Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

Il metodo Nelder Mead per funzioni non differenziabili

Tesi di Laurea in Analisi Numerica

Relatore:
Chiar.ma Prof.ssa
Germana Landi

Presentata da:
Eleonora Cinquegrani

Sessione Straordinaria
Anno Accademico 2015-2016

Indice

Introduzione

Nell'ambito della matematica applicata, ed in particolare dell'analisi numerica, l'ottimizzazione si occupa dello studio della teoria e dei metodi per la ricerca dei punti di massimo e minimo di una funzione. Dovendo indagare su massimi e minimi, la maggior parte degli algoritmi di ottimizzazione noti è basata sul concetto di derivata e sulle informazioni che possono essere dedotte dal gradiente. Tuttavia molti problemi di ottimizzazione derivanti da applicazioni reali sono caratterizzati dal fatto che l'espressione analitica della funzione obiettivo non è nota, cosa che rende impossibile calcolarne le derivate, oppure è particolarmente complessa, per cui codificare le derivate potrebbe richiedere troppo tempo. Per risolvere questo tipo di problemi sono stati sviluppati diversi algoritmi che non tentano di approssimare il gradiente ma piuttosto utilizzano i valori della funzione in un insieme di punti di campionamento per determinare una nuova iterata con altri mezzi.
Questa tesi analizza e sperimenta uno di tali algoritmi: il metodo Nelder Mead, il cui scopo è appunto quello di minimizzare una funzione non lineare attraverso la sua valutazione in alcuni punti di prova che costituiscono una particolare forma geometrica detta simplesso. In particolare l'obiettivo di questo studio è stato quello di valutare tale metodo, le sue proprietà di convergenza e la possibilità di applicarlo efficaciemente in diversi contesti.
Per potere fare un bilancio più preciso, nell'analisi sperimentale, l'algoritmo è stato messo a confronto con un altro metodo basato sempre sul simplesso, ma che sfrutta informazioni sul gradiente. I risultati ottenuti hanno messo in luce sostanzialmente due aspetti importanti dell'algoritmo Nelder Mead:

- l'accuratezza nell'approssimare il punto di minimo è più che accettabile;
- i tempi di esecuzione sono confrontabili con quelli dell'altro metodo.

La tesi si articola in sei capitoli.
Il Capitolo 1 è di carattere prettamente introduttivo. Si introduce infatti il tema dell'ottimizzazione senza derivate e, all'interno di questo contesto, la filosofia che sta alla base dell'algoritmo Nelder Mead, il quale viene descritto in tutte le sue possibili

fasi, esponendo poi alcune proprietà di carattere generale.
Il Capitolo 2 illustra le proprietà di convergenza nel caso di funzioni strettamente convesse sia per quanto riguarda la dimensione uno che la dimensione due.
Il Capitolo 3 descrive ed analizza casi in cui l'algoritmo si trova ad eseguire una contrazione interna ripetuta per poi convergere ad un punto che non risulta essere il minimo della funzione in oggetto. Viene anche accennato un esempio di non convergenza che verrà poi ripreso nella parte sperimentale.
Nel Capitolo 4 viene proposta una leggera modifica del metodo che prevede un riavvio orientato nel momento in cui si arrivi alla situazione di contrazione interna ripetuta descritta nel capitolo precedente. Vengono quindi forniti alcuni risultati di convergenza ottenuti grazie a tale strategia.
Il Capitolo 5 riporta i risultati numerici dei test effettuati su alcune funzioni note ed un confronto con quanto ottenuto tramite l'uso della funzione *fminsearch* di Matlab.
Infine nel Capitolo 6 vi è un'ulteriore applicazione dell'algoritmo Nelder Mead, stavolta riguardante un problema reale, quale lo studio clinico RSA.

Capitolo 1

L'algoritmo Nelder Mead

1.1 Introduzione

Un problema di programmazione matematica assume la forma generale

$$min_{x \in X} f(x) \tag{1.1}$$

dove

- $f : \mathbb{R}^n \to \mathbb{R}$ è la funzione obiettivo;
- $X \subseteq \mathbb{R}^n$ è l'insieme ammissibile delle soluzioni.

Nei problemi vincolati risulta $X \subset \mathbb{R}^n$, in quelli non vincolati si ha $X = \mathbb{R}^n$. Molti problemi di ottimizzazione derivanti da applicazioni reali sono caratterizzati dal fatto che l'espressione analitica della funzione obiettivo e/o dei vincoli non è nota. Tale situazione si verifica, ad esempio, quando sistemi fisici complessi vengono descritti, analizzati e controllati ottimizzando risultati di simulazioni al computer. La valutazione di $f(x)$ può, per esempio, essere il risultato di una misura sperimentale o di una simulazione stocastica, con la sottostante forma analitica di f sconosciuta. Anche se la funzione oggetto f è nota in forma analitica, codificare le sue derivate può richiedere molto tempo o essere impraticabile. Questo tipo di problemi possono essere risolti, in linea di principio, approssimando il gradiente (e possibilmente l'Hessiano) [14] usando le differenze finite ma, nonostante questo approccio con le differenze finite sia efficace in alcune applicazioni, non può essere considerato una tecnica generale per l'ottimizzazione senza derivate perché il numero di valutazioni della funzione richiesto può essere eccessivo e l'approccio può essere inaffidabile in presenza di rumore. A causa di questi inconvenienti, sono stati sviluppati vari algoritmi che non tentano di approssimare il gradiente. Piuttosto, essi utilizzano i valori

della funzione in un insieme di punti di campionamento per determinare una nuova iterata con altri mezzi.
In questo contesto quindi le grandezze necessarie al processo di ottimizzazione vengono calcolate mediante simulazioni ripetute ed ogni simulazione può coinvolgere diversi programmi tra loro indipendenti. I dati così ottenuti vengono poi ulteriormente processati per calcolare la funzione obiettivo e/o i vincoli. E' chiaro che in questo caso non è possibile (o comunque richiede un costo troppo elevato) calcolare le derivate, anche quando i fenomeni fisici considerati potrebbero essere rappresentati per loro natura tramite funzioni "smooth" (in generale i fenomeni naturali sono continui). L'interesse applicativo ha motivato quindi lo sviluppo di metodi di ottimizzazione che non richiedano la conoscenza delle derivate. Storicamente i primi metodi senza derivate sono stati introdotti già negli anni '50, ma sono poi stati abbandonati nei primi anni '70 per mancanza di un'analisi teorica rigorosa e per la bassa velocità di convergenza dimostrata. Solo ultimamente l'interesse della comunità scientifica si è risvegliato grazie ad una serie di articoli che dimostrano proprietà teoriche di convergenza globale per algoritmi senza derivate.
In generale, gli algoritmi (con e senza derivate) proposti in letteratura consentono soltanto la determinazione di punti stazionari di f, cioè di punti che soddisfano le condizioni di ottimalità del primo ordine e che quindi appartengono all'insieme

$$\Omega = \{x \in \mathbb{R}^n : \nabla f(x) = 0\}$$

Notiamo che non avendo a disposizione le derivate non è possibile verificare direttamente l'appartenenza di un punto all'insieme Ω tramite la valutazione del gradiente. Si deve quindi utilizzare un criterio diverso per stabilire se un punto appartenga all'insieme Ω o meno. La potenza dei metodi senza derivate è proprio quella di riuscire a garantire la convergenza a punti stazionari senza fare esplicitamente uso del valore del gradiente. Anche nel caso di algoritmi senza derivate lo schema generale di un algoritmo di minimizzazione è il seguente:

1. si fissa un punto iniziale $x_0 \in \mathbb{R}^n$;
2. Se $x_k \in \Omega$ stop
3. Si calcola una direzione di ricerca $d_k \in \mathbb{R}^n$;
4. Si calcola un passo $\alpha_k \in \mathbb{R}$ lungo d_k;
5. Si determina un nuovo punto $x_{k+1} = x_k + \alpha_k d_k$, si pone $k = k + 1$ e si ritorna al Passo 2.

Quello che caratterizza il singolo metodo è la scelta della direzione di ricerca $d_k \in \mathbb{R}^n$ e la scelta del passo $\alpha_k \in \mathbb{R}$.
I metodi senza derivate possono essere raggruppati in tre classi:

- metodi che fanno uso di approssimazioni alle differenze finite;
- metodi di ricerca diretta;
- metodi di modellizzazione.

In particolare, con il nome di metodi di ricerca diretta ("direct search") [7] si indica un insieme di metodi accomunati dall'idea di basare la minimizzazione sul confronto diretto dei valori della funzione obiettivo nei punti generati dall'algoritmo. All'interno di questa classe di metodi si trovano algoritmi euristici, che non hanno cioè proprietà teoriche, e algoritmi per cui invece si può dimostrare la convergenza a punti stazionari. In particolare, si possono distinguere tre classi di metodi di ricerca diretta:

- metodi di tipo simplesso;
- metodi di tipo pattern search;
- metodi di tipo line search.

L'algoritmo base dei metodi di tipo simplesso è l'algoritmo di Nelder e Mead [12], introdotto nel 1965 e rimasto uno dei più popolari algoritmi senza derivate per l'efficienza dimostrata soprattutto per problemi di piccole dimensioni. L'algoritmo Nelder Mead non deve essere confuso con il (probabilmente) più famoso algoritmo del simplesso di Dantzig di programmazione lineare; entrambi gli algoritmi impiegano una sequenza di simplessi ma sono comunque completamente differenti e non correlati- in particolare, il metodo Nelder Mead è destinato per l'ottimizzazione non vincolata.
L'algoritmo Nelder Mead è particolarmente popolare nei campi della chimica, ingegneria chimica e medicina. Il metodo Nelder Mead tenta di minimizzare una funzione non lineare di n variabili reali utilizzando solo valori della funzione, senza alcuna informazione sulla derivata (esplicita o implicita). Per fare ciò ad ogni passo mantiene una figura geometrica, detta *simplesso*, che soddisfa la seguente definizione:

Definizione 1.1. Si definisce simplesso S l'involucro convesso di $n+1$ punti $\{x_i \in \mathbb{R}^n\}_{i=1}^{n+1}$ (detti vertici del simplesso), cioè

$$S = \{y \in \mathbb{R}^n : y = \sum_{i=1}^{n+1} \lambda_i x_i, \lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i = 1\}$$

Un simplesso S si dice non singolare se gli n vettori $\{x_2 - x_1, ..., x_{n+1} - x_1\}$ sono tra loro linearmente indipendenti.

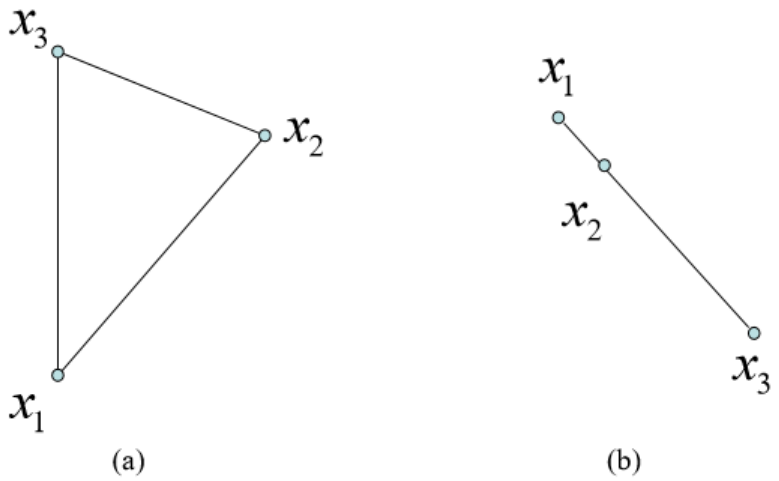


Figura 1.1

In figura (a) e (b) sono riportati rispettivamente un esempio di simplesso non singolare e uno di simplesso singolare.
La scelta della figura geometrica del simplesso è dovuta principalmente a due motivi: la capacità del simplesso di adattare la sua forma all'andamento nello spazio della funzione obiettivo deformandosi (allungandosi o schiacciandosi), e il fatto che richiede la memorizzazione di soli $n + 1$ punti.
Ogni iterazione di un metodo di ricerca diretta basato sul simplesso inizia con un simplesso, specificato dai suoi $n+1$ vertici ed i valori delle funzioni associati. Vengono calcolati uno o più punti di prova ed i rispettivi valori della funzione, e l'iterazione termina con un nuovo simplesso (differente) tale che i valori della funzione nei suoi vertici soddisfano alcune forme di condizione di discesa rispetto al simplesso precedente. Tra tali algoritmi, l'algoritmo Nelder Mead è particolarmente parsimonioso nelle valutazioni della funzione ad ogni iterazione, dato che in pratica richiede tipicamente solo uno o due valutazioni della funzione per costruire un nuovo simplesso. Sorprendentemente, non è stata pubblicata alcuna analisi teorica riguardante esplicitamente l'algoritmo originale Nelder Mead negli oltre 30 anni dalla sua pubblicazione. In sostanza sono stati dimostrati risultati di non convergenza, anche se nel 1985 Woods studiò un algoritmo Nelder Mead modificato applicato ad una funzione strettamente convessa. I pochi fatti conosciuti circa l'algoritmo originale Nelder Mead consistono principalmente di risultati negativi. Woods mostrò un esempio non convesso in due dimensioni per il quale l'algoritmo Nelder Mead converge a un

punto non minimizzante. Più recentemente, McKinnon [11] ha fornito una famiglia di funzioni strettamente convesse e una configurazione iniziale in due dimensioni per cui tutti i vertici nel metodo Nelder Mead convergono ad un punto non minimizzante.

Il quadro teorico di altri metodi di ricerca diretta è molto più chiaro. Torczon [18] ha dimostrato che gli algoritmi di ricerca pattern convergono ad un punto stazionario quando vengono applicati ad una generale funzione liscia in n dimensioni. I metodi di ricerca pattern, tra cui i metodi di ricerca multidirezionali, mantengono l'indipendenza lineare uniforme dei bordi del simplesso (cioè gli angoli diedri sono uniformemente delimitati distanti zero e π) e richiedono solo semplice diminuzione del valore migliore della funzione ad ogni iterazione. Rykov ha introdotto diversi metodi di ricerca diretta che convergono ad un minimizzante per funzioni strettamente convesse. Nei metodi proposti da Tseng è richiesta una condizione di "discesa fortificata" - più forte di una semplice discesa- insieme all'indipendenza lineare uniforme dei bordi del simplesso. A seconda di un parametro specificato da utente, i metodi di Tseng possono coinvolgere solo un piccolo numero di valutazioni funzionali ad ogni data iterazione e sono mostrati convergere verso un punto stazionario per funzioni generali non rumorose in n dimensioni.

Analisi di convergenza pubblicate dei metodi di ricerca diretta basati su simplessi impongono uno o entrambi i seguenti requisiti: (i) i bordi del simplesso rimangono uniformemente linearmente indipendenti ad ogni iterazione; (ii) una condizione di discesa più forte della semplice diminuzione è soddisfatta ad ogni iterazione. In generale, l'algoritmo Nelder Mead non riesce ad avere entrambe queste proprietà; le difficoltà risultanti nell'analisi possono spiegare la mancanza di vecchia data di risultati di convergenza.

Poiché il metodo Nelder Mead è così ampiamente usato in diversi contesti applicativi per risolvere importanti problemi di ottimizzazione, riteniamo che le sue proprietà teoriche devono essere capite più appieno possibile. Saranno quindi presentati risultati di convergenza in uno e due dimensioni per l'algoritmo Nelder Mead originale applicato a funzioni strettamente convesse con insiemi di livello limitati. Il nostro approccio è quello di considerare l'algoritmo Nelder Mead come un sistema dinamico discreto le cui iterazioni sono "guidate" dai valori della funzione. Combinata con la stretta convessità della funzione, questa interpretazione implica restrizioni sulle sequenze consentite delle mosse dell'algoritmo, da cui possono essere derivati risultati di convergenza. I risultati principali che analizzeremo sono i seguenti:

1. In dimensione 1, il metodo Nelder Mead converge a un minimizzzante, e la convergenza è infine lineare in M-passi quando il parametro di riflessione ρ=

1.

2. In dimensione 2, i valori della funzione in tutti i vertici del simplesso nell'algoritmo Nelder Mead standard convergono allo stesso valore.

3. In dimensione 2, simplessi nell'algoritmo Nelder Mead standard hanno diametri convergenti a zero.

Si noti che il risultato 3 non asserisce che i simplessi convergono in un unico punto x_*. Nessun esempio è noto in cui le iterate falliscono nel convergere verso un unico punto, ma la questione non è risolta.
Date tutte le carenze note e i fallimenti dell'algoritmo Nelder Mead, ci si potrebbe chiedere perché è così straordinariamente popolare. Vi possono essere diverse risposte; innanzitutto, in molte applicazioni, ad esempio nel controllo dei processi industriali, uno semplicemente vuole trovare valori di parametri che migliorano in qualche misura le prestazioni; l'algoritmo Nelder Mead produce tipicamente un miglioramento significativo dalle prime poche iterazioni. In secondo luogo, vi sono importanti applicazioni in cui una valutazione della funzione è estremamente costosa o consuma tempo, ma le derivate non possono essere calcolate. In tali problemi, un metodo che richiede almeno n valutazioni funzionali ad ogni iterazione è troppo costoso o troppo lento. Quando riesce, il metodo Nelder Mead tende a richiedere sostanzialmente un minor numero di valutazioni della funzione di queste alternative, e il suo relativo "Caso-migliore di efficienza" spesso supera la mancanza di una teoria della convergenza. In terzo luogo, il metodo Nelder Mead è attraente perché i suoi passaggi sono facili da spiegare e semplici da programmare.
L'algoritmo Nelder Mead [12] fu proposto come metodo per minimizzare una funzione di valori reali $f(x)$ per $x \in \mathbb{R}^n$. Per definire un metodo Nelder Mead completo devono essere specificati quattro parametri scalari, i coefficienti di: *riflessione* (ρ), *espansione* (χ), *contrazione* (γ) e *restringimento* (σ). Secondo il documento Nelder Mead originale, i parametri devono soddisfare

$$\rho > 0,\ \chi > 1,\ \chi > \rho,\ 0 < \gamma < 1,\ e\ 0 < \sigma < 1 \tag{1.2}$$

(La relazione $\chi > \rho$, sebbene non dichiarata esplicitamente nel documento originale, è implicita nella descrizione dell'algoritmo e nella terminologia). Le scelte ormai universali usate nell'algoritmo standard sono:

$$\rho = 1,\ \chi = 2,\ \gamma = \frac{1}{2},\ e\ \sigma = \frac{1}{2} \tag{1.3}$$

Assumiamo le condizioni generali (1.2) per il caso unidimensionale ma ci restringeremo al caso standard (1.3) nell'analisi bidimensionale.

1.2 Esposizione dell'algoritmo

L'algoritmo tiene traccia degli $n+1$ vertici del simplesso corrente con i rispettivi valori della funzione obiettivo e ad ogni iterazione tenta di generare un nuovo simplesso sostituendo il punto a cui corrisponde il valore massimo di f con un nuovo punto, scelto in maniera opportuna, in cui la funzione obiettivo abbia valore inferiore. In particolare, all'inizio della $k-$esima iterazione, con $k \geq 0$, si ha un simplesso non degenere Δ_k, insieme ai suoi $n+1$ vertici, ognuno dei quali è un punto di $\mathbb{R}^n$. Si presuppone sempre che l'iterazione k inizi ordinando ed etichettando questi vertici come $x_1^{(k)}, ..., x_{n+1}^{(k)}$, in modo che

$$f_1^{(k)} \leq f_2^{(k)} \leq ... \leq f_{n+1}^{(k)}, \tag{1.4}$$

dove $f_i^{(k)}$ denota $f(x_i^{(k)})$. La k-esima iterazione genera quindi un insieme di $n+1$ vertici che definiscono un simplesso differente per la successiva iterazione, così che $\Delta_{k+1} \neq \Delta_k$. Poichè cerchiamo di minimizzare f, ci riferiamo ad $x_1^{(k)}$ come il miglior punto o vertice, ad $x_{n+1}^{(k)}$ come il peggiore, e ad $x_n^{(k)}$ come il "secondo peggiore". Similmente, ci riferiamo a $f_{n+1}^{(k)}$ come il peggiore valore della funzione, e così via.
Viene specificata una singola iterazione generica, omettendo l'apice k per evitare confusione. Il risultato di ogni iterazione è o un singolo nuovo vertice - il punto accettato- che sostituisce x_{n+1} nell'insieme dei vertici per la successiva iterazione, o, se viene effettuato un restringimento, un insieme di n nuovi punti che, insieme ad x_1, formano il simplesso dell'iterazione successiva.

1.2.1 Un'iterazione dell'algoritmo Nelder Mead

Ad ogni iterazione dell'algoritmo, i possibili passi effettuati possono essere schematizzati come segue:

1. **Ordinamento** Calcolare la funzione f negli $n+1$ vertici ed ordinarli in modo tale che soddisfino $f_1 \leq f_2 \leq ... \leq f_{n+1}$,

2. **Riflessione** Calcolare il punto di riflessione x_r da

$$x_r = \overline{x} + \rho(\overline{x} - x_{n+1}) = (1+\rho)\overline{x} - \rho x_{n+1} \tag{1.5}$$

dove $\overline{x} = \sum_{i=1}^{n} x_i/n$ è il baricentro degli n punti migliori (tutti i vertici eccetto x_{n+1}). Calcolare $f_r = f(x_r)$.
Se $f_1 \leq f_r < f_n$, accettare il punto riflesso x_r e terminare l'iterazione.

3. **Espansione** Se $f_r < f_1$, calcolare il punto di espansione x_e,

$$x_e = \overline{x} + \chi(x_r - \overline{x}) = \overline{x} + \rho\chi(\overline{x} - x_{n+1}) = (1 + \rho\chi)\overline{x} - \rho\chi x_{n+1}, \qquad (1.6)$$

e calcolare $f_e = f(x_e)$. Se $f_e < f_r$, accettare x_e e terminare l'iterazione; altrimenti (se $f_e \geq f_r$), accettare x_r e terminare l'iterazione.

4. **Contrazione** Se $f_r \geq f_n$,
eseguire una contrazione tra $\overline{x}$ ed il migliore tra x_{n+1} e x_r.

   - **Esterna** Se $f_n \leq f_r < f_{n+1}$ eseguire una contrazione esterna: calcolare

$$x_c = \overline{x} + \gamma(x_r - \overline{x}) = \overline{x} + \gamma\rho(\overline{x} - x_{n+1}) = (1 + \rho\gamma)\overline{x} - \rho\gamma x_{n+1} \qquad (1.7)$$

   e calcolare $f_c = f(x_c)$. Se $f_c \leq f_r$, accettare x_c e terminare l'iterazione, altrimenti andare al passo 5 (effettuare un restringimento).

   - **Interna** Se $f_r \geq f_{n+1}$, effettuare una contrazione interna: calcolare

$$x_{cc} = \overline{x} - \gamma(\overline{x} - x_{n+1}) = (1 - \gamma)\overline{x} + \gamma x_{n+1} \qquad (1.8)$$

   e calcolare $f_{cc} = f(x_{cc})$. Se $f_{cc} < f_{n+1}$, accettare x_{cc} e terminare l'iterazione; altrimentiandare al passo 5 (effettuare un restringimento).

5. **Effettuare un passo di restringimento**. Calcolare f negli n punti
$v_i = x_1 + \sigma(x_i - x_1)$ $i = 2, ..., n + 1$. I vertici (disordinati) del simplesso all'iterazione successiva consistono di $x_1, v_2, ..., v_{n+1}$.

Si vede dall'algoritmo che l'idea è quella di cercare di espandere il simplesso se si trovano valori buoni della funzione obiettivo e contrarlo se non se ne trovano.

Le figure 1.2 e 1.3 mostrano gli effetti di riflessione, espansione, contrazione e restringimento per un simplesso in due dimensioni (ovvero un triangolo), usando i coefficienti standard $\rho = 1,\ \chi = 2,\ \gamma = 1/2,\ e\ \sigma = 1/1$. Osserviamo che, eccetto nel restringimento, l'unico nuovo vertice giace sempre sulla linea congiungente $\overline{x}$ e x_{n+1}. Per di più, è visivamente evidente che la forma del simplesso subisce un notevole cambiamento durante un'espansione o contrazione con i coefficienti standard.

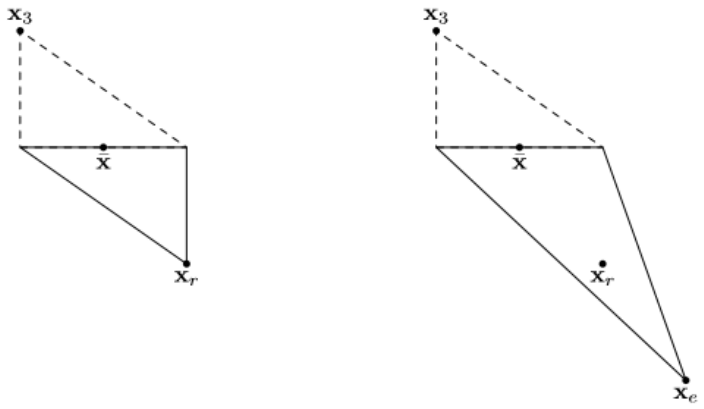


Figura 1.2: I simplessi Nelder Mead dopo un passo di riflessione e un'espansione. Il simpleso iniziale è quello tratteggiato.

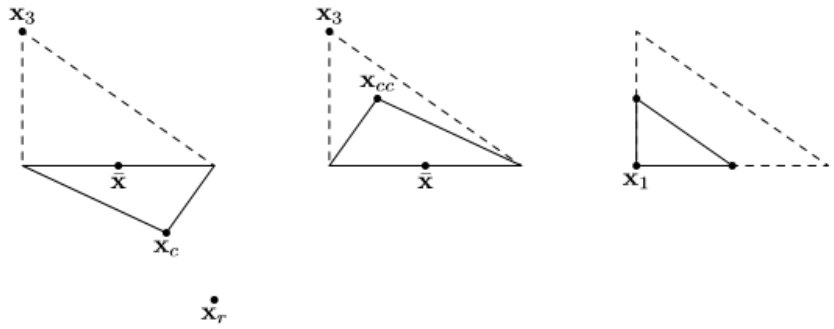


Figura 1.3: I simplessi Nelder Mead dopo una contrazione esterna, una interna, ed un passo di restringimento. Il simplesso iniziale è sempre quello tratteggiato.

Il documento Nelder Mead originale [12] non descriveva come ordinare i punti nel caso di valori delle funzioni uguali. Adottiamo quindi le seguenti regole tie-breaking, che assegnano al nuovo vertice il più alto indice possibile compatibile con la relazione $f(x_1^{(k+1)}) \leq f(x_2^{(k+1)}) \leq ... \leq f(x_{n+1}^{(k+1)})$.

Regola di ordinamento senza restringimento. Quando si verifica un passo di non restringimento, il vertice peggiore $x_{n+1}^{(k)}$ viene scartato. Il punto accettato creato durante l'iterazione k, indicato con $v^{(k)}$, diventa un nuovo vertice e prende la posizione $j+1$ nei vertici di Δ_{k+1}, dove

$$j = max_{0 \leq l \leq n}\{l \mid f(v^{(k)}) < f(x_{l+1}^{(k)})\};$$

tutti gli altri vertici mantengono il loro relativo ordinamento dall'iterazione k.

Regola di ordinamento con restringimento. Se si verifica un passo di restringimento, l'unico vertice trasportato da Δ_k a Δ_{k+1} è $x_1^{(k)}$. Solo una regola tie-breaking è specificata, per il caso in cui $x_1^{(k)}$ ed uno o più dei nuovi punti sono legati con il punto migliore: se

$$min\{f(v_2^{(k)}), ..., f(v_{n+1}^{(k)})\} = f(x_1^{(k)}),$$

allora $x_1^{(k+1)} = x_1^{(k)}$. Al di là di questo, qualunque regola venga utilizzata per definire l'ordinamento originale può essere applicata dopo un restringimento.

Definiamo il *cambio di indice* k dell'iterazione k come il più piccolo indice di un vertice che differisce tra le iterazioni k e $k+1$:

$$k^* = min\{i \mid x_i^{(k)} \neq x_i^{(k+1)}\} \tag{1.9}$$

(Le regole tie-breaking sono necessarie per definire un unico valore di k^*.) Quando l'algoritmo Nelder Mead termina al passo 2, $1 < k^* \leq n$; quando termina al passo 3, $k^* = 1$; al passo 4, $1 \leq k^* \leq n+1$, e quando termina al passo 5, $k^* = 1$ o 2. L'affermazione "x cambia" significa che j è l'indice di cambiamento all'iterazione rilevante.

Le regole e definizioni date finora implicano che, per un'iterazione di non restringimento,

$$\begin{aligned} f_j^{(k+1)} = f_j^{(k)} \quad e \quad x_j^{(k+1)} = x_j^{(k)}, \quad j < k^*; \\ f_{k^*}^{(k+1)} = f_{k^*}^{(k)} \quad e \quad x_{k^*}^{(k+1)} \neq x_{k^*}^{(k)}; \\ f_j^{(k+1)} = f_{j-1}^{(k)} \quad e \quad x_j^{(k+1)} = x_{j-1}^{(k)}, \quad j > k^*. \end{aligned} \tag{1.10}$$

1.3 Notazione matriciale

Il simplesso generato dal metodo passo dopo passo può essere rappresentato anche attraverso una matrice, notazione che risulta più conveniente per descrivere meglio il comportamento ad ogni iterazione [9]. Nello specifico, il simplesso Δ_k può essere rappresentato come una matrice nx$(n+1)$ le cui colonne sono i vertici

$$\Delta_k = (x_1^{(k)} \cdots x_{n+1}^{(k)}) = (B_k \, x_{n+1}^{(k)}), \quad dove \; B_k = (x_1^{(k)} \cdots x_n^{(k)}).$$

Per qualsiasi simplesso Δ_k in $\mathbb{R}^n$, definiamo M_k come la matrice nxn la cui colonna j-esima rappresenta il "bordo "di Δ_k tra $x_j^{(k)}$ e $x_{n+1}^{(k)}$:

$$M_k \equiv (x_1^{(k)} - x_{n+1}^{(k)} \; x_2^{(k)} - x_{n+1}^{(k)} \cdots x_n^{(k)} - x_{n+1}^{(k)}) = B_k - x_{n+1}^{(k)} e^T, \tag{1.11}$$

dove $e = (1, 1, ..., 1)^T$. Il volume n-dimensionale di Δ_k è dato da

$$vol(\Delta_k) = \frac{|det(M_k)|}{n!} \tag{1.12}$$

Un simplesso Δ_k è non degenere se M_k è non singolare o, equivalentemente, se $vol(\Delta_k) > 0$. Il volume del simplesso dipende ovviamente solo dalle coordinate dei vertici, non dal loro ordinamento. Definiamo il diametro di Δ_k come

$$diam(\Delta_k) = max_{i \neq j} \|x_i^{(k)} - x_j^{(k)}\|,$$

dove $\|\cdot\|$ indica la norma 2. Durante un'iterazione di non restringimento, la funzione viene valutata solo nei punti di prova della forma

$$z^{(k)}(\tau) := \overline{x}^{(k)} + \tau(\overline{x}^{(k)} - x_{n+1}^{(k)}) = (1+\tau)\overline{x}^{(k)} - \tau x_{n+1}^{(k)} \tag{1.13}$$

dove il coefficiente τ ha uno dei quattro possibili valori:

$$\begin{array}{llll} \tau = \rho & (riflessione) & \tau = \rho\chi & (espansione); \\ \tau = \rho\gamma & (contrazione \; esterna) & \tau = -\gamma & (contrazione \; interna). \end{array} \tag{1.14}$$

In un passo di non restringimento, l'unico punto accettato è uno dei punti di prova, e denotiamo con τ_k il coefficiente associato al punto accettato all'iterazione k. Così il nuovo vertice $v^{(k)}$ prodotto durante l'iterazione k, che sostituirà $x_{n+1}^{(k)}$, è dato da $v^{(k)} = z^{(k)}(\tau_k)$. A volte chiamiamo τ_k il tipo di spostamento per una iterazione k di non contrazione. Durante la k-esima iterazione Nelder Mead, (1.13) mostra che

ogni punto di prova (riflessione, espansione, contrazione) può essere scritto come

$$z^{(k)}(\tau) = \Delta_k t(\tau), \quad dove \quad t(\tau) = (\frac{1+\tau}{n}, ..., \frac{1+\tau}{n}, -\tau)^T \tag{1.15}$$

A seguito della k-esima iterazione Nelder Mead, i vertici (non ordinati) del simplesso successivo sono le colonne di $\Delta_k S_k$, dove S_k è una matrice $(n+1)$x$(n+1)$ data da

$$\begin{pmatrix} I_n & \frac{(1+\tau_k)}{n} e \\ 0^T & -\tau_k \end{pmatrix}$$

per un passo di tipo τ e da

$$\begin{pmatrix} 1 & (1-\sigma)e^T \\ 0 & \sigma I_n \end{pmatrix}$$

per un passo di restringimento, con 0 colonna n-dimensionale di zeri ed I_n matrice identità n-dimensionale. Dopo essere ordinati all'inizio dell'iterazione $k+1$, i vertici di Δ_{k+1} soddisfano

$$\Delta_{k+1} = \Delta_k T_k, \quad con \quad T_k = S_k P_k, \tag{1.16}$$

dove P_k è una matrice di permutazione scelta per far rispettare le regole di ordinamento (così P_k dipende dai valori della funzione ai vertici).
Il simplesso aggiornato Δ_{k+1} ha un interno disgiunto da Δ_k per una riflessione, un'espansione o una contrazione esterna, mentre $\Delta_{k+1} \subseteq \Delta_k$ per una contrazione interna o una riduzione.
Con la forma di un simplesso non degenere, intendiamo la sua classe di equivalenza secondo similarità, cioè Δ e $\lambda\Delta$ hanno la stessa forma quando $\lambda > 0$. La forma di una simplesso è determinata dai suoi angoli, o equivalentemente dai valori singolari della matrice associata M (1.11), dopo il ridimensionamento in modo che Δ ha volume unitario. Il metodo di Mead Nelder è stato volutamente progettato con l'idea che le forme dei simplessi dovrebbero "adattarsi alle caratteristiche del paesaggio locale" [12]. I movimenti Nelder Mead apparentemente consentono a qualsiasi forma di simplesso di essere approssimata.

1.4 Proprietà dell'algoritmo Nelder Mead

1.4.1 Risultati generali

Dalla definizione dell'algoritmo e dalla descrizione dei passi possibili ad ogni ierazione descritti nella sezione precedente, si nota che possono essere tratte subito le segenti proprietà:

1. Un'iterazione Nelder Mead richiede una valutazione della funzione quando l'iterazione termina al passo 2, due valutazioni della funzione quando si verifica la cessazione al passo 3 o 4 ed $n+2$ valutazioni della funzione se si verifica un passo di restringimento.

2. Il passo di riflessione è così chiamato perché il punto di riflessione x_r (1.5) è una riflessione (in scala) del punto peggiore x_{n+1} intorno al punto $\overline{x}$ sulla linea attraverso x_{n+1} e $\overline{x}$. Si tratta di una vera e proprio riflessione su questa linea, quando $\rho = 1$, che è la scelta standard per il coefficiente di riflessione.

3. Per funzioni generali, una fase di restringimento può plausibilmente portare ad un incremento in ogni vertice valore della funzione tranne f_1, cioè, è possibile che $f_i^{(k+1)} > f_i^{(k)}$ per $2 \leq i \leq n+1$. Inoltre, osserviamo che con una contrazione esterna, l'algoritmo fa un passo di restringimento se $f(x_c) > f(x_r)$, anche se è già stato trovato un nuovo punto x_r che migliora rigorosamente il vertice peggiore, poiché $f(x_r) < f(x_{n+1})$.

4. Nel passo di espansione, il metodo descritto nel documento originale Nelder Mead accetta x_e se $f(x_e) < f_1$ e accetta x_r altrimenti. La pratica standard oggi (che seguiamo) accetta il migliore di x_r e x_e se entrambi danno un miglioramento rispetto x_1.

E' comunemente (e correttamente) presunto che la non degeneranza del simplesso iniziale Δ_0 implica non degeneranza di tutti i successivi simplessi Nelder Mead. Per costruzione, ogni punto di prova (1.13) nel metodo Nelder Mead si trova rigorosamente al di fuori della faccia definita dagli n migliori vertici, lungo la congiungente il vertice peggiore con il baricentro di tale faccia. Se si verifica un'iterazione di non restringimento, il vertice peggiore viene sostituito da uno dei punti di prova. Se si verifica un'iterazione di restringimento, ogni vertice corrente, tranne il migliore, è sostituito da un punto che trova una frazione del passo al migliore vertice corrente. In entrambi i casi è evidente dalla geometria che il nuovo simplesso deve essere non degenere. Tuttavia, una prova della non degeneranza, basata su un risultato utile sui volumi dei simplessi successivi, è data dal seguente [9]:

Lemma 1.1. *(Volume e non degeneranza dei simplessi Nelder Mead.)*

1. *Se il simplesso iniziale Δ_0 è non degenere, lo sono anche tutti i simplessi successivi Nelder Mead.*

2. *A seguito di una fase di non restringimento di tipo τ, $vol(\Delta_{k+1}) = \mid vol(\Delta_k) \mid$.*

3. *A seguito di una fase di restringimento nell'iterazione k,*
   $vol(\Delta_{k+1}) = \sigma^n vol(\Delta_k)$.

Tale lemma mostra che, in qualsiasi dimensione, una passo di riflessione con $\rho = 1$ conserva il volume. La scelta $\rho = 1$ è geometricamente naturale, dato che un passo di riflessione è poi una vera e propria riflessione. Una simplesso riflesso con $\rho = 1$ è necessariamente congruente al simplesso originale per $n = 1$ ed $n = 2$, ma ciò non è più vero per $n \geq 3$.
Si noti che, anche se i simplessi Nelder Mead sono non degeneri in aritmetica esatta, non vi è, in generale, nessun limite superiore sulla cond (M_k). Infatti, l'algoritmo permette alla cond(M_k) di diventare arbitrariamente grande, come avviene nell'esempio di McKinnon che verrà analizzato in seguito [11].

