Università degli Studi di Firenze

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Specialistica
in Matematica per le Applicazioni

Anno Accademico 2005-2006

Un nuovo metodo trust-region per sistemi non lineari rettangolari con vincoli semplici

Laureando: Margherita Porcelli

Relatore: Prof. Benedetta Morini

Indice

Introduzione **3**

1 Sistemi non lineari vincolati **6**
1.1 Il problema 6
1.1.1 Sistemi non lineari con soluzioni spurie 7
1.1.2 Sistemi di equazioni e disequazioni non lineari 9
1.1.3 Problemi di complementarità misti 11
1.2 Metodi numerici 11

2 Preliminari **15**
2.1 Il metodo di Newton 15
2.2 Metodi trust-region 17

3 Il metodo TRESNORE **21**
3.1 Descrizione del metodo 21
3.2 Analisi di convergenza 28
3.2.1 Terminazione finita del calcolo di una iterata 29
3.2.2 Proprietà di convergenza globale 32
3.2.3 Proprietà di convergenza locale 34

4 Implementazione del metodo TRESNORE e risultati numerici **43**
4.1 L'algoritmo e la sua implementazione 43
4.1.1 Calcolo del passo di trust-region $p_{tr}(\Delta_k)$ 43
4.1.2 Calcolo del passo di prova $p(\Delta_k)$ 45
4.1.3 Accettazione del passo e aggiornamento della trust-region . . . 46
4.1.4 L' algoritmo implementato 47
4.2 Risultati Numerici 49
4.2.1 Risoluzione di problemi quadrati 51
4.2.2 Sistemi sovradeterminati 51
4.2.3 Sistemi sottodeterminati 52
4.2.4 Sistemi di equazioni e disequazioni non lineari 52
4.3 Considerazioni finali 53

A Funzioni vettoriali e successioni **59**

B Decomposizione ai valori singolari di una matrice reale e pseudoinversa di Moore-Penrose **60**

Bibliografia **62**

Introduzione

Un tema importante ed attuale nell'ambito dell'analisi numerica è lo studio della risoluzione di sistemi di equazioni non lineari con vincoli sulle variabili. Tale interesse è dovuto al fatto che i sistemi vincolati costituiscono il modello matematico per numerosi problemi, ad esempio problemi di equilibrio, discretizzazione di problemi differenziali di tipo ordinario o alle derivate parziali, problemi non lineari di complementarità.

In questa tesi ci occupiamo di *sistemi di equazioni non lineari con vincoli semplici*, i.e. problemi della forma

$$F(x) = 0, \qquad x \in \Omega, \tag{1}$$

dove $F : \mathbb{R}^n \to \mathbb{R}^m$ è una funzione non lineare differenziabile con continuità e la regione di ammissibilità Ω è definita come

$$\Omega = \{x \in \mathbb{R}^n : \ l_i \leq x_i \leq u_i \quad i = 1, \ldots, n\}.$$

L'imposizione del vincolo sulla variabile ha grande rilevanza applicativa perché consente di limitare la ricerca delle radici della funzione F ad una specificata regione Ω. Tale regione può ad esempio rappresentare il dominio di F oppure la regione a cui appartengono le soluzioni del sistema non lineare che hanno significato fisico.

Il sistema non lineare (1) viene detto *quadrato* se $m = n$, *rettangolare* altrimenti. Inoltre il sistema è detto sovradeterminato se $m > n$, sottodeterminato se $m < n$.

La risoluzione di sistemi rettangolari è richiesta in numerosi problemi, tra i quali ricordiamo il calcolo di punti ammissibili per un problema di programmazione non lineare e il calcolo di soluzioni di problemi ai minimi quadrati a residuo nullo.

Nonostante l'importanza applicativa dei sistemi rettangolari vincolati, la loro risoluzione numerica è ancora un problema aperto. Per questo motivo, il nostro lavoro di tesi si occupa principalmente dello studio di sistemi rettangolari.

I metodi iterativi per sistemi non vincolati non sono adatti alla risoluzione del problema (1) poiché possono convergere a soluzioni che non appartengono alla regione Ω pur disponendo di un'accurata approssimazione iniziale di una soluzione ammissibile. L'approccio comunemente usato nel passato per la risoluzione dei sistemi non lineari vincolati non sfruttava la struttura del problema bensì una riformulazione mediante un problema di ottimizzazione ai minimi quadrati [9, 12, 17, 29]. Tuttavia, le ben note differenze fra i metodi per i sistemi non lineari e problemi di minimo hanno indotto a definire procedimenti che sfruttano la struttura del problema (1), [12].

Recentemente sono stati proposti procedimenti specifici per la risoluzione di sistemi vincolati. Tali procedimenti possono essere suddivisi in tre categorie: metodi ai *vincoli attivi* [20, 31], metodi che richiedono ad ogni iterazione la risoluzione di un problema di *programmazione quadratica e convessa* con vincoli semplici [15, 23, 34] e metodi *trust-region* [2, 3, 4, 5, 14, 21, 33]. In particolare, i metodi trust-region hanno proprietà teoriche simili a quelle degli altri procedimenti ma prestazioni numeriche superiori. Eccettuato il metodo proposto in [14], i procedimenti citati sono utilizzabili solo nella risoluzione di sistemi quadrati. Il metodo fornito nel lavoro [14] è applicabile a problemi sottodeterminati.

In questo lavoro di tesi proponiamo la risoluzione di sistemi vincolati mediante un procedimento generale, cioè applicabile al problema (1) indipendentemente dal numero di equazioni m e dal numero di incognite n. Il contributo di questo lavoro consiste nella definizione del procedimento, nell'analisi delle sue proprietà teoriche e nella sua validazione sperimentale.

Il nuovo metodo appartiene alla classe dei metodi trust-region con scalatura affine e può essere applicato a problemi quadrati e rettangolari. Nell'ambito dei problemi quadrati il nuovo metodo ha applicabilità superiore a quella del metodi proposti in [2, 3, 5] in quanto risolve anche sistemi in cui la matrice Jacobiana è singolare in qualche iterata. L'analisi teorica mostra che il metodo ha buone proprietà di convergenza. Infatti il metodo genera una successione di iterate la cui convergenza non dipende in modo critico dall'approssimazione iniziale e asintoticamente la convergenza è veloce se la matrice Jacobiana calcolata nella soluzione ha rango massimo per colonne. Le proprietà del metodo sono state verificate sperimentalmente risolvendo un gran numero di problemi noti in letteratura. I risultati ottenuti evidenziano l'efficienza del procedimento.

I Capitoli 1 e 2 sono introduttivi e contengono una descrizione dettagliata del problema considerato e dei metodi iterativi comunemente usati nella risoluzione di sistemi non lineari. I Capitoli 3 e 4 contengono i risultati originali della tesi.

Nel Capitolo 1 mostriamo problemi reali che possono essere espressi mediante la formulazione (1) e presentiamo metodi numerici esistenti per la risoluzione di sistemi non lineari vincolati. Nel Capitolo 2 presentiamo le principali caratterisiche del metodo di Newton e dei metodi trust-region per la risoluzione di sistemi non lineari non vincolati e quadrati. Nel Capitolo 3 presentiamo il nuovo metodo e ne studiamo le proprietà di convergenza. Infine nel Capitolo 4, forniamo una descrizione dettagliata dell'implementazione del metodo e i risultati ottenuti nella risoluzione di un elevato numero di problemi reali.

Ringraziamenti

Desidero ringraziare con affetto la Prof.ssa Benedetta Morini per il supporto e la stima che mi hanno permesso di lavorare con entusiasmo a questa tesi.

Ringrazio inoltre la Dott.ssa Stefania Bellavia per la grande disponibilità e per i preziosi consigli.

Un ringraziamento particolare va alla Prof.ssa Maria Macconi per la fiducia e per avermi permesso di usufruire di tutte le attrezzature del Dipartimento di Energetica "S. Stecco" dell'Università degli Studi di Firenze.

Desidero infine ringraziare tutti coloro che, con il loro appoggio, mi hanno permesso di crescere durante l'iter universitario e di portare a termine questa tesi serenamente.

Capitolo 1

Sistemi non lineari vincolati

1.1 Il problema

Consideriamo il problema

$$F(x) = 0, \qquad x \in \Omega, \tag{1.1}$$

dove $F : D \to \mathbb{R}^m$ è una funzione non lineare differenziabile con continuità e $D \subseteq \mathbb{R}^n$ è un insieme aperto e convesso contenente la regione di ammissibilità Ω. Nel seguito supporremo che Ω sia una regione della forma

$$\Omega = \{x \in \mathbb{R}^n : \ l \leq x \leq u\},$$

dove le disuguaglianze sono intese componente a componente e componenti di x possono essere non vincolate, ovvero $l \in (\mathbb{R} \cup -\infty)^n$, $u \in (\mathbb{R} \cup +\infty)^n$. Questo tipo di vincoli è definito in letteratura di tipo "box" o "bound" poiché Ω è una "scatola" di dimensione n, oppure di tipo "semplice", [29].

Il problema (1.1) consiste quindi nel trovare le radici del sistema non lineare che soddisfano i vincoli definiti in Ω. Se $m = n$ il sistema viene detto quadrato, se $m > n$ viene detto sovradeterminato, sottodeterminato altrimenti.

I sistemi non lineari con vincoli semplici costituiscono i modelli matematici per problemi di reale interesse. Infatti i sistemi non lineari descrivono molti problemi che nascono in contesti applicativi (ad esempio problemi di equilibrio, discretizzazione di problemi differenziali di tipo ordinario o alle derivate parziali, problemi non lineari di complementarità) e la formulazione (1.1) consente di limitare la ricerca delle radici di F a quelle che appartengono ad una specificata regione Ω. In particolare, Ω può rappresentare il dominio di F, oppure la regione di $\mathbb{R}^n$ a cui appartengono le soluzioni di F che hanno significato fisico. In quest'ultimo caso, imponendo il vincolo, si evita il calcolo di soluzioni dette spurie, ovvero soluzioni prive di significato fisico.

Presentiamo di seguito vari problemi che sono descritti mediante la formulazione (1.1).

1.1.1 Sistemi non lineari con soluzioni spurie

Frequentemente il sistema non lineare che descrive un problema reale ammette varie soluzioni e non tutte hanno significato per il problema originario. È quindi necessario imporre vincoli sulle variabili in modo da individuare solo le soluzioni di reale interesse. Mostriamo due esempi significativi di problemi di questo tipo che nascono nell'ambito della risoluzione di un problema ai limiti e di un problema di equilibrio chimico.

Esempio 1 ([32]) Vogliamo determinare una soluzione $w \in C^2((0,1)) \cap C([0,1])$ del seguente problema ai limiti (BVP):

$$w'' = \frac{3}{2}w^2, \quad w(0) = 4, \quad w(1) = 1. \tag{1.2}$$

Questo problema ammette due soluzioni

$$w_1(t) = \frac{4}{(1+t)^2} \quad \text{e} \quad w_2(t) = C_1^2\Big(\frac{1 - cn(C_1 t - C_2, k)}{1 + cn(C_1 t - C_2, k)} - \frac{1}{\sqrt{3}}\Big),$$

dove $cn(\xi, k)$ denota la funzione jacobiana ellittica di modulo k, i.e.

$$cn(\xi, k) = \cos(\phi(\xi, k)), \qquad \xi = \int_0^{\phi} \frac{dt}{\sqrt{1 - k^2 \sin^2(t)}} \ ,$$

e le costanti che compaiono in ω_2 sono date da $k = \frac{1}{2}\sqrt{2+\sqrt{3}}$, $C_1 = 4.30310990$, $C_2 = 2.33464196$. Le soluzioni ω_1 e ω_2 sono rappresentate nella Figura 1.1.

Per risolvere numericamente il BVP, introduciamo un insieme I di n punti

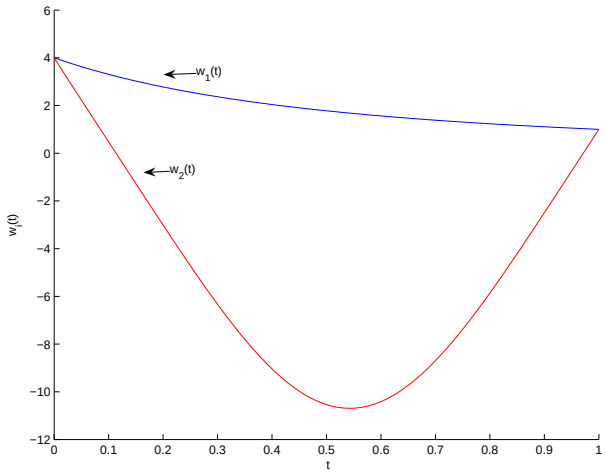

Figura 1.1: Grafico delle soluzioni del BVP (1.2).

equidistanti in $[0,1]$, $I = \{t_k \in [0,1] : t_k = \frac{k-1}{n-1},\ k = 1, \ldots, n\}$. Si osservi che

$t_{k+1} = t_k + h$, $h = \frac{1}{n-1}$ per $k = 1, \ldots, n-1$. Quindi indichiamo con $\{x_i\}$, $i = 1, \ldots, n$, una funzione di rete il cui valore x_i è assunto come approssimazione del valore della soluzione $\omega(t)$ del BVP nel punto $t_i \in I$. Banalmente, da (1.2) abbiamo $x_1 = 4$, $x_n = 1$.

Considerate approssimazioni alle differenze finite centrali per $\omega''(t)$ su I

$$\omega''(t_k) \simeq \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2}, \quad k = 2, \ldots, n-1,$$

e sostituite tali approssimazioni nell'equazione differenziale (1.2), si ottiene un sistema di equazioni non lineari $F(x) = 0$ dove $x = (x_2, \ldots, x_{n-1})^T$, $F = (F_1, \ldots, F_{n-2})^T$ e

$$F_{k-1}(x) \quad = \quad 2x_k - x_{k-1} - x_{k+1} + \frac{3}{2}h^2 x_k^2, \quad k = 2, \ldots, n-1.$$

Poiché il BVP (1.2) ha due soluzioni, il calcolo approssimato di ω_1 è garantito solo risolvendo il sistema vincolato (1.1) nel quale $l = (0, \ldots, 0) \in \mathbb{R}^{n-2}$ e $u = (+\infty, \ldots, +\infty) \in \mathbb{R}^{n-2}$.

Esempio 2 ([25]) La combustione del propano (C_3H_8) nell'aria (O_2 e N_2) produce i 10 prodotti elencati nella Tabella 1.1. In particolare l'equazione stechiometrica ha la forma

$$C_3H_8 + \frac{R}{2}(O_2 + 4N_2) \rightarrow prodotti$$

dove $R \geq 3$ è un parametro che rappresenta il rapporto tra la quantità di aria e di combustibile.

Ci poniamo il problema di determinare il numero n_i di moli del prodotto i-esimo

prodotti	moli	descrizione	cambio variabili per la mole
CO_2	n_1	anidride carbonica	x_1x_2/x_5
H_2O	n_2	acqua	$x_2x_3^2/x_5$
N_2	n_3	azoto	x_4^2/x_5
CO	n_4	monossido di carbonio	x_1/x_5
H_2	n_5	idrogeno	$R_5x_3^2$
H	n_6	idrogeno in forma atomica	R_6x_3/x_5
OH	n_7	radicale idrossile	$R_7x_2x_3/x_5$
O	n_8	ossigeno in forma atomica	R_8x_2/x_5
NO	n_9	monossido di azoto	$R_9x_2x_4/x_5$
O_2	n_{10}	ossigeno	$R_{10}x_2^2/x_5$

Tabella 1.1: Prodotti della combustione del propano e cambio di variabili.

$i = 1, \ldots, 10$, generato dalla combustione di una mole di propano. La IV colonna della Tabella 1.1 mostra che le incognite n_i, $i = 1, \ldots, 10$, possono essere espresse mediante le variabili $x_1, x_2, \ldots, x_5$ e sei costanti R_i, $i = 5, \ldots, 10$ dipendenti dalla pressione atmosferica p. È facile osservare che la condizione di tipo fisico $n_i \geq 0$, $i = 1, \ldots, 10$, implica $x_i \geq 0$, $i = 1, \ldots, 5$.

L'equilibrio della combustione del propano è descritto mediante le seguenti equazioni non lineari $F(x) = 0$ dove $x = (x_1, x_2, \ldots, x_5)^T$, $F = (F_1, F_2, \ldots, F_5)^T$ e

$$
\begin{aligned}
F_1(x) &= x_1x_2 + x_1 - 3x_5 \\
F_2(x) &= 2x_1x_2 + x_1 + 2R_{10}x_2^2 + x_2x_3^2 + R_7x_2x_3 + R_9x_2x_4 + R_8x_2 - Rx_5 \\
F_3(x) &= 2x_2x_3^2 + R_7x_2x_3 + 2R_5x_3^2 + R_6x_3 - 8x_5 \\
F_4(x) &= R_9x_2x_4 + 2x_4^2 + 4Rx_5 \\
F_5(x) &= x_1 + x_2(x_1 + 2R_{10}x_2 + x_3^2 + +R_7x_3 + R_9x_4 + R_8) + x_3(R_5x_3 + R_6) + x_4^2 - 1
\end{aligned}
$$

ed è facile verificare che il sistema ammette soluzioni prive di significato fisico. Ad esempio, se $p = 40$ ed $R = 10$, il sistema ammette quattro soluzioni reali e dodici soluzioni complesse. Delle quattro soluzioni reali, solo una ha componenti positive. Le rimanenti hanno x_3 e/o x_4 negative. È dunque necessario introdurre un vincolo di positività sulle variabili x. Il problema non lineare risultante ha quindi la forma (1.1) con $l = (0, \ldots, 0) \in \mathbb{R}^5$ e $u = (+\infty, \ldots, +\infty) \in \mathbb{R}^5$.

1.1.2 Sistemi di equazioni e disequazioni non lineari

In questo paragrafo mostriamo che il problema (1.1) è generale nonostante la semplicità del vincolo considerato. È dato un sistema di equazioni e disequazioni non lineari della forma:

$$
\begin{aligned}
&c_i(x) = 0 \quad \text{se} \quad i \in E, \\
&c_i(x) \leq 0 \quad \text{se} \quad i \in I, \\
&x \in \Omega,
\end{aligned}
\tag{1.3}
$$

dove I ed E sono insiemi di indici tali che $I \cup E = \{1, \ldots, m\}$ e $I \cap E = \emptyset$ e $c_i : \mathbb{R}^n \to \mathbb{R}$ sono funzioni differenziabili con continuità per $i = 1, \ldots, m$.

Questo problema si può interpretare come una generalizzazione di (1.1). Infatti, in (1.3) abbiamo un sistema di equazioni non lineari, $c_i(x) = 0$, $i \in E$, ed un insieme di vincoli in cui quelli semplici, $x \in \Omega$, sono distinti da quelli più generali rappresentati dalle equazioni $c_i(x) \leq 0$, $i \in I$. Nel seguito mostreremo che questa classe di problemi è riconducibile alla forma (1.1).

Indichiamo con m_e e m_i il numero di equazioni e disequazioni rispettivamente. Sia $C(x) = (c_1(x), \ldots, c_m(x))^T$, $C_E : \mathbb{R}^n \to \mathbb{R}^{m_e}$ la funzione vettoriale le cui componenti sono $c_i(x)$ per $i \in E$, e $C_I : \mathbb{R}^n \to \mathbb{R}^{m_i}$ la funzione vettoriale le cui componenti sono $c_i(x)$ per $i \in I$. Allora, (1.3) può essere riscritto in forma compatta come

$$
\begin{aligned}
&C_E(x) = 0, \\
&C_I(x) \leq 0, \\
&x \in \Omega,
\end{aligned}
$$

ed una prima riformulazione del problema si ottiene ponendo

$$
F(x, z) = \begin{pmatrix} C_E(x) \\ z - C_I(x) \end{pmatrix} = 0, \quad x \in \Omega,\ z \leq 0, \tag{1.4}
$$

dove $z \in \mathbb{R}^{m_i}$. Chiaramente questo problema ha la forma (1.1) e $F : \mathbb{R}^{n+m_i} \to \mathbb{R}^{m_e+m_i}$.

In base al lavoro [11], presentiamo riformulazioni alternative. Definiamo la funzione $G : \mathbb{R}^n \to \mathbb{R}^m$ le cui componenti hanno la forma

$$(G(x))_i = \begin{cases} c_i(x) & \text{se} \quad i \in E \\ \max\{c_i(x), 0\} & \text{se} \quad i \in I \end{cases} . \tag{1.5}$$

È facile osservare che una soluzione di (1.3) risolve anche il sistema non lineare

$$G(x) = 0, \quad x \in \Omega.$$

Tuttavia, è evidente che la funzione G non è differenziabile con continuità in $\mathbb{R}^n$ dato che la funzione $t_+ = \max\{t, 0\}$ è solo continua. In alternativa, osserviamo che la funzione $\frac{1}{2}t_+^2$ soddisfa

$$\frac{d}{dt}\left(\frac{1}{2}t_+^2\right) = t_+,$$

cioè $t_+^2 \in C^1(\mathbb{R})$ (l'utilizzo dello scalare 1/2 ha solo motivazioni algebriche). Quindi proponiamo la riformulazione

$$F(x) = \begin{pmatrix} C_E(x) \\ \frac{1}{2}[\max\{C_I(x), 0\}]^2 \end{pmatrix} = 0, \quad x \in \Omega. \tag{1.6}$$

Il problema (1.6) risulta equivalente al problema di partenza (1.3) e la funzione F è differenziabile con continuità. In particolare, se indichiamo con $J(x) \in \mathbb{R}^{m \times n}$ la matrice Jacobiana di F in x, la riga i-esima di $J(x)$ è data da

$$(J(x))_i = \begin{cases} \nabla c_i(x)^T & \text{se} \quad i \in E \\ \max\{c_i(x), 0\}\nabla c_i(x)^T & \text{se} \quad i \in I \end{cases} , \tag{1.7}$$

per $i = 1, \ldots, m$.

Le due riformulazioni (1.4) e (1.6) sono formalmente equivalenti ma differiscono nella dimensione. Infatti osserviamo che la riformulazione dei vincoli di uguaglianza e disuguaglianza operata in (1.4) comporta un aumento nel numero delle variabili pari al numero di vincoli trasformati, mentre in (1.6) il numero di variabili è pari a quello del problema originario. In generale, dal punto di vista computazionale, è vantaggioso non aumentare la dimensione del problema assegnato e quindi utilizzare la forma (1.6).

Concludiamo questa sezione, osservando che comunemente i sistemi (1.4) o (1.6) sono rettangolari. Ad esempio, nel calcolo di un punto appartenente ad una regione di ammissibilità per il problema di ottimizzazione

$$\begin{aligned} \min \quad & f(x) \\ \text{t.c.} \quad & C_I(x) \leq 0 \\ & x \in \Omega, \end{aligned}$$

si deve risolvere il sistema rettangolare $C_I(x) \leq 0, \quad x \in \Omega$, [19], mentre il calcolo di punti stazionari di un problema ai minimi quadrati

$$\min f(x) = \sum_{i=1}^{m} f_i(x)^2,$$

tipicamente dà luogo ad un sistema sovradeterminato, [26].

1.1.3 Problemi di complementarità misti

I problemi di complementarità misti (MCP) sono sistemi di equazioni non lineari vincolati della forma

$$F(v,s,z) = \begin{pmatrix} G(v,s,z) \\ SZe \end{pmatrix} = 0, \quad s \geq 0, \; z \geq 0, \tag{1.8}$$

dove $v \in \mathbb{R}^{m'}$, $s, z \in \mathbb{R}^{n'}$, $G : \mathbb{R}^{m'+2n'} \to \mathbb{R}^{m'+n'}$, $S = diag(s_1, \ldots, s_{n'})$, $Z = diag(z_1, \ldots, z_{n'})$, $e = (1, \ldots, 1) \in \mathbb{R}^{n'}$ e tutte le componenti dei vettori s e z sono vincolate ad essere non negative. Chiaramente un problema MCP è un problema della forma (1.1) nel quale $n = m' + 2n'$, $m = n$ e l'insieme Ω è definito dai vettori l e u tali che $l_i = -\infty$ per $i = 1, \ldots, m'$, $l_i = 0$ per $i = m'+1, \ldots, n$ e $u_i = +\infty$ per $i = 1, \ldots, n$.

