1}^{n} |x_i|^2}$. Per verificare che questa è una norma, l'unico punto complicato da dimostrare è la disuguaglianza triangolare, cioè l'ultima condizione; è conveniente usare la disuguaglianza di Cauchy-Schwartz. Come nel caso della lunghezza euclidea, si verifica che un qualunque prodotto scalare induce una norma $\|x\| \coloneqq \sqrt{\langle x, x \rangle}$. Dentro $\mathbb{R}^2$, si denota con S_2 l'insieme dei vettori di norma unitaria rispetto alla norma euclidea; chiaramente questa è una circonferenza attorno all'origine.

Sul modello della norma euclidea, si definisce $\|x\|_1 \coloneqq \sum_{i=1}^{n} |x_i|$; anche questa è una norma. Ancora, si definisce $\|x\|_\infty \coloneqq \max\{\, |x_i| \mid 1 \leq i \leq n \,\}$. I corrispondenti insiemi di norma unitaria in $\mathbb{R}^2$ sono S_1, un quadrato attorno all'origine con vertici nei punti $(0, \pm 1)$ e $(\pm 1, 0)$, e S_∞, un quadrato attorno all'origine con vertici nei punti $(\pm 1, \pm 1)$ (figura 2). In generale, si definisce la norma $\|\bullet\|_p$ con

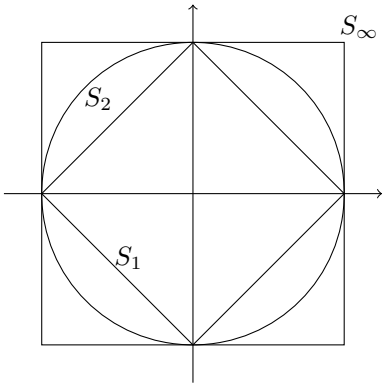
$$\|x\|_p \coloneqq \left(\sum |x_i|^p \right)^{1/p}.$$

Per $p \geq 1$, questa è una norma, mentre per $p < 1$ non lo è.

A partire dalla disuguaglianza triangolare, si sostituisce x con $a - b$ e y con b; si ottiene $\|a\| \leq \|a - b\| + \|b\|$. Ancora, si può manipolare ottenendo $\|a\| - \|b\| \leq \|a - b\|$; scambiando di nome le variabili, si ottiene

$$\|b\| - \|a\| \leq \|-(a - b)\| = |-1| \, \|a - b\| = \|a - b\| .$$

Figura 2: Insiemi di norma unitaria



In definitiva,

$$\begin{aligned} |\|a\| - \|b\|| \leq \|a-b\| = \left\| \sum (a_i - b_i) e_i \right\| \leq \\ \leq \sum \|(a_i - b_i) e_i\| = \sum |a_i - b_i| \, \|e_i\| \, . \end{aligned}$$

Se $|a_i - b_i| \leq \delta$, questa relazione asserisce che $|\|a\| - \|b\|| \leq \delta \sum \|e_i\|$. Quindi per ogni $\varepsilon > 0$, esiste $\delta \coloneqq \frac{\varepsilon}{\sum \|e_i\|}$ tale che per ogni $a, b \in \mathbb{C}^n$, $|a_i - b_i| \leq \delta$ implica $|\|a\| - \|b\|| \leq \varepsilon$, cioè la norma è uniformemente continua.

L'uniforme continuità permette di dimostrare la seguente.

Proposizione 3.2. *Per ogni coppia di norme su $\mathbb{C}^n$, $\|\bullet\|'$ e $\|\bullet\|''$, esistono $\alpha, \beta > 0$ tali che per ogni $x \in \mathbb{C}^n$, $\alpha \|x\|'' \leq \|x\|' \leq \beta \|x\|''$ (cioè le proprietà di convergenze tra le varie norme sono le stesse, e le maggiorazioni con una norma valgono anche per le altre norme a meno di conoscere le costanti α e β).*

Dimostrazione. Per $x = 0$ la disuguaglianza chiaramente vale; si suppone inoltre $\|\bullet\|'' = \|\bullet\|_\infty$; se si dimostra con questo assunto, varrà per qualsiasi coppia di norme: infatti, se si hanno le stime

$$\begin{aligned} \alpha' \|x\|_\infty \leq \|x\|' \leq \beta' \|x\|_\infty \, , \\ \alpha'' \|x\|_\infty \leq \|x\|'' \leq \beta'' \|x\|_\infty \, , \end{aligned}$$

allora si avrà anche ${}^{\alpha'}\!/_{\beta''} \|x\|'' \leq \|x\|' \leq {}^{\beta'}\!/_{\alpha''} \|x\|''$.

Ora, l'insieme S_∞ dentro $\mathbb{C}^n$ è chiuso e limitato, quindi la funzione $\|\bullet\|'$ avrà un massimo e un minimo su S_∞; si denotino rispettivamente con β e α. Allora $\alpha \leq \|{}^{x}\!/_{\|x\|_\infty}\|' \leq \beta$, da cui la tesi. □

Esempio 3.3. Si hanno le seguenti:

$$\begin{aligned} \|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty \, , \\ \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty \, . \end{aligned}$$

10/10/2007

Esercizio 3.4. Tutte le norme che si sono viste finora soddisfano la relazione $\|(x_1, \dots, x_n)\| = \|(|x_1|, \dots, |x_n|)\|$; non è detto che una norma debba soddisfare questa condizione, e quelle che lo fanno sono dette norme *assolute*. Ancora, non è detto che se $|x_i| \geq |y_i|$ per ogni i allora $\|x\| \geq \|y\|$, come avviene per le norme viste; quelle che soddisfano la relazione sono dette norme *monotone*. Dimostrare l'esistenza di norme non assolute e norme non monotone e mostrare che i due concetti coincidono.

3.2 Norme matriciali

Definizione 3.5. Una *norma matriciale* è un'applicazione $\|\bullet\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ tale che:

1. $\|A\| \geq 0$ per ogni $A \in \mathbb{C}^{n\times n}$;
2. $\|A\| = 0$ se e solo se $A = 0$;
3. $\|\alpha A\| = |\alpha| \, \|A\|$ per ogni $\alpha \in \mathbb{C}$, $A \in \mathbb{C}^{n\times n}$;
4. $\|A + B\| \leq \|A\| + \|B\|$ per ogni $A, B \in \mathbb{C}^{n\times n}$;
5. $\|AB\| \leq \|A\| \, \|B\|$ per ogni $A, B \in \mathbb{C}^{n\times n}$.

L'ultima proprietà, che si aggiunge alle richieste di una norma vettoriale, è detta *submoltiplicatività*.

Se si vedono le matrici come applicazioni lineari da $\mathbb{C}^n$ in sé, fissata una norma vettoriale, si può considerare l'applicazione che manda una matrice A nel numero $\|A\| \coloneqq \max\{\, \|Ax\| \mid \|x\| = 1 \,\}$. L'applicazione risultante è ben definita, perché si fa il massimo di una funzione continua su un chiuso (perché preimmagine di $\{1\}$ tramite una funzione continua) e limitato (per la maggiorazione con la norma infinito), misura quanto A amplifica i vettori di norma unitaria ed è detta *norma matriciale indotta* dalla norma vettoriale $\|\bullet\|$. Per verificare che questa è davvero una norma matriciale, l'unico problema è la submoltiplicatività: per $AB = 0$ la disuguaglianza è soddisfatta, quindi si può supporre $AB \neq 0$; si ha $\|AB\| = \max\{\, \|ABx\| \mid \|x\| = 1 \,\} = \|ABy\|$ per qualche y di norma 1; inoltre $z \coloneqq By \neq 0$, altrimenti $AB = 0$; ora,

$$\|ABy\| = \|Az\| = \frac{\|Az\|}{\|z\|} \|By\| = \left\| A \frac{z}{\|z\|} \right\| \|By\| \leq \|A\| \, \|B\| \,,$$

perché $z/\|z\|$ e y hanno norma 1, quindi le due norme saranno minori del massimo nascosto nella norma matriciale.

Si ricavano alcune proprietà delle norme matriciali.

- Poiché $II = I$, presa una qualsiasi norma si ha $\|I\| = \|II\| \leq \|I\|^2$, cioè $\|I\| \geq 1$. Inoltre, per una qualsiasi norma indotta, $\|I\| = 1$.
- Se A è non singolare, $AA^{-1} = I$, da cui $1 \leq \|I\| \leq \|A\| \left\|A^{-1}\right\|$, cioè $\left\|A^{-1}\right\| \geq \|A\|^{-1}$.
- Applicando la submoltiplicatività varie volte, $\left\|A^k\right\| \leq \|A\|^k$.

- Lavorando con una norma matriciale indotta, se $y = Ax$, si ha $\|y\| \leq \|A\| \, \|x\|$. In particolare, se x è un autovettore di autovalore λ e norma 1, allora $Ax = \lambda x$, da cui $|\lambda| = \|Ax\| \leq \|A\| \, \|x\| = \|A\|$. Si ottiene che per una norma matriciale indotta, $\|A\| \geq |\lambda|$ per ogni autovalore di A.

Si definisce $\|A\|_p \coloneqq \max \left\{ \|Ax\|_p \mid \|x\|_p = 1 \right\}$; allora vale

$$\|A\|_1 = \max \left\{ \sum_{i=1}^{n} |a_{i,j}| \mid 1 \leq j \leq n \right\}.$$

Infatti, se $\|x\|_1 = 1$, vale

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{i,j} x_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}| \, |x_j| = \\ &= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{i,j}| \leq \sum_{j=1}^{n} |x_j| \max \left\{ \sum_{i=1}^{n} |a_{i,k}| \mid 1 \leq k \leq n \right\} = \\ &= \max \left\{ \sum_{i=1}^{n} |a_{i,k}| \mid 1 \leq k \leq n \right\}; \end{aligned}$$

si è ottenuto che $\|A\|_1 \leq \max \{ \sum_{i=1}^{n} |a_i, k| \mid 1 \leq k \leq n \}$. Per ottenere l'uguaglianza, si esibisce un vettore per cui vale: se h è la colonna che massimizza, allora Ae_h è il vettore corrispondente alla colonna h e

$$\|Ae_h\|_1 = \max \left\{ \sum_{i=1}^{n} |a_{i,k}| \mid 1 \leq k \leq n \right\}.$$

Le proprietà analoghe per $\|\bullet\|_\infty$ e $\|\bullet\|_2$ sono le seguenti:

$$\begin{aligned} \|A\|_\infty &= \max \left\{ \sum_{j=1}^{n} |a_{i,j}| \mid 1 \leq i \leq n \right\}, \\ \|A\|_2 &= \sqrt{\rho(A^h A)}, \end{aligned}$$

dove ρ è il *raggio spettrale* di una matrice, cioè il massimo dei moduli dei suoi autovalori. La proprietà di $\|\bullet\|_1$ è analoga a quella di $\|\bullet\|_\infty$, ma si prende il massimo al variare delle colonne, mentre per la seconda il massimo è al variare delle righe. In particolare, $\|A\|_1 = \|A^t\|_\infty$. Per $\|\bullet\|_2$, si osserva che calcolare gli autovalori di una matrice è complicato, perciò questa norma non ha un uso pratico. Se A è hermitiana, gli autovalori di $A^h A$ sono gli autovalori di A elevati al quadrato, da cui $\|A\|_2 = \rho(A)$.

3.3 Norme matriciali e raggio spettrale

Ci si può chiedere quali siano i legami tra le norme matriciali e il raggio spettrale.

- Il raggio spettrale è una norma per le matrici hermitiane, ma non è una norma per le matrici in generale (per esempio, $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ è una matrice non nulla con raggio spettrale nullo).

- Si può riformulare una proprietà già vista in precedenza per le norme indotte: da $\|A\| \geq |\lambda|$ per ogni autovalore λ di A, si deduce che $\|A\| \geq \rho(A)$.
- La proprietà più importante è la seguente: per ogni A e per ogni $\varepsilon > 0$, esiste una norma matriciale indotta $\|\bullet\|$ tale che $\rho(A) \leq \|A\| \leq \rho(A)+\varepsilon$. In particolare, il raggio spettrale non è una norma, ma è un limite di norme (è il limite inferiore tra tutte le norme indotte).

Si dimostrerà quest'ultima in vari passi. Innanzitutto, se S è una matrice, e $\|\bullet\|$ è una norma vettoriale, allora l'applicazione $x \mapsto \|Sx\| =: \|x\|_S$ è una norma se e solo se S è non singolare. La norma matriciale indotta da $\|\bullet\|_S$ è data da $\|A\|_S = \max\{\, \|SAx\| \mid \|Sx\| = 1 \,\}$; posto $y \coloneqq Sx$, si ha $\|A\|_S = \max\left\{\, \left\|SAS^{-1}y\right\| \mid \|y\| = 1 \,\right\} = \left\|SAS^{-1}\right\|$. Ora si può dimostrare la proprietà: data A, si considera la sua forma di Jordan $J \coloneqq V^{-1}AV$; se si prende la matrice $D_\varepsilon \coloneqq \operatorname{diag}(1, \varepsilon, \ldots, \varepsilon^{n-1})$, la matrice $J' \coloneqq D_\varepsilon^{-1} J D_\varepsilon$ è una matrice con $j'_{i,j} = \varepsilon^{-i+1} j_{i,j} \varepsilon^{j-1}$, cioè sulla diagonale si hanno sempre gli autovalori, mentre sulla sopradiagonale si ha ε al posto di 1. A questo punto, se λ è un autovalore massimo in modulo, $\|J'\|_\infty \leq |\lambda| + \varepsilon$ (che è il massimo tra gli autovalori con blocchi 1×1 e gli autovalori sommati a ε per gli autovalori con blocchi di dimensione maggiore). Quindi si ha $\|J'\|_\infty = \|A\|_{D_\varepsilon^{-1} V^{-1}} \leq \rho(A) + \varepsilon$. In particolare, il raggio spettrale è una norma per quella classe di matrici per cui gli autovalori di modulo massimo compaiono in blocchi 1×1.

Presa una norma matriciale, si ha $\lim \left(\left\|A^k\right\|\right)^{1/k} = \rho(A)$. Per dimostrarlo, si osserva che anche per le norme matriciali vale la proprietà per cui una norma è maggiorata e minorata da una qualsiasi altra norma a meno di moltiplicare per una costante. Si ha quindi $\alpha \left\|A^k\right\|'' \leq \left\|A^k\right\|' \leq \beta \left\|A^k\right\|''$, da cui

$$\sqrt[k]{\alpha} \sqrt[k]{\|A^k\|''} \leq \sqrt[k]{\|A^k\|'} \leq \sqrt[k]{\beta} \sqrt[k]{\|A^k\|''}.$$

Poiché le radici k-esime di numeri non nulli tendono a 1, per il teorema dei due carabinieri il limite al variare della norma è lo stesso e si può lavorare con la norma infinito. Ora, se $J \coloneqq V^{-1}AV$ è una forma di Jordan di A, $J^k = V^{-1}A^kV$ e basta stimare $\left\|J^k\right\|_\infty$. Posto H la matrice con la sola sopradiagonale uguale a 1, si ha che un blocco di Jordan è dato da $\lambda I + H$; poiché I commuta con H, si può applicare il binomio di Newton e $(\lambda I + H)^k = \sum_{i=0}^{k} \binom{k}{i} H^i \lambda^{k-i}$; ma la sommatoria si può fermare a $m - 1$, dove m è la dimensione del blocco di Jordan, perché H è nilpotente.

Esercizio 3.6. Terminare la dimostrazione, calcolando la forma di J^k e la sua norma infinito.

4 Sistemi lineari: metodi diretti

23/10/2007

Si studiano i sistemi lineari del tipo $Ax = b$, dove $A \in \mathbb{C}^{n \times n}$ è non singolare e $b \in \mathbb{C}^n$. Si svolgerà questo studio in tre parti: prima si studierà il condizionamento, cioè la dipendenza della soluzione da piccole perturbazioni del problema; successivamente si vedranno i metodi di risoluzione diretti; infine, si studieranno dei metodi di risoluzione detti iterativi, dove non si trova la soluzione ma una successione che vi converge.

4.1 Condizionamento

Si considera una perturbazione del vettore dei termini noti: al posto di b si prende $b+\delta_b$: se $A(x+\delta_x) = b+\delta_b$, si vuole studiare come varia δ_x al variare di δ_b. Si può intraprendere questo studio con gli strumenti dell'analisi degli errori visti all'inizio, ma l'analisi sarebbe pesante. Invece, si useranno le proprietà delle norme per avere delle stime sugli errori.

Innanzitutto, da $Ax = b$, si ottiene anche $A\delta_x = \delta_b$, cioè $\delta_x = A^{-1}\delta_b$. Questo implica $\|\delta_x\| \leq \left\|A^{-1}\right\| \|\delta_b\|$. Interessa però l'errore relativo, cioè $\|\delta_x\|/\|x\|$ (si suppone $b \neq 0$, da cui $x \neq 0$). Per raggiungere una stima dell'errore relativo, si osserva che $\|b\| = \|Ax\| \leq \|A\| \|x\|$ e che questa stima e la precedente sono le migliori possibili, cioè il segno di uguaglianza viene effettivamente raggiunto per qualche valore di x. A partire da queste due disuguaglianze, si ottiene

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\| \|\delta_b\|}{\|A\|^{-1} \|b\|} = \|A\| \left\|A^{-1}\right\| \frac{\|\delta_b\|}{\|b\|}.$$

Si ottiene che l'errore relativo non può superare l'errore relativo di b per il numero $\|A\| \left\|A^{-1}\right\|$, che ha il significato di un coefficiente di amplificazione; data la sua importanza, gli viene assegnato un simbolo, $\mu(A)$, e viene detto *numero di condizionamento* di A.

Con calcoli analoghi, si analizza il condizionamento nel caso che sia perturbata anche la matrice del sistema, cioè se si ha il sistema $(A+\delta_A)(x+\delta_x) = (b+\delta_b)$: si ottiene che se $\|\delta_A\| \left\|A^{-1}\right\| < 1$ (cioè se la perturbazione è piccola rispetto alla matrice), anche la matrice perturbata è non singolare, e

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\mu(A)}{1 - \|\delta_A\| \|A^{-1}\|} \left(\frac{\|\delta_b\|}{\|b\|} + \frac{\|\delta_A\|}{\|A\|} \right).$$

Più la perturbazione di A è piccola, più l'analisi si riduce al caso precedente; in pratica si può pensare che $\mu(A)$ sia il coefficiente di amplificazione anche in questo caso.

Esempio 4.1. La matrice H di entrate $h_{i,j} \coloneqq (i+j-1)^{-1}$ è detta *matrice di Hilbert*. Si calcola che il numero di condizionamento di H cresce in modo esponenziale con la dimensione di H.

Esempio 4.2. Sia V la matrice di entrate $v_{i,j} \coloneqq x_{i-1}^{j-1}$, detta *matrice di Vandermonde*. Se gli x_i sono numeri reali, V è mal condizionata, e il suo numero di condizionamento cresce ancora in modo esponenziale con la dimensione. Se come x_i si scelgono n numeri complessi sulla circonferenza unitaria equispaziati (come per le radici delle unità) il numero di condizionamento è 1.

I risultati trovati sono stime in norma, mentre in generale si cercherebbe un risultato componente per componente: in particolare, $\|\delta_x\|/\|x\| \leq \varepsilon$ non implica $|\delta_{x_i}/x_i| \leq \varepsilon$. Per esempio, se $x \coloneqq \binom{1}{\varepsilon}$ e $\delta_x \coloneqq \binom{\varepsilon}{\varepsilon}$, in norma infinito l'errore relativo è uguale a ε; tuttavia, l'errore relativo della seconda componente è 1.

In alcuni casi, i numeri di condizionamento sono facili da calcolare. Se Q è una matrice unitaria, $\|Q\|_2$ e $\left\|Q^{-1}\right\|_2$ sono entrambe 1, quindi $\mu(Q) = 1$. Se A è hermitiana, $\|A\|_2 = \rho(A)$ e $\left\|A^{-1}\right\|_2 = \rho(A^{-1})$; se gli autovalori di A sono $\lambda_1, \ldots, \lambda_n$ con $|\lambda_1| \geq \cdots \geq |\lambda_n|$, si ha $\|A\|_2 = |\lambda_1|$ e $\left\|A^{-1}\right\|_2 = |\lambda_n|^{-1}$, cioè $\mu(A) = |\lambda_1/\lambda_n|$.

4.2 Metodi diretti

Se A è unitaria, cioè $AA^h = I$, risolvere il sistema è semplice: $x = A^{-1}b = A^h b$, e il costo è asintoticamente n^2. Anche se A è una matrice triangolare inferiore è facile risolvere il sistema, basta usare la sostituzione in avanti: si ottiene subito $x_1 = {}^{b_1}/_{a_{1,1}}$ e le successive incognite si ottengono come

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j}{a_{i,i}};$$

il costo computazionale è ancora asintoticamente n^2. Si può dimostrare che questo algoritmo è numericamente stabile all'indietro (se il numero di condizionamento è alto, questo comunque non aiuta). *24/10/2007*

Quindi se la matrice del sistema appartiene a particolari classi, risolvere il sistema è semplice. Inoltre, se si vuole risolvere il sistema $Ax = b$, dove $A = BC$, si può risolvere il sistema costituito dai sottosistemi $Cx = y$ e $By = b$: in particolare, se A si decompone come prodotto di due matrici "semplici", anche il sistema $Ax = b$ si potrà risolvere in modo semplice. L'idea quindi è di trovare fattorizzazioni della matrice A mediante matrici triangolari o unitarie.

4.3 Fattorizzazioni LU e QR

Particolare importanza assumono le fattorizzazioni del tipo $A = LU$, dove L è triangolare inferiore con diagonale uguale a 1 e U è triangolare superiore, e quelle del tipo $A = QR$, dove Q è unitaria e R è triangolare destra (cioè, superiore).

Una sottomatrice si dice *principale* se è definita dagli stessi indici per le righe e per le colonne; una sottomatrice principale è detta *di testa* se gli indici sono $1, \dots, k$. Data A, la sottomatrice principale di testa di dimensione k si indicherà con A_k.

Teorema 4.3 (esistenza e unicità della fattorizzazione LU)**.** *Se $A \in \mathbb{C}^{n\times n}$ è tale che tutte le sue sottomatrici principali di testa A_k con $1 \le k < n$ sono non singolari, allora esiste ed è unica la fattorizzazione $A = LU$.*

Dimostrazione. Si dimostrerà costruttivamente per induzione su n. Se $n = 1$ non c'è nulla da dimostrare. Se il teorema vale per ogni matrice di dimensione minore di n, basta dimostrare che esistono due matrici di dimensione $n-1$, L_{n-1} e U_{n-1} rispettivamente triangolare inferiore con diagonale uguale a 1 e triangolare superiore, due vettori v e u di lunghezza $n-1$ e uno scalare α tali che

$$A = \begin{pmatrix} A_{n-1} & b \\ a^t & a_{n,n} \end{pmatrix} = \begin{pmatrix} L_{n-1} & 0 \\ v^t & 1 \end{pmatrix} \begin{pmatrix} U_{n-1} & u \\ 0 & \alpha \end{pmatrix}.$$

Perché accada ciò, si hanno le condizioni $A_{n-1} = L_{n-1}U_{n-1}$, $b = L_{n-1}u$, $a^t = v^t U_{n-1}$ e $a_{n,n} = v^t u + \alpha$. La prima condizione è verificata in modo unico per ipotesi induttiva; le successive due permettono di ricavare v e u come soluzioni di sistemi lineari di dimensione $n-1$ non singolari e dall'ultima si ricava α. □

4.4 Matrici elementari

Questa è una dimostrazione costruttiva, quindi può essere usata per calcolare la fattorizzazione; tuttavia ci sono metodi migliori. Si definisce una *matrice elementare* come una matrice $E \coloneqq I - \sigma uv^h$, dove $\sigma \in \mathbb{C}$ e u e v sono vettori

colonna di lunghezza n. Il rango della matrice uv^h è necessariamente 1, dato che tutte le colonne sono parallele a u e tutte le righe sono parallele a v. Geometricamente è evidente: $(uv^h)(x) = u(v^h x) \in \langle u \rangle$, cioè tutti i vettori hanno immagine parallela a u (e il nucleo dell'applicazione è costituito dai vettori ortogonali a v^h). Se E è elementare, allora $Ex = x - \sigma u v^h(x)$, da cui $Eu = u(1 - \sigma v^h u)$ e $Ex = x$ se $v^h x = 0$.

Esercizio 4.4. Se E è una matrice elementare non singolare, allora la sua inversa è ancora elementare, definita dagli stessi vettori u e v e scalare

$$\tau := \frac{\sigma}{(v^h u)\sigma - 1}.$$

In particolare, si osserva che il costo del calcolo dell'inversa di una matrice elementare è lineare (il prodotto $v^h u$ ha costo asintoticamente uguale a n) e che E è non singolare se e solo se $(v^h u)\sigma \neq 1$.

Un'altra proprietà interessante delle matrici elementari è che per ogni $x, y \in \mathbb{C}^n$ non nulli, esiste una matrice elementare E non singolare tale che $Ex = y$. Infatti, ponendo $y = (I - \sigma u v^h)x$, si ricava che v non deve essere ortogonale a x; dalla relazione si ha anche che $x - \sigma u(v^h x) = y$, da cui si ha che

$$\sigma u = \frac{x - y}{v^h x}.$$

Inoltre questi valori soddisfano la condizione della non singolarità a meno di scegliere v non ortogonale nemmeno a y.

4.5 Decomposizioni con matrici elementari

Sia ora $A = A_1 = (a, B)$ una matrice, a la sua prima colonna. Per quanto detto, esiste una matrice elementare E_1 tale che $E_1 a = (\alpha_1, 0, \ldots, 0)^t$, e si pone $A_2 := E_1 A$; questa è una matrice che ha $E_1 a$ come prima colonna e $E_1 B$ nel resto. In pratica, A_2 è triangolare sulla prima colonna. Chiamata C_2 la sottomatrice principale di coda di A_2 di dimensione $n-1$, si può iterare il processo, trovando una matrice elementare $\hat{E}_2$ tale che $\hat{E}_2 C$ è triangolare nella prima colonna. Se E_2 è $\hat{E}_2$ a cui viene aggiunta una prima riga e una prima colonna come nella matrice identità, allora $A_3 := E_2 E_1 A_1$ è triangolare nelle prime due colonne, e si dimostra facilmente che anche E_2 è elementare. Iterando il procedimento, si ha una successione A_k con $A_{k+1} = E_k A_k$, per $k = 1, \ldots, n$. Infine, $A_n = E_{n-1} \cdots E_1 A$ è triangolare, quindi $A = E_1^{-1} \cdots E_{n-1}^{-1} A_n$ è una fattorizzazione. Se si scelgono le E_k come matrici unitarie, si ha una fattorizzazione QR; se le E_k sono triangolari inferiori, si ha una fattorizzazione LU. Per i sistemi lineari, $Ax = b$, si fa lo stesso procedimento: $E_{n-1} \cdots E_1 A x = E_{n-1} \cdots E_1 b$ è un sistema con matrice triangolare.

