Università degli studi di Padova

Dipartimento di matematica "Tullio Levi-Civita"

Corso di Laurea Triennale in Matematica

Formule di cubatura attraverso programmazione lineare semi-infinita

Relatore:
Prof. Alvise Sommariva

Laureando:
Riccardo Viero
Matricola:
1069366

21 Luglio 2017

Anno Accademico 2016-2017

Indice

Introduzione

Considerato un insieme $\Omega \subset \mathbb{R}^d$ e μ una misura su di esso, desideriamo valutare per ogni funzione continua e μ-integrabile $f : \Omega \to \mathbb{R}$, l'integrale

$$I(f) = \int_\Omega f d\mu.$$

Dipendentemente da f e μ, la determinazione di $I(f)$ potrebbe risultare difficile analiticamente, giustificando quindi la necessità di approssimarne il valore numericamente mediante *formule di cubatura*. Queste formule sono descritte come una combinazione lineare di valutazioni di f in N punti, ovvero

$$I(f) = \int_\Omega f d\mu \approx \sum_{k=1}^{N} w_k f(t_k) \tag{1}$$

dove t_k e w_k sono detti rispettivamente i *nodi* e i *pesi* della formula.

Nel caso univariato, per $\Omega \subset \mathbb{R}$ compatto, si parla di "*formule di quadratura*" e si è cercato storicamente di determinare quelle con un numero minimo di nodi N per un certo grado di esattezza polinomiale prefissato m. In altre parole, dato un qualsiasi polinomio p appartenente allo spazio vettoriale $\mathbb{P}_m$ dei polinomi di grado m, si richiede che la formula determini in modo esatto il valore $I(p)$, ovvero

$$I(p) = \sum_{k=1}^{N} w_k p(t_k).$$

Dato $m = 2N - 1$, è noto da un teorema di Gauss che tali formule esistano per molte misure μ e domini Ω, quali intervalli limitati ma anche semirette o rette reali. Una loro peculiarità è che i pesi $\{w_k\}$ sono positivi, aspetto essenziale per questioni di stabilità numerica, ed inoltre non esistono formule capaci di raggiungere tale grado di precisione m con un numero minore di punti. Per determinare queste formule sono noti in letteratura molti metodi come quello di Golub-Welsch [11], Glaser-Liu-Rokhlin [5] e più recentemente quello descritto nell'articolo [12].

Nel caso multivariato il problema è ben più complesso e le formule dipendono fortemente dal dominio Ω e dalla misura μ. Nel 1957, nella sua opera intitolata

Formules de cubature mécaniques à coefficients non négatifs, V. Tchakaloff dimostrò l'esistenza di almeno una formula di cubatura con grado di precisione m, assumendo che Ω sia compatto e μ sia positiva ed assolutamente continua rispetto la misura di Lebesgue. In particolare, provò che tra di esse ne esiste almeno una con pesi positivi e descritta da al più n nodi appartenenti alla regione, dove n indica la dimensione dello spazio vettoriale $\mathbb{P}_m^d$. Tuttavia per arrivare ad una prima descrizione costruttiva per la misura di Lebesgue, si dovrà attendere il lavoro di P.J. Davis nell'articolo *A construction of nonnegative approximate quadratures* del 1967. Successivamente altri metodi furono pubblicati per misure più generali da Davis, Wilson, Dannis e Goldstein. Dal punto di vista teorico, si proposero in seguito estensioni al Teorema di Tchakaloff: nel 1975 I.P. Mysovskikh provò l'enunciato anche per domini non limitati mentre nel 1997 M. Putinar determinò l'estensione per misure di Borel positive.

In questa tesi ci proponiamo di studiare un metodo per ottenere formule di cubatura per misure di Borel positive su domini compatti a partire dai momenti di una qualsiasi base dello spazio $\mathbb{P}_m^d$. L'idea che intendiamo discutere è ripresa dall'articolo *Extensions of Gauss Quadrature Via Linear Programming* pubblicato nel 2015 da E. K. Ryu e S. P. Boyd, in cui si mostrano alcuni esempi senza però descrivere esplicitamente il metodo. In particolare intendiamo fare un'analisi critica dell'articolo mostrandone alcune lacune di natura computazionale.

I metodi utilizzati al momento non sono da intendersi come stato dell'arte. Ciò nonostante risultano un'interessante e poco nota alternativa a quanto usualmente utilizzato dai ricercatori, evidenziando nuovi filoni di studio. Il punto chiave per la determinazione di formule di cubatura è la risoluzione di un problema di programmazione lineare descritto da un insieme finito di vincoli e da una variabile che appartiene ad uno spazio di dimensione infinita. Questi problemi sono noti nell'ambito dell'ottimizzazione matematica come *Linear Semi-Infinite Programs*, o comunemente denotati con l'acronimo *LSIP*, di cui diamo una introduzione tecnica nel Capitolo 1. Il Capitolo 2 darà spazio ad una spiegazione più esauriente di questa intepretazione mentre nel Capitolo 3 proporremo un algoritmo esplicito sulla base della teoria più generale degli LSIP, presentando infine i risultati ottenuti nel Capitolo 4.

Capitolo 1

Introduzione ai LSIP

Un *Linear Semi-Infinite Program* (denotato con l'acronimo *LSIP*) è un problema di ottimizzazione dove la funzione obbiettivo e i vincoli che lo formalizzano sono lineari, tuttavia esso o è costituito da un numero non finito di vincoli oppure il suo spazio dei parametri ha dimensione infinita.

Possiamo pertanto interpretare i LSIP come una generalizzazione degli ordinari problemi di *Programmazzione Lineare* (*LP*), nei quali le quantità sopra coinvolte sono finite. La teoria di questi ultimi problemi è stata ampiamente studiata durante il ventesimo secolo ed ha portato il matematico G.B. Dantzig nel 1947 alla formulazione dell'algoritmo del simplesso, ancor oggi classificato come uno dei risultati matematici più importanti dello scorso secolo. Le numerose applicazioni dirette dei LSIP e le proprietà teoriche ad essi collegate, spiegano il perchè tali problemi costituiscano un ambito attivo della ricerca operativa fin dai primi anni '60 grazie ai lavori di Charnes, Cooper e Kortanek. Tuttavia si tratta di una matematica "giovane" che ha ampio margine di sviluppo, in particolar modo sotto l'aspetto numerico.

Ci baseremo sui lavori monografici pubblicati da M. A. Goberna e M. A. López (rispettivamente nel 1998 [8] e nel 2014 [9]), per presentare gli elementi necessari alla nostra discussione.

1.1 Forma Primale e Forma Duale

Definizione 1.1 (Problema Primale)**.** *Si definisce* Problema in forma primale *il problema di ottimizzazione formulato come segue:*

$$\begin{aligned} \mathcal{P}: \quad \inf_{x\in\mathbb{R}^n} \quad & \langle c, x\rangle \\ \text{s.a} \quad & \langle a(t), x\rangle \geq b(t),\ \forall t \in T \end{aligned} \tag{1.1}$$

dove $c \in \mathbb{R}^n$, T è un insieme di cardinalità non finita, $a(t) = (a_1(t), \ldots, a_n(t))^T$ è una mappa $T \to \mathbb{R}^n$, $b(t)$ è una funzione scalare $T \to \mathbb{R}$ e $\langle \cdot, \cdot \rangle$ denota il prodotto scalare usuale.

Come per i LP, dato un problema primale $\mathcal{P}$, possiamo definire differenti forme duali. Nel seguito, enunciamo e focalizziamo la nostra attenzione alla più diffusa in letteratura introducendo preliminarmente la definizione di *generalized finite sequences*.

Definizione 1.2. *Dato un insieme $X \subseteq \mathbb{R}$ tale che $0 \in X$, denotiamo con $X^{(T)}$ lo spazio di tutte le funzioni $\lambda : T \to X$ tale che il suo supporto*

$$supp\ \lambda := \{t \in T : \lambda(t) \neq 0\}$$

abbia cardinalità finita. Gli elementi di $X^{(T)}$ sono chiamati generalized finite sequences *in X e denoteremo la loro valutazione in t attraverso il simbolo λ_t.*

A titolo storico, vogliamo ricordare che il primo lavoro sui LSIP redatto dai matematici A. Charnes, W.W. Cooper e K. Kortanek nel 1962, impiegò un risultato del 1924 ad opera di A. Haar sui sistemi semi-infiniti di disequazioni lineari. In riconoscenza di tale impiego, il problema duale da loro principalmente discusso assunse il nome di problema duale nel senso di Haar.

Definizione 1.3. *Si definisce* Problema duale nel senso di Haar *il problema di ottimizzazione formulato come segue:*

$$\begin{aligned} \mathcal{D}: \quad \sup_{\lambda\in\mathbb{R}_+^{(T)}} \quad & \sum_{t\in T} \lambda_t\, b(t) \\ s.a \quad & \sum_{t\in T} \lambda_t\, a(t) = c, \end{aligned} \tag{1.2}$$

Si osservi che anch'esso appartiene alla classe dei problemi LSIP poiché lo spazio $\mathbb{R}_+^{(T)}$ in cui varia il parametro ha dimensione infinita mentre il numero di vincoli è finito. Ci rivolgeremo spesso a questo problema indicandolo semplicemente come *Problema Duale*.

In conclusione a questa sezione, vogliamo riproporre gli elementi fondamentali che costituiscono lo studio di un generico problema di ottimizzazione.

Definizione 1.4. *Si definisce* insieme delle soluzioni ammissibili *di un problema di ottimizzazione, l'insieme dei parametri che soddisfano i vincoli dati.*
Durante la nostra discussione denoteremo come segue tali insiemi:

$$F := \{x \in \mathbb{R}^n : \langle a(t), x\rangle \geq b(t)\ \forall t \in T\},$$

$$\Lambda := \{\lambda \in \mathbb{R}_+^{(T)} : \sum_{t\in T} \lambda_t\, a(t) = c\},$$

rispettivamente l'insieme delle soluzioni ammissibili di $\mathcal{P}$ e $\mathcal{D}$.
Associamo al simbolo $\nu(\mathcal{P})$ il valore ottimo *del problema di ottimizzazione $\mathcal{P}$ mentre con F^* indichiamo l'*insieme delle soluzioni ottime *ovvero l'insieme delle soluzioni ammissibili che lo realizzano. Con $\nu(\mathcal{D})$ e Λ^* si denotano gli analoghi elementi del problema $\mathcal{D}$.*

Si osservi, a titolo divulgativo, che l'insieme F è un sottoinsieme chiuso e convesso di $\mathbb{R}^n$ perché costituito dall'intersezione di famiglie di semi-spazi chiusi. Ciò permette di asserire che $\mathcal{P}$ è un problema di ottimizzazione convessa.

1.2 Teoremi di dualità

In questa sezione siamo interessati al legame che intercorre tra $\mathcal{P}$ e $\mathcal{D}$ e a questo scopo introduciamo la funzione *duality gap*,

$$\delta(\mathcal{P}, \mathcal{D}) := \nu(\mathcal{P}) - \nu(\mathcal{D})$$

Nostro è l'obbiettivo di determinare delle condizioni che garantiscano la nullità della mappa $\delta(\mathcal{P}, \mathcal{D})$. Impiegheremo alcuni sviluppi della cosiddetta teoria della dualità che ha radici nella teoria dei sistemi di disequazioni, in quella sviluppata dal problema classico dei momenti e nella teoria dell'approssimazione uniforme delle funzioni [6, pag.1].

1.2.1 Dualità debole

Sarà importante nel seguito della nostra discussione la classificazione dei problemi di ottimizzazione in base all'esistenza e la cardinalità dell'insieme delle sue soluzioni ottime [7].

Definizione 1.5. *Un problema di ottimizzazione si dice* inconsistente *se l'insieme delle soluzioni ammissibili è vuoto. Esso si dice* limitato *se il problema è consistente mentre il valore ottimo è finito. Infine esso si dice* illimitato *se il problema è consistente tuttavia il valore ottimo non è finito.*

Teorema 1.1. *Siano $\mathcal{P}$ il problema definito in* (1.1) *e $\mathcal{D}$ il problema definito in* (1.2)*. Allora la coppia $(\mathcal{P}, \mathcal{D})$ soddisfa la proprietà di* dualità debole

$$\nu(\mathcal{D}) \leq \nu(\mathcal{P}) \tag{1.3}$$

con la convenzione che venga posto $\nu(\mathcal{P}) = +\infty$ quando il problema $\mathcal{P}$ è inammissibile e $\nu(\mathcal{D}) = -\infty$ quando lo è $\mathcal{D}$.

Dimostrazione. Supponiamo inizialmente che entrambi i problemi $\mathcal{P}$ e $\mathcal{D}$ siano ammissibili. Perciò possiamo considerare una soluzione ammissibile arbitraria del problema $\mathcal{P}$, $\overline{x} \in F = \{x \in \mathbb{R}^n : \langle a(t), x\rangle \geq b_t \ \forall t \in T\}$, e una del problema $\mathcal{D}$, $\overline{\lambda} \in \Lambda = \{\lambda \in \mathbb{R}_+^{(T)} : \sum_{t\in T} \lambda_t\, a(t) = c\}$. Allora vale quanto segue:

$$\langle c, \overline{x}\rangle = \langle \sum_{t\in T} \lambda_t\, a(t), \overline{x}\rangle = \sum_{t\in T} \lambda_t\, \langle a(t), \overline{x}\rangle \geq \sum_{t\in T} \lambda_t\, b(t)$$

Pertanto, per l'arbitrarietà che ci siamo concessi, è possibile asserire che

$$\nu(\mathcal{D}) = \sup_{\lambda\in\Lambda} \sum_{t\in T} \lambda_t\, b_t \leq \inf_{x\in F} \langle c, x\rangle = \nu(\mathcal{P})$$

Infine, grazie alla convenzione sopra esposta, è naturale che tale relazione sia estesa anche per i problemi non ammissibili. □

I possibili valori assunti dalla mappa $\delta(\mathcal{P}, \mathcal{D})$ in base alla classificazione sopra menzionata dei problemi $\mathcal{P}$ e $\mathcal{D}$, sono deducibili interpretando la tabella illustrata in [8, pag. 51] che riportiamo di seguito. Si convenga che gli spazi vuoti vogliano indicare le combinazioni non possibili.

		$\mathcal{D}$		
		INCONSISTENTE	LIMITATO	ILLIMITATO
$\mathcal{P}$	INCONSISTENTE	$+\infty$	$+\infty$	0
	LIMITATO	$+\infty$	$\delta \geq 0$	
	ILLIMITATO	0		

Tabella 1.1: possibili valori di $\delta(\mathcal{P}, \mathcal{D})$

Dalla Tabella 1.1 possiamo osservare come la mappa $\delta(\mathcal{P}, \mathcal{D})$ possa annullarsi nel caso consistente solamente se entrambi i problemi sono limitati. In questa tesi non utilizzeremo direttamente questa affermazione e pertanto ci giustifichiamo dal non dimostrarla anche se risulta in modo immediato dal teorema precedente.

1.2.2 Dualità forte

Allo scopo di discutere dei criteri di sufficienza tali da garantire la proprietà di dualità forte, siamo interessati ad approfondire la geometria su cui è definito il problema $\mathcal{D}$. Introduciamo i seguenti insiemi immagine di T tramite le funzioni $a(\cdot)$ e $b(\cdot)$:

$$A_T := \{a(t) : t \in T\} \subset \mathbb{R}^n$$
$$\tilde{A}_T := \left\{ \begin{pmatrix} b(t) \\ a(t) \end{pmatrix} : t \in T \right\} \subset \mathbb{R}^{n+1}$$

Consideriamo allora gli inviluppi conici convessi di entrambi gli insiemi (si veda la Definizione A.3 ed in particolare la relazione (A.2)). Tenendo conto della Definizione 1.2 possiamo riscrivere come segue la loro forma:

$$\begin{aligned} \mathcal{M} :=& CC(A_T) \qquad (1.4) \\ =& \left\{ \sum_{t \in T} \lambda_t a(t) : \lambda \in \mathbb{R}_+^{(T)} \right\} \\ \mathcal{N} :=& CC(\tilde{A}_T) \qquad (1.5) \\ =& \left\{ \sum_{t \in T} \lambda_t \begin{pmatrix} b(t) \\ a(t) \end{pmatrix} : \lambda \in \mathbb{R}_+^{(T)} \right\} \end{aligned}$$

Tali insiemi sono rispettivamente chiamati *first moment cone* e *second moment cone*. Grazie ad essi possiamo dare una formulazione equivalente del problema $\mathcal{D}$ nei termini di questi due insiemi.

Una prima osservazione triviale è che il problema $\mathcal{D}$ è consistente se, e solo se, $c \in \mathcal{M}$. Inoltre è altrettanto immediato determinare il valore ottimo del problema $\mathcal{D}$ cercando tra le soluzioni ammissibili dell'insieme $\mathcal{N}$. Attraverso questa idea possiamo proporre la seguente formulazione geometrica del problema duale $\mathcal{D}$ nei termini di $\mathcal{N}$:

$$\begin{aligned} \mathcal{D}_G : \quad & \sup_{(c_0,c)^T \in \mathbb{R}^{n+1}} \quad c_0 \\ & \text{s.a} \qquad \begin{pmatrix} c_0 \\ c \end{pmatrix} \in \mathcal{N} \end{aligned} \qquad (1.6)$$

Ci proponiamo ora di dare seguito a due teoremi che attraverso condizioni sufficienti distinte ci garantiscano la proprietà di dualità forte per i problemi $\mathcal{P}$ e $\mathcal{D}$. Ovvero che permettano di asserire che $\nu(\mathcal{P}) = \nu(\mathcal{D})$.

Teorema 1.2 (Teorema di Dualità Forte - prima versione)**.** *Siano $\mathcal{P}$ il problema definito in* (1.1) *e $\mathcal{D}$ il problema definito in* (1.2)*. Supponiamo che il problema $\mathcal{D}$ sia limitato e che il cono convesso, $\mathcal{N}$, sia chiuso. Allora il problema $\mathcal{P}$ è limitato e vale la proprietà di* dualità forte

$$\nu(\mathcal{D}) = \nu(\mathcal{P})$$

Dimostrazione. La dimostrazione che segue fa riferimento a [7, pag. 79].

Per la condizione di limitatezza del problema $\mathcal{D}$, possiamo osservare che $\nu(\mathcal{D}) < +\infty$. Pertanto, data una soluzione ottima del problema $\mathcal{D}_G$, ovvero un punto di coordinate $z = (\nu(\mathcal{D}), c_1, \ldots, c_n)^T$, è ben definito per ogni scelta di $\Delta > 0$ il punto $z_\Delta := (\nu(\mathcal{D}) + \Delta, c_1, \ldots, c_n)^T$. Si osservi che necessariamente z_Δ non può appartenere all'insieme $\mathcal{N}$, altrimenti verrebbe violata l'ottimalità di z. Poiché $\mathcal{N}$ è un cono convesso non vuoto (per la consistenza del problema $\mathcal{D}$) e chiuso per ipotesi, possiamo applicare il Corollario A.2 sul punto z_Δ. Esistono perciò dei coefficienti $y_0, y_1, \ldots, y_n$ tali che

$$\begin{cases} y_0(\nu(\mathcal{D}) + \Delta) & + \sum_{i=1}^n y_i\, c_i > 0 \\ y_0 x_0 & + \sum_{i=1}^n y_i\, x_i \leq 0,\ \forall x \in \mathcal{N} \end{cases} \tag{1.7}$$

Banalmente, se z è un punto di frontiera dell'insieme chiuso $\mathcal{N}$, allora z appartiene ad $\mathcal{N}$. Applicando quindi la seconda disuguaglianza su questo punto si ottiene:

$$\left(y_0\, \nu(\mathcal{D}) + \sum_{i=1}^n y_i\, c_i \right) \leq 0 < y_0 \Delta + \left(y_0\, \nu(\mathcal{D}) + \sum_{i=1}^n y_i\, c_i \right)$$

da cui deduco che $y_0 > 0$.
In particolare, fissato $t \in T$, è facile mostrare che il punto $(b(t), a_1(t), \ldots, a_n(t))^T$ appartiene ad $\mathcal{N}$ prendendo la mappa $\lambda \in \mathbb{R}_+^{(T)}$ che si annulla ovunque tranne nel punto t dove vale 1. Applicando al variare di $t \in T$ la seconda disuguaglianza del sistema (1.7), si ottiene:

$$y_0\, b(t) + \sum_{i=1}^n y_i\, a_i(t) \leq 0,\ \forall t \in T$$

equivalentemente

$$\sum_{i=1}^n y_i\, a_i(t) \leq -y_0\, b(t),\ \forall t \in T$$

Allora dividendo entrambi i membri per $-y_0 < 0$ si ottiene la seguente relazione:

$$\sum_{i=1}^n -\frac{y_i}{y_0}\, a(t) \geq b(t),\ \forall t \in T$$

Quindi il vettore $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_n)$, dove $\overline{x}_i = -\frac{p_i}{p_0}$, risulta essere una soluzione ammissibile per $\mathcal{P}$. Per la prima disuguaglianza del sistema (1.7) si ha:

$$\sum_{i=1}^{n} c_i \, \overline{x}_i < \nu(\mathcal{D}) + \Delta$$

Essendo $\overline{x}$ una soluzione ammissibile del problema di minimo $\mathcal{P}$, allora il valore della funzione obbiettivo ad essa associata sarà superiore. Pertanto, applicando l'asserto appena dimostrato e il Teorema 1.1, si ottiene

$$\nu(\mathcal{P}) \leq \sum_{i=1}^{n} c_i \, \overline{x_i} < \nu(\mathcal{D}) + \Delta \leq \nu(\mathcal{P}) + \Delta$$

equivalentemente

$$\nu(\mathcal{P}) - \Delta \leq \nu(\mathcal{D}) \leq \nu(\mathcal{P})$$

Per l'arbitrarietà di $\Delta > 0$, la proprietà di dualità forte è verificata. □

Volendo enunciare il secondo teorema di dualità forte e darne una forma con condizioni più deboli, introduciamo la seguente nozione topologica.

Definizione 1.6. *Considerato un insieme E contenuto in uno spazio tolopologico X, definiamo l'*interno relativo *di E, rint E, l'interno di E nel più piccolo insieme affine che lo contiene.*

Illustriamo un rapido esempio per rendere consistente la nostra discussione. Se definiamo un disco unitario immerso in uno spazio $\mathbb{R}^3$, dotato della topologia naturale,

$$R = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 \leq 1 \text{ e } z = 0\}$$

si avrebbe int $R = \emptyset$ mentre

$$\text{rint } R = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 < 1 \text{ e } z = 0\}.$$

Viceversa, se l'insieme ammette interno topopologico, allora la nozione di interno relativo coincide ad esso. Questo aspetto non è significativo nella dimostrazione tuttavia amplia l'insieme dei problemi LSIP a cui il teorema può essere applicato.

La seconda versione del teorema di dualità forte che esponiamo in questa discussione "costituisce un'importante sviluppo della teoria dei momenti e fu dimostrata indipendentemente nel 1960 da JSII e Karlin" cit. [6, pag. 4].

Teorema 1.3 (Teorema di Dualità Forte - seconda versione)**.** *Siano $\mathcal{P}$ il problema definito in* (1.1) *e $\mathcal{D}$ il problema definito in* (1.2)*. Supponiamo che il problema $\mathcal{D}$ sia limitato e il vettore c appartenga all'interno relativo di $\mathcal{M}$, $c \in rint\ (\mathcal{M})$. Allora il problema $\mathcal{P}$ è limitato e vale la proprietà di* dualità forte

$$\nu(\mathcal{D}) = \nu(\mathcal{P}) \tag{1.8}$$

Dimostrazione. La dimostrazione che segue fa riferimento a quella proposta in [6, pag. 8]- [7, pag.85] integrata da [13, pag. 473].

Senza perdere di generalità, possiamo supporre che il cono $\mathcal{N}$ (e di conseguenza $\mathcal{M}$) non sia contenuto in un sottospazio più piccolo. In caso contrario ci possiamo restringere ad esso. Allora possiamo limitare la nostra dimostrazione nell'ipotesi in cui $c \in \text{int } \mathcal{M}$.
Per la definizione della forma duale geometrica $\mathcal{D}_G$, $s := (\nu(D), c_1, \ldots, c_n)^T$ è un punto di frontiera dell'insieme $\mathcal{N}$. Osservando inoltre che $\mathcal{N}$ è un cono convesso non vuoto (per la consistenza del problema $\mathcal{D}$), possiamo applicare il Corollario A.3 sul punto s. Allora esistono dei coefficienti $y_0, y_1, \ldots, y_n$ tali che

$$\begin{cases} y_0\,\nu(\mathcal{D}) & + \sum_{i=1}^{n} y_i\,c_i = 0 \\ y_0\,x_0 & + \sum_{i=1}^{n} y_i\,x_i \leq 0,\ \forall x \in \mathcal{N} \end{cases} \tag{1.9}$$

Ci proponiamo ora di dimostrare che $p_0 > 0$.
Fissato uno scalare Δ strettamente positivo, definiamo il punto dello spazio $\mathbb{R}^{n+1}$, $s_\Delta := (\nu(\mathcal{D}) - \Delta, c_1, \ldots, c_n)$. È evidente che s_Δ appartiene al semispazio inferiore desritto dall'iperpiano $\mathcal{H}_{n+1}((y_0, \ldots, y_n)^T, 0)$ che descrive l'iperpiano di supporto di $\mathcal{N}$ in s. Pertanto si otterrà per definizione

$$y_0\,\nu(\mathcal{D}) - y_0\,\Delta + \sum_{i=1}^{n} y_i\,c_i \leq 0$$

Da tale relazione, applicando l'equazione del sistema (1.9), si deduce $p_0 \geq 0$. Ora supponiamo per assurdo che $p_0 = 0$, quindi il sistema (1.9) può essere così riscritto

$$\begin{cases} \sum_{i=1}^{n} y_i\,c_i = 0 \\ \sum_{i=1}^{n} y_i\,x_i \leq 0,\ \forall x \in \mathcal{M} \end{cases} \tag{1.10}$$

Esso però coincide con la definizione di un iperpiano di supporto di M in $c \in M$. Ma per la Proposizione A.2 ciò non è possibile e arriviamo all'assurdo cercato.

Allora dividendo ciascun elemento del sistema (1.9) per $-p_0 < 0$, si ottiene il sistema

$$\begin{cases} \sum_{i=1}^{n} -\frac{y_i}{y_0}\,c_i = \nu(\mathcal{D}) \\ \sum_{i=1}^{n} -\frac{y_i}{y_0}\,x_i \geq x_0,\ \forall x \in \mathcal{N} \end{cases} \tag{1.11}$$

In particolare, fissato $t \in T$, è semplice mostrare che il punto $(b(t), a_1(t), \ldots, a_n(t))^T$ appartiene ad $\mathcal{N}$ prendendo banalmente la mappa $\lambda \in \mathbb{R}_+^{(T)}$ che si annulla ovunque tranne in t. Applicando al variare di $t \in T$ la disuguaglianza del sistema (1.11), si ottiene il nuovo sistema

$$\begin{cases} \sum_{i=1}^n -\frac{y_i}{y_0}\, c_i = \nu(\mathcal{D}) \\ \sum_{i=1}^n -\frac{y_i}{y_0}\, a_i(t) \geq b(t),\ \forall t \in T \end{cases} \tag{1.12}$$

Queste due relazioni permettono di asserire che il vettore $\overline{x} = (\overline{x_1}, \ldots, \overline{x_n})$, dove $\overline{x_i} = -\frac{p_i}{p_0}$, è una soluzione ammissibile per il problema $\mathcal{P}$ che raggiunge il lower-bound imposto dal Teorema 1.1. Da ciò possiamo desumere che $\overline{x}$ è soluzione ottima del problema $\mathcal{P}$ a cui corrisponde il valore ottimo $\nu(\mathcal{P}) = \nu(\mathcal{D})$. □

1.3 Legame tra le soluzioni ottime

In questa sezione indendiamo comprendere il legame tra la soluzione ottima del problema $\mathcal{P}$ e quella del problema $\mathcal{D}$ quando vale la proprietà di dualità forte.

