Appunti di Calcolo Numerico

Matteo Lisotto, Tobia Tesan

Indice

Licenza e Prefazione

Segnalazioni di errori e omissioni all'indirizzo degli autori (o, più precisamente, degli umili scribi che presero questi appunti nelle aule dell'Università di Padova e che ora li rendono disponibili in formato elettronico) sono benvenute.

1 Introduzione

Lo scopo del calcolo numerico è lo studio e l'implementazione di algoritmi per la soluzione **approssimata** (con errore) di problemi matematici (tipicamente riguardanti applicazioni scientifiche e tecnologiche).
Le caratteristiche che interessano nello studio degli algoritmi numerici sono:

- **Convergenza:** velocità di convergenza e stima dell'errore.
- **Stabilità**: propagazione degli errori (problemi mal condizionati)
- **Efficienza**: costo computazionale, misurato in termini di numero di operazioni ma anche di velocità di esecuzione delle stesse sull'hardware di destinazione (vale in questo senso l'esempio di BLAS).

Il calcolo numerico ha svariati campi di applicazione, tra cui:

- **Elaborazione dei segnali** monodimensionali piuttosto che bi- e più dimensionali (ad esempio elaborazione di suono e immagini)
- **Data Mining**
- **Diagnostica non invasiva** (TAC, risonanza magnetica)
- Soluzione di **modelli fisici** - ad esempio CFD (computational fluid dynamics, una branca della fluidodinamica utilizzata in progettazione navale e aerospaziale, meteorologia computazionale, simulazione del sistema cardiocircolatorio.

2 Fondamenti del Calcolo Numerico

2.1 Errore

2.1.1 L'errore assoluto

Si supponga di avere a disposizione un metodo numerico che dia come risultato un numero $\tilde{x} \in \mathbb{R}$ approssimazione del numero del risultato *esatto* $x \in \mathbb{R}$. É possibile definire:

Definizione 2.1 (Errore assoluto)**.**

$$\Delta x = |x - \tilde{x}|$$

Questa grandezza, benchè fornisca un'indicazione precisa del valore dell'errore commesso dal metodo matematico preso in esame, non permette di valutare quanta influenza abbia l'errore *sul risultato ottenuto.*

Ad esempio, un errore assoluto $\Delta x = O(10^{-5})$ potrebbe avere un peso accettabile se il nostro risultato corretto x fosse dell'ordine di grandezza $O(1)$ o superiore, ma non si avrebbe un risultato valutabile se x fosse anch'esso $O(10^{-5})$.

2.1.2 L'errore relativo

Volendo conoscere quanto un errore influenzi il risultato quando $x \neq 0$, si definisce:

Definizione 2.2 (Errore relativo)**.**

$$\varepsilon_x = \frac{\Delta x}{|x|} = \frac{|x - \tilde{x}|}{|x|}$$

Questa grandezza, a differenza dell'errore assoluto, permette di valutare quanta influenza abbia l'errore sul risultato ottenuto.

2.2 Rappresentazione dei numeri reali su base arbitraria

Teorema 2.1. *Dato $x \in \mathbb{R}$, fissata una base $b \in \mathbb{N}$ con $b > 1$ è sempre possibile riscrivere x come:*

$$x = sign(x)\Bigg(\underbrace{\sum_{j=0}^{m} a_j b^j}_{parte\ reale} + \underbrace{\sum_{j=1}^{\infty} a_{-j} b^{-j}}_{parte\ frazionaria}\Bigg)$$

avendo la parte reale $\sum_{j=0}^{m} a_j b^j \in \mathbb{Z}$, la parte frazionaria $\sum_{j=1}^{\infty} a_{-j} b^{-j} \in [0, 1]$ e le cifre $a_j \in [0, b-1]$

La rappresentazione può essere riscritta come:

$$x = sign(x)(\alpha_m \alpha_{m-1} ... \alpha_1 \alpha_0 . \alpha_{-1} \alpha_{-2} ...).$$

Lemma 2.1. La parte frazionaria $\sum_{j=1}^{+\infty} a_{-j} b^{-j}$ converge.

Dimostrazione 2.1. Poichè $a_j \in [0, b-1]$ vale:

$$a_{-j}b^{-j} \leq (b-1)(b^{-j})$$

È noto:

Teorema 2.2 (Criterio del confronto)**.**

$$a_n \leq b_n \Rightarrow \sum a_n \leq \sum b_n$$

Allora:

$$\sum a_{-j}b^{-j} \leq \sum (b-1)(b^{-j}) = (b-1)\sum b^{-j}$$

È nota la seguente proprietà delle serie geometriche:

$$S_n = \sum_{j=0}^{N} a_j, a \in \mathbb{R}^+$$

$$\begin{cases} a = 1 \Rightarrow S_n = N+1 \\ a = -1 \Rightarrow S_n \text{ oscilla} \\ a \neq 1 \Rightarrow S_n = \frac{a^{N+1}-1}{a-1} = \begin{cases} |a| > 1 : \text{diverge} \\ |a| < 1 : \frac{1}{1-a} \end{cases} \end{cases}$$

Allora, poichè $b > 1 \Rightarrow \forall j > 1 : b^{-j} < 1$

$$\sum a_{-j}b^{-j} \leq \sum (b-1)(b^{-j}) = (b-1)\sum b^{-j} < \infty$$

□

2.3 Troncamento di un numero

La necessità della rappresentazione con troncamento è data dalle seguenti osservazioni:

Lemma 2.2. I numeri con parte frazionaria finita, in qualche base, sono razionali.

Lemma 2.3. Per i numeri razionali il numero di cifre dopo la virgola dipende dalla base.

Ad esempio: $\frac{1}{3} = (0.\bar{3})_{10}$ e $(0.1)_3$

Lemma 2.4. Se x è irrazionale (ad esempio $\sqrt{2}$ o π) allora, necessariamente, su qualsiasi base deve avere infinite cifre dopo la virgola.

Allora si definisce:

Definizione 2.3 (Troncamento $\tilde{x}_n$)**.** Sia

$$x = sign(x)(\alpha_m \alpha_{m-1} ... \alpha_1 \alpha_0 . \alpha_{-1} \alpha_{-2} ...)$$

Il troncamento all'n-sima cifra $\tilde{x}_n \in \mathbb{Q}$ è:

$$\tilde{x}_n = sign(x)\left(\sum_{j=0}^{m} a_j b^j + \sum_{j=1}^{n} a_{-j}b^{-j}\right)$$

Lemma 2.5 (Errore assoluto nel troncamento)**.** È possibile scrivere:

$$x = sign(x)\left(\sum_{j=0}^{m} a_j b^j + \sum_{j=1}^{n} a_{-j} b^{-j}\right) + sign(x) \sum_{j=n+1}^{\infty} a_{-j} b^{-j}$$

e denotare Δx così:

$$\Delta x = \left| sign(x) \sum_{j=n+1}^{\infty} a_{-j} b^{-j} \right| = \sum_{j=n+1}^{\infty} a_{-j} b^{-j}$$

2.3.1 Stima dell'errore nel troncamento

Teorema 2.3.

$$\Delta x = |x - \tilde{x}_n| \leq b^{-n}$$

Dimostrazione 2.2. È possibile scrivere:

$$\Delta x = |x - \tilde{x}_n| = \sum_{j=n+1}^{\infty} a_{-j} b^{-j} \leq (b-1) \sum_{j=n+1}^{\infty} b^{-j}$$

Dove:

$$\begin{aligned} \sum_{j=n+1}^{\infty} b^{-j} \leq \sum_{j=0}^{\infty} b^{-j} - \sum_{j=0}^{n} b^{-j} &= \frac{1}{1-\frac{1}{b}} - \frac{1-(\frac{1}{b})^{n+1}}{1-\frac{1}{b}} \\ &= \frac{b}{b-1} - \frac{b(1-b^{n+1})}{b-1} = \frac{b-b+b^{-n}}{b-1} \\ &= \frac{b^{-n}}{b-1} \end{aligned}$$

Allora:

$$\Delta x \leq (b-1) \sum_{j=n+1}^{\infty} b^{-j} \leq (b-1)\frac{b^{-n}}{b-1} = b^{-n}$$

□

Esempio 2.1. Si verifica che $(0.\bar{3})_{10} = \frac{1}{3}$:

$$\begin{aligned} (0.\bar{3})_{10} &= \sum_{j=1}^{\infty} 3 \cdot \frac{1}{10^j} = 3 \sum_{j=1}^{\infty} \frac{1}{10^j} \\ &= 3\left(\frac{1}{1-\frac{1}{10}} - 1\right) \\ &= 3\left(\frac{10}{9} - 1\right) = 3\left(\frac{1}{9}\right) = \frac{1}{3} \end{aligned} \tag{1}$$

2.4 Arrotondamento di un numero

Definizione 2.4 (Arrotondamento $\tilde{x}_n$)**.** Dato $x = sign(x)(\alpha_m\alpha_{m-1}...\alpha_1\alpha_0.\alpha_{-1}\alpha_{-2}...)$ si definisce:

$$\tilde{x}_n = sign(x)(\alpha_m\alpha_{m-1}...\alpha_1\alpha_0.\alpha_{-1}\alpha_{-2}...\tilde{\alpha}_{-n})$$

dove

$$\tilde{\alpha}_{-n} = \begin{cases} \alpha_{-n} & \text{se } \alpha_{n+1} \leq \frac{b}{2} \text{ (b pari)} \\ (\alpha+1)_{-n} & \text{se } \alpha_{n+1} > \frac{b}{2} \end{cases}$$

Teorema 2.4 (Limite superiore per l'errore di arrotondamento)**.**

$$\Delta x = |x - \tilde{x}_n| \leq \frac{b^{-n}}{2}$$

Dimostrazione 2.3. Omessa poichè estremamente difficile.

2.5 Rappresentazione posizionale normalizzata

Definizione 2.5 (Rappresentazione posizionale normalizzata)**.** Ogni $x \in \mathbb{R}$ si può scrivere come:

$$x = sign(x)(0.d_1d_2....)b^p$$

dove:

- $p \in \mathbb{Z}$ è l'esponente
- $d_1 \neq 0$ [1]
- $d_j \in \{0, 1.....b-1\}$
- $0.d_1d_2...$ è la mantissa

Tale rappresentazione è detta rappresentazione posizionale normalizzata o *virgola mobile*.

Esempio 2.2.

$$x = 1278.635\ldots = \underbrace{0.1278635\ldots}_{\text{mantissa}} \cdot \underbrace{10^4}_{\text{esponente}}$$

2.6 I numeri macchina $\mathbb{F}$

Definizione 2.6 (Insieme dei reali macchina $\mathbb{F}$)**.**

$$\mathbb{F}(b,t,L,U) = \left\{\mu \in \mathbb{Q} \mid \mu = \pm(0.\mu_1\mu_2...\mu_t)\,b^p\right\}$$

dove:

- $\mu_i = \{0, 1...b-1\}$
- $\mu_1 \neq 0$

[1] La necessità di avere la prima cifra decimale $\neq 0$ serve a evitare di avere multiple rappresentazioni per lo *stesso* numero, ad esempio $0.01 \cdot b^1$ e $0.1 \cdot b^{10}$

- $t \geq 1$
- $p \in [L, U] \in \mathbb{Z}$

Si indica con b la base, t le cifre di mantissa (che deve appartenere a $\mathbb{Q}$, poichè in un calcolatore la memoria per la mantissa è finita), L il più piccolo esponente disponibile e con U il più grande esponente. [2]

Lo standard IEEE754 definisce le seguenti precisioni:

- **Semplice**: 32 bit $\varepsilon_m \approx 10^{-8}$
- **Doppia**: 64 bit $\varepsilon_m \approx 10^{-16}$
- **Quadrupla**: 128 bit $\varepsilon_m \approx 10^{-32}$

Definizione 2.7 (Insieme dei reali rappresentabili ν)**.**

$$\nu = \{0\} \cup [\min \mathbb{F}^+, \max \mathbb{F}^+] \cup [-\max \mathbb{F}^+, -\min \mathbb{F}^+]$$

2.6.1 Condizioni di errore con i numeri macchina $\mathbb{F}$

Può capitare di dover rappresentare un numero non contenuto in ν, e in questo caso ci troveremmo in una condizione di errore.

Definizione 2.8 (Overflow)**.** Si parla di overflow quando si ha x t.c. $|x| \geq b^U$.

La gestione di una siffatta situazione dipende dal sistema di calcolo utilizzato. Matlab e lo standard IEEE754 prevedono la definizione di una quantità indicata con `Inf`, con eventuale segno, per denotare l'overflow.

Definizione 2.9 (Underflow)**.** Si parla di underflow quando si ha x t.c. $0 \leq |x| \leq b^L$.

L'underflow è comunemente gestito ponendo $x = 0$.

2.6.2 Funzione floating

In quanto $\mathbb{F}$ non può rappresentare i numeri irrazionali e ha un numero finito di elementi, è necessario definire una funzione $fl(x)$, detta **funzione floating** che associ ad ogni numero reale $x \in \mathbb{R}$ un corrispondente *rappresentante* in $\mathbb{F}$ tale che

$$x \approx \mu = fl^t(x) \qquad x \in \mathbb{R}, \mu \in \mathbb{F}$$

Definizione 2.10 (Funzione floating)**.**

$$fl^t : \mathbb{R} \to \mathbb{F}$$
$$x \mapsto fl^t(x) = sign(x)(0.d_1 d_2 \ldots \tilde{d}_t \cdot b^p)$$

[2] In Matlab l'insieme dei numeri macchina è $\mathbb{F}(2, 53, -1021, 1024)$. Utilizza variabili a 64 bit (definita precisione doppia dallo standard IEEE754). I bit vengono utilizzati nel seguente modo:

- 1 bit di segno
- 52 (53) bit di mantissa. 53 perchè il primo numero dopo la virgola è $\neq 0$ (viene ignorato nella memorizzazione) e quindi per forza 1.
- 11 bit per l'esponente

2.6.3 Stima dell'errore di rappresentazione

Teorema 2.5 (Maggiorazione dell'errore assoluto di rappresentazione)**.**

$$\Delta x = |x - fl^t(x)| \leq \frac{b^{p-t}}{2}$$

Dimostrazione 2.4. Si vuole calcolare un limite superiore per l'errore assoluto derivante dall'impiego di un rappresentante $fl^t(x)$ con t cifre di mantissa in luogo di x:

$$\Delta x = \big|x - fl^t(x)\big|$$

Per definizione, è possibile scrivere:

$$\begin{aligned}\Delta x &= |x - fl^t(x)| \\ &= (0.d_1d_2...d_t...)b^p - (0.d_1d_2...\tilde{d_t})b^p\end{aligned}$$

Per il Teorema 2.4:

$$\begin{aligned}\Delta x &= b^p \underbrace{((0.d_1d_2...d_t...) - (0.d_1d_2...\tilde{d_t}))}_{\leq \frac{b^{-t}}{2}} \\ &\leq \frac{b^{p-t}}{2}\end{aligned}$$

□

L'errore dipende da p. Si nota che se $p \to L$ l'errore aumenta mentre se $p \to U$ l'errore diminuisce. Questo è desiderabile, poichè mantiene l'errore relativo costante.

Teorema 2.6 (Maggiorazione dell'errore relativo di rappresentazione)**.** *Esiste un maggiorante dell'errore relativo di rappresentazione ε_m t.c.:*

$$e = \frac{|x - fl^t(x)|}{|x|} \leq \varepsilon_m = \frac{b^{1-t}}{2}$$

Tale maggiorante ε_m è chiamato precisione di macchina *ed è il massimo errore relativo di arrotondamento.*

Lemma 2.6.

$$\varepsilon_m \gg \min \mathbb{F}^+$$

⋆ **Dimostrazione 2.1.**

$$e = \frac{|x - fl^t(x)|}{|x|} \leq \frac{b^{p-t}}{2|x|} \qquad x \neq 0$$

Per la definizione 2.5, la più piccola mantissa è $(0.1)_b$.

Allora, per p fissato:

$$|x| \geq (0.1)_b b^p = b^{-1}b^p = b^{p-1}$$

quindi:

$$\frac{1}{|x|} \leq b^{1-p}$$

Segue:

$$e = \frac{b^{p-t}}{2|x|} \leq \frac{b^{p-t}}{2} b^{1-p} = \frac{b^{1-t}}{2} = \varepsilon_m$$

□

2.7 Operazioni con numeri macchina e propagazione degli errori

2.7.1 Proprietà algebriche dei numeri macchina

É importante soffermarsi sul comportamento delle operazioni elementari in aritmetica finita.

Lavorando su numeri interi, se il risultato dell'operazione cade all'interno dell'insieme di rappresentabilità, le operazioni coincidono con quelle algebriche.

Considerando invece i numeri reali, le operazioni saranno definite solo su numeri di macchina e dovranno avere per risultato ancora un numero di macchina; ovvero ad esempio nel caso dell'addizione si ha che $x \oplus y = fl(fl(x) + fl(y))$.

In particolare:

Teorema 2.7. *In generale, nelle operazioni in macchina sui numeri reali non valgono la proprietà associativa e distributiva. L'unica proprietà algebrica che continua a valere è la proprietà commutativa.*

Si userà la seguente notazione per indicare un'operazione macchina:

$$\odot = \begin{cases} \oplus \\ \otimes \\ \ominus \\ \oslash \end{cases}$$

$$x, y \in \nu, \quad x \odot y = fl^t(fl^t(x) \odot fl^t(y))$$

Teorema 2.8. *Nell'insieme dei numeri macchina $\mathbb{F}$ non vige l'unicità dell'elemento neutro dell'addizione.*

Esempio 2.3. Un possibile controesempio si può costruire ponendo $b = 10$ e $t = 16$ e considerando:

$$1 + 10^{-16}$$

Per rappresentare il risultato servirebbero 17 cifre di mantissa, ma $t = 16$.
Allora, in macchina:

$$1 \oplus 10^{-16} = 1$$

Poichè l'ultima cifra viene arrotondata a 0.
10^{-16} si comporta dunque da elemento neutro, in aggiunta a 0.

□

Lemma 2.7.

$$\varepsilon_m = min\left\{\mu \in \mathbb{F}^+ : 1 \oplus \mu > 1\right\}$$

ε_m è ossia il più piccolo numero macchina che non si comporta come elemento neutro.

Teorema 2.9. *Nell'insieme dei numeri macchina $\mathbb{F}$ non vige la proprietà associativa.*

Esempio 2.4. Siano $b = 10$, $U = 308$.