4.5.1 Decomposizione LU con matrici di Gauss

Si esaminano ora le due fattorizzazioni. Le *matrici elementari di Gauss* sono matrici tali che $v = e_1$ e $u_1 = 0$. In questo caso, la matrice elementare è

$$E = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -u_2 & & & \\ \vdots & & I_{n-1} & \\ -u_n & & & \end{pmatrix}.$$

In particolare, $Ex = (x_1, x_2 - u_2x_1, \ldots, x_n - u_nx_1)^t$; questo vettore può essere del tipo $(\alpha, 0, \ldots, 0)^t$ con $\alpha \neq 0$ se e solo se $x_1 \neq 0$. Tuttavia, se ci si pone nelle ipotesi del teorema di esistenza e unicità, il primo passo si può fare, perché $(a_{1,1})$ è una sottomatrice principale di testa, quindi non singolare per ipotesi; con un'analisi un po' più attenta, si dimostra che sotto quelle ipotesi si può costruire la fattorizzazione LU con matrici elementari di Gauss. Il processo che si ottiene si dice *metodo di eliminazione gaussiana* e si può usare per calcolare la fattorizzazione LU o per risolvere un sistema lineare.

4.5.2 Decomposizione QR con matrici di Householder

Un'altra possibilità è usare le *matrici di Householder*, ovvero matrici hermitiane $P \coloneqq I - \sigma uu^h$ con $\sigma \in \mathbb{R}$. Un rapido calcolo mostra che P è una matrice hermitiana se e solo se $\sigma = 0$ o $\sigma = {}^2/_{u^hu}$. L'interpretazione geometrica di una matrice di Householder è la seguente: i vettori ortogonali a u sono mandati in sé stessi, mentre quelli paralleli a u vengono mandati nel loro opposto: in pratica, una matrice di Householder rappresenta una riflessione per l'iperpiano ortogonale a u. Si deve però trovare un modo, dato $x \in \mathbb{C}^n$ non nullo, di trovare P tale che $Px = (\alpha, 0, \ldots, 0)^t$. Necessariamente, dato che P è unitaria, $|\alpha| = \|x\|_2$. Inoltre, dato che P è hermitiana, $x^hPx \in \mathbb{R}$, cioè $\overline{x_1}\alpha$ deve essere reale; in definitiva,

$$\alpha = \begin{cases} \pm \|x\|_2 \frac{x_1}{|x_1|} & \text{se } x_1 \neq 0 \\ \pm \|x\|_2 & \text{altrimenti.} \end{cases}$$

Per calcolare u, da $Px = (I - \sigma uu^h)x = (\alpha, 0, \ldots, 0)^t$, si ha $\sigma(u^hx)u = x - (\alpha, 0, \ldots, 0)^t$, da cui si può scegliere $u(x_1 - \alpha, x_2, \ldots, x_n)^t$. La scelta per il segno di α si fa in modo che il calcolo di u sia il più stabile possibile, cioè che non si abbia cancellazione: si sceglie il segno negativo in modo che $u_1 = x_1 + \|x\|_2 {}^{x_1}/_{|x_1|}$. Il calcolo di P è lineare nella dimensione.

4.6 Costo del calcolo della fattorizzazione LU

30/10/2007

Si costruisce la fattorizzazione LU con le matrici elementari di Gauss. La matrice A_k costruita per induzione, ha una struttura particolare: infatti è triangolare nelle prime $k-1$ colonne, cioè $a_{i,j}^{(k)} = 0$ per $j < k, i$. La matrice E_k che si deve moltiplicare per A_k in modo da ottenere A_{k+1} è la matrice identità con $(0, \ldots, 0, 1, -m_{k+1,k}, \ldots, -m_{n,k})^t$ al posto della colonna k. Infatti, prendendo questa E_k, con $m_{i,k} \coloneqq a_{i,k}^{(k)} {a_{k,k}^{(k)}}^{-1}$, si ha:

$$a_{i,j}^{(k+1)} = \begin{cases} 0 & \text{se } j = k,\ i > k; \\ a_{i,j}^{(k)} - m_{i,k}a_{k,j}^{(k)} & \text{se } i, j > k; \\ a_{i,j}^{(k)} & \text{altrove.} \end{cases}$$

Il costo computazionale del calcolo di A_{k+1} è dato da:

- $n - k$ divisioni per trovare $m_{i,k}$ con $k < i \leq n$;
- $2(n-k)^2$ operazioni (metà moltiplicazioni e metà differenze) per trovare $a_{i,j}$ con $k < i, j \leq n$.

Il costo totale per trovare U è quindi

$$\sum_{k=1}^{n-1}((n-k)+2(n-k)^2) = \sum_{k=1}^{n-1}(k+2k^2) = \frac{n(n-1)}{2} + 2\frac{2n^3+3n^2+n}{6} = \\ = \frac{2}{3}n^3 + O(n^2).$$

Se la matrice A è una matrice a banda (cioè è nulla al di fuori di alcune diagonali attorno a quella principale), si verifica che il costo computazionale è quadratico (se la larghezza della banda è costante al variare di n).

Esercizio 4.5. Le matrici di Hessenberg sono matrici con entrate nulle per $i > j+1$. Per queste matrici, il calcolo della fattorizzazione è quadratico.

Ora, $A_n = E_{n-1}\cdots E_1 A$ è triangolare superiore, e sarà la matrice U. La matrice L invece è $E_1^{-1}\cdots E_{n-1}^{-1}$ (che infatti è triangolare inferiore in quanto prodotto di triangolari inferiori). Poiché le E_i sono matrici elementari di Gauss, le loro inverse si ottengono cambiando di segno agli elementi fuori dalla diagonale, cioè prendendo $m_{i,k}$ come entrate invece che $-m_{i,k}$. Si verifica facilmente che il prodotto di queste inverse è la matrice che ha entrate $m_{i,k}$ nel posto (i,k) se $i > k$, e 1 sulla diagonale: infatti, se $m_k \coloneqq (0,\dots,0,m_{k+1,k},\dots,m_{n,k})$, la matrice E_i è $I + m_k e_k^t$; di conseguenza, $E_1^{-1}E_2^{-1} = (I+m_1e_1^t)(I+m_2e_2^t) = I + m_1e_1^t + m_2e_2^t + \cancel{(m_1e_1^t)(m_2e_2^t)}$ dove il cancellamento si ottiene riparentesizzando in modo da calcolare all'inizio $e_1^t m_2 = 0$ (perché e_1^t stacca la prima componente di m_2, che è nulla); allo stesso modo, proseguendo, tutti i prodotti misti si annullano per lo stesso motivo.

4.7 Stabilità numerica del calcolo della fattorizzazione LU

Il punto debole dell'algoritmo è la divisione per $a_{k,k}^{(k)}$: ovviamente se è nullo non si può proseguire; ma anche se è molto prossimo a 0, gli errori possono essere molto amplificati. Bisogna quindi assicurarsi a ogni passo che questi elementi non solo siano non nulli, ma che abbiano anche modulo maggiore possibile. Questo risultato si può vedere usando l'analisi all'indietro, i cui risultati vengono ora esposti.

Esercizio 4.6. Se A è triangolare inferiore, e $\tilde{x}$ è il risultato calcolato con la sostituzione in avanti nell'aritmetica floating point, allora $\tilde{x}$ soddisfa $(A+\delta_A)\tilde{x} = b$, con $|\delta_A| \le nu\,|A|$, dove la disuguaglianza si intende termine a termine.

Proposizione 4.7. *Sia A una matrice non singolare; calcolando la fattorizzazione LU in aritmetica floating point, si trovano delle matrici $\tilde{L}$ e $\tilde{U}$ che costituiscono una fattorizzazione LU della matrice $A+\delta_A$, con*

$$|\delta_A| \le 2nu(|A| + |L|\,|U|).$$

La proposizione dice che se i valori che si ottengono calcolando L e U sono grandi, non c'è garanzia che l'errore sia limitato.

4.8 Pivoting della fattorizzazione LU

31/10/2007

Gli elementi che assumono il ruolo principale sono gli $a_{k,k}^{(k)}$, che vengono chiamati *pivot*. Se a un certo passo l'elemento pivot è nullo, il metodo per la fattorizzazione fallisce. Se però lo scopo è quello di risolvere un sistema $Ax = b$, si è

osservato che non è necessario calcolare $A = LU$, ma basta applicare le matrici elementari sia a A che a b; se a un certo passo il pivot è nullo, si può proseguire il metodo scambiando la riga k con una riga $j > k$ tale che $a_{j,k}^{(k)}$ non sia nullo (se lo sono tutti questi elementi, la matrice è singolare). Ci si può chiedere se questo metodo, che contempla lo scambio di righe in certe occasioni, corrisponde a una qualche fattorizzazione. La regola induttiva è $A_{k+1} = E_k P_k A_k$, dove E_k è la stessa matrice elementare mentre P_k è la matrice di permutazione che, eventualmente, scambia la riga k con una riga successiva. Arrivando alla fine si ottiene $A_n = E_{n-1} P_{n-1} \cdots E_1 P_1 A$, ma si possono trasportare tutte le matrici di permutazione alla fine, prima di A: infatti, per esempio, $P_2 E_1$ è la matrice elementare E_1 con la seconda riga scambiata con la j-esima riga, $j \geq 2$; ma se a questa nuova matrice si scambia la seconda colonna con la j-esima, si ritrova una matrice elementare, solo con la prima colonna permutata; in definitiva, $P_2 E_1 = \hat{E}_1 P_2$. Allo stesso modo si possono far commutare le matrici di permutazione con le matrici elementari, a meno di considerare le $\hat{E}_i$ al posto delle E_i. Cioè, $A_n = \hat{E}_{n-1} \cdots \hat{E}_1 (P_{n-1} \cdots P_1) A$; in altri termini,

$$P_{n-1} \cdots P_1 A = (\hat{E}_1^{-1} \cdots \hat{E}_{n-1}^{-1}) A_n,$$

cioè si è ottenuta una fattorizzazione LU della matrice A con le righe permutate (il che implica che la fattorizzazione LU esiste sempre a meno di permutazioni).

Si è visto che l'analisi all'indietro contiene un termine $2nu\,|L|\,|U|$, che a priori non si sa limitare. Però, gli elementi della matrice L sono $m_{i,k} = a_{i,k}^{(k)} {a_{k,k}^{(k)}}^{-1}$; se $a_{k,k}^{(k)}$ è piccolo, questi possono essere molto grandi; tuttavia, per migliorare la situazione, si può applicare ancora il metodo della permutazione delle righe, portando come pivot l'elemento della colonna di modulo massimo. Con questo accorgimento, il modulo di ogni elemento di L è minore o uguale a 1. Si deve però capire che cosa cambia negli elementi di U: da

$$\left| a_{i,j}^{(k+1)} \right| \leq \left| a_{i,j}^{(k)} \right| + \left| a_{k,j}^{(k)} \right|,$$

indicando con $a_M^{(k)}$ il massimo modulo di una entrata di A_k, nel caso peggiore si ha $a_M^{(k+1)} \leq 2 a_M^{(k)}$. Di conseguenza,

$$|u_{i,j}| \leq a_M^{(n)} \leq 2^{n-1} a_M^{(1)}.$$

In pratica, per molte matrici questa crescita esponenziale non si verifica; tuttavia, alcune matrici invece la presentano.

Esempio 4.8. Sia A la matrice

$$I - \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & 0 & 0 \\ 1 & \cdots & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Il massimo modulo è 1; il primo passo dell'eliminazione gaussiana è dato dalla matrice elementare con prima colonna con tutti 1; l'effetto sull'ultima colonna è quello di mettere 2 su tutta la colonna a meno del primo elemento. Di passo in passo, l'ultima colonna aumenta esponenzialmente; alla fine si ha in posizione (n, n) il numero 2^{n-1}. La maggiorazione quindi è la migliore possibile.

La situazione però si può ancora migliorare: invece di permutare solo le righe per ottenere come pivot l'elemento di massimo modulo nella colonna in esame, si mette come pivot l'elemento di massimo modulo tra tutta la sottomatrice in esame. La maggiorazione è la stessa ($|m_{i,k}| \leq 1$), ma in generale il valore sarà più piccolo perché il massimo è stato fatto su un insieme più grande. Quello che si può dimostrare per U è che

$$a_M^{(k)} \leq a_M^{(1)} \sqrt{k \prod_{j=2}^{k} j^{\frac{1}{j-1}}}.$$

Con strumenti informatici si scopre che la crescita di questa funzione è poco più che lineare; inoltre non si conoscono casi in cui l'effettiva crescita sia più che lineare. Questo metodo porta a una fattorizzazione del tipo $P_1 A P_2 = LU$, dove P_i sono matrici di permutazione.

Un'altra proprietà è che si interrompe il metodo se e solo se tutta la sottomatrice è nulla, cioè se si è trovata una base del nucleo della matrice. L'inconveniente è che i numeri che dovrebbero essere nulli in realtà sono nell'ordine della precisione di macchina: numericamente non si può determinare il rango della matrice[1].

Per l'implementazione della fattorizzazione LU, si osserva che si può effettuare l'algoritmo utilizzando solo n^2 aree di memoria, sovrascrivendo di volta in volta la parte triangolare superiore e utilizzando la parte triangolare inferiore per memorizzare gli elementi $m_{i,k}$.

4.9 Stabilità numerica del calcolo della fattorizzazione QR

Per la risoluzione di sistemi tramite la fattorizzazione QR, se si ha il sistema $Ax = b$, l'analisi all'indietro mostra che le matrici effettivamente trovate, $\tilde{Q}$ e $\tilde{R}$ danno una soluzione $\tilde{x}$ (calcolata tramite $\tilde{R}x = \tilde{Q}^h b$) che è la soluzione esatta del sistema con la matrice A perturbata da una matrice δ_A con

$$\|\delta_A\|_F \leq u \left(\gamma n^2 \|A\|_F + n \left\| \tilde{R} \right\|_F \right) + O(u^2),$$

e i termini noti perturbati da δ_b con

$$\|\delta_b\|_2 \leq \gamma n^2 u \|b\|_2 + O(u^2),$$

dove γ è una costante e $\|\bullet\|_F$ è la *norma di Frobenius*, definita da

$$\|A\|_F := \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}|^2 \right)^{1/2}.$$

Si osserva che la disuguaglianza è in norma, e non termine a termine come nella fattorizzazione LU; inoltre si ha un termine quadratico in n invece che lineare; tuttavia, nel complesso si ha di "pericoloso" solo il termine $\left\| \tilde{R} \right\|_F$, che però si disinnesca osservando che $\|A\|_F = \|QR\|_F = \|R\|_F$, e R differisce da $\tilde{R}$ nell'ordine della precisione di macchina.

