Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

Algoritmi di Regolarizzazione con Variazione Totale nella Ricostruzione di Immagini di Tomosintesi

Tesi di Laurea in Analisi Numerica

Relatore:
Chiar.ma Prof.ssa
Elena Loli Piccolomini
Correlatore:
Dott.ssa Elena Morotti

Presentata da:
Lucia Traini

Sessione III
Anno Accademico 2012/2013

Io penso che non potrei più vivere
se non Lo sentissi più parlare
J. A. Moehler

A Federico e Paolo

Indice

Elenco delle figure

Introduzione

La Tomografia Computerizzata (CT) è una metodica di diagnostica per immagini che consiste in una particolare applicazione dei Raggi X. Essa, grazie ad una valutazione statistico-matematica (computerizzata) dell'assorbimento di tali raggi da parte delle strutture corporee esaminate, consente di ottenere immagini di sezioni assiali del corpo umano. Nonostante le dosi di radiazioni somministrate durante un esame siano contenute, questa tecnica rimane sconsigliata per pazienti sensibili, perciò nasce il tentativo della tomosintesi: ottenere lo stesso risultato della CT diminuendo le dosi somministrate. Infatti un esame di tomosintesi richiede poche proiezioni distribuite su un range angolare di appena 40°.
Se da una parte la possibilità di una ingente diminuzione di dosi e tempi di somministrazione costituisce un grosso vantaggio dal punto di vista medico-diagnostico, dal punto di vista matematico esso comporta un nuovo problema di ricostruzione di immagini: infatti non è banale ottenere risultati validi come per la CT con un numero basso di proiezioni, cosa che va a infierire sulla mal posizione del problema di ricostruzione.
Un possibile approccio al problema della ricostruzioni di immagini è considerarlo un problema inverso mal posto e studiare tecniche di regolarizzazione opportune. In questa tesi viene discussa la regolarizzazione tramite variazione totale: verranno presentati tre algoritmi che implementano questa tecnica in modi differenti. Tali algoritmi verranno mostrati dal punto di vista matematico, e valutati dal punto di vista qualitativo, attraverso alcune immagini ricostruite.
Lo scopo è quindi stabilire se ci sia un metodo più vantaggioso e se si possano ottenere buoni risultati in tempi brevi, condizione necessaria per una applicazione diffusa della to-

mosintesi a livello diagnostico. Per la ricostruzione si è fatto riferimento a problemi-test cui è stato aggiunto del rumore, così da conoscere l'immagine originale e poter vedere quanto la ricostruzione sia ad essa fedele.

Il lavoro principale dunque è stata la sperimentazione degli algoritmi su problemi test e la valutazione dei risultati, in particolare per gli algoritmi SGP e di Vogel, che in letteratura sono proposti per problemi di image deblurring.

Nel dettaglio, il primo capitolo presenta la tecnica di tomosintesi, i concetti che ne stanno alla base, le sue evoluzioni fino ad oggi e l'utilizzo attuale. Sarà fatto un breve accenno alla costruzione della matrice del sistema di tomosintesi e ai più noti metodi di ricostruzione. Nel secondo capitolo si trattano le basi matematiche di questo elaborato, ossia metodi di regolarizzazione e ottimizzazione e in particolare la regolarizzazione con variazione totale. Saranno analizzati i tre metodi trattati sia dal punto di vista algoritmico sia dal punto di vista della convergenza teorica alla soluzione esatta. Infine il terzo capitolo contiene parte delle ricostruzioni effettuate e tabelle di dati che facilitano il paragone tra i metodi, al fine di raggiungere lo scopo prefissato.

Capitolo 1

La Tomosintesi

La tomosintesi è una forma di tomografia ad angoli limitati che usa proiezioni bidimensionali per ricostruire immagini volumetriche.
Nel capitolo verranno visti i vantaggi e gli svantaggi di questa tecnica: essa viene applicata in radio-diagnostica e consente una riduzione del tempo d'esame e della dose di radiazioni somministrata al paziente; saranno inoltre valutate le differenze con gli approcci tradizionali in tale ambito.

1.1 Breve Cronologia

La tomosintesi viene studiata a partire dagli anni '70 per ovviare ad alcune lacune della tomografia tradizionale quali la confusione nell'immagine ricostruita, dovuta alla presenza di dettagli sfocati fuori dal piano di interesse, l'elevato tempo di acquisizione e l'eccessiva dose di radiazioni necessari per realizzare la ricostruzione volumetrica, che rendono questa tecnica poco adatta a soggetti sensibili e a parti del corpo delicate.
Lo sviluppo della tecnica di tomosintesi avviene negli anni '90, quando la tecnologia rende possibile la sostituzione del detector radiografico tradizionale (il sistema schermo/pellicola) con detector digitali come i *flat-panel*; questa modifica permette di acquisire i dati di trasmissione dei raggi X relativi a un numero discreto di proiezioni dell'intera regione

in studio con un solo movimento combinato del sistema emettitore/detector.
Se per la ricostruzione di un'immagine volumetrica con la tomografia computerizzata si devono acquisire proiezioni a 360° attorno al paziente e ciascuna sezione di interesse del volume da ricostruire viene calcolata sfruttando tali proiezioni, la tomosintesi è invece una forma di tomografia ad angoli limitati che produce sezioni (o *slice*) dell'immagine da una serie di proiezioni acquisite con un movimento dell'emettitore/ricevitore su un preciso percorso. Il range angolare totale si riduce al massimo a 40°, consentendo la diminuzione del tempo di esposizione ai raggi X.
L'idea di base è prendere varie proiezioni 2D dell'oggetto 3D a diverse angolazioni in modo che ogni proiezione fornisca una diversa informazione sull'oggetto in esame. La tomosintesi quindi riduce di molto il tempo e la quantità di esposizione ai raggi X da parte del paziente; inoltre in quanto tecnica di ricostruzione volumetrica diminuisce i casi di falsi negativi e falsi positivi tipici della valutazione di immagini tridimensionali con tecniche bidimensionali.

1.2 Apparato di Tomosintesi

Per capire bene il metodo della ricostruzione volumetrica in esame si deve prima di tutto studiare la geometria degli apparati utilizzati: un sistema di tomosintesi è formato da un emettitore di raggi X e un detector che si muovono in senso opposto attorno al paziente con un range angolare limitato. I movimenti che il sistema emettitore/detector svolge attorno al piano di fulcro possono essere di diverso tipo, uno di essi ha una rilevanza storica particolare perché venne presentato in un lavoro di Grant riconosciuto da quasi tutti gli esperti come la base della tomosintesi. Si tratta della geometria circolare, in cui l'emettitore ruota attorno all'oggetto rispetto all'asse z e la retta che idealmente connette emettitore e ricevitore ha inclinazione fissa rispetto allo stesso asse. Emettitore e detector realizzano moti circolari rispettivamente al di sopra ed al di sotto dell'oggetto di interesse, come illustra la figura 1.1:

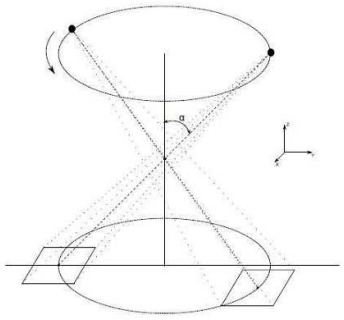

Figura 1.1: Geometria circolare

Un altro tipo di geometria è quello lineare, in letteratura se ne possono trovare tre diverse tipologie:

1) *Parallel Geometry*, nota anche come *Twinning Geometry*. In questo caso l'emettitore si muove su un piano parallelo al piano del detector. Quest'ultimo può essere fisso o mobile lungo il piano, in direzione opposta all'emettitore.

2) *Complete-Isocentric Geometry*, anche nota come *Grossman Geometry*. In questo schema detector ed emettitore sono connessi allo stesso braccio e si muovono contemporaneamente con traiettorie circolari attorno al paziente, naturalmente con movimenti opposti l'uno all'altro.

3) *Partial-Isocentric Geometry*. In quest'ultimo schema il detector o è fisso o si muove con moto lineare parallelo al piano dove giace, mentre l'emettitore si muove lungo un arco, quindi con una traiettoria circolare.

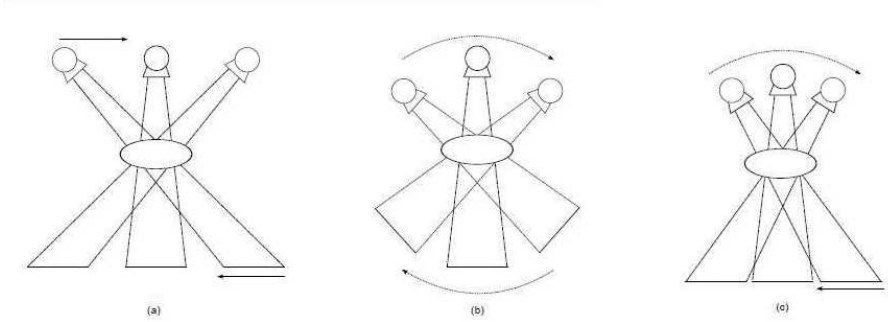

Figura 1.2: (a) Parallel Geometry; (b) Complete-Isocentric Geometry; (c) Partial-Isocentric Geometry

Per acquisire i dati dal paziente l'emettitore a raggi X manda un fascio di raggi attraverso il paziente fino a raggiungere il recevitore, quest'ultimo potrà così costruire il segnale che definisce la proiezione dell'oggetto per una determinata angolazione. Questo procedimento verrà poi attuato per tutte le angolazioni necessarie fino ad ottenere un set di proiezioni utili per la ricostruzione dell'oggetto 3D.

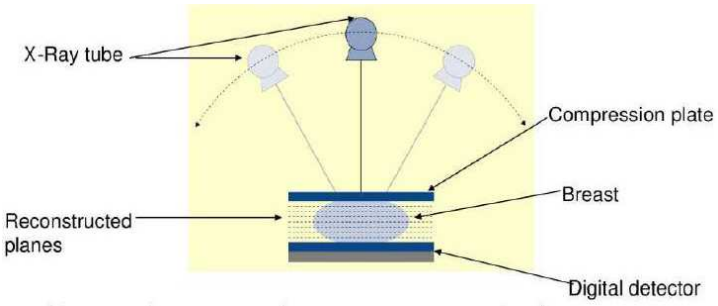

Figura 1.3: Movimento tubo-detector

In tomografia esistono due modi di lavorare con oggetti 3D:

- si esamina una *slice* per volta. In questo caso il fascio di Raggi X è piatto e il detector è formato da una striscia di ricevitori che acquisiscono il segnale. La proiezione che si ottiene è in questo modo una funzione unidimensionale rappresentante la sezione presa in considerazione. Viene poi eseguita la proiezione unidimensionale della 'fetta' per ogni angolazione, il tutto ripetuto per ogni *slice*.

- Il secondo è un metodo proprio della tomosintesi e prevede la considerazione dell'intero oggetto 3D. In questo caso viene utilizzato un fascio tridimensionale di raggi, il ricevitore è un piano e la proiezione che si ottiene è una funzione a due variabili. Basta a questo punto ripetere l'acquisizione del segnale 2D per ogni angolazione desiderata ottenendo così il set di proiezioni di tutto l'oggetto.

In entrambi i casi la proiezione è funzione della densità dei tessuti. L'intensità del segnale è proporzionale all'assorbimento del raggio da parte del tessuto: un tessuto più denso assorbe maggiormente il raggio e di conseguenza in quel punto l'immagine apparirà meno intensa. Ogni punto del segnale è la somma dei valori di assorbimento lungo un singolo raggio del fascio che corrisponde spazialmente a quel punto. Utilizzando una sola proiezione non si riesce a determinare ogni dettaglio dell'oggetto, per questo si ripete il procedimento per più angolazioni, naturalmente più angoli si utilizzano maggiore sarà la precisione, ma in questo modo crescerà anche il tempo di esposizione alle radiazioni e la quantità di radiazioni assorbita.

1.3 Applicazioni in mammografia

Figura 1.4: Sistema per tomosintesi presente all'Ospedale Maggiore di Bologna

Un possibile utilizzo della tomosintesi è dato dalla prevenzione dei tumori al seno: in questo ambito la mammografia classica spesso incorre in falsi positivi, dati per esempio dal sovrapporsi di più tessuti normali, o in falsi negativi, dovuti ad una massa tumorale nascosta da tessuti ghiandolari densi.
La tomosintesi digitale della mammella (*Digital Breast Tomosynthesis* o *DBT*) permette di ricostruire immagini volumetriche della mammella a partire da poche proiezioni bidimensionali a bassa dose ottenute con angolazioni diverse del tubo radiogeno. La ricostruzione volumetrica, in linea di principio, consente di superare uno dei limiti principali dell'imaging bidimensionale, ovvero il mascheramento di lesioni (come masse e microcalcificazioni) causato dalla sovrapposizione di strutture normali; è proprio l'opportunità di dissociare piani diversi che rende possibile una riduzione del numero di falsi negativi e di falsi positivi dovuti alla sovrapposizione.

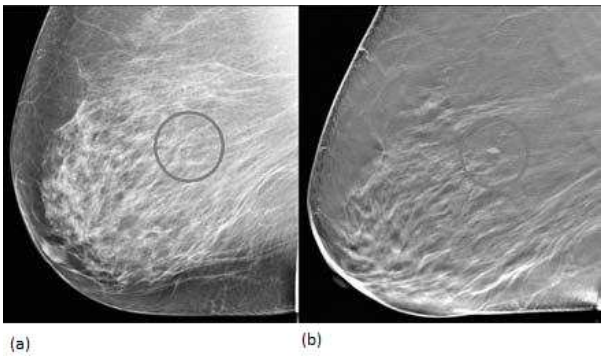

Figura 1.5: Confornto tra mammografia (a) e tomosintesi mammaria (b): il rumore apportato dalla sommazione di ombre di strutture ubicate nei diversi piani della mammella maschera nell'immagine mammografica la piccola formazione nodulare presente nell'area cerchiata, bene evidenziata nell'immagine tomografica.

È intuitivo che l'elevato spessore può produrre su una superficie un effetto di elevata densità anche se tra le strutture ghiandolari esistono piani di grasso più o meno estesi, in presenza di tali condizioni consegue la mancata o scadente visualizzazione e la non percezione delle lesioni espansive (masse e distorsioni) e delle calcificazioni. In virtù di una nitida rappresentazione in assenza di sovrapposizioni, la tomosintesi è in grado di rendere visibili e meglio analizzabili nelle forme, nei contorni, nella disposizione e nel numero le lesioni non rappresentate o mal rappresentate dalla mammografia tradiziona-

le. La DBT permette un sostanziale miglioramento nel rilevamento e nell'analisi delle lesioni, influendo nella certezza tanto della loro presenza quanto della loro assenza.
Tuttavia la DBT non è in grado, allo stato attuale, di definire con sufficiente accuratezza il quadrante nel quale la lesione è situata: il rilevamento tomografico consente di apprezzare solo isolatamente le lesioni che più soffrono della sovrapposizione e quindi della confusione dei piani nella mammografia standard. Alcuni centri in Italia utilizzano questa tecnica per la rilevazione di eventuali noduli al seno, essa però ancora non è un'alternativa alla mammografia ma solo un esame da affiancare a quest'ultima; se anche i primi risultati comparativi dimostreranno la non inferiorità della tomosintesi rispetto alle tecniche classiche, la DBT potrà sostituire le tecniche attualmente in uso.

1.4 Algoritmi di Ricostruzione

Inizialmente la ricostruzione delle immagini di tomosintesi avveniva con la tecnica detta *Shift And Add*, basata sul fatto che oggetti ad altezze differenti all'interno del volume da ricostruire, al muoversi dell'emettitore vengono proiettati sul detector in posizioni che dipendono dalle altezze tra i piani; dunque la tecnica consiste nello spostare ogni proiezione acquisita di una quantità che dipende solamente dall'altezza del piano (fase di Shift) e fare poi una media fra le nuove proiezioni ottenute (fase di Add). Questo metodo, molto vantaggioso dal punto di vista computazionale, risente del fatto che le sezioni ricostruite contengono, oltre ai dettagli del piano di interesse, anche dei contributi degli altri piani, che appaiono come rumore. Nel tempo si è cercata una soluzione a tali limitazioni, in particolare alla fine degli anni '60 Edholm e Quiding proposero la rimozione dello sfocamento in tomografia mediante l'applicazione di filtri *passa-alto* al tomogramma originale.
Un'altra tecnica per eliminare lo sfocamento è introdotta nel 1985 da Ghosh Roy: si prevede la conoscenza delle funzioni che regolano lo sfocamento per togliere completamente la distorsione generata dai piani adiacenti a quello di messa a fuoco; è una tecnica computazionalmente molto veloce ed efficace ma molto suscettibile all'amplificazione del rumore a frequenze molto basse. Nel 1984 Ruttimann introduce un metodo in cui le

stime dello sfocamento tomografico sono calcolate coinvolgendo un numero limitato di piani con le loro funzioni di sfocamento, in questo modo la tecnica non viene influenzata dall'amplificazione del rumore a basse frequenze ma richiede un alto costo computazionale.
Altri metodi per ridurre lo sfocamento nella tomosintesi fanno uso delle *wavelets*, forme d'onda oscillanti che permettono di mettere in evidenza oggetti piccoli ed ad alto contrasto.

1.4.1 Metodi diretti: FBP

La retroproiezione è uno dei metodi in uso per la ricostruzione di oggetti 3D a partire da loro proiezioni; le basi matematiche di questo metodo vengono attribuite al matematico viennese Johann Radon e alla sua *Trasformata di Radon* (1917), che permette di proiettare un oggetto 2D lungo raggi paralleli come parte del lavoro dell'oggetto stesso sugli integrali di linea.
La *Filterd Back Projection* o FBP nasce per la classica Tomografia Computerizzata e il fatto che le proiezioni siano prese in un range angolare completo attorno al paziente è rilevante per la ricostruzione dello spettro dell'oggetto considerato e penalizza l'applicazione di questo metodo in tomosintesi, dove le proiezioni acquisite non forniscono informazioni sufficienti per una ricostruzione corretta. Nelle immagini seguenti possiamo confrontare il dominio delle frequenze relativo alla ricostruzione con TC e quello relativo alla tomosintesi: in quest'ultimo caso l'informazione in frequenza in direzione z è carente e ciò comporta una bassa risoluzione sulla profondità dell'oggetto ricostruito. Tale fenomeno si attenua all'aumentare del range angolare delle proiezioni, fino a sparire quando esso diventa completo (caso TC).

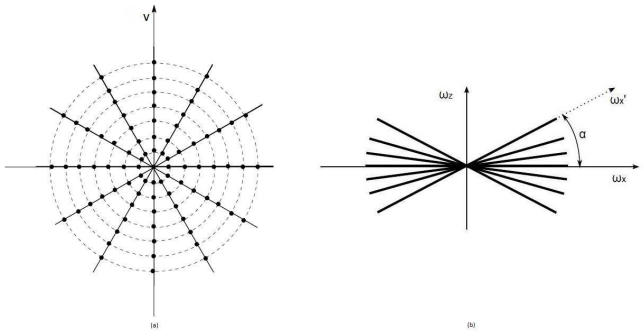

Figura 1.6: Spettro di Fourier per ricostruzione con TC (a) e con tomosintesi (b)

Nonostante la limitazione detta questo metodo è tuttora molto studiato per l'applicazione in tomosintesi.

1.4.2 Metodi Iterativi

Un altro possibile approccio è considerare la ricostruzione di immagini di tomosintesi come un problema mal-posto:

- la mal-posizione nasce dal basso numero di proiezioni utilizzate e dalla mal-posizione dell'operatore di proiezione;
- a causa del basso numero di proiezioni la ricostruzione delle immagini consiste nella risoluzione di un sitema lineare indeterminato (con infinite soluzioni possibili in quanto il sistema ottenuto ha meno equazioni e più incognite).

Tra i metodi numerici di risoluzione dei sistemi lineari mal-posti si considerano in particolare i metodi iterativi; in questo ambito gli algoritmi di ricostruzione iterativi si basano sulla risoluzione di sistemi lineari le cui incognite rappresentano i coefficienti di attenuazione dei voxel del volume.
La figura mostra lo schema di Partial-Isocentric Geometry da cui si ricava il sistema:

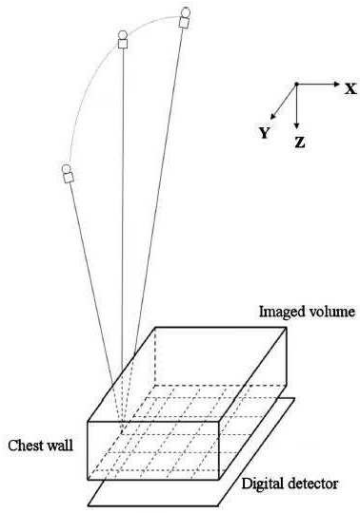

Figura 1.7: Geometria a rilevatore fisso e sorgente in movimento circolare su un range di angoli limitato

- L'oggetto in esame viene diviso in J voxel;
- il coefficiente di attenuazione del j-esimo voxel è indicato con x_j, $1 \leq j \leq J$;
- il detector è suddiviso in una griglia di I pixel;
- i raggi X sono costituiti da un segmento di estremi la sorgente degli stessi e il centro di ogni pixel;
- per ogni angolazione si hanno tanti raggi quanti sono i pixel del detector.

Ad esempio, se il detector è monodimensionale, esso si presenta come un array di 512 rilevatori o *detector-bins* ed è grande abbastanza perché il campo visivo sia la circonferenza inscritta nella matrice 256×256 dell'immagine.
Le misure della TC possono essere correlate al cammino integrale del coefficiente di attenuazione dei raggi X lungo i raggi definiti dalla sorgente e da un singolo *detector-bin* del detector (si veda la figura (1.7)). Nel discreto gli integrali possono essere scritti come somme pesate sui pixel attraversati da ogni raggio, per modellare le proiezioni di un'immagine discreta si valutano i pesi calcolando l'intersezione del raggio i-esimo col pixel

j-esimo, come in figura (1.8).