L'importanza del problema (1.8) risiede nel fatto che problemi non lineari di vario tipo possono essere espressi mediante un MCP. Questa formulazione comprende infatti problemi di complementarità standard, disuguaglianze variazionali e le condizioni di Karush-Kuhn-Tucker per problemi di programmazione non lineare. Poiché il concetto di complementarità può essere interpretato come sinonimo della nozione di sistema di equilibrio, i problemi di complementarità sono di notevole interesse nelle applicazioni ingegneristiche ed economiche, [13].

1.2 Metodi numerici

I metodi numerici per la risoluzione di sistemi non lineari sono iterativi. Quindi, assegnato un vettore iniziale $x_0 \in \mathbb{R}^n$, sotto opportune condizioni essi generano una successione di vettori $\{x_k\}$ che converge ad una soluzione del sistema.

La risoluzione del sistema non lineare vincolato (1.1) deve essere svolta utilizzando metodi che gestiscono i vincoli poiché i metodi numerici per sistemi non vincolati non sono adeguati alla risoluzione di (1.1). In particolare, i metodi per sistemi non vincolati possono convergere a soluzioni non ammissibili pur disponendo di un'accurata approssimazione di una soluzione, [25].

L'approccio comunemente usato nel passato non sfruttava la struttura del problema originario bensì associava ad (1.1) un problema di ottimizzazione ai minimi quadrati. In pratica, sono state considerate le due riformulazioni

$$\min_{x \in \Omega} f(x) = \frac{1}{2}\|F(x)\|^2, \tag{1.9}$$

e

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|F(x)\|^2 + \sum_{i=1}^{n}[\max\{l_i - x_i, 0\}]^2 + \sum_{i=1}^{n}[\max\{x_i - u_i, 0\}]^2. \qquad (1.10)$$

Evidentemente (1.9) è un problema ai minimi quadrati con vincoli semplici mentre (1.10) è un problema ai minimi quadrati non vincolato poiché i vincoli sono inglobati nella funzione obiettivo f. Le soluzioni del sistema (1.1) sono minimi globali per (1.9) e (1.10) ma i problemi ai minimi quadrati possono ammettere soluzioni che non soddisfano il sistema. Poiché le soluzioni x^* del sistema (1.1) sono soluzioni a "residuo nullo", cioè $f(x^*) = 0$, i problemi (1.9) e (1.10) possono essere risolti con metodi per problemi ai minimi quadrati denominati metodi di tipo Gauss-Newton, [9, 12, 17, 29]. Questi metodi richiedono la conoscenza della matrice Jacobiana J di F contrariamente ai metodi per il problemi ai minimi quadrati con "residuo non piccolo" che necessitano della matrice hessiana di f, [29].

Nonostante le buone prestazioni dei metodi di tipo Gauss-Newton per (1.9) e (1.10), ben note differenze fra i metodi per i sistemi non lineari e problemi di minimo hanno indotto a definire procedimenti specifici per (1.1), [12]. Ricordiamo che esiste una vasta letteratura riguardo alla risoluzione di sistemi non lineari che descrivono problemi di complementarità, [35]. Al contrario, lo studio dei metodi numerici per sistemi non lineari vincolati con arbitraria funzione non lineare F è stato oggetto di ricerca solo recentemente.

Relativamente ai metodi per la risoluzione del sistema (1.1) con funzione F arbitraria, in letteratura si hanno i seguenti contributi [2, 3, 4, 5, 14, 15, 20, 21, 23, 31, 33, 34]. Tutti i procedimenti proposti gestiscono in modo automatico i vincoli e generano una successione di iterate appartenenti ad Ω. In questo modo i metodi sono applicabili nella risoluzione di problemi in cui Ω coincide con il dominio della funzione non lineare. Inoltre questi metodi iterativi sono globalmente convergenti, cioè convergono per una scelta arbitraria dell'approssimazione iniziale. La funzione F deve essere differenziabile con continuità per tutti i metodi eccettuati quelli presentati in [21, 33, 34], nei quali F può essere non differenziabile (funzione "nonsmooth"). Nel primo caso, ad ogni iterazione è utilizzata la matrice Jacobiana, nel secondo è previsto l'utilizzo di una matrice Jacobiana di F generalizzata, [8].

L'unico metodo sviluppato per problemi vincolati rettangolari è [14] ed è applicabile a problemi sottodeterminati.

I metodi esistenti possono essere divisi in tre classi come segue. La prima classe consiste di metodi ai vincoli attivi, [20, 31]. Ad ogni iterazione viene predetto l'insieme dei vincoli attivi di una soluzione x^* del sistema (1.1); quindi viene risolto un sottoproblema di dimensione minore o uguale a quello originario ed espresso nelle variabili libere. Il principale costo computazionale di una iterazione è la risoluzione di un sistema lineare algebrico con dimensioni uguale al numero predetto di variabili non attive di x^*.

I metodi appartenenti alla seconda classe sono proposti in [15, 23, 34] e richiedono ad ogni iterazione la risoluzione di un problema di programmazione quadratica e convessa con vincoli semplici. Le difficoltà connesse all'implementazione di questi metodi

dipendono dalla risoluzione del problema di programmazione quadratica. Nonostante l'utilizzo di efficienti metodi per tale problema, generalmente il costo computazionale di questi procedimenti è superiore a quello degli altri approcci.

La terza classe di procedimenti è costituita dai metodi trust-region con scalatura affine ([2, 3, 4, 5, 14, 21]) e senza scalatura ([33]). I procedimenti trust-region che utilizzano scalatura affine realizzano una strategia di globalizzazione di tipo trust-region nella quale le variabili sono opportunamente scalate tenendo conto dei vincoli. I metodi risultanti hanno proprietà teoriche analoghe a quelle di tutti i procedimenti sopra citati ma prestazioni numeriche superiori. In particolare, nel confronto numerico con i metodi [20] e [23] per (1.1) e con metodi Gauss-Newton per la risoluzione di (1.9), si sono rilevati più convenienti riguardo al costo computazionale e più "robusti", cioè capaci di risolvere un maggior numero di problemi, vedi [4].

Il principale costo computazionale di una iterazione di un metodo trust-region con scalatura affine è dato dalla risoluzione di un sistema lineare algebrico con dimensione pari a quella del problema assegnato. A questo proposito osserviamo che il metodo fornito in [5] è l'unico metodo sviluppato in modo da differenziare la risoluzione dei problemi in base alla loro dimensione. Infatti il metodo è formulato in modo generale e consente di risolvere il sistema lineare algebrico che nasce ad ogni iterazione in modo approssimato se il problema è di grandi dimensioni (metodo "inesatto"), in modo esatto altrimenti (metodo "esatto").

Considerate le buone proprietà teoriche e numeriche dei metodi trust-region con scalatura affine, in questa tesi proponiamo un nuovo procedimento che appartiene a tale classe e può essere applicato a sistemi rettangolari nei quali la funzione F è differenziabile con continuità. Il procedimento sviluppato è un'estensione della versione "esatta" del metodo proposto nel recente lavoro [5] ed è formulato in modo generale. In particolare, il nuovo procedimento può essere applicato a problemi sottodeterminati e sovradeterminati ed ha applicabilità superiore a quella del metodo [5] nell'ambito dei problemi quadrati poiché consente la risoluzione di problemi in cui la matrice Jacobiana è singolare in una iterata. Il nuovo procedimento preserva le buone proprietà teoriche dei metodi esistenti. Esso è globalmente convergente e se la successione generata converge ad una soluzione di (1.1) in cui la la matrice Jacobiana ha rango massimo per colonne, allora la convergenza asintotica è veloce, i.e. quadratica.

Notazioni

Indichiamo con $\mathbb{R}^{m\times n}$ lo spazio delle matrici a coefficienti reali di dimensione $m \times n$.

Scriviamo $int(\Omega)$ per indicare l'interno non vuoto di Ω. Per ogni funzione $F : X \to \mathbb{R}^m$, differenziabile nel punto $x \in X \subset \mathbb{R}^n$, la matrice Jacobiana di F in x è identificata con $J(x) \in \mathbb{R}^{m\times n}$; $J(x)^+$ denota invece la pseudoinversa di Moore-Penrose di $J(x) \in \mathbb{R}^{n\times m}$ (vedi Definizione B.1). Quando è chiaro dal contesto, viene omesso l'argomento della funzione e viene usata la notazione F_k per indicare $F(x_k)$.

La norma-2 di una matrice $m \times n$, è indicata con $\|\cdot\|$. Per ogni vettore $y \in \mathbb{R}^n$, la sfera aperta centrata in y e di raggio ρ è indicata con $B_\rho(y)$, i.e. $B_\rho(y) = \{x : \|x - y\| < \rho\}$.

Come regola generale, viene usato il simbolo $(x)_i$ per rappresentare l'i-esima componente di un vettore x e $(A)_{ij}$ per l'elemento (i, j) di una matrice A, ma, quando è chiaro dal contesto, vengono omesse le parentesi.

Per le etichette alfabetiche, si faccia riferimento alle Appendici.

Capitolo 2

Preliminari

In questo capitolo presentiamo alcuni importanti metodi numerici per la risoluzione di sistemi quadrati e non vincolati. Nel prossimo capitolo, i procedimenti descritti saranno generalizzati ed adattati alla risoluzione di sistemi rettangolari e vincolati.

Consideriamo il sistema non lineare quadrato e non vincolato

$$F(x) = 0, \tag{2.1}$$

dove $F : D \to \mathbb{R}^n$ è una funzione differenziabile con continuità e $D \subseteq \mathbb{R}^n$ è un insieme aperto e convesso.

I metodi numerici per la risoluzione di (1.1) sono iterativi, ovvero assegnato un vettore iniziale $x_0 \in \mathbb{R}^n$, generano una successione di vettori $\{x_k\}$. Un metodo iterativo è *convergente* se la successione $\{x_k\}$ generata converge ad una soluzione x^* del problema, i.e. $\lim_{k\to+\infty} x_k = x^*$; è *localmente convergente* se la convergenza della successione generata dipende in modo critico dalla vicinanza dell'approssimazione iniziale x_0 a x^*. Se la convergenza non dipende dalla bontà dell'approssimazione iniziale il metodo è detto *globalmente convergente*.

Algoritmi efficaci per risolvere il problema (2.1) sono di solito basati sul metodo di Newton, o sue varianti, a motivo delle sue buone proprietà di convergenza locale. Per aumentare la regione di convergenza ed ottenere procedimenti globalmente convergenti, il metodo di Newton viene combinato con opportune strategie dette di globalizzazione. Nel seguito presenteremo le principali caratteristiche del metodo di Newton e di una classe di strategie di globalizzazione, dette metodi trust-region.

2.1 Il metodo di Newton

Il metodo di Newton è un procedimento iterativo per la risoluzione del problema (2.1). Data un'approssimazione iniziale x_0, esso genera una successione $\{x_k\}$ nel modo seguente. Alla k-esima iterazione, dato il punto x_k, la funzione F viene approssimata in un intorno di x_k mediante il seguente modello lineare

$$\mathcal{L}_k(p) = F_k + J_k p, \tag{2.2}$$

dove $p = x - x_k$ e J è la matrice jacobiana di F. Tale modello è ottenuto con lo sviluppo di Taylor di F in x_k, troncato al prim'ordine. In particolare dal Teorema A.1 si ottiene la relazione

$$F(x_k + p) = \mathcal{L}_k(p) + \int_0^1 (J(x_k + tp) - J(x_k))\, p\, dt,$$

che indica la distanza tra il modello lineare $\mathcal{L}_k(p)$ e la funzione $F(x_k+p)$. Dato che la matrice jacobiana $J(x)$ è una funzione continua, abbiamo che $\|J(x_k+tp)-J(x_k)\| \to 0$ quando $p \to 0$ per ogni $t \in [0, 1]$.

Per calcolare la nuova iterata x_{k+1}, viene calcolato il passo $p \in \mathbb{R}^n$ tale che $\mathcal{L}_k(p) = 0$. In particolare, la nuova iterata è $x_{k+1} = x_k + p_k^N$, dove p_k^N è il cosiddetto passo di Newton, cioè la soluzione del sistema lineare

$$J_k p_k^N = -F_k. \tag{2.3}$$

Il metodo di Newton può essere schematizzato come segue:

Metodo di Newton
Dato: $x_0 \in \mathbb{R}^n$
Per k=0, 1, ...
1. *Se* J_k *è invertibile*
   calcola la soluzione p_k^N *dell'equazione di Newton (2.3)*
   altrimenti
   stop
2. *Poni* $x_{k+1} = x_k + p_k^N$

Il seguente Teorema mostra le principali proprietà di convergenza del metodo di Newton.

Teorema 2.1.1 *[22] Sia F una funzione differenziabile con continuità in un insieme aperto convesso $D \subseteq \mathbb{R}^n$. Sia $J : D \to \mathbb{R}^{n\times n}$ lipschitziana e $x^* \in D$ tale che $F(x^*) = 0$ e $J(x^*)$ invertibile. Allora esiste $\epsilon > 0$ tale che se $x_0 \in B_\epsilon(x^*)$, la successione $\{x_k\}$ generata dal metodo di Newton è ben definita, converge ad x^* e*

$$\|x_{k+1} - x^*\| = O(\|x_k - x^*\|^2), \qquad k \geq 0,$$

ovvero la convergenza è quadratica.

Alla luce del precedente teorema osserviamo che è garantita l'esistenza della successione $\{x_k\}$ e la sua convergenza se x_0 è sufficientemente vicino a x^* e $J(x^*)$ è invertibile. In questo caso il metodo ha convergenza veloce, cioè quadratica. Se l'approssimazione iniziale x_0 è distante dalla soluzione, il metodo di Newton può fallire perché la successione $\{x_k\}$ è ben definita ma non converge ad x^* oppure J_k è singolare per un certo k ed il procedimento iterativo si interrompe.

2.2 Metodi trust-region

I metodi trust-region sono procedimenti utilizzati per ampliare il dominio di convergenza del metodo di Newton applicato al problema (2.1). Questi metodi sono applicati al problema di ottimizzazione non vincolata

$$\min_x f(x) = \frac{1}{2}\|F(x)\|^2, \tag{2.4}$$

di cui sono soluzione tutte le radici x^* di (2.1). In particolare, tali radici sono minimi globali per la funzione f, detta funzione di merito.

I metodi trust-region garantiscono ad ogni iterazione la decrescita della funzione di merito, i.e.

$$f(x_{k+1}) < f(x_k),$$

e generano una successione $\{x_k\}$ la cui convergenza ad un punto $\bar{x}$ soddisfacente le condizioni di ottimalità del prim'ordine, cioè ad un punto tale che $\nabla f(\bar{x}) = 0$, non dipende dalla vicinanza del punto iniziale x_0 ad $\bar{x}$.

Le iterate sono costruite sfruttando un opportuno modello quadratico della funzione di merito ed una regione di affidabilità, detta appunto trust-region. La trust-region contiene l'iterata corrente ed è una regione in cui si ritiene che il modello quadratico approssimi bene la funzione di merito. La nuova iterata è quindi calcolata risolvendo il cosiddetto problema di trust-region, ovvero minimizzando il modello quadratico all'interno della regione di affidabilità. Presentiamo il paradigma per l'iterata k-esima di un metodo trust-region:

a) Si costruisce la funzione modello $m_k(p) = \frac{1}{2}\|\mathcal{L}_k(p)\|^2$ di f in un intorno di x_k, dove $\mathcal{L}_k$ è definito in (2.2). Quindi

$$m_k(p) = f_k + \nabla {f_k}^T p + \frac{1}{2} p^T J_k^T J_k p,$$

dove $p = x - x_k$ e $\nabla f_k = J_k^T F_k$.
Poiché m_k ed f_k si accordano al prim'ordine, possiamo ritenere che il modello quadratico $m_k(p)$ sia una buona approssimazione di $f(x)$ in un opportuno intorno di x_k.

b) Si calcola il minimo del modello quadratico m_k nella trust-region. Questa regione è definita come l'insieme $\{x \in \mathbb{R}^n : x = x + p,\ \|p\| \leq \Delta_k\}$, cioè è una sfera centrata in x_k con raggio $\Delta_k > 0$, detto raggio di trust-region. Quindi il problema di trust-region ha la forma

$$\min_p \{m_k(p) :\ \|p\| \leq \Delta_k\}. \tag{2.5}$$

Indicata con $p_{tr}(\Delta_k)$ la sua soluzione

$$p_{tr}(\Delta_k) = \text{argmin}\{m_k(p),\ p \in \mathbb{R}^n,\ \|p\| \leq \Delta_k\},$$

per costruzione $m_k(p_{tr}(\Delta_k)) < m_k(p)$ per ogni p tale che $\|p\| \leq \Delta_k$ e in particolare si ha $m_k(p_{tr}(\Delta_k)) < m_k(0) = f(x_k)$. Dato che m_k è il modello quadratico di f, se la trust-region è stata scelta correttamente avremo $f(x_k + p_{tr}(\Delta_k)) < f(x_k)$.

c) Si esamina la nuova iterata della forma

$$x_{k+1} = x_k + p_{tr}(\Delta_k),$$

e si accetta x_{k+1} se si ha una "sufficiente" riduzione nel valore della funzione f. Specificatamente, posto

$$\rho_f(p_{tr}(\Delta_k)) = \frac{f(x_k) - f(x_k + p_{tr}(\Delta_k))}{m_k(0) - m_k(p_{tr}(\Delta_k))},$$

si ha riduzione nel valore di f rispetto a $f(x_k)$ solo se $\rho_f(p_{tr}(\Delta_k)) > 0$.
In pratica si accetta $p_{tr}(\Delta_k)$ se fornisce una decrescita "sufficiente" ovvero se $\rho_f(p_{tr}(\Delta_k)) \geq \eta_1$, $\eta_1 \in (0, \frac{1}{4})$. In questo caso si ritiene che il valore $\rho_f(p_{tr}(\Delta_k))$ indichi un buon accordo tra m_k e f nella trust-region. Se $\rho_f(p_{tr}(\Delta_k)) < \eta_1$, si rifiuta x_{k+1}, si riduce Δ_k e si ripete il procedimento partendo dal passo b).

d) Il valore di $\rho_f(p_{tr}(\Delta_k))$ è utilizzato anche per scegliere il raggio Δ_{k+1} per l'iterata successiva. Solitamente, data una costante $\eta_2 \in [\frac{1}{4}, 1)$, se $\rho_f(p_{tr}(\Delta_k)) \in [\eta_1, \eta_2)$ si pone $\Delta_{k+1} = \Delta_k$. Se $\rho_f(p_{tr}(\Delta_k)) \geq \eta_2$ si pone $\Delta_{k+1} \geq \Delta_k$.

Il seguente teorema caratterizza la soluzione del problema di trust-region (2.5).

Teorema 2.2.1 *[12, Lemma 6.4.1] Sia $J_k \in \mathbb{R}^{n\times n}$ invertibile. Allora, il problema (2.5) è risolto da $p_{tr}(\Delta_k) = p(\lambda)$ dove*

$$p(\lambda) = -(J_k^T J_k + \lambda I)^{-1} \nabla f_k,$$

per l'unico $\lambda \geq 0$ tale che $\|p(\lambda)\| = \Delta_k$ a meno che $p(0) = -J_k^{-1} F_k$ soddisfi $\|p(0)\| \leq \Delta_k$. In questo caso, $p_{tr}(\Delta_k) = p(0)$. La funzione $\|p(\lambda)\|$, $\lambda \geq 0$, è una funzione monotona decrescente di λ.

Utilizzando il precedente teorema, studiamo la curva $\Gamma(\lambda) = \{p(\lambda), \lambda \geq 0\}$ ed analizziamo come varia la soluzione del problema di trust-region (2.5) al variare di Δ_k.

- Il passo $p(0)$ coincide con il passo di Newton p_k^N definito in (2.3). Quindi se la trust-region è sufficientemente ampia da soddisfare $\|p_k^N\| \leq \Delta_k$, si ha $p_{tr}(\Delta_k) = p_k^N$.

- Se $\lambda \neq 0$, la soluzione $p(\lambda)$ ha norma pari a Δ_k e $\lim_{\lambda\to+\infty} \|p(\lambda)\| = 0$. Inoltre al crescere di λ e per λ sufficientemente grande, otteniamo $p(\lambda) \cong -\frac{1}{\lambda}\nabla f_k$. Allora la curva $\Gamma(\lambda)$ connette p_k^N a 0 tendendo a diventare tangente a $-\nabla f_k$ e si può concludere che $p_{tr}(\Delta_k)$ tende alla direzione del gradiente per $\Delta_k \to 0$, vedi Figura 2.1.

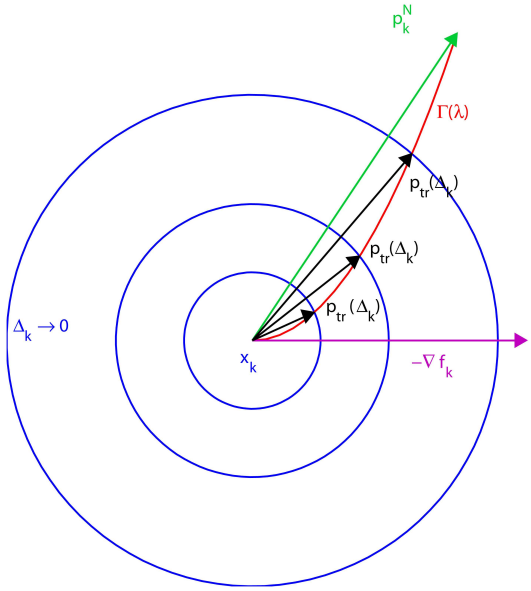

Figura 2.1: Curva $\Gamma(\lambda) = \{p(\lambda),\ \lambda \geq 0\}$ e comportamento di $p_{tr}(\Delta_k)$ per $\Delta_k \to 0$.

La fase più complessa nella realizzazione di un metodo trust-region è la risoluzione del problema (2.5). Infatti il Teorema 2.2.1 ne caratterizza la soluzione $p_{tr}(\Delta_k)$ ma non fornisce la forma analitica. Non essendo disponibile la soluzione esatta, comunemente nei metodi trust-region si utilizzano procedimenti numerici che forniscono una soluzione approssimata del problema (2.5).

Presentiamo adesso risultati di convergenza globale per i metodi trust-region. Premettiamo la definizione del cosiddetto punto di Cauchy.

Definizione 2.2.1 *Il punto di Cauchy $p^C(\Delta_k)$ è il minimo di m_k all'interno della trust-region e lungo la direzione di più ripida discesa $-\nabla f_k$, cioè*

$$p^C(\Delta_k) = \text{argmin}\{m_k(p),\ \ p = -\tau\nabla f_k,\ \ \tau > 0,\ \ \|p\| \leq \Delta_k\}.$$

Osservando che $m_k(-\tau\nabla f_k) = f_k - \tau\|\nabla f_k\|^2 + \frac{1}{2}\tau^2\|J_k\nabla f_k\|^2$, il minimo p_k^U non vincolato è dato da

$$p_k^U = -\bar{\tau}\nabla f_k, \quad \text{dove } \bar{\tau} = \frac{\|\nabla f_k\|^2}{\|J_k\nabla f_k\|^2}. \tag{2.6}$$

Se p_k^U sta dentro la trust-region, allora $p^C(\Delta_k)$ coincide con p_k^U; se invece p_k^U non appartiene alla trust-region, allora $p^C(\Delta_k)$ coincide con uno dei due punti nei quali la retta passante per l'origine e con direzione data dalla direzione di più ripida discesa interseca la frontiera della trust-region. Riassumendo:

$$p^C(\Delta_k) = \begin{cases} -\bar{\tau}\nabla f_k, & \text{se} \quad \bar{\tau}\|\nabla f_k\| \leq \Delta_k \\ -\Delta_k \frac{\nabla f_k}{\|\nabla f_k\|} & \text{altrimenti}\ , \end{cases}$$

ed in forma compatta:

$$p^C(\Delta_k) = -\hat{\tau}_k \nabla f_k \quad \text{dove} \quad \hat{\tau}_k = \min\{\frac{\|\nabla f_k\|^2}{\|J_k \nabla f_k\|^2}, \frac{\Delta_k}{\|\nabla f_k\|}\}. \tag{2.7}$$

L'importanza del punto di Cauchy è dovuta al fatto che esso permette di caratterizzare i metodi trust-region convergenti. Infatti, posto

$$\rho^C(p_{tr}(\Delta_k)) = \frac{m_k(0) - m_k(p_{tr}(\Delta_k))}{m_k(0) - m_k(p^C(\Delta_k))},$$

il seguente risultato mostra che l'uso di soluzioni approssimate $p_{tr}(\Delta_k)$ del problema (2.5) che producono una riduzione in m_k almeno pari ad una frazione fissata della riduzione raggiunta da $p^C(\Delta_k)$, dà luogo a procedimenti convergenti.