Lemma 1.2. *(invarianza affine) Il metodo NM è invariante per moti affini di* $\mathbb{R}^n$*, cioè, sotto un cambiamento di variabile* $\phi(x) = Ax + b$ *in cui* A *è invertibile, nel senso seguente: quando si minimizza* $f(x)$ *a partire da un simplesso* Δ_0*, la sequenza completa dei passi NM e i valori della funzione è la stessa di quando minimizziamo la funzione* $\tilde{f}(z) = f(\phi(z))$ *con simplesso iniziale* $\tilde{\Delta_0}$ *definito da*

$$\tilde{\Delta_0} = \phi^{-1}(\Delta_0) = A^{-1}(\Delta_0) - A^{-1}b$$

Dimostrazione. Nei vertici di $\tilde{\Delta_0}$, $\tilde{f}(\tilde{x_i}^{(0)}) = f(x_i^{(0)})$. Si procede per induzione, ipotizzando per semplicità che b = 0. Se $\tilde{\Delta_k} = A^{-1}\Delta_k$, e $\tilde{f}(\tilde{x_i}^{(k)}) = f(x_i^{(k)})$ per $1 \leq i \leq n+1$, allora la relazione (1.15) mostra che i punti di prova generati da $\tilde{\Delta_k}$ soddisfano $\tilde{z}(\tau) = A^{-1}z(\tau)$, il che significa che $\tilde{f}(\tilde{z}(\tau)) = f(z(\tau))$. La matrice T_k di (1.16) sarà quindi la stessa per entrambi Δ_k e $\tilde{\Delta_k}$, in modo che $\tilde{\Delta}_{k+1} = A^{-1}\Delta_{k+1}$. Segue che $\tilde{f}(\tilde{x}_i^{(k+1)}) = f(x_i^{(k+1)})$. Un discorso analogo si applica quando $b \neq 0$. □

Utilizzando il Lemma 1.2, possiamo ridurre lo studio dell'algoritmo Nelder Mead per una generale funzione quadratica strettamente convessa su $\mathbb{R}^n$ allo studio di $f(x) = \|x\|^2 = x_1^2 + ... + x_n^2$ Il prossimo lemma riassume diversi semplici risultati.

Lemma 1.3. *Sia* f *una funzione delimitata inferiormente su* $\mathbb{R}^n$*. Quando l'algoritmo NM è applicato per minimizzare* f*, iniziando con un simplesso non degenere* Δ_0*, allora*

1. *la sequenza di* $\{f_1^{(k)}\}$ *converge sempre;*
2. *ad ogni iterazione di non restringimento k,* $f_i^{(k+1)} \leq f_i^{(k)}$ *per* $1 \leq i \leq n+1$*, con disuguaglianza stretta per almeno un valore di i;*
3. *se ci sono solo un numero finito di iterazioni di restringimento, allora*

(a) ciascuna sequenza $\{f_i^{(k)}\}$ *converge come* $k \to \infty$ *per* $1 \leq i \leq n+1$,

(b) $f_i^* \leq f_i^{(k)}$ *per* $1 \leq i \leq n+1$ *e tutti i* k, *dove* $f_i^* = \lim_{k\to\infty} f_i^{(k)}$,

(c) $f_1^* \leq f_2^* \leq ... \leq f_{n+1}^*$;

4. *se ci sono solo un numero finito di iterazioni di non restringimento, allora tutti i vertici del simplesso convergono ad un unico punto.*

Analizziamo ora l'algoritmo Nelder Mead nel caso in cui si verificano solo passi di non restringimento (nella pratica i passi di restringimento si verificano molto raramente, come osservato in diversi esperimenti numerici, e nel lemma 1.6 mostreremo che quando il metodo viene applicato a funzioni strettamente convesse non vengono adottati passi di restringimento). Partendo dal presupposto che non vengono effettuati passi di restringimento, il prossimo lemma fornisce un'importante proprietà degli $n+1$ vertici limitanti i valori della funzione la cui esistenza è verificata nella parte (3) del Lemma 1.3.

Lemma 1.4. *(Rottura di convergenza.) Supponiamo che la funzione* f *è limitata inferiormente su* $\mathbb{R}^n$, *che l'algoritmo Nelder Mead è applicato ad* f *iniziando con un simplesso iniziale non degenere* Δ_0, *e che non si verifichino passaggi di restringimento. Se vi è un numero intero* j, $1 \leq j \leq n$, *per cui*

$$f_j^* < f_{j+1}^*, \quad dove \quad f_i^* = lim_{k\to\infty} f_j^{(k)}, \tag{1.17}$$

Allora c'è un indice di iterazione K *tale che per tutti i* $k \geq K$, *il cambiamento dell'indice soddisfa*

$$k^* > j \tag{1.18}$$

Cioè, i primi j *vertici di tutti i simplessi rimangono fissi dopo l'iterazione* K. *(Ci riferiamo alla proprietà* (1.17) *come rottura della convergenza per il vertice* j*)*

Dimostrazione. Il lemma è dimostrato per assurdo. Dall'ipotesi (1.17), $f_j^* + \delta = f_{j+1}^*$ per qualche $\delta > 0$. Sia $\epsilon > 0$ tale che $\delta - \epsilon > 0$. Dal momento che $f_j^* = \lim_{k\to\infty} f_j^{(k)}$, esiste K tale che per ogni $k \geq K$, $f_j^{(k)} - \epsilon \leq f_j^*$. Allora, per ogni $k \geq K$,

$$f_j^{(k)} < f_j^{(k)} - \epsilon + \delta \leq f_j^* + \delta = f_{j+1}^*$$

Ma, dal Lemma 1.3, parte (3), per qualsiasi indice l, $f_{j+1}^* \leq f_{j+1}^{(l)}$. Pertanto, per ogni $k \geq K$ ed ogni l,

$$f_j^{(k)} < f_{j+1}^{(l)} \tag{1.19}$$

Ma se $k^* \leq j$ per ogni $k \geq K$, allora, utilizzando la terza relazione in (1.10), deve essere vero che $f_{j+1}^{(k+1)} = f_i^{(k)}$, che contraddice (1.19). Così $k^* > j$ per ogni $k \geq K$. □

Una conseguenza immediata del Lemma 1.4 è:

Corollario 1.5. *Si supponga che f sia limitata inferiormente su $\mathbb{R}^n$, l'algoritmo NM è applicato a partire da una simplesso iniziale non degenere Δ_0, e non si verificano passi di restringimento. Se l'indice di cambiamento è di 1 infinitamente spesso, cioè, il punto migliore cambia infinitamente molte volte, allora $f_1^* = \cdots = f_{n+1}^*$.*

1.4.2 Risultati per funzioni strettamente convesse

Oltre ai risultati già descritti, assumendo che f sia una funzione strettamente convessa se ne possono ottenere degli altri. A tal proposito si ricorda quindi la nozione di convessità.

Definizione 1.2. (Convessità rigorosa.) La funzione f è strettamente convessa su $\mathbb{R}^n$ se, per ogni coppia di punti y, z con $y \neq z$ ed ogni λ soddisfacente $0 < \lambda < 1$,

$$f(\lambda y + (1-\lambda)z) < \lambda f(y) + (1-\lambda) f(z). \tag{1.20}$$

Quando f è strettamente convessa su $\mathbb{R}^n$ e

$$c = \sum_{i=1}^{l} \lambda_i z_i, \quad dove \quad 0 < \lambda_i < 1 \quad e \quad \sum_{i=1}^{l} \lambda_i = 1,$$

$$allora \quad f(c) < \sum_{i=1}^{l} \lambda_i f(z_i) \quad e\ quindi \quad f(c) < max\{f(z_1), ..., f(z_l)\}. \tag{1.21}$$

Grazie a questa proprietà si può dimostrare che, quando si applica il metodo NM a una funzione strettamente convessa, non possono verificarsi passi di restringimento [17].

Lemma 1.6. *Si supponga che f sia strettamente convessa su $\mathbb{R}^n$ e che l'algoritmo NM venga applicato ad f con un simplesso iniziale non degenere Δ_0. Allora non saranno effettuati passi di restringimento.*

Dimostrazione. Passi di restringimento possono avvenire solo se l'algoritmo raggiunge il punto 4 della descrizione esposta sopra e non riesce ad accettare il punto di contrazione rilevante. Quando $n = 1$, $f(\overline{x}) = f_n$. Quando $n > 1$, l'applicazione di (1.21) ad $x_1, ..., x_n$ mostra che $f(\overline{x}) < f_n$.

Consideriamo una contrazione esterna, che è provata se $f_n \leq f_r < f_{n+1}$. Poiché il coefficiente di contrazione γ soddisfa $0 < \gamma < 1$, x_c definito da (1.7) è una combinazione convessa di $\overline{x}$ e il punto riflesso x_r. Così, da (1.21),

$$f(x_c) < max\{f(\overline{x}), f_r\}$$

Sappiamo che $f(\overline{x}) \leq f_n$ e $f_n \leq f_r$, cosicché $max\{f(\overline{x}), f_r\} = f_r$. Quindi $f\ (x_c) < f_r$, x_c sarà accettato, e non sarà preso un passo di restringimento.
Un discorso analogo vale per una contrazione interna, dal momento che $f_{n+1} \leq f_r$ ed x_{cc} è una combinazione convessa di $\overline{x}$ *e* x_{n+1}. □

Lemma 1.7. *Si supponga che f sia strettamente convessa su $\mathbb{R}^n$ e delimitata inferiormente. Se, in aggiunta alle proprietà $\rho > 0$ e $0 < \gamma < 1$, il coefficiente di riflessione ρ e il coefficiente di contrazione γ soddisfano $\rho\gamma < 1$, allora*

1. *$f_n^* = f_{n+1}^*$; e*
2. *ci sono infinitamente molte iterazioni per cui $x_n^{(k+1)} \neq x_n^{(k)}$.*

Dimostrazione. La dimostrazione è per assurdo. Si supponga che $f_n^* < f_{n+1}^*$. Dal Lemma 1.4, ciò significa che esiste un indice di iterazione K tale che l'indice di cambiamento $k^* = n+1$ per $k \geq K$. Senza perdita di generalità, possiamo prendere $K = 0$. Poiché $k^* = n+1$ per tutti $k \geq 0$, i migliori n vertici, che devono essere distinti, rimangono costanti per tutte le iterazioni; così il baricentro $\overline{x}^{(k)} = \overline{x}$, un vettore costante, e $f(x_n)$ è uguale al valore limite f_n^*. Poiché f è strettamente convessa, $f(\overline{x}) \leq f(x_n) = f_n^*$. (Questa disuguaglianza è stretta se $n > 1$.) L'indice di cambiamento sarà $n+1$ ad ogni iterazione solo se un punto di contrazione è accettato e diventa il nuovo punto peggiore. Pertanto, il vertice $x_{n+1}^{(k+1)}$ soddisfa una delle ricorrenze

$$x_{n+1}^{(k+1)} = (1+\rho\gamma)\overline{x} - \rho\gamma x_{n+1}^{(k)} \quad oppure \quad x_{n+1}^{(k+1)} = (1-\gamma)\overline{x} + \gamma x_{n+1}^{(k)} \tag{1.22}$$

Le forme omogenee di queste equazioni sono

$$y_{n+1}^{(k+1)} = -\rho\gamma y_{n+1}^{(k)} \quad oppure \quad y_{n+1}^{(k+1)} = \gamma y_{n+1}^{(k)} \tag{1.23}$$

Dal momento che $0 < \gamma < 1$ e $0 < \rho\gamma < 1$, abbiamo $\lim_{k\to\infty} y_{n+1}^{(k)} = 0$, in modo che le soluzioni di entrambe le equazioni in (1.23) sono pari a zero per $k \to \infty$. Ora dobbiamo solo trovare una soluzione particolare alle forme disomogenee di (1.22). Entrambe sono verificate dal vettore costante $\overline{x}$, in modo che le loro soluzioni generali

sono date da $x_{n+1}^{(k)} = y_{n+1}^{(k)} + \overline{x}$, dove $y_{n+1}^{(k)}$ soddisfa una delle relazioni (1.23). Poichè $\lim_{k\to\infty} y_{n+1}^{(k)} = 0$, segue che

$$\lim_{k\to\infty} x_{n+1}^{(k)} = x_{n+1}^{*} = \overline{x}, \quad con\ f_{n+1}^{*} = f(\overline{x}).$$

Ma sappiamo fin dall'inizio della dimostrazione che $f(\overline{x}) \leq f^{*}n$, il che significa che $f_{n+1}^{*} \leq f_{n}^{*}$. Il lemma 1.3, parte (3), mostra che questo può essere vero solo se $f_{n}^{*} = f_{n+1}^{*}$, che dà la parte (1). Il risultato della parte (2) è immediato perché abbiamo già dimostrato una contraddizione se esiste K tale che $x_1^{(k)}, ..., x_n^{(k)}$ rimane costante per $k \geq K$. □

Tale lemma dunque, utilizzando la definizione di una iterazione NM, il Lemma 1.4, e una leggera ulteriore limitazione relativa ai coefficienti di riflessione e contrazione, dimostra che il valore di limitazione della funzione peggiore e il successivo peggiore sono uguali (cosa che per $n = 1$ vale, senza l'ulteriore restrizione).
Nell'analizzare la convergenza, sappiamo dal Lemma 1.4 che, in caso di rottura di convergenza, esiste un indice j tale che tutti i vertici $\{x_i^{(k)}\}$ *per* $1 \leq i \leq j$, rimangono costanti da qualche punto. Se questo accade, il miglior punto $x_1^{(k)}$ non sarà cambiato, e quindi non possono verificarsi stadi di espansione. (Né passi di riflessione in cui viene rilevato in f_1 un miglioramento stretto). Per questo motivo, è interessante considerare un algoritmo Nelder Mead ristretto in cui non vengono adottate misure di espansione; l'analisi dell'algoritmo ristretto è più semplice perché sia $vol(\Delta_k)$ che $diam(\Delta_k)$ sono non crescenti se $\rho \leq 1$ [10]. Da ora in poi considereremo funzioni strettamente convesse f con insiemi di livello limitati. L'insieme di livello $\Gamma_\mu(f)$ viene definito come

$$\Gamma_\mu(f) = \{x : f(x) \leq \mu\} \tag{1.24}$$

Una funzione f ha insiemi di livello limitati se $\Gamma_\mu(f)$ è delimitata per ogni μ; questa restrizione esclude funzioni strettamente convesse come e^{-x}. Il punto di questa restrizione è che una funzione strettamente convessa con insiemi di livello limitati ha un unico minimizzante x_{min}.

Capitolo 2

Analisi di convergenza

2.1 Nelder Mead in dimensione 1 per funzioni strettamente convesse

Come è stato ricordato nel capitolo precedente, una funzione strettamente convessa con inisiemi di livello limitati è caratterizzata dall'avere un solo punto di minimo. Adesso verrà analizzato quindi il comportamento dell'algoritmo in dimensione 1 su funzioni di questo tipo, deducendone altre proprietà oltre quelle già descritte precedentemente.
Il comportamento dell'algoritmo Nelder Mead in dimensione 1 dipende non banalmente dai valori del coefficiente di riflessione ρ, di espansione χ e di contrazione γ. Il coefficiente di restringimento σ è irrilevante perché non si possono verificare passi di restringimento per una funzione strettamente convessa. Mostriamo che la convergenza ad x_{min} avviene sempre finché $\rho\chi \geq 1$ e che la convergenza è M-passi lineare quando $\rho = 1$. L'algoritmo non sempre converge al minimizzante x_{min} se $\rho\chi < 1$. Una caratteristica interessante dell'analisi è che la convergenza M-passo lineare può essere garantita, anche se possono occorrere infinitamente molti passi di espansione [9].

2.1.1 Proprietà speciali in una dimensione

In una dimensione, il "secondo peggiore" e il migliore vertice sono lo stesso punto, il che significa che il baricentro $\overline{x}^{(k)}$ è uguale a $x_1^{(k)}$ ad ogni iterazione. Un simplesso Nelder Mead è un segmento di linea, in modo che, data l'iterazione k di tipo τ_k,

$$diam(\Delta_{k+1}) = \mid \tau_k \mid diam(\Delta_k). \tag{2.1}$$

Così, nel caso particolare dei parametri standard $\rho = 1$ *e* $\chi = 2$, un passo di riflessione mantiene lo stesso diametro ed una fase di espansione raddoppia il diametro del simplesso. Per procedere con i diversi ordinamenti dei punti finali, usiamo la notazione $int(y, z)$ per indicare l'intervallo aperto con punti finali y e z (anche se $y > z$), con analoga notazione per intervalli chiusi o semiaperti. Il seguente lemma riassume tre importanti proprietà, di funzioni strettamente convesse in $\mathbb{R}^1$ con insiemi di livello limitati.

Lemma 2.1. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con un unico minimizzante x_{min}.*

1. *Siano y_1, y_2, y_3 tre punti distinti in modo tale che $y_2 \in int(y_1, y_3)$. Allora*

$$f(y_1) \geq f(y_2) \quad e \quad f(y_2) \leq f(y_3) \;\Rightarrow x_{min} \in int(y_1, y_3).$$

2. *Se $x_{min} \in int[y_1, y_3]$, allora $f(y_2 + \xi_2(y_1 - y_2)) > f(y_2 + \xi_1(y_1 - y_2))$ se $\xi_2 > \xi_1 \geq 1$.*

3. *f è continua.*

Una proprietà caratteristica del caso monodimensionale è che un'iterazione Nelder Mead non può mai terminare al passo 2 dell'algoritmo: o sarà fatta una contrazione (passo 4), o un passo di espansione (passo 3). Utilizzando la regola al passo 3 secondo cui dobbiamo accettare i migliori punti di riflessione ed espansione, un passo di riflessione sarà preso solo se $f_r < f_1$ *e* $f_e \geq f_r$.

2.1.2 Convergenza verso il minimizzante

Consideriamo prima i parametri generali Nelder Mead soddisfacenti (1.2) e dimostriamo che la condizione $\rho\chi \geq 1$ è necessaria per la convergenza globale dell'algoritmo ad x_{min}. Se $\rho\chi < 1$, il cosiddetto step di " espansione " in realtà riduce il diametro del simplesso, e i punti finali dell'intervallo NM possono spostarsi di una distanza di al più $diam(\Delta_0)/(1 - \rho\chi)$ dal vertice iniziale $x_1^{(0)}$. Così la convergenza ad x_{min} non si verificherà ogni volta che

$$\rho\chi < 1 \quad e \quad | \, x_{min} - x_1^{(0)} \, | > diam(\Delta_0)/(1 - \rho\chi).$$

Mostriamo poi il risultato generale che la condizione $\rho\chi \geq 1$, combinata con i requisiti (1.2), è sufficiente per la convergenza globale ad x_{min} dell'algoritmo NM in una dimensione.

Teorema 2.2. *(Convergenza del metodo unidimensionale Nelder Mead.) Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo NM venga applicato ad f con parametri soddisfacenti $\rho > 0,\ \chi > 1,\ \chi > \rho,\ \rho\chi \geq 1$ e $0 < \gamma < 1$, cominciando con un simplesso iniziale non degenere Δ_0. Allora entrambi i punti finali dell'intervallo NM convergono a x_{min}.*

La dimostrazione di questo teorema dipende da vari lemmi intermedi. In primo luogo dimostriamo che l'algoritmo Mead trova, in un numero finito di iterazioni, un "intervallo di incertezza" in cui il minimizzante deve giacere.

Lemma 2.3. *(Collegamento di x_{min}) Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead venga applicato ad f con un simplesso iniziale non degenere Δ_0 e che i coefficienti di riflessione ed espansione soddisfano $\rho > 0,\ \chi > 1,\ \chi > \rho,\ \rho\chi \geq 1$. Quindi c'è un più piccolo intero K soddisfacente*

$$K \leq \frac{\mid x_{min} - x_1^{(0)} \mid}{diam(\Delta_0)}, \quad tale\ che \quad f_2^{(K)} \geq f_1^{(K)}\ e\ f_1^{(K)} \leq f_e^{(K)}. \tag{2.2}$$

In questo caso, $x_{min} \in int(x_2^{(K)}, x_e^{(K)})$, e diciamo che x_{min} è collegato da $x_2^{(K)}$ e $x_e^{(K)}$.

Dimostrazione. Per evitare confusione, lasciamo cadere l'indice k e usiamo un primo per indicare quantità associate con l'iterazione $k+1$. Per definizione, $f_2 > f_1$, in modo che la prima disuguaglianza nella relazione "up-down-up" che coinvolge f in (2.2) vale automaticamente per ogni intervallo NM. Ci sono due possibilità:

1. Se $f_1 \leq f_e$, il pattern up down up di f da (2.2) vale alla corrente iterazione.

2. Se $f_1 > f_e$, sappiamo dalla stretta convessità che $f_r < f_1$, e il punto di espansione viene accettato. Alla successiva iterazione, $x_2' = x_1$ e $x_1' = x_e$. Ci sono due casi da considerare.
   In primo luogo, supponiamo che x_{min} si trova in $int(x_2', x_1'] = int(x_1, x_e]$. Usando il risultato (2) del Lemma 2.1, sia $f(x_r')$ che $f(x_e')$ deve essere rigorosamente maggiore di $f(x_1')$. Quindi il pattern "up down up" di (2.2) vale in occasione della prossima iterazione.
   In alternativa, si supponga che x_{min} si trovi oltre x_e, vale a dire, oltre x_1'. Allora
   $$\mid x_{min} - x_1' \mid = \mid x_{min} - x_1 \mid - diam(\Delta').$$
   Risulta da (2.1) e dalla disuguaglianza $\rho\chi \geq 1$ che $diam(\Delta') = \rho\chi diam(\Delta) \geq diam(\Delta)$. Così la distanza tra x_{min} e il miglior punto corrente viene ridotta di

una quantità limitata inferiormente da Δ_0, il diametro dell'intervallo iniziale. Questo dà il limite superiore su K di (2.2).

□

Il prossimo risultato mostra che, una volta che x_{min} si trova in un intervallo specifico definito dal corrente intervallo di NM e da un numero che dipende solo dai coefficienti di riflessione, espansione, e contrazione, esso si trova in un intervallo analogo a tutte le successive iterazioni.

Lemma 2.4. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con i parametri soddisfacenti $\rho > 0,\ \chi > 1,\ \chi > \rho,\ \rho\chi \geq 1$ e $0 < \gamma < 1$, sia applicato a f a partire da un simplesso iniziale non degenere. Definiamo N_{NM} come*

$$N_{NM} = max(\frac{1}{\rho\gamma}, \frac{\rho}{\gamma}, \rho\chi, \chi - 1), \tag{2.3}$$

e diciamo che la proprietà di vicinanza vale all'iterazione k se

$$x_{min} \in int(x_2^{(k)}, x_1^{(k)} + N_{NM}(x_1^{(k)} - x_2^{(k)})]. \tag{2.4}$$

Allora, se la proprietà di vicinanza vale all'iterazione k, vale all'iterazione $k+1$.

Dimostrazione. Come nella dimostrazione del precedente lemma, usiamo un primo per indicare le quantità associate all'iterazione $k+1$. La dimostrazione considera tutti i casi possibili per la posizione di x_{min} nell'intervallo definito da (2.4). Abbiamo o $x_2 < x_1 < x_r < x_e$ *o* $x_e < x_r < x_1 < x_2$.
Caso 1. $x_{min} \in int(x_2, x_1]$.
Il lemma 2.1, parte (2), implica che $f_r > f_1$, il che significa che sarà effettuato un passo di contrazione.
1a. Se $f_r \geq f_2$, si verificherà una contrazione interna, con $x_{cc} = x_1 - \gamma(x_1 - x_2)$. La convessità stretta implica che $f_{cc} < f_2$.
(i) Se $f_{cc} \geq f_1,\ x_{min}$ sta nell' $int(x_{cc}, x_1]$. L'intervallo successivo NM è dato da $x_2' = x_{cc}$ *e* $x_1' = x_1$, il che significa che $x_{min} \in int(x_2', x_1']$, e la proprietà di vicinanza vale all'iterazione successiva.
(ii) Se $f_{cc} < f_1$, il successivo intervallo Nelder Mead è $x_2' = x_1$ *e* $x_1' = x_{cc}$. Sappiamo anche che $x_{min} \neq x_1$, in modo che $x_{min} \in int(x_2, x_1) = int(x_2, x_2')$. Per verificare se (2.4) vale, esprimiamo x_2 in termini del nuovo intervallo Nelder Mead come $x_2 = x_1' + \xi(x_1' - x_2')$. Utilizzando la definizione di x_{cc} dà

$$x_2 = x_{cc} + \xi(x_{cc} - x_1) = x_1 + \gamma(x_2 - x_1) + \xi\gamma(x_2 - x_1) \quad \textit{in modo che } \xi = 1/\gamma - 1.$$

Per $\rho > 1$, abbiamo $1/\gamma - 1 < \rho/\gamma \leq N_{NM}$, mentre per $0 < \rho \leq 1$ abbiamo $1/\gamma - 1 < 1/(\rho/\gamma) \leq N_{NM}$, di modo che la proprietà di vicinanza (2.4) vale alla successiva iterazione.
1b. Se $f_r < f_2$, si verificherà una contrazione esterna, con $x_c = x_1 + \rho\gamma(x_1 - x_2)$. Poiché $x_{min} \in int(x_2, x_1]$, la parte (2) del Lemma 2.1 implica che $f_c > f_1$ Il nuovo intervallo è dato da $x'_2 = x_c$ e $x'_1 = x_1$, e l'intervallo di incertezza resta $int(x_2, x'_1]$. Esprimendo x_2 come $x'_1 + \xi(x'_1 - x'_2)$ si ha

$$x_2 = x_1 + \xi(x_1 - x_c) = x_1 - \xi\rho\gamma(x_1 - x_2), \quad \textit{in modo che } \xi = 1/\rho\gamma \leq N_{NM},$$

e (2.4) vale alla successiva iterazione.
Caso 2. $x_{min} \in int(x_1, x_r]$.
2a. Se $f_r < f_1$, proviamo il passo di espansione x_e. La parte (2) del Lemma 2.1 implica che $f_e > f_r$, che significa che il passo di riflessione viene accettato, e il nuovo intervallo NM è $x'_2 = x_1$ *e* $x'_1 = x_r$. Allora $x_{min} \in int(x'_2, x'_1]$ e (2.4) vale alla successiva iterazione.
2b. Se $f_r \geq f_2$, sarà presa una contrazione interna, $x_{cc} = x_1 - \gamma(x_1 - x_2)$. Sappiamo anche che $x_{min} \neq x_r$, in modo che $x_{min} \in int(x_1, x_r)$. La parte (2) del Lemma 2.1 implica che $f_{cc} > f_1$, e il prossimo intervallo NM è $x'_2 = x_{cc}$ *e* $x'_1 = x_1$, con $x_{min} \in int(x'_1, x_r)$. Esprimiamo x_r come $x'_1 + \xi(x'_1 - x'_2)$, che dà

$$x_r = x_1 + \rho(x_1 - x_2) = x_1 + \xi(x_1 - x_{cc}) = x_1 + \xi\gamma(x_1 - x_2)$$

in modo che $\xi = \rho/\gamma \leq N_{NM}$, e (2.4) vale alla successiva iterazione.
2c. Se $f_r \geq f_1$ *e* $f_r < f_2$, sarà presa una contrazione esterna, $x_c = x_1 + \rho\gamma(x_1 - x_2)$. Sappiamo anche che $x_{min} \neq x_r$, in modo che $x_{min} \in int(x_1, x_r)$.
(i) Se $f_c > f_1$, il nuovo intervallo di NM è $x'_2 = x_c$ *e* $x'_1 = x_1$. Poiché $f_c > f_1$, $x_{min} \in int(x_1, x_c) = int(x'_2, x'_1)$. e (2.4) vale alla successiva iterazione.
(ii) Se $f_c < f_1$, il nuovo intervallo di NMead è $x'_2 = x_1$ *e* $x'_1 = x_c$, e $x_{min} \neq x_1$. L'intervallo di incertezza resta $int(x_1, x_r) = int(x'_2, x_r)$ Scriviamo quindi x_r come $x'_1 + \xi(x'_1 - x'_2)$:

$$x_r = x_c + \xi(x_c - x_1) = x_1 + \rho\gamma(x_1 - x_2) + \xi\rho\gamma(x_1 - x_2), \textit{in modo che } \xi = 1/\gamma - 1 < N_{NM},$$

e (2.4) vale alla successiva iterazione.
Caso 3. $x_{min} \in int(x_r, x_e]$
3a. Se $f_e \geq f_r$, il nuovo intervallo di NM è $x'_2 = x_1$ *e* $x'_1 = x_r$; per di più, $x_{min} \neq$

x_e *e* $x_{min} \in int(x'_1, x_e)$. Esprimendo x_e come $x'_1 + \xi(x'_1 - x'_2)$ si ha

$$x_e = x_1 + \rho\chi(x_1 - x_2) = x_1 + \rho(x_1 - x_2) + \xi\rho(x_1 - x_2), \textit{in modo che } \xi = \chi - 1.$$

Poiché $\xi \leq N_{NM}$, (2.4) vale all'iterazione successiva.
3b. Se $f_e < f_r$, accettiamo x_e. Il nuovo intervallo di NM è $x'_2 = x_1$ *e* $x'_1 = x_e$. Dal momento che x_r si trova tra x_1 *e* x_e, $x_{min} \in int(x'_2, x'_1)$ e (2.4) vale alla successiva iterazione.
Caso 4. $x_{min} \in int(x_e, x_1 + N_{NM}(x_1 - x_2)]$.
Il caso 4 può avvenire solo se $N_{NM} > \rho\chi$, dal momento che $x_e = x_1 + \rho\chi(x_1 - x_2)$. Così deve essere vero che $f_1 > f_r > f_e$, e il punto di espansione sarà accettato. Il nuovo intervallo Nelder Mead è definito da $x'_2 = x_1$ *e* $x'_1 = x_e$. Scrivendo $x_1 + N_{NM}(x_1 - x_2)$ come $x_e + \xi(x_e - x_1)$ si ottiene

$$x_1 + N_{NM}(x_1 - x_2) = x_1 + \rho\chi(x_1 - x_2) + \xi\rho\chi(x_1 - x_2), \textit{in modo che } \xi = (N_{NM} - \rho\chi)/\rho\chi.$$

Poichè $\rho\chi \geq 1$, $\xi < N_{NM}$ e la proprietà di prossimità vale all'iterazione successiva. I casi 1 -4 sono esaustivi, e il lemma è dimostrato. □

Dimostriamo adesso che il risultato del lemma 1.7 vale anche quando $\rho\gamma \geq 1$.

Lemma 2.5. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con i parametri soddisfacenti $\rho > 0$ e $0 < \gamma < 1$ venga applicato ad f iniziando con un simplesso iniziale non degenere. Allora $f_1^* = f_2^*$.*

Dimostrazione. Se $\rho\gamma < 1$, il risultato segue dal Lemma 1.7. Quindi assumiamo che $\rho\gamma \geq 1$, il che significa che $\rho > 1$. La dimostrazione è per assurdo, iniziando come nella dimostrazione del Lemma 1.7. Se $f_1^* < f_2^*$, vi è un indice di iterazione K tale che, per $k \geq K$, ogni iterazione k è una contrazione e x_1 non cambia. (Senza perdita di generalità, possiamo prendere $K = 0$). Se l'iterazione k è una contrazione interna, $diam(\Delta_{K+1}) = \gamma diam(\Delta_K) < diam(\Delta_K)$. Se l'iterazione k è una contrazione esterna, $diam(\Delta_{K+1}) = \rho\gamma diam(\Delta_K) \geq diam(\Delta_K)$. Così $\lim_{k\to\infty} diam(\Delta_k) \to 0$ se ci sono un numero finito di contrazioni esterne, e così abbiamo bisogno di considerare solo il caso di un numero infinito di contrazioni esterne. Supponiamo che l'iterazione k è una contrazione esterna. Allora $f_r^{(k)} \geq f_1^{(k)}$, $f_r^{(k)} < f_2^{(k)}$, e il punto di contrazione è $x_c^{(k)} = x_1^{(k)} + \rho\gamma(x_1^{(k)} - x_2^{(k)})$. Dal momento che il miglior punto non cambia, $f_c^{(k)} \geq f_1^{(k)}$ *ed* $x_2^{(k+1)} = x_c^{(k)}$. Dalla stretta convessità, $f_c^{(k)} < f_r^{(k)}$. Definiamo $z(\xi)$ come

$$z(\xi) \equiv x_1^{(k)} + \xi(x_1^{(k)} - x_2^{(k)}),$$

in modo che $x_2^{(k)} = z(-1)$ *e* $x_r^{(k)} = z(\rho)$. Esprimendo $f_2^{(k)}, f_1^{(k)}$ *e* $f_c^{(k)}$ in questo modo, abbiamo

$$f(z(-1)) > f(z(0)) \leq f(z(\rho\gamma)) = f_2^{(k+1)}, \tag{2.5}$$

in modo che $x_{min} \in int(z(-1), z(\rho\gamma))$. La relazione $f(z(-1)) = f_2^{(k)} > f_2^{(k+1)}$ e il risultato (2) del Lemma 2.1 quindi implicano che

$$f(z(\xi)) > f_2^{(k+1)} \quad se\ \xi \leq -1. \tag{2.6}$$

Il punto di riflessione successivo $x_r^{(k+1)}$ è dato da

$$x_r^{(k+1)} = x_1^{(k)} + \rho(x_1^{(k)} - x_2^{(k+1)}) = x_1^{(k)} - \rho^2\gamma(x_1^{(k)} - x_2^{(k)}) = z(-\rho^2\gamma).$$

Poiché $\rho\gamma \geq 1$ *e* $\rho > 1$, abbiamo $\rho^2\gamma > 1$, e concludiamo da (2.6) che $f_r^{(k+1)}$ supera strettamente $f_2^{(k+1)}$. L'iterazione $k+1$ deve essere quindi essere una contrazione interna, con

$$x_{cc}^{(k+1)} = x_1^{(k+1)} + \gamma(x_1^{(k+1)} - x_2^{(k+1)}) = x_1^{(k)} + \rho\gamma^2(x_1^{(k)} - x_2^{(k)}) = z(\rho\gamma^2).$$

Poiché x_1 non cambia, $x_2^{(k+2)} = x_{cc}^{(k+1)}$ e il punto di riflessione all'iterazione $k+2$ è dato da

$$x_r^{(k+2)} = x_1^{(k)} + \rho(x_1^{(k)} - x_2^{(k+2)}) = x_1^{(k)} - \rho^2\gamma^2(x_1^{(k)} - x_2^{(k)}) = z(-\rho^2\gamma^2).$$

Dal momento che $\rho^2\gamma^2 \geq 1$, (2.6) implica ancora una volta che il valore di f in $x_r^{(k+2)}$ *supera* $f_2^{(k+2)}$, e l'iterazione $k+2$ deve essere una contrazione interna. Continuando, se l'iterazione k è una contrazione esterna seguita da j contrazioni interne, il (respinto) punto di riflessione all'iterazione $k+j$ è $z(-\rho^2\gamma^j)$ e l' (accettato) punto di contrazione è $z(\rho\gamma^{j+1})$. A causa della (2.6), l'iterazione $k+j$ deve essere una contrazione interna finché $\rho^2\gamma^j \geq 1$. Sia c* il più piccolo numero intero tale che $\rho^2\gamma^{c^*} < 1$; notiamo che $c^* > 2$. Risulta che la sequenza delle contrazioni divide in blocchi, in cui il blocco j-esimo è costituito da una sola contrazione esterna seguita da un numero c_j di contrazioni interne, con $c_j \geq c^*$ in ogni caso. Denotando k_j l'indice di iterazione all'inizio del j-esimo tale blocco, abbiamo

$$diam(\Delta_{k_j}) = \rho\gamma^{c_j} diam(\Delta_{k_{j-1}}) \leq \theta diam(\Delta_{k_{j-1}}), \quad con\ \theta = \rho\gamma^{c^*} < 1.$$

Il simplesso di diametro più grande all'interno di ogni blocco si verifica dopo la

contrazione esterna, e ha un diametro $\rho\gamma diam(\Delta_{k_j})$. Così abbiamo

$$\lim_{k\to\infty} diam(\Delta_k) \to 0, \lim_{k\to\infty} x_2^{(k)} = x_1^{(k)}, e\ f_2^* = f_1^*,$$

contraddicendo la nostra ipotesi che $f_1^* < f_2^*$ e dando il risultato desiderato. □

Grazie a questo risultato possiamo mostrare quindi che in tutti i casi il diametro del simplesso converge a zero, cioè, il simplesso si riduce ad un punto.

Lemma 2.6. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con i parametri soddisfacenti $\rho > 0$ e $0 < \gamma < 1$ viene applicato ad f iniziando con un simplesso iniziale non degenere Δ_0. Allora $\lim_{k\to\infty} diam(\Delta_k) = 0$.*

Dimostrazione. Il lemma 2.5 mostra che $f_1^* = f_2^*$. Se $f_1^* = f_{min}$, questo valore della funzione è assunto in esattamente un punto, x_{min}, e il risultato desiderato è immediato. Se $f_1^* > f_{min}$, sappiamo dalla stretta convessità che f assume il valore f_1^* in esattamente due punti distinti, indicati con x_1^* *ed* x_2^*, con $x_1^* < x_{min} < x_2^*$. I valori vertici della funzione convergono dall'alto ai limiti ed f è continua. Così per qualsiasi $\epsilon > 0$ c'è un indice di iterazione $\tilde{K}$ tale che, per $k \geq \tilde{K}$, $x_1^{(k)}$ *ed* $x_2^{(k)}$ sono confinati a $I_1^\epsilon \cup I_2^\epsilon$, dove

$$I_1^\epsilon = [x_1^* - \epsilon, x_1^*] \quad e\ I_2* = [x_2^*, x_2^* + \epsilon]. \tag{2.7}$$

Ci sono due casi da considerare.
Caso 1. entrambi i punti finali $x_1^{(k)}$ *ed* $x_2^{(k)}$ giacciono nello stesso intervallo per infinitamente molte iterazioni, cioè, per una delle j = 1, 2, la relazione

$$x_1^{(k)} \in I_j^\epsilon \quad e\ x_2^{(k)} \in I_j^\epsilon \tag{2.8}$$

vale per infiniti k. In questo caso, affermiamo che entrambi i punti finali rimangano in uno di questi intervalli per tutti i k sufficientemente grandi. Questo risultato è dimostrato per assurdo: assumiamo che per ogni $\epsilon > 0$ e l'iterazione K_1 dove (2.8) vale, vi è un'iterazione K_2 successiva in cui $x_1^{(K_2)}$ *e* $x_2^{(K_2)}$ sono in intervalli diversi. Allora, dal momento che $diam(\Delta_{K_1}) \leq \epsilon\ e\ diam(\Delta_{K_2}) \geq x_2^* - x_1^*$, possiamo scegliere ϵ così piccolo che $diam(\Delta_{K_2}) > max(1, \rho\chi)diam(\Delta_{K_1})$. Il diametro del simplesso può essere aumentato solo con riflessione, espansione o contrazione esterna, e il massimo fattore con cui il diametro può aumentare in una singola iterazione è $\rho\chi$. Se $x_1^{(K_1)}$ ed $x_2^{(K_1)}$ sono entrambi in I_1^ϵ, allora la stretta convessità implica che ogni riflessione, espansione o contrazione esterna deve muoversi attraverso I_2^ϵ (e viceversa se i due

vertici si trovano in I_2^ϵ). Ma se ϵ è abbastanza piccolo in modo che $\epsilon\rho\chi < x_2^* - x_1^*$, allora alcuni punti di prova tra le iterazioni K_1 e K_2 devono trovarsi nell'intervallo aperto (x_1^*, x_2^*), e dalla convessità stretta il valore della funzione associato è minore di f_1^*, una contraddizione. Concludiamo che, poiché i punti finali Nelder Mead $x_1^{(k)}$ ed $x_2^{(k)}$ sono in I_j^ϵ per tutti i k sufficientemente grandi, e poiché $f_2^{(k)} \to f_1^{(k)} \to f_1^*$, entrambi i punti finali devono convergere al punto x_j^* e $diam(\Delta_k) \to 0$.
Caso 2. entrambi i punti finali $x_1^{(k)}$ ed $x_2^{(k)}$ sono in intervalli separati I_1^ϵ e I_2^ϵ per tutti i $k \geq K_1$. Mostriamo per assurdo che questo non può accadere perché alla fine si verifica una contrazione interna che genera un punto interno (x_1^*, x_2^*). Sia x_r^* il punto di riflessione per lintervallo Nelder Mead $[x_1^*, x_2^*]$, in cui l'uno o l'altro punto può essere preso come "punto migliore "; sappiamo dalla convessità stretta che $f(x_r^*) > f_1^*$, con $f_r^* = f_1^* + \delta_r$ per alcuni $\delta_r > 0$. Poiché f è continua e $x_r^{(k)}$ è una funzione continua di $x_1^{(k)}$ e $x_2^{(k)}$, ne consegue che, dato un qualsiasi $\delta > 0$, alla fine $f_1^{(k)}, f_2^{(k)},\ e\ f_r^{(k)}$ sono all'interno δ dei loro valori limite. Così, per k sufficientemente grande, $f_r^{(k)} > f_2^{(k)} \geq f_1^{(k)}$ e sarà effettuata una contrazione interna. Dal momento che $x_1^{(k)}$ ed $x_2^{(k)}$ sono in intervalli differenti, il punto di contrazione interna $x_{cc}^{(k)}$ soddisfa

$$x_1^* - \epsilon + \gamma(x_2^* - (x_1^* - \epsilon)) \leq x_{cc}^{(k)} \leq x_2^* + \epsilon + \gamma(x_1^* - (x_2^* + \epsilon)).$$

Se ϵ è abbastanza piccolo, cioè $\epsilon < \gamma(x_2^* - x_1^*)/(1-\gamma)$, allora

$$x_1^* < x_1^* + \gamma(x_2^* - x_1^*) - (1-\gamma)\epsilon \leq x_{cc}^{(k)} \leq x_2^* - \gamma(x_2^* - x_1^*) + (1-\gamma)\epsilon < x_2^*,$$

cioè $x_{cc}^{(k)}$ giace nell'intervallo aperto (x_1^*, x_2^*) ed $f(x_{cc}^{(k)}) < f_1^*$ una contraddizione. □

Tramite questi lemmi dimostriamo quindi il Teorema 2.2 .