Teorema 1.4. *Siano $\mathcal{P}$ il problema definito in* (1.1) *e $\mathcal{D}$ il problema definito in* (1.2) *tali da soddisfare la proprietà di dualità forte.*
Consideriamo $x^* \in \mathbb{R}^n$ e $\lambda^* \in \mathbb{R}_+^{(T)}$ soluzioni ammissibili dei problemi $\mathcal{P}$ e $\mathcal{D}$. Allora x^* e λ^* sono soluzioni ottime dei rispettivi problemi se, e solo se, per ogni punto appartenente al supporto di λ^*, $\tilde{t}$, si ha

$$\langle a(\tilde{t}), x^* \rangle = b(\tilde{t}) \tag{1.13}$$

Dimostrazione. Dalle condizioni di ammissibilità di ciascun problema ricaviamo che $x^* \in \mathbb{R}^n$ e $\lambda^* \in \mathbb{R}_+^{(T)}$ sono soluzioni ammissibili se, e solo se, valgono le seguenti condizioni:

$$\begin{cases} \langle a(t), x^* \rangle \geq b(t)\ \forall t \in T \\ \sum_{t \in T} \lambda_t^*\, a(t) = c \end{cases} \tag{1.14}$$

Poiché vale la proprietà di dualità forte tra i problemi $\mathcal{P}$ e $\mathcal{D}$, deduciamo che $\nu := \nu(\mathcal{P}) = \nu(\mathcal{D})$. Allora le soluzioni ammissibili x^* e λ^* sono soluzioni ottime se, e solo se, valgono le seguenti condizioni

$$\begin{cases} \langle c, x^* \rangle = \nu \\ \sum_{t \in T} \lambda_t^*\, b(t) = \nu \end{cases}$$

Tuttavia, riprendendo la dimostrazione del Teorema 1.1, è evidente che tale sistema è equivalente alla relazione più semplice

$$\langle c, x^* \rangle = \sum_{t \in T} \lambda_t^*\, b(t) \tag{1.15}$$

Da essa si ricava, attraverso semplici passaggi algebrici, le seguenti uguaglianze

$$\langle c, x^* \rangle - \sum_{t \in T} \lambda_t^* \, b(t) \overset{(1.14)}{=} \langle \sum_{t \in T} \lambda_t^* \, a(t), x^* \rangle - \sum_{t \in T} \lambda_t^* \, b(t)$$
$$= \sum_{t \in T} \lambda_t^* \left(\langle a(t), x^* \rangle - b(t) \right)$$

Da tale relazione segue immediatamente che se x^* e λ^* sono soluzioni ottime per i rispettivi problemi allora, per la relazione (1.15), si deve avere

$$\sum_{t \in T} \lambda_t^* \left(\langle a(t), x^* \rangle - b(t) \right) = 0$$

Osservando però che la mappa λ_t^* è non-negativa e per la prima condizione del sistema (1.14) lo è anche l'argomento della sommatoria, si deve avere

$$\langle a(\tilde{t}), x \rangle - b(\tilde{t}) = 0, \ \forall \tilde{t} \in \text{supp } \lambda^*$$

Viceversa, se vale la relazione appena descritta, allora necessariamente si ha verificata la relazione (1.15) che prova l'ottimalità delle soluzioni. □

Allora deduciamo la seguente condizione di ottimalità per una soluzione ammissibile del problema primale $\mathcal{P}$ [8, pag.255]. Essa avrà un ruolo chiave nell'algoritmo numerico discusso nella sezione 3.2.

Corollario 1.1. *Siano $\mathcal{P}$ il problema definito in* (1.1) *e $\mathcal{D}$ il problema definito in* (1.2) *tali da soddisfare la proprietà di dualità forte.*
Allora condizione necessaria e sufficiente affinché una soluzione ammissibile del problema $\mathcal{P}$, $x \in \mathbb{R}^n$, sia una soluzione ottima di $\mathcal{P}$ è l'esistenza di una soluzione ammissibile del problema duale $\mathcal{D}$ tale che ogni punto del suo supporto soddisfi la relazione (1.13).

1.4 Problemi discretizzabili

Formalizziamo ora l'esistenza di una classe di problemi che godono di proprietà particolari. Si osserverà nel Capitolo 3 che i metodi numerici più efficaci per risolvere i LSIP richiedono l'appartenenza a questa classe. Si osservi che qualora un problema non fosse discretizzabile, esistono due metodi per trattarlo. Il primo consiste nel considerare una perturbazione della funzione obbiettivo. Il secondo è invece determinato dall'intersezione dell'insieme ammissibile con un politopo sufficiente grande e vicino all'origine [10, pag. 188].

Definizione 1.7. *Un problema $\mathcal{P}$ si dice* discretizzabile *se esiste una sequenza di sotto-problemi LP*

$$\begin{aligned} \mathcal{P}_k : \quad & \inf_{x\in\mathbb{R}^n} \quad \langle c, x\rangle \\ & \text{s.a} \quad \langle a(t), x\rangle \geq b(t),\ \forall t \in T_k, \end{aligned} \tag{1.16}$$

dove T_k sono sottoinsiemi finiti di T, $\forall k \in \mathbb{N}_0$, tali che

$$\nu(\mathcal{P}) = \lim_{k\to\infty} \nu(\mathcal{P}_k) \tag{1.17}$$

dove $\nu(\mathcal{P}_k)$ indica il valore ottimo del problema $\mathcal{P}_k$.
In particolare $\mathcal{P}$ si dice riducibile *se esiste un insieme finito $S \subset T$ tale che $\nu(\mathcal{P}) = \nu(\mathcal{P}_S)$, dove quest'ultimo denota il valore ottimo dell'ordinario LP che ha come insieme dei vincoli quelli di indice appartenente ad S.*

Teorema 1.5. *Siano $\mathcal{P}$ il problema definito in* (1.1) *e $\mathcal{D}$ il problema definito in* (1.2) *tali da soddisfare la proprietà di dualità forte.*
Allora il problema $\mathcal{P}$ è un problema discretizzabile.

Dimostrazione. Allo scopo di provare che $\mathcal{P}$ sia discretizzabile, consideriamo una sequenza di elementi di Λ, ovvero $\left\{\lambda^k\right\}_{k\in\mathbb{N}} \subset \Lambda$, tale che

$$\lim_{k\to\infty} \sum_{t\in T} \lambda_t^k\, b(t) = \nu(\mathcal{D}). \tag{1.18}$$

L'esistenza di tale successione è garantita dalla definizione di punto di frontiera osservando, come già detto nelle sezioni precedenti, che $(\nu(\mathcal{D}), c_1, \ldots, c_n)^T$ è un punto appartenente al bordo del cono convesso $\mathcal{N}$. Definiamo $T_k = \operatorname{supp} \lambda^k$ e $\mathcal{P}_k$ i corrispondenti sotto-problemi. È nostra intenzione provare che $\mathcal{P}_k$ è la successione di sotto-problemi finiti ricercata.

Osserviamo che per k fissato, si ha $\nu(\mathcal{P}_k) \geq \sum_{t\in T_k} \lambda_t^k\, b(t)$ per il teorema di dualità debole negli ordinari problemi PL. Mentre, selezionando solo una parte dei vincoli di un problema di ottimizzazione di minimo, si osserva facilmente che il valore ottimo sarà inferiore rispetto al problema iniziale, ovvero $\nu(\mathcal{P}) \geq \nu(\mathcal{P}_k)$. Allora si verifica che

$$\nu(\mathcal{P}) \geq \nu(\mathcal{P}_k) \geq \sum_{t\in T_k} \lambda_t^k\, b(t)$$

Passando al limite per k che tende a $+\infty$ ed utilizzando la relazione (1.18), si ottiene

$$\nu(\mathcal{P}) \geq \lim_{k\to\infty} \nu(\mathcal{P}_k) \geq \nu(\mathcal{D})$$

L'applicazione del principio di dualità forte ci permette di determinare che le quantità agli estremi coincidono e pertanto che $\lim_{k\to\infty} \nu(\mathcal{P}_k) = \nu(\mathcal{D})$. □

Capitolo 2

Formule di cubatura attraverso i LSIP

In questo capitolo intendiamo proporre la formulazione del *Linear Semi-Infinite Program* discusso da E. K. Ryu e S. P. Boyd nel loro articolo [15] pubblicato nel 2015. Grazie alle nozioni introdotte nel Capitolo 1, saremo in grado di dimostrare la bontà della definizione e una interpretazione duale altrettanto interessante nel campo dell'analisi numerica.

2.1 Definizione del problema LSIP

Sia $\mu : \mathcal{P}(\Omega) \to [0, +\infty]$ una misura Boreliana positiva sull'insieme $\Omega \subset \mathbb{R}^d$. Allora per una qualsiasi funzione $f : \Omega \to \mathbb{R}$ μ-integrabile, descriviamo una formula di cubatura a pesi positivi attraverso la definizione dei suoi N nodi $\{t_1, \dots, t_N\} \subset \Omega$ e i pesi a loro associati $w_i \geq 0$,

$$\int_\Omega f d\mu \approx \sum_{k=1}^{N} w_k f(t_k)$$

Per discutere in modo lineare le argomentazioni successive, preferiamo ridefinire la formula di cubatura nel seguente modo

$$\int_\Omega f d\mu \approx \sum_{t \in T} \lambda_t f(t) \tag{2.1}$$

dove $\lambda \in \mathbb{R}_+^{(\Omega)}$ è la *generalized finite sequence*

$$\lambda_t = \begin{cases} w_k & \text{se } t = t_k \\ 0 & \text{altrimenti} \end{cases} \tag{2.2}$$

Spesso ci riferiremo alla *formula di cubatura* λ intendendo la formula di cubatura descritta dalla (2.1).

Denotiamo inoltre con $\mathbb{P}_m^d$ lo spazio vettoriale costituito dai polinomi d-variati di grado totale al più m che ricordiamo avere dimensione

$$\dim_{\mathbb{R}} \mathbb{P}_m^d = \binom{m+d}{d}. \tag{2.3}$$

Tale quantità la richiameremo numerose volte attraverso la lettera n.

Definizione 2.1. *Diremo che la formula di cubatura λ è di ordine almeno m se è esatta per ogni polinomio di grado totale al più m.*
Equivalentemente, se $\nu_1, \dots, \nu_n$ costituisce una base per lo spazio $\mathbb{P}_m^d$, una formula di cubatura λ è di ordine almeno m se, e solo se, $\forall i = 1, \dots, n$ si ha

$$\underbrace{\int_\Omega \nu_i d\mu}_{c_i} = \sum_{t\in\Omega} \lambda_t \, \nu_i(t). \tag{2.4}$$

I termini c_i sono chiamati momento i-esimo. *Perciò ad ogni base di $\mathbb{P}_m^d$ possiamo associare il* vettore dei momenti $c := (c_1, \dots, c_n)^T$.

Allora, richiedere che una formula di cubatura λ sia di ordine almeno m, equivale ad imporre un numero $n = \binom{m+d}{d}$ di equazioni linearmente indipendenti, ovvero

$$\begin{cases} \sum_{t\in\Omega} \lambda_t \, \nu_1(t) = c_1 \\ \dots \\ \sum_{t\in\Omega} \lambda_t \, \nu_i(t) = c_i \\ \dots \\ \sum_{t\in\Omega} \lambda_t \, \nu_n(t) = c_n \end{cases} \tag{2.5}$$

Definendo la funzione vettoriale $\nu : \Omega \to \mathbb{R}^n$ che ha come componenti i polinomi che costituiscono la base

$$\nu(t) = \begin{pmatrix} \nu_1(t) \\ \nu_2(t) \\ \dots \\ \nu_n(t) \end{pmatrix},$$

possiamo riscrivere la (2.5) in notazione vettoriale

$$\sum_{t\in\Omega} \lambda_t \, \nu(t) = c. \tag{2.6}$$

Allora

$$\Lambda := \left\{ \lambda \in \mathbb{R}_+^{(\Omega)} : \sum_{t\in\Omega} \lambda_t \, \nu(t) = c \right\} \tag{2.7}$$

è l'insieme di tutte le formule di cubatura λ_t di ordine almeno m con pesi positivi. Ricordiamo che per garantire la stabilità di una formula di cubatura, quest'ultima ipotesi è di fondamentale importanza. Altrettanto importante è la cardinalità dei nodi poiché per molteplici aspetti è interessante ridurre quanto possibile la loro numerosità.

Definizione 2.2. *Fissato m, diremo che una formula di cubatura λ di ordine m è* efficiente *se la cardinalità del suo supporto è strettamente minore della dimensione dello spazio $\mathbb{P}_n^d$, ovvero $n > |supp\ \lambda|$. Equivalentemente, definito il* grado di efficienza *della formula di cubatura λ*

$$\rho := \frac{|supp\ \lambda|}{dim\ \mathbb{P}_m^d},$$

diremo che la formula di cubatura λ è efficiente se, e solo se, $\rho < 1$.
Diremo invece che una formula di cubatura λ^* è minimale *per m se il suo supporto ha cardinalità minore tra tutte le formule di ordine m, ovvero*

$$|supp\ \lambda^*| = \min_{\lambda_t\in\Lambda} |supp\ \lambda|.$$

Seguendo l'approccio suggerito dall'articolo [15], viene introdotta una funzione $r : \Omega \to \mathbb{R}$ detta *funzione di sensibilità*. Ad essa si associa il seguente problema LSIP,

$$\begin{aligned} \mathcal{D}(r) : \quad \sup_{\lambda\in\mathbb{R}_+^{(\Omega)}} \quad & \sum_{t\in\Omega} \lambda_t \, r(t) \\ \text{s.a} \quad & \sum_{t\in\Omega} \lambda_t \, \nu(t) = c, \end{aligned} \tag{2.8}$$

ovvero si cerca tra tutte le formule di cubatura di grado almeno m, la formula che massimizza la valutazione integrale di r. Daremo successivamente una interpretazione più appropriata. Nel frattempo, in virtù di quanto costruito nel capitolo precedente, possiamo definire il problema nella forma primale associato a $\mathcal{D}(r)$:

$$\begin{aligned} \mathcal{P}(r) : \quad \inf_{x\in\mathbb{R}^n} \quad & \langle c, x\rangle \\ \text{s.a} \quad & \langle \nu(t), x\rangle \geq r(t), \ \forall t \in \Omega, \end{aligned} \tag{2.9}$$

2.2 Formulazione delle ipotesi

In questa sezione vogliamo discutere progressivamente le ipotesi da assumere al fine di garantire la consistenza del problema $\mathcal{D}(r)$ e la proprietà di dualità forte tra i problemi $\mathcal{P}(r)$ e $\mathcal{D}(r)$.

Innanzitutto ci permettiamo di riscrivere gli insiemi fondamentali che costituiscono la geometria del problema $\mathcal{D}(r)$, introdotti nella sezione 1.2.2:

$$A_\Omega := \{\nu(t) : t \in \Omega\} \subset \mathbb{R}^n \tag{2.10}$$

$$\tilde{A}_\Omega := \left\{ \begin{pmatrix} r(t) \\ \nu(t) \end{pmatrix} : t \in \Omega \right\} \subset \mathbb{R}^{n+1} \tag{2.11}$$

$$\begin{aligned} \mathcal{M} :=& CC(A_\Omega) \\ =& \left\{ \sum_{t\in\Omega} \lambda_t \, \nu(t) : \lambda \in \mathbb{R}_+^{(\Omega)} \right\} \end{aligned} \tag{2.12}$$

$$\begin{aligned} \mathcal{N} :=& CC(\tilde{A}_\Omega) \\ =& \left\{ \sum_{t\in\Omega} \lambda_t \begin{pmatrix} r(t) \\ \nu(t) \end{pmatrix} : \lambda \in \mathbb{R}_+^{(\Omega)} \right\} \end{aligned} \tag{2.13}$$

Appare evidente che confrontando la definizione delle soluzioni ammissibili del problema $\mathcal{D}(r)$ data in (2.7) con la definizione di $\mathcal{M}$, il problema $\mathcal{D}(r)$ è ammissibile se, e solo se, il vettore dei momenti c appartiene al cono convesso $\mathcal{M}$. Allora ci occupiamo di dimostrare la seguente proposizione.

Proposizione 2.1. *Sia $\Omega \subset \mathbb{R}^d$ un insieme non nullo e sia il cono convesso $\mathcal{M}$, definito in (2.12), un insieme chiuso. Siano inoltre $\nu_1(\cdot), \ldots, \nu_n(\cdot)$ funzioni μ-misurabili. Allora si deve avere che il vettore dei momenti c gli appartiene, ovvero che il problema $\mathcal{D}(r)$ è consistente.*

Dimostrazione. Supponiamo per assurdo che c non appartenga al cono convesso chiuso $\mathcal{M}$. Dalla prima ipotesi, $\Omega \neq \emptyset$, e ricordando che $\nu_1(\cdot), \ldots, \nu_n(\cdot)$ costituisce una base dello spazio vettoriale $\mathbb{P}_m^d$, possiamo dedurre che l'insieme A_Ω è non vuoto. Di conseguenza anche il suo inviluppo conico convesso, $\mathcal{M}$, sarà non vuoto. In virtù del Corollario A.2, possiamo affermare che esiste un iperpiano passante per l'origine che separa il punto $c \notin \mathcal{M}$ dal cono $\mathcal{M}$. Ovvero esistono dei coefficienti reali $y_1, \ldots, y_n$ tali che

$$\begin{cases} \sum_{i=1}^n y_i c_i > 0 \\ \sum_{i=1}^n y_i x_i \leq 0, \forall x \in \mathcal{M} \end{cases} \tag{2.14}$$

Sviluppando la prima disequazione si ottiene:

$$0 < \sum_{i=1}^n y_i \, c_i = \int_\Omega \sum_{i=1}^n y_i \, \nu_i(t) \, d\mu(t) \tag{2.15}$$

Allora deve esistere almeno un punto $\tilde{t} \in \Omega$ tale che

$$\sum_{i=1}^{n} y_i \, \nu_i(\tilde{t}) > 0 \tag{2.16}$$

Se per assurdo così non fosse, allora l'insieme $E := \{t \in \Omega : \sum_{i=1}^{n} y_i \nu_i(t) > 0\}$ sarebbe vuoto ed in particolare avrebbe misura nulla. Poiché $\nu_1(\cdot), \ldots, \nu_n(\cdot)$ sono funzioni μ-misurabili, lo è anche una loro combinazione lineare. Allora, per la proprietà della monotonia della misura μ, si ottiene

$$\int_{\Omega} \sum_{i=1}^{n} y_i \, \nu_i(t) \, d\mu(t) = \int_{\Omega \setminus E} \sum_{i=1}^{n} y_i \, \nu_i(t) \, d\mu(t) \leq 0$$

che vìola chiaramente la relazione (2.15) e prova l'esistenza del punto $\tilde{t}$.

Infine, fissato $t \in \Omega$, è facile constatare che il punto $x_t = (\nu_1(t), \ldots, \nu_n(t))^T$ appartiene al cono $\mathcal{M}$. Infatti basta prendere banalmente la mappa $\lambda \in \mathbb{R}_+^{(\Omega)}$ che ha t come unico punto di supporto. Allora, per la seconda disequazione del sistema (2.14), si ottiene che per ogni $t \in \Omega$ vale la relazione seguente:

$$\sum_{i=1}^{n} y_i \nu_i(t) \leq 0$$

Essa però contraddice l'esistenza di $\tilde{t} \in \Omega$ per la quale vale la relazione (2.16). Si giunge perciò all'assurdo cercato e si conclude che la tesi è verificata nelle ipotesi assunte. □

Per quanto invece riguarda il principio di dualità forte, nella sezione 1.2.2 abbiamo anticipato due diversi teoremi che grazie a distinte condizioni provano la relazione cercata. Ci proponiamo nel seguito di applicare la prima versione e a tale scopo, siamo interessati a studiare le seguenti asserzioni:

i) il cono convesso $\mathcal{N}$ è chiuso;

ii) il problema $\mathcal{D}(r)$ è limitato.

Si osservi che la prima richiesta coincide verosimilmente ad una condizione necessaria per la Proposizione 2.1 e ne dimostra l'importanza chiave nella nostra analisi. Definiamo di seguito una "condizione di regolarità" comunemente applicata nei teoremi di esistenza e quelli di dualità nella teoria dell'ottimizzazione.

Definizione 2.3. *Un problema $\mathcal{P}(r)$ è detto* superconsistente *se esiste un vettore $\overline{x} \in \mathbb{R}^n$ tale che*

$$\langle \nu(t), \overline{x} \rangle > r(t); \forall t \in \Omega$$

Tale condizione è anche detta Condizione di Slater*, e $\overline{x}$ chiamato* punto di Slater.

Grazie a tale definizione possiamo enunciare il teorema che segue.

Proposizione 2.2. *Supponiamo che Ω sia un insieme compatto di $\mathbb{R}^d$ e che le funzioni scalari $\nu_1(\cdot), \ldots, \nu_n(\cdot), r(\cdot)$ siano continue in Ω.*
Se $\mathcal{P}(r)$ è superconsistente, allora il cono $\mathcal{N}$ è chiuso.

Dimostrazione. La dimostrazione che segue fa riferimento a [7, pag. 71].

Sia z un elemento arbitrario della chiusura di $\mathcal{N}$, $z \in \overline{\mathcal{N}}$. La nostra tesi è verificata se riusciamo a dimostrare che esso appartiene anche a $\mathcal{N}$.

Per la relazione (2.12) sappiamo che $\mathcal{N} = CC(\tilde{A}_\Omega)$. Quindi ogni suo elemento può essere descritto tramite il prodotto degli elementi che costituiscono la coppia $(h, \alpha) \in \mathrm{Conv}(\tilde{A}_\Omega) \times \mathbb{R}_+$. Allora, per definizione di punto di chiusura, possiamo associare al punto z una successione $\{(h_i, \alpha_i)\}_{i\in\mathbb{N}}$ di elementi appartenenti al prodotto cartesiano $\mathrm{Conv}(\tilde{A}_\Omega) \times \mathbb{R}_+$ tali che

$$z = \lim_{i\to\infty} \alpha_i \, h_i$$

L'insieme $\tilde{A}_\Omega$ è un insieme compatto poiché ciascuna componente è immagine di un compatto attraverso una funzione continua. Per la Proposizione A.1 si deduce allora che $\mathrm{Conv}(\tilde{A}_\Omega)$ è anch'esso compatto. Possiamo perciò scegliere una sottosuccessione di $\{h_i\}_{i\in\mathbb{N}}$ che converge ad un vettore $h \in \mathrm{Conv}(\tilde{A}_\Omega)$. Senza perdere di generalità consideriamo come $\{(h_i, \alpha_i)\}_{i\in\mathbb{N}}$ tale sua sottosuccessione. Se la successione degli scalari $\{\alpha_i\}_{i\in\mathbb{N}}$ è limitata allora possiamo desumere che essa converga ad un certo $\alpha > 0$. Si ha perciò

$$z = \lim_{i\to\infty} \alpha_i \, h_i = \alpha \, h$$

che appartiene al cono $\mathcal{N}$ per la Definizione A.3 di cono convesso.
Se invece la successione degli scalari $\{\alpha_i\}_{i\in\mathbb{N}}$ non è limitata, vogliamo provare che la condizione di superconsistenza di $\mathcal{P}(r)$ viene violata. Se la successione non è limitata allora possiamo assumere che esiste una sottosuccessione tale che $\frac{1}{\alpha}$ converga a 0. Come prima, senza perdere di generalità, consideriamo come $\{(h_i, \alpha_i)\}_{i\in\mathbb{N}}$ tale sua sottosuccessione. Si otterrà che

$$h = \lim_{i\to\infty} h_i = \lim_{i\to\infty} \frac{1}{\alpha_i} \alpha_i \, h_i = \lim_{i\to\infty} \frac{1}{\alpha_i} \lim_{i\to\infty} \alpha_i \, h_i = 0 \, z = 0$$

Ciò significa che il vettore nullo appartiene a $\mathrm{Conv}(\tilde{A}_\Omega)$.
Per la Definizione A.2 di inviluppo convesso, deduciamo che esiste un intero $q \geq 1$, $t_1, \ldots, t_q$ e γ_i tali che

$$\sum_{j=1}^{q} \gamma_j \begin{pmatrix} r(t_j) \\ \nu(t_j) \end{pmatrix} = 0 \in \mathbb{R}^{n+1}$$

dove $\sum_{j=1}^{q} \gamma_j = 1$ e dove ciascuna componente è non-negativa.
Suddividendo questa relazione si ottiene il seguente sistema:

$$\begin{cases} \sum_{j=1}^{q} \gamma_j \, r(t_j) = 0 \\ \sum_{j=1}^{q} \gamma_j \, \nu_r(t_j) = 0 \in \mathbb{R}^n \end{cases}$$

Consideriamo allora un arbitrario vettore $y \in \mathbb{R}^n$.
Dall'ultimo sistema ricaviamo la seguente relazione:

$$\sum_{j=1}^{q} \gamma_j \left(\langle \nu(t_j), y \rangle - r(t_j) \right) = \langle \sum_{j=1}^{q} \gamma_j \, \nu(t_j), y \rangle - \sum_{j=1}^{q} \gamma_j \, r(t_j) = 0$$

Poiché le componenti γ_j sono non-negative, si deduce che

$$\langle \nu(t_j), y \rangle - r(t_j) = 0$$

in contraddizione con l'ipotesi che esiste un punto di Slater. □

Ora dobbiamo provare che il problema $\mathcal{D}(r)$ è limitato, ovvero che il suo valore ottimo, $\nu(\mathcal{D}(r))$, è finito.