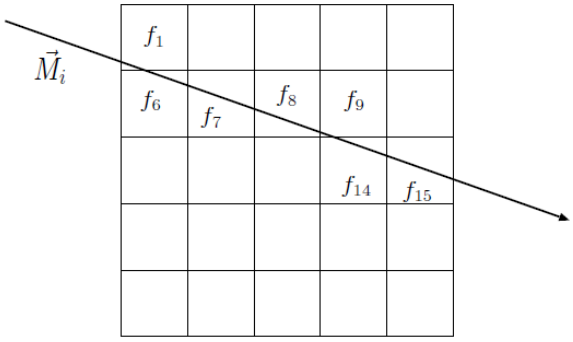

Figura 1.8: Intersezione del raggio i-esimo col pixel j-esimo: in questo esempio 5×5 il peso $M_{i,j}$ è non nullo solo nei pixel f_1, f_6, f_7, f_8, f_9, f_{14}, f_{15} perché sono i soli ad intersecare il raggio.

La matrice A dei coefficienti del sistema è data proprio dall'intersezione dei raggi con i pixel, dunque considerando N proiezioni, ossia N angoli presi in un certo range limitato, per ogni angolo n il modello di ricostruzione è dato dal sistema:

$$A_n x = y_n \tag{1.1}$$

dove y_n rappresenta la i-esima proiezione, A_n è la matrice di proiezione per l'n-esimo angolo e $A_{ij,n}$ rappresenta il contributo della lunghezza del tratto del raggio i-esimo che interseca il voxel j-esimo. Dette $I_{0,n}$ l'intensità del raggio incidente e $I_{i,n}$ l'intensità del raggio trasmesso si ha la seguente relazione tra il valore della i-esima proiezione e le intensità:

$$y_{i,n} = k \ln \left(\frac{I_{0,n}}{I_{i,n}} \right) \tag{1.2}$$

Dunque si ottiene il sistema lineare che modella la ricostruzione di immagini di tomosintesi:

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_N \end{bmatrix} x = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \longrightarrow Ax = y \tag{1.3}$$

Il sistema 1.3 è la base della tecnica di ricostruzione dell'immagine in tomosintesi. In condizioni ideali il sistema può essere risolto esattamente invertendo la matrice A; in realtà quest'approccio è irrealizzabile a causa dell'enorme dimensione del sistema. A questo proposito entrano in gioco i metodi iterativi: l'idea è quella di migliorare il risultato ad ogni iterazione, aggiornando ad ogni passo i coefficienti di attenuazione lineare dei voxel e cercando di minimizzare l'errore fra le proiezioni calcolate e quelle reali. Al termine del processo si avrà una soluzione approssimata del problema.
In letteratura possiamo trovare numerosi metodi che si differenziano per l'approccio matematico utilizzato per incrementare le incognite. Il primo di questi algoritmi ad essere implementato è chiamato *Algebraic Reconstruction technique* (*ART*): i coefficienti di attenuazione sono aggiornati considerando singolarmente ogni equazione del sistema 1.3; in pratica l'aggiornamento avviene raggio per raggio. *ART* converge velocemente ma le ricostruzioni risentono della presenza di rumore *salt and pepper*, soprattutto nel presente caso, in cui il problema è mal condizionato.
Per risolvere tale inconveniente sono stati proposti altri due metodi di tipo algebrico, il *SIRT* (*Simultaneous Iterative Recostruction Technique*) ed il *SART* (*Simultaneous* (*Superior*) *Algebraic Reconstruction Technique*). L'idea di fondo è, in entrambi i casi, quella di *ART*, cambia solo il metodo di aggiornamento dei valori dell'approssimazione di x. Nel primo l'aggionamento del volume approssimato avviene solo dopo aver preso in considerazione il contributo di tutti i raggi X in tutte le proiezioni. Si ha dunque che in ogni singola iterazione *SIRT* prevede un'unica fase di aggiornamento dei coefficienti. *ART* invece prevede, ad ogni iterazione, un numero di aggiornamenti pari al numero complessivo di raggi X, ovvero $I \times N$ con I numero di pixel e N numero di proiezioni. *SIRT* produce risultati migliori dell'*ART*, ma converge con un numero maggiore di iterazioni.

Questo aspetto lo rende difficilmente utilizzabile per la tomosintesi perché le proiezioni vengono acquisite ad alta risoluzione, provocando un elevato tempo di esecuzione per ogni iterazione.
Per questo motivo si preferiscono metodi che permettono di arrivare ad una soluzione con un numero basso di iterazioni, in quanto aumentando anche di poco questo numero, aumenta di molto il tempo di esecuzione: *SART* rappresenta un compromesso tra *ART* e *SIRT*. L'aggiornamento dei coeffiecienti avviene sfruttando il contributo di tutti i raggi X di un'intera proiezione. In un'iterazione vengono quindi effettuati N aggiornamenti.

Capitolo 2

Regolarizzazione mediante Variazione Totale

In questo capitolo si affronta l'aspetto matematico e formale della ricostruzione di immagini: verranno presentati i problemi mal posti e i metodi di regolarizzazione più adatti alla loro risoluzione. Si entrerà poi nello specifico dei metodi che usano il funzionale di Variazione Totale e in particolare srarnno esposti i tre algoritmi testati per questo lavoro.

2.1 Problemi Inversi e Mal Posti

Il problema della ricostruzione di immagini appartiene, dal punto di vista matematico, alla classe dei problemi inversi.
Di questa tipologia di problemi non esiste una definizione precisa come invece accade per altre strutture matematiche, per capire di cosa si tratta è necessario fare riferimento ad alcuni esempi e considerare che si presuppone l'esistenza di un altro problema, detto diretto, al quale il problema inverso è strettamente correlato: si dice che due problemi sono uno l'inverso dell'altro quando la formulazione del primo coinvolge necessariamente il secondo. Di questa coppia di problemi viene chiamato *problema diretto* quello che è stato studiato più nel dettaglio e per primo, mentre viene detto *problema inverso* quello

meno considerato o studiato solo in tempi recenti. Ad esempio è di facile risoluzione il problema di determinare il prodotto di due numeri interi assegnati; l'inverso di questo problema (dato un numero intero, trovare una coppia di fattori di cui esso sia il prodotto) si presenta invece più complicato, anche perché non è detto che abbia una soluzione unica. Per garantire l'unicità della soluzione bisognerebbe restringersi alla classe dei numeri con fattorizzazione unica, i numeri primi, complicando la situazione.
I problemi diretti sono problemi nei quali si forniscono sufficienti informazioni per poter avviare un procedimento ben definito e stabile che porta ad una unica soluzione:

$$\text{informazioni} \xrightarrow{procedimento} \text{soluzione}$$
$$(2,3) \xrightarrow{moltiplicazione} 6$$

Se il processo descrive un fenomeno fisico, o comunque del mondo reale, si può schematizzare il problema diretto come:

$$\text{causa} \rightarrow \text{modello} \rightarrow \text{effetto}$$
$$x \longrightarrow K \longrightarrow y$$

ossia

$$Kx = y \tag{2.1}$$

Il problema diretto consiste nell'assegnare la causa x e il modello K e calcolare l'effetto y, ma questo è solo uno dei tre modi nei quali si può leggere l'equazione (2.1): ogni problema diretto suggerisce immediatamente due problemi inversi:

1) dati il modello K e l'effetto y, risalire alla causa x;

2) dati la causa x e l'effetto y, costruire un modello K.

Queste ultime due letture dell'equazione (2.1) corrispondono a due tipi di problemi inversi e dal punto di vista applicativo sono studiati, rispettivamente, per due ragioni:

1) conoscere lo stato passato o i parametri che regolano un sistema (es: diagnosi mediche, ricostruzione di immagini)

2) controllare lo stato finale del sistema modificando lo stato presente o i parametri del modello (es: produzione industriale di manufatti)

Il problema della ricostruzione di immagini corrisponde al primo caso e può essere schematizzato dalla seguente equazione:

$$\int_{\Omega} input \times sistema \; d\Omega = output \tag{2.2}$$

Esso ha come scopo la determinazione dell'oggetto da scansionare (*input*) conoscendo

- la misura della densità dei raggi X che arrivano in ogni punto del ricevitore (*output*)
- il modello matematico che descrive l'apparecchio di emissione ed acquisizione dei raggi X (*sistema*)

Dunque per i problemi originati dalle applicazioni si può dire che si affronta un problema inverso quando si cercano le cause di un determinato effetto osservato o desiderato.
Dal punto di vista puramente matematico esiste una ulteriore e decisiva distinzione tra problema diretto e inverso: il problema diretto gode di certe buone proprietà che corrispondono alla definizione di problema *ben posto*, mentre il problema inverso è solitamente *mal posto*.

2.1.1 I Problemi Mal Posti

La forma matematica che meglio modellizza la (2.2) è l'*Equazione Integrale di Fredholm di Prima Specie*:

$$\int_{\Omega} K(s,t) f(t) \; dt = g(s) \tag{2.3}$$

dove:

- $\mathcal{L}^2(\Omega)$ è uno spazio di Hilbert di funzioni reali;
- $K \in \mathcal{L}^2(\Omega \times \Omega)$, è detto *nucleo* dell'equazione ed è noto esattamente attraverso un modello matematico;
- $f \in \mathcal{L}^2(\Omega)$;
- $g \in \mathcal{L}^2(\Omega)$, solitamente è una quantità misurata ed è nota solo in alcuni punti, per questo è necessario parlare di discretizzazione del problema.

Il problema (2.3) discretizzato viene indicato con:

$$K\mathbf{f} = \mathbf{g} \tag{2.4}$$

dove $\mathbf{f}$ e $\mathbf{g}$ sono vettori di dimensione N e K è un operatore compatto.
I prblemi con questo tipo di mal posizione vengono risolti attraverso *metodi di regolarizzazione*, il cui scopo è evitare le soluzioni fortemente instabili, anche a scapito della precisione.

Notazioni

- $\Omega \subseteq \mathbb{R}^n$ insieme non vuoto, semplicemente connesso e aperto con bordo continuo e Lipschitziano indicato con $\partial\Omega$;
- $C^1(\Omega)$ spazio delle funzioni $f : \Omega \longrightarrow \mathbb{R}$ per cui le derivate parziali $\dfrac{\partial f}{\partial x_i}$ esistono e sono continue;
- $\mathcal{H}$ spazio di Hilbert reale con prodotto interno $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ e norma indotta $\|\cdot\|_{\mathcal{H}}$, se il contesto è chiaro viene omessa la $\mathcal{H}$ a pedice;
- la convergenza si intende in senso forte:

$$f_n \longrightarrow f$$

significa che

$$\lim_{n\longrightarrow\infty}\|f_n - f\| = 0$$

- dato un insieme S, $S^{\perp}$ indica il suo complementare ortogonale, vale a dire l'insieme degli elementi $f \in \mathcal{H}$ tali che $\langle f, s\rangle = 0 \ \forall \ s \in S$;
- dati due insiemi S e T si intende $S+T = \{s+t | s \in S, t \in T\}$, $s+T$ è equivalente a $\{s\}+T$;
- $\mathbb{R}^n_+ = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n | x_i \geq 0 \ \forall \ i\}$
- si scrive

$$f(\alpha) = o(g(\alpha))_{\alpha\longrightarrow\alpha_*} \Longleftrightarrow \lim_{\alpha\longrightarrow\alpha_*}\frac{f(\alpha)}{g(\alpha)} = 0 \tag{2.5}$$

- La delta di Kronecker è definita da

$$\delta_{i,j} = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

Operatori e Spazi di Hilbert

Sia $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ un operatore, si indica con Range(A) la sua immagine in $\mathcal{H}_2$. Si dice che A è continuo se $A(f_n) \longrightarrow A(f_*)$ per $f_n \longrightarrow f_*$; nel caso in cui A sia lineare si sostituisce ad $A(f)$ la scrittura meno pesante Af.
Si definisce il *Null Space* di A: Null$(A) = \{f \in \mathbb{H}_1 | Af = 0\}$.
A è limitato se e solo se la norma indotta dall'operatore

$$\|A\| \stackrel{def}{=} \sup_{\|f\|_{\mathcal{H}_1}=1}\|Af\|_{\mathcal{H}_2}$$

è finita. Operatori lineari e limitati sono anche continui; lo spazio degli operatori lineari limitati da $\mathcal{H}_1$ a $\mathcal{H}_2$ si indica con $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$.
L'aggiunto di un operatore A è $A^* \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$ tale che

$$\langle Af, g\rangle_{\mathcal{H}_2} = \langle f, A^*g\rangle_{\mathcal{H}_1} \tag{2.6}$$

dove $f \in \mathcal{H}_1$ e $g \in \mathcal{H}_2$.
A si dice autoaggiunto se $A = A^*$ (e dunque $\mathcal{H}_1 = \mathcal{H}_2 =: \mathcal{H}$), in questo caso

$$\lambda_{min}(A) \stackrel{def}{=} \inf_{\|f\|_{\mathcal{H}}=1} \langle Af, f\rangle_{\mathcal{H}} \tag{2.7}$$

$$\lambda_{max}(A) \stackrel{def}{=} \sup_{\|f\|_{\mathcal{H}}=1} \langle Af, f\rangle_{\mathcal{H}} \tag{2.8}$$

sono numeri reali finiti. A è semidefinito positivo se $\lambda_{min}(A) \geq 0$ e definito positivo se $\langle Af, f\rangle_{\mathcal{H}} > 0 \ \ \forall\ f \neq 0$; si dice che A è coercivo se $\lambda_{min}(A) > 0$.
Se A è autoaggiunto allora $\|A\| = \max\{|\lambda_{min}(A)|, |\lambda_{max}(A)|\}$; in generale invece $\|A\| = \sqrt{\lambda_{max}(A^*A)}$.

Osservazione 1. Sia $\mathcal{M}(\mathbb{R}^{m\times n})$ l'insieme delle matrici di dimensione $m \times n$ a valori reali. Esso è uno spazio di Hilbert con il prodotto di Frobenius:

$$\langle A, B\rangle_{Fro} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}b_{i,j} \tag{2.9}$$

La norma indotta è quella di Frobenius:

$$\|A\|_{Fro} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}^2} \tag{2.10}$$

Migliore approssimazione negli spazi di Hilbert

Sia $f \in \mathcal{H}$ e sia $\mathcal{S} \subseteq \mathcal{H}$. Si dice che s_* è la migliore approssimazione per f in $\mathcal{S}$ se

$$s_* = \arg\min_{s\in\mathcal{S}} \|s - f\|_{\mathcal{H}} \tag{2.11}$$

vale a dire: $s_* \in \mathcal{S}$ e $\|s_* - f\|_{\mathcal{H}} \leq \|s - f\|_{\mathcal{H}} \ \ \forall\ s \in \mathcal{S}$. Se esiste un tale s_*, esso è unico; l'esistenza è garantita solo se $\mathcal{S}$ è chiuso in $\mathcal{H}$, cioè se $\mathcal{S}$ è un sottospazio di $\mathcal{H}$ di dimensione finita.

Teorema 2.1.1. *Se s_* è la migliore approssimazione di f in un sottospazio $\mathcal{S}$, allora*

$$\langle s_* - f, s\rangle_{\mathcal{H}} = 0 \qquad \forall\ s \in \mathcal{S} \tag{2.12}$$

Se $\mathcal{S}$ ha dimensione finita e una sua base è $(\phi_1, \ldots, \phi_n)$, dal teorema precedente viene la seguente formula per il calcolo della migliore approssimazione:

$$s_* = \sum_{j=1}^{n} \hat{a}_j \phi_j$$

dove il vettore dei coefficienti $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_n)$ risolve il sistema lineare $G\hat{a} = b$ con

$$[G]_{i,j} = \langle \phi_j, \phi_i \rangle_{\mathcal{H}}$$
$$[b]_i = \langle f, \phi_i \rangle \mathcal{H}$$

2.1.2 Mal Posizione per Operatori e Regolarizzazione

Definizione 2.1. Sia $K : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$; l'equazione

$$K(f) = g \tag{2.13}$$

si dice ben posta se

(i) $\forall\ g \in \mathcal{H}_2\ \exists\ f \in \mathcal{H}_1$, detta *soluzione*, per cui valga la (2.13);

(ii) la soluzione f è unica;

(iii) la soluzione f è stabile rispetto alle perturbazioni in g. Questo significa che se $Kf_* = g_*$ e $Kf = g$ allora $f \longrightarrow f_*$ se $g \longrightarrow g_*$

Un'equazione non ben posta è detta mal posta.
Se la (2.13) è ben posta, allora K ha un operatore inverso, continuo e ben definito K^{-1} tale che $K^{-1}(K(f)) = f\ \ \forall\ f \in \mathcal{H}_1$; chiaramente in questo caso $\text{Range}(K) = \mathcal{H}_2$

Operatori Compatti, Sistemi Singolari, SVD

Nel caso trattato l'operatore coinvolto nel problema è un operatore compatto, vale a dire un operatore $K : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ per cui la chiusura dell'immagine è compatta in $\mathcal{H}_2$.

Teorema 2.1.2 (Mal posizione per operatori compatti)**.**
Siano $\mathcal{H}_1$, $\mathcal{H}_2$ spazi di Hilbert di dimensione infinita e sia $K : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ un operatore compatto lineare. Se Range(K) ha dimensione infinita allora la (2.13) è mal posta in quanto sono violate le condizioni (i) e (iii) della definizione 2.1. In questo caso Range(K) non è chiuso. Se invece Range(K) ha dimensione finita è violata la condizione (ii).

Se K è compatto, K^*K è compatto e autoaggiunto, dunque esistono autovalori positivi e un corrispondente insieme di autovettori ortonormali che formano una base per $\text{Null}(K^*K)^\perp = \text{Null}(K)^\perp$. Da questa scomposizione si costruisce un *sistema singolare.*

Definizione 2.2 (Sistema singolare)**.**
Un sistema singolare per un operatore lineare compatto $K : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ è un insieme numerabile di terne $\{u_j, s_j, v_j, \}_j$ con le seguenti proprietà:

- I vettori singolari destri v_j formano una base ortonormale per $\text{Null}(K)^\perp$;
- i vettori singolari sinistri u_j formano una base ortonormale per la chiusura di Range(K);
- i valori singolari s_j sono numeri reali positivi in ordine non crescente:

$$s_1 \geq s_2 \geq \ldots > 0;$$

- $\forall\ j$ vale $Kv_j = s_j u_j$;
- $\forall\ j$ vale $K^* u_j = s_j v_j$;
- se Range(K) ha dimensione infinita, allora

$$\lim_{j \longrightarrow \infty} s_j = 0$$

Sia ora K una matrice $m \times n$, K può essere visto come un operatore compatto $K : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ con $\mathcal{H}_1 = \mathbb{R}^n$ e $\mathcal{H}_2 = \mathbb{R}^m$; questo operatore ammette un sistema singolare $\{\mathbf{u}_j, s_j, \mathbf{v}_j, \}_{j=1}^r$ con $r = rank(K)$.
Siano U_r la matrice di dimensione $m \times r$ e di colonne $\mathbf{u}_j$ e V_r la matrice di dimensione $n \times r$ e di colonne $\mathbf{v}_j$, se $r < n$ allora Null(K) è un sottospazio non banale di $\mathbb{R}^n$ ed è possibile costruire una matrice V_0 di dimensione $n \times (n - r)$ le cui colonne siano una base ortonormale per Null(K). Analogamente, se $r < m$, Range$(K)^\perp$ è un sottospazio non banale di $\mathbb{R}^m$ e si può costruire una matrice $U_\perp$ di dimensione $m \times (m - r)$ le cui colonne formino una base ortonormale per Range$(K)^\perp$.
La SVD (Singular Value Decomposition) di K è la scomposizione

$$K = UDV^T \tag{2.14}$$

dove

$$U = [U_r U_\perp], \quad V = [V_r V_0], \quad D = \begin{bmatrix} \text{diag}(s_1, \ldots, s_r) & 0 \\ 0 & 0 \end{bmatrix} \tag{2.15}$$

Soluzioni ai Minimi Quadrati e Pseudo-Inversa

Sia K un operatore compatto con sistema singolare $\{u_j, s_j, v_j, \}_j$, allora K ammette la seguente rappresentazione:

$$Kf = \sum_j s_j \langle f, v_j \rangle_{\mathcal{H}_1} u_j \tag{2.16}$$

e dato $g \in$ Range(K) è possibile costruire il vettore

$$K^\dagger g = \sum_j \frac{\langle g, u_j \rangle_{\mathcal{H}_2}}{s_j} v_j \tag{2.17}$$

che appartiene allo spazio $\text{Null}(K)^\perp$ e per cui vale $K(K^\dagger g)$. Questo operatore è chiamato *pseudo-inversa* di K ed è definito sugli spazi

$$K^\dagger : \text{Range}(K) + \text{Range}(K)^\perp \longrightarrow \text{Null}(K)^\perp$$

Se K è una matrice si applica la SVD:

$$K^\dagger = V D^\dagger U^T$$

dove

$$[D^\dagger]_{i,j} = \begin{cases} \dfrac{1}{s_j} & \text{se } i = j \text{ e } 1 \leq i \leq r \\ 0 & \text{altrimenti} \end{cases}$$

La pseudo-inversa di un operatore può essere caratterizzata in un altro modo, usando le soluzioni ai minimi quadrati:

Definizione 2.3 (Soluzione ai minimi quadrati)**.**
Sia $K : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ un operatore limitato e lineare. Una soluzione ai minimi quadrati di $Kf = g$ è f_{ls} tale che:

$$\|Kf_{ls} - g\|_{\mathcal{H}_2} \leq \|Kf - g\|_{\mathcal{H}_2} \qquad \forall\ f \in \mathcal{H}_1 \tag{2.18}$$

Non necessariamente esiste una soluzione ai minimi quadrati, se esiste l'insieme di tutte le possibili soluzioni ai minimi quadrati è dato dal sottospazio affine $f_{ls} + \text{Null}(K)$ e tra tutte quella di minima norma è data da

$$f_{ls_{mn}} = \arg \min_{f \in f_{ls} + \text{Null}(K)} \|f\|_{\mathcal{H}_1} \tag{2.19}$$

Teorema 2.1.3. *Se $g \in Range(K) + Range(K)^\perp$ allora*

$$f_{ls_{mn}} = K^\dagger g$$

$K^\dagger$ è limitato se e solo se $Range(K)$ è chiuso. In questo caso $f_{ls_{mn}}$ esiste per ogni $g \in \mathcal{H}_2$.

