▷ Lezione I

Teoria dell'approssimazione e aritmetica finita

Libro Capitolo 1, sezione 1.2

Teoria dell'approssimazione

Uno degli aspetti principali del calcolo numerico è saper controllare e, dove possibile, quantificare gli errori che vengono commessi nel risolvere un problema fisico, o meglio, il corrispondente modello matematico, al calcolatore.

Chiamiamo PF il fenomeno o problema fisico che vogliamo risolvere. Assumiamo che sia possibile misurare la soluzione di tale problema fisico, e chiamiamo tale soluzione T_f. Solitamente, tramite le leggi della fisica è possibile associare a PF un modello o problema matematico PM, ad esempio tale modello può essere formato da un set di equazioni differenziali o un sistema algebrico. Necessariamente tale modello non potrà includere tutti i dettagli del processo fisico ma solamente quelli che riteniamo più importanti. La soluzione del problema matematico è data da $T = T(x, y, z, t)$, ottenuta, ad esempio, risolvendo un'equazione alle derivate parziali come

$$\frac{\partial T}{\partial t} - D\left(\frac{\partial T}{\partial x^2} + \frac{\partial T}{\partial y^2} + \frac{\partial T}{\partial z^2}\right) = f \tag{1.1}$$

dove D e f rappresentano dati fisici del problema. Se interpretiamo T come temperatura l'equazione (1.1) descrive la diffusione del calore secondo la legge di Fourier, dove D descrive la conducibilità e f rappresenta possibili sorgenti di calore. Lo stesso tipo di equazione può descrivere la diffusione di un inquinante di concentrazione T secondo la legge di Fick. Il calcolo di T è solitamente limitato ad una porzione finita dello spazio (in questo caso $\mathbb{R}^3$). Tale regione viene solitamente chiamata *dominio* in cui (1.1) viene risolta e spesso indicata con Ω.

Tuttavia, l'equazione (1.1) può essere risolta analiticamente solo per pochi casi, che solitamente hanno un impatto applicativo limitato a causa della loro semplicità. Negli altri casi dobbiamo quindi affidarci ad un problema numerico PN, approssimazione di $PN \approx PM$. Un esempio, che verrà approfondito nel seguito, è approssimare la derivata in tempo con

$$\left.\frac{\partial T}{\partial t}\right|_{t^n+1} \approx \frac{T(t^{n+1}) - T(t^n)}{t^{n+1} - t^n} \tag{1.2}$$

dove t^n è un determinato instante di tempo in cui vogliamo calcolare la soluzione numerica T_n, approssimazione di T. La formula a destra in (1.2) *approssima* il termine a sinistra, la derivata, e tale approssimazione è tanto più accurata quanto più t^{n+1} è vicino a t^n, o, in altre parole se $\Delta t = t^{n+1} - t^n$ è piccolo. Idealmente per $t^{n+1} \to t^n$ si ottiene l'uguaglianza.

Possiamo quindi sintetizzare il processo che parte dal problema fisico per arrivare alla sua risoluzione numerica come

$$T_f \xrightarrow{\text{errore di modello}} T \xrightarrow{\text{errore numerico}} T_n,$$

Supponendo note T_f, T e T_n è possibile quantificare due tipi di errore:

- l'errore *di modello*, $e_m = \|T_f - T\|$ che misura la differenza fra il fenomeno fisico e il suo modello matematico. Inevitabilmente, infatti, il modello matematico trascura alcuni aspetti della realtà e introduce idealizzazioni. *"Essentially, all models are wrong, but some are useful"* è una famosa citazione dello statistico George Box. Più in dettaglio, spiega che

  Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

  For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?"

- l'errore *numerico*, $e_n = \|T - T_n\|$ che misura la differenza fra la soluzione esatta del modello matematico, e la soluzione del problema numerico.

Nelle definizioni precedenti $\|\cdot\|$ indica la norma degli oggetti considerati, concetto che verrà approfondito in seguito. L'errore numerico e_n si può dividere a sua volta in due contributi distinti, il primo associato allo schema stesso e detto errore *di troncamento*, il secondo dovuto all'implementazione dello schema in un programma software. Quest'ultimo viene detto errore di *arrotondamento* e sarà approfondito in seguito.

Consistenza e convergenza di un metodo numerico

Come accennato, l'approssimazione numerica della derivata prima rispetto al tempo "migliora" quando il passo temporale Δt tende a zero, ossia tale approssimazione *converge* per $\Delta t \to 0$. Proviamo a definire il concetto di convergenza in astratto, per poi poter precisare tale definizione per i singoli metodi numerici che affronteremo nel corso.