Proposizione 2.3. *Siano $\mathcal{P}(r)$ e $\mathcal{D}(r)$ definiti in* (2.9) *e* (2.8)*. Se il problema $\mathcal{D}(r)$ è consistente e contemporaneamente il problema $\mathcal{P}(r)$ è superconsistente, allora il problema $\mathcal{D}(r)$ è limitato.*

Dimostrazione. Sia $\lambda \in \mathbb{R}_+^{(\Omega)}$ una arbitraria soluzione ammissibile del problema $\mathcal{D}(r)$. Si avrà pertanto che

$$\sum_{t \in \Omega} \lambda_t \, \nu(t) = c$$

Allora il valore associato alla soluzione λ, unito all'ipotesi di esistenza di un punto di Slater, x, per il problema $\mathcal{P}(r)$ ci fornisce la seguente relazione:

$$\begin{aligned} \nu_\lambda(\mathcal{D}(r)) &:= \sum_{t \in \Omega} \lambda_t \, r(t) \\ &< \sum_{t \in \Omega} \lambda_t \, \langle \nu(t), x \rangle \\ &= \langle \sum_{t \in \Omega} \lambda_t \, \nu(t), x \rangle \\ &= \langle c, x \rangle \end{aligned}$$

Da quest'ultima segue la seguente stima per il valore ottimo:

$$\nu(\mathcal{D}(r)) = \sup_{\lambda \in \mathbb{R}_+^{(\Omega)}} \nu_\lambda(\mathcal{D}(r)) \leq \langle c, x \rangle < +\infty$$

Una dimostrazione alternativa è proposta in [17][pag. 60] □

A seguito di questi enunciati ci permettiamo di sintetizzare le ipotesi che assumeremo da qui sino alla fine della discussione.

Definizione 2.4 (Ipotesi)**.** *Definiamo le seguenti ipotesi:*

i) $\Omega \subset \mathbb{R}^d$ *sia un insieme compatto;*

ii) le funzioni $\nu_1 \equiv 1, \nu_2(\cdot), \ldots, \nu_n(\cdot)$ *costituiscono una base dello spazio* $\mathbb{P}_m^d$*;*

iii) la funzione di sensibilità $r(\cdot)$ *è una funzione continua, linearmente indipendente dalla base* $\{\nu_1 \equiv 1, \nu_2(\cdot), \ldots, \nu_n(\cdot)\}$*,* μ*-misurabile e limitata superiormente nell'insieme* Ω*.*

Sotto le ipotesi appena descritte, possiamo enunciare il seguente corollario che sintetizza i risultati principali del capitolo precedente .

Corollario 2.1. *Siano* $\mathcal{P}(r)$ *il problema definito in* (2.9) *e* $\mathcal{D}(r)$ *quello definito in* (2.8)*. Supponiamo verificate le ipotesi della Definizione 2.4.*
Allora sono vere le seguenti affermazioni:

a) i problemi $\mathcal{P}(r)$ *e* $\mathcal{D}(r)$ *soddisfano la proprietà di dualità forte*

$$\nu(\mathcal{P}(r)) = \nu(\mathcal{D}(r)). \tag{2.17}$$

b) Consideriamo $x^* \in \mathbb{R}^n$ *e* $\lambda^* \in \mathbb{R}_+^{(\Omega)}$ *soluzioni ammissibili dei problemi* $\mathcal{P}(r)$ *e* $\mathcal{D}(r)$*. Allora* x^* *e* λ^* *sono soluzioni ottime dei rispettivi problemi se, e solo se, per ogni punto* $\tilde{t}$ *appartenente al supporto della mappa* λ^**, si ha*

$$\langle \nu(\tilde{t}), x^* \rangle = r(\tilde{t})$$

c) Ad una soluzione ottima del problema $\mathcal{P}(r)$*,* $x^* \in \mathbb{R}^n$*, possiamo associare una soluzione ottima del problema* $\mathcal{D}(r)$ *cercando una soluzione ammissibile il cui supporto sia contenuto nell'insieme descritto dai punti di minimo globale della mappa* $t \mapsto \langle \nu(t), x^* \rangle = r(t)$*.*

d) Il problema $\mathcal{P}(r)$ *è discretizzabile.*

Dimostrazione. a) Dall'ipotesi sulla limitatezza della funzione $r(\cdot)$ deduco che $\exists M \in \mathbb{R}$ tale che $r(t) < M$ per ogni $t \in \Omega$. Pertanto il punto $\overline{x} = (M, 0, \ldots, 0)^T \in \mathbb{R}^n$ è un pnto di Slater per il problema $\mathbb{P}(r)$. Poiché le funzioni $\nu_1, \ldots, \nu_n(\cdot)$ sono polinomi, è semplice verificare che sono funzioni continue e μ-misurabili. Per la Proposizione 2.2 sappiamo che il cono convesso $\mathcal{N}$, definito in (2.13), è un insieme chiuso. Inoltre, essendo il cono convesso $\mathcal{M}$ immagine della mappa continua che esclude la prima componente dell'insieme chiuso $\mathcal{N}$, è anch'esso convesso. Allora per la Proposizione 2.1 si ha che il problema $\mathcal{D}(r)$ è consistente. Infine applicando la Proposizione 2.3 si conclude che il problema $\mathcal{D}(r)$ è limitato. Possiamo perciò affermare che la proprietà di dualità forte è garantita dal Teorema 1.2.

b) Risulta banalmente provata unendo la relazione espressa nel punto *a)* con la definizione di ammissibilità per il problema $\mathcal{P}(r)$. Infatti essi ci permettono di affermare che ogni punto $\tilde{t}$ appartenente al supporto di una soluzione ottima del problema $\mathcal{D}(r)$ è un minimo globale della mappa $t \mapsto \langle \nu(t), x^* \rangle = r(t)$.

c) Poiché vale la proprietà di dualità forte, segue dal Teorema 1.4.

d) Poiché vale la proprietà di dualità forte, segue dal Teorema 1.5.

□

Consapevoli dell'importanza del punto *b)* proponiamo la seguente definizione.

Definizione 2.5. *Sia* $r : \Omega \to \mathbb{R}$ *una funzione continua.*
Dato $Q(t)$ *un polinomio tale che*

$$Q(t) \leq r(t), \ \forall t \in \Omega,$$

un punto $t_0 \in \Omega$ *si dice* punto di contatto *se* $Q(t_0) = r(t_0)$.

2.3 Interpretazione dei problemi $\mathcal{P}(r)$ e $\mathcal{D}(r)$

La proprietà di dualità forte fornisce un chiaro legame tra i problemi $\mathcal{P}(r)$ e $\mathcal{D}(r)$. Tuttavia è indispensabile ora dare una interpretazione concreta sulle definizioni dei due problemi poiché, anche se abbiamo ben argomentato che una soluzione del problema $\mathcal{D}(r)$ ci fornisce una formula di cubatura di ordine almeno m, non abbiamo ancora compreso quali caratteristiche abbiano le soluzioni ottime. Nella sezione successiva caratterizzeremo le funzioni di sensibilità che, nel caso univariato, permettono di ottenere la formula di quadratura di Gauss sulla misura μ come unica soluzione ottima del problema $\mathcal{D}(r)$. Tuttavia anticipiamo già ora che essa non rappresenta una concreta generalizzazione della formula di quadratura di Gauss poiché nella sua costruzione non riconosciamo la struttura dei polinomi ortogonali attraverso cui essa è usualmente descritta e ad oggi non si è stati in grado di provare una correlazione all'apparente efficienza (vedi Definizione 2.2) che si può constatare in modo empirico. Questa è la medesima conclusione che gli stessi E.K. Ryu e S.P. Boyd giungono a scrivere alla fine del loro articolo [15].

Una interpretazione del problema $\mathcal{P}(r)$ è stata esposta da R. Bojanic e da R. DeVore nel loro lavoro del 1966 [1]. Esso discute la determinazione del polinomio di miglior approssimazione di $r(\cdot)$ rispetto la norma $\mathcal{L}^1_\mu$ tra i polinomi che sono o superiori o inferiori alla funzione $r(\cdot)$ nel dominio Ω. Definiamo in modo rigorosa questa idea.

Definizione 2.6. *Definiamo il* polinomio superiore di miglior $\mathcal{L}^1_\mu$-approssimazione *di una funzione misurabile $r(\cdot)$ la soluzione ottima del seguente problema:*

$$\begin{aligned} \mathcal{P}_0(r): \quad & \inf_{Q\in\mathbb{P}^d_m} \quad \|Q-r\|_{\mathcal{L}^1_\mu} \\ & s.a \qquad Q(t)\geq r(t),\ t\in\Omega \end{aligned} \tag{2.18}$$

Poiché l'insieme $\{\nu_1(\cdot),\dots,\nu_n(\cdot)\}$ costituisce una base dello spazio vettoriale $\mathbb{P}^d_m$, ogni polinomio di questo spazio può essere scritto in modo unico come combinazione lineare degli elementi appartenenti a tale base. Questa considerazione ci permette di operare un cambio di variabile e, in virtù della definizione di norma $\mathcal{L}^1_\mu$, riscrivere il problema (2.18) come segue

$$\begin{aligned} \inf_{x\in\mathbb{R}^d} \quad & \int_\Omega \langle\nu(t),x\rangle - r(t)\ d\mu(t) \\ \text{s.a} \quad & \langle\nu(t),x\rangle \geq r(t),\ t\in\Omega \end{aligned}$$

In questo modo è già evidente come l'insieme delle soluzioni ammissibili sia in corrispondenza biunivoca con quello del problema $\mathcal{P}(r)$. Ponendo attenzione alla funzione obbiettivo, si può osservare il seguente sviluppo algebrico

$$\int_\Omega \langle\nu(t),x\rangle - r(t)\ d\mu(t) = \langle\int_\Omega \nu(t)d\mu(t),x\rangle - \int_\Omega r(t)\ d\mu(t) = \langle c,x\rangle - \int_\Omega r(t)\ d\mu(t)$$

che ci permette di constatare come la funzione obbiettivo del problema $\mathcal{P}_0(r)$ differisca da quella del problema $\mathcal{P}(r)$ per una costante. Possiamo concluderne che anche gli insiemi delle soluzioni ottime sono in corrispondenza biunivoca (per il medesimo isomorfismo) e quindi affermare che i problemi $\mathcal{P}(r)$ e $\mathcal{P}_0(r)$ sono in un certo senso equivalenti tenendo conto però che $\nu(\mathcal{P}(r)) = \nu(\mathcal{P}_0(r)) - \int_\Omega r(t)\ d\mu(t)$.

Ispirati da questa idea, sottraiamo ora alla funzione obbiettivo del problema $\mathcal{D}(r)$ la costante $\int_\Omega P^*(t)d\mu(t)$, dove $P^*\in\mathbb{P}^d_m$ è il polinomio superiore di miglior $\mathcal{L}^1_\mu$-approssimazione di $r(\cdot)$. Ma essendo la formula di cubatura λ di ordine almeno m, si ha per definizione che $\int_\Omega Q(t)d\mu(t) = \sum_{t\in\Omega}\lambda_t Q(t)$. Allora il problema $\mathcal{D}(r)$ assume la seguente forma:

$$\begin{aligned} \sup_{\lambda\in\mathbb{R}^{(\Omega)}_+} \quad & \sum_{t\in\Omega}\lambda_t\left(r(t)-P^*(t)\right) \\ \text{s.a} \quad & \sum_{t\in\Omega}\lambda_t\,\nu(t) = c \end{aligned}$$

Per la costruzione data in (2.18), possiamo asserire $r(t) - P^*(t) \leq 0 \, \forall t \in \Omega$ e di conseguenza è evidente che il seguente problema equivale ad $\mathcal{D}(r)$

$$\begin{aligned} \mathcal{D}_0(r): \quad \inf_{\lambda \in \mathbb{R}_+^{(\Omega)}} \quad & \left| \sum_{t \in \Omega} \lambda_t \, r(t) - \int_\Omega Q(t) d\mu(t) \right| \\ \text{s.a} \quad & \sum_{t \in \Omega} \lambda_t \, \nu(t) = c \end{aligned} \tag{2.19}$$

Consideriamo ora una soluzione ottima $\lambda^* \in \mathbb{R}_+^{(\Omega)}$ del problema $\mathcal{D}_r$ (e quindi anche di $\mathcal{D}_0(r)$) e ricordiamo che al polinomio $P^*(\cdot)$ possiamo associare una soluzione x^* ottima per il problema $\mathcal{P}(r)$ per l'isomorfismo già discusso. Allora, in virtù del punto *b)* del Corollario 2.1, possiamo constatare che per ogni $\tilde{t} \in \text{supp}\lambda^*$ si ha $P^*(\tilde{t}) = r(\tilde{t})$. Poiché la formula λ^* è di ordine almeno m, il valore assunto dalla funzione obbiettivo $\mathcal{D}_0(r)$ è

$$\begin{aligned} \nu(\mathcal{D}_0(r)) &= \left| \sum_{t \in \Omega} \lambda_t \, r(t) - \int_\Omega P^*(t) d\mu(t) \right| = \sum_{t \in \Omega} \lambda_t \, |r(t) - P^*(t)| \\ &= \sum_{t \in \Omega} \lambda_t \, |r(t) - r(t)| = 0 \end{aligned}$$

Riassumendo, la soluzione ottima del problema $\mathcal{D}(r)$ descrive una formula di cubatura di ordine almeno m tale che l'approssimazione integrale di $r(\cdot)$ coincida con quella (esatta) di $P^*(\cdot)$, il suo polinomio superiore di miglior $\mathcal{L}^1_\mu$-approssimazione. Pertanto l'errore commesso dall'approssimazione integrale sulla funzione $r(\cdot)$ è

$$\begin{aligned} \left| \sum_{t \in \Omega} \lambda_t^* \, r(t) - \int_\Omega r(t) \, d\mu(t) \right| &\leq \int_\Omega |P^*(t) - r(t)| \; d\mu(t) \\ &= \nu(\mathcal{P}_0(r)) \\ &= \nu(\mathcal{P}(r)) + \int_\Omega r(t) \, d\mu(t) \end{aligned}$$

2.4 Caso Univariato: Formula di quadratura di Gauss

Come anticipato, in questa sezione siamo intenzionati ad esibire una caratterizzazione della funzione di sensibilità $r(\cdot)$ tale da determinare come unica soluzione ottima per il problema $\mathcal{D}(r)$ la formula di quadratura di Gauss.
La formula di quadratura di Gauss, sviluppata nel 1814, è una formula di quadratura definita sull'intervallo di integrazione $[-1, 1]$ tradizionalmente presentata attraverso la teoria dei polinomi ortogonali [14, par. 9.2]. Ci proponiamo di darle una semplice definizione sufficiente a provare quanto voluto.

Definizione 2.7. *Fissato un intero m dispari, la* Formula di quadratura di Gauss *è una formula di ordine almeno m descritta da* $N = \left[\frac{m}{2}\right] + 1 = \frac{m+1}{2}$.
I nodi coincidono con gli zeri del polinomio ortogonale N-esimo rispetto la misura μ *mentre i pesi sono la valutazione dei nodi nel rispettivo polinomio caratteristico di Lagrange [14, 9.3].*

Basandoci sull'interpretazione data nella sezione precedente, intendiamo proporre uno sviluppo simile a quello dato dal lavoro già citato di R. Bojanic e da R. DeVore [1]. Esso rappresenta una valida alternativa alla dimostrazione proposta dal [15, Teorema 1] poiché esprime la costruzione attraverso cui si giunge effettivamente a tale conclusione.

Focalizziamo il nostro studio sul problema $\mathcal{P}_0(r)$, definito nella sezione 2.3. Innanzitutto osserviamo che sotto le ipotesi formulate nella Definizione 2.4, il problema $\mathcal{P}(r)$ è limitato e di conseguenza lo è anche $\mathcal{P}_0(r)$. Inoltre è evidente che i nodi descritti dal punto *b)* del Corollario 2.1 siano dei punti di contatto per $Q(t)$ con il polinomio $\langle \nu(t), x^* \rangle$. Quindi la chiave per dimostrare la tesi che ci siamo proposti di discutere sarà trovare delle condizioni quantitative sui punti di contatto.

Proposizione 2.4. *Sia* $\Omega = [-1, 1]$ *e siano verificate le ipotesi della Definizione 2.4. Allora una soluzione* $P^*(\cdot)$ *del problema* $\mathcal{P}_0(r)$ *ha almeno* $\left[\frac{m}{2}\right] + 1$ *punti di contatto distinti con* $r(\cdot)$.

Dimostrazione. La tesi è banalmente vera per il caso $m = 0$ e $m = 1$ e pertanto assumiamo $m \geq 2$. Osservando che $P^*(\cdot)$ è necessariamente una soluzione ammissibile del problema $\mathcal{P}_0(r)$, si deve avere che $P^*(t) \geq r(t)\ \forall t \in [-1, 1]$.

Supponiamo ora per assurdo che vi siano $N < \left[\frac{m}{2}\right] + 1$ punti di contatto (equivalentemente $N \leq \left[\frac{m}{2}\right]$), ovvero che vi siano $t_1 < \ldots < t_N$ punti di contatto distinti nell'intervallo $[-1, 1]$. Proveremo in tali circostanze che l'ottimalità della soluzione $P^*(t)$ è violata. Senza perdere di generalità possiamo supporre che $\{t_i\}_i \in (-1, 1)$. Sarà naturale adattare il ragionamento che segue qualora il primo o l'ultimo nodo coincidano con un estremo dell'intervallo.
Fissato $\epsilon > 0$ tale che $2\epsilon < \min_{i=1,\ldots,N-1} |t_{i+1} - t_i|$, possiamo definire il seguente polinomio

$$Q_\epsilon(t) := (t - (t_1 - \epsilon))(t - (t_1 + \epsilon)) \ldots (t - (t_N - \epsilon))(t - (t_N + \epsilon))$$

Per costruzione, abbiamo che il grado del polinomio Q_ϵ è

$$\deg Q_\epsilon = 2N \leq 2 \left[\frac{m}{2}\right] \leq m$$

È evidente che la successione $Q_\epsilon(\cdot)$ converge uniformemente,

$$\lim_{\epsilon\to 0^+} Q_\epsilon(t) = (t-t_1)^2\cdots(t-t_N)^2$$

Poiché $\int_{-1}^{1}(t-t_1)^2\ldots(t-t_N)^2 d\mu(t) > 0$, allora possiamo scegliere per la convergenza uniforme $\bar\epsilon > 0$ tale che

$$\int_{-1}^{1} Q_{\bar\epsilon}(t)d\mu(t) > 0$$

Definiamo ora l'unione degli $\bar\epsilon$-intorni aperti

$$\Pi_{\bar\epsilon} := \bigcup_{i=1}^{k}(t_i-\bar\epsilon, t_i+\bar\epsilon)$$

Poiché la mappa $t\mapsto P^*(t)-r(t)$ è strettamente positiva e continua su ciascun sottointervallo compatto di $[-1,1]\setminus\Pi_{\bar\epsilon}$, per il teorema di Weierstrass, esiste un numero reale $d>0$ tale che $P^*(t)-r(t)\geq d$ per ogni $t\in[-1,1]\setminus\Pi_{\bar\epsilon}$.
Definiamo allora il polinomio

$$\widetilde{P}(t) := P^*(t) - \eta Q_{\bar\epsilon}(t)$$

dove η è uno scalare fissato nel seguente modo:

$$\eta := \frac{d}{\max_{t\in[-1,1]}|Q_{\bar\epsilon}(t)|}$$

Intendiamo ora provare che tale polinomio violi l'ottimalità di $P^*(\cdot)$.
E' facile constatare che $\widetilde{P}(\cdot)$ appartiene a $\mathbb{P}^1_m$. Inoltre osservado che

$$\eta Q_{\bar\epsilon}(t) \leq d \leq P^*(t)-r(t),\ \forall t\in[-1,1]\setminus\Pi_{\bar\epsilon}$$
$$\eta Q_{\bar\epsilon}(t) \leq 0 \leq P^*(t)-r(t),\ \forall t\in[-1,1]\cap\Pi_{\bar\epsilon}$$

ne segue che $\widetilde{P}(t)\geq r(t)\ \forall t\in[-1,1]$, ovvero che $\widetilde{P}(\cdot)$ è ammissibile per il problema $\mathcal{P}_0(r)$. Tuttavia, come anticipato inizialmente, si può osservare che $\widetilde{P}(\cdot)$ viola l'ottimalità di $P^*(\cdot)$:

$$\int_{-1}^{1}\widetilde{P}(t)d\mu(t) = \int_{-1}^{1}P^*(t)d\mu(t) - \eta\int_{-1}^{1}Q_\epsilon(t)d\mu(t) < \int_{-1}^{1}P^*(t)d\mu(t)$$

□

Siamo intenzionati ora a provare l'unicità della soluzione ottima del problema $\mathcal{P}_0(r)$. In generale ciò non sempre si realizza ma mostreremo come la richiesta della differenziabilità della funzione di sensibilità, $r(\cdot)$ garantisce tale proprietà.

Proposizione 2.5. *Sia $\Omega = [-1, 1]$ e siano verificate le ipotesi della Definizione 2.4. Supponiamo inoltre che la funzione di sensibilità, $r(\cdot)$, sia differenziabile in $(-1, 1)$. Allora la soluzione del problema $\mathcal{P}_0(r)$ è unica.*

Dimostrazione. Circoscriviamo il nostro interesse al solo caso in cui m sia un intero dispari, anche se è possibile dare una dimostrazione del tutto analoga se m fosse un intero pari (si veda [1, Lemma 4]).

Supponiamo che $P_1(\cdot)$ e $P_2(\cdot)$ siano due soluzioni ottime del problema $\mathcal{P}_0(r)$. Allora anche

$$P(t) := \frac{P_1(t) + P_2(t)}{2}$$

è una soluzione ottima del problema $\mathcal{P}(r)$. Allora per la Proposizione 2.4, esistono $N \geq \left[\frac{m}{2}\right] + 1$ punti di contatto per $P(\cdot)$, ovvero N punti distinti $t_1 < \ldots < t_N$ tali che per ogni $i = 1, \ldots, N$ si abbia

$$r(t_i) = \frac{P_1(t_i) + P_2(t_i)}{2}$$

Tuttavia in virtù dell'ammissibilità delle soluzioni ottime, possiamo dedurre che $P_1(t) \geq r(t)$ e $P_2(t) \geq r(t)$ per ogni $t \in [-1, 1]$. Ed in particolare, applicando la disuguaglianza ai punti t_i sopra definiti, si ottiene:

$$r(t_i) = P_1(t_i) = P_2(t_i),\ i = 1, \ldots, N \tag{2.20}$$

Per la stessa ragione, possiamo asserire che le mappe continue $t \mapsto P_1(t) - r(t)$ e $t \mapsto P_2(t) - r(t)$ sono non-negative. Tenuto conto della relazione (2.20), possiamo desumere che tali mappe assumono il valore minimo (ovvero il valore nullo) nei punti $t_1, \ldots, t_N$.

Supponiamo inizialmente che $N > \left[\frac{m}{2}\right] + 1$ (ovvero che $N \geq \left[\frac{m}{2}\right] + 2$).
Si avranno quindi almeno $\left[\frac{m}{2}\right]$ punti di contatto in $(-1, 1)$. Poiché $r(t)$ è differenziabile e $\{t_i\}_{i=2,\ldots,N-1}$ sono contemporaneamente punti stazionari delle mappe sopra citate, allora

$$P_1'(t_i) = r'(t_i) = P_2'(t_i),\ i = 2, \ldots, N-1 \tag{2.21}$$

Quindi il sistema costituito dalle equazioni (2.20) e (2.21),

$$\begin{cases} P_1(t_i) - P_2(t_i) = 0 & i = 1, \ldots, N \\ P_1'(t_i) - P_2'(t_i) = 0 & i = 2, \ldots, N-1 \end{cases} \tag{2.22}$$

, è costituito da $2N - 2$ equazioni in $m + 1$ incognite rappresentanti i coefficienti del polinomio $P_1(\cdot) - P_2(\cdot)$. Poiché m è dispari allora $m = 2l + 1$ per qualche $l \geq 0$ e si ha

$$2N - 2 \geq 2\left(\left[\frac{m}{2} + 2\right]\right) - 2 = 2\left[l + \frac{1}{2}\right] + 2 = 2l + 2 = m + 1$$

ne segue che il sistema (2.22) ammette l'unica soluzione $P_1(t) = P_2(t)$.

Supponiamo ora che $N = \left[\frac{m}{2}\right] + 1$ ed m dispari.
Ci proponiamo di provare innanzitutto che, in tali ipotesi, tutti gli N punti di contatto $-1 \leq t_1 < \ldots < t_N \leq 1$ tra $P(t)$ e $r(t)$ siano in $(-1, 1)$. Supponiamo allora per assurdo che $t_1 = -1$ e procediamo analogamente alla dimostrazione della Proposizione 2.4.
Fissato $\epsilon > 0$ tale che $2\epsilon < \min_{i=1,\ldots,N-1} |t_{i+1} - t_i|$, possiamo definire il seguente polinomio

$$Q_\epsilon(t) := (t - (-1 + \epsilon))(t - (t_2 - \epsilon))(t - (t_2 + \epsilon)) \cdots (t - (t_N - \epsilon))(t - (t_N - \epsilon))$$

Per costruzione, abbiamo che il grado del polinomio $Q_\epsilon(\cdot)$ è

$$\deg Q_\epsilon = 2N - 1 = 2\left[l + \frac{1}{2}\right] - 1 = 2l - 1 = m - 2$$

E' evidente che la successione $Q_\epsilon(\cdot)$ converge uniformemente,

$$\lim_{\epsilon \to 0^+} Q_\epsilon(t) = (t + 1)(t - t_2)^2 \cdots (t - t_N)^2$$

Poiché $\int_{-1}^{1} (t + 1)(t - t_2)^2 \ldots (t - t_N)^2 d\mu(t) > 0$, allora possiamo scegliere per la convergenza uniforme $\bar{\epsilon} > 0$ tale che

$$\int_{-1}^{1} Q_{\bar{\epsilon}}(t) d\mu(t) > 0$$

Definiamo ora l'unione degli $\bar{\epsilon}$-intorni aperti

$$\Pi_{\bar{\epsilon}} := \bigcup_{i=2}^{N} (t_i - \bar{\epsilon}, t_i + \bar{\epsilon})$$

Poiché la mappa $t \mapsto P(t) - r(t)$ è strettamente positiva e continua su ciascun sottointervallo compatto di $[-1, 1] \setminus \Pi_{\bar{\epsilon}}$, per il teorema di Weierstrass, esiste un numero reale $d > 0$ tale che $P(t) - r(t) \geq d$ per ogni $t \in [-1, 1] \setminus \Pi_{\bar{\epsilon}}$.
Definiamo allora il polinomio

$$\widetilde{P}(t) := P(t) - \eta Q_{\bar{\epsilon}}(t)$$

dove η è uno scalare fissato nel seguete modo:

$$\eta := \frac{d}{\max_{t \in [-1,1]} |Q_{\bar{\epsilon}}(t)|}$$

Intendiamo ora provare che tale polinomio vìoli l'ottimalità di $P(\cdot)$.
E' facile constatare che $\widetilde{P}(\cdot)$ appartenga a $\mathbb{P}^1_m$. Inoltre osservando che

$$\eta Q_{\bar{\epsilon}}(t) \leq d \leq P(t) - r(t),\ \forall t \in [-1,1] \setminus \Pi_{\bar{\epsilon}}$$
$$\eta Q_{\bar{\epsilon}}(t) \leq 0 \leq P(t) - r(t),\ \forall t \in [-1,1] \cap \Pi_{\bar{\epsilon}}$$

ne segue che $\widetilde{P}(t) \geq r(t)\ \forall t \in [-1,1]$, ovvero che il polinomio $\widetilde{P}(\cdot)$ è ammissibile per il problema $\mathcal{P}_0(r)$. Tuttavia si può osservare che $\widetilde{P}(\cdot)$ vìola l'ottimalità di $P(\cdot)$

$$\int_{-1}^{1} \widetilde{P}(t)d\mu(t) = \int_{-1}^{1} P(t)d\mu(t) - \eta \int_{-1}^{1} Q_\epsilon(t)d\mu(t) < \int_{-1}^{1} P(t)d\mu(t)$$

Perciò ne deduciamo che $-1 < t_1$ e analogamente si può provare che $t_N < 1$, ovvero che tutti i punti di contatto appartengono all'interno del dominio di intragrazione, ovvero in $(-1,1)$.