2.2 Regolarizzazione

Si considera l'equazione (2.13), $Kf = g$, e si assume che esista un operatore R_* che porti ogni $g \in \text{Range}((K)$ in un unico $R_*(g) \in \mathcal{H}_1$ tale che $K(R_*(g)) = g$, se K è lineare generalmente si pone $R_* = K^T$. Si considera una famiglia di operatori regolarizzanti $R_\alpha : \mathcal{H}_2 \longrightarrow \mathcal{H}_1$, dove $\alpha \in I$ è detto *parametro di regolarizzazione* (I insieme di indici).

Definizione 2.4 (Schema di regolarizzazione)**.**
$\{R_\alpha\}_{\alpha \in I}$ è uno schema di regolarizzazione che converge a R_* se

(i) $\forall\ \alpha \in I$, R_α è un operatore continuo;

(ii) dato $g \in \text{Range}(K)$, per ogni successione $\{g_n\} \subset \mathcal{H}_2$ che converge a g è possibile estrarre una successione $\{\alpha_n\} \subset I$ tale che $R_{\alpha_n}(g_n) \longrightarrow R_*(g)$ per $n \longrightarrow \infty$

Si dice che $\{R_\alpha\}_{\alpha \in I}$ è lineare se lo è ogni suo elemento.

Di particolare interesse sono gli schemi di regolarizzazione con rappresentazione filtrata

$$R_\alpha(g) = \sum_j \frac{w_\alpha s_j^2}{s_j} \langle g, u_j \rangle_{\mathcal{H}_2} v_j \tag{2.20}$$

che converge a $R_* = K^\dagger$. Esempi noti di filtri di regolarizzazione w_α sono il filtro TSVD e quello di Tikhonov, per i quali $\alpha \in I = (0, \infty)$.
A partire dalla (2.20) si può dimostrare che

$$\|R_\alpha\| = \sup_j \frac{w_\alpha(s_j^2)}{s_j} \tag{2.21}$$

Sia $g \in \text{Range}(K)$, se $g_n \in \mathcal{H}_2$ e $\delta_n > 0$ sono tali che

$$\|g_n - g\| \leq \delta_n \tag{2.22}$$

allora per la disuguaglianza triangolare e per la (2.21) vale:

$$\|R_{\alpha_n} g_n - R_* g\| \leq \|R_{\alpha_n} g - R_* g\| + \|R_{\alpha_n}\| \delta_n \tag{2.23}$$

Il prossimo teorema garantisce l'esistenza di un $\alpha = \alpha(\delta)$ per cui i due termini della parte destra della (2.23) convergono a 0 per $\delta_n \longrightarrow 0$. Si indica con α_* il valore limite del parametro di regolarizzazione per cui il limite della (2.4) è R_*.

Teorema 2.2.1 (Convergenza per schemi di regolarizzazione)**.**
Si assuma che per ogni $\alpha \in I$, $\sup_{s>0} \left| \dfrac{w_\alpha(s^2)}{s} \right| < \infty$ *e che* $\forall\ s > 0$

$$\lim_{\alpha \longrightarrow \alpha_*} w_\alpha(s^2) = 1 \tag{2.24}$$

Si assuma anche che esista una funzione $\alpha = \alpha(\delta) : \mathbb{R}_+ \longrightarrow I$ *tale che*

$$\lim_{\delta \longrightarrow 0} \alpha(\delta) = \alpha_* \tag{2.25}$$

$$\lim_{\delta \longrightarrow 0} \|R_{\alpha(\delta)}\| \delta = 0 \tag{2.26}$$

Allora la (2.20) definisce uno schema di regolarizzazione che converge a $K^\dagger$.

2.3 Ottimizzazione

Sia $J : \mathcal{H} \longrightarrow \mathbb{R}$ e sia $\mathcal{C}$ un sottoinsieme di $\mathcal{H}$, si vuole calcolare il minimo di J su $\mathcal{C}$, indicato con

$$f_* = \arg\min_{f \in \mathcal{C}} J(f)$$

Se $\mathcal{C} = \mathcal{H}$ il problema e detto di minimizzazione *non vincolata*, se invece $\mathcal{C} \subset \mathcal{H}$ si parla di minimizzazione *vincolata*.
Se esiste $\delta > 0$ tale che

$$J(f_*) \leq J(f) \qquad \text{per } f \in \mathcal{C},\ \|f - f_*\|_{\mathcal{H}} < \delta$$

allora si dice che f_* è un minimo locale.
Si presentano ora alcune condizioni che garantiscono l'esistenza e l'unicità del minimo.

Definizione 2.5 (Convergenza debole)**.**
Una successione $\{f_n\}$ in uno spazio di Hilbert $\mathcal{H}$ converge debolmente a f_* se

$$\lim_{n\longrightarrow\infty}\langle f_n - f_*, f_*\rangle_{\mathcal{H}} \quad \forall\ f \in \mathcal{H}$$

e in questo caso si scrive $f_n \rightharpoonup f_*$.

La convergenza forte implica la debole e in spazi finito-dimensionali le due si equivalgono.

Definizione 2.6 (Semicontinuità debole dal basso)**.**
$J : \mathcal{H} \longrightarrow \mathbb{R}$ è debolmente semicontiuno dal basso se

$$J(f_*) \leq \liminf_{n\longrightarrow\infty} J(f_n) \qquad \text{quando } f_n \rightharpoonup f_* \tag{2.27}$$

Definizione 2.7 (Funzionale convesso)**.**
$J : \mathcal{C} \subset \mathcal{H} \longrightarrow \mathbb{R}$ è un funzionale convesso se

$$J(\tau f_1 + (1-\tau)f_2) \leq \tau J(f_1) + (1-\tau)J(f_2) \tag{2.28}$$

con $f_1, f_2 \in \mathcal{C}$ e $0 < \tau < 1$. J si dice strettamente convesso se la (2.7) è stretta per $f_1 \neq f_2$.

I funzionali convessi sono sempre debolmente semicontinui dal basso.
Si ricorda che un insieme $\mathcal{C}$ è convesso se per ogni sua coppia di elementi (f_1, f_2) una combinazione lineare convessa $\tau f_1 + (1-\tau)f_2$, con $0 < \tau < 1$, appartiene ancora allo spazio $\mathcal{C}$. $\mathcal{C}$ è chiuso se ogni successione convergente ha limite nello spazio stesso.

Definizione 2.8 (Funzionale coercivo)**.**
Un funzionale $J : \mathcal{H} \longrightarrow \mathbb{R}$ è coercivo se

$$J(f_n) \longrightarrow \infty \qquad \text{quando } \|f_n\| \rightharpoonup \infty$$

Teorema 2.3.1. *Sia $J : \mathcal{H} \longrightarrow \mathbb{R}$ debolmente semicontinuo dal basso e coercivo e sia $\mathcal{C}$ un sottoinsieme di $\mathcal{H}$ chiuso e convesso. Allora J ha un minimo su $\mathcal{C}$ e se $\mathcal{C}$ è strettamente convesso tale minimo è unico.*

Dimostrazione. Sia $\{f_n\}$ una successione minimizzante per J in $\mathcal{C}$, ossia ogni suo elemento $f_n \in \mathcal{C}$ e $J(f_n) \longrightarrow J_* \stackrel{def}{=} \inf_{f\in\mathcal{C}} J(f)$. Poichè J è coercivo $\{f_n\}$ è limitata, e la limitatezza di una successione in uno spazio di Hilbert implica l'esistenza di una sottosuccessione $\{f_{n_j}\}$ debolmente convergente, sia f_* il suo limite; $f_* \in \mathcal{C}$ perché $\mathcal{C}$ è un sottoinsieme chiuso e convesso in uno spazio di Hilbert e dunque è anche chiuso per la convergenza debole. Si ha, per la semicontinuità debole dal basso di J:

$$J(f_*) \leq \liminf J(f_{n_j}) = \lim J(f_n) = J_*$$

e dunque $J(f_*) = J_*$. Sia ora J strettamente convesso e $J(f_0) = J_*$ con $f_0 \neq f_*$. Prendendo $\tau = \frac{1}{2}$ nella (2.28) si ottiene $J(\frac{f_0+f_*}{2}) < J_*$ che è in contraddizione con quanto detto in precedenza. □

Si provvede ora alla caratterizzazione del minimo.

Definizione 2.9 (Differenziabilità secondo Fréchet)**.**
Un operatore $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ è differenziabile secondo Fréchet in $f \in \mathcal{H}_1$ se e solo se esiste $A'(f) \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$, detto *derivata di Fréchet di A in f*, tale che

$$A(f+h) = A(f) + A'(f)h + o(\|h\|_{\mathcal{H}}) \qquad \text{per } \|h\|_{\mathcal{H}} \longrightarrow 0 \tag{2.29}$$

Si definiscono ricorsivamente le derivate di Fréchet di ordini maggiori, ad esempio

$$A'(f+k) = A'(f) + A''(f)k + o(\|k\|_{\mathcal{H}})$$

definisce la derivata seconda di Fréchet, con $A''(f) \in \mathcal{L}(\mathcal{H}, \mathcal{L}(\mathcal{H}, \mathcal{H}_2))$. La mappa $(h,k) \longmapsto (A''(f)k)h$ è limitata e bilineare e va da $\mathcal{H} \times \mathcal{H}$ in $\mathcal{H}$, inoltre è simmetrica:

$$(A''(f)k)h = (A''(f)h)k \tag{2.30}$$

Definizione 2.10 (Gradiente di J in f)**.**
Sia $J : \mathcal{H} \longrightarrow \mathbb{R}$ differenziabile secondo Fréchet in f. Per il teorema di rappresentazione di Riesz esiste $\text{grad}J(f) \in \mathcal{H}$, detto il gradiente di J in f tale che

$$J'(f)h = \langle \text{grad}J(f), h \rangle_{\mathcal{H}} \tag{2.31}$$

Proposizione 2.3.2. *Se $J : \mathcal{H} \longrightarrow \mathbb{R}$ è differenziabile secondo Fréchet in f, allora $\forall\, h \in \mathcal{H}$ la mappa da $\mathbb{R}$ in $\mathbb{R}$ che porta $\tau \longmapsto J(f + \tau h)$ è differenziabile in $\tau = 0$ con*

$$\frac{d}{d\tau} J(f + \tau h)|_{\tau=0} = \langle gradJ(f), h \rangle_{\mathcal{H}} \tag{2.32}$$

Osservazione 2. La parte a sinistra dell'uguaglianza in (2.32) definisce la derivata direzionale o prima variazione di J in f nella direzione h ed è spesso indicata con $\delta J(f, h)$.

Teorema 2.3.3. *Sia $J : \mathcal{H} \longrightarrow \mathbb{R}$ differenziabile secondo Fréchet. Se ha minimo locale non vincolato in f_* allora $gradJ(f_*) = 0$.*

Dimostrazione. Sia $g = \text{grad}J(f_*)$. Per la (2.29) e la (2.31) si ha:

$$J(f_* - \tau g) = J(f_*) - \tau \|g\|^2 + o(\tau) \qquad \text{per } \tau \longrightarrow 0 \tag{2.33}$$

Se $g \neq 0$ la parte destra della (2.33) diventa minore di $J(f_*)$ prendendo $\tau > 0$ e sufficintemente piccolo. □

Teorema 2.3.4. *Sia $\mathcal{C}$ un sottoinsieme di $\mathcal{H}$ chiuso e convesso. Se $f_* \in \mathcal{C}$ è un minimo locale vincolato per J e J è differenziabile secondo Fréchet in f_*, allora*

$$\langle gradJ(f_*), f - f_* \rangle_{\mathcal{H}} \geq 0 \qquad \forall\, f \in \mathcal{C} \tag{2.34}$$

Dimostrazione. Si fissi un generico $f \in \mathcal{C}$. Dato che $\mathcal{C}$ è convesso, $f_* + \tau(f - f_*) = \tau f + (1 - \tau) f_* \in \mathcal{C}$ per $0 \leq \tau \leq 1$. Dunque, siccome f_* è minimo vincolato:

$$\lim_{\tau \longrightarrow 0^+} \frac{J(f_* + \tau(f - f_*)) - J(f_*)}{\tau} \geq 0$$

A questo punto la tesi segue dalla proposizione 2.3.2. □

Di seguito sono riportati alcuni teoremi che caratterizzano i funzionali convessi differenziabili.

Teorema 2.3.5. *Sia J differenziabile secondo Fréchet su un insieme convesso $\mathcal{C}$. J è convesso se e solo se*

$$\langle gradJ(f_1) - gradJ(f_2), f_1 - f_2 \rangle \geq 0 \tag{2.35}$$

per $f_1, f_2 \in \mathcal{C}$. J è strettamente convesso se la (2.35) è una disuguaglianza stretta per ogni f_1, f_2 tali che $f_1 \neq f_2$.

Definizione 2.11 (Derivata seconda di Fréchet)**.**
Se la derivata seconda di Fréchet esiste per J in f, allora per la (2.30) ha la seguente rappresentazione:

$$(J''(f)h)k = \langle \mathrm{Hess}J(f)h, k \rangle \tag{2.36}$$

dove $\mathrm{Hess}J(f)$ è un operatore su $\mathcal{H}$ lineare, limitato ed autoaggiunto ed è detto Hessiana di J in f.

Teorema 2.3.6. *Sia J differenziabile due volte secondo Fréchet su un convesso $\mathcal{C}$. Allora J è convesso se e solo se $HessJ(f)$ è semidefinita positiva per ogni f nell'interno di $\mathcal{C}$. Se $HessJ(f)$ è definita positiva allora J è strettamente convesso.*

Se J è due volte differenziabile secondo Fréchet in f, allora

$$J(f+h) = J(f) + \langle \mathrm{grad}J(f), h \rangle + \frac{1}{2}\langle \mathrm{Hess}J(f)h, h \rangle + o(\|h\|^2_{\mathcal{H}}) \tag{2.37}$$

per $\|h\|_{\mathcal{H}} \longrightarrow 0$. Questo fornisce le basi per condizioni del second'ordine sufficienti per minimizzazioni non vincolate.

Teorema 2.3.7. *Se J è due volte differenziabile secondo Fréchet in f_*, $gradJ(f_*) = 0$ e $HessJ(f_*)$ è strettamente positiva, allora f_* è un minimo locale stretto per J.*

2.4 Regolarizzazione secondo Tikhonov

Un generico funzionale di Tikhonov per il problema $K(f) = g$ in (2.13) ha la forma

$$\Im_\alpha(f; g) = \rho(K(f), g) + \alpha J(f) \tag{2.38}$$

dove

- $\alpha > 0$ è il parametro di regolarizzazione;
- $J : \mathcal{H}_1 \longrightarrow \mathbb{R}$ è il penalty functional;
- $\rho : \mathcal{H}_2 \times \mathcal{H}_2 \longrightarrow \mathbb{R}$ è il fit-to-data functional.

2.4.1 Penalty Functional

Lo scopo del penalty functional è quello di indurre stabilità e permettere l'inserimento di informazioni a priori sulla soluzione desiderata f. Il penalty functional standard di Tikhonov su uno spazio di Hilbert $\mathcal{H}_1$ è

$$J(f)) = \frac{1}{2}\|f\|^2_{\mathcal{H}_1}$$

Un altro esempio è il penalty functional di Sobolev H^1:

$$J_{H^1}(f) = \frac{1}{2}\int_\Omega \sum_{i=1}^{d} \left(\frac{\partial f}{\partial x_i}\right)^2$$

Funzionali di questo tipo penalizzano le soluzioni non lisce.

2.4.2 Fit-to-data Functional

Il compito del funzionale ρ nella (2.38) è quello di valutare quanto bene la previsione $K(f)$ interpola i dati osservati g. L'esempio più conosciuto è la norma al quadrato in

uno spazio di Hilbert,

$$\rho_{LS}(g_1, g_2) = \frac{1}{2}\|g_1 - g_2\|^2_{\mathcal{H}_2}, \qquad g_1, g_2 \in \mathcal{H}_2$$

2.4.3 Analisi

Si ricordi, dalla definizione 2.4, che uno schema regolarizzante è una famiglia di mappe $R_\alpha : \mathcal{H}_2 \longrightarrow \mathcal{H}_1$. Sia ora

$$\mathcal{T}_\alpha(f; g) = \frac{1}{2}\|Kf - g\|^2_{\mathcal{H}_2} + \alpha\langle Lf, f\rangle_{\mathcal{H}_1} \tag{2.39}$$

si definisce

$$R_\alpha(g) = \arg\min_{f\in\mathcal{C}} \mathcal{T}_\alpha(f; g) \qquad \alpha > 0$$

Si assume che:

A1. $\mathcal{C}$ è un sottoinsieme chiuso e convesso di$\mathcal{H}_1$;

A2. L è un operatore autoaggiunto, lineare e strettamente positivo su $\mathcal{H}_1$, questo implica che esiste una costante $c_0 > 0$ tale che

$$\langle Lf, f\rangle_{\mathcal{H}_1} \geq c_0\|f\|^2_{\mathcal{H}_1} \tag{2.40}$$

A3. K è limitato e lineare.

Il teorema seguente è un risultato di esistenza e continuità per R_α.

Teorema 2.4.1 (Esistenza e continuità di uno schema regolarizzante)**.**
Sotto le precedenti ipotesi R_α esiste ed è continuo per ogni $\alpha > 0$.

Dimostrazione. Dapprima si prova che l'operatore R_α è ben definito. Siano $g \in \mathcal{H}_2$ e $\alpha > 0$ fissati. Per alleggerire la notazione sia $\mathcal{T}(f) = \mathcal{T}_\alpha(f; g)$.
$\mathcal{T}$ è convesso e dunque è debolmente semicontinuo dal basso, inoltre per la A2 è anche coercivo, dal momento che $\mathcal{T}(f) \geq \alpha c_0\|f\|^2_{\mathcal{H}_2}$. Il teorema 2.3.1 garantisce che la (2.39)

ha un unico minimo.
Per dimostrare la continuità si fissa $g_0 \in \mathcal{H}_2$. Sia $g_n \longrightarrow g_0$. Siano $f_0 = R_\alpha(g_0)$ e $f_n = R_\alpha(g_n)$; come in precedenza si semplifica la notazione indicando $\mathcal{T}_\alpha(f; g_0)$ con $\mathcal{T}_0(f)$ e $\mathcal{T}_\alpha(f; g_n)$ con $\mathcal{T}_n(f)$. Dal teorema 2.3.4 si ha:

$$\begin{aligned} 0 &\geq \langle \text{grad}\mathcal{T}_n(f_n), f_n - f_0\rangle_{\mathcal{H}_1} \\ &= \langle (K^*K + \alpha L)(f_n - f_0), f_n - f_0\rangle_{\mathcal{H}_1} + \langle K^*(g_n - g_0), f_n - f_0\rangle_{\mathcal{H}_1} \\ &\geq \alpha c_0 \|f_n - f_0\|^2_{\mathcal{H}_1} - \|K^*(g_n - g_0)\|_{\mathcal{H}_1} \|f_n - f_0\|_{\mathcal{H}_1} \end{aligned}$$

dove l'ultima disuguaglianza segue dalla (2.40). Di conseguenza

$$\|f_n - f_0\|_{\mathcal{H}_1} \leq \frac{1}{\alpha c_0} \|K^*(g_n - g_0)\|_{\mathcal{H}_1}$$

□

2.5 Regolarizzazione con Variazione Totale

Si indica con ∇ il fradiente di una funzione regolare $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ e con $C_0^1(\Omega, \mathbb{R}^d)$ lo spazio dei vettori $\vec{v} = (v_1, \ldots, v_d)$ con componenti continuamente differenziabili e a supporto compatto su Ω. La divergena di $\vec{v}$ è data da:

$$\text{div}\vec{v} = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}$$

La norma euclidea è denotata da $|\cdot|$; lo spazio di Sobolev $W^{1,1}(\Omega)$ è la chiusura di $C_0^1(\Omega)$ rispetto alla norma

$$\|f\|_{1,1} = \int_\Omega \left[|f| + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} \right]$$

Definizione 2.12 (Funzionale TV)**.**
La variazione totale di una funzione $f \in L^1(\Omega)$ è definita come:

$$TV(f) = \sup_{\vec{v} \in V} f \text{div} \vec{v} dx \tag{2.41}$$

dove V è lo spazio delle funzioni test,

$$V = \{\vec{v} \in C_0^1(\Omega, \mathbb{R}^d) || \vec{v}(x)| \leq 1 \ \forall \ x \in \Omega\} \tag{2.42}$$

La (2.41) si generalizza in 2 e 3 dimensioni:

$$TV(f) = \sup_{\mathbf{v} \in V_2} \int_0^1 \int_0^1 f(x,y) \text{div} \mathbf{v} \ dxdy$$
$$TV(f) = \sup_{\mathbf{v} \in V_3} \int_0^1 \int_0^1 \int_0^1 f(x,y,z) \text{div} \mathbf{v} \ dxdydz$$

dove V_2 e V_3 sono analoghi a V ma in 2 e 3 dimensioni.
L'equazione (2.41) ha la seguente forma debole per funzioni lisce:

$$TV(f) = \int_\Omega |\nabla f| dx \tag{2.43}$$

da cui derivano le generalizzazioni in due e tre dimensioni:

$$TV(f) = \int_0^1 \int_0^1 \|\nabla f\| dxdy \tag{2.44}$$
$$TV(f) = \int_0^1 \int_0^1 \int_0^1 \|\nabla f\| dxdydz \tag{2.45}$$

Proposizione 2.5.1. *Se* $f \in W^{1,1}(\Omega)$ *allora* $TV(f) = \int_\Omega |\nabla f|$

Definizione 2.13 (Spazio BV(Ω))**.**
Lo spazio delle funzioni con variazione limitata, indicato con BV(Ω), ha per elementi

funzioni $f \in L^1(\Omega)$ tali che

$$\|f\|_{\mathrm{BV}} \stackrel{def}{=} \|f\|_{L^1(\Omega)} + TV(f) < \infty \tag{2.46}$$

Teorema 2.5.2. *$\|\cdot\|_{BV}$ è una norma e con questa $BV(\Omega)$ è uno spazio di Banach. In questo spazio il funzionale TV è una seminorma.*

Si enunciano di seguito importanti risultati sulla compattezza, convessità e semicontinuità di $TV(f)$.