Dimostrazione. (Convergenza di Nelder Mead in una dimensione)
Il lemma 2.3 mostra che x_{min} è collegato col tempo con il peggiore vertice e il punto di espansione, cioè, per alcune iterazioni K,

$$x_{min} \in int(x_2^{(K)}, x_1^{(K)} + \rho\chi(x_1^{(K)} - x_2^{(K)})).$$

Dal momento che la costante N_{NM} di (2.3) soddisfa $N_{NM} \geq \rho\chi$, il lemma 2.4 mostra che, per tutti i $k \geq K$, x_{min} soddisfa la proprietà di vicinanza (2.4),

$$x_{min} \in int(x_2^{(k)}, x_1^{(k)} + N_{NM}(x_1^{(k)} - x_2^{(k)})).$$

che implica che

$$| x_{min} - x_1^{(k)} | \leq N_{NM} diam(\Delta_k). \tag{2.9}$$

Il lemma 2.6 mostra che $diam(\Delta_k) \to 0$. In combinazione con (2.9), questo dà il risultato desiderato. □

2.1.3 Convergenza lineare con $\rho = 1$

Attenzioniamo adesso il caso in cui il coefficiente di riflessione è la scelta standard $\rho = 1$, situazione in cui il metodo Nelder Mead non solo converge al minimizzante, ma il suo tasso di convergenza è infine M-passo lineare, cioè la distanza dal migliore vertice al punto ottimale diminuisce ogni M passi con almeno una costante moltiplicativa fissa inferiore a uno. Questo risultato deriva dall'analisi della particolare struttura delle sequenze di movimento consentite Nelder Mead [9].

Teorema 2.7. *(Convergenza lineare di Nelder Mead in una dimensione con $\rho = 1$.) Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo NM con coefficiente di riflessione $\rho = 1$, e coefficienti di espansione e contrazione soddisfacenti $\chi > 1$ e $0 < \gamma < 1$, venga applicato ad f iniziando con un simplesso non degenere Δ_0. Allora vi è un numero intero M dipendente solo da χ e γ tale che*

$$diam(\Delta_{k+M}) \leq 1/2 diam(\Delta_k) \quad \textit{per ogni } k \geq K,$$

dove K è l'indice di iterazione definito nel Lemma 2.3.

Come primo passo per dimostrare questo teorema, si ottengono due risultati relativi alla dimensione 1 sulle sequenze di iterazioni NM.

Lemma 2.8. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati, e supponiamo che il metodo Nelder Mead con parametri $\rho = 1$, $\chi > 1$ e $0 < \gamma < 1$, sia applicato ad f iniziando con un simplesso iniziale non degenere. Allora*

1. *il numero di riflessioni consecutive è delimitato da $r^* = \lceil \chi - 1 \rceil$;*
2. *l'iterazione immediatamente successiva ad una riflessione non può essere un'espansione.*

Dimostrazione. Per ogni iterazione k, definiamo $z^{(k)}(\xi)$ come

$$z^{(k)}(\xi) \equiv x_1^{(k)} + \xi(x_1^{(k)} - x_2^{(k)}), \tag{2.10}$$

in modo che $x_2^{(k)} = z^{(k)}(-1),\ x_r^{(k)} = z^{(k)}(1),\ e\ x_e^{(k)} = z^{(k)}(\chi)$. Se l'iterazione k è una riflessione,

$$f_r^{(k)} < f_1^{(k)},\ f_e^{(k)} \geq f_r^{(k)},\ x_1^{(k+1)} = x_r^{(k)}\ e\ x_2^{(k+1)} = x_1^{(k)}. \tag{2.11}$$

Applicando il lemma 2.1 alle prime due relazioni in (2.11), possiamo vedere che $x_{min} \in int(x_1^{(k)}, x_e^{(k)})$ e

$$f(z^{(k)}(\xi)) \geq f_1^{(k+1)}\ se\ \xi \geq \chi \tag{2.12}$$

A partire dall'iterazione k, il (potenziale) punto l-esimo della riflessione consecutiva è dato da

$$x_r^{(k+l-1)} = x_1^{(k)} + l(x_1^{(k)} - x_2^{(k)}) = z^{(k)}(l) \tag{2.13}$$

che può essere accettato solo se il suo valore della funzione è strettamente minore di $f(x_1^{(k+l-1)})$. La convessità stretta e (2.12) mostrano che ogni punto $z^{(k)}(\xi)$ con $\xi \geq \chi$ non può essere accettato come punto di riflessione. Così il numero di riflessioni consecutive è delimitato dal numero intero r^* soddisfacente

$$r^* < \chi\ e\ r^* + 1 \geq \chi, \quad cioe'\ r^* = \lceil \chi - 1 \rceil$$

Questo completa la dimostrazione di (1).
Se l'iterazione k è una riflessione, il punto di espansione all'iterazione $k+1$ è dato da

$$x_e^{(k+1)} = x_1^{(k+1)} + \chi(x_1^{(k+1)} - x_2^{(k+1)}) = x_1^{(k)} + (1+\chi)(x_1^{(k)} - x_2^{(k)}) = z^{(k)}(1+\chi)$$

La relazione (2.12) implica che il valore della funzione a $x_e^{(k+1)}$ supera $f_1^{(k+1)}$, in modo che $x_e^{(k+1)}$ non sarà accettato. Questo prova il risultato (2) e mostra che l'iterazione immediatamente seguente una riflessione di successo deve essere o una riflessione o una contrazione. □

Si noti che $r^* = 1$ ogni volta che il coefficiente di espansione $\chi \leq 2$; quindi non ci possono essere due riflessioni consecutive con i coefficienti standard di Nelder Mead (1.3) per $n = 1$. Come corollario, si ha che una contrazione deve avvenire entro e non oltre l'iterazione $K + r^*$, dove K è la prima iterazione alla quale il minimizzante è collegato con x_2 e il punto di espansione (Lemma 2.3).

Corollario 2.9. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con $\rho = 1$ venga applicato ad f con un simplesso iniziale non degenere Δ_0, e sia K l'indice di iterazione definito dal Lemma 2.3 in cui, per il tempo prima, $f_1^{(K)} \leq f_e^{(K)}$. Allora deve avvenire una contrazione entro e non oltre l'iterazione $K + r^*$.*

Il prossimo lemma deriva un limite al numero di espansioni consecutive immediatamente dopo una contrazione.

Lemma 2.10. *(espansioni consecutive limitate.) Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con $\rho = 1, \chi > 1,$ e $0 < \gamma < 1$ venga applicato ad f con un simplesso iniziale non degenere Δ_0. Sia $N_{NM} = max(\chi; 1/\gamma)$, che è equivalente alla sua definizione generale* (2.3) *quando $\rho = 1$. Se l'iterazione k è una contrazione, allora per tutte le iterazioni successive lì non possono esserci più di j^* passi consecutivi di espansione, dove j^* è definito come segue:*
(a) se $\chi = N_{NM}, \ j^* = 0$;
(b) se $\chi < N_{NM}, \ j^*$ è il più grande intero soddisfacente $\chi+\chi^2+\chi^3+...+\chi^{j^*} < N_{NM}$.

Dimostrazione. Dal momento che l'iterazione k è una contrazione,
$x_{min} \in int(x_2^{(k)}, x_r^{(k)})$. Così la proprietà di vicinanza (2.4) è soddisfatta all'iterazione k e, dal Lemma 2.4, per tutte le successive iterazioni. La prima espansione in una sequenza di espansioni consecutive deve immediatamente seguire una contrazione (vedi risultato (2) del Lemma 2.8), e la convessità stretta impone un limite al numero di successive espansioni consecutive. Utilizzando la notazione di (2.10), si considerano le disuguaglianze che si applicano al migliore valore della funzione $f_1^{(k+1)}$ alla successiva iterazione, che è (eventualmente) il primo passo di espansione in una sequenza di espansioni consecutive.
Caso 1. Se $f_r^{(k)} < f_2^{(k)}$, l'iterazione k è una contrazione esterna con
$x_c^{(k)} = x_1^{(k)} + \gamma(x_1^{(k)} - x_2^{(k)})$.
(i) Se $f_c^{(k)} \geq f_1^{(k)}$, il prossimo Nelder intervallo di Mead è definito da $x_2^{(k+1)} = x_c^{(k)}$, $x_1^{(k+1)} = x_1^{(k)}$ e $x_{min} \in int(x_2^{(k)}, x_2^{(k+1)})$. (è richiamata la regola tie-breaking esposta nel primo capitolo se $f_c^{(k)} = f_1^{(k)}$). Se si verifica una espansione, l'intervallo si espanderà verso $x_2^{(k)}$, che soddisfa

$$x_2^{(k)} = x_1^{(k+1)} + (x_1^{(k+1)} - x_2^{(k+1)})/\gamma = z^{(k+1)}(1/\gamma), \ con \ f_2^{(k)} > f_1^{(k+1)}. \tag{2.14}$$

(ii) Se $f_c^{(k)} < f_1^{(k)}$, il successivo intervallo di Mead è definito da $x_2^{(k+1)} = x_1^{(k)}$, $x_1^{(k+1)} = x_c^{(k)}$ e $x_{min} \in int(x_2^{(k+1)}, x_r^{(k)})$. Ogni espansione sarà verso $x_r^{(k)}$, che soddisfa

$$x_r^{(k)} = x_1^{(k+1)} + (1/\gamma + 1)(x_1^{(k+1)} - x_2^{(k+1)}) = z^{(k+1)}(1/\gamma - 1) \tag{2.15}$$

con $f_r^{(k)} > f_1^{(k+1)}$.
Caso 2. Se $f_r^{(k)} \geq f_2^{(k)}$, l'iterazione k è una contrazione interna
$x_{cc}^{(k)} = x_1^{(k)} - \gamma(x_1^{(k)} - x_2^{(k)})$.

(i) Se $f_{cc}^{(k)} \geq f_1^{(k)}$,il successivo intervallo di Mead è definito da $x_2^{(k+1)} = x_{cc}^{(k)}$, $x_1^{(k+1)} = x_1^{(k)}$, e $x_{min} \in int(x_2^{(k+1)}, x_r^{(k)})$. (La regola tie-breaking del capitolo 1 è richiamata se $f_{cc}^{(k)} = f_1^{(k)}$). Se si verifica una espansione, l'intervallo si espanderà verso $x_r^{(k)}$, che soddisfa

$$x_r^{(k)} = x_1^{(k+1)} + (x_1^{(k+1)} - x_2^{(k+1)})/\gamma = z^{(k+1)}(1/\gamma), \tag{2.16}$$

con $f_r^{(k)} > f_1^{(k+1)}$.
(ii) Se $f_{cc}^{(k)} < f_1^{(k)}$, il successivo intervallo di Mead è definito da $x_2^{(k+1)} = x_1^{(k)}$, $x_1^{(k+1)} = x_{cc}^{(k)}$ e $x_{min} \in int(x_2^{(k)}, x_2^{(k+1)})$. Ogni espansione sarà verso $x_2^{(k)}$, che soddisfa

$$x_2^{(k)} = x_1^{(k+1)} + (1/\gamma - 1)(x_1^{(k+1)} - x_2^{(k+1)}) = z^{(k+1)}(1/\gamma - 1) \tag{2.17}$$

con $f_2^{(k)} > f_1^{(k+1)}$. Per ciascuno dei quattro casi 1(i) -2 (ii), il valore di f in $z^{(k+1)}(\xi)$ supera $f_1^{(k+1)}$ per alcuni ξ che sono uguali o limitati superiormente da N_{NM}. Applicando il risultato (2) del Lemma 2.1 all'l'intervallo in cui giace x_{min} e la corrispondente espressione da (2.14)-(2.17), concludiamo che, se una sequenza di espansioni consecutive comincia all'iterazione $k+1$, allora

$$f(z^{(k+1)}(\xi)) > f(x_1^{(k+1)}) \; quando \; \xi \geq N_{NM}. \tag{2.18}$$

Il resto della dimostrazione è simile a quella del Lemma 2.8. Il punto di espansione all'iterazione $k+1$ è $x_e^{(k+1)} = z^{(k+1)}(\chi)$. Se $\chi = N_{NM}$, dalla (2.18) segue che questo punto non sarà accettato, e di conseguenza l'iterazione $k+1$ non può essere una espansione; questo corrisponde al caso $j^* = 0$. Se $\chi < N_{NM}$, allora, a partire dall'iterazione $k+1$, il (potenziale) j-esimo punto di espansione consecutiva per $j \geq 1$ è dato da

$$x_e^{(k+j)} = z^{(k+1)}(\chi + \chi^2 + ... + \chi^j) \tag{2.19}$$

Questo punto può essere accettato solo se il suo valore di funzione è strettamente minore di $f(x_1^{(k+j)})$, che diminuisce strettamente dopo ogni espansione accettata. Le relazioni (2.18) e (2.19) insieme dimostrano che, per $j \geq 1$, $x_e^{(k+j)}$ potrebbe essere accettato solo se

$$\chi + \chi^2 + ... + \chi^j < N_{NM}$$

Applicando la definizione di j^*, ne consegue che il valore di j deve essere limitato superiormente da j^*. □

Per il coefficiente di espansione standard $\chi = 2$, il valore di N_{NM} è $max(2, 1/\gamma)$ e i valori di j^* per diverse gamme di γ sono

$j^* = 0$ quando $1/2 \leq \gamma < 1$;
$j^* = 1$ quando $1/6 \leq \gamma < 1/2$;
$j^* = 2$ quando $1/14 \leq \gamma < 1/6$.
Nell'algoritmo standard NM con coefficiente di contrazione $\gamma = 1/2$, il valore zero di j^* significa che non si possono verificare passi di espansione una volta che il minimizzante è collegato con il punto peggiore e il punto di riflessione in ogni iterazione. Esaminiamo ora gli effetti di valide sequenze di movimento NM sul diametro del simplesso.

Lemma 2.11. *Sia f una funzione strettamente convessa su $\mathbb{R}^1$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con $\rho = 1$, $\chi > 1$, e $0 < \gamma < 1$ venga applicato ad f a partire da un simplesso iniziale non degenere Δ_0. Sia Δ il simplesso immediatamente seguente qualsiasi contrazione, e Δ' il simplesso immediatamente dopo la contrazione successiva. Allora esiste un valore ϕ dipendente solo da χ e γ in modo tale che $diam(\Delta') \leq \phi diam(\Delta)$, dove $\phi < 1$.*

Combinando tutti questi risultati, dimostriamo il teorema 2.7, ovvero la convergenza M-passo lineare di Nelder Mead in dimensione 1, quando $\rho = 1$.

Dimostrazione. Nella dimostrazione della convergenza lineare M-passo, usiamo un grafo diretto per descrivere la struttura delle sequenze di movimento valide Nelder Mead. Abbiamo dimostrato finora che il minimizzante viene collegato all'iterazione K (Lemma 2.3) e che una contrazione deve avvenire entro e non oltre l'iterazione $K + r^*$ (Lemmi 2.8 e Corollario 2.9). Successivamente, non più di j^* espansioni consecutive possono verificarsi (Lemma 2.10), e qualsiasi sequenza di espansioni consecutive deve terminare con una sola contrazione o una sequenza di al massimo r^* riflessioni consecutive seguite da una contrazione (Lemma 2.8). La struttura delle sequenze di iterazione possibili seguenti una contrazione può essere pertanto rappresentata da un grafo orientato con quattro stati (nodi): espansione, riflessione, e le due forme di contrazione. Ogni stato è etichettato con il valore assoluto del suo tipo di mossa, in modo che una contrazione interna è etichettata "γ", una contrazione esterna è etichettata "$\rho\gamma$", una riflessione è etichettata "ρ", e una espansione è etichettata "$\rho\chi$". Ad esempio, la Figura 2.1 mostra il grafo corrispondente a $\rho = 1$, $\chi = 2$, e qualsiasi coefficiente di contrazione soddisfacente $1/14 \leq \gamma < 1/6$. Per questi coefficienti, al massimo due fasi di espansione consecutive possono verificarsi ($j^* = 2$), e al massimo una riflessione consecutiva ($r^* = 1$). (Poiché $\rho = 1$, non abbiamo distinto tra le contrazioni interna ed esterna.)

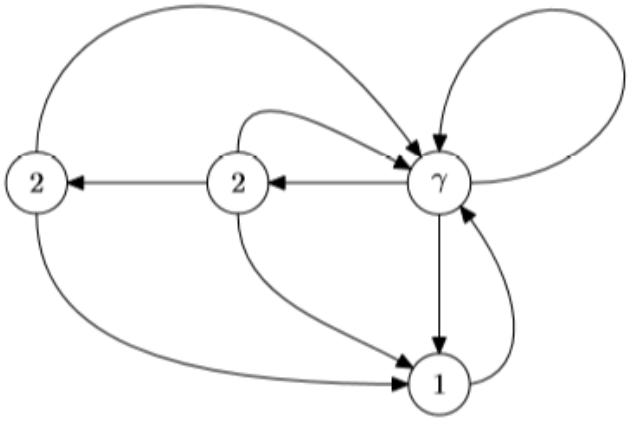


Figura 2.1

Secondo (2.1), il diametro del simplesso viene moltiplicato per ρ per una riflessione, $\rho\chi$ per un'espansione, $\rho\gamma$ per una contrazione esterna, e γ per una contrazione interna. Partendo dallo stato di contrazione con diametro iniziale 1, il diametro dell'intervallo NM dopo ogni sequenza di movimenti è quindi il prodotto delle etichette degli stati incontrati. La prima contrazione nel metodo Nelder Mead può avvenire non più tardi dell'iterazione $K + r^*$. Da allora in poi, i Lemmi 2.8 e 2.10 mostrano che ogni ciclo minimo nel grafo dei movimenti Nelder Mead validi (cioè un ciclo che non passa attraverso qualsiasi nodo due volte) ha lunghezza al massimo $j^* + r^* + 1$; Lemma 2.11 indica che il prodotto delle etichette degli stati oltre qualsiasi ciclo nel grafo Nelder Mead non può superare ϕ . Per ogni intero m, un percorso di lunghezza $m(j^*+r^*+1)$ deve contenere almeno m cicli minimi. Dato un tale percorso, possiamo rimuovere cicli minimi finchè sono rimasti al massimo $j^* + r^*$ bordi. Di conseguenza, il diametro del simplesso alla fine della sequenza associata delle iterazioni NM deve essere moltiplicato per un fattore non più grande di $\chi^{j^*+r^*}\phi^m$. Se si sceglie m come il più piccolo valore tale che

$$\chi^{j^*+r^*}\phi^m \leq 1/2,$$

allora $M = m(j^* + r^* + 1)$ soddisfa $diam(\Delta_{k+M}) \leq 1/2 diam(\Delta_k)$, che dà il risultato desiderato.

□

La convergenza lineare M-passo può anche essere dimostrata per alcune gamme di valori di parametro con $\rho \neq 1$ imponendo restrizioni che garantiscono, per esempio, che $j^* = 0$ ed $r^* = 1$.

2.2 Nelder Mead standard in dimensione 2 per funzioni strettamente convesse

Dopo avere analizzato i risultati di convergenza in dimensione uno per funzioni strettamente convesse con insiemi di livello limitati, verrà esaminato adesso ciò che accade in dimensione due, ipotizzando sempre di lavorare con funzioni strettamente convesse con insiemi di livello limitati. I coefficienti considerati sono $\chi = 2$,$\gamma = 1/2$ e soprattutto $\rho = 1$, ipotesi che risulta essenziale nell'analisi.
Indichiamo il minimizzante (necessariamente unico) di f con x_{min}, e sia $f_{min} = f(x_{min})$. Si noti che l'insieme di livello $\{x \mid f(x) \leq \mu\}$, è vuoto se $\mu < f_{min}$, è il singolo punto x_{min} se $\mu = f_{min}$, e un insieme convesso chiuso se $\mu > f_{min}$.

2.2.1 La convergenza dei valori della funzione al vertice

Il primo risultato che viene messo in luce mostra che, per l'algoritmo standard Nelder Mead, i valori della funzione di limitazione ai vertici sono uguali [9].

Teorema 2.12. *Sia f una funzione strettamente convessa su $\mathbb{R}^2$ con insiemi di livello limitati. Si supponga che l'aloritmo NM con coefficiente di riflessione $\rho = 1$ e coefficiente di contrazione $\gamma = 1/2$ sia applicato ad f a partire da un simplesso iniziale non degenere Δ_0. Allora i tre vertici limitanti i valori di funzione sono identici, cioè,*

$$f_1^* = f_2^* = f_3^*.$$

Dimostrazione. Il Corollario 1.5, che si applica in qualsiasi dimensione, dà subito il risultato se il miglior vertice $x_1^{(k)}$ cambia infinitamente spesso. Il seguente lemma considera l'unico caso restante, in cui $x_1^{(k)}$ alla fine diventa costante. □

Lemma 2.13. *Sia f una funzione strettamente convessa su $\mathbb{R}^2$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con $\rho = 1$ e $\gamma = 1/2$ sia applicato ad f con un simplesso iniziale non degenere Δ_0. Se il migliore vertice $x_1^{(k)}$ è costante per ogni k, allora i simplessi Δ_k convergono al punto $x_1^{(0)}$ come $k \to \infty$.*

.

Dimostrazione. Senza perdita di generalità, il (costante) migliore vertice x_1 può essere preso come origine. La prova che x_2 e x_3 convergono all'origine ha quattro elementi: (i) una matrice di ricorsione che definisce i vertici Nelder Mead all'infinita sottosequenza di iterazioni quando cambia x_2; (ii) una norma speciale che misura il progresso verso l'origine; (iii) limiti su questa norma ottenuti dai valori singolari di

una matrice vincolata ad un sottospazio; e (iv) l'impossibilità di alcuni pattern di tipi di movimenti NM nella sottosequenza dell'iterazione.
(i) *La matrice di ricorsione.* Sappiamo dal Lemma 1.7 che il vertice-peggiore successivo $x_2^{(k)}$ deve cambiare infinitamente spesso. Vi è quindi una sottosequenza di iterazioni $\{k_l\},\ l = 0, 1, ...,\ con\ k_0 = 0,\ dove\ x_2$ cambia, vale a dire,

$$x_2^{(k_{l+1})} \neq x_2^{(l)}\ e\ x_2^{(i)} = x_2^{(i-1)}, \quad i = k_l + 1, ..., k_{l+1} - 1.$$

Definiamo poi nuove sequenze $\tilde{x}_2\ e\ \tilde{x}_3$ da

$$\tilde{x}_2^{(l)} = x_2^{(k_l)}\ e\ \tilde{x}_3^{(l)} = x_3^{(k_l)} \tag{2.20}$$

Poiché x_1 è costante e x_2 cambia all'iterazione k_l, x_3 diventa a quel punto il "vecchio" x_2, cioè,

$$\tilde{x}_3^{(l)} = \tilde{x}_2^{(l-1)} \tag{2.21}$$

Per ogni iterazione strettamente tra $k_l\ e\ k_{l+1}$, solo x_3 cambia, in modo che

$$x_3^{(i)} = 1/2x_2^{(i-1)} + \tau_{i-1}(1/2x_2^{(i-1)} - x_3^{(i-1)})\ per\ i = k_l + 1, ..., k_{l+1} - 1 \tag{2.22}$$

Dove τ_i è il tipo di iterazione i. Si noti che ogni iterazione in cui solo x_3 cambia deve essere una contrazione, in modo che τ_i è necessariamente $\pm 1/2$ quando $k_l < i < k_{l+1}$; il valore di k_l, tuttavia, può essere 1 o $\pm 1/2$. Poiché solo x_3 sta cambiando tra le iterazioni k_l e k_{l+1}, la relazione (2.22) implica che

$$x_3^{(k_l+j)} = 1/2x_2^{(k_l)} + (-1)^{j-1}\Pi_{i=0}^{j-1}\tau_{k_l+i}(1/2x_2^{(k_l)} - x_3^{(k_l)}) \tag{2.23}$$

per $j = 1, ..., k_{l+1} - k_l - 1$.
Utilizzando (2.20), (2.21) e (2.23), si ottiene un'espressione che rappresenta $\tilde{x}_2^{(l+1)}$ del tutto in termini di $\tilde{x}_2^{(l)}\ e\ \tilde{x}_2^{(l-1)}$:

$$\tilde{x}_2^{(l+1)} = 1/2\tilde{x}_2^{(l)} + \tilde{\tau}_l(1/2\tilde{x}_2^{(l)} - \tilde{x}_2^{(l-1)}) \tag{2.24}$$

dove

$$\tilde{\tau}_l = (-1)^{\tilde{l}}\Pi_{i=0}^{\tilde{l}}\tau_{k_l+i}, \quad con\ \tilde{l} = k_{l+1} - k_l - 1$$

Perché riflessioni non possono verificarsi fra le iterazioni k_l e k_{l+1}, sappiamo che $|\ \tilde{\tau}_l\ | \leq 1/2$ o $\tilde{\tau}_l = 1$. (Quest'ultima si verifica solo quando le iterazioni k_l e k_{l+1} sono

consecutive). Utilizzando la notazione matriciale, abbiamo

$$\tilde{x}_2^{(l)} = \begin{pmatrix} \tilde{x}_{21}^{(l)} \\ \tilde{x}_{22}^{(l)} \end{pmatrix} = \begin{pmatrix} u_l \\ v_l \end{pmatrix}; \quad (2.20) da' \; allora \; \tilde{x}_3^{(l)} = \tilde{x}_2^{(l-1)} = \begin{pmatrix} u_{l-1} \\ v_{l-1} \end{pmatrix} \tag{2.25}$$

L'aggiornamento Nelder Mead incluso nella (2.24) può essere scritto come una matrice di ricorsione in u e v:

$$\begin{pmatrix} u_{l+1} & v_{l+1} \\ u_l & v_l \end{pmatrix} = \begin{pmatrix} 1/2(1+\tilde{\tau}_l) & -\tilde{\tau}_l \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u_l & v_l \\ u_{l-1} & v_{l-1} \end{pmatrix} \tag{2.26}$$

Definiamo u_l e v_l con

$$u_l = \begin{pmatrix} u_l \\ u_{l-1} \end{pmatrix} \quad e \; v_l = \begin{pmatrix} v_l \\ v_{l-1} \end{pmatrix},$$

in modo che u_l contiene le coordinate x degli attuali vertici secondo peggiore e peggiore, $\tilde{x}_2^{(l)}$ e $\tilde{x}_3^{(l)}$, e v_l contiene le loro coordinate y. La conclusione desiderata del Lemma 2.13 segue se siamo in grado di dimostrare che

$$\lim_{l\to\infty} u_l = 0 \quad e \quad \lim_{l\to\infty} v_l = 0. \tag{2.27}$$

Dimostreremo solo la prima relazione in (2.27); la prova della seconda è simile.
(ii) *misurare il progresso verso l'origine.* Per dimostrare la convergenza di u_l all'origine, potrebbe sembrare che basti semplicemente applicare le disuguaglianze di norma al l'equazione matriciale (2.26). Purtroppo, la norma due della matrice (2.26) supera di uno per tutti i validi $\tilde{\tau}_l$, il che significa che $\|u_{l+1}\|$ può essere maggiore di $\|u_l\|$. Quindi abbiamo bisogno di trovare un'adatta misura delle dimensioni non crescente associata ad ogni iterazione Nelder Mead (2.26). Tale misura di dimensione è data da una funzione quadratica definita positiva Q di due argomenti scalari (o, equivalentemente, di un vettore bidimensionale):

$$Q(a,b) = 2(a^2 - ab + b^2) = a^2 + b^2 + (a-b)^2. \tag{2.28}$$

Valutando $Q(_{ul+1)}$ con (2.26) dà

$$Q(u_{l+1}) = (3/2 + 1/2\tilde{\tau}_l^2)u_l^2 - 2\tilde{\tau}_l^2 u_l u_{l-1} + 2\tilde{\tau}_l^2 u_{l-1}^2$$

Dopo la sostituzione e la manipolazione, otteniamo

$$Q(u_l) - Q(u_{l+1}) = 2(1-\tilde{\tau}_l^2)(1/2u_l - u_{l-1})^2 \tag{2.29}$$

che dimostra che

$$Q(u_{l+1}) \leq Q(u_l) \quad quando \quad -1 \leq \tilde{\tau}_l \leq 1 \tag{2.30}$$

Ne consegue che, come desiderato, la misura delle dimensioni di Q è non crescente per tutti i valori validi di $\tilde{\tau}_l$. Inoltre, poiché Q è definita positiva, possiamo dimostrare che $u_l \to 0$ mostrando che $Q(u_l) \to 0$. Un evidente e accattivante interpretazione geometrica di Q in termini di simplessi NM è che la quantità $Q(u_l) + Q(v_l)$ è la somma delle lunghezze del lato quadrato del triangolo Nelder Mead all'iterazione k_l, con vertici all'origine, $\tilde{x}_2^{(l)}$, *e* $\tilde{x}_3^{(l)}$. La relazione (2.30) indica che, dopo una riflessione o contrazione in cui x_2 cambia, la somma delle lunghezze dei lati quadrati del nuovo triangolo Nelder Mead non può aumentare, anche se $\|u_{l+1}\|$ potrebbe essere più grande.

(iii) *valori singolari in un sottospazio.* Per ottenere i limiti del caso peggiore sulla dimensione di Q, è conveniente interpretare Q come la norma due di un appositamente strutturato vettore tridimensionale derivato da u_l. Nell'ambito di un'iterazione Nelder Mead (2.25), utilizziamo la notazione

$$\xi_l = \begin{pmatrix} u_l \\ u_{l-1} \\ u_l - u_{l-1} \end{pmatrix}, \; in\ modo\ che\ Q(u_l) = \|\xi_l\|^2 \tag{2.31}$$

La struttura di ξ (2.31) può essere formalizzata osservando che si trova nello spazio bidimensionale nullo del vettore (1, -1, -1). Sia z la seguente matrice 3x2 le cui colonne formano una (non univoca) base ortonormale per questo spazio nullo:

$$Z = (z_1\ z_2),\ dove\ z_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\ e\ z_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

Sia q_l l'unico vettore bidimensionale soddisfacente

$$\xi_l = Zq_l = \begin{pmatrix} u_l \\ u_{l-1} \\ u_l - u_{l-1} \end{pmatrix} \tag{2.32}$$

Dal momento che $Z^T Z = I$, abbiamo

$$\|\xi_l\| = \|q_l\| \quad e\ Q(u_l) = \|\xi_l\|^2 = \|q_l\|^2 \tag{2.33}$$

in modo che possiamo usare $\|q_l\|$ per misurare Q. La mossa Nelder Mead (2.26) può essere scritta in termini di una matrice 3x3 M_l applicata a ξ_l

$$\xi_{l+1} = M_l \xi_l, \quad dove\ M_l = \begin{pmatrix} 1/2(1+\tilde{\tau}_l) & -\tilde{\tau}_l & 0 \\ 1 & 0 & 0 \\ -1/2 & -1/2\tilde{\tau}_l & 1/2\tilde{\tau}_l \end{pmatrix} \tag{2.34}$$

Come abbiamo già dimostrato, la particolare struttura del vettore ξ_l vincola gli effetti della trasformazione M_l ad un sottospazio. Per analizzare questi effetti, si noti che, per costruzione di M_l, la sua applicazione a qualsiasi vettore nello spazio colonna di Z produce un vettore nello stesso spazio colonna, cioè

$$M_l Z = Z W_l, \quad dove\ W_l = Z^T M_l Z \tag{2.35}$$

Una sola mossa Nelder Mead (2.26) è quindi data da

$$\xi_{l+1} = M_l \xi_l = M_l Z q_l = Z W_l q_l,$$

in modo che, usando (2.33),

$$Q(u_{l+1}) = \|\xi_{l+1}\|^2 = \|W_l q_l\|^2$$

e possiamo dedurre informazioni circa il comportamento di Q dalla struttura della matrice 2x2 W. Un calcolo diretto mostra che, per qualsiasi $\tilde{\tau}_l$, W_l è il prodotto di una matrice ortonormale $\tilde{Z}$ e una matrice diagonale:

$$W_l = \tilde{Z}\Sigma_l,\ dove\ \tilde{Z} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}\ e\ \Sigma_l = \begin{pmatrix} 1 & 0 \\ 0 & -\tilde{\tau}_l \end{pmatrix} \tag{2.36}$$

con $\tilde{Z}$ rappresentante una rotazione di 60 gradi. La forma (2.36), analoga al valore di decomposizione singolare a parte il possibile elemento diagonale negativo di Σ_l, rivela che i valori estremi di $\|W_l q_l\|$ sono

$$\begin{aligned} max_{\|q\|=1}\|W_l q\| = 1 \qquad & quando \qquad q = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ min_{\|q\|=1}\|W_l q\| = |\tilde{\tau}_l| \qquad & quando \qquad q = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{aligned} \tag{2.37}$$

Per una riflessione ($\tilde{\tau}_l = 1$), il valore di Q è invariato per tutti i q e quindi per tutti gli u. Quando $|\ \tilde{\tau}_l\ | = 1/2$, la relazione (2.32) indica come gli estremi di (2.37) mappano

in un u-spazio. Il valore di Q rimane costante, cioè, $Q(u_{l+1}) = Q(u_l)$, solo quando ha u_l la forma $(2\alpha, \alpha)$ per qualche alfa diverso da zero; questo può anche essere visto direttamente in (2.29). La riduzione massima in Q, di un fattore $\tilde{\tau}_l^2$, si verifica solo quando u_l ha la forma $(0, \alpha)$ per alcuni alfa diversi da zero. Un'interpretazione geometrica delle mosse di riflessione e di contrazione è mostrata in figura 2.2. Il piano in ogni caso rappresenta u-spazio. La prima figura mostra una curva di livello ellittica di punti (u_l, u_{l-1}) per cui Q = 2; tre punti particolari sulla curva di livello sono etichettati come u_i. La seconda figura mostra l'immagine di questa curva di livello successiva la mossa di riflessione (2.26) con $\tilde{\tau} = 1$. I punti sulla curva di livello sono trasformati da un riflessione in punti ruotati sulla stessa curva di livello; i punti immagine di u_i sono etichettati come u_i'. La terza figura mostra l'immagine della curva di livello nella prima figura dopo una mossa di Nelder Mead contrazione (2.26) con $\tilde{\tau} = 1/2$. I punti trasformati sono non solo ruotati, ma i loro Q-valori sono (ad eccezione di due punti) ridotti. I punti $u_2 = (2/\sqrt{3}, 1/\sqrt{3})$ *e* $u_3 = (0, 1)$ rappresentano gli effetti estremi di contrazione, poiché $Q(u_2') = Q(u_2),\ e\ Q(u_3') = 1/4Q(u_3)$.

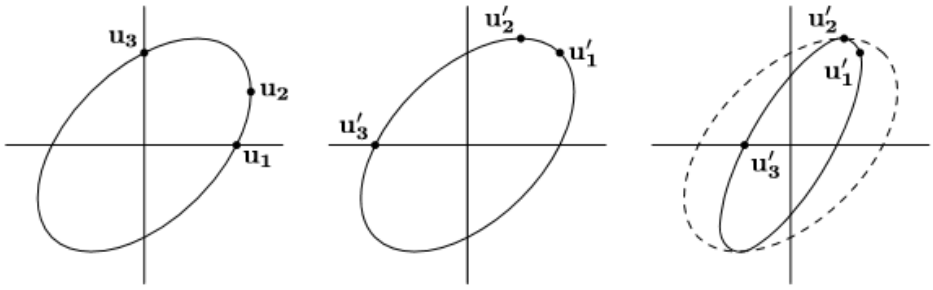


Figura 2.2: Nell'ordine, curva di livello originale, immagine in seguito a rilessione, immagine in seguito a contrazione.

Analizziamo adesso ciò che può accadere al valore di Q dopo una sequenza di iterazioni NM e dimostriamo che anche nel peggiore dei casi Q alla fine deve andare a zero. La relazione (2.36) implica che per ogni vettore q,

$$\|W_j q\| \leq \|W_k q\| \quad se \quad |\ \tilde{\tau}_j\ | \leq |\ \tilde{\tau}_l\ |$$

Nel determinare i limiti superiori su Q, abbiamo quindi bisogno di prendere in considerazione solo i due valori $\tilde{\tau}_l = 1\ e\ \tilde{\tau}_l = 1/2$ (quest'ultimo corrispondente al valore massimo possibile di $|\ \tilde{\tau}\ |$. Quando $\tilde{\tau} \neq 1$). Utilizzando (2.35) ripetutamente per muovere Z verso sinistra, esprimiamo una sequenza di N movimenti NM (2.26)

a partire dall'iterazione l come

$$\xi_{l+N} = M_{l+N-1} \cdots M_l Z q_l = Z W_{l+N-1} \cdots W_l q_l$$

Sostituendo per ogni W da (2.36), la lunghezza euclidea q_{l+N} è delimitata da

$$\|q_{l+N}\| \leq \|\tilde{Z}\Sigma_{l+N-1} \cdots \tilde{Z}\Sigma_l\| \|q_l\|\| \tag{2.38}$$

Un calcolo relativamente semplice mostra che $\|q_{l+N}\|$ è strettamente più piccolo di $\|q_l\|$. dopo ognuna delle sequenze di movimento:

$$\begin{array}{cc} (c,c)\ per\ N=2 & (c,1,c)\ per\ N=3 \\ (c,1,1,1,c)\ per\ N=5 & (c,1,1,1,1,c)\ per\ N=6 \end{array} \tag{2.39}$$

dove " c "denota $\tilde{\tau} = 1/2$ e "1 "indica $\tilde{\tau} = 1$. Per queste sequenze,

$$\|q_{l+N}\| \leq \beta_{cc}\|q_l\|, \quad dove\ \beta_{cc} \approx 0.7215$$

(la quantità β_{cc} è la più grande radice della quadratica $\lambda^2 + 41/64\lambda + 1/16$) A seguito di qualsiasi tipo di pattern NM (2.39), la misura della dimensione di Q deve essere diminuita di un fattore di almeno $\beta_{cc}^2 \approx 0.5206$.