Teorema 2.2.2 *[29, Teorema 11.9] Sia $\beta_1 \in (0,1)$, $\beta_2 \in (0, \frac{1}{4})$. Supponiamo che J sia continua e lipschitziana in un intorno $\mathcal{D}$ dell'insieme di livello $\{x : f(x) < f(x_0)\}$ e che $\|J(x)\|$ sia limitata superiormente in $\mathcal{D}$. Sia $\{x_k\}$ la successione generata da un metodo trust-region e tale che $x_{k+1} = x_k + p_{tr}(\Delta_k)$ dove*

$$\begin{aligned} \rho^C(p_{tr}(\Delta_k)) &\geq \beta_1, \\ \rho_f(p_{tr}(\Delta_k)) &\geq \beta_2, \end{aligned} \tag{2.8}$$

per ogni $k \geq 0$. Allora

$$\lim_{k\to\infty} \nabla f_k = 0. \tag{2.9}$$

È importante sottolineare che se $p_{tr}(\Delta_k)$ è soluzione esatta di (2.5) allora esso soddisfa (2.8) poiché $\rho^C(p_{tr}(\Delta_k)) \geq 1$.

Si noti che se x^* è un punto limite di $\{x_k\}$ e $J(x^*)$ è non singolare, la condizione (2.9) facilmente implica $F(x^*) = 0$ e quindi x^* è soluzione del problema (2.1). D'altro canto, possono esistere punti limite $\bar{x}$ di $\{x_k\}$ che sono minimi per f ma non risolvono il sistema (2.1). In questo caso $J(\bar{x})$ è singolare.

Concludiamo questo capitolo con il risultato più rilevante. Esso mostra le condizioni sotto le quali il metodo trust-region asintoticamente si riduce al metodo di Newton. In questo caso, il metodo trust-region consente di rendere globale la convergenza del metodo di Newton e di recuperarne asintoticamente le buone proprietà di convergenza locali.

Teorema 2.2.3 *[29, Teorema 11.10] Assumiamo le ipotesi del Teorema 2.2.2. Supponiamo che la successione $\{x_k\}$ generata dal metodo trust-region abbia un punto limite x^* che risolve il problema (2.1) ed sia tale che $J(x^*)$ è non singolare. Allora,*

a) se il problema di trust-region è risolto esattamente, per k sufficientemente grande $x_{k+1} = x_k + p_k^N$ e la successione converge ad x^* quadraticamente;

b) se il problema di trust-region è risolto in modo approssimato e tale che $p_{tr}(\Delta_k) = p_k^N$ ogni volta che $\|p_k^N\| \leq \Delta_k$, allora per k sufficientemente grande $x_{k+1} = x_k + p_k^N$ e la successione converge ad x^* quadraticamente.

Capitolo 3

Il metodo TRESNORE

In questo capitolo presentiamo un nuovo metodo per risolvere sistemi non lineari rettangolari con vincoli semplici. Il metodo è di tipo trust-region con scalatura affine e genera una successione di iterate $\{x_k\}$ strettamente ammissibili, $x_k \in int(\Omega)\ \forall\ k \geq 0$. Nel seguito sarà denominato metodo TRESNORE (metodo Trust-REgion per Sistemi NOn lineari REttangolari) e la presentazione del metodo sarà indipendente dalle dimensioni del problema.

3.1 Descrizione del metodo

Il metodo TRESNORE costituisce una generalizzazione del metodo proposto in [5] per problemi quadrati di piccole e medie dimensioni. Nella definizione del procedimento si utilizza una caratterizzazione delle soluzioni del problema

$$F(x) = 0, \qquad x \in \Omega, \tag{3.1}$$

basata sulla riformulazione

$$\min_{x \in \Omega} f(x) = \frac{1}{2}\|F(x)\|^2. \tag{3.2}$$

Ricordiamo che le condizioni necessarie affinché x^* soddisfi (3.2) sono

$$\begin{array}{lll} (\nabla f(x^*))_i = 0 & \text{se} & l_i < x_i < u_i \\ (\nabla f(x^*))_i \leq 0 & \text{se} & x_i = u_i \\ (\nabla f(x^*))_i \geq 0 & \text{se} & x_i = l_i \end{array}$$

dove $\nabla f(x) = J(x)^T F(x)$, [29]. Queste condizioni sono anche dette di ottimalità del prim'ordine ed equivalgono a richiedere che x^* soddisfi il sistema non lineare

$$D^{-2}(x^*)\nabla f(x^*) = 0, \tag{3.3}$$

dove $D(x)$ è la matrice di scalatura

$$D(x) = diag(|v_1(x)|^{-1/2}, |v_2(x)|^{-1/2}, \ldots, |v_n(x)|^{-1/2}), \tag{3.4}$$

con

$$v_i(x) = \begin{cases} x_i - u_i & \text{se} \quad (\nabla f(x))_i < 0 \quad \text{e} \;\; u_i < \infty \\ x_i - l_i & \text{se} \quad (\nabla f(x))_i > 0 \quad \text{e} \;\; l_i > -\infty \\ \min\{x_i - l_i,\, u_i - x_i\} & \text{se} \quad (\nabla f(x))_i = 0 \quad \text{e} \;\; l_i > -\infty \text{ o } u_i < \infty \\ 1 & \text{altrimenti} \end{cases} \tag{3.5}$$

per $i = 1, \ldots, n$, [9]. Osserviamo che la matrice D non è definita sul bordo di Ω ma la matrice D^{-1} può essere estesa con continuità fin sul bordo e tale estensione verrà ancora indicata con D^{-1}. Il sistema (3.3) è continuo, ma non è differenziabile nei punti x tali che $v_i(x) = 0$ per qualche $1 \leq i \leq n$. Tale situazione può essere facilmente evitata, mantenendo la successione $\{x_k\}$ interna a Ω.

I metodi iterativi per la risoluzione del sistema (3.1) devono utilizzare direzioni di movimento ben angolate rispetto ai vincoli. In pratica, le direzioni di movimento utilizzate per aggiornare le iterate devono consentire un passo "ampio" all'interno della scatola Ω prima che siano violati i vincoli. Al contrario una direzione di movimento che punta verso un vincolo, costringe a selezionare un passo di ampiezza molto piccola al fine di restare all'interno della regione di ammissibilità e ciò può pregiudicare la convergenza del metodo iterativo. In questo ambito, la matrice di scalatura D ha notevole rilevanza perché permette di definire direzioni di movimento ben angolate. In particolare, Coleman e Li in [9] hanno mostrato che la direzione del gradiente scalato

$$d_k = -D_k^{-2} \nabla f_k, \tag{3.6}$$

risulta ben angolata rispetto ai bordi di Ω. Vediamo nei dettagli come D_k^{-2} agisce sull'antigradiente $-\nabla f_k$. Sia $x_k \in int(\Omega)$. Ogni elemento diagonale di D_k^{-2} rappresenta la distanza della corrispondente componente di x_k dal vincolo finito in direzione del quale il gradiente ∇f_k indica una diminuzione di f. Per esempio se l ed u sono finite e $(\nabla f_k)_i > 0$, allora $(D_k^{-2})_{ii} = (x_k)_i - l_i$, mentre se $(\nabla f_k)_i < 0$ allora $(D_k^{-2})_{ii} = u_i - (x_k)_i$. Per valutare l'effetto di D_k^{-2} su $-\nabla f_k$, supponiamo che x_k sia vicino al bordo di Ω e distinguiamo due casi.

- Nel primo caso, supponiamo che ∇f_k indichi una riduzione di f in direzione di un vincolo vicino. Nella parte A della Figura 3.1 viene rappresentato il caso in cui $(x_k)_1$ è vicino a l_1 e $(\nabla f_k)_1 > 0$. Poiché f diminuisce muovendosi verso l_1, l'elemento $(D_k^{-2})_{11}$ è piccolo, e più x_k si avvicina al contorno di Ω, più la direzione di d_k diventa tangenziale al bordo. Quindi d_k risulta ben angolata.

- Nel secondo caso, supponiamo che ∇f_k indichi una riduzione di f in direzione interna a Ω. Nella parte B della Figura 3.1 è rappresentato il caso particolare in cui $(x_k)_1$ è vicino a l_1 e $(\nabla f_k)_1 < 0$. Allora l'elemento $(D_k^{-2})_{11}$ è "grande" e ciò rende d_k più angolata verso l'interno di Ω rispetto a $-\nabla f_k$. Quindi, il passo che si può compiere lungo d_k prima di violare i vincoli è molto più ampio di quello lungo $-\nabla f_k$, i.e. d_k è ben angolata.

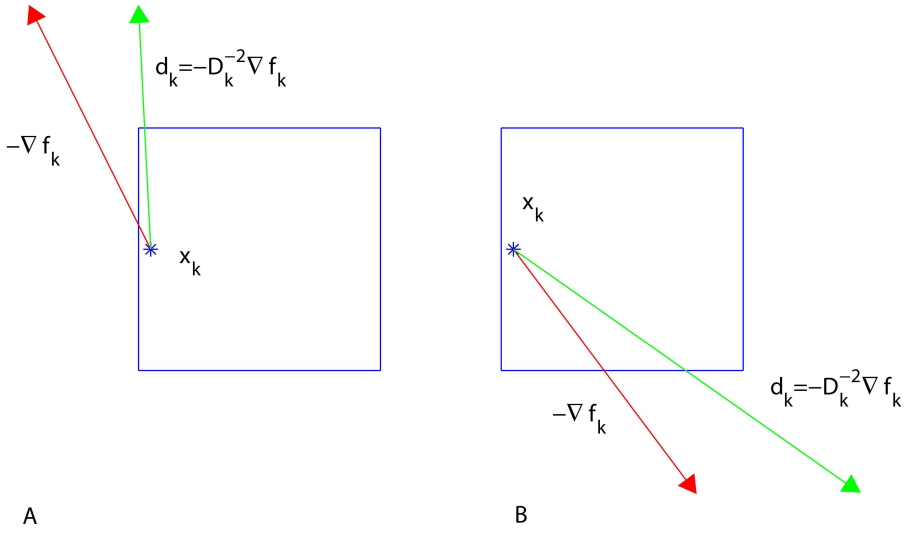

Figura 3.1: Effetti della matrice D_k^{-2} sull'antigradiente $-\nabla f_k$.

Sulla base delle osservazioni precedenti, dato $x_k \in int(\Omega)$, costruiamo l'iterata successiva utilizzando il problema di trust-region ellittico

$$\min_p \{m_k(p) : \quad \|D_k\, p\| \leq \Delta_k\}, \tag{3.7}$$

dove $m_k(p)$ è il modello quadratico

$$m_k(p) = \frac{1}{2}\|J_k p + F_k\|^2 = f_k + \nabla {f_k}^T p + \frac{1}{2} p^T J_k^T J_k p,$$

con $p = x - x_k$, e Δ_k è il raggio corrente di trust-region. Sottolineiamo che l'iterata x_k deve essere strettamente ammissibile affinché D_k esista.
Per giustificare l'utilizzo della trust-region ellittica, mostriamo che, al tendere di Δ_k a zero, la soluzione di (3.7) tende a diventare parallela a d_k e quindi ben angolata. Eseguiamo il cambio di variabile

$$\tilde{p} = D_k\, p. \tag{3.8}$$

Allora il problema (3.7) espresso nella variabile $\tilde{p}$, diventa il problema di trust-region sferico:

$$\min_{\tilde{p}} \{\tilde{m}_k(\tilde{p}) : \quad \|\tilde{p}\| \leq \Delta_k\}, \tag{3.9}$$

dove $\tilde{m}_k(\tilde{p})$ è dato da:

$$\tilde{m}_k(\tilde{p}) = \frac{1}{2}\|J_k D_k^{-1}\tilde{p} + F_k\|^2 = f_k + \nabla {f_k}^T D_k^{-1}\tilde{p} + \frac{1}{2}\tilde{p}^T D_k^{-1} J_k^T J_k D_k^{-1}\tilde{p}. \tag{3.10}$$

Le soluzioni di (3.7) e (3.9) sono fornite dal seguente teorema.

Teorema 3.1.1 *[6, Corollario 7.2.2] Una soluzione $\tilde{p}_{tr}(\Delta_k)$ di (3.9) soddisfa*

$$(D_k^{-1} J_k^T J_k D_k^{-1} + \lambda I)\tilde{p}_{tr}(\Delta_k) = -D_k^{-1}\nabla f_k,$$

dove $\lambda \geq 0$ e $\lambda(\Delta_k - \|\tilde{p}_{tr}(\Delta_k)\|) = 0$. Se $D_k^{-1} J_k^T J_k D_k^{-1} + \lambda I$ è definita positiva allora $\tilde{p}_{tr}(\Delta_k)$ è unica.
Una soluzione $p_{tr}(\Delta_k)$ di (3.7) soddisfa

$$(J_k^T J_k + \lambda D_k^2)p_{tr}(\Delta_k) = -\nabla f_k, \tag{3.11}$$

dove $\lambda \geq 0$ e $\lambda(\Delta_k - \|D_k\, p_{tr}(\Delta_k)\|) = 0$. Se $J_k^T J_k + \lambda D_k^2$ è definita positiva allora $p_{tr}(\Delta_k)$ è unica.

Relativamente alla (3.11), osserviamo che se $J_k^T J_k$ è invertibile, allora $J_k^T J_k + \lambda D_k^2$ è definita positiva per ogni $\lambda \geq 0$. In particolare, ciò si verifica se J_k ha rango massimo per colonne, cioè se il problema è quadrato e J_k è invertibile, oppure se il problema è sovradeterminato e J_k ha rango massimo. Inoltre si ricava da (3.11) che, al crescere di λ, la norma di $p_{tr}(\Delta_k)$ tende a zero e $p_{tr}(\Delta_k)$ tende a diventare parallela a d_k. Quindi se Δ_k tende a zero, $p_{tr}(\Delta_k)$ tende ad una direzione ben angolata.

Analizziamo ulteriormente la soluzione del problema (3.7) e la sua dipendenza da Δ_k. A tale fine è necessario introdurre il minimizzatore non vincolato p_k^{min} di m_k:

$$p_k^{min} = \operatorname*{argmin}_p m_k(p).$$

Il vettore p_k^{min} risolve il seguente sistema lineare detto sistema normale

$$J_k^T J_k p_k^{min} = -J_k^T F_k,$$

vedi [1, Teorema 7.1], ed è unico se il problema è quadrato e J_k è invertibile, oppure se il problema è sovradeterminato e J_k ha rango massimo. Nel primo caso p_k^{min} coincide con il passo di Newton definito in (2.3), nel secondo caso

$$p_k^{min} = -(J_k^T J_k)^{-1} J_k^T F_k = -J_k^+ F_k,$$

dove J_k^+ è la matrice pseudoinversa di J_k, vedi Appendice B. Al contrario esistono infinite soluzioni p_k^{min} se J_k è quadrata e singolare, oppure se il problema è sovradeterminato e J_k non ha rango massimo, o infine se il problema è sottodeterminato. Fra le infinite soluzioni, una sola soluzione ha minima norma. Nel seguito, indicheremo con p_k^N il minimizzatore globale di $m_k(p)$ di norma minima e lo chiameremo *passo di Newton di minima norma*, [14]. Sappiamo che [12]:

$$p_k^N = \operatorname{argmin}\{\|\, p\,\|,\ \ J_k^T J_k p = -J_k^T F_k\} = -J_k^+ F_k. \tag{3.12}$$

Chiaramente p_k^N risolve (3.11) per $\lambda = 0$ ed è soluzione del problema di trust-region (3.7) se $\|D_k p_k^N\| \leq \Delta_k$.

È molto importante osservare che i problemi (3.7) e (3.9) sono equivalenti ma le rispettive soluzioni di minima norma possono differire. In particolare, considerato

il modello $\tilde{m}_k$ definito in (3.10), i suoi minimizzatori globali soddisfano il sistema normale $D_k^{-1} J_k^T J_k D_k^{-1} \tilde{p} = -D_k^{-1} J_k^T F_k$ e quello di minima norma è dato da

$$\tilde{p}_k^{MN} = -(J_k D_k^{-1})^+ F_k.$$

Banalmente, $p_k^N = D_k^{-1} \tilde{p}_k^{MN}$ se la soluzione del problema di trust-region è unica, i.e. se il problema è quadrato o sovradeterminato e J_k ha rango massimo. Altrimenti, in generale si ha $D_k^{-1} \tilde{p}_k^{MN} \neq p_k^N$.
Riepilogando, possiamo concludere che:

- se $J_k^T J_k$ è invertibile e $\|D_k p_k^N\| \leq \Delta_k$, allora $p_{tr}(\Delta_k) = p_k^N$ è l'unica soluzione del problema di trust-region (3.7) e soddisfa (3.11) per $\lambda = 0$;
- se $J_k^T J_k$ è invertibile e $\|D_k p_k^N\| > \Delta_k$, allora l'unica soluzione del problema di trust-region (3.7) ha norma $\|p_{tr}(\Delta_k)\| = \Delta_k$ e risolve (3.11) con $\lambda > 0$;
- se $J_k^T J_k$ è singolare e $\|\tilde{p}_k^{MN}\| \leq \Delta_k$, allora $p_{tr}(\Delta_k) = D_k^{-1} \tilde{p}_k^{MN}$ risolve il problema di trust-region (3.7) e risolve (3.11) con $\lambda = 0$. Possono esistere più soluzioni nella trust-region.
- se $J_k^T J_k$ è singolare e $\|\tilde{p}_k^{MN}\| > \Delta_k$, la soluzione del problema di trust-region (3.7) ha norma $\|p_{tr}(\Delta_k)\| = \Delta_k$ e risolve il sistema (3.11) per $\lambda > 0$. La soluzione è unica dato che la matrice dei coefficienti del sistema (3.11) è definita positiva.

Poiché ogni iterata deve essere strettamente ammissibile, dopo aver calcolato la soluzione del problema di trust-region, definiamo

$$\hat{p}(\Delta_k) = P_\Omega(x_k + p_{tr}(\Delta_k)) - x_k, \tag{3.13}$$

$$\bar{p}_{tr}(\Delta_k) = \alpha_k \hat{p}(\Delta_k), \tag{3.14}$$

dove $P_\Omega(x)$ è la proiezione di x su Ω, i.e. $(P_\Omega(x))_i = \max\{l_i, \min\{x_i, u_i\}\}$, $i = 1, \ldots, n$ e $\alpha_k \in (0, 1)$. È facile verificare che $x_k + \bar{p}_{tr}(\Delta_k) \in int(\Omega)$, $(p_{tr}(\Delta_k))_i$ e $(\bar{p}_{tr}(\Delta_k))_i$ hanno lo stesso segno e

$$|(\bar{p}_{tr}(\Delta_k))_i| < |(p_{tr}(\Delta_k))_i|, \tag{3.15}$$

per $i = 1, \ldots, n$, vedi Figura 3.2.

Formiamo ora la nuova iterata e stabiliamo se è accettabile. La scelta del passo $p(\Delta_k)$ usato per aggiornare x_k deve essere compiuta in modo da garantire la convergenza a punti stazionari di (3.2), ovvero una proprietà fondamentale per la risoluzione del sistema non lineare con vincoli. A tal fine si introduce il cosiddetto *passo di Cauchy generalizzato*:

$$p_c(\Delta_k) = \text{argmin}\{m_k(p),\ p = \tau\, d_k,\ \tau > 0,\ \|D_k p\| \leq \Delta_k,\ x_k + p \in int(\Omega)\}.$$

Il vettore $p_c(\Delta_k)$ è costruito nel modo seguente. Consideriamo il problema di trust-region sferico (3.9). Il vettore

$$\tilde{d}_k = -D_k^{-1} \nabla f_k, \tag{3.16}$$

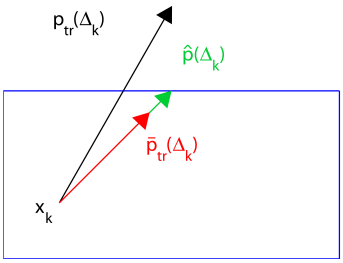

Figura 3.2: I passi $p_{tr}(\Delta_k)$, $\hat{p}(\Delta_k)$, $\bar{p}_{tr}(\Delta_k)$.

rappresenta la direzione di più ripida discesa e dalla (2.7), il punto di Cauchy $\tilde{p}^C(\Delta_k)$ per il problema (3.9) ha la forma $\tilde{p}^C(\Delta_k) = \hat{\tau}_k \tilde{d}_k$ con

$$\hat{\tau}_k = \min\{\frac{\|D_k^{-1}\nabla f_k\|^2}{\|J_k D_k^{-2}\nabla f_k\|^2}, \frac{\Delta_k}{\|D_k^{-1}\nabla f_k\|}\}. \tag{3.17}$$

Quindi, utilizzando il cambio di variabile (3.8), il punto di Cauchy del problema di trust-region ellittico (3.7) ha la seguente espressione:

$$p^C(\Delta_k) = D_k^{-1}\,\hat{\tau}_k\,\tilde{d}_k = \hat{\tau}_k\,d_k. \tag{3.18}$$

Il passo di Cauchy generalizzato $p_c(\Delta_k)$ è definito usando $p^C(\Delta_k)$ e imponendo la stretta ammissibilità di $x_k + p_c(\Delta_k)$. A tal fine, calcoliamo il più lungo passo λ_k che si può compiere da x_k lungo d_k senza violare l'ammissibilità, i.e.

$$\lambda_k = \min_{1\le i\le n} \Lambda_i \quad \text{dove} \quad \Lambda_i = \begin{cases} \max\{\frac{l_i-(x_k)_i}{(d_k)_i}, \frac{u_i-(x_k)_i}{(d_k)_i}\} & \text{se } (d_k)_i \neq 0 \\ \infty & \text{se } (d_k)_i = 0 \end{cases}. \tag{3.19}$$

Quindi, il punto di Cauchy generalizzato è tale che

$$p_c(\Delta_k) = \tau_k\,d_k, \tag{3.20}$$

dove

$$\tau_k = \begin{cases} \hat{\tau}_k & \text{se } x_k + \hat{\tau}_k d_k \in int(\Omega) \\ \theta\lambda_k & \theta \in (0,\,1), \text{ altrimenti.} \end{cases} \tag{3.21}$$

Dall'analisi svolta in [9] sappiamo che per ottenere convergenza globale a punti stazionari per (3.2), il passo $p(\Delta_k)$ usato per aggiornare x_k, deve fornire una decrescita nel modello quadratico m_k almeno pari di una fissata frazione della decrescita che si ottiene in m_k con il passo $p_c(\Delta_k)$, i.e. deve soddisfare

$$\rho_c(p(\Delta_k)) = \frac{m_k(0) - m_k(p(\Delta_k))}{m_k(0) - m_k(p_c(\Delta_k))} \ge \beta_1, \tag{3.22}$$

dove $\beta_1 \in (0,1)$. Questa condizione è analoga alla condizione (2.8) introdotta per i problemi non vincolati. Si osservi che $p_c(\Delta_k)$ soddisfa banalmente (3.22). Al contrario, operando le trasformazioni (3.13)-(3.14) necessarie per la stretta ammissibilità, $\bar{p}_{tr}(\Delta_k)$ può non soddisfare la condizione (3.22).