Ora si riprende la dimostrazione del caso precedente.
Poiché $r(t)$ è differenziabile e $\{t_i\}_{i=1,\dots,N}$ sono contemporaneamente punti stazionari delle mappe non-negative $t \mapsto P_1(t) - r(t)$ e $t \mapsto P_2(t) - r(t)$, allora

$$P_1'(t_i) = r'(t_i) = P_2'(t_i),\ i = 1,\dots,N \tag{2.23}$$

Quindi il sistema costituito dalle equazioni (2.20) e (2.21),

$$\begin{cases} P_1(t_i) - P_2(t_i) = 0 & i = 1,\dots,N \\ P_1'(t_i) - P_2'(t_i) = 0 & i = 2,\dots,N \end{cases} \tag{2.24}$$

, è costituito da $2N$ equazioni in $m+1$ incognite rappresentanti i coefficienti del polinomio $P_1(\cdot) - P_2(\cdot)$. Poiché m è dispari allora $m = 2l+1$ per qualche $l \geq 0$ e si ha

$$2N \geq 2\left(\left[\frac{m}{2}\right] + 1\right) = 2\left[l + \frac{1}{2}\right] + 2 = 2l + 2 = m + 1$$

ne segue che il sistema (2.22) ammette l'unica soluzione $P_1(t) = P_2(t)$. □

Provata l'unicità, siamo ora interessati a caratterizzare i punti di contatto tra il polinomio superiore di miglior $\mathcal{L}^1_\mu$-approssimazione di $r(\cdot)$, $P^*(\cdot)$, e la funzione $r(\cdot)$ stessa.

Teorema 2.1. *Fissato m un intero dispari, supponiamo che la funzione di sensibilità, $r(\cdot)$, sia $m+1$-volte differenziabile ed inoltre che abbia $r^{(m+1)}(t) \leq 0$ per ogni $t \in (-1,1)$. Allora i punti di contatto tra il polinomio $P(\cdot)$, soluzione del problema $\mathcal{P}_0(r)$, e la funzione sensibilità $r(\cdot)$ coincidono con gli zeri del polinomio ortogonale N-esimo secondo la misura μ, con $N = \left[\frac{m}{2}\right] + 1$.*

Dimostrazione. Consideriamo il polinomio $\tilde{P}(\cdot)$ definito da

$$\begin{cases} \tilde{P}(t_i) = r(t_i) & i = 1, \dots, N \\ \tilde{P}'(t_i) = r'(t_i) & i = 1, \dots, N \end{cases} \tag{2.25}$$

dove $t_1 < t_2 < \dots, t_N$ sono gli zeri del polinomio ortogonale N-esimo secondo la misura μ. Osservando che il sistema (2.25) è costituito da $2N$ equazioni, allora esso individua un polinomio con altrettanti coefficienti. Pertanto possiamo asserire che il polinomio $\tilde{P}(\cdot)$ appartiene $\mathbb{P}^1_{2N-1} = \mathbb{P}^1_m$. Inoltre, per il teorema del resto di Cauchy [3, Teorema 3.5.1, pag. 67], si ha che $\exists \xi \in (\min(t, t_1), \max(t, t_l))$ tale che

$$r(t) - \tilde{P}(t) = \frac{r^{(m+1)}(\xi)}{(m+1)!}(t - t_1)^2 \dots (t - t_N)^2 \tag{2.26}$$

Poiché per ipotesi abbiamo supposto $r^{(m+1)}(t) \leq 0$, otteniamo $\tilde{P}(t) \geq r(t)$ per ogni $t \in [-1, 1]$. In altre parole, $\tilde{P}(\cdot)$ è soluzione ammissibile del problema $\mathcal{P}_0(r)$. Dalla Proposizione 2.5 sappiamo che il problema $\mathcal{P}_0(r)$ ammette un'unica soluzione ottima. Pertanto è nostra intenzione provare che $\tilde{P}(\cdot)$ è soluzione ottima. Per la definizione della formula di quadratura di Gauss, si ha che $\forall Q(t) \in \mathbb{P}^d_m$,

$$\int_{-1}^{1} Q(t) d\mu(t) = \sum_{k=1}^{N} w_{G_k} Q(t_k) \tag{2.27}$$

Considerando allora i soli polinomi $Q(t) \geq r(t),\ t \in [-1, 1]$, si ottiene

$$\int_{-1}^{1} Q d\mu = \sum_{k=1}^{N} w_{G_k} Q(t_k) \geq \sum_{k=1}^{N} w_{G_k} r(t_k) = \sum_{k=1}^{N} w_{G_k} \tilde{P}(t_k) \tag{2.28}$$

Questo prova l'ottimalità di $\tilde{P}(t)$ per il problema $\mathcal{P}_0(t)$. □

Corollario 2.2. *Fissato m un intero dispari, supponiamo che la funzione di sensibilità, $r(\cdot)$, sia $(m+1)$-volte differenziabile ed inoltre che abbia $r^{(m+1)}(t) \leq 0$ per ogni $t \in (-1, 1)$. Allora la soluzione del problema $\mathcal{D}(r)$ coincide con la formula di quadratura di Gauss.*

Dimostrazione. Questo asserto segue immediatamente dalla sezione precedente. Infatti abbiamo pazientemente provato che le soluzioni del problema $\mathcal{P}_0(r)$ sono in corrispondenza biunivoca con quelle del problema $\mathcal{D}_0(r)$. Pertanto, per la Propozione 2.5, sappiamo che tale insieme delle soluzioni ottime è costituito da un unico elemento. Per il Teorema 2.1, quest'ultimo descrive un polinomio che ha come punti di contatto con la funzione $r(\cdot)$ gli zeri del N-esimo polinomio ortogonale secondo la misura μ. Poiché la formula di Gauss è una soluzione ammissibile del problema $\mathcal{D}(r)$, per il punto *b)* del Corollario 2.1, si deduce che essa è l'unica soluzione ottima del problema. □

L'articolo [1] discute anche il caso in cui si ponga la condizione $r^{(m+1)} \geq 0$ per ogni $t \in [-1, 1]$. In tale lavoro si mostra analogamente che la soluzione ottima coincide con la formula di quadratura di Gauss-Lobatto.

2.5 Caso Multivariato

Dalle sezioni precedenti è evidente che la scelta di una funzione di sensibilità $r(t) : \Omega \to \mathbb{R}$ è determinante. Il vero obbiettivo sarebbe per l'appunto determinare quali proprietà sono garantite in base a ipotesi che si possono assumere su questa funzione. Ad esempio, la scelta della funzione $r(t)$ e la numerosità dei nodi sono apparentemente caratteristiche dipendenti. Tuttavia determinare a priori delle caratteristiche simili a quelle discusse nella sezione precedente, come sottolineato dagli stessi autori dell'articolo [15], non è scontato. Ciò che invece è immediato da dimostrare, è una stima superiore della cardinalità dei nodi.

Teorema 2.2. *Le soluzioni del problema $\mathcal{D}(r)$ hanno supporto di cardinalità al più n. Ovvero una soluzione ottima di $\mathcal{D}(r)$ descrive una formula di cubatura il cui grado di efficienza non è maggiore di 1, $\rho \leq 1$.*

Dimostrazione. La dimostrazione fa riferimento a quella proposta in [6, pag. 14]. Se tutte le soluzioni ottime hanno cardinalità al più n, l'asserto è provato. In caso contrario consideriamo una soluzione ottima del problema $\mathcal{D}(r)$, $\lambda^* \in \Lambda$, tale da aver supporto di cardinalità $N \geq n + 1$. Allora esiste l'insieme che costituisce il supporto di λ^*, $t_1, \ldots, t_N \subset T$ e i relativi pesi non-negativi $w_1, \ldots, w_N > 0$.
Per la formulazione (1.6) si ha che

$$\sum_{k=1}^{N} \begin{pmatrix} r(t_k) \\ \nu(t_k) \end{pmatrix} w_k = \begin{pmatrix} \nu(\mathcal{D}(r)) \\ c \end{pmatrix} =: v \tag{2.29}$$

e come sappiamo questo vettore sta sulla frontiera di $\mathcal{N}$.
Senza perdere di generalità consideriamo $N = n + 1$ e, seguendo il Lemma presentato in [7, pag. 89], intendiamo provare che il vettore v è un punto interno, giungendo all'assurdo. Definiamo quindi la matrice A che ha come colonne i vettori

$$\begin{pmatrix} r(t_1) \\ \nu(t_1) \end{pmatrix}, \ldots, \begin{pmatrix} r(t_N) \\ \nu(t_N) \end{pmatrix}$$

Allora possiamo riscrivere la relazione (2.29) come

$$v = A\ w$$

Essendo una matrice di tipo Vandermonde ed i nodi punti distinti, ricaviamo che tale matrice è non-singolare. Pertanto possiamo determinare un'espressione per calcolare i pesi:

$$w = A^{-1} v$$

Poiché i pesi $w_1, \ldots, w_N$ sono strettamente positivi, esiste un $\epsilon > 0$ tale che per ogni vettore $\tilde{v}$ che soddisfi $|v - \tilde{v}| \leq \epsilon$ permette di asserire

$$\tilde{x} := A^{-1}\tilde{v} > 0 \tag{2.30}$$

Perciò il vettore v possiamo scriverlo come $\tilde{v} = A\tilde{x}$ e questo ci permette di concludere che appartiene al cono $\mathcal{N}$.

Come avevamo anticipato, possiamo perciò asserire che il vettore v costituisce un punto interno dell'insieme $\mathcal{N}$ ma, contemporaneamente, anche un punto di frontiera. Poiché tale considerazione non può essere vera, deduciamo l'assurdo e quindi concludiamo che la tesi enunciata è vera. □

Capitolo 3

Proposta di un algoritmo numerico

Lo scopo di questo capitolo è di determinare un metodo per calcolare una formula di cubatura di un prestabilito ordine, risolvendo il problema $\mathcal{D}(r)$. Tuttavia le difficoltà che si presentano nella trattazione numerica per la determinazione di una tale formula, si riscontrano comprensibilmente anche nella risoluzione diretta dei LSIP di questa forma. La letteratura relativa ai LSIP non per caso conferisce alla forma principale una maggiore importanza poiché permette di ideare dei metodi numerici più efficaci e semplici che sono usualmente classificati in cinque categorie: *Discretization Methods*, *Local Reduction Methods*, *Exchange Methods*, *Simplex-like Methods* e *Descent Methods*. L'ordine di questa lista non è casuale e rispecchia l'ordine decrescente di efficienza computazionale attribuita da molti esperti del settore, come riportato in [8, Capitolo 11].

Analizzeremo in questo capitolo un algoritmo costituito da due fasi, sfruttando in maniera efficace le proprietà della teoria dei LSIP in virtù delle ipotesi formulate nella Definizione 2.4. La prima fase costituisce una prima approssimazione e risolve il problema $\mathcal{P}(r)$ direttamente, ignorando la geometria del suo duale, $\mathcal{D}(r)$. In virtù della sua lenta convergenza e dell'importante costo computazionale necessario, siamo motivati ad innestare successivamente un metodo iterativo più rapido, appartenente alla categoria dei *Local Reduction Methods*. Come suggerisce il nome, questi ultimi sono metodi che offrono una maggiore velocità di convergenza ma ne garantiscono la validità solo quando sono applicati ad una buona approssimazione della soluzione del problema $\mathcal{P}(r)$. In particolare, sfruttano la proprietà di dualità forte e quindi coinvolgono anche le caratteristiche del problema $\mathcal{D}(r)$ offrendo, come vedremo sperimentalmente, un'ottima decrescrescita dell'errore sui momenti prodotto dalla formula di cubatura associata. Sfortunatamente, ad oggi, nessun risultato teorico permette di stabilire un criterio con cui decidere quando applicare la seconda fase [9, pag. 19].

3.1 Prima fase: Discretization Method

Dalla letteratura [9, Sezione 1.3] segue che i *Discretization Methods* sono algoritmi che generano sequenze di vettori $x^{(k)} \in \mathbb{R}^n$ risolvendo il problema LP

$$\begin{aligned} \mathcal{P}_k(r): \quad & \inf_{x\in\mathbb{R}^n} \quad \langle c, x\rangle \\ & \text{s.a} \qquad \langle \nu(t), x\rangle \geq r(t),\ \forall t \in T_k \end{aligned} \tag{3.1}$$

dove T_k è una sequenza di sottoinsiemi finiti di Ω.
Se T_k è una sequenza espansiva ($T_k \subset T_{k+1}$ per $k = 1, \ldots$), allora la successione $x^{(k)}$ convergerà ad una soluzione ottima di $\mathcal{P}(r)$ per la proprietà di essere discretizzabile, verificata nel Corollario 2.1.

In generale, un elemento della sequenza $x^{(k)}$ non è ammissibile per il problema $\mathcal{P}(r)$. Infatti, se lo fosse, sarebbe in realtà soluzione ottima di $\mathcal{P}(r)$ e quest'ultimo risulterebbe essere un problema riducibile. Quindi possiamo indicare con $\mathcal{E}_F^{(k)}$ l'errore sull'ammissibilità del vettore $x^{(k)}$, ovvero il valore minimo della mappa $t \mapsto \langle \nu(t), x^{(k)}\rangle - r(t)$. In base a queste constatazioni, possiamo fissare uno scalare $\epsilon > 0$ che chiamiamo *margine di accuratezza* attraverso il quale stabilire un criterio di arresto. Segue allora lo schema risolutivo di un generico *Discretization Method*:

Algoritmo 3.1. *Inizializzato k=0,*

Step-1) esplicitare il sottoinsieme T_k;

Step-2) determinare una soluzione $x^{(k)}$ del problema $\mathcal{P}_k(r)$, se il problema $\mathcal{P}_k(r)$ è inammissibile allora anche $\mathcal{P}(r)$ lo è;

Step-3) trovare il minimo $\mathcal{E}_F^{(k)}$ della mappa $t \mapsto \langle \nu(t), x^{(k)}\rangle - r(t)$;

Step-4) se $\mathcal{E}_F^{(k)} \geq -\epsilon$ allora è stata trovata una soluzione ottima; altrimenti ritorno allo Step-1 sostituendo a k, k+1.

In particolare, i *Grid Discretization Method* costituiscono un sottoinsieme di questi metodi dove gli insiemi T_k sono determinati a priori. Solitamente è scelta una sequenza espansiva, $T_k \subset T_{k+1}$, per ogni $k = 1, 2, \ldots$ per le questioni di convergenza già sollevate. Generalmente, le sequenze T_k sono determinate induttivamente riducendo il numero dei vincoli del PL da studiare. Per esempio, stabilita una sequenza ϵ_k non-negativa tale da convergere a zero e una griglia di partenza T_0, il classico *Cutting Plane Discretization Method* prevede che

$$T_{k+1} = T_k \cup \left\{ t \in \mathcal{T}(\subset \Omega) : \langle \nu(t), x^{(k)}\rangle - r(t) \geq \mathcal{E}_F^{(k)} + \epsilon_k \right\} \tag{3.2}$$

dove il sottoinsieme $\mathcal{T}$ può essere a sua volta una griglia di punti fissa o espansiva. Una particolare scelta dei parametri costituisce l'*Alternating Algorithm* dove, stabilita la griglia di partenza T_0, prevede che

$$T_{k+1} = T_k \cup \{t_{\min}\} \tag{3.3}$$

dove $t_{\min}$ è un punto che realizza il minimo $\mathcal{E}_F^{(k)}$. La terminazione di questi particolari casi è dimostrata nel Teorema 11.1 e 11.2 del libro [8, pag. 259 e 263)]. Motivati dalla sua intuitiva efficienza, utilizzeremo prevalentemente un *Cutting Plane Discretization Method* nei nostri esperimenti numerici.

Determinata una soluzione approssimata $\tilde{x}$ del problema $\mathcal{P}(r)$, cioé il vettore dei coefficienti (rispetto alla base $\nu_1(\cdot), \ldots, \nu_n(\cdot)$) del polinomio $p(\cdot)$ di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra di $r(\cdot)$, intendiamo determinare la formula di cubatura associata alla soluzione del duale $\mathcal{D}(r)$ descritta dal punto *c)* del Corollario 2.1. In virtù di quest'ultimo, i nodi che descrivono tale formula di cubatura devono essere dei punti di contatto tra il polinomio $p(\cdot)$ e la funzione di sensibilità $r(\cdot)$. In modo equivalente, possiamo affermare che i punti di minimo globale della mappa che descrive la loro differenza, $t \to p(t) - r(t) = \langle \nu(t), \tilde{x} \rangle - r(t)$, sono candidati ad essere i nodi della formula di cubatura. Tuttavia, considerata l'approssimazione di $\tilde{x}$ (e quindi di $p(\cdot)$), è opportuno introdurre una tolleranza nella selezione dei punti di minimo globale tra quelli di minimo locale. Ragionevolmente considereremo solo i punti di minimo locale che abbiano un valore inferiore ad una certa soglia di tolleranza $\tau > 0$, ovvero quelli che, in modo intuitivo, riconosciamo essere dei punti di contatto approssimati. È opportuno non scegliere un parametro τ troppo piccolo altrimenti si potrebbero escludere dei nodi della formula di cubatura ottenendo degli errori sui momenti molto importanti. Viceversa, scegliere un parametro non piccolo, può determinare molti più candidati, di conseguenza più calcoli. Sperimentalmente si è constatato un buon compresso il valore $\tau = 0.05$. Costruita allora la matrice di Vandermonde V (ovvero la matrice le cui colonne sono le valutazione di $\nu(\cdot) = (\nu_1(\cdot), \ldots, \nu_n(\cdot))^T$ in ciascun candidato), per determinare i pesi sarà sufficiente definire il sistema lineare costituito dai vincoli sui momenti, $Vw = c$, imponendo la condizione di positività dei nodi, $w \geq 0$. A causa dell'approssimazione con cui abbiamo determinato $\tilde{x}$, ci aspettiamo che il sistema $V\,w = c$ non ammetta soluzione. Pertanto, fissata una norma vettoriale, ci interessiamo alla soluzione $w \geq 0$ che comporti il minor errore $\mathcal{E}_{\Lambda;<norma>}$, corrispondente alla norma prescelta del residuo $Vw - c$. Questa soluzione è agevolmente determinata per la norma $\|\cdot\|_2$ dalla routine `lsqnonneg` di MATLAB che risolve il seguente problema di ottimo

$$\begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \mathcal{E}_{\Lambda;2} := \|Vw - c\|_2 \\ \text{s.a} \quad & w \geq 0 \end{aligned} \tag{3.4}$$

Possiamo riassumere questo procedimento nell'algoritmo che segue.

Algoritmo 3.2. *Data una soluzione approssimata del problema $\mathcal{P}(r)$, $\tilde{x}$, e fissato il margine di tolleranza τ,*

Step-1) determinare i punti di contatto approssimati tra il polinomio $p(\cdot)$ e la funzione di sensibilità $r(t)$, determinando i minimi globali della mappa $t \mapsto \langle \nu(t), \tilde{x} \rangle - r(t)$ (selezionati con tolleranza τ);

Step-2) calcolare la matrice di Vandermonde V e risolvere il problema di ottimizzazione (3.4) *per determinare i pesi non-negativi associati.*

È evidente che richiedere un margine di accuratezza più basso nell'Algoritmo 3.1 permette di raggiungere, con il metodo sopra descritto, un errore sui momenti, $\mathcal{E}_{\Lambda;2}$, asintoticamente nullo. Tuttavia, poiché la matrice di Vandermonde V dipende della base $\nu_1(\cdot), \ldots, \nu_n(\cdot)$, una sua scelta trascurata può detereminare un mal-condizionamento dellla matrice V. Ciò significa che l'errore dovuto all'approssimazione dei nodi verebbe amplificato nel calcolo dei pesi e pertanto lo sforzo computazionale verrebbe vanificato.

3.2 Seconda Fase: Local Reduction Method

Nell'introduzione di questo capitolo abbiamo anticipato l'intenzione di definire un metodo che in una prima fase impieghi il *Cutting Plane Discretization Method* e di seguito un *Local Reduction Method*, allo scopo di migliorare la convergenza. A tale proposito, vogliamo mettere a confronto, tramite la Figura 3.1, la decrescita (ad ogni iterazione) dell'errore $\mathcal{E}_{\Lambda;\infty}$ tra la sola prima fase e il metodo più efficiente che combina a quest'ultima, la seconda fase descritta nella sezione corrente.

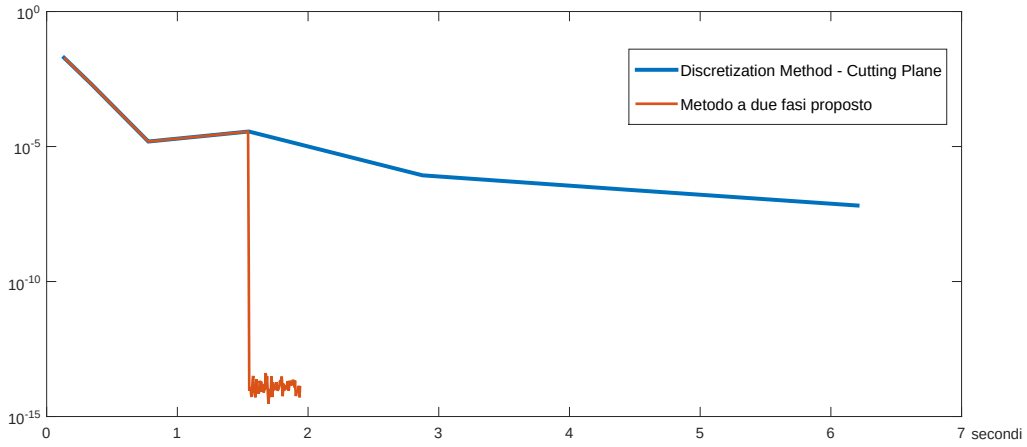

Figura 3.1: Confronto dell'andamento di $\mathcal{E}_{\Lambda;\infty}$ nel tempo del *Cutting Plane Discretization Method* rispetto a quello proposto, costituito da due fasi.

Il grafico, in scala semilogaritmica, mette in evidenza un importante miglioramento. Si è considerato lo studio della formula di cubatura di grado $m = 9$ sull'intervallo $\Omega = [-1, 1]$ rispetto la funzione di sensibilità $r(t) = -T_{m+1}(t)$ avendo scelto la base di Chebyshev. La prima fase raggiunge un errore minimo di $\mathcal{E}_{\Lambda;\infty} = 6.527742 \cdot 10^{-8}$ dopo 20 iterazioni (71.580 secondi), mentre la seconda fase alla sua terza iterazione (dalla durata totale di 1.990 secondi) raggiunge un errore pari a $\mathcal{E}_{\Lambda;\infty} = 9.103829 \cdot 10^{-15}$. Queste prestazioni sono sufficienti per creare l'interesse a sviluppare le argomentazioni seguenti con la premessa di non poter scendere in modo dettagliato su alcuni aspetti e rimandando il lettore più attento alla bibliografia citata.

Dalla letteratura [9, Sezione 1.3.3], segue che i *Local Reduction Methods*, meglio conosciuti come approssimazione di Chebyshev, sono metodi che sostituiscono il problema $\mathcal{P}(r)$ con un sistema (non necessariamente lineare) costituito da condizioni di ottimalità. In generale, se non sono note delle condizioni necessarie e sufficienti di ottimalità, vengono impiegate le *condizioni di Karush-Kuhn-Tucker* (si veda [8, pag. 78]). Queste sono condizioni necessarie (e non sufficienti) per la soluzione di un problema di programmazione non lineare purché i vincoli soddisfino certe condizioni di regolarità dette *condizioni di qualificazioni dei vincoli*.

Tuttavia la teoria dei precedenti capitoli ci ha permesso di dedurre condizioni di ottimalità necessarie e anche sufficienti (si veda il Corollario 2.1) che risultano più appropriate per il nostro studio.

Corollario 3.1. *Sia $x \in \mathbb{R}^n$ un vettore di coefficienti tale da determinare, nella base prescelta $\nu(\cdot) = (\nu_1(t), \dots, \nu_n(t))^T$, un polinomio $p(\cdot) = \langle \nu(\cdot), x \rangle$ che sia al di sopra della funzione di sensibilità $r(\cdot)$. Sia inoltre data una formula di cubatura di ordine almeno m. Allora, i nodi della formula di cubatura sono punti di contatto tra $p(\cdot)$ e $r(\cdot)$ se, e solo se, x e la formula di cubatura sono rispettivamente soluzioni dei problemi $\mathcal{P}(r)$ e $\mathcal{D}(r)$.*

Possiamo allora tradurre in modo algebrico il Corollario appena presentato. Descrivendo la condizione su x, i vincoli di essere una formula di cubatura di grado almeno m con pesi positivi e la condizione di ottimalità, possiamo affermare che un vettore $x \in \mathbb{R}^n$, un insieme di nodi $t = (t_1, \dots, t_N)$ e pesi $w = (w_1, \dots, w_N)$ costituiscono rispettivamente le soluzioni ottime dei problemi $\mathcal{P}(r)$ e $\mathcal{D}(r)$ se, e solo se, sono soluzioni del seguente sistema:

$$\begin{cases} \langle \nu(t), x \rangle \geq r(t) & \forall t \in \Omega \\ \sum_{k=1}^{N} w_k \, \nu(t_k) = c & \\ w_k \geq 0 & k = 1, \dots, N \\ \langle \nu(t_k), x \rangle = r(t_k) & k = 1, \dots, N \end{cases} \tag{3.5}$$

Il sistema pone delle condizioni su ciascuna componente delle variabili sopra citate (ovvero $x_1, \dots, x_n$, $t_1, \dots, t_N$ e $w_1, \dots, w_N$) e possiamo perciò intepretarlo come

un sistema costituito da $n+N$ equazioni e dalle due disequazioni in $n+(d\cdot N)+N$ variabili. Allo scopo di ricavare un maggior numero di equazioni che in modo equivalente descrivano il sistema (3.5), poniamo la nostra attenzione sulla prima condizione. Questa dovrebbe essere verificata sull'intero insieme compatto Ω. Tuttavia si osserva facilmente che unita all'ultima condizione, i punti $t_1, \dots, t_N$ risultano essere i punti di minimo globale della mappa $t \to \langle \nu(t), x\rangle - r(t)$. Imponendo, d'ora in avanti, la costruzione di questi punti attraverso questa interpretazione, l'ultima condizione ci fornisce in modo tautologico la verifica della prima condizione. Quindi, presupponendo questa costruzione, il sistema (3.5) è equivalente a quello che segue:

$$\begin{cases} \sum_{k=1}^{N} w_k\, \nu(t_k) = c & \\ w_k \geq 0 & k = 1, \dots, N \\ \langle \nu(t_k), x\rangle = r(t_k) & k = 1, \dots, N \end{cases} \tag{3.6}$$

Nonostante tutte le componenti siano deducibili univocamente dalla determinazione dei punti come minimi globali e le restanti variabili dalle equazioni del sistema (3.6), una sintesi efficace di questo metodo richiede la descrizione del maggior numero di equazioni che però dovrà essere pari alla dimensione delle variabili che stiamo studiando. Le motivazioni saranno poi chiarite al lettore.