(iv) *i modelli illegali di tipi di pattern Nelder Mead.* A questo punto aggiungiamo l'elemento finale di prova: alcuni pattern dei tipi di movimento NM non possono verificarsi nella sottosequenza (2.20). Ricordiamo che un nuovo punto può essere accettato solo se il suo valore della funzione è strettamente minore del peggiore valore della funzione corrente. Consideriamo ora cinque consecutive iterazioni Nelder Mead (2.26) di tipi $(1, 1, \tilde{\tau}_3, 1, 1)$ in cui x_2 cambia. Dopo un tale pattern, il vertice appena accettato è definito da

$$\begin{pmatrix} u_{l+5} & v_{l+5} \\ u_{l+4} & v_{l+4} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 10 & \end{pmatrix}^2 \begin{pmatrix} 1/2(1+\tilde{\tau}_3) & -\tilde{\tau}_3 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}^2 \begin{pmatrix} u_l & v_l \\ u_{l-1} & v_{l-1} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 1 \\ -\tilde{\tau}_3 & 1/2(1+\tilde{\tau}_3) \end{pmatrix} \begin{pmatrix} u_l & v_l \\ u_{l-1} & v_{l-1} \end{pmatrix} \tag{2.40}$$

La prima riga di questa relazione dà

$$(u_{l+5}, v_{l+5}) = (u_{l-1}, v_{l-1}), \text{in modo che } \tilde{x}_2^{(l+5)} = \tilde{x}_3^{(l)},$$

che implica il risultato impossibile che il vertice appena accettato è uguale al peg-

giore vertice in un simplesso precedente. Da qui la sequenza di tipo $(1, 1, \tilde{\tau}_3, 1, 1)$ non può verificarsi.

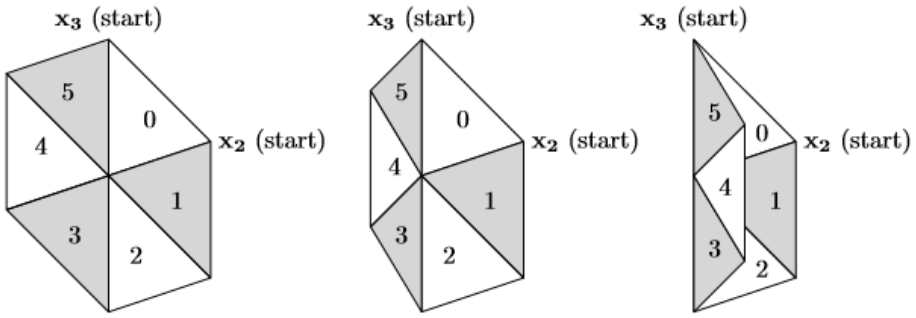


Figura 2.3

La figura 2.3 illustra queste sequenze di movimento inaccettabili geometricamente. Da sinistra a destra, vediamo 5 riflessioni consecutive; due riflessioni, una contrazione esterna, e due ulteriori riflessioni; e due riflessioni, una contrazione interna, e altre due riflessioni. Se eliminiamo entrambi i modelli di riduzione della norma (2.39) e il pattern inammissibile $(1, 1, *, 1, 1)$, solo tre valide sequenze di 6-mosse rimangono durante il quale Q potrebbe rimanere invariato:

$$(1, 1, 1, 1, c, 1);\ (1, c, 1, 1, 1, 1)\ e\ (1, c, 1, 1, c, 1).$$

L'esame di questi tre casi dimostra subito che nessuna sequenza ammissibile di 7 passi esiste per i quali Q può rimanere costante, dal momento che la mossa successiva crea o una norma riducente o un pattern illegale. In particolare, per tutte le sequenze ammissibile di 7 passi vale che

$$\|q_{l+7}\| \leq \beta_{cc}\|q_l\| < 0.7216\|q_l\|.$$

Concludiamo che $\|q_l\| \to 0$ e quindi, usando (2.33), che $Q(u_l) \to 0$, come desiderato. Questo completa la dimostrazione del Lemma 2.13. □

Per finire dimostrazione del Teorema 2.12, notiamo che, nel caso in cui $x_1^{(k)}$ alla fine diventa costante, la prova appena completata del Lemma 2.13 implica la convergenza di x_2, x_3 *ad* x_1, che dà $f_1^* = f_2^* = f_3^*$, come desiderato.

2.2.2 La convergenza dei diametri del simplesso a zero

Sapere che i valori della funzione nel vertice convergono ad un valore comune non implica che i vertici stessi convergono. Analizziamo quindi l'evoluzione delle forme dei triangoli Δ_k prodotti dall'algoritmo Nelder Mead su una funzione strettamente convessa in $\mathbb{R}^2$.

Lemma 2.14. *(Convergenza dei volumi dei simplessi a zero.) Supponiamo che f sia una funzione strettamente convessa su $\mathbb{R}^2$ con insiemi di livello limitati e che l'algoritmo di NM con coefficiente di riflessione $\rho = 1$, coefficiente di espansione $\chi = 2$, e coefficiente di contrazione $\gamma = 1/2$ venga applicato ad f con un simplesso iniziale non degenere Δ_0. Allora i simplessi $\{\Delta_k\}$ generati dall'algoritmo soddisfano*

$$\lim_{k\to\infty} vol(\Delta_k) = 0 \tag{2.41}$$

A questo punto possiamo dimostrare che i diametri convergono a zero, in modo che i simplessi NM crollano in un punto.

Teorema 2.15. *(Convergenza dei diametri dei simplessi a zero.) Sia f una funzione strettamente convessa su $\mathbb{R}^2$ con insiemi di livello limitati. Si supponga che l'algoritmo Nelder Mead con coefficiente di riflessione $\rho = 1$, coefficiente di espansione $\chi = 2$, e coefficiente di contrazione $\gamma = 1/2$ venga applicato ad f con un simplesso iniziale non degenere Δ_0. Allora i simplessi $\{\Delta_k\}$ generati dall'algoritmo soddisfano*

$$\lim_{k\to\infty} diam(\Delta_k) = 0 \tag{2.42}$$

Dimostrazione. La dimostrazione è per assurdo. Il Lemma precedente mostra che $vol(\Delta_k) \to 0$ Poiché la riflessione preserva il volume, devono verificarsi infinitamente molti passi di non riflessione. Supponiamo che la conclusione del teorema non sia vera, cioè che $diam(\Delta_k)$ non converge a zero. Allora possiamo trovare un'infinita sottosequenza $\{k_j\}$ per i quali i simplessi associati Δ_{k_j} hanno diametri delimitati lontano da zero, in modo che

$$diam(\Delta_{k_j}) \geq \alpha > 0 \tag{2.43}$$

Per ogni k_j in questa sottosequenza, si consideri la sequenza di iterazioni $k_j, k_j+1, ...,$ e sia k'_j la prima iterazione in questa sequenza che precede immediatamente un passo di non riflessione. Allora il simplesso $\Delta_{k'_j}$ è congruente a Δ_{k_j}, così che il $diam(\Delta_{k'_j}) \geq \alpha$, e un passo di non riflessione si verifica quando si passa da $\Delta_{k'_j}$ *a* $\Delta_{k'_j+1}$. Ora definiamo una sottosequenza k''_j di k'_j con le seguenti proprietà:

1. $\Delta_{k_j''}$ converge a un segmento di linea fisso $[v_0, v_1]$, con $v_0 \neq v_1$ e $\|v_1 - v_0\|_2 \geq \alpha$;

2. ogni passo Nelder Mead da $\Delta_{k_j''}$ a $\Delta_{k_j''+1}$ ha la stessa combinazione di distinti (peggiori) vertici e il tipo si muove tra le nove possibili coppie dei tre vertici e delle tre mosse di non riflessione.

Si noti che i vertici di $\Delta_{k_j''+1}$ sono funzioni continue dei vertici di $\Delta_{k_j''}$ e che i valori di f in tutti i vertici di $\Delta_{k_j''+1}$ devono convergere monotonicamente dall'alto ad f^*. I punti v_0 e v_1 devono trovarsi sul confine dell'insieme di livello strettamente convesso $L_* = \{x \mid f(x) \leq f^*\}$. Se i vertici di $\Delta_{k_j''}$ convergono a tre punti distinti sul segmento di linea $[v_0, v_1]$, la stretta convessità implicherebbe che il valore della funzione nel punto interno è strettamente minore di f^*, che è impossibile. Così due dei tre vertici devono convergere ad uno di v_0 e v_1, il che significa che due dei vertici di $\Delta_{k_j''}$ alla fine si troveranno vicino ad uno di v_0 o v_1. Senza perdita di generalità supponiamo che due dei vertici sono vicino a v_0 e il vertice rimanente è vicino v_1. Per ottenere una contraddizione, dimostriamo che tutti i passi di non riflessione sono inaccettabili.
(i) una contrazione interna applicata a $\Delta_{k_j''}$ con vertice distinto vicino v_0 produce un vertice (limite) per $\Delta_{k_j''+1}$ a $3/4v_0 + 1/4v_1$; una contrazione interna con vertice distinto nei pressi di v_1 produce un vertice limite a $1/2v_0 + 1/2v_1$. In entrambi i casi, il vertice limite per $\Delta_{k_j''+1}$ si trova strettamente tra v_0 *e* v_1, dando un valore di funzione più piccolo di f^*, una contraddizione.
(ii) una contrazione esterna applicata a $\Delta_{k_j''}$ con vertice distinto vicino v_0 produce un vertice limite per $\Delta_{k_j''+1}$ a $1/4v_0 + 3/4v_1$,dando una contraddizione come in (i). Con vertice distinto nei pressi di v_1, una contrazione esterna produce un vertice limite a $-1/2v_1 + 3/2v_0$. Poiché v_0 *e* v_1 giacciono sul confine dell'insieme strettamente convesso L_*, questo punto di vertice limite si trova al di fuori dell'insieme di livello e quindi ha valore di funzione maggiore di f^*. Questo contraddice il fatto che il valore del vertice della funzione associato in $\Delta_{k_j''+1}$ deve convergere a f^*.
(iii) un passo di espansione con vertice distinto vicino v_0 produce un vertice limite per $\Delta_{k_j''+1}$ a $3v_1 - 2v_0$, e un passo di espansione con vertice distinto vicino v_1 produce un vertice limite a $3v_0 - 2v_1$. In entrambi i casi, il vertice limite giace fuori L_*. Ciò significa che il suo valore di funzione supera f^*, dando una contraddizione. Dal momento che una contraddizione nasce dall'applicare ogni possibile non mossa di non riflessione al simplesso $\Delta_{k_j''}$, la sequenza k_j di (2.43) non può esistere. Così abbiamo dimostrato che $\lim diam(\Delta_k) \to 0$, vale a dire che ciascun simplesso Nelder Mead alla fine collassa a un punto. □

Notiamo che questo teorema non implica che la sequenza di simplessi $\{\Delta_k\}$ converge ad un punto limite x_*. Sappiamo, però, che tutti i vertici convergono ad x_1 se questo vertice rimane costante (vedi Lemma 2.13); questa situazione si verifica negli esempi di McKinnon che analizzeremo.
Riepilogando, si è visto che l'algoritmo Nelder Mead generico in dimensione 1 converge al minimizzante di una funzione strettamente convessa con insiemi di livello limitati se e solo se il passo di espansione è una vera espansione (cioè, se $\rho\chi \geq 1$). È interessante notare che, a parte questo ulteriore requisito, le condizioni (1.2) fornite nel documento originale Nelder Mead bastano per garantire la convergenza in una dimensione. Il comportamento dell'algoritmo in dimensione 1 può comunque essere molto complicato; per esempio, ci può essere un numero infinito di espansioni anche quando la convergenza è M-passo lineare. In due dimensioni invece, anche il comportamento del metodo NM standard (con $\rho = 1, \chi = 2,\ e\ \gamma = 1/2$) è più difficile da analizzare per due ragioni:

1. Lo spazio delle forme dei simplessi non è compatto. Sembra che le mosse Nelder Mead siano dense in questo spazio, cioè, qualsiasi simplesso può essere trasformato da qualche sequenza di mosse Nelder Mead per essere arbitrariamente vicino a qualsiasi altra forma di simplesso; questa proprietà riflette l'intento espresso da Nelder e Mead [12] che la forma del simplesso dovrebbe "adattarsi" al paesaggio locale, ma contrasta fortemente con la natura di molti metodi di ricerca pattern, in cui le forme del simplesso rimangono costanti.

2. La presenza della fase di espansione significa che vol (Δ) non è una funzione di Lyapunov per l'iterazione.

I risultati bidimensionali dimostrati in questo capitolo sembrano molto deboli ma plausibilmente rappresentano i limiti di ciò che può essere dimostrato per arbitrarie funzioni strettamente convesse. Una domanda ovvia riguarda come il metodo Nelder Mead può non convergere ad un punto di minimo nel caso bidimensionale. Ulteriore analisi suggerisce che, per funzioni rigorosamente convesse adatte (C^1 sembra sufficientece), il fallimento può avvenire solo se i simplessi si allungano indefinitamente e la loro forma va a "infinito" nello spazio delle forme del simplesso (come nel controesempio di McKinnon [11]).
Di sicuro uno dei punti ancora da risolvere non è se in definitiva converge ad un punto di minimo- per generali funzioni (non convesse), non lo fa - ma piuttosto perché tende a lavorare così bene in pratica producendo una diminuzione iniziale rapida nei valori della funzione.

Capitolo 3

La convergenza del metodo del simplesso Nelder Mead ad un punto non stazionario

In letteratura, come già osservato, si trovano spesso risultati in cui il metodo Nelder Mead non converge ad un minimo locale, anche nel caso di funzioni regolari di dimensione bassa: è stato notato infatti già per funzioni con solo otto variabili. Per approfondire meglio la questione, viene presentata in questo capitolo una famiglia di esempi di funzioni di due variabili, per la quale la convergenza si verifica ad un punto non stazionario per una serie di simplessi di partenza. Alcuni esempi hanno una derivata prima discontinua ed altri sono strettamente convessi con derivate continue tra uno e tre. I simplessi convergono ad una retta che è ortogonale alla direzione di discesa ripida e hanno angoli interni che tendono a zero. Gli esempi riportati portano il metodo Nelder Mead ad applicare il passo di contrazione interna ripetutamente lasciando fisso il vertice migliore. Questo comportamento sarà indicato come contrazione interna ripetuta focalizzata (RFIC) [11]. Nessun altro tipo di passo si verifica per questi esempi, e questo semplifica notevolmente la loro analisi. Gli esempi sono molto semplici ed evidenziano un serio difetto nel metodo: il crollo dei simplessi lungo la direzione di discesa ripida, una direzione lungo la quale vorremmo invece che si ingrandissero.

Finora abbiamo considerato il caso in cui la fase di espansione viene accettata se $f_e < f_r$ ma poichè gli esempi seguenti sono costruiti in modo che $f_r > f_1$, vale a dire, il punto riflesso non è mai un miglioramento del punto migliore, il passo di espansione non è mai considerato. Quindi entrambe le versioni funzionano allo stesso modo.

Sono noti altri esempi in cui il metodo di Nelder Mead o le sue varianti falliscono:

Dennis e Woods danno un esempio strettamente convesso, dove una variante del metodo esegue una sequenza ininterrotta di passaggi di restringimento verso un unico punto che è una discontinuità del gradiente e in cui c'è un sottogradiente non nullo. Nella loro variante la condizione per accettare una fase di contrazione è che $f_c < f_s$, che è più restrittiva rispetto al metodo originale Nelder Mead, quindi sono eseguiti più passaggi di restringimento. Questo comportamento non può verificarsi per la versione originale del metodo Nelder Mead perché questo non esegue mai passi di restringimento sulle funzioni strettamente convesse. Woods dà anche uno schizzo di una funzione derivabile non convessa per la quale il metodo Nelder Mead converge verso un punto non minimizzante da una sequenza di passi di restringimento. Tuttavia, si può dimostrare che, affinchè questo comportamento si verifichi con la forma originale del metodo Nelder Mead, il punto a cui il simplesso si contrae deve essere un punto stazionario.

E' anche possibile costruire esempi di funzioni non convesse differenziabili per le quali la forma originale del metodo in aritmetica esatta converge con contrazioni ripetute ad un simplesso degenere di lunghezza finita, nessuno dei cui vertici sono punti stazionari. Un esempio di questo caso è la funzione $f(x, y) = x^2 - y(y - 2)$ con simplesso iniziale (1,0), (0, -3), (0,3), che tende al limite a (0,0), (0, -3), (0,3). Gli esempi riportati qui sono, tuttavia, i primi esempi noti in cui il metodo Nelder Mead non riesce a convergere ad un punto di minimo di una funzione strettamente convessa differenziabile. Vi è anche un'ampia varietà di metodi del simplesso che permettono al simplesso di variare nella forma in modo simile al metodo Nelder Mead : questi metodi accettano certi passaggi di prova solo se vi è una diminuzione sufficiente in una funzione obiettivo; in questo differiscono dunque dal metodo Nelder Mead e dai metodi di Torczon che richiedono soltanto una rigorosa diminuzione e il cui comportamento dipende solo dall'ordine dei valori della funzione nei punti di prova, non sui valori reali. I risultati di convergenza per tali metodi, esposti da Rykov e Tseng, contano su questa diminuzione sufficiente. Una delle varianti del metodo di Tseng [19] è la stessa del metodo Nelder Mead tranne per la condizione di diminuzione sufficiente e una condizione che delimita gli angoli interni del simplesso lontano da zero. Per questo motivo, quando la variante di Tseng viene applicata agli esempi riportati, alla fine esegue passi di riduzione al posto dei passi di contrazione interna eseguiti dal metodo originale Nelder Mead, cosa che gli permette di evitare il punto non stazionario che è al centro della RFIC nel metodo originale Nelder Mead.

3.1 Analisi del comportamento di contrazione interna ripetuta

Si consideri un simplesso in due dimensioni con vertici a 0 (cioè, l'origine), $v^{(n+1)}$, $v^{(n)}$. Assumiamo che

$$f(0) < f(v^{(n+1)}) < f(v^{(n)}) \tag{3.1}$$

Dopo la fase di ordinamento dell'algoritmo, $x_1 = 0, x_2 = v^{(n+1)}$, e $x_3 = v^{(n)}$. Il metodo Nelder Mead calcola $m^{(n)} = v^{(n+1)}/2$, il punto medio della linea che unisce il primo e il secondo punto peggiore, e poi riflette il punto peggiore, $v^{(n)}$, in $m^{(n)}$ con un fattore di riflessione $\rho = 1$ dando il punto

$$r^{(n)} = m^{(n)} + \rho(m^{(n)} - v^{(n)}) = v^{(n+1)} - v^{(n)}. \tag{3.2}$$

Supponiamo che

$$f(v^{(n)}) < f(r^{(n)}) \tag{3.3}$$

In questo caso il punto $r^{(n)}$ viene respinto e il punto $v^{(n+2)}$ viene calcolato utilizzando un fattore di riflessione $\rho = -0.5$ in

$$v^{(n+2)} = m^{(n)} + \rho(m^{(n)} - v^{(n)}) = 1/4v^{(n+1)} + 1/2v^{(n)}$$

$v^{(n+2)}$ è il punto medio della linea che unisce $m^{(n)}$ e $v^{(n)}$. Purchè $f(v^{(n+2)}) < f(v^{(n+1)})$, cioè (3.1) mantiene con n sostituito da $n + 1$, il metodo Nelder Mead fa il passo di contrazione all'interno piuttosto che un passo di restringimento. Il passo di contrazione interna sostituisce $v^{(n)}$ con il punto $v^{(n+2)}$, in modo che il nuovo simplesso consiste di $v^{(n+1)}, v^{(n+2)}$, e l'origine. Purchè questa schema si ripeta, i successivi vertici del simplesso soddisferanno la relazione di ricorrenza lineare

$$4v^{(n+2)} - v^{(n+1)} - 2v^{(n)} = 0$$

Questa ha soluzione generale

$$v^{(n)} = A_1\lambda_1^n + A_2\lambda_2^n \tag{3.4}$$

dove $A_i \in \mathbb{R}^n$ e

$$\lambda_1 = \frac{1 + \sqrt{33}}{8}, \quad \lambda_2 = \frac{1 - \sqrt{33}}{8} \tag{3.5}$$

Quindi $\lambda_1 \cong 0.84307$ *e* $\lambda_2 \cong -0.59307$. Segue da (3.2) e (3.4) che

$$r^{(n)} = -A_1\lambda_1^n(1-\lambda_1) - A_2\lambda_2^n(1-\lambda_2). \tag{3.6}$$

E' questa contrazione interna ripetuta verso lo stesso vertice fissato ad essere denominata contrazione interna focalizzata ripetuta (RFIC). Il concetto dimostrato formalmente da Lagarias [9] che nessuna fase del metodo Nelder Mead può trasformare un simplesso non degenere in uno degenere, nel caso bidimensionale corrisponde al fatto che l'area del simplesso o aumenta di un fattore 2, o rimane la stessa, o diminuisce di un fattore di 2 o 4. Pertanto, purché il metodo Mead Nelder venga avviato da un simplesso iniziale non degenere, il simplesso successivo non può essere degenere e se si verifica RFIC, allora il simplesso iniziale per RFIC è non degenere. Ciò implica che A_1 *e* A_2 in (3.4) sono linearmente indipendenti. Consideriamo ora il simplesso iniziale con vertici $v^{(0)} = (1,1); v^{(1)} = (\lambda_1, \lambda_2),\ e\ (0,0)$. Sostituendo nella (3.4) si ottiene $A_1 = (1,0)$ *e* $A_2 = (0,1)$. Ne consegue che la soluzione particolare per queste condizioni iniziali è $v^{(n)} = (\lambda_1^n, \lambda_2^n)$. Questa soluzione è mostrata nella Figura 3.1. La forma generale dei tre punti necessari ad un passo del Nelder metodo Mead è quindi

$$v^{(n)} = (\lambda_1^n, \lambda_2^n) \tag{3.7}$$

$$v^{(n+1)} = (\lambda_1^{n+1}, \lambda_2^{n+1}) \tag{3.8}$$

$$r^{(n)} = (-\lambda_1^n(1-\lambda_1), -\lambda_2(1-\lambda_2)). \tag{3.9}$$

Purchè (3.1) e (3.3) valgano in questi punti, il metodo del simplesso prenderà il passo di contrazione interna assunto in precedenza. Si noti che le coordinate x di $v^{(n)}$ *e* $v^{(n+1)}$ sono positive e la coordinata x di $r^{(n)}$ è negativa.

3.2 Funzioni che causano RFIC

Una famiglia di funzioni che causano RFIC, introdotta da Mc Kinnon [11], ha la seguente espressione:

$$f(x,y) = \begin{cases} \theta\phi|x|^\tau + y + y^2, & x \le 0, \\ \theta x^\tau + y + y^2, & x \ge 0, \end{cases} \tag{3.10}$$

dove θ *e* ϕ sono costanti positive. Si noti che (0, -1) è una direzione di discesa dall'origine (0,0) e che f è strettamente convessa purchè $\tau > 1$. Per diversi valori

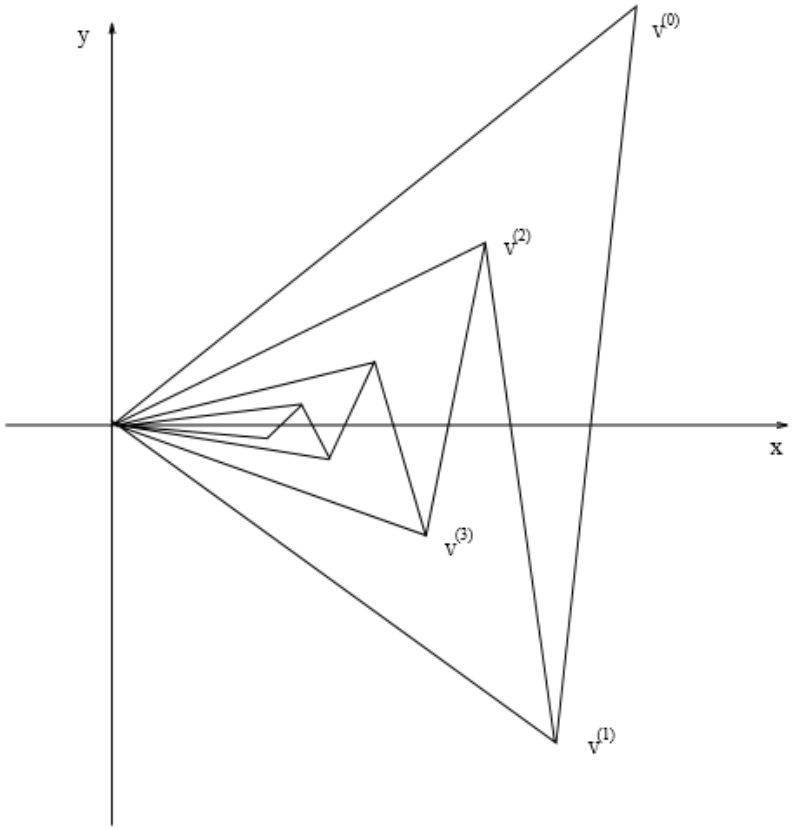


Figura 3.1: Simplessi successivi con RFIC

dei parametri θ, ϕ, τ si ha, in corrispondenza di un determinato simplesso iniziale, convergenza dell'algoritmo di Nelder Mead al punto di coordinate (0,0), che non è un punto stazionario del problema.
La f ha derivata prima continua se $\tau > 1$ derivate seconde continue se $\tau > 2$, e derivate terze continue se $\tau > 3$. La figura 3.2 mostra la trama di contorno di questa funzione e i primi due passi del metodo Nelder Mead per il caso $\tau = 2, \tau = 6,\ e\ \tau = 60$. Entrambi i passi sono di contrazioni interne.
Definiamo $\hat{\tau}$ in modo tale che

$$\lambda_1^{\hat{\tau}} = |\lambda_2|, \tag{3.11}$$

Così $\hat{\tau}$ è dato da

$$\hat{\tau} = \frac{ln|\lambda_2|}{ln\lambda_1} \cong 3.0605 \tag{3.12}$$

In ciò che segue assumiamo che τ soddisfi

$$0 < \tau < \hat{\tau} \tag{3.13}$$

Poiché $0 < \lambda_1 < 1$, pertanto segue che

$$\lambda_1^{\tau} > \lambda_1^{\hat{\tau}} = |\lambda_2| \tag{3.14}$$

Usando (3.7) e (3.9) segue che

$$f(v^{(n)}) = \theta\lambda_1^{\tau n} + \lambda_2^n + \lambda_2^{2n}$$

$$e\ f(r^{(n)}) = \phi\theta(\lambda_1^{\tau n}(1-\lambda_1)^{\tau}) - \lambda_2^n(1-\lambda_2) + \lambda_2^{2n}(1-\lambda_2)^2.$$

Poiché $f(v^{(n)}) > f(v^{(n+1)})$ se e solo se

$$\theta\lambda_1^{\tau n}(1-\lambda_1^{\tau}) > \lambda_2^n(\lambda_2 - 1) + \lambda_2^{2n}(\lambda_2^2 - 1).$$

Poiché $\lambda_1^{\tau} > |\lambda_2|$ e $\lambda_2^2 - 1 < 0$ questo è vero per ogni $n \geq 0$ se θ è tale che

$$\theta(1-\lambda_1^{\tau}) > |\lambda_2 - 1| \tag{3.15}$$

Anche $f(v^{(n+1)}) > f(0)$ se e solo se

$$\theta\lambda_1^{\tau(n+1)} + \lambda_2^{n+1} + \lambda_2^{2(n+1)} > 0.$$

Poiché $\lambda_1^t au > |\lambda_2|$, questo è vero per ogni $n \geq 0$ se

$$\theta > 1 \tag{3.16}$$

Anche $f(r^{(n)}) > f(v^{(n)})$ se e solo se

$$\phi\theta(\lambda_1^{\tau n}(1-\lambda_1)^{\tau}) - \lambda_2^n(1-\lambda_2) + \lambda_2^{2n}(1-\lambda_2)^2 > \theta\lambda_1^{\tau n} + \lambda_2^n + \lambda_2^{2n}$$

$$\Leftrightarrow \theta\lambda_1^{\tau n}(\phi(1-\lambda_1)^{\tau} - 1) > \lambda_2^n(2-\lambda_2) - \lambda_2^{2n}((1-\lambda_2)^2 - 1).$$

Poiché $\lambda_2 < 0\ e\ \lambda_1^{\tau} > |\lambda_2|$, questo è vero per ogni $n \geq 0$ se $\theta\ e\ \phi$ sono tali che

$$\theta(\phi(1-\lambda_1)^{\tau} - 1) > (2-\lambda_2). \tag{3.17}$$

Per ogni τ nell'insieme dato dalla (3.13), può essere scelto in modo che siano mantenute (3.15) e (3.16) e può essere scelto in modo che (3.17) vale. Ne deriva che valgono (3.1) e (3.3), quindi ad ogni passo l'algoritmo contrae internamente il simplesso corrente senza mai cambiare il miglior vertice, e procede in questo modo fino a convergere proprio sul punto x_1, ossia l'origine, che, come già anticipato, è un

punto non stazionario. In particolare, una direzione di discesa in questo punto è costituita dal secondo asse coordinato. Esempi di valori di θ e ϕ che rendono (3.15), (3.16), e (3.17) valide sono i seguenti: per $\tau = 1, \theta = 15$ *e* $\phi = 10$; *per* $\tau = 2, \theta = 6$ *e* $\phi = 60$; *per* $\tau = 3, \theta = 6$ *e* $\phi = 400$.
Tale famiglia di funzioni verrà comunque ripresa nel capitolo 5, in quanto costituisce uno degli oggetti di studio dei test numerici effettuati.

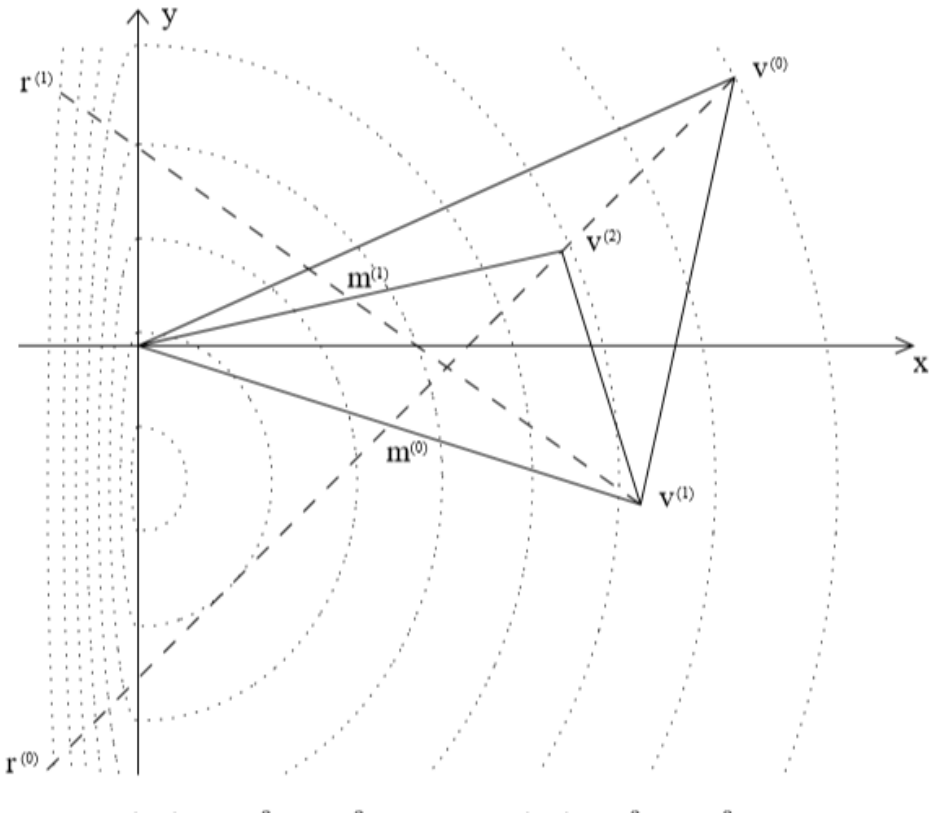


Figura 3.2

3.3 Condizioni necessarie perché si verifichi RFIC

Come già detto, vi sono dei casi in cui l'algoritmo converge ad un punto che non è il minimo della funzione in esame e si trova inoltre ad effettuare una contrazione interna ripetuta focalizzata nello stesso vertice. In questa sezione verranno dedotte delle condizioni necessarie affinchè si verifichi RFIC. Per comodità di notazione i risultati sono dati per RFIC con l'origine come fuoco, ma con un cambiamento dell'origine possono essere applicati a qualsiasi punto. Risulta dalla descrizione dell'algoritmo che una condizione necessaria affinchè si verifichi RFIC [11] è che

$$f_0 = f(0) \leq f(v^{(n+1)}) \leq f(v^{(n)}) \leq f(r^{(n)}) \tag{3.18}$$

(Gli esempi della sezione precedente soddisfano la forma rigorosa delle relazioni (3.18) come indicate nella (3.1) e (3.3).) Se f è s volte differenziabile all'origine, allora f può essere scritta nella forma $f(v) = p_s(v) + o(\|v\|^s)$, dove p_s è un polinomio di grado al più s, e $D^i f(0) = D^i p_s(0)$ per $i = 0, ..., s$, cioè le derivate di f e p_s coincidono. Facendo un cambiamento di variabile allo z-spazio usando $v = A_1 z_1 + A_2 z_2$, f e p_s possono essere viste come funzioni di $(z_1, z_2) \in \mathbb{R}^2$. Quando le derivate necessarie esistono, definiamo

$$f_0 = f(0), \quad g_i = \frac{\delta f}{\delta z_i}(0), \quad h = \frac{1}{2}\frac{\delta^2 f}{\delta z_1^2}(0), \text{ e } k = \frac{1}{6}\frac{\delta^3 f}{\delta z_1^3}(0)$$

Poi (g_1, g_2) è il gradiente di f nello z-spazio, e g_i, h, k sono i $z_i, z_1^2,$ *e* z_1^3 coefficienti nella sviluppo di Taylor di f nello z-spazio. Dal momento che $|\lambda_2| < \lambda_1$ e (3.4) vale, $\left\|v^{(n)}\right\| = O(\lambda_1^n)$, così

$$f(v^{(n)} = p_s(v^{(n)}) + o(\lambda_1^{sn}) \tag{3.19}$$

Teorema 3.1. *Se l'origine è il fuoco della contrazione interna ripetuta partendo da un simplesso con direzione limitante A_1, allora*

1. *se f è derivabile all'origine, allora $g_1 = 0$;*
2. *se f è 2 volte derivabile all'origine, allora $h = 0$;*
3. *se f è 3 volte derivabile all'origine, allora $k = 0$.*

Dimostrazione. 1. da (3.18) segue che una condizione necessaria perchè RFIC avvenga è che $f_0 \leq f(v^{(n)})$ *e* $f_0 \leq f(r^{(n)})$. Questo è vero se e solo se

$$f_0 \leq f_0 + g_1\lambda_1^n + g_2\lambda_2^n + o(\lambda_1^n),$$

$$e\ f_0 \leq f_0 - g_1\lambda_1^n(1 - \lambda_1) - g_2\lambda_2^n(1 - \lambda_2) + o(\lambda_1^n),$$

Dal momento che $|\lambda_2| < \lambda_1 < 1$, questo non può avvenire per ogni n a meno che $g_1 = 0$.

2. Poiché f è 2 volte derivabile all'origine, vale la parte (1), così $g_1 = 0$. Quindi $p_2(v^{(n)}) - (f_0 + g_2\lambda_2^n + h\lambda_1^{2n}) = O(|\lambda_1\lambda_2|^n) = o(\lambda_1^{4n})$, poichè $|\lambda_2| < \lambda_1^3$. Da questo e da (3.19) segue che

$$f(v^{(n)}) = f_0 + g_2\lambda_2^n + h\lambda_1^{2n} + o(\lambda_1^{2n}),$$

Dalla (3.18) segue che una condizione necessaria perchè RFIC avvenga è che $f_0 \leq f(v^{(n)})$ *e* $f(v^{(n)}) \leq f(r^{(n)})$. Questo è vero se e solo se

$$f_0 \leq f_0 + g_2\lambda_2^n + h\lambda_1^{2n} + o(\lambda_1^{2n}),$$

$$e\ 0 \leq -g_2\lambda_2^n(2-\lambda_2) - h\lambda_1^{2n+1}(2-\lambda_1) + o(\lambda_1^{2n}),$$

Dal momento che $|\lambda_2| < \lambda_1^2 < 1$, questo non può verificarsi per ogni n a meno che h=0.

3. Poiché f è 3 volte derivabile all'origine, valgono le parti (1) e (2), quindi $g_1 = 0$ *e* $h = 0$. Quindi $p_3(v^{(n)}) - (f_0 + g_2\lambda_2^n + k\lambda_1^{3n}) = O(|\lambda_1\lambda_2|^n) = o(\lambda_1^{4n})$. Da questa e da (3.19) segue che

$$f(v^{(n)}) = f_0 + g_2\lambda_2^n + k\lambda_1^{3n} + o(\lambda_1^{3n}),$$

Dalla (3.18) segue che una condizione necessaria perchè RFIC avvenga è che $f_0 \leq f(v^{(n)})$ *e* $f_0 \leq f(r^{(n)})$. Questo è vero se e solo se

$$f_0 \leq f_0 + g_2\lambda_2^n + k\lambda_1^{3n} + o(\lambda_1^{3n}),$$

$$e\ f_0 \leq f_0 - g_2\lambda_2^n(1-\lambda_2) - k\lambda_1^{3n}(1-\lambda_1)^3 + o(\lambda_1^{3n}),$$

Dal momento che $\lambda_1^3 > |\lambda_2|$, questo non può verificarsi per ogni n a meno che $k = 0$.

□

Teorema 3.2. *Se f ha un gradiente non nullo all'origine e in un intorno dell'origine può essere espressa nella forma*

$$f(v) = p_4(v) + o(\|v\|^{\hat{\tau}}), \tag{3.20}$$

dove p_4 è almeno 4 volte derivabile all'origine, e se il simplesso iniziale è non degenere, allora l'origine non può essere il centro di ripetute contrazioni interne.

