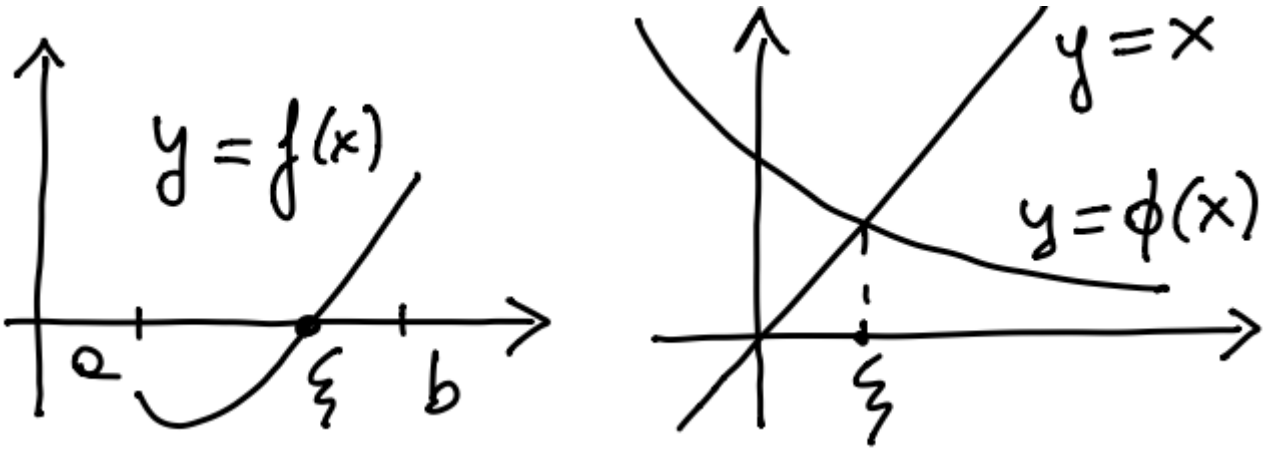
8. Introduzione alla soluzione numerica di equazioni non lineari, metodo di bisezione

In questo e nei prossimi tre capitoli ci occuperemo di un argomento classico del calcolo numerico, ovvero la soluzione numerica, cioe' approssimata, di equazioni non lineari.
Tratteremo due tipi di equazioni:

- $f(x) = 0$ *zeri* di funzione;
- $x = \phi(x)$ equazioni di *punto fisso*;

che possiamo schematizzare coi seguenti disegni

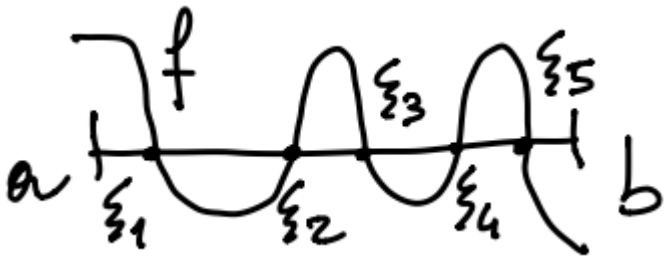


Il primo tipo corrisponde al calcolo di uno zero di una funzione (continua), cioe' di un punto ξ in cui f si annulla.
Il secondo tipo corrisponde invece al calcolo del punto fisso ξ di una funzione (continua) ϕ e si puo' interpretare come calcolo dell'ascissa dell'itersezione del grafico della bisettrice del primo e terzo quadrante $y = x$ col grafico di $y = \phi(x)$.
In entrambi i casi daremo condizioni *sufficienti* per l'*esistenza* e l'*unicita'* della soluzione in un dato intervallo e discuteremo, analizzandoli in dettaglio, i tre metodi classici di soluzione: i metodi di **bisezione** e di **Newton** (tangenti) per gli zeri e il metodo delle **iterazioni di punto fisso**.