Un procedimento semplice ed economico dal punto di vista computazionale che permette di soddisfare (3.22) è il seguente. Costruiamo $p(\Delta_k)$ usando una combinazione convessa di $\bar{p}_{tr}(\Delta_k)$ e del punto di Cauchy generalizzato $p_c(\Delta_k)$

$$p(\Delta_k) = t\, p_c(\Delta_k) + (1-t)\bar{p}_{tr}(\Delta_k), \tag{3.23}$$

e scegliamo $t \in [0,1)$ in modo che valga (3.22). In particolare, se $\rho_c(\bar{p}_{tr}(\Delta_k)) \geq \beta_1$, prendiamo $t = 0$, i.e. $p(\Delta_k) = \bar{p}_{tr}(\Delta_k)$. Altrimenti, dal momento che $m_k(t\, p_c(\Delta_k) + (1-t)\, \bar{p}_{tr}(\Delta_k))$ è una funzione quadratica in t e $\rho_c(p_c(\Delta_k)) = 1$, è facile vedere che esiste un $t \in (0,1)$ tale che $p(\Delta_k)$ soddisfa $\rho_c(p(\Delta_k)) = \beta_1^*$ per ogni $\beta_1^* \in [\beta_1, 1]$.

L'analisi teorica del procedimento evidenzierà che l'utilizzo di $p_c(\Delta_k)$ in (3.23) assicura l'esistenza di un vettore $p(\Delta_k)$ che soddisfa (3.22) e rende il metodo globalmente convergente a punti stazionari di (3.2). Invece l'uso di $\bar{p}_{tr}(\Delta_k)$ e un'opportuna scelta degli scalari α_k in (3.14) possono dare luogo ad una convergenza asintotica rapida.

Infine, analogamente allo studio dei problemi non vincolati, il passo $p(\Delta_k)$ è accettato se è soddisfatta la condizione

$$\rho_f(p(\Delta_k)) = \frac{f(x_k) - f(x_k + p(\Delta_k))}{m_k(0) - m_k(p(\Delta_k))} \geq \beta_2, \tag{3.24}$$

con $\beta_2 \in (0,1)$. Il soddisfacimento di questa condizione assicura un buon accordo tra la funzione modello m_k e la funzione obiettivo f. Se (3.24) non è soddisfatta, $p(\Delta_k)$ viene rifiutato e il raggio di trust-region viene modificato mediante riduzioni successive finché la richiesta di accuratezza (3.24) non è soddisfatta.

Presentiamo uno schema riassuntivo del metodo risultante.

METODO TRESNORE

Dati $x_0 \in int(\Omega)$, $\bar{\Delta}_0 \geq \Delta_{min} > 0$, β_1, β_2, δ, $\theta \in (0,1)$.

Per $k = 0, 1, \ldots$

1. *Poni* $\Delta_k = \bar{\Delta}_k/\delta$.
2. *Scegli* $\alpha_k \in (0,1)$.
3. *Ripeti fintanto* $\rho_f(p(\Delta_k)) \geq \beta_2$.
   - *3.1 Poni* $\Delta_k = \delta\Delta_k$.
   - *3.2 Calcola* $p_{tr}(\Delta_k)$ *soluzione di (3.7).*
   - *3.3 Costruisci* $\bar{p}_{tr}(\Delta_k)$ *usando (3.13) e (3.14).*
   - *3.4 Calcola* $p_c(\Delta_k)$ *definito in (3.20) e (3.21).*
   - *3.5 Trova* $p(\Delta_k)$ *tale che* $\rho_c(p(\Delta_k)) \geq \beta_1$.
4. *Poni* $x_{k+1} = x_k + p(\Delta_k)$.
5. *Scegli* $\bar{\Delta}_{k+1} \geq \Delta_{min}$.

Nel metodo TRESNORE, $\bar{\Delta}_k$ rappresenta il raggio iniziale di trust-region all'iterazione k-esima. Per sviluppare l'analisi di convergenza del metodo, imporremo a $\bar{\Delta}_k$ di essere limitato dal basso da un prefissato valore Δ_{min} per ogni $k \geq 0$. I dettagli dell'implementazione del metodo saranno forniti nel Capitolo 4.

3.2 Analisi di convergenza

In questa sezione studiamo le proprietà di convergenza del nuovo metodo TRESNORE assumendo la funzione $F : D \to \mathbb{R}^m$ differenziabile con continuità e le seguenti condizioni sulla matrice Jacobiana J. Sia $r > 0$ e chiamiamo

$$L = \bigcup_{k=0}^{\infty} \{x \in D \ : \ \|x - x_k\| \leq r\}, \tag{3.25}$$

un intorno della successione $\{x_k\}$ generata dal metodo. Supponiamo che

(H1) J sia Lipschitziana su L, con costante di Lipschitz $2\gamma_L$;

(H2) $\|J(x)\|$ sia limitata su L e $\chi_J = \sup_{x \in L} \|J(x)\|$.

Nel seguito mostreremo che il metodo TRESNORE è ben definito e converge globalmente. Se inoltre la successione $\{x_k\}$ generata converge ad una soluzione x^* di (3.1) tale che $J(x^*)$ ha rango massimo per colonne, allora la convergenza asintotica è quadratica.

Iniziamo l'analisi teorica del metodo TRESNORE studiando le proprietà dei passi $p_{tr}(\Delta_k)$, $\hat{p}(\Delta_k)$ e $\bar{p}_{tr}(\Delta_k)$.

Lemma 3.2.1 *Siano $x_k \in int(\Omega)$, $p_{tr}(\Delta_k)$ soluzione del problema di trust-region (3.7), $\hat{p}(\Delta_k)$ e $\bar{p}_{tr}(\Delta_k)$ definiti in (3.13) e in (3.14). Allora*

$$\|\bar{p}_{tr}(\Delta_k)\| < \|\hat{p}(\Delta_k)\| \leq \|p_{tr}(\Delta_k)\|, \tag{3.26}$$

$$\|D_k \bar{p}_{tr}(\Delta_k)\| < \Delta_k, \tag{3.27}$$

$$\|x_k + \hat{p}(\Delta_k) - z\| \leq \|x_k + p_{tr}(\Delta_k) - z\| \quad \forall \ z \in \Omega. \tag{3.28}$$

Dim. La funzione proiezione $P_\Omega(x)$ soddisfa la seguente proprietà di non estensività [3]:

$$\|P_\Omega(x) - P_\Omega(z)\| \leq \|x - z\|, \qquad \text{per ogni } x \in \mathbb{R}^n, \ z \in \Omega.$$

Ciò implica la (3.26) poiché

$$\|\bar{p}_{tr}(\Delta_k)\| = \|\alpha_k \hat{p}(\Delta_k)\| < \|\hat{p}(\Delta_k)\| = \|P_\Omega(x_k + p_{tr}(\Delta_k)) - x_k\| \leq \|p_{tr}(\Delta_k)\|.$$

La disuguaglianza (3.27) può essere ottenuta notando che $\|D_k p_{tr}(\Delta_k)\| \leq \Delta_k$ per costruzione e che dalla (3.15) deriva

$$\|D_k \bar{p}_{tr}(\Delta_k)\| = \sqrt{\sum_{i=1}^{n} \frac{(\bar{p}_{tr}(\Delta_k))_i^2}{|v_i(x_k)|}} < \sqrt{\sum_{i=1}^{n} \frac{(p_{tr}(\Delta_k))_i^2}{|v_i(x_k)|}} = \|D_k p_{tr}(\Delta_k)\|.$$

Anche la (3.28) deriva direttamente dalla proprietà della proiezione richiamata sopra. Infatti per ogni $z \in \Omega$

$$\|x_k + \hat{p}(\Delta_k) - z\| = \|x_k + P_\Omega(x_k + p_{tr}(\Delta_k)) - x_k - z\| \leq \|x_k + p_{tr}(\Delta_k) - z\|.$$

□

3.2.1 Terminazione finita del calcolo di una iterata

Un importante risultato nell'analisi del nostro metodo è che la condizione (3.24) valga se il raggio di trust-region è sufficientemente piccolo. Questo implica che l'istruzione "ripeti" al Passo 3 del metodo TRESNORE termina in un numero finito di ripetizioni e che di conseguenza il metodo TRESNORE è ben definito. Questo risultato può essere dimostrato procedendo come in [3]. Per completezza della presentazione, riportiamo le dimostrazioni sebbene non siano originali.

Definiamo la riduzione "attuale" $ared(p(\Delta_k))$ di f in $x_k + p(\Delta_k)$ rispetto ad x_k

$$ared(p(\Delta_k)) = f(x_k) - f(x_k + p(\Delta_k)),$$

e la riduzione "predetta" $pred(p(\Delta_k))$, i.e. la decrescita nel modello quadratico m_k in $x_k + p(\Delta_k)$ rispetto ad x_k

$$pred(p(\Delta_k)) = m_k(0) - m_k(p(\Delta_k)).$$

Il seguente lemma mostra come $pred(p(\Delta_k))$, cioè il progresso che può essere fatto all'iterata k-esima sul modello m_k, sia almeno proporzionale a $\|D_k^{-1}\nabla f_k\|$, quantità che rappresenta quanto x_k violi la condizione (3.3).

Lemma 3.2.2 *Supponiamo che l'Ipotesi (H1) sia soddisfatta e che* $\|\nabla f_k\| \neq 0$. *Allora il vettore* $p(\Delta_k)$ *soddisfa*

$$ared(p(\Delta_k)) \geq pred(p(\Delta_k)) - (\gamma_L\|F_k\| + \frac{1}{2}\gamma_L^2\|p(\Delta_k)\|^2)\|p(\Delta_k)\|^2, \tag{3.29}$$

$$pred(p(\Delta_k)) \geq \frac{\beta_1}{2}\|D_k^{-1}\nabla f_k\| \min\{\Delta_k, \frac{\|D_k^{-1}\nabla f_k\|}{\|D_k^{-1}J_k^T J_k D_k^{-1}\|}, \frac{\theta\|D_k^{-1}\nabla f_k\|}{\|\nabla f_k\|_\infty}\}, \tag{3.30}$$

dove θ *è la costante usata in (3.21).*

Dim. Dimostriamo innanzitutto la (3.29). Dal Teorema A.1 del Valor Medio segue che

$$\|F(x_k + p(\Delta_k))\|^2 = \|F_k + J_k p(\Delta_k) + w(x_k, p_k)\|^2,$$

dove

$$w(x_k, p_k) = \int_0^1 (J(x_k + tp(\Delta_k)) - J(x_k))\, p(\Delta_k)\, dt.$$

Quindi, abbiamo che

$$\begin{aligned} |m_k(p(\Delta_k)) - f(x_k + p(\Delta_k))| &= \frac{1}{2}\Big|\, \|F_k + J_k p(\Delta_k)\|^2 - \|F(x_k + p(\Delta_k))\|^2 \,\Big| \\ &\leq \|F_k + J_k p(\Delta_k)\|\|w(x_k, p_k)\| \\ &\quad + \frac{1}{2}\|w(x_k, p_k)\|^2. \end{aligned} \tag{3.31}$$

Dall'Ipotesi (H1), otteniamo

$$\|w(x_k, p_k)\| \leq \gamma_L\|p(\Delta_k)\|^2,$$

e da $m_k(p(\Delta_k)) \leq m_k(0)$ ricaviamo $\|F_k + J_k p(\Delta_k)\| \leq \|F_k\|$. Allora la disuguaglianza (3.31) diventa

$$|m_k(p(\Delta_k)) - f(x_k + p(\Delta_k))| \leq (\gamma_L \|F_k\| + \frac{1}{2}\gamma_L^2 \|p(\Delta_k)\|^2)\|p(\Delta_k)\|^2.$$

Poiché $m_k(0) = f(x_k)$, da questa disuguaglianza segue la (3.29), in quanto

$$\begin{aligned} ared(p(\Delta_k)) &= f(x_k) - m_k(p(\Delta_k)) + m_k(p(\Delta_k)) - f(x_k + p(\Delta_k)) \\ &\geq pred(p(\Delta_k)) - |m_k(p(\Delta_k)) - f(x_k + p(\Delta_k))| \\ &\geq pred(p(\Delta_k)) - (\gamma_L \|F_k\| + \frac{1}{2}\gamma_L^2 \|p(\Delta_k)\|^2)\|p(\Delta_k)\|^2. \end{aligned}$$

Per dimostrare la (3.30), determiniamo una limitazione inferiore per $pred(p_c(\Delta_k))$. Ricordiamo che $p_c(\Delta_k) = \tau_k d_k$, dove d_k e τ_k sono definiti in (3.6) e (3.21).
Se $p_c(\Delta_k) = \hat{\tau}_k d_k$ e da (3.17) si ha $\hat{\tau}_k = \dfrac{\|D_k^{-1}\nabla f_k\|^2}{\|J_k D_k^{-2}\nabla f_k\|^2}$, allora

$$\begin{aligned} pred(p_c(\Delta_k)) &= \hat{\tau}_k \|D_k^{-1}\nabla f_k\|^2 - \frac{1}{2}\hat{\tau}_k^2 \|J_k D_k^{-2}\nabla f_k\|^2 = \frac{1}{2}\frac{\|D_k^{-1}\nabla f_k\|^4}{\|J_k D_k^{-2}\nabla f_k\|^2} \\ &= \frac{1}{2}\frac{\|D_k^{-1}\nabla f_k\|^4}{(D_k^{-1}\nabla f_k)^T (D_k^{-1} J_k^T J_k D_k^{-1})(D_k^{-1}\nabla f_k)} \\ &\geq \frac{1}{2}\frac{\|D_k^{-1}\nabla f_k\|^2}{\|D_k^{-1} J_k^T J_k D_k^{-1}\|}. \end{aligned} \tag{3.32}$$

Se invece $p_c(\Delta_k) = \hat{\tau}_k d_k$ con $\hat{\tau}_k = \Delta_k / \|D_k^{-1}\nabla f_k\|$, allora $\hat{\tau}_k \leq \dfrac{\|D_k^{-1}\nabla f_k\|^2}{\|J_k D_k^{-2}\nabla f_k\|^2}$ e si ottiene

$$\begin{aligned} pred(p_c(\Delta_k)) &= \hat{\tau}_k (\|D_k^{-1}\nabla f_k\|^2 - \frac{1}{2}\hat{\tau}_k \|J_k D_k^{-2}\nabla f_k\|^2) \\ &\geq \frac{1}{2}\hat{\tau}_k \|D_k^{-1}\nabla f_k\|^2 = \frac{1}{2}\Delta_k \|D_k^{-1}\nabla f_k\|. \end{aligned} \tag{3.33}$$

Infine, consideriamo il caso in cui $\tau_k = \theta\lambda_k$ con λ_k definito in (3.19). Il passo di Cauchy generalizzato ha la forma $p_c(\Delta_k) = \theta\lambda_k d_k$ e $\theta\lambda_k \leq \hat{\tau}_k \leq \dfrac{\|D_k^{-1}\nabla f_k\|^2}{\|J_k D_k^{-2}\nabla f_k\|^2}$ per costruzione. Inoltre, da [9] sappiamo che $\lambda_k \geq \frac{1}{\|\nabla f_k\|_\infty}$. Allora otteniamo

$$\begin{aligned} pred(p_c(\Delta_k)) &= \theta\lambda_k \|D_k^{-1}\nabla f_k\|^2 - \frac{1}{2}(\theta\lambda_k)^2 \|J_k D_k^{-2}\nabla f_k\|^2 \\ &\geq \frac{1}{2}\theta\lambda_k \|D_k^{-1}\nabla f_k\|^2 \geq \frac{1}{2}\theta\frac{\|D_k^{-1}\nabla f_k\|^2}{\|\nabla f_k\|_\infty}. \end{aligned} \tag{3.34}$$

Poiché $pred(p(\Delta_k)) \geq \beta_1 pred(p_c(\Delta_k))$, da (3.32), (3.33) e (3.34) segue la (3.30). $\square$

Mostriamo ora che la condizione (3.24) è soddisfatta in un numero finito di tentativi.

Lemma 3.2.3 *Supponiamo che l'Ipotesi (H1) sia soddisfatta e che* $\|\nabla f_k\| \neq 0$. *Allora il Passo 3 del metodo TRESNORE termina in un numero finito di ripetizioni.*

Dim. Prima di tutto, troviamo una limitazione superiore per $\|p(\Delta_k)\|$. Poiché $\|D_k p_{tr}(\Delta_k)\| \leq \Delta_k$, $\|p_{tr}(\Delta_k)\| \leq \|D_k^{-1}\|\Delta_k$ e da (3.27) otteniamo $\|\bar{p}_{tr}(\Delta_k)\| \leq \|D_k^{-1}\|\Delta_k$. In modo analogo $\|p_c(\Delta_k)\| \leq \|D_k^{-1}\|\Delta_k$, in quanto $\|p_c(\Delta_k)\| \leq \|p^C(\Delta_k)\|$, dove $p^C(\Delta_k)$ è il punto di Cauchy definito in (3.18), e per definizione $\|D_k p^C(\Delta_k)\| \leq \Delta_k$. Di conseguenza, da (3.23) possiamo concludere che

$$\|p(\Delta_k)\| \leq \|D_k^{-1}\|\Delta_k. \tag{3.35}$$

Per ipotesi $\|D_k^{-1}\nabla f_k\| \neq 0$. Quindi, ponendo

$$\Delta_k \leq \min\{\frac{\|D_k^{-1}\nabla f_k\|}{\|D_k^{-1}J_k^T J_k D_k^{-1}\|}, \frac{\theta\|D_k^{-1}\nabla f_k\|}{\|\nabla f_k\|_\infty}\},$$

(3.30) implica $\Delta_k \leq \tilde{C}_k\, pred(p(\Delta_k))$, con $\tilde{C}_k = 2/(\beta_1\|D_k^{-1}\nabla f_k\|)$ e da (3.35) otteniamo

$$\|p(\Delta_k)\| \leq \|D_k^{-1}\|\, \tilde{C}_k\, pred(p(\Delta_k)). \tag{3.36}$$

Si osservi che (3.29) e (3.35) implicano

$$\begin{aligned} ared(p(\Delta_k)) \quad \geq \quad & pred(p(\Delta_k)) - (\,\gamma_L\,\|F_k\|\,\|D_k^{-1}\|\Delta_k + \\ & (\gamma_L^2/2)\,\|D_k^{-1}\|^3\Delta_k^3\,)\|p(\Delta_k)\|, \end{aligned}$$

e usando (3.36) si ricava

$$ared(p(\Delta_k)) \geq g_k(\Delta_k)\, pred(p(\Delta_k)), \tag{3.37}$$

dove

$$g_k(\Delta_k) = 1 - \|D_k^{-1}\|\tilde{C}_k(\gamma_L\|F_k\|\|D_k^{-1}\|\Delta_k + (\gamma_L^2/2)\,\|D_k^{-1}\|^3\Delta_k^3). \tag{3.38}$$

Poichè $g_k(0) = 1$, $\beta_2 \in (0,1)$, e $g_k(\Delta_k)$ è continua, esiste $\tilde{\Delta}_k > 0$ tale che $g_k(\Delta_k) \geq \beta_2$ quando $\Delta_k \leq \tilde{\Delta}_k$. Quindi, se $\Delta_k \leq \min\{\tilde{\Delta}_k, \frac{\|D_k^{-1}\nabla f_k\|}{\|D_k^{-1}J_k^T J_k D_k^{-1}\|}, \frac{\theta\|D_k^{-1}\nabla f_k\|}{\|\nabla f_k\|_\infty}\}$ otteniamo

$$ared(p(\Delta_k)) \geq \beta_2\, pred(p(\Delta_k)).$$

Così la condizione (3.24) è soddisfatta e il Passo 3 del metodo TRESNORE termina o con Δ_k uguale al raggio iniziale, $\Delta_k = \bar{\Delta}_k$, o con Δ_k tale che

$$\Delta_k \geq \min\{\delta\tilde{\Delta}_k, \frac{\|D_k^{-1}\nabla f_k\|}{\|D_k^{-1}J_k^T J_k D_k^{-1}\|}, \frac{\theta\|D_k^{-1}\nabla f_k\|}{\|\nabla f_k\|_\infty}\}.$$

□

3.2.2 Proprietà di convergenza globale

Dimostriamo che indipendentemente da x_0, la successione $\{x_k\}$ converge ad un punto stazionario x^* del problema di minimo vincolato (3.2).

Teorema 3.2.1 *Supponiamo che le Ipotesi (H1) e (H2) siano soddisfatte. Se la successione $\{x_k\}$ generata dal metodo TRESNORE è limitata, allora tutti i punti limite di $\{x_k\}$ sono punti stazionari per il problema (3.2) i.e.*

$$\lim_{k\to\infty} \|D_k^{-1}\nabla f_k\| = 0.$$

Inoltre se $\{x_k\}$ ammette un punto limite x^* tale che $F(x^*) = 0$, allora tutti i punti limite di $\{x_k\}$ sono soluzione del problema (3.1).

Dim. Poiché $\{x_k\}$ è limitata, esiste una costante $\chi_D > 0$ tale che $\|D(x)^{-1}\| < \chi_D$, $x \in L$ ed L definito in (3.25). Per l'Ipotesi (H1), $\nabla f_k = J(x)^T F(x)$ è lipschitziana in L con costante $2\gamma_L\tilde{\gamma} + \tilde{\gamma}^2$ dove $\tilde{\gamma} = \max\{\sup_{x\in L} f(x), \sup_{x\in L} \|J(x)\|\}$ (vedi [29]). Inoltre dall'Ipotesi (H2), segue che esiste uno scalare positivo χ_g tale che $\|\nabla f_k\|_\infty < \chi_g$ per ogni $x \in L$.

Proviamo che

$$\liminf_{k\to\infty} \|D_k^{-1}\nabla f_k\| = 0. \tag{3.39}$$

Procediamo per assurdo. Supponiamo che esista un $\epsilon > 0$ tale che

$$\liminf_{k\to\infty} \|D_k^{-1}\nabla f_k\| > \epsilon.$$

Allora esiste $\bar{k}$ tale che $\|D_k^{-1}\nabla f_k\| > \epsilon$ se $k > \bar{k}$. Assumiamo $k > \bar{k}$.

La successione $\{f_k\}$ è monotona decrescente e inferiormente limitata. Quindi, è convergente e $f_k - f_{k+1} \to 0$ per $k \to +\infty$. Per costruzione, ad ogni iterazione (3.24) è soddisfatta, i.e.

$$f_k - f_{k+1} \geq \beta_2\ pred(p(\Delta_k)).$$

Allora, da (3.30)

$$\begin{aligned} f_k - f_{k+1} &\geq \frac{\beta_1\beta_2}{2}\|D_k^{-1}\nabla f_k\| \min\{\Delta_k, \frac{\|D_k^{-1}\nabla f_k\|}{\|D_k^{-1}J_k^T J_k D_k^{-1}\|^2}, \frac{\theta\|D_k^{-1}\nabla f_k\|}{\|\nabla f_k\|_\infty}\}, \\ &\geq \frac{\beta_1\,\beta_2}{2}\,\epsilon \min\{\Delta_k, \frac{\epsilon}{\chi_D^2\chi_J^2}, \frac{\theta\epsilon}{\chi_g}\}, \end{aligned} \tag{3.40}$$

con χ_J data nell'Ipotesi (H2). Quindi, $f_k - f_{k+1} \to 0$ implica $\Delta_k \to 0$ e che esiste $\hat{k} > \bar{k}$ tale che $\Delta_k \leq \min\{\epsilon/(\chi_D^2\chi_J^2), \theta\epsilon/\chi_g\}$ se $k \geq \hat{k}$. Assumiamo $k > \hat{k}$.