Dimostrazione. Si supponga che l'origine sia al centro di contrazioni ripetute. Le prime tre derivate di f e p_4 all'origine sono uguali. Il Teorema 3.1 mostra che $g_1 = h = k = 0$. Quindi $p_4(v^{(n)}) - (f_0 + g_2\lambda_2^n) = O(|\lambda_1\lambda_2|^n) = o(\lambda^{4n})$. Poichè $\hat{\tau} < 4$ *e* $o(\left\|v^{(n)}\right\|^{\hat{\tau}}) = o(\lambda_1^{\hat{\tau}n})$ *e* $\lambda_1^{\hat{\tau}} = |\lambda_2|$ (dalla definizione di $\hat{\tau}$), segue che

$$f(v^{(n)}) = f_0 + g_2\lambda_2^n + o(|\lambda_2|^n).$$

Dalla (3.18) segue che una condizione necessaria perchè RFIC avvenga è che $f_0 \leq f(v^{(n)})$ *e* $f_0 \leq f(v^{(n+1)})$. Poiché $\lambda_2 < 0$, ciò non può verificarsi per ogni n a meno che $g_2 = 0$. Tuttavia, poiché una condizione del teorema è che il gradiente è non nullo all'origine e poiché $g_1 = 0$, non è possibile che $g_2 = 0$. Questo contraddice l'ipotesi originale e così dimostra che l'origine non può essere al centro delle contrazioni ripetute. □

Il teorema 3.2 mostra che RFIC non può verificarsi per funzioni sufficientemente lisce, essendo il limite un po' più di 3 volte derivabile. Gli esempi nella sezione precedente mostrano che, se le condizioni del Teorema 3.2 non valgono, allora è possibile RFIC.

3.4 Perturbazioni del simplesso iniziale

In questa sezione viene analizzato il comportamento degli esempi per perturbazioni del simplesso di partenza [11]. La posizione perturbata per il vertice nell'origine deve essere sull'asse y, altrimenti il simplesso contratto alla fine si troverà all'interno di una regione dove esistono tutte le derivate della funzione, ed i teoremi 3.1 e 3.2 mostrano che un punto non stazionario non può essere il centro di RFIC in tale regione. Anche se $\tau > 1$, esiste il gradiente dove x = 0 e la sua direzione è parallela all'asse y. Risulta dal Teorema 3.1 che i soli simplessi iniziali che possono produrre RFIC sono quelli con l'autovettore dominante A_1 perpendicolare all'asse y. Consideriamo quindi solo perturbazioni dove il vertice nell'origine è perturbato a $(0, y_0)$ avente la forma generale

$$v^{(n)} = \begin{bmatrix} 0 \\ y_0 \end{bmatrix} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \lambda_1^n + \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \lambda_2^n, \tag{3.21}$$

E quando $\tau > 1$ prendiamo $y_1 = 0$. Il punto riflesso è dato da

$$r^{(n)} = \begin{bmatrix} 0 \\ y_0 \end{bmatrix} - \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \lambda_1^n(1 - \lambda_1) - \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \lambda_2^n(1 - \lambda_2), \tag{3.22}$$

Stiamo considerando $y_0, x_1 - 1, y_1, x_2,\ e\ y_2 - 1$ vicini a 0. Ripetendo l'analisi della sezione 2 di questo capitolo abbiamo $f(v^{(n)}) > f(v^{(n+1)})$ se e solo se

$$\theta\lambda_1^{\tau n} x_1^\tau((1 + \frac{x_2}{x_1}(\frac{\lambda_2}{\lambda_1})^n)^\tau - (1 + \frac{x_2}{x_1}(\frac{\lambda_2}{\lambda_1})^{n+1})^\tau\lambda_1^\tau)$$

$$+\lambda_1^n(1 - \lambda_1)y_1(1 + 2y_0 + \lambda_1^n(1 + \lambda_1)y_1 + \lambda_2^n(1 + \lambda_2)y_2)$$

$$> \lambda_2^n(1-\lambda_2)y_2(1+2y_0+\lambda_1^n(1+\lambda_1)y_1)+\lambda_2^{2n}(\lambda_2^2-1)y_2^2.$$

Anche $f(v^{(n)}) > f(0,y_0)$ se e solo se

$$\theta\lambda_1^{\tau(n+1)}x_1^\tau(1+\frac{x_2}{x_1}(\frac{\lambda_2}{\lambda_1})^{n+1})^\tau + y_1\lambda_1^{n+1}(1+2y_0+y_1\lambda_1^{n+1}+y_2\lambda_2^{n+1}).$$

$$+y_2\lambda_2^{n+1}(1+2y_0+y_1\lambda_1^{n+1})+y_2^2\lambda_2^{n+1}>0 \tag{3.23}$$

Notiamo che per x_1-1 *ed* x_2 sufficientemente vicini a 0, la coordinata x di $r^{(n)}$ è negativa, così il caso di x negativa per la forma di f vale. Poiché $f(r^{(n)}) > f(v^{(n)})$ se e solo se

$$\theta\lambda_1^{\tau n}x_1^\tau(\phi(1-\lambda_1-\frac{x_2}{x_1}(\frac{\lambda_2}{\lambda_1})^n(1-\lambda_2))^\tau-(1+\frac{x_2}{x_1}(\frac{\lambda_2}{\lambda_1})^n)^\tau$$

$$-y_1\lambda_1^n(2-\lambda_1)(1+2y_0+y_1\lambda_1^{n+1}+y_2\lambda_2^{n+1})$$

$$> y_2\lambda_2^n(2-\lambda_2)(1+2y_0+y_1\lambda_1^{n+1})+y_2^2\lambda_2^n(2-\lambda_2).$$

Dal momento che le disuguaglianze corrispondenti nella sezione precedente sono strette e tutte le funzioni sono continue, segue che esiste un intorno simmetrico di $y_0=0, x_1=1, y_1=0, x_2=0,\ e\ y_2=1$ in cui le suddette tre relazioni valgono per $n=0$. Poiché $|\lambda_1|<1\ e\ |\lambda_2|<1$, segue che se $\tau\leq 1$, le disuguaglianze continuano a valere per ogni $n\geq 0$. Se $\tau>1$, allora il comportamento RFIC non cambierà nell'intorno purchè $y_1=0$. L'insieme delle possibili perturbazioni che mantengono il comportamento RFIC è pertanto di dimensione 4 per $\tau>1$ e di dimensione 5 per $\tau\leq 1$. Per questo ci aspettiamo che il comportamento degli esempi sia stabile contro piccole perturbazioni numeriche causate da un errore di arrotondamento quando $\tau\leq 1$ e non sia stabile quando $\tau>1$. Questo comportamento è in effetti confermato da test numerici. L'errore di arrotondamento introduce una componente dell'autovettore più grande nella direzione y, e questo è sufficiente per impedire che l'algoritmo converga all'origine quando $\tau>1$, ma non è sufficiente per disturbare la convergenza all'origine quando $\tau\leq 1$. Si noti, tuttavia, che nel caso $\tau>1$ il comportamento è molto sensibile alla rappresentazione del problema e ai dettagli dell'implementazione del metodo Mead Nelder e della funzione. Per esempio, una traslazione o rotazione degli assi può influenzare se il metodo converge al minimizzante o no. L'esempio con $\tau=1$ non è strettamente convesso; però, un esempio strettamente convesso che è numericamente stabile può essere costruito prendendo la media degli esempi con $\tau=1\ e\ con\ \tau=2$.

In conclusione, è stata presentata una famiglia di funzioni di due variabili che porta il metodo di Mead Nelder a convergere verso un punto non stazionario, i cui membri

sono funzioni strettamente convesse con un massimo di tre derivate continue. Gli esempi visti portano il metodo Nelder Mead ad eseguire il passo di contrazione interna ripetutamente mantenendo fisso il miglior vertice, ma è stato dimostrato che questo comportamento non può verificarsi per funzioni più lisce. Questi esempi forniscono dunque un limite a ciò che può essere provato sulla convergenza del metodo Nelder Mead.

Ci sono sei valori necessari per specificare il simplesso iniziale per le funzioni di due variabili. È stato dimostrato che, per gli esempi nella famiglia che ha una derivata prima discontinua, esiste un intorno del simplesso iniziale di dimensione 5 in cui tutti i simplessi esibiscono lo stesso comportamento. Questi esempi sembrano essere numericamente stabili. Per tali esempi nella famiglia dove esiste il gradiente, la dimensione dell'intorno è solo 4. Questi esempi sono spesso numericamente instabili ed è quindi meno probabile che si verifichino nella pratica, a causa di errori di arrotondamento, anche per simplessi di partenza all'interno dell'intorno. Tuttavia, anche nei casi in cui gli errori numerici eventualmente perturbano il simplesso abbastanza da sfuggire dal punto di fuoco non stazionario, il metodo può passare un gran numero di passi vicino a questo punto prima dell'allontanamento. Questi risultati evidenziano la necessità di varianti del metodo Nelder Mead originale che hanno garantito proprietà di convergenza.

Capitolo 4

Rilevamento e rimedio della stagnazione del metodo Nelder Mead

Come già osservato nei capitoli precedenti, diversamente dagli algoritmi di ricerca pattern, che mantengono la forma del simplesso, l'algoritmo Nelder-Mead può trovarsi ad un certo punto in una situazione di RFIC, che ricordiamo essere una contrazione interna ripetuta focalizzata, arrivando quindi a ristagnare e convergere ad un punto non ottimale anche per funzioni obiettivo convesse molto semplici e lisce. Adesso, partendo dal concetto di gradiente del simplesso, questo verrà utilizzato per monitorare le prestazioni dell'iterazione Nelder-Mead e per la modifica del passo di restringimento. Verrà descritta quindi una condizione per il sufficiente decremento e mostrato che se la funzione obiettivo f è sufficientemente liscia, le iterate Nelder-Mead soddisfano questa condizione di sufficiente decremento, e i diametri del simplesso convergono a zero in un certo modo, allora qualsiasi punto limite dei vertici del simplesso è stazionario. Verrà proposta inoltre un'alternativa al passo di restringimento che deve essere utilizzata quando la condizione precedente non vale. Questo nuovo passo, che viene chiamato riavvio orientato, reinizializza il simplesso ad uno più piccolo con bordi ortogonali che contiene un passo di discesa ripida approssimata dal miglior punto corrente. Nel capitolo successivo si mostrerà poi come una versione modificata dell'algoritmo Nelder-Mead che incorpora tali idee si comporta sugli esempi già in parte menzionati nel capitolo 3.

4.1 Notazione

Indicheremo con $\|\cdot\|$ la norma l^2 o la norma di matrice indotta e considereremo algoritmi che mantengono un simplesso S di potenziali ottimi con vertici $\{x_j\}_{j=1}^{N+1}$ che soddisfano l'equazione (1.4) definita nel primo capitolo. Sempre da tale capitolo riprendiamo la definizione della matrice nxn di direzioni del simplesso (precedentemente indicata con M_k, adesso per semplicità con V, o $V(S)$):

$$V(S) = (x_2 - x_1, x_3 - x_1, ..., x_{N+1} - x_1) = (v_1, ..., v_N),$$

e del diametro del simplesso:

$$diam(S) = max_{1 \leq i,j \leq N+1} \|x_i - x_j\| .$$

Ci riferiremo all'l^2 numero di condizione $\kappa(V)$ di V [6] come la condizione del simplesso. Sia $\delta(f, S)$ il vettore delle differenze della funzione obiettivo

$$\delta(f, S) = (f(x_2) - f(x_1), f(x_3) - f(x_1), ..., f(x_{N+1}) - f(x_1))^T.$$

Non useremo il diametro del simplesso direttamente nelle nostre stime o algoritmi. Piuttosto utilizzeremo due lunghezze orientate

$$\sigma_+(V) = max_{2 \leq j \leq N+1} \|x_1 - x_j\| \ \ e \ \sigma_-(V) = min_{2 \leq j \leq N+1} \|x_1 - x_j\| .$$

Chiaramente,

$$\sigma_+(S) \leq diam(S) \leq 2\sigma_+(S),$$

Definizione 4.1. Sia S un simplesso con vertici $x_j{}_{j=1}^{N+1}$ ordinati in modo chel'equazione (1.4) valga e $V(S)$ sia non singolare. Il gradiente del simplesso $D(f : S)$ è

$$D(f : S) = V^{-T}\delta(f : S).$$

Questa definizione di gradiente del simplesso è motivata dalla prima stima di ordine:

Lemma 4.1. *Sia S un simplesso con i vertici ordinati in modo che* (1.4) *vale. Sia ∇f continua e Lipschitziana in un intorno di S con costante di Lipschitz $2K$. Allora*

$$\|\nabla f(x_1) - D(f : S)\| \leq K\kappa(V)\sigma_+(S) \tag{4.1}$$

Dimostrazione. Le nostre ipotesi di levigatezza su f e il teorema di Taylor implicano che per ogni $1 \leq j \leq N$,

$$\left\| f(x_1) - f(x_j) + \frac{\delta f(x_1)}{\delta v_j} v_j \right\| = \left\| f(x_1) - f(x_j) + v_j^T \nabla f(x_1) \right\| \leq K \left\| v_j \right\|^2 \leq K \sigma_+(S)^2.$$

Quindi

$$\left\| \delta(f:S) - V^T \nabla f(x_1) \right\| \leq K \sigma_+(S)^2$$

e quindi

$$\left\| \nabla f(x_1) - D(f:S) \right\| \leq K \left\| V^{-T} \right\| \sigma_+(S)^2.$$

La conclusione segue dal fatto che $\sigma_+(S) \leq \left\| V \right\|$. □

Funzioni obiettivo della forma

$$f(x) = g(x) + \phi(x) \tag{4.2}$$

dove g è pensata come una funzione liscia e facile da ottimizzare e $\phi \in L^\infty$ una perturbazione di bassa ampiezza, sono presenti in molte applicazioni. Gli algoritmi che usano differenze di approssimazioni al gradiente di f sono state proposte per problemi di limitazioni vincolati [5] e non vincolati come un modo per evitare l'intrappolamento in minimi locali causati dalla perturbazione. Come gli algoritmi di ricerca pattern, questi metodi differenza richiedono $O(N)$ valutazioni / iterazioni della funzione e di conseguenza possono essere molto meno efficienti nella fase iniziale dell'iterazione rispetto ad un algoritmo simplesso come Nelder-Mead che richiede solo O(1) valutazioni / iterazioni. Uno degli scopi che ci proponiamo è quello di applicare algoritmi simpliciali che utilizzano meno di $O(N)$ valutazioni/ iterazioni di funzione a problemi con tali funzioni obiettivo. Si potrebbe sperare che le diverse dimensioni dei simplessi durante l'iterazione potrebbero contribuire ad evitare minimi locali. Noi avremo bisogno di misurare le perturbazioni su ogni simplesso. A tal fine si definisce per un simplesso S

$$\left\| \phi \right\|_S = ess\,sup_{x \in S} \left\| \phi(x) \right\|.$$

L'analogo del Lemma 4.1 per funzioni obiettivo che soddisfano (4.2) è

Lemma 4.2. *Sia S un simplesso con i vertici ordinati in modo che* (1.4) *vale. Sia f soddisfacente* (4.2) *e sia ∇g continuamente differenziabile in un intorno di S con*

costante di Lipschitz $2K_g$. *Allora, c'è* $K > 0$ *tale*

$$\|\nabla g(x_1) - D(f:S)\| \leq K\kappa(V)(\sigma_+(S) + \frac{\|\phi\|_S}{\sigma_+(S)}) \tag{4.3}$$

Dimostrazione. Il Lemma 4.1 (applicato a g) implica che

$$\|\nabla g(x_1) - D(g:S)\| \leq K_g\kappa(V)\sigma_+(S).$$

Ora, dal momento che $\|\delta(\phi, S)\| \leq 2\sqrt{N}\,\|\phi\|_S$, e $\sigma_+(V) \leq \|V\|$,

$$\|D(f:S) - D(g:s)\| \leq \left\|V^{-T}\right\| \|\delta(f:S) - \delta(g:S)\| = \left\|V^{-T}\right\| \|\delta(\phi:S)\|$$

$$\leq 2N^{1/2}\left\|V^{-T}\right\| \|\phi\|_S \leq 2N^{1/2}\kappa(V)\frac{\|\phi\|_S}{\sigma_+(S)}$$

Questo completa la dimostrazione con $K = K_g + 2N^{1/2}$. □

Le costanti K in (4.1) e (4.3) dipendono da S ed esprimeremo questa dipendenza come $K = K(S)$ quando necessario. Indicheremo i vertici del simplesso S^k all'iterazione k-esima con $\{x_j^k\}_{j=1}^N$. Semplificheremo la notazione sopprimendo la menzione esplicita di S^k in quanto segue denotando

$$V^k = V(S^k), \delta^k = \delta(f:S^k), K^k = K(S^k),\ e\ D^k(f) = D(f:S^k).$$

Se V^0 è non singolare allora V^k è non singolare per ogni $k > 0$. Quindi se V^0 è non singolare, $D^k(f)$ è definito per ogni k. Partiamo dal presupposto che la nostra sequenza di simplessi soddisfa la seguente:
ASSUNZIONE 1.

- V^0 è non singolare.
- I vertici soddisfano (1.4)
- Per ogni k, $\underline{f}^{k+1} \leq \underline{f}^k$.

L'assunzione 1 è soddisfatta dalla sequenza Nelder-Mead, se non vengono presi passi di restringimento e le direzioni del simplesso iniziale sono linearmente indipendenti, mentre non necessita di essere soddisfatta dai metodi di ricerca pattern, in cui vengono imposte solo le condizioni sul miglior valore.

4.2 Decremento sufficiente e il riavvio orientato

Chiederemo che la $k+1^\circ$ iterazione soddisfi [1]:

$$\underline{f}^{k+1} - \underline{f}^{k} < \alpha \left\| D^k f \right\|^2 \tag{4.4}$$

Qui $\alpha > 0$ è un parametro piccolo, una scelta tipica nei metodi di ricerca di linea è $\alpha = 10^{-4}$. Si propone di utilizzare il fallimento di (4.4) come test per imminente stagnazione in un non-minimizzante. Chiaramente, se $\{\left\| D^k f \right\|\}$ non converge a zero, la sequenza di Nelder- Mead è anche non convergente verso un minimizzante.

4.3 Risultati di convergenza

Una conseguenza immediata del Lemma 4.1 [6], valida per f regolare, è

Teorema 4.3. *Sia una sequenza di simplessi soddisfacente l'assunzione 1 e siano valide le ipotesi del Lemma 2.1, con costanti di Lipschitz K^k uniformemente limitate. Allora se* (4.4) *vale per tutti, ma un numero finito di k e il prodotto $\sigma_+(S)^k \kappa(V^k) \to 0$, allora ogni punto di accumulazione dei simplessi è un punto critico di f.*

Dimostrazione. L'assunzione 1 e (4.4) implicano che $\lim_{k\to\infty} D^k f = 0$. Quindi (4.1) implica che

$$\lim_{k\to\infty} \left\| \nabla f(x_1^k) \right\| \leq \lim_{k\to\infty} (K^k \kappa(V^k) \sigma_+(S^k) + \left\| D^k f \right\|) = 0.$$

Quindi, se x^* è un qualsiasi punto di accumulazione della sequenza $\{x_1^k\}$ allora $\nabla f(x^*) = 0$. Questo completa la dimostrazione perché $\sigma_+(V^k) \to 0$. □

Si noti che la conclusione del Teorema 4.3 vale anche se la condizione di sufficiente diminuzione (4.4) è sostituita da

$$\underline{f}^{k+1} - \underline{f}^{k} < -\Phi(D^k f), \tag{4.5}$$

Dove Φ è una funzione monotona crescente su $[0, \infty)$ con $\Phi(0) = 0$. Il risultato per le funzioni rumorose che soddisfano (4.2) con g liscia riflette il fatto che la risoluzione è limitata dalla dimensione di ϕ. Infatti, se $\sigma_+(S^k)$ è molto più piccolo di $\|\phi\|_{S^k}$, nessuna informazione su g può essere ottenuta valutando f ai vertici di S^k e una volta che $\sigma_+(S^k)$ è più piccolo di $\|\phi\|_{S^k}^{1/2}$ nessuna conclusione su ∇g può essere tratta.

Teorema 4.4. *Sia una sequenza di simplessi soddisfacente l'assunzione 1 e siano le ipotesi del Lemma 4.2 valide con costanti di Lipschitz K_g^k uniformemente limitate.*

Allora se (4.4) *vale per tutti, ma un numero finito di* k *e che*

$$\lim_{k\to\infty} \kappa(V^k)(\sigma_+(S^k) + \frac{\|\phi\|_{S^k}}{\sigma_+(S^k)} = 0, \tag{4.6}$$

allora ogni punto di accumulazione dei simplessi è un punto critico di g.

Dimostrazione. Le nostre ipotesi, come nella dimostrazione del Teorema 4.3 implicano che $D^k f \to 0$. Il lemma 4.2 implica che

$$\left\|D^k g\right\| \leq \left\|D^k f\right\| + K^k \kappa(V^k)(\sigma_+(S^k) + \frac{\|\phi\|_{S^k}}{\sigma_+(S^k)}), \tag{4.7}$$

e la sequenza $\{K^k\}$ è limitata. Quindi, per la (4.6), $D^k g \to 0$ *come* $k \to \infty$. □

4.4 Riavvio orientato

Si può monitorare un'iterazione basata sul simplesso per vedere se (4.4) vale, tuttavia, a differenza del caso di un metodo di ricerca di linea basato sul gradiente, ridurre semplicemente la dimensione del simplesso (per esempio, una passo di restringimento in Nelder-Mead) non risolverà il problema. Viene proposta l'esecuzione di un riavvio orientato [6] quando (4.4) non riesce ma $\underline{f}^{k+1} - \underline{f}^k < 0$. Ciò significa sostituire il simplesso corrente con vertici $\{x_j\}_{j=1}^{N+1}$, ordinati in modo che (1.4) vale, con un nuovo simplesso più piccolo avente vertici (prima di ordinare!) $\{y_j\}_{j=1}^{N+1}$ con $y_1 = x_1$ e

$$y_j = y_1 + \beta_{j-1} e_{j-1}, \ per \ 2 \leq j \leq N+1 \tag{4.8}$$

dove e_l è il k-esimo vettore di coordinate

$$\beta_l = \frac{1}{2} \begin{cases} \sigma_-(S^k) sign((D^k f)_l) & (D^k f)_l \neq 0 \\ \sigma_-(S^k) & (D^k f)_l = 0 \end{cases}$$

e $(D^k f)_l$ la componente l-esima di $D^k f$. Se $D^k f = 0$ supponiamo che l'iterazione Nelder-Mead sarebbe stata terminata all'iterazione k perché non ci sarebbe stata nessuna differenza tra il valore migliore e il peggiore. Quindi, prima di ordinare, il nuovo simplesso ha lo stesso primo punto del vecchio. Il diametro del nuovo simplesso non è stato aumentato perché il diametro del nuovo simplesso è al massimo $\sigma_+(S^k)$. Inoltre tutte le lunghezze dei bordi sono state ridotte. Così, dopo il riordino $\sigma_+(S^{k+1}) \leq \sigma_-(S^k)$. Quanto a κ, dopo il restringimento orientato, ma prima del riordino, $\kappa(V) = 1$. Dopo il riordino, naturalmente, il punto migliore potrebbe non essere più x_1. Se il punto migliore è invariato $V^{k+1} = I$. Se il punto migliore è stato

modificato, allora, fino alla permutazione delle righe e moltiplicazione per lo scalare $\pm\sigma_+(S^k)/2, V^{k+1}$, è dato dalla matrice triangolare inferiore

$$V^{k+1} = (V^{k+1})^{-1} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & -1 & 0 & \cdots & 0 \\ \vdots & 0 & -1 & \ddots & \vdots \\ 0 & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & -1 \end{pmatrix}$$

Qui o $V^{k+1} = I$ *e* $\kappa(V) = 1$ oppure l'l^1 numero di condizionamento è $\kappa_1(V^{k+1}) = \left\|V^{k+1}\right\|_1^2 = 4$. Il numero di condizionamento l^2 può essere stimato da

$$\kappa(V^{k+1}) = \left\|V^{k+1}\right\|^2 \leq (1 + \sqrt{N})^2.$$

In ogni caso, il nuovo simplesso è ben condizionato. Il nuovo orientamento del simplesso è destinato a compensare il tipo di stagnazione che era esposto nel capitolo precedente [11], in cui il migliore vertice, che non era un minimizzante, rimaneva invariato durante l'intera iterazione, ed i simplessi convergevano a tale vertice. L'aspettativa è che una volta che il simplesso è abbastanza piccolo, il gradiente del simplesso punti in una direzione che approssimava il vero gradiente e che il simplesso riavviato avrebbe un nuovo migliore vertice. La riduzione è la dimensione del simplesso e, come la riduzione della lunghezza del passo in un metodo di ricerca di linea, dovrebbe essere più facile da soddisfare (4.4).

Nel capitolo seguente, che riporta i test numerici effettuati, verranno applicati sia l'algoritmo tradizionale Nelder Mead che quello modificato, il quale prevede, come abbiamo visto, un riavvio orientato. Verrà altresì messo in luce che quest'ultimo costituisce effettivamente un rimedio al problema della non convergenza.

Capitolo 5

Test numerici

In questo capitolo vengono descritti alcuni dei problemi test analizzati: i primi tre sono delle funzioni riportate anche nel documento originale di Nelder e Mead [12] e tipicamente utilizzate per testare il funzionamento e l'efficienza degli algoritmi di ottimizzazione. E' analizzata poi una funzione in due variabili con due punti di minimo, affrontata per cercare eventuali anomalie o diversità rispetto ai casi con un solo minimizzante, ed infine viene ripreso ed approfondito dal punto di vista pratico il controesempio di Mc Kinnon [11] già citato precedentemente.
Nell'analisi, sono stati messi a confronto due metodi di ottimizzazione, la funzione *fminsearch* di Matlab ed un'implementazione dell'algoritmo Nelder Mead proposta da Kelley [7]. Quest'ultima è stata testata sia nella sua forma basilare, realizzata sulla base dell'idea originale di Nelder Mead, sia nella versione che prevede un riavvio orientato in caso di stagnazione. A questo proposito, nei casi in cui siano stati trovati risultati coincidenti tra le due versioni, questi sono stati riportati per brevità soltanto una volta, classificandoli semplicemente sotto il metodo: "Nelder Mead".
Per ogni problema test, vi è una tabella riassuntiva ed esplicativa contenente:

- il tipo di metodo utilizzato;
- il punto iniziale dato in input all'algoritmo, o nel caso dell'algoritmo Nelder Mead i vertici del simplesso iniziale (indicato con x_0);
- l'errore relativo o assoluto calcolato rispetto al valore esatto del punto di minimo della funzione obiettivo (nel caso delle funzioni test trattate infatti tale punto poteva essere calcolato facilmente con i metodi classici);
- l'errore relativo o assoluto calcolato rispetto al valore esatto della funzione obiettivo nel punto di minimo;

- il numero di iterazioni compiute dall'algoritmo prima di terminare (indicato con K);
- il tempo impiegato dall'algoritmo per arrivare al termine dell'esecuzione (indicato con T).

Il criterio di arresto in entrambi i metodi riguarda la tolleranza sul valore della funzione nel miglior vertice, e cioè $1e^{-10}$, mentre per la scelta dei punti iniziali si è seguita fondamentalmente questa linea: nel caso di *fminsearch*, si sono dati punti iniziali via via più lontani dal valore esatto del punto di minimo, invece con Nelder Mead si è dato ad uno dei vertici lo stesso valore usato nel precedente algoritmo oppure dei punti in modo tale che il simplesso iniziale che si veniva a creare contenesse al suo interno uno degli x_0 della funzione *fminsearch*. Sono presenti inoltre dei grafici che riportano, per ogni iterazione effettuata, il valore della funzione obiettivo nel miglior vertice, e nel caso dell'algoritmo Nelder Mead, anche l'errore relativo rispetto al valore esatto del minimo.
Tutti i risultati riportati sono stati ottenuti utilizzando una licenza accademica di Matlab R2016b su un sistema operativo Windows10 a 64 bit, con processore Intel Core i7-4710HQ.

5.1 Funzione di Rosenbrock

$$y = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

Valore esatto del punto di minimo: (1,1)
Valore esatto della funzione nel punto di minimo: 0.
Nella tabella 5.1 i risultati trovati. Si può notare che in tutti i casi analizzati entrambi gli algoritmi convergono al punto di minimo atteso, con un errore dell'ordine di 10^{-11} nel caso di *fminsearch*, e di 10^{-6} con NM. L'errore relativo al valore della funzione nel punto è invece dell'ordine di 10^{-11} tranne nei primi due punti in cui si è analizzato la funzione *fminsearch*, ma bisogna tenere anche conto del fatto che entrambi fossero molto vicini al valore esatto del minimo. Si osserva a tal proposito che dando un valore iniziale molto lontano da quello da trovare, come nel caso del punto (-8,-300), la funzione *fminsearch* è sempre convergente, ma l'errore in questo caso è decisamente peggiore rispetto agli altri e soprattutto l'algoritmo termina non perchè sia stata soddisfatta la condizione sulla tolleranza, bensì perchè raggiunto il limite massimo delle iterazioni previste.
Nelle figure 5.1 e 5.2 vediamo l'fminsearch applicato nei due casi opposti, cioè con

$x_0 = (-1.2, 1)$, in cui una delle coordinate coincide con quella del valore esatto del minimo, e con $(-8, -300)$, molto lontano da esso, con i rispettivi errori relativi. Nella 5.3 e 5.4 il comportamento dell'algoritmo Nelder Mead.

Metodo	x_0	Err. rel. x^*	Err. ass. f^*	K	T
fminsearch	(-1.2,1)	3.8167e-11	5.8326e-22	132	0.024211
	(1,-1.2)	2.3309e-11	2.3673e-22	127	0.009097
	(5,3)	2.1264e-11	2.3793e-11	138	0.004126
	(-8,-300)	5.5025e-06	2.3793e-11	215 (max)	0.005542
Nelder Mead	(-1.2,1) (0,0) (3,3)	4.7864e-06	5.9920e-11	78	0.006680
	(3,3) (5,6) (5,-1)	4.9743e-0.6	1.6121e-11	94	0.006308
	(5,-1) (-2,-1) (0,6)	1.2226e-06	3.1538e-12	74	0.004277

Tabella 5.1: Funzione di Rosenbrock

5.2 Funzione di Powell

$$y = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

Valore esatto del punto di minimo: (0,0,0,0)
Valore esatto della funzione nel punto di minimo: 0.
Anche in questo caso entrambi gli algoritmi convergono al minimo effettivo della funzione, anche se con un errore complessivamente peggiore rispetto al problema test precedente. Inoltre, come si vede dalla tabella 5.2, per tutti i valori dati in input alla funzione *fminsearch* il metodo termina non perchè sia stato soddisfatto il requisito sulla tolleranza, bensì perchè raggiunto il numero massimo di iterazioni previste. Nelle figure 5.5 e 5.6 i risultati ottenuti con il primo algoritmo, nelle 5.7 e 5.8 quelli relativi a Nelder Mead.

5.3 Funzione di Fletcher

$$100(x_3 - \frac{10atan(x_2, x_1)}{2\pi})^2 + (sqrt(x_1^2 + x_2^2) - 1)^2 + x_3^2$$

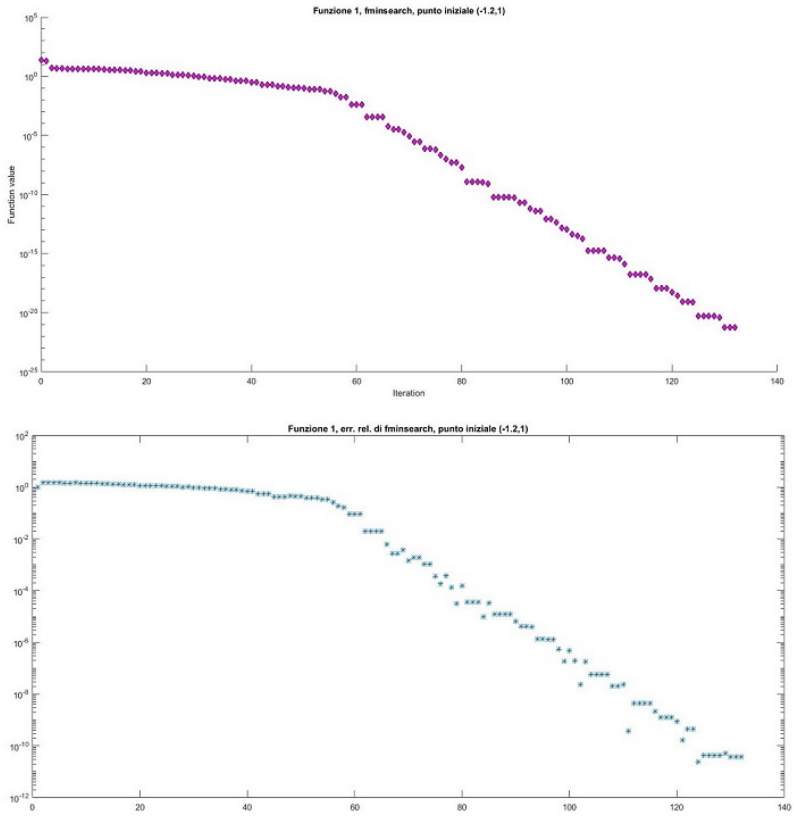


Figura 5.1: Valori della funzione ed errore relativo fminsearch, funzione di Rosenbrock, primo esempio

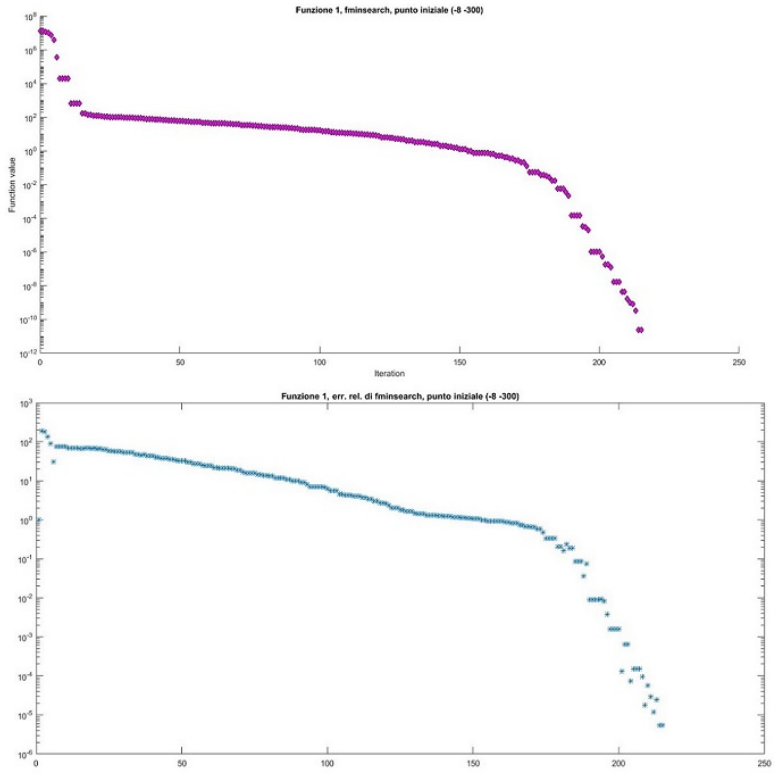


Figura 5.2: Valori della funzione ed errore relativo fminsearch, funzione di Rosenbrock, secondo esempio

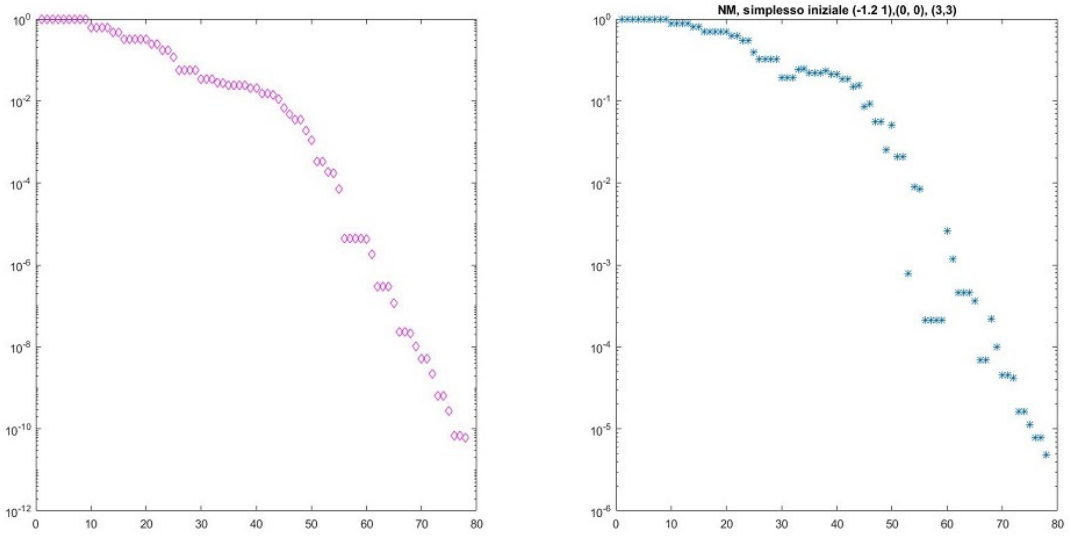


Figura 5.3: Valori della funzione ed errore relativo Nelder Mead, funzione di Rosenbrock, primo esempio

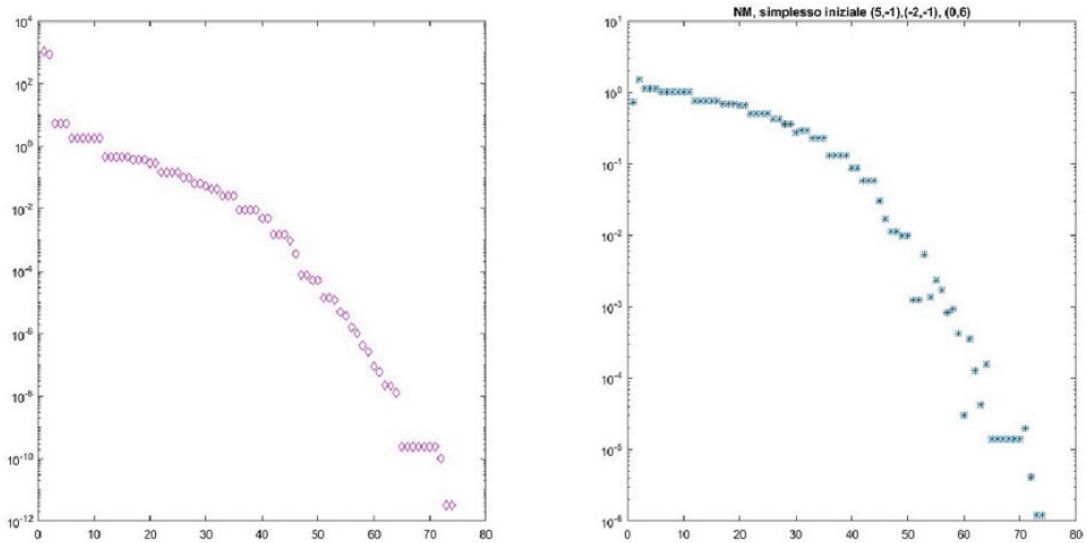


Figura 5.4: Valori della funzione ed errore relativo Nelder Mead, funzione di Rosenbrock, secondo esempio