E' evidente che

$$\frac{10^{200}10^{150}}{10^{100}} = 10^{250}$$

Ma sebbene, correttamente,

$$10^{200} \otimes (10^{150} \oslash 10^{100}) = 10^{250}$$

Eliminando le parentesi e dunque alterando l'ordine in cui vengono effettuate le operazioni, il risultato viene inficiato da un overflow che prima non si manifestava:

$$\underbrace{10^{200} \otimes 10^{150}}_{=10^{350}\ \Rightarrow \text{overflow}} \oslash 10^{100} = \texttt{Inf} \oslash 10^{100} = \texttt{Inf}$$

□

2.8 Condizionamento

2.8.1 Condizionamento delle operazioni elementari

Interessa sapere in che modo un errore su (la rappresentazione approssimata de) i dati iniziali del problema può influenzare il risultato finale.

Si vuole ossia ossia studiare il condizionamento di un'operazione - se piccoli errori nei dati in ingresso causino piccoli o grandi errori nel risultato e se le operazioni siano dunque stabili o meno.

Siano $x, y \in \mathbb{R}$ e $\tilde{x} \approx x,\ \tilde{y} \approx y$, si definisce allora:

$$\varepsilon_x = \frac{|x - \tilde{x}|}{|x|} \qquad \varepsilon_y = \frac{|y - \tilde{y}|}{|y|} \qquad x, y \neq 0$$

Poichè $\tilde{x} = fl^t(x)$ segue per definizione di ε_m che $\varepsilon_x \leq \varepsilon_m$.

Si vuole trovare, per ciascuna operazione, un limite superiore per

$$\frac{|(x \odot y) - (\tilde{x} \odot \tilde{y})|}{|x \odot y|} \qquad x \odot y \neq 0$$

L'errore necessariamente introdotto dalla rappresentazione del risultato è trascurabile:

$$x \odot y = \underbrace{fl^t}_{\text{trascurabile}} (fl^t(x) \odot fl^t(y))$$

Teorema 2.10 (Stabilità del prodotto)**.** *Il prodotto è un'operazione stabile:*

$$\frac{|xy - \tilde{x}\tilde{y}|}{|xy|} \leq \varepsilon_x + \frac{|\tilde{x}|}{|x|}\varepsilon_y \approx \varepsilon_x + \varepsilon_y \qquad xy \neq 0$$

⋆ **Dimostrazione 2.2.** Si può riscrivere

$$\frac{|xy - \tilde{x}\tilde{y}|}{|xy|} = \frac{|xy - \tilde{x}y + \tilde{x}y - \tilde{x}\tilde{y}|}{xy}$$

È possibile applicare la disuguaglianza triangolare ($|a+b| \leq |a| + |b|$) su

$$\frac{|\overbrace{xy - \tilde{x}y}^{a} + \overbrace{\tilde{x}y - \tilde{x}\tilde{y}}^{b}|}{xy}$$

e scrivere

$$\begin{aligned}\frac{|xy - \tilde{x}y + \tilde{x}y - \tilde{x}\tilde{y}|}{xy} &\leq \frac{|xy - \tilde{x}y|}{|xy|} + \frac{|\tilde{x}y - \tilde{x}\tilde{y}|}{|xy|} \\ &= \frac{|x - \tilde{x}||y|}{|xy|} + \frac{|\tilde{x}||y - \tilde{y}|}{|xy|} \\ &= \varepsilon_x + \frac{|\tilde{x}|}{|x|}\varepsilon_y\end{aligned}$$

E poichè

$$\frac{|\tilde{x}|}{|x|} = \frac{|\tilde{x} - x + x|}{|x|} \overset{D.T}{\leq} \frac{|\tilde{x} - x|}{|x|} + \frac{|x|}{|x|} = \varepsilon_x + 1$$

Si può scrivere

$$\frac{|xy - \tilde{x}\tilde{y}|}{|xy|} \leq \varepsilon_x + \frac{|\tilde{x}|}{|x|}\varepsilon_y \approx \varepsilon_x + \varepsilon_y$$

L'errore sul prodotto resta dunque dello stesso ordine di grandezza degli errori sui dati.

□

Teorema 2.11. *Il reciproco è un'operazione stabile:*

$$\frac{\frac{1}{x} - \frac{1}{\tilde{x}}}{\frac{1}{x}} \approx \varepsilon_x$$

Teorema 2.12 (Stabilità dell'addizione)**.** *La somma algebrica è stabile quando rappresenta un'addizione:*

$$\frac{|(x+y) - (\tilde{x} + \tilde{y})|}{|x+y|} \leq \varepsilon_x + \varepsilon_y \qquad sign(x) = sign(y)$$

Teorema 2.13 (Stabilità della sottrazione)**.** *La somma algebrica non è stabile quando rappresenta una sottrazione.*

Corollario 2.1. *La sottrazione tra due numeri prossimi in termini relativi è in particolare estremamente perturbata.*

⋆ **Dimostrazione 2.3** (Addizione)**.** Si riscrive per prima cosa:

$$\begin{aligned}\frac{|(x+y)-(\tilde{x}+\tilde{y})|}{|x+y|} &= \frac{|(x-\tilde{x})+(y-\tilde{y})|}{|x+y|} \\ &\overset{D.T}{\leq} \frac{|x-\tilde{x}|}{|x+y|}\frac{|x|}{|x|} + \frac{|y-\tilde{y}|}{|x+y|}\frac{|y|}{|y|} \\ &= w_1(x,y)\varepsilon_x + w_2(x,y)\varepsilon_y \qquad w_1 = \frac{|x|}{|x+y|} \quad w_2 = \frac{|y|}{|x+y|}\end{aligned}$$

Allora, $sign(x) = sign(y) \Rightarrow w_1, w_2 \leq 1$, e dunque:

$$\frac{|(x+y)-(\tilde{x}+\tilde{y})|}{|x+y|} \leq 1 \cdot \varepsilon_x + 1 \cdot \varepsilon_y$$

□

⋆ **Dimostrazione 2.4** (Sottrazione)**.**

$$\begin{aligned}\frac{|(x+y)-(\tilde{x}+\tilde{y})|}{|x+y|} &= \frac{|(x-\tilde{x})+(y-\tilde{y})|}{|x+y|} \\ &\overset{D.T}{\leq} \frac{|x-\tilde{x}|}{|x+y|}\frac{|x|}{|x|} + \frac{|y-\tilde{y}|}{|x+y|}\frac{|y|}{|y|} \\ &= w_1(x,y)\varepsilon_x + w_2(x,y)\varepsilon_y \qquad w_1 = \frac{|x|}{|x+y|} \quad w_2 = \frac{|y|}{|x+y|}\end{aligned}$$

In caso di addendi di segno discorde: $sign(x) \neq sign(y) \Rightarrow 0 \leq w_1, w_2 < \infty$. In particolare, se $|x| \approx |y|$ allora $w_1, w_2 \gg 1$.

Poichè:

$$|x+y| \to 0 \Rightarrow \frac{|x|}{|x+y|}, \frac{|y|}{|x+y|} \to \infty$$

Allora:

$$|x+y| \to 0 \Rightarrow \frac{|(x+y)-(\tilde{x}+\tilde{y})|}{|x+y|} \leq w_1(x,y)\varepsilon_x + w_2(x,y)\varepsilon_y \to \infty$$

□

2.8.2 Condizionamento del calcolo soluzioni di equazioni di secondo grado

La soluzione delle equazioni di secondo grado è un problema ben condizionato, ma l'usuale formula può portare a perdita di precisione.

Si consideri l'equazione di secondo grado:

$$ax^2 + bx + c = 0 \qquad a \neq 0$$

$$\Delta = b^2 - 4ac > 0 \Rightarrow x_\pm = \frac{-b \pm \sqrt{\Delta}}{2a}$$

Esistono due possibili casi:

$$\begin{cases} b > 0 : x_+ = \frac{\sqrt{\Delta}-b}{2a} \\ b < 0 : x_- = \frac{-b-\sqrt{\Delta}}{|2a|} \end{cases}$$

Nel primo caso al numeratore si presenta una sottrazione.
Se dovessero essere b, a, c tali che $b^2 \gg |4ac|$ allora (in termini relativi):

$$\Delta = b^2 \ominus 4ac \approx b^2$$

Allora $\sqrt{\Delta} \approx b$ e a causa delle proprietà della sottrazione in macchina $\ominus$ si avrebbe perdita di precisione nel calcolo di $\sqrt{\Delta} - b$:

$$\sqrt{\Delta} \ominus b \to 0$$

È però sufficiente manipolare algebricamente l'x_+ problematico per rimuovere l'operazione instabile:

$$x_+ = \frac{-b+\sqrt{\Delta}}{2a} = \frac{\sqrt{\Delta}-b}{2a}\frac{(\sqrt{\Delta}+b)}{\sqrt{\Delta}+b} = \frac{\Delta - b^2}{2a(\sqrt{\Delta}+b)} = \frac{-2c}{\sqrt{\Delta}+b}$$

Dunque

Definizione 2.11 (Formula risolutiva stabilizzata per le equazioni di secondo grado)**.**

$$x_1 = -sign(b)\frac{2c}{|b|+\sqrt{\Delta}}$$

$$x_2 = -sign(b)\frac{|b|+\sqrt{\Delta}}{2a}$$

Esempio 2.5. Siano $b = 10$, $t = 4$:

$$x^2 + 10x - 1 = 0$$

$$\tilde{x}_+ = \frac{\overbrace{\sqrt{10^4+4}}^{10^4+4\approx 10^4} \ominus 10^2}{2} = 0$$

$$e = \frac{|x_+ - \overbrace{\tilde{x}_+}^{\to 0}|}{|x_+|} = 1 \Rightarrow 100\%$$

Esempio 2.6. Siano $t = 16$, $\varepsilon_m = \frac{10^{-15}}{2}$ (condizioni simili a Matlab):

$$10^{-2}x^2 + 10^4 x - 10^{-2} = 0$$

$$\tilde{x}_+ = \frac{\sqrt{10^8 + 4 \cdot 10^{-4}} - 10^4}{2 \cdot 10^{-2}}$$

$$e = \frac{|x_+ - \tilde{x}_+|}{|x_+|} = 1.1 \cdot 10^{-5}$$

Si perdono ossia 10 ordini di grandezza.

2.8.3 Calcolo di π

Esempio 2.7 (Formula di Viète o Archimede per il calcolo di π)**.** Si consideri la seguente successione, dovuta a François Viète e basata sul metodo di Archimede per il calcolo di π:

$$\begin{cases} z_2 = 2 \\ z_{n+1} = 2^{n-\frac{1}{2}}\sqrt{1-\sqrt{1-4^{1-n}z_n^2}} \end{cases}$$

Vale:

$$\lim_{n\to\infty} z_n = \pi$$

Il calcolo della successione è caratterizzato da un errore che oltre una certa soglia cresce, a causa dell'influenza crescente delle sottrazioni necessarie.

Poichè $\alpha_n \to 0$ per $n \to +\infty$, si ottiene una sottrazione tra due termini x e y che si avvicinano sempre di più.

L'errore che ne deriva cresce esponenzialmente con n:

$$\omega_2(x,\, y) = \frac{|y|}{|x+y|} = \frac{\sqrt{1-\alpha_n}}{1-\sqrt{1-\alpha_n}} =$$

$$= \frac{1+\sqrt{1-\alpha_n}}{1+\sqrt{1-\alpha_n}}\,\frac{\sqrt{1-\alpha_n}}{1-\sqrt{1-\alpha_n}} = \frac{\sqrt{1-\alpha_n}(1+\sqrt{1-\alpha_n})}{1-(1-\alpha_n)} = \frac{2}{\alpha_n} = \frac{4^n}{2z_n^2}$$

È necessario rendere stabile la sottrazione $1-\sqrt{1-\alpha_n}$:

$$1-\sqrt{1-\alpha_n}\,\frac{1+\sqrt{1-\alpha_n}}{1+\sqrt{1-\alpha_n}} = \frac{1-(1-\alpha_n)}{1+\sqrt{1-\alpha_n}} = \frac{\alpha_n}{1+\sqrt{1-\alpha_n}}$$

Sostituendo, si ottiene una nuova successione algebricamente equivalente, ma scritta in modo stabile.

$$\begin{cases} z_2 = 2 \\ z_{n+1} = \frac{\sqrt{2}(z_n)}{\sqrt{1+\sqrt{1-\alpha_n}}} \end{cases}$$

Una successione alternativa deriva dalla serie geometrica:

$$\sum_{j=1}^{+\infty} \frac{1}{j^2}$$

Essa converge a $\frac{\pi^2}{6}$, e la sua successione di somme parziali è esprimibile come:

$$\begin{cases} S_1 = 1 \\ S_{n+1} = S_n + \frac{1}{(n+1)^2} \end{cases} \qquad n = 1,\, 2,\, \ldots$$

Da essa si può costruire una seconda successione $\{u_n\}$ con $u_n = \sqrt{6S_n}$, convergente a π:

$$\lim_{n \to +\infty} u_n = \pi$$

L'algoritmo è stabile, ma non è utilizzato poichè converge molto lentamente.

In particolare, l'errore commesso rispetto a π decade con un fattore $\frac{1}{n}$: per ottenere, ad esempio, una precisione di 10^{-16}, sono necessarie 10^{16} iterazioni.

2.8.4 Condizionamento di una funzione

Si supponga di dover risolvere un problema rappresentabile nella forma $x = f(x)$, dove x sono i dati in ingresso, y rappresenta il risultato e f è un metodo numerico.

Quello che ci si troverà a risolvere in macchina è in reltà un problema:

$$\tilde{y} = \tilde{f}(\tilde{x})$$

$\tilde{f}$ è una funzione *perturbata* (a causa di operazioni eseguite in aritmetica finita, con possibili errori di di discretizzazione e/o convergenza) su un dato a sua volta *perturbato* (a causa dell'errore di rappresentazione sempre presente o dell'origine sperimentale del dato).

Per semplicità si studierà il problema $\tilde{y} = f(\tilde{x})$, considerando il metodo numerico come eseguito in artimetica esatta.

Definizione 2.12 (Condizionamento di un problema)**.** Un problema è *ben condizionato* quando piccole perturbazioni nei dati iniziali hanno un piccolo effetto sul risultato finale.

Contrariamente, un problema *mal condizionato* è un problema in cui la soluzione è molto sensibile a piccole perturbazioni nei dati.

Osservazione 2.1. Il condizionamento è una proprietà intrinseca della funzione, indipendente dall'algoritmo utilizzato (coincidente con la forma in cui è scritta), mentre la stabilità è una proprietà degli algoritmi.

Definizione 2.13 (Funzione di condizionamento)**.**

$$cond f(x) = \frac{\varepsilon_f}{\varepsilon_x}$$

Dove ε_x è l'errore relativo sull'input, ε_f è l'errore relativo sul risultato.

Teorema 2.14. *Sia $f : \mathbb{R} \to \mathbb{R}, f \in C'$ (ad una variabile, differenziabile).*

$$cond f(x) = \left| \frac{x f'(x)}{f(x)} \right|$$

Dimostrazione 2.5. Se $f : \mathbb{R} \to \mathbb{R}, f \in C'$ è noto che:

$$f'(x) = \lim_{h \to 0} f(x+h) - f(x)h = 0$$

E quindi nelle vicinanze di x:

$$f(x+h) \approx f(x) + f'(x) \cdot h$$

Allora

$$f(\tilde{x}) \approx f(x) + f'(x)(\tilde{x} - x)$$

Applicando la consueta definizione di errore relativo ε:

$$\varepsilon_f = \frac{|f(\tilde{x}) - f(x)|}{|f(x)|} \qquad f(x) \neq 0$$

È possibile riscrivere:

$$\begin{aligned}\varepsilon_f &\approx \frac{|f'(x)(\tilde{x} - x)|}{|f(x)|}\frac{|x|}{|x|} \\ &= \left|\frac{xf'(x)}{f(x)}\right|\frac{|\tilde{x} - x|}{|x|} = condf(x)\varepsilon_x\end{aligned}$$

□

Definizione 2.14. Sia $f : \mathbb{R}^2 \to R, f \in C'$.

$$f(\tilde{x}, \tilde{y}) \approx f(x, y) + \frac{\delta f}{\delta x}(x, y)(\tilde{x} - x) + \frac{\delta f}{\delta x}(x, y)(\tilde{y} - y)$$

$$\varepsilon_f = \frac{|f(\tilde{x}, \tilde{y}) - f(x, y)|}{|f(x, y)|} \approx \frac{|\frac{\delta f}{\delta x}(\tilde{x} - x) - \frac{\delta f}{\delta y}(\tilde{y} - y)|}{|f|} \leq \frac{|x\frac{\delta f}{\delta x}|}{|f|}\underbrace{\frac{|\tilde{x} - x|}{|x|}}_{\varepsilon_x} + \left|y\frac{\delta f}{\delta y}\right|\underbrace{\frac{|\tilde{y} - y|}{|y|}}_{\varepsilon_y}$$

Definizione 2.15 (Formula degli errori)**.**

$$\varepsilon_{f(x,y)} \approx \underbrace{\left|x\frac{\delta f(x, y)}{\delta x}\right|}_{w_1}\varepsilon_x + \underbrace{\left|y\frac{\delta f(x, y)}{\delta y}\right|}_{w_2}\varepsilon_y$$

Esempio 2.8.

$$f(x, y) = x + y$$

$$\varepsilon_{x+y} \lesssim \frac{|x|}{|x + y|}\varepsilon_x + \frac{|y|}{|x + y|}\varepsilon_y$$

$$\frac{\delta f}{\delta x} = 1 + 0 = 1$$

$$\frac{\delta f}{\delta y} = 1$$

$$w_1 = \frac{|x \cdot 1|}{|x + y|}$$

$$w_2 = \frac{|y \cdot 1|}{|x + y|}$$

Come visto in precedenza, la somma algebrica è una funzione ben condizionata se non è vero che $|x + y| \to 0$.

Esempio 2.9.

$$f(x,y) = xy$$
$$\frac{\delta f}{\delta x} = y$$
$$\frac{\delta f}{\delta y} = x$$
$$w_1 = \frac{|xy|}{|xy|} = 1$$
$$w_2 = \frac{|yx|}{|xy|} = 1$$

Il prodotto risponde bene agli errori sui dati.

Esempio 2.10. Si consideri

$$f(x) = 1 - x$$

Calcolando $f(x)$ in $x \approx 1$:

$$condf(x) = \left|\frac{x(-1)}{1-x}\right| = \left|\frac{x}{1-x}\right|$$
$$lim_{x\to 1}\left|\frac{x}{1-x}\right| = \infty$$

$f(x)$ è una funzione mal condizionata (per $x \approx 1$)

Esempio 2.11.

$$f(x) = \frac{(1+x)-1}{x} \equiv 1 \qquad x \neq 0$$
$$condf(x) = \left|\frac{x \cdot 0}{f(x)}\right| = 0$$

Va osservato che $f(x)$ è diversa da $x \longmapsto 1$. Molte funzioni possono non essere mal condizionate ma possiamo avere problemi nel processo di calcolo a causa dell'*algoritmo* utilizzato.