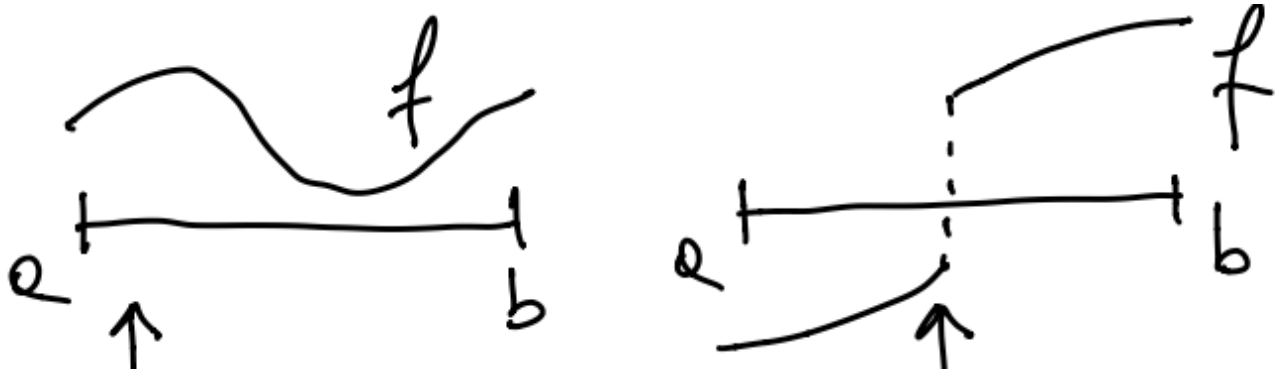
Cominciamo col ricordare alcuni risultati di esistenza e unicita' nel caso della ricerca di zeri.

> **Teorema** (degli zeri di funzioni continue)
> Sia $f \in C[a,b]$, cioe' f continua nell'intervallo chiuso e limitato $[a,b]$ e $f(a)f(b) < 0$, cioe' f cambia segno agli estremi $\implies \exists\, \xi \in (a,b) : f(\xi) = 0$.

Osserviamo che tale zero puo' non essere unico



D'altra parte togliendo una delle ipotesi, la condizione restante non basta a garantire l'esistenza



Diamo anche due classiche condizioni *sufficienti*, ciascuna delle quali garantisce l'*unicita'* dello zero:

- f *strettamente* monotona;
- $f \in C[a,b], f(a)f(b) < 0$ e f *strettamente* convessa o concava in $[a,b]$.

Nel caso in cui f e' derivabile in $[a,b]$ la monotonia stretta e' legata al segno di f', cosi' come concavita' e convessita' strette sono legate al segno di f''.

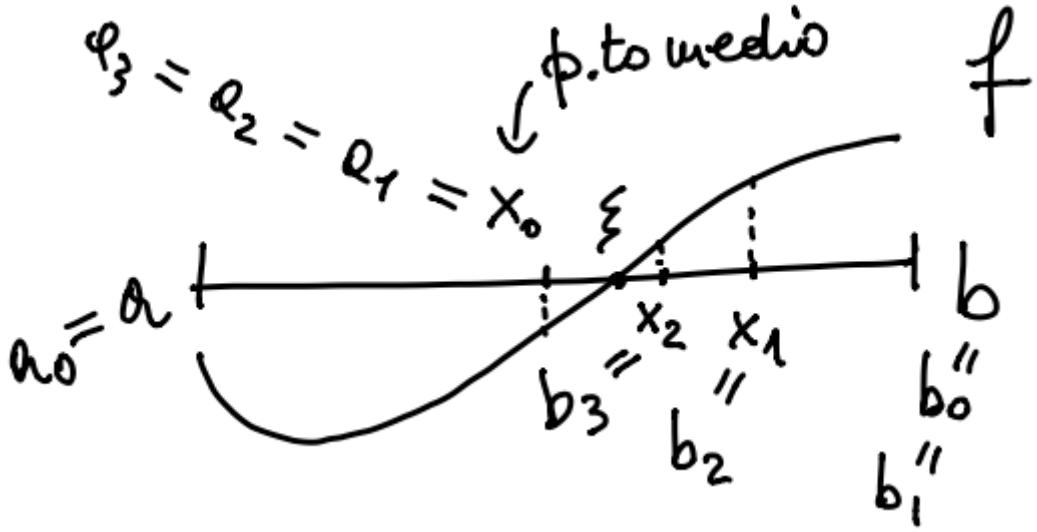
Fatti questi brevi richiami teorici su esistenza e unicita' delle soluzioni di equazioni scritte nella forma $f(x) = 0, x \in [a,b]$, introduciamo uno dei metodi piu' semplici per la soluzione numerica, il metodo di **bisezione**.