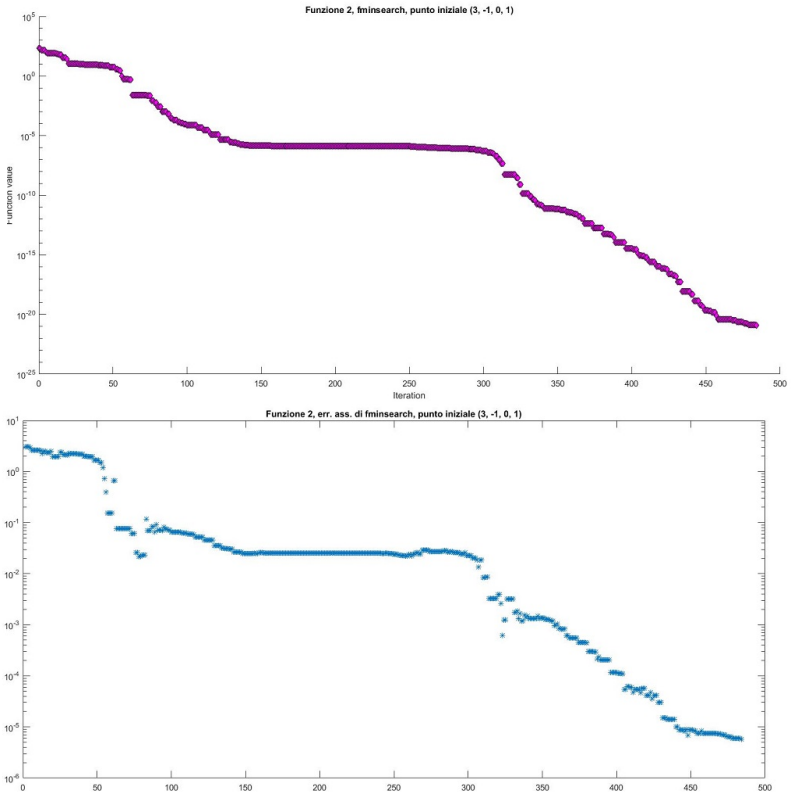


Figura 5.5: Valori della funzione ed errore assoluto fminsearch, funzione di Powell, primo esempio

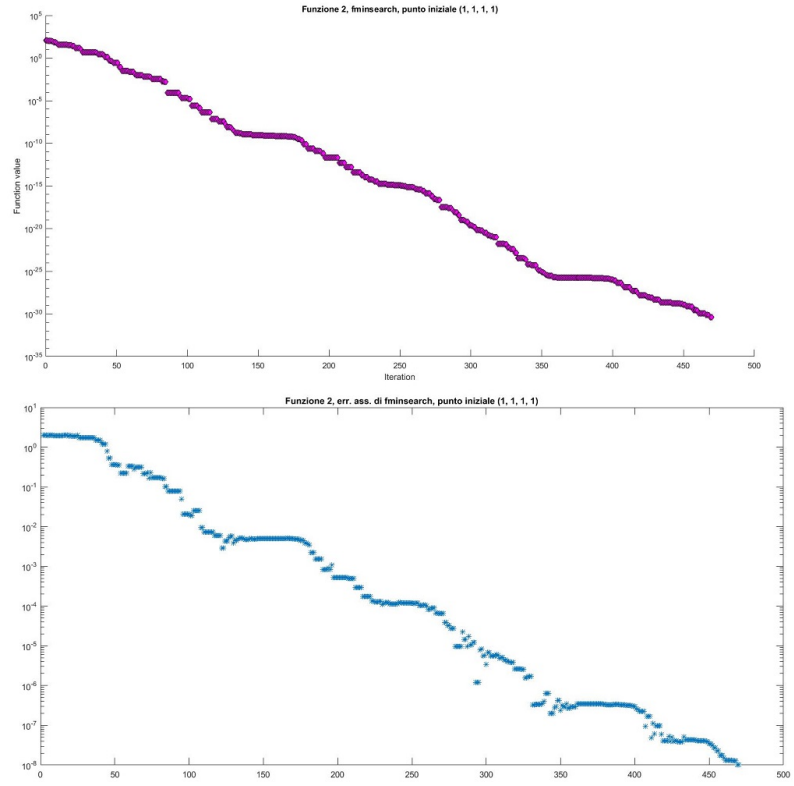


Figura 5.6: Valori della funzione ed errore assoluto fmnisearch, funzione di Powell, secondo esempio

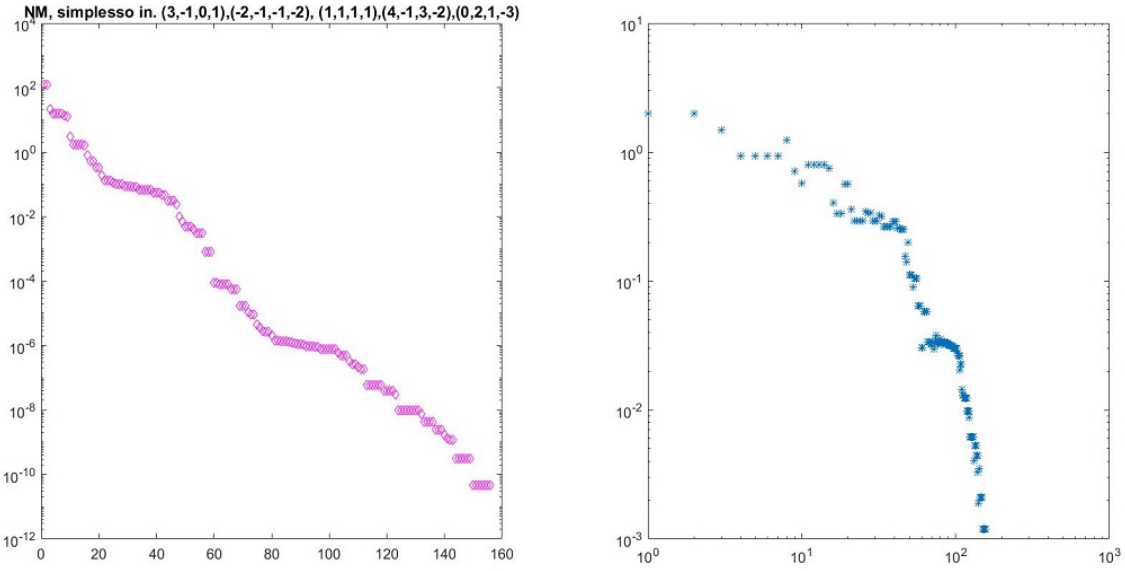


Figura 5.7: Valori della funzione ed errore assoluto Nelder Mead, funzione di Powell, primo esempio

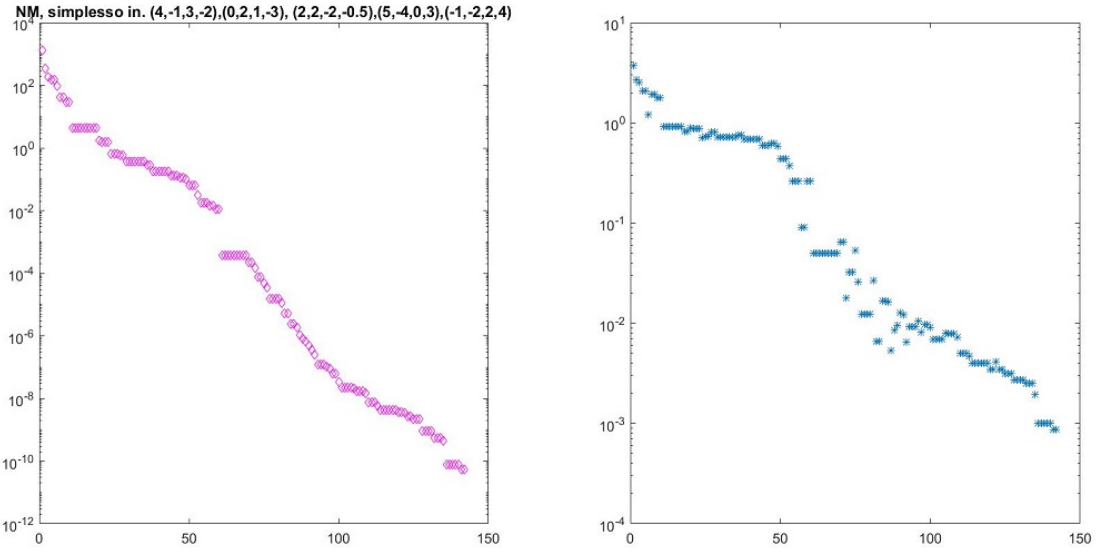


Figura 5.8: Valori della funzione ed errore assoluto Nelder Mead, funzione di Powell, secondo esempio

Metodo	x_0	Err. ass. x^*	Err. ass. f^*	K	T
fminsearch	(3,-1,0,1)	5.7642e-06	1.2457e-21	484 (max)	0.010921
	(0,-5,1,2)	1.5984e-04	2.1219e-15	474 (max)	0.0110015
	(1,1,1,1)	1.0250e-08	4.1242e-31	470 (max)	0.008169
	(-10,-10,-10,-10)	2.8454e-08	6.8832e-31	472 (max)	0.012942
Nelder Mead	(3,-1,0,1) (-2,-1,-1,-2) (1,1,1,1) (4,-1,3,-2) (0,2,1,-3)	0.0012	4.5806e-11	156	0.013162
	(4,-1,3,-2) (0,2,1,-3) (2,2,-2,-0.5) (5,-4,0,3) (-1,-2,2,4)	8.7034e-04	5.2861e-11	142	0.011646
	(3,-1,0,1) (-2,-1,-1,-2) (0,2,0,10) (-1,1,2,-5) (1,-3,3,-7)	0.0014	2.1870e-11	164	0.011791

Tabella 5.2: Funzione di Powell

Valore esatto del punto di minimo: (1,0,0)
Valore esatto della funzione nel punto di minimo: 0.
Anche in questo caso entrambi gli algoritmi sono convergenti, e l'ordine di grandezza deli errori è più che accettabile. La tabella relativa ai risultati è la 5.3; inoltre la figura 5.9 riguarda la funzione *fminsearch* e il corrispondente errore relativo con x_0 abbastanza vicino al minimo effettivo, la 5.10 con x_0 abbastanza lontano, ed infine i grafici 5.11 e 5.12 rappresentano l'esecuzione Nelder Mead.

5.4 Funzione con due punti di minimo

$$y = 2(x_1^4 + x_2^4 + 1) - (x_1 + x_2)^2$$

Valori esatti dei punti di minimo: $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}), (-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$
Valore esatto della funzione nei punti di minimo: 1.
La prima cosa che si può notare dalla tabella riassuntiva relativa a tale funzione, la 5.4, è che pur essendoci due punti di minimo, entrambi gli algoritmi convergono ad uno dei due di tali punti impiegando un numero di iterazioni decisamente inferiore rispetto agli esempi visti finora, fatto che ovviamente si riflette anche sul tempo

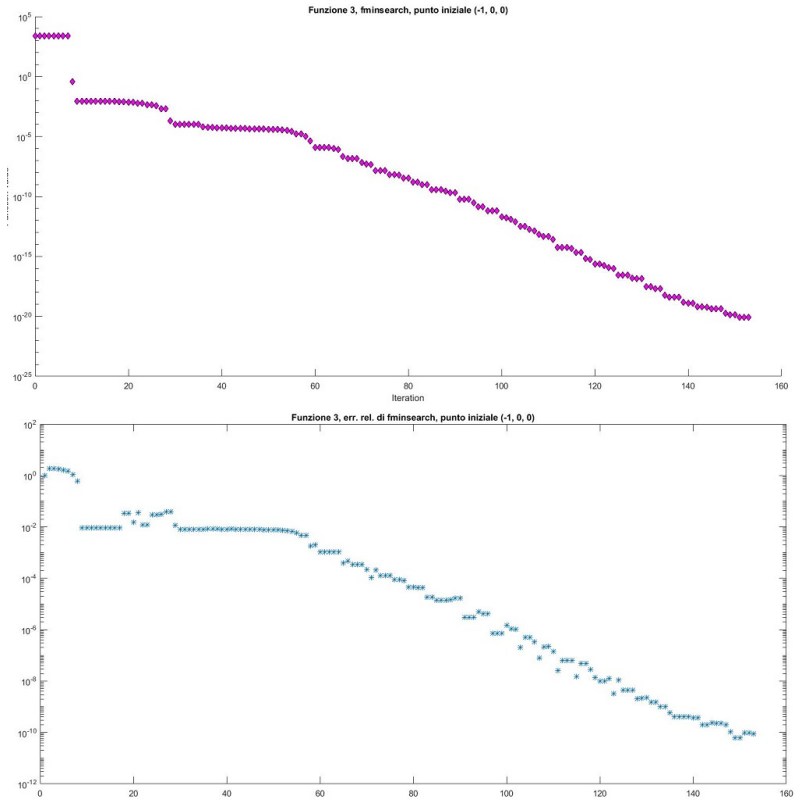


Figura 5.9: Valori della funzione ed errore relativo fminsearch, funzione di Fletcher, primo esempio

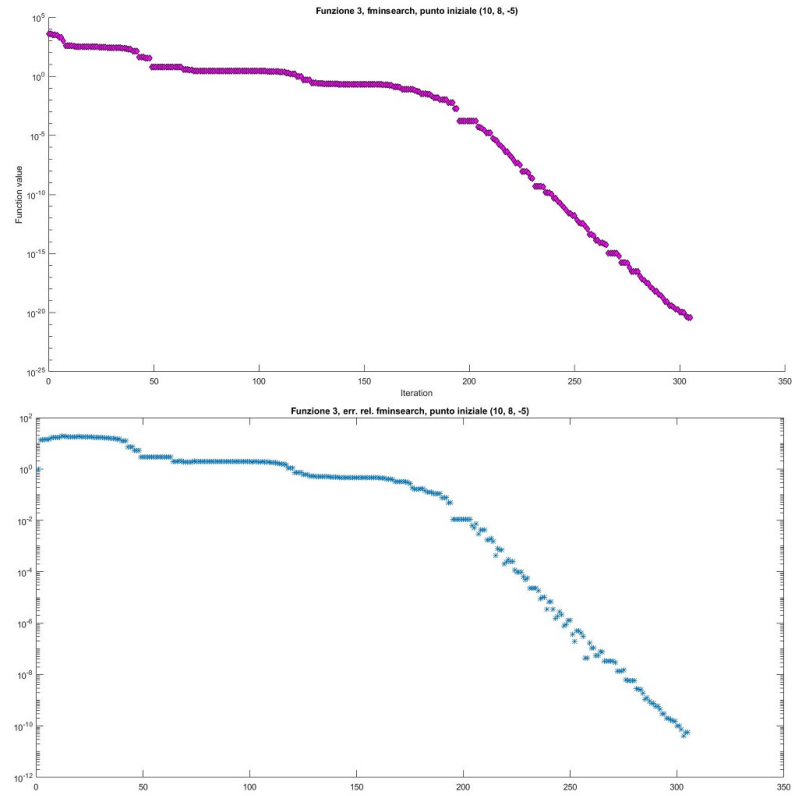


Figura 5.10: Valori della funzione ed errore relativo fminsearch, funzione di Fletcher, secondo esempio

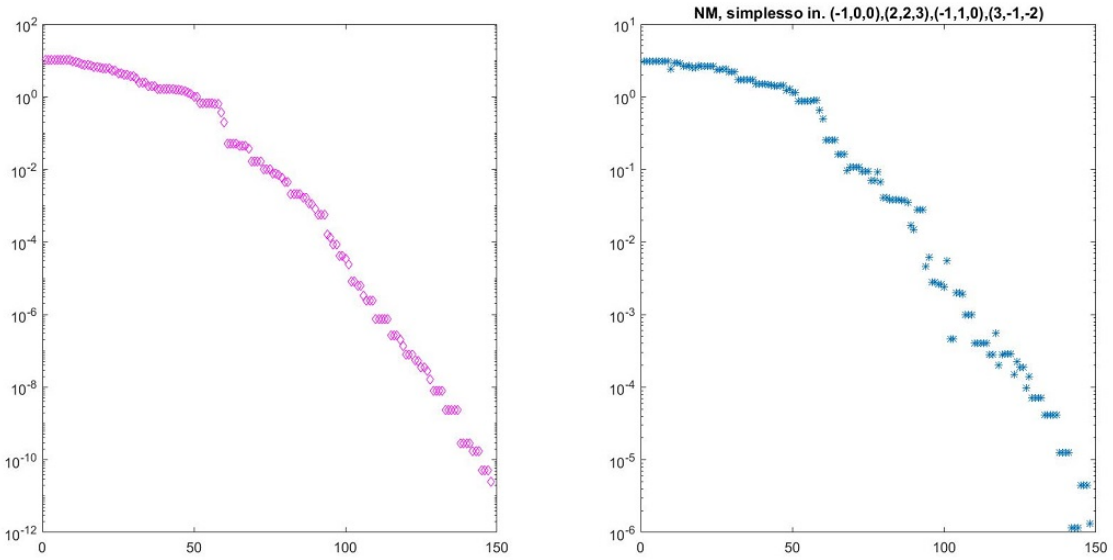


Figura 5.11: Valori della funzione ed errore relativo Nelder Mead, funzione di Fletcher, primo esempio

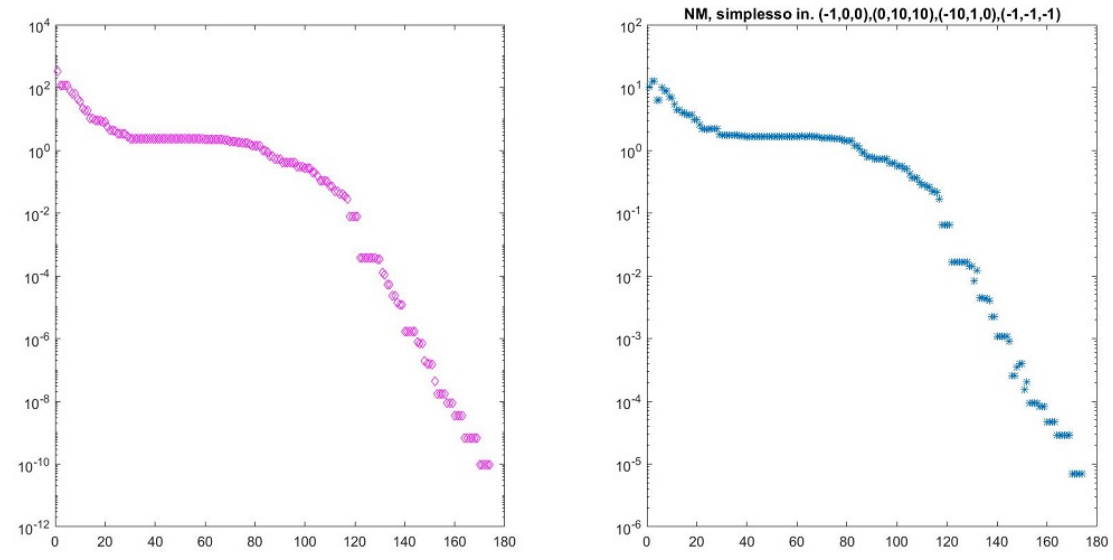


Figura 5.12: Valori della funzione ed errore relativo Nelder Mead, funzione di Fletcher, secondo esempio

Metodo	x_0	Err. ass. x^*	Err. ass. f^*	K	T
fminsearch	(-1,0,0)	8.9248e-11	8.0777e-21	153	0.004371
	(-1,-1,-1)	4.2104e-11	1.8432e-21	280	0.005645
	(10,8,-5)	5.6244e-11	3.9843e-21	305	0.005982
	(0,1,2)	1.9912e-11	2.4536e-21	287	0.005570
Nelder Mead	(-1,-1,-1) (-2,-5,-2) (-1,-0.7,-0.2) (-3,-2,-3,-2)	6.1755e-06	6.4049e-11	168	0.011984
	(-1,0,0) (2,2,3) (-1,1,0) (3,-1,-2)	1.3241e-0.6	2.4589e-11	148	0.010593
	(-1,0,0) (0,10,10) (-10,1,0) (-1,-1,-1)	7.0640e-06	9.6039e-11	174	0.011592

Tabella 5.3: Funzione di Fletcher

di esecuzione. La convergenza ad uno dei minimi piuttosto che all'altro dipende ovviamente dalla scelta del punto iniziale: si è osservato che in linea di massima se entrambe le coordinate di tale punto sono positive, il punto di minimo trovato sarà $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$, altrimenti l'altro. Nello specifico, per gli esempi riportati in tabella, *fminsearch* converge a $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ solo nel caso in cui $x_0 = (3, 5)$, mentre invece l'algoritmo Nelder Mead solo quando il simplesso iniziale ha vertici $(0, 0), (2, 10), (10, 15)$. Le figure dalle 5.13 alla 5.16 sono relative ad alcuni dei risultati ottenuti.

5.5 Il controesempio di Mc Kinnon

$$f(x,y) = \begin{cases} \theta\phi|x|^\tau + y + y^2 \ x \leq 0 \\ \theta x^\tau + y + y^2 \ x \geq 0 \end{cases}$$

Negli esempi analizzati sono divisi in tre sottocasi, in cui i parametri τ, ϕ, θ assumono rispettivamente i valori:

$$(\tau, \theta, \phi) = \begin{cases} (1, 15, 10), \\ (2, 6, 60), \\ (3, 6, 400) \end{cases}$$

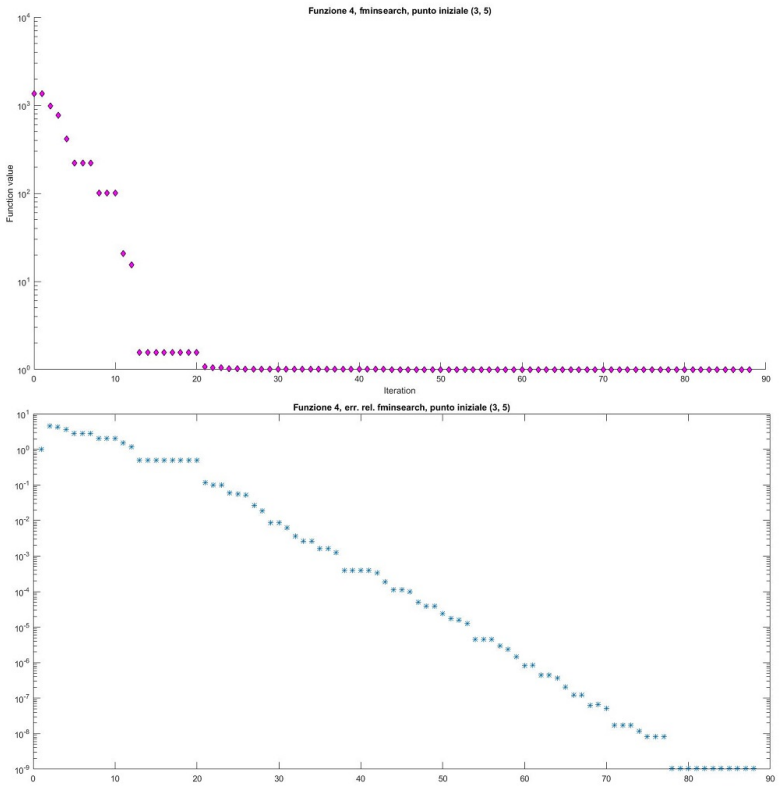


Figura 5.13: Valori della funzione ed errore relativo fminsearch, funzione con due punti di minimo, primo esempio

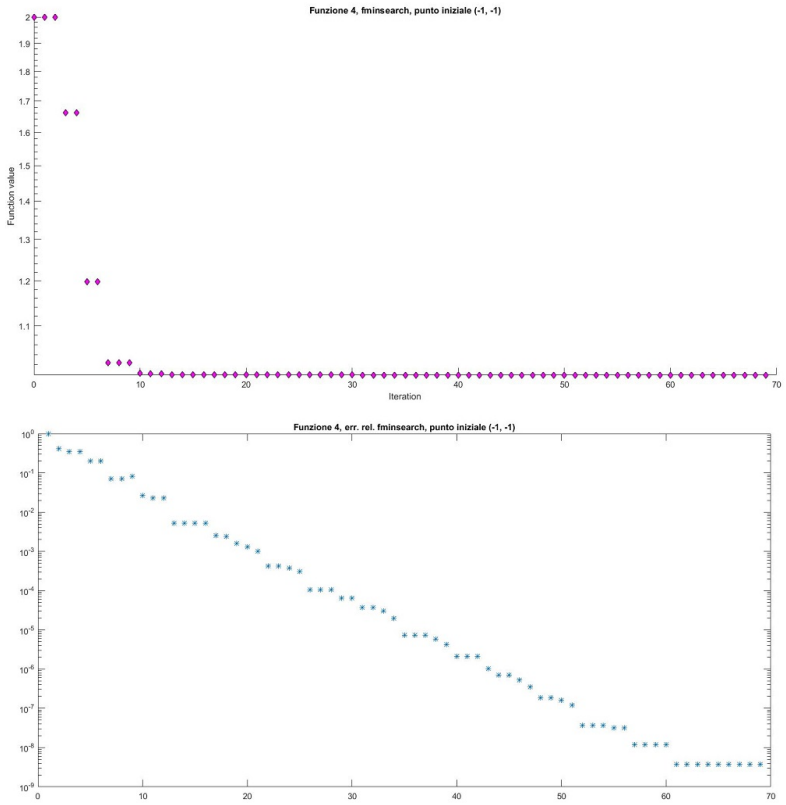


Figura 5.14: Valori della funzione ed errore relativo fminsearch, funzione con due punti di minimo, secondo esempio

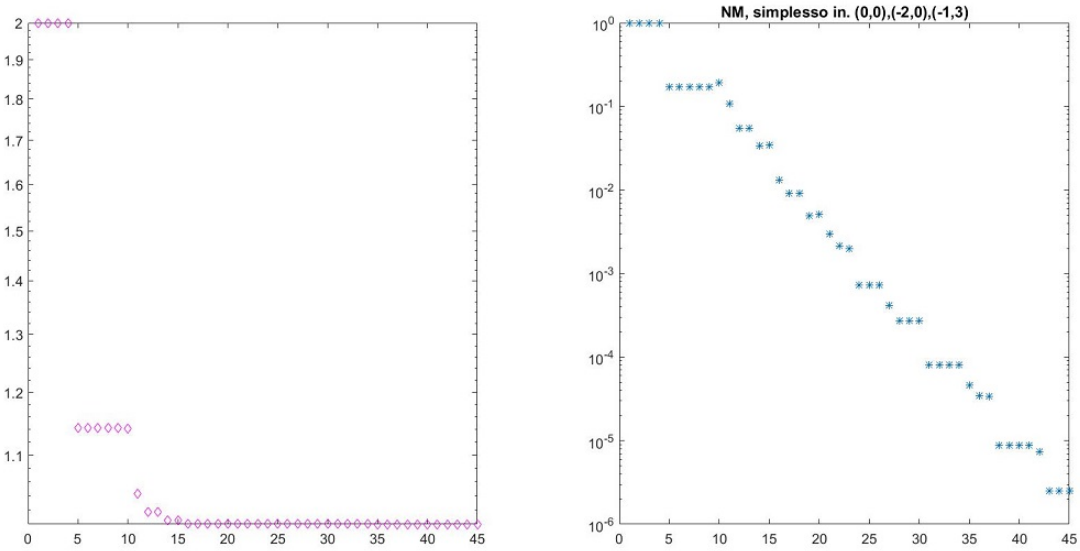


Figura 5.15: Valori della funzione ed errore relativo Nelder Mead, funzione con due punti di minimo, primo esempio

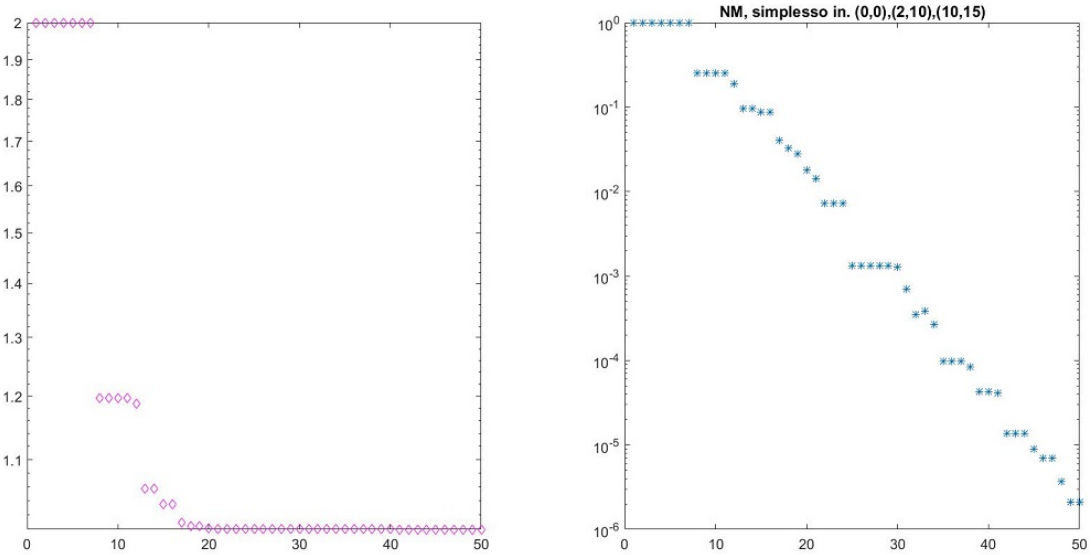


Figura 5.16: Valori della funzione ed errore relativo Nelder Mead, funzione con due punti di minimo, secondo esempio

Metodo	x_0	Err. ass. x_k	Err. ass. f_k	K	T
fminsearch	(-1,1)	1.0287e-08	0	86	0.003588
	(3,-5)	2.9669e-09	4.4409e-16	86	0.003141
	(3,5)	1.0357e-09	4.4409e-16	88	0.003265
	(-1,-1)	3.7511e-09	4.4409e-16	69	0.003006
Nelder Mead	(-1,1) (-1,-3) (2,-3)	2.6702e-06	2.8522e-11	46	0.003238
	(0,0) (-2,0) (-1,3)	2.15135e-06	3.5437e-11	45	0.003174
	(0,0) (2,10) (10,15)	2.0964e-06	2.2326e-11	50	0.002868

Tabella 5.4: Funzione con due punti di minimo

Un'osservazione generale valida in tutte e tre le situazioni è che, a partire dal simplesso iniziale avente vertici

$$x_1 = (1,1)^T, x_2 = (\lambda_+, \lambda_-)^T, x_3 = (0,0)^T, \ dove\ \lambda_\pm = (1 \pm \sqrt{33})/8 \tag{5.1}$$

l'iterazione Nelder-Mead sarà ristagnante all'origine, che non è un punto critico per f.

Nell'analisi effettuata da Mc Kinnon [11], sono presenti anche delle figure nelle quali viene tracciato, come funzioni dell'indice di iterazione, la differenza tra il migliore ed il peggiore valore della funzione, σ_+, la lunghezza del massimo orientato, la norma del gradiente del simplesso, e l'l^2 numero di condizionamento della matrice di direzioni del simplesso. In tutti e tre i problemi la stagnazione è evidente dal comportamento dei gradienti del simplesso. Si noti anche come il numero di condizionamento del simplesso è in rapida crescita. Tali figure vengono riportate, per completezza, suddivise nei vari sottocasi studiati.

5.5.1 $(\tau, \theta, \phi) = (1, 15, 10)$

In questo caso, essendo la funzione non differenziabile, vengono messi a confronto soltanto l'algoritmo originale Nelder Mead e la sua versione modificata, non utilizzando invece l'fminsearch. Di conseguenza nella tabella 5.5, relativa ai risultati ottenuti, sono stati riportati il punto di convergenza e il corrispondente valore della funzione in tale punto, piuttosto che l'errore.

Nella figura 5.18, si vede come l'algoritmo modificato termini con il fallimento dopo

il riavvio sulle iterazioni 30, 31, e 32. Poiché l'obiettivo è non regolare al punto di stagnazione, questo è il meglio che possiamo aspettarci ed è di gran lunga migliore del comportamento dell'algoritmo non modificato, che ristagna senza avvertimento dell'errore.
Nella figura 5.19 i valori della funzione nel migliore vertice rispettivamente nel caso dell'algoritmo non modificato e in quello modificato. Nel caso del secondo simplesso riportato in tabella, si ha un risultato finale che è pressocchè lo stesso con le due implementazioni, motivo per cui è stato riportato un solo grafico relativo a questo esempio, cioè la figura 5.20.

Metodo	Simplesso iniziale	x^*	$f(x^*)$	K	T
NM	(0,0) $(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$ (1,1)	(0,0)	0	44	0.010140
	(0,0) (-2,0) (-1,5)	(0,-0.5087)	-0.2495	36	0.010988
N.M modificato	(0,0) $(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$ (1,1)	e-0.3(0,-0.4103)	-4.1010e-04	46	0.006048
	(0,0) (-2,0) (-1,5)	(0,-0.5082)	-0.2495	42	0.005547

Tabella 5.5: Controesempio di MC Kinnon, caso 1

5.5.2 $(\tau, \theta, \phi) = (2, 6, 60)$

In questo caso, come nel successivo, la funzione è differenziabile e si ha che il valore esatto del punto di minimo è $(0, \frac{1}{2})$, ed il valore esatto della funzione nel punto di minimo $\frac{1}{4}$.
In questo caso la funzione *fminsearch* converge al punto di minimo con un errore dell'ordine di 10^{-9}, mentre nel caso dell'algoritmo NM si ha che:

- col simplesso iniziale di vertici (5.1) NM converge all'origine che come è stato più volte detto non è un punto di minimo, invece la sua forma modificata converge al valore corretto, con un errore dell'ordine di 10^{-5}, effettuando un riavvio orientato già alla diciassettesima iterazione;

- cambiando vertici iniziali entrambi riescono a convergere al valore coretto.

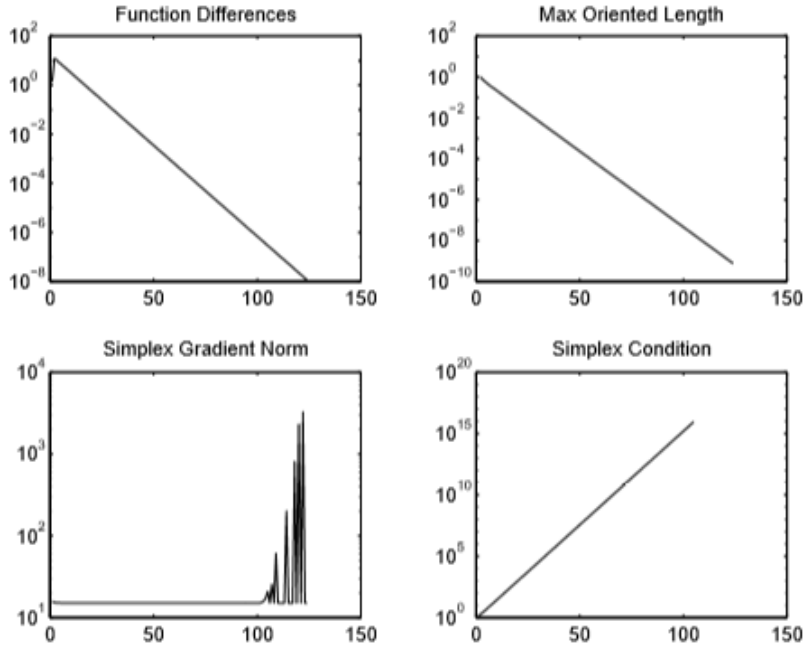


Figura 5.17: Nelder Mead non modificato, caso 1

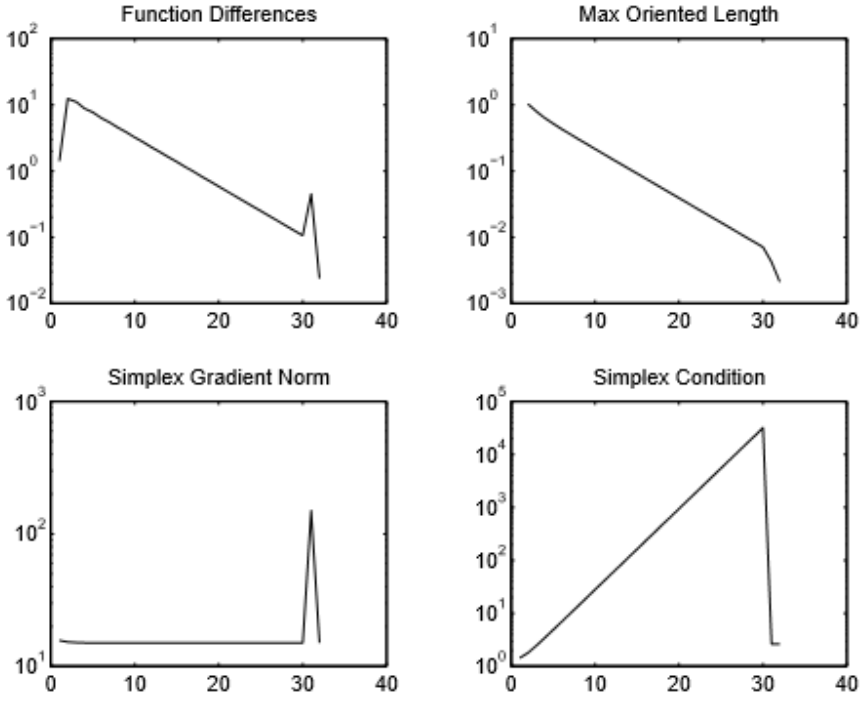


Figura 5.18: Nelder Mead modificato, caso 1

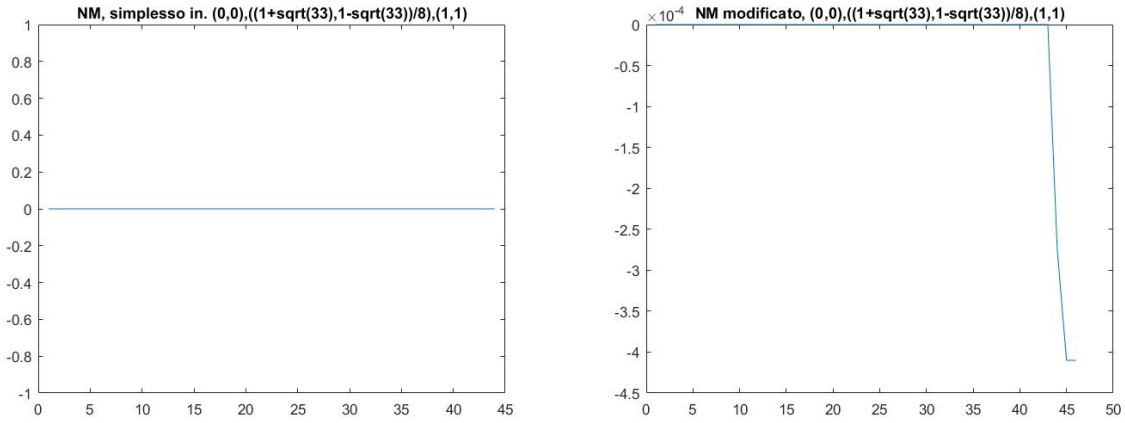


Figura 5.19: Valori della funzione di Mc Kinnon, caso 1, NM e NM modificato, primo esempio

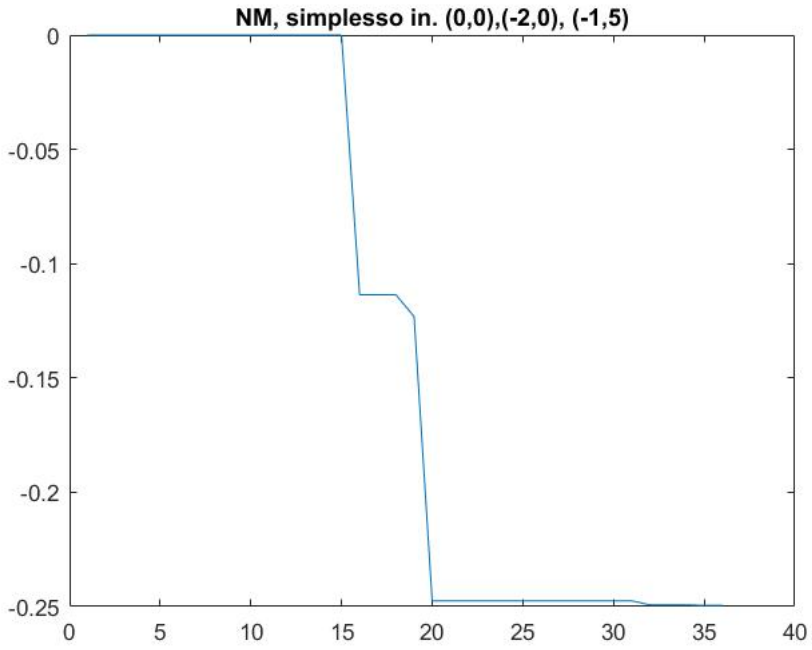


Figura 5.20: Valori della funzione di Mc Kinnon, caso 1,Nelder Mead, secondo esempio

In tabella 5.6 i risultati ottenuti. Nelle figure 5.21 e 5.22 i grafici riportati da Kinnon, nella 5.23, 5.24 il comportamento di *fminsearch* con il corrispondente errore relativo, nella 5.25 e 5.26 l'applicazione rispettivamente di NM e NM modificato, con i rispettivi errori, e infine nella 5.27 ciò che accade col secondo simplesso iniziale, con cui come abbiamo visto il comportamento dei due metodi coincide.