[1] Se la matrice è di interi, un metodo per determinare il rango è quello di svolgere i calcoli modulo p: se si trova che la matrice è non singolare modulo p allora è non singolare anche su $\mathbb{C}$; se si trova che è singolare, significa che il determinante è divisibile per p; scegliendo a caso un certo numero di primi, sufficientemente grandi, si può essere ragionevolmente sicuri che se la matrice è non singolare, si sia trovato un primo che non divide il suo determinante.

5 Sistemi lineari: metodi iterativi

Si considera una matrice sparsa; il metodo gaussiano non sfrutta tale caratteristica (a meno che la sparsità non abbia una qualche struttura, per esempio nelle matrici a banda), infatti si verifica il fenomeno detto *fill-in*, che consiste nella perdita della sparsità durante lo svolgimento dell'algoritmo. Per sfruttare una proprietà di questo tipo si deve usare un *metodo iterativo*, cioè un metodo che costruisce una successione $(x_k)_{k\in\mathbb{N}}$ che converge alla soluzione x del sistema; ovviamente un metodo di questo tipo deve soddisfare due richieste: il passo per calcolare la successiva x_k deve essere veloce e la convergenza deve essere rapida.

Dato un sistema $Ax = b$, si spezza A come $A = M - N$, con la richiesta che $\det M \neq 0$ (e possibilmente che M sia facilmente invertibile); il sistema si può dunque riscrivere come $Mx = Nx+b$, da cui $x = M^{-1}Nx+M^{-1}b$. Formulato in questo modo, il sistema diventa un problema di punto fisso, cioè si cerca x tale che $M^{-1}Nx + M^{-1}b$ valga ancora x. Per risolverlo, si costruisce la successione $x_{k+1} \coloneqq M^{-1}Nx_k + M^{-1}b$, a partire da un x_0 qualsiasi. Si cercano le condizioni per cui questa successione converga e una stima della velocità di convergenza, in modo da poter confrontare due successioni diverse.

5.1 Convergenza del metodo iterativo

06/11/2007

Se il passo d'iterazione è $x_{k+1} = Px_k + q$ e la successione converge, converge necessariamente a $x^\star$ tale che $x^\star = Px^\star + q$ (cioè alla soluzione del sistema); se $e_k \coloneqq x_k - x^\star$, si ha $e_{k+1} = Pe_k$, cioè $e_k = P^k e_0$. La matrice P è detta *matrice di iterazione*; usando le norme, presa una qualsiasi norma vettoriale si ha $\|e_k\| \leq \|P\|^k \|e_0\|$; da questo, si ha che se esiste una norma vettoriale tale che con la norma matriciale indotta si ha $\|P\| < 1$, allora per ogni x_0, $\lim x_k = x^\star$. Questa è una condizione sufficiente per la convergenza, ma non è necessaria.

Proposizione 5.1. *Per ogni x_0, $\lim x_k = x^\star$ se e solo se $\rho(P) < 1$.*

Dimostrazione.

($\Leftarrow$) Se il raggio spettrale è minore di 1, esiste $\varepsilon > 0$ tale che $\rho(P) + \varepsilon < 1$ e si sa esistere una norma indotta tale che $\rho(P) \leq \|P\| \leq \rho(P) + \varepsilon$, da cui per la condizione sufficiente già vista, si ha la tesi.

($\Rightarrow$) Se la successione converge per ogni x_0, si può scegliere x_0 in modo che e_0 sia un autovettore di P, cioè $Pe_0 = \lambda e_0$; allora $e_k = P^k e_0 = \lambda^k e_0$ e per ipotesi $e_k \to 0$, da cui $\lambda^k \to 0$, cioè $\lambda < 1$. □

Definizione 5.2. Un metodo iterativo è *convergente* se per ogni $x_0 \in \mathbb{C}^n$, $\lim x_k = x^\star$, dove $x^\star$ è la soluzione del sistema.

5.2 Riduzione asintotica dell'errore

Se P_1 e P_2 sono matrici di iterazione, si vogliono confrontare le velocità di convergenza. Innanzitutto, si potrebbe vedere quale delle due matrici ha norma più piccola; ma questo approccio dipende dalla norma scelta. Intuitivamente, il concetto migliore per valutare la velocità di convergenza è il raggio spettrale. In effetti è così: fissata una norma, al primo passo l'errore si moltiplica per

un fattore $\|e_1\|/\|e_0\|$, e così via per i passi successivi fino al k-esimo. La media geometrica di queste riduzioni è

$$\left(\frac{\|e_1\|}{\|e_0\|} \cdot \dots \cdot \frac{\|e_k\|}{\|e_{k-1}\|}\right)^{1/k} = \left(\frac{\|e_k\|}{\|e_0\|}\right)^{1/k} = \left(\frac{\left\|P^k e_0\right\|}{\|e_0\|}\right)^{1/k}.$$

Questa media dipende da e_0; si vorrebbe vedere cosa succede nel caso peggiore; per fare questo, si può sicuramente maggiorare con $\left\|P^k\right\|^{1/k}$, che indipendentemente dalla norma scelta, converge al raggio spettrale di P. Allora il raggio spettrale fornisce la riduzione media dell'errore asintoticamente.

Tutte queste considerazioni valgono anche nel caso in cui la matrice A del sistema sia singolare. Se A è singolare e $A = M - N$, $P = M^{-1}N$, allora esiste $v \neq 0$ tale che $Av = 0$, da cui $Mv = Nv$ e $Pv = v$, cioè il raggio spettrale di P è maggiore o uguale a 1: se A è singolare, il metodo iterativo non è convergente.

Però, la stima col raggio spettrale vale nel caso peggiore, cioè quando si sceglie il vettore iniziale parallelo all'autovettore di modulo massimo; se invece si sceglie un altro vettore, la velocità di convergenza può variare.

Esempio 5.3. Si considera una matrice P con autovalori $\lambda_1 = 0.99$ e $\lambda_2 = \dots = \lambda_n = 0.01$, dopo pochi passi l'errore è concentrato lungo una direzione. Si può quindi cercare una strategia mista, in cui si usano insieme metodi iterativi e diretti.

5.3 Condizioni di arresto dell'iterazione

07/11/2007

Ora, se P è una matrice con $\rho(P) < 1$, ci si può chiedere quando arrestare il processo iterativo. I due modi più intuitivi sono i seguenti: fermarsi quando $\|x_k - x_{k-1}\| \leq \varepsilon$ per un qualche ε fissato oppure fermarsi quando $\|b - Ax_k\| \leq \varepsilon$. Si vorrebbe che queste condizioni implicassero che $\|x_k - x^\star\| \leq \delta$, dove δ è una costante dipendente da ε.

5.3.1 Valutazione del residuo

Per la seconda condizione, si definisce il *residuo* come $r_k := b - Ax_k$; allora $A^{-1}r_k = A^{-1}b - x_k = x^\star - x_k$; di conseguenza,

$$\|x^\star - x_k\| = \left\|A^{-1}r_k\right\| \leq \left\|A^{-1}\right\| \|r_k\| \leq \varepsilon \left\|A^{-1}\right\|.$$

Ora, non si può sapere $\left\|A^{-1}\right\|$ (calcolarlo sarebbe più pesante di risolvere il sistema), quindi si scriverà $\|x_k - x^\star\| \leq \varepsilon^{\mu(A)}/\|A\|$. L'errore relativo è

$$\frac{\|x^\star - x_k\|}{\|x\|} \leq \varepsilon \frac{\mu(A)}{\|b\|},$$

sfruttando la relazione $\|b\| \leq \|A\| \|x\|$.

5.3.2 Valutazione del passo

Per il primo metodo, $x_k - x_{k-1} = (x_k - x^\star) - (x_{k-1} - x^\star)$; per quanto visto in precedenza, l'errore al passo k è l'errore al passo $k-1$ moltiplicato per P, da cui

$$x_k - x_{k-1} = P(x_{k-1} - x^\star) - (x_{k-1} - x^\star) = (I - P)(x^\star - x_{k-1}).$$

Ricavando l'errore al passo $k-1$, $x^\star - x_{k-1} = (I-P)^{-1}(x_k - x_{k-1})$ ($I-P$ è invertibile perché $\rho(P) < 1$). Passando alle norme, $\|x^\star - x_{k-1}\| \leq \left\|(I-P)^{-1}\right\| \varepsilon$. Ancora ci sarebbe da calcolare un'inversa, ma in questo caso, se $\|P\| < 1$,

$$\left\|(I-P)^{-1}\right\| \leq (1 - \|P\|)^{-1},$$

da cui $\|x^\star - x_{k-1}\| \leq (1-\|P\|)^{-1}\varepsilon$. Si osserva che se $\|P\|$ è vicina a 0, tutto va bene: il metodo converge velocemente e la dipendenza di δ da ε ha una costante di amplificazione ridotta; al contrario, se $\|P\|$ è prossima a 1, la convergenza è lenta e la costante di amplificazione è alta.

5.4 Metodo di Jacobi

Si esaminano ora alcuni metodi iterativi. Un metodo iterativo è semplicemente una scelta dello spezzamento $A = M - N$ con $\det M \neq 0$. Il primo che si vedrà è il *metodo di Jacobi*; si spezza A come $D - B - C$, dove D è diagonale, B è strettamente diagonale inferiore e C è strettamente diagonale superiore, e si pongono $M = D$, $N = B + C$ (supponendo che gli elementi sulla diagonale di A siano tutti non nulli). La matrice P è $M^{-1}N = D^{-1}(B+C)$ e il passo iterativo è $x_{k+1} = D^{-1}((B+C)x_k + b)$; dal punto di vista computazionale, il prodotto $(B+C)x_k$ costa come il prodotto Ax_k (in particolare, se A è sparsa, il prodotto costerà poco); l'inversa di D viene gratuitamente. Operativamente, il calcolo da effettuare per ottenere la componente i-esima è il seguente:

$$x_{k+1,i} = \frac{1}{a_{i,i}} \left(-\sum_{j=1}^{i-1} a_{i,j} x_{k,j} - \sum_{j=i+1}^{n} a_{i,j} x_{k,j} + b_i \right).$$

Questo metodo è chiamato anche *metodo degli spostamenti simultanei* per il fatto che il calcolo della componente i-esima è indipendente dal calcolo delle altre componenti e in particolare è parallelizzabile.

5.5 Metodo di Gauss-Seidel

Un altro metodo, che richiede sempre che la diagonale non abbia elementi nulli, è prendere come M la matrice $D - B$ e come N la matrice C. Questo metodo è detto *metodo di Gauss-Seidel*. L'iterazione di questo metodo è $x_{k+1} = (D-B)^{-1}(Cx_k + b)$; moltiplicando per $D-B$, si ottiene $(D-B)x_{k+1} = Cx_k + b$, da cui $Dx_{k+1} = Bx_{k+1} + Cx_k + b$ e infine $x_{k+1} = D^{-1}(Bx_{k+1} + Cx_k + b)$. Formalmente, l'unica cosa diversa tra questa espressione e quella del metodo di Jacobi è la presenza di un x_{k+1} invece di un x_k. Sembra strano che x_{k+1} stia nel membro a destra, ma calcolando in componenti si ottiene

$$x_{k+1,i} = \frac{1}{a_{i,i}} \left(-\sum_{j=1}^{i-1} a_{i,j} x_{k+1,j} - \sum_{j=i+1}^{n} a_{i,j} x_{k,j} + b \right):$$

nel calcolo di $x_{k+1,i}$, si usano le componenti di x_{k+1} di indice minore di i, cioè quelle già calcolate. In pratica, è lo stesso metodo di Jacobi dove però si usano delle informazioni più aggiornate, che si presumono essere migliori. In realtà, ci

sono esempi in cui il metodo di Jacobi è più veloce, ma si vedranno ampie classi di matrici per cui vince il metodo di Gauss-Seidel. Uno svantaggio del metodo di Gauss-Seidel è che non è parallelizzabile.

5.6 Teoremi di convergenza

Data una matrice A con diagonale non nulla, si denoteranno con G la matrice $(D-B)^{-1}C$ del metodo di Gauss-Seidel e con J la matrice $D^{-1}(B+C)$ del metodo di Jacobi. Si ha il seguente risultato di convergenza.

Teorema 5.4. *Se A soddisfa a una delle seguenti condizioni, allora $\rho(J) < 1$ e $\rho(G) < 1$:*

1. *A è fortemente dominante diagonale;*
2. *A è dominante diagonale e irriducibile;*
3. *A^t è fortemente dominante diagonale;*
4. *A^t è dominante diagonale e irriducibile.*

Dimostrazione. Per studiare il raggio spettrale di J, si devono studiare gli autovalori, cioè le radici di $\det(J-\lambda I)=0$, cioè di $\det(D^{-1}(B+C)-\lambda I)=0$. Ma la condizione non cambia moltiplicando tutto per D, cioè è equivalente a $\det(B+C-\lambda D)=0$. La matrice (a meno di cambiare di segno a ogni entrata) è la matrice A con la diagonale moltiplicata per λ. Se $|\lambda| \geq 1$, anche questa matrice è fortemente dominante diagonale o dominante diagonale e irriducibile e di conseguenza è non singolare, cioè λ non è un autovalore.