8.1. Metodo di Bisezione

Il metodo di bisezione consiste nell'applicazione iterativa del teorema degli zeri di funzioni continue, quindi si assume che

$$f \in C[a,b], f(a)f(b) < 0$$

Illustriamo graficamente la costruzione delle iterazioni



L'idea e' la seguente: si parte da $[a_0, b_0] = [a, b]$, si calcola il punto medio $x_0 = \frac{(a_0 + b_0)}{2}$. Ora, se $f(x_0) = 0$ siamo su uno zero altrimenti, siccome $f(a_0)$ e $f(b_0)$ hanno segno definito, $f(x_0)$ sara' discorde con uno solo dei due e quindi sicuramente ci sara' uno zero in (a_0, x_0) se $f(a_0)f(b_0) < 0$ altrimenti ci sara' uno zero in (x_0, b_0) visto che $f(x_0)f(b_0) < 0$, sempre per il teorema degli zeri.
Nel primo caso si definisce $a_1 = a_0, b_1 = x_0$, mentre nel secondo $a_1 = x_0, b_1 = b_0$, con la garanzia che $\exists\, \xi \in (a_1, b_1)$ tale che $f(\xi) = 0$ visto che $f(a_1)f(b_1) < 0$.

Il procedimento viene iterato applicando ripetutamente il teorema degli zeri per passare da $[a_n, b_n]$ ad $[a_{n+1}, b_{n+1}]$ in cui uno degli estremi e' diventato il punto medio $x_n = \frac{(a_n + b_n)}{2}$ di $[a_n, b_n]$.

Si tratta in generale di un processo infinito (a meno che qualche n non riuslti $f(x_n) = 0$) che permette di costruire tre successioni $\{a_n\}, \{b_n\}, \{x_n\}$ tali che:

- $\exists\, \xi : f(\xi) = 0, \xi \in (a_n, b_n)$;
- $|\xi - a_n|, |\xi - b_n| \leq b_n - a_n = \frac{b-a}{2^n}$;
- $|\xi - x_n| < \frac{b_n - a_n}{2} = \frac{b-a}{2^{n+1}}$;
  con $n = 0, 1, 2, \ldots$.

Il nome "bisezione" viene dal fatto che l'intervallo viene diviso iterativamente a meta', "buttando via" ad ogni iterazione mezzo intervallo per restare nella meta' dove f cambia segno e dove quindi c'e' sicuramente uno zero ("lo" zero nel caso questo sia unico in (a, b), ma in generale il metodo funziona anche con vari zeri, calcolandone uno).

Si vede subito che il metodo e' convergente, cioe' che tutte e tre le successioni convergono ad uno zero $\xi \in (a, b)$. Infatti

$$0 \leq |\xi - a_n|, |\xi - b_n| < \frac{b-a}{2^n} \to 0, n \to \infty$$

e per il teroema dei due carabinieri

$$|\xi - a_n|, |\xi - b_n| \to 0, n \to \infty$$

analogamente si vede che

$$|\xi - x_n| \to 0, n \to \infty$$

visto che

$$0 \leq |\xi - x_n| < \frac{b-a}{2^{n+1}} \to 0, n \to \infty$$

Quest'ultima disuguaglianza e' immediatamente comprensibile dal momento in cui sappiamo che ξ zero di f sta in (a_n, b_n), non sappiamo dove, ma sicuramente la sua distanza dal punto medio x_n e' minore dell'intervallo $[a_n, b_n]$ cioe' e' $< \frac{b_n - a_n}{2}$.
In altri termini, ξ sta nell'intorno aperto di centro x_n e raggio $\frac{b_n - a_n}{2}$.
Nel metodo di bisezione si sceglie x_n come successione di approssimazione, visto che la *stima* dell'errore e' migliore di un fattore $\frac{1}{2}$ rispetto a quella delle successioni a_n e b_n.

Volendo garantire una tolleranza $\epsilon > 0$ nel calcolo approssimato dello zero ξ, basta quindi risolvere la disuguaglianza

$$e_n = |\xi - x_n| < \frac{b-a}{2^{n+1}} \leq \epsilon$$

in modo che $x_n \in (\xi - \epsilon, \xi + \epsilon)$, ovvero $2^{n+1} \geq \frac{b-a}{\epsilon}$, cioe

$$n + 1 \geq \log_2 \left(\frac{b-a}{\epsilon} \right) = \log_2 (b-a) + \log_2 \left(\frac{1}{\epsilon} \right)$$

Questa disuguaglianza permette "a priori", cioe' prima di iniziare il processo di calcolo, di decidere a quale iterazione fermarsi in modo da garantire la tolleranza $\epsilon > 0$, basta prendere

$$n(\epsilon) = \text{parte intera} \left(\log_2 (b-a) + \log_2 \left(\frac{1}{\epsilon} \right) \right)$$

In effeti una stima del tipo $e_n \leq \text{stima}(n)$ dove la stima non dipende dalle quantita' calcolate si chiama usualmente *stima a priori*.