Metodo	x_0	Err. rel. x_k	Err. rel. f_k	K	T
fminsearch	(1,1)	2.4761e-09	0	114	0.004514
	(10,10)	5.0530e-09	0	117	0.004350
	$(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$	1.5809e-09	0	99	0.003937
NM	(0,0) $(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$ (1,1)	1	1	19	0.019944
	(0,0) (2,0) (1,3)	1.1761e-05	1.5396e-10	56	0.003472
NM modificato	(0,0) $(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$ (1,1)	3.1748e-05	1.0313e-09	68	0.011058
	(0,0) (2,0) (1,3)	1.1761e-05	1.5396e-10	56	0.003472

Tabella 5.6: Controesempio di MC Kinnon, caso 2

5.5.3 $(\tau, \theta, \phi) = (3, 6, 400)$

Anche in questo caso il punto di minimo esatto è $(0, \frac{1}{2})$, e la funzione *fminsearch* converge a tale punto, stavolta con un errore dell'ordine di 10^{-6}. Per quanto riguarda invece le due versioni dell'algoritmo NM, valgono le stesse considerazioni fatte nel caso precedente. Stavolta la forma modificata ha effettuato un solo riavvio orientato alla diciannovesima iterazione. Come si può vedere dalle figure 5.28 e 5.29 il riavvio ha avuto un effetto immediato sulla norma del gradiente del simplesso e ha superato la stagnazione. In tabella 5.7 i risultati ottenuti. Nelle figure 5.30, 5.31 il comportamento di *fminsearch*, nella 5.32 e 5.33 l'applicazione rispettivamente di NM e NM modificato, con i rispettivi errori, e infine nella 5.34 ciò che accade col secondo simplesso iniziale, con il quale il comportamento dei due metodi coincide.

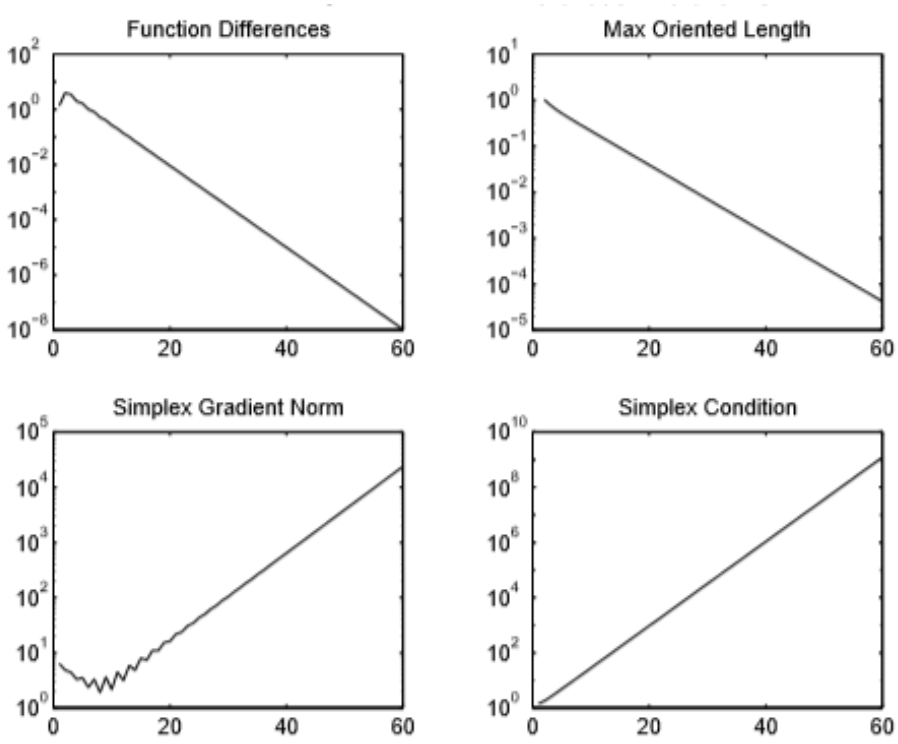


Figura 5.21: Nelder Mead non modificato, caso 2

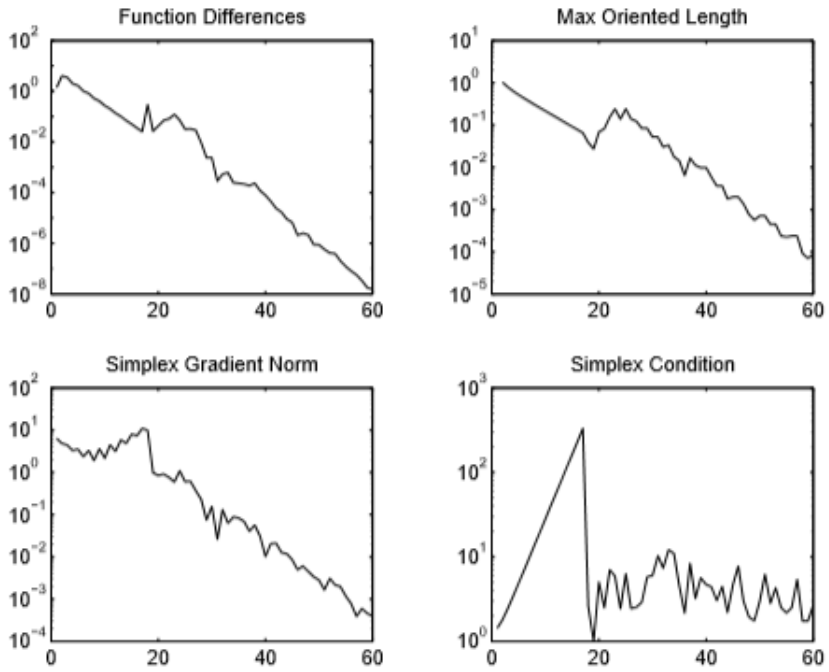


Figura 5.22: Nelder Mead modificato, caso 2

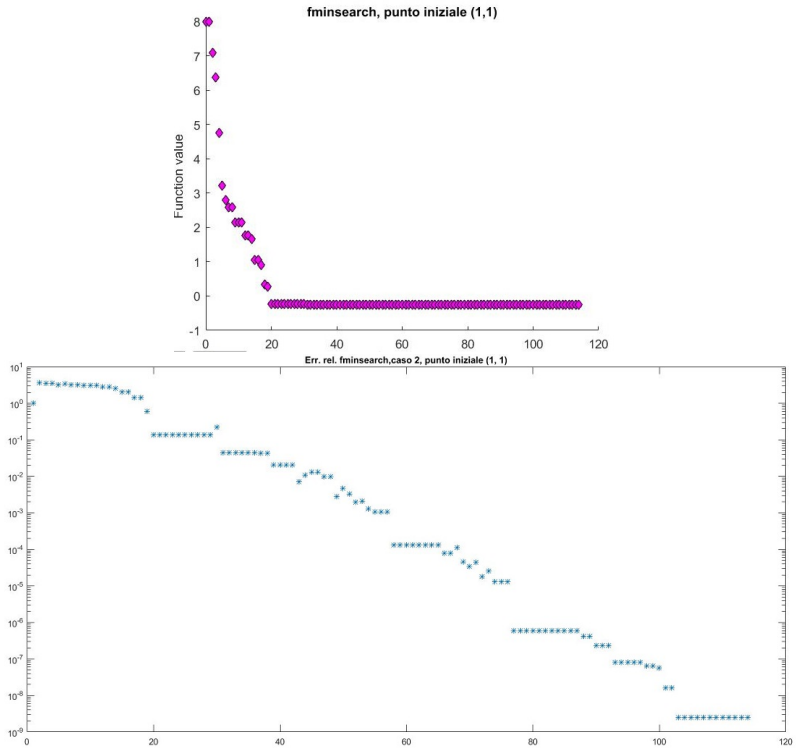


Figura 5.23: Valori della funzione ed errore relativo fminsearch, funzione di MC Kinnon,caso 2, primo esempio

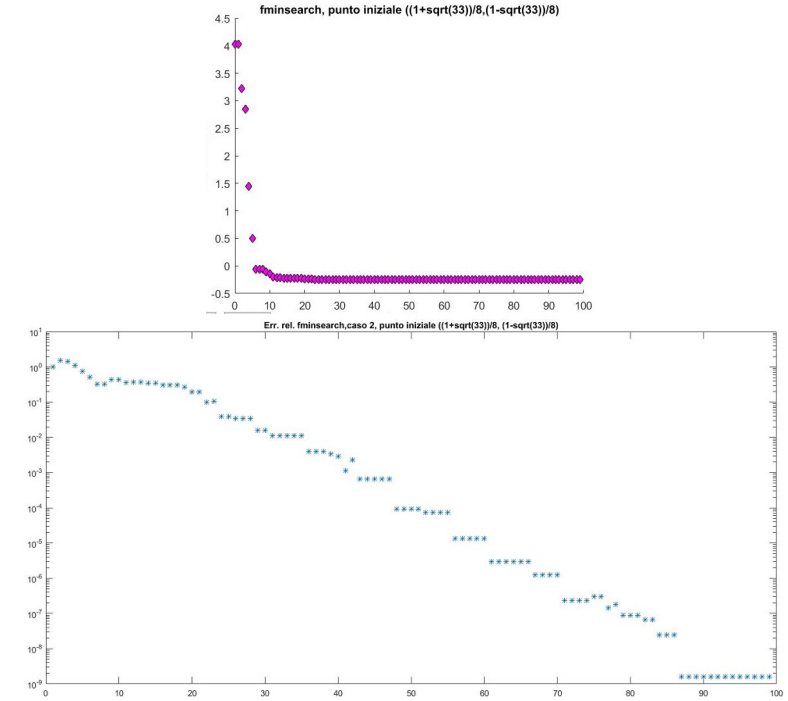


Figura 5.24: Valori della funzione ed errore relativo fminsearch, funzione di Mc Kinnon, caso 2, secondo esempio

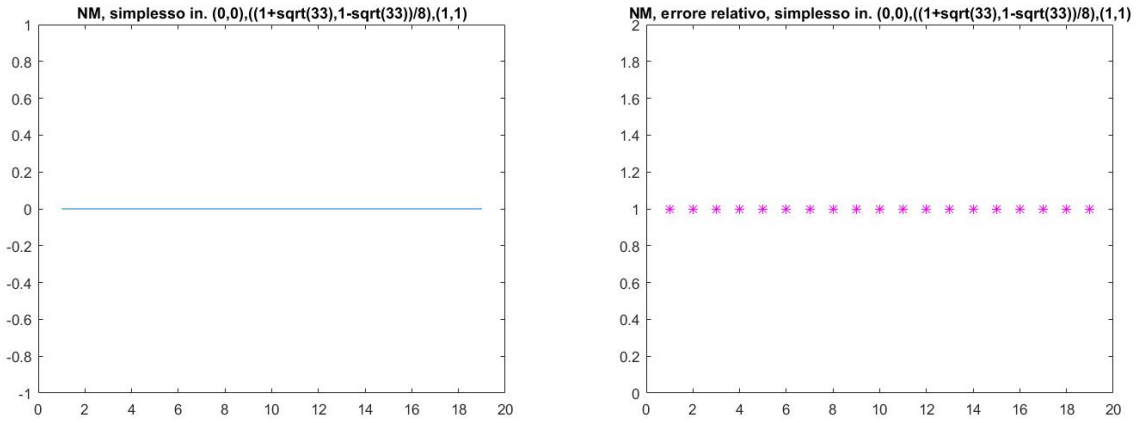


Figura 5.25: Valori della funzione ed errore relativo Nelder Mead, funzione di MC Kinnon, caso2, primo esempio

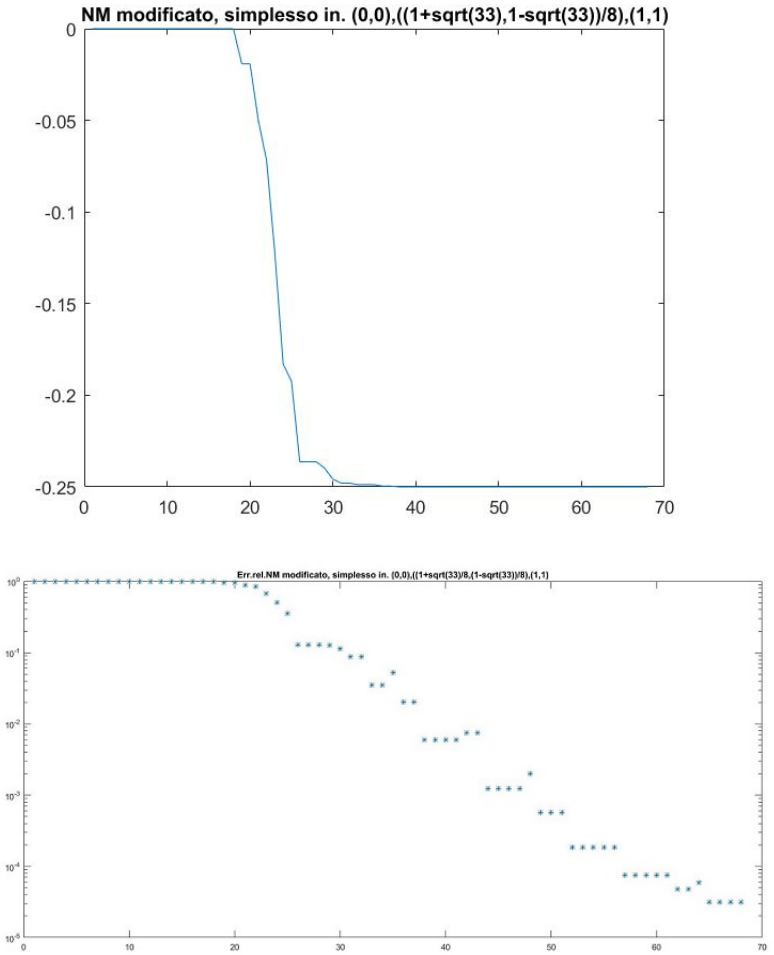


Figura 5.26: Valori della funzione ed errore relativo Nelder Mead modificato, funzione di MC Kinnon, caso 2, primo esempio

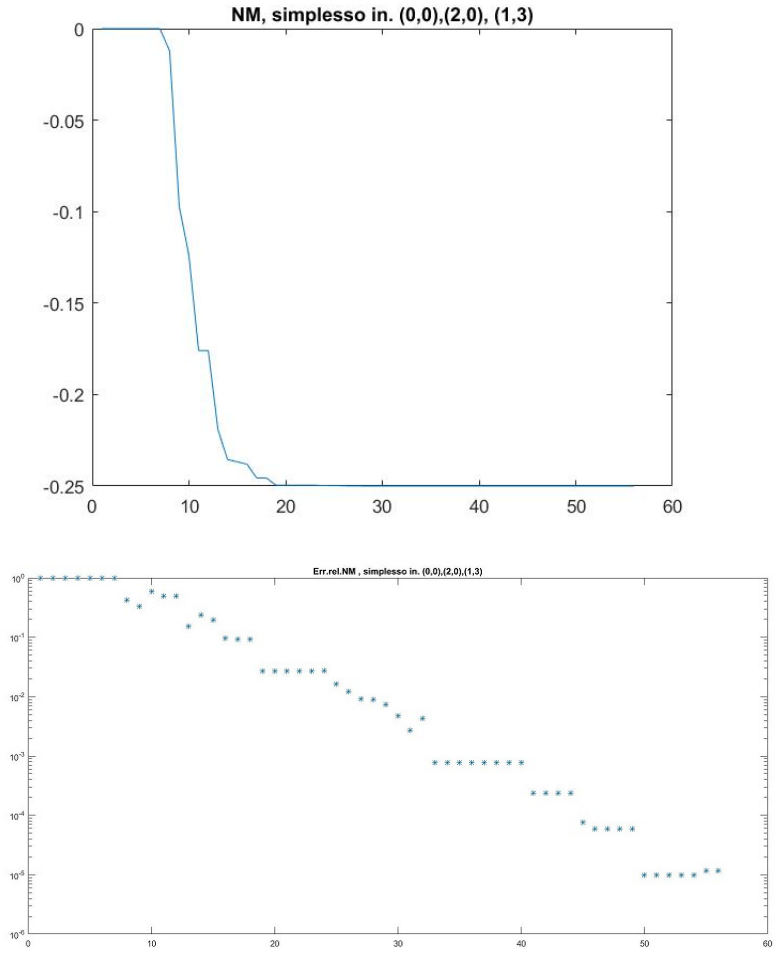


Figura 5.27: Valori della funzione ed errore relativo Nelder Mead, funzione di MC Kinnon, caso2, secondo esempio

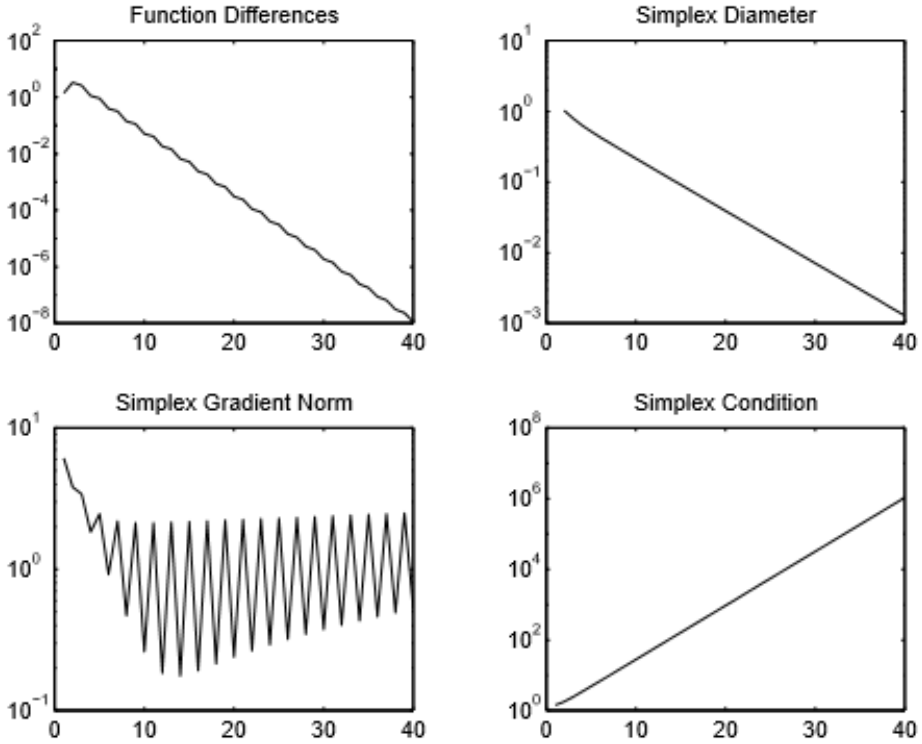


Figura 5.28: Nelder Mead non modificato, caso 3

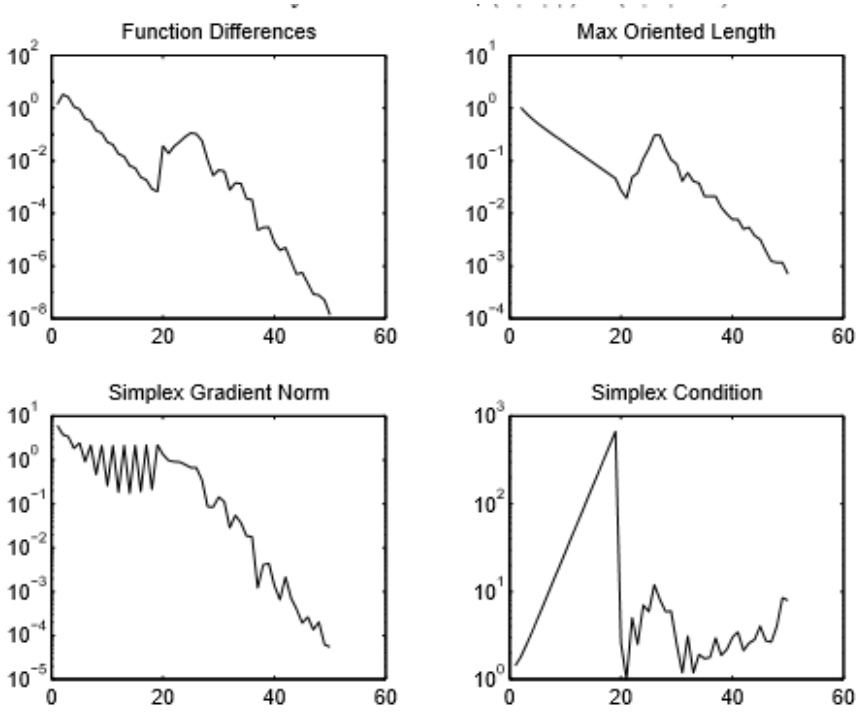


Figura 5.29: Nelder Mead modificato, caso 3

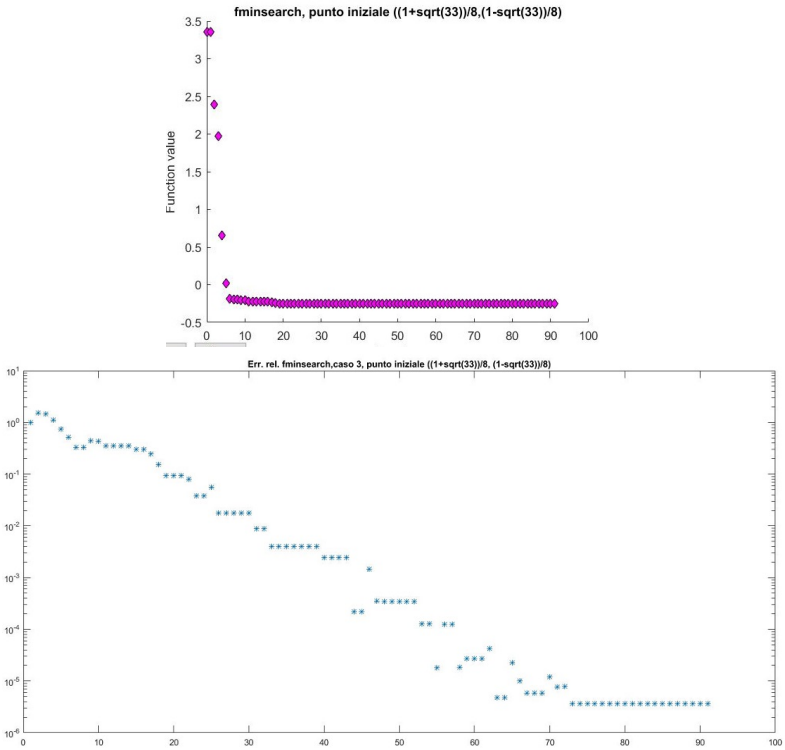


Figura 5.30: Valori della funzione ed errore relativo fminsearch, funzione di MC Kinnon,caso 3, primo esempio

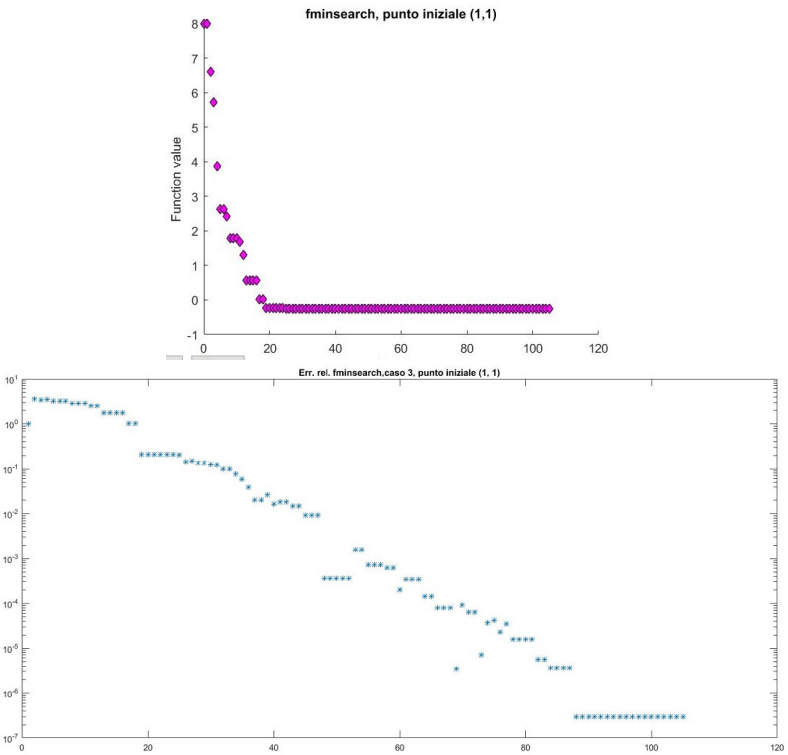
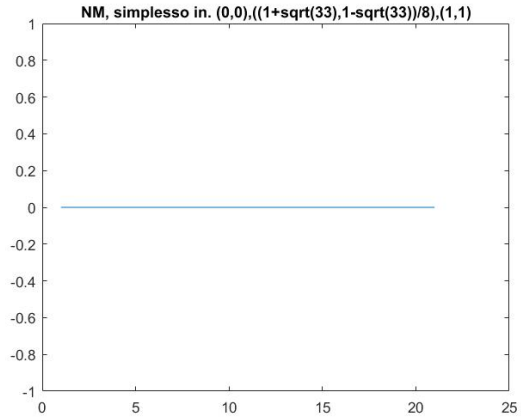


Figura 5.31: Valori della funzione ed errore relativo fminsearch, funzione di MC Kinnon,caso 3, secondo esempio



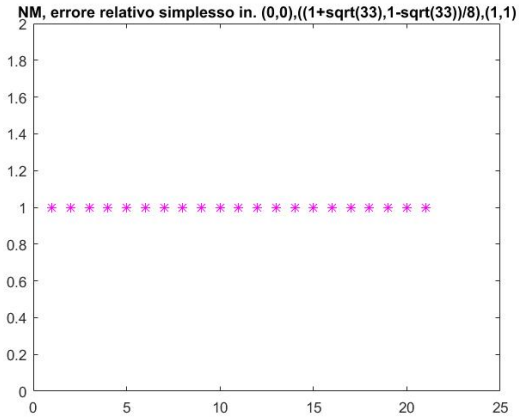
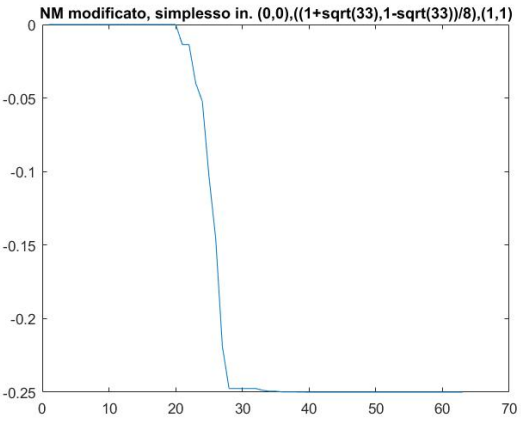


Figura 5.32: Valori della funzione ed errore relativo Nelder Mead, funzione di MC Kinnon, caso3, primo esempio



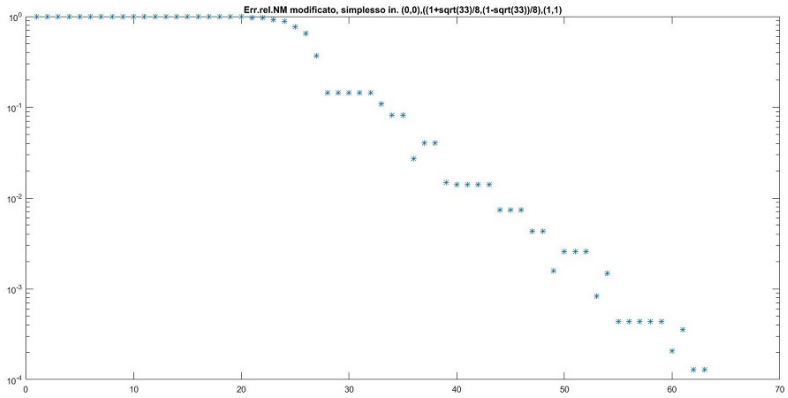


Figura 5.33: Valori della funzione ed errore relativo Nelder Mead modificato, funzione di MC Kinnon, caso3, primo esempio

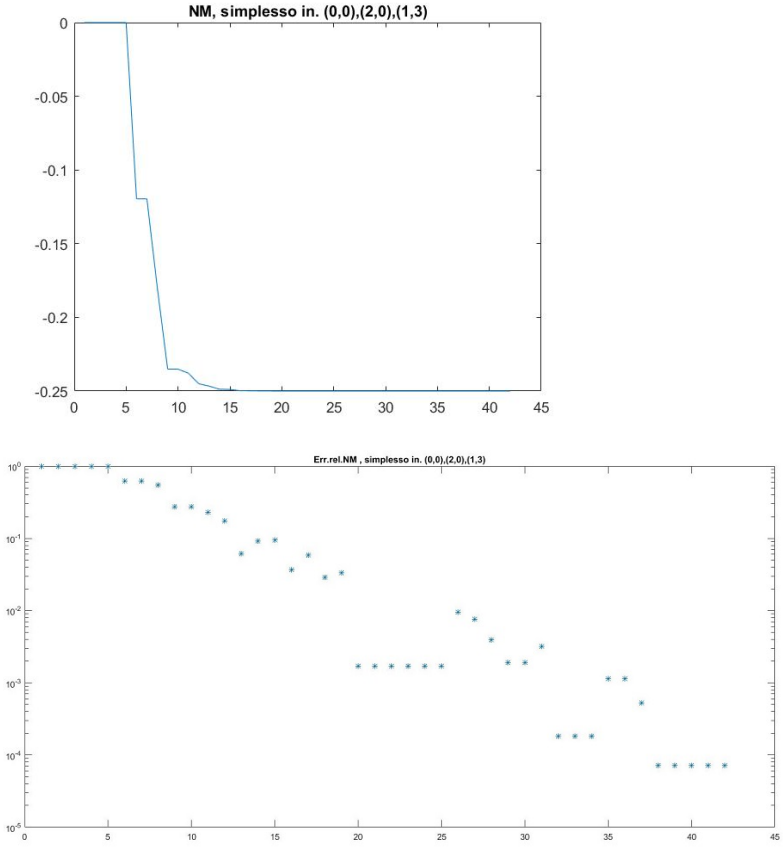


Figura 5.34: Valori della funzione ed errore relativo Nelder Mead, funzione di MC Kinnon, caso3, secondo esempio

Metodo	x_0	Err. rel. x_k	Err. rel. f_k	K	T
fminsearch	$(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$	3.6154e-06	0	91	0.003998
	(4,5)	2.1676e-06	0	106	0.004234
	(1,1)	3.0223e-07	0	105	0.004044
NM	(0,0) $(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$ (1,1)	1	1	21	0.007189
	(0,0) (2,0) (1,3)	7.1620e-05	1.7321e-10	42	0.003047
NM modificato	(0,0) $(\frac{1+\sqrt{33}}{8}, \frac{1-\sqrt{33}}{8})$ (1,1)	1.2793e-04	2.4207e-11	63	0.005055
	(0,0) (2,0) (1,3)	7.1620e-05	1.7321e-10	42	0.003047

Tabella 5.7: Controesempio di MC Kinnon, caso3

5.6 Conclusioni

Complessivamente, alla luce di quanto osservato sperimentalmente ed in parte riportato in questo capitolo, le conclusioni che possono essere tratte dal confronto di questi due metodi, la funzione *fminsearch* e il Nelder Mead, sono principalmente le seguenti:

- l'errore relativo è sicuramente migliore nel primo metodo, sia per quanto riguarda il punto trovato che il valore assunto dalla funzione in tale punto. Tuttavia un errore dell'ordine di 10^{-6}, come quello riportato in media dall'algoritmo Nelder Mead, è più che accettabile;

- il numero di iterazioni effettuate è, nella maggior parte delle volte, inferiore nel caso di Nelder Mead, e molto spesso tale numero è addirittura minore di 100;

- i tempi di esecuzione sono tutto sommato confrontabili.

Inoltre, la versione modificata dell'algoritmo Nelder Mead riesce, come si è visto, a sopperire al problema della stagnazione garantendo anche in questi casi dei buoni risultati di convergenza. Malgrado la mancanza di una robusta teoria sulla convergenza, tale algoritmo risulta quindi comunque abbastanza efficace nel fornire un'adeguata approssimazione del punto di minimo di una funzione.

Capitolo 6

Un'applicazione pratica

Come già evidenziato in precedenza, l'algoritmo Nelder Mead è particolarmente popolare in diversi contesti applicativi, come la chimica, l'ingegneria chimica e la medicina. In questo capitolo verrà dato quindi un esempio concreto di come questo metodo possa effettivamente essere applicato nella realtà, e nello specifico nello studio clinico RSA, tecnica per la valutazione di micromovimenti delle protesi ortopediche. Prima di mostrare i risultati ottenuti, è necessario però fornire una descrizione dettagliata del contesto in cui tale studio va ad operare.

6.1 Introduzione del problema

La sostituzione dell'innesto artificiale è un trattamento comune per articolazioni che sono state affette da traumi, artrosi o artrite reumatoide. In tutto il mondo vengono eseguite ogni anno circa un milione di sostituzioni totali dell'anca e 500.000 protesi totali del ginocchio [20]. La durata massima di una protesi sarà circa dai 15 ai 20 anni. Purtroppo, alcune protesi devono essere riviste prima della durata massima prevista. Questo può essere necessario quando la protesi è usurata, o quando è allentata rispetto all'osso circostante. In generale, l'allentamento inizia con un micromovimento progressivo, nell'intervallo di $0,2 - 1\ mm$, della protesi rispetto all'osso circostante. Una volta avviato, è un processo continuo che rovinerà l'osso, e come risultato, la protesi inizierà a migrare su lunghe distanze. Attualmente, questo processo di allentamento protesico e distruzione ossea può essere fermato solo mediante revisione della protesi.
Un'operazione di revisione è molto più impegnativa per il paziente che l'impianto di una protesi primaria, ed i risultati sono inferiori rispetto a quelli delle protesi primarie: molti pazienti affetti da dolore, possibilità di movimento ridotta, e un tasso di allentamento della protesi revisionata molto più alto rispetto a quelle primarie. Per-

tanto, è della massima importanza sviluppare protesi che hanno una lunga durata. Dal momento che l'allentamento di una protesi inizia con un micromovimento, la conoscenza dei micromovimenti è importante, in quanto può predire il futuro allentamento [4]. Studiando un micromovimento, si può ottenere una visione del processo di allentamento della protesi, che può essere utilizzata per migliorarla. Nella pratica clinica, l'allentamento della protesi è valutato indirettamente attraverso radiografie successive misurando linee radiotrasparenti-scure intorno alla protesi e valutando le differenze di posizione della protesi rispetto all'osso. Le linee radiotrasparenti indicano la presenza di uno strato fibroso attorno alla protesi che è sempre presente quando una protesi è allentata. Nella Figura 6.1 sono indicate alcune misure di base nelle radiografie convenzionali per la valutazione della posizione della protesi ed il suo orientamento. Queste misurazioni non sono molto precise: la radiotrasparenza può verificarsi in aree che sono proiettate oltre il metallo dell'impianto e quindi non può essere osservata; di conseguenza, la quantità di radiotrasparenza potrebbe essere sottostimata. La migrazione della protesi viene valutata misurando variazioni nel tempo della posizione relativa di determinati punti di riferimento protesici e punti di riferimento ossei. Tuttavia questi ultimi non sono sufficientemente distintivi e sono quindi difficili da misurare in modo riproducibile. Per queste ragioni, le misurazioni su radiografie piane non sono precise. Nell'artroplastica totale dell'anca, per esempio, le misurazioni di migrazione possono avere una precisione tra 5 e 12 mm (intervallo di confidenza del 95%), a seconda della scelta dei punti di riferimento ossei. Sono stati fatti molti tentativi per aumentare la precisione delle misurazioni nelle radiografie; si sono ottenuti dei miglioramenti dalla standardizzazione della posizione del paziente, dall'uso di punti di riferimento aggiuntivi, e dall'uso di un software che esegue le misure in modo riproducibile ed oggettivo [3]. Una tecnica di misurazione che combina questi tre miglioramenti è l'Einzel Bild Roentgen Analyse che ha una precisione di 1 mm [8]. Tuttavia, per misurare un micromovimento sub-millimetrico degli impianti nel corso del primo semestre dopo l'innesto o per rilevare le differenze tra pattern di migrazione di piccoli gruppi di impianti, questa precisione non sarà sufficiente. Pertanto, nel 1974, Selvik sviluppò una tecnica molto accurata per la valutazione della migrazione tridimensionale della protesi, Roentgen Stereophotogrammetric Analysis (RSA). La precisione riportata dall'RSA varia tra 0,05 e 0,5 mm per la traslazione e tra 0.15◦ e 1.15◦ per le rotazioni (intervallo di confidenza del 95%).