Per G la tecnica è la stessa: λ è autovalore di G se e solo se $\det(G-\lambda I)=0$ e questa equazione è equivalente a $\det(C-\lambda(D-B))=0$. La matrice è A, cambiata di segno e con la parte triangolare inferiore moltiplicata per λ, e ancora, se A è fortemente dominante diagonale o dominante diagonale e irriducibile, lo è anche $C-\lambda(D-B)$. □

Si vedrà ora che il metodo di Gauss-Seidel è migliore di quello di Jacobi per alcune classi di matrici utili nella pratica.

Proposizione 5.5. *Se la matrice A è tridiagonale, con elementi sulla diagonale diversi da 0, allora $\rho(G) = \rho(J)^2$.*

Dimostrazione. Si considera la matrice diagonale $D_\alpha := \mathrm{diag}(1, \alpha, \ldots, \alpha^{n-1})$ con $\alpha \neq 0$; coniugare J con D_α ha l'effetto di moltiplicare la sottodiagonale per α e la sopradiagonale per α^{-1}, cioè $D_\alpha J D_\alpha^{-1} = D^{-1}(\alpha B + \alpha^{-1}C)$. Gli autovalori di J e di $D_\alpha J D_\alpha^{-1}$ sono gli stessi, quindi λ è un autovalore di J se e solo se esiste (per ogni) α con $\det(D^{-1}(\alpha B+\alpha^{-1}C)-\lambda I)=0$ se e solo se esiste (per ogni) α con $\det(\alpha B+\alpha^{-1}C-\lambda D)=0$ se e solo se esiste (per ogni) α con $\det(\alpha^2 B+C-\lambda\alpha D)=0$.

Invece, μ è un autovalore di G se e solo se $\det((D-B)^{-1}C-\mu I)=0$ se e solo se $\det(C+\mu B-\mu D)=0$.

Sia quindi λ un autovalore non nullo di J: allora per $\alpha=\lambda$, si ha $\det(\lambda^2 B+B-\lambda^2 D)=0$, cioè λ^2 è autovalore di G; viceversa, sia μ un autovalore non nullo di G: allora $\det(C+\mu B-\mu D)=0$, cioè se λ è una radice di μ, λ è autovalore di $D_\lambda J D_\lambda^{-1}$ e quindi anche di J. Di conseguenza, $\rho(G)=\rho(J)^2$. □

La proposizione dice che il metodo di Gauss-Seidel converge due volte più velocemente del metodo di Jacobi per le matrici tridiagonali. Per un'altra classe di matrici si ha il seguente.

Teorema 5.6 (Stein-Rosenberg)**.** *Se A è tale che $a_{i,i} \neq 0$ e $j_{i,j} \geq 0$ (cioè se $a_{i,i}a_{i,j} \leq 0$ per ogni $i \neq j$), allora vale una e una sola tra le seguenti:*

1. $\rho(G) = \rho(J) = 0$;
2. $\rho(G) = \rho(J) = 1$;
3. $\rho(G) < \rho(J) < 1$;
4. $1 < \rho(J) < \rho(G)$;

In particolare, se il metodo di Jacobi converge, converge anche il metodo di Gauss-Seidel e più velocemente, anche se non si sa di quanto.

Nel caso particolare in cui la matrice d'iterazione ha raggio spettrale nullo, il metodo iterativo è in realtà un metodo diretto, dato che tutti i blocchi di Jordan sono nilpotenti e quindi la successione coinciderà definitivamente la soluzione.

In molti problemi si presentano matrici che si descrivono più facilmente con strutture a blocchi. I metodi di Gauss-Seidel e di Jacobi funzionano anche in questo caso, se D invece di essere la diagonale è la diagonale rispetto ai blocchi, e B e C sono di conseguenza.

6 Problemi di punto fisso

21/11/2007

Si cerca una soluzione al seguente problema: data una funzione g con una certa regolarità, trovare un valore x tale che $g(x) = x$. Un problema di questo tipo è la risoluzione di equazioni polinomiali, scegliendo come g una funzione che dipende dal polinomio e ha un punto fisso in x se e solo se il polinomio si annulla in x.

6.1 Convergenza

La formulazione del problema sotto forma di punto fisso permette di risolverlo costruendo una successione data da $x_m \coloneqq g(x_{m-1})$ a partire da un x_0 qualsiasi. Se g è continua in $[a, b]$, $x_i \in [a, b]$ per ogni i e $\lim x_i = \alpha$, allora α è un punto fisso per g: infatti, $\alpha = \lim x_i = \lim g(x_{i-1})$ e per continuità questo è uguale a $g(\lim x_{i-1}) = g(\alpha)$.

Teorema 6.1 (condizione sufficiente di convergenza)**.** *Sia α un punto fisso per g, una funzione di classe $\mathscr{C}^1$ in un intervallo chiuso $I \coloneqq [\alpha - \rho, \alpha + \rho]$ con $\rho > 0$ e tale che $|g'(x)| < 1$ per ogni $x \in I$; sia inoltre $x_0 \in I$; allora, $x_i \in I$ per ogni i e $\lim x_i = \alpha$.*

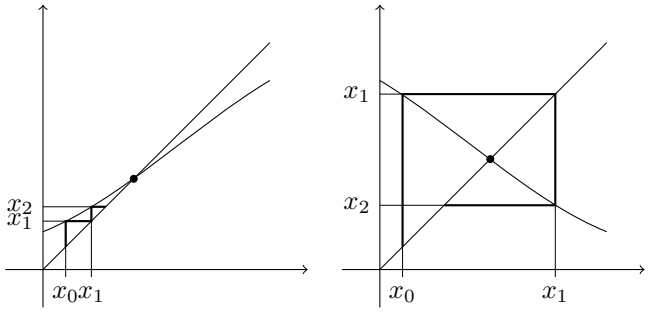
Dimostrazione. Sia $\lambda \coloneqq \max\{\, |g'(x)| \mid x \in I \,\} < 1$; si dimostrerà che $|x_i - \alpha| \leq \lambda^i \rho$: questa proprietà, dato che $\lambda < 1$, implica innanzitutto che $x_i \in I$, ma anche che $\lim x_i = \alpha$.

Si dimostrerà la proprietà per induzione: se $i = 0$ vale per ipotesi; se vale per ogni indice minore o uguale di i, allora per il teorema del valor medio

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

dove $x_i \leq c_i \leq \alpha$; in particolare, $c_i \in I$. Passando ai moduli, $|x_{i+1} - \alpha| = |g'(x_i)|\,|x_i - \alpha| \leq \lambda \cdot \lambda^i \rho$. □

Figura 3: Convergenza al punto fisso.



(a) Caso della derivata positiva. (b) Caso della derivata negativa.

Teorema 6.2. *Se $g \in \mathscr{C}^1([a,b])$ con $|g'(x)| < 1$ per ogni $x \in [a,b]$, esiste al più un punto fisso α per g in $[a,b]$.*

Dimostrazione. Se per assurdo $\beta \in [a,b]$ è un punto fisso per g, diverso da α, allora $\alpha - \beta = g(\alpha) - g(\beta) = g'(c)(\alpha - \beta)$, da cui $g'(c) = 1$, assurdo. □

Ci si pone nelle ipotesi del teorema, supponendo inoltre che $0 < g'(x) < 1$; trovare i punti fissi di g significa trovare le intersezioni con la bisettrice $y = x$. Graficamente, si parte da un x_0 qualsiasi sull'asse delle ascisse; si trova il corrispondente $g(x_0)$ sull'asse delle ordinate; questo valore si proietta sulla bisettrice in modo da trovare il prossimo elemento della successione, x_1; quindi si itera il processo. In questo caso x_i converge a α e la successione è crescente se $x_0 < \alpha$, decrescente altrimenti. Nel caso opposto, $-1 < g'(x) < 0$, lo stesso procedimento porta a una successione che si alterna tra $[\alpha, \alpha + \rho]$ e $[\alpha - \rho, \alpha]$ (si veda figura 3).

Ci si può chiedere quando fermare il calcolo nella pratica. Il modo più ovvio è fermarsi quando $|x_{i+1} - x_i| < \varepsilon$; tuttavia, questo modo dipende dal segno della derivata: per l'alternarsi della successione, nel caso $g'(x) < 0$, si sa che α sta nell'intervallo di estremi x_i e x_{i+1} (che sarà lungo al più ε); invece nel caso della derivata positiva, cioè della successione monotòna, non c'è questa garanzia.

Il teorema sulla convergenza vale in aritmetica esatta; nella pratica, ci si può aspettare che $\tilde{x}_1$, il valore effettivamente calcolato al posto di x_1, non sia contenuto nell'intervallo I. Tuttavia, si dimostra che anche se la successione effettivamente calcolata non converge a α, converge a un valore che sta in un opportuno intorno di α. Sia $(\tilde{x}_i)_{i \in \mathbb{N}}$ la successione effettivamente calcolata, con $\tilde{x}_{i+1} = g(\tilde{x}_i) + \delta_i$ e $|\delta_i| < \delta$ (cioè, gli errori sono limitati).

Teorema 6.3. *Siano α un punto fisso per la funzione $g \in \mathscr{C}^1(I)$, $\lambda \coloneqq \max\{\, |g'(x)| \mid x \in I \,\} < 1$ e $\sigma \coloneqq \delta/1 - \lambda$; se $\sigma < \rho$ e $|x_0 - \alpha| \leq \rho$, allora $|\tilde{x}_i - \alpha| \leq \sigma + \lambda^i(\rho - \sigma)$.*

Dimostrazione. Per induzione: se $i = 0$, $\tilde{x}_0 = x_0$, allora $|\tilde{x}_0 - \alpha| \leq \rho$, quindi non ci sono problemi. Se vale l'asserto per ogni numero minore o uguale a i, $|\tilde{x}_i - \alpha| \leq \sigma + \lambda^i(\rho - \sigma) < \sigma + (\rho - \sigma) = \rho$, perciò $\tilde{x}_i \in I$. Allora $\tilde{x}_{i+1} - \alpha =$

$g(\tilde{x}_i) + \delta_i - g(\alpha) = g'(c_i)(\tilde{x}_i - \alpha) + \delta_i$, con $c_i \in I$ perché $\tilde{x}_i \in I$. Passando ai moduli,

$$|\tilde{x}_{i+1} - \alpha| = |g'(c_i)|\,|\tilde{x}_i - \alpha| + \delta \leq \lambda(\sigma + \lambda^i(\rho - \sigma)) + (1-\lambda)\sigma = \sigma + \lambda^{i+1}(\rho - \sigma). \quad \square$$

6.2 Calcolo degli zeri di una funzione continua

23/11/2007

Sia $f\colon [a,b] \to \mathbb{R}$ una funzione continua, di cui si vogliono trovare gli zeri (si suppone che ne esista almeno uno). Per farlo si può usare il metodo di bisezione, basato sulla filosofia del divide et impera. Se $f(a)f(b) < 0$, per la continuità esiste almeno un punto interno in cui f si annulla; allora si considera $c \coloneqq {}^1\!/_2(a+b)$; a seconda del suo segno, si sceglie di continuare la ricerca a destra o a sinistra (in modo che il segno di f agli estremi del nuovo intervallo sia ancora diverso). Iterando questo procedimento si costruisce una successione di intervalli (a_i, b_i) che contengono sicuramente uno zero di f. Il metodo può essere fermato quando $b_i - a_i \leq \varepsilon \min\{|a_i|, |b_i|\}$. Si osserva che $b_i - a_i = 2^{-i}(b-a)$.

Considerando i calcoli pratici, il vero valore calcolato non è $f(c)$, ma $\tilde{f}(c)$, compreso tra $f(c) - \delta$ e $f(c) + \delta$. Di conseguenza il valore di α può essere calcolato solo a meno di un'incertezza di ${}^{\delta}\!/_{f'(\alpha)}$ a destra e a sinistra.

Per applicare i metodi per trovare i punti fissi a questo caso, bisogna trovare una funzione g tale che $g(\alpha) = \alpha$ se e solo se $f(\alpha) = 0$. Una possibilità è quella di considerare $g(x) \coloneqq x - {}^{f(x)}\!/_{f'(x)}$. Si è visto che se $|g'(x)| \leq \lambda < 1$ in $[a,b]$, esiste un intervallo $I \coloneqq [\alpha - \rho, \alpha + \rho]$ tale che, se $x_0 \in I$, $|x_k - \alpha| \leq \lambda^i \rho$. In particolare, questo metodo funziona bene quando la derivata è piccola vicino al punto fisso.

Bisogna sapere quando terminare il calcolo della successione. I principali criteri di arresto sono due: $|x_{k+1} - x_k| < \varepsilon$ o $|f(x_k)| < \varepsilon$.

Con la prima condizione, se

$$|x_{k+1} - x_k| = |g(x_k) - g(\alpha) + \alpha - x_k| = |(g'(\xi) - 1)(x_k - \alpha)| \leq \varepsilon,$$

si ottiene che

$$|\alpha - x_k| = \left|\frac{x_{k+1} - x_k}{1 - g'(\xi)}\right| \leq \frac{\varepsilon}{|1 - g'(\xi)|};$$

quindi se $g' < 0$ non ci sono problemi, che invece compaiono per g' vicina a 1: in questo caso, l'intervallo di incertezza è ampio.

Con la seconda condizione, se

$$|f(x_k)| = |f(x_k) - f(\alpha)| = |f'(\eta)(x_k - \alpha)| \leq \varepsilon,$$

si ha che

$$|x_k - \alpha| = \left|\frac{f(x_k)}{f'(x_k)}\right| \leq \frac{\varepsilon}{|f'(\eta)|};$$

anche in questo caso ci sono problemi, quando f' è prossimo a 0.