Il problema con le stime a priori e' che sono spesso **sovrastime**, cioe' non sono vicine all'errore effettivo ma ne danno solo un confine superiore garantito in modo teorico, che pero' puo' portare a un aumento del numero di iterazioni e quindi del costo computazionale rispetto a quello che sarebbe sufficiente ad ottenere la tolleranza richiesta.

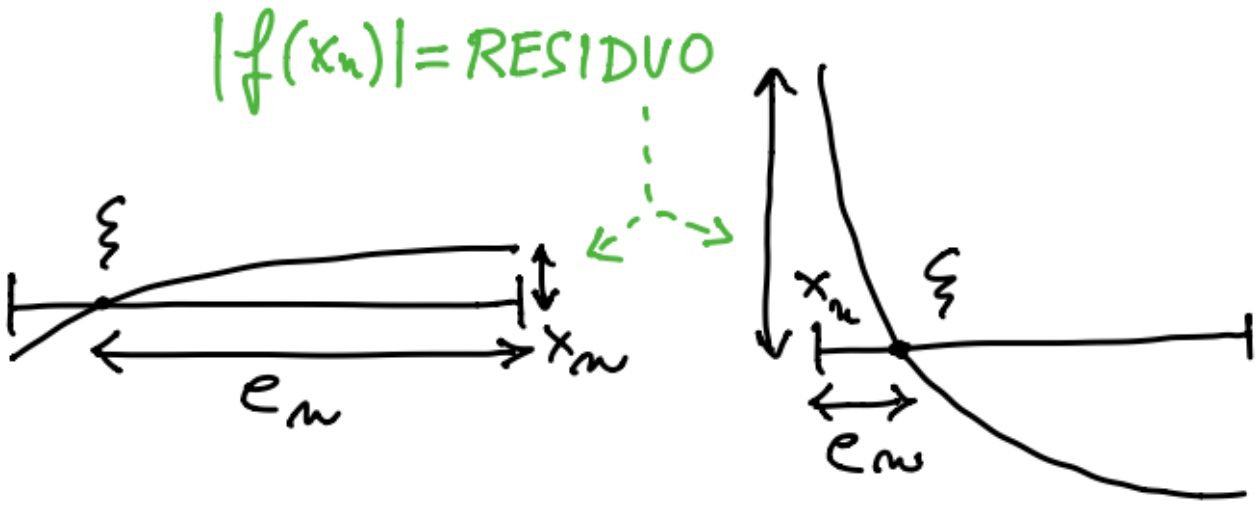
Per ottenere una stima dell'errore piu' aderente, cominciamo col fare la seguente osservazione: visto che $x_n \to \xi, n \to \infty$ e che f e' continua, si avra' che

$f(x_n) \to f(\xi) = 0, n \to \infty$.
Quella che stiamo usando qui in realta' e' una caratterizzazione della continuita di una funzione in analisi matematica che ci dice che "f e' continua se e solo se il limite si puo' trasportare dentro la funzione".
Nel nostro caso $f(\xi) = 0$ quindi $f(x_n) \to 0, n \to \infty$ e anche $|f(x_n)| \to 0, n \to \infty$.
La quantita' $|f(x_n)|$ si chiama **residuo** perche' dice quanto "resta" ad f per annullarsi.

Viene allora spontanea la seguente domanda: siccome $f(x_n) \to 0$, $n \to \infty$, possiamo arrestare il processo di calcolo quando il residuo $|f(x_n)|$ e' piccolo? In altre parole

$$|f(x_n)| \leq \epsilon \overset{?}{\implies} e_n \leq \epsilon$$

La risposta e' **NO** perche', come vedremo subito, la grandezza del residuo non e' in se' un buon indicatore dell'errore, ma va opportunamente "pesata": per capirlo consideriamo i seguenti grafici



Nel primo caso il residuo e' piccolo ma l'errore e' grande, cioe' il residuo e' una *sottostima* dell'errore. Nel secondo caso il residuo e' grande ma l'errore e' piccolo, cioe' il residuo e' una *sovrastima* dell'errore.