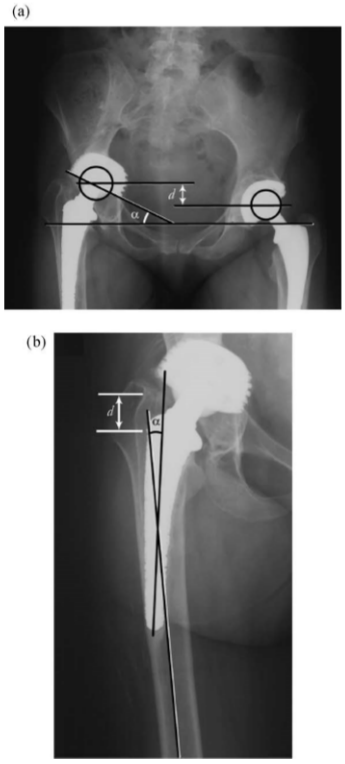


Figura 6.1

6.2 Nozioni di base

RSA è la tecnica Roentgen più accurata per la valutazione di micromovimenti di protesi ortopediche. Tuttavia, per ottenere tale accuratezza, devono essere fatti diversi passaggi che rendono la tecnica piuttosto complicata in un ambiente clinico.

6.2.1 Marcatori ossei e protesi

Per misurare con precisione la migrazione nelle radiografie RSA, i punti di riferimento ossei non sono sufficientemente distintivi. Per ottenere punti di misurazione ben definiti, vengono inserite nell'osso gocce di tantalio con uno speciale strumento di inserimento. A causa delle loro piccole dimensioni e della forma sferica, la loro proiezione non sarà influenzata dai cambiamenti nella posizione del paziente o in quella del fuoco Roentgen. Pertanto, la posizione di questi marcatori può essere misurata con grande precisione. Poiché la maggior parte delle protesi non hanno punti di riferimento che possono essere misurati in modo riproducibile, devono essere marcate con almeno tre marcatori non collineari (Figura 6.2).

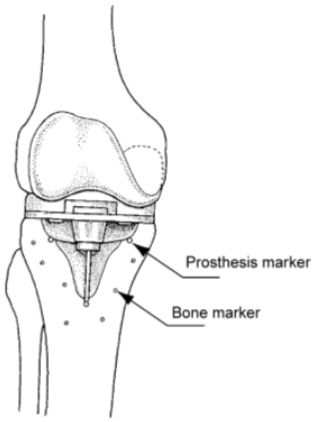


Figura 6.2

6.2.2 Impostazione Roentgen

Nella RSA, due tubi Roentgen sincronizzati vengono utilizzati per ottenere due proiezioni di una zona di interesse di un paziente. Utilizzando le informazioni in queste due proiezioni, è possibile ricostruire la posizione tridimensionale dei marcatori in quella zona. I due tubi Roentgen sono posizionati circa 1,60 m sopra la cassetta Roentgen con un angolo di 20 gradi rispetto all'asse verticale. Una scatola di calibrazione viene utilizzata per calibrare l'impostazione Roentgen. La scatola di calibrazione che viene usata principalmente ha due piani che contengono marcatori tantalio la cui posizione tridimensionale è nota con precisione. I marcatori nel piano vicino alla pellicola radiografica sono detti marcatori fiduciali, quelli nel piano distanti dalla pellicola radiografica marcatori di controllo. I primi definiscono il sistema tridimensionale di coordinate fiduciali mentre i marcatori di controllo sono utilizzati per valutare la posizione dei fuochi Roentgen.
Dopo avere fatto la radiografia RSA, le coordinate dei marcatori ossei e della protesi devono essere misurate con precisione. Nella ricerca RSA convenzionale, questo viene fatto usando una tabella di misura azionata manualmente con una precisione di 0,02 mm. La digitalizzazione manuale delle coordinate dei marcatori è un compito piuttosto noioso che potrebbe richiedere fino a 45 minuti per film. Pertanto, è stato sviluppato un software che consente di automatizzare questo compito, e riduce drasticamente i tempi di analisi.

6.2.3 Calibrazione

Per poter calcolare le posizioni tridimensionali dei marcatori e i punti di riferimento sull'impianto, le coordinate misurate devono essere trasformate al piano inferiore della scatola di calibrazione (piano fiduciale). Questo viene fatto tramite

queste due equazioni:

$$x_{fid,i} = \frac{l_1 x_{rad,i} + l_2 y_{rad,i} + l_3}{l_7 x_{rad,i} + l_8 y_{rad,i} + 1} \quad i = 1, ..., n \tag{6.1}$$

$$y_{fid,i} = \frac{l_4 x_{rad,i} + l_5 y_{rad,i} + l_6}{l_7 x_{rad,i} + l_8 y_{rad,i} + 1} \quad i = 1, ..., n \tag{6.2}$$

dove $(x_{fid,i}, y_{fid,i})$ sono le coordinate fiduciali bidimensionali di un punto, $(x_{rad,i}, y_{rad,i})$ sono le coordinate bidimensionali radiografiche di un punto, ed n è il numero di punti. I parametri l vengono valutati utilizzando le posizioni conosciute dei marcatori fiduciali sulla scatola di calibrazione e le loro proiezioni misurate sulla radiografia. Per calcolare i parametri l sono necessari almeno quattro marcatori fiduciali non lineari e le loro proiezioni. I parametri vengono dapprima stimati in modo lineare. Quindi, le equazioni (6.1) e (6.2) vengono riscritte in modo che siano lineari per i parametri l:

$$x_{fid} = l_1 x_{rad} + l_2 y_{rad} + l_3 - l_7 x_{rad} x_{fid} - l_8 y_{rad} x_{fid}; \tag{6.3}$$

$$y_{fid} = l_4 x_{rad} + l_5 y_{rad} + l_6 - l_7 x_{rad} y_{fid} - l_8 y_{rad} y_{fid}. \tag{6.4}$$

La soluzione di questo problema può essere trovata eseguendo una decomposizione QR. Questa stima lineare viene usato come punto di partenza per un algoritmo Gauss-Newton non lineare. La funzione costo che viene utilizzata è:

$$J = \underline{e}^t \underline{e} \tag{6.5}$$

dove, $\underline{e}$ è un vettore di errore e sia questo che la funzione costo J sono funzioni dei parametri l. Al fine di determinare la direzione di ricerca per l'ottimizzazione, deve essere determinato lo jacobiano:

$$Jac = \frac{\partial J}{\partial \underline{e}} \frac{\partial \underline{e}}{\partial l} \tag{6.6}$$

L'espressione per l'ottimizzazione Gauss-Newton è:

$$\underline{l}_{new} = \underline{l}_{old} - (Jac^t Jac)^{-1} (Jac^t \underline{e}), \tag{6.7}$$

dove l_{new} contiene gli l parametri che sono il risultato della fase di ottimizzazione corrente ed l_{old} è il risultato del passo di ottimizzazione precedente. Dopo che sono stati valutati i parametri l, i punti che sono stati misurati nella radiografia vengono trasformati nel sistema di coordinate fiduciali mediante le equazioni (6.1) e (6.2). La qualità di questa trasformazione è espressa come la distanza tra i marcatori fiduciali

e loro proiezioni trasformate.
Il passo successivo della calibrazione è la valutazione della posizione di entrambi i fuochi Roentgen. Per questa valutazione vengono utilizzati i marcatori nel piano superiore della scatola di calibrazione, cioè quelli di controllo. Attraverso questi marcatori ($\underline{c}_i$) e le loro proiezioni trasformate ($\underline{c}'_i$), vengono determinate le linee di proiezione:

$$\underline{r}_i(\alpha_i) = \underline{c}_i + \alpha_i(\underline{c}_i - \underline{c}'_i) \quad -\infty < \alpha_i < \alpha < \infty \quad i = 1, ...n \tag{6.8}$$

dove n indica il numero di marcatori che vengono utilizzati. Nella situazione ideale, queste linee si intersecheranno in un punto. Tuttavia, si verificano errori di misurazione e deve essere determinata la posizione del fuoco f risolvendo il problema di minimi quadrati:

$$min_{\underline{f},\alpha_i} = \begin{bmatrix} (\underline{c}'_1 - \underline{c}_1) & 0 & \cdots & 0 & I_3 \\ 0 & (\underline{c}'_2 - \underline{c}_2) & & \vdots & I_3 \\ \vdots & & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & (\underline{c}'_n - \underline{c}_n) & I_3 \end{bmatrix} \tag{6.9}$$

che è risolto tramite una decomposizione QR.
La determinazione della posizione tridimensionale dei marcatori in tantalio è simile alla determinazione della posizione del fuoco. Le proiezioni del marcatore nelle due immagini sono denotate con $\underline{t}_1$ e $\underline{t}_2$. La posizione dei due fuochi Roentgen è indicata con $\underline{f}_1$ ed $\underline{f}_2$. Le equazioni delle linee di proiezione che collegano le proiezioni trasformate ed i fuochi corrispondenti sono:

$$\begin{aligned} l_1(\alpha) &= \underline{f}_1 + \alpha(\underline{f}_1 - \underline{t}_1) \\ l_2(\beta) &= \underline{f}_1 + \beta(\underline{f}_2 - \underline{t}_2) \end{aligned} \tag{6.10}$$

La posizione tridimensionale del marcatore è la posizione in cui queste due rette si intersecano. Tuttavia, a causa di errori di misura, le rette non si intersecheranno ma si incroceranno a breve distanza. La posizione tridimensionale del marcatore p si presume essere al centro della retta più breve che collega le due linee di proiezione. La lunghezza di questa retta più breve è detta errore lineare di attraversamento. La posizione tridimensionale di p è la soluzione del problema di minimi quadrati:

$$min_{p,\alpha,\beta} = \| \begin{bmatrix} (\underline{t}_1 - \underline{f}_1) & 0 & I_3 \\ 0 & (\underline{t}_2 - \underline{f}_2) & I_3 \end{bmatrix} \| \tag{6.11}$$

che è risolto tramite una decomposizione QR.

6.2.4 Moto dei corpi rigidi

Dopo che è stata valutata la posizione tridimensionale dei marcatori, può essere calcolato il moto relativo della protesi rispetto all'osso. I marcatori ossei funzionano come corpo di riferimento rigido relativamente al quale viene calcolato il movimento del secondo corpo rigido, la protesi. Dal momento che i pazienti non possono essere posizionati esattamente nella stessa posizione, la posizione e l'orientamento di questo corpo rigido di riferimento cambia tra i seguiti. Per ottenere la stessa posizione ed orientamento dell'osso, i marcatori ossei nella prima radiografia ed in quelle successive sono mappati uno sull'altro; successivamente, può essere calcolato il moto relativo della protesi rispetto all'osso. I risultati dei calcoli del movimento sono una matrice di rotazione ed un vettore di traslazione.
Supponiamo di avere n punti in un corpo rigido, siano $\underline{a}_1, ..., \underline{a}_n$ le posizioni di questi punti sulla prima istanza e $\underline{b}_1, ..., \underline{b}_n$ le posizioni sulla seconda. Per valutare una matrice di rotazione M ed un vettore di traslazione d, deve essere risolta la seguente equazione:

$$min_{M,d} = \Sigma_{i=1}^{n} \|M\underline{a}_i + \underline{d} - \underline{b}_i\|^2 \tag{6.12}$$

in modo che M sia una matrice ortogonale.Questo problema può essere risolto in molti modi ma quello più elegante è stato descritto da Soderkvist (1990), e si basa sulla decomposizione in valori singolari. M è la matrice di rotazione ortogonale e d è il vettore di traslazione. Un'espressione per $\underline{d}$ è, secondo Soderkvist:

$$d = \frac{1}{n}\Sigma_{i=1}^{n}(\underline{b}_i - M\underline{a}_i) = \overline{b} - M\overline{a} \tag{6.13}$$

Quando questa espressione è sostituita nell'equazione (6.12), l'unica incognita rimane M:

$$min_M = \Sigma_{i=1}^{n} \|M(\underline{a}_i - \overline{a}) - (\underline{b}_i - \overline{b})\|^2. \tag{6.14}$$

Quando definiamo $A = [\underline{a}_1 - \overline{a}, ..., \underline{a}_n - \overline{a}]$ e $B = [\underline{b}_1 - \overline{b}, ..., \underline{b}_n - \overline{b}]$, il problema può essere scritto come:

$$min_M \|MA - B\|, \tag{6.15}$$

in modo che M è una matrice ortogonale. La soluzione della matrice di rotazione è:

$$M = UV^t, \tag{6.16}$$

in cui

$$BA^t = U\Sigma V^t \tag{6.17}$$

è la decomposizione in valori singolari. La soluzione di $\underline{d}$ è trovata quando M è sostituita nell'equazione (6.13).
Negli studi clinici RSA, solitamente si è interessati al moto di una protesi rispetto all'osso circostante. Questo significa che ci interessa il moto relativo tra due corpi rigidi. Quando questi corpi rigidi sono indicati con A e B ed il movimento relativo viene calcolato tra il tempo t_0 e t_1, il moto relativo può essere calcolato come

$$A_{t_1} \approx M_A A_{t_0} + \underline{d}_A \underline{u}^t \quad e \quad B_{t_1} \approx M_B B_{t_0} + \underline{d}_B \underline{u}^t \tag{6.18}$$

dove $\underline{u}^t = [1, ..., 1]$. Il moto relativo può essere espresso come

$$M_{rel} = M_A^t M_B \tag{6.19}$$

In seguito l'origine $\underline{o}$ viene posizionata nel centro geometrico di A_{t_0}:

$$\underline{o} = \frac{1}{n_A} \Sigma_{i=1}^{n_A} A_{t_0,i}, \tag{6.20}$$

la traslazione relativa può essere espressa come:

$$\underline{d}_{rel} = M_A^t(\underline{d}_b - \underline{d}_A) + (M_{rel} - I_3)\underline{o} \tag{6.21}$$

dove I_3 è una matrice unità. Il vettore di traslazione $\underline{d}_{rel}$ che è stato calcolato con l'equazione (6.21) potrebbe essere difficile da capire per il clinico. Pertanto, è spesso presentata la differenza nella posizione del centro geometrico di un corpo rigido in due punti nel tempo piuttosto che $\underline{d}_{rel}$.

6.3 Studi clinici

A causa dell'elevata precisione dell'RSA, piccoli gruppi di pazienti sono in genere sufficienti per studiare l'effetto sul fissaggio protesico in seguito a cambiamenti nel design dell'impianto, o a rivestimenti alla protesi, o a nuovi cementi ossei. L'RSA è stata applicata in molti studi che sono stati principalmente condotti in Svezia e in altri paesi scandinavi. Più di 3000 pazienti sono stati inclusi nei diversi studi e più di 150 articoli scientifici sono stati pubblicati. Gli argomenti che sono stati studiati dall'RSA sono: la fissazione della protesi, la stabilità articolare e quella cinematica, la stabilità della frattura, la crescita scheletrica, i movimenti vertebrali, e la fusione

spinale.
L'importanza di valutare i nuovi sviluppi in piccoli gruppi di pazienti prima di una introduzione di massa sul mercato può essere illustrato con l'introduzione del cemento Boneloc nel 1991. Il cemento è usato per fissare una protesi nell'osso. Il cemento osseo è un polimero fatto mescolando un monomero liquido e polvere. Durante il processo di polimerizzazione che segue la miscelazione, viene formato il polimero e creato calore: il vantaggio del cemento Boneloc era la temperatura di polimerizzazione inferiore, 17° anziché 80° come nei cementi tradizionali. Questa temperatura inferiore era stata prevista per ridurre la morte cellulare locale e per ottenere un migliore collegamento osso-cemento, migliorando così il fissaggio della protesi. Tuttavia, nella pratica clinica, è stato osservato un deterioramento del fissaggio protesico: diverse cliniche hanno riportato un allentamento della protesi dopo aver usato il cemento Boneloc. Di conseguenza, sono stati avviati due studi clinici RSA: uno studio totale del ginocchio con 19 pazienti [13] e uno studio totale dell'anca con 11 pazienti [15]. Questi studi RSA sostennero le osservazioni cliniche: le protesi fissate con Boneloc migravano significativamente di più rispetto a quelle fissate con il cemento convenzionale. Pertanto, il cemento Boneloc porta ad un aumento del rischio di revisione causato dalla mobilizzazione asettica. Purtroppo, a quel punto, Boneloc era già stato utilizzato in più di 1000 casi solo in Norvegia. Dopo un periodo di 4 anni e mezzo, il tasso di revisione delle protesi era 14 volte superiore per protesi fissate con cemento convenzionale. Questo sarebbe potuto essere impedito da un test clinico di pre-commercializzazione del fissaggio con un adeguato esame RSA.

6.4 L'effetto dell'idrossiapatite sul fissaggio della protesi del ginocchio

Ci sono due approcci per il fissaggio della protesi nell'osso. Nel primo approccio, il cemento viene utilizzato per formare un forte intreccio tra l'osso e la protesi. Nel secondo, la protesi viene introdotta in una cavità preformata nell'osso che corrisponde esattamente alla superficie della protesi; la ricrescita dell'osso nella superficie protesica fornirà il fissaggio. La crescita ossea può essere stimolata con speciali rivestimenti superficiali che vengono spruzzati sulla superficie protesica. Uno di questi rivestimenti è l'idrossiapatite. Alla Leiden University Medical Center, è stato eseguito un potenziale studio randomizzato per valutare tre diversi mezzi di fissaggio della tibia; lo scopo era quello di studiare l'effetto dell'aggiunta dell'idrossiapatite sul fissaggio delle protesi del ginocchio non cementate. Sono state studiate undici protesi

fissate con il cemento, 10 con rivestimento in idrossiapatite fissate senza cemento, e 10 protesi non rivestite fissate senza cemento. L'RSA è stata utilizzata per valutare micromovimenti dei componenti durante un periodo di 2 anni. Con questo piccolo insieme di pazienti e nel breve periodo dei 2 anni, i componenti tibiali cementati e quelli con rivestimento in idrossiapatite fissati senza cemento sono risultati avere subito molto meno micromovimenti lungo i tre assi ortogonali rispetto ai componenti tibiali non rivestiti fissati senza cemento: alla valutazione del secondo anno, il cedimento dei componenti non rivestiti era $-0.73 \pm 0.924\ mm$, quello dei componenti cementati $-0.05 \pm 0.109\ mm$, e quello dei componenti rivestiti con idrossiapatite $-0.06 \pm 0169\ mm$. Così, dopo 2 anni, il cedimento dei componenti non rivestiti era significativamente maggiore di quello degli altri due gruppi. A causa delle sue piccole dimensioni, questa differenza non sarebbe potuta essere stata rilevata con una misurazione Roentgen convenzionale. In conclusione, i micromovimenti dei componenti tibiali rivestiti con idrossiapatite fissati senza cemento era simile a quello dei componenti tibiali fissati con cemento. Pertanto, l'idrossiapatite, un mediatore biologico, può essere necessario per l'adeguato fissaggio dei componenti tibiali quando il cemento non viene utilizzato. L'azienda ortopedica che vende la protesi ha preso l'esito dello studio molto sul serio; la promozione della protesi non cementata nè rivestita terminò.

6.5 Recenti sviluppi nell'RSA

6.5.1 Digital RSA

Uno svantaggio della pellicola convenzionale RSA è che richiede molta interazione con l'utente. In ogni misurazione radiografica, i punti devono essere etichettati. Successivamente, le coordinate di tutti questi punti devono essere misurati manualmente utilizzando una tavola di misura altamente accurata. Per ridurre il tempo di analisi totale delle radiografie RSA, è stato sviluppato un pacchetto software che è in grado di eseguire le misurazioni delle coordinate automaticamente nelle immagini digitali RSA. L'RSA-CMS può occuparsi di radiografie digitalizzate convenzionali o radiografie digitali dirette Digital Imaging Communications in Medicine in formato (DICOM): uno standard mondiale per le immagini digitali in medicina. Questo pacchetto software gira su un PC con sistema operativo Windows NT.L'RSA-CMS utilizza algoritmi di elaborazione delle immagini sviluppati specificamente per il rilevamento automatico e l'identificazione dei marcatori RSA, cioè i marcatori di calibrazione, i marcatori ossei, e quelli della protesi nelle radiografie RSA. Le posi-

zioni dei centri dei marcatori sono determinate automaticamente Successivamente, le posizioni dei marcatori sono migliorate con precisione sub-pixel stimando un paraboloide attraverso il profilo a scala di grigi dei marcatori proiettati. Così, non solo viene utilizzata l'informazione dei pixel di contorno del marcatore, ma l'informazione dei pixel nell'area proiettata viene utilizzata per un risultato ottimale. Per mezzo di un algoritmo di raccordo, i marcatori di calibrazione,fiduciali e di controllo, vengono estratti dal gruppo totale dei marcatori ed automaticamente etichettati. Inoltre, il software confronta i marcatori nelle due radiografie, ricostruisce le coordinate spaziali dei marcatori, e infine calcola il micromovimento dell'endoprotesi. Con RSA-CMS, la procedura RSA può essere eseguita in modo completamente automatico. Se necessario, l'utente può interattivamente correggere risultati intermedi su scoperte non corrette come artefatti nella superficie della pellicola, fili e viti. La precisione del sistema RSA digitale è stata confrontata con la precisione di un sistema RSA azionato manualmente dalla Svezia. A questo scopo, sono state usate le radiografie di un arto fantasma e quelle di pazienti fornite dall'Ospedale Lund University, Lund, Svezia. Nell'esperimento fantasma, il sistema che operava manualmente ha prodotto risultati significativamente migliori rispetto al sistema digitale, anche se la differenza massima tra i valori mediani del sistema manuale e di quello digitale era circa di 0,013 millimetri per la traslazione e 0.033 gradi per le rotazioni. Questi risultati leggermente meno accurati sono stati probabilmente causati dal digitalizzatore del film utilizzato. Nelle radiografie dei pazienti, è stato utilizzato uno scanner migliore e il sistema manuale e quello digitale hanno prodotto risultati altrettanto precisi: non è stata trovata nessuna differenza significativa. Ancora una volta, è stato dimostrato che l'RSA digitale fornisce un'elevata precisione. I primi risultati della RSA digitale sono stati pubblicati da Ostgaard nel 1997; in questo sistema semi-automatico, le posizioni dei marcatori devono essere indicate manualmente e il software affina quelle posizioni. Con questo sistema è stato effettuato solo uno studio di validazione , ma non è mai stato utilizzato in ambito clinico. Ad Oxford è stato sviluppato un sistema RSA digitale usato poi in ambiente clinico. Il sistema di Oxford è anche in grado di valutare la posizione e l'orientamento delle protesi dell'anca utilizzando punti di riferimento protesici. Un altro sviluppo è il modulo di misura digitale che è stato creato per il sistema UmRSA disponibile in commercio, mentre per un altro sistema RSA disponibile in commercio, il WinRSA-sistema è stato recentemente sviluppato un modulo di misura digitale.

Nessun risultato di questo sistema è stato ancora pubblicato. Entrambi i sistemi-UmRSA e WinRSA-operano in modo semi-automatico, mentre RSA-CMS ha il vantaggio che è completamente automatico, cioè il software trova i marcatori senza l'in-

tervento dell'utente. Lo sviluppo di tutti questi sistemi digitali RSA dimostra che vi è una chiara necessità di sistemi RSA che sono veloci e più facili da usare rispetto a quelli convenzionali.

6.5.2 RSA basata sul modello

Il fissaggio di marcatori di tantalio alla protesi è un prerequisito per l'RSA, ma può essere difficile e talvolta è addirittura impossibile. Inoltre, la marcatura degli impianti è una procedura costosa e, in alcuni paesi consentita solo dagli organismi di regolamentazione, dopo numerosi test ed una documentazione completa. Pertanto, è stato sviluppato un metodo RSA basato sul modello che utilizza un modello di superficie triangolare della protesi. Viene calcolato un contorno proiettato di questo modello e questo modello di contorno calcolato è accoppiato al contorno rilevato dell'impianto attuale (Fig. 6.3a) nella radiografia RSA (Fig 6.3b). La differenza tra i due contorni viene minimizzata dalla variazione della posizione e dell'orientamento del modello (Fig. 6.3c). Quando viene trovata una differenza minima tra i contorni, sono stati ottenuti una posizione ottimale e orientamento del modello (Fig. 6.3d). Il metodo è stato convalidato mediante un esperimento fantasma. Questo consisteva di un cilindro di plexiglas con 12 1-min marcatori sferici incorporati nella sua superficie. In questo esperimento sono stati usati tre componenti protesiche : la componente femorale e tibiale di una protesi totale del ginocchio Interax e la componente femorale di una protesi totale del ginocchio Profix. Per ogni esperimento, uno dei componenti è stato rigidamente fissato al piano di base di questo cilindro. La posizione calcolata del modello e l'orientamento sono stati confrontati con la posizione e l'orientamento del cilindro. Poiché il moto reale tra la protesi ed il cilindro era zero, qualsiasi cambiamento nella posizione relativa indica un errore nella valutazione dei parametri del micromovimento con il metodo RSA basato sul modello. Per i componenti protesici utilizzati in questo studio, la precisione del metodo basato sul modello è risultata essere inferiore a quella dell'RSA tradizionale. Per i componenti Interax femorali e tibiali, sono state trovate significative tolleranze dimensionali che, probabilmente, sono state causate dal processo di fusione e lucidatura manuale delle superfici dei componenti. Per questi componenti, sono stati trovati errori sistematici per le traslazioni e le rotazioni. La deviazione standard più grande per ogni traslazione era di 0,19 mm e per ogni rotazione 0.52°. Per la componente femorale Profix che non aveva grandi tolleranze dimensionali, la deviazione standard più grande per le traslazioni era di 0,22 millimetri e per le rotazioni 0.22°. Da questo studio pilota, possiamo concludere che la precisione del corrente metodo RSA basato sul modello è sensibile alle tolleranze dimensionali dell'impianto. L'obiettivo futuro è ottenere

un miglioramento di questo metodo RSA basato sul modello in modo che sia insensibile a grandi tolleranze dimensionali e che fornisca una precisione paragonabile alla precisione dell'RSA tradizionale.
In conclusione si è visto che l'RSA è una tecnica di misurazione estremamente precisa, ma piuttosto complicata per la valutazione di micromovimenti della protesi. Dal momento che un micromovimento progressivo è un indicatore importante di un fissaggio inadeguato (cioè allentamento) della protesi, ed estesi micromovimenti a breve termine potrebbero indicare una futura operazione di revisione della protesi, gli studi clinici RSA sono importanti. Ciò è illustrato dai risultati degli studi clinici RSA dopo l'introduzione del cemento Boneloc, ed in quello in cui si è studiato l'effetto dell'idrossiapatite sul fissaggio della protesi al ginocchio. Con le tecniche Roentgen convenzionali, i risultati negativi trovati in questi studi non potrebbero essere stati ottenuti in tale breve periodo di valutazione, e con un talmente piccolo numero di pazienti coinvolti. Con l'esecuzione di studi clinici RSA, un sacco di inutili sofferenze potrebbero essere prevenute. Presso l'Università di Leiden Medical Centre (LUMC), la semplificazione della tecnica RSA per l'automazione delle misure e l'introduzione dell'RSA basata sul modello sono due importanti temi di ricerca. I test di validazione hanno dimostrato che il nuovo sistema automatico RSA ha una elevata precisione, e, successivamente, viene ora utilizzato in diversi studi clinici RSA. Tuttavia, quando si confrontano l'RSA basata sul modello con quella convenzionale, questa nuova tecnica è stata meno precisa quando sono stati utilizzati impianti con grandi tolleranze dimensionali. Il metodo RSA basato su modello deve essere modificato in modo da ottenere una elevata precisione su questi impianti. Queste modifiche sono attualmente oggetto di ricerca.

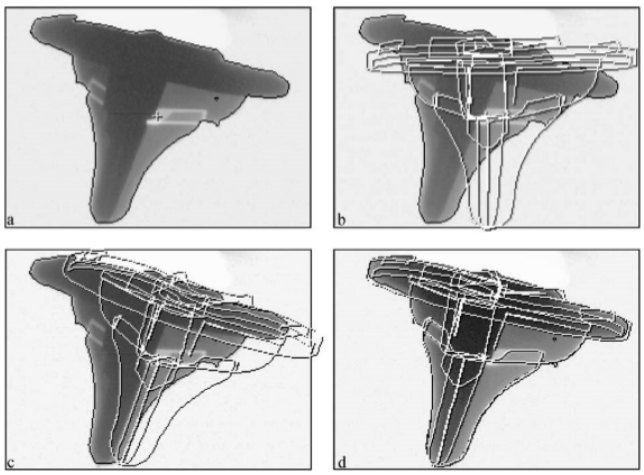


Figura 6.3

6.6 Nelder Mead e l'RSA

Dopo questa panoramica generale sullo scopo dello studio RSA e sui suoi sviluppi, vediamo come l'algoritmo Nelder Mead può essere applicato in tale ricerca. Nella sezione precedente si è visto che, calcolata la proiezione del contorno del modello della protesi e dato il contorno nell'impianto attuale della radiografia RSA, va minimizzata la differenza tra questi due contorni.
Di conseguenza, a partire da dati reali forniti dall'Istituto Ortopedico Rizzoli, che costituisce uno dei centri di sviluppo di un progetto europeo attualmente in fase di svolgimento, si è provato ad usare l'algoritmo Nelder Mead proprio per effettuare tale minimizzazione e valutarne l'eventuale uso per contribuire ad un esito proficuo dello studio. La traduzione matematica del problema è la seguente: la funzione obiettivo in esame è una funzione non differenziabile e non nota in forma analitica; le incognite del problema sono 6, di cui 3 corrispondono alle coordinate delle traslazioni e 3 a quelle delle rotazioni. Partendo quindi da una radiografia RSA come quella in figura 6.4, l'algoritmo è stato applicato con un simplesso iniziale centrato in un determinato punto di riferimento X_0. Nelle figure 6.5 e 6.6 si vede il comportamento dell'algoritmo durante alcune delle iterazioni effettuate nei due diversi casi analizzati, il femore e la tibia. E' evidente come in entrambi i casi all'inizio il simplesso sia lontano dal contorno di riferimento mentre dopo si vada via via sovrapponendo.
Malgrado i risultati numerici ottenuti non siano stati particolarmete soddisfacenti, restano aperte le possibilità di migliorie. Inoltre tale studio costituisce comunque un'interessante e concreta applicazione del metodo e lascia intravedere le sue numerose opportunità di impiego.

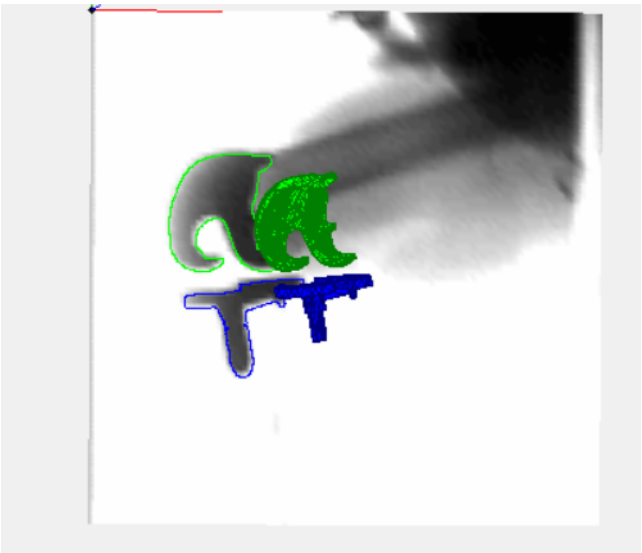


Figura 6.4: Radiografia RSA

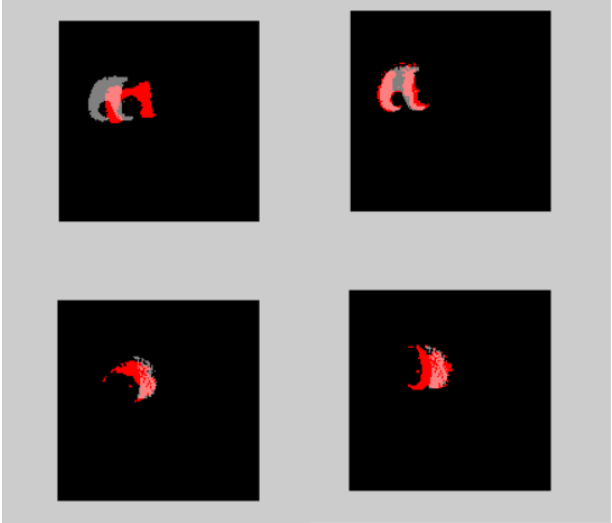


Figura 6.5: Analisi sul femore

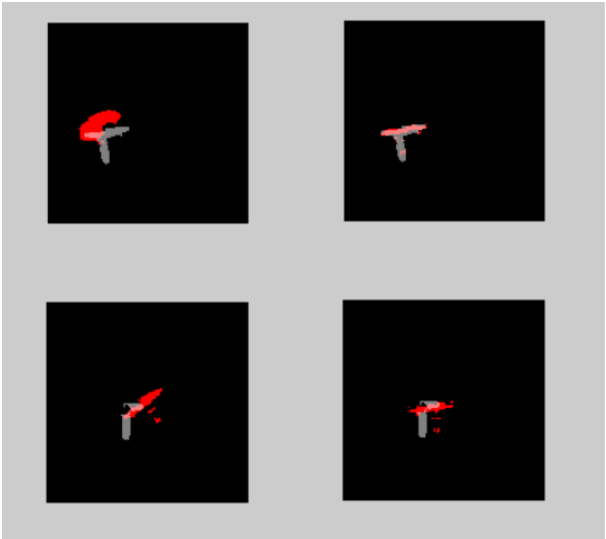


Figura 6.6: Analisi sulla tibia

Conclusioni

L'obiettivo principale di questa tesi è stato quello di mettere in evidenza l'utilità, anche in diversi contesti più generali, dell'algoritmo di ottimizzazione Nelder Mead per funzioni non differenziabili. In quest'ottica la tesi è stata sviluppata partendo da una base teorica, volta a comprendere appieno il meccanismo di funzionamento di tale algoritmo e le sue proprietà di convergenza.
A questa prima fase è seguita poi quella sperimentale, in cui sono stati messi a confronto due algoritmi basati sulla filosofia del simplesso: il metodo Nelder Mead oggetto della tesi e la funzione *fminsearch* di Matlab. Dai risultati numerici ottenuti su alcune funzioni test è emerso come l'algoritmo Nelder Mead ottenga dei risultati più che apprezzabili impiegando un minor numero di valutazioni della funzione e riportando un errore che risulta essere non peggiore rispetto a quello ottenuto dall'altro metodo.
Per ultimo si è visto come l'algoritmo possa essere utilizzato anche in applicazioni concrete, come ad esempio in ambito medico ed in particolare negli studi clinici RSA. Malgrado i risultati ottenuti a questo proposito non siano ancora ottimali, questo lascia intuire comunque le numerose possibilità di miglioramento e soprattutto di impiego di quest'algoritmo anche in contesti applicativi.

Bibliografia

[1] J. E. Dennis and R. B. Scnabel, *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*, no. 16 in Classics in Applied Mathematics, SIAM, Philadelphia, 1996.

[2] J.E. Dennis, D.J. Woods, *Optimization on microcomputers: the Nelder-Mead simplex algorithm.* New Computing Environments: Microcomputers in Large-Scale Computing, A. Wouk, ed., Philadelphia, 1987, SIAM, pp. 116-122.

[3] Hardinge K., Porter M.L., Jones P.R., Hukins D.W., Taylor C.J. *Measurement of hip prostheses using image analysis. The MAXIMA hip technique* Journal of Bone and Joint Surgery 73 (5), 724?728, 1991.

[4] Kärrholm J., Borssén B., Lowenhielm G., Snorrason F. *Does early micromotion of femoral stem prostheses matter 4?7-year stereoradiographic follow-up of 84 cemented prostheses* Journal of Bone and Joint Surgery 76 (6), 912?917,1994.

[5] P. Gilmore and C-T.Kelley, *An implicit Filtering algorithm for optimization of functions with many local minima*, SIAM J. Optim., 5 (1995), pp. 269?285.

[6] C. T. Kelley, *Detection and Remediation of Stagnation in the Nelder-Mead Algorithm Using a Sufficient Decrease Condition*, Technical report, Department of Mathematics, North Carolina State University, Raleigh, NC, 1997.

[7] C. T. Kelley, *Iterative Methods for Optimization*, no. 18 in Frontiers in Applied Mathematics, SIAM Publications, Philadelphia, PA, 1999

[8] Krismer M., Bauer R., Tschupik J., Mayrhofer P. *EBRA a method to measure migration of acetabular components* Journal of Biomechanics 28 (10), 1225?1236, 1995

[9] J.C. Lagarias, J.A. Reeds, M. H. Wright, e P.E. Wright, *Convergence properties of the NelderMead simplex algorithm in low dimensions*, Tech. Rep. 96-4-07, AT&T Bell Laboratories, April 1996.

[10] J. C. Lagarias, B. Poonen, and M. H. Wright, *Convergence of the restricted Nelder-Mead algorithm in two dimensions*, in preparation, 1998.

[11] K. I. M. McKinnon, *Convergence of the Nelder-Mead simplex method to a non-stationary point*, Tech. Rep., Department of Mathematics and Computer Science, University of Edinburgh, Edinburgh, 1996.

[12] J. A. Nelder and R. Mead, *A simplex method for function minimization*, Comput. J., 7 (1965), pp. 308? 313.

[13] Nilsson K.G., Da ´len T. *Inferior performance of Boneloc bone cement in total knee arthroplasty: a prospective randomized study comparing Boneloc with Palacos using radiostereometry (RSA) in 19 patients* Acta Orthopaedica Scandinavica 69 (5), 479?483,1998

[14] Jorge Nocedal, Stephen J. Wright, *Numerical optimization*, New York : Springer, 2006

[15] Thanner J., Freij Larsson C., Ka¨rrholm J., Malchau, H., Wesslen B. *Evaluation of Boneloc. Chemical and mechanical properties, and randomized clinical study of 30 total hip arthroplasties* Acta Orthopaedica Scandinavica 66 (3), 207?214,1995

[16] Tomick, J.J., Arnold, S.F. and Barton, *Sample Size Selection for Improved Nelder Mead Performance*. In: Proceeding of the 1995 Winter Simulation Conference, December 3?6, pp. 341-345.

[17] V. Torczon, *Multi-directional Search: A Direct Search Algorithm for Parallel Machines*, Ph.D. thesis, Rice University, Houston, TX, 1989

[18] V. Torczon, *On the convergence of the multidimensional direct search*, SIAM J. Optim., 1 (1991), pp. 123-145.

[19] P. Tseng, *Fortified-descent simplicial search method: A general approach*, SIAM J. Optim., to appear.

[20] Valstar ER, Nelissen RGHH, Reiber JHC, Rozing PM, *The use of Roentgen stereophotogrammetry tot study micromotion of orthopaedic implants.* ISPRS Journal of Photogrammetry & remote sensing. 2002;56:376?389.

[21] D. J. Woods, *An Interactive Approach for Solving Multi-Objective Optimization Problems*, PhD thesis, Rice University, 1985