6.3 Tipi di convergenza

27/11/2007

Data una successione $(x_k)_{k\in\mathbb{N}} \to \alpha$, se

$$\gamma \coloneqq \lim_{k\to+\infty} \left|\frac{x_{k+1} - \alpha}{x_k - \alpha}\right|$$

esiste, si dice che la convergenza è:

- *geometrica* (o *lineare*) se $0 < \gamma < 1$ (per esempio, succede quando $|x_k - \alpha| = \delta^k$);
- *sublineare* se $\gamma = 1$;
- *superlineare* se $\gamma = 0$.

In quest'ultimo caso, si può affinare la classificazione: preso

$$\vartheta := \lim \left| \frac{x_{k+1} - \alpha}{(x_k - \alpha)^p} \right|,$$

si dice che la convergenza ha *ordine* p se il limite esiste ed è compreso tra 0 e $+\infty$ estremi esclusi. In questo modo si può distinguere, per esempio, tra $|x_k - \alpha| = 2^{-2^k}$ (che ha ordine 2) e $|x_k - \alpha| = 2^{-k}$ (che ha ordine 1).

Esempio 6.4. Sia (x_k) una successione per cui $x_k - \alpha = 2^{-k}$. Si sa che l'errore che si commette troncando un numero binario dopo la t-esima cifra è 2^{1-t}; viceversa, se l'errore è inferiore a 2^{1-d}, si sa che ci sono d cifre esatte. Quindi in questo caso si sta aggiungendo una cifra significativa a ogni passaggio, perciò si fanno esattamente tanti passaggi quante sono le cifre richieste. Se invece $x_k - \alpha = 2^{-2^k}$, il numero di passi è il logaritmo delle cifre richieste, dato che a ogni passo si raddoppiano le cifre esatte.

Si considera ora una successione del tipo $x_{k+1} = g(x_k)$, con g di classe $\mathscr{C}^1$; allora applicando il teorema del valor medio

$$\lim \left| \frac{x_{k+1} - \alpha}{x_k - \alpha} \right| = \lim \left| \frac{g(x_k) - g(\alpha)}{x_k - \alpha} \right| = \lim |g'(\xi_k)| ;$$

se x_k converge a α, anche ξ_k converge a α, perciò il limite risulta $|g'(\alpha)|$.

Se esiste x_0 tale che $\lim x_k = \alpha$, allora una condizione necessaria per la convergenza lineare è che $0 < |g'(\alpha)| < 1$; una condizione necessaria per la convergenza sublineare è che $|g'(\alpha)| = 1$; per la convergenza superlineare che $|g'(\alpha)| = 0$.

Queste condizioni sono anche sufficienti: se $0 < |g'(\alpha)| < 1$, allora questa caratteristica vale in un intorno di α e partendo da uno dei punti di questo intervallo la successione converge a α linearmente. Se $|g'(\alpha)| = 0$, ancora si ha un intorno in cui la derivata prima è minore di 1 in modulo, quindi se x_0 è nell'intorno, la successione converge a α in modo superlineare. Se $|g'(\alpha)| = 1$, non è più vero che esiste un siffatto intervallo (per esempio, se la derivata è maggiore di 1 al di fuori di α); però si può dire che se $|g'(x)| < 1$ per $x \neq \alpha$ in un intervallo, allora esiste un sottointervallo partendo dal quale la successione converge in modo sublineare.

Sembra che per calcolare il punto fisso α si debba sapere la derivata prima in α (cioè, si debba sapere α); tuttavia non è sempre così: se $g(x) := x - {}^{f(x)}/_{f'(x)}$, con $f(\alpha) = 0$, allora $g'(\alpha) = 1 - 1 + {}^{f(\alpha)}/_{f'(\alpha)^2} f''(\alpha) = 0$, quindi si sa che il metodo convergerà superlinearmente, anche se non si conosce α.

28/11/2007

Proposizione 6.5. *Se $g \in \mathscr{C}^p([a,b])$, $g^{(1)}(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0$ e $g^{(p)}(x) \neq 0$, allora esiste un intervallo $I := [\alpha - \rho, \alpha + \rho]$ tale che se $x_0 \in I$, la successione converge in modo superlineare a α con ordine p. Viceversa, se $g \in \mathscr{C}^p([a,b])$ (con $p > 1$), $\alpha = g(\alpha) \in [a,b]$ ed esiste $x_0 \in [a,b]$ tale che la successione converge a α con ordine p, allora $0 = g^{(1)}(\alpha) = \cdots = g^{(p-1)}(\alpha)$ e $g^{(p)}(\alpha) \neq 0$.*

Dimostrazione. Se $g'(\alpha) = 0$, allora esiste un intervallo (centrato in α) in cui la derivata è in modulo minore di 1; in questo intervallo per definizione la convergenza è superlineare. Per l'ordine, sviluppando in serie di Taylor si ha

$$g(x_k) = g(\alpha) + \cancel{\frac{(x_k-\alpha)g'(\alpha)}{1!}} + \cdots + \\ + \cancel{\frac{(x_k-\alpha)^{p-1}g^{(p-1)}(\alpha)}{(p-1)!}} + \frac{(x_k-\alpha)^p g^{(p)}(\xi_k)}{p!},$$

da cui

$$\lim \left| \frac{g(x_k)-\alpha}{(x_k-p)^p} \right| = \frac{\left|g^{(p)}(\alpha)\right|}{p!} \in (0, +\infty).$$

Per il secondo asserto: sia $1 \leq r < p$; si deve dimostrare che $g^{(r)}(\alpha) = 0$. Sicuramente,

$$\lim \left| \frac{g(x_k)-g(\alpha)}{(x_k-\alpha)^r} \right| = 0.$$

Per $r = 1$,

$$0 = \lim \left| \frac{g(x_k)-g(\alpha)}{x_k-\alpha} \right| = \lim \left| \frac{x_k-\alpha}{x_k-\alpha} g'(\xi_k) \right| = \lim |g'(\xi_k)| = |g'(\alpha)|,$$

da cui $g'(\alpha) = 0$. Ora, supponendo $g^{(1)}(\alpha) = \cdots = g^{(r-1)}(\alpha) = 0$, si dimostra $g^{(r)}(\alpha) = 0$:

$$g(x_k) - g(\alpha) = \cancel{\frac{(x_k-\alpha)g'(\alpha)}{1!}} + \cdots + \cancel{\frac{(x_k-\alpha)^{r-1}g^{(r-1)}(\alpha)}{(r-1)!}} + \frac{(x_k-\alpha)^r g^{(r)}(\xi_k)}{r!},$$

da cui

$$0 = \lim \left| \frac{g(x_k)-\alpha}{(x_k-\alpha)^r} \right| = \lim \frac{\left|g^{(r)}(\xi_k)\right|}{r!} = \frac{\left|g^{(r)}(\alpha)\right|}{r!},$$

cioè $g^{(r)}(\alpha) = 0$. Per induzione segue la tesi. □

Notevole è il fatto che basta un solo x_0 che soddisfa le ipotesi per avere la tesi. Questo non è vero quando la funzione non è sufficientemente regolare.

Esempio 6.6. Si considera una funzione che arriva al punto fisso da sinistra come una retta con $0 < m < 1$, mentre da destra come una parabola con vertice nel punto fisso. La funzione non è derivabile in α; si osserva che da sinistra la convergenza è lineare, mentre da destra la convergenza è superlineare.

Esempio 6.7. Se la funzione g non è sufficientemente regolare, l'ordine di convergenza può non essere intero. Per esempio, sia $g(x) \coloneqq x^{5/3}$; 0 è un punto fisso per g; per calcolare la velocità di convergenza, si calcola $\lim |x_k^{5/3}/x_k^p|$: il valore di p che rende il limite finito e non nullo è $5/3$, che perciò è l'ordine della convergenza.

I metodi visti chiaramente non funzionano, perché darebbero un risultato intero. Il motivo è da ricercarsi nella regolarità di g: $g'(x) = 5/3 x^{2/3}$, da cui $g'(0) = 0$; $g''(x) = 10/3 x^{-1/3}$, che non è definita in 0. In particolare, $g \in \mathscr{C}^1$ e $g'(0) = 0$, ma $g \notin \mathscr{C}^2$. Per applicare il teorema, serve una derivata che esista e non sia nulla nel punto fisso, ma in questo caso non c'è.

Altri esempi sono le funzioni $x \log x$ e $x/\log x$.

6.4 Metodo di Newton

Per trovare uno zero di una funzione f si può pensare di scegliere un punto x_0 opportunamente vicino allo zero α, considerare la tangente in x_0 e prendere come x_1 l'intersezione della tangente con l'asse delle ascisse. Facendo i calcoli, quello che risulta è $x_1 = x_0 - {}^{f(x_0)}/_{f'(x_0)}$.

Quello che si fa è considerare quanto visto in precedenza con la funzione $g(x) := x - {}^{f(x)}/_{f'(x)}$: si era già visto che la convergenza di questo metodo è superlineare (anche non conoscendo il punto fisso), infatti $g'(x) = {}^{f(x)f''(x)}/_{f'(x)^2}$, che si annulla nel punto fisso di g che è lo zero di f. La derivata seconda di g in α è $g''(\alpha) = {}^{f''(\alpha)}/_{f(x)}$. Se questa derivata esiste e $f''(x) \neq 0$, allora il metodo converge con ordine 2. Ma chiedere che $g \in \mathscr{C}^2$, implica chiedere che $f \in \mathscr{C}^3$, perché nella scrittura della derivata seconda di g nel punto generico x, compare la derivata terza di f. In realtà, questa richiesta è sovradimensionata: se non si applicano i teoremi già visti e si dimostra questo caso specifico, basta l'ipotesi $f \in \mathscr{C}^2$.

Proposizione 6.8. *Sia $f \in \mathscr{C}^2([a,b])$ con $f(\alpha) = 0$ per un opportuno $\alpha \in [a,b]$; se $f'(\alpha) \neq 0$, allora esiste un intorno $I := [\alpha - \rho, \alpha + \rho] \subseteq [a,b]$ tale che per ogni $x_0 \in I$ la successione converge a α con ordine almeno 2 e con ordine 2 se $f''(\alpha) \neq 0$.*

Dimostrazione. Sia $g(x) := x - {}^{f(x)}/_{f'(x)}$; poiché $g'(\alpha) = 0$, esiste un intorno $I := [\alpha - \rho, \alpha + \rho]$ tale che $|g'(x)| < 1$ per $x \in I$; quindi sono soddisfatte le ipotesi del teorema del punto fisso, perciò per ogni $x_0 \in I$, la successione generata converge a α.

Ponendo per brevità $x := x_k$,

$$\frac{g(x) - \alpha}{(x-\alpha)^2} = \frac{x - \alpha - {}^{f(x)}/_{f'(x)}}{(x-\alpha)^2};$$

sviluppando in serie attorno a x,

$$0 = f(\alpha) = f(x) + \frac{(\alpha - x)f'(x)}{1!} + \frac{(\alpha - x)^2 f''(\xi)}{2!};$$

da questa, si ricava

$$-\frac{f(x)}{f'(x)} = (\alpha - x) + \frac{(\alpha - x)^2}{2}\frac{f''(\xi)}{f'(x)}.$$

Andando a sostituire,

$$\frac{g(x) - \alpha}{(x-\alpha)^2} = \frac{\frac{(\alpha-x)^2}{2}\frac{f''(x)}{f'(x)}}{(x-\alpha)^2} = \frac{1}{2}\frac{f''(x)}{f'(x)}.$$

Poiché fare il limite per $k \to +\infty$ o per $x \to \alpha$ è lo stesso, la dimostrazione è conclusa. □

Esempio 6.9. Sia $p(x) := x^3 - 1$: questo polinomio si annulla in 1, ma nel campo complesso ha altre due radici (le radici terze primitive dell'unità). Si vorrebbe che il metodo di Newton funzionasse anche per trovare le radici complesse. Applicando formalmente il metodo, si ottiene $g(x) := x - \frac{x^3-1}{3x^2}$.

Per l'analisi, si considera un polinomio p che ha α come radice semplice, cioè $p(x) = (x-\alpha)q(x)$; in questo caso, il metodo di Newton dà

$$g(x) - \alpha = \left(x - \frac{(x-\alpha)q(x)}{q(x) + (x-\alpha)q'(x)}\right) - \alpha = \frac{(x-\alpha)^2 q'(x)}{q(x) + (x-\alpha)q'(x)},$$

da cui

$$\lim \left| \frac{g(x)-\alpha}{(x-\alpha)^2} \right| = \lim \left| \frac{q'(x)}{q(x) + (x-\alpha)q'(x)} \right| = \left| \frac{q'(\alpha)}{q(\alpha)} \right|.$$

Tornando al polinomio $x^3 - 1$, preso un punto qualsiasi di $\mathbb{C}$ si può annotare se la successione converge e in caso positivo a quale delle tre radici converge. Quello che si ottiene colorando ciascun esito con un colore diverso è una figura frattale.

6.4.1 Comportamento vicino a zeri multipli

04/12/2007

Si è visto che se α è uno zero semplice di f, allora il metodo di Newton converge in modo almeno quadratico. Si vedrà ora cosa succede nel caso in cui le prime $r-1$ derivate si annullano in α. Questo è un caso sfortunato, in quanto più derivate si annullano, più l'andamento della funzione f è orizzontale, più il problema sarà mal condizionato.

Si suppone quindi che $0 = f(\alpha) = f'(\alpha) = \cdots = f^{(r-1)}(\alpha)$ e che $f^{(r)}(\alpha) \neq 0$, con $r > 1$. In questo caso, la funzione $g(x) \coloneqq x - {}^{f(x)}/_{f'(x)}$ non è definita in α; però, si può definire la funzione g allo stesso modo per $x \neq \alpha$, e porre $g(\alpha) \coloneqq \alpha$. Computazionalmente, non cambia nulla, in quanto se $g(x) = x$ significa che si è arrivati al termine dell'algoritmo. Si dimostrerà che comunque $g(x) \in \mathscr{C}^1$ e che quindi si possono applicare i teoremi visti.