E' importante osservare che una sottostima dell'errore in pratica e' la cosa piu' *pericolosa*, perche' induce a fermare le iterazioni quando x_n non e' ancora nell'intorno del limite individuato dalla tolleranza: si pensa di aver approssimato la quantita' limite a meno della tolleranza e invece *l'errore e' piu' grande della tolleranza*.

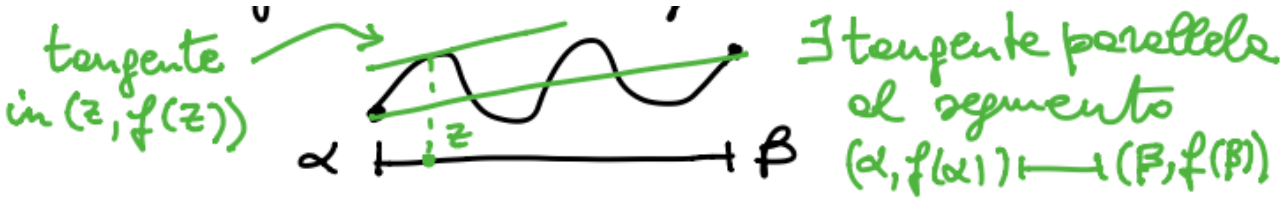
Questo puo' chiaramente portare a conseguenze gravi in applicazioni in cui il rispetto della tolleranza e' decisivo.
D'altra parte, una sovrastima, pur essendo meno grave, ha come conseguenza un aumento del numero di iterazioni rispetto a quello che sarebbe sufficiente e quindi un incremento del costo computazionale.
Nei due grafici disegnati sopra si nota che il residuo e' una sottostima dell'errore quando la funzione e' "piatta", cioe' la variazione e' lenta in un intorno dello zero,

mentre e' una sovrastima quando la funzione e' "ripida" cioe' ha una variazione veloce in un intorno dello zero.

Si capisce allora che il residuo va in qualche modo "pesato" per tener conto della velocita' di variazione: se f e' derivabile, bisogna quindi tenere conto della grandezza della derivata, che per definizione misura la velocita' di variazione di una funzione. Per fare questo in modo rigoroso possiamo ricorrere a un teorema chiave del calcolo differenziale, il teorema del **valor medio** (detto anche teorema di Lagrange).

> **Teorema** (del valor medio)
> Sia $f \in C[\alpha, \beta]$ derivabile in $(\alpha, \beta) \implies \exists\, z \in (\alpha, \beta)$: $\dfrac{f(\beta) - f(\alpha)}{\beta - \alpha} = f'(z)$

L'interpretazione geometrica del teorema e' la seguente



Tornando all'analisi del residuo nel metodo di bisezione, mettiamoci nelle seguenti ipotesi:

- $f \in C^1[a, b]$, cioe' f e' derivabile con derivata prima continua in $[a, b]$;
- $\{x_n\} \subset [c, d] \subseteq [a, b]$, almento per $n \geq n_0$, cioe' per n abbastanza grande;
- $x_n \to \xi, n \to \infty, f(\xi) = 0$ e $f'(x) \neq 0 \forall\, x \in [c, d]$.

Allora vale la rappresentazaione:

$$e_n = |x_n - \xi| = \frac{|f(x_n)|}{|f'(z_n)|}, n \geq n_0$$

con $z_n \in \text{int}(x_n, \xi)$, l'intervallo aperto che ha per estremi x_n e ξ.

Prima di domostrare la stima, osserviamo che:

1. la rappresentazione dell'errore mostra chiaramente che l'*errore* e' un *residuo pesato* dalla derivata;
2. l'ipotesi che $f'(x) \neq 0$ in $[c, d] \subseteq [a, b]$ e' equivalente all'ipotesi che lo zero sia *semplice*, ovvero che $f'(\xi) \neq 0$.

Infatti se $x_n \in [c, d], \forall\, n > n_0$ allora $\xi = \lim x_n \in [c, d]$ perche' $[c, d]$ e' chiuso e quindi contiene i limiti delle successioni li' contenute, quindi $f'(\xi) \neq 0$. Viceversa, se $f'(\xi) \neq 0$ siccome f' e' continua, per il teorema della *permanenza del segno*

$$\exists\, \delta > 0 \colon f'(x) \neq 0\ \forall\, x \in [\xi - \delta, \xi + \delta] = [c, d]$$

e siccome $x_n \to \xi, n \to \infty$ allora $\exists\ n_0$ tale che $|x_n - \xi| \leq \delta\ \forall\ n \geq n_0$.
Si noti che $f'(z_n) \neq 0, n \geq n_0$ perche' $z_n \in \text{int}(x_n, \xi) \subset [c, d]$.
E' importante osservare che la rappresentazione dell'errore come residuo pesato non vale solo per il metodo di bisezione, ma per ogni metodo convergente ad uno zero semplice se $f \in C^1$ (applicheremo infatti questo risultato piu' avanti col metodo di Newton);

3. dalla rappresentazione siamo in grado di ricavare delle *stime a posteriori* dell'errore. A posteriori perche' si utilizza il residuo che e' calcolabile solo a posteriori dopo aver prodotto x_n nel processo di calcolo.

> **Dimostrazione** della rappresentazione
> Utiliziamo il teorema del valor medio, supponendo che $x_n > \xi$ (l'altro caso e' del tutto analogo), con $\alpha = \xi, \beta = x_n$.
> Sappiamo che $\exists\ z_n$ tale che
>
> $$\frac{f(x_n) - \overset{=0}{f(\xi)}}{x_n - \xi} = \frac{f(x_n)}{x_n - \xi} = f'(z_n)$$
>
> allora $f(x_n) - \underbrace{f(\xi)}_{=0} = f'(z_n)(x_n - \xi), z_n \in (\xi, x_n)$ cioe' $|f(x_n)| = |f'(z_n)||x_n - \xi|$ che si puo' riscrivere come $e_n = |x_n - \xi| = \frac{|f(x_n)|}{|f'(z_n)|}$.

Ora, siccome il teorema del valor medio non ci dice chi sia z_n, ma solo che esiste almeno un z_n in $\text{int}(x_n, \xi)$, cerchiamo di ricavare delle stime "pratiche" dell'errore utilizzando il residuo opportunamente pesato:

1. se e' noto che $f'(x) \geq k > 0\ \forall\ x \in [a, b]$ (ma basta $\forall\ x \in [c, d]$) allora $e_n = \frac{|f(x_n)|}{|f'(z_n)|} \leq \frac{f(x_n)}{k}$;
2. se f' e' nota o calcolabile, siccome $z_n \to \xi, n \to \infty$ per il teorema dei due carabinieri visto che z_n sta fra ξ e x_n, per la continuita' di f' si ha che $|f'(x_n)|, |f'(z_n)| \to |f'(\xi)| \neq 0, n \to \infty$.``

Quindi, almeno per n abbastanza grande, $|f'(x_n)|$ e $|f'(z_n)|$ saranno entrambi dell'ordine di grandezza di $|f'(\xi)|$ (notiamo che nel residuo pesato quello che ci interessa e' essenzialmente l'*ordine di grandezza del peso* $f'(z_n)$).
Abbiamo quindi una **stima empirica** $e_n = |x_n - \xi| \approx \frac{|f(x_n)|}{|f'(x_n)|}$ valida almeno per $n \geq n_0$, dove n_0 corrisponde ad un controllo empirico che l'ordine di grandezza di $|f'(x_n)|$ si stia "stabilizzando", cioe valga $\left| \frac{|f'(x_n)|}{|f'(x_{n-1})|} - 1 \right| \leq \delta$.
Ad esempio con $\delta = 10^{-1}$ o 10^{-2} ($\frac{|f'(x_n)|}{|f'(x_{n-1})|} \to 1, n \to \infty$).
Quindi ha senso controllare quando il rapporto si sta stabilizzando intorno a 1, tenendo

presente che questo e' un criterio sensato ma non completamente rigoroso, diciamo una "linea guida" per l'utilizzo del peso;

3. se f' non e' nota esplicitamente, va in qualche modo approssimata.

Oserviamo infatti che non sempre abbiamo a disposizione una formula analitica per f (come ad esempio per l'equazione algebrica $x^2 - 2 = 0$ corrispondente al calcolo di $\sqrt{2}$ o per l'equazione non algebrica $x - e^{-x} = 0$).
Infatti f potrebbe essere nota in forma di "scatola nera"

$$x \to \boxed{\text{f}} \to f(x)$$

cioe' potremmo averne a disposizione solo i valori da misure o altri algoritmi.
Se pero' sappiamo almeno che f e' derivabile, possiamo approssimare f' con un rapporto incrementale costruito con le quantita' calcolate $f'(z_n) \approx \dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$ (qui ad esempio prendiamo le ultime due iterazioni).