In questo caso, ad esempio la funzione è in una forma notoriamente instabile per $x \approx 0$.

Esempio 2.12.

$$f(x) = 1 - \sqrt{1-x^2} \qquad |x| \leq 1$$
$$f'(x) = \not{-}\frac{1}{\not{2}\sqrt{1-x^2}} \not{-} \not{2}x = \frac{x}{\sqrt{1-x^2}}$$
$$condf(x) = \frac{\left|x \cdot \frac{x}{\sqrt{1-x^2}}\right|}{|1 \cdot \sqrt{1-x^2}|} = \frac{x^2}{\sqrt{1-x^2}} \cdot \frac{1}{1-\sqrt{1-x^2}} \cdot \frac{1+\sqrt{1-x^2}}{1+\sqrt{1-x^2}}$$
$$= \frac{\not{x}^2}{\sqrt{1-x^2}} \cdot \frac{1+\sqrt{1-x^2}}{\not{1} - (\not{1} - \not{x}^2)} = \frac{1+\sqrt{1-x^2}}{\sqrt{1-x^2}} \underset{x\to 0}{\to} 2$$

Questa f è ben condizionata, ma l'*algoritmo* è instabile. E' possibile correggerlo cosi':

$$f(x) = 1 - \sqrt{1-x^2} \cdot \frac{1+\sqrt{\cdots}}{1+\sqrt{\cdots}} = \frac{x^2}{1+\sqrt{1-x^2}}$$

L'algoritmo è allora stabile (per $x \approx 0$).

2.9 Complessità computazionale

Definizione 2.16. La complessità computazionale $C(n)$ è il numero di operazioni in virgola mobile (FLOP) coinvolte nell'esecuzione di un algoritmo in funzione della dimensione n dell'input.

Il costo computazionale espresso in operazioni piuttosto che in *tempo di esecuzione* e' una buona misura perchè è slegato dalla potenza del calcolatore di destinazione, ma non tiene conto del *flusso* delle istruzioni.

Algoritmi diversi per risolvere lo stesso problema hanno potenzialmente complessità diverse.

Esempio 2.13 (Algoritmo di Hörner)**.** Sia $p(x)$ un polinomio qualsiasi

$$p(x) = a_0 + a_1x + a_2x^2 + ... + a_nx^n$$

ll costo del calcolo "ingenuo" di $p(x)$ è:

$$C(n) = 3n$$

L'algoritmo di Hörner consiste nel rappresentare il polinomio come:

$$p(x) = ((a_nx + a_{n-1}) + a_{n-2})x + ...a - 0$$

Allora è evidente che ogni iterazione costa solo 2 operazioni in virgola mobile al posto di 3, e quindi:

$$C(n) = 2n$$

Esempio 2.14 (Calcolo delle potenze)**.** Si vuole calcolare una potenza a^n, dove $a \in \mathbb{R}$, $n \in \mathbb{N}$.

$$a^n = \underbrace{a \cdot a \cdot \ldots \cdot a}_{n-1\ prodotti}$$

Una soluzione veloce, che impieghi meno di $n-1$ prodotti, consiste nel suddividere i prodotti in potenze di due.

$$\left.\begin{array}{ll} a \cdot a & = a^2 \\ a^2 \cdot a^2 & = a^4 \\ \vdots & \\ a^{n/2} \cdot a^{n/2} & = a^n \end{array}\right\} \log_2 n \text{ operazioni}$$

Se n è una potenza di 2, la complessità è

$$C(n) = log_2 n$$

In caso contrario, si può considerare la rappresentazione in base 2 dell'esponente:

$$n = \sum_{j=0}^{m} c_j 2^j$$

$$a^n = a^{c_0} \cdot a^{c_1 2} \cdot a^{c_2 2^2} \cdot \cdots \cdot a^{c_m 2^m}$$

$$\left.\begin{array}{ll} a \cdot a & = a^2 \\ a^2 \cdot a^2 & = a^4 \\ \vdots & \\ a^{2^{m-1}} \cdot a^{2^{m-1}} & = a^{2^m} \end{array}\right\} m \text{ operazioni}$$

L'algoritmo impiega m operazioni per calcolare i singoli fattori $a^{c_j 2^j}$, e un certo numero di prodotti tra essi per ottenere a^n, al più m (se $c_j = 0$, si risparmia un prodotto).

La complessità $C(n)$ è quindi al più $2m$.

Esempio 2.15. Data la seguente matrice

$$A^n, A \in \mathcal{R}^{m \times m}$$

Abbiamo che

$$A^n = \prod_{j=0}^{m} A^{2^j} + \text{eventuali m prodotti di matrici}$$

Per $m = 10^2$ il calcolo del prodotto ha costo $O(10^6)$ FLOPS, mentre per $m = 10^3$ il calcolo del prodotto ha costo $O(10^9)$ FLOPS.

$$(AB)_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$$

Ha costo ≈ 2 MFLOPS.

Esempio 2.16. Consideriamo il problema di calcolare e^x. Sia $a > 0$, allora è noto che

$$a^n = e^{n \log a}$$

$$f(x) = e^x, x > 0$$

Consideriamo la serie di Taylor corrispondente (con resto di Lagrange):

$$e^x = e^0 + e^0 x + e^0 \frac{x^2}{2} + e^0 \frac{x^3}{3!} + \ldots + e^0 \frac{x^m}{m!} + \overbrace{R_m(x)}^{\frac{e^\xi x^{m+1}}{(m+1)!}}$$

$R_m(x) = \frac{e^\xi x^{m+1}}{(m+1)!}$ con $\xi \in (0, x)$ converge a 0 per $m \to \infty$, $x \in [0, 1]$.

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)(x - x_0)^2}{2} + \frac{f'''(x_0)(x - x_0)^3}{3!} + \ldots + \frac{f^m(x_0)(x - x_0)^m}{n} + \underbrace{R_m(x)}_{\text{errore}}$$

Con $f \in C^{n+1}(I_{x_0})$ (f derivabile n+1 volte nell'intorno di x_0).

$$R_m(x) = \frac{f^{m+1}(\xi)}{(m+1)!}(x - x_0)^{m+1} \text{ con } \xi \in (x_0, x)$$

Vicino a 0 l'espansione di Taylor funziona bene:

$$0 \leq R_m(k) \leq \xi e^b \underbrace{\frac{b^{m+1}}{(m+1)!}}_{\to 0}$$

Stima dell'errore:

$$e^x = T_m(x) + R_m(x)$$

$$R_m(x) = \frac{e^\xi x^{m+1}}{(m+1)!} \leq \underbrace{\frac{e^x x^{m+1}}{(m+1)!}}_{\substack{\to 0 \\ m \to \infty}}$$

$$\frac{|e^x - T_m(x)|}{|e^x|} = \frac{e^\xi x^{m+1}}{(m+1)!}$$

$$\begin{cases} 0 < x < 1 : \varepsilon_x \leq \underbrace{\frac{1}{(m+1)!}}_{\substack{\to 0 \\ \text{presto}}} \\ x > 1 : \frac{1}{m!} \approx \varepsilon_m \text{ per } m = 18 \end{cases}$$

Il caso due si puo' risolvere agevolmente con la seguente astuzia:

$$e^x = (e^{\frac{x}{n}})^n = a^n = (e^y)^n$$

dove

$$a^{\frac{x}{n}} = a^y$$

con $0 \leq y \leq 1$, $n = [x] + 1$ Il che equivale a trovarsi sempre nel caso 1 con l'aggiunta di una potenza rapida.

Esempio 2.17 (Sviluppo di Laplace)**.** Consideriamo il problema di calcolare il determinante di una matrice $A \in \mathbb{R}^{n \times n}$.

La sviluppo di Laplace è un metodo particolarmente inefficiente di calcolare il determinante[3]:

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \det(A_{i,j})$$

dove $A_{i,j}$ è una matrice $\mathbb{R}^{(n-1)\times(n-1)}$ ottenuta da A cancellando la riga i-esima e la colonna j-esima.

Il costo del calcolo del determinante di A con Laplace è $C(n) \approx 2n!$ e il numero di moltiplicazioni necessario è $\approx n^n$.

n	t @ 1GFLOPS	t @ 1TFLOPS
10	4 ms	–
15	24 minuti	–
20	154 anni	$\approx$ 2 mesi
25	10^9 anni	10^6 anni
100	10^{141} anni	10^{13} anni

[3] Metodi inefficienti nell'applicazione pratica possono avere dei risvolti teorici importanti - è questo il caso dello sviluppo di Laplace

Esempio 2.18 (MEG)**.** Il metodo di eliminazione di Gauss è un algoritmo per risolvere sistemi di equazioni lineari per mezzo di una *sequenza di operazioni* (mosse di Gauss) operate sulla matrice associata.

$$A \overset{MEG}{\to} U$$

Dove $\overset{MEG}{\to}$ è una sequenza di mosse di Gauss risulta in una matrice triangolare superiore U tale che $\det A = \pm \det U = \pm \prod_{i=1}^{n} a_{ii}$.

Ricordiamo le proprietà del determinante:

- Dato $R_k = R_k + \alpha R_i$ il determinante non cambia. Significa che se sommo o sottraggo ad una riga (colonna) una qualunque riga (colonna) moltiplicata per un numero α il determinante non cambia;
- Scambiando due righe R_i (oppure due colonne R_j) il determinante cambia segno.

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 2 & 3 \\ -1 & -3 & 0 \end{pmatrix} \overset{R_2=R_2+(-2)R_1}{\to} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 1 \\ -1 & -3 & 0 \end{pmatrix}$$

$$\overset{R_3=R_3+R_1}{\to} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 1 \\ 0 & -1 & 1 \end{pmatrix} \overset{R_3=R_3+(-1/2)R_2}{\to} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 1 \\ 0 & 0 & 1/2 \end{pmatrix} = U$$

Non c'è stato scambio di righe, dunque il determinante di U vale

$$\det U = -1 = \det A$$

L'algoritmo del MEG è un algoritmo *finito*, ma delicato da trattare in aritmetica di macchina.

```
MEG(A ∈ R^{n×n}, n)
 1  s_1 ← 0
 2  B ← A
 3  for i ∈ (1, n − 1)
 4       do
 5          cerca indice (p : |b_pi| = max(b_ti, i ≤ t ≤ n)).
 6          if b_pi = 0
 7             then
 8                  det = 0
 9                  return det
10             else
11                  if p ≠ i
12                     then
13                          R_i ↔ R_p
14                  s ← s + 1
15                  for k ∈ (i + 1, n)
16                       do
17                          _k ← R_k + R_i (b_ki / b_ii)
18                  det B = −πb_ii
19                  return det B
```

3 Soluzione numerica di equazioni non lineari

Sia data un'equazione della forma:

$$f(x) = 0, \qquad x \in \mathbb{R}, \quad f : \mathbb{R} \to \mathbb{R}$$

In generale si possono presentare tre diverse situazioni:

- $f(x)$ ha un numero finito di soluzioni. Si consideri ad esempio: $f(x) = (x-1)(x-2)$
- $f(x)$ non ha soluzioni reali. Ad esempio: $f(x) = x^2 + 1$
- $f(x)$ ha un numero infinito di soluzioni. Ad esempio: $f(x) = \sin(x)$

L'esistenza di una soluzione è garantita dal teorema degli zeri per quelle funzioni che ne soddisfano i requisiti:

Teorema 3.1 (Teorema degli zeri)**.** *Siano due punti a, b e una funzione f tale che:*

$$a, b \in \mathbb{R}, \quad f \in C[a,b], \quad f(a)f(b) < 0$$

Allora

$$\exists \xi \in (a,b) : f(\xi) = 0$$

Intuitivamente, la funzione, passando almeno una volta da positiva a negativa all'interno dell'intervallo $[a,b]$, deve necessariamente attraversare l'asse delle ascisse in almeno un punto.

Lemma 3.1. Se f è strettamente crescente o decrescente allora sappiamo che la soluzione è unica.

3.1 Il metodo della bisezione

Per trovare uno zero di $f(x)$ con il metodo della bisezione occorre in primo luogo poter scegliere un intervallo $[a,b]$ contenente *sicuramente* uno zero della funzione.

Si sceglie poi un punto interno all'intervallo che sia una "migliore possibile" approssimazione della soluzione (ξ) - di regola questa coincide con il punto medio: $x_0 = \frac{a+b}{2}$

Possono di conseguenza verificarsi solo tre casi:

- $f(x_0) = 0$: x_0 è la radice cercata.
- $f(a)f(x_0) < 0$: la funzione interseca l'asse delle ascisse prima del punto x_1: si può allora ripetere il procedimento sull'intervallo $[a, x_0]$
- $f(x_0)(b) < 0$: la funzione interseca l'asse delle ascisse dopo il punto x_1: si può ripetere il procedimento sull'intervallo $[x_0, b]$

Il procedimento equivale a costruire tre successioni $\{a_n\}$, $\{b_n\}$, $\{x_n\}$ (estremi sinistri, estremi destri e punti medi).

3.1.1 Esistenza di soluzioni

Teorema 3.2. *Esiste certamente una soluzione a cui il metodo di bisezione converge se la funzione soddisfa le condizioni del teorema degli zeri:*

$$f \in C[a,b], \quad f(a)f(b) < 0 \Rightarrow \exists \xi \in [a,b] : f(\xi) = 0$$

Inoltre:

Teorema 3.3. *La soluzione è anche unica se la funzione è strettamente crescente o decrescente nell'intervallo (a,b):*

$$\forall x \in (a,b) : \begin{cases} f'(x) > 0 \\ f'(x) < 0 \end{cases} \Rightarrow \exists! \xi \in (a,b) : f(\xi) = 0$$

3.1.2 Velocità di Convergenza

Le successioni a_n, b_n, x_n godono delle seguenti proprietà:

Teorema 3.4.

$$b_n - a_n = \frac{b-a}{2^n}$$

Teorema 3.5.

$$|a_n - \xi|, |b_n - \xi| \leq \max(|a_n - \xi|, |b_n - \xi|) < b_n - a_n = \frac{b-a}{2^n}$$

Teorema 3.6.

$$0 < e_n = |x_n - \xi| < \frac{b_n - a_n}{2} = \frac{b-a}{2^{n+1}}$$

Teorema 3.7.

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \lim_{n\to\infty} x_n = \xi$$

La convergenza del metodo di bisezione è molto lenta.

È difficile trarre conclusioni generali oltre ad affermare che l'errore e_n è maggiorato da una successione s_n tale che $s_{n+1} = \frac{1}{2}s_n$.

Lemma 3.2 (Velocità di guadagno delle cifre significative)**.** La velocità di convergenza del metodo di bisezione è logaritmica.

Per guadagnare una cifra significativa nell'approssimazione di ξ, ossia per avere

$$|x_n - \xi| = \frac{|x_m - \xi|}{10}$$

Occorrono $n - m = log_2(10) \approx 3.32$ iterazioni.

3.1.3 Criteri di Arresto

Arresto per stima a priori dell'errore

Teorema 3.8. *Data una tolleranza ε decisa arbitrariamente, è possibile individuare a priori un n t.c. e_n possa essere maggiorata da ε:*

$$e_n = |x_n - \xi| \leq \frac{b_n - a_n}{2} \leq \varepsilon$$

Lemma 3.3. Per garantire

$$e_n = |x_n - \xi| \leq \varepsilon$$

Occorre

$$n \geq \log(b-a) - \log(\varepsilon)$$

È importante rilevare che:

Teorema 3.9. *Con il metodo di bisezione l'errore non è in generale monotono decrescente.*

Dimostrazione 3.1. Come controesempio si consideri una funzione con un unico zero in $2 + \frac{3}{4}$.

Preso un intervallo iniziale $[a, b] = [1, 5]$, $|x_0 - \xi| = e_0 < e_1 = |x_1 - \xi|$, anche se la situazione migliorerà con n abbastanza grande:

n	a_n	b_n	x_n
0	1	5	3
1	1	3	2
...	...	...	...

□

Arresto per stima a posteriori tramite residuo pesato Una stima a posteriori è una stima dell'errore eseguita *durante il processo di calcolo.*

Ingenuamente, può essere invogliante maggiorare il modulo del residuo con un ε scelto:

$$|f(x_n)| < \varepsilon$$

Non è in generale una buona idea: se la funzione è molto piatta in prossimità di ξ la condizione può essere soddisfatta anche in un punto molto distante da ξ causando una **sottostima** dell'errore.

Può accadere anche il contrario: una buona approssimazione di ξ può non soddisfare la condizione (se la funzione è molto pendente in prossimità della radice) - si avrebbe una **sovrastima** dell'errore.

La stima tramite il **residuo pesato** è una stima **a posteriori**, ovvero calcolata all'interno del processo di calcolo.

Poggia sui seguenti, noti, teoremi dell'Analisi:

Teorema 3.10. *Teorema del valor medio*

$$f \in C[c,d], \quad \exists f' \in (\alpha, \beta) \Rightarrow \exists z \in (c,d) : f'(z) = \frac{f(c) - f(d)}{c - d}$$

Teorema 3.11. *Teorema della permanenza del segno*

$$\{a_n\} \to a_0 > 0 : \exists N : a_n > 0 : \forall n > N$$

Assumendo $f \in \mathcal{C}^1$ (continua e derivabile con derivata prima continua) e $f'(\xi) = 0$ (zero semplice), per la permanenza del segno di f' si può affermare che esiste un intorno in cui $f'(x_n)$ è $\neq 0$:

$$\exists N : f'(x) \neq 0, \ \forall x \in (x_n, \xi), \ n > N$$

Possiamo applicare il teorema del valore medio all'intervallo (x_n, ξ) per poter affermare, per qualche $z_n \in (x_n, \xi)$:

$$\frac{f(x_n) - \overbrace{f(\xi)}^{0}}{x_n - \xi} = \frac{f(x_n)}{x_n - \xi} = f'(z_n)$$

Allora, algebricamente:

$$f(x_n) = f'(z_n)(x_n - \xi)$$
$$\Rightarrow e_n = |x_n - \xi| = \frac{|f(x_n)|}{|f'(z_n)|}$$

Allora è sufficiente arrestare il calcolo quando:

$$e_n = \frac{|f(x_n)|}{|f'(z_n)|} < \varepsilon$$

ovvero, intuitivamente, l'arresto avviene quando il residuo è abbastanza piccolo e la pendenza abbastanza grande.

L'approccio nella pratica dipende se si abbiano informazioni su f' (si potrebbero non avere affatto informazioni sulla funzione, vista come una *black box*).

Test di arresto basato su f' Si consideri l'equazione:

$$e_n = \frac{|f(x_n)|}{|f'(z_n)|} < \varepsilon$$

Se si conosce f' è possibile scegliere una costante $C > 0$ predeterminata t.c. $\forall x \in (a, b) : |f'(x)| \geq C$.