Seguendo l'articolo [18] di G.A. Watson, distinguiamo i punti di minimo globale della mappa $t \to \langle \nu(t), x\rangle - r(t)$ che appartengono all'interno dell'insieme compatto Ω da quelli che fanno parte della sua frontiera. Per rendere più agevole la nostra notazione, supponiamo d'ora in poi che degli N nodi, i primi $L \leq N$ indichino quelli appartenenti all'interno dell'insieme Ω, $t_i \in \text{int}\ \Omega$. Dalla teoria elementare dell'analisi, sappiamo che una condizione necessaria affinché un punto interno al dominio Ω sia di minimo, è quella di essere un punto stazionario, ovvero che annulli il gradiente della mappa $t \to \langle \nu(t), x\rangle - r(t)$. Allora è nostro scopo considerare i nodi interni delle variabili mentre i nodi esterni dei parametri stabiliti e quindi giungere alla descrizione del seguente sistema nelle variabili $x_1, \dots, x_n$, $t_1, \dots, t_L$ e $w_1, \dots, w_N$:

$$\begin{cases} \sum_{k=1}^{N} w_k\, \nu(t_k) = c & \\ \langle \nu(t_k), x\rangle = r(t_k) & k = 1, \dots, N \\ \langle \nabla\nu(t_k), x\rangle = \nabla r(t_k) & k = 1, \dots, L \\ w_k \geq 0 & k = 1, \dots, N \end{cases} \tag{3.7}$$

Una rapida analisi mostra che tale sistema è costituito da $n+N+(L\cdot d)$ equazioni in $n + (L \cdot d) + N$ variabili (si ricorda $t_{L+1}, \dots, t_N$ sono considerati noti), infatti si osserva che la terza condizione descrive, per ogni k, un insieme di d vincoli di uguaglianza, costituendo complessivamente $L \cdot d$ equazioni. L'articolo allora propone di determinare una successione di soluzioni del sistema descritto dalle sole

3 prime condizioni di uguaglianza tramite il metodo di Newton per poter convergere ad una soluzione per entrambi i problemi. Ora dovrebbero essere chiare le motivazioni che ci hanno spinto a voler richiedere il maggior numero di equazioni: infatti, intuitivamente, queste determinano dei maggiori vincoli all'iterazione successiva migliorando l'efficienza e la convergenza di un metodo ad iterazioni a punto fisso.

A titolo divulgativo, vogliamo sottolineare che quella proposta non è l'unica strategia discussa in letteratura. Ad esempio si veda [9, pag. 18], dove nella trattazione delle condizioni necessarie di KKT si propone il coinvolgimento anche dei nodi appartenenti alla frontiera che noi non abbiamo considerato per ragioni presto chiare. Inoltre questo metodo richiede che il dominio di integrazione sia descrivibile tramite delle disuguaglianze di classe $\mathcal{C}^2$, restringendo le ipotesi del nostro problema. In base a queste ragioni abbiamo ritenuto più idoneo mantenere questa trattazione ma non si intendono escludere altre metodologie che possono essere utilizzate al fine di migliorare ulteriormente la velocità di convergenza del metodo.

Il sistema (3.8) può essere risolto tramite il metodo di Newton e, a tale scopo, vogliamo riscrivere le equazioni come un'unica funzione di cui intendiamo trovare gli zeri. Ricordando ancora una volta che i nodi $t_L, \dots, t_N$ sono dei parametri noti, introduciamo la mappa $\mathcal{F} : \mathbb{R}^n \times (\mathbb{R}^d \times \mathbb{R})^L \times \mathbb{R}^{N-L} \to \mathbb{R}^n \times \mathbb{R}^N \times (\mathbb{R}^d)^L$ le cui componenti raggruppate si possono così esplicitare

$$\begin{aligned}
\mathcal{F}_1 : \quad (\mathbb{R}^d \times \mathbb{R})^L \times \mathbb{R}^{N-L} &\longrightarrow \mathbb{R}^n \\
\{(t_k, w_k)\}_{k\le N} &\longmapsto \sum_{k=1}^{N} w_k\, \nu(t_k) - c \\
\mathcal{F}_2 : \quad \mathbb{R}^n \times (\mathbb{R}^d)^L &\longrightarrow \mathbb{R}^N \\
x, \{t_k\}_{k\le N} &\longmapsto (\langle \nu(t_k), x\rangle - r(t_k))_{k\le N} \\
\mathcal{F}_3 : \quad \mathbb{R}^n \times (\mathbb{R}^d)^L &\longrightarrow (\mathbb{R}^d)^L \\
x, \{t_k\}_{k\le L} &\longmapsto (\langle \nabla\nu(t_k), x\rangle - \nabla r(t_k))_{k\le L}
\end{aligned}$$

Allora il sistema (3.8) può essere riscritto nel seguente modo:

$$\begin{cases} \mathcal{F}(x, t_1, \dots, t_L, w_1, \dots, w_N) = 0 & \\ w_k \ge 0 & k = 1, \dots, N \end{cases} \tag{3.8}$$

Come anticipato, trascuriamo momentaneamente il vincolo di non negatività sui pesi, seguendo l'articolo [18]. Dalla teoria (si veda [14, sezione 6.7]) emerge la

necessità di verificare che la funzione $\mathcal{F}$ sia di classe $\mathcal{C}^1$.

Essa è verificata poiché, per ipotesi, i polinomi che costituiscono la base e la funzione di sensibilità sono almeno di classe $\mathcal{C}^2(\Omega;\mathbb{R})$ e i nodi coinvolti nella derivazione sono interni. Viceversa se avessimo coinvolto quelli di frontiera non avremmo potuto dire che la derivata fosse ben definita e pertanto sarebbe stato necessario valutare un altro metodo per il raggiungimento del medesimo scopo.

Per poter impiegare il metodo di Newton, ordiniamo le variabili in modo da rappresentarle come un unico vettore colonna. In particolare, visto che $t_1, \dots, t_L \in \Omega \subseteq \mathbb{R}^d$, definiamo il vettore

$$\mathring{\underline{t}} := (t_1; \dots; t_L) \in \mathbb{R}^{Ld};$$

I pesi invece li immagaziniamo in un unico vettore colonna $w = (w_1; \dots; w_n)$.
Dato il vettore $(x^{(0)}; \mathring{\underline{t}}^{(0)}; w^{(0)})$ sufficientemente vicino alla soluzione ottima $(x^*; \mathring{\underline{t}}^*; w^*)$, il metodo di Newton generalizzato afferma che possiamo determinare iterativamente una successione $(x^{(q)}; \mathring{\underline{t}}^{(q)}; w^{(q)})$ che converge sotto opportune ipotesi alla soluzione ottima:

$$\mathcal{J}(x^{(q)}, \mathring{\underline{t}}^{(q)}, w^{(q)}) \begin{pmatrix} \delta_x \\ \delta_{\mathring{\underline{t}}} \\ \delta_w \end{pmatrix} = -\mathcal{F}(x^{(q)}, \mathring{\underline{t}}^{(q)}, w^{(q)}) \tag{3.9}$$

dove $\delta_x = x^{(q+1)} - x^{(q)}$, $\delta_{\mathring{\underline{t}}} = \mathring{\underline{t}}^{(q+1)} - \mathring{\underline{t}}^{(q)}$ e $\delta_x = w^{(q+1)} - w^{(q)}$ mentre la matrice $\mathcal{J}(\cdots)$ è la matrice Jacobiana associata alla mappa $\mathcal{F}$.

Allo scopo di esplicitare una serie di passaggi algebrici per determinare il vettore δ_x, vogliamo riformulare la matrice Jacobiana attraverso la struttura a blocchi suggerita dalle componenti $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ che abbiamo introdotto in precedenza e dalla derivazione secondo le componenti delle variabili x, $\mathring{\underline{t}}$ e w:

$$\mathcal{J}(x^{(q)}, \mathring{\underline{t}}^{(q)}, w^{(q)}) = \begin{pmatrix} \partial_x\mathcal{F}_1 & \partial_{\mathring{\underline{t}}}\mathcal{F}_1 & \partial_w\mathcal{F}_1 \\ \partial_x\mathcal{F}_2 & \partial_{\mathring{\underline{t}}}\mathcal{F}_2 & \partial_w\mathcal{F}_2 \\ \partial_x\mathcal{F}_3 & \partial_{\mathring{\underline{t}}}\mathcal{F}_3 & \partial_w\mathcal{F}_3 \end{pmatrix}$$

Intendiamo ora calcolare ogni singolo blocco, partendo dalla prima riga e scorrendoli da sinistra a destra. Per non appesantire la notazione, ci permettiamo d'ora in avanti di omettere l'indice relativo all'iterazione, (q). Inoltre introduciamo subito le matrici e vettori ricorrenti nelle nostre espressioni, al fine di rendere più sintentica e lineare l'esposizione.

Definizione 3.1. *Introduciamo i seguenti elementi preferendo indicarli secondo l'ordine che si ritiene più logico per calcolarli numericamente:*

$$
\begin{aligned}
V \in \mathbb{R}_{n\times N}, \quad & (V)_{i,j} := \nu_i(t_j) \\
U \in \mathbb{R}_{n\times Ld}, \quad & U := \begin{pmatrix} \partial_1\nu_1(t_1) & \cdots & \partial_d\nu_1(t_1) & \partial_1\nu_1(t_2) & \dots & \partial_d\nu_1(t_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \partial_1\nu_n(t_1) & \cdots & \partial_d\nu_n(t_1) & \partial_1\nu_n(t_2) & \dots & \partial_d\nu_n(t_N) \end{pmatrix} \\
W \in \mathbb{R}_{Ld\times Ld}, \quad & W := Diag\ (\underbrace{w_1,\dots,w_1}_{d\text{-volte}},\dots,\underbrace{w_L,\dots,w_L}_{d\text{-volte}}) \\
r \in \mathbb{R}^N, \quad & r_i := r(t_i) \\
r' \in \mathbb{R}^{Ld}, \quad & r' := \{\partial_1 r(t_1),\dots,\partial_d r(t_1),\partial_1 r(t_2),\dots,\partial_d r(t_L)\} \\
g \in \mathbb{R}^N, \quad & g_i := -(\mathcal{F}_2(x,t))_i = r(t_i) - \langle \nu(t_i), x\rangle \\
& g = r - V^T x \\
g' \in \mathbb{R}^L d, & g' := -\mathcal{F}_3(x,\mathring{\underline{t}}) = r' - U^T x
\end{aligned}
$$

Inoltre definiamo le matrici $D_1 \in \mathbb{R}_{L\times Ld}$ e $D_2 \in \mathbb{R}_{L\times L}$.

$$
D_1 := \begin{pmatrix} \nabla(\langle \nu(t_1), x\rangle - r(t_1)) & & & \mathbb{O} \\ & \nabla(\langle \nu(t_2), x\rangle - r(t_2)) & & \\ & & \ddots & \\ \mathbb{O} & & & \nabla(\langle \nu(t_L), x\rangle - r(t_L)) \end{pmatrix}
$$

La matrice D_1, nel caso univariato, è una matrice diagonale dove l'i-esimo elemento diagonale è un vettore riga che esprime la derivata della mappa $t \mapsto \langle \nu(t), x\rangle - r(t)$ calcolato nel punto t_i. Nel caso più generale, per $d > 1$, la matrice assume una forma rettangolare, dettata dal fatto che gli elementi espressi ora nella diagonale costutiscono ciascuno un vettore riga d-dimensionale.
Invece la matrice D_2 è una matrice a blocchi diagonale dove l'i-blocco è definito dalla matrice Hessiana della mappa $t \mapsto \langle \nu(t), x\rangle - r(t)$ calcolata nel punto t_i.

Si noti che la matrice D_1 è intuitivamente il vettore g' descritto componente per componente in riga e spostandosi a quella successiva quando si cambia il nodo a cui ci si sta riferendo. Inoltre, poiché i punti $\{t_k\}_{k\le L}$ sono interni al dominio di integrazione e la mappa in esame è di classe $\mathcal{C}^2$ (si vedano le ipotesi della Definizione 2.4), allora, per il Teorema di Schwarz nella forma più generale, la matrice Hessiana è una matrice simmetrica. Queste due osservazioni rendono il calcolo di queste due matrici molto più semplice di quanto sembri.

- Il primo blocco, $\partial_x \mathcal{F}_1$, è una matrice $n \times n$.
  Ogni sua componente $(\partial_x \mathcal{F}_1)_{i,j}$ è la derivata della mappa

$$\{(t_k, w_k)\}_{k \leq N} \longmapsto \sum_{k=1}^{N} w_k \nu_i(t_k) - c_i$$

  per la componente j-esima di x.
  Tuttavia, non essendo dipendente da x, tale blocco è nullo.
  Allora $\partial_x \mathcal{F}_1 = \mathbb{O}_{n \times n}$.

- Il secondo blocco, $\partial_{\underline{t}} \mathcal{F}_1$, è una matrice $n \times Ld$.
  Ogni sua componente $(\partial_{\underline{t}} \mathcal{F}_1)_{i,j}$ è la derivata della mappa

$$\{(t_k, w_k)\}_{k \leq N} \longmapsto \sum_{k=1}^{N} w_k \nu_i(t_k) - c_i$$

  per la componenge $t_{k;s}$ tale che $j = (k-1)d + s$.
  Osserviamo che la derivata rispetto a $t_{k;s}$ è uguale a $w_k \partial_s \nu_i(t_k)$.
  Allora, tenuto conto che $\mathring{\underline{t}} = \{t_{1;1}, \ldots, t_{1;d}, t_{2;1}, \ldots, t_{L,d}\}$, si ottiene che $\partial_{\underline{t}} \mathcal{F}_1 = U\,W$.

- Il terzo blocco, $\partial_w \mathcal{F}_1$, è una matrice $n \times N$.
  Ogni sua componente $(\partial_w \mathcal{F}_1)_{i,j}$ è la derivata della mappa

$$\{(t_k, w_k)\}_{k \leq N} \longmapsto \sum_{k=1}^{N} w_k \nu_i(t_k) - c_i$$

  per il peso j-esimo.
  Quindi $(\partial_w \mathcal{F}_1)_{i,j} = \nu_i(t_j)$. Allora $\partial_w \mathcal{F}_1 = V$.

- Il quarto blocco, $\partial_x \mathcal{F}_2$, è una matrice $N \times n$.
  Ogni sua componente $(\partial_x \mathcal{F}_2)_{i,j}$ è la derivata della mappa

$$x, t_i \longmapsto \langle \nu(t_i), x \rangle - r(t_i)$$

  per la componente j-esima di x.
  Quindi $(\partial_x \mathcal{F}_2)_{i,j} = \nu_j(t_i)$. Allora $\partial_x \mathcal{F}_2 = V^T$.

- Il quinto blocco, $\partial_{\underline{t}} \mathcal{F}_2$, è una matrice $N \times Ld$.
  Ogni sua componente $(\partial_{\underline{t}} \mathcal{F}_2)_{i,j}$ è la derivata della mappa

$$x, t_i \longmapsto \langle \nu(t_i), x \rangle - r(t_i)$$

  per la componenge $t_{k;s}$ tale che $j = (k-1)d + s$.
  Osserviamo che la derivata rispetto a $t_{k;s}$ quando $i = k$ è

uguale a $\sum_{p=1}^{n} x_p \partial_s \nu_p(t_k) - \partial_s r(t_k) = \langle \partial_s \nu(t_k), x \rangle - \partial_s r(t_k)$. Quando invece ci riferiamo ad una riga associata ad un nodo t_i e desideriamo fare la derivata rispetto ad una qualche componente del nodo t_k è evidente che il risultato sia zero. In questo modo possiamo quindi osservare che le ultime $N - L$ righe saranno nulle. Allora $\partial_t \mathcal{F}_2 = \begin{pmatrix} D_1 \\ \mathbb{O}_{(N-L)\times N} \end{pmatrix}$.

- Il sesto blocco, $\partial_w \mathcal{F}_2$, è una matrice $N \times N$.
  Ogni sua componente $(\partial_w \mathcal{F}_2)_{i,j}$ è la derivata della mappa

  $$x, t_i \longmapsto \langle \nu(t_i), x \rangle - r(t_i)$$

  per il peso j-esimo.
  Tuttavia, non essendo dipendente da w, tale blocco è nullo.
  Allora $\partial_w \mathcal{F}_2 = \mathcal{O}_{N\times N}$.

- Il settimo blocco, $\partial_x \mathcal{F}_3$, è una matrice $Ld \times N$.
  Ogni sua componente $(\partial_x \mathcal{F}_3)_{i,j}$ è la derivata della mappa

  $$x, t_k \longmapsto \langle \partial_s \nu(t_k), x \rangle - \partial_s r(t_k)$$

  (dove $i = (k-1)d + s$) per la componente j-esima di x.
  Quindi per la riga associata alla coppia di indici (k, s) si ottiene $(\partial_x \mathcal{F}_3)_{i,j} = \partial_s \nu_j(t_k)$. Allora $\partial_x \mathcal{F}_3 = U^T$.

- L'ottavo blocco, $\partial_{\underline{t}} \mathcal{F}_3$, è una matrice $Ld \times Ld$.
  Ogni sua componente $(\partial_{\underline{t}} \mathcal{F}_3)_{i,j}$ è la derivata della mappa

  $$x, t_k \longmapsto \langle \partial_s \nu(t_k), x \rangle - \partial_s r(t_k)$$

  (dove $i = (k-1)d + s$) per la componenge $t_{\overline{k};\overline{s}}$ tale che $j = (\overline{k} - 1)d + \overline{s}$.
  Consideriamo la riga associata alla coppia di indici (k, s). Attraverso un'argomentazione simile ai precedenti, se ci stiamo riferendo a nodi diversi $(k \neq \overline{k})$ è naturale che il risultato sia zero. Otteniamo perciò che l'ottavo blocco è una matrice diagonale a blocchi. Nel blocco diagonale k-esimo, dove $k = \overline{k}$, si ha la valutazione della matrice Hessiana $d \times d$ nel nodo interno t_k. Da cui si giunge a definire $\partial_t \mathcal{F}_3 = D_2$.

- Il nono blocco, $\partial_w \mathcal{F}_3$, è una matrice $N \times Ld$.
  Ogni sua componente $(\partial_w \mathcal{F}_3)_{i,j}$ è la derivata della mappa

  $$x, t_k \longmapsto \langle \partial_s \nu(t_k), x \rangle - \partial_s r(t_k)$$

  (dove $i = (k-1)d + s$) per il peso j-esimo.
  Tuttavia non essendo dipendente da w, tale blocco è nullo.
  Allora $\partial_w \mathcal{F}_3 = \mathcal{O}_{N\times N}$.

Esplicitando la matrice Jacobiana e la valutazione della funzione $\mathcal{F}$, possiamo riformulare l'equazione (3.9) propria del metodo di Newton nella seguente forma

$$\begin{pmatrix} \mathbb{O} & U\,W & V \\ V^T & \begin{pmatrix} D_1 \\ \mathbb{O} \end{pmatrix} & \mathbb{O} \\ U^T & D_2 & \mathbb{O} \end{pmatrix} \begin{pmatrix} \delta_x \\ \delta_t \\ \delta_w \end{pmatrix} = \begin{pmatrix} c - V\ w \\ g \\ g' \end{pmatrix}$$

dove le matrici V, W, U, D_1, D_2 ed i vettori g, g' corrispondo a quelli descritti nella Definizione 3.1 mentre i vettori c e w esprimono rispettivamente i momenti della base scelta rispetto la misura e il dominio prescelto ed i pesi della soluzione corrente. Calcolando ad ogni iterazione i nodi come punti di minimo, è evidente che i nodi interni siano punti stazionari della mappa $t \mapsto \langle \nu(t), x \rangle - r(t)$ perciò risulta $g' = 0$ e $D_1 = \mathbb{O}$. Allora si può riscrivere il precedente sistema come segue:

$$\begin{pmatrix} \mathbb{O} & U\,W & V \\ V^T & \mathbb{O} & \mathbb{O} \\ U^T & D_2 & \mathbb{O} \end{pmatrix} \begin{pmatrix} \delta_x \\ \delta_t \\ \delta_w \end{pmatrix} = \begin{pmatrix} c - V\ w \\ g \\ 0_L \end{pmatrix} \tag{3.10}$$

Nel seguito intendiamo esplicitare i conti in modo da ottenere una serie di passaggi sequenziali per la determinazione di δ_x sfruttando la struttura a blocchi della matrice Jacobiana. Attraverso una fattorizzare QR, possiamo associare alla matrice V, le matrici $Y \in \mathbb{R}_{n\times N}$, $Z \in \mathbb{R}_{n\times(n-N)}$ e $R \in \mathbb{R}_{N\times N}$ tali che

$$C = \begin{pmatrix} Y & Z \end{pmatrix} \begin{pmatrix} R \\ \mathbb{O}_{(n-N)\times N} \end{pmatrix}$$

dove la matrice $\begin{pmatrix} Y & Z \end{pmatrix}$ è ortogonale. Definiamo d_1 e d_2, due vettori tali che $\delta_x = Y d_1 + Z d_2$. Per il secondo gruppo di equazioni del sistema (3.10) si ha:

$$g = V^T \delta_x = \begin{pmatrix} R^T & \mathbb{O} \end{pmatrix} \begin{pmatrix} Y^T \\ Z^T \end{pmatrix} \begin{pmatrix} Y & Z \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} R^T & \mathbb{O} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \tag{3.11}$$

Da cui risulta che il termine d_1 può essere facilmente ricavato dal sistema

$$R^T d_1 = g \tag{3.12}$$

È fondamentale ora supporre che la matrice D_2 sia non singolare. In altre parole stiamo richiedendo che la matrice Hessiana calcolata in ciascun nodo sia definita positiva. Tuttavia è risaputo che questa condizione è vera nel caso univariato ($d = 1$) ma è non lo è necessariamente nel caso più generale. Grazie a questa nuova ipotesi possiamo dire che esiste D_2^{-1}. Descriviamo allora il sistema delle equazioni restanti da quello emerso in (3.10):

$$\begin{cases} U^T \delta_x + D_2 \delta_t = 0 \\ U\ W \delta_t + V \delta_w = c - V w \end{cases}$$

Dalla prima equazione possiamo ricavare un'espressione per δ_t:

$$\delta_t = -D_2^{-1}U^T\delta_x$$

Sostituendo il termine δ_t nella seconda espressione si ottiene:

$$-UWD_2^{-1}U^T\delta_x + V(w+\delta_w) = c$$

Definendo la matrice $H := UWD_2^{-1}U^T$, riformuliamo l'equazione precedente:

$$H\delta_x = V(w+\delta_w) - c$$

Sostituendo attraverso la relazione $\delta_x = Yd_1 + Zd_2$, segue:

$$HYd_1 + HZd_2 = V(w+\delta_w) - c$$

Moltiplicando per la matrice di nucleo banale, Z^T, si ottiene che la relazione precedente è verificata se, e solo se,

$$Z^THYd_1 + Z^THZd_2 = Z^T[V(w+\delta_w) - c]$$

da cui otteniamo un'espressione tramite la quale esplicitare d_2 in funzione di d_1:

$$Z^THZd_2 = Z^T[V(w+\delta_w) - HYd_1 - c] \tag{3.13}$$

Ponendo attenzione al termine $Z^TV(w+\delta_w)$ si scopre:

$$\begin{aligned} Z^TV(w+\delta_w) &= Z^T\begin{pmatrix} Y & Z\end{pmatrix}\begin{pmatrix} R \\ \mathbb{O}\end{pmatrix}(w+\delta_w) \\ &= \begin{pmatrix} Z^TY & Z^TZ\end{pmatrix}\begin{pmatrix} R(w+\delta_w) \\ \mathbb{O}\end{pmatrix} \\ &= Z^TYR(w+\delta_w) \end{aligned}$$

Poiché la matrice $\begin{pmatrix} Y & Z\end{pmatrix}$ è ortogonale, è evidente che $Z^TY = \mathbb{O}$.
Perciò l'espressione (3.13) diviene

$$Z^THZd_2 = -Z^T[HYd_1 + c]$$

ricavando anche la seconda componente di δ_x che ci permette di definire l'iterazione successiva del vettore x, $x^{(q+1)} = x^{(q)} + \delta_x = Yd_1 + Zd_2$.

Giungiamo all'algoritmo proposto anche nell'articolo [18, pag. 96].

Algoritmo 3.3. *Fissato il margine di accuratezza ϵ e il margine di tolleranza τ, determinare $x^{(0)}$ tramite l'Algoritmo 3.1,*

Step-1) calcolare i punti di minimo assoluto in Ω della mappa $t \mapsto \langle \nu(t), x \rangle - r(t)$ con una tolleranza τ ed ordinarli affinché i primi L siano i punti appartenenti all'interno di Ω;

Step-2) determinare la matrice di Vandermonde V e ricavare i pesi dal problema di ottimizzazione

$$\begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \mathcal{E}_{\Lambda;2} := \|Vw - c\|_2 \\ s.a \quad & w \geq 0 \end{aligned}$$

Step-3) se $\mathcal{E}_{\Lambda;2}$ è minore della precisione richiesta allora l'algoritmo termina raggiungendo una soluzione utile altrimenti si prosegue;

Step-4) determinare le matrici Z, Y, R, U, W, D_2 ed il vettore g associati all'istanza corrente;

Step-5) se la matrice D_2 ha determinante nullo, l'algoritmo termina altrimenti calcolo la sua inversa e la matrice $H = UWD_2^{-1}U^T$;

Step-6) ricavare d_1 dal sistema $R^T d_1 = g$;

Step-7) ricavare d_2 dal sistema $Z^T H Z d_2 = -Z^T(c + HY d_1)$;

Step-8) calcolare quindi $\delta_x = Y d_1 + Z d_2$ e tornare allo Step-1 con l'istanza successiva $x^{(q+1)} = x^{(q)} + \delta_x$.

Si osservi che richiedendo allo *Step-2* la risoluzione di quel particolare problema di ottimizzazione, si vincola la soluzione a rispettare la disequazione del sistema (3.8). Questo fa intuitivamente pensare che l'algoritmo converga alla soluzione ottima tuttavia ciò non è semplice da dimostrare. Pertanto ci permettiamo di non affrontare questo delicato aspetto ma di osservare sperimentalmente quanto succede sottolineando però la fondamentale importanza che ricopre il partire da una soluzione $x^{(0)}$ vicina alla soluzione ottima. [1]

[1] Si potrebbe pensare di applicare il metodo di Newton in modo "puro", ma esso evidentemente modificherebbe soltanto la posizione dei nodi interni lasciando invariati quelli appartenenti alla frontiera di Ω. Questa idea potrebbe essere presa in considerazione qualora riuscissimo a caratterizzare una funzione di sensibilità in grado di garantire che i nodi ottimi non si trovino nella frontiera. Inoltre, in questo modo non si richiede la positività dei pesi: vincolo costituente le condizioni di ottimalità date dal sistema 3.8. Il vero vantaggio di tale idea, risulterebbe quello di impiegare l'informazione δ_t nella determinazione dei nodi all'istanza successiva tramite l'espressione $\underset{\sim}{t}^{(q+1)} = \underset{\sim}{t}^{(q)} + \delta_{\underset{\sim}{t}} = \underset{\sim}{t}^{(q)} - D_2^{-1}U^T\delta_x$, riducendo di conseguenza la complessità computazionale. Il lettore può affrontare tale argomento attraverso l'articolo [18].