Innanzitutto, g è continua: sicuramente, lo è al di fuori di α; in α, poiché lo sviluppo in serie di f comincia dal termine r-esimo e quello di f' comincia dal termine $(r-1)$-esimo,

$$\lim_{x\to\alpha} \left(x - \frac{f(x)}{f'(x)}\right) = \lim_{x\to\alpha} x - \frac{\frac{(x-\alpha)^r}{r!} f^{(r)}(\xi_1)}{\frac{(x-\alpha)^{r-1}}{(r-1)!} f^{(r)}(\xi_2)} = \alpha.$$

Per mostrare che $g \in \mathscr{C}^1$, si calcola la derivata in α:

$$\begin{aligned} g'(\alpha) &= \lim_{h\to 0} \frac{g(\alpha+h) - g(\alpha)}{h} = \lim_{h\to 0} \frac{h - \frac{f(\alpha+h)}{f'(\alpha+h)}}{h} = \\ &= \lim_{h\to 0} 1 - \frac{1}{h}\frac{f(\alpha+h)}{f'(\alpha+h)} = \lim_{h\to 0} 1 - \frac{1}{r}\frac{f^{(r)}(\xi_1)}{f^{(r)}(\xi_2)} = 1 - \frac{1}{r}. \end{aligned}$$

Quindi la derivata prima esiste in α e vale $1 - {}^1/_r$; analogamente a quanto fatto in precedenza, ma considerando anche lo sviluppo in serie di f'' in α, si scopre che la derivata prima è continua anche in α.

Allora, applicando i teoremi, si scopre che il metodo di Newton funziona anche in questo caso, e la convergenza è del tipo $1 - {}^1/_r$, che peggiora all'aumentare di r. Se $f \in \mathscr{C}^\infty$ e tutte le sue derivate si annullano in α, ci si può aspettare che il metodo di Newton abbia una convergenza sublineare.

Esempio 6.10. Dato un numero $a \in [1/2, 1]$, si vuole calcolare a^{-1} senza usare la divisione. Per farlo, si può per esempio risolvere con il metodo di Newton l'equazione $x^{-1} - a = 0$. La funzione g sarà

$$g(x) \coloneqq x - \frac{x^{-1} - a}{-x^2} = 2x - ax^2,$$

che in effetti non contiene divisioni ma solo moltiplicazioni. L'errore dopo un passo a partire da x, è

$$g(x) - a^{-1} = 2x - ax^2 - a^{-1} = -a(a^{-2} + x^2 - 2a^{-1}x) = -a(x - a^{-1})^2.$$

Perciò, a partire da x_0, l'errore al passo k è dato da $e_k \coloneqq -ae_{k-1}^2$. Tenendo presente che $1/2 \leq a \leq 1$, $|e_k| \leq e_{k-1}^2 \leq \cdots \leq e_0^{2^k}$: la convergenza ha ordine (almeno) 2. Se inoltre $x_0 = 3/2$, cioè si sceglie il punto iniziale a metà dell'intervallo $[1, 2]$ in cui può cadere a^{-1}, l'errore e_0 è maggiorato in modulo da $1/2$, da cui $\left|e^k\right| \leq 2^{-2^k}$: con $k = 6$ già si hanno 18 cifre decimali corrette.

Si suppone ora di avere un polinomio $a(t)$ e di voler risolvere $x(t)a(t) \equiv 1$ (t^n) (si sta cercando una serie di potenze che sia l'inverso di $a(t)$). Applicando il metodo di Newton, si ottiene una iterazione del tipo $x_{k+1}(t) = 2x_k(t) - a(t)x_k(t)^2$. Allora, per quanto visto prima, $x_k(t) - a^{-1}(t) \equiv (x_{k-1}(t) - a^{-1}(t))^2(-a(t))$ (t^n). Se al passo $k-1$ l'errore è 0 modulo t^q, grazie all'elevamento al quadrato, al passo successivo l'errore sarà 0 modulo t^{2q}: si raddoppiano le cifre corrette. Se per esempio si sceglie $x_0(t) = a_0$ (il polinomio costante), l'errore iniziale è $e_0 \equiv 0$ (t), quindi $e_k(t) \equiv 0$ (t^{2^k}). Questa tecnica non è di analisi numerica, ma di algebra computazionale; il risultato è noto come sollevamento di Newton-Hensel.

Osservazione 6.11. Se f è una funzione crescente, positiva, convessa, per $x \in [\alpha, \alpha + \rho]$, si annulla in α e $f'(x) \neq 0$, allora prendendo $x_0 \in [\alpha, \alpha + \rho]$ la successione che si ottiene è monotòna. In effetti si dimostra questo: se $f(x)f''(x) \geq 0$ per $x \geq \alpha$ e $f'(x) \neq 0$ per $x \neq \alpha$, allora per ogni $x_0 > \alpha$, la successione di Newton è decrescente e converge a α. Infatti, $x_{k+1} = x_k - f(x)/f'(x) < x_n$, e $x_{k+1} - \alpha = g'(x_k)(x_k - \alpha) = \frac{f(x)f''(x)}{f'(x)^2}(x_k - \alpha) \geq \alpha$.

Osservazione 6.12. Il metodo di Newton può essere applicato anche nel caso multidimensionale, sostituendo a $1/f'(x)$ l'inversa della matrice jacobiana.

6.5 Velocità di convergenza

05/12/2007

Si studierà ora uno strumento per aumentare la convergenza di un qualsiasi metodo, al prezzo di diminuire la stabilità dell'algoritmo. Si suppone che $x_k = \alpha + \lambda^k$, dove α è un punto fisso per g (cioè che la successione converga in modo geometrico). Allora $x_{k+1} - x_k = \lambda^{k+1} - \lambda^k = \lambda^k(\lambda - 1)$, da cui $x_{k+2} - 2x_{k+1} + x_k = \lambda^{k+1}(\lambda - 1) - \lambda^k(\lambda - 1) = \lambda^k(\lambda - 1)^2$. Si ha quindi

$$\frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k} = \frac{\lambda^{2k}(\lambda - 1)^2}{\lambda^k(\lambda - 1)^2} = \lambda^k,$$

da cui $\alpha = x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k}$.

Allora, si può considerare la successione data da $z_0 \coloneqq x_0$ e

$$z_{k+1} \coloneqq z_k - \frac{(g(z_k) - z_k)^2}{g(g(z_k)) - 2g(z_k) + z_k};$$

in particolare, si sta calcolando il punto fisso della funzione G definita da

$$G(z) \coloneqq z - \frac{(g(z) - z)^2}{g(g(z)) - 2g(z) + z}$$

per $z \neq \alpha$ e da $G(\alpha) \coloneqq \alpha$. Si può dimostrare che se g è sufficientemente regolare, G sarà continua o anche di classe $\mathscr{C}^1$. Inoltre, se g dà successioni con convergenza sublineare, G dà convergenza lineare; se g dà convergenza lineare, G dà convergenza almeno quadratica; in generale, se g dà convergenza di ordine p, G dà convergenza di ordine almeno $2p - 1$.

Supponendo che il costo computazionale per il calcolo di g sia molto maggiore delle altre operazioni aritmetiche, per avere questo miglioramento si paga sostanzialmente un calcolo di g in più. In particolare, se il metodo aveva convergenza di ordine p, non è conveniente usare questo miglioramento, dato che il semplice uso della funzione $g \circ g$ ha lo stesso costo e ordine p^2 invece che $2p - 1$. Negli altri casi, apparentemente c'è vantaggio; tuttavia, vicino al punto fisso la computazione floating point dà problemi di cancellazione sia al numeratore che al denominatore che rendono inutile il metodo, almeno in questa forma.

Osservazione 6.13. Supponendo di avere due metodi, si vuole capire quale dei due metodi sia più vantaggioso. Se un metodo è dato da una funzione g, con convergenza di ordine $p > 1$ e costo per passo c, e l'errore è $e_k \leq \gamma\lambda^{p^k}$ con $0 < \gamma < 1$; sia k il numero di passi per avere un errore minore o uguale a ε; allora $\log\gamma + p^k \log\lambda \leq \log\varepsilon$, da cui $p^k \geq {}^{\log\varepsilon}/_{\log\lambda} - {}^{\log\gamma}/_{\log\lambda}$ e, ricavando k,

$$k \geq \frac{1}{\log p}\left(\frac{\log\varepsilon}{\log\lambda} - \frac{\log\gamma}{\log\lambda}\right) \geq \frac{1}{\log p}\log\log\varepsilon^{-1} + \mathrm{O}(1).$$

Di conseguenza, il costo per arrivare a un errore inferiore a ε è ${}^{c}/_{\log p}\log\log\varepsilon^{-1}$.

7 Interpolazione

Il problema è il seguente: si hanno $n + 1$ punti distinti assegnati, $x_0, \ldots, x_n$, associati a degli altri valori qualsiasi $y_0, \ldots, y_n$, e un insieme $\mathscr{F}$ di funzioni da $\mathbb{R}$ in sé; si vuole determinare una funzione $\varphi \in \mathscr{F}$ che soddisfi la condizione $\varphi(x_i) = y_i$ per ogni i. Gli x_i si chiameranno *nodi* e le condizioni $\varphi(x_i) = y_i$ *condizioni di interpolazione*.

La soluzione, se esiste, dipende fortemente dall'insieme $\mathscr{F}$. Alcuni esempi sono interpolazioni con $\mathscr{F}$ l'insieme dei polinomi, oppure l'insieme delle funzioni razionali, oppure lo spazio vettoriale generato dalle funzioni $\sin jx$ e $\cos jx$ per $j \in \mathbb{N}$; si dicono rispettivamente interpolazione polinomiale, razionale e trigonometrica. Si studieranno il primo e l'ultimo caso.

7.1 Interpolazione polinomiale

Si considera la base $(1, x, \ldots, x^k, \ldots)$ dello spazio vettoriale dei polinomi, così che un polinomio p di grado n scritto sulla base abbia la forma $\sum_{j=0}^{n} \alpha_j x^j$. La

condizione di interpolazione $p(x_i) = y_i$ dà $\sum_{j\geq 0}^{n} \alpha_j x_i^j = y_j$; se i va da 0 a n, cioè se si hanno $n+1$ punti, si hanno $n+1$ equazioni e $n+1$ incognite, le α_j. Il sistema corrispondente ha come vettore dei termini noti $(y_0, \dots, y_n)^t$ e come matrice V_n la matrice che ha come riga i-esima il vettore $(1, x_i, \dots, x_i^n)$, detta *matrice di Vandermonde*. Se la matrice è non singolare si può risolvere il sistema con uno dei metodi visti. Si vorrebbe anche avere qualche informazione sul condizionamento della matrice.

Lemma 7.1. *Sia V_n la matrice di Vandermonde $V_n \coloneqq (x_i^j)_{i,j}$; allora* $\det V_n = \prod_{i>j}(x_i - x_j)$.

Dimostrazione. Si dimostra per induzione su n: per $n = 1$ la formula funziona; supponendo che funzioni per V_{n-1}, si considera la matrice $V_n(\lambda)$, uguale a V_n tranne per l'ultima riga che diventa $(1, \lambda, \dots, \lambda^n)$; chiaramente $V_n = V_n(x_n)$, e il determinante di $V_n(\lambda)$ come funzione di λ è un polinomio con coefficiente di testa pari a $\det V_{n-1}$. Allora, raccogliendo $\det V_{n-1}$ e fattorizzando, $\det V_n(\lambda) = \det V_{n-1}(\lambda - \xi_1) \cdots (\lambda - \xi_n)$. Ma gli zeri di $\det V_n(\lambda)$ comprendono sicuramente $x_0, \dots, x_{n-1}$, quindi sono esattamente questi perché il numero è quello giusto. Perciò, $\det V_n(\lambda) = \det V_{n-1}(\lambda - x_0) \cdots (\lambda - x_{n-1})$; di conseguenza, $\det V_n = \det V_n(x_n) = \det V_{n-1} \prod_{i=0}^{n-1}(x_n - x_i)$ e si conclude grazie al passo induttivo. □

Grazie al lemma, e dato che si sono scelti gli x_i distinti, si ha che il sistema è non singolare, e quindi tra i polinomi di grado al più n esiste ed è unica la soluzione. Il problema è che il numero di condizionamento di V_n, se $x_i \in \mathbb{R}$ sono qualunque, cresce esponenzialmente con n. Si vedrà invece che nel caso complesso il numero di condizionamento è basso (per esempio, se gli x_i sono le radici $(n+1)$-esime dell'unità, è 1).

Ignorando i problemi di condizionamento, utilizzando i metodi generici per risolvere i sistemi si può risolvere il sistema con un costo di $2/3n^3$; esistono però algoritmi specifici per le matrici di Vandermonde che pagano n^2 operazioni, o addirittura solo $n \log^2 n$ (però con problemi di stabilità).

Moltiplicare V_n per un vettore α, significa trovare un vettore con i valori del polinomio che ha come coefficienti α nei punti $x_0, \dots, x_n$. Applicando ripetutamente lo schema di Horner, per calcolare questo vettore si pagano n^2 operazioni. Anche qui, esistono algoritmi più veloci (almeno asintoticamente), che arrivano a un costo di $n \log^2 n$, anche se con problemi di stabilità.

7.1.1 Resto dell'interpolazione

Per valutare la bontà dell'approssimazione polinomiale calcolata come abbiamo visto, consideriamo i resti $r_n(x) = f(x) - p_n(x)$, dove $p_n(x)$ è l'n-esimo polinomio di interpolazione, calcolato sui nodi $x_0, \dots, x_n \in [a, b]$.