Teorema 2.5.3 (Compattezza in L^p)**.**
Sia $\mathcal{S}$ un insieme BV-limitato di funzioni. Per $\Omega \subseteq \mathbb{R}^d$, $\mathcal{S}$ è un sottinsieme relativamente compatto di $L^p(\Omega)$ per $1 \leq p \leq \dfrac{d}{d-1}$ ed è debolmente relativamente compatto in $L^{\frac{d}{d-1}}(\Omega)$. Se $d = 1$ si considera $\dfrac{d}{d-1} = +\infty$.

Teorema 2.5.4 (Convessità di $TV(f)$)**.**
Il funzionale TV in (2.41) definito sullo spazio $BV(\Omega)$ è convesso ma non strettamente convesso. La restrizione in $W^{1,1}(\Omega)$ è strettamente convessa.

Teorema 2.5.5 (Semicontinuità)**.**
Il funzionale TV in (2.41) è debolmente semicontinuo dal basso rispetto alla norma L^p per $1 \leq p < \infty$.

Ora si esaminano l'esistenza, l'unicità e la stabilità della minimizzazione del funzionale TV con penalizzazione ai minimi quadrati, ossia del funzionale

$$\mathcal{T}(f) = \|Kf - g\|^2_{L^2(\Omega)} + \alpha\|f\|_{\mathrm{BV}} \tag{2.47}$$

dove $\alpha > 0$.

Teorema 2.5.6 (Unicità del minimo vincolato di $\mathcal{T}$)**.**
Sia $1 \leq p < \dfrac{d}{d-1}$ e sia $\mathcal{C}$ un sottinsieme di $L^p(\Omega)$ chiuso e convesso.

Sia $K : L^p(\Omega) \longrightarrow L^2(\Omega)$ lineare, limitato e tale che Null$(K) = \{0\}$. Allora per ogni $g \in L^2(\Omega)$ fissato, il funzionale (2.47) ha un unico minimo vincolato,

$$f_* = \arg\min_{f \in \mathcal{C}} \mathcal{T}(f)$$

Dimostrazione. La dimostrazione dell'esistenza è analoga alla dimostrazione del teorema 2.3.1. Siccome K è lineare e con *null space* banale, e poichè la norma al quadrato di uno spazio di Hilbert è strettamente convessa, la mappa $f \longmapsto \|Kf - g\|^2_{L^2(\Omega)}$ è strettamente convessa. L'unicità segue da questo. □

Teorema 2.5.7 (Stabilità)**.**
Siano valide le ipotesi del teorema 2.5.6. Allora il minimo f_* è stabile rispetto a

(i) perturbazioni g_n del dato g tali che $\|g_n - g\|_{L^2(\Omega)} \longrightarrow 0$;

(ii) perturbazioni K_n dell'operatore K tali che $\|K_n(f) - K(f)\|_{L^2(\Omega)} \longrightarrow 0$ uniformemente sui sottinsiemi compatti di $L^p(\Omega)$;

(iii) perturbazioni α_n del parametro di regolarizzaizone $\alpha > 0$.

Risultati analoghi di esistenza, unicità e stabilità si ottengono sostituendo la norma BV in (2.46) con il funzionale TV in (2.41):

$$\mathcal{T}(f) = \|Kf - g\|^2_{L^2(\Omega)} + \alpha TV(f) \tag{2.48}$$

La condizione Null$(K) = \{0\}$ può essere in qualche modo indebolita, il seguente teorema ne è un esempio.

Teorema 2.5.8. *Sia $\mathcal{C}$ un sottinsieme di $L^p(\Omega)$ chiuso e convesso con $1 \leq p < \dfrac{d}{d-1}$. Sia $K : L^p(\Omega) \longrightarrow L^2(\Omega)$ limitato e lineare e sia $K1 \neq 0$, dove 1 indica la funzione $1(x) = 1 \ \forall \ x \in \Omega$. Allora il funzionale in (2.48) ha un unico minimo vincolato su $\mathcal{C}$.*

Dunque la minimizzazione del funzionale di Tikhonov-TV nel discreto:

$$\mathcal{T}_\alpha(\mathbf{f}) = \frac{1}{2}\|K\mathbf{f} - \mathbf{g}\|^2 + \alpha TV(\mathbf{f}) \tag{2.49}$$

è la regolarizzazione dell'equazione $K\mathbf{f} = \mathbf{g}$ in (2.4).
La variazione totale di f, $TV(f)$, può essere interpretata geometricamente come l'area della superficie laterale del grafico di f. In particolare, sia S una regione con un bordo ∂S liscio contenuto nel quadrato unitario $[0,1] \times [0,1]$; sia:

$$f(x,y) = \begin{cases} H > 0 & se(x,y) \in S \\ 0 & altrimenti \end{cases}$$

$TV(f)$ è allora la lunghezza del bordo ∂S moltiplicato per l'altezza H del salto in f.

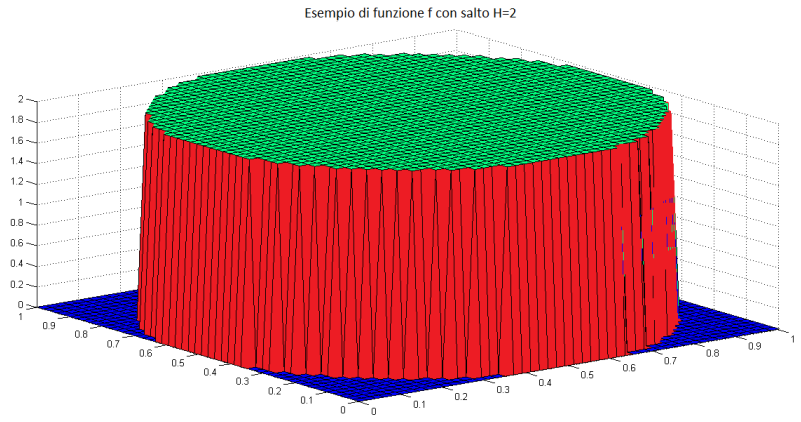

Figura 2.1: Significato geometrico della TV: in rosso è rappresentata la superficie laterale della funzione f, che corrisponde alla lunghezza del bordo del dominio per il salto di f.

Se f ha delle oscillazioni molto ampie ha anche una superfice laterale ampia e dunque un'alta variazione totale. Questa è una proprietà condivisa con il funzionale di regolarizzazione di Sobolev H^1 *norma quadrata del gradiente*, ma a differenza di esso la regolarizzazione con la variazione totale può ricostruire funzioni con dei salti. Nel caso dell'image deblurring la regolarizzazione con variazione totale produce delle ricostruzioni qualitativamente corrette di imagini a blocchi, cioè un'immagine quasi costante con dei salti. Utilizzando la regolarizzazione con la TV si ottengono risultati migliori rispetto alle tecniche di regolarizzazione tradizionali.

2.5.1 Metodi numerici per la Variazione Totale

Per ottenere la soluzione regolarizzata dell'equazione $K\mathbf{f} = \mathbf{g}$ in principio veniva minimizzato il funzionale di Tikhonov-TV (2.49). Tuttavia le rappresentazioni (2.43), (2.44) e (2.45) non sono adatte per l'implementazione di metodi numerici per l'ottimizzazione di funzioni, a causa della non differenziabilità della norma euclidea nell'origine. Per superare questo ostacolo si considera la seguente approssimazione della norma euclidea:

$$\|x\| = \sqrt{\|x\|^2 + \beta^2} \tag{2.50}$$

dove β è un parametro positivo e sufficientemente piccolo. Si ha dunque un'approssimazione di $TV(f)$ valida per funzioni smorzate f definite in $[0, 1]$:

$$J_\beta(f) = \int_0^1 \sqrt{\left(\frac{df}{dx}\right)^2 + \beta^2}\, dx \tag{2.51}$$

da cui le generalizzazioni in 2 e 3 dimensioni:

$$J_\beta(f) = \int_0^1 \int_0^1 \sqrt{\left(\frac{\partial f}{dx}\right)^2 + \left(\frac{\partial f}{dy}\right)^2 + \beta^2}\, dxdy \tag{2.52}$$

$$J_\beta(f) = \int_0^1 \int_0^1 \int_0^1 \sqrt{\left(\frac{\partial f}{dx}\right)^2 + \left(\frac{\partial f}{dy}\right)^2 + \left(\frac{\partial f}{dz}\right)^2 + \beta^2}\, dxdydz \tag{2.53}$$

Dunque il problema consiste nel minimizzare il funzionale

$$\mathcal{T}(\mathbf{f}) = \frac{1}{2}\|K\mathbf{f} - \mathbf{g}\|^2 + \alpha J_\beta(\mathbf{f}) \tag{2.54}$$

ed è un problema ben posto, che possiede un'unica soluzione stabile rispetto alle perturbazioni nei dati e nell'operatore K.
Il problema della ricerca del minimo del funzionale $\mathcal{T}(f)$ è un problema di minimo vincolato

$$\min_{\mathbf{f}} J_\beta(\mathbf{f}), \qquad \|K\mathbf{f} - \mathbf{g}\|^2 = \sigma^2 \tag{2.55}$$

dove σ è il livello d'errore.

2.5.2 Discretizzazione

La discretizzazione, come già è stato detto a proposito dei problemi mal posti, è un passaggio fondamentale per permettere la risoluzione numerica degli stessi; è necessario che tale operazione venga fatta anche sui modelli di risoluzione.

Discretizzazione in una Dimensione

Sia $f(x)$ una funzione regolare definita su $[0,1]$ e sia $\mathbf{f} = (f_0, \ldots, f_n)$ dove $f_i \approx f(x_i)$, $x_i = i\Delta x$ e $\Delta x = \dfrac{1}{n}$.
Si consideri la seguente approssimazione della derivata:

$$D_i\mathbf{f} = \frac{(f_i - f_{i-1})}{\Delta x} \qquad \forall\ i = 1, \ldots, n \tag{2.56}$$

Dunque $D_i = \left[0, \ldots, 0, -\dfrac{1}{\Delta x}, \dfrac{1}{\Delta x}, 0, \ldots, 0\right] \in \mathcal{M}_{n+1}(\mathbb{R})$
Si considera un penalty functional di tipo

$$J(\mathbf{f}) = \frac{1}{2}\sum_{i=1}^{n} \psi((D_i\mathbf{f})^2)\Delta x \tag{2.57}$$

dove ψ è una funzione liscia che approssima il doppio della radice quadrata e ha la proprietà

$$\psi'(t) > 0 \qquad \forall\ t > 0 \tag{2.58}$$

Ad esempio una buona approssimazione per (2.51) è:

$$\psi(t) = 2\sqrt{t + \beta^2}$$

Per minimizzare (2.49) occorre il gradiente di J (2.51 e seguenti). Per ogni $\mathbf{v} \in \mathbb{R}^{n+1}$

$$\frac{d}{d\tau} J(\mathbf{f} + \tau\mathbf{v}) = \sum_{i=1}^{n} \psi'([D_i\mathbf{f}]^2)(D_i\mathbf{f})(D_i\mathbf{v})\Delta x \tag{2.59}$$

$$= \Delta x (D\mathbf{v})^T \mathrm{diag}(\psi'(\mathbf{f}))(D\mathbf{f}) \tag{2.60}$$

$$= \langle \Delta x D^T \mathrm{diag}(\psi'(\mathbf{f}))D\mathbf{f}, \mathbf{v}\rangle \tag{2.61}$$

dove:

- $\mathrm{diag}(\psi'(\mathbf{f}))$ indica la matrice diagonale di dimensione $n \times n$ di elementi $\psi'([D_i\mathbf{f}]^2)$;
- D è la matrice di dimensione $n \times (n+1)$ di righe D_i (si veda (2.56));
- $\langle \cdot, \cdot \rangle$ denota il prodotto interno di $\mathbb{R}^{n+1}$

Dunque si ottiene il gradiente

$$\text{grad } J(\mathbf{f}) = L(\mathbf{f})\mathbf{f} \tag{2.62}$$

dove

$$L(\mathbf{f}) = \Delta x D^T \mathrm{diag}(\psi'(\mathbf{f}))D \tag{2.63}$$

è una matrice simmetrica di dimensioni $(n+1) \times (n+1)$.
Per avere l'Hessiana di J si parte dalla (2.59):

$$\frac{\partial^2 J}{\partial\tau\partial\xi}(\mathbf{f} + \tau\mathbf{v} + \xi\mathbf{w})|_{\tau,\xi=0} = \sum_{i=1}^{n} \psi'([D_i\mathbf{f}]^2)(D_i\mathbf{w})(D_i\mathbf{v})\Delta x + \\ + \sum_{i=1}^{n} \psi''([D_i\mathbf{f}]^2)(D_i\mathbf{f})(D_i\mathbf{v})2(D_i\mathbf{f})(D_i\mathbf{w})\Delta x \tag{2.64}$$

ossia:

$$\frac{\partial^2 J}{\partial\tau\partial\xi}(\mathbf{f} + \tau\mathbf{v} + \xi\mathbf{w})|_{\tau,\xi=0} = \langle \Delta x[\mathrm{diag}(\psi'(\mathbf{f})) + \mathrm{diag}(2(D\mathbf{f})^2\psi_i''(\mathbf{f}))]D\mathbf{v}, D\mathbf{w}\rangle \tag{2.65}$$

dove diag$(2(D\mathbf{f})^2\psi_i''(\mathbf{f}))$ indica la matrice diagonale $n \times n$ di elementi $2(D_i\mathbf{f})^2\psi_i''([D_i\mathbf{f}]^2)$. Di conseguenza

$$\text{Hess } J(\mathbf{f}) = L(\mathbf{f}) + L'(\mathbf{f})\mathbf{f} \tag{2.66}$$

dove $L(\mathbf{f})$ è dato da (2.63) e

$$L'(\mathbf{f})\mathbf{f} = \Delta x D^T \text{diag}(2(D\mathbf{f})^2\psi_i''(\mathbf{f}))D \tag{2.67}$$

Dalle equazioni (2.49), (2.62) e (2.63) si ottiene il gradiente del funzionale (2.49)

$$\text{grad } \mathcal{T}_\alpha(\mathbf{f}) = K^T(K\mathbf{f} - \mathbf{g}) + \alpha L(\mathbf{f})\mathbf{f} \tag{2.68}$$

Infine, dalle (2.66), (2.67) e (2.68), si ricava l'Hessiana:

$$\text{Hess } \mathcal{T}_\alpha(\mathbf{f}) = K^T K + \alpha L(\mathbf{f}) + \alpha L'(\mathbf{f})\mathbf{f} \tag{2.69}$$

Discretizzazione in due Dimensioni

Definizione 2.14. Dato un array $v \in \mathbb{C}^{n_x \times n_y}$, è possibile ottenere un vettore $\mathbf{v} \in \mathbb{C}^{n_x n_y}$ impilando le colonne di v. Questo definisce un operatore lineare $\mathbf{vec} : \mathbb{C}^{n_x \times n_y} \longrightarrow \mathbb{C}^{n_x n_y}$ che agisce nel seguente modo:

$$\mathbf{vec}(v) = [v_{1,1} \dots v_{n_x,1} v_{1,2} \dots v_{n_x,2} \dots\dots v_{1,n_y} \dots v_{n_x,n_y}]^T \tag{2.70}$$

Si considera ora una minimizzazione di (2.49) dove l'approssimazione del funzionale TV è bidimensionale, come in (2.52). La matrice K corrisponde alla discretizzazione di un operatore lineare che agisce su funzioni di due variabili e $\mathbf{g}$ indica i dati discreti.
Sia $f = f_{i,j}$ una funzione definita su una griglia equispaziata in uno spazio bidimensionale:

$$\{(x_i, y_i) | x_i = i\Delta x,\ y_i = j\Delta y;\ i = 1, \dots n_x;\ j = 1, \dots n_y\} \tag{2.71}$$

Analogamente a quanto fatto in una dimensione, si definisce il penalty functional discreto

$J : \mathbb{R}^{(n_x+1)\times(n_y+1)} \longrightarrow \mathbb{R}$ come:

$$J(f) = \frac{1}{2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\psi((D^x_{i,j}f)^2 + (D^y_{i,j}f)^2) \tag{2.72}$$

dove

$$D^x_{i,j}f = \frac{f_{i,j} - f_{i-1,j}}{\Delta x} \tag{2.73}$$

$$D^y_{i,j}f = \frac{f_{i,j} - f_{i,j-1}}{\Delta y} \tag{2.74}$$

Al fine di semplificare la notazione è possibile tralasciare un fattore $\Delta x \Delta y$ nella parte destra di (2.72), esso verrà assorbito dal parametro di regolarizzazione α in (2.49).
Il calcolo del gradiente è analogo al caso monodimensionale:

$$\frac{d}{d\tau}J(f+\tau v)|_{\tau=0} = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\psi'_{i,j}[(D^x_{i,j}f)(D^x_{i,j}v) + (D^y_{i,j}f)(D^y_{i,j}v)] \tag{2.75}$$

dove

$$\psi'_{i,j} = \psi'((D^x_{i,j}f)^2 + (D^y_{i,j}f)^2) \tag{2.76}$$

Ora siano $\mathbf{f} = vec(f)$ e $\mathbf{v} = vec(v)$ come in definizione 2.14; siano D_x e $_y$ le matrici di dimensioni $n_x n_y \times (n_x+1)(n_y+1)$ corrispondenti agli operatori in (2.73) e (2.74); sia $\mathrm{diag}(\psi'(\mathbf{f}))$ la matrice diagonale di dimensione $n_x n_y \times n_x n_y$ di elementi $\psi'_{i,j}(s)$; infine sia $\langle\cdot,\cdot\rangle$ il prodotto scalare interno di $\mathbb{R}^{(n_x+1)(n_y+1)}$. Allora:

$$\frac{d}{d\tau}J(f+\tau v)|_{\tau=0} = \langle \mathrm{diag}(\psi'(\mathbf{f}))D_x\mathbf{f}, D_x\mathbf{v}\rangle + \langle \mathrm{diag}(\psi'(\mathbf{f}))D_y\mathbf{f}, D_y\mathbf{v}\rangle \tag{2.77}$$

Da qui si ha la rappresentazione del gradiente (2.62), ma in questo caso

$$L(\mathbf{f}) = D_x^T\mathrm{diag}(\psi'(\mathbf{f}))D_x + D_y^T\mathrm{diag}(\psi'(\mathbf{f}))D_y$$

ossia

$$L(\mathbf{f}) = [D_x^T D_y^T] \begin{bmatrix} \text{diag}(\psi'(\mathbf{f})) & 0 \\ 0 & \text{diag}(\psi'(\mathbf{f})) \end{bmatrix} \begin{bmatrix} D_x \\ D_y \end{bmatrix} \tag{2.78}$$

Osservazione 3. La matrice $L(\mathbf{f})$ può essere vista come una discretizzazione dell'operatore di diffusione costante

$$\mathcal{L}(f)u = -\nabla \cdot (\psi' \nabla u) \tag{2.79}$$

$$= -\frac{\partial}{\partial x}\left(\psi' \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\psi' \frac{\partial u}{\partial y}\right) \tag{2.80}$$

con coefficiente di diffusione

$$\psi' = \psi'(|\nabla f|^2) = \psi'\left(bigg(\frac{\partial f}{\partial x}\right)^2 + bigg(\frac{\partial f}{\partial y}\right)^2\Big)$$

e con condizioni al bordo omogenee di Neumann[1]. L'espressione (2.80) fornisce le derivate direzionali lungo u del funzionale

$$J(f) = \frac{1}{2}\int_0^1 \int_0^1 \psi(|\nabla f|^2) dxdy \tag{2.81}$$

È possibile arrivare alla discretizzazione di $L(\mathbf{f})$ anche discretizzando direttamente l'operatore continuo $\mathcal{L}(f)$ definito in (2.80).
Come nel caso monodimensionale (2.66), si può calcolare l'Hessiana del penalty functional, con $L(\mathbf{f})$ dato in (2.78) e

$$L'(\mathbf{f})\mathbf{f} = [D_x^T D_y^T] \begin{bmatrix} \text{diag}(2(D_x\mathbf{f})^2\psi''(\mathbf{f})) & \text{diag}(2(D_x\mathbf{f})(D_y\mathbf{f})\psi''(\mathbf{f})) \\ \text{diag}(2(D_x\mathbf{f})(D_y\mathbf{f})\psi''(\mathbf{f})) & \text{diag}(2(D_y\mathbf{f})^2\psi''(\mathbf{f})) \end{bmatrix} \begin{bmatrix} D_x \\ D_y \end{bmatrix} \tag{2.82}$$

[1] *Le condizioni al bordo di Neumann di un problema*
$y'' = f(x, y, y'),\ a \le x \le b$
sono del tipo
$y'(a) = \alpha;\ y'(b) = \beta$

2.6 Lagged Diffusivity Fixed Point

Il primo dei metodi testati è il **Lagged Diffusivity Fixed Point Iteration**. Esso minimizza il problema (2.49), vale a dire:

$$\mathcal{T}_\alpha(\mathbf{f}) = \frac{1}{2}\|K\mathbf{f} - \mathbf{g}\|^2 + \alpha TV(\mathbf{f})$$

Gli iterati si calcolano nel seguente modo:

$$\mathbf{f}_{k+1} = [K^T K + \alpha L(\mathbf{f}_k)]^{-1} K^T \mathbf{g} \tag{2.83}$$

$$= \mathbf{f} - [K^T K + \alpha L(\mathbf{f}_k)]^{-1} \text{grad}(\mathcal{T}_\alpha(\mathbf{f}_k)) \tag{2.84}$$

L'espressione(2.83) si ottiene ponendo $\text{grad}\mathcal{T}_\alpha(\mathbf{f}) = 0$, così dalla (2.68) si ha

$$(K^T K + \alpha L(\mathbf{f}))\mathbf{f} = K^T \mathbf{g}$$

Dunque il coefficiente di diffusione discretizzato $\psi'(\mathbf{f})$ viene valutato in $\mathbf{f}_k$ per ottenere $L(\mathbf{f}_k)$, si vedano in proposito le espressioni (2.63) e (2.78) e l'osservazione 3; da qui viene anche il nome del metodo: *lagged diffusivity*, diffusione rallentata. La formula quasi-Newton equivalente, la (2.84) si ottiene tralasciando il termine $\alpha L'(\mathbf{f}))\mathbf{f}$ dalla Hessiana (2.69).
Di seguito viene presentato l'algoritmo, esso è basato sulla formula quasi-Newton (2.84) in quanto essa è meno sensibile agli errori inerenti[2] rispetto alla formula del punto fisso (2.83).