Detto k_n il peso calcolato con uno degli approcci 1, 2 o 3, siamo allora in grado di scrivere un **test di arresto** per il metodo di bisezione che *combina* stima a priori e stima a posteriori

$$\min \left\{ \frac{b-a}{2^{n+1}}, \frac{|f(x_n)|}{k_n} \right\} \leq \epsilon$$

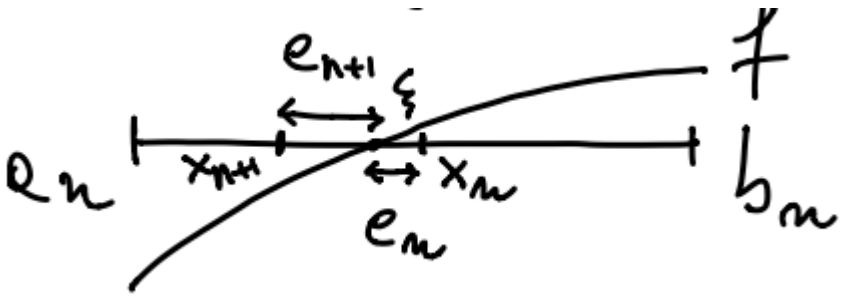
Possiamo a questo punto fare alcune considerazioni di carattere computazionale:

- il metodo di bisezione funziona con richieste analitiche e computazionali minime;

La versione base con la stima a priori chiede solo che siano soddisfatte le ipotesi del teorema degli zeri (continuita' e cambio di segno agli estremi) e unicamente la possibilita' di calcolare correttamente il segno di $f(x_n)$ (per cui e' sufficiente un errore relativo $< 100\%$ su $f(x_n)$).

Infatti in generale se $\tilde{\alpha} \approx \alpha \neq 0$ e $\dfrac{|\alpha - \tilde{\alpha}|}{|\alpha|} < 1 \implies \text{sgn}(\tilde{\alpha}) = \text{sgn}(\alpha)$

- mentre la stima a priori e' descrescente, l'errore effettivo in generale non lo e', come si vede da questo disegno



dove $e_{n+1} > e_n$ (anche se comunque $e_n \to 0, n \to \infty$); qui si vede anche che $e_n \ll \frac{b_n - a_n}{2}$ e la stima del residuo pesato e' tendenzialmente piu' accurata, a differenza della stima a priori che si dimezza, $e_{n+1} \approx \frac{e_n}{2}$, solo "in media", su un po' di iterazioni.

Concludiamo il capitolo con un esempio, il calcolo approssimato di $\sqrt{2}$ risolvendo l'equazione algebrica $x^2 - 2 = 0$ con il metodo di bisezione.

Esempio (calcolo di $\sqrt{2}$ alla precisione di macchina)

Consideriamo l'equazione algebrica $f(x) = x^2 - 2 = 0$. Calcolarne la soluzione positiva significa calcolare $\sqrt{2}$.
Le ipotesi del teorema degli zeri sono soddisfatte in $[a, b] = [1, 2]$. Infatti $f \in C[a, b]$ (anzi a dirla tutta $f \in C^\infty(\mathbb{R})$ cioe' e' derivabile infinite volte in $\mathbb{R}$ con derivate tutte continue, perche' f e' un polinomio).

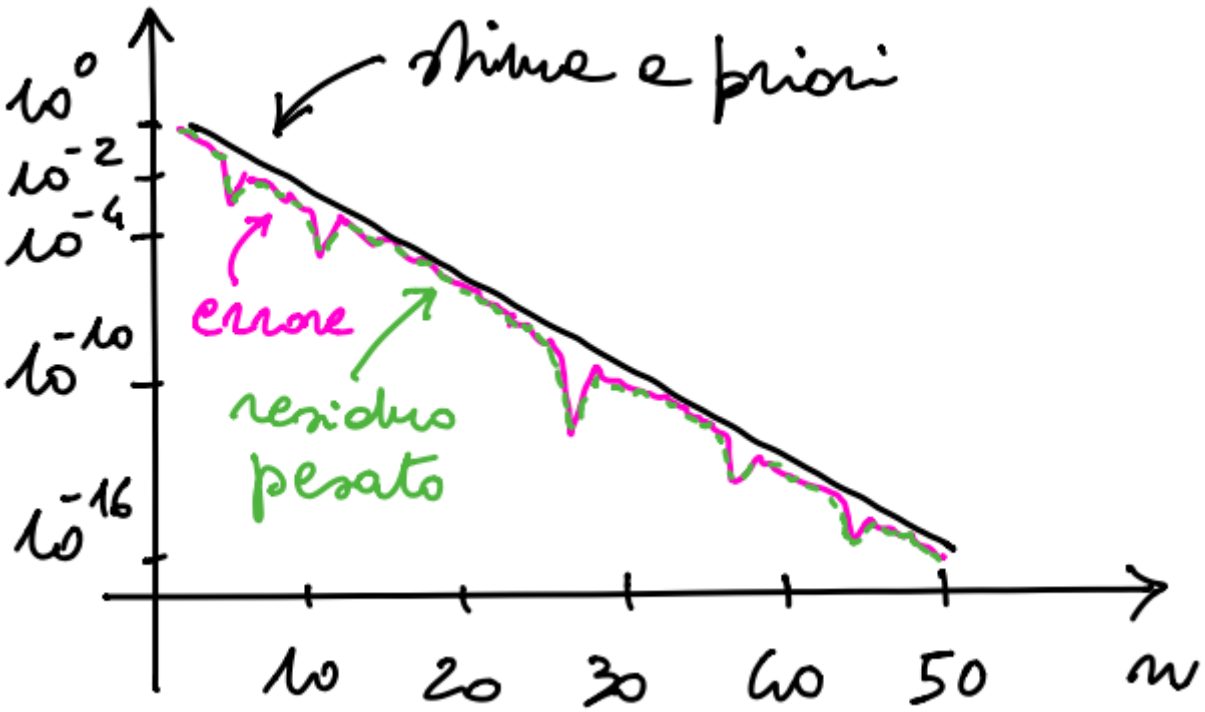
$$f(a) = f(1) = 1 - 2 = -1 < 0$$
$$f(b) = f(2) = 2^2 - 2 = 2 > 0$$

Inoltre $f'(x) = 2x \geq 2 \; \forall \; x \in [1, 2]$ quindi $\sqrt{2} \in (1, 2)$ ed e' l'unico zero in tale intervallo (in effetti noi sappiamo che e' l'unico zero in $\mathbb{R}^+$).
Possiamo applicare il metodo di bisezione che comincia in questo modo:

$$x_0 = \frac{1 + 2}{2} = 1.5$$
$$x_1 = \frac{1 + 1.5}{2} = 1.25$$
$$x_2 = \frac{1.25 + 1.5}{2} = 1.375$$
$$x_3 = \frac{1.375 + 1.5}{2} = 1.4375$$
$$x_4 = \cdots$$

ricordando che $\sqrt{2} = 1.4142\ldots$.

Nel grafico sottostante (in scala logaritmica) riportiamo l'errore effettivo, la stima a priori $\frac{b_n - a_n}{2} = \frac{1}{2^{n+1}}$ e la stima a posteriori $\frac{|f(x_n)|}{k} = \frac{|x_n^2| - 2}{2}$ (visto che $f'(x) \geq k = 2 \; \forall \; x \in [1, 2]$) tutte relativizzate a $|\xi| = \sqrt{2}$ (in questo caso comunque $|\xi|$ e' dell'ordine dell'unita' e quindi errore assoluto e relativo sono vicini).



Come sempre per comodita' i valori discreti sono interpolati con linee continue o tratteggiate.

Si vede che l'errore segue solo "in media" l'andamentodella stima a priori, che lo sovrastima a volte di vari ordini di grandezza (nei picchi dell'errore verso il basso ad esempio tra $n = 20$ e $n = 30$, l'errore va circa a 10^{-11} mentre la stima a priori ha bisogno di una decina di terazioni in piu').

D'altra parte la stima del residuo pesato e' praticamente sovrapposta all'errore effettivo.
Si noti infine che per raggiungere un errore dell'ordine della precision di macchina $\epsilon_M = 2^{-53}$ servono circa 50 iterazioni, il che non e' sorprendente visto che il fattore medio di riduzione dell'errore e' $\frac{1}{2}$.