# Richiami di algebra lineare

## Norma di un vettore

La norma è una funzione che associa ad ogni vettore di componenti , uno scalare in modo che valgano le seguenti proprietà:

* *=*
* *≤ +*
* *=>*

Una delle norme più usate è la **norma euclidea**, che è capace di quantificare la lunghezza del vettore ed è definita come segue:



essendo la radice quadrata della somma dei quadrati delle componenti del vettore, ovvero l’enunciato del teorema di Pitagora.

Esistono anche altre norme comunemente usate:



Per esempio la norma-1 è la somma dei moduli delle componenti del vettore. Quindi l’applicazione della norma-1, oppure della norma-2 ad un vettore, fornisce valori anche molto diversi dello scalare che ne deriva.

E’ possibile valutare graficamente l’effetto dell’applicazione delle due norme prendendo come riferimento la sfera. Si definisce sfera di norma-p, raggio r e centro **x0** l’insieme:



La superficie esterna di questo insieme costituisce il luogo dei punti con . La forma del luogo dei punti può essere diversa a seconda del tipo di norma usata. Per esempio, nel caso bidimensionale e raggio unitario, la circonferenza appare come in Figura 1 (a) o (b) se calcolata rispettivamente utilizzando la norma 2 oppure la norma 1.



(a) (b)

Figura 1

Da questo risultato, seppur ottenuto tramite un semplice esempio, si evidenza che la norma-2 restituisce il modulo delle due componenti indipendentemente dai valori assunti dalle coppie delle coordinate scelte. In altre parole la norma-2 non si presta per essere utilizzata per discriminare la diversità dei valori di ciascuna delle due componenti del vettore.

La norma-1 invece esalta alcuni valori rispetto ad altri, in particolare mette in evidenza situazioni in cui emerge una delle due componenti del vettore. Per questi ed altri motivi nel seguito verrà utilizzata sia la norma-2 che la norma-1.

### Applicazione della norma-2: il metodo dei minimi quadrati (Least Square – LS)

Nella soluzione di problemi lineari è frequente che si presenti un problema nella forma: , con *n* equazioni in *m* incognite (con n > m), e quello che si vuole ottenere è una stima delle *x* dopo aver misurato *y* e conoscendo il modello di misura rappresentato da *A*.

Se le misure fossero ideali, ovvero senza alcun disturbo, effettivamente il sistema sarebbe risolubile come problema inverso, ma dal momento che il modello reale di misura è:

la risoluzione algebrica per trovare x non è praticabile.

Sappiamo che l’errore potrebbe provenire oltre che da un disturbo esterno che si sovrappone alla misura di *y,* anche dalla scelta di un modello di misura non ideale (errori sulla *A*). A questo proposito si distinguono due tipi di problemi:

**Problema mal-condizionato**: è un problema in cui piccole variazioni dell’ingresso determinano variazioni inaccettabili sull’uscita, ma **in modo continuo**. In altri termini variando l’ingresso in modo infinitesimo l’uscita è modificata in modo pesante, ma proporzionalmente all’ingresso.

**Problema mal-posto**: è un problema in cui piccole variazioni dell’ingresso determinano inaccettabili variazioni sull’uscita **in modo** **non continuo**. Ciò significa che se si modifica l’ingresso anche in modo infinitesimo, l’uscita può cambiare drasticamente.

Il problema lineare di rintracciare le *x* note le *y* e la *A* è un problema mal posto perché anche una minima perturbazione dei coefficienti di *y* o di *A* può determinare il passaggio da sistema risolubile a sistema non risolubile.

Supponendo che *y* non faccia parte dell’immagine di *A* (nel caso 2D ciò significa che y non si trova sulla retta individuata da Ax, vedi Figura 2), **il problema di trovare *x* non può essere risolto con esattezza**.

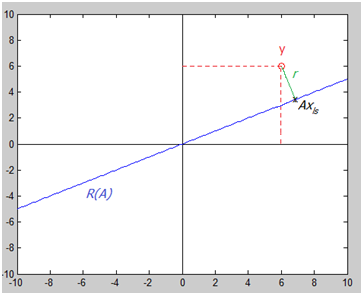


Figura 2: Axls proiezione di y sul range di A

Un possibile modo di procedere consiste nel cercare di risolvere il problema in maniera approssimata, cercando un vettore *x* che sia capace di minimizzare l’errore tra la vera misura (*y*) e quella ottenuta con il modello:

Ci troviamo di fronte a un problema di minimizzazione e in generale in questo tipo di problema la funzione da minimizzare prende il nome di ***funzione obiettivo*** (o **funzione di costo**). In questo caso la funzione obiettivo è rappresentata dalla norma del ***residuo*** -y; tipicamente la norma che si usa è quella euclidea perché permette di evitare problemi nella ricerca del minimo mediante l’imposizione a zero della derivata, ma potrebbero essere usati anche altri tipi di norme.

Usando la norma euclidea il problema di minimizzazione si può scrivere come:

Per questo motivo il problema prende anche il nome di Least Squares.

In realtà il metodo dei minimi quadrati per la risoluzione del sistema non è l’unico possibile, per esempio si può usare anche il **metodo del gradiente o del gradiente coniugato**, tecniche iterative che richiedono la positività della matrice A.

Supponiamo, in primo luogo, che la matrice *A* sia di rango massimo . Quello che stiamo cercando è il vettore *x* capace di minimizzare la distanza tra il vettore *Ax* e il vettore *y* (vedi Figura 6). In altri termini stiamo cercando la *proiezione* di *y* nel range di *A*, ovvero Axls.

Imponiamo la minimizzazione della norma del residuo (si calcola la derivata e si uguaglia a zero):

L’equazione ottenuta prende il nome di *normal equation*. Quindi:

La matrice *Bls* rappresenta una generalizzazione della matrice inversa tale che . Nel caso che A sia quadrata la pseudo inversa ricade nella semplice inversa, ovvero .

Quindi, tornando all’esempio precedente, la matrice rappresenta la **matrice di proiezione** di *A*. Il residuo vale dunque , e *r* è perpendicolare all’immagine di *A* come è ben visibile in 2 e quindi rappresenta la minima distanza tra y e Ax.

## Regolarizzazione

In un problema di minimizzazione mediante i minimi quadrati si può porre il problema **della non unicità del minimo, oppure della sua instabilità rispetto al rumore**. La regolarizzazione consiste nell’aggiunta di un funzionale detto di ‘penalizzazione’, pesato con un parametro di peso positivo, detto parametro di regolarizzazione, scelti in modo tale che il nuovo funzionale abbia sempre un solo punto di minimo (diverso da quello senza regolarizzazione), ma che sia meno sensibile al rumore. Il fatto che il punto di minimo sia diverso da quello senza regolarizzazione significa che con la regolarizzazione ci si può allontanare dalla soluzione vera, però con vantaggi sulla riduzione anche drastica della fonte di errore o rumore.

Con la regolarizzazione vengono minimizzate due funzioni obiettivo, la prima costituita dal termine LS e la seconda dalla funzione di regolarizzazione. Utilizzando il formalismo matematico, si può scrivere:



dove è la matrice di regolarizzazione ovvero una matrice che determina quali componenti del segnale vorremmo penalizzare,  è il parametro di regolarizzazione scalare,  è la norma 1 ().

# Algoritmi iterativi

## Metodo del gradiente

### Premessa

Il metodo del gradiente è una delle tecniche iterative più conosciute e diffuse per la soluzione di problemi inversi della forma:



dove x è il vettore incognito da stimare, y è il vettore di misura e H è una matrice che descrive il sistema (di misura).

Esistono delle situazioni in cui non esiste l’inversa: . Abbiamo appena dimostrato che in questi casi si cerca la soluzione basata sulla minimizzazione della norma-2 del residuo:



che porta al calcolo della pseudo-inversa di Moore-Penrose nella forma:



**Tuttavia nel caso di immagini può non convenire invertire il prodotto**  **per questioni computazionali,** considerata la notevole dimensione di H. In questi casi un metodo molto usato per trovare la soluzione ottimale  è quello basato sui metodi iterativi.

### Razionale del metodo iterativo

*Soluzione I.*

Si voglia trovare la soluzione al seguente sistema: Hx = y. Un modo semplice per trovare il vettore incognito x senza dover invertire la matrice H o senza ricorrere alla pseudo-inversa di Moore Penrose, è quello di trovare il minimo della seguente funzione quadratica f(x):



il cui grafico è riportato in Figura 3.



Figura 3

Si può dimostrare che il primo e il terzo termine alla destra del segno di uguaglianza sono la rappresentazione vettoriale della funzione quadratica del tipo: ax^2 + by^2 + 2dxy + c, con H simmetrica.

Si dimostra che il minimo  della precedente funzione quadratica coincide con la soluzione del sistema: Hx = y. A questo proposito calcoliamo il gradiente di f(x) nell’ipotesi che H sia simmetrica:



Quindi coincide con il residuo del modello Hx = y. Se  è il punto di minimo, cioè , il residuo è nullo e:



Quindi il valore di  che annulla il gradiente è la soluzione cercata.

Se H non è simmetrica la soluzione sarà:



essendo  simmetrica.

Pertanto un modo per trovare il vettore incognito x senza dover invertire la matrice A o senza ricorrere alla pseudo-inversa di Moore Penrose è quello di trovare il minimo della funzione f(x).