Allora l'errore è così maggiorato:

$$e_n = \frac{|f(x_n)|}{|f'(z_n)|} < \frac{|f(x_n)|}{C} < \varepsilon$$

L'arresto può avvenire dunque quando

$$|f(x_n)| < \varepsilon \cdot C$$

Esempio 3.1 (Calcolo di $\sqrt{2}$)**.** Si consideri:

$$f(x) = x^2 - 2$$

La funzione ha una soluzione ξ in $\sqrt{2}$.

Si osserva che $f(1) = -1$, $f(2) = 2$, dunque è possibile scegliere $(a, b) = (1, 2)$.

Poichè $f'(x) = 2x$, $f'(x) > 0 \quad \forall x > 0$.

Dunque:

$$|f'(x)| \geq 2 \quad x \in [1, 2]$$

Allora la scelta di $C = 2$ garantisce che $f'(x) \geq C$ in $(1, 2)$ e:

$$e_n = |x_n - \xi| = |x_n - \sqrt{2}| \leq \frac{f(x_n)}{C} = \frac{x_n^2 - 2}{2} \leq \varepsilon$$

Una volta scelto un ε sarà dunque sufficiente arrestare il calcolo quando

$$\frac{f(x_n)}{C} = \frac{x_n^2 - 2}{2} \leq \varepsilon$$

Si avrà così la certezza di avere un'approssimazione di $\sqrt{2}$ con errore non superiore a ε.

Test di arresto in condizione di ignoranza di f' Se non si hanno informazioni su f' (il calcolo di $f(x)$ è ridotto ad una *black box*) è possibile usare un test empirico.

Non essendo possibile conoscere f' non si può applicare direttamente:

$$e_n = \frac{|f(x_n)|}{|f'(z_n)|}, z_n \in (x_n, \xi)$$

È possibile però approssimare $f'(z_n)$:

$$f \in C^1 \Rightarrow f'(z_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = f'(u_n) \quad u_n \in (x_n, x_{n-1})$$

Se f' è continua:

$$\lim_{n\to\infty} \begin{cases} f'(z_n) \\ f'(u_n) \end{cases} = f'(\xi)(\neq 0)$$

Infatti $z_n \in (x_n, \xi), u_n \in (x_n, x_{n-1})$, e poichè $x_n \to \xi$ e $x_{n-1} \to \xi$ per il teorema dei Carabinieri:

$$\lim_{x\to\infty} \begin{cases} z_n \\ u_n \end{cases} = \xi$$

3.1.4 Vantaggi e svantaggi del metodo di bisezione

Il metodo di bisezione ha diversi vantaggi, tra cui:

- Il metodo della bisezione lavora con ipotesi minime, ovvero quelle del teorema degli zeri.
- Al passo n può essere sufficiente calcolare solo $sign(f(x_n))$ per scegliere il successivo intervallo (a_{n+1}, b_{n+1}).

Il metodo è però affetto da un singolo, grande svantaggio: il suo fattore di convergenza è non maggiore di $\frac{1}{2}$, ed è dunque un metodo **lento**.

3.2 Metodo di Newton

Il metodo di Newton si basa sull'idea di sfruttare, come successiva approssimazione della radice, il punto di intersezione della retta tangente alla funzione del punto $(x_0, f(x_0))$ con l'asse delle ascisse, considerando x_0 un'approssimazione iniziale "sufficientemente buona" (ottenuta per ispezione grafica o con un paio di iterazioni del metodo di bisezione).

L'idea fondamentale è che, anche in una funzione non lineare, la tangente in un punto possa "puntare" all'incirca verso lo zero. Banalmente allora occorre che $f \in C'$.

Poichè l'equazione della retta tangente al punto $(x_0, f(x_0))$ è:

$$y = f(x_0) + f'(x_0)(x - x_0)$$

L'equazione dell'intersezione della retta con l'asse x si ottiene mettendo a sistema a risolvendo:

$$\begin{cases} y = f(x_0) + f'(x_0)(x - x_0) \\ y = 0 \end{cases}$$

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Più in generale si costruirà una successione $\{x_n\}$ t.c.:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad n = 0, 1, 2, ...$$

3.2.1 Convergenza del Metodo di Newton

Il metodo di Newton richiede che

$$f \in C'[a, b]$$
$$\forall x \in [a, b] : f'(x) \neq 0$$

La convergenza del metodo di Newton non è sempre garantita. Esistono diversi teoremi che sotto particolari ipotesi permettono di provarne la convergenza. Alcuni sono detti di *convergenza locale* e dicono che se $x_o \in \mathscr{I}$ intorno di ξ sufficientemente piccolo allora il metodo converge. Altri sono detti di *convergenza globale* e dicono che se x_0 appartiene a un ben definito intorno $\mathscr{I}$ di ξ allora il metodo converge.

Teorema 3.12 (Convergenza del metodo di Newton)**.** *Sia $f \in \mathcal{C}^2\,[a,\, b]$ tale che:*

- $f(a)f(b) < 0$.
- $f''(x) > 0 \vee f''(x) < 0,\ \forall x \in [a,\, b]$

  La funzione è ossia strettamente concava o strettamente convessa.
- $f'(x) \neq 0,\ \forall x \in [a,\, b]$.
- $x_0 \,:\, f(x_0)f''(x_0) > 0$

  La funzione e la derivata seconda hanno ossia segni concordi.

Sotto queste ipotesi di convergenza globale, il metodo di Newton è ben definito e si ha

$$\lim_{n\to+\infty} e_n = \lim_{n\to+\infty} |x_0 - \xi| = 0$$

Lemma 3.4. Sotto le stesse ipotesi, la successione x_n è decrescente.

Dimostrazione 3.2. Quattro casi possibili soddisfano le condizioni iniziali:

$$\begin{cases} f''(x) > 0 \begin{cases} f(a) < 0, f(b) > 0 \\ f(a) > 0, f(b) < 0 \end{cases} \\ f''(x) < 0 \begin{cases} f(a) < 0, f(b) > 0 \\ f(a) > 0, f(b) < 0 \end{cases} \end{cases}$$

Si considererà il caso $f''(x) > 0,\ f(a) < 0, f(b) > 0$ (gli altri casi sono simmetrici).

Se $f(a) < 0$ e $f(\xi) = 0$ allora $f' > 0$ in un intorno sinistro sufficientemente piccolo di ξ.

Se $f' > 0$ in un intorno sinistro sufficientemente piccolo di ξ, poichè $f'' > 0$ allora $f' > 0$ in $[\xi, b]$.

Allora in $[\xi, b]$

$$\frac{f(x_n)}{f'(x_n)} > 0$$

Quindi

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} < x_n$$

□

Dimostrazione 3.3. Si considererà ancora il caso $f''(x) > 0, f(a) < 0, f(b) > 0$ (gli altri casi sono simmetrici).

Poichè $f'' > 0$ l'intersezione della tangente con l'asse delle x è sempre a destra dell'intersezione della f con l'asse medesima, dunque x_n è inferiormente limitata da ξ: $x_n \geq \xi$.

Per questo e poichè $\{x_n\}$ è decrescente è vero: $\forall n : x_n \in [\xi, b]$.

Posto allora che:

$$\eta = \inf\{x_n\} \qquad (\eta \geq \xi)$$

Poichè x_n è strettamente decrescente:

$$\lim_{n\to\infty} x_n = \eta$$

$$\begin{aligned} x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \Rightarrow \lim_{n\to\infty} x_{n+1} &= \lim_{n\to\infty} x_n - \lim_{n\to\infty} \frac{f(x_n)}{f'(x_n)} \\ \eta &= \eta - \lim_{n\to\infty} \frac{f(x_n)}{f'(x_n)} \\ \eta &= \eta - \frac{f(\lim x_n)}{f'(\lim x_n)} \textit{(per continuità)} \\ \eta &= \eta - \frac{f(\eta)}{f'(\eta)} \end{aligned}$$

$$\frac{f(\eta)}{f'(\eta)} = 0 \Rightarrow f(\eta) = 0$$

Allora η è soluzione, e per unicità:

$$\eta = \xi$$

□

Teorema 3.13 (Maggiorazione dell'errore nel metodo di Newton)**.** *Sia $f \in C^2(\mathscr{I})$ tale che, con $e_n = |x_n - \xi|$:*

- $\lim_{n\to\infty} x_n = \xi \Leftrightarrow \lim_{n\to\infty} e_n = 0$
- $\{x_n\} \subset \mathscr{I}, \xi \in \mathscr{I}$ *con $\mathscr{I}$ intervallo chiuso e limitato e* $f'(x) \neq 0 \ \forall x \in I$

Allora

$$\exists\ C > 0 : e_{n+1} \leq C e_n^2$$

⋆ **Dimostrazione 3.1.** Si rammenti:

Teorema 3.14 (Polinomio di Taylor (con resto di Lagrange))**.** *$f \in C^2(x, x_0)$ è esprimibile da un polinomio di Taylor di secondo ordine centrato su x_0:*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(z)}{2}(x - x_0)^2 \qquad z \in int(x_0, x)$$

Pertanto si può esprimere $f(\xi)$ come

$$f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{f''(z_n)}{2}(\xi - x_n)^2$$

Dunque, manipolando algebricamente:

$$\begin{aligned}
f(\xi) = 0 &= f(x_n) + f'(x_n)(\xi - x_n) + \frac{f''(z_n)}{2}(\xi - x_n)^2 \\
-f(x_n) &= f'(x_n)(\xi - x_n) + \frac{f''(z)}{2}(\xi - x_n)^2 \\
-\frac{f(x_n)}{f'(x_n)} &= (\xi - x_n) + \frac{f''(z_n)}{2}\frac{1}{f'(z_n)}(\xi - x_n)^2 \\
\underbrace{x_n - \frac{f(x_n)}{f'(x_n)}}_{x_{n+1}} - \xi &= \frac{f''(z_n)}{2f'(x_n)}(\xi - x_n)^2 \\
\underbrace{|x_{n+1} - \xi|}_{e_{n+1}} &= \left|\frac{f''(z_n)}{2f'(x_n)}\right| \cdot \left|(\xi - x_n)^2\right| \\
e_{n+1} &= C_n e_n^2
\end{aligned}$$

Si ricordi a questo punto il teorema di Weierstrass:

Teorema 3.15 (Teorema di Weierstrass)**.** *Sia $f \in C[a, b]$ chiuso e limitato. Allora*

$$\exists (c, d) \in [a, b] : f(c) \leq f(x) \leq f(d) \qquad \forall\ x \in [a, b]$$

Pertanto, $\left|f''(x)\right|$ ha un massimo assoluto in $\mathscr{I}$ e $\left|f'(x)\right|$ ha un minimo assoluto in I:

$$|f''(z_n)| \leq \max_{x\in[a,b]} |f''(x)| = M$$
$$|f'(x_n)| \geq \min_{x\in[a,b]} |f'(x)| = m > 0$$

Dunque, scelto $C = \frac{M}{2m}$:

$$C_n = \left|\frac{f''(z_n)}{2f'(x_n)}\right| \leq \frac{M}{2m} = C$$

E segue dunque:

$$e_{n+1} = C_n e_n^2 \leq C e_n^2$$

□

3.2.2 Ordine di convergenza del metodo di Newton

Definizione 3.1 (Ordine di convergenza)**.** Sia $\{x_n\}$ una successione convergente a μ e sia $e_n = x_n - \mu$ l'errore al passo k.

Se $\exists p > 0, C \neq 0$ t.c.:

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = C$$

Allora p è chiamato *ordine di convergenza* della successione e C è la *costante asintotica di errore*.

Per $p = 1$ la convergenza si dice *lineare*, per $p = 2$ si dice *quadratica*.

Teorema 3.16. *Il metodo di Newton converge con ordine almeno 2*

Lemma 3.5. Se $f''(\xi) \neq 0$ l'ordine di convergenza del metodo di Newton è esattamente 2.

⋆ **Dimostrazione 3.2.**

$$\begin{aligned}\lim_{n\to\infty} \frac{|e_{n+1}|}{\left|e_n^p\right|} &= \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n^2|} \\ &= \lim_{n\to\infty} \frac{\left|C_n e_n^2\right|}{|e_n^2|} \\ &= \lim_{n\to\infty} C_n \\ &= \lim_{n\to\infty} \left|\frac{f''(z_n)}{2f'(x_n)}\right| = \left|\frac{f''(\xi)}{2f'(\xi)}\right|\end{aligned}$$

Per definizione di ordine di convergenza, $p = 2$ se $f''(\xi) \neq 0$.

□

3.2.3 Confronto con il metodo di bisezione

Se il metodo di Newton converge, ovvero $\{x_n\} \to 0$ è triviale che:

$$\exists \bar{n} : C\, e_n \leq \Theta, \quad \forall n \geq \bar{n},\, \Theta > 0$$

Fissato un $\Theta \in [0,\, 1]$, si ha:

$$\begin{aligned} e_{n+1} &\leq C\, e_n^2 \\ C\, e_{\bar{n}+1} &\leq C^2\, e_{\bar{n}}^2 \\ C\, e_{\bar{n}+1} &\leq (C\, e_{\bar{n}})^2 = \Theta^2 \\ C\, e_{\bar{n}+2} &\leq (C\, e_{\bar{n}+1})^2 = \Theta^4 \\ C\, e_{\bar{n}+k} &\leq \Theta^{2^k} \end{aligned}$$

Ponendo $\Theta = 1/2$ è possibile confrontare il metodo di Newton con il metodo di bisezione:

$$C\, e_{\bar{n}+k} \leq \left(\frac{1}{2}\right)^{2^k} \qquad \text{Newton}$$

$$C\, e_{\bar{n}+k} \leq \left(\frac{1}{2}\right)^{k} \qquad \text{Bisezione}$$

3.2.4 Convergenza globale e locale

Le ipotesi di convergenza globale garantiscono la convergenza del metodo di Newton per qualsiasi errore iniziale $e_0 = |x_0 - \xi|$, ma in certi casi possono essere troppo restrittive. Si vorrebbe trovare un valore di x_0 sufficientemente vicino alla soluzione da garantire la convergenza in ipotesi meno restrittive.

Sia $\bar{n} = 0$ e $C\, e_0 < 1$ nella (**??**):

$$C\, e_k \leq (C\, e_0)^{2^k}, \quad \forall k \geq 0$$

Passando al limite:

$$\begin{aligned} \lim_{k \to +\infty} C\, e_k &\leq \lim_{k \to +\infty} (C\, e_0)^{2^k} \\ &\leq 0 \end{aligned}$$

Il metodo converge per $C\, e_0 < 1$, quindi:

$$C\, e_0 < 1 \implies |x_0 - \xi| < \frac{1}{C}$$

Ipotesi di convergenza locale:

- $f \in \mathcal{C}^2\,[\xi - \delta,\, \xi + \delta] = \mathscr{I}$
- $f'(x) \neq 0, \quad \forall x \in \mathscr{I}$
- $x_0 \in \mathscr{I} : |x_0 - \xi| < \min\{\frac{1}{C},\, \delta\}$

Si noti che le ipotesi di convergenza locale sono meno restrittive in quanto non richiedono convessità e/o concavità strette.

3.2.5 Stima dell'errore

Stimare l'errore con il passo $|x_{n+1} - x_n|$ è pericoloso in generale, ma nel caso del metodo di Newton porta a buoni risultati per costruzione stessa del metodo:

$$|x_{n+1} - x_n| = \left| x_n - \frac{f(x_n)}{f'(x_n)} - x_n \right| = \frac{f(x_n)}{f'(x_n)}$$

Si noti che il passo $|x_{n+1} - x_n|$ assume la forma di un residuo pesato. L'arresto del metodo, fissata una tolleranza $\varepsilon > 0$, avverrà quando:

$$|x_{n+1} - x_n| < \varepsilon$$

3.2.6 Esempi

Esempio 3.2 (Calcolo $\sqrt{2}$)**.** Si consideri

$$f(x) : x \mapsto x^2 - 2$$

La funzione ha una soluzione $\xi = \sqrt{2}$.

È immediatamente verificato che $f \in C^2$, inoltre scegliendo $(a, b, x_0) = (1, 2, 2)$ le ipotesi del teorema ???sono verificate, come pure $\forall x \in [a, b] : f \neq 0$

Sono sufficienti 5 iterazioni per arrivare ad una precisione di 16 cifre decimali, contro 50 del metodo di bisezione:

x_0	x_1	x_2	x_3	x_4	x_5
2	1.5	1.41...	1.41421...	1.41421356237...	1.41421356237095... $= fl^{16}(\sqrt{2})$

3.3 Altri metodi

Oltre al metodo della bisezione e di Newton esistono altri metodi più o meno efficienti per la ricerca degli zeri. Di seguito si riportano il metodo delle **corde** e delle **secanti**

3.3.1 Metodo delle corde

Il metodo delle corde è un metodo iterativo più efficiente per calcolare le radici di un'equazione non lineare reale e continua in un intervallo chiuso e limitato $[a, b]$ che assuma valori di segno opposto agli estremi dell'intervallo.

Il metodo consiste nel costruire una *successione* di punti dove, assegnato un punto iniziale x_0 $\forall n \leq 0$ il punto x_{n+1} sia lo zero della retta passante per il punto iniziale $(x_n, f(x_n))$ e di coefficiente angolare:

$$m = \frac{f(a) - f(x_n)}{a - x_n}$$

ovvero quello della retta passante per $(x_n, f(x_n)), (a, f(a))$. Iterando il procedimento del calcolo dell'intersezione delle varie rette con l'asse delle ascisse, si ottiene la relazione di ricorrenza

$$x_{n+1} = x_n - f(x_n)\frac{a - x_n}{f(a) - f(x_n)}$$

Teorema 3.17 (Convergenza del metodo delle corde.)**.** *Il metodo delle corde converge linearmente se, detta α la soluzione corretta vale*

$$0 < \frac{f'(\alpha)}{m} < 2$$

In altri termini α e $f'(\alpha)$ devono avere lo stesso segno e l'intervallo $[a, b]$ deve soddisfare la condizione

$$b - a < 2\frac{f(b) - f(a)}{f'(\alpha)}$$

Negli altri casi il metodo potrebbe non convergere affatto.

3.3.2 Metodo delle secanti

Come i metodi descritti precedentemente, il metodo delle secanti è un metodo per il calcolo delle radici di un'equazione. A differenza del metodo delle corde si applica ad un intervallo $[a, b]$ contenente una *sola* radice.

Il metodo consiste nel costruire una successione di punti con il seguente criterio: assegnati due punti iniziali x_0, x_1, $\forall n \geq 1$ il punto x_{n+1} sia lo zero della retta passante nei punti $(x_{n-1}, f(x_{n-1})), (x_n, f(x_n))$. Si ottiene

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)$$

Rispetto al metodo delle corde, quello delle secanti richiede un punto iniziale in più e ad ogni passo, il calcolo del rapporto che compare nella formula. Inoltre la convergenza è locale, cioè dipende dalla scelta dei punti iniziali x_0, x_1. Il guadagno è però una maggiore velocità di convergenza, che risulta superlineare.