[2] *L'errore inerente è l'errore dovuto al condizionamento del problema e indipendente dall'algoritmo utilizzato. Insieme all'errore algoritmico, che è legato al particolare algoritmo usato, forma l'errore totale contenuto nella soluzione numerica di un sistema.*

$\mathbf{f}_0 :=$ valore iniziale;
for $k = 0, 1, \ldots$ **do**
  Step 1. $L_k := L(\mathbf{f}_k)$ % operatore di diffusione discretizzato;
  Step 2. $\mathbf{G}_k := K^T(K\mathbf{f}_k - \mathbf{g}) + \alpha L_k \mathbf{f}_k$ % gradiente;
  Step 3. $H = K^T K + \alpha L_k$ % approssimazione Hessiana;
  Step 4. $\mathbf{s}_{k+1} := -H^{-1}\mathbf{G}_k$ % quasi-Newton;
  Step 5. $\mathbf{f}_{k+1} := \mathbf{f}_k + \mathbf{s}_k$ % soluzione approssimata;
end

Osservazione 4. Se $K^T K$ è definita positiva si dimostra che il metodo converge globalmente senza alcuna ricerca in linea. L'Hessiana approssimata differisce da quella reale (2.69) per il termine $\alpha L'(\mathbf{f}_k))\mathbf{f}_k$, che non si annulla annulla al procedere delle iterazioni, dunque ci si aspetta che l'andamento della convergenza sia lineare.

2.6.1 Convergenza

Una volta approssimato il funzionale $TV(f)$ con $J_\beta(f)$, come in (2.51), si considera $\mathrm{BV}(\Omega) := \{u \in L^1(\Omega) | J_0(u) < \infty\}$, dove $\Omega \subseteq \mathbb{R}^d$ è una regione limitata e convessa. Allo stesso modo si definisce $\|u\|_{\mathrm{BV}} := \|u\|_{L^1(\Omega)} + J_0(u)$.
Dato un problema

$$Kf = g \tag{2.85}$$

che ha un'unica soluzione $f_* \in \mathrm{BV}(\Omega)$, si consideri una sequenza di problemi perturbati

$$K_n f = g_n \tag{2.86}$$

avente soluzioni approssimate g_n (non necessariamente uniche) ottenute minimizzando i funzionali

$$\mathcal{T}_n(f) = \|K_n f - g_n\|_{\mathcal{H}}^2 + \alpha_n \|f\|_{\mathrm{BV}} \tag{2.87}$$

Il teorema seguente fornisce condizioni che garantiscono la convergenza degli f_n a f_*.

Teorema 2.6.1. *Sia* $1 \leq p \leq \dfrac{d}{d-1}$. *Si supponga:*

$$\begin{aligned} &\|g_n - g\|_{\mathcal{H}} \longrightarrow 0 \\ &\qquad K_n \longrightarrow K \quad \textit{puntualmente in } L^p(\Omega) \\ &\qquad \alpha_n \longrightarrow 0 \quad \textit{a una velocità per la quale } \frac{\|A_n f_* - g_n\|^2}{\alpha_n} \quad \textit{rimane limitato} \end{aligned}$$

Quindi $f_n \longrightarrow f_*$ *in modo forte in* $L^p(\Omega)$ *se* $1 \leq p < \dfrac{d}{d-1}$.
La convergenza è debole in $L^p(\Omega)$ *se* $p = \dfrac{d}{d-1}$.

Dimostrazione. Si osservi che

$$\begin{aligned} \|K_n f_n - g_n\|_{\mathcal{H}}^2 &\leq \mathcal{T}_n(f_n) \\ &\leq \mathcal{T}_n(f_*) \\ &= \|A_n f_* - g_n\|_{\mathcal{H}}^2 + \alpha_n \|f_*\|_{\mathrm{BV}} \end{aligned}$$

Così, dalle ipotesi che $\dfrac{\|A_n f_* - g_n\|^2}{\alpha_n}$ rimane limitato e che $\alpha_n \longrightarrow 0$, vale

$$\|A_n f_n - g_n\|_{\mathcal{H}}^2 \longrightarrow 0 \tag{2.88}$$

Allo stesso modo

$$\begin{aligned} \|f_n\|_{\mathrm{BV}} &\leq \frac{\mathcal{T}_n f_n}{\alpha_n} \\ &\leq \frac{\mathcal{T}_n f_*}{\alpha_n} \\ &= \frac{\|A_n f_* - g_n\|^2}{\alpha_n} + \|f_*\|_{\mathrm{BV}} \end{aligned}$$

da qui, gli f_n sono BV-limitati. Si supponga che essi non convergano fortemente (debolmente, se $p = \dfrac{d}{d-1}$) a f_*. Dal teorema 2.5.3 esiste una sottosuccessione f_{n_j} che

converge in maniera forte (o, rispettivamente, debole) in $L^p(\Omega)$ a un certo $\hat{f} \neq f_*$. Per ogni $v \in \mathcal{H}$,

$$|\langle K\hat{f} - g, v\rangle_{\mathcal{H}}| \leq |\langle K(\hat{f} - f_{n_j}), v\rangle_{\mathcal{H}}| + |\langle (K - K_{n_j})f_{n_j}, v\rangle_{\mathcal{H}}| + \\ + |\langle K_{n_j}f_{n_j} - g_{n_j}, v\rangle_{\mathcal{H}}| + |\langle g_{n_j} - g, v\rangle_{\mathcal{H}}| \quad (2.89)$$

Il terzo e il quarto termine del membro di destra si annullano per $j \longrightarrow \infty$ a motivo della (2.88) e dell'ipotesi per cui $g_n \longrightarrow g$. Il secondo termine si annulla anch'esso, poichè

$$|\langle (K - K_{n_j})f_{n_j}, v\rangle_{\mathcal{H}}|\rangle \leq \|f_{n_j}\|_{L^p(\Omega)}\|(K^* - K^*_{n_j})v\|_{L^p(\Omega)} \longrightarrow 0$$

per la convergenza puntuale assunta sugli A_n (e quindi sui loro aggiunti) e la limitatezza di $\|f_n\|_{L^p(\Omega)}$. Il primo termine si annulla, considerando gli aggiunti e usando la convergenza (debole) di f_{n_j} a $\hat{f}$. Conseguentemente $\langle K\hat{f} - g, v\rangle_{\mathcal{H}} = 0 \ \ \forall\, v \in \mathcal{H}$ e quindi $K\hat{f} = g$. Ma questo viola l'unicità della soluzione f_* di (2.85) □

Si può anche considerare, invece del funzionale (2.87), il funzionale

$$\mathcal{T}_n(f) = \|K_n f - g_n\|^2_{\mathcal{H}} + \alpha_n J_\beta(f) \quad (2.90)$$

e ottenere lo stesso risultato del teorema precedente.

Lemma 2.6.2. *Sia* $1 \leq p \leq \dfrac{d}{d-1}$ *e sia* K *un operatore che non annulla funzioni costanti. In altre parole*

$$K_{\chi\Omega} \neq 0$$

Allora l'operatore

$$\mathcal{T}(f) = \|Kf - g\|^2_{\mathcal{H}} + \alpha J_\beta(f)$$

è BV-coercivo.

Teorema 2.6.3. *Sia* $1 \leq p \leq \dfrac{d}{d-1}$, *e sia* $\|g_n - g\|_{\mathcal{H}} \longrightarrow 0$. *Siano gli operatori* K_n *limitati, lineari e puntualmente convergenti a* K *e, per ogni* n *valga*

$$\|K_{n\chi\Omega}\|_{\mathcal{Z}} \geq \gamma > 0$$

Se $\mathcal{T}_n$ *ha un unico minimo* u_n *e* $\mathcal{T}$ *ha un unico minimo* $\bar{u}$ *allora*

$$\|f_n - \bar{f}\|_{L^p(\Omega)} \longrightarrow 0$$

Teorema 2.6.4. *Nel teorema 2.6.1, sostituendo* $\mathcal{T}_n$ *con il (2.90) e assumendo le stesse ipotesi per* K_n, α_n, g_n *e* p, *assumendo inoltre che* $|K_n\chi_\Omega| \geq \gamma > 0$, *si raggiungono le stesse conclusioni del teorema 2.6.1.*

Dimostrazione. Vale che

$$J_0(f) \leq J_\beta(f) \leq J_0(f) + \sqrt{\beta}|\Omega|$$

Da qui si assume $\beta = 0$, e, come nella dimostrazione del teorema 2.6.1, si ottiene

$$\|K_n f_n - g_n\|^2 \leq \|K_n f_* - g_n\|^2 + \alpha J_0 f_*$$

e questo implica la (2.88).
D'altra parte, ponendo $f_n = v_n + w_n$ e facendo riferimento al lemma 2.6.2 e al teorema 2.6.3, si ha

$$\mathcal{T}_n f_n \geq \gamma\|w_n\|_{L^1(\Omega)} - 2(MC_1 J_0(v_n) + m) + \alpha J_0(v_n)$$

dove M è il limite superiore di $\|A\|$, $C_1 \stackrel{def}{=} (|\Omega| + 1)^{\frac{1}{d}} C$ e C è tale che

$$\begin{aligned}\|v\|_{L^p(\Omega)} &\leq |\Omega|^{\frac{1}{p} - \frac{1}{q}} \|v\|_{L^q(\Omega)} \\ &\leq (|\Omega| + 1)^{1-\frac{1}{q}} C J_0(v)\end{aligned}$$

per $q = \dfrac{d}{d-1}$. Come nel lemma 2.6.2 ciò comporta che le f_n sono uniformemente BV-

limitate.
La dimostrazione termina come nel teorema 2.6.1. □

2.7 Algoritmo SGP

Viene di seguito introdotto il metodo **Scaled Gradient Projection (SGP)** per la risoluzione di problemi di minimo vincolato analoghi a (2.55), ossia della forma:

$$\min(f(x)), \qquad x \in \Omega \tag{2.91}$$

dove si intende:

- $\Omega \subseteq \mathbb{R}^n$ insieme chiuso e convesso;
- $f : \Omega \longrightarrow \mathbb{R}$ funzione continua e differenziabile.

2.7.1 Definizioni e Proprietà Preliminari

Definizione 2.15. La norma 2 di vettori e matrici è indicata con $\|\cdot\|$; si definisce la norma di un vettore $\mathbf{v}$ associata a una matrice simmetrica e definita positiva D come:

$$\|\mathbf{v}\|_D = \sqrt{\mathbf{v}^T D \mathbf{v}} \tag{2.92}$$

Definizione 2.16 (Punto stazionario)**.**
In riferimento al problema di ottimizzazione (2.91) si ricorda che $x_* \in \Omega$ è definito punto stazionario di f in Ω se $\forall y \in \Omega$

$$-\nabla f(x_*)^T (y - x_*) \leq 0 \tag{2.93}$$

Nel caso presente, in cui Ω è convesso, si ha una definizione equivalente:

$$-\nabla f(x_*)^T \mathbf{w} \leq 0 \tag{2.94}$$

per ogni vettore **w** appartenente al cono tangente di Ω in x_*.

Definizione 2.17 (Operatore di Proiezione $P_{\Omega,D}$)**.**
Sia $\Omega \subseteq \mathbb{R}^n$ un insieme chiuso e convesso, come precedentemente detto; sia D una matrice simmetrica e definita positiva di dimensione $n \times n$. Si definisce l'operatore di proiezione:

$$\begin{aligned} P_{\Omega,D} : \mathbb{R}^n &\longrightarrow \Omega \\ x &\longmapsto \arg\min_{y\in\Omega} \|y - x\|_D = \arg\min_{y\in\Omega} \left(\phi(y) \equiv \frac{1}{2} y^T D y - y^T D x\right) \end{aligned}$$

Questo operatore è una funzione continua rispetto agli elementi di D, dunque a partire dalla definizione (2.16) e dalla convessità di ϕ, $P_{\Omega,D}$ può essere definito come:

$$(P_{\Omega,D}(x) - x)^T D(P_{\Omega,D}(x) - y) \leq 0 \qquad \forall\ y \in \Omega \tag{2.95}$$

Definizione 2.18. Sia $\mathcal{D}_L \subseteq \mathbb{R}^{n\times n}$ l'insieme compatto delle matrici simmetriche e definite positive, di dimensione $n \times n$ e tali che $\|D\| \leq L$ e $\|D^{-1}\| \leq L$ per una soglia L fissata.

Lemma 2.7.1 (Condizione di continuità Lipschitzianaziana per $P_{\Omega,D}$)**.**
Se $D \in \mathcal{D}_L$, allora

$$\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\| \leq L^2 \|x - z\| \qquad \forall\ x, z \in \mathbb{R}^n \tag{2.96}$$

Dimostrazione. Usando la condizione (2.95) si ottiene:

$$\begin{aligned} (P_{\Omega,D}(x) - x)^T D(P_{\Omega,D}(x) - P_{\Omega,D}(z)) \leq 0 \\ (P_{\Omega,D}(z) - z)^T D(P_{\Omega,D}(z) - P_{\Omega,D}(x)) \leq 0 \end{aligned}$$

Sommando le precedenti disuguaglianze:

$$((P_{\Omega,D}(x) - x) - (P_{\Omega,D}(z) - z))^T D(P_{\Omega,D}(x) - P_{\Omega,D}(z)) \leq 0$$

ossia

$$\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|_D^2 \leq (P_{\Omega,D}(x) - P_{\Omega,D}(z))^T D(x - z) \tag{2.97}$$

Denotando con σ_{min} il minimo autovalore di D si ha la seguente maggiorazione della parte sinistra di (2.97):

$$\begin{aligned}\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|_D^2 &\geq \sigma_{min}\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|^2 \\ &= \frac{1}{\|D^{-1}\|}\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|^2 \\ &\geq \frac{1}{L}\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|^2\end{aligned}$$

Dunque la (2.97) diventa:

$$\begin{aligned}\frac{1}{L}\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|^2 &\leq \|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|_D^2 \\ &\leq (P_{\Omega,D}(x) - P_{\Omega,D}(z))^T D(x - z) \\ &\leq \|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|\|D\|\|x - z\| \\ &\leq L\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|\|x - z\|\end{aligned}$$

Da cui si ricava immediatamente la tesi. □

Lemma 2.7.2 (Caratterizzazione dei punti stazionari del problema (2.91))**.**
Un vettore $x_* \in \Omega$ è un punto stazionario per il problema (2.91) se e solo se $\forall\ \alpha > 0$ e $\forall\ D$ matrice simmetrica e definita positiva vale:

$$x_* = P_{\Omega,D^{-1}}(x_* - \alpha D\nabla f(x_*)) \tag{2.98}$$

Dimostrazione. Sia $\alpha \in \mathbb{R}^+$ e sia D una matrice simmetrica e definita positiva; sia $x_* = P_{\Omega,D^{-1}}(x_* - \alpha D\nabla f(x_*))$, dalla disuguaglianza (2.95) si ottiene:

$$(x_* - x_* + \alpha D\nabla f(x_*))^T D^{-1}(x_* - x) \leq 0 \qquad \forall\ x \in \Omega,$$

che implica la condizione di stazionarietà per x_*:

$$\nabla f(x_*))^T(x_* - x) \leq 0 \qquad \forall\, x \in \Omega$$

Viceversa, si assuma che $x_* \in \Omega$ sia un punto stazionario per (2.91) e si prenda $\bar{x} = P_{\Omega,D^{-1}}(x_* - \alpha D\nabla f(x_*))$ con $\bar{x} \neq x_*$. Ora, dalla (2.95) viene:

$$(\bar{x} - x_* + \alpha D\nabla f(x_*))^T D^{-1}(\bar{x} - x_*) \leq 0$$

ossia

$$\|\bar{x} - x_*\|^2_{D^{-1}} + \alpha\nabla f(x_*)^T(\bar{x} - x_*) \leq 0$$

Ciò implica:

$$\nabla f(x_*)^T(\bar{x} - x_*) \leq \frac{\|\bar{x} - x_*\|^2_{D^{-1}}}{\alpha} > 0$$

che nega l'ipotesi di x_* punto stazionario; dunque $\bar{x} = x_*$ □

2.7.2 Il Metodo

L'algoritmo del metodo SGP si presenta come segue:

Si scelga un punto di partenza, sia $x^{(0)} \in \Omega$, si impostino i parametri $\beta, \theta \in (0,1)$, $0 < \alpha_{min} < \alpha_{max}$ e si fissi $M \in \mathbb{Z}^+$;
for $k = 0, 1, \ldots$ **do**
  Step 1. Scegliere il parametro $\alpha_k \in [\alpha_{min}, \alpha_{max}]$ e la matrice $\mathcal{D}_k \in \mathcal{D}_L$;
  Step 2. proiezione: $y^{(k)} = P_{\Omega, D_k^{-1}}(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)}))$;
  if $y^{(k)} = x^{(k)}$ **then**
    l'algoritmo si ferma perché $x^{(k)}$ è punto stazionario
  end
  Step 3. direzione di discesa: $d^{(k)} = y^{(k)} - x^{(k)}$;
  Step 4. impostare $\lambda_k = 1$ e $f_{max} = \max_{0 \le j \le \min\{k, M-1\}} f(x^{k-j})$;
  Step 5. ciclo di backtracking:
  if $f(x^k + \lambda_k d^{(k)}) \le f_{max} + \beta \lambda_k \nabla f(x^k)^T d^{(k)}$ **then**
    proseguire con lo step 6;
  else
    impostare $\lambda_k = \theta \lambda_k$ e ritornare allo step 5;
  end
  Step 6. impostare $x^{k+1} = x^k + \lambda_k d^{(k)}$
end

Il lemma 2.7.2 mostra l'azione dell'operatore $P_{\Omega, D^{-1}}$ sui punti $(x_* - \alpha D \nabla f(x_*))$, dove $\alpha > 0$ e x_* è punto stazionario per il problema (2.91). Se $\bar{x} \in \Omega$ non è punto stazionario si può usare $P_{\Omega, D^{-1}}(\bar{x} - \alpha D \nabla f(\bar{x}))$ per generare una direzione di decrescita per la funzione f in $\bar{x}$: questa è l'idea alla base del metodo descritto dall'algoritmo SGP precedentemente illustrato.
Vale la pena sottolineare che è permessa qualsiasi scelta sia del parametro α_k in un intervallo chiuso sia della matrice D_k nell'insieme compatto $\mathcal{D}_L$, questo ha una grande rilevanza dal punto di vista pratico perché rende l'aggiornamento di α e D problemi orientati all'ottimizzazione delle prestazioni.

Osservazione 5. Si osserva che

$$\begin{aligned} d^{(k)} &= y^{(k)} - x^{(k)} \\ &= P_{\Omega, D_k^{-1}}(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})) - x^{(k)} \\ &= \left(\arg\min_{y \in \Omega} \frac{1}{2} y^T D_k^{-1} y - y^T D_k^{-1}(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})) \right) - x^{(k)} \end{aligned}$$

Ora, ponendo $B_k := \dfrac{D_k^{-1}}{\alpha_k}$ e introducendo una nuova variabile d tale che $y = x^{(k)} + d$ si può scrivere:

$$d^{(k)} = \arg\min_{x^{(k)}+d \in \Omega} \frac{1}{2} d^T B_k d + \nabla f(x^{(k)})^T d \tag{2.99}$$

Lemma 2.7.3 (Condizione di discesa per la direzione $d^{(k)}$)**.**
Sia $d^{(k)} \neq 0$. Allora $d^{(k)}$ è direzione di discesa per la funzione f in $x^{(k)}$, ossia vale

$$\nabla f(x^{(k)})^T d^{(k)} < 0$$

Dimostrazione. Dalla disuguaglianza (2.95) ponendo $x = x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})$, $D = D_k^{-1}$ e $y = x^{(k)}$ segue che

$$(d^{(k)} + \alpha_k D_k \nabla f(x^{(k)}))^T D_k^{-1} d^{(k)} \leq 0$$

e dunque

$$\nabla f(x^{(k)})^T d^{(k)} \leq -\frac{d^{(k)^T} D_k^{-1} d^{(k)}}{\alpha_k} < 0 \tag{2.100}$$

□

Lemma 2.7.4 (Limitatezza della successione $d^{(k)}$)**.**
Se la successione $\{x^{(k)}\}$ è limitata, allora lo è anche la successione $\{d^{(k)}\}$.

Dimostrazione. Dalla definizione di $d^{(k)}$ e dalla (2.96) si ha, $\forall\ k$:

$$\begin{aligned}\|d^{(k)}\|^2 &= \|P_{\Omega,D_k^{-1}}(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})) - x^{(k)}\| \\ &= \|P_{\Omega,D_k^{-1}}(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})) - P_{\Omega,D_k^{-1}}(x^{(k)})\| \\ &\leq L^2 \|\alpha_k D_k \nabla f(x^{(k)})\| \\ &\leq \alpha_{max} L^3 \|\nabla f(x^{(k)})\|\end{aligned}$$

Sia $\bar{\Omega} \subset \Omega$ un insieme chiuso e limitato tale che gli iterati $x^{(k)} \in \bar{\Omega}$; poichè ∇f è una funzione continua su Ω, allora essa è limitata su $\bar{\Omega}$ e dunque lo è anche $d^{(k)}$. □

L'algoritmo proposto è ben definito, infatti:

- se la proiezione calcolata nello Step 2 restituisce un vettore $y^{(k)} = x^{(k)}$, per il lemma 2.7.2 $x^{(k)}$ è un punto stazionario e l'algoritmo si ferma;
- diversamente, se $y^{(k)} \neq x^{(k)}$, $d^{(k)}$ è una direzione di decrescita per f in $x^{(k)}$ e il ciclo nello step 5 termina in un numero finito di iterazioni (si vedano in proposito i lemmi 2.7.3 e 2.7.4).