*Soluzione II.*

In alternativa alla prima soluzione potremmo partire da una diversa tipologia di funzioni quadratiche ottenute direttamente dalla definizione di norma-2 come segue:



Con semplici passaggi si ottiene la seguente formula per il gradiente:



essendo H simmetrica. Anche in questo caso il valore di  che annulla il gradiente è la soluzione di f(x) ed è anche la soluzione di:



I valori di x tali che il gradiente è diverso da zero rappresentano l’errore tra la misura e il modello.

Quindi utilizzando le due precedenti soluzioni è possibile trovare la soluzione all’equazione Hx = y ricorrendo alla ricerca del minimo di una funzione quadratica che abbia il minimo coincidente con quello cercato.

### Algoritmo iterativo basato sulla discesa rapida del gradiente

In questo paragrafo si descrive una strategia efficiente dal punto di vista computazionale per ottimizzare la ricerca della soluzione per f(x). Il metodo della discesa rapida del gradiente è il più semplice e intuitivo.

Sappiamo che per calcolare il gradiente di una funzione si applica il seguente operatore gradiente:



Il gradiente risulta essere un campo vettoriale che per ciascun punto x punta nella direzione di maggiore incremento di f(x). La Fig. 4 riporta la mappa di gradiente ottenuta dalla forma quadratica f(x) di Figura 3, in cui si osserva che il gradiente è nullo nel punto più basso del paraboloide.



Fig. 4

Il metodo di discesa rapida è basato sulla seguente strategia:

1) per qualsiasi punto  si calcola il gradiente e si individua la direzione del massimo gradiente;

2) successivamente si procede nella direzione negativa rispetto a quella di massimo gradiente , cioè: . Ovviamente la direzione individuata dalla pendenza negativa del massimo gradiente individua la direzione di discesa rapida cercata.

Pertanto il metodo consiste nella scelta di un primo punto *k* a caso nello spazio multidimensionale x e nel calcolo del gradiente con segno negativo della funzione . I successivi punti della funzione quadratica vengono scelti lungo la direzione indicata dai gradienti sulla base della seguente condizione: . Essendo funzioni quadratiche, i gradienti successivi avranno un valore inferiore al precedente fino a quando non si raggiungerà il punto di minimo.

Supponiamo di inizializzare l’algoritmo nel punto n-dimensionale: . Nell’applicazione del metodo della massima discesa del gradiente si individua il punto successivo mediante la seguente relazione:



Nell’algoritmo del gradiente si definisce anche l’errore al passo k trae il corrispondente valore misurato y, , e successivamente si procede con le iterazioni fino a minimizzare tale errore.



Fig. 5

In Fig. 5 è mostrato un grafico con sovrapposto un esempio di traiettoria percorsa per raggiungere il punto di minimo. Un problema che si pone è quello di trovare il valore ottimale del peso in modo da raggiungere il punto più vicino al punto di minimo. Valori troppo bassi aumentano il numero di iterazioni per raggiungere il minimo, mentre valori troppo elevati rischiano di superare il minimo, rischiando di non convergere mai.