4 Approssimazione di funzioni e di dati

Può accadere di necessitare una conveniente approssimazione di una generica funzione f t.c.:

$$f \ : \ [a,b] \subset \mathbb{R} \to \mathbb{R}$$

Ciò può rendersi necessario per diversi motivi:

- la forma funzionale di f potrebbe essere troppo complessa;
- la forma funzionale di f potrebbe essere non nota, e in questo casi si ipotizza di essere a conoscenza dei valori assunti su un insieme di ascisse tra loro distinte: $a \leq x_0 < x_1 < x_2 < x_3 < \cdots < x_n \leq b$
- operazioni funzionali (integrazioni, derivazioni) da dati discreti

4.1 Successioni di funzioni

Definizione 4.1. Dato un insieme di funzioni F tra due insiemi fissati X e Y, una successione di funzioni è una applicazione da $\mathbb{N}$ a F.

$$\{f_n\}_{n\in\mathbb{N}}$$

Definizione 4.2 (Convergenza puntuale)**.** Sia $\{f_n\}$ una successione di funzioni da X a Y e sia $f : X \to Y$. La successione $\{f_n\}$ converge puntualmente a f se, per x fissato:

$$\lim_{n\to\infty} f_n(x) = f(x)$$

Definizione 4.3 (Convergenza uniforme)**.** Sia $\{f_n\}$ una successione di funzioni da X a $\mathbb{R}$ e sia $f : X \to \mathbb{R}$. La successione $\{f_n\}$ converge uniformemente a f se:

$$\forall\, \varepsilon > 0,\ \exists N \in \mathbb{N} : |f_n(x) - f(x)| < \varepsilon \quad \forall x \in X \tag{2}$$

La convergenza uniforme è indipendente dal valore di x e pertanto un concetto più forte di quella puntuale.

Lemma 4.1. La convergenza uniforme implica quella puntuale, ma non è vero il contrario.

Dimostrazione 4.1. Si consideri come controesempio la successione di funzioni $\{f_n\}$, dove le $f_n(x)$ sono definite nell'intervallo [0, 1] e sono descritte dal grafico: Ogni funzione $f_n(x)$ è nulla per $x \geq 2/n$. La successione converge puntualmente a $f(x) = 0$, poichè, fissato un certo x, si riesce sempre a trovare un numero N tale che:

$$\frac{2}{n} \leq x, \quad \forall\, n \geq N \implies f_n(x) = 0$$

Non vi è, però, convergenza uniforme. Qualsiasi sia n, infatti, $|f_n(\frac{1}{n}) - 0| = 1$, quindi basta prendere $\varepsilon \in [0,\ 1)$ per mostrare che la (2) non è verificata.

Definizione 4.4 (Convergenza in media)**.**

$$dist(f_n, f) = \int_a^b |f_n(x) - f(x)|\ dx$$

Definizione 4.5 (Convergenza in media quadratica)**.**

$$dist(f_n, f) = \sqrt{\int_a^b f_n(x) - f(x)^2 \, dx} \leq \int_b^a \max_{x\in[a,b]} |f_n(x) - f(x)| \, dx$$

$$= \max_{x\in[a.b]} |f_n(x) - f(x)|(b-a)$$

4.2 Interpolazione polinomiale

Si immagini di disporre di:

$$(x_i, y_i), \quad i = 0, \ldots, n \quad y = f(x_i)$$

Si vuole ricercare un cosiddetto *polinomio interpolatore* $\phi_n \in \Pi_n$ interpolante la $f(x)$ sulle ascisse $x_i \in [a, b]$, $x_i < x_j$ se $i < j$, $x_i \neq x_j$ se $i \neq j$, tale che:

$$\phi(x) \in \Pi_n \quad \phi(x_i) = f(x_i) \quad i = 0 \ldots n$$

Teorema 4.1 (Esistenza di un unico polinomio interpolatore)**.** *Per ogni insieme di coppie* (x_i, y_i), $i = 0, 1, \ldots, n$, *con nodi* x_i *distinti fra loro,* $\exists!$ $\phi_n \in \Pi_n$ *tale che*

$$\phi_n(x_i) = y_i, \quad i = 0, \ldots, n$$

⋆ **Dimostrazione 4.1** (Unicità del polinomio interpolatore)**.** Si supponga per assurdo che che esistano due polinomi p, q di grado n, entrambi interpolatori per i punti (x_i, y_i), $i = 0, 1, \ldots, n$. Allora,

$$p(x_i) = q(x_i), \quad i = 0, 1, \ldots, n$$

Per il teorema fondamentale dell'algebra p e q hanno esattamente n zeri complessi e al più n zeri reali.

Si consideri il polinomio r differenza tra i due:

$$r(x) = p(x) - q(x)$$

$r(x)$ è di grado n, in quanto differenza di due polinomi di grado n, e ha $n + 1$ zeri, poichè

$$r(x_i) = p(x_i) - q(x_i) = y_i - y_i = 0, \quad i = 0, \ldots, n$$

Ma per il teorema fondamentale $r(x)$ non può esistere, dunque:

$$r(x) = p(x) - q(x) = 0 \quad \Longrightarrow \quad p(x) = q(x)$$

□

⋆ **Dimostrazione 4.2** (Esistenza del polinomio interpolatore)**.** Si dimostra l'esistenza di tale polinomio mostrando che è sempre possibile costruirne uno.

Il polinomio interpolatore è nella forma:

$$\phi_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

Le condizioni di interpolazione sono rappresentabili da un sistema di $n+1$ equazioni (una per ogni campione) e $n+1$ incognite $a_0, \dots a_n$, che costituiscono i coefficienti del polinomio.

$$\begin{cases} \phi_n(x_0) = a_0 + a_1 x_0 + \dots a_n x_0^n = y_0 \\ \phi_n(x_1) = a_0 + a_1 x_1 + \dots a_n x_1^n = y_1 \\ \vdots \\ \phi_n(x_n) = a_0 + a_1 x_n + \dots a_n x_n^n = y_n \end{cases}$$

Il sistema è nella forma $Ax = b$, con A matrice di Vandermonde:

Definizione 4.6 (Matrice di Vandermonde)**.** Una matrice di Vandermonde è una matrice in cui gli elementi delle righe costituiscono una progressione geometrica, ovvero:

$$V_{i,j} = \alpha_i^{j-1} \qquad V = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \dots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \dots & \alpha_2^{n-1} \\ 1 & \alpha_3 & \alpha_3^2 & \dots & \alpha_3^{n-1} \\ \vdots & & & & \\ 1 & \alpha_m & \alpha_m^2 & \dots & \alpha_m^{n-1} \end{pmatrix}$$

Il determinante di A può essere espresso come:

$$\det A = \prod_{0 \le i < j \le n} (x_j - x_i)$$

Se $\det A = 0$, la matrice è singolare e ha infinite soluzioni. Per ipotesi, però, i nodi x_i sono distinti tra loro, quindi $\det A \neq 0$, la matrice è invertibile e il sistema ammette soluzione unica $x = A^{-1}b$.

□

Dimostrazione 4.2. Un'ulteriore dimostrazione di esistenza consiste nel costruire il polinomio nella cosiddetta *forma di Lagrange*. Siano $\varphi_i(x) \in \Pi_n$ polinomi definiti come segue:

$$\varphi_i(x) = \prod_{j=0,\, j \neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

Tali particolari polinomi sono definiti *polinomi caratteristici di Lagrange*. Si noti che il loro comportamento è analogo al *delta di Kronecker* δ_{ij}:

$$\varphi_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

Dai polinomi $\varphi_i(x)$ si ottiene $\phi_n(x)$ per sovrapposizione degli effetti:

$$\phi_n(x) = \sum_{i=0}^{n} y_i \, \varphi_i(x)$$

Tale polinomio soddisfa le condizioni di interpolarizzazione ed è pertanto un polinomio interpolatore, espresso in *forma di Lagrange*.

Definizione 4.7 (Errore di interpolazione)**.** Si chiama errore di interpolazione $E_n(x)$ la distanza tra una funzione f e il polinomio interpolatore π_n al passo n:

$$E_n(x) = dist(f, \Pi_n) = \max\{\, |\pi_n(x) - f(x)|, \quad \forall x \in [a, b]\,\}$$

Teorema 4.2. *Sia I un intervallo limitato e siano $\{x_i : i = 0, \dots, n\}$ nodi di interpolazione distinti in I. Sia $f \in \mathcal{C}^{n+1}[a, b]$ in I. Allora $\forall x \in I : (\exists\ \xi \in I)$ t.c.*

$$E_n = f(x) - \Pi_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x - x_0)(x - x_1)\dots(x - x_n) \tag{3}$$

Tale espressione rappresenta l'errore di interpolazione di ordine n nella forma di Lagrange.

Dimostrazione 4.3. Si vuole dimostrare la (3) per ogni $x \in [a, b]$. Se $x = x_i$, $i = 0, \dots, n$, la tesi è triviale.

$$f(x_i) - \Pi_n(x_i) = 0 \Leftrightarrow f(x_i) = \Pi_n(x_i) \quad \text{per definizione di } \Pi_n$$

$$\frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x_i - x_0)\dots\underbrace{(x_i - x_i)}_{0}\dots(x_i - x_n) = 0$$

Sia $x \notin \{x_0, \dots, x_n\}$ fissato, e sia $z \in [a, b]$. Definiamo $G(z)$ come:

$$G(z) = E_n(z) - \omega(z)\,\frac{E_n(x)}{\omega(x)}$$

dove

$$\omega(z) = \prod_{i=0}^{n}(x - x_i)$$

Si ricordi allora il teorema di Rolle:

Teorema 4.3 (Teorema di Rolle)**.** *Sia f continua in $[a, b]$ e derivabile in (a, b) tale che $f(a) = f(b)$. Allora, $\exists\, c \in (a, b)$ tale che*

$$f'(c) = 0$$

I nodi x_i e x suddividono l'intervallo $[a, b]$ in $n + 1$ intervalli, dove G(z) si annulla trivialmente. Allora, per il teorema di Rolle, in ogni intervallo i-esimo $\exists\, z_{1,i} : G'(z_{1,i}) = 0$. Analogamente, i punti $z_{1,i}$ suddividono l'intervallo int $(z_{1,0}, z_{1,1}, \dots, z_{1,n})$ in n intervalli dove $G''(z_{2,i}) = 0$. Reiterando Rolle, si ottiene:

$$\exists\, \xi : G^{(n+1)}(\xi) = 0$$

Per la linearità dell'operatore di derivazione, si ha:

$$\begin{aligned} G^{(n+1)}(z) &= \mathcal{D}^{n+1}\left(f(z) - \Pi_n(z)\right) - \frac{E_n(x)}{\omega(x)}\mathcal{D}^{n+1}\left(\omega(z)\right) \\ &= f^{(n+1)}(z) - \Pi_n^{(n+1)}(z) - \frac{E_n(x)}{\omega(x)}\,\omega^{(n+1)}(z) \end{aligned}$$

$\Pi_n(z)$ ha grado n, pertanto $\Pi_n^{(n+1)}(z) = 0$. $\omega(z)$ ha termine di grado massimo z^{n+1}, quindi $\omega^{(n+1)}(z) = (n+1)!$. Quindi, si ha

$$\underbrace{G^{(n+1)}(\xi)}_{0} = f^{(n+1)}(\xi) - \frac{E_n(x)}{\omega(x)}\,(n+1)!$$

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

□

Lemma 4.2 (Maggiorazione dell'errore)**.**

$$\begin{aligned} E_n = \left|f(x) - \Pi_n(x)\right| &= \left|f^{(n+1)}(\xi)\right| \frac{\prod_{i=0}^{n}|x - x_i|}{(n+1)!} \\ &\leq \max_{x\in[a,b]}\left|f^{(n+1)}(x)\right| \cdot \frac{(b-a)^{n+1}}{(n+1)!} \\ &= M_{n+1}\frac{(b-a)^{n+1}}{(n+1)!} \end{aligned}$$

Dove M_{n+1} è la massima derivata $n+1$-esima.

Lemma 4.3 (Convergenza)**.** Poichè

$$\lim_{n\to\infty} \frac{(b-a)^{n+1}}{(n+1)!} = 0$$

Se $\exists M : M_n \leq M$

$$\lim_{n\to\infty} E_n(x) \leq \lim_{n\to\infty} M\frac{(b-a)^{n+1}}{(n+1)!} = 0$$

Esempio 4.1. Si consideri $f(x) = e^x$. Si ha che

$$M_{n+1} = \max_{x\in[a,b]}\left|f^{(n+1)}(x)\right| \leq e^b$$

M_{n+1} è limitata dalla costante e^b, pertanto $E_n(x) \to 0$. Tale stima vale per ogni distribuzione dei nodi.

Lemma 4.4. Come si può dimostrare, in caso di distribuzione esclusivamente equidistante, l'errore è inoltre maggiorabile come segue:

$$|E_n(x)| \leq M_{n+1}\,\frac{h^{n+1}}{4\,(n+1)} \qquad h = \frac{b-a}{n}$$

Lemma 4.5. In generale, con l'interpolazione polinomiale, non c'è convergenza uniforme (neppure convergenza puntuale).

Dimostrazione 4.4 (Esempio di Runge)**.** Il cosiddetto esempio di Runge vale come controesempio di 4.5.

Sia $f(x) = \frac{1}{1+x^2}$ e $f \in C^\infty$

Eseguiamo un campionamento a passo costante, ovvero prendiamo dei punti equispaziati tra loro. Notiamo che più aumentiamo di grado il nostro polinomio interpolatore più le oscillazioni agli estremi aumentano. Questo problema è dovuto alla scelta di campionamento a passo costante. Sicuramente non c'è convergenza uniforme e, vicino agli estremi non c'è neppure quella puntuale.

4.3 Tecniche di interpolazione polinomiale

Esistono due grandi famiglie di tecniche di interpolazione:

- Interpolazione con un unico polinomio ma scegliendo punti speciali. **NON** devono essere punti equispaziati.
- Interpolazione polinomiale a tratti: si fissa il grado del polinomio e si uniscono polinomi del grado fissato sui vari pezzetti.

Valgono:

Lemma 4.6. Sia $f \in C^{n+1}$ per qualsiasi distribuzione di nodi

$$\max_{x\in[a,b]} |E_n(x)| \leq M_{n+1} \frac{(b-a)^{n+1}}{(n+1)!}$$

Lemma 4.7. Per nodi equispaziati

$$\max_{x\in[a,b]} |E_n(x)| \leq M_{n+1} \frac{h^{n+1}}{4(n+1)} \qquad h = \frac{b-a}{n}$$

4.4 Nodi di Chebychev

Definizione 4.8 (Nodi di Chebychev)**.** Siano n angoli equispaziati compresi tra 0 e π:

$$\Theta_i = i\frac{\pi}{n} \qquad 0 \leq i \leq n$$

Si chiamano nodi di Chebychev l'insieme delle loro proiezioni sulla circonferenza:

$$t_i = -\cos(\Theta_i) = -\cos\left(\frac{i\pi}{n}\right) \qquad 0 \leq i \leq n$$

Su un intervallo $[a, b]$:

$$x_i^{\text{cheb}} = \underbrace{\frac{b+a}{2}}_{\text{centro}} + \underbrace{\frac{b-a}{2}}_{\text{raggio}} t_i^{\text{cheb}} = \frac{b+a}{2} - \frac{b-a}{2}\cos\left(\frac{i\pi}{n}\right)$$

Osservazione 4.1. È apparente che in prossimità degli estremi i nodi si addensano.

Definizione 4.9 (Polinomio interpolatore di Chebychev)**.** È possibile scrivere il polinomio interpolatore in forma di Lagrange nel seguente modo, ponendo $l_i^{cheb} = \varphi_i$:

$$\Pi_n^{\text{cheb}}(x) = \sum_{i=0}^{n} f(x_i^c) l_i^c(x)$$

Teorema 4.4 (Esistenza e unicità del polinomio interpolatore di Chebychev)**.** *Come banale conseguenza dell'essere in forma di Lagrange, il polinomio interpolatore esiste ed è unico.*

Teorema 4.5 (Convergenza dell'interpolazione con nodi di Chebychev)**.** *Se $f \in C^K[a, b], K \geq 1$ l'interpolazione con i nodi di Chebychev converge.*

Lemma 4.8. La convergenza con i nodi di Chebychev è uniforme.

Lemma 4.9. Più volte la f è derivabile, più sarà veloce la sua convergenza.

Dimostrazione 4.5. Se $f \in C^K[a,b], K \geq 1$:

$$\left|E_n(x)\right| = \left|dist(f, \Pi_n^{\text{cheb}})\right| \leq M_k \frac{\log n}{n^k}$$

Poichè M_k è costante:

$$\lim_{n\to\infty} \left|E_n(x)\right| = \frac{\log n}{n^k} = 0$$

□

4.5 Stabilità dell'interpolazione polinomiale

Posto $y_i = f(x_i)$, spesso si avranno a disposizione dei valori approssimati $\widetilde{y_i}$ invece di y_i, dovuto ad esempio all'errore commesso dal calcolatore nel valutare $f(x_i)$.

Si supponga di poter stimare un valore massimo ε di tale errore sui dati:

$$\max|\widetilde{y_i} - y_i| \leq \varepsilon$$

Si indicherà con $\widetilde{\Pi}_n$ il polinomio interpolatore calcolato con $\widetilde{y_i}$.
Allora:

$$\begin{aligned}\max_{x\in[a,\,b]} \left|\,\Pi_n - \widetilde{\Pi}_n\right| &= \left|\sum_{i=0}^{n} y_i\, l_i(x) - \sum_{i=0}^{n} \widetilde{y_i}\, l_i(x)\right| \\ &= \left|\sum_{i=0}^{n} (y_i - \widetilde{y_i})\; l_i(x)\right|\end{aligned}$$

Sapendo che $|\widetilde{y_i} - y_i| \leq \varepsilon$ e applicando la disuguaglianza triangolare, si ha:

$$\begin{aligned}\left|\sum_{i=0}^{n} (y_i - \widetilde{y_i})\; l_i(x)\right| &\overset{D.T.}{\leq} \sum_{i=0}^{n} |y_i - \widetilde{y_i}|\;\left|l_i(x)\right| \\ &\leq \varepsilon \underbrace{\max_{x\in[a,\,b]} \sum_{i=0}^{n} \left|l_i(x)\right|}_{\Lambda_n} \qquad (4)\end{aligned}$$

Il valore Λ_n è chiamato *costante di Lebesgue*. Allora,

$$\max_{x\in[a,\,b]} \left|\,\Pi_n - \widetilde{\Pi}_n\right| \leq \varepsilon\, \Lambda_n$$

Lemma 4.10. Λ_n dipende unicamente dal prodotto di LaGrange, e quindi dai nodi.