Usando (3.37)-(3.38) e osservando che $\|D_k^{-1}\| < \chi_D$, $\tilde{C}_k = 2/(\beta_1\|D_k^{-1}\nabla f_k\|) < 2/(\beta_1\epsilon)$, otteniamo la seguente disuguaglianza

$$ared(p_k(\Delta_k)) \geq g(\Delta_k)\ pred(p_k(\Delta_k)),$$

dove

$$g(\Delta_k) = 1 - \chi_D \frac{2}{\beta_1 \epsilon}(\gamma_L \, \|F_0\| \, \chi_D \Delta_k + (\gamma_L^2/2) \, \chi_D^3 \Delta_k^3).$$

Poiché $g(0) = 1$, g è continua e $\beta_2 \in (0,1)$, esiste $\tilde{\Delta} > 0$, indipendente da k, tale che $g(\Delta_k) \geq \beta_2$ quando $\Delta_k \leq \tilde{\Delta}$. Di conseguenza, vale $ared(p_k(\Delta_k)) \geq \beta_2 \; pred(p_k(\Delta_k))$ per $\Delta_k \leq \tilde{\Delta}$.

Procedendo come nel Lemma 3.2.3 è facile vedere che, al termine del Passo 3, abbiamo o $\Delta_k = \bar{\Delta}_k \geq \Delta_{min}$ oppure

$$\Delta_k \geq \min\{\delta\tilde{\Delta}, \frac{\epsilon}{\chi_D^2 \chi_J^2}, \frac{\theta\epsilon}{\chi_g}\}.$$

Quindi, Δ_k è uniformemente limitato da zero. Questa è una contraddizione e dobbiamo dunque avere $\liminf_{k\to\infty} \|D_k^{-1}\nabla f_k\| = 0$.

Proviamo ora che $\lim_{k\to\infty} \|D_k^{-1}\nabla f_k\| = 0$. Per assurdo, supponiamo che esista una successione $\{m_i\}$ tale che $\|D_{m_i}^{-1}\nabla f_{m_i}\| \geq \epsilon_1$ per qualche $\epsilon_1 \in (0,1)$. Poiché abbiamo dimostrato (3.39), allora per ogni $\epsilon_2 \in (0, \epsilon_1)$ esiste una sottosuccessione di $\{m_i\}$, senza perdere di generalità considereremo l'intera successione, e una successione $\{l_i\}$ tale che

$$\|D_k^{-1}\nabla f_k\| \geq \epsilon_2, \qquad m_i \leq k < l_i \qquad \|D_{l_i}^{-1}\nabla f_{l_i}\| < \epsilon_2. \tag{3.41}$$

Quindi, da (3.40) segue

$$f_k - f_{k+1} \geq \frac{1}{2}\beta_1 \, \beta_2 \, \epsilon_2 \min\{\Delta_k, \frac{\epsilon_2}{\chi_D^2 \chi_J^2}, \frac{\theta \, \epsilon_2}{\chi_g}\}, \quad m_i \leq k < l_i,$$

e da (3.35) abbiamo $\|x_{k+1} - x_k\| \leq \chi_D \, \Delta_k$. Possiamo dunque concludere che

$$f_k - f_{k+1} \geq \frac{1}{2}\beta_1 \, \beta_2 \, \epsilon_2 \min\{\frac{\|x_{k+1} - x_k\|}{\chi_D}, \frac{\epsilon_2}{\chi_D^2 \chi_J^2}, \frac{\theta\epsilon_2}{\chi_g}\}, \quad m_i \leq k < l_i. \tag{3.42}$$

Poiché, $f_k - f_{k+1} \to 0$, da (3.42) segue che

$$f_k - f_{k+1} \geq \epsilon_3 \|x_{k+1} - x_k\|, \quad m_i \leq k < l_i, \tag{3.43}$$

per i sufficientemente grande e $\epsilon_3 = \frac{1}{2}\beta_1 \, \beta_2 \, \epsilon_2/\chi_D$. Quindi, usando la disuguaglianza triangolare otteniamo

$$f_{m_i} - f_{k_i} \geq \epsilon_3 \|x_{m_i} - x_{k_i}\|, \quad m_i \leq k_i \leq l_i, \tag{3.44}$$

e possiamo concludere che $\|x_{m_i} - x_{k_i}\|$ tende a zero. Inoltre, per la lipschitzianità di ∇f e poiché $\|x_{m_i} - x_{k_i}\|$ tende a zero, segue

$$\|\nabla f_{m_i} - \nabla f_{k_i}\| \leq \epsilon_2, \tag{3.45}$$

per i sufficientemente grande.

Senza perdita di generalità, supponiamo che l'intera successione $\{x_{l_i}\}$ converga ad un punto, diciamo x^*. Da (3.44) abbiamo anche che $\{x_{m_i}\}$ converge a x^*.

Se $(\nabla f(x^*))_j \neq 0$ per qualche $j \in \{1, \ldots, n\}$, allora (3.5) implica $|(v_{m_i})_j - (v_{l_i})_j| \leq |(x_{m_i})_j - (x_{l_i})_j|$ per i sufficientemente grande. Di conseguenza $\|(D_{m_i}^{-1} - D_{l_i}^{-1})\nabla f_{l_i}\| \to 0$ e quindi

$$\|(D_{m_i}^{-1} - D_{l_i}^{-1})\nabla f_{l_i}\| \leq \epsilon_2, \tag{3.46}$$

per i sufficientemente grande. Infine, da $\|D_{m_i}^{-1}\nabla f_{m_i}\| \geq \epsilon_1$, (3.41), (3.45), (3.46) e

$$\begin{aligned} \|D_{m_i}^{-1}\nabla f_{m_i}\| \quad \leq \quad & \|D_{m_i}^{-1}\| \, \|\nabla f_{m_i} - \nabla f_{l_i}\| + \\ & \|(D_{m_i}^{-1} - D_{l_i}^{-1})\nabla f_{l_i}\| + \|D_{l_i}^{-1}\nabla f_{l_i}\|, \end{aligned}$$

otteniamo

$$\epsilon_1 \leq (\chi_D + 2)\, \epsilon_2,$$

i.e. una contraddizione dal momento che $\epsilon_2 \in (0, \epsilon_1)$ può essere preso arbitrariamente piccolo.

Dimostriamo ora che se $\{x_k\}$ ammette un punto limite x^* tale che $F(x^*) = 0$, allora tutti i punti limite di $\{x_k\}$ sono soluzione del problema (3.1). Abbiamo visto che la successione $\{\|F_k\|\}$ è monotona decrescente, inferiormente limitata e quindi convergente. L'ipotesi $F(x^*) = 0$ implica $\lim_{k\to+\infty} \|F_k\| = 0$ e quindi la tesi. $\square$

3.2.3 Proprietà di convergenza locale

In questa sezione, studiamo le proprietà di convergenza locale del metodo TRESNORE e ci limitiamo al caso di sistemi quadrati o sovradeterminati che ammettono una soluzione x^* tale che $J(x^*)$ ha rango massimo. Iniziamo il nostro studio con alcuni risultati intermedi.

Lemma 3.2.4 *Siano soddisfatte le Ipotesi (H1) e (H2) e sia $x^* \in \Omega$ un punto limite della successione $\{x_k\}$ generata dal metodo TRESNORE tale che $F(x^*) = 0$ e $J(x^*)$ abbia rango n. Allora esistono $K, \rho > 0$ tali che, se $x \in B_\rho(x^*)$, abbiamo $x \in L$ e*

$$\|x - x^*\| \leq K\, \|F(x)\| \qquad \forall\, x \in B_\rho(x^*), \tag{3.47}$$

$$\|F(x)\| \leq K\, \|x - x^*\| \qquad \forall\, x \in B_\rho(x^*), \tag{3.48}$$

$$\|F(x) - F(z) - J(z)(x - z)\| \leq K\|x - z\|^2 \qquad \forall\, x, z \in B_\rho(x^*), \tag{3.49}$$

$$\textit{rango di } J(x)^+ = n \quad e \quad \|J(x)^+\| \leq K \qquad \forall\, x \in B_\rho(x^*), \tag{3.50}$$

$$\|D(x)(x - x^*)\| \leq \Delta_{min} \quad \forall\, x \in B_\rho(x^*) \cap int(\Omega). \tag{3.51}$$

Dim. Sia ρ_1 tale che $0 < \rho_1 < r$, dove $r > 0$ è lo scalare definito in (3.25). Mostriamo che se $\|x - x^*\| \leq \rho_1$, allora $x \in L$. Infatti, dal momento che x^* è un punto limite di $\{x_k\}$, esiste x_k tale che $\|x_k - x^*\| \leq r - \rho_1$. Quindi, $\|x - x_k\| \leq \|x - x^*\| + \|x_k - x^*\| \leq \rho_1 + r - \rho_1 = r$, i.e. $x \in L$.

Nel seguito dimostreremo che per $K = \max\{2\|J(x^*)^+\|, \chi_J, \gamma_L\}$, le disuguaglianze (3.47)-(3.49) sono soddisfatte. Ricordiamo che per il Teorema A.1 del Valor Medio abbiamo

$$F(x) = F(x^*) + \int_0^1 J(x^* + t(x - x^*))\, (x - x^*)\, dt\,. \tag{3.52}$$

Quindi, otteniamo

$$J(x^*)^+F(x) = (x - x^*) - \int_0^1 [I - J(x^*)^+J(x^* + t(x - x^*))]\,(x - x^*)\,dt\,. \qquad (3.53)$$

Sia $\rho_2 < \min\{\rho_1, \frac{1}{4\gamma_L\|J(x^*)^+\|}\}$. Osserviamo che dal Lemma B.2 si ha $J(x^*)^+\,J(x^*) = I$. Allora se $x \in B_{\rho_2}(x^*)$, usando l'Ipotesi (H1), otteniamo

$$\begin{aligned} \|I - J(x^*)^+J(x)\| &= \|J(x^*)^+(J(x^*) - J(x))\| \le \|J(x^*)^+\|\,\|J(x^*) - J(x)\| \\ &\le 2\gamma_L\|J(x^*)^+\|\,\|x - x^*\| \le 2\gamma_L\|J(x^*)^+\|\,\rho_2 < \frac{1}{2}, \end{aligned} \qquad (3.54)$$

e dalla (3.53)

$$\begin{aligned} \|J(x^*)^+\|\,\|F(x)\| &\ge \|J(x^*)^+F(x)\| \\ &\ge \|x - x^*\| - \int_0^1 \|I - J(x^*)^+J(x^* + t(x - x^*))\|\,\|x - x^*\|\,dt \\ &> \|x - x^*\| - \int_0^1 \frac{1}{2}\|x - x^*\|\,dt = \frac{1}{2}\|x - x^*\|. \end{aligned}$$

Pertanto la (3.47) è vera in $B_{\rho_2}(x^*)$.

Passando alle norme in (3.52) e usando l'Ipotesi (H2) ricaviamo

$$\|F(x)\| \le \int_0^1 \|J(x^* + t(x - x^*))\,\|\,\|x - x^*\,\|dt \le \chi_J\,\|x - x^*\|, \qquad (3.55)$$

e quindi la (3.48) è vera per ogni $x \in B_{\rho_2}(x^*)$.

Per dimostrare la (3.49), applichiamo il Teorema A.1 del Valor Medio. Per ogni $x, z \in B_{\rho_2}(x^*)$, usando l'Ipotesi (H1) e la definizione di K, si ottiene la tesi come segue

$$\begin{aligned} \Big\|F(x) - F(z) - J(z)(x - z)\Big\| &= \Big\|\int_0^1 (J(z + t(x - z)) - J(z))\,(x - z)\,dt\Big\| \\ &\le \int_0^1 \|J(z + t(x - z)) - J(z)\|\,\|x - z\|\,dt \\ &\le \int_0^1 2\gamma_L\|x - z\|^2\,t\,dt = \gamma_L\|x - z\|^2. \end{aligned} \qquad (3.56)$$

Per provare la (3.50), posto $x \in B_{\rho_2}(x^*)$, dal Teorema B.3 e dall'Ipotesi (H1) otteniamo

$$|\sigma_n(J(x)) - \sigma_n(J(x^*))| \le \|J(x) - J(x^*)\| \le 2\gamma_L\|x - x^*\|,$$

dove $\sigma_n(J(x))$ e $\sigma_n(J(x^*))$ sono il più piccolo valore singolare di $J(x)$ e di $J(x^*)$ rispettivamente. Notiamo che $\sigma_n(J(x^*)) > 0$ dal momento che per ipotesi $J(x^*)$ ha rango massimo. Quindi, se $x \in B_{\rho_3}(x^*)$, con $\rho_3 < \min\{\rho_2, \frac{\sigma_n(J(x^*))}{2\gamma_L}\}$, segue che

$$\sigma_n(J(x)) \ge \sigma_n(J(x^*)) - 2\gamma_L\|x - x^*\| > 0,$$

cioè $J(x)$ ha rango massimo per ogni $x \in B_{\rho_3}(x^*)$. Posto allora $\mu = \frac{1}{(\sigma_n(J(x^*))-2\gamma_L\rho_3)}$, dal Lemma B.1 otteniamo che per ogni $x \in B_{\rho_3}(x^*)$

$$\|J(x)^+\| = \frac{1}{\sigma_n(J(x))} < \mu. \tag{3.57}$$

Ponendo infine $K = \max\{2\|J(x^*)^+\|, \chi_J, \gamma_L, \mu\}$ dalla (3.57) segue la (3.50).

Dimostriamo ora la disuguaglianza (3.51). Sia I_* l'insieme degli indici inattivi in x^*, i.e. $I_* = \{i \mid x_i^* \in (l_i, u_i)\}$ e A_* sia l'insieme degli indici attivi in x^*, cioè $A_* = \{1, 2, \ldots, n\} \backslash I_*$. Ricordando la (3.4), $\|D(x)(x - x^*)\|$ può essere scritta come

$$\|D(x)(x - x^*)\| = \sqrt{\sum_{i \in I^*} \frac{(x - x^*)_i^2}{|v_i(x)|} + \sum_{i \in A^*} \frac{(x - x^*)_i^2}{|v_i(x)|}}.$$

Sia $\rho_4 < \rho_3$ tale che per $x \in B_{\rho_4}(x^*) \cap int(\Omega)$ ed $i \in I_*$ esiste $\delta > 0$ che soddisfa $(x - l)_i > \delta$ e $(u - x)_i > \delta$. Allora, per $x \in B_{\rho_4}(x^*) \cap int(\Omega)$ abbiamo

$$\frac{(x - x^*)_i^2}{|v_i(x)|} \leq \frac{|(x - x^*)_i|^2}{\delta} \qquad \text{per } i \in I^*, \tag{3.58}$$

$$\frac{(x - x^*)_i^2}{|v_i(x)|} = |(x - x^*)_i| \qquad \text{per } i \in A^*. \tag{3.59}$$

La disuguaglianza (3.58) vale banalmente in quanto $|v_i(x)| > \delta$. Per quanto riguarda (3.59), per semplicità supponiamo $x_i^* = l_i$. Sotto la nostra ipotesi, abbiamo che il punto x^* è un minimo di f su Ω. Quindi, per x sufficientemente vicino a x^*, abbiamo che $(\nabla f)_i \geq 0$ e per la definizione di $D(x)$ segue che $v_i(x) = (x - x^*)_i$. Allora riducendo se necessario il valore di ρ_4 vale la (3.59). Concludendo esiste $\rho < \rho_4$ sufficientemente piccolo e tale che la tesi è soddisfatta. □

Teorema 3.2.2 *Siano soddisfatte le Ipotesi (H1) e (H2). Supponiamo che x^* sia un punto limite della successione $\{x_k\}$ generata dal metodo TRESNORE.*

(i) Se $x^* \in int(\Omega)$ e $J(x^*)$ ha rango m, allora $\|F_k\| \to 0$ e tutti i punti di acccumulazione di $\{x_k\}$ risolvono il problema (3.1).

(ii) Se $F(x^*) = 0$ e $J(x^*)$ ha rango n, allora $\{x_k\}$ converge a x^*.

Dim. Dimostriamo innanzitutto la parte (i). Sia $\rho \in (0, 2]$ sufficientemente piccolo tale che $B_\rho(x^*) \subset int(\Omega)$ e per $x \in B_\rho(x^*)$, $\|J(x)^+\| \leq 2\|J(x^*)^+\|$. Inoltre, sia $\{x_{k_j}\}$ una sottosuccessione di $\{x_k\}$ tale che $x_{k_j} \to x^*$ e j_0 sia l'indice tale che per $k_j > k_{j_0}$, $\{x_{k_j}\} \in B_{\rho/2}(x^*)$. Sia $k_j > k_{j_0}$. Allora $|l_i - (x_{k_j})_i| > \rho/2$ e $|u_i - (x_{k_j})_i| > \rho/2$ per $i = 1, \ldots, n$. Quindi, $\|D_{k_j}\| \leq \sqrt{2/\rho}$ dove $\sqrt{2/\rho} \geq 1$. Inoltre, la successione $\{\|F_k\|\}$

è monotona decrescente e quindi è convergente. Allora, da $\|D_{k_j}^{-1}\nabla f_{k_j}\| \to 0$ e dal Lemma B.3, si ottiene

$$\|D_{k_j}^{-1}\nabla f_{k_j}\| \geq \frac{\|\nabla f_{k_j}\|}{\|D_{k_j}\|} \geq \frac{\|F_{k_j}\|}{2\|J(x^*)^+\|\sqrt{2/\rho}},$$

e la tesi segue.

Proviamo ora la parte (ii). Osserviamo che per la (3.47) esiste un intorno di x^* nel quale $\|F(x)\| > 0$ se $x \neq x^*$. Questo implica che x^* è una soluzione isolata di (3.1). Inoltre, poiché la successione $\{\|F_k\|\}$ è monotona decrescente, essa è convergente. L'ipotesi $F(x^*) = 0$ implica $\|F_k\| \to 0$ e ogni punto limite di $\{x_k\}$ è una soluzione (3.1). Essendo x^* una soluzione isolata di (3.1), x^* è un punto limite isolato della successione $\{x_k\}$.

Sia ρ come nel Lemma 3.2.4, $\{x_{k_j}\}$ una sottosuccessione tale che $x_{k_j} \to x^*$ e j_0 l'indice tale che $x_{k_j} \in B_\rho(x^*) \cap int(\Omega)$ per $k_j \geq k_{j_0}$. Notiamo che $\lim_{k_j\to\infty} \|\nabla f_{k_j}\| = 0$. Sia $k_j \geq k_{j_0}$. Per provare la tesi, studiamo il comportamento asintotico di $p_c(\Delta_{k_j})$ e $\bar{p}_{tr}(\Delta_{k_j})$. Dalla (3.21), $\tau_{k_j} \leq \dfrac{\|D_{k_j}^{-1}\nabla f_{k_j}\|^2}{\|J_{k_j}D_{k_j}^{-2}\nabla f_{k_j}\|^2}$, e dal Lemma B.3, seguono le disuguaglianze

$$\begin{aligned}
\|p_c(\Delta_{k_j})\| &\leq \frac{\|D_{k_j}^{-1}\nabla f_{k_j}\|^2}{\|J_{k_j}D_{k_j}^{-2}\nabla f_{k_j}\|^2}\|D_{k_j}^{-2}\nabla f_{k_j}\| \\
&\leq \frac{\|D_{k_j}^{-1}\nabla f_{k_j}\|^2}{\|J_{k_j}D_{k_j}^{-2}\nabla f_{k_j}\|^2}\|J_{k_j}{}^+\|\|J_{k_j}D_{k_j}^{-2}\nabla f_{k_j}\| \\
&\leq \frac{\|D_{k_j}^{-1}\nabla f_{k_j}\|^2}{\|D_{k_j}^{-2}\nabla f_{k_j}\|}\|J_{k_j}{}^+\|^2 = \frac{\nabla f_{k_j}^T(D_{k_j}^{-2}\nabla f_{k_j})}{\|D_{k_j}^{-2}\nabla f_{k_j}\|}\|J_{k_j}{}^+\|^2.
\end{aligned}$$

Di conseguenza, la (3.50) implica

$$\|p_c(\Delta_{k_j})\| \leq \|\nabla f_{k_j}\|\|J_{k_j}{}^+\|^2 \leq K^2\|\nabla f_{k_j}\|. \tag{3.60}$$

Poichè $\lim_{k_j\to\infty} \|\nabla f_{k_j}\| = 0$, possiamo concludere che $\lim_{k_j\to\infty} \|p_c(\Delta_{k_j})\| = 0$.

Per quanto riguarda $p_{tr}(\Delta_{k_j})$, $m_k(p_{tr}(\Delta_{k_j})) \leq m_k(0)$ i.e. $\|F_{k_j} + J_{k_j}p_{tr}(\Delta_{k_j})\| \leq \|F_{k_j}\|$ per costruzione. Quindi, dalla (3.50)

$$\begin{aligned}
\|p_{tr}(\Delta_{k_j})\| &= \|J_{k_j}{}^+(-F_{k_j} + (F_{k_j} + J_{k_j}p_{tr}(\Delta_{k_j})))\| \\
&\leq \|J_{k_j}^+\|(\|F_{k_j}\| + \|F_{k_j} + J_{k_j}p_{tr}(\Delta_{k_j})\|) \leq 2\,K\,\|F_{k_j}\|.
\end{aligned}$$

Abbiamo dunque $\lim_{k_j\to\infty} \|p_{tr}(\Delta_{k_j})\| = 0$ e per la (3.26) $\lim_{k_j\to\infty} \|\bar{p}_{tr}(\Delta_{k_j})\| = 0$.

In conclusione, $\lim_{k_j\to\infty} \|p(\Delta_{k_j})\| \leq \lim_{k_j\to\infty}(\|\bar{p}_{tr}(\Delta_{k_j})\| + \|p_c(\Delta_{k_j})\|) = 0$ e per il Lemma A.1 possiamo concludere che $\{x_k\}$ converge a x^*. □

Nel lemma seguente mostriamo il comportamento asintotico del passo $\bar{p}_{tr}(\bar{\Delta}_k)$.

Lemma 3.2.5 *Siano soddisfatte le Ipotesi (H1) e (H2). Sia $x^* \in \Omega$ un punto limite della successione $\{x_k\}$ generata dal metodo TRESNORE tale che $F(x^*) = 0$ e $J(x^*)$ abbia rango n Allora, per ogni $x_k \in B_\rho(x^*) \cap int(\Omega)$, esiste una costante $\nu > 0$ tale che*

$$\|\bar{p}_{tr}(\bar{\Delta}_k)\| < \|p_{tr}(\bar{\Delta}_k)\| \leq \nu \|x_k - x^*\|, \tag{3.61}$$

dove $\bar{\Delta}_k$ è il raggio iniziale di trust-region alla k-esima iterazione.

Dim. Sia ρ lo scalare definito nel Lemma 3.2.4. Poiché $p_{tr}(\bar{\Delta}_k)$ risolve il problema di trust-region $\min_p\{m_k(p) : \quad \|D_k\, p\| \leq \bar{\Delta}_k\}$, esso risolve il seguente sistema lineare

$$({J_k}^T J_k + \lambda D_k^2) p_{tr}(\bar{\Delta}_k) = -\nabla f_k,$$

dove $\lambda \geq 0$ (vedi Teorema 3.1.1). Di conseguenza

$$p_{tr}(\bar{\Delta}_k)^T {J_k}^T J_k p_{tr}(\bar{\Delta}_k) + \lambda p_{tr}(\bar{\Delta}_k)^T D_k^2 p_{tr}(\bar{\Delta}_k) = -p_{tr}(\bar{\Delta}_k)^T \nabla f_k,$$

i.e.

$$\|J_k p_{tr}(\bar{\Delta}_k)\|^2 + \lambda p_{tr}(\bar{\Delta}_k)^T D_k^2 p_{tr}(\bar{\Delta}_k) = -p_{tr}(\bar{\Delta}_k)^T \nabla f_k. \tag{3.62}$$

Dal Lemma B.3 e dalla (3.62) abbiamo che

$$\frac{\|p_{tr}(\bar{\Delta}_k)\|^2}{\|{J_k}^+\|^2} \quad \leq \quad \|J_k p_{tr}(\bar{\Delta}_k)\|^2 + \lambda p_{tr}(\bar{\Delta}_k)^T D_k^2 p_{tr}(\bar{\Delta}_k) = -p_{tr}(\bar{\Delta}_k)^T \nabla f_k.$$

Da questo segue

$$\|p_{tr}(\bar{\Delta}_k)\|^2 \leq -\|{J_k}^+\|^2 p_{tr}(\bar{\Delta}_k)^T \nabla f_k \leq \|{J_k}^+\|^2 \|p_{tr}(\bar{\Delta}_k)\| \, \|J_k\| \, \|F_k\|.$$

Così per l'Ipotesi (H2) e da (3.26), (3.48) e (3.50) otteniamo

$$\|\bar{p}_{tr}(\bar{\Delta}_k)\| < \|p_{tr}(\bar{\Delta}_k)\| \leq K^3 \, \chi_J \, \|x_k - x^*\|. \tag{3.63}$$

Ponendo $\nu = K^3 \chi_J$, otteniamo la tesi. □

Lemma 3.2.6 *Siano soddisfatte le Ipotesi (H1) e (H2). Sia $x^* \in \Omega$ un punto limite della successione $\{x_k\}$ generata dal metodo TRESNORE tale che $F(x^*) = 0$ e $J(x^*)$ abbia rango n. Allora, per ogni $x_k \in B_\tau(x^*) \cap int(\Omega)$, esistono $\tau > 0$ e $\sigma_k > 0$ tali che*

$$\|F_k + J_k \, \bar{p}_{tr}(\bar{\Delta}_k)\| \leq \sigma_k \|x_k - x^*\|, \tag{3.64}$$

$$\|x_k + \bar{p}_{tr}(\bar{\Delta}_k) - x^*\| \leq \sigma_k \|x_k - x^*\|, \tag{3.65}$$

dove $\bar{\Delta}_k$ è il raggio iniziale di trust-region alla k-esima iterazione.