La ricerca in linea non monotona implementata al passo 5 assicura che $f(x^{(k)})$ sia minore del massimo della funzione obiettivo nelle ultime M iterazioni, è chiaro che se si pone $M = 1$ si implementa la regola di Armijo[3].

2.7.3 Convergenza per il metodo SGP

Vengono ora dimostrate alcune proprietà dei punti di accumulazione della successione $\{x^{(k)}\}$ generata dal metodo SGP.

[3] *La regola di Armijo è una procedura di backtracking in cui si effettuano successive riduzioni del passo a partire da un valore assegnato $\bar{\alpha} > 0$ fino a determinare un valore α_k che soddisfa una condizione di sufficiente riduzione della funzione obiettivo: $f(x_k + \alpha p_k) \leq f(x_k) + \gamma\alpha\nabla f(x_k)^T p_k,\ \ \gamma \in (0,1)$*

Lemma 2.7.5. *Sia $K \subset \mathbb{N}$; si assuma che la successione $\{x^k\}_{k\in K}$ converga ad un punto $x_* \in \Omega$. Allora x_* è un punto stazionario di (2.91) se e solo se*

$$\lim_{k\in K} \nabla f(x^{(k)})^T d^{(k)} = 0$$

Dimostrazione. Sia x_* un punto stazionario di (2.91), allora $\nabla f(x_*)^T d \geq 0 \ \forall\ d$ tale che $x_* + d \in \Omega$. Si supponga che $\{\nabla f(x^{(k)})^T d^{(k)}\} \nrightarrow 0$ $(k \in K)$; in tal caso tenendo presente il lemma 2.7.3 esistono $\varepsilon > 0$ e un insieme infinito $K_1 \subset K$ tali che

$$\nabla f(x^{(k)})^T d^{(k)} \leq -\varepsilon \leq 0 \qquad \forall\ k \in K_1$$

Poichè l'intervallo $[\alpha_{min}, \alpha_{max}]$ e l'insieme $\mathcal{D}_L$ sono compatti è possibile estrarre da K_1 un insieme di indici $K_2 \subset K_1$ tale che per $k \in K_2$ si ha $\alpha_k \longrightarrow \alpha_* \in [\alpha_{min}, \alpha_{max}]$ e $\mathcal{D}_k \longrightarrow \mathcal{D}_* \in \mathcal{D}_L$; quindi per continuità si può scrivere

$$\lim_{k\in K_2} d^{(k)} = d_*$$

dove

$$d_* := P_{\Omega, D_*^{-1}}(x_* - \alpha_* D_* \nabla f(x_*)) - x_* \tag{2.101}$$

Dunque

$$\lim_{k\in K_2} \nabla f(x^{(k)})^T d^{(k)} = \nabla f(x_*)^T d_* \leq -\varepsilon < 0 \tag{2.102}$$

Dalla definizione (2.101) si evince che $x_* + d_* \in \Omega$, pertanto la (2.102) contraddice l'ipotesi di stazionarietà di x_*.
D'altra parte si assuma che

$$\lim_{k\in K_2} \nabla f(x^{(k)})^T d^{(k)} = 0 \tag{2.103}$$

e per assurdo x_* non sia punto stazionario. Sia $K_3 \subset K$ un insieme di indici tale che per $k \in K_2$ si ha $\alpha_k \longrightarrow \alpha_*$ e $\mathcal{D}_k \longrightarrow \mathcal{D}_*$. Dunque

$$\lim_{k\in K_3} d^{(k)} = P_{\Omega, D_*^{-1}}(x_* - \alpha_* D_* \nabla f(x_*)) - x_*$$

e per il lemma 2.7.2 $\exists\ \delta > 0$ tale che $\|P_{\Omega,D_*^{-1}}(x_* - \alpha_* D_* \nabla f(x_*)) - x_*\|^2 = \delta$. Usando la disuguaglianza (2.100) si può scrivere, per un $\bar{k} \in K_3$ sufficientemente grande:

$$\nabla f(x^{(k)})^T d^{(k)} \leq -\frac{d^{(k)^T} D_k^{-1} d^{(k)}}{\alpha_k} \leq -\frac{\delta}{2\alpha_{max} L} < 0 \qquad \forall\ k \in K_3,\quad k > \bar{k}$$

Questo contraddice l'ipotesi (2.103) e permette di concludere che x_* deve essere punto stazionario. □

Lemma 2.7.6. *Sia $x_* \in \Omega$ un punto di accumulazione per la successione $\{x^{(k)}\}$ tale che per qualche $K \in \mathbb{N}$*

$$\lim_{k \in K} x^{(k)} = x_*$$

Se x_* è punto stazionario per (2.91) allora x_* è un punto di accumulazione anche per la successione $\{x^{(k+r)}\}_{k \in K},\ \forall\ r \in \mathbb{N}$. Di più:

$$\lim_{k \in K} \|d^{(k+r)}\| = 0 \qquad \forall\ r \in \mathbb{N}$$

Dimostrazione. Dal lemma 2.7.5 si ha

$$\lim_{k \in K} \nabla f(x^{(k)})^T d^{(k)} = 0$$

e dalla (2.100) si ottiene:

$$\lim_{k \in K} \|d^{(k)}\| = 0$$

Ora,

$$\lim_{k \in K} \|x^{(k+1)} - x^{(k)}\| = 0$$

e ciò implica che x_* è un punto di accumulazione anche per la successione $x^{(k+1)}$. Ricordando il lemma 2.7.5 si aggiunge che

$$\lim_{k \in K} \nabla f(x^{(k+1)})^T d^{(k+1)} = 0$$

e, per le stesse ragioni addotte in precedenza, si conclude che

$$\lim_{k\in K}\|d^{(k+1)}\| = 0$$

La tesi segue per induzione. □

Teorema 2.7.7 (Risultato di convergenza per il metodo SGP)**.**
Sia l'insieme di livello $\Omega_0 = \{x \in \Omega | f(x) \leq f(x^{(0)})\}$ *limitato. Ogni punto di accumulazione della successione* $\{x^{(k)}\}$ *generata dall'algoritmo SGP è un punto stazionario di (2.91).*

Dimostrazione. Dato che ogni iterato $x^{(k)} \in \Omega_0$, la successione $\{x^{(k)}\}$ è limitata e ha almeno un punto di accumulazione. Sia $x_* \in \Omega$ tale che

$$\lim_{k\in K} x^{(k)} = x_*$$

per un insieme di indici $K \subset \mathbb{N}$.
Si considerano separatamente due casi:

(i) sia

$$\inf_{k\in K} \lambda_k = 0$$

(si rimanda allo pseudo-codice dell'algoritmo SGP per la definizione di λ_k).
Si considera un insieme di indici K_1, $K_1 \subset K$, tale che

$$\inf_{k\in K_1} \lambda_k = 0$$

Questo significa che per k sufficientemente grande, $k \in K_1$, almeno una volta la condizione del ciclo di backtracking non viene soddisfatta e dunque al penultimo step del ciclo si ha

$$f\left(x^{(k)} + \frac{\lambda_k}{\theta} d^{(k)}\right) > f(x^{(k)}) + \beta \frac{\lambda_k}{\theta} \nabla f(x^{(k)})^T d^{(k)}$$

da cui

$$\frac{f\left(x^{(k)}+\frac{\lambda_k}{\theta}d^{(k)}\right)}{\frac{\lambda_k}{\theta}} > \beta\nabla f(x^{(k)})^T d^{(k)} \tag{2.104}$$

Per il teorema del valor medio esiste uno scalare $t_k \in \left[0, \frac{\lambda_k}{\theta}\right]$ tale che

$$\frac{f\left(x^{(k)}+\frac{\lambda_k}{\theta}d^{(k)}\right)}{\frac{\lambda_k}{\theta}} = \nabla f(x^{(k)}+t_k d^{(k)})^T d^{(k)}$$

e la (2.104) diventa:

$$\nabla f(x^{(k)}+t_k d^{(k)})^T d^{(k)} > beta\nabla f(x^{(k)})^T d^{(k)} \tag{2.105}$$

Poichè α_k e D_k sono limitati, è possibile estrarre un insieme di indici $K_2 \subset K_1$ tale che

$$\lim_{k\in K_2} \alpha_k = \alpha_*$$
$$\lim_{k\in K_2} D_k = D_*$$

In questo modo la successione $\{d^{(k)}\}_{k\in K_2}$ converge al vettore

$$d_* = P_{\Omega, D_*^{-1}}(x_* - \alpha_* D_* \nabla f(x_*)) - x_*$$

e inoltre quando $k \in K_2$ diverge si ha $t_k d^{(k)} \longrightarrow 0$; dunque mandando al limite per $k \longrightarrow \infty$, $k \in K_2$, si ottiene

$$(1-\beta)\nabla f(x_*)^T d_* \geq 0$$

Ora, dato che $(1-\beta) \geq 0$ e $\nabla f(x^{(k)})^T d^{(k)} < 0$ $(\forall\ k)$, deve necessariamente valere:

$$\lim_{k \in K_2} \nabla f(x^{(k)})^T d^{(k)} = \nabla f(x_*)^T d_* = 0$$

Questo, per il lemma 2.7.5, conclude la dimostrazione provando che x_* è un punto stazionario.

(ii) sia

$$\inf_{k \in K} \lambda_k = \rho > 0$$

Si definisce il punto $x^{(l(k))}$ come quel punto tale che

$$f(x^{(l(k))}) := f_{max} = \max_{0 \leq j \leq min\{k, M-1\}} f(x^{(k-j)})$$

Per $k > M-1$, $k \in \mathbb{N}$, vale la seguente condizione:

$$f(x^{(l(k))}) \leq f(x^{(l(l(k)-1))}) + \beta \lambda_{l(k)-1} \nabla f(x^{(l(k)-1)})^T d^{(l(k)-1)} \tag{2.106}$$

Poichè gli iterati $x^{(k)}$, dove $k \in \mathbb{N}$, stanno in un insieme limitato, la successione monotona non crescente $\{f(x^{(l(k))})\}$ ammette, per $k \in K$, un limite finito $\mathcal{L} \in \mathbb{R}$. Sia $K_3 \subset K$ un insieme di indici tale che

$$\lim_{k \in K_3} \lambda_{l(k)-1} = \rho_1 \geq \rho > 0$$
$$\exists \lim_{k \in K_3} \nabla f(x^{(l(k)-1)})^T d^{(l(k)-1)}$$

(si ricorda che, per il lemma 2.7.4, la successione $\{d^{(k)}\}_{k \in \mathbb{N}}$ è limitata); mandando al limite per $k \in K_3$ la (2.106) si ottiene:

$$\mathcal{L} \leq \mathcal{L} + \beta \rho_1 \lim_{k \in K_3} \nabla f(x^{(l(k)-1)})^T d^{(l(k)-1)}$$

vale a dire

$$\lim_{k \in K_3} \nabla f(x^{(l(k)-1)})^T d^{(l(k)-1)} \geq 0$$

Ora, dal momento che $\nabla f(x^{(k)})^T d^{(k)} < 0, \ \forall\, k$, la disuguaglianza precedente implica che

$$\lim_{k \in K_3} \nabla f(x^{(l(k)-1)})^T d^{(l(k)-1)} = 0 \tag{2.107}$$

Tale equazione implica, per il lemma 2.7.5, che ogni punto di accumulazione per la successione $\{x^{(l(k)-1)}\}_{k \in K_3}$ è punto stazionario per (2.91).
Si dimostra ora che x_* è un punto di accumulazione per $\{x^{(l(k)-1)}\}_{k \in K_3}$.
La definizione di $x^{(l(k))}$ implica $k - M + 1 \leq l(k) \leq k$ e dunque è lecito scrivere

$$\|x^{(k)} - x^{(l(k)-1)}\| \leq \sum_{j=1}^{k-l(k)} \lambda_{l(k)-1+j} \|d^{(l(k)-1+j)}\| \qquad k \in K \tag{2.108}$$

Sia $K_4 \subseteq K_3$ un insieme di indici tale che la successione $\{x^{(l(k)-1)}\}_{k \in K_4}$ converge a un punto di accumulazione $\bar{x} \in \Omega$. Dato che, come visto nella (2.107) e nel lemma 2.7.5, $\bar{x}$ è punto stazionario di (2.91), è possibile applicare il lemma 2.7.6 per ottenere, $\forall\, j \in \mathbb{N}$:

$$\lim_{k \in K_4} \|d^{(l(k)-1+j)}\| = 0$$

Usando la (2.108) si conclude che

$$\lim_{k \in K_4} \|x^{(k)} - x^{(l(k)-1)}\| = 0 \tag{2.109}$$

Da cui

$$\|x_* - x^{(l(k)-1)}\| \leq \|x^{(k)} - x^{(l(k)-1)}\| + \|x^{(k)} - x_*\|$$

e

$$\lim_{k \in K} x^{(k)} = x_*$$

dunque la (2.109) implica che x_* è punto di accumulazione anche per $\{x^{(l(k)-1)}\}_{k \in K_3}$ e si conclude che x_* è punto stazionario per (2.91).

□

Questo algoritmmo, ora presentato in forma teorica, è stato applicato per la mini-

mizzazione del funzionale

$$\mathcal{T}_\alpha(\mathbf{f}) = \frac{1}{2}\|K\mathbf{f} - \mathbf{g}\|^2 + \alpha TV(\mathbf{f})$$

dunque la successione generata dal metodo è una successione di vettori $\mathbf{f}^{(k)}$ convergenti alla soluzione, ossia il vettore $\mathbf{f}_*$ che minimizza $\mathcal{T}_\alpha(\mathbf{f})$.

2.8 Un Algoritmo per la Ricostruzione di Immagini di Tomosintesi con TV

Questa sezione tratta del terzo algoritmo testato. Esso è stato implementato per CT con sorgente a raggi divergenti ma può essere applicato anche a sorgenti con raggi a cono. L'immagine verrà rappresentata nella sua forma discreta attraverso il vettore $\mathbf{f}$ di lunghezza N_{image} e componenti $\mathbf{f}_j, \quad \forall\, j = 1, \ldots, N_{\text{image}}$; se sarà necessario farvi riferimento nel contesto di un'immagine 2D si farà uso del doppio indice, $\mathbf{f}_{s,t}$, dove

$$j = (s-1)W + t, \qquad s = 1, \ldots, H, \qquad t = 1, \ldots, W \tag{2.110}$$

e gli interi W e H sono rispettivamente la larghezza e l'altezza dell'immagine 2D, tali che $W \times H = N_{\text{image}}$.
I dati delle proiezioni sono forniti dal vettore $\mathbf{g}$, di lunghezza N_{data} e di componenti, relative alle singole misure, $\mathbf{g}_i, \quad \forall\, i = 1, \ldots, N_{\text{data}}$.
L'impostazione teorica generale di questo algoritmo TV comporta l'inversione della trasformazione lineare discreta

$$\mathbf{g} = K\mathbf{f} \tag{2.111}$$

dove la matrice del sistema K, di elementi $K_{i,j}$, è composta da N_{data} vettori-riga K_i che forniscono ogni informazione nota: $\mathbf{g}_i = K_i\mathbf{f}$. Si cerca di ottenere una rappresentazione dell'immagine attraverso un vettore $\mathbf{f}$ a partire dalla conoscenza del vettore dei dati $\mathbf{g}$ e della matrice del sistema K. Dal punto di vista matematico il problema coinvolge un numero troppo piccolo di dati: N_{data} campioni sono pochi per determinare in modo

univoco gli N_{image} elementi del vettore-immagine $\mathbf{f}$ semplicemente invertendo la (2.111); la strategia utilizzata è di incorporare l'assunzione della sparsità del gradiente dell'immagine alla funzione $\mathbf{f}$ per arrivare a una soluzione a partire dalla conoscenza dei dati $\mathbf{g}$. Per risolvere il sistema lineare in (2.111) è stato sviluppato un algoritmo TV che implementa la seguente ottimizzazione: trovare $\mathbf{f}$ tale che

$$\min \|\mathbf{f}\|_{TV} \qquad \text{per } K\mathbf{f} = \mathbf{g}, \quad \mathbf{f}_j \geq 0 \tag{2.112}$$

dove

$$\|\mathbf{f}_{s,t}\|_{TV} = \sum_{s,t} |\vec{\nabla}\mathbf{f}_{s,t}| = \sum_{s,t} \sqrt{(\mathbf{f}_{s,t} - \mathbf{f}_{s-1,t})^2 + (\mathbf{f}_{s,t} - \mathbf{f}_{s,t-1})^2}$$

Nell'algoritmo la minimizzazione del gradiente è realizzata attraverso il metodo di discesa del gradiente e il vincolo, imposto dalla conoscenza dei dati, è inserito nelle proiezioni su insiemi convessi (Projection on Convex Sets o POCS).

2.8.1 Cacolo del Gradiente del TV e Realizzazione dei Vincoli

L'algoritmo TV minimizza la Variazione Totale dell'immagine calcolata, che può essere realizzato dal metodo di discesa del gradiente o da altri metodi di ottimizzazione, questo richiede l'espressione del gradiente della TV dell'immagine. Il gradiente può essere visto esso stesso come un'immagine, in cui il valore di ogni pixel è la derivata parziale della TV dell'immagine rispetto a quel pixel. Si usa la seguente approssimazione per la derivata:

$$\begin{aligned} v_{s,t} = \frac{\partial \|\mathbf{f}\|_{TV}}{\partial \mathbf{f}_{s,t}} \approx & \frac{2(\mathbf{f}_{s,t} - \mathbf{f}_{s-1,t}) + 2(\mathbf{f}_{s,t} - \mathbf{f}_{s,t-1})}{\sqrt{\varepsilon + (\mathbf{f}_{s,t} - \mathbf{f}_{s-1,t})^2 + (\mathbf{f}_{s,t} - \mathbf{f}_{s,t-1})^2}} \\ & - \frac{2(\mathbf{f}_{s+1,t} - \mathbf{f}_{s,t})}{\sqrt{\varepsilon + (\mathbf{f}_{s+1,t} - \mathbf{f}_{s,t})^2 + (\mathbf{f}_{s+1,t} - \mathbf{f}_{s+1,t-1})^2}} \\ & - \frac{2(\mathbf{f}_{s,t+1} - \mathbf{f}_{s,t})}{\sqrt{\varepsilon + (\mathbf{f}_{s,t+1} - \mathbf{f}_{s,t})^2 + (\mathbf{f}_{s,t+1} - \mathbf{f}_{s-1,t+1})^2}} \end{aligned} \tag{2.113}$$

dove ε è un numero positivo sufficientemente piccolo.
Si fa riferimento al vettore gradiente con $\mathbf{v}$ e, come nel caso dell'immagine $\mathbf{f}$, le sue componenti sono indicate con uno o due indici: $\mathbf{v}_j$ o $\mathbf{v}_{s,t}$. Nell'algoritmo si usa il gradiente normalizzato, indicato con $\hat{\mathbf{v}}$.
Si usa il metodo POCS per realizzare il sistema lineare vincolato in (2.112); ogni componente $\mathbf{g}_i$ di $\mathbf{g}$ individua un iperpiano nello spazio N_{image}-dimensionale di tutte le possibili soluzioni $\mathbf{f}$. Il metodo POCS proietta la stima corrente di $\mathbf{f}$ sugli iperpiani, che sono convessi, in corrispondenza di ciascun punto $\mathbf{g}_i$ nell'ordine corrispondente. Ripetendo questo procedimento l'immagine calcolata si sposta nell'intersezione di tutti questi iperpiani, che è il sottospazio delle soluzioni valide per il sistema lineare. Nell'implementazione usata, il POCS comprende anche il vincolo di positività.

2.8.2 L'algoritmo

Viene ora descritto l'algoritmo TV che implementa l'ottimizzazione in (2.112) per la ricostruzione di immagini da CT con sorgente a raggi divergenti. Ogni iterazione, indicizzata con n, consiste di due fasi: POCS e discesa del gradiente; la fase POCS è ulteriormente divisa in due step che forzano rispettivamente la consistenza e la positività. Dunque ogni iterazione si compone di:

- DATA-step: vincola la consistenza con le proiezioni, il vettore-immagine è $\mathbf{f}^{(TV-DATA)}[n,m]$;
- POS-step: garantisce la non-negatività dell'immagine (vettore $\mathbf{f}^{(TV-POS)}[n]$);
- GRAD-step: riduce la TV dell'immagine calcolata restituendo un vettore $\mathbf{f}^{(TV-GRAD)}[n,m]$.

Si noti che sono esplicitati i due livelli di iterazioni.

Si inizializzano $n = 1$ e $\mathbf{f}^{(TV-DATA)}[n, 1]$;

for $m = 2, \ldots, N_{data}$ **do**

Step 1. Proiezioni:

$$\mathbf{f}^{(TV-DATA)}[n,m] = \mathbf{f}^{(TV-DATA)}[n,m-1] - K_{m-1}\frac{g_{m-1} - K_{m-1}\cdot \mathbf{f}^{(TV-DATA)}[n,m-1]}{K_{m-1}\cdot K_{m-1}}$$

Step 2. Positività:

$$(\mathbf{f}_j)^{(TV-POS)}[n] = \begin{cases} (\mathbf{f}_j)^{(TV-DATA)}[n, N_{\text{data}}] & \text{se } (\mathbf{f}_j)^{(TV-DATA)}[n, N_{\text{data}}] \geq 0 \\ 0 & \text{se } (\mathbf{f}_j)^{(TV-DATA)}[n, N_{\text{data}}] < 0 \end{cases}$$

Step 3. Inizializzazione della discesa del gradiente:

$$\mathbf{f}^{(TV-GRAD)}[n,1] = \mathbf{f}^{(TV-POS)}[n];$$
$$d_A(n) = \|\mathbf{f}^{(TV-DATA)}[n,1] - \mathbf{f}^{(TV-POS)}[n]\|_2$$

for $m = 2, \ldots, N_{grad}$ **do**

Step 4. Discesa del gradiente:

$$v_{s,t}[n,m-1] = \left.\frac{\partial \|\mathbf{f}\|_{TV}}{\partial \mathbf{f}_{s,t}}\right|_{\mathbf{f}_{s,t}=\mathbf{f}^{(TV-GRAD)}_{s,t}[n,m-1]};$$
$$\hat{\mathbf{v}}[n,m-1] = \frac{\mathbf{v}[n,m-1]}{|\mathbf{v}[n,m-1]|};$$
$$\mathbf{f}^{(TV-GRAD)}[n,m] = \mathbf{f}^{(TV-GRAD)}[n,m-1] - ad_A(n)\hat{\mathbf{v}}[n,m-1]$$

Step 5. Prossimo giro:

$$\mathbf{f}^{(TV-DATA)}[n+1,1] = \mathbf{f}^{(TV-GRAD)}[n,m] = [n, N_{\text{grad}}]$$

end

Step 6. Si incrementa n e si ritorna allo Step 1.

end

L'algoritmo si ferma quando non si verificano cambiamenti apprezzabili nell'immagine dopo il POCS-step, cioè quando $\mathbf{f}^{(TV-POS)}[n] - \mathbf{f}^{(TV-POS)}[n-1]$ è piccola.
La distanza $d_A(n)$ misura la differenza tra l'immagine prima del DATA-step e la stima dopo limposizione di positività. La discesa del gradiente è controllata specificando: il parametro a, la distanza $d_A(n)$ lungo cui l'immagine è sviluppata e N_{grad}, il numero totale di step di discesa del gradiente effettuati. L'algoritmo si basa sull'equilibrio tra la fase POCS e la discesa del gradiente. Ridimensionando il salto della discesa del gradiente con la distanza d_A, l'importanza relativa del POCS e della discesa del gradiente permette all'algoritmo di mantenersi equilibrato. Finchè il cambiamento totale nell'immagine dovuto alla discesa del gradiente non supera quello dovuto al POCS, la procedura generale diiterazione porta le stime dell'immagine più vicine allo spazio delle soluzioni del sistema lineare. Se i salti di decrescita del gradiente sono troppo alti l'immagine diventa uniforme e inconsistente con i dati delle proiezioni; d'altra parte se il gradiente scende troppo lentamente, l'algoritmo si riduce all'ART standard con un vincolo di positività.

Capitolo 3

Risultati Numerici

In questo capitolo verrano mostrati i risultati ottenuti applicando i tre metodi precedentemente spiegati alla ricostruzione di immagini tridimensionali, al fine di determinare pregi e difetti di ognuno.

3.1 I Problemi Test

Le immagini usate sono fantocci digitali tridimensionali di base quadrata e spessore 15 voxel, il *cirs-mini* ad esempio è costituito da un quadrato di 64×64 pixel di base; con questo si vuole simulare la presenza di masse tumorali e calcificazioni all'interno del tessuto, pertanto si trovano, nei tre piani centrali, dei pixel ad intensità maggiore.

Figura 3.1: *Cirs-mini*: fantoccio con gli oggetti del piano centrale e, a destra, differenze di intensità tra gli oggetti e lo sfondo

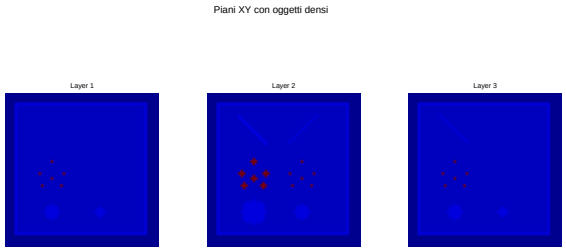

Figura 3.2: *Cirs-mini*: oggetti densi nei tre piani centrali.

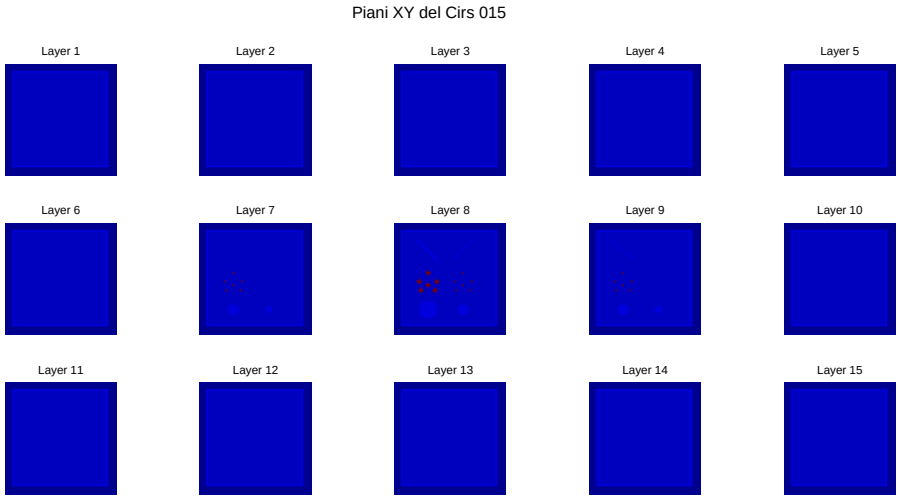

Figura 3.3: *Cirs-mini* completo.

In particolare, i pixel rossi sulla sinistra rappresentano calcificazioni di dimensioni notevoli, infatti appaiono in tutti e tre i piani, al loro fianco si trovano altri pixel che simulano microcalcificazioni, questi due gruppi hanno la stessa intensità ma dimensioni e profondità diverse; in alto e in basso stanno invece macchie più chiare, rafiguranti masse, formazioni nodulari e filamenti, anche questi ultimi oggetti hanno tra loro la stessa

intensità ma forme e dimensioni differenti. In questo modo è possibile saggiare la bontà della ricostruzione su oggetti diversi in forme, dimensioni e natura. Inoltre i test sono stati eseguiti aggiungendo del rumore all'immagine esatta, così da ottenere una ricostruzione non banale. Sono stati adottati due diversi *Random Noise Level*, uno pari a 10^{-3} e l'altro pari a 5×10^{-3}.

Un altro fantoccio usato è il *cirs*, analogo al *cirs-mini* ma di base 128×128 pixel. Nei piani centrali si trovano più elementi rispetto al caso precedente, in modo da simulare i medesimi oggetti ma in un corpo più grande.

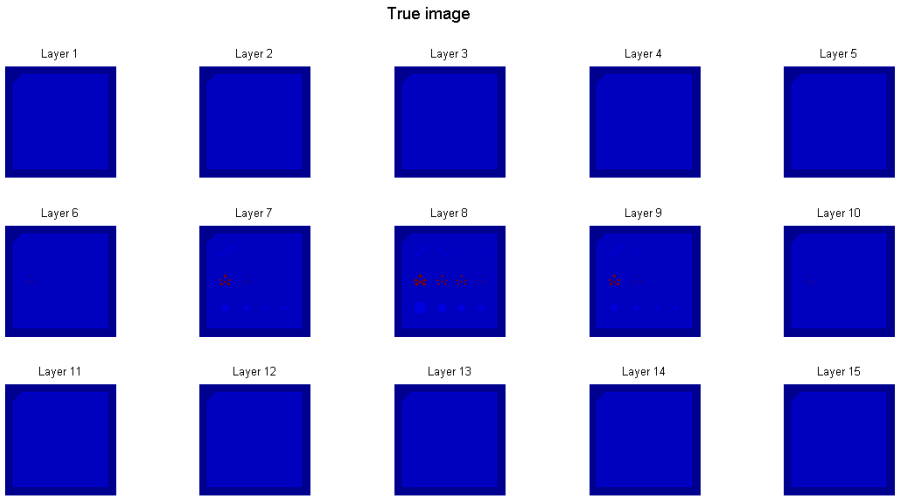

Figura 3.4: *Cirs*.

L'ultimo fantoccio è lo *Shepp-Logan* e appare decisamente più complesso, esso infatti simula la presenza di più corpi di forme, natura e intensità differenti disposti uno dentro l'altro. Si vedrà che risulta il più difficile da ricostruire.

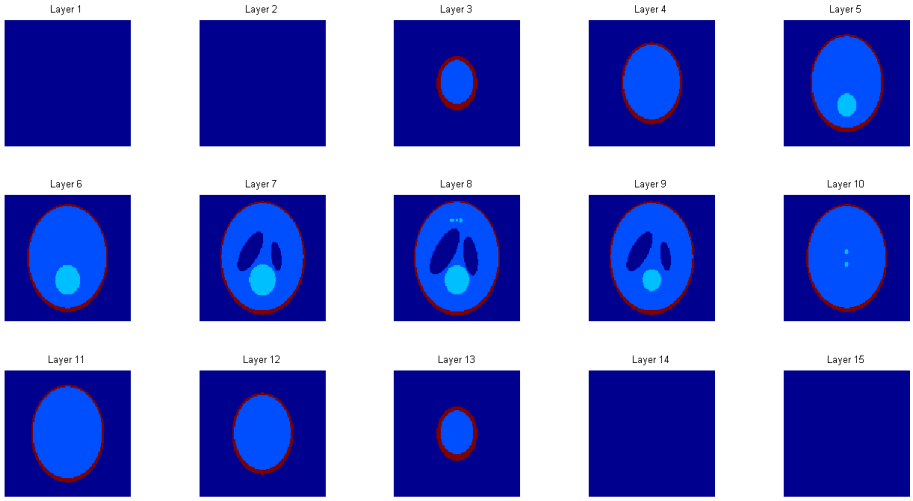

Figura 3.5: *Shepp-Logan.*

3.2 Il Parametro α

Come detto, la procedura implementata consiste nella minimizzazione del funzionale (2.54):

$$\mathcal{T}_\alpha(\mathbf{f}) = \frac{1}{2}\|K\mathbf{f} - \mathbf{g}\|^2 + \alpha J_\beta(\mathbf{f})$$

dove il parametro β è fissato a 10^{-6}.
Il parametro α invece è il parametro del modello e può variare a seconda dell'algoritmo, del problema test e del rumore usati. Questo è un valore molto importante perchè gestisce il rapporto tra il penalty functional e il fit-to-data functional, valori sbagliati di α possono provocare distorsioni rilevanti nelle immagini ricostruite, come accade nell'esempio in figura 3.6, dove si osserva l'ultimo piano del fantoccio, in teoria privo di oggetti.

Figura 3.6: L'immagine di destra è realizzata con un valore di α troppo alto, $\alpha = 10^{-3}$, la ricostruzione è distorta in maniera evidente in quanto presenta macchie di dimensioni considerevoli non presenti nell'immagine originale. Sulla sinistra, un valore di α più appropriato, pari a 10^{-6}, realizza un'immagine decisamente migliore.

Un buon valore di α produce una ricostruzione valida, come quella riportata in figura 3.8, essa è stata ottenuta con il metodo di Vogel aggiungendo all'immagine esatta un rnl pari a 5×10^{-3} e imponendo un errore massimo di 0.09. Una soglia così bassa è stata ben tollerata grazie alle ridotte dimensioni dell'immagine, per fantocci più grandi è necessario usare soglie molto più alte, come nell'esempio di figura (3.10) in cui l'errore massimo è pari a 0.2, si notano nei piani non centrali distorsioni dell'immagine dovute all'errore non piccolo, tuttavia per motivi di tempi di calcolo è necessario tenere l'errore intorno a questo valore.

Figura 3.7: *Cirs-mini*: ricostruzione con errore massimo pari a 0.09, dettaglio del piano centrale.

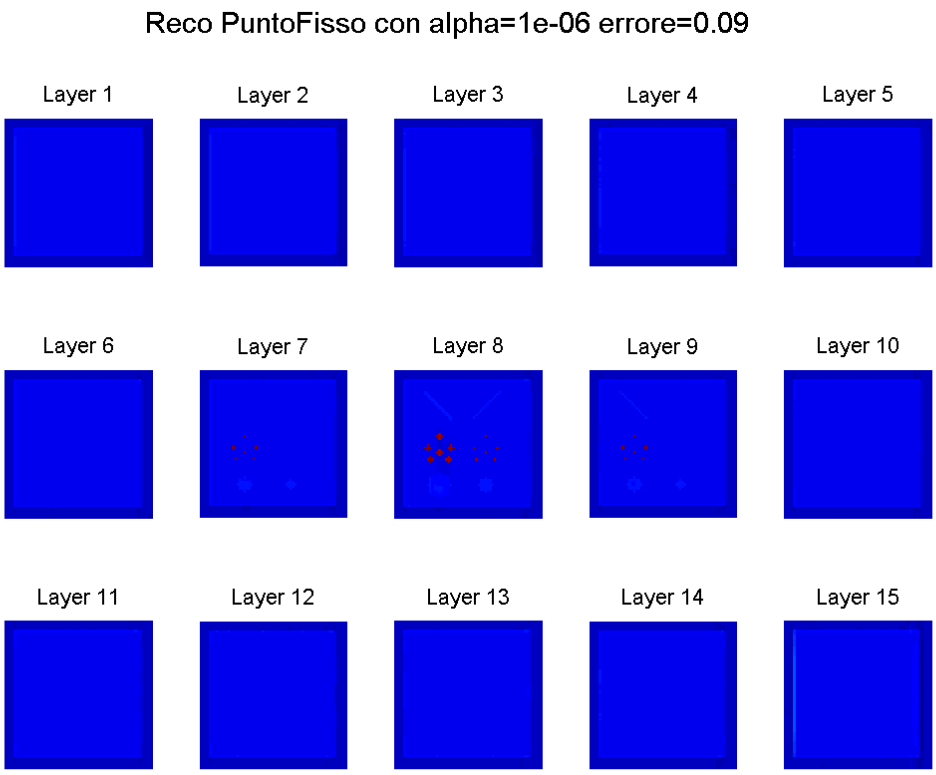

Figura 3.8: *Cirs-mini*: ricostruzione con errore massimo pari a 0.09.

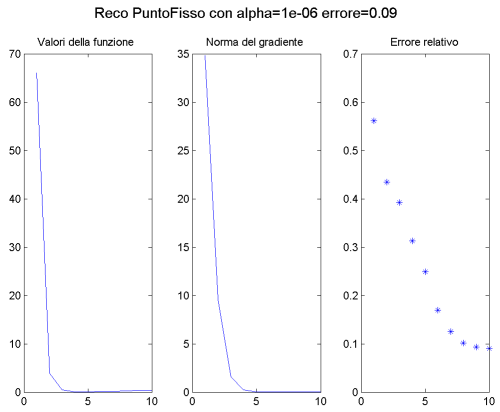

Figura 3.9: *Cirs-mini*: grafici per errore massimo 0.09.

Nei grafici riportati si osserva la corretta decrescita della funzione-obiettivo, della norma del gradiente e dell'errore relativo.

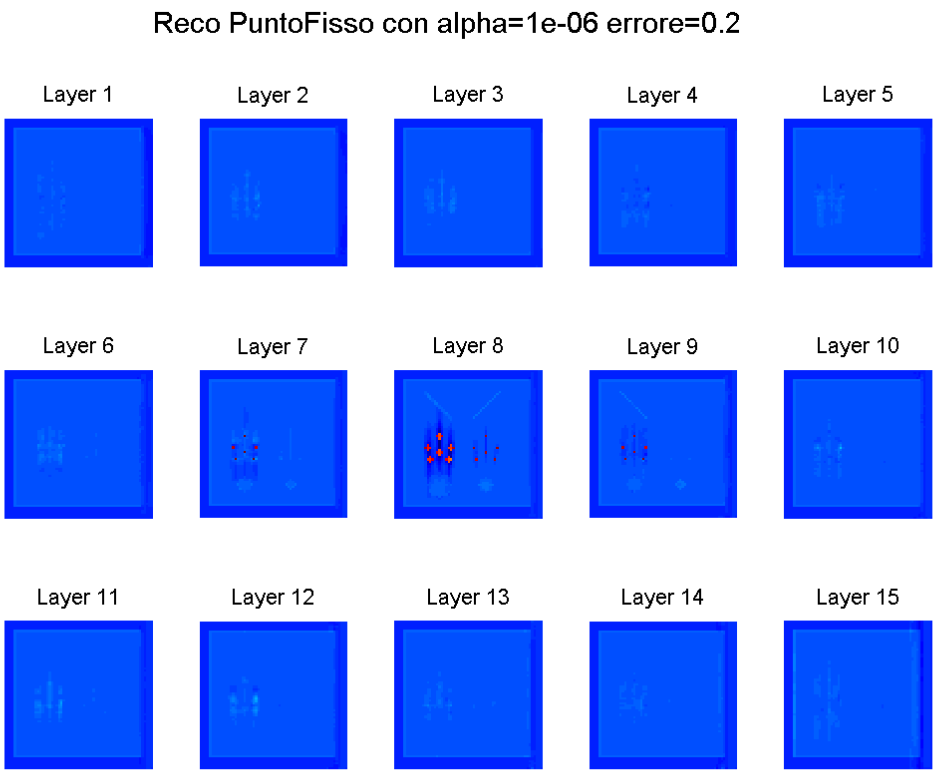

Figura 3.10: *Cirs-mini*: ricostruzione con errore massimo pari a 0.2.

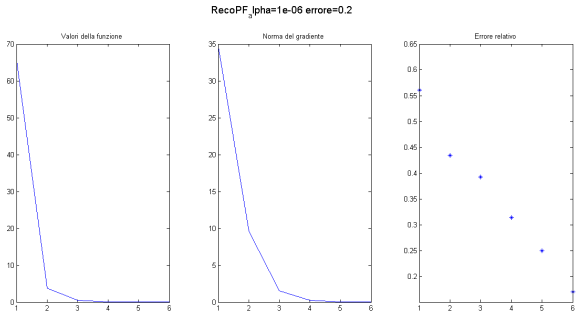

Figura 3.11: *Cirs-mini*: grafici per errore massimo 0.2.

Come appare dal grafico, in questo caso la decrescita dell'errore è più ripida rispetto al caso precedente. Questo è un fattore non secondario nella scelta del valore di α in previsione dell'utilizzo di questi metodi nella diagnostica: come detto, vengono privilegiati quei valori del parametro che ad una buona ricostruzione associano tempi di elaborazione ridotti, dunque la velocità di discesa dell'errore costituisce un'indicazione importante in questo senso.

3.3 Confronto tra i tre Metodi

Un confronto tra tutti e tre i metodi proposti è stato possibile solo usando il fantoccio $cirs - mini$, infatti l'implementazione del terzo algoritmo, proposta da Sidky, è pensata per un linguaggio C e non sfrutta appieno le capacità numeriche di MatLab, richiedendo tempi di calcolo troppo lunghi; probabilmente si può ovviare a questa difficoltà realizzando un codice che permetta il calcolo in parallelo, consentendo una diminuzione del tempo necessario per il calcolo delle proiezioni.

- FANTOCCIO USATO: *cirs-mini*, dimensioni della base: 64×64 pixel
- RNL=10^{-3}
- CRITERIO DI ARRESTO: distanza degli iterati successivi minore di 2×10^{-2} o errore minore di 0.1 (0.2 per il terzo metodo)

Metodo	**Complessità Computazionale**	**Tempo Impiegato**
SGP, $\alpha = 10^{-7}$	84 iterazioni compiute	$148,226$ sec
Punto Fisso, $\alpha = 10^{-6}$	485 iterazioni compiute dal CG	$80,387$ sec
III metodo, $\alpha = 10^{-7}$	3 iterazioni esterne compiute	> 1 giorno

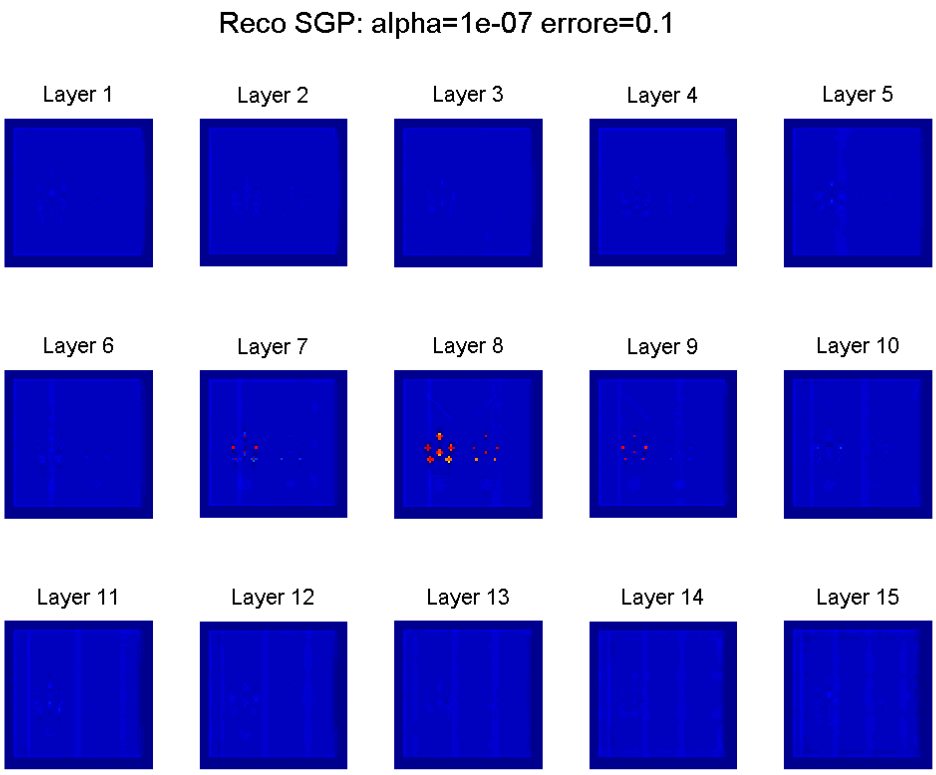

Figura 3.12: *Cirs-mini*: ricostruzione con SGP.

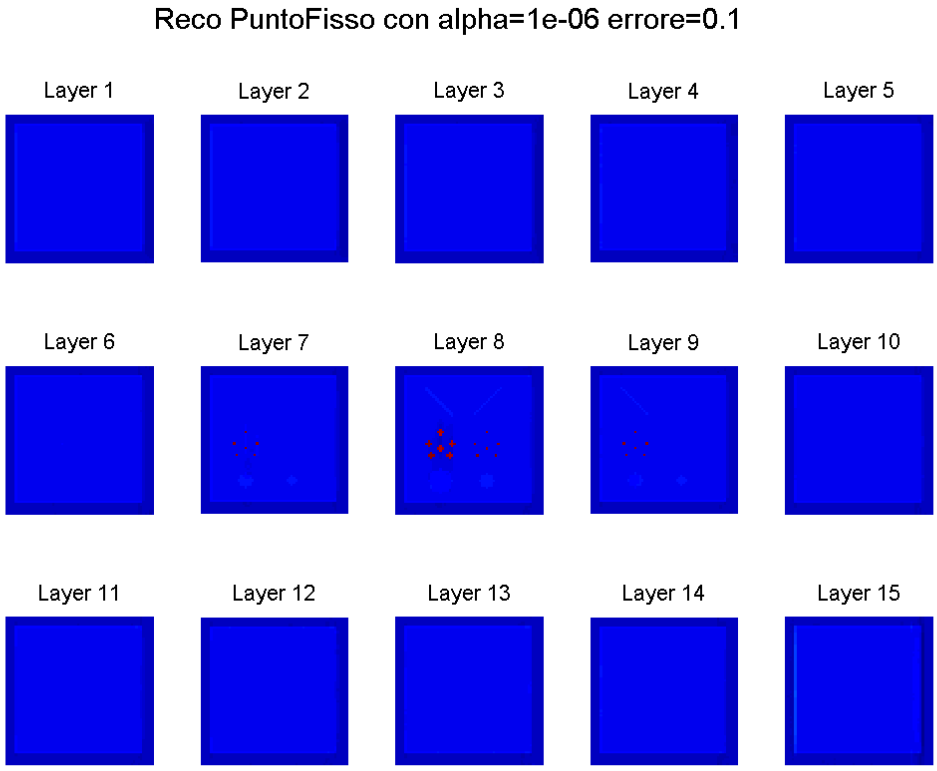

Figura 3.13: *Cirs-mini*: ricostruzione con punto fisso.

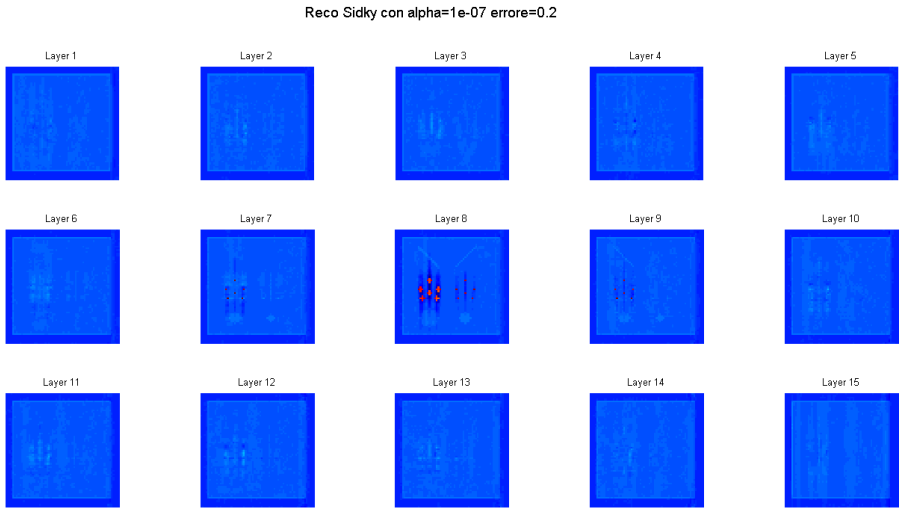

Figura 3.14: *Cirs-mini*: ricostruzione con terzo metodo.

La ricostruzione migliore è quella ottenuta con il metodo del punto fisso, che si rivela il metodo più vantaggioso anche dal punto di vista del tempo di elaborazione. A livello di costo computazionale invece risulta più funzionale il metodo SGP, che richiede molte meno iterazioni, ma per arrivare a immagini pulite ha bisogno di un errore più basso di quello imposto. Il terzo metodo infine produce immagini con distorsioni evidenti, ma questo probabilmente dipende dal numero massimo di iterazioni imposte: per evitare tempi di calcolo ancora più lunghi è stato imposto un ulteriore criterio di arresto, ossia il numero di iterazioni esterne non superiore a 3.

- FANTOCCIO USATO: *Shepp-Logan*, dimensioni della base: 127×127 pixel
- RNL=10^{-3}
- CRITERIO DI ARRESTO: distanza degli iterati successivi minore di 2×10^{-2} o errore minore di 0.2 (0.1 per lo SGP)

Metodo	Complessità Computazionale	Tempo Impiegato
SGP, $\alpha = 10^{-9}$	370 iterazioni compiute	$1436,102$ sec
Punto Fisso, $\alpha = 10^{-7}$	14612 iterazioni compiute dal CG	$1016,383$ sec

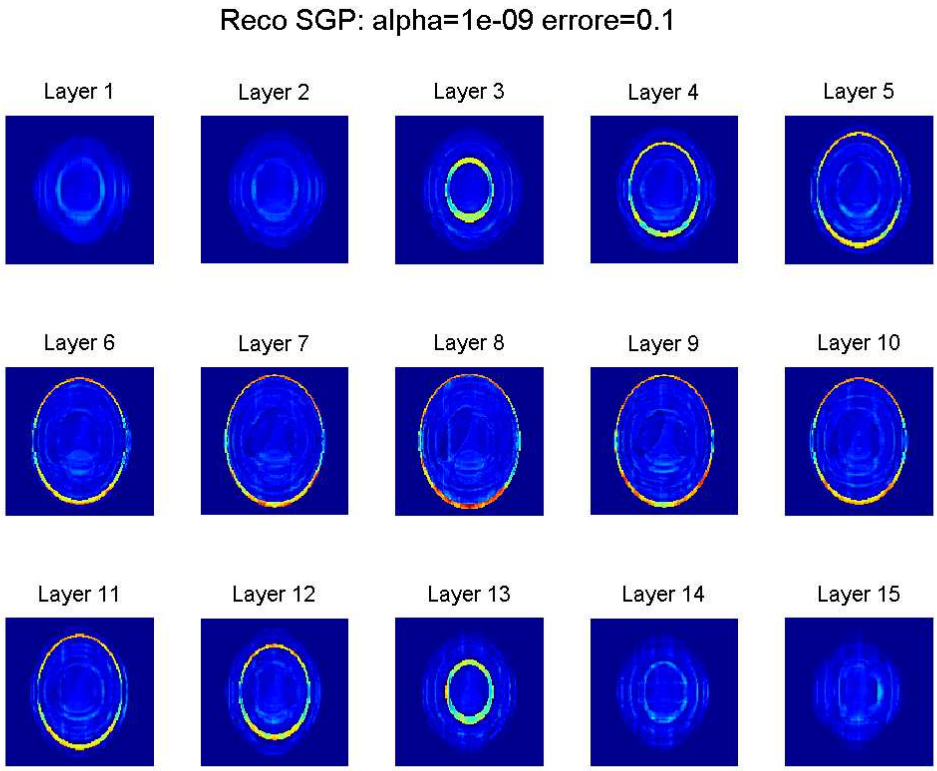

Figura 3.15: *Shepp-Logan*: ricostruzione con SGP.

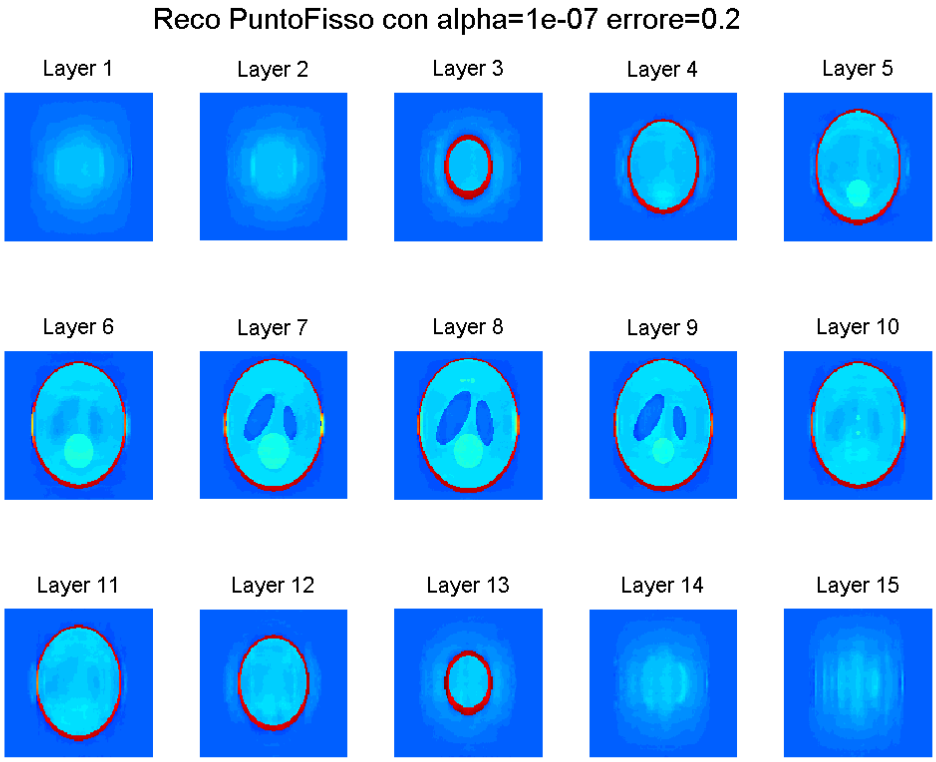

Figura 3.16: *Shepp-Logan*: ricostruzione con punto fisso.

Anche in questo caso il metodo migliore è il punto fisso, che restituisce in tempi minori rispetto al gradiente scalato proiettato (per via delle dimensioni del fantoccio non è stato possibile testare il terzo metodo) un'immagine in cui i piani centrali sono abbastanza definiti, anche se quelli alle estremità risentono di distorsioni molto evidenti. Lo SGP anche in questo caso non riesce a riprodurre un'immagine nitida, nonostante l'errore imposto sia più basso di quello usato per il punto fisso; nei piani centrali si distinguono gli oggetti principali ma non è possibile realizzare una ricostruzione fedele o almeno vicina all'originale.
I grafici riportati in figura 3.17 mostrano che nel metodo di Vogel l'errore decresce in modo regolare e abbastanza lentamente nelle ultime $50 - 70$ iterazioni, mentre nel me-

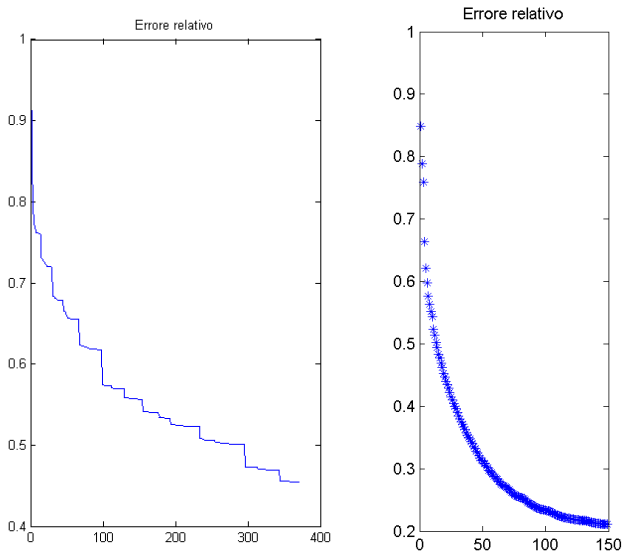

Figura 3.17: *Shepp-Logan*: sulla sinistra, grafico dell'error relativo con il metodo SGP; sulla destra, errore relativo per il punto fisso.

todo SGP la decrescita è fortemente irregolare. Questo motiva ulteriormente i risultati ottenuti.

- FANTOCCIO USATO: *Cirs*, dimensioni della base: 128×128 pixel
- RNL=5×10^{-3}
- CRITERIO DI ARRESTO: norma del gradiente minore di 5×10^{-3} o errore minore di 0.2

Metodo	**Complessità Computazionale**	**Tempo Impiegato**
SGP, $\alpha = 10^{-6}$	9 iterazioni compiute	$73,9343$ sec
Punto Fisso, $\alpha = 10^{-7}$	833 iterazioni compiute dal CG	$78,806$ sec

NB: In entrambe i casi l'arresto è per errore< 0.2, non per norma del gradiente inferiore alla soglia.

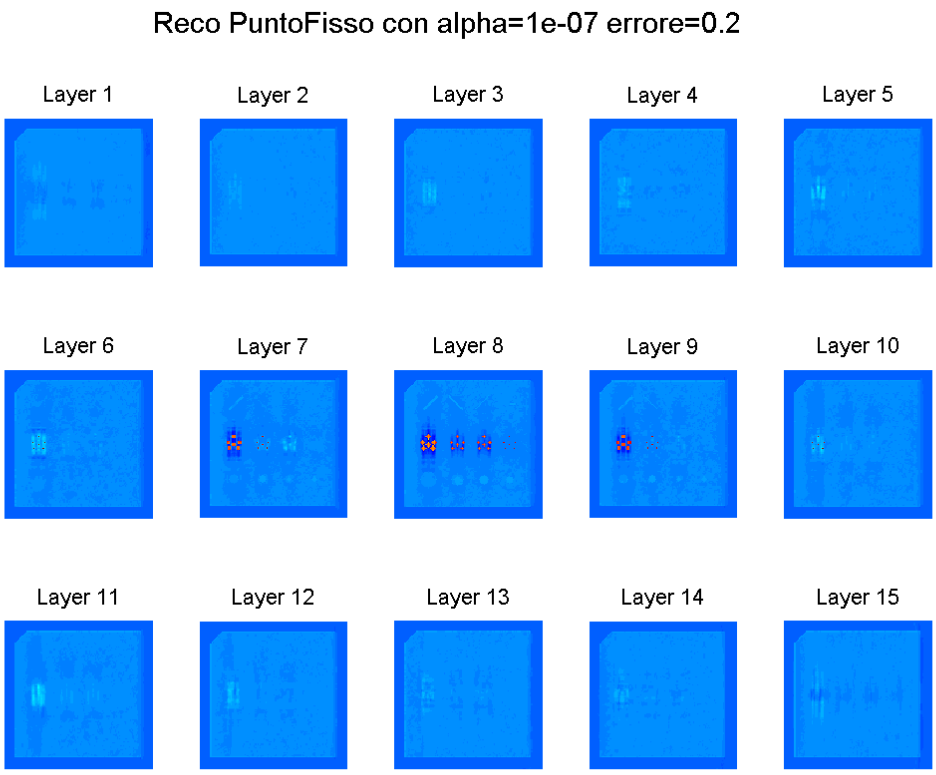

Figura 3.18: *Cirs*: ricostruzione con punto fisso.

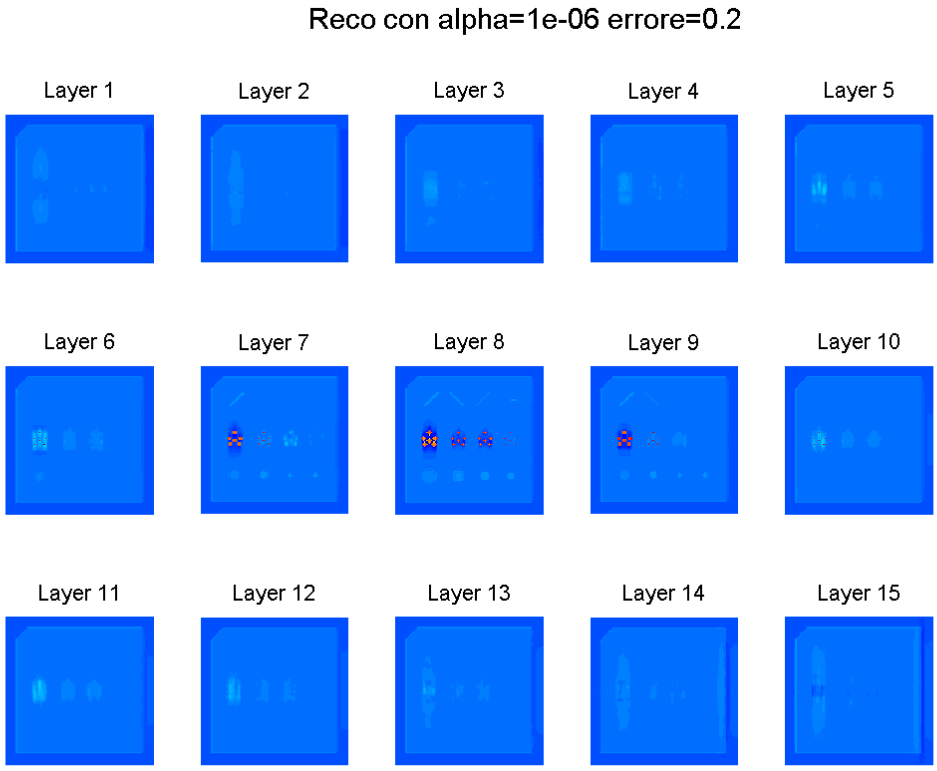

Figura 3.19: *Cirs*: ricostruzione con SGP.

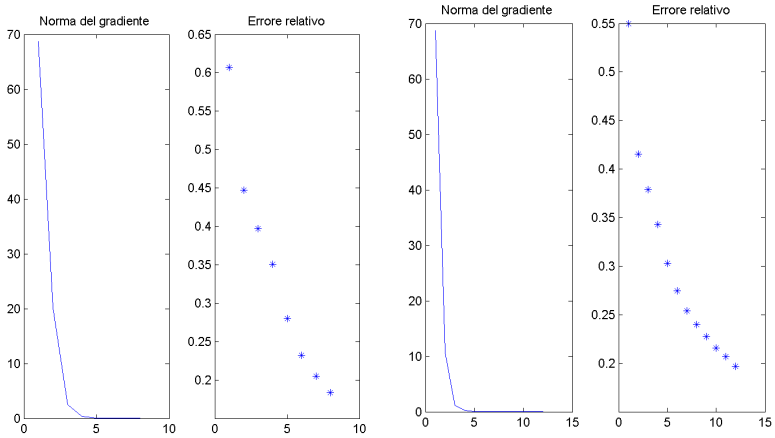

Figura 3.20: *Cirs*: sulla sinistra, grafico dell'errore relativo con il metodo SGP; sulla destra, errore relativo per il punto fisso.

Anche in quest'ultimo caso riportato il metodo che risponde meglio è il punto fisso, anche se lo SGP si rivela, come sempre, più vantaggioso in termini di costo computazionale. L'immagine del punto fisso, seppur macchiata e debolmente distorta, è fedele all'originale e permette di individuare correttamente gli oggetti densi all'interno del corpo. Il metodo SGP restituisce un'immagine meno buona ma comunque abbastanza vicina all'originale. I grafici degli errori indicano una velocità di decrescita piuttosto sostenuta in entrambe i casi, questo conferma la bontà del valore di α e la ragionevolezza dei criteri di arresto.

Conclusioni

Nel presente lavoro di tesi sono stati studiati tre diversi algoritmi per la risoluzione di problemi mal posti. Sono stati inoltre valutati dal punto di vista della qualità della ricostruzione effettuata, e le conclusioni che si possono trarre in questo senso sono inevitabilmente condizionate dalla lentezza del terzo algoritmo, che rende impossibile un confronto alla pari tra i tre metodi: non è ragionevole confrontare immagini ottenute con tempi di elaborazione così differenti.
Gli altri due algoritmi, lo SGP e il punto fisso proposto da Vogel, dal punto di vista matematico differiscono per la proiezione dei dati, presente nel primo ma non nel secondo, tuttavia restituiscono entrambi immagini abbastanza buone e in tempi simili. È difficile determinare quale sia il migliore tra i due, infatti questo dipende anche dai criteri d'arresto imposti e dal livello del rumore sui dati, accusato in maniera differente dai due algoritmi.

Un possibile sviluppo per il futuro è certamente la sperimentazione degli algoritmi su immagini reali, confrontando i risultati con le ricostruzioni effettuate dai metodi normalmente in uso per questo tipo di diagnosi.

Bibliografia

[1] Bonettini, S. - Zanella, R. - Zanni, L. *A Scaled Gradient Projection Method for Constrained Image Deblurring*, Università di Ferrara, 2009

[2] Comincioli, V. *Metodi Numerici e Statistici per le Scienze Applicate*, 2004, F.A.R Università degli Studi di Pavia

[3] Francini, E. *Problemi Inversi*, dispense del corso di Problemi Inversi, Università degli Studi di Firenze, A.A. 2005-2006

[4] Lemmo, A. *Ricostruzione di Immagini di Tomosintesi con Metodi Iterativi*, CdS Matematica, sessione terza, A.A. 2010-2011

[5] Morotti, E. *Tecniche di Regolarizzazione per Analisi Perfusionali da Immagini Tomografiche*, CdS Matematica, sessione terza, A.A. 2011-2012

[6] Sidky, E. Y. - Kao, C. - Xiaochuan P. *Accurate Image Reconstruction from Few-Views and Limited-Angle Data in Divergent-Beam CT*, University of Chicago, 2009

[7] Tartuferi, S. *Metodi di Regolarizzazione nella Ricostruzione di Immagini di Tomosintesi*, CdS Matematica, sessione seconda, A.A. 2010-2011

[8] Vogel, C. R. *Computational Methods for Inverse Problems*, 2002, SIAM

[9] Vogel, C.R. - Oman, M.E. *Iterative Methods for Total Variation Denoising*

[10] Vogel, C.R. - Acar, R. *Analysis of Bounded Variation Penalty Methods for Ill-Posed Problems*, in *Inverse Problems 10*, 1994

[11] http://www.artoi.it/tomosintesi-mammaria/

[12] http://rsb.info.nih.gov/ij/

[13] http://www.math.montana.edu/ vogel/

[14] http://www.rad.unipi.it/index.php/area-pazienti/tomografia-computerizzata