Inoltre, si ha che:

Lemma 4.11. Su nodi equidistanti, $\Lambda_n \approx \frac{2^n}{n \log n} \approx 2^n$

Lemma 4.12. Su nodi di Chebyshev, $\Lambda_n \approx \log n$.

Di conseguenza:

Lemma 4.13. Con distribuzione equidistante, l'interpolazione polinomiale reagisce in modo instabile alle perturbazioni sul campionamento della funzione da interpolare

Lemma 4.14. L'interpolazione polinomiale risulta sostanzialmente stabile con la distribuzione di Chebychev.

4.6 Interpolazione polinomiale a tratti

L'idea di base dell'interpolazione polinomiale a tratti consiste nel prendere un Π interpolante costituito da un polinomio per ciascuna coppia di punti consecutivi.

Dato un intervallo $[a, b]$, si procede dunque a suddividerlo in una serie di n intervallini $[x_i, x_{i+1}]$ tali che:

$$x_i = a + i \cdot \qquad h = \frac{b-a}{n} \qquad 0 \leq i \leq n$$

Il caso più semplice è quello con grado 1, ovvero l'interpolazione lineare composita.

In tal caso $\Pi_1^C(x)$ è lineare in $[x_i, x_{i+1}]$ e la funzione ottenuta è una spezzata lineare a tratti nonchè continua.

In questo caso $\Pi_2^C(x)$ consiste in pezzi di parabola "incollati" insieme.

Il Π ottenuto è una funzione quadratica a tratti.

Per tre punti non allineati passa una ed una sola parabola, dunque bisogna organizzare i punti a "pacchetti" di 3 - ovvero, organizzare gli intervallini a "pacchetti" di 2.

Occorre dunque scegliere $n = 2k$.

Più in generale vale:

Lemma 4.15. Con $\Pi_S^C(x)$ di grado S il numero di punti n deve essere multiplo di S, avendo:

$$x_i = \{(x_0, x_1, \ldots, x_s), (x_{S_1}, x_{S_1+1}, \ldots, x_{2s}), (x_{n-s}, x_{n-s+1}, \ldots, x_n)\}$$

4.6.1 Convergenza dell'interpolazione polinomiale a tratti

Teorema 4.6 (Convergenza uniforme dell'interpolazione polinomiale a tratti)**.** *Sia $f \in C^{s+1}[a, b]$, con n multiplo di S, allora:*

$$dist\,(f, \Pi_s^c) \leq C_{s+1} \cdot h^{s+1}$$

dove in C_{s+1} compare $f^{(s+1)}$.

Lemma 4.16 (Convergenza uniforme dell'interpolazione lineare a tratti)**.** Sia $f \in \mathcal{C}^2[a, b]$. Allora,

$$\text{dist}\left(f, \Pi_1^C\right) \leq c \cdot h^2, \qquad h = \max_{0 \leq i < n} \Delta x_i$$

⋆ **Dimostrazione 4.3** (Per grado 1)**.** Sia $I_i = [x_i,\, x_{i+1}]$. Allora,

$$\begin{aligned}\operatorname{dist}(f,\, \Pi_1^C) &= \max_{x\in[a,\,b]} \left| f(x) - \Pi_1^C(x) \right| \\ &= \max_{0\le i<n} \max_{x\in I_i} \left| f(x) - \Pi_{1,\,i}^C(x) \right| \end{aligned} \tag{5}$$

Si rammenti che l'errore in forma di Lagrange (vedi 3) è:

$$E_n = f(x) - \Pi_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x-x_0)(x-x_1)\dots(x-x_n)$$

Allora, con $n = 1$:

$$\begin{aligned} f(x) - \Pi_{1,\,i}^C(x) &= \frac{f^{(2)}(\xi_i)}{2!}\,(x-x_i)\,(x-x_{i+1}) \\ \max_{x\in I_i} \left| f(x) - \Pi_{1,\,i}^C(x) \right| &\le \frac{1}{2} \max_{x\in I_i} f''(x)\ (\Delta x_i)^2 \end{aligned}$$

Sostituendo nella (5), si ha:

$$\begin{aligned} \max_{0\le i<n} \frac{1}{2} \max_{x\in I_i} f''(x)\ (\Delta x_i)^2 &\le \frac{1}{2} \left(\max_{x\in(a,\,b)} f''(x) \right) h^2 \\ &\le c\cdot h^2 = O(h^2) \end{aligned}$$

□

4.6.2 Stabilità dell'interpolazione polinomiale a tratti

$$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|$$

A grado basso non ci sono problemi (e infatti di solito *si usa* un grado basso).
Con punti equispaziati

$$\Lambda_n \approx C \frac{2^n}{n \log n}$$

Questo è un problema.

4.7 Interpolazione Spline

Utilizzando l'interpolazione polinomiale a tratti si ha che, nei raccordi, in generale, non è nemmeno possibile derivare Π_S^C. Si ha una moltitudine di punti angolosi e ciò è specialmente indesiderabile nella grafica.

Per ovviare a questo problema si usa l'interpolazione spline[4].

Si cercano funzioni interpolanti *lisce*, senza punti angolosi, ossia:

Definizione 4.10 (Interpolante Spline)**.** Si chiama interpolante spline una funzione $S_k(x) \in C^{k-1}[a,b]$ tale che:

1. $S_k(x)\big|_{I_i} \in P_k$

[4] *Spline* è la riga flessibile usata da artigiani e disegnatori industriali prima dell'avvento del CAD

2. $S_k(x_i) = y_i, \qquad 0 \leq i \leq n$

3. $S_k^{(j)}(x_i^+) = S_k^{(j)}(x_i^-) \qquad 0 \leq i \leq n-1 \quad 0 \leq j \leq k-1$

Distinguiamo le funzioni spline in base al loro grado k; con $k = 1$ avremo spline lineari, mentre con $k = 3$ avremo spline cubiche.

Le spline cubiche sono il tipo più usato nella pratica:

Definizione 4.11 (Interpolante spline cubica)**.** Si chiama interpolante spline una funzione $S_k(x) \in C^{k-1}[a,b]$ con $k = 3$, dunque:

1. $S_3|_I \in P_3$
2. $S_3(x_i) = y_i \qquad 0 \leq i \leq n$
3. $S_3 \in C^2[a,b]$

Osservazione 4.2. Si osservi che $S_3|_I \in P_3$ equivale a dire che $S_3|_I = a_{0i} + a_{1i}x + a_{2i}x^2 + a_{3i}x^3$ con $0 \leq i \leq n-1$: si hanno ossia n intervallini, ciascuno dei quali "vede" quattro incognite.

Osservazione 4.3. $S_3(x_i) = y_i$ implica trivialmente che agli estremi deve valere $S_3(x_0) = y_0, S_3(x_n) = y_n$.

Osservazione 4.4. Inoltre, poichè $S_3 \in C^2[a,b]$

$$\begin{cases} S_3(x_i^+) = S_3(x_i^-), & 1 \leq i \leq n-1 \text{(cioè nei nodi interni o di raccordo)} \\ S_3'(x_i^+) = S_3'(x_i^-) \\ S_3''(x_1^+) = S_3''(x^-) \end{cases}$$

Osservazione 4.5. Il sistema che risulta dei vincoli è sottodimensionato e dunque singolare.

Mettendo a sistema i vincoli si vede che si hanno $4n$ incognite (4 incognite $a_{0..3}$ per ogni polinomio $S_{3,i}$ e n polinomi).

Si hanno inoltre $4n-2$ equazioni: $n+1$ equazioni per le condizioni di interpolazione, $n-1$ equazioni per la continuità di $S_3(x)$ su $n-1$ nodi interni e $2(n-1)$ equazioni per la continuità della derivata prima e seconda sui nodi interni.

È possibile ottenere un sistema non singolare aggiungendo arbitrariamente, ad esempio:

$$S_3''(x_0) = 0 \qquad S_3''(x_n) = 0$$

La soluzione del sistema risultante permette di determinare la funzione spline cubica interpolatoria.

4.8 Approssimazione Polinomiale dei Minimi Quadrati

L'interpolazione polinomiale di LaGrange non migliora l'interpolazione al crescere del grado. Tale inconveniente può essere superato con l'interpolazione a tratti e spline, che però poco si prestano ad estrapolare previsioni sui dati, ovvero valori di cui non si ha un campionamento.

Supponiamo di disporre di un insieme di N dati campione $\{(x_i, y_i), i = 0, \dots, n\}$. Dato un grado $m \geq 1$ (tipicamente $\ll N$), si vuole trovare un polinomio $p(x) \in P_m$ tale che:

$$\sum_{i=0}^{n} \left[p(x) - y_i\right]^2 \leq \sum_{i=0}^{n} \left[q(x) - y_i\right]^2, \qquad \forall q \in P_m$$

Trovare $p(x)$ equivale a trovare un vettore di coefficenti $a \in \mathbb{R}^{m+1}$ tale che

$$a = \min_{v \in \mathbb{R}^{m+1}} \sum_{i=0}^{n} \left[(v_0 + v_1 x_i + v_2 x_i^2 + \cdots + v_m x_i^m) - y_i\right]^2$$

Quando $m = 1$, si parla di *retta dei minimi quadrati* o *retta di regressione*.

Sia $V \in \mathbb{R}^{n \times (m+1)}$ matrice rettangolare tale che $v_{i,\,j} = x_i^j$, e sia $a \in \mathbb{R}^{m+1}$ il vettore colonna dei coefficenti:

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^m \\ 1 & x_1 & x_1^2 & \dots & x_1^m \\ & & \vdots & & \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix}_{\in \mathbb{R}^{n \times (m+1)}} \qquad a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}_{\in \mathbb{R}^{m+1}} \qquad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}_{\in \mathbb{R}^N}$$

Come si può dimostrare, tale problema ha un'unica soluzione, calcolabile risolvendo il sistema lineare $V^\top V a = V^\top y$.

Caso lineare: retta dei minimi quadrati Con $m = 1$, l'insieme di punti viene approssimato da una retta. Il sistema $V^\top V a = V^\top y$ è quindi formato dalle seguenti componenti:

$$V = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}, \qquad V^\top = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$

$$V^\top V = \begin{bmatrix} N & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix}$$

Il sistema risulta quindi:

$$\begin{bmatrix} N & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Nel problema del minimo quadrato, si vuole trovare il vettore dei coefficenti $a \in \mathbb{R}^{m+1}$ per cui $\phi(a) = \sum_{i=0}^{n} (Va - y_i)^2$ è minimo.

Per definizione di prodotto scalare:

$$\phi(a) = \langle Va - y,\, Va - y \rangle$$

Se a è minimo, allora

$$\phi(a+b) \geq \phi(a) \quad \forall\, b \in \mathbb{R}^{m+1} \tag{6}$$

$$\begin{aligned}
\phi(a+b) &= \langle V(a+b) - y,\, V(a+b) - y \rangle \\
&= \langle Va + Vb - y,\, Va + Vb - y \rangle \\
&= \langle Va - y,\, Va - y \rangle + \langle Va - y,\, Vb \rangle + \langle Vb,\, Va - y \rangle + \langle Vb,\, Vb \rangle \\
&= \phi(a) + 2\langle Va - y,\, Vb \rangle + \langle Vb,\, Vb \rangle
\end{aligned}$$

Quindi, la (6) diventa:

$$2\langle Va - y,\, Vb \rangle + \langle Vb,\, Vb \rangle \geq 0 \quad \forall\, b \in \mathbb{R}^{m+1}$$

Fissato $v \in \mathbb{R}^N$ e posto $b = \varepsilon\, v$, si ha:

$$\begin{aligned}
2\,\langle Va - y,\, V(\varepsilon\, v) \rangle + \langle V(\varepsilon\, v),\, V(\varepsilon\, v)\,\rangle \geq 0, &\qquad \forall \varepsilon > 0 \\
2\varepsilon\,\langle Va - y,\, Vv \rangle + \varepsilon^2 \langle Vv,\, Vv \rangle \geq 0 & \\
2\,\langle Va - y,\, Vv \rangle + \varepsilon\,\langle Vv,\, Vv \rangle \geq 0, &\qquad \forall\, v \in \mathbb{R}^N, \quad \forall\, \varepsilon > 0
\end{aligned}$$

Passando al limite per $\varepsilon \to 0$, si ha:

$$\langle Va - y,\, Vv \rangle \geq 0, \qquad \forall\, v \in \mathbb{R}^N$$

Si noti che, con $-v$ al posto di v, si ha $\langle Va - y,\, Vv \rangle \leq 0$, pertanto:

$$\begin{cases} \langle Va - y,\, Vv \rangle \geq 0 \\ \langle Va - y,\, Vv \rangle \leq 0 \end{cases} \quad \Longrightarrow \quad \langle Va - y,\, Vv \rangle = 0$$

5 Integrazione Numerica

L'integrazione numerica, nota anche come *quadratura* numerica, consta nel calcolo approssimato di un'integrale - per mezzo dell'integrazione di un'approssimazione della funzione integranda.

Si rammentino i teoremi fondamentali del calcolo:

Teorema 5.1 (Primo teorema fondamentale del calcolo integrale)**.** *Sia* $f\colon [a,b] \to \mathbb{R}$ *una funzione integrabile.*

Si definisce "funzione integrale" di f *la funzione* F *tale che:*

$$F(x) = \int_a^x f(t)dt \qquad a \leq x \leq b$$

Se f *è limitata, allora* F *è una funzione continua in* $[a,b]$*.*

Se inoltre f *è una funzione continua in* (a,b)*, allora* F *è differenziabile in tutti i punti in cui* f *è continua e si ha:*

$$F'(x) = f(x)$$

cioè la F *risulta essere una primitiva di* f*.*

Teorema 5.2 (Secondo teorema fondamentale del calcolo integrale)**.** *Sia* $f\colon [a,b] \to \mathbb{R}$ *una funzione che ammette una primitiva* G *su* $[a,b]$*.*

Sia cioè $G(x)$ *tale che:*

$$G'(x) = f(x)$$

Se f *è integrabile si ha:*

$$\int_a^b f(x)dx = G(b) - G(a)$$

Teorema 5.3. *Non tutte le funzioni hanno come primitive delle funzioni elementari.*

Esempio 5.1. Ad esempio, la funzione $f(x) = e^{-x^2}$ non ha una funzione elementare come primitiva.

Teorema 5.4. *L'operazione di integrazione numerica è stabile.*

Dimostrazione 5.1. Si supponga che $dist(f, \tilde{f}) \leq \varepsilon$.

Si ricordi che:

$$dist(f, \tilde{f}) = \max_{x \in [a,b]} \Big| f(x) - \tilde{f}(x) \Big|$$

Allora

$$\begin{aligned}\left|I(f) - I(\tilde{f})\right| &= \left|\int_a^b f(x)\ dx - \int_b^a \tilde{f}(x)\ dx\right| \\ &= \left|\int_a^b f(x) - \tilde{f}(x)\ dx\right| \\ &\leq \int_a^b f(x) - \tilde{f}(x)\ dx \\ &\leq \int_a^b \varepsilon\ dx = \varepsilon \int_a^b 1\ dx = \varepsilon(b-a)\end{aligned}$$

□

Lemma 5.1.

$$\left|I(f) - I(\phi_n)\right| \leq (b-a)dist(f, \phi_n)$$

Lemma 5.2. Se c'è convergenza uniforme, allora

$$\lim_{n\to\infty} \int_a^b \phi_n = \int_a^b f$$

5.1 Formule di quadratura

Esistono due grandi famiglie di formule di quadratura: algebriche e composte. Ci aspettiamo che le formule composte convergano sempre, ma non è garantita la convergenza delle formule algebriche.

5.1.1 Formule di quadratura algebriche

Le formule di quadratura algebriche utilizzano, per l'approssimazione un polinomio interpolatore.

$$I(f) \approx I(\phi_n) \qquad \phi_n \in \{x_i, y_i\},\ \ 0 \leq i \leq n \text{ e } y_i = f(x_i)$$

Sia $\phi_n = \Pi_n$ allora

$$\begin{aligned}I_n(f) &= \int_a^b \Pi_n(x)\ dx = \int_a^b \sum_{i=0}^n y_i l_i(x)\ dx = \sum_{i=0}^n \int_a^b y_i l_i(x)\ dx \\ &= \sum_{i=0}^n y_i \underbrace{\int_a^b l_i(x)}_{w_i}\ dx = \sum_{i=0}^n y_i w_i \quad \text{con } w_i = \text{ pesi}\end{aligned}$$

L'integrale definito del prodotto di Lagrange è detto *peso*, e il risultato è chiamato *somma pesata*.

Le formule di quadratura si dividono nelle seguenti famiglie:

- Formule di Newton-Cotes: utilizzano nodi equispaziati. Non sono in generale convergenti.
- Formula di Clenshaw-Curtis: utilizzano i nodi di Chebychev. Convergono per $f \in C^k$ con $k \geq 1$.

– Formule Gaussiane: richiedono una distribuzione speciale dei nodi.

Le formule algebriche sono, per costruzione, esatte (l'errore è nullo) sui polinomi di grado $\leq n$ (la funzione interpolatoria è il polinomio stesso).

Le formule gaussiane sono esatte fino al grado $2n+1$ ma chiedono una scelta "specialissima" dei nodi di campionamento.

Nelle formule di Clenshaw-Curtis e in quelle Gaussiane i pesi sono positivi ($w_i \geq 0$).

5.1.2 Formule di quadratura composte

Le formule di quadratura composte utilizzano, per l'approssimazione, un'insieme di polinomi.

$$\phi_n = \Pi_s^c \qquad n \text{ multiplo di } s$$

Anche le formule composte sono espresse nella forma di somme pesate.

$$I_n(f) = I(\Pi_s^c) = \int_b^a \Pi_s^c(x)dx = \sum_{i=0}^{n} y_i w_i$$

Nel caso di interpolazione a tratti, tipicamente lineare o quadratica, si hanno:

– Se $s = 1$ sono dette formule dei trapezi;

– Se $s = 2$ sono dette formule delle parabole.

5.1.3 Caso lineare (Formule dei trapezi)

Nelle formule dei trapezi ($s = 1$) ogni polinomio interpolatore Π_s^c è una retta e l'integrale definito equivale all'area del trapezio di vertici $x_i,\ x_{i+1},\ y_i,\ y_{i=1}$.

$$\int_b^a \Pi_n^c(x)dx = \sum_{i=0}^{n-1}(y_i + y_{i+1})\frac{\Delta x_i}{2}$$

Dove l'area del trapezio i-esimo è:

$$(y_i + y_{i+1})\frac{\Delta x_i}{2} \qquad \Delta x_i \equiv h$$

L'integrale approssimato risulta:

$$\begin{aligned} I(\Pi_1^c) &= (y_0 + y_1)\frac{h}{2} + (y_1 + y_2)\frac{h}{2} + \cdots + (y_{n-1} + y_n)\frac{h}{2} \\ &= y_0 + \frac{h}{2} + ny_1 + ny_2 + \cdots + \frac{h}{2}y_n \end{aligned}$$

Nella forma di somme pesate si ha:

$$w_i = \begin{cases} \frac{h}{2} & \text{per } i = 0 \\ n & \text{per } 1 \leq i \leq n-1 \end{cases}$$

5.1.4 Caso quadratico (Formule delle parabole)

Dalla regola di Cavalieri-Simpson (o formula delle parabole) con nodi equidistanti, si ha:

$$w_i = \begin{cases} h/3 & i = 0,\ i = n \\ 4h/3 & i \text{ dispari} \\ 2h/3 & i \text{ pari} \end{cases}$$

5.2 Convergenza dell'integrazione numerica

Teorema 5.5. *Le formule di Newton-Cotes non sono in generale convergenti*

Si desidera:

$$\lim_{n\to\infty} I_n(f) = I(f)$$

È possibile allora vedere che:

- $s = 1$ (formula dei trapezi)

$$\begin{aligned} & f \in C^2[a,b],\ dist(f, \Pi_1^C) = O(h^2) \\ \Rightarrow & \left| I(f) - I(\Pi_1^C) \right| = O(h^2) \end{aligned} \qquad \Delta x_i = h$$

- $s = 2$ (formula delle parabole)

$$\begin{aligned} & f \in C^3[a,b],\ dist(f, \Pi_2^C) = O(h^3) \\ \Rightarrow & \left| I(f) - I(\Pi_2^C) \right| = O(h^3) \end{aligned} \qquad \Delta x_i = h$$

$$f \in C^4 \Rightarrow \left| I(f) - I(\Pi_2^C) \right| = O(h^4)$$

5.3 Integrazione numerica con dati perturbati

La stabilità dell'integrazione numerica è data da:

$$\left| \int_b^a f(x)dx - \int_b^a \tilde{f}(x)dx \right| \leq dist(f, \tilde{f})(b-a)$$

Se la funzione approssimatrice è polinomiale a tratti Π_1^c si ha

$$\left| \int_b^a f(x)dx - \int_b^a \Pi_1^c(x)dx \right| \leq dist(f, \Pi_1^c)(b-a)$$

Spesso si hanno dati perturbati; la funzione da integrare f non è nota analiticamente, ma si conosce una sua approssimazione $\tilde{f}$

$$I_n(f) = \sum_{i=0}^{n} w_i y_i$$

$$I_n(\tilde{f}) = \sum_{i=0}^{n} w_i \tilde{y}_i$$

Analogamente, al posto dei valori y_i si conoscono dei valori approssimati $\tilde{y}_i$ tali che

$$|\tilde{y}_i - y_i| \leq \varepsilon$$

Si ha che

$$\begin{aligned} I(f) - I_n(\tilde{f}) &= I(f) - I_n(\tilde{f}) + I_n(f) - I_n(f) \\ &= \underbrace{I(f) - I_n(f)}_{\text{convergenza}} + \underbrace{I_n(f) - I_n(\tilde{f})}_{\text{stabilità}} \end{aligned}$$

Si deve allora verificare la stabilità dell'operazione che dipende dall'espressione $I_n(f) - I_n(\tilde{f})$

$$\begin{aligned} \left|I_n(f) - I_n(\tilde{f})\right| &= \left|\sum w_i y_i - \sum w_i \tilde{y}_i\right| \\ &= \left|\sum w_i (y_i - \tilde{y}_i)\right| \\ &\overset{D.T}{\leq} \sum |w_i| \underbrace{|y_i - \tilde{y}_i|}_{\leq \varepsilon} \\ &\leq \varepsilon \underbrace{\sum_{i=0}^{n} |w_i|}_{\alpha_n} \end{aligned}$$

Nel caso di funzione a tratti

$$dist(\Pi_n, \tilde{\Pi}_n) \leq \varepsilon \Lambda_n \qquad \text{con } \Lambda_n = \max_{x \in [a,b]} \sum_{i=0}^{n} \left|l_i(x)\right|$$

Si può allora dimostrare:

Teorema 5.6. *Le formule a pesi positivi sono stabili*

Dimostrazione 5.2. Avendo $\alpha_n = \sum_{i=0}^{n} |w_i|$ positiva

$$\sum_{i=0}^{n} |w_i| = \sum_{i=0}^{n} w_i = \sum_{i=0}^{n} w_i \cdot 1 = I_n(1)$$

e dunque

$$\begin{aligned} \left|I_n(f) - I_n(\tilde{f})\right| &= \left|\sum w_i y_i - \sum w_i y_i\right| \\ &= \left|\sum w_i (y_i - \tilde{y}_i)\right| \\ &\overset{D.T}{\leq} \sum |w_i| \underbrace{|y_i - \tilde{y}_i|}_{\leq \varepsilon} \\ &\leq \varepsilon \sum_{i=0}^{n} |w_i| = \varepsilon \cdot C \end{aligned}$$

□

Osservazione 5.1. Le formule di quadratura algebriche sono esatte su tutti i polinomi di grado $\leq n$ (l'interpolatoria è il polinomio stesso); Le formule di quadratura composte sono esatte su tutti i polinomi di grado $\leq s$ (lineari per grado ≤ 1, quadratiche per grado ≤ 2). In generale le formule di integrazione su $y = 1$ (grado $= 0$) sono esatte, pertanto

$$\sum_{i=0}^{n} w_i = \int_b^a 1dx = b - a$$

α_n non solo è limitata, ma è addirittura costante $(b - a)$. L'errore può quindi aumentare al massimo della lunghezza dell'intervallo.

5.4 Derivazione numerica

La derivazione numerica si occupa di derivare funzioni a partire da approssimazioni di esse.

Teorema 5.7. *La derivazione numerica non è in generale stabile.*

Vi sono due approcci principali alla derivazione numerica:

1. Globale: si cerca di approssimare f' su tutto l'intervallo $[a, b]$ tramite interpolazione spline.
2. Locale: calcolo f' tramite rapporto incrementale su un punto.

Per definizione di derivata, si ha:

$$\delta_+(h) = \frac{f(x+h) - f(x)}{h}$$

$$f'(x) = \lim_{h \to 0} \delta_+(h)$$

Sia $f \in \mathcal{C}^2$ in un intorno di x. Applicando la formula di Taylor si ha:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(z), \qquad z \in (x, x+h)$$

Il rapporto incrementale $\delta_+(h)$ si può esprimere come:

$$\delta_+(h) = f'(x) + \underbrace{\frac{h}{2} f''(z)}_{O(h)}$$

Sia f approssimata da $\tilde{f}$, con un errore maggiorabile con un ε fissato:

$$\left| f(x) - \tilde{f}(x) \right| \leq \varepsilon$$

Si ha di conseguenza un rapporto incrementale approssimato $\tilde{\delta}_+$:

$$\tilde{\delta}_+(h) = \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h}$$

Si può analizzare la stabilità sviluppando l'espressione dell'errore:

$$f'(x) - \tilde{\delta}_+(h) = \underbrace{f'(x) - \delta_+(h)}_{\text{convergenza}} + \underbrace{\delta_+(h) - \tilde{\delta}_+(h)}_{\text{stabilità}}$$

$$\begin{aligned}
\left|\delta_+(h) - \tilde{\delta}_+(h)\right| &= \left|\frac{f(x+h) - f(x)}{h} - \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h}\right| \\
&= \left|\frac{f(x+h) - \tilde{f}(x+h) + \tilde{f}(x) - f(x)}{h}\right| \\
&\overset{D.T.}{\leq} \left|\frac{f(x+h) - \tilde{f}(x+h)}{h}\right| + \left|\frac{\tilde{f}(x) - f(x)}{h}\right| \\
&\leq \frac{\varepsilon}{h} + \frac{\varepsilon}{h} = \frac{2\,\varepsilon}{h}
\end{aligned}$$

Quindi:

$$\left|f'(x) - \tilde{\delta}_+(h)\right| \leq \left|f'(x) - \delta_+(h)\right| + \left|\delta_+(h) - \tilde{\delta}_+(h)\right| \leq \frac{2\,\varepsilon}{h} + O(h)$$

Sia $g(h) \overset{def.}{=} \frac{2\,\varepsilon}{h} + O(h)$. Si noti che per h molto piccolo, l'errore diventa molto grande.

Si deve trovare $h^* = \min_h g(h) = \min_h ch + \frac{2\,\varepsilon}{h}$:

$$\begin{aligned}
g'(h) = c - \frac{2\,\varepsilon}{h^2} = 0 &\implies \frac{h^2}{2\,\varepsilon} = \frac{1}{c} \\
&\implies h^2 = \frac{2\,\varepsilon}{c} \\
&\implies h = h^* = \sqrt{\frac{2\,\varepsilon}{c}} = O(\sqrt{3})
\end{aligned}$$

Sia quindi $\min_h g(h) = g(h^*)$.

$$\begin{aligned}
g(h^*) &= ch^* + \frac{2\,\varepsilon}{h^*} \\
&= 2\sqrt{2c}\sqrt{\varepsilon} \\
&= O\left(\sqrt{\varepsilon}\right)
\end{aligned}$$

Funzioni arbitrariamente vicine possono avere derivata molto diversa.

Si può ottenere una stima migliore considerando un ordine in più nella formula di Taylor:

$$\begin{aligned}
f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f^{(3)}(z) \\
f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{3!}f^{(3)}(z)
\end{aligned} \qquad z \in (x-h,\, x+h)$$

$$f(x+h) - f(x-h) = 2hf'(x) + \frac{h^3}{3}f^{(3)}(z)$$

Calcolando il rapporto incrementale simmetrico $\delta(h)$, risulta:

$$\delta(h) = \frac{f(x+h) - f(x-h)}{2h} = \frac{1}{2h}\, 2hf'(x) + \frac{1}{2h}\,\frac{h^3}{3} f^{(3)}(z)$$

$$= f'(x) + O(h^2)$$

Inoltre, si vede che:

$$\delta(h) - f'(x) = O(h^2)$$

$$\left|\delta(h) - \tilde{\delta}(h)\right| \leq \frac{2\,\varepsilon}{2h} = \frac{\varepsilon}{h}$$

L'errore risulta:

$$\left|f'(x) - \tilde{\delta}(h)\right| \leq \left|f'(x) - \delta(h)\right| + \left|\delta(h) - \tilde{\delta}(h)\right|$$

$$\leq O(h^2) + \frac{\varepsilon}{h} = g(h)$$

$$h^* = O\left(\sqrt[3]{\varepsilon}\right), \qquad g(h^*) = O\left(\sqrt[3]{\varepsilon^2}\right)$$

6 Algebra Lineare Numerica

6.1 Cenni di Algebra Lineare

Definizione 6.1 (Sistema lineare)**.** Un **sistema lineare** è un sistema nella forma

$$Ax = b$$

Dove

$$A \in \mathbb{R}^{m \times n}, \quad x \in \mathbb{R}^n, \quad b \in \mathbb{R}^m$$

Definizione 6.2 (Sistema quadrato)**.** Un **sistema quadrato** è un **sistema lineare** con

$$m = n$$

Definizione 6.3 (Sistema sovradeterminato)**.** Un **sistema sovradeterminato** è un sistema con

$$m > n$$

Un sistema sovradeterminato possiede più equazioni che incognite - ovvero, ci sono "troppi vincoli". Per questo in generale non esiste una soluzione che soddisfi il sistema. Si può però cercare una soluzione approssimata, che minimizzi la distanza rispetto ai vincoli:

Definizione 6.4 (Soluzione nel senso dei minimi quadrati)**.** Una soluzione nel senso dei minimi quadrati di un sistema $Ax = b$ è una $\tilde{x}$ tale che

$$\tilde{x} = \min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} (b_i - \sum_{j=1}^{n} a_{ij} x_j)^2$$

Definizione 6.5 (Sistema sottodeterminato)**.** Se $m < n$, il sistema è sottodeterminato. Ha in genere infinite soluzioni.

Esistono una serie di approcci alla soluzione dei sistemi lineari, in generale riconducibili a due famiglie:

- Metodi diretti, ovvero metodi in cui la soluzione viene calcolata dopo un numero finito di passi (ad es. il Metodo di Eliminazione di Gauss)
- Metodi iterativi, in cui la soluzione si ottiene in un numero teoricamente infinito di passi. In essi si cerca una successione $\{x_k\}, x_k \in \mathbb{R}^n$ tale che $\lim_{k \to \infty} x_k = x$, con x soluzione del sistema, e dunque. Si tratta di metodi perlopiù costruiti *ad hoc* rispetto a un determinato problema che costituiscono attivo campo di ricerca.

6.1.1 Norme

Definizione 6.6 (Norma vettoriale)**.** Una norma su uno spazio vettoriale reale R^n è una funzione

$$\| \cdot \| : \quad \mathbb{R}^n \quad \longrightarrow \quad [0, +\infty)$$

tale che $\forall x, y \in \mathbb{R}^n$:

1. $\|x+y\| \leq \|x\| + \|y\|$ (D.T.)
2. $\|x\| = 0 \Leftrightarrow x = 0$
3. $\|\alpha x\| = |\alpha| \, \|x\|$

Definizione 6.7 (Distanza indotta da una norma)**.**

$$dist\|\cdot\| = dist(x,y) = \|x-y\|$$

Per definizione la distanza è simmetrica

Definizione 6.8 (Norma uno)**.**

$$\|x\|_1 = \sum_{i=1}^{m} |x_i|$$

Con $m = 2$ trivialmente $\|x\|_1 = |x_1| + |x_2|$

Definizione 6.9 (Norma due o Euclidea)**.**

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Definizione 6.10 (2-distanza)**.**

$$dist_2(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$$

Definizione 6.11 (Norma infinito)**.**

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Definizione 6.12 (∞-distanza)**.**

$$dist_\infty(x,y) = \max\{||x_1-y_1|, |x_2-y_2|\} = \max_{1 \leq i \leq n} \{|x_i - y_i|\}$$

Si consideri un intorno $\mathscr{I}_\varepsilon$ di x con raggio ε rispetto a una distanza:

$$\mathscr{I}_\varepsilon(x) = \{y \in \mathbb{R}^n : dist(y,x) \leq \varepsilon\}$$

Visivamente, possiamo immaginare:

- Un intorno in norma due come cerchi di raggio ε
- Un intorno in norma infinito come quadratini di lato 2ε
- Un intorno in norma uno come rombi di lato 2ε

Definizione 6.13 (Norme equivalenti)**.** Siano due norme

$$\|x\|_a \|x\|_b$$

Esse si dicono equivalenti se:

$$\exists c_1, c_2 > 0 : c_2\|x\|_b \leq \|x\|_a \leq c_1\|x\|_b$$

Teorema 6.1. *In uno spazio vettoriale finito-dimensionale tutte le norme sono equivalenti*

Teorema 6.2.

$$\lim_{k\to\infty} x^{(k)} = x \Leftrightarrow dist(x^{(k)}, x) \to_{k\to\infty} 0$$

ossia

$$\|x^{(k)} - x\| = \|x - x^{(k)}\|$$

Definizione 6.14 (Limite di una successione)**.** Sia una successione

$$(x^{(k)})$$

allora

$$\forall \varepsilon > 0 : \exists N(\varepsilon) : dist(x^{(k)}, x) \leq \varepsilon \qquad \forall n \geq N(\varepsilon)$$

x è limite della successione.

Teorema 6.3 (Proprietà fondamentale delle successioni)**.** *Una successione è convergente se e solo se ogni sua sottosuccessione è convergente.*

$$\lim_{k\to\infty} x^{(k)} = x \Leftrightarrow \lim_{k\to\infty} x_i^{(k)} = x_i \qquad 1 \leq i \leq n$$

Ciò implica la convergenza elemento per elemento.

Lemma 6.1. Siano

$$c_1 = \frac{1}{\sqrt{n}},\ c_2 = 1$$

Allora

$$\frac{\|x\|_2}{\sqrt{n}} \leq \|x\|_\infty \leq \|x\|_2$$

Lemma 6.2.

$$\begin{aligned}\|x\|_2 &= \sqrt{x_1^2 + x_2^2 + \ldots + x_\mu^2} \\ &\leq \sqrt{\max(|x_i|)^2 + \ldots + \max(|x_i|)^2} \\ &= \sqrt{n}\max\|x_i\|\end{aligned}$$

6.1.2 Norma di matrici

Definizione 6.15 (Applicazione lineare)**.** Sia $A \in \mathbb{R}^{m\times n}$. Allora è applicazione lineare L_A:

$$\begin{aligned}L_A : \mathbb{R}^n &\to \mathbb{R}^m \\ x &\mapsto y = Ax\end{aligned}$$

t.c.:

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay$$

Definizione 6.16 (Norma di matrice)**.** Sia $A \in \mathbb{R}^{n\times m}$, fissata una norma vettoriale, si definisce la norma $\|A\|$ come:

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

Da tale definizione si ricava una proprietà fondamentale delle norme matriciali:

Teorema 6.4 (Proprietà fondamentale)**.** *Sia $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$:*

$$y = \|Ax\| \leq \|A\| \cdot \|x\|$$

Dimostrazione 6.1. Triviale: significa che, per $x \neq 0$,

$$\frac{\|Ax\|}{\|x\|} \leq \|A\|$$

In pratica:

$$\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

$$|(Ax)_i| = |y_i| = |\sum_{j=1}^{n} a_{ij}x_j| \leq \sum_{j=1}^{n} |a_{ij}||x_j| \leq \|x\|_\infty \sum_{j=1}^{n} |a_{ij}|$$

Quindi

$$\|Ax\|_\infty = \max_{1\leq i\leq n} |(Ax)_i| \geq \frac{\|Ax\|_\infty}{\|x\|_\infty}$$

□

Definizione 6.17. La distanza tra due matrici è la norma della differenza:

$$dist(A, B) = \|A - B\|$$

Esempio 6.1. Con norma infinito la norma matriciale è:

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

Esempio 6.2. Con norma euclidea la norma matriciale è:

$$\|A\|_2 = \sqrt{\rho(A^t A)}$$

dove ρ è il raggio spettrale.

$$\rho(B) = \max\{|\lambda| : \lambda \textit{ autovalore di } B\}$$

$B = A^t A$ simmetrico (autovalore reale) e ha autovalori ≥ 0.