Proposizione 7.2. *Supponiamo $f \in \mathscr{C}^{n+1}([a, b])$ e r_n definiti come sopra. Allora, preso $x \in [a, b]$, esiste $\xi \in [a, b]$ tale che*

$$r_n(x) = \prod_{i=0}^{n}(x - x_i)\frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Dimostrazione. La proposizione è evidente per $x = x_i$, dunque supponiamo $x \neq x_i$ per ogni i. Consideriamo

$$g(y) = r_n(y) - r_n(x) \cdot \frac{\prod(y - x_i)}{\prod(x - x_i)},$$

che è di classe $\mathscr{C}^{n+1}$.

La funzione g si annulla su x e su tutti gli x_i, ossia su $n+2$ punti in totale. Dunque g' si annulla perlomeno su $n+1$ punti, g'' su n, e così via. In particolare, $g^{(n+1)}$ si annulla su almeno un punto.

$$g^{(n+1)}(y) = \Big[f^{(n+1)}(y) - 0\Big] - \frac{(n+1)! \cdot r_n(x)}{\prod(x - x_i)}.$$

Imponendo l'annullamento su $\xi \in [a, b]$ si ottiene la tesi. □

Vale la pena notare che questa stima non è molto stringente: basta allontanarsi poco dagli x_i perché la produttoria possa assumere valori molto alti. In effetti, raramente l'interpolazione polinomiali è in grado di dare buoni risultati.

7.1.2 Interpolazione polinomiale di Lagrange

Per risolvere i problemi di mal condizionamento della matrice di Vandermonde, si può provare a cambiare la base dello spazio vettoriale dei polinomi. Si considera allora la famiglia dei polinomi del tipo

$$L_i(x) \coloneqq \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j};$$

si ha che $L_i(x_k) = \delta_{i,k}$. I polinomi $L_0, \ldots, L_n$ sono linearmente indipendenti e formano una base per lo spazio dei polinomi di grado al più n; sono chiamati *base di Lagrange*. Sulla base di Lagrange, il polinomio d'interpolazione risulta essere semplicemente $p(x) \coloneqq \sum_{i=0}^{n} y_i L_i(x)$. In un certo senso, si sta risolvendo un sistema che ha come matrice l'identità (e di conseguenza, i termini noti sono la soluzione). Se si vuole calcolare p in un punto z, l'espressione risulta

$$p(z) = \sum_{i=0}^{n} y_i \prod_{j \neq i} \frac{z - x_j}{x_i - x_j} = \prod_{j=0}^{n} (z - x_k) \sum_{i=0}^{n} \frac{y_i}{(z - x_i) \prod_{j \neq i} (x_i - x_j)};$$

se le produttorie al denominatore si sono precalcolate, il calcolo di $p(z)$ risulta lineare in n.

7.2 Trasformata discreta di Fourier (DFT)

18/12/2007

Per $n \geq 2$, sia ω_n una *radice n-esima primitiva dell'unità*, cioè un numero complesso tale che $\{\, \omega_n^j \mid 0 \leq j < n \,\}$ è l'insieme di tutte le radici n-esime dell'unità. Per esempio, si può prendere

$$\omega_n = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n} = e^{2\pi i / n}.$$

Questa radice gode di una proprietà di ortogonalità: infatti,

$$\sum_{j=0}^{n-1} \omega_n^{jr} = \begin{cases} n & \text{per } r \equiv 0 \quad (n) \\ 0 & \text{per } r \not\equiv 0 \quad (n) \end{cases}.$$

Si considera ora il problema dell'interpolazione polinomiale sui nodi ω_n^j per $0 \le j < n$; la matrice di Vandermonde associata è $F \coloneqq (\omega_n^{kj})_{0\le k,j<n}$; per la proprietà vista in precedenza, questa matrice soddisfa

$$F^h F = \left(\sum_{r=0}^{n-1} \overline{\omega}_n^{kr}\omega_n^{jr}\right)_{0\le k,j<n} = \left(\sum_{r=0}^{n-1} \omega_n^{-kr}\omega_n^{jr}\right)_{0\le k,j<n} = nI,$$

cioè $F^{-1} = {}^1\!/_n F^h$. In particolare, $\Omega \coloneqq {}^1\!/_{\sqrt{n}} F$ è unitaria e ha numero di condizionamento 1. Di conseguenza,

$$\|F\|_2 \|F^{-1}\|_2 = \left\|\frac{1}{\sqrt{n}}\Omega\right\|_2 \left\|\sqrt{n}\Omega^{-1}\right\|_2 = \|\Omega\|_2 \|\Omega^h\|_2 = 1.$$

Definizione 7.3. Sia F la matrice di Vandermonde appena vista; la funzione $\mathrm{IDFT}\colon \mathbb{C}^n \to \mathbb{C}^n$ con $\mathrm{IDFT}(x) \coloneqq Fx$ è detta *trasformata discreta inversa di Fourier*; la funzione $\mathrm{DFT}\colon \mathbb{C}^n \to \mathbb{C}^n$ con $\mathrm{DFT}(y) \coloneqq F^{-1}(y)$ è detta *trasformata discreta di Fourier*.

7.2.1 Algoritmi per il calcolo della DFT

Calcolare la trasformata discreta di Fourier corrisponde a calcolare i coefficienti del polinomio interpolante; viceversa, dato il vettore x, l'espressione di $y \coloneqq \mathrm{IDFT}(x)$ è data da $y_k = \sum_{j=0}^{n-1} \omega_n^{kj} x_j$: se si indica con $p(z)$ il polinomio $\sum_{j=0}^{n-1} z^j x_j$, la trasformata discreta inversa di Fourier è semplicemente la valutazione di p nei punti ω_n^k e perciò si può calcolare valutando n volte un polinomio di grado $n-1$, per esempio con l'algoritmo di Horner. Ci sono però metodi più veloci.

Si suppone per semplicità $n = 2^a$, $m \coloneqq {}^n\!/_2$; si vedrà l'algoritmo di Cooley-Tukey, basato sul *divide et impera*, separando la somma negli indici pari e in quelli dispari. Dati $x \in \mathbb{C}^n$ e $y \coloneqq \mathrm{IDFT}(x)$, poiché $\omega_n^2 = \omega_m$, si ha:

$$\begin{aligned} y_k &= \sum_{j=0}^{n-1} \omega_n^{kj} x_j = \sum_{j=0}^{m-1} \omega_n^{2jk} x_{2j} + \sum_{j=0}^{m-1} \omega_n^{(2j+1)k} x_{2j+1} = \\ &= \sum_{j=0}^{m-1} \omega_m^{jk} x_{2j} + \omega_n^k \sum_{j=0}^{m-1} \omega_m^{jk} x_{2j+1}. \end{aligned}$$

Si osserva che per calcolare le prime m componenti di y basta calcolare due trasformate discrete inverse di rango m, m moltiplicazioni per ω_n^k e m addizioni. Per le ultime m componenti, le sommatorie sono uguali, mentre $\omega_n^{k+m} = -\omega_n^k$,
19/12/2007
perciò sono necessarie solo altre m sottrazioni. Se IDFT_n è il numero di operazioni necessarie a calcolare la trasformata discreta inversa di rango n, allora si ha la formula ricorsiva

$$\begin{cases} \mathrm{IDFT}_n = 2\,\mathrm{IDFT}_{n/2} + \frac{1}{2}nM + nA \\ \mathrm{IDFT}_1 = 0 \end{cases},$$

da cui risulta $\mathrm{IDFT}_n = \log_2 n({}^1\!/_2 nM + nA)$. Gli algoritmi con questa complessità si dicono algoritmi FFT (*fast Fourier transform*).

L'algoritmo di Saude-Tukey non usa la ricorsione, ma il seguente procedimento:

$$\begin{pmatrix} y_0 \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} I & I \\ I & -I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} F_{n/2} & 0 \\ 0 & F_{n/2} \end{pmatrix} \Pi,$$

dove Π è una matrice di permutazione che serve per riportare le in testa le righe pari. Riapplicando il procedimento a $F_{n/2}$, si ottiene una composizione di matrici di permutazione che risulta corrispondere alla cosidetta permutazione *bit-reversal*: al numero i che si scrive in binario come $b_a \dots b_0$, si associa il numero che si scrive in binario come $b_0 \dots b_a$.

Osservazione 7.4. Per costruire le due trasformate discrete si sono usate solo le proprietà formali delle radici, quindi si possono applicare gli stessi procedimenti a un qualunque campo in cui esistano le radici primitive dell'unità. In realtà, può bastare un dominio d'integrità (o un anello qualsiasi dove i divisori di 0 non diano fastidio).

Esempio 7.5. In $\mathbb{Z}/5\mathbb{Z}$, 2 è una radice 4 primitiva dell'unità (questo in generale accade in tutti i campi della forma $\mathbb{Z}/(2^n+1)\mathbb{Z}$).

Sia ora $y \coloneqq \mathrm{IDFT}(x) = Fx$ e siano $\tilde{y}$ e $\hat{y}$ i valori ottenuti calcolando y in aritmetica floating point mediante l'algoritmo banale (il prodotto matrice-vettore) e mediante l'algoritmo di Cooley-Tuckey. Allora si hanno le seguenti stime sugli errori, dove γ_i sono costanti:

$$\frac{\|y-\tilde{y}\|_2}{\|y\|_2} \leq u\gamma_1 n^{3/2}$$
$$\frac{\|y-\hat{y}\|_2}{\|y\|_2} \leq u\gamma_2 n^{1/2} \log_2 n$$

7.3 Applicazioni della DFT

Le applicazioni della trasformata discreta di Fourier sono numerose; come esempi si porteranno la moltiplicazione veloce di polinomi o numeri e l'interpolazione trigonometrica.

7.3.1 Moltiplicazione di polinomi

Siano $a(x) \coloneqq \sum_{i=0}^{n_a} a_i x^i$ e $b(x) \coloneqq \sum_{i=0}^{n_b} b_i x^i$ due polinomi, e sia $c(x) \coloneqq a(x)b(x) = \sum_{i=0}^{n_a+n_b} c_i x^i$. Il metodo ovvio per calcolare i coefficienti di c è la moltiplicazione matrice-vettore

$$\begin{pmatrix} a_0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ a_{n_a} & \cdots & \cdots & a_0 \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n_a} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_{n_b} \end{pmatrix},$$

che però necessita di $2n_an_b$ moltiplicazioni, cioè di tempo $O(n^2)$ (se $n_a = n = n_b$). Con la trasformata discreta di Fourier, si considera $N \coloneqq 2^a \geq n_a + n_b + 1$; con la trasformata discreta inversa si calcolano $\alpha_i \coloneqq a(\omega_N^i)$ e $\beta_i \coloneqq b(\omega_N^i)$ con $2(3/2 N \log_2 N)$ moltiplicazioni; successivamente si calcolano $\gamma_i \coloneqq \alpha_i\beta_i = c(\omega_N^i)$ con altre N moltiplicazioni e infine si ricavano i coefficienti c_i tramite la trasformata discreta di $(\gamma_0, \ldots, \gamma_{N-1})$, con altre $3/2 N \log_2 N$ moltiplicazioni. In totale, il numero di moltiplicazioni è $9/2 N \log_2 N \in O(n \log_2 n)$.

7.3.2 Moltiplicazione di numeri

Siano a e b due numeri interi di n cifre. L'algoritmo standard di moltiplicazione richiede $n^2 + 2n$ operazioni; il metodo migliore finora conosciuto è quello di Schönhage-Strassen, che ha complessità $O(n \log_2 n \log_2 \log_2 n)$, anche se per valori non troppo grandi ha risultati migliori l'algoritmo di Karatsuba-Hoffmann, che ha complessità asintotica $O(n^{\log_2 3})$. L'algoritmo di Schönhage-Strassen funziona in questo modo: si scrivono a e b in base β, in modo che siano $A(\beta)$ e $B(\beta)$ se $A(x) \coloneqq \sum_{i=0}^{n} a_i x^i$ e $B(x) \coloneqq \sum_{i=0}^{n} b_i x^i$. Ora si può usare il metodo di calcolo veloce del prodotto di polinomi per trovare $C(x) \coloneqq A(x)B(x)$, che è della forma $\sum c_i x^i$; allora $c \coloneqq ab$ è uguale a $\sum c_i \beta^i$. Il problema è che i c_i non sono necessariamente compresi tra 0 e $\beta - 1$, cioè non sono le cifre della scrittura in base β di c; dato che i numeri hanno circa $\log_\beta n$ cifre, ritrovando la scrittura di c in base β col modo ovvio si riesce a ottenere un algoritmo con complessità $O(n \log n \log^2 n)$; usando formule più sofisticate, l'algoritmo di Schönhage-Strassen riesce invece a raggiungere $O(n \log n \log \log n)$.

7.3.3 Interpolazione trigonometrica

Un polinomio trigonometrico è una funzione del tipo

$$\frac{a_0}{2} + \sum_{j=1}^{m-1} (a_j \cos jx + b_j \sin jx) + \frac{a_m}{2} \cos mx.$$

Lo scopo dell'interpolazione trigonometrica è quello di calcolare i coefficienti $a_0, \ldots, a_m, b_1, \ldots, b_{m-1}$ di un polinomio trigonometrico p tale che $p(x_k) = y_k$, dove $x_k \coloneqq 2\pi k/n$ e $y \coloneqq (y_0, \ldots, y_{n-1})$ è un vettore di numeri complessi assegnato. Si può dimostrare che esiste un unico polinomio trigonometrico con $m = n$ e che soddisfa le condizioni d'interpolazione; i coefficienti si possono trovare grazie alla trasformata discreta: se $z \coloneqq \mathrm{DFT}(y)$, risulta $a_j = 2\Re(z_j)$ e $b_j = -2\Im(z_j)$.

Riferimenti bibliografici

[BBCM92] Bevilacqua, Roberto, Dario Bini, Milvio Capovani e Ornella Menchi: *Metodi numerici*. Zanichelli, 1992.

[BCM88] Bini, Dario, Milvio Capovani e Ornella Menchi: *Metodi numerici per l'algebra lineare*. Zanichelli, 1988.