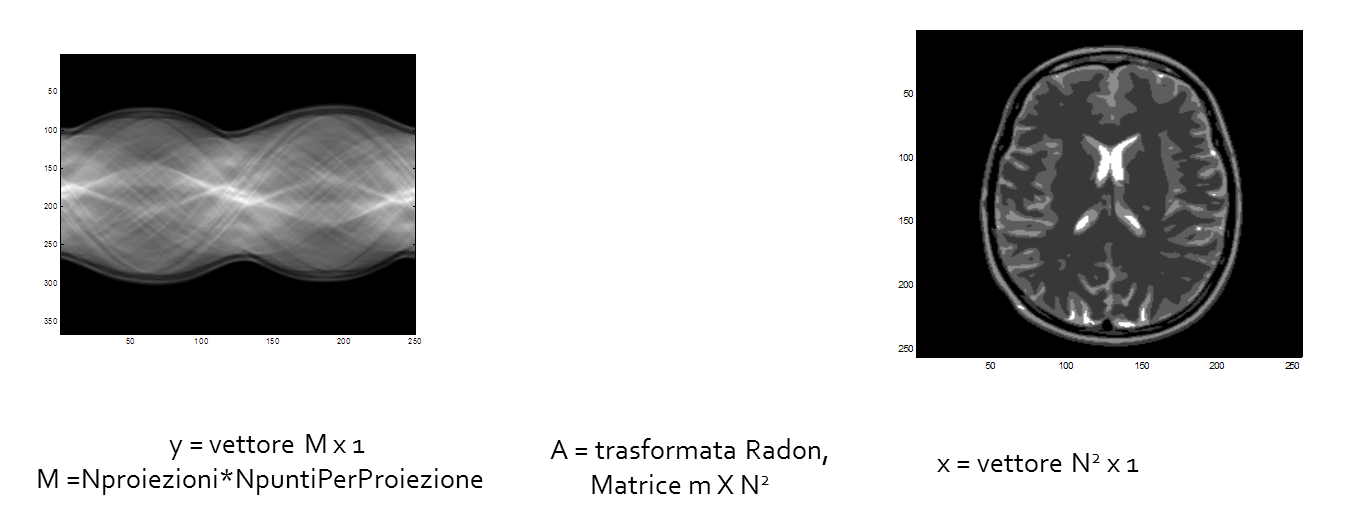
## Gradiente coniugato

Il metodo precedente è in generale molto lento. Mediante l’assunzione di condizioni di ortogonalità si può modificare il metodo del gradiente in modo da ottenere una convergenza molto più veloce. Il nuovo metodo che usa tali condizioni è denominato ‘metodo del gradiente coniugato’. Esso non viene trattato in questi appunti. Si sappia che l’idea principale consiste nel costruire una base di vettori ortogonali e utilizzare tale base per una ricerca più efficiente della soluzione.

##### Esercitazione n. 2.6: Applicazione del metodo del gradiente alla ricostruzione di immagini da sinogramma

In un problema descritto dal seguente modello:

**y = A \* x**



dove x è l’immagine da acquisire, A è la trasformata di Radon e y rappresenta il sinogramma. Quindi x è ciò che vogliamo stimare partendo dal sinogramma.

Si tratta di risolvere la seguente equazione in norma-2:



per trovare una stima di x.

In questa applicazione A è tipicamente rettangolare e quindi non esiste l’inversa: . D’altra parte si potrebbe ricorrere alla pseudo inversa di Moore Penrose, ma non conviene invertire la matrice per il carico computazionale. Pertanto si ricorre al metodo iterativo.

Come abbiamo già discusso, la formula iterativa al passo k è data da:



*Algoritmo*

* Caricare in matlab il file I0.mat
* Definire il vettore contenente gli angoli di proiezione da inserire nella funzione ‘radon’:
  + Theta = linspace(-180,180, numero proiezioni)
* Usare 150 o più proiezioni
* Creare due funzioni inline che rappresentano le operazioni di trasformata dirette A\*x e inversa At\*x, nel caso di trasformata radon:
  + A = @(x) radon(x, theta);
  + At = @(x) iradon(x, theta, 'nearest', 'None', size(I0,1));
  + Creare il sinogramma e aggiungere rumore nel modo seguente:
    - y = A(x) + noise;
* Effettuare la retroproiezione e la retroproiezione filtrata di y:
  + Ibp = At(y)
  + Ifbp = iradon(x, theta, 'nearest', ’Ram-Lak', size(I0,1))
* 1) Effettuare la stima ai minimi quadrati usando il metodo del gradiente (descritto nelle slides 10 e 26) e le funzioni A(x) e At(x). Usare almeno 150 iterazioni.
* 2) Effettuare la stima ai minimi quadrati usando il metodo del gradiente coniugato (funzione cgRecNE fornita, 20-30 iterazioni, lambda=0). Questo, a convergenza, produce lo stesso risultato del metodo del gradiente a convergenza, ma è molto più veloce.
* Confrontare i risultati dei metodi iterativi con la retroproiezione filtrata con due differenti filtri (Ram-Lak e Hann)
* Ripetere il tutto usando solo 40 proiezioni

**Per sviluppare l’algoritmo iterativo si devono calcolare:**

1) l’errore al passo k: , dove:

y valore misurato del sinogramma (nel nostro caso è il valore ottenuto mediante radon)

xk immagine al passo k

A trasformata di Radon

AT trasformata inversa di Radon

Si osservi che la componente dentro la parentesi è definita nel dominio del sinogramma; dopo l’applicazione di AT siamo nel dominio dell’immagine.

res = At(A(xk)) - At(y); % residuo nel dominio dell'immagine

2) il termine xk+1 e quindi il valore di  alla i-esima iterazione;



xk = xk - stepSize\*res;

stepSize = 5/25;

3) il numero di iterazioni può essere scelto sulla base della bontà della ricostruzione (es. visiva), oppure sulla base dell’errore calcolato al punto precedente. In ogni caso conviene mettere un controllo del tipo:

Obj = norm(y1(:)-y(:)); % distanza nel dominio del sinogramma

if Obj < prevObj

prevObj = Obj;

stepSize = st;

break;

end

Alcuni risultati attesi:

a) metodo del gradiente



Residuo con il metodo del gradiente Immagine di partenza

b) metodo del gradiente coniugato





Da sinistra immagine: fornita, con gradiente, con gradiente coniugato, con Iradon con filtro Ram-Lak e con filtro Hann



Da sinistra: immagine fornita, immagine residuo con gradiente, con gradiente coniugato, con Iradon con filtro Ram-Lak e con filtro Hann