POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Problemi di controllo ottimo stocastico modellati da PDE soggetti a vincoli sul Conditional Value-at-Risk

Relatori

Sandra Pieraccini
Tommaso Vanzan

Candidato

Martino Cavo

Anno Accademico 2024-2025

Ai miei genitori
Ai miei nonni

Sommario

In questo elaborato si affrontano problemi di controllo ottimo vincolati da equazioni alle derivate parziali (PDE) in presenza di incertezza modellistica. Nello specifico, viene analizzato l'impatto dell'incertezza su alcuni parametri del problema e vengono proposte strategie per il controllo del rischio associato, sia in termini di formulazione matematica sia dal punto di vista numerico. Dopo una preliminare analisi del caso deterministico, utile come base teorica, viene analizzato il problema stocastico, studiandone il comportamento in presenza di variabili di controllo, definite sull'intero dominio oppure sul contorno, minimizzando un funzionale obiettivo ridotto di tipo risk-neutral. Nella seconda parte, si approfondisce l'impiego del Conditional Value-at-Risk (CVaR) come strumento per modellare l'avversione al rischio. L'obiettivo di questa tesi è rappresentato dal confronto di due approcci computazionali, basati su strategie di tipo interior-point, per la risoluzione di problemi di controllo ottimo che includono, come vincolo, la misura di rischio CVaR valutata rispetto ad un funzionale di interesse. Nel primo approccio viene considerata una riformulazione di tipo epigrafico del vincolo mentre nel secondo viene implementata una strategia di smoothing della misura di rischio, unitamente al metodo della funzione implicita per eliminare la variabile aggiuntiva derivante dalla riformulazione del CVaR proposta da Rockafellar e Uryasev. Sono presentati, infine, alcuni risultati per evidenziare le caratteristiche di questi metodi, considerando sia un funzionale obiettivo di tipo risk-neutral che uno di tipo risk-averse, utilizzando sempre il CVaR come misura di rischio. Le metodologie sviluppate costituiscono una base solida per futuri approfondimenti su problemi più complessi o ad alta dimensionalità.

Ringraziamenti

Desidero rivolgere un sincero ringraziamento alla Professoressa Pieraccini, per aver supervisionato il mio lavoro con costante attenzione, guidandomi nella scelta dell'argomento e offrendomi interessanti spunti e orientamenti.
Vorrei esprimere un sentito ringraziamento al Professor Vanzan, per i preziosi consigli ricevuti, per le conoscenze condivise e per la sua costante disponibilità e tempestività, che hanno contribuito in maniera significativa alla realizzazione di questo elaborato.

Con profonda gratitudine ringrazio i miei genitori e i miei familiari, per il loro sostegno incondizionato, la fiducia che hanno sempre riposto in me e l'amore con cui mi hanno accompagnato in ogni fase del mio percorso. A loro devo la forza, la determinazione e la serenità che mi hanno permesso di affrontare le sfide di questi anni. Il mio ringraziamento va anche a tutte le persone che mi sono state vicine, incoraggiandomi nelle scelte intraprese, condividendo con me gioie e difficoltà e contribuendo, ciascuno a suo modo, al raggiungimento di questo traguardo.

Indice

Figure

Elenco delle tabelle

Algoritmi

Capitolo 1

Introduzione

Il capitolo introduttivo di questo elaborato persegue l'obiettivo di fornire una visione d'insieme sul significato e sulle applicazioni dell'analisi di problemi di controllo ottimo vincolati da equazioni differenziali alle derivate parziali, in condizioni di incertezza. Inoltre, sono presentati alcuni concetti matematici rilevanti nella descrizione di tali problemi, al fine di introdurre gli strumenti necessari alla loro comprensione.

1.1 Motivazione generale e contesto applicativo

Le equazioni differenziali alle derivate parziali rappresentano uno strumento matematico fondamentale nella descrizione modellistica di un vasto insieme di fenomeni fisici, biologici, meccanici e socio-economici. In questo framework, risulta di particolare rilevanza applicativa considerare problemi di ottimizzazione, in cui si richiede la minimizzazione di una determinata quantità di interesse, che dipende dalla soluzione dell'equazione differenziale che regola il sistema in esame.
Una formulazione molto comune di questo tipo di problemi consiste nel forzare le soluzioni della PDE ad avvicinarsi il più possibile ad una soluzione di riferimento. Per perseguire questo scopo, il sistema viene influenzato tramite una variabile aggiuntiva, detta *variabile di controllo*, modificando la quale è possibile "spingere" la soluzione della PDE verso la soluzione target. Risulta pertanto sfidante la ricerca di una discrepanza accettabile tra la *variabile di stato* (cioè la soluzione della PDE) e il riferimento, senza tuttavia dover richiedere una configurazione eccessivamente costosa per la variabile di controllo.
Una difficoltà aggiuntiva, tipica delle applicazioni empiriche, consiste nell'introduzione di incertezza su alcuni parametri del problema, quali ad esempio coefficienti di diffusione e/o termini forzanti. Nella presente trattazione, saranno analizzati problemi di controllo ottimo *stocastici*, nei quali il funzionale da minimizzare è rappresentato da una data *misura di rischio* legata alla stocasticità del sistema. Rispetto al caso puramente deterministico, questa generalizzazione del problema introduce ulteriori difficoltà dal punto di vista computazionale, in quanto viene notevolmente aumentata la dimensionalità del sistema e di conseguenza le tempistiche di calcolo dei risultati. Anche in questo caso, quindi, è necessario trovare un compromesso tra accuratezza numerica, che aumenta con il numero

di campioni considerati, ed efficienza computazionale.
I problemi di controllo ottimo stocastici saranno analizzati seguendo due approcci con differenti applicazioni pratiche, basati su due diversi funzionali di interesse, detti misure di rischio. In particolare, in alcuni ambiti può essere interessante ottimizzare il valore atteso della distanza tra la soluzione della PDE e il target, applicando un termine di regolarizzazione del controllo. Tuttavia, questo approccio, detto *risk-neutral*, può comportare notevoli rischi qualora il verificarsi di eventi estremi, rappresentati da realizzazioni appartenenti alle code di una distribuzione di probabilità, influenzi eccessivamente in negativo il fenomeno in esame. In tal caso, risulta più opportuno ottimizzare un funzionale di tipo *risk-averse*, ad esempio basato su misure di rischio quali il *Conditional Value-at-Risk (CVaR)*, che rappresenta il valore medio delle perdite oltre una certa soglia di confidenza, con l'obiettivo di limitare gli effetti distruttivi di questi eventi estremi. Il vantaggio modellistico consiste nella possibilità di evitare configurazioni del controllo inaffidabili all'atto pratico, garantendo stabilità del sistema anche in presenza di deviazioni significative dalla media dei dati.
I problemi di controllo stocastici descritti sono comuni in vari contesti applicativi, quali ad esempio l'ingegneria civile e meccanica, in cui si ha la necessità di progettare strutture e sistemi vibranti capaci di mantenere elevati standard di sicurezza in condizioni avverse o in presenza di carichi, oppure in ambito gestionale, per organizzare ad esempio la produzione energetica in base all'equilibrio domanda-offerta o per la pianificazione della manutenzione di macchinari. Un classico esempio applicativo consiste nella progettazione di un processo termico, in cui la PDE da risolvere è l'equazione del calore stazionaria, in cui il controllo rappresenta una sorgente di energia che influenza la variabile di stato, cioè la temperatura del sistema, con l'obiettivo di avvicinarla ad un valore target. Il CVaR, inoltre, ha rilevante importanza nell'ambito economico-finanziario, come nella gestione di un investimento in un portafoglio di titoli azionari e obbligazionari. In particolare, un approccio risk-neutral consentirebbe di massimizzare il rendimento del portafoglio, mentre una strategia risk-averse tenderebbe a sacrificare la performance, per garantire una protezione in caso di eventi estremi o scenari catastrofici, fornendo però una soluzione molto più conservativa. Una strategia bilanciata rispetto a questo tradeoff è rappresentata dal principale problema in esame nel Capitolo 3, in cui viene mantenuto un funzionale obiettivo di tipo risk-neutral, per massimizzare il rendimento, rispettando tuttavia un vincolo sul CVaR, per mantenere il rischio di eventi estremi sotto una soglia fissata.

Il tema del controllo ottimo in condizioni di incertezza è un ambito di ricerca molto popolare nella letteratura scientifica.
In maniera più generale, l'oggetto di studio di questa tesi rientra nell'ambito dello *Stochastic Programming*, trattato in maniera completa da Shapiro et al. [27].
L'ottimizzazione vincolata da equazioni differenziali alle derivate parziali è stata sviluppata già dagli anni '70 grazie al lavoro di J.L. Lions [12], autore di un testo pionieristico riguardo questo argomento. Sviluppi moderni del tema sono stati proposti da Hinze et al. [9] e Tröltzsch [30].
Un riferimento obbligato in termini di completezza argomentativa è rappresentato dal libro di Manzoni et al. [14], che fornisce una base teorica fondamentale per l'analisi dei problemi di controllo ottimo, combinando risultati analitici e applicazioni numeriche.
I fondamenti matematici e teorici dell'ottimizzazione in condizioni di incertezza sono

accuratamente introdotti nel testo di J. M. Frutos e F. P. Esparza [19], dove inoltre vengono presentati metodi numerici basati su discretizzazioni Monte Carlo, Stochastic Collocation o Stochastic Galerkin, per affrontare problemi sia di tipo risk-neutral che risk-averse. In aggiunta, viene analizzata un'interessante applicazione alla meccanica strutturale, volta all'obiettivo di ottimizzare la distribuzione di un materiale, minimizzando un certo funzionale di interesse, tenendo presenti gli effetti dell'incertezza, dovuta a difetti strutturali o variazioni nelle proprietà del materiale.
Riguardo la trattazione computazionale del problema, vengono considerati i risultati ottenuti da Martin et al. [18], che propongono un confronto tra la risoluzione del sistema completo generato dal problema con Steepest Descent e con un metodo del gradiente stocastico. Un altro contributo essenziale all'analisi del problema di controllo stocastico è rappresentato dal lavoro di Nobile e Vanzan [21], in cui viene, inoltre, proposto un approccio avanzato di discretizzazione in probabilità basato su formule di quadratura *multilevel Monte Carlo*, con l'obiettivo di ridurre il costo computazionale. Altri importanti lavori riguardanti problemi di controllo ottimo stocastico sono di Tiesler et al. [29], Babuška et al. [1], Mateos [20] e Gunzburger et al. [8], quest'ultimo focalizzato su problemi con controllo definiti sul bordo di Neumann del dominio.
Nella seconda parte dell'elaborato vengono analizzati problemi di controllo vincolati dalla CVaR, con l'obiettivo di imporre un limite alla media dei valori della coda di una distribuzione di interesse. La descrizione matematica e lo studio delle proprietà del CVaR, con applicazioni numeriche relative all'ottimizzazione in ambito finanziario, si deve a R. T. Rockafellar e S. P. Uryasev [25, 24], che proposero una riformulazione di questa misura di rischio, rendendola computazionalmente trattabile in problemi di ottimizzazione convessa. Un contributo essenziale all'analisi di problemi di controllo risk-averse si deve a D. P. Kouri e T. M. Surowiec [11], nel cui lavoro la riformulazione del CVaR di Rockafellar e Uryasev viene applicata all'analisi di problemi vincolati da equazioni differenziali alle derivate parziali, applicando tecniche di regolarizzazione per approssimare il CVaR e costruire algoritmi numerici. L'obiettivo finale di analizzare problemi di controllo vincolati da CVaR può essere raggiunto adattando strategie computazionali già implementate per la risoluzione di problemi di controllo risk-averse *unconstrained*. In particolare, una panoramica d'insieme di questi metodi è fornita dall'articolo di J. O. Royset [26], in cui viene presentato l'approccio epigrafico, che consiste nella riformulazione del vincolo CVaR tramite un set di variabili ausiliarie. Fondamentali nell'analisi di problemi relativi al Conditional Value-at-Risk sono i contributi di M. Markowski [15, 16], in cui vengono descritti alcuni approcci per la regolarizzazione del CVaR, tramite i quali costruire metodi *Newton-based* per la risoluzione del sistema algebrico derivato dalla discretizzazione a elementi finiti di una PDE. Inoltre, un'applicazione del metodo smoothing-splitting al caso non vincolato con funzionale risk-averse è presentata negli articoli di S. Pieraccini e T. Vanzan [23], in cui viene introdotto un metodo di campionamento adattivo per ridurre il costo computazionale dell'algoritmo, e di Ciaramella et al. [5], dove sono sviluppati solutori multigrid per i sistemi KKT associati alle condizioni di ottimalità di problemi risk-neutral e risk-averse, possibilmente anche con vincoli di tipo box sulla variabile di controllo. Infine, è opportuno citare alcuni testi di riferimento riguardo le basi matematiche dell'ottimizzazione convessa (Boyd, Vandenberghe [3]) e l'analisi di metodi numerici in

ottimizzazione (Nocedal, Wright [22]), in particolare per la definizione dei metodi interior-point utilizzati per l'ottimizzazione vincolata. A tale scopo, si cita anche l'articolo dello stesso S. J. Wright [31], specificamente dedicato all'analisi dei metodi barriera e completo di una dettagliata analisi di convergenza.