6.2 Soluzione approssimata di sistemi di equazioni

Consideriamo il caso più comune, in cui interessa la soluzione *approssimata* $\tilde{b}$ di un sistema quadrato:

$$A\tilde{x} = \tilde{b} \qquad \tilde{b} \approx b$$
$$\tilde{\tilde{A}}\tilde{x} = \tilde{b}$$

6.2.1 Risoluzione di sistemi con errori nel termine noto

Si rammenti per prima cosa:

Teorema 6.5. $\det(A) = 0$ *se e solo se* A *è singolare*

Si consideri:

$$Ax = b \qquad A \in R^{n \times m}, \det(A) \neq 0, x \in R^n, b \in R^n$$

Si supponga ora che il termine noto sia tale solo in una forma approssimata $\tilde{b}$. Ci si troverebbe dunque in realtà a risolvere:

$$A\tilde{x} = \tilde{b}$$

È possibile scrivere l'errore sul risultato come:

$$\delta x = \tilde{x} - x \qquad \delta b = \tilde{b} - b$$

$$A(x + \delta x) = b + \delta b$$

Interessa naturalmente sapere in che modo l'errore sul termine noto δb influenzi l'errore sul risultato δx.

Definizione 6.18 (Indice di condizionamento)**.**

$$k(A) = \|A\|\|A^{-1}\|$$

Lemma 6.3. $k(A)$ dipende solo dalla matrice A una volta fissata la norma vettoriale.

Lemma 6.4. $k(A) \geq 1$ (=1 quando $A = \mathscr{I}$)

Può succedere che $k(A)$ sia *molto grande*. In questo caso si parla di **sistemi malcondizionati**, ovvero di instabilità intrinseca del problema: la risoluzione di siffatti sistemi interessa fino a un certo punto.

Teorema 6.6. *Siano* $e_r(b) = \frac{\|\delta b\|}{\|b\|}$ *e* $e_r(x) = \frac{\|\delta x\|}{\|x\|}$ *errori relativi fissata una certa norma* $\|\cdot\|$.
Allora:

$$e_r(x) \leq k(A)\, e_r(b)$$

⋆ **Dimostrazione 6.1.** Per un sistema

$$Ax = b$$

vale:

$$\delta x = A^{-1}\delta b$$

Infatti:

$$\begin{aligned} A(x + \delta x) &= b + \delta b \\ Ax + A\delta x &= b + \delta b \\ AA^{-1}b + A\delta x &= b + \delta b \\ \delta x &= A^{-1}\delta b \end{aligned}$$

Dunque fissata una norma $\|\cdot\|$ è possibile scrivere:

$$\begin{aligned} \delta x = A^{-1}\delta b \Rightarrow \|\delta x\| &= \|A^{-1}\delta b\| \\ \|\delta x\| &\leq \|A^{-1}\| \cdot \|\delta b\| \end{aligned}$$

Ora, poichè

$$\|x\| \geq k > 0 \Rightarrow \frac{1}{\|x\|} \leq \frac{1}{k}$$

Dunque:

$$\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta b\| \Rightarrow \frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\delta b\|}{\|x\|}$$

Poichè inoltre $Ax = b \Rightarrow \|Ax\| = \|b\|$:

$$\begin{aligned} \underbrace{\|Ax\|}_{\leq \|A\| \cdot \|x\|} &= \|b\| \\ \|x\| &\geq \frac{\|b\|}{\|A\|} \\ \frac{1}{\|x\|} &\leq \frac{\|A\|}{\|b\|} \end{aligned}$$

Sostituendo:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\delta b\|}{\|x\|} \leq \frac{\|A\|\|A^{-1}\|\|\delta b\|}{\|b\|} = k(A)\frac{\|\delta b\|}{\|b\|}$$

□

Esempio 6.3. Si consideri il sistema

$$\begin{cases} 7x_1 + 10x_2 = 1 \\ 5x_1 + 7x_2 = 0.7 \end{cases}$$

con:

$$A = \begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 0.7 \end{bmatrix}$$

Sia $\tilde{b}$ il vettore dei dati perturbati:

$$\tilde{b} = \begin{bmatrix} 1 \\ 0.069 \end{bmatrix}, \qquad \delta b = \begin{bmatrix} 0.01 \\ 0.01 \end{bmatrix}$$

Scelta come norma la norma infinito si ottiene:

$$\|b\|_\infty = \max\{|1|, |0.7|\} = 1$$

$$\|\delta b\|_\infty = \max\{10^{-2}, 10^{-2}\} = 10^{-2}$$

L'errore sul termine noto è il seguente:

$$\frac{\|\delta b\|_\infty}{\|b\|_\infty} = \frac{10^{-2}}{1} = 10^{-2} = 1\%$$

Si calcola allora la soluzione:

$$x = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} \qquad \tilde{x} = \begin{bmatrix} -0.17 \\ 0.22 \end{bmatrix} \qquad \delta x = \begin{bmatrix} -0.17 \\ 0.12 \end{bmatrix}$$

Le norme sono le seguenti:

$$\|x\|_\infty = 0.1$$

$$\|\delta x\|_\infty = \max\{0.17, 0.12\} = 0.17$$

Si ottiene un'errore sulla soluzione pari a:

$$\frac{\|\delta x\|}{\|x\|} = \frac{0.17}{0.1} = 1.7 = 170\%$$

$$\|A\|_\infty = \{17, 12\} = 17$$

Calcolando A^{-1} si ha:

$$A^{-1} = \begin{bmatrix} -7 & 10 \\ 5 & -7 \end{bmatrix}$$

$$\|A^{-1}\| = \max\{17, 12\} = 17$$

L'indice di condizionamento è:

$$k(A) = 17 \cdot 17 = 17^2 = 289$$

La soluzione è quindi inaccettabile.

6.2.2 Cenni su risoluzione sistemi con errori sulla matrice

Lemma 6.5. A è una matrice simmetrica definita positiva se e solo se i suoi autovalori sono tutti positivi.

Si suppone di avere:

$$\tilde{A}\tilde{x} = \tilde{b}$$

con errore:

$$(A + \delta A)(x + \delta x) = b + \delta b$$

Teorema 6.7. *Avendo* $k(A) \cdot \frac{\|\delta A\|}{\|A\|} < 1$, *scegliendo una norma vettoriale e la sua norma matriciale indotta si ottiene:*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{k(A)}{1 - K(A) \cdot \frac{\|\delta A\|}{\|A\|}} \cdot \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

Lemma 6.6.

$$\|A\|_2 = \frac{\lambda \max}{\lambda \min}$$

6.3 Metodo di eliminazione di Gauss

Sia A matrice quadrata non singolare. Nella soluzione con MEG di un sistema $Ax = b$, la matrice A viene trasformata in una matrice triangolare superiore U e il sistema viene trasformato nell'equivalente $Ux = \beta$ (soluzione identica):

$$[A|b] \rightarrow [U|\beta]$$

L'algoritmo realizza una decomposizione LU, la quale fattorizza A in un prodotto di due matrici L e U rispettivamente triangolare inferiore e superiore.

$$A = LU$$

Quindi il sistema $Ax = b$ diventa

$$\begin{aligned} LUx &= b \\ L^{-1}LUx &= \underbrace{L^{-1}b}_{\beta} \\ Ux &= \beta \end{aligned}$$

Il sistema triangolare $Ux = \beta$ è facilmente risolvibile con sostituzioni all'indietro, di complessità $\approx n^2$.

$$x_i = \frac{\beta_i - \sum_{j=i+1}^{n} u_{ij} x_j}{u_{ii}}$$

Questo approccio è estremamente utile quando si ha necessità di trovare la soluzione per diversi sistemi con matrice A fissa in cui varia solo il termine noto $b_{1,2,\ldots,n}$: è sufficiente calcolare LU una volta sola a costo $O(n^3)$ e poi risolvere ciascun sistema a costo $O(n^2)$.

$$LUx = b \Leftrightarrow \begin{cases} Ly = b \\ Ux = y \end{cases}$$

6.3.1 Pivoting e stabilizzazione

Il pivoting - ossia lo scambio preliminare tra righe - è importante per la stabilizzazione dell'algoritmo:

Si consideri:

$$Ax = b \Leftrightarrow PAx = Pb$$

Poichè $PA = LU$:

$$\begin{aligned} Ax = b &\Leftrightarrow PAx = Pb \\ &= LUx = Pb \\ &\Leftrightarrow Ux = \underbrace{PL^{-1}b}_{\beta} \end{aligned}$$

È allora possibile risolvere il sistema trasformandolo in due sistemi triangolari nel seguente modo:

$$\begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

6.3.2 MEG e sistemi malcondizionati

Un esempio canonico di sistema malcondizionato è la matrice di Hilbert.

Definizione 6.19 (Matrice di Hilbert)**.**

$$H = (h_{ij})$$

$$h_{ij} = \frac{1}{i+j-1} \qquad 1 \leq i, j \leq n$$

Si consideri la soluzione del sistema $Hx = b$:

$$H \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \hat{b}$$

Per $n = 13$, $e_n = 10$ - ovvero, abbiamo un errore del1000%

Il condizionamento in $\|\cdot\|_2$, $k_2(H)$ cresce esponenzialmente e il risultato del MEG è inservibile, come è possibile osservare con Matlab.

6.3.3 Soluzione di sistemi con Matrice Triangolare

Come visto in precedenza, il MEG opera una trasformazione $PA = LU$.

Posto A non singolare:

$$\begin{aligned} Ax = b &\Leftrightarrow PAx = Pb \\ &\Leftrightarrow LUx = Pb \\ &\Leftrightarrow \begin{cases} Ly = Pb \\ Ux = y \end{cases} \end{aligned}$$

Trasforma ossia un sistema $Ax = b$, denotato in breve come $[A|b]$, in $[U|\beta]$.
Il sistema $Ux = \beta$ si risolve, ovviamente, con una serie di sostituzioni all'indietro:

$$\underbrace{x_i = \beta_i - \sum_{j=i+1}^{n} u_{ij}x_j}_{u_{ii}} \qquad i = n, n-1, \ldots, 1$$

Il costo computazionale è di $\approx n^2$ FLOPs, e il costo asintotico è:

$$C = \not{2}\frac{n(n-1)}{\not{2}} = O(n^2)$$

Dunque il costo per risolvere

$$\begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

è

$$C = 2n^2$$

6.3.4 Applicazioni del MEG

Tra le applicazioni del MEG si ricorda:

- Calcolo $\det(A)$ (A quadrata)
- Calcolo del rango di A (A rettangolare)
- Soluzione di un sistema $Ax = b$
- Calcolo dell'inversa di una matrice quadrata, non singolare.

6.3.5 Calcolo di A^{-1} con fattorizzazione LU

Si considerino k sistemi con matrice A e termine noto $b^{(i)}$:

$$Ax^{(i)} = b^{(i)} \qquad 1 \leq i \leq k$$

Per risolverli il modo ingenuo può essere, con costo $O(n^3) \cdot k$:

$$[A|b^{(i)}] \overset{MEG}{\rightarrow} [U|\beta] \overset{\text{sost. indietro}}{\rightarrow} x^{(i)}$$

Non è però il metodo più intelligente per procedere.
Ricordiamo:

$$A \overset{MEG}{\rightarrow} U \Rightarrow PA = LU$$

Allora possiamo riscrivere i sistemi come coppie di sistemi triangolari:

$$\begin{cases} Ly^{(i)} = Pb^{(i)} \\ Ux^{(i)} = y^{(i)} \end{cases} \qquad 1 \leq i \leq k$$

Con k significativi il costo è vantaggioso:

$$C = \underbrace{O(n^3)}_{\text{Calcola una sola volta L, U}} + \underbrace{k \cdot O(n^2)}_{k \text{ risoluzioni di sistemi con matrici triangolari}}$$

Consideriamo una base B:

$$B \in \mathbb{R}^{m \times n} \alpha \in \mathbb{R}^n$$

È notorialmente possibile rappresentare qualunque vettore v dato come:

$$v = Bc = \alpha_1 C_1 + \alpha_2 C_2 + \ldots + \alpha_n C_n \qquad c_i \in \mathbb{R}^n, c = col_i(B)$$

Definiamo $e^{(i)}$ come elemento i-esimo della base canonica:

$$e^{(i)} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Dunque $Be^{(i)} = col_i(B)$.
Osserviamo che data una matrice A invertibile vale allora:

$$A^{-1} e^{(i)} = col_i(A^{-1}) \qquad \text{per } m = n$$

Moltiplicando a sinistra entrambi gli elementi per la matrice A:

$$\underbrace{AA^{-1}}_{I} e^{(i)} = A col_i(A^{-1}) \qquad 1 \leq i \leq n$$

Poichè:

$$A^{-1} = [col_1(A^{-1}), col_2(A^{-1}), \ldots, col_n(A^{-1})$$

Allora è possibile riformulare il problema come:

$$[A|e^{(i)}] \overset{MEG}{\rightarrow} [U|\beta^{(i)}]$$

$$\begin{cases} Ly^{(i)} = Pe^{(i)} \\ Ucol_i(A) = y^{(i)} \end{cases}$$

6.3.6 Cenni sulla soluzione di sistemi fortemente malcondizionati

Si rammenti per prima cosa:

Definizione 6.20 (Sistema mal condizionato)**.** Un sistema mal condizionato è un sistema $Ax = b$ t.c.

$$k(A) \gg 1$$

Si consideri:

$$PA = LU \qquad PA \approx \tilde{L}\tilde{U} \qquad \tilde{L}\tilde{U}x = Pb \qquad \|PA - \tilde{L}\tilde{U}\| \approx eps$$

La fattorizzazione $\tilde{L}\tilde{U}$ è molto buona poichè la distanza da PA è vicina alla precisione di macchina, ma se il sistema è mal condizionato gli effetti possono essere comunque deleteri sulla soluzione.

Si può allora osservare:

$$\frac{\|\delta b\|}{\|b\|} \leq \varepsilon \rightarrow \frac{\|\delta x\|}{\|x\|} \approx k(A)\varepsilon$$

Invece di risolvere $A\tilde{x} = \tilde{b}$, ottenendo una soluzione sbagliata, è possibile risolvere con una famiglia di sistemi tale che:

$$A_h x_h = b \qquad x_h \rightarrow x \qquad k(A_h) < k(A)$$

Con:

$$k(A_h) \rightarrow_{h \rightarrow 0} k(A)$$

Allora:

$$A_h \tilde{x}_h = \tilde{b}$$

Errore relativo Si consideri la distanza tra x e $\tilde{x}$:

$$\|x - \tilde{x}_h\| = \|x - x_h + x_h - \tilde{x}_h\| \leq \|x - x_h\| + \|x_h - \tilde{x}_h\|$$

Dividendo per $\|x\|$ si ottiene l'errore relativo:

$$\|x - \tilde{x}_h\| \leq \frac{\|x - x_h\|}{\|x\|} + k(A_h)\frac{\|\delta b\|}{\|b\|} \cdot \underbrace{\frac{\|x_h\|}{\|x\|}}_{\rightarrow 1}$$

$$= \underbrace{e(h)}_{\rightarrow 0} + k(A_h) \cdot \frac{\|\delta b\|}{\|b\|}$$

La scelta di h deve essere tale che h non sia troppo piccolo per non amplificare l'errore, ma neanche troppo grande: occorre un compromesso.

In generale, in ogni caso, usare un metodo classico per risolvere un sistema fortemente malcondizionato significa avere una **soluzione completamente inattendibile**.

6.3.7 Cenno ai sistemi sovradeterminati

Definizione 6.21 (Sistema sovradeterminato)**.** Un **sistema sovradeterminato** è un sistema in cui ci sono più equazioni che incognite:

$$Ax = b$$

$$A \in \mathbb{R}^{m \times n} \qquad x \in \mathbb{R}^n \quad b \in R^m \quad n > m$$

In generale il sistema non ha soluzione.

Quello che si può fare è circoscrivere un intorno della soluzione, minimizzando $\|Ax - b\|_2^2$ con il metodo dei minimi quadrati.

$$\|Ax - b\|_2^2 = \min dist(Ax, b) = \min \|\{\sum_{j=1}^{n} a_{ij}x_j\} - \{b_i\}$$

Un caso particolare è:

$$Va = y$$

$V = (v_{ij})$ è matrice di Vandermonde, $\lambda = a$ coefficiente del polinomio dei minimi quadrati, $b = y$ valori campionati.

Allora:

$$Va = y \Leftrightarrow V^tVa = V^ty \qquad \text{(equazioni normali)}$$

$$\min x \in \mathbb{R}^n \|Ax - b\|_2^2 \Leftrightarrow A^tAx = A^tb$$

Se A ha rango n (cioè se le colonne di A sono vettori n-dimensionali indipendenti) allora A^tA è simmetrica definita positiva e quindi nonè singolare.

Esiste allora, ed è unica, una soluzione x - A^tA che potrebbe però essere estremamente malcondizionata.

6.4 Cenni su Fattorizzazione QR

Una matrice $A \in \mathbb{R}^{m \times n}$ con $m \geq n$ si può fattorizzare in

$$A = QR$$

Con R matrice triangolare superiore $n \times n$ non singolare e Q matrice ortogonale $m \times n$.

Si rammenta che:

Definizione 6.22 (Matrice ortogonale)**.** Q ortogonale significa $Q^tQ = I$.

Si può altrimenti caratterizzare:

Lemma 6.7. Se Q è matrice ortogonale le colonne di Q sono vettori ortonormali:

$$col_i(Q^t)^t \cdot col_\delta(Q) = \delta_{ij} = \begin{cases} i = \delta : 0 \\ i = \delta : 1 \end{cases}$$

(La norma euclidea è 1).

Si assuma che $rango(A) = n$, ovvero il massimo possibile.

$$A = QR$$
$$AR^{-1} = QRR^{-1} = Q$$

Se R è triangolare superiore allora R^{-1} è anch'essa triangolare superiore.

$$[col_1(A)col_2(A)\dots col_n(A)] \cdot \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ 0 & & & \\ 0 & 0 & \dots & p_{n,n} \end{bmatrix}$$

$$= \begin{bmatrix} p_{11}\,col(A) & p_{11}\,col_1(A) + p_{22}\,col_2(A) & \cdots & P_{1n}\,col_1(A) + \ldots + p_{nm}\,col_n(A) \end{bmatrix}$$

Risultano vettori ortonormali da vettori indipendenti - il procedimento è equivalente ad applicare Gram-Schmidt sulle colonne di A).

$$A = QR \qquad rango(A) = n$$

$$A^tA = A^tb$$

$$A^tA = (QR)^t(QR) = R^t \underbrace{(Q^tQ)}_{\mathscr{I}} R = R^tR$$

$$A^tb = R^tQ^tb$$

$$R^tRx = Q^tQ^tb$$

R è una matrice non singolare, per cui il sistema di partenza diventa equivalente a:

$$Rx = Q^tb$$

Questo è il sistema dell'equazione normali scritto in un'altra forma.
La fattorizzazione QR è il secondo algoritmo più usato al mondo (dopo FFT).

Indice analitico