Dim. Sia $0 < \tau \leq \rho/(1+\nu)$, dove ρ e ν sono le costanti usate rispettivamente nei Lemma 3.2.4 e 3.2.5. Sia $x_k \in B_\tau(x^*) \cap int(\Omega)$. Dal Lemma 3.2.5 e dalla definizione di τ si ha che $x_k + p_{tr}(\bar{\Delta}_k)$ e $x_k + \bar{p}_{tr}(\bar{\Delta}_k)$ sono contenuti nella $B_\rho(x^*)$. Infatti

$$\|x_k + p_{tr}(\bar{\Delta}_k) - x^*\| \leq \|x_k - x^*\| + \|p_{tr}(\bar{\Delta}_k)\| \leq \|x_k - x^*\| + \nu\|x_k - x^*\| \leq \tau + \nu\tau \leq \rho$$

e analogamente per $x_k + \bar{p}_{tr}(\bar{\Delta}_k)$.

Si noti che la disuguaglianza (3.51) implica $\|D_k(x_k - x^*)\| \leq \bar{\Delta}_k$, ovvero che il passo $x^* - x_k$ è all'interno alla trust-region. Di conseguenza, poiché $p_{tr}(\bar{\Delta}_k)$ risolve $\min_p\{m_k(p) : \quad \|D_k p\| \leq \bar{\Delta}_k\}$, abbiamo $m_k(p_{tr}(\bar{\Delta}_k)) \leq m_k(x^* - x_k)$, i.e.

$$\|F_k + J_k p_{tr}(\bar{\Delta}_k)\|^2 \leq \|F_k + J_k(x^* - x_k)\|^2.$$

Dalla (3.49) otteniamo

$$\|F_k + J_k(x^* - x_k)\| = \|F_k + J_k(x^* - x_k) - F(x^*)\| \leq K\|x^* - x_k\|^2,$$

e quindi

$$\|F_k + J_k\, p_{tr}(\bar{\Delta}_k)\| \leq K\|x_k - x^*\|^2. \tag{3.66}$$

Per ogni $q \in \mathbb{R}^n$ tale che $x_k + q \in B_\rho(x^*)$ usando (3.47) e (3.49) otteniamo

$$\begin{aligned} \|x_k + q - x^*\| &\leq K\|F(x_k + q)\| \leq K(\|F(x_k + q) - F_k - J_k q\| + \|F_k + J_k q\|) \\ &\leq K(K\|q\|^2 + \|F_k + J_k q\|). \end{aligned} \tag{3.67}$$

Quindi, posto $q = p_{tr}(\bar{\Delta}_k)$ e usando la (3.61) e la (3.66), otteniamo

$$\begin{aligned} \|x_k + p_{tr}(\bar{\Delta}_k) - x^*\| &\leq K(K\|p_{tr}(\bar{\Delta}_k)\|^2 + \|F_k + J_k p_{tr}(\bar{\Delta}_k)\|) \\ &\leq K^2(\nu^2 + 1)\|x_k - x^*\|^2 \\ &= \delta_k\|x_k - x^*\|, \end{aligned} \tag{3.68}$$

con $\delta_k = K^2(\nu^2 + 1)\|x_k - x^*\|$.

Per provare la (3.64), consideriamo inizialmente $\hat{p}(\bar{\Delta}_k)$ e determiniamo una limitazione superiore per $\|F_k + J_k\hat{p}(\bar{\Delta}_k)\|$:

$$\begin{aligned} F_k + J_k\hat{p}(\bar{\Delta}_k) &= F(x_k + \hat{p}(\bar{\Delta}_k)) + J_k\hat{p}(\bar{\Delta}_k) + F_k - F(x_k + \hat{p}(\bar{\Delta}_k)) \\ &= F(x_k + \hat{p}(\bar{\Delta}_k)) - F(x^*) + \int_0^1 (J(x_k) - J(x_k + t\hat{p}(\bar{\Delta}_k)))\hat{p}(\bar{\Delta}_k)dt \\ &= \int_0^1 J(x^* + t(x_k + \hat{p}(\bar{\Delta}_k) - x^*))(x_k + \hat{p}(\bar{\Delta}_k) - x^*)dt \\ &\quad + \int_0^1 (J(x_k) - J(x_k + t\hat{p}(\bar{\Delta}_k)))\hat{p}(\bar{\Delta}_k)dt. \end{aligned}$$

Passando alle norme, le relazioni (3.26), (3.28), (3.61), (3.68) e le Ipotesi (H1) e (H2) implicano

$$\begin{aligned}
\|F_k + J_k\hat{p}(\bar{\Delta}_k)\| &\leq \chi_J\|x_k + \hat{p}(\bar{\Delta}_k) - x^*\| + \gamma_L\|\hat{p}(\bar{\Delta}_k)\|^2 \\
&\leq \chi_J\|x_k + p_{tr}(\bar{\Delta}_k) - x^*\| + \gamma_L\|p_{tr}(\bar{\Delta}_k)\|^2 \\
&\leq \chi_J\delta_k\|x_k - x^*\| + \gamma_L\nu^2\|x_k - x^*\|^2 \\
&= (\chi_J\delta_k + \gamma_L\nu^2\|x_k - x^*\|)\|x_k - x^*\|. \qquad (3.69)
\end{aligned}$$

Infine dalla (3.14), (3.48) e (3.69) otteniamo

$$\begin{aligned}
\|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\| &= \|F_k + \alpha_k J_k\hat{p}(\bar{\Delta}_k)\| \\
&\leq \alpha_k\|F_k + J_k\hat{p}(\bar{\Delta}_k)\| + (1-\alpha_k)\|F_k\| \\
&\leq (\chi_J\delta_k + \gamma_L\nu^2\|x_k - x^*\| + (1-\alpha_k)K)\|x_k - x^*\|. \qquad (3.70)
\end{aligned}$$

Poniamo $\psi_k = \chi_J\delta_k + \gamma_L\nu^2\|x_k - x^*\| + (1-\alpha_k)K$. Per provare la (3.65) sfruttiamo la (3.67), la (3.61) e la (3.70) nel seguente modo

$$\begin{aligned}
\|x_k + \bar{p}_{tr}(\bar{\Delta}_k) - x^*\| &\leq K(K\|\bar{p}_{tr}(\bar{\Delta}_k)\|^2 + \|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\|) \\
&\leq K(K\nu^2\|x_k - x^*\| + \psi_k)\|x_k - x^*\|. \qquad (3.71)
\end{aligned}$$

Concludendo, posto $\sigma_k = \max\{\psi_k, K(K\nu^2\|x_k - x^*\| + \psi_k)\}$ dalla (3.70) e dalla (3.71) seguono rispettivamente la (3.64) e la (3.65). □

Concludiamo questa sezione, dimostrando che, per opportune scelte degli scalari α_k dati in (3.14), la successione $\{x_k\}$ converge con velocità asintotica superlineare o quadratica. In particolare, il passo selezionato per k sufficientemente grande è $\bar{p}_{tr}(\bar{\Delta}_k)$, cioè il passo definito applicando (3.13) e (3.14) al passo di trust-region calcolato con raggio iniziale $\bar{\Delta}_k$.

Teorema 3.2.3 *Siano soddisfatte le Ipotesi (H1) e (H2). Sia x^* un punto limite della successione $\{x_k\}$ generata dal metodo TRESNORE tale che $F(x^*) = 0$ e $J(x^*)$ abbia rango n. Supponiamo inoltre che gli scalari α_k in (3.14) soddisfino $\alpha_k \to 1$ per $k \to \infty$. Allora per k sufficientemente grande, $x_{k+1} = x_k + \bar{p}_{tr}(\bar{\Delta}_k)$ e la successione $\{x_k\}$ converge a x^* superlinearmente, i.e.*

$$\|x_{k+1} - x^*\| = o(\|x_k - x^*\|).$$

Se inoltre $\alpha_k = 1 - O(\|F_k\|)$ per $k \to \infty$, la successione $\{x_k\}$ converge a x^* con convergenza quadratica, i.e.

$$\|x_{k+1} - x^*\| = O(\|x_k - x^*\|^2).$$

Dim. Dal Teorema 3.2.2 sappiamo che $\{x_k\}$ converge a x^*. Sia ζ tale che

$$\zeta < \min\left\{\tau, \frac{1-\beta_2}{2K^2\nu^2}\right\}, \qquad (3.72)$$

dove τ, ν e K sono gli scalari definiti nel Lemma 3.2.6, Lemma 3.2.5 e Lemma 3.2.4 rispettivamente. Sia k sufficientemente grande da avere $x_k \in B_\zeta(x^*) \cap int(\Omega)$ e

$$\sigma_k < \frac{1}{K} \min\{\sqrt{1-\beta_1}, \frac{1}{2}\}, \tag{3.73}$$

dove σ_k è dato nel Lemma 3.2.6. Questa condizione è soddisfatta per k sufficientemente grande dal momento che

$$\sigma_k = O(\|x_k - x^*\| + (1-\alpha_k)), \qquad k \to \infty. \tag{3.74}$$

Prima di tutto proviamo che $\bar{p}_{tr}(\bar{\Delta}_k)$ soddisfa (3.22), i.e. $p(\bar{\Delta}_k) = \bar{p}_{tr}(\bar{\Delta}_k)$. Da (3.22) e (3.64) segue

$$\begin{aligned} \rho_c(\bar{p}_{tr}(\bar{\Delta}_k)) &\geq \frac{m_k(0) - m_k(\bar{p}_{tr}(\bar{\Delta}_k))}{m_k(0)} \\ &= 1 - \frac{\|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\|^2}{\|F_k\|^2} \\ &\geq 1 - \frac{\sigma_k^2\|x_k - x^*\|^2}{\|F_k\|^2}. \end{aligned}$$

Questo implica, usando (3.47) e (3.73),

$$\rho_c(\bar{p}_{tr}(\bar{\Delta}_k)) \geq 1 - \frac{\sigma_k^2\|x_k - x^*\|^2 K^2}{\|x_k - x^*\|^2} \geq 1 - \sigma_k^2 K^2 > \beta_1. \tag{3.75}$$

Proviamo ora che $\rho_f(\bar{p}_{tr}(\bar{\Delta}_k)) \geq \beta_2$. Osserviamo che

$$\rho_f(\bar{p}_{tr}(\bar{\Delta}_k)) = 1 - \frac{\|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k))\|^2 - \|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\|^2}{\|F_k\|^2 - \|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\|^2}, \tag{3.76}$$

e che (3.72) implica $\zeta < 1/(2K^2\nu^2)$, mentre (3.73) implica $2\sigma_k < 1/K$.
Analizziamo separatamente il numeratore e il denominatore di (3.76). Osservando che

$$\begin{aligned} \|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k))\|^2 - \|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\|^2 &= \|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k)) - F_k - J_k\bar{p}_{tr}(\bar{\Delta}_k)\|^2 \\ &\quad +2(F(x_k + \bar{p}_{tr}(\bar{\Delta}_k)) - F_k - J_k\bar{p}_{tr}(\bar{\Delta}_k))^T \\ &\quad \times(F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)), \end{aligned}$$

da (3.49), (3.64) e (3.61) si ricava

$$\begin{aligned} \|F(x_k + \bar{p}_{tr}(\bar{\Delta}_k))\|^2 - \|F_k + J_k\bar{p}_{tr}(\bar{\Delta}_k)\|^2 &\leq K^2\|\bar{p}_{tr}(\bar{\Delta}_k)\|^4 + 2\sigma_k K\|\bar{p}_{tr}(\bar{\Delta}_k)\|^2\|x_k - x^*\| \\ &\leq K\nu^2(K\nu^2\|x_k - x^*\| + 2\sigma_k)\|x_k - x^*\|^3 \\ &\leq K\nu^2(K\nu^2\zeta + 2\sigma_k)\|x_k - x^*\|^3 \\ &\leq K\nu^2(K\nu^2\frac{1}{2K^2\nu^2} + \frac{1}{K})\|x_k - x^*\|^3 \\ &= \frac{3\nu^2}{2}\|x_k - x^*\|^3. \end{aligned} \tag{3.77}$$

Inoltre, da (3.47), (3.64) e (3.73) otteniamo

$$\|F_k\|^2 - \|F_k + J_k \bar{p}_{tr}(\bar{\Delta}_k)\|^2 \geq \left(\frac{1}{K^2} - \sigma_k^2\right)\|x_k - x^*\|^2 > \frac{3}{4K^2}\|x_k - x^*\|^2. \qquad (3.78)$$

Infine, da (3.76), (3.77), (3.78) e (3.72) si ha

$$\begin{aligned} \rho_f(\bar{p}_{tr}(\bar{\Delta}_k)) &\geq 1 - \frac{3/2\nu^2 K^2}{3/4}\|x_k - x^*\| \\ &= 1 - 2K^2\nu^2\|x_k - x^*\| \geq 1 - 2K^2\nu^2\zeta > 1 - 2K^2\nu^2\,\frac{1-\beta_2}{2K^2\nu^2} = \beta_2. \end{aligned}$$

Quindi, $x_{k+1} = x_k + \bar{p}_{tr}(\bar{\Delta}_k)$ e da (3.65) concludiamo che $\|x_{k+1} - x^*\| \leq \sigma_k\|x_k - x^*\|$. La forma di σ_k data in (3.74) assicura la convergenza superlineare se $\alpha_k \to 1$ per $k \to \infty$. Inoltre, se $1 - \alpha_k = O(\|F_k\|)$, da (3.47)-(3.48), otteniamo $\sigma_k = O(\|x_k - x^*\|)$ e questo dà luogo a velocità di convergenza quadratica. $\square$

Capitolo 4

Implementazione del metodo TRESNORE e risultati numerici

Lo scopo di questo capitolo è fornire un algoritmo per il metodo TRESNORE e mostrarne il comportamento numerico nella risoluzione di sistemi quadrati e rettangolari.

4.1 L'algoritmo e la sua implementazione

In questo paragrafo forniamo un algoritmo dettagliato per il metodo TRESNORE e descriviamo la sua implementazione in un codice MATLAB. Analizziamo in particolare i Passi 3.2 e 3.5 del metodo che sono stati indicati nel capitolo precedente senza fornire dettagli sulla loro effettiva realizzazione.

4.1.1 Calcolo del passo di trust-region $p_{tr}(\Delta_k)$

Al Passo 3.2 del metodo TRESNORE, si richiede il calcolo del passo $p_{tr}(\Delta_k)$, cioè la risoluzione del problema di trust-region (3.7). Poiché il Teorema 3.1.1 dà una caratterizzazione delle soluzioni di (3.7) ma non fornisce indicazioni pratiche per il loro calcolo, proponiamo un procedimento per il calcolo di una soluzione approssimata.

Iniziamo valutando il passo di Newton di minima norma $p_k^N = -J_k^+ F_k$ e per contenere il costo computazionale, se il problema è quadrato e numericamente non singolare, calcoliamo p_k^N usando il metodo di Eliminazione di Gauss con pivoting parziale. Quindi, se Δ_k è sufficientemente grande da soddisfare $\|D_k p_k^N\| \leq \Delta_k$, poniamo $p_{tr}(\Delta_k) = p_k^N$. Se invece $\|D_k p_k^N\| > \Delta_k$, calcoliamo una soluzione approssimata di (3.7) utilizzando il problema di trust-region sferico descritto in (3.9) e (3.10). Di fatto, il problema (3.9) è risolto con la tecnica dogleg. Essa consiste nel costruire un cammino $\tilde{p}^D(\pi)$, $\pi \in [0, \pi_u]$ che parte dall'origine e connette un prescelto insieme di punti detti "nodi" in modo che $\tilde{m}_k(\tilde{p}^D(\pi))$ sia una funzione monotona decrescente all'aumentare di π. Quindi, il problema (3.9) è risolto in modo approssimato

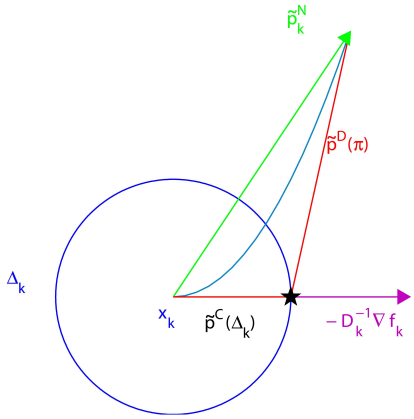

Figura 4.1: Dogleg: caso $\|\tilde{p}^C(\Delta_k)\| = \Delta_k$.

considerando il problema

$$\min_{0\le\pi\le\pi_u} \{\tilde{m}_k(\tilde{p}^D(\pi)),\ \|\tilde{p}^D(\pi)\| \le \Delta_k\}. \tag{4.1}$$

Per realizzare la strategia dogleg, posto

$$\tilde{p}^C(\Delta_k) = D_k p^C(\Delta_k) \quad \text{e} \quad \tilde{p}_k^N = D_k p_k^N, \tag{4.2}$$

dove $p^C(\Delta_k)$ è definito in (3.18), consideriamo il cammino

$$\tilde{p}^D(\pi) = \begin{cases} \pi\tilde{p}^C(\Delta_k), & \pi \in [0,1] \\ \tilde{p}^C(\Delta_k) + (\pi-1)(\tilde{p}_k^N - \tilde{p}^C(\Delta_k)), & \pi \in [1,2] \end{cases}. \tag{4.3}$$

Il cammino $\tilde{p}^D(\pi)$ è una spezzata formata da due segmenti: il primo connette l'origine con $\tilde{p}^C(\Delta_k)$ lungo la direzione di più ripida discesa; il secondo connette $\tilde{p}^C(\Delta_k)$ e $\tilde{p}_k^N$.

Mostriamo che lungo il cammino $\tilde{p}^D(\pi)$, $\tilde{m}_k$ è monotona decrescente al crescere di π. Banalmente questo è vero se $\pi \in [0,1]$, cioè sul segmento dall'origine a $\tilde{p}^C(\Delta_k)$. Consideriamo allora il segmento da $\tilde{p}^C(\Delta_k)$ a $\tilde{p}_k^N$ ponendo $\tilde{p}^D(1+\pi)$, $\pi \in [0,1]$ e $h(\pi) = \tilde{m}_k(\tilde{p}^D(1+\pi))$ e dimostriamo che $h'(\pi) \le 0$. Osserviamo che

$$\begin{aligned} h(\pi) \;=\; & \frac{1}{2}\pi^2(\tilde{p}_k^N - \tilde{p}^C(\Delta_k))^T D_k^{-1} J_k^T J_k D_k^{-1}(\tilde{p}_k^N - \tilde{p}^C(\Delta_k)) + \\ & +\pi[(\tilde{p}_k^N - \tilde{p}^C(\Delta_k))^T D_k^{-1} J_k^T J_k D_k^{-1}\tilde{p}^C(\Delta_k) + \nabla f_k{}^T D_k^{-1}(\tilde{p}_k^N - \tilde{p}^C(\Delta_k))] + \\ & +\frac{1}{2}\tilde{p}^C(\Delta_k)^T D_k^{-1} J_k^T J_k D_k^{-1}\tilde{p}^C(\Delta_k) + \nabla f_k{}^T D_k^{-1}\tilde{p}^C(\Delta_k) + f_k. \end{aligned}$$

Passando alla variabile del problema di trust-region ellittico mediante la relazione (3.8), la derivata di $h(\pi)$ risulta

$$\begin{aligned} h'(\pi) \;=\; & \pi(p_k^N - p^C(\Delta_k))^T J_k^T J_k(p_k^N - p^C(\Delta_k)) + \\ & +(p_k^N - p^C(\Delta_k))^T J_k^T J_k p^C(\Delta_k) + \nabla f_k{}^T(p_k^N - p^C(\Delta_k)). \end{aligned}$$

Poiché $\pi \in [0,1]$ e vale la (3.12), abbiamo che

$$\begin{aligned} h'(\pi) &\leq (p_k^N - p^C(\Delta_k))^T [J_k^T J_k (p_k^N - p^C(\Delta_k)) + J_k^T J_k p^C(\Delta_k) + \nabla f_k] \\ &= (p_k^N - p^C(\Delta_k))^T [J_k^T J_k p_k^N + J_k^T F_k] = 0. \end{aligned}$$

Osserviamo inoltre che esiste un solo punto del cammino $\tilde{p}^D(\pi)$ tale che $\|\tilde{p}^D(\pi)\| = \Delta_k$. Questo punto è $\tilde{p}^D(1) = \tilde{p}^C(\Delta_k)$ se $\|\tilde{p}^C(\Delta_k)\| = \Delta_k$ (vedi Figura 4.1) oppure è il punto sul cammino da $\tilde{p}^C(\Delta_k)$ a $\tilde{p}_k^N$ ottenuto calcolando la radice positiva della seguente equazione quadratica scalare:

$$\|\tilde{p}^C(\Delta_k) + (\pi - 1)(\tilde{p}_k^N - \tilde{p}^C(\Delta_k))\|^2 = \Delta_k^2. \tag{4.4}$$

È facile verificare che questa equazione ha una sola radice positiva $\pi \in (1,2)$, vedi Figura 4.2. Osserviamo che se J_k ha rango massimo per colonne, allora procedendo come in [29, Lemma 4.1] si può dimostrare che $\|\tilde{p}^D(\pi)\|$ è una funzione crescente di π.

Considerando le precedenti proprietà ed il Teorema 3.1.1 possiamo definire la soluzione del problema (4.1) come segue. Poniamo $\pi = 1$, i.e. $\tilde{p}^D(1) = \tilde{p}^C(\Delta_k)$ se $\|\tilde{p}^C(\Delta_k)\| = \Delta_k$. Altrimenti calcoliamo la soluzione positiva di (4.4) e utilizziamo $\tilde{p}^D(\pi)$ come approssimazione della soluzione del problema di trust-region. Infine, dalla (3.8), il passo $p_{tr}(\Delta_k) = D_k^{-1}\tilde{p}^D(\pi)$ rappresenta una soluzione approssimata di (3.7).

Riassumiamo il procedimento appena descritto nel seguente algoritmo:

Algoritmo 1 METODO DOGLEG PER IL CALCOLO DI $p_{tr}(\Delta_k)$
Dati: Δ_k, $p^C(\Delta_k)$, p_k^N.
Se $\|D_k\, p_k^N\| \leq \Delta_k$
    poni $p_{tr}(\Delta_k) = p_k^N$,
altrimenti
    calcola $\tilde{p}_k^N$ *e* $\tilde{p}^C(\Delta_k)$ *usando la (4.2).*
    Se $\|\tilde{p}^C(\Delta_k)\| = \Delta_k$
        poni $\pi = 1$,
    altrimenti
        calcola la radice positiva di (4.4).
    Poni $\tilde{p}^D(\pi) = \tilde{p}^C(\Delta_k) + (\pi - 1)(\tilde{p}_k^N - \tilde{p}^C(\Delta_k))$.
    Poni $p_{tr}(\Delta_k) = D_k^{-1}\tilde{p}^D(\pi)$.