Consideriamo un problema in cui x è l'incognita, d sono i dati, e F è la relazione che li lega. Il problema è quindi: trovare x tale che

$$F(x, d) = 0. \tag{1.3}$$

Quando risolviamo tale problema con un metodo numerico consideriamo la sua forma approssimata, ossia

$$F_n(x_n, d_n) = 0, \tag{1.4}$$

dove il parametro n può indicare, ad esempio, il numero di intervalli nella quadratura numerica, il numero di time step nella soluzione di un'equazione differenziale, o altro. In generale vogliamo che $x_n \to x$ per $n \to \infty$: perché questo accada è necessario che $d_n \to d$ e inoltre che F_n sia *consistente* con F. Per il momento trascuriamo per semplicità la questione dell'approssimazione dei dati e concentriamoci su F_n.

Definition. *Un metodo numerico è detto* consistente *se la soluzione esatta x è anche soluzione del problema numerico quando $n \to \infty$. Più precisamente, se e solo se*

$$F_n(x, d) \to 0 \text{ per } n \to \infty.$$

Inoltre, un metodo è detto fortemente consistente *se $F_n(x, d) = 0$ per qualsiasi n.*

Ora possiamo definire più precisamente cosa si intende per *convergenza*.

Definition. *Un metodo numerico definito come* (1.4) *è* convergente *se e solo se*

$\forall \epsilon > 0 \; \exists n_0 = n_0(\epsilon)$ *tale che*

$\forall n > n_0(\epsilon) \Rightarrow \|x - x_n\| \leq \epsilon,$

dove x è la soluzione esatta e x_n è la soluzione numerica.

Significa che possiamo sempre trovare un valore di n_0 e una corrispondente soglia per il valore di δd_n (la perturbazione dei dati) tale per cui, per ogni $n > n_0$, la soluzione numerica è vicina a piacere alla soluzione esatta.

Questa definizione è piuttosto generale e si può applicare a diversi problemi. Nel caso delle equazioni differenziali ordinarie possiamo ridurre Δt, o in altre parole aumentare il numero di passi temporali fino a $n > n_0$ per ottenere l'accuratezza desiderata, mentre nel caso di metodi iterativi possiamo costruire una sequenza di soluzioni approssimate x_n che tende a x quando n aumenta.

Aritmetica floating point

L'errore di arrotondamento citato nella precedente sezione è dovuto al fatto che un computer non può, per sua natura, rappresentare ogni numero $x \in \mathbb{R}$. Infatti, $\pi \in \mathbb{R}$ ha infinite cifre decimali e richiederebbe quindi uno spazio disco infinito per essere memorizzato. I numeri in $\mathbb{R}$ devono quindi essere approssimati e rappresentati nella aritmetica floating point.

Assumiamo quindi che esista una funzione fl che, dato un numero $x \in \mathbb{R}$ restituisca un numero $fl(x)$ rappresentabile dal calcolatore, detto numero macchina, e appartenente all'insieme $\mathbb{F}$, detto appunto insieme dei floating point. In generale $x \neq fl(x)$ e avremo che $\mathbb{F} \subset \mathbb{R}$. Un numero $y \in \mathbb{F}$ può essere scritto nel seguente modo

$$y = (-1)^s (0.a_1 a_2 \dots a_t)\beta^e,$$

dove

- s è utilizzato per definire il *segno*, in particolare $s = 0$ se $y > 0$, $s = 1$ se $y < 0$
- β è la *base*, ad esempio 2 o 10
- e è l'esponente
- i numeri che seguono la "virgola" (point) formano la *mantissa* $m = a_1 a_2 \dots a_t$, in cui si assume che $a_1 \neq 0$ e $0 \leq a_1 \leq \beta - 1$.

Esempio 1.1

Consideriamo il seguente numero $x = 0.04573 \in \mathbb{R}$ e scegliamo $\beta = 10$, la sua rappresentazione in $\mathbb{F}$ è data da

$$y = fl(x) = (-1)^0 (0.4573) 10^{-1},$$

abbiamo quindi la mantissa data da $m = 4573$, con $a_1 = 4$, $a_2 = 5$, $a_3 = 7$ e $a_4 = 3$. Inoltre $s = 0$ essendo x un numero positivo.

L'insieme $\mathbb{F}$ risulta quindi definito data la base β, il numero di cifre significative t e il massimo e minimo esponente dati da (U, L), rispettivamente. Scriviamo quindi $\mathbb{F} = \mathbb{F}(\beta, t, L, U)$.

L'occupazione di memoria di un numero floating point dipende dalla lunghezza della mantissa (numero di cifre memorizzate) e dal range di esponenti considerato.

In MATLAB , Python e in generale nei codici di calcolo numerico i calcoli vengono effettuati considerando numeri floating point in *doppia precisione* (double) per cui ad ogni numero sono dedicati 64 bit, così ripartiti:

- 1 bit per il segno
- $\log_2(U - L + 1) = 11$ bit per l'esponente, che varia da -1022 a 1023;
- $t \log_2(\beta) = 53$ bit per la mantissa (con t numero di cifre), di cui solo 52 esplicitamente memorizzati, che corrisponde a 16 cifre in base 10.

L'insieme dei floating point in precisione doppia è quindi:

$$\mathbb{F}(2, 52, -1022, 1023),$$

abbiamo quindi che in $\mathbb{F}$ il numero più grande, in modulo, che si può rappresentare è dato da circa $\beta^U \approx 10^{308}$. Se un'operazione produce un risultato maggiore di 10^{308} si ottiene un errore di tipo overflow e il computer tipicamente restituisce il valore `Inf`. Un'altra conseguenza del numero limitato di cifre a disposizione è il fatto che la distribuzione di numeri nell'insieme F non è uniforme. Per comprendere meglio, consideriamo un caso "giocattolo" con $\mathbb{F}(2, 2, -2, 2)$: le uniche due mantisse possibili sono 1 0 e 1 1, che in formato decimale risultano $0.10 \xrightarrow{\text{base } 10} 0 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} = 0.5$ e $0.11 \xrightarrow{\text{base } 10} 0 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} = 0.75$. Considerando tutti i 5 possibili esponenti (interi) generano in totale 10 numeri, così distribuiti:
`0.1250 0.1875 0.2500 0.3750 0.5000 0.7500 1.0000 1.5000 2.0000 3.0000`

Possiamo notare che i numeri più piccoli sono vicini tra loro, mentre i numeri grandi sono molto distanti tra loro dato che il numero di cifre significative che possono essere usate per rappresentare i numeri è fissato. Consideriamo la rappresentazione in doppia precisione del numero 1: il primo numero in $\mathbb{F}$ maggiore di 1 si ottiene cambiando l'ultima cifra della mantissa e corrisponde a $1 + 2^{-52}$; la risoluzione intorno a 1 è quindi pari a circa $2e^{-16}$. Questo ci porta a concludere che se dati due numeri u, $v \in \mathbb{R}$ tali che $v \neq u$, è possibile che $fl(u) = fl(v)$.

Un altro esempio associato all'aritmetica floating point è dato dal fenomeno della cancellazione di cifre significative. Consideriamo la seguente espressione

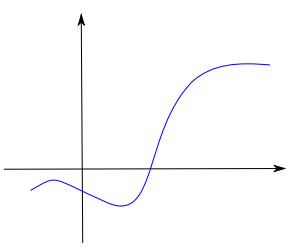
$$l = \frac{(1 + x) - 1}{x} \quad \text{con} \quad x \neq 0,$$

chiaramente se $x \in \mathbb{R}$ otteniamo $l = 1$, tuttavia se operiamo in aritmetica finita e scegliendo $x = 10^{-15}$ otteniamo che $l \approx 1.11$. Questo perché il valore di x per essere sommato ad 1 viene rappresentato solo dalle ultime cifre significative. Andiamo quindi a commettere un errore dell'11%, dovuto al fatto che il numeratore è dato da circa $1.11 \cdot 10^{-15}$.

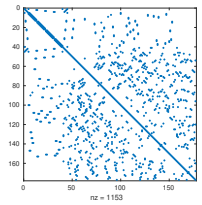
Metodi numerici

Esistono diverse classi di metodi numerici sviluppati per la risoluzione di differenti problemi. In particolare citiamo

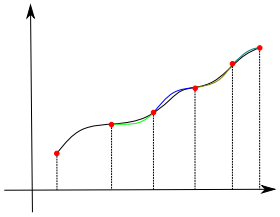
- risoluzione di equazioni e sistemi non-lineari



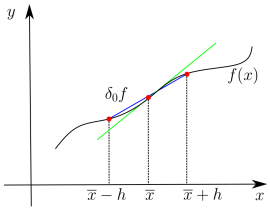
- risoluzione di grandi sistemi lineari



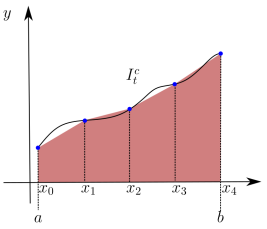
- interpolazione di dati e calcolo della funzione approssimante



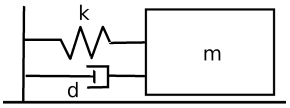
- approssimazione di derivate



- approssimazione di integrali



- risoluzione di equazioni differenziali ordinarie



- risoluzione di equazioni differenziali alle derivate parziali.