Tuttavia siamo intenzionati ad introdurre un miglioramento all'Algoritmo 3.3 oggetto di un lavoro sempre dovuto a G.A. Watson [2]. Rimandiamo al testo originale il suo studio e ci limitiamo alla sua esposizione. L'obbiettivo è modificare lo *Step-7* affinché la matrice del sistema lineare sia invertibile. L'idea è sommare alla matrice H una matrice unitaria moltiplicata per uno scalare μ (quanto più piccolo possibile) in grado da rendere la matrice $Z^T(H+\mu\mathbb{I})Z$ definita positiva. Tuttavia, considerata l'imprecisione introdotta, sarà opportuno limitare lo spostamento δ_x di questa iterazione introducendo un fattore γ per il quale $x^{(q+1)} = x^{(q)} + \gamma\delta_x$. L'articolo discute quindi di come scegliere la coppia μ,γ e introduce una funzione di penalità $P_\sigma : \mathbb{R}^n \to \mathbb{R}$ di parametro σ che per questioni di stabilità, deve essere più grande di tutti i pesi w_k,

$$P_\sigma(\tilde{x}) = \langle c, \tilde{x}\rangle + \sigma \sum_{k=1}^{L} \left[r(\tilde{t}_k) - \langle \nu(\tilde{t}_k), \tilde{x}\rangle\right]_+$$

dove $\tilde{t}_k$ sono i punti di minimo assoluto associato alla mappa $t \mapsto \langle \nu(t_k), \tilde{x}\rangle - r(t_k)$. Allora si richiedere di determinare il più piccolo σ e il più grande γ tali che

$$T(\gamma, x) = \frac{P(x+\gamma\delta_x) - P(x)}{\gamma \partial_{\delta_x} P(x)} \geq \rho$$

dove ∂_{δ_x} indica la derivata direzionale di direzione δ_x mentre il parametro ρ è impostato al valore $\rho = 0.0001$ per motivi sperimentali.

Algoritmo 3.4. *Si descrive la seguente modifica dello Step-7 dell'Algoritmo 3.3.*

Step-7 bis) fissare $\mu = 0$ e $\mu_{min} = 1$,

i) ricavare d_2 dal sistema $Z^T(H+\mu\mathbb{I})Zd_2 = -Z^T(c + HYd_1)$ e calcolare $\delta_x = Yd_1 + Zd_2$;

ii) se la matrice $Z^T(H+\mu\mathbb{I})Z$ è definita positiva allora passare allo Step-8 ponendo $\delta_x = \gamma\,\delta_x$, altrimenti proseguire;

iii) porre $\gamma = 1$ e calcolare $T(\gamma, x)$;

iv) se quest'ultimo valore non è maggiore di ρ allora dimezzo γ e ripeto così sino a raggiungere quella disuguaglianza;

v) se accade che $\mu_{min} = \mu$ allora si aggiorna il primo come segue

$$\mu_{min} = \begin{cases} \frac{1}{4}\ \mu_{min} & \text{se } T(\gamma, x) > 0.5 \text{ e } \gamma = 1 \\ 4\ \mu_{min} & \text{se } \gamma \leq \frac{1}{4} \end{cases}$$

vi) aggiornare l'istanza $\mu = 4\mu + \mu_{min}$ e ripetere dal punto i);

Capitolo 4

Risultati Numerici

In questo capitolo si intende dare spazio ai risultati numerici, ovvero ad analizzare le formule di cubatura di ordine m risultanti dal metodo discusso nelle pagine precedenti. Ricordiamo brevemente al lettore che tale metodo si riduce al voler risolvere il problema di ottimizzazione LSIP

$$\begin{aligned} \mathcal{D}(r): \quad & \sup_{\lambda \in \mathbb{R}_+^{(\Omega)}} \quad \sum_{t \in \Omega} \lambda_t \, r(t) \\ & \text{s.a} \qquad \sum_{t \in \Omega} \lambda_t \, \nu(t) = c, \end{aligned}$$

in cui c rappresenta il vettore dei momenti associati alla base $\{\nu_1(\cdot), \ldots, \nu_n(\cdot)\}$ dello spazio vettoriale $\mathbb{P}_m^d$ ed $r(\cdot)$ è la funzione di sensibilità. Pertanto la scelta della base e di tale funzione rappresentano un aspetto chiave che affronteremo attraverso gli esempi di questo capitolo. Tuttavia, nel corso dello studio teorico, abbiamo assunto alcune ipotesi quali la compattezza del dominio di integrazione e la continuità, la non appartenenza allo spazio $\mathbb{P}_m^d$, la μ-integrabilità e la superiore limitatezza della funzione di sensibilità (si veda la Sezione 2.2), le quali restringono la libertà di scelta. Si è visto inoltre che gli algoritmi discussi nel Capitolo 3 vertono sulla risoluzione del problema nella forma primale associato a $\mathcal{D}(r)$ e per questo motivo abbiamo più volte sottolineato come la formula risultante possieda un errore sui momenti che abbiamo indicato attraverso la notazione

$$\mathcal{E}_{\Lambda;\cdot} := \|Vw - c\|$$

dove V indica la matrice di Vandermonde calcolata nei nodi rispetto la base scelta. In concordanza con tali osservazioni, è nostro interesse valutare l'errore dei momenti e la cardinalità dei nodi della formula. Studieremo in primo luogo il caso univariato soffermandoci in modo particolare sulla valutazione di differenti basi e funzioni di sensibilità. Successivamente, porremo la nostra attenzione sul quadrato unitario, oggetto di esempio nell'articolo [15], del quale vogliamo evidenziare l'aspetto critico che è emerso dal nostro studio.

4.1 Caso Univariato

Nel caso di un dominio univariato abbiamo già evidenziato dagli studi di R. Bojanic e R. DeVore [1] una scelta per la funzione di sensibilità. Come abbiamo dimostrato nel Corollario 2.2, se la funzione di sensibilità è $(m+1)$-volte differenziabile e si ha nel dominio (ove definita) $r^{(m+1)} \leq 0$, allora la formula di quadratura coincide con quella di Gauss associata alla misura μ. Mentre se $r^{(m+1)} \geq 0$ nel dominio, allora coincide con quella di Gauss-Lobatto. Ci proponiamo pertanto di studiare differenti funzioni di sensibilità, basi e misure. Premettiamo alle tabelle e illustrazioni, mediante le quali sintetizzeremo i risultati ottenuti, qualche appunto sulle idee applicate per implementare gli algoritmi proposti nel Capitolo 3. Come si è visto, è di fondamentale importanza la determinazione dei punti di minimo di una mappa reale sull'insieme Ω. Per quanto riguarda il caso univariato, abbiamo distinto due situazioni alle quali abbiamo applicato due differenti metodi. Se la funzione di sensibilità è un polinomio, allora la mappa da minimizzare risulterà essere anch'essa un polinomio, di conseguenza i punti stazionari possono essere facilmente ricavabili mediante la routine `root` di MATLAB. In caso contrario, questi punti vengono calcolati attraverso la libreria `chebfun` di nota conoscenza. Gli esperimenti numerici di questa sezione sono stati svolti attraverso l'ambiente di calcolo MATLAB versione 2015a, adoperando un processore Intel® Core™2 Quad Q6600 a 2.4GHz. Inoltre la griglia iniziale della prima fase è stata impostata per avere una distanza tra i punti di 0.01 mentre come margine di tolleranza è stato scelto $\tau = 0.05$ e come criterio di arresto per la prima fase $\epsilon = 10^{-8}$ (si veda la Sezione 3.1 per l'intepretazione di tali parametri).

4.1.1 Formula di quadratura di Gauss

Affrontiamo innanzitutto lo studio della misura di Lebesgue sull'intervallo [-1,1]. Con l'intenzione di ottenere la formula gaussiana, utilizziamo inizialmente, come suggerito da Ryu e Boyd, la funzione di sensibilità $r(t) = -t^{m+1}$. Quale base dello spazio $\mathbb{P}_m$, impieghiamo la base canonica ottenendo, mediante l'Algoritmo 3.3, i risultati sintetizzati nella Tabella 4.1.
Il valore $\mathcal{E}_F$ indica l'errore sulla condizione di ammissibilità rispetto al problema primale commesso dalla soluzione: si tratta di una quantità che, se negativa, esprime di quanto il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazone da sopra è sottostante ad essa. Si tratta di un errore indicativo per comprendere il criterio di arresto della prima fase e per questo motivo, ai lettori più interessati, può rappresentare un dato significativo. Mentre i valori di $\mathcal{E}_{\Lambda;\infty}$ e $\mathcal{E}_{\Lambda;2}$ indicano l'errore commesso dalla formula rispetto i momenti nella norma infinito e nella norma 2. In questo esempio, come pure nei successivi, il tempo di calcolo non è competitivo con quelli più noti quali Golub-Welsch [11], Glaser-Liu-Rokhlin [5] e più recentemente quello descritto nell'articolo [12].

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	2	0.000000e+00	2.220446e-16	2.989367e-16	1.598 sec.
5	3	-1.387779e-16	5.828671e-16	9.226381e-16	0.535 sec.
7	4	-5.551115e-17	5.551115e-16	6.674336e-16	0.785 sec.
9	7	-1.900815e-09	9.661304e-08	1.565212e-07	1.277 sec.
11	8	-1.024886e-09	7.722608e-07	1.130152e-06	1.825 sec.
13	8	-4.520905e-10	1.149515e-06	2.369576e-06	2.396 sec.
15	10	-2.427203e-09	4.627417e-06	8.137540e-06	1.245 sec.
17	11	-9.143014e-10	2.753767e-09	5.647998e-09	1.564 sec.
19	11	-1.211129e-09	8.464116e-07	1.916237e-06	1.771 sec.

Tabella 4.1: $\Omega = [-1, 1]$: base canonica e $r(t) = -t^{m+1}$.

Dalla Tabella 4.1 si osservano degli errori sui momenti non trascurabili per $m \geq 9$. Visto che il numero di nodi delle formule gaussiane è $N = \left[\frac{m}{2}\right] + 1$, è evidente che la scelta della base per questi gradi di precisione non è opportuna. Infatti, come anticipato nel Capitolo 3, la seconda fase implementata opera risolvendo sistemi lineari connessi alla matrice di Vandermonde riferita alla base prefissata. Pertanto, una sua scelta non attenta, può caratterizzare sistemi mal condizionati e potenzialmente una cattiva convergenza alla soluzione del problema $\mathcal{D}(r)$. È noto infatti che la base canonica determina proprio questo effetto e questo giustifica le criticità osservate da questo primo test.

Sulla base di questa osservazione, differendo quindi dall'esposizione di Ryu-Boyd, siamo maggiormente propensi a studiare una diversa base e, a questo scopo, abbiamo individuato quella costituita dai polinomi di Chebyshev $T_n \in \mathbb{P}_n$ definiti attraverso la struttura ricorsiva seguente:

$$\begin{cases} T_0(t) \equiv 1,\ T_1(t) = t, \\ T_n(t) = 2\ T_{n-1}(t) - T_{n-2}(t),\ n \geq 2 \end{cases}$$

Applicando il medesimo Algoritmo 3.3 ed utilizzando ora la base di Chebyshev otteniamo i risultati descritti nella Tabella 4.2.

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	2	1.387779e-17	3.330669e-16	4.002966e-16	1.210 sec.
5	3	0.000000e+00	9.436896e-16	1.253618e-15	0.689 sec.
7	4	-5.551115e-17	8.899131e-16	1.410902e-15	0.780 sec.
9	5	-1.301043e-17	4.940492e-15	9.495744e-15	1.339 sec.
11	6	4.211342e-18	1.523781e-14	3.027503e-14	1.665 sec.

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
13	7	-1.831868e-15	5.973000e-14	9.578831e-14	1.976 sec.
15	8	-6.217249e-15	2.031583e-12	1.303987e-23	1.978 sec.
17	9	-5.773160e-15	9.852924e-12	3.661438e-22	1.569 sec.
19	10	-1.143530e-14	9.224835e-11	2.224641e-20	1.812 sec.

Tabella 4.2: $\Omega = [-1, 1]$: base di Chebyshev e $r(t) = -t^{m+1}$.

Emergono dei risultati sensibilmente migliori in termini di errore sul momento mentre i tempi di esecuzione rimangono invariati. Si osserva quindi la consistenza tra la teoria studiata nella Sezione 2.4 e quanto emerge dalla sua applicazione. Tuttavia, allo scopo di confrontare differenti funzioni di sensibilità, impieghiamo ora il polinomio $(m + 1)$-esimo di Chebyshev (moltiplicato per -1 per garantire che $r^{(m+1)} \leq 0$).

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	2	-3.330669e-16	6.661338e-16	7.529899e-16	1.092 sec.
5	3	-1.776357e-15	5.551115e-16	8.308148e-16	0.924 sec.
7	4	-1.421085e-14	1.054712e-15	1.763312e-15	1.420 sec.
9	5	-2.842171e-14	2.997602e-15	5.225687e-15	1.766 sec.
11	6	-1.023182e-12	6.591949e-15	1.256064e-14	4.060 sec.
13	7	-2.273737e-12	3.111400e-14	5.915914e-14	5.074 sec.
15	8	-3.637979e-12	2.066403e-13	3.920105e-13	5.798 sec.
17	9	-1.818989e-12	1.169287e-12	2.941721e-12	8.097 sec.
19	10	-7.566996e-10	5.620490e-12	1.262790e-11	7.296 sec.

Tabella 4.3: $\Omega = [-1, 1]$: base di Chebyshev e $r(t) = -T_{m+1}$.

In questa situazione si evidenziano dei risultati (descritti nella Tabella 4.3) leggermente migliori (in funzione degli errori $\mathcal{E}_{\Lambda;\cdot}$) rispetto ai precedenti ma con tempi di esecuzione in diversi casi più alti. L'aspetto però più importante è constatare, attraverso gli ultimi due casi studiati, la validità del metodo dal punto di vista pratico rendendolo un'alternativa interessante rispetto ai metodi noti in letteratura, anche se ad oggi meno rilevante in quanto a tempi di calcolo.

Concludiamo questa sottosezione con la Figura 4.1 che rappresenta la soluzione del problema primale per il caso $m = 13$ dell'ultimo test. In tale illustrazione sono rappresentate la funzione di sensibilità, il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra e i punti di contatto, le cui ascisse coincidono con i nodi della formula di cubatura, in virtù della dualità (si veda il Corollario 2.1).

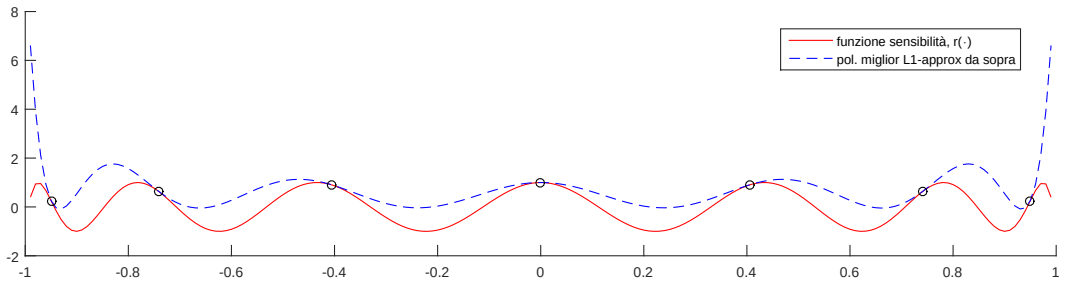

Figura 4.1: Rappresentazione di $r(t)$ e del relativo polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione per $m = 13$ e con $r(t) = -T_{14}(t)$, con μ misura di Lebesgue.

4.1.2 Formula di quadratura di Gauss-Lobatto

Intendiamo ora riprodurre le prove relativamente alle formule di Gauss-Lobatto, aspettandoci un comportamento del tutto analogo. Ricordiamo che queste sono caratterizzate dal possedere un grado di precisione pari a $2N-3$, dove N indica il numero di nodi ed inoltre gli estremi dell'intervallo sono nodi di quadratura. Per la teoria introdotta, è sufficiente considerare quali funzioni di sensibilità, funzioni $(m+1)$-volte differenziabili tali che $r^{(m+1)} \leq 0$ in Ω. Considereremo quindi, similmente allo studio precedente, $r(t) = t^{m+1}$ e $r(t) = T_{m+1}(t)$.

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	3	-9.861143e-14	7.771561e-16	9.036561e-16	2.169 sec.
4	3	-2.775558e-17	4.440892e-16	6.672893e-16	4.893 sec.
5	4	1.387779e-17	4.440892e-16	6.826969e-16	3.650 sec.
6	4	-2.428613e-17	5.551115e-16	8.498375e-16	3.255 sec.
7	5	-4.163336e-17	2.220446e-16	3.127895e-16	3.625 sec.
8	6	-2.542966e-13	4.452549e-13	6.961727e-13	5.359 sec.
9	6	-2.775558e-17	4.440892e-16	6.140823e-16	6.037 sec.
10	7	-1.726420e-09	4.053753e-08	6.790791e-08	8.840 sec.
11	7	-3.330669e-16	6.591949e-16	1.033499e-15	7.091 sec.
12	8	-6.815711e-09	5.355120e-08	8.964637e-08	7.934 sec.
13	8	-5.551115e-17	6.175616e-16	1.061401e-15	6.843 sec.
14	9	-1.564085e-12	5.226930e-12	8.950847e-12	13.008 sec.
15	9	-4.440892e-16	3.851086e-16	6.734660e-16	9.275 sec.
16	9	-1.110223e-16	1.415534e-15	2.998710e-15	9.439 sec.
17	10	-1.110223e-16	7.216450e-16	1.737580e-15	8.433 sec.
18	11	-2.921321e-09	5.944290e-08	1.251946e-07	14.795 sec.
19	11	-1.221245e-15	8.604228e-16	1.931632e-15	41.497 sec.

Tabella 4.4: $\Omega = [-1, 1]$: base canonica e $r(t) = t^{m+1}$.

Scelta la base canonica e la funzione di sensibilità $r(t) = t^{m+1}$, si ottengono, mediante l'Algoritmo 3.3, i risultati illustrati nella Tabella 4.4. Come nel caso precedente, si evidenzia la presunta influenza negativa della base canonica che, per gli m pari e maggiori di 8, determina una cardinalità dei nodi differente da quella nota per le formule di Gauss-Lobatto ed un errore relativamente alto.

Allo scopo di sopperire a tale difficoltà, risolviamo il problema di ottimizzazione descritto dalla medesima funzione di sensibilità rispetto alla base di Chebyshev, diversamente da prima. Tuttavia i risultati ottenuti, illustrati dalla Tabella 4.5, dimostrano invece che questa scelta non è interessante quanto lo era nello studio delle formule gaussiane. Infatti si evidenzia un comportamento del tutto simile a quello della prova precedente ma per $m \geq 12$ ed inoltre si rilevano tempi di esecuzione più elevati.

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	3	-2.462189e-32	2.775558e-16	3.723801e-16	2.538 sec.
4	3	-1.665335e-16	5.828671e-16	8.657782e-16	4.833 sec.
5	4	-2.220446e-16	2.220446e-16	3.015027e-16	3.664 sec.
6	4	-4.907709e-18	8.881784e-16	1.275927e-15	4.404 sec.
7	5	-2.220446e-16	1.526557e-16	2.799738e-16	5.088 sec.
8	5	-4.163336e-17	7.632783e-16	1.154362e-15	4.660 sec.
9	6	-4.440892e-16	4.440892e-16	7.616365e-16	5.171 sec.
10	6	-3.330669e-16	5.551115e-16	1.050618e-15	7.828 sec.
11	7	-1.653408e-18	7.216450e-16	1.435949e-15	7.801 sec.
12	8	-1.361755e-09	1.916158e-08	3.406383e-08	12.166 sec.
13	8	-3.330669e-16	4.440892e-16	9.154608e-16	8.454 sec.
14	9	-4.480899e-09	1.312550e-07	2.318912e-07	10.689 sec.
15	9	-2.220446e-16	7.771561e-16	1.793593e-15	141.303 sec.
16	10	-1.794705e-09	3.607312e-08	7.856356e-08	22.471 sec.
17	10	-5.551115e-17	7.771561e-16	1.733655e-15	160.275 sec.
18	11	-6.217249e-15	9.200973e-15	2.099598e-14	61.458 sec.
19	11	-2.220446e-16	6.557255e-16	1.902858e-15	99.151 sec.

Tabella 4.5: $\Omega = [-1, 1]$: base di Chebyshev e $r(t) = t^{m+1}$.

Risultati molto più soddisfacenti si ottengono considerando la base di Chebyshev e scegliendo come funzione di sensibilità $r(t) = T_{m+1}(t)$. In questo caso, i risultati descritti dalla Tabella 4.6, forniscono dei risultati più consistenti a quanto ci aspettiamo fornendoci in modo empirico una dipendenza quasi lineare del tempo e degli errori sui momenti in funzione del grado di precisione m richiesto.

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	3	-1.549871e-13	4.440892e-16	5.467214e-16	2.069 sec.
4	3	-1.776357e-15	2.775558e-16	3.040471e-16	5.374 sec.
5	4	0.000000e+00	6.661338e-16	1.177569e-15	5.857 sec.
6	4	-3.552714e-15	8.881784e-16	1.605511e-15	5.575 sec.
7	5	-1.421085e-14	7.910339e-16	1.473988e-15	5.530 sec.
8	5	-1.598721e-14	1.332268e-15	1.988595e-15	15.315 sec.
9	6	-3.183218e-16	3.372302e-15	6.935257e-15	9.797 sec.
10	6	-1.519618e-15	8.111567e-15	1.675014e-14	35.546 sec.
11	7	-1.136868e-13	6.918077e-15	1.317960e-14	41.620 sec.
12	7	-6.821210e-13	3.685247e-14	7.518127e-14	38.899 sec.
13	8	-2.728484e-12	4.531792e-14	8.843043e-14	44.921 sec.
14	8	-4.014566e-13	1.540434e-13	2.925140e-13	76.225 sec.
15	9	-4.405365e-13	1.734723e-13	3.956618e-13	77.156 sec.
16	9	-8.731149e-11	4.226203e-13	8.938054e-13	79.373 sec.
17	10	-1.018634e-10	1.685992e-12	3.239545e-12	88.955 sec.
18	10	-6.984919e-10	4.453590e-12	9.828228e-12	84.756 sec.
19	11	-3.492460e-10	6.785426e-12	1.630039e-11	95.018 sec.

Tabella 4.6: $\Omega = [-1, 1]$: base di Chebyshev e $r(t) = T_{m+1}(t)$.

Tuttavia è interessante osservare che considerando il primo e l'ultimo test, i risultati ottenuti nel primo sono incostanti per m pari ma per i gradi di precisioni dispari offrono una miglior precisione di quello appena commentato.

Concludiamo questa sottosezione con la Figura 4.2 che rappresenta la soluzione del problema primale per il caso $m = 13$ dell'ultimo test. In tale illustrazione sono rappresentate la funzione di sensibilità, il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra e i punti di contatto, le cui ascisse coincidono con i nodi della formula di cubatura, in virtù della dualità (si veda il Corollario 2.1).

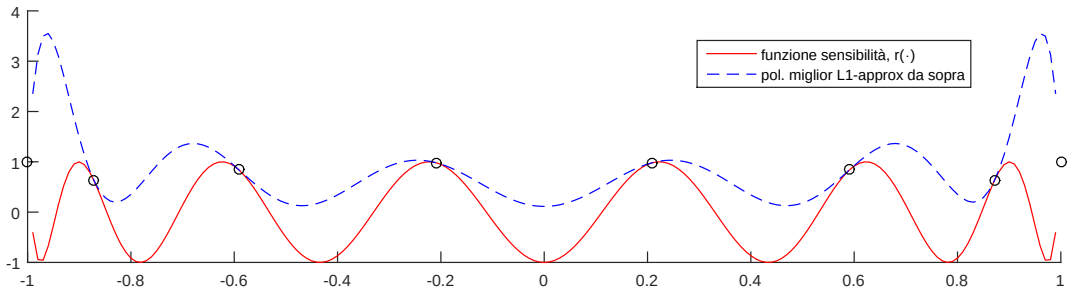

Figura 4.2: Rappresentazione di $r(t)$ e del relativo polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione per $m = 13$ e con $r(t) = T_{14}(t)$, con μ misura di Lebesgue.

4.1.3 Formula di quadratura per la misura di Chebyshev

Intendiamo ora analizzare il comportamento dell'algoritmo proposto nella determinazione di una formula di quadratura gaussiana per una misura differente da quella di Lebesgue. Consideriamo ad esempio quella determinata dalla funzione peso $\omega(t) = \frac{1}{\sqrt{1-t^2}}$. Scegliendo la base di Chebyshev e come funzione di sensibilità $r(t) = -T_{m+1}(t)$, abbiamo riscontrato i risultati sintetizzati nella Tabella 4.7 dall'Algoritmo 3.3. I risultati ottenuti rispecchiano quanto visto per la misura di

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	2	-8.881784e-16	6.661338e-16	9.420555e-16	1.298 sec.
5	3	-1.110223e-16	1.099847e-15	1.382839e-15	1.175 sec.
7	4	0.000000e+00	2.553513e-15	4.150579e-15	1.770 sec.
9	5	-5.684342e-14	1.068649e-14	1.528409e-14	2.995 sec.
13	7	-1.000444e-11	1.668110e-13	4.665920e-13	4.858 sec.
15	8	-4.365575e-11	1.040945e-12	2.610753e-12	5.575 sec.
17	9	-2.182787e-10	3.780781e-12	1.037112e-11	6.824 sec.
19	10	-6.106227e-16	1.443140e-11	3.942981e-11	10.089 sec.

Tabella 4.7: $\Omega = [-1, 1]$ e $\omega(t) = \frac{1}{\sqrt{1-t^2}}$: base di Chebyshev e $r(t) = -T_{m+1}(t)$.

Lebesgue, confermando ancora una volta che l'algoritmo è coerente con quanto atteso. Abbiamo constatato i fenomeni che hanno accomunato i due casi precedenti quando si sceglie o la base canonica e la funzione di sensibilità $r(t) = -t^{m+1}$ oppure la base di Chebyshev e $r(t) = -t^{m+1}$, tuttavia abbiamo preferito ometterli in questa tesi per non sembrare ripetitivi. Facendo invece un confronto con la misura di Lebesgue, notiamo che il comportamento in termini di tempo ed errori è da considerarsi analogo e non vi sono altri particolari da osservare.

4.1.4 Insieme non connesso

Studiamo come ultimo esempio dell'ambito univariato, un dominio compatto ma non connesso e a tale scopo si è scelto l'insieme $\Omega = \left[-1, -\frac{1}{2}\right] \cup \left[\frac{1}{2}, 1\right]$ a cui associamo nuovamente la misura di Lebesgue. Limitandoci alla scelta della base di Chebyshev e al polinomio $-T_{m+1}(\cdot)$ come funzione di sensibilità, l'Algoritmo 3.3 ha dato luogo ai risultati descritti nella Tabella 4.8. Emerge come le formule

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
3	2	-1.332268e-15	2.220446e-16	2.306595e-16	0.641 sec.
4	3	-1.332268e-15	8.881784e-16	1.027800e-15	10.274 sec.
5	4	5.551115e-16	5.551115e-16	5.847742e-16	3.797 sec.
6	4	0.000000e+00	7.077672e-16	1.163781e-15	7.784 sec.

m	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
7	4	-7.815970e-14	1.592476e-15	1.735525e-15	2.617 sec.
8	5	2.220446e-15	4.831552e-14	1.000732e-13	8.943 sec.
9	6	-1.705303e-13	4.690692e-15	7.952413e-15	13.818 sec.
10	6	-3.410605e-13	1.510913e-12	3.552918e-12	12.420 sec.
11	6	-1.364242e-12	3.708145e-14	7.316136e-14	2.536 sec.
12	7	-3.637979e-12	4.656596e-13	8.731339e-13	48.567 sec.
13	8	-8.185452e-12	2.111089e-13	3.936798e-13	64.505 sec.
14	8	-1.273293e-11	2.240644e-11	5.778620e-11	168.469 sec.
15	8	-6.912160e-11	9.726664e-13	2.043960e-12	6.194 sec.
16	9	-1.600711e-10	1.874483e-11	4.672148e-11	830.919 sec.
17	10	-6.257324e-10	1.852574e-12	4.364341e-12	118.729 sec.
18	10	-7.858034e-10	8.171993e-11	2.013658e-10	460.141 sec.
19	10	-2.677552e-09	6.175149e-11	1.239021e-10	15.824 sec.

Tabella 4.8: $\Omega = \left[-1, -\frac{1}{2}\right] \cup \left[\frac{1}{2}, 1\right]$: base di Chebyshev e $r(t) = -T_{m+1}(t)$.

risultanti da questo metodo non possano coincidere con le formule di quadrature gaussiane note poiché quest'ultime possiedono esattamente $N = \left[\frac{m}{2}\right] + 1$ nodi. Invece la cardinalità dei nodi che si osserva dalla Tabella 4.8 non rispetta tale caratteristica per tutti i gradi di libertà: curiosamente solo per gli m appartenenti all'insieme $4\mathbb{Z} + 1$ e considerati nello studio, questa proprietà non vale. Tuttavia questo risultato non contraria la teoria affrontata nella Sezione 2.4 poiché il lettore osserverà che le costruzioni usate per dimostrare ad esempio la Proposizione 2.4 impiegano la caratteristica dell'intervallo $[-1, 1]$ di essere connesso. Valutando quindi i dati sintetizzati nella Tabella 4.8, si può constatare che questo metodo risulta ancora una volta poco efficace se comparato a quelli già noti sia in termini di tempi di esecuzione, sia analizzando gli errori $\mathcal{E}_{\Lambda;\cdot}$ determinati.

Intendiamo infine concludere questo esempio attraverso la Figura 4.3 che rappresenta la soluzione del problema primale per il caso $m = 9$ del test. Ricordiamo che esso rappresenta uno dei casi in cui non vale la minimalità e proprio per questo vogliamo dare al lettore l'opportunità di confrontare visivamente la soluzione.

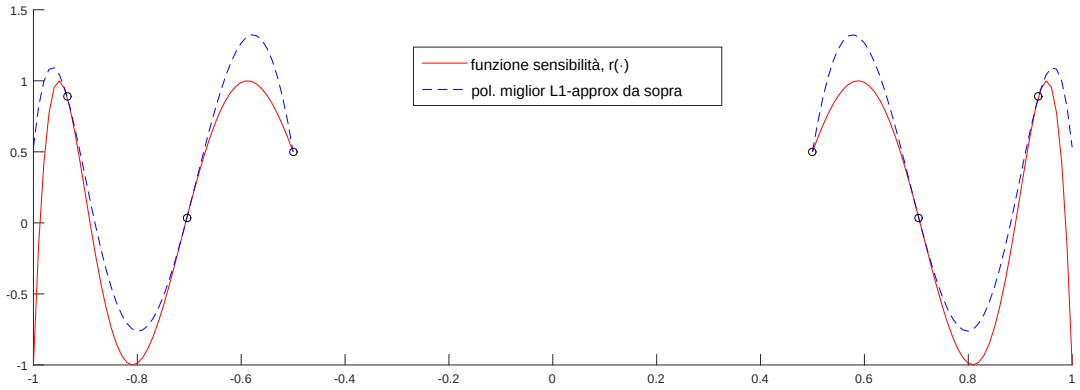

Figura 4.3: Rappresentazione dei punti di contatto tra $r(t) = T_{10}(t)$, $m = 9$, e il suo polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione, sull'insieme $\Omega = \left[-1, -\frac{1}{2}\right] \cup \left[\frac{1}{2}, 1\right]$ con la misura di Lebesgue.

4.2 Caso Multivariato

Studiato il caso univariato, siamo interessati ad estendere questa analisi al caso multivariato. La generalità con cui abbiamo elaborato il Capitolo 3 ci permette di applicare in modo immediato gli algoritmi in esso descritti. In virtù di questa proprietà, i programmi che abbiamo implementato in ambiente MATLAB e consultabili nell'Appendice B, sono orientati alla programmazione ad oggetti. Questo paradigma della programmazione, permette all'utilizzatore di considerare nuove casistiche, modificando solo la classe dell'oggetto di partenza (si veda l'implementazione della classe *class univariate poly sensitive*, *class bivariate poly sensitive* e le relative classi per i domini di prova). Ogni classe possiede dei metodi richiamati dalle routine che implementano il *Discretization Method* e il *Local Reduction Method*. Tra di essi, vi è quello che determina i minimi per una mappa multivariata su un certo dominio. Questo problema non è di facile risoluzione in ambito del tutto generale ma focalizzando la nostra attenzione nel caso bivariato, abbiamo distinto due situazioni, analoghe al caso univariato. Se la funzione di sensibilità scelta è un polinomio, la libreria `tensorlab` offre una concreta ed efficace implementazione per la determinazione dei punti stazionari. Nel caso di una funzione di sensibilità più generale, l'impiego della libreria `chebfun2` ci ha permesso di ricavare dei risultati anche se con tempi di esecuzione molto più alti. Gli esperimenti numerici di questa sezione sono stati svolti attraverso l'ambiente di calcolo MATLAB versione 2015a, adoperando un processore Intel® Core™2 Quad Q6600 a 2.4GHz. Inoltre la griglia iniziale della prima fase è stata costruita combinando coordinate equidistanti di parametro $h = 0.1$ mentre il margine di tolleranza è stato scelto dipendentemente dalla geometria del problema e come criterio di arresto per la prima fase $\epsilon = 10^{-8}$ (si veda la Sezione 3.1 per l'intepretazione di tali parametri).

4.2.1 Studio del quadrato unitario

Come anticipato nell'introduzione, è nostra intenzione mettere in discussione gli esempi dell'articolo pubblicato da E. K. Ryu e S. P. Boyd. Dalle osservazione che si sono susseguite in questo capitolo, è ormai chiaro il ruolo importante che riveste l'errore sui momenti $\mathcal{E}_{\Lambda;\cdot}$. Negli esempi del loro articolo non viene analizzato o messo in evidenza tale quantità, affermando con molta disinvoltura che le formule rappresentate siano valide. Inoltre non vengono resi noti gli effettivi nodi e pesi di cubatura e neppure un'implementazione attraverso la quale il lettore possa riscontrare la veridicità.

È nostra intenzione ripercorrere criticamente i risultati illustrati in Figura 5 di [15] relativamente a delle formule di cubatura di grado 5, nel quadrato unitario $\Omega = [-1, 1] \times [-1, 1]$. La prima rappresentazione è data da una funzione di sensibilità $r(x, y) = -x^6 - y^6$ e ad essa hanno associato una formula di cubatura costituita da 9 nodi, mentre la seconda è data da una funzione di sensibilità $r(x, y) = -\cos\left(\frac{\pi}{2}\sqrt{(x+1)^2 + (y+1)^2}\right)$ a cui corrispondono 8 nodi.

Scelta la base canonica, abbiamo riprodotto lo studio per la prima funzione di sensibilità proposta attraverso l'Algoritmo 3.1, con l'impiego dell'Algoritmo 3.2 per determinare la soluzione del problema $\mathcal{D}(r)$ associata alla soluzione primale corrente. A seguito di considerazioni sperimentali che presto giustificheremo, abbiamo scelto come margine di tolleranza $\tau = 0.025$. I risultati ottenuti sono descritti descritti nella Tabella 4.9.

iterazione	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
I	9	-6.389050e-03	3.028004e-02	2.429565e-03	1.282 sec.
II	9	-7.114229e-05	7.050764e-03	1.823580e-04	24.626 sec.
III	9	-6.530039e-07	2.580611e-04	2.855708e-07	76.447 sec.
IV	9	-6.793016e-09	4.652257e-05	4.688662e-09	123.077 sec.
V	9	-7.181582e-11	8.209197e-06	1.978464e-10	186.933 sec.
VI	9	-2.431574e-12	7.114470e-07	1.654703e-12	291.401 sec.
VII	9	-1.828849e-12	1.166260e-06	4.693013e-12	420.945 sec.

Tabella 4.9: $m = 5$ e $\Omega = [-1, 1] \times [-1, 1]$: base canonica e $r(x, y) = -x^6 - y^6$.

Emerge innanzitutto che la complessità algoritmica è cresciuta notevolmente rispetto al caso univariato determinando tempi di esecuzione molto lunghi. Si nota che i tempi di esecuzione crescono esponenzialmente ad ogni iterazioni mentre l'errore non sembra diminuire sensibilmente. Questo aspetto lo avevamo già osservato nel caso univariato (si veda Figura 3.1) e per questo motivo abbiamo introdotto la seconda fase discussa nella Sezione 3.2. Purtroppo non siamo stati

in grado, sperimentalmente, di raggiungere tramite la prima fase una soluzione alla quale applicare la seconda, capace di convergere più rapidamente.

Tramite la Figura 4.4 intendiamo rappresentare i pesi e i nodi che costituiscono la formula di quadratura risultante dalla settima iterazione descritta nella Tabella 4.9. Vista la somiglianza della Figura 4.4 con la Figura 2a dell'articolo [15] ed in assenza di un'analisi dell'errore $\mathcal{E}_{\Lambda;\cdot}$, possiamo solo ipotizzare che Ryu-Boyd abbiano impiegato un algoritmo di simile natura. Emerge allora dai dati della Tabella 4.9 la criticità che avevamo sottolineato fin dall'introduzione: i tempi e gli errori costituiscono un quadro di insieme che permette di dare una visione completa alla discussione, differente da quella espressa dagli autori. È evidente la non competitività rispetto ad altri metodi già noti, sia in termini di errore sia in quelli di tempo di esecuzione. Si consideri ad esempio l'articolo [4], nel quale per il grado di precisione $m = 5$ è stata determinata una formula minimale composta da 7 nodi, in tempi di esecuzione ben più brevi.

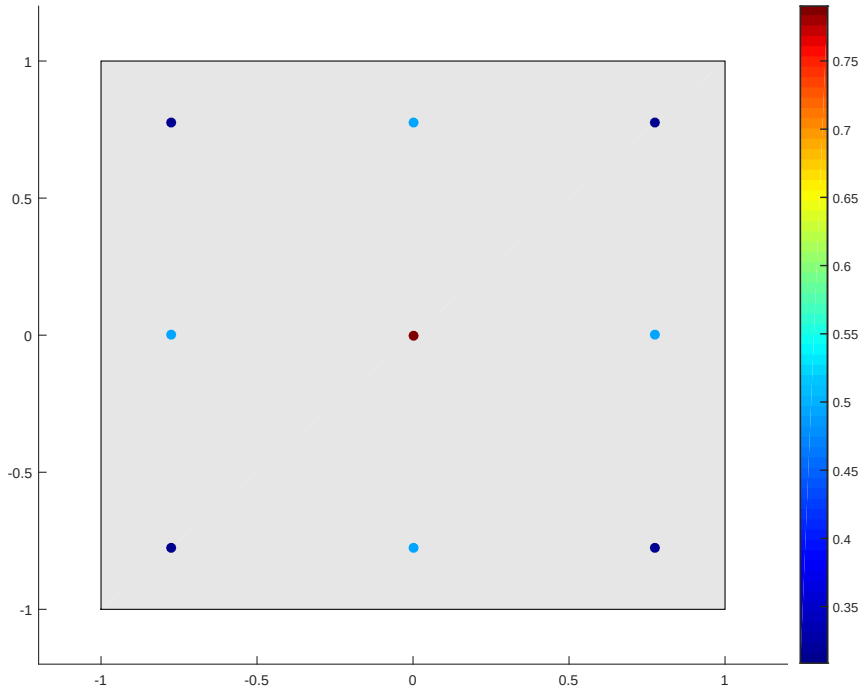

Figura 4.4: Rappresentazione della formula di cubatura su $\Omega = [-1, 1] \times [-1, 1]$ determinata dalla VII iterazione impiegando la base canonica e $r(x, y) = -x^6 - y^6$.

Ora vogliamo invece discutere della scelta della tolleranza, fondamentale per l'individuazione corretta dei nodi. Per facilitare la spiegazione, rappresentiamo la mappa $p_i(t) - r(t)$, dove $p_i(t)$ è il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra calcolato all'i-esima iterazione dell'Algoritmo 3.1-3.2. La Figura 4.5 mette

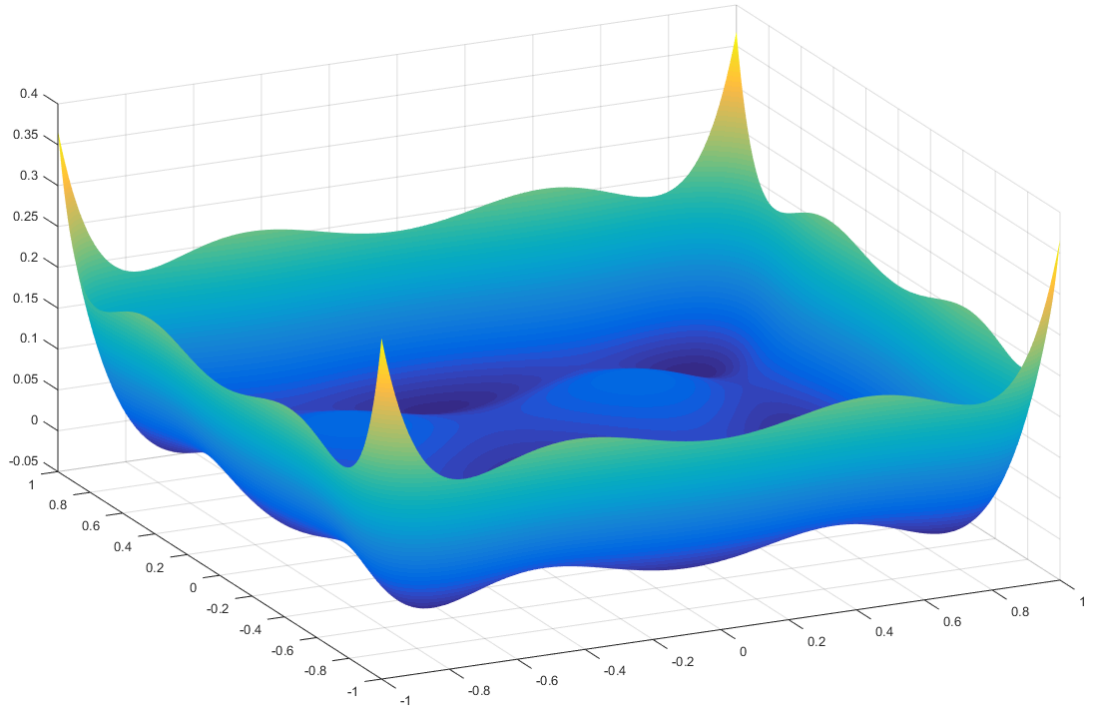

Figura 4.5: Rappresentazione $p_7(t) - r(t)$, dove $p_7(t)$ è il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra calcolato alla VII iterazione del metodo.

già in chiara evidenza i 9 punti che costituiscono la formula di cubatura, tuttavia il grafico ci permette anche di osservare che i punti stazionari possiedono valori molto vicini (< 0.05) a quelli che costituiscono i minimi. Per non introdurre nodi ulteriori, è necessario selezionare un margine di tolleranza empirico: non inferiore al valore di minimo ma tale da filtrare solo questi ultimi. Purtroppo questo aspetto rappresenta un evidente limite del metodo a cui non abbiamo trovato una diversa implementazione.

Visto il comportamento dei precedenti esempi, abbiamo voluto considerare una differente base dello spazio $\mathbb{P}^d_m$. Con le medesime ipotesi, abbiamo scelto la base generalizzata di Chebyshev, ottenendo i risultati descritti nella Tabella 4.10.

iterazione	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
I	9	-6.389602e-03	1.104744e-01	4.124923e-02	1.244 sec.
II	9	-7.113424e-05	3.266059e-02	7.142542e-03	15.689 sec.
III	9	-6.560300e-07	1.078402e-03	3.786601e-06	46.493 sec.
IV	9	-7.091156e-09	2.085935e-04	1.723887e-07	93.065 sec.

iterazione	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
V	9	-6.477793e-11	3.484919e-05	3.352780e-09	155.381 sec.
VI	9	-4.994387e-14	7.278363e-07	1.032206e-12	223.877 sec.
VII	9	3.403786e-13	3.556780e-07	4.615152e-13	337.695 sec.

Tabella 4.10: $m = 5$ e $\Omega = [-1, 1] \times [-1, 1]$: base di Cheb. e $r(x, y) = -x^6 - y^6$.

Non emergono differenze apprezzabili in termini di tempo di esecuzione o di quelli determinati dall'errore $\mathcal{E}_{\Lambda;\cdot}$ rispetto alla Tabella 4.9. Questo è dovuto al fatto che non viene impiegata la seconda fase, per la quale la scelta della base riveste un ruolo importante come abbiamo potuto osservare nella Sezione 4.1.

Analizzando la seconda funzione di sensibilità proposta dall'articolo [15],

$$r(x, y) = -\cos\left(\frac{\pi}{2}\sqrt{(x+1)^2 + (y+1)^2}\right), \tag{4.1}$$

abbiamo appurato risultati analoghi. Infatti, considerando la base canonica e una tolleranza pari a $\tau = 0.6001$, l'Algoritmo 3.1-3.2 ci ha fornito i risultati descritti nella Tabella 4.11. Come gli autori stessi mettono in evidenza, una scelta differen-

iterazione	nodi	$\mathcal{E}_F$	$\mathcal{E}_{\Lambda;\infty}$	$\mathcal{E}_{\Lambda;2}$	tempo
I	8	5.999524e-01	5.822229e-02	1.007446e-02	3.045 sec.
II	8	5.999990e-01	5.349710e-03	8.840854e-05	12.050 sec.
III	8	5.999998e-01	2.058322e-04	2.352434e-07	28.509 sec.
IV	8	5.999998e-01	1.068761e-04	3.495401e-08	48.330 sec.
V	8	5.999999e-01	1.525832e-05	1.013794e-09	83.237 sec.
VI	8	5.999998e-01	5.392475e-05	4.064090e-09	138.859 sec.
VII	8	5.999998e-01	3.622371e-05	2.090283e-09	205.358 sec.

Tabella 4.11: $m = 5$ e $\Omega = [-1, 1] \times [-1, 1]$: base di Chebyshev e r come (4.1).

te della funzione di sensibilità, determina una diversa formula di cubatura. Infatti non esiste alcuna teoria grazie alla quale sia possibile caratterizzare scelte preferibili di tale funzione, diversamente da quanto era emerso dagli studi di R. Bojanic e R. DeVore [1] per il caso univariato. Questa determinata scelta, differentemente dall'esempio precedente, permette di ottenere una formula di cubatura costituita da un numero inferiore di nodi, ovvero 8 anziché 9. La disposizione dei nodi è ben diversa (si veda la Figura 4.4), come differente è la mappa $p_i(t) - r(t)$, dove $p_i(t)$ è il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra calcolato all'i-esima iterazione (si veda Figura 4.7). Infine, anche in questo studio, emergono tempi di esecuzione ed errori non competitivi rispetto a quelli determinati dai metodi oggi diffusi.

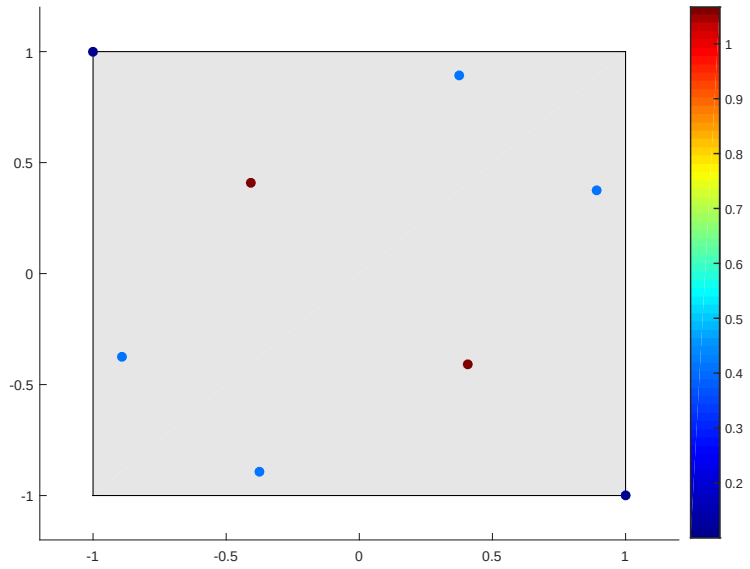

Figura 4.6: Rappresentazione della formula di cubatura su $\Omega = [-1, 1] \times [-1, 1]$ determinata dalla VII iterazione impiegando la base canonica e r come (4.1).

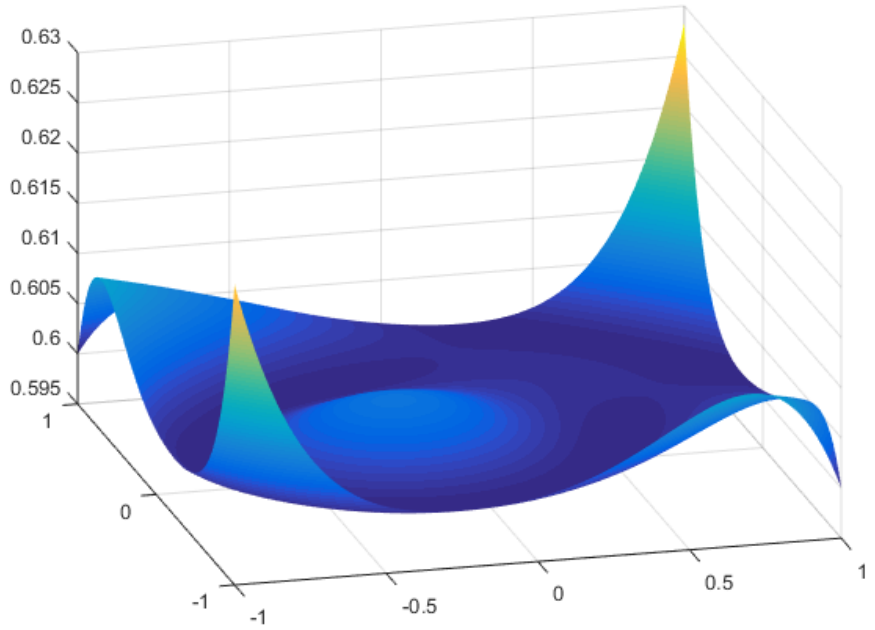

Figura 4.7: Rappresentazione $p_7(t) - r(t)$, dove $p_7(t)$ è il polinomio di miglior $\mathcal{L}^1_\mu$-approssimazione da sopra calcolato alla settima iterazione.

Appendice A

Nozioni di Analisi Convessa

Intendiamo qui esporre la teoria legata all'analisi convessa utilizzata nel corso di questa tesi, presentando definizioni e dimostrazioni ben note in letteratura.

Definizione A.1. *Un insieme $K \subseteq \mathbb{R}^n$ è detto* convesso *se dati due elementi k_1 e k_2 appartenenti a K, ogni loro combinazione convessa è un elemento di K, ovvero se $\forall k_1, k_2 \in K \; \forall \alpha \in [0,1]$ si ha*

$$\alpha k_1 + (1-\alpha) k_2 \in K$$

Per induzione è possibile dare una definizione equivalente che coinvolge un numero arbitrario (necessariamente finito) di elementi di K.

Corollario A.1. *Un insieme $K \subseteq \mathbb{R}^n$ è un insieme convesso se, e solo se, per un qualsiasi sottoinsieme finito di K, $\{k_1, \ldots, k_q\}$, vale*

$$\sum_{i=1}^{q} \gamma_i k_i \in K$$

al variare di $(\gamma_1, \ldots, \gamma_q)$ tale che $\sum_{i=1}^{q} \gamma_i = 1$ e $\gamma \geq 0$.

Conseguenza del precedente corollario è la seguente rappresentazione algebrica.

Definizione A.2. *Dato un insieme arbitrario $E \subseteq \mathbb{R}^n$, definiamo l'*inviluppo convesso *di E, Conv(E), il più piccolo (cardinalmente) insieme convesso contenente E. Equivalentemente tale insieme è rappresentato come segue:*

$$\textit{Conv(E)} := \left\{ \sum_{i=1}^{q} \gamma_i \, e_i : \sum_{i=1}^{q} \gamma_i = 1, \; \gamma \geq 0, \; \{e_1, \ldots, e_q\} \subset E \; e \; q \geq 1 \right\} \quad \text{(A.1)}$$

Proposizione A.1. *Dato un insieme compatto $E \subset \mathbb{R}^n$. Allora l'inviluppo convesso, Conv(E), è anch'esso compatto.*

Dimostrazione. Dimostrazione ripresa da [7, pag.71].

Per un lemma dovuto a Carathéodory si può asserire che l'inviluppo convesso dell'insieme E è generato attraverso tutte le possibili combinazioni convesse di al più $q = n + 1$. Dalla definizione data dalla relazione (A.1)e fissato $q = n + 1$ si può dedurre che, definito l'insieme

$$D := \left\{ \gamma \in \mathbb{R}^{n+1} : \gamma \geq 0 \text{ e } \sum_{j=1}^{n+1} \gamma_j \right\},$$

allora l'inviluppo convesso di E è l'immagine dell'insieme compatto $E^{n+1} \times D$ tramite la mappa continua:

$$(e_1, \ldots, e_{n+1}, \gamma) \longmapsto \sum_{i=1}^{n+1} \gamma_i \, e_i$$

Pertanto è anch'esso un compatto. □

Un'altra nozione insiemistica per noi fondamentale è la seguente.

Definizione A.3. *Un insieme $C \subseteq \mathbb{R}^n$ è detto* cono convesso *se è un insieme convesso tale che $\forall x \in C \; \forall \alpha \in [0, +\infty)$ si ha*

$$\alpha \, x \in C$$

*Dato un insieme convesso $K \subseteq \mathbb{R}^n$, denotiamo con il simbolo Cone(K) l'*inviluppo conico *di K, ossia il più piccolo (cardinalmente) cono convesso contenente K.*

*Allo stesso modo, dato un insieme arbitrario $E \subseteq \mathbb{R}^n$, denotiamo con CC(E) l'*inviluppo conico convesso *di E ovvero l'inviluppo conico dell'inviluppo convesso di E. In altre parole possiamo asserire che*

$$\begin{aligned} CC(E) :=& Cone(Conv(E)) \\ =& \left\{ \sum_{i=1}^{q} \gamma_i e_i : \gamma \geq 0, \; \{e_1, \ldots, e_q\} \subset E \text{ e dove } q \geq 1 \right\} \end{aligned}$$

Teoremi del supporto e di separazioni

Una nozione geometrica associata agli insiemi convessi è l'iperpiano di supporto. Nel seguito denotiamo un iperpiano dello spazio $\mathbb{R}^n$ come

$$\mathcal{H}_n(y,\eta) := \{x \in \mathbb{R}^n : \langle y, x\rangle = \eta\}$$

dove, di fatto, le componenti del vettore y descrivono i coefficienti della sua equazione e η il termine noto annesso. Si noti fin da ora che affinché si abbia una buona definizione le componenti non devono essere tutte nulle, ovvero $y \neq 0$. Definiamo inoltre i due semispazi delimitati da tale iperpiano come segue:

$$\mathcal{H}_n^-(y,\eta) := \{x \in \mathbb{R}^n : \langle y, x\rangle \leq \eta\}$$

$$\mathcal{H}_n^+(y,\eta) := \{x \in \mathbb{R}^n : \langle y, x\rangle \geq \eta\}$$

Ci proponiamo ora di dare una descrizione rigorosa all'idea intuitiva che, considerato un punto non appartenente ad un insieme convesso chiuso, esiste un iperpiano che li divide.

Definizione A.4. *Sia $K \subset \mathbb{R}^n$ un insieme convesso e non vuoto. Fissato un punto $z \in \mathbb{R}^n$ non appartenente a K, si dice che l'iperpiano $\mathcal{H}_n(y,\eta)$ separa z da K se $K \subset \mathcal{H}_n^-(y,\eta)$ e $z \in \mathcal{H}_n^+(y,\eta) \setminus \mathcal{H}_n(y,\eta)$. In altre parole se dati i parametri dell'iperpiano $\mathcal{H}_n(y,\eta)$, valgono le seguenti relazioni:*

$$\begin{cases} \sum_{i=1}^n y_i z_i > \eta \\ \sum_{i=1}^n y_i x_i \leq \eta, \ \forall x \in K \end{cases}$$

Teorema A.1 (Teorema iperpiano separazione per punto-insieme convesso)**.** *Sia $K \subseteq \mathbb{R}^n$ un insieme convesso, chiuso e non vuoto. Fissato un punto $z \in \mathbb{R}^n$ non appartente a K, denotiamo con $x_0 \in K$ un punto che realizza la minima distanza tra z e K. Allora l'iperpiano $\mathcal{H}(z - x_0, \langle z - x_0, x_0\rangle)$ separa z da K.*

Dimostrazione. Ci ispiriamo alla dimostrazione proposta in [7, pag. 78].

Definiamo x_0 il punto di minima distanza nell'insieme K da z,

$$|z - x_0| \leq |z - x|, \ \forall x \in K \tag{A.2}$$

Tale punto esiste per il teorema di Weierstrass applicato alla mappa $x \mapsto |x - z|$ sul dominio compatto $K \cap \overline{B(z,r)}$ per $r > 0$ sufficientemente grande. Attraverso altre considerazioni si potrebbe provare anche l'unicità di tale punto ma non è necessario affrontare questo aspetto nella nostra discussione.

Fissiamo un elemento x di K e consideriamo una sua combinazione convessa con $x_0 \in K$ di parametro $\xi \in (0, 1]$ applicato alla relazione (A.2),

$$\begin{aligned} |z - x_0|^2 \leq & \quad |z - (x_0 + \xi(x - x_0))|^2 \\ = & \quad |(z - x_0) - \xi(x - x_0)|^2 \\ = & \quad |z - x_0|^2 - 2\xi\langle z - x_0, x - x_0\rangle + \xi^2|x - x_0|^2 \end{aligned}$$

Osservando che il primo addendo coincide con il primo termine dell'ultima relazione, posso semplificarli ottenendo

$$\langle z - x_0, x\rangle \leq \langle z - x_0, x_0\rangle + \frac{\xi}{2}|x - x_0|^2$$

Per l'arbitrarietà di $\xi \in (0, 1]$ è ragionevole porre il limite $\xi \to 0$ trovando che $\forall x \in K$ vale

$$\langle z - x_0, x\rangle \leq \langle z - x_0, x_0\rangle$$

Ne consegue la seguente relazione:

$$K \subseteq \{x \in \mathbb{R}^n : \langle z - x_0, x\rangle \leq \langle z - x_0, x_0\rangle\}$$

Allora l'iperpiano $\mathcal{H}^n(z - x_0, \langle z - x_0, x_0\rangle)$ è iperpiano di supporto di K ed inoltre, per le proprietà fondamentali del prodotto scalare euclideo, possiamo dedurre che $x_0 \in \mathcal{H}^n(z - x_0, \langle z - x_0, x_0\rangle)$.

Ci rimane da provare che z appartenga al semispazio

$$\mathcal{H}_n^+(z - x_0, \langle z - x_0, x_0\rangle) \setminus \mathcal{H}_n(z - x_0, \langle z - x_0, x_0\rangle)$$

In altre parole dobbiamo dimostrare che valga la relazione

$$\langle z - x_0, z\rangle > \langle z - x_0, x_0\rangle$$

Poiché z non appartiene all'insieme chiuso K possiamo asserire che $|x_0 - z| > 0$. Pertanto possiamo dedurre che

$$0 < |z - x_0|^2 = \langle z - x_0, z - x_0\rangle = \langle z - x_0, z\rangle - \langle z - x_0, x_0\rangle$$

□

Corollario A.2 (Corollario iperpiano separazione per punto-cono convesso)**.** *Sia $C \subseteq \mathbb{R}^n$ un cono convesso, chiuso e non vuoto. Se $z \in \mathbb{R}^n$ è un punto che non appartiene a C allora un iperpiano che separa z da C è del tipo $\mathcal{H}_n(y, 0)$, ovvero esistono dei coefficienti reali $y_1, \ldots, y_n$ tali che*

$$\begin{cases} \sum_{i=1}^n y_i z_i > 0 \\ \sum_{i=1}^n y_i x_i \leq 0, \ \forall x \in K \end{cases}$$

Dimostrazione. Dal Teorema A.1 si deduce che esistono dei coefficienti reali $y_1, \ldots, y_n$ tali che

$$\begin{cases} \sum_{i=1}^{n} y_i z_i > \eta \\ \sum_{i=1}^{n} y_i x_i \leq \eta, \ \forall x \in K \end{cases} \tag{A.3}$$

dove $\eta = \sum_{i=1}^{n} y_i x_{0i}$ ed x_0 un punto di C. Pertanto è sufficiente provare che $\eta = 0$. Per la Definizione A.3, sappiamo che essendo C un cono convesso e x_0 un suo elemento, allora $\forall \alpha \geq 0$ si ha $\alpha x_0 \in C$. In particolare applicando la seconda disequazione del sistema (A.3) si ottiene che $\forall \alpha \geq 0$ si deve avere

$$0 \geq \sum_{i=1}^{n} y_i \alpha x_{0i} - \eta = \sum_{i=1}^{n} y_i \alpha x_0 - \sum_{i=1}^{n} y_i x_{0i} = (\alpha - 1) \left(\sum_{i=1}^{n} y_i x_{0i} \right)$$

Per l'arbitrarietà di γ si deve avere necessariamente che il secondo fattore sia nullo, ovvero che η sia nullo. □

Come abbiamo visto dal procedente teorema, la condizione sull'insieme K di essere chiuso è necessaria per la bontà della costruzione dimostrativa. Con l'intenzione di porre condizioni meno stringenti diamo la seguente definizione.

Definizione A.5. *Sia $K \subseteq \mathbb{R}^n$ un insieme convesso e non vuoto.*
Fissato un punto $s \in K$ appartenente all'insieme K, si dice che l'iperpiano $\mathcal{H}^n(y, \eta)$ è un iperpiano di supporto *di K in s se $s \in \mathcal{H}^n(y, \eta)$ e $K \subseteq \mathcal{H}_n^-(y, \eta)$.*
In altre parole se dati i parametri dell'iperpiano $\mathcal{H}_n(y, \eta)$, valgono le seguenti relazioni:

$$\begin{cases} \langle y, s \rangle = \eta \\ \langle y, x \rangle \leq \eta, \ \forall x \in K \end{cases}$$

Teorema A.2 (Teorema di esistenza del piano di supporto per i convessi)**.** *Sia $K \subseteq \mathbb{R}^n$ un insieme convesso e non vuoto. Se $s \in \mathbb{R}^n$ è un punto di frontiera di K allora esiste almeno un iperpiano di supporto di K in s.*

Dimostrazione. Dimostrazione ripresa da [7, pag. 84].

Poiché per gli insiemi convessi vale int $M =$ int $\overline{M}$, allora risulta la seguente uguaglianza insiemistica fra le frontiere di M e la sua chiusura:

$$\partial M = \overline{M} - \text{int } M = \partial \overline{M}$$

Pertanto, tenuto conto che $s \in \overline{M}$, possiamo dire che esiste una successione di punti $\{s_i\}_{i \in \mathbb{N}}$ non appartenenti all'insieme $\overline{M}$ convergenti però all'elemento s.

Applicando allora per ogni punto della successione il Teorema A.1 sull'insieme convesso e chiuso $\overline{M}$, si ottiene che denotando con $\overline{z_i}$ il punto di minima distanza nell'insieme $\overline{M}$ da s_i, si ottiene che per ogni $i \in \mathbb{N}$:

$$\begin{cases} \langle s_i - \overline{z_i}, s_i \rangle > \langle s_i - \overline{z_i}, \overline{z_i} \rangle \\ \langle s_i - \overline{z_i}, x \rangle \leq \langle s_i - \overline{z_i}, \overline{z_i} \rangle, \ \forall x \in K \end{cases} \tag{A.4}$$

Poiché $\overline{z_i}$ appartiene ad $\overline{M}$ mentre s_i no, allora $\overline{z_i} - s_i \neq 0$.
Allora possiamo definire la seguente successione di versori

$$y_i := \frac{\overline{z_i} - s_i}{|\overline{z_i} - s_i|}$$

per la quale esiste una sotto-successione convergente ad un certo versore y (segue dal fatto che l'insieme degli elementi $|y| = 1$ è un compatto). Dividendo ciascuna disequazione del sistema (A.4) per la norma di $\overline{z_i} - s_i$ si ottiene per ciascun $i \in \mathbb{N}$ la seguente relazione:

$$\langle y_i, s_i \rangle > \langle y_i, \overline{z_i} \rangle, \ \forall x \in K$$

Passando al limite $i \to +\infty$ si ottiene

$$\langle y, s \rangle \geq \langle y, x \rangle, \ \forall x \in K$$

Appare evidente che l'iperpiano $\mathcal{H}_n(y, \langle y, s \rangle)$, contenendo tautologicamente l'elemento s, è un iperpiano di supporto di K in s. □

Corollario A.3 (Corollario esistenza iperpiano supporto per cono convesso)**.** *Sia $C \subseteq \mathbb{R}^n$ un cono convesso e non vuoto. Se $s \in C$ è un punto di frontiera di C allora un iperpiano di supporto di C in s è del tipo $\mathcal{H}_n(y, 0)$. Ovvero esistono dei coefficienti reali $y_1, \ldots, y_n$ tali che*

$$\begin{cases} \sum_{i=1}^n y_i s_i = 0 \\ \sum_{i=1}^n y_i x_i \leq 0, \ \forall x \in K \end{cases}$$

Dimostrazione. Si procede in modo del tutto analogo al precedente corollario. Dal Teorema A.2 si deduce che esistono dei coefficienti reali $y_1, \ldots, y_n$ e termine noto η tali che

$$\begin{cases} \sum_{i=1}^n y_i s_i = \eta \\ \sum_{i=1}^n y_i x_i \leq \eta, \ \forall x \in K \end{cases} \tag{A.5}$$

Rimane perciò da dimostrare che η sia nullo. Per la Definizione A.3 sappiamo che, essendo C un cono convesso, considerato un suo elemento $s \in C$, allora per ogni scalare $\alpha \geq 0$ si ha $\alpha s \in C$.

In particolare applicando la disequazione del sistema (A.5) si ottiene che $\forall \alpha \geq 0$ si deve avere

$$0 \geq \sum_{i=1}^{n} y_i \alpha s_i - \eta = \sum_{i=1}^{n} y_i \alpha s_i - \sum_{i=1}^{n} y_i s_i = (\alpha - 1) \left(\sum_{i=1}^{n} y_i s_i \right)$$

Per l'arbitrarietà di α si deve avere necessariamente che il secondo fattore sia nullo, ovvero che η sia nullo. □

Dimostriamo infine una proposizione utile alla nostra discussione.

Proposizione A.2. *Sia $K \subseteq \mathbb{R}^n$ un insieme convesso e non vuoto che ammette interno, int K. Se s è un punto interno di K, allora non esistono iperpiano di supporto di K in s.*

Dimostrazione. Dimostraione ripresa da [7, pag.83].

Supponiamo per assurdo che esista un iperpiano di supporto di K in s, ovvero $\mathcal{H}_n(y, \eta)$. Poiché s è un punto interno, allora esiste $\alpha > 0$ tale che $s_\alpha := s + \alpha y$ sia ancora un elemento di K. Si nota allora che per la linearità del prodotto scalare

$$\langle y, s_\alpha \rangle = \langle y, s \rangle + \alpha \langle y, y \rangle$$

ed essendo $s_\alpha \in K$ allora per la definizione di $\mathcal{H}_n(y, \eta)$ si ottiene:

$$\langle y, s \rangle + \alpha \sqrt{|y|} \leq \eta = \langle y, s \rangle$$

Pertanto si ottiene che $\alpha \sqrt{|y|} \leq 0$. Tuttavia, tenendo conto che α è strettamente positivo, si realizza solo quando il vettore y è nullo. Questo ovviamente vìola la bontà della definizione di $\mathcal{H}_n(y, \eta)$ portando alla violazione cercata. □

Appendice B

Programmi

In questo appendice desideriamo riportare solo le implementazioni delle due fasi discusse nel Capitolo 3. I sorgenti delle classi e degli esempi sono resi disponibili attraverso il supporto multimediale allegato.

La prima fase è implementata nel file **discretization methods.m**.

```
function [x, time] = discretization_methods(LSIP, type_grid, ...
        s_soglia, max_iterazioni,debug)

%PARAMETRI-------------------------------------------------------------
h = LSIP.h_start;  % stabilisce la griglia iniziale
dh = 0.1;          % stabilisce il fattore molt.
                   %  di h ad ogni iterazione
flag = 1;
max_point_linear_system = 500000;
if nargin < 5, max_iterazioni=10; end
if nargin < 4, debug = 0; end
if nargin < 3, s_soglia = -10^-10; end

options = optimoptions('linprog');
options.MaxIter = 10^5;
options.Display = 'off'; %disabilita le stampe di routine
ind_iter = 1;
%----------------------------------------------------------------------

if debug > 0
    ...
end

tic

try
%-----determiniamo la griglia iniziale
T0 = LSIP.equigrid(h);
%---risoluzione del sottoproblema LP
Au = -LSIP.val_base(T0)';
```

```
bu = -LSIP.val_r(T0)';

[x,fval,exitflag] = linprog((LSIP.c)',Au,bu,[],[],[],[],[],options);

if exitflag ~= 1
    print_error_linprog(exitflag);
else
    [min_pts,s] = LSIP.min_points_g_map(x);
    time = toc;
    flag = s < s_soglia;
    if debug > 0
        ...
    end
    ind_iter = ind_iter +1;
end
Tr = T0;
while flag & ind_iter<=max_iterazioni
    h = h*dh;
    switch type_grid
        case 'pure_cutting_plane'
            Tr = LSIP.pure_cutting_plane_grid(h,Tr,x,s,ind_iter);
        case 'cutting_plane'
            Tr = LSIP.cutting_plane_grid(h,Tr,x,s,min_pts,ind_iter);
        otherwise
            Tr = LSIP.equigrid(h);
    end
    %scatter(Tr(1,:)',Tr(2,:)'); pause;
    if size(Tr,2) > max_point_linear_system
        error('Il problema PL ha troppi vincoli %d',size(Tr,2));
    end

    %---risoluzione del sottoproblema LP
    Au = -LSIP.val_base(Tr)';
    bu = -LSIP.val_r(Tr)';

    [x,fval,exitflag] = linprog((LSIP.c)',Au, bu,...
                                        [],[],[],[],[],options);

    if exitflag ~= 1
        print_error_linprog(exitflag);
    else
        [min_pts,s] = LSIP.min_points_g_map(x);
        time_pre = time;
        time = toc;
        flag = s < s_soglia;
        if debug > 0
            ...
        end
        ind_iter = ind_iter +1;
    end
end
```

```
catch exception
    fprintf('ERRORE: %s',getReport(exception));
end
time = toc;
end
```

La seconda fase è implementata nel file `local reduction method.m`.

```
function [ott_t,ott_w,ott_x_ass,ott_err_D,ott_err_P,...
    ott_ind_iter, time] = local_reduction_method(LSIP,...
    x,max_iterazioni,prec,debug, time_fst_ps,...
    option_robustness)

%PARAMETRI-------------------------------------------------------------
if nargin < 7, option_robustness = 1; end
if nargin < 6, time_fst_ps = 0; end
if nargin < 5, debug = 0; end
if nargin < 4, prec = 10^-15; end
if nargin < 3, max_iterazioni = 100; end
%----------------------------------------------------------------------

ott_t = zeros(1,1); ott_w = zeros(1,1); ott_err_D = 10;
ott_x_ass = zeros(1,1); ott_err_P = 10;

flg_recalc_solution = 1;
ind_iter = 1; time = 0;

if debug > 0
    fprintf('LOCAL REDUCTION_METHOD_2\n');
    fprintf('soglia ammissibilità problema D: %e \n\n', prec);
end

tic
while flg_recalc_solution && ind_iter <= max_iterazioni

    %STEP 1
    [pts,s,L] = LSIP.min_points_g_map(x);
    N = size(pts,2);
    int_pts = pts(:,1:L);   %punti interni

    %STEP 2
    V = LSIP.val_base(pts);
    [w,err_2] = lsqnonneg(V,LSIP.c);
    err = norm(V*w-LSIP.c,inf);

    %CANCELLO I PUNTI CON W=0
    L_tp = 0; index = false(size(w));
    for i=1:length(w)
        if w(i) > 10^-9
            if i <= L, L_tp = L_tp + 1; end
            index(i) = 1;
        end
    end
```

```
46      L = L_tp;
47      pts = pts(:,index);
48      w = w(index);
49      N = size(pts,2);
50      int_pts = pts(:,1:L);    %punti interni
51      V = V(:,index);
52
53      if err < ott_err_D
54          ott_t = pts;
55          ott_w = w;
56          ott_x_ass = x;
57          ott_err_D = err;
58          ott_err_P = s;
59          ott_ind_iter = ind_iter;
60      end
61
62      if debug > 0
63         ...
64      end
65
66      %STEP 3
67      if err < prec
68          flg_recalc_solution = 0;
69      else
70          %STEP 4
71          [Q,R] = qr(V);
72          if N > LSIP.n
73              error('Il metodo di calcolo dei punti'...
74                      ` di minimo non è soddisfacente');
75          end
76          %Q è n*n e R è triangolare n*N
77          Y = Q(:,1:N);
78          Z = Q(:,(N+1):end);
79          R=R(1:N,:);
80
81          U = LSIP.calc_U(int_pts);
82          W = zeros(L*LSIP.d);
83          k=1;
84          for i=1:L
85          for j=1:LSIP.d
86              W(k,k) = w(i);
87              k = k+1;
88          end
89          end
90          %W = diag(w(1:L));
91          D2 = LSIP.calc_D2(x,int_pts);
92          g = - (LSIP.val_g_map(x,pts))';
93
94          %STEP 5
95          if det(D2)==0
96              fprintf('LA MATRICE D2 NON E'' INVERTIBILE \n');
```

```
97              flg_recalc_solution = 0;
98          else
99              D2_inv = inv(D2);
100             H = U*W*D2_inv*(U');
101
102             %STEP 6
103             d1 = R'\g;
104
105             %STEP 7
106             M1 = Z'*H*Z;
107             M2 = - (Z'*(LSIP.c+H*Y*d1));
108             d2 = M1\M2;
109             dx = Y*d1 + Z*d2;
110
111             if any(eig(Z'*H*Z)<10^-14)
112                 %in caso applicare il programma di Watson
113                 if option_robustness == 1
114                     dx = robustness_adjustment(LSIP,Y,Z,H,M2,d1,...
115                                                x,g(1:L),V,dx,max(w),debug);
116                 else
117                     fprintf('WARNING'...);
118                     pause;
119                 end
120             end
121
122             %STEP 8
123             x = x+dx;
124
125             ind_iter = ind_iter + 1;
126         end
127     end
128 end
129
130 time = toc+time_fst_ps;
131
132 end
133
134 function dx = robustness_adjustment(LSIP,Y,Z,H,M2,d1,x,...
135                                                g_int,V,dx,max_w,debug)
136      flg_not_defpos = 1;
137      mu = 0; mu_min = 1;
138      max_iter = 20; ro = 0.0001;
139      while flg_not_defpos & max_iter > 0
140         max_iter = max_iter - 1;
141         % -> Determino le parti che non dipendono da gamma
142         %g è la valutazione della mappa r(t) -  <\nu(t),x>
143         T_p1 = (LSIP.c)'*x;
144         krn = find(g_int > 0); %dimensione N
145         T_p2 = sum(g_int(krn));
146         T_p3 = (sum(V(:,krn),2))'*dx;
147
```

```
148         flg_dont_find_descent_direction = 1;
149         sigma = max(max_w + 0.00001,1); ind_iter_sigma = 50;
150         while flg_dont_find_descent_direction && ind_iter_sigma > 0
151             ind_iter_sigma = ind_iter_sigma - 1;
152             gamma = 1; int_iter_gamma = 50;
153             T = ro - 1;
154             while T < ro && int_iter_gamma > 0
155                 int_iter_gamma = int_iter_gamma - 1;
156                 gamma = gamma / 2;
157
158                 %Calcolo T_p4
159                 x_p4 = x + gamma*dx;
160                 [t,s,L,vals] = LSIP.min_points_g_map(x_p4,0);
161                 krn_p4 = vals < 0;
162                 T_p4 = sum(-vals(krn_p4));
163
164                 T = (gamma*T_p1 + sigma*(T_p4 - T_p2)) /...
165                     (gamma*(T_p1 + sigma* T_p3));
166             end
167             if int_iter_gamma==0
168                 sigma = sigma + 30; %Lo aumento temporaneamente
169             else
170                 flg_dont_find_descent_direction = 0;
171             end
172         end
173         %Ricalcolo di mu e mu_min
174         if mu == mu_min
175             if T>0.5 && gamma == 1
176                 mu_min = mu_min/4;
177             else
178                 if gamma<=0.25
179                     mu_min = mu_min*4;
180                 end
181             end
182         end
183         mu = 4*mu + mu_min;
184         M1 = Z'*(H + mu*eye(size(H)))*Z;
185         d2 = M1\M2;
186         dx = Y*d1 + Z*d2;
187         flg_not_defpos = any(eig(M1)<10^-14);
188         if debug > 0
189             ...
190         end
191     end
192     dx = gamma*dx;
193 end
```

Indice analitico

Bibliografia

[1] R. Bojanic, R. DeVore, "On polynomials of best one sided approximation", *L'Enseignement Mathématique*, Vol.12 (1966), pag 139-164.

[2] I.D. Coope, G.A. Watson, "A projected Lagrangian Algorithm for Semi-Infinite Programming", *Mathematical Programming*, Springer Berlin Heidelberg, Vol.32 (1985), pag 337-356.

[3] P.J. Davis, *Interpolation and Approximation*, Dover Publications, Waltham, 1975.

[4] Mattia Festa, Alvise Sommariva, "Computing almost minimal formulas on the square", *Journal of Computational and Applied Mathematics*, Vol.236 (2012), pag 4296-4302.

[5] A. Glaser, X. Liu, V. Rokhlin, "A fast algorithm for the calculation of the roots of special functions", *SIAM Journal on Scientific Computing*, Elsevier, Vol.29 (2007), pag 1420-1438.

[6] "Duality theory of semi-infinite programming", *Lecture Notes in Control and Information Sciences*, Springer, 1978.

[7] K. Glashoff, S. Gustafson, *Linear Optimization and Approximation: an introduction to the Theoretical Analysis and Numerical Treatment of Semi-infinite Programs*, Springer, New York, 1978.

[8] Miguel A. Goberna, Marco A. López, *Linear Semi-Infinite Optimization*, John Wiley, 1998.

[9] Miguel A. Goberna, Marco A. López, *Post-Optimal Analysis in Linear Semi-Infinite Optimization*, Springer, 2014.

[10] Miguel A. Goberna, Marco A. López, "On Duality in Semi-Infinite Programing and Existence Theorems for Linear Inequalities", *Journal of Mathematical Analysis and Application*, Elsevier, Vol.230 (1999), pag 173-192.

[11] G.H. Golub, J.H. Welsch, "Calculation of Gauss quadrature rules", *Mathematics of Computation*, JSTOR, Vol.23 (1969), pag 221-230.

[12] N. Hale, L.N. Trefethen, *Chebfun and Numerical quadrature*, Science in China, Vol.55 (2012), pag 1749-1760.

[13] S.J. Karlin, W.J. Studden, *Tchebycheff system: with applications in analysis and statistics*, John Wiley, New York, 1966.

[14] A. Quarteroni, R. Sacco, F. Saleri, P. Gervasio, *Matematica Numerica*, Springer, Milano, 2014 (4a Edizione).

[15] E. K. Ryu, S. P. Boyd, "Extensions of Gauss Quadrature Via Linear Programming", *Foundations of Computational Mathematics*, Springer US, Vol.15 (2015), pag 953-971.

[16] S.L. Sobolev, V.L. Vaskevich, *The Theory of Cubature Formulas*, Springer, Novosibirsk, 1997.

[17] A.H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall, 1971.

[18] G.A. Watson, "Lagrangian Methods for Semi-Infinite Programming Problems", *Infinite Dimensional Linear Programming, Lecture Notes in Economics and Mathematical System*, Springer-Verlag, Vol.259 (1985), pag 90-107.