4.1.2 Calcolo del passo di prova $p(\Delta_k)$

Il passo $p(\Delta_k)$ usato per aggiornare l'iterata corrente x_k è costruito al Passo 3.5 del metodo TRESNORE. Per soddisfare la condizione (3.22), cerchiamo un passo $p(\Delta_k)$ della forma (3.23), i.e. $p(\Delta_k) = t\, p_c(\Delta_k) + (1-t)\bar{p}_{tr}(\Delta_k)$, $t \in [0,1]$, come segue. Se $\bar{p}_{tr}(\Delta_k)$ soddisfa la (3.22), allora poniamo $t = 0$, i.e. $p(\Delta_k) = \bar{p}_{tr}(\Delta_k)$. Altrimenti scegliamo il parametro $t \in (0,1)$ in modo che $\rho_c(p(\Delta_k)) = \beta_1$.

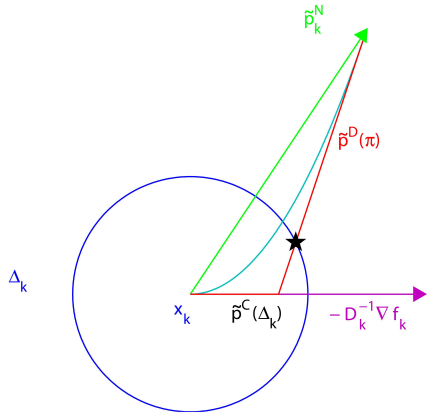

Figura 4.2: Dogleg: caso $\|\tilde{p}^C(\Delta_k)\| < \Delta_k$.

Vediamo in pratica come ottenere il parametro t. Osserviamo che la funzione $g_k(t) = \rho_c(tp_c(\Delta_k) + (1-t)\bar{p}_{tr}(\Delta_k)) - \beta_1$ è una funzione quadratica in t della forma

$$g_k(t) = at^2 + bt + c, \tag{4.5}$$

dove, posto $p_1 = J_k\, p_c(\Delta_k)$ e $p_2 = J_k\bar{p}_{tr}(\Delta_k)$, abbiamo

$$\begin{aligned} a &= -\tfrac{1}{2}\|p_1 - p_2\|^2, \\ b &= -(F_k + \bar{p}_{tr}(\Delta_k))^T(p_1 - p_2), \\ c &= \tfrac{1}{2}\|p_2\|^2 - \nabla f_k{}^T\bar{p}_{tr}(\Delta_k) + \beta_1(\tfrac{1}{2}\|p_1\|^2 + \nabla f_k{}^T p_c(\Delta_k)). \end{aligned} \tag{4.6}$$

Poichè $g_k(0) = \rho_c(\bar{p}_{tr}(\Delta_k)) - \beta_1 < 0$ e $g_k(1) = \rho_c(p_c(\Delta_k)) - \beta_1 > 0$, allora esiste $t \in (0,1)$ tale che $g_k(t) = 0$. Di seguito, schematizziamo il procedimento per trovare $p(\Delta_k)$.

Algoritmo 2 Trovare un passo che soddisfi (3.22).

Dati : $x_k \in int(\Omega)$, $\bar{p}_{tr}(\Delta_k)$, $p_c(\Delta_k)$.
1. Poni $t = 0$.
2. Se $\rho_c(\bar{p}_{tr}(\Delta_k)) < \beta_1$
Calcola la radice positiva dell'equazione definita in (4.5) e (4.6).
3. Poni $p(\Delta_k) = tp_c(\Delta_k) + (1-t)\bar{p}_{tr}(\Delta_k)$.

4.1.3 Accettazione del passo e aggiornamento della trust-region

Il passo di prova $p(\Delta_k)$ viene accettato se soddisfa la condizione (3.24). Altrimenti, il passo $p(\Delta_k)$ viene rifiutato e il raggio di trust-region viene ridotto ponendo:

$$\Delta_k = \min\{\delta_1\Delta_k, \delta_2\|D_k\, p(\Delta_k)\|\}, \tag{4.7}$$

dove δ_1, δ_2 sono scalari tali che $0 < \delta_1 \leq \delta_2 < 1$. Quindi viene calcolato un nuovo passo di prova mediante gli Algoritmi 1 e 2. Ricordiamo che il Lemma 3.2.3 garantisce

che nella strategia di accettazione del passo non può verificarsi un "break-down", cioè un passo $p(\Delta_k)$ accettabile viene costruito operando un numero finito di riduzioni del raggio di trust-region.

Accettato il passo $p(\Delta_k)$, si pone $x_{k+1} = x_k + p(\Delta_k)$ e si eseguono strategie standard di aggiornamento del raggio di trust-region per l'iterazione $(k+1)$-esima. Osserviamo che se il raggio di trust-region è troppo piccolo in relazione all'accordo che c'è tra il modello e la funzione di merito, allora il metodo può perdere l'occasione di fare una sostanziale miglioramento nella stima della soluzione. Quindi, alla fine di ogni iterazione, controlliamo la seguente condizione:

$$\rho_f(p(\Delta_k)) = \frac{f(x_k) - f(x_k + p(\Delta_k))}{m_k(0) - m_k(p(\Delta_k))} \geq \beta_3, \tag{4.8}$$

dove $\beta_3 \in (0, 1]$ è una costante tale che $\beta_2 < \beta_3 < 1$. Se (4.8) vale, allora si consente un eventuale aumento del raggio di trust-region ponendo

$$\Delta_{k+1} = \max\{\Delta_k, 2\|D_k\, p(\Delta_k)\|\}, \tag{4.9}$$

altrimenti, poniamo $\Delta_{k+1} = \Delta_k$.

4.1.4 L' algoritmo implementato

Dopo aver descritto gli aspetti implementativi dei passi più importanti del metodo TRESNORE, presentiamo un algoritmo per il metodo.

Algoritmo per il metodo TRESNORE

Dati: $x_0 \in int(\Omega)$, $\bar{\Delta}_0 \geq \Delta_{min} > 0$, $0 < \delta_1 \leq \delta_2 < 1$ $\beta_1 \in (0, 1]$, $0 < \beta_2 < \beta_3 < 1$, $\theta \in (0, 1)$.

Ripeti fintanto non sono soddisfatti i criteri d'arresto

1. *Poni* $\Delta_k = \bar{\Delta}_k$.
2. *Calcola* F_k.
3. *Calcola la matrice* D_k *usando la (3.4).*
4. *Calcola la matrice jacobiana* J_k.
5. *Calcola il passo di Newton di norma minima dato in (3.12).*
6. *Ripeti fintanto* $\rho_f(p(\Delta_k)) \geq \beta_2$
   - *6.1 Calcola* $p_{tr}(\Delta_k)$ *usando l'Algoritmo 1.*
   - *6.2 Calcola il passo di Cauchy generalizzato* $p_c(\Delta_k)$ *definito in (3.20).*
   - *6.3 Calcola il passo* $\bar{p}_{tr}(\Delta_k)$ *usando (3.14).*
   - *6.4 Calcola il passo* $p(\Delta_k)$ *usando l'Algoritmo 2.*
   - *6.5 Poni* $\Delta_k^* = \Delta_k$ *e riduci* Δ_k *usando (4.7).*
   - *6.6 Calcola* $\rho_f(p(\Delta_k))$ *da (3.24).*
7. *Poni* $x_{k+1} = x_k + p(\Delta_k)$ *e* $\Delta_k = \Delta_k^*$.
8. *Se* $\rho_f(p(\Delta_k)) \geq \beta_3$

   aggiorna $\bar{\Delta}_{k+1}$ *usando (4.9)*

   altrimenti

   poni $\bar{\Delta}_{k+1} = \Delta_k$.

Nell'implementazione MATLAB di questo algoritmo, al Passo 4 la matrice Jacobiana J_k è approssimata con differenze finite. Al Passo 6.1, il calcolo del passo di Newton di minima norma p_k^N è realizzato come segue.

- Se J_k è quadrata e numericamente invertibile, si risolve il sistema lineare (2.3) con il metodo di eliminazione di Gauss mediante la funzione predefinita in MATLAB (operatore "left matrix divide"). Osserviamo che questa funzione stima il reciproco del numero di condizionamento di J_k e considera J_k numericamente invertibile se tale stima è superiore alla precisione di macchina. Altrimenti la matrice J_k è dichiarata numericamente singolare.

- Se J_k è numericamente singolare oppure è rettangolare, si calcola la pseudoinversa di J_k mediante la funzione MATLAB `pinv` e si forma il passo p_k^N usando (3.12).

Riguardo ai criteri d'arresto, viene dichiarata la convergenza quando è verificata la seguente condizione:

$$\|F_{k+1}\| \leq \texttt{atol} + \texttt{rtol}\,\|F_k\|, \tag{4.10}$$

dove `atol` e `rtol` sono tolleranze fornite in ingresso.
Viene invece dichiarato fallimento quando si presentano le seguenti situazioni:

Fallimento E1: viene raggiunto il numero massimo di iterazioni `maxit`;

Fallimento E2: viene raggiunto il numero massimo `maxnF` di valutazioni di F;

Fallimento E3: il raggio di trust-region Δ_k è stato ridotto sotto $\sqrt{\epsilon_m}$, dove ϵ_m è la precisione di macchina;

Fallimento E4: il cambiamento relativo nel valore di F soddisfa:

$$\|F_{k+1} - F_k\| \leq 100\epsilon_m \|F_k\|;$$

Fallimento E5: è stato calcolato un punto stazionario per f, i.e.

$$\|D_k^{-1}\nabla f_k\| \leq 100\epsilon_m;$$

Fallimento E6: nel calcolo della matrice di scalatura D_k si è verificato un overflow.

I fallimenti E4 ed E5 possono indicare che la successione $\{x_k\}$ converge ad un minimo locale della funzione di merito f che non è soluzione del nostro problema $F(x) = 0$. Osserviamo che questa situazione è in accordo con i risultati teorici presentati nel Teorema 3.2.1 ma rappresenta un fallimento nella risoluzione del problema (3.1). Il fallimento E6 si verifica quando la succesione tende ad un punto sul bordo di Ω e l'iterata è così vicina al bordo da provocare un overflow nel calcolo di D_k.

Operando con precisione di macchina $\epsilon_m \cong 2 \cdot 10^{-16}$, i parametri usati nell'algoritmo sono stati scelti come segue:

$$\begin{aligned}
&\theta = 0.99995,\ \beta_1 = 0.1,\ \beta_2 = 0.25,\ \beta_3 = 0.75,\ \delta_1 = 0.25,\ \delta_2 = 0.5,\\
&\Delta_{min} = \sqrt{\epsilon_m},\ \sigma = 0.95,\ \alpha_k = \max\{\sigma, 1 - \|F_k\|\},\\
&\texttt{rtol} = \texttt{0},\ \texttt{atol} = \texttt{10}^{-8},\ \texttt{maxit} = \texttt{1000},\ \texttt{maxnF} = \texttt{1000}.
\end{aligned}$$

La scelta di questi parametri è stata fatta in base alle sperimentazioni numeriche svolte in [5].

Concludiamo questo paragrafo, quantificando il costo computazionale di una iterazione. Il costo predominante è dato dal calcolo di J_k e del passo di Newton di minima norma. In particolare, notiamo che:

- Il calcolo di J_k con il metodo delle differenze finite richiede n valutazioni della funzione F.
- Per ogni ampiezza di trust-region Δ_k considerata all'iterata k-esima, è necessaria una valutazione di funzione F nel punto $x_k + p(\Delta_k)$.
- Se J_k è quadrata e non singolare il costo della risoluzione del sistema lineare (2.3) con il metodo di eliminazione di Gauss con pivoting parziale è dell'ordine $O(\frac{1}{3}n^3)$, [18].
- Se J_k è quadrata e singolare o rettangolare, il costo del calcolo del passo di Newton di minima norma è principalmente determinato dal calcolo della pseudoinversa di J_k mediante la sua decomposizione ai valori singolari. Questo costo è tipicamente elevato, ad esempio nel caso sovradeterminato è dell'ordine $O(2mn^2 - \frac{2}{3}n^3)$, [18].

4.2 Risultati Numerici

In questa sezione mostriamo le prestazioni del codice TRESNORE applicato ad una vasta gamma di problemi noti in letteratura. Nella Tabella 4.1 sono elencati i 46 problemi considerati ed è specificato il nome del problema e il corrispondente riferimento bibliografico. In particolare, 15 problemi sono sistemi quadrati, 7 problemi sono sistemi sovradeterminati, 12 problemi sono sistemi sottodeterminati e i restanti 12 problemi corrispondono a sistemi di equazioni e disequazioni. Le dimensioni dei problemi considerati sono tali che $2 \leq m \leq 500$, $2 \leq n < 500$.

Per ogni problema sono eseguite tre risoluzioni variando l'approssimazione iniziale x_0. Se i vincoli l ed u sono finiti, il problema è risolto scegliendo tre approssimazioni iniziali x_0 della forma:

$$x_0 = l + 0.25\gamma(u - l), \quad \text{per} \quad \gamma = 1, 2, 3. \tag{4.11}$$

Se $\Omega = \mathbb{R}^n_+$, ovvero Ω è l'ottante positivo $\Omega = \{x \in \mathbb{R}^n \ t.c.\ x \geq 0\}$, viene usata l'approssimazione iniziale

$$x_0 = 10^{\gamma-1}, \quad \text{per} \quad \gamma = 1, 2, 3. \tag{4.12}$$

Inoltre tutti i problemi sono stati risolti con due raggi iniziali di trust-region:

$$\Delta_0 = 1 \quad \text{e} \quad \Delta_0 = \|D_0^{-1}\nabla f_0\|. \tag{4.13}$$

Nelle tabelle seguenti verranno riportate: le dimensioni m ed n del problema (solo n nel caso quadrato), il numero di iterazioni `it` ed il numero `nvF` di valutazioni di funzione eseguite, la norma della funzione F nell'ultima iterata calcolata. Inoltre verrà riportata la scelta del parametro γ che determina l'approssimazione iniziale x_0 ed il valore di Δ_0 se $\Delta_0 = \|D_0^{-1}\nabla f_0\|$. I fallimenti descritti nella sezione precedente verranno indicati con i simboli E1-E6. La notazione $\omega(q)$ significa $\omega 10^q$. È importante osservare che `it` $\leq$ `nvF` $- 1$ e che nel conteggio del numero di valutazioni di funzione eseguite non vengono considerate le valutazioni di funzione effettuate nel calcolo della matrice Jacobiana J_k. Tuttavia il costo complessivo delle valutazioni della matrice Jacobiana è facilmente ricavabile dal numero di iterazioni eseguite.

Pb#	Nome e riferimento bibliografico	Pb#	Nome e riferimento bibliografico
1	Himmelblau [16]	24	HS-53 [19]
2	Equilibrium Combustion [16]	25	HS-56 [19]
3	Bullard-Biegler [16]	26	HS-63 [19]
4	Ferraris-Tronconi [16]	27	HS-77 [19]
5	Azeotropes-Wilson [16]	28	HS-79 [19]
6	Azeotropes-NRTL [16]	29	HS-81 [19]
7	BVP3 [32]	30	HS-87 [19]
8	Rosenbrock [26]	31	HS-107 [19]
9	Freudenstein and Roth [26]	32	HS-111 [19]
10	Helical valley [26]	33	Linear system [10]
11	Brown's almost linear [16]	34	Quadratic system [10]
12	HS-8 [19]	35	HS-14 [19]
13	HS-55 [19]	36	HS-15 [19]
14	Powell badly scaled [16]	37	HS-23 [19]
15	Robot kinematics [16]	38	HS-24 [19]
16	Beale [26]	39	HS-44 [19]
17	Biggs EXP6 [26]	40	HS-59 [19]
18	Box three-dimensional [26]	41	HS-74 [19]
19	Brown badly scaled [26]	42	HS-83 [19]
20	Gulf research & development [26]	43	HS-86 [19]
21	Variably dimensioned [26]	44	HS-32 [19]
22	Wood [26]	45	HS-43 [19]
23	HS-46 [19]	46	HS-100 [19]

Tabella 4.1: Problemi risolti

4.2.1 Risoluzione di problemi quadrati

I problemi quadrati `Pb1-Pb15` nascono in contesti applicativi. Notiamo in particolare che i problemi `Pb2` e `Pb7` corrispondono ai problemi descritti negli esempi del Capitolo 1 e sono rispettivamente il problema di equilibrio chimico presentato nell'Esempio 2 e il problema ai limiti descritto nell'Esempio 1.

I problemi `Pb1-Pb4, Pb11 e Pb15` hanno vincoli inferiormente e superiormente limitati. Quindi, sono stati risolti con approssimazioni iniziali date in (4.11). Per il problema `Pb13` si utilizza l'approssimazione iniziale (4.11) per le componenti che hanno vincoli limitati, l'approssimazione iniziale (4.12) per le rimanenti. Gli altri problemi hanno regione di ammissibilità $\Omega = \mathbb{R}^n_+$, quindi si utilizza x_0 definito in (4.12).

Nella Tabella 4.2 vengono presentati i risultati numerici relativi ai problemi `Pb1-Pb10`. Nella risoluzione di questi problemi la matrice Jacobiana J_k è risultata non singolare per ogni k. Complessivamente, il metodo ha risolto 53 test su 60 e la scelta $\Delta_0 = \|D_0^{-1}\nabla f_0\|$ è risultata meno efficiente di $\Delta_0 = 1$ poiché TRESNORE ha fallito nel risolvere `Pb3` per ogni scelta di γ.

La Tabella 4.3 presenta i risultati relativi ai test nella cui esecuzione la matrice Jacobiana risulta numericamente singolare per un $k \geq 0$. Mostriamo questi risultati perché i test presentati nella tavola non sono risolubili utilizzando la versione "esatta" del metodo SIATR [5] di cui TRESNORE è l'estensione. In particolare il metodo SIATR risolve sistemi non lineari quadrati vincolati e consente la risoluzione del sistema lineare algebrico che nasce ad ogni iterazione con un metodo diretto o in modo approssimato mediante un procedimento iterativo. In letteratura la prima versione è detta "esatta", la seconda è detta "inesatta". Il metodo SIATR non gestisce il caso in cui la matrice Jacobiana è singolare ad una iterazione. Infatti il calcolo del passo di Newton mediante il metodo di eliminazione di Gauss viene svolto stimando il reciproco del numero di condizionamento di J_k e se esso è inferiore alla precisione di macchina, si dichiara fallimento. La versione esatta del codice SIATR, reso disponibile dagli autori, fallisce nella risoluzione di tutti i test presentati nella Tabella 4.3 con un fallimento dovuto alla singolarità della matrice Jacobiana. Al contrario, il codice TRESNORE gestisce la singolarità di J_k come descritto nella sezione precedente e mostra superiorità numerica poiché risolve con successo un numero elevato di test. In particolare, su un totale di 22 test, TRESNORE risolve 19 test.

Le Tabelle 4.2 e 4.3 mostrano le buone prestazioni del metodo. Numericamente è stata osservata convergenza asintotica veloce sia nei problemi `Pb1-Pb10`, che nei problemi `Pb11-Pb15`. In particolare, in quest'ultimi è stato osservato sperimentalmente che J_k è singolare in punti lontani dalla soluzione, mentre è invertibile in un intorno della soluzione.

4.2.2 Sistemi sovradeterminati

Il metodo TRESNORE è stato applicato nella risoluzione dei sistemi non lineari sovradeterminati `Pb16-Pb22`. I problemi sono contenuti in [26] e rappresentano il

calcolo di punti stazionari per problemi di minimo. Per tutti i problemi, $\Omega = \mathbb{R}^n_+$.

Nella risoluzione del problema `Pb20`, l'approssimazione iniziale non è definita come in (4.12). È stato usato $x_0 = \gamma x_m$ dove $\gamma = 1, 2, 3$ e x_m è l'approssimazione iniziale suggerita in [26]. Tale scelta è motivata dal fatto che le equazioni del sistema non lineare hanno forma esponenziale e nei punti x_0 della forma (4.12) la matrice Jacobiana risulta numericamente nulla. Questo provoca inevitabilmente un fallimento di tipo E5 alla prima iterazione.

La scelta di $\gamma = 0$ in `Pb21` e in `Pb22` è dovuta al fatto che per $\gamma = 1$ il punto iniziale coincide con soluzioni esatte dei problemi.

I risultati riportati nella Tabella 4.4 mostrano l'efficienza del metodo TRESNORE che ha risolto 50 test su 60. Si osservi che il raggio iniziale di trust-region $\Delta_0 = 1$ dà luogo ad un maggior numero di successi.

4.2.3 Sistemi sottodeterminati

Per valutare il comportamento del nostro metodo nella risoluzione dei problemi sottodeterminati abbiamo risolto i problemi `Pb23-Pb34` che rappresentano il calcolo di un punto ammissibile per un problema di programmazione non lineare, [19].

I problemi considerati hanno regione di ammissibilità $\Omega = \mathbb{R}^n_+$ ad eccezione di `Pb24, Pb29-Pb32` che hanno l ed u finiti e di `Pb31` in cui solo alcune componenti dei vettori l ed u sono finite. Quindi, nel problema `Pb31` si è costruita l'approssimazione iniziale usando (4.11) per le componenti vincolate e (4.12) per le rimanenti. Nella Tabella 4.5 sono riassunti i risultati ottenuti. La scelta di $\gamma = 0$ in `Pb23` e $\gamma = 2.5$ in `Pb24` è dovuta al fatto che per $\gamma = 1$ e $\gamma = 2$ rispettivamente, x_0 coincide con una soluzione dei problemi.

Il metodo TRESNORE risulta efficiente anche per questa classe di problemi in quanto si ottengono 54 successi su 72 test. I fallimenti E6 nella risoluzione del problema `Pb30` e `Pb31` sono imputabili al fatto che per alcune componenti l'ampiezza $u_i - l_i$ della scatola è minore di uno e considerevolmente inferiore a quelle di altre componenti.

4.2.4 Sistemi di equazioni e disequazioni non lineari

Nell'ambito dei sistemi di equazioni e disequazioni non lineari abbiamo considerato problemi tratti da [19] che rappresentano il calcolo di punti ammissibili per problemi di programmazione non lineare. Dodici problemi di [19] sono stati trasformati in sistemi rettangolari utilizzando la riformulazione definita in (1.6). I problemi risultanti sono `Pb35-Pb46`.

I problemi `Pb35-Pb36` sono quadrati, `Pb37-Pb43` sono sovradeterminati e i restanti problemi sono sottodeterminati. I problemi `Pb37, Pb40-Pb42` hanno vincoli l ed u finiti, in `Pb36` l ed u sono tali che $l = (-\infty, -\infty)^T$ e $u = (0.5, +\infty)^T$ e si sono utilizzate approssimazioni iniziali $x_0 = (-10^{\gamma-1}, 10^{\gamma-1})^T$ con $\gamma = 1, 2, 3$. I rimanenti problemi hanno $\Omega = \mathbb{R}^n_+$. La scelta di $\gamma = 1.5$ nei problemi `Pb37, Pb39, Pb40,`

`Pb45` e `Pb46` è dovuta al fatto che per $\gamma = 2$ in `Pb37, Pb40` e $\gamma = 1$ in `Pb39, Pb45` e `Pb46` x_0 coincide con una soluzione del sistema non lineare.

Nella Tabella 4.6, riportiamo i risultati ottenuti e indichiamo con m_e e m_i rispettivamente il numero di equazioni e di disequazioni del problema originario. Il metodo TRESNORE ha ottime prestazioni e si riscontrano 69 successi su 72 . È importante osservare che per tutti i problemi la matrice Jacobiana non ha rango massimo nella soluzione calcolata.