1.2 Risultati preliminari

In questa sezione vengono introdotti alcuni fondamenti matematici propedeutici all'analisi del problema di ottimizzazione vincolata in esame. In particolare, sono presentate alcune nozioni base relative alla teoria della misura, all'analisi funzionale, all'analisi delle equazioni differenziali alle derivate parziali e alla teoria dell'ottimizzazione convessa.

1.2.1 Teoria della misura e integrazione secondo Lebesgue

Per affrontare una trattazione stocastica del problema di controllo ottimo, è opportuno introdurre alcuni concetti della teoria della misura. Le strutture fondamentali di questa trattazione sono le sigma-algebre e gli spazi misurabili, presentate nelle seguenti definizioni.

Definizione 1.2.1. *Sia X un insieme non vuoto e $\mathcal{P}(X)$ l'insieme delle parti di X. Una σ-algebra $\mathcal{A}$ di X è un sottoinsieme di $\mathcal{P}(X)$ tale che*

- $X \in \mathcal{A}$;
- *se $A \in \mathcal{A}$ allora $A^C \in \mathcal{A}$;*
- *se $\{A_n\}_{n\in\mathbb{N}^*}$ è una famiglia numerabile di elementi di $\mathcal{A}$ allora $\bigcup_{n\in\mathbb{N}^*} A_n \in \mathcal{A}$.*

Definizione 1.2.2. *La coppia $(X, \mathcal{A})$ si dice spazio misurabile e gli insiemi della σ-algebra $A \in \mathcal{A}$ vengono detti insiemi misurabili.*

Un esempio notevole è rappresentato dalla σ-algebra generata dagli aperti di X (costruita come l'intersezione di tutte le σ-algebre che li contengono), che viene detta σ-algebra di Borel e indicata con $\mathfrak{B}(X)$. Su una σ-algebra, è possibile introdurre la seguente definizione.

Definizione 1.2.3. *Sia $\mathcal{A}$ una σ-algebra di parti di X. Si definisce misura una funzione $\mu : \mathcal{A} \to [0, +\infty]$ che sia σ-additiva, cioè tale per cui*

- $\mu(\varnothing) = 0$, *dove $\varnothing$ è l'insieme vuoto;*
- *se $\{A_n\}_{n\in\mathbb{N}^*}$ è una famiglia numerabile di elementi di $\mathcal{A}$, tali che $A_i \cap A_j = \varnothing$ se $i \neq j$ allora $\mu(\bigcup_{n\in\mathbb{N}^*} A_n) = \sum_{n\in\mathbb{N}^*} \mu(A_n)$.*

Pertanto, la terna $(X, \mathcal{A}, \mu)$ si dice *spazio di misura*. Inoltre, se $\mu(X) = 1$ (misura di probabilità), tale spazio si definisce *spazio di probabilità* o *modello aleatorio*. Questi concetti astratti vengono riformulati con una differente notazione e significato nell'ambito del calcolo delle probabilità, dove si indica con Ω lo spazio campionario, con $\mathcal{F}$ lo spazio degli eventi e con $\mathbb{P}$ la misura di probabilità. Sulla base di queste definizioni preliminari,

per definire l'integrazione secondo Lebesgue, è necessario costruire una misura naturale (detta appunto *misura di Lebesgue*) su $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. A tale scopo è utile enunciare il seguente teorema.

Teorema 1.2.1 (di estensione o prolungamento (Carathéodory)). *Sia λ una funzione σ-additiva su un'algebra $\mathcal{A}$. Allora $\exists\, \mu$ misura definita su $\sigma(\mathcal{A})$ che estende λ, cioè tale per cui $\mu(A) = \lambda(A)$, $\forall A \in \mathcal{A}$.*

Per costruire la misura di Lebesgue sono necessarie alcune operazioni preliminari.

Step 1. Si definisce inizialmente una misura sull'intervallo [0,1), detta *misura di Borel-Lebesgue*, a partire dalla funzione σ-additiva

$$m : \mathcal{E} \to [0, +\infty],\ t.\,c.\ m([a,b]) = \begin{cases} b-a & se\ a < b, \\ 0 & se\ a = b, \end{cases} \tag{1.2.1}$$

su $\mathcal{E} = \{[a,b) \,|\, 0 \le a \le b \le 1\}$. Per il teorema di estensione, è possibile definire una misura che estende m su $\sigma(\mathcal{A})$, che si indica con $m_{[0,1)}$.

Step 2. In maniera analoga si costruisce una misura su un generico intervallo $[a,b)$.

Step 3. Si definisce quindi una misura su $\mathbb{R}$, come

$$m(A) = \lim_{n\to\infty} m_{[-n,n)}(A \cap [-n,n)), \quad \forall A \in \mathfrak{B}(\mathbb{R}). \tag{1.2.2}$$

Si dimostra che m così definita è una misura e viene detta *misura di Lebesgue* su $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$.

Step 4. Infine è necessario completare la σ-algebra $\mathfrak{B}(\mathbb{R})$ rispetto alla misura appena definita m. A tale scopo si definisce la *σ-algebra di Lebesgue*

$$\mathcal{L}(\mathbb{R}) = \mathcal{Z}_m(\mathfrak{B}(\mathbb{R})) = \{A \cup N \,|\, A \in \mathfrak{B}(\mathbb{R}), N\ sottoinsieme\ di\ un\ insieme\ a\ misura\ nulla\}, \tag{1.2.3}$$

su cui si definisce la misura $\overline{m}(B) = \overline{m}(A \cup N) = m(A)$, $\forall B \in \mathcal{L}(\mathbb{R})$.

Il collegamento naturale tra spazi misurabili è stabilito dalla seguente definizione.

Definizione 1.2.4. *Siano $(X, \mathcal{A})$ e $(Y, \mathcal{F})$ due spazi misurabili. Una funzione $f : X \to Y$ si dice misurabile se la controimmagine di A mediante f*

$$f^{-1}(A) = \{x \in X \,|\, f(x) \in A\} \in \mathcal{A}, \quad \forall A \in \mathcal{F}.$$

Nel calcolo delle probabilità una funzione misurabile $X : \Omega \to \mathbb{R}$ si dice *variabile aleatoria*.

Osservazione 1.2.1. *Una funzione continua tra spazi metrici, su cui sono definite le relative σ-algebre di parti, è anche misurabile. Inoltre sono misurabili la composizione di funzioni misurabili, il limite puntuale di funzioni misurabili ed una serie di operazioni elementari che coinvolgono funzioni misurabili, quali somma, prodotto, parte positiva, valore assoluto, etc.*

Inoltre, è possibile definire un concetto di convergenza per funzioni misurabili, come dalle seguenti definizioni.

Definizione 1.2.5. *Sia $f : X \to \overline{\mathbb{R}}$ una funzione misurabile su uno spazio misurabile $(X, \mathcal{A}, \mu)$. f si dice nulla quasi ovunque se $\{f \neq 0\} \coloneqq \{x \in X \mid f(x) \neq 0\}$ è tale per cui $\mu(\{f \neq 0\}) = 0$ e si indica con $f = 0$ q.o..*

Definizione 1.2.6. *Siano $f_n : X \to \overline{\mathbb{R}}$, per $n \geq 1$. Si dice quindi che $\{f_n\}_{n\in\mathbb{N}^*}$ converge quasi ovunque ad una funzione f, cioè $f_n \overset{q.o.}{\to} f$, se*

$$\exists N \in \mathcal{A} \text{ a misura nulla t.c. } \lim_{n\to\infty} f_n(x) = f(x), \ \forall x \in N^C.$$

Infine, per dare un'interpretazione matematica all'integrazione rispetto alla misura di Lebesgue, è necessario introdurre la seguente definizione.

Definizione 1.2.7. *Sia $f : X \to \overline{\mathbb{R}}$ una funzione misurabile. Allora f è una funzione semplice se $Im(f) = f(X)$ è un insieme finito, e si indica con $f \in \mathcal{S}(x)$.*

La caratterizzazione delle funzioni semplici è basata su una partizione di X di elementi A_i della σ-algebra $\mathcal{A}$ tale che $\bigcup_{i=1}^n A_i = X$ e $A_i \cap A_j = \varnothing$, se $i \neq j$. Si ha che

$$f \in \mathcal{S}(X) \iff \exists a_1, \dots, a_n \in \overline{\mathbb{R}} \text{ t.c. } f(x) = \sum_{i=1}^n a_i I_{A_i}(x), \quad \forall x \in X, \tag{1.2.4}$$

dove I è la funzione indicatrice. Il concetto di funzione semplice è utile per definire l'integrazione secondo Lebesgue. Infatti è opportuno applicare tale operazione a funzioni semplici, per poi estenderla a funzioni misurabili. Valgono pertanto le seguenti definizioni.

Definizione 1.2.8. *Sia $(X, \mathcal{A}, \mu)$ uno spazio di misura. Sia definita una funzione semplice non negativa $\phi \in \mathcal{S}^+(X)$, per cui vale la rappresentazione data dalla definizione precedente, $\phi = \sum_{i=1}^n a_i I_{A_i}$. Si dice integrale secondo Lebesgue di ϕ rispetto alla misura μ*

$$\int_X \phi d\mu = \sum_{i=1}^n a_i \mu(A_i), \tag{1.2.5}$$

con ϕ integrabile se tale integrale è finito.

Definizione 1.2.9. *Sia $f : X \to [0, \infty]$ una funzione misurabile. Si dice integrale secondo Lebesgue di ϕ rispetto alla misura μ*

$$\int_X f d\mu = \sup \left\{ \int_X \phi d\mu \,\middle|\, \phi \in \mathcal{S}^+(X), 0 \leq \phi \leq f \right\}, \tag{1.2.6}$$

con f integrabile se tale integrale è finito.

Osservazione 1.2.2. *L'estensione a funzioni misurabili $\overline{f} = f^+ - f^-$ di segno variabile segue dalla definizione precedente, come*

$$\int_X \overline{f} d\mu = \int_X f^+ d\mu - \int_X f^- d\mu,$$

se almeno uno degli integrali a secondo termine è finito.

Inoltre è utile introdurre il seguente importante risultato.

Teorema 1.2.2 (di convergenza dominata). *Sia $\{f_n\}_{n\in\mathbb{N}^*}$ una successione di funzioni misurabili tale che $f_n \to f$ q.o., con f misurabile. Si supponga inoltre che $\exists g$ integrabile tale che $|f_n(x)| \leq g(x)$, $\forall n \in \mathbb{N}^*, \forall x \in X$. Allora f_n ed f sono integrabili ed esiste*

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X f d\mu. \tag{1.2.7}$$

Infine, per motivare la trattazione su spazi prodotto è opportuno introdurre i concetti di σ-algebra prodotto e di misura prodotto.