4.3 Considerazioni finali

Il nuovo metodo TRESNORE è stato applicato nella risoluzione di un elevato numero di problemi rettangolari. Le approssimazioni iniziali sono state fissate indipendentemente dal problema e quindi in generale non sono buone approssimazioni delle soluzioni. Pertanto, dai risultati ottenuti possiamo concludere che la strategia di globalizzazione proposta è efficiente e che il metodo TRESNORE è un metodo "robusto", cioè un metodo capace di risolvere un elevato numero di test. In particolare, TRESNORE ha risolto 245 test su un totale di 286 test e le scelte del raggio di trust-region iniziale $\Delta_0 = 1$ e $\Delta_0 = \|D_0^{-1}\nabla f_0\|$, hanno fornito 124 e 121 successi rispettivamente. Riguardo ai fallimenti è importante osservare che solo in 17 casi, fissati un problema e una approssimazione iniziale x_0, si è avuto fallimento per entrambe le scelte di Δ_0 e che 13 fallimenti sono del tipo E4-E5.

L'analisi dei valori di `it` e `nvF` mostra inoltre che nella risoluzione di numerosi problemi, il metodo TRESNORE ha un costo computazionale contenuto. Infatti il metodo richiede al più 20 valutazioni di funzione nel 71% dei test (173 test su 245) e al più 30 valutazioni di funzione nel 83% dei test (204 test su 245).

Abbiamo dimostrato che il metodo TRESNORE ha garanzia di convergenza asintotica veloce nel caso di problemi quadrati e sovradeterminati se la matrice Jacobiana ha rango massimo per colonne nella soluzione calcolata. Tale convergenza asintotica veloce è stata verificata sperimentalmente nella risoluzione dei problemi `Pb1-Pb22` e può essere parzialmente confermata dai risultati presentati nelle Tavole 4.2-4.4. È inoltre rilevante il fatto che anche per i problemi sottodeterminati `Pb23-Pb34` è stata osservata convergenza asintotica veloce a punti con Jacobiana avente rango massimo.

Infine osserviamo che nonostante nei problemi quadrati o sovradeterminati `Pb35-Pb43` la matrice Jacobiana sia singolare nella soluzione calcolata, il metodo TRESNORE risulta economico.

			$\Delta_0 = 1$			$\Delta_0 = \|D_0^{-1}\nabla f_0\|$			
`Pb#`	n	γ	`it`	`nvF`	$\|F\|$	`it`	`nvF`	$\|F\|$	Δ_0
1	2	1	7	8	7(-15)	7	9	3(-13)	2(3)
		2	6	7	0(+00)	6	7	0(+00)	1(3)
		3	6	7	7(-11)	6	7	7(-11)	1(3)
2	5	1	13	14	2(-11)	13	14	9(-13)	6(8)
		2	16	17	4(-11)	15	16	1(-11)	3(10)
		3	17	18	7(-9)	16	17	1(-10)	2(11)
3	2	1	9	10	1(-12)	E3			2(9)
		2	E6			E3			3(10)
		3	E6			E3			1(11)
4	2	1	5	6	1(-15)	5	6	2(-14)	5(-1)
		2	6	7	3(-14)	6	7	3(-14)	1(0)
		3	7	9	5(-14)	6	8	6(-9)	9(0)
5	4	1	12	13	1(-9)	8	9	5(-13)	9(0)
		2	11	13	1(-12)	10	11	9(-9)	1(2)
		3	14	15	2(-10)	51	52	7(-14)	5(3)
6	4	1	13	14	6(-13)	8	9	2(-12)	3(0)
		2	10	11	8(-9)	8	9	8(-14)	1(2)
		3	13	14	3(-11)	10	11	1(-9)	5(3)
7	500	1	8	9	4(-9)	7	8	6(-13)	3(0)
		2	10	11	6(-10)	5	6	8(-10)	3(1)
		3	16	17	1(-12)	8	9	4(-11)	1(3)
8	2	0	13	19	3(-10)	14	21	6(-10)	3(1)
		2	9	10	2(-11)	6	7	2(-9)	5(5)
		3	14	16	1(-12)	7	8	9(-11)	1(9)
9	2	1	E4			E4			
		2	9	10	2(-14)	9	10	8(-15)	1(6)
		3	16	17	1(-10)	15	16	1(-11)	5(11)
10	3	1	6	7	6(-14)	6	8	3(-10)	5(1)
		2	8	9	8(-11)	7	8	8(-18)	5(3)
		3	11	12	6(-12)	7	8	9(-10)	1(5)

Tabella 4.2: Problemi quadrati con matrice Jacobiana non singolare.

			$\Delta_0 = 1$			$\Delta_0 = \|D_0^{-1}\nabla f_0\|$			
`Pb#`	n	γ	`it`	`nvF`	$\|F\|$	`it`	`nvF`	$\|F\|$	Δ_0
11	5	2	7	8	4(-14)	5	6	6(-10)	9(1)
12	2	1	E4			E4			7(1)
		2	E4			31	33	0(+00)	1(4)
		3	16	21	5(-11)	11	13	4(-9)	7(7)
13	6	1	4	5	2(-10)	4	5	2(-10)	6(1)
		2	8	9	1(-13)	5	6	7(-9)	1(3)
		3	11	12	3(-9)	7	8	4(-11)	4(4)
14	2	1	54	63	7(-12)	74	92	7(-11)	1(8)
		2	25	32	4(-11)	15	16	1(-9)	4(11)
		3	17	18	7(-9)	50	59	9(-11)	1(15)
15	8	2	7	8	4(-10)	7	8	4(-9)	8(-1)

Tabella 4.3: Problemi quadrati con matrice Jacobiana singolare.

				$\Delta_0 = 1$			$\Delta_0 = \|D_0^{-1}\nabla f_0\|$			
`Pb#`	m	n	γ	`it`	`nvF`	$\|F\|$	`it`	`nvF`	$\|F\|$	Δ_0
16	3	2	1	7	8	2(-11)	6	8	8(-9)	1(1)
			2	E4			E4			1(8)
			3	E4			E4			3(15)
17	80	6	1	E2			E2			3(0)
			2	19	23	2(-12)	16	22	1(-15)	4(0)
			3	E2			E2			6(-3)
18	4	3	1	4	5	1(-10)	4	5	1(-10)	1(0)
			2	6	7	2(-9)	5	6	1(-12)	5(2)
			3	8	9	2(-9)	6	7	4(-15)	1(3)
	400	3	1	4	5	5(-10)	4	5	5(-10)	5(0)
			2	7	8	3(-16)	5	6	2(-11)	1(2)
			3	9	10	1(-16)	6	7	1(-13)	3(3)
19	3	2	1	29	30	2(-10)	9	10	1(-9)	1(6)
			2	29	30	2(-10)	10	11	0(+00)	9(5)
			3	29	30	2(-10)	10	11	0(+00)	9(6)
20	7	3	1	104	126	2(-12)	83	101	7(-10)	4(0)
			2	87	101	1(-10)	113	122	6(-16)	4(0)
			3	76	92	8(-10)	95	124	2(-10)	4(0)
	70	3	1	22	27	2(-13)	E3			2(1)
			2	21	26	4(-14)	23	27	1(-12)	2(1)
			3	27	35	3(-13)	E3			2(1)
21	11	9	0	11	12	5(-10)	11	12	3(-11)	2(6)
			2	16	17	5(-12)	15	16	0(+00)	7(9)
			3	21	22	4(-9)	18	19	1(-9)	2(13)
	22	20	0	14	15	4(-12)	13	14	4(-9)	7(8)
			2	19	20	1(-12)	17	18	1(-11)	2(12)
			3	24	25	4(-9)	21	22	3(-14)	9(15)
22	6	4	0	8	10	1(-14)	7	10	5(-12)	1(1)
			2	10	11	4(-10)	9	10	7(-13)	7(5)
			3	15	16	4(-9)	12	13	5(-9)	2(9)

Tabella 4.4: Sistemi non lineari sovradeterminati.

				$\Delta_0 = 1$			$\Delta_0 = \|D_0^{-1}\nabla f_0\|$			
Pb#	m	n	γ	it	nvF	$\|F\|$	it	nvF	$\|F\|$	Δ_0
23	2	5	0	13	14	3(-9)	14	15	1(-9)	2(0)
			2	17	18	2(-9)	18	19	2(-15)	1(12)
			3	43	50	4(-10)	39	44	4(-9)	4(23)
24	3	5	1	6	7	2(-9)	5	6	3(-11)	2(2)
			2.5	5	6	4(-10)	5	6	1(-13)	1(2)
			3	6	7	2(-10)	5	6	3(-11)	2(2)
25	4	7	1	4	5	4(-15)	4	5	4(-15)	1(1)
			2	14	17	1(-13)	7	8	1(-9)	5(2)
			3	25	31	1(-14)	16	17	2(-10)	3(4)
26	2	3	1	8	9	3(-14)	7	9	1(-14)	5(2)
			2	9	10	8(-11)	8	10	6(-10)	4(4)
			3	15	17	0(+00)	12	14	5(-9)	5(-9)
27	2	5	1	7	9	2(-14)	6	8	2(-11)	3(1)
			2	19	20	1(-15)	18	19	8(-9)	1(12)
			3	134	175	2(-13)	44	50	1(-9)	4(23)
28	3	5	1	5	6	3(-9)	5	6	5(-11)	1(1)
			2	75	76	8(-9)	57	58	7(-9)	2(4)
			3	E6			E6			
29	3	5	1	6	7	1(-10)	6	7	4(-11)	2(2)
			2	E5			E5			
			3	10	12	1(-15)	9	11	1(-14)	2(2)
30	4	6	1	E6			E6			1(5)
			2	E6			E6			5(5)
			3	E6			E6			1(6)
31	6	9	1	E6			E6			3(0)
			2	E6			E6			4(1)
			3	E6			E6			1(3)
32	3	10	1	E5			E5			1(1)
			2	7	8	1(-15)	7	8	1(-15)	2(2)
			3	59	60	2(-9)	59	60	2(-9)	1(46)
33	150	300	1	17	18	0(+00)	19	20	1(-10)	1(5)
			2	16	17	3(-15)	7	8	7(-11)	1(5)
			3	14	15	5(-15)	7	8	1(-10)	1(6)
34	150	300	1	13	14	1(-13)	8	11	3(-9)	5(3)
			2	13	14	6(-11)	9	10	1(-10)	7(5)
			3	19	20	1(-13)	13	14	1(-13)	2(9)

Tabella 4.5: Sistemi non lineari sottodeterminati.

						$\Delta_0 = 1$			$\Delta_0 = \|D_0^{-1}\nabla f_0\|$			
Pb#	m	m_e	m_i	n	γ	it	nvF	$\|F\|$	it	nvF	$\|F\|$	Δ_0
35	2	1	1	2	1	11	12	9(-9)	13	14	6(-9)	2(-2)
					2	23	24	2(-9)	23	24	2(-9)	6(7)
					3	32	33	6(-9)	31	32	6(-9)	2(15)
36	2	0	2	2	1	E6			3	4	0(+00)	6(0)
					2	22	23	2(-9)	E6			1(7)
					3	29	30	7(-9)	E6			5(14)
37	5	0	5	2	1	22	25	7(-9)	40	49	5(-9)	8(5)
					1.5	25	30	3(-9)	24	28	5(-9)	9(4)
					2	14	15	2(-9)	14	15	2(-9)	5(0)
38	3	0	3	2	1	12	13	6(-9)	14	15	2(-9)	4(-2)
					2	19	20	3(-9)	18	19	7(-9)	3(4)
					3	24	25	7(-9)	22	23	6(-9)	1(8)
39	6	0	6	4	1.5	17	18	5(-9)	17	18	5(-9)	4(3)
					2	21	22	4(-9)	20	21	4(-9)	1(6)
					3	27	28	3(-9)	24	25	4(-9)	1(10)
40	3	0	3	2	1	23	24	2(-9)	22	23	8(-9)	5(9)
					1.5	18	19	2(-9)	18	19	2(-9)	3(5)
					3	16	17	4(-9)	16	17	4(-9)	2(3)
41	5	3	2	4	1	11	12	0(+00)	7	8	2(-13)	1(6)
					2	9	10	1(-9)	7	8	2(-13)	1(6)
					3	8	9	7(-13)	7	8	7(-13)	2(6)
42	6	0	6	5	1	13	14	6(-9)	13	14	6(-9)	3(-1)
					2	12	13	8(-9)	15	16	3(-9)	4(-2)
					3	15	16	2(-9)	15	16	2(-9)	2(0)
43	10	0	10	5	1	17	18	4(-9)	17	18	4(-9)	2(3)
					2	26	27	3(-9)	28	30	2(-9)	3(7)
					3	31	32	4(-9)	42	43	8(-9)	1(11)
44	2	1	1	3	1	5	6	8(-15)	4	5	6(-10)	3(1)
					2	25	26	7(-9)	25	26	3(-9)	3(11)
					3	38	39	9(-9)	38	39	5(-9)	1(23)
45	2	0	2	4	1.5	20	21	6(-9)	20	21	6(-9)	2(6)
					2	25	26	6(-9)	25	26	3(-9)	2(10)
					3	35	36	5(-9)	33	34	8(-9)	9(17)
46	4	0	4	7	1.5	23	24	4(-9)	23	24	4(-9)	5(9)
					2	33	34	4(-9)	33	34	4(-9)	5(17)
					3	52	53	3(-9)	52	53	2(-9)	1(33)

Tabella 4.6: Sistemi di equazioni e disequazioni non lineari.

Appendice A

Funzioni vettoriali e successioni

Enunciamo il Teorema del Valor Medio per funzioni $F : \mathbb{R}^n \to \mathbb{R}^m$ ed un risultato che permette il passaggio della norma sotto il segno di integrale.

Teorema A.1 *[12, Lemma 4.1.9] Sia $F : \mathbb{R}^n \to \mathbb{R}^m$ una funzione differenziabile con continuità in un aperto convesso $D \subset \mathbb{R}^n$. Per ogni x, $x + p \in D$,*

$$F(x+p) - F(x) = \int_0^1 J(x+tp)p\,dt.$$

Teorema A.2 *[12, Lemma 4.1.10] Sia $J : D \to \mathbb{R}^{m\times n}$, dove $D \subset \mathbb{R}^n$ è un aperto convesso contenente x e $x+p$. Allora, se J è integrabile sul segmento $[x, x+p]$*

$$\left\| \int_0^1 J(x+tp)p\,dt \right\| \leq \int_0^1 \|J(x+tp)p\|\,dt,$$

per ogni norma vettoriale definita in $\mathbb{R}^m$.

Il seguente lemma caratterizza la convergenza di una succesione in $\mathbb{R}^n$ aventi punti limite isolati.

Lemma A.1 *[28, Lemma 4.10] Sia x^* un punto limite isolato della successione $\{x_k\}$ in $\mathbb{R}^n$. Se $\{x_k\}$ non converge allora esiste una sottosuccessione $\{x_{l_j}\}$ che converge a x^* e un $\epsilon > 0$ tale che $\|x_{l_j+1} - x_{l_j}\| \geq \epsilon$.*

Appendice B

Decomposizione ai valori singolari di una matrice reale e pseudoinversa di Moore-Penrose

Teorema B.1 *[1, Teorema 7.8] Sia $A \in \mathbb{R}^{m\times n}$. Allora esistono una matrice ortogonale $U \in \mathbb{R}^{m\times m}$ e una matrice ortogonale $V \in \mathbb{R}^{n\times n}$ tali che*

$$A = U\Sigma V^T, \tag{B.1}$$

dove la matrice $\Sigma \in \mathbb{R}^{m\times n}$ ha elementi σ_{ij} nulli per $i \neq j$ ed elementi diagonali $\sigma_{ii} = \sigma_i$ tali che

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0, \quad p = \min\{m, n\}.$$

La decomposizione (B.1) è detta *decomposizione ai valori singolari* della matrice A, mentre i valori σ_i, per $i = 1, \ldots, p$, sono detti *valori singolari* di A.

Dal Teorema B.1 segue il prossimo risultato che in particolare lega i valori singolari di A e gli autovalori di $A^T A$.

Teorema B.2 *[1, Teorema 7.10] Sia $A \in \mathbb{R}^{m\times n}$ e sia*

$$A = U\Sigma V^T,$$

la sua decomposizione ai valori singolari, dove

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Allora valgono le proprietà

a) rango di $A = r$,

b) σ_i^2, $i = 1, \ldots, p$, sono gli autovalori della matrice $A^T A$ (se $m < n$ i restanti autovalori sono nulli) e quindi

$$\|A\| = \sigma_1.$$

Si può dimostrare il seguente teorema di perturbazione per i valori singolari.

Teorema B.3 *[1, Teorema 7.24] Siano A e $\delta A \in \mathbb{R}^{m\times n}$. Se σ_i e ψ_i, $i = 1, \ldots, p$, sono i valori singolari di A e di $A + \delta A$ rispettivamente, risulta*

$$|\psi_i - \sigma_i| \leq \|\delta A\|, \quad i = 1, \ldots, p.$$

Definizione B.1 *[1, Definizione 7.17] Sia $A \in \mathbb{R}^{m\times n}$ una matrice di rango r, $A = U\Sigma V^T$ la sua decomposizione ai valori singolari. La matrice $A^+ \in \mathbb{R}^{n\times m}$ tale che*

$$A^+ = V\Sigma^+ U^T,$$

dove $\Sigma^+ \in \mathbb{R}^{n\times m}$ è la matrice che ha elementi σ_{ij}^+ nulli per $i \neq j$ e per $i = j$ ha elementi

$$\sigma_{ii}^+ = \begin{cases} \dfrac{1}{\sigma_i} & \text{per } i = 1, \ldots, r \\ 0 & \text{per } i = r+1, \ldots, p \end{cases},$$

è detta pseudoinversa di Moore-Penrose di A.

Dalla Definizione B.1 e dal Teorema B.2 segue il

Lemma B.1 *[1] Sia $A \in \mathbb{R}^{m\times n}$ una matrice di rango r e siano $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ i suoi valori singolari non nulli. Allora*

$$\|A^+\| = \frac{1}{\sigma_r}.$$

Se il numero delle righe supera il numero di colonne di A, vale la seguente proprietà.

Lemma B.2 *[1, Esercizio 7.11] Sia $A \in \mathbb{R}^{m\times n}$, con $m \geq n$. Se A ha rango massimo, allora*

$$A^+ = (A^T A)^{-1} A^T. \tag{B.2}$$

Lemma B.3 *Sia $A \in \mathbb{R}^{m\times n}$, con $m \geq n$. Se A ha rango massimo, allora*

$$\|Av\| \geq \frac{\|v\|}{\|A^+\|}, \quad \forall\ v \in \mathbb{R}^n.$$

Dim. Sia $v \in \mathbb{R}^n$. Dalla (B.2) abbiamo che

$$\|Av\| = \frac{\|A^+\|}{\|A^+\|}\|Av\| \geq \frac{\|A^+Av\|}{\|A^+\|} = \frac{\|(A^TA)^{-1}A^TAv\|}{\|A^+\|}.$$

Poiché $(A^TA)^{-1}A^TA = I$, segue la tesi. □

Bibliografia

[1] D. Bini, M. Capovani, O. Menchi, *Metodi Numerici per L'Algebra Lineare*, Zanichelli, Bologna, 1993.

[2] S. Bellavia, M.Macconi, B. Morini, *An affine scaling trust-region approach to bound-constrained nonlinear systems*, Appl. Numer. Math., 44 (2003), pp. 257-280.

[3] S. Bellavia, B. Morini, *An interior global method for nonlinear systems with simple bounds*, Optim. Methods Softw., 20 (2005), pp. 1-22.

[4] S. Bellavia, M. Macconi, B. Morini, *STRSCNE: A Scaled Trust Region Solver for Constrained Nonlinear Equations*, Comput. Optim. Appl., 98 (2004), pp. 31-50.

[5] S. Bellavia, B. Morini, *Subspace trust-region method for large bound-constrained nonlinear equations*, SIAM J. Numer. Anal., in corso di pubblicazione.

[6] A. Conn, N. Gould, Ph. Toint, *Trust Region Methods*, SIAM, 2000.

[7] J.W. Chinneck, *The constraint consensus method for finding approximately feasible points in nonlinear programs*, Informs Journal on Computing, 16 (2004), pp. 255-265.

[8] F.H. Clarke, *Optimization and nonsmooth analysis*, John Wiley and Sons, New York, 1983.

[9] T.F. Coleman, Y. Li, *An interior trust-region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418-445.

[10] H. Dan, N. Yamashita, M. Fukushima, *Convergence properties of the inexact Levenberg-Marquardt method under local error bound conditions*, Technical Report 2001-001, Department of Applied Mathematics and Physics, Kyoto University.

[11] J.E. Dennis, Jr., M. El-Alem, K. Williamson, *A trust-region approach to nonlinear systems of equalities and inequalities*, SIAM J. Optim., 9 (1999), pp. 291-315.

[12] J.E. Dennis, R.B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice Hall, Englewood Cliffs, NJ, 1983.

[13] M.C. Ferris, J.S. Pang, *Engineering and Economic Applications of Complementarity Problems*, SIAM Review, 39, (1997), pp. 669-713.

[14] J.B. Francisco, N. Krejić, J.M. Martínez, *An interior-point method for solving box-constrained underdetermined nonlinear systems*, Journal of Computational and Applied Mathematics, 177 (2005), pp.67-88.

[15] R. Fletcher, S. Leyffer, *Filter-type Algorithms for Solving Systems of Algebraic Equations and Inequalities*, High Performance Algorithms and Software for Nonlinear Optimization, G. Di Pillo and A. Murli, editors, Kluwer Academic Publishers, (2003), pp. 259-278.

[16] C.A. Floudas et al., *Handbook of test problems in local and global optimization*, Kluwer Academic Publishers, Nonconvex Optimization and its Applications, 33, 1999.

[17] P.E. Gill, W. Murray, M.H. Wright, *Practical Optimization*, Academic Press, 1981.

[18] G.H. Golub, C.F. Van Loan, *Matrix Computations*, The J. Hopkins University Press, 1988.

[19] W. Hock, K. Schittkowski, *Test examples for nonlinear programming codes*, Lecture Notes in Economics and Mathematical Systems, Vol. 187, 1981.

[20] C. Kanzow, *An active set-type Newton method for constrained nonlinear systems*, Complementarity: Applications, Algorithms and Extensions, M.C. Ferris, O.L.Mangasarian, J.S. Pang eds, Kluwer Academic Publishers, (2001), pp.179-200.

[21] C. Kanzow, A. Klug, *An interior-point affine-scaling trust-region method for semismooth equations with box constraints*, Computational Optimization and Applications, in corso di pubblicazione.

[22] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers in Applied Mathematics, SIAM, 1995.

[23] D.N. Kozakevich, J.M. Martinez, S.A. Santos, *Solving nonlinear systems of equations with simple bounds*, Comput. Appl. Math. 16 (1997), pp. 215-235.

[24] MATLAB 7, The Math Works, Natick, MA.

[25] K. Meintjes, A.P. Morgan, *Chemical Equilibrium systems as numerical tests problems*, ACM Trans. Math. Soft., 16 (1990), pp. 143-151.

[26] J.J. Moré, B.S. Garbow, K.E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Soft., 7 (1981), 17-41.

[27] J.J. Moré, *The Levenberg-Marquardt algorithm, implementation and theory*, In G. A. Watson, editor, Numerical Analysis, Lecture Notes in Mathematics 630, SpringerVerlag, 1977.

[28] J.J. Moré, D.C. Sorensen, *Computing a trust-region step*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 553-572.

[29] J. Nocedal, S.J. Wright, *Numerical Optimization*, Springer Series in Operations Research, 1999.

[30] J.M. Ortega, W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[31] L. Qi, X.J. Tong, D.H.Li, *An active-set projected trust-region algorithm for box-constrained nonsmooth equations*, J. Optim. Theory Appl., 120 (2004), pp. 627-649.

[32] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Springer, New-York, NY, 2nd ed., 1993.

[33] X.J. Tong, L.Qi, *On the convergence of a trust-region method for solving constrained nonlinear equations with degenerate solutions*, J. Optim. Theory Appl., 123 (2004), pp. 187-211.

[34] M. Ulbrich, *Nonmonotone trust-region methods for bound-constrained semi-smooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim., 11 (2000), pp. 889-917.

[35] S.J. Wright, *Primal-dual interior point methods*, SIAM, 1997.