Definizione 1.2.10. *Siano $(X_1, \mathcal{A}_1, \mu_1)$ e $(X_2, \mathcal{A}_2, \mu_2)$ due spazi misurabili. Siano $A \in \mathcal{A}_1$ e $B \in \mathcal{A}_2$, allora un sottoinsieme di $X_1 \times X_2$ definito da $A \times B$ si dice rettangolo misurabile. La σ-algebra prodotto è pertanto la σ-algebra generata dai rettangoli misurabili*

$$\mathcal{A}_1 \times \mathcal{A}_2 = \sigma(\{A \times B \,|\, A \in \mathcal{A}_1, B \in \mathcal{A}_2\}).$$

Si dimostra inoltre che la funzione $\mu_1 \times \mu_2 : \mathcal{A}_1 \times \mathcal{A}_2 \to [0, \infty]$ definita da

$$\mu_1 \times \mu_2(A \times B) = \int_{X_1} \mu_2(E_x^2) d\mu_1(x) = \int_{X_2} \mu_1(E_y^1) d\mu_2(y),$$

dove

$$E_x^2 = \begin{cases} B & \text{se } x \in A \\ \varnothing & \text{se } x \notin A \end{cases} \in \mathcal{A}_2, \qquad E_y^1 = \begin{cases} A & \text{se } y \in B \\ \varnothing & \text{se } y \notin B \end{cases} \in \mathcal{A}_1,$$

è una misura σ-finita su $(X_1 \times X_2, \mathcal{A}_1 \times \mathcal{A}_2)$, detta misura prodotto, date due misure σ-finite μ_1 e μ_2.

In questo framework di spazi prodotto, è opportuno citare due importanti risultati, il Teorema di Tonelli ed il Teorema di Fubini [10], che permettono rispettivamente di integrare funzioni misurabili non negative tramite un integrale doppio e di scambiare l'ordine di integrazione per funzioni integrabili secondo Lebesgue.

1.2.2 Framework funzionale

L'analisi variazionale di problemi differenziali si fonda su una descrizione astratta degli spazi a dimensione infinita di definizione delle variabili. In questo ambito, è fondamentale introdurre alcuni concetti e teoremi dell'Analisi Funzionale, su cui si basa l'intera trattazione.

Spazi di Banach e di Hilbert

Una delle strutture fondamentali dell'Analisi Funzionale è lo spazio di Banach, definito dalla seguente definizione.

Definizione 1.2.11. *Si dice spazio di Banach ogni spazio vettoriale normato $(X, ||\cdot||_X)$ completo rispetto alla metrica indotta dalla sua norma.*

Si ricorda che uno spazio vettoriale normato è completo se tutte le successioni di Cauchy $\{x_n\}_{n\in\mathbb{N}}$ sono convergenti in X, quindi se $\forall\{x_n\}_{n\in\mathbb{N}}$ di Cauchy, $\exists x \in X$ tale che $||x_n - x||_X \to 0$. Di seguito sono presentati alcuni esempi noti di spazi di Banach.

- Lo spazio delle successioni p-sommabili

$$\ell^p = \left\{ \underline{x} = \{x_n\}_{n\in\mathbb{N}}, x_n \in \mathbb{R} \,\middle|\, \sum_{n=1}^{\infty} |x_n|^p < \infty \right\}, \tag{1.2.8}$$

  per $1 \leq p < \infty$ e il caso particolare

$$\ell^\infty = \left\{ \underline{x} = \{x_n\}_{n\in\mathbb{N}}, x_n \in \mathbb{R} \,\middle|\, \sup_{n\in\mathbb{N}} |x_n| < \infty \right\}. \tag{1.2.9}$$

- Lo spazio delle funzioni p-sommabili, su uno spazio di misura $(D, \mathcal{A}, \mu)$, dove $\mathcal{A}$ è una σ-algebra e μ è una misura

$$L^p(D) = \left\{ f : D \to \mathbb{R} \text{ misurabili} \,\middle|\, ||f||_{L^p(D)} = \left(\int_D |f|^p \right)^{1/p} < \infty \right\}, \tag{1.2.10}$$

  per $1 \leq p < \infty$ e il caso particolare

$$L^\infty(D) = \left\{ f : D \to \mathbb{R} \text{ misurabili} \,\middle|\, ||f||_{L^\infty(D)} = \operatorname{ess\,sup}_D |f| < \infty \right\}. \tag{1.2.11}$$

- Lo spazio delle funzioni continue su un intervallo chiuso $C([a, b])$, rispetto alla norma del massimo.

Un sottoinsieme di notevole importanza concettuale degli spazi di Banach è rappresentato dai cosiddetti spazi di Hilbert, ossia spazi normati tali per cui la norma è indotta da un prodotto scalare, cioè da una mappa bilineare $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ tale che

$$\begin{aligned}
&\langle \alpha x + \beta y, z \rangle = \langle \alpha x, z \rangle + \langle \beta y, z \rangle, \quad \forall x, y, z \in X, \forall \alpha, \beta \in \mathbb{R}, \\
&\langle x, y \rangle = \langle y, x \rangle, \quad \forall x, y \in X, \\
&\langle x, x \rangle \geq 0, \quad \forall x \in X, \\
&\langle x, x \rangle = 0 \iff x = 0.
\end{aligned}$$

Ad esempio, sono spazi di Hilbert $L^2(D)$ e ℓ^2.
Una proprietà importante relativa al prodotto scalare è rappresentata dalla seguente disuguaglianza

$$| \langle x, y \rangle | \leq ||x||_X ||y||_X, \quad \forall x, y \in X, \tag{1.2.12}$$

detta *disuguaglianza di Cauchy-Schwarz*.
Per descrivere gli elementi di un generico spazio di Hilbert H è necessario definire una base in tale spazio. Una scelta comune, per la semplicità di rappresentazione degli elementi, è l'introduzione di una base ortonormale, ossia un insieme $\{e_n \,|\, n \in \mathbb{N}^*\}$ che sia

- ortonormale, cioè tale per cui

$$\langle e_n, e_m \rangle = \delta_{nm} = \begin{cases} 1, & n = m, \\ 0, & n \neq m. \end{cases}$$

- completo, ossia tale per cui $\overline{\text{Span}\{e_n \,|\, n \in \mathbb{N}^*\}} = H$.

Si dimostra che una base ortonormale permette di rappresentare ogni elemento di H come

$$x = \sum_{n=1}^{\infty} \langle x, e_n \rangle \, e_n. \tag{1.2.13}$$

Osservazione 1.2.3. *Ad esempio, è possibile mostrare che lo spazio di Hilbert $L^2(-\pi, \pi)$ ammette la base di Fourier ortonormale*

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos(nx), \frac{1}{\sqrt{\pi}} \sin(nx) \right\}_{n \in \mathbb{N}^*}.$$

Operatori limitati e dualità

Il legame tra elementi di spazi di Banach è descritto tramite la definizione di operatori. In particolare verranno considerati quelli lineari e limitati, che mappano un elemento di uno spazio in un altro, eventualmente appartenente ad un altro spazio. Vale allora la seguente definizione.

Definizione 1.2.12. *Un operatore $T : X \to Y$ tra spazi di Banach X, Y si dice*

- *lineare se*
$$T(\alpha x + \beta y) = \alpha T x + \beta T y, \quad \forall x, y \in X, \, \forall \alpha, \beta \in \mathbb{R};$$
- *limitato se $\exists C > 0$ tale che*
$$||Tx||_Y \leq C ||x||_X, \quad \forall x \in X.$$

La norma di T è definita da $||T|| = \inf\{C > 0 \,|\, ||Tx||_Y \leq C||x||_X, \forall x \in X\}$ o, equivalentemente, come

$$||T|| = \sup_{x \in X, \, x \neq 0} \frac{||Tx||_Y}{||x||_X}. \tag{1.2.14}$$

L'insieme degli operatori lineari e limitati tra spazi di Banach forma lo spazio vettoriale $\mathcal{B}(X, Y)$, che, munito della suddetta norma, è a sua volta uno spazio di Banach. Quindi, è possibile stabilire la seguente definizione.

Definizione 1.2.13. *Dato uno spazio di Banach X, il suo spazio duale è*

$$X' = \mathcal{B}(X, \mathbb{R}) = \{ f : X \to \mathbb{R} \text{ lineari e limitati } \}. \tag{1.2.15}$$

I suoi elementi $F \in X'$ si dicono funzionali lineari continui.

Ad esempio, lo spazio duale H' di uno spazio di Hilbert H è costituito dalle applicazioni lineari e limitate $f : H \to \mathbb{R}$, definite da $x \in H \to \langle x, x_0 \rangle = f(x)$, dato $x_0 \in H$. In particolare, per gli spazi di Hilbert reali vale il seguente teorema.

Teorema 1.2.3 (**di Riesz-Fréchet**). *Sia definita un'applicazione $R_H : H \to H'$, tale per cui $y \in H \to f_y = R_H(y) \in H'$, dove f_y è un funzionale definito da $f_y(x) = \langle x, y \rangle$. Allora R_H è un'isometria lineare suriettiva. Di conseguenza, è un isomorfismo e quindi si ha che $H \cong H'$. L'applicazione R_H assume il nome di operatore di Riesz.*

Più in generale, per uno spazio di Banach, vale il seguente teorema.

Teorema 1.2.4 (**di Hahn-Banach**). *Sia X uno spazio di Banach, e $W \subset X$ un suo sottospazio. Se $f : W \to \mathbb{R}$ è lineare e limitato, allora $\exists \tilde{f} \in X'$ tale per cui $\tilde{f}|_W = f$ e $||\tilde{f}||_{X'} = ||f||$.*

Nell'ambito dell'ottimizzazione convessa, è interessante considerare alcune implicazioni geometriche di questo teorema, come la separazione tra punti o tra insiemi convessi appartenenti allo spazio di Banach, tramite funzionali.

Corollario 1.2.1. *Sia X uno spazio di Banach non nullo. Allora $\forall x, y \in X$ tali che $x \neq y$ si ha che $\exists f \in X'$ tale che $f(x) \neq f(y)$.*

Corollario 1.2.2. *Sia X uno spazio di Banach non nullo. Allora se A, B sono sottoinsiemi non vuoti, convessi e disgiunti in X, dei quali almeno uno è aperto, allora $\exists f \in X'$ tale per cui l'iperpiano $f = a$, con $a \in \mathbb{R}$ separa i due sottoinsiemi.*

Oltre alla dualità semplice, è possibile definire il concetto di *bidualità* come

Definizione 1.2.14. *Sia X uno spazio di Banach, lo spazio biduale di X è definito come*

$$X'' = (X')' = \{F : X' \to \mathbb{R} \,|\, F \text{ lineare e limitato } \}, \tag{1.2.16}$$

con norma

$$||F||_{X''} = \sup_{f \in X', f \neq 0} \frac{|F(f)|}{||f||_{X'}}. \tag{1.2.17}$$

Vale pertanto il seguente teorema.

Teorema 1.2.5. *Sia definita sullo spazio di Banach X l'applicazione $J : X \to X''$ tale per cui $x \in X \to J(x) \in X''$, con $J(x)(f) = f(x)$ per ogni $f \in X'$, è un'isometria lineare. Pertanto X si identifica isometricamente con $J(X) \subset X''$.*

In generale, quindi l'applicazione J non è suriettiva. Uno spazio di Banach per cui questa ipotesi è tuttavia valida si dice *riflessivo* e si ha che X si identifica isometricamente con X''.

Osservazione 1.2.4. *Ogni spazio di Hilbert H è sempre riflessivo. Questo segue dal teorema di Riesz-Fréchet, applicato due volte per gli spazi biduali $H \cong H' \cong H''$.*

Basandosi su questa descrizione degli spazi duali e biduali, è possibile introdurre, oltre alla convergenza classica in norma sullo spazio di Banach, alcune nozioni di convergenza in senso debole, definite sugli spazi X e X'. Ad esempio, si considerano per $\{x_n\}_{n\in\mathbb{N}^*} \subset X$ e per $\{f_n\}_{n\in\mathbb{N}^*} \subset X'$ le seguenti convergenze.

- Convergenza in norma, indicata con il simbolo $\overset{n\to\infty}{\longrightarrow}$.

  Si dice che

  $$x_n \overset{n\to\infty}{\longrightarrow} x \in X \iff ||x_n - x||_X \to 0, \quad f_n \overset{n\to\infty}{\longrightarrow} f \in X' \iff ||f_n - f||_{X'} \to 0.$$

- Convergenza debole, indicata con il simbolo $\overset{n\to\infty}{\rightharpoonup}$.

  Si dice che

  $$x_n \overset{n\to\infty}{\rightharpoonup} x \in X \text{ se } f(x_n) \overset{n\to\infty}{\longrightarrow} f(x), \quad \forall f \in X',$$
  $$f_n \overset{n\to\infty}{\rightharpoonup} f \in X' \text{ se } F(f_n) \overset{n\to\infty}{\longrightarrow} F(f) \in X', \quad \forall F \in X''.$$

- Convergenza debole-star, indicata con il simbolo $\underset{n\to\infty}{\overset{*}{\rightharpoonup}}$.

  Si dice che

  $$f_n \underset{n\to\infty}{\overset{*}{\rightharpoonup}} f \in X' \text{ se } J(x)(f_n) \overset{n\to\infty}{\longrightarrow} J(x)(f) \in X'', \quad \forall x \in X.$$

Osservazione 1.2.5. *La convergenza in norma implica la convergenza debole (ma non viceversa) e analogamente la convergenza debole implica quella debole-star ma non viceversa. Tuttavia, su uno spazio di Banach riflessivo (a maggior ragione se spazio di Hilbert), si ha che X e X'' sono isomorfi, pertanto i concetti di convergenza debole e debole-star coincidono.*

Vale inoltre il seguente importante risultato, con il relativo corollario, entrambi utili nella dimostrazione dell'esistenza del minimo del problema di ottimizzazione convessa in esame nei capitoli seguenti.

Teorema 1.2.6 (di Banach-Alaouglu). *Sia X uno spazio di Banach separabile, cioè tale per cui sia ammesso un sottoinsieme numerabile denso*

$$\exists \{x_n\}_{n\in\mathbb{N}} \subset X \text{ t.c. } \overline{\{x_n \,|\, n \in \mathbb{N}\}} = X.$$

Sia inoltre $\{f_n\}_{n\in\mathbb{N}} \subset X'$ limitata.
Allora esiste una sottosuccessione $\{f_{n_k}\}_{k\in\mathbb{N}}$ che converge debolmente-star a $f \in X'$.

Corollario 1.2.3. *Sia X uno spazio di Banach riflessivo con duale X' separabile. Se $\{x_n\}_{n\in\mathbb{N}}$ è una successione limitata in X, allora esiste una sottosuccessione $\{x_{n_k}\}_{k\in\mathbb{N}}$ debolmente convergente a $x \in X$.*

Dimostrazione. Questo risultato deriva dall'applicazione del Teorema di Banach-Alaouglu a X' e X''. Per la limitatezza di $\{x_n\}$, si può definire una successione $\{J(x_n)\}_{n\in\mathbb{N}}$ limitata in X'', che quindi ammette una sottosuccessione $\{J(x_{n_k})\}_{k\in\mathbb{N}}$ che converge debolmente-star a $F \in X''$, cioè tale che

$$\exists \lim_{k\to\infty} J(x_{n_k})(f) = F(f), \quad \forall f \in X'.$$

Quindi, poiché X è riflessivo, si ha che $\exists x \in X$ tale che $F = J(x)$. In definitiva pertanto

$$\exists \lim_{k\to\infty} f(x_{n_k}) = f(x), \quad \forall f \in X'.$$

da cui si conclude che $x_{n_k} \xrightarrow{k\to\infty} x$. □

Osservazione 1.2.6. *Quest'ultimo risultato è fondamentale per la dimostrazione dell'esistenza di minimi di funzionali a valori reali su spazi di funzioni, poichè permette l'estrazione di sottosuccessioni minimizzanti debolmente convergenti (in spazi di Hilbert).*

Infine, visto il suo utilizzo nella definizione del gradiente del funzionale obiettivo del problema di controllo ottimo, è opportuno introdurre anche una nozione di derivata su spazi di Banach.

Definizione 1.2.15. *Siano X e Y due spazi di Banach e sia definita $f : X \to Y$. Tale funzione si dice derivabile secondo Fréchet in $x \in X$ se esiste un operatore lineare e limitato $D_f(x) \in \mathcal{B}(X, Y)$ tale che*

$$\lim_{||h||_X\to 0} \frac{||f(x+h) - f(x) - D_f(x)(h)||_Y}{||h||_X} = 0, \quad \forall h \in X. \tag{1.2.18}$$

Inoltre, è necessario introdurne anche una nozione più debole tramite la definizione seguente.

Definizione 1.2.16. *Siano X e Y due spazi di Banach e sia definita $f : X \to Y$. Tale funzione si dice derivabile secondo Gâteaux in $x \in X$ se per ogni direzione $h \in X$ esiste la derivata direzionale*

$$\partial_h f(x) = \lim_{\tau\to 0} \frac{f(x+\tau h) - f(x)}{\tau}. \tag{1.2.19}$$

Pertanto la mappa $h \to \partial_h f(x)$ viene detta derivata di Gâteaux di f.

1.2.3 Spazi di Sobolev

Sia definito un dominio $D \subset \mathbb{R}^d$, con $d = 1,2,3$, aperto, limitato e con bordo di Lipschitz, cioè tale per cui esiste un intorno U di x_0 su cui è definita una funzione Lipschitz continua $\phi : \mathbb{R}^{d-1} \to \mathbb{R}$ tale che

$$\partial D \cap U = \{x = (x_d, x_0) \in U \,|\, x_0 = \phi(x_d)\}, \qquad x_d \in \mathbb{R}^{d-1}. \tag{1.2.20}$$

Sia introdotto, su questo dominio, lo spazio delle funzioni localmente integrabili $L^1_{loc}(D)$, definito come

$$L^1_{loc}(D) = \left\{ u : D \to \mathbb{R} \text{ misurabile} \,\middle|\, \int_K |u(x)|dx < \infty, \forall K \subset D \text{ compatto} \right\}. \quad (1.2.21)$$

La definizione del concetto di derivata debole è quindi propedeutica all'introduzione degli spazi di Sobolev.

Definizione 1.2.17. *Una funzione $u \in L^1_{loc}(D)$ su un aperto $D \in \mathbb{R}^d$ si dice derivabile in senso debole rispetto ad una componente x_j se $\exists v \in L^1_{loc}(D)$ tale che*

$$\int_D u \frac{\partial \phi}{\partial x_j} dx = -\int_D v\phi dx, \qquad \forall \phi \in C^\infty_c(D), \quad (1.2.22)$$

dove è definito lo spazio delle funzioni C^∞ a supporto compatto

$$C^\infty_c(D) = \{\phi : D \to \mathbb{R}^n \,|\, \phi \in C^\infty(D),\ supp(\phi)\ compatto\ su\ D\}. \quad (1.2.23)$$

Pertanto si dice che v è la derivata in senso debole di u e si scrive $v = \partial u/\partial x_j$.

Si osserva che, nel caso la funzione $u \in C^1(D)$, i concetti di derivata debole e derivata classica coincidono.
Alla luce di questa definizione, è possibile introdurre lo spazio di Sobolev $H^1(D)$ delle funzioni di $L^2(D)$ con anche le relative derivate deboli in $L^2(D)$,

$$H^1(D) = \left\{ u \in L^2(D) \,\middle|\, \frac{\partial u}{\partial x_j} \in L^2(D), \forall j = 1, \dots, d \right\}, \quad (1.2.24)$$

su cui è definita la norma indotta dal prodotto scalare

$$||u||_{H^1(D)} = \left(||u||^2_{L^2(D)} + ||\nabla u||^2_{L^2(D)} \right)^{1/2}. \quad (1.2.25)$$

Poichè $|\nabla u|^2 = \sum_{j=1}^d (\partial u/\partial x_j)^2$, si osserva che la condizione $\partial u/\partial x_j \in L^2(D)$ equivale a richiedere che $|\nabla u| \in L^2(D)$. Chiaramente dalla definizione segue che

$$||u||_{L^2(D)} \leq ||u||_{H^1(D)}, \qquad ||\,|\nabla u|\,||_{L^2(D)} \leq ||u||_{H^1(D)}.$$

Pertanto si ha che $H^1(D) \subset L^2(D)$ con *immersione continua*, cioè l'applicazione identità

$$\begin{aligned} i : H^1(D) &\to L^2(D), \\ u &\mapsto i(u)=u, \end{aligned}$$

è continua. Valgono, inoltre, le proprietà di immersione per funzioni regolari

- $\forall D \in \mathbb{R}^d$ limitato si ha che $C^1(\overline{D}) \subset H^1(D)$;
- $\forall D \in \mathbb{R}^d$ vale $C^\infty_c(D) \subset H^1(D)$.

In particolare, da questo risultato si osserva che il set delle funzioni C^∞ a supporto compatto, che risulta denso in $L^2(D)$, non lo è invece nello spazio di Sobolev $H^1(D)$. Si ricorda che la proprietà di densità di un sottospazio consiste nel fatto che

$$\forall v \in L^2(D),\ \exists\{\phi_n\}_{n\in\mathbb{N}} \in C_c^\infty(D) \text{ tale che } \phi_n \to u \text{ in } L^2(D). \tag{1.2.26}$$

Pertanto ha senso definire il sottospazio vettoriale chiuso, rispetto alla norma di $H^1(D)$

$$H_0^1(D) \coloneqq \overline{C_c^\infty(D)}_{||\cdot||_{H^1(D)}}, \tag{1.2.27}$$

tale per cui ora $C_c^\infty(D)$ è denso rispetto ad $H_0^1(D)$.
L'interpretazione matematica delle funzioni appartenenti a questo sottospazio segue dalla teoria delle tracce, dalla quale viene richiamato il seguente importante risultato.

Teorema 1.2.7. *Sia $D \subset \mathbb{R}^d$, con frontiera ∂D Lipschitziana, allora esiste un operatore lineare e continuo, detto operatore di traccia, definito da*

$$\begin{aligned} \tau_{\partial D} : H^1(D) &\to L^2(\partial D), \\ u &\mapsto \tau_{\partial D}(u)=u|_{\partial D},\ \textit{con } u\in C_c^\infty(D). \end{aligned} \tag{1.2.28}$$

Osservazione 1.2.7. *La continuità dell'operatore $\tau_{\partial D}$ segue dalla disuguaglianza di traccia*

$$\exists C > 0 \ \textit{t.c.}\ ||\tau_{\partial D}(u)||_{L^2(\partial D)} \le C||u||_{H^1(D)}, \quad \forall u \in H^1(D). \tag{1.2.29}$$

In aggiunta, è possibile dimostrare che l'operatore di traccia non risulta essere né iniettivo (il suo nucleo è proprio $H_0^1(D)$), né suriettivo (la sua immagine è $H^{1/2}(\partial D) \subset L^2(\partial D)$).

Rimane da introdurre una norma sullo spazio $H_0^1(D)$. A tale scopo si richiama il seguente teorema.

Teorema 1.2.8. *Sia $D \subset \mathbb{R}^d$ aperto e limitato. Allora si ha che*

$$\exists C_P > 0 \ \textit{t.c.}\ ||u||_{L^2(D)} \le C_P||\,|\nabla u|\,||_{L^2(D)}, \quad \forall u \in H_0^1(D). \tag{1.2.30}$$

Ogni costante che soddisfa tale disuguaglianza, detta disuguaglianza di Poincaré, viene denominata costante di Poincaré. Inoltre si fa riferimento al minimo valore di C_P come costante di Poincaré del dominio D.

Dimostrazione. Poichè $C_c^\infty(D)$ è denso in $H_0^1(D)$ è sufficiente mostrare che

$$||\phi||_{L^2(D)} \le C_P||\,|\nabla\phi|\,||_{L^2(D)}, \quad \forall\phi \in C_c^\infty(D).$$

Poichè $D \subset \mathbb{R}^d$ è limitato, è possibile introdurre un intervallo $[a,b]$ ed un punto $x_d \in \mathbb{R}^{d-1}$ tale per cui

$$x = (x_d, x_0), \quad D = \mathbb{R}^{d-1} \times [a,b].$$

Estendendo $C_c^\infty(D)$ tramite il suo prolungamento banale a $C_c^\infty(\mathbb{R}^d)$, dal Teorema Fondamentale del Calcolo segue che, $\forall x_d \in \mathbb{R}^{d-1}$, $\forall x_0 \in [a,b]$,

$$\phi(x_d, x_0) = \int_a^{x_0} \frac{\partial\phi}{\partial x_0}(x_d, s)ds.$$

Dalla disuguaglianza di Cauchy-Schwartz segue che

$$|\phi(x_d,x_0)| = \left|\int_a^{x_0} 1\cdot\frac{\partial\phi}{\partial x_0}(x_d,s)ds\right| \leq \left(\int_a^{x_0} 1^2 ds\right)^{1/2}\left(\int_a^{x_0}\left|\frac{\partial\phi}{\partial x_0}(x_d,s)\right|^2 ds\right)^{1/2}$$
$$\leq (x_0-a)^{1/2}\left(\int_a^b\left|\frac{\partial\phi}{\partial x_0}(x_d,s)\right|^2 ds\right)^{1/2}.$$

Elevando a quadrato ed integrando ambo i membri in $[a,b]$ si ha

$$\int_a^b \phi^2(x_d,x_0)dx_0 \leq \int_a^b (x_0-a)dx_0\int_a^b\left|\frac{\partial\phi}{\partial x_0}(x_d,s)\right|^2 ds = \frac{1}{2}(b-a)^2\int_a^b\left|\frac{\partial\phi}{\partial x_0}(x_d,s)\right|^2 ds.$$

Integrando quindi rispetto a x_d

$$\int_{\mathbb{R}^d\times[a,b]}\phi^2(x)dx \leq \frac{1}{2}(b-a)^2\int_{\mathbb{R}^d\times[a,b]}\left|\frac{\partial\phi}{\partial x_0}(x)\right|^2 dx.$$

Tuttavia, poichè $\phi\in C_c^\infty(D)$, si ha che $\phi\equiv 0$ fuori dal dominio D, e pertanto

$$||\phi||^2_{L^2(D)} = \int_D\phi^2(x)dx \leq \frac{1}{2}(b-a)^2\int_D\left|\frac{\partial\phi}{\partial x_0}(x)\right|^2 dx \leq \frac{1}{2}(b-a)^2||\,|\nabla\phi|\,||^2_{L^2(D)}.$$

In conclusione si osserva che questa espressione rappresenta la disuguaglianza richiesta all'inizio della dimostrazione, ponendo $C_P=(b-a)/\sqrt{2}$. □

Pertanto, ricordando che

$$||u||^2_{H^1(D)} = ||u||^2_{L^2(D)} + ||\,|\nabla u|\,||^2_{L^2(D)} \geq ||\,|\nabla\phi|\,||^2_{L^2(D)}, \qquad \forall u\in H^1(D),$$

una importante conseguenza del teorema appena dimostrato, consiste nel fatto che, se $u\in H_0^1(D)$, allora

$$||u||^2_{H^1(D)} \leq (1+C_P^2)||\,|\nabla u|\,||^2_{L^2(D)}.$$

Si conclude che la norma $||\,|\nabla u|\,||^2_{L^2(D)}$ risulta equivalente alla norma di $H^1(D)$ in $H_0^1(D)$.

Risulta quindi possibile enunciare la seguente definizione.

Definizione 1.2.18. *Su D limitato, $H_0^1(D)$ è uno spazio di Hilbert, dotato del prodotto scalare*

$$\langle u,v\rangle_{H_0^1(D)} = \int_D\nabla u\cdot\nabla v dx, \qquad \forall u,v\in H_0^1(D), \tag{1.2.31}$$

che induce la norma

$$||u||^2_{H_0^1(D)} \coloneqq ||\,|\nabla u|\,||^2_{L^2(D)}. \tag{1.2.32}$$

1.2.4 Trattazione debole dei problemi ellittici

In questa sottosezione, viene analizzato un problema ellittico modello, con condizioni al bordo miste di Dirichlet/Neumann. Durante la trattazione dei problemi di controllo ottimo, tale analisi sarà applicata ad alcune formulazioni particolari utilizzate per i test numerici. Sia pertanto definito il problema differenziale al contorno

$$\begin{cases} -\nabla\cdot(A\nabla y)+\nabla\cdot(\mathbf{b}y)+a_0 y = f & \text{in } D,\\ y = g & \text{su } \Gamma_D,\\ \dfrac{\partial y}{\partial n_a} = \psi & \text{su } \Gamma_N. \end{cases} \tag{1.2.33}$$

dove il bordo $\partial D = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ limitato, connesso e con frontiera Lipschitziana, è diviso in due parti la cui intersezione è l'insieme vuoto: una su cui sono applicate condizioni di Dirichlet non omogenee Γ_D e una su cui vale una condizione di Neumann Γ_N. Per quanto concerne i coefficienti, si ha che A è una matrice simmetrica e definita positiva sul dominio D tale che $A \in L^\infty(D)^{n\times n}$, $\mathbf{b} \in L^\infty(D)^n$, $a_0 \in L^\infty(D)$ ed infine si ha per la forzante $f \in L^2(D)$. Sia inoltre non vuoto il bordo di Neumann, così da poter definire la derivata conormale $\partial y/\partial n_A = n_A \cdot (A\nabla y)$, con $n_A = A\mathbf{n}$. La condizione di Dirichlet non omogenea è espressa tramite la funzione g, che rappresenta la traccia sulla frontiera di una funzione di $H^1(D)$, quindi $g \in H^{1/2}(D)$, mentre la condizione di Neumann contiene la funzione $\psi \in L^2(\Gamma_N)$. Per scrivere la formulazione debole del problema in esame, è necessario estendere la funzione g ad una funzione $\tilde{g} \in H^1(D)$ la cui traccia su ∂D coincida con g, $\tilde{g}|_{\partial D} = g$. Introdotta quindi

$$y_0 \in H^1_{0,\Gamma_D}(D) \coloneqq \{y \in H^1(D) | y = 0 \text{ su } \Gamma_D\}, \tag{1.2.34}$$

è possibile sostituire $y = y_0 + \tilde{g}$.
Il problema variazionale seguente si ottiene quindi applicando le formule di integrazione per parti.
Trovare $y_0 \in H^1_{0,\Gamma_D}(D)$ *tale che:*

$$\begin{aligned} \int_D (A\nabla y)\cdot\nabla\phi\,dx - \int_D (y_0+\tilde{g})(\mathbf{b}\cdot\nabla\phi)dx + \int_D a_0(y_0+\tilde{g})\phi\,dx + \int_{\Gamma_N} \mathbf{b}\cdot\mathbf{n}\,(y_0+\tilde{g})\phi d\sigma = \\ = \int_D f\phi\,dx + \int_{\Gamma_N}\psi\phi\,d\sigma, \qquad \forall\phi \in H^1_{0,\Gamma_D}(D). \end{aligned} \tag{1.2.35}$$

Una formulazione più compatta richiede la definizione della forma bilineare $a : H^1(D) \times H^1(D) \to \mathbb{R}$ tale per cui $a(y,\phi)$ coincide con il primo membro della precedente e della forma lineare $F : H^1(D) \to \mathbb{R}$, pari al secondo membro della (1.2.35). Si ha

$$a(y,\phi) = F(\phi), \qquad \forall\phi \in H^1_{0,\Gamma_D}(D). \tag{1.2.36}$$

Esistenza e unicità della soluzione di questo problema differenziale sono garantite da un risultato fondamentale dell'Analisi Funzionale, il Teorema di Lax-Milgram, di seguito enunciato.

Teorema 1.2.9 (**Lax-Milgram**). *Sia V uno spazio di Hilbert con norma $||\cdot||_V$. Siano $a(u,v) : V \times V \to \mathbb{R}$ una forma bilineare e $F(v) : V \to \mathbb{R}$ una forma lineare in V aventi le seguenti proprietà:*

- *$a(u,v)$ è continua, cioè esiste $C > 0$ tale per cui*

$$|a(u,v)| \leq C||u||_V||v||_V, \qquad \forall u, v \in V. \tag{1.2.37}$$

- *$a(u,v)$ è coerciva, cioè esiste $\alpha > 0$ tale per cui*

$$a(v,v) \geq \alpha||v||_V^2, \qquad \forall v \in V. \tag{1.2.38}$$

- *$F \in V'$, dove la notazione $(\cdot)'$ rappresenta lo spazio duale di V.*

Allora il problema variazionale: trovare $y \in V$ tale per cui $a(u,v) = F(v)$, $\forall v \in V$ ammette una ed un'unica soluzione, che soddisfa la seguente stima,

$$||u||_V \leq \frac{1}{\alpha}||F||_{V'}. \tag{1.2.39}$$

Osservazione 1.2.8. *Il Teorema di Lax-Milgram fornisce condizioni sufficienti per stabilire la buona positura del problema nel senso di Hadamard, quali esistenza, unicità della soluzione e dipendenza continua dai dati.*

Dimostrazione. Per dimostrare l'esistenza di una soluzione, è necessario mostrare che l'operatore lineare

$$\begin{aligned} A : V &\to V', \\ u &\mapsto Au, \end{aligned}$$

dove $Au \in V'$ rappresenta la mappa lineare e continua $u \in V \to a(u,v)$. Quindi A è definito da $(Au)(v) = a(u,v)$, $\forall u, v \in V$, che si può scrivere equivalentemente con il prodotto scalare in V come $\langle Au, v\rangle = a(u,v)$, $\forall u, v \in V$, è un isomorfismo tra V e V'. A tale scopo è necessario dimostrare che è continuo, iniettivo e che il suo inverso, rispetto alla sua immagine $Z = \{G \in V' \,|\, G = Au, \forall u \in V\}$, che coincide con V', è a sua volta continuo. Dalla continuità di a, segue che

$$||Au||_{V'} = \sup_{v\in V}\frac{\langle Au, v\rangle}{||v||_V} = \sup_{v\in V}\frac{a(u,v)}{||v||_V} \leq \frac{||a||||u||_V||v||_V}{||v||_V} = ||a||\,||u||_V, \quad \forall v \in V,$$

dove $||a||$ è la norma della forma bilineare a, quindi l'operatore A è continuo. Esso è anche iniettivo, poichè dalla proprietà di coercività di a si deduce che se per un qualche $u \in V$ si ha $Au = 0$, allora $0 = a(u,u) \geq \alpha||u||_V^2$, che implica necessariamente $u = 0$. Per quanto riguarda l'operatore inverso $A^{-1} : Z \to V$, ancora una volta è possibile mostrare la sua continuità utilizzando l'ipotesi su a. Infatti, dato $u \in V$ tale per cui $G = Au \in Z$ si ha che tale u soddisfa il problema $a(u,v) = G(v)$, $\forall v \in V$ e pertanto vale la stima $||A^{-1}G||_V \leq \frac{1}{\alpha}||G||_{V'}$. Infine, si osserva che Z è un sottospazio chiuso di V', in quanto se

$F \in \overline{Z}$ in V', allora esiste una sequenza $\{G_n\}_{n\in\mathbb{N}}$ che converge a F in norma V'. Pertanto è possibile definire una successione $\{w_n\}_{n\in\mathbb{N}}$, come $w_n = A^{-1}G_n \in V$, tale per cui

$$||w_n - w_m||_V \leq \frac{1}{\alpha}||G_n - G_m||_{V'}, \quad \forall n, m \in \mathbb{N}.$$

Segue che $\{w_n\}_{n\in\mathbb{N}}$ è una successione di Cauchy e converge a $w \in V$, poiché tale spazio è completo. Quindi $G = Aw \in Z$ e la successione $\{G_n\}_{n\in\mathbb{N}}$ converge a $G \in Z$. Per l'unicità del limite quindi si ha che $F \equiv G \in Z$. Infine, si mostra che l'immagine Z e lo spazio duale V' sono coincidenti. Sia per assurdo Z un sottospazio proprio (cioè non vuoto nè coincidente con l'intero V'). Per il Teorema di Hahn-Banach $\exists\mathcal{Z}$, forma lineare e continua su V' tale per cui $\mathcal{Z}(G) = 0$ ma $\mathcal{Z}(F) \neq 0$ per qualche $F \in V' \setminus Z$. Quindi $\mathcal{Z} \in V''$ e $||\mathcal{Z}||_{V''} > 0$. Per riflessività dello spazio V, pertanto, si ha che $\exists! w \in V$ che identifica $\mathcal{Z}$, nel senso $\mathcal{Z}(F) = F(w)$, $\forall F \in V'$ e $||\mathcal{Z}||_{V''} = ||w||_V > 0$. Scelto quindi $F = Av$, ricordando la proprietà di coercività di a si ha

$$0 = \mathcal{Z}(Aw) = (Aw)(w) = a(w, w) \geq \alpha||w||^2, \quad \forall v \in V,$$

da cui si deduce che $w = 0$. Ciò contraddice quindi l'ipotesi fatta che $||w||_V > 0$, pertanto si conclude che $Z \equiv V'$. Quindi $A : V \to V'$ è un isomorfismo ed esiste una soluzione al problema variazionale di riferimento. Dimostrata, l'esistenza, è immediato verificare che, scegliendo $u = v$ nella condizione di coercività segue che

$$\alpha||u||_V^2 \leq a(u, u) = F(u) \leq ||F||_{V'}||u||_V,$$

da cui la (1.2.39), ed inoltre, scelte due differenti soluzioni $u_1, u_2 \in V$ si ha

$$a(u_2 - u_1, v) = F(v) - F(v) = 0, \, \forall v \in V \Rightarrow ||u_1 - u_2||_V \leq \frac{1}{\alpha}||F - F||_{V'} = 0,$$

da cui si conclude che $u_1 \equiv u_2$. □

1.2.5 Teoria dell'ottimizzazione

In quest'ultima sezione introduttiva, vengono presentati alcuni risultati e definizioni nell'ambito dell'ottimizzazione convessa, utili per la ricerca delle soluzioni del problema di controllo ottimo in esame. Inizialmente, è utile ricordare la definizione di convessità, applicata ad un insieme e ad una funzione.

Definizione 1.2.19. *Un insieme $X \subset \mathbb{R}^n$ si dice convesso se contiene ogni segmento tra due suoi punti*

$$x_1, x_2 \in X, \lambda \in [0,1] \Rightarrow \lambda x_1 + (1 - \lambda)x_2 \in X. \tag{1.2.40}$$

Definizione 1.2.20. *Una funzione $f : \mathbb{R}^n \to \mathbb{R}$ si dice convessa se il suo dominio è un insieme convesso e se vale*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \tag{1.2.41}$$

Inoltre f si dice strettamente convessa se tale disuguaglianza è strettamente soddisfatta. Infine, f si dice fortemente convessa se $\exists m > 0$ tale che $\tilde{f}(x) = f(x) - \frac{m}{2}||x||_2^2$ è convessa, cioè

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{m}{2}\lambda(1 - \lambda)||x - y||_2^2. \tag{1.2.42}$$

Sia, infine, introdotto l'*epigrafo* di f, come il set dei punti giacenti sotto il grafico di tale funzione.

$$\text{epi}(f) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R}, x \in \text{dom}\, f, t \in \mathbb{R} \,|\, f(x) \leq t\}. \tag{1.2.43}$$

Osservazione 1.2.9. *Alcune funzioni elementari, quali combinazioni lineari, trasformazioni affini, parte positiva, limite puntuale applicate a funzioni convesse ne preservano la convessità.*

Al fine di dimostrare l'esistenza di soluzioni per problemi di ottimizzazione convessa, come nel teorema (2.1.1), è opportuno introdurre il concetto di semicontinuità inferiore, tramite la seguente definizione.

Definizione 1.2.21. *Sia $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Si dice che f è inferiormente semicontinua (s.c.i.) in $x_0 \in X$, se $\forall \epsilon > 0\, \exists U(x_0)$ intorno di x_0 tale per cui $f(x) > f(x_0) - \epsilon$, $\forall x \in U(x_0)$. Equivalentemente, f è s.c.i. se*

$$\liminf_{x \to x_0} f(x) \geq f(x_0). \tag{1.2.44}$$

I seguenti risultati mostrano le condizioni del primo e del secondo ordine per la convessità.

Proposizione 1.2.1 (**Condizioni del primo ordine**). *Sia f una funzione differenziabile, cioè con dominio aperto su cui esiste ovunque il gradiente. Allora vale*

$$f \text{ è convessa} \iff f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in dom\, f. \tag{1.2.45}$$

f è inoltre strettamente convessa se tale disuguaglianza è strettamente soddisfatta.

Proposizione 1.2.2 (**Condizioni del secondo ordine**). *Sia f una funzione due volte differenziabile, allora è convessa se e solo se la sua matrice Hessiana è semi-definita positiva, cioè*

$$f \text{ è convessa} \iff \nabla^2 f \succeq 0, \quad \forall x \in dom\, f. \tag{1.2.46}$$

f è strettamente convessa se la sua Hessiana è definita positiva, $\nabla^2 f \succ 0, \forall x \in dom\, f$. Infine si ha che se $\nabla^2 f \succeq mI, \forall x \in dom\, f$, per qualche $m > 0$, con I matrice identità, allora f è fortemente convessa.

Sia definito il seguente problema di ottimizzazione modello

$$\begin{aligned} p^\star = \min_{x \in \mathbb{R}^n} \; & f(x) \\ \text{s.t. } & g_i(x) \leq 0, \qquad i = 1, \ldots, N, \\ & h_j(x) = 0, \qquad j = 1, \ldots, M. \end{aligned} \tag{1.2.47}$$

Tale problema si dice *convesso* se il funzionale obiettivo e le funzioni g_i sono convessi e le h_i sono affini. Si dice *feasible set* l'insieme degli $x \in \mathbb{R}^n$ che soddisfano i vincoli di uguaglianza e disuguaglianza imposti, e si indica con $\mathcal{X}$. In forma più compatta, è pertanto possibile riscrivere tale problema come

$$p^\star = \min_{x \in \mathcal{X}} f(x). \tag{1.2.48}$$

L'obiettivo della minimizzazione consiste nel determinare il *minimo* $p^\star$ e possibilmente l'elemento del feasible set $x^\star \in \mathcal{X}$ per cui $f(x^\star) = p^\star$, detto *punto di minimo*. Nel caso in cui il feasible set sia vuoto, si fissa $p^\star = +\infty$ e il problema viene detto *infeasible*; se invece si ottiene che $p^\star = -\infty$, il problema è *inferiormente illimitato*. Nel caso di problemi convessi vale il seguente teorema.

Teorema 1.2.10. *Sia definito un problema di ottimizzazione con un funzionale obiettivo convesso f su un feasibile set convesso $\mathcal{X}$ come nella* (1.2.48)*. Allora ogni soluzione ottima locale di tale problema è anche globale.*

Dimostrazione. Sia $x^\star$ un punto di minimo locale per f, tale per cui $p^\star = f(x^\star)$. Bisogna mostrare che $f(y) \geq f(x^\star) = p^\star$, $\forall x \in \mathcal{X}$ per un generico y. Dalla definizione di convessità applicata a $x^\star$ e y si ha che $\exists \theta \in [0,1]$ tale che

$$f(\theta y + (1-\theta)x^\star) \leq \theta f(y) + (1-\theta) f(x^\star).$$

Sottraendo $f(x^*)$ ad entrambi i membri,

$$f(\theta y + (1-\theta)x^\star) - f(x^\star) \leq \theta(f(y) - f(x^\star)).$$

Poichè $x^\star$ è un minimo locale, il termine a sinistra è non-negativo per valori opportunamente piccoli di $\theta > 0$. Si conclude pertanto che anche il secondo termine è non negativo e quindi $f(y) \geq f(x^\star)$. □

Le condizioni di ottimalità (del primo ordine) per il problema (1.2.48) sono garantite dalla seguente proposizione.

Proposizione 1.2.3. *Per* (1.2.48) *si ha che*

$$x \in \mathcal{X} \text{ ottimo} \iff \nabla f(x)^\top (y - x) \geq 0, \quad \forall y \in \mathcal{X}. \tag{1.2.49}$$

Dimostrazione. Si ha che, $\forall x, y \in \operatorname{dom} f$, vale che $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$.

$\boxed{\Longleftarrow}$

Questa implicazione è immediata, in quanto data come ipotesi la condizione a destra, si ha che $f(y) \geq f(x)$ dalla precedente espressione.

$\boxed{\Longrightarrow}$

Viceversa, è necessario dimostrare che se x è ottimo allora $\nabla f(x)^\top (y - x) \geq 0$, $\forall y \in \mathcal{X}$. Chiaramente, ciò è immediato se $\nabla f(x) = 0$, mentre se $\nabla f(x) \neq 0$ si supponga per assurdo che, con x ottimo valga $\nabla f(x)^\top (y - x) < 0$, $\forall y \in \mathcal{X}$. Allora per un generico punto $x_\theta = \theta y + (1-\theta)x$, $\theta \in [0,1]$ si ha che, per θ sufficientemente piccolo, x_θ giace in un intorno di x dove il segno del termine del primo ordine della seguente espansione di Taylor prevale sugli altri,

$$\begin{aligned} f(x_\theta) =& f(x) + \nabla f(x)^\top (x_\theta - x) + o(||x_\theta - x||) \\ =& f(x) + \theta \nabla f(x)^\top (y - x) + o(||y - x||) \\ =& f(x) + \text{ termine negativo.} \end{aligned}$$

Ciò implicherebbe che per tale θ, $f(x_\theta) \geq f(x)$, che è assurdo perchè contraddice l'ottimalità di x. □

L'approccio considerato per i problemi di ottimizzazione del tipo (1.2.47) è basato sulla costruzione di una funzione detta *Lagrangiana* introducendo *moltiplicatori di Lagrange* relativi ai vincoli, da cui ricavare condizioni di ottimalità. Tale funzione si scrive come

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_{i=1}^{N} \lambda_i g_i(x) + \sum_{j=1}^{M} \nu_j h_j(x), \tag{1.2.50}$$

dove si introducono i vettori di moltiplicatori di Lagrange $\lambda = [\lambda_1, \ldots, \lambda_N]$ e $\nu = [\nu_1, \ldots, \nu_M]$. Una strategia efficace per definire condizioni di ottimalità e introdurre un limite inferiore per il valore ottimo consiste nella definizione del *problema duale* di quello in esame, a partire dalla funzione Lagrangiana.

Definizione 1.2.22. *Si definisce funzione duale una funzione* $g(\lambda, \nu) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ *tale per cui posto* $\lambda \geq 0$*, si ha*

$$g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu) = \inf_x f(x) + \sum_{i=1}^{N} \lambda_i g_i(x) + \sum_{j=1}^{M} \nu_j h_j(x). \tag{1.2.51}$$

Vale, quindi, la seguente proposizione.

Proposizione 1.2.4. *La funzione duale* $g(\lambda, \nu)$ *è concava in* (λ, ν)*. Inoltre si ha che* $g(\lambda, \nu) \leq p^\star, \forall \lambda \geq 0, \forall \nu$*.*

Poichè la funzione duale risulta essere in generale un limite inferiore al valore ottimo, è naturale ricercare il "migliore possibile" di tali valori. Pertanto si definisce il problema

$$d^\star = \max_{\lambda, \nu} g(\lambda, \nu), \quad \text{s.t. } \lambda \geq 0, \tag{1.2.52}$$

detto problema duale.

Osservazione 1.2.10. *In virtù del risultato precedente, il problema duale risulta convesso anche nel caso in cui il problema primale non lo sia.*

Un'altra conseguenza della Proposizione (1.2.4) consiste nel fatto che $d^\star \leq p^\star$. Questa proprietà si denota con l'espressione *dualità debole*. Inoltre la quantità $\delta = p^\star - d^\star$ si dice *duality gap*. Sotto alcune ipotesi aggiuntive, tra cui la convessità del problema primale, è possibile mostrare la validità di una relazione di uguaglianza $d^\star = p^\star$, detta *dualità forte*. In particolare vale la seguente proposizione.

Proposizione 1.2.5 (Condizioni di Slater). *Siano* g_i*,* $i = 1, \ldots, N$ *funzioni convesse e* h_j*,* $j = 1, \ldots, M$ *funzioni affini. Siano inoltre le prime* $k \leq N$ *funzioni* g_i *affini. Sia inoltre introdotto l'interno relativo del dominio* D*, definito da*

$$\textit{relint } D = \{x \in D \,|\, \exists U \textit{ intorno di } x \in \textit{aff}(D) \textit{ t.c. } U \subset D\}, \tag{1.2.53}$$

con

$$\mathit{aff}(D) = \left\{ \sum_{i=1}^{k} \theta_i x_i \,|\, x_i \in D, \sum_{i=1}^{k} \theta_i = 1, \theta_i \in \mathbb{R}, k \in \mathbb{N} \right\} \tag{1.2.54}$$

insieme di tutte le combinazioni affini su D. Allora $\exists x \in$ relint D tale che

$$\begin{aligned} g_1(x) \leq 0, \ldots, g_k(x) \leq 0, \quad g_{k+1}(x) < 0, \ldots, g_N(x) < 0, \\ h_1(x) = 0, \ldots, h_M(x) = 0, \end{aligned}$$

allora vale la dualità forte $d^\star = p^\star$. Inoltre se $p^\star > \infty$, anche il valore ottimo duale viene raggiunto e si ha che $g(\lambda^\star, \nu^\star) = d^\star = p^\star$.

Sotto queste ipotesi, le condizioni che caratterizzano l'ottimalità sono

$$\nabla_x (f(x) + \lambda^\top g(x) + \nu^\top h(x)) = 0, \qquad \text{stazionarietà delle Lagrangiana,} \tag{1.2.55}$$

$$\lambda_i g_i(x) = 0, \forall i = 1, \ldots, N, \qquad \text{complementary slackness,} \tag{1.2.56}$$

$$g(x) \leq 0, h(x) = 0, \qquad \text{feasibility del problema primale,} \tag{1.2.57}$$

$$\lambda \geq 0, \qquad \text{feasibility del problema duale.} \tag{1.2.58}$$

che assumono il nome di *condizioni di Karush-Kuhn-Tucker (KKT).*

Osservazione 1.2.11. *La condizione di complementary slackness espressa dalla* (1.2.56) *ha un significato cruciale nel trattamento dei vincoli di disuguaglianza $g_i(x) \leq 0$. Infatti tale condizione implica che o il vincolo è attivo, e quindi $g_i(x) = 0$, oppure il moltiplicatore di Lagrange associato è nullo $\lambda_i = 0$.*

Capitolo 2

Caso Deterministico

Nel presente capitolo, il problema del controllo ottimo di un funzionale vincolato da equazioni differenziali alle derivate parziali è trattato dal punto di vista puramente deterministico, in maniera propedeutica allo studio del caso stocastico. L'approccio seguito è in generale quello proposto da Manzoni et al. [14].

2.1 Analisi del problema

2.1.1 Framework generale per funzionali lineari-quadratici

Siano definite la variabile di stato $y \in V$ e una funzione target $z_d \in V$, su uno spazio di Hilbert V. L'obiettivo dell'analisi è la ricerca della soluzione del problema di ottimizzazione di un funzionale $J(y,u)$ rispetto alla variabile di controllo $u \in \mathcal{U}$, definita sullo spazio di Hilbert $\mathcal{U}$. Il funzionale lineare-quadratico da ottimizzare è

$$J(y,u) = \frac{1}{2}||y - z_d||_V^2 + \frac{\nu}{2}||u||_{\mathcal{U}}^2. \tag{2.1.1}$$

Il controllo u influenza la variabile di stato tramite il seguente problema differenziale *Trovare* $y = y(u) \in V$ *tale che*

$$a(y,\phi) = {}_{V'}\langle F, \phi\rangle_V + {}_{V'}\langle Bu, \phi\rangle_V, \qquad \forall \phi \in V. \tag{2.1.2}$$

dove $F \in V'$ e ${}_{V'}\langle\cdot,\cdot\rangle_V$ rappresenta la dualità tra V e V'. Inoltre è definito l'operatore continuo lineare $B : \mathcal{U} \to V'$ tale che

$$||Bu||_{V'} \leq b||u||_{\mathcal{U}}, \qquad \forall u \in \mathcal{U}. \tag{2.1.3}$$

Esistenza e unicità della soluzione $y(u)$ del problema di stato sono garantite dal Teorema di Lax-Milgram, imponendo le condizioni di continuità e coercività della forma bilineare $a(\cdot,\cdot) : V \times V \to \mathbb{R}$

$$|a(\phi,\psi)| \leq C||\phi||_V||\psi||_V, \quad \forall \phi,\psi \in V; \qquad a(\phi,\phi) \geq \alpha||\phi||_V^2, \quad \forall \phi \in V,$$

e di continuità della forma lineare $F \in V'$. Dal Teorema di Lax-Milgram segue quindi la stima

$$||y(u)||_V \leq \frac{1}{\alpha}(||F||_{V'} + b||u||_{\mathcal{U}}). \tag{2.1.4}$$

Alla luce dell'unicità della soluzione del problema di stato, è possibile formulare il funzionale obiettivo *ridotto*

$$\hat{J}(u) = \frac{1}{2}||y(u) - z_d||_V^2 + \frac{\nu}{2}||u||_{\mathcal{U}}^2, \tag{2.1.5}$$

da cui deriva il problema di minimizzazione ridotto

$$\min_{u \in \mathcal{U}} \hat{J}(u). \tag{2.1.6}$$

Nella sezione seguente, questo modello generale di analisi è applicato ad un problema di convezione-diffusione, al fine di ricavare un sistema di condizioni di ottimalità.

2.1.2 Formulazione variazionale

Sia $D \subset \mathbb{R}^d$ un dominio spaziale limitato e si consideri il funzionale

$$J(y, u) = \frac{1}{2}\int_D (y - z_d)^2 d\mathbf{x} + \frac{\nu}{2}\int_D u^2 d\mathbf{x}, \tag{2.1.7}$$

dove $\nu \geq 0$ è detto parametro di regolarizzazione. Nel caso in esame, la variabile di stato rappresenta la soluzione di un problema differenziale di convezione-diffusione-reazione

$$\begin{cases} -\nabla \cdot (\alpha \nabla y) + \mathbf{b} \cdot \nabla y + c\, y = f + u & \text{in } D, \\ y = 0 & \text{su } \partial D. \end{cases} \tag{2.1.8}$$

In generale $\mathbf{b} \in (L^\infty(D))^d$ è un campo di velocità con divergenza $\nabla \cdot \mathbf{b} \in L^\infty(D)$ sul dominio e $f \in L^2(D)$ è una funzione sorgente, mentre $\alpha, c \in L^\infty(D)$.
Per dimostrare buona positura e unicità della soluzione del problema è necessario scegliere uno spazio di funzioni test e considerare la formulazione debole dell'equazione differenziale. In particolare, si considerino funzioni a derivata con modulo quadrato integrabile e a traccia nulla $\phi \in H_0^1(D)$. Pertanto, integrando sul dominio il prodotto tra i termini dell'equazione e le funzioni test si ottiene

$$\int_D (-\nabla \cdot (\alpha \nabla y) + \mathbf{b} \cdot \nabla y + c\, y)\, \phi \, d\mathbf{x} = \int_D (f + u)\, \phi \, d\mathbf{x}.$$

Integrando per parti il primo termine e ricordando che le funzioni test si annullano sul bordo ∂D si ha

$$\int_D -\nabla \cdot (\alpha \nabla y)\, \phi \, d\mathbf{x} = \int_D \alpha \nabla y \cdot \nabla \phi \, d\mathbf{x} - \int_{\partial D} \alpha \nabla y \cdot \mathbf{n}\, \phi \, ds = \int_D \alpha \nabla y \cdot \nabla \phi \, d\mathbf{x}.$$

In conclusione si ottiene il seguente problema in senso debole.

Trovare $y \in H_0^1(D)$ tale che

$$\int_D (\alpha \nabla y \cdot \nabla \phi + (\mathbf{b} \cdot \nabla y)\, \phi + c\, y \phi)\, d\mathbf{x} = \int_D (f + u)\, \phi \, d\mathbf{x}, \qquad \forall \phi \in H_0^1(D). \tag{2.1.9}$$

In alternativa, è possibile riscrivere il problema introducendo la forma bilineare

$$a(y,\phi) = \int_D (\alpha \nabla y \cdot \nabla \phi + (\mathbf{b} \cdot \nabla y)\,\phi + c\,y\phi)\,d\mathbf{x}, \tag{2.1.10}$$

e il funzionale lineare

$$F(\phi) = \int_D (f+u)\,\phi\,d\mathbf{x}. \tag{2.1.11}$$

Quindi si riformula il problema in forma compatta come segue.

Trovare $y \in H_0^1(D)$ *tale che*

$$a(y,\phi) = F(\phi), \qquad \forall \phi \in H_0^1(D). \tag{2.1.12}$$

L'esistenza e l'unicità della soluzione, oltre alla buona positura del problema, sono garantite dal Teorema di Lax-Milgram, per applicare il quale è necessario mostrare la continuità e la coercività della forma bilineare a, oltre alla continuità del funzionale F. Utilizzando le disuguaglianze di Cauchy-Schwarz e di Poincaré, si ha

- Continuità di a

$$\begin{aligned}
&|a(y,\phi)| \le \left|\int_D \alpha \nabla y \cdot \nabla \phi\,d\mathbf{x}\right| + \left|\int_D (\mathbf{b}\cdot\nabla y)\,\phi\,d\mathbf{x}\right| + \left|\int_D c\,y\phi\,d\mathbf{x}\right| \\
&\le ||\alpha||_{L^\infty(D)}||\nabla y||_{L^2(D)}||\nabla\phi||_{L^2(D)} + ||\mathbf{b}||_{L^\infty(D)}||\nabla y||_{L^2(D)}||\phi||_{L^2(D)} + \\
&\quad ||c||_{L^\infty(D)}||y||_{L^2(D)}||\phi||_{L^2(D)} \\
&\le ||\alpha||_{L^\infty(D)}||y||_{H_0^1(D)}||\phi||_{H_0^1(D)} + C_P||\mathbf{b}||_{L^\infty(D)}||y||_{H_0^1(D)}||\phi||_{H_0^1(D)} + \\
&\quad C_P^2||c||_{L^\infty(D)}||y||_{H_0^1(D)}||\phi||_{H_0^1(D)} \\
&\le (||\alpha||_{L^\infty(D)} + C_P||\mathbf{b}||_{L^\infty(D)} + C_P^2||c||_{L^\infty(D)})||y||_{H_0^1(D)}||\phi||_{H_0^1(D)}, \qquad \forall y,\phi \in H_0^1(D).
\end{aligned}$$

- Coercività di a

$$a(y,y) = \int_D \alpha|\nabla y|^2\,d\mathbf{x} + \int_D (\mathbf{b}\cdot\nabla y)\,y\,d\mathbf{x} + \int_D c\,y^2\,d\mathbf{x}.$$

Per il termine convettivo, utilizzando l'identità $(\nabla y)y = \nabla(\frac{1}{2}y^2)$ si ottiene

$$\int_D (\mathbf{b}\cdot\nabla y)y\,d\mathbf{x} = \frac{1}{2}\int_D \mathbf{b}\cdot\nabla(y^2)\,d\mathbf{x} = -\frac{1}{2}\int_D (\nabla\cdot\mathbf{b})\,y^2\,d\mathbf{x}.$$

In definitiva, sostituendo nell'espressione per $a(y,y)$ si ha

$$a(y,y) = \int_D \alpha|\nabla y|^2\,d\mathbf{x} + \int_D \left(-\frac{1}{2}\nabla\cdot\mathbf{b} + c\right) y^2\,d\mathbf{x}.$$

Pertanto, supponendo che

$$\alpha \ge \alpha_0 > 0, \qquad -\frac{1}{2}\nabla\cdot\mathbf{b} + c \ge c_0 > 0,$$

si conclude che la forma bilineare è coerciva

$$a(y,y) \ge \int_D \alpha|\nabla y|^2\,d\mathbf{x} = \alpha_0||y||^2_{H_0^1(D)}, \qquad \forall y \in H_0^1(D).$$

- Continuità di F
  Supponendo $u \in L^2(D)$ si ha

$$\begin{aligned}|F(\phi)| \leq & (||f||_{L^2(D)} + ||u||_{L^2(D)})||\phi||_{L^2(D)} \\ \leq & C_P(||f||_{L^2(D)} + ||u||_{L^2(D)})||\phi||_{H_0^1(D)}, \quad \forall \phi \in H_0^1(D).\end{aligned}$$

Sulla base di queste supposizioni, è possibile enunciare la seguente proposizione.

Proposizione 2.1.1. *Per ogni generica $u \in L^2(D)$, il problema* (2.1.12) *ha soluzione unica $y = y(u) \in H_0^1(D)$. Inoltre vale la stima di limitatezza*

$$||\nabla y||_{L^2(D)} \leq \frac{C_P}{\alpha_0}(||f||_{L^2(D)} + ||u||_{L^2(D)}), \tag{2.1.13}$$

dove la costante $C_P > 0$, detta costante di Poincaré, è definita dalla relazione tra norme espressa dalla (1.2.30).

2.1.3 Derivazione delle condizioni di ottimalità

In ragione del risultato di unicità ottenuto nella sezione precedente è naturale riscrivere il funzionale obiettivo esplicitando la dipendenza della variabile di stato dal controllo $u \in \mathcal{U} \equiv L^2(D)$, ottenendo il problema di ottimizzazione

$$u^\star = \min_{u \in L^2(D)} J(y(u), u) = \min_{u \in L^2(D)} \frac{1}{2}\int_D (y(u) - z_d)^2 d\mathbf{x} + \frac{\nu}{2}\int_D u^2 d\mathbf{x}. \tag{2.1.14}$$

Il problema in esame risulta essere lineare-quadratico, quindi è garantita la forte convessità del funzionale obiettivo J. Pertanto il problema ammette una ed un'unica soluzione, ottenuta imponendo $J'(u) = 0$. Tale asserzione è legittimata dal seguente risultato, generalizzato al caso di $u^\star \in \mathcal{U}_{ad} \subset \mathcal{U}$.

Teorema 2.1.1. *Sia $\mathcal{U}_{ad}$ un sottospazio non vuoto, chiuso e convesso di uno spazio di Hilbert $\mathcal{U}$. Allora, se $\nu > 0$ il problema di ottimizzazione* (2.1.14) *ha una ed un'unica soluzione $u^\star$.*

Dimostrazione. Definito il *feasible set* $\mathcal{U}_{ad}$, poichè $\hat{J} \geq 0$ e $\mathcal{U}_{ad}$ è non vuoto, allora esiste

$$\hat{J}^\star = \inf_{u \in \mathcal{U}_{ad}} \hat{J}(u),$$

e di conseguenza è possibile definire una successione minimizzante $\{u_k\}_{k \in \mathbb{N}} \subset \mathcal{U}_{ad}$ tale che

$$\lim_{k \to \infty} \hat{J}(u_k) = \hat{J}^\star.$$

Per definizione del funzionale $\hat{J}$, si ha

$$\frac{\nu}{2}||u_k||_{\mathcal{U}}^2 \leq \hat{J}(u_k) \leq C, \quad \forall k \in \mathbb{N},$$

da cui la limitatezza della successione minimizzante $||u_k||_{\mathcal{U}} \leq \sqrt{2C/\nu} =: K$.
Poichè in ogni spazio di Banach riflessivo si ha che ogni successione limitata ammette una sottosuccessione debolmente convergente, è possibile definire $\{u_{k_i}\}_{k\in\mathbb{N}}$ tale che

$$u_{k_i} \xrightarrow{i\to\infty} u^\star.$$

L'esistenza di questa sottosuccessione e la sua convergenza debole ad un elemento $u^\star \in \mathcal{U}_{ad}$ è garantita dal Teorema di Banach-Alaoglu. Infatti, considerata la palla chiusa, limitata e convessa

$$B_{\mathcal{U}}(K) = \{u \in \mathcal{U} \,|\, ||u||_{\mathcal{U}} \leq K\}, \tag{2.1.15}$$

si ha che, data la riflessività dello spazio di Hilbert $\mathcal{U}$, essa è compatta rispetto alla topologia debole di $\mathcal{U}$. Inoltre, è possibile generalizzare al sottospazio $\mathcal{U}_{ad}$, poichè essendo chiuso e convesso, è anche debolmente chiuso.
L'ottimalità della soluzione $u^\star$ è garantita dalla semicontinuità inferiore del funzionale $\hat{J}$,

$$\hat{J}(u^\star) \leq \hat{J}^\star = \liminf_{i\to\infty} \hat{J}(u_{k_i}).$$

L'unicità della soluzione ottenuta segue infine dal fatto che il funzionale obiettivo risulta strettamente convesso per $\nu > 0$. □

Dimostrate esistenza e unicità del controllo ottimo, è necessario determinare le condizioni di ottimalità relative al problema in esame. Indicata con $y_0(u)$ la soluzione del problema di stato con forzante nulla, si ha che, in generale $y(u) = y_0(u) + y(0)$. Per ottenere delle condizioni di ottimalità che permettano di determinare $u^\star$ e la relativa funzione di stato $y^\star = y(u^\star)$, si definisca la forma bilineare simmetrica, continua e coerciva rispetto alla norma di $L^2(D)$

$$q(u,v) = \int_D (y_0(u)y_0(v) + \nu uv)d\mathbf{x}, \tag{2.1.16}$$

e il funzionale lineare e continuo in $L^2(D)$

$$Lu = \int_D y_0(u)(z_d - y(0))d\mathbf{x}. \tag{2.1.17}$$

Questa rappresentazione permette di riscrivere il funzionale obiettivo nella forma

$$u^\star = \min_u J(u) = \min_u \frac{1}{2}q(u,u) - Lu + c, \tag{2.1.18}$$

dove $c = 1/2\int_D (y(0) - z_d)^2 d\mathbf{x}$ è un termine costante, irrilevante nell'ottimizzazione.
Si confrontino ora i valori di J per $u = u^\star + \epsilon v$ e $u = u^\star$, con $\epsilon > 0$.
Svolgendo i prodotti scalari si ha

$$\frac{J(u^\star + \epsilon v) - J(u^\star)}{\epsilon} = q(u^\star, v) - Lv + \frac{1}{2}\epsilon q(v,v).$$

Facendo tendere $\epsilon \to 0$, risulta in generale che

$$\frac{d}{d\epsilon} J(u^\star + \epsilon v)\Big|_{\epsilon=0} = q(u^\star, v) - Lv = 0, \qquad \forall v \in L^2(D). \tag{2.1.19}$$

Poichè il primo membro dell'equazione è un elemento dello spazio duale di $L^2(D)$, detto derivata nel senso di Gateaux del funzionale J, è possibile concludere che, definita una direzione v, vale la seguente espressione

$$\frac{d}{d\epsilon}J(u^\star+\epsilon v)\Big|_{\epsilon=0} = q(u^\star,v) - Lv = J'(u^\star)v = 0, \qquad \forall v \in L^2(D). \tag{2.1.20}$$

Esplicitando l'ultima uguaglianza, nota come *equazione di Eulero*, per il caso in esame

$$J'(u^\star)v = \int_D ((y^\star - z_d)y_0(v) + \nu u^\star v)\, d\mathbf{x} = 0, \qquad \forall v \in L^2(D). \tag{2.1.21}$$

In conclusione vale il seguente teorema.

Teorema 2.1.2. *Per il problema di ottimizzazione* (2.1.1)*, vincolato da* (2.1.2)*, esiste unica la soluzione ottima $u^\star \in L^2(D)$. Inoltre, tale soluzione è ottima $\Leftrightarrow$ vale l'equazione di Eulero.*

Tuttavia, poichè la ricerca di un'espressione analitica per $J'(u^\star)$ è complessa e dispendiosa, una strategia efficace per determinare le condizioni di ottimalità consiste nel definire una variabile aggiunta $p \in H_0^1(D)$, e scrivere la Lagrangiana $\mathcal{L}$

$$\mathcal{L}(y,u,p) = J(y,u) - a(y,p) + (f+u,p)_{L^2(D)}. \tag{2.1.22}$$

Analogamente alla definizione di J', calcolando le derivate parziali della Lagrangiana rispetto alle tre variabili (y,u,p) si ottengono rispettivamente

$$\begin{aligned}
\mathcal{L}'_p(y,u,p)\phi &= \frac{d}{d\epsilon}\mathcal{L}(y^\star,u^\star,p^\star+\epsilon\phi)\Big|_{\epsilon=0} \qquad \forall \phi \in H_0^1(D),\\
&= -a(y^\star,\phi) + (f+u,\phi)_{L^2(D)} = 0. \qquad \textit{Equazione di stato.}
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}'_y(y,u,p)\psi &= \frac{d}{d\epsilon}\mathcal{L}(y^\star+\epsilon\psi,u^\star,p^\star)\Big|_{\epsilon=0} \qquad \forall \psi \in H_0^1(D),\\
&= J'_y(y^\star,u^\star)\psi - a(\psi,p^\star)\\
&= -a^*(p^\star,\psi) + (y^\star - z_d,\psi)_{L^2(D)} = 0. \qquad \textit{Equazione aggiunta.}
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}'_u(y,u,p)v &= \frac{d}{d\epsilon}\mathcal{L}(y^\star,u^\star+\epsilon v,p^\star)\Big|_{\epsilon=0} \qquad \forall v \in L^2(D),\\
&= J'_u(y^\star,u^\star)v + (v,p^\star)_{L^2(D)}\\
&= (v,\nu u^\star)_{L^2(D)} + (v,p^\star)_{L^2(D)} = 0. \qquad \textit{Equazione di Eulero.}
\end{aligned}$$

In definitiva si ottiene il sistema di condizioni

$$\begin{cases}
a(y^\star,\phi) = (f+u,\phi)_{L^2(D)},\\
a^*(p^\star,\psi) = (y^\star - z_d,\psi)_{L^2(D)},\\
(v,\nu u^\star)_{L^2(D)} + (v,p^\star)_{L^2(D)} = 0.
\end{cases} \tag{2.1.23}$$

Esplicitando i prodotti scalari dell'equazione di Eulero si ha

$$J'(u^\star)v = \int_D (\nu u^\star + p^\star)\, d\mathbf{x} = 0, \qquad \forall v \in L^2(D), \tag{2.1.24}$$

che implica

$$\nu u^\star + p^\star = 0, \qquad q.o.\ in\ D. \tag{2.1.25}$$

Infine, ricordando la definizione di operatore di Riesz, si può introdurre il gradiente del funzionale $\nabla J(u^\star)$ come il rappresentante di Riesz associato a $J'(u^\star)$

$$\nabla J(u^\star) = \nu u^\star + p^\star. \tag{2.1.26}$$

Il gradiente dell'operatore fornisce la direzione di massima ascesa nell'ottimizzazione di J, pertanto questa definizione sarà utile per la costruzione di metodi iterativi di minimizzazione per la trattazione numerica del problema, quali il metodo *steepest descent* o il metodo *BFGS*.

2.2 Discretizzazione

In questa sezione viene introdotta una discretizzazione agli elementi finiti per la risoluzione di problemi differenziali quali (2.1.2). Inoltre, vengono presentati due differenti approcci per l'ottimizzazione del funzionale obiettivo J: nel primo (*optimize then discretize*) la discretizzazione viene applicata alle condizioni di ottimalità ricavate nella sezione precedente, mentre nella seconda (*discretize then optimize*) si discretizzano direttamente funzionale obiettivo e vincolo differenziale, ottenendo un problema di ottimizzazione discreto.
I due approcci proposti, nel caso in cui si adotti la stessa discretizzazione per lo spazio di definizione delle variabili di stato e aggiunta e per quello della variabile di controllo, generano equivalenti sistemi lineari algebrici di condizioni di ottimalità, per il caso lineare-quadratico.

2.2.1 Caratterizzazione degli elementi finiti

Si consideri una partizione del dominio D in un set di triangoli tali che la loro intersezione consista in un punto o in un singolo lato. L'insieme che contiene tali elementi viene detto *triangolazione* del dominio e indicato con $\mathcal{T}$. La dimensione caratteristica di ogni elemento della triangolazione, definita come la lunghezza del suo lato maggiore h_e, è una misura della finezza della partizione

$$h = \max_{e \in \mathcal{T}} h_e. \tag{2.2.1}$$

Sia definito il triangolo di riferimento

$$\hat{e} = \{\hat{\mathbf{x}} \subset \mathbb{R}^2 \,|\, 0 \leq \hat{x}, \hat{y} \leq 1 \cap 0 \leq 1 - \hat{x} - \hat{y} \leq 1\}. \tag{2.2.2}$$

Ogni elemento contenuto nella triangolazione può essere rappresentato come una trasformazione lineare del triangolo di riferimento, tramite la mappa affine

$$\mathcal{F}(\hat{x}, \hat{y}) = \mathbf{a}_1 \hat{\phi_1}(\hat{x}, \hat{y}) + \mathbf{a}_2 \hat{\phi_2}(\hat{x}, \hat{y}) + \mathbf{a}_3 \hat{\phi_3}(\hat{x}, \hat{y}), \tag{2.2.3}$$

dove $\mathbf{a}_i$, $i = 1,2,3$ sono i vertici del generico triangolo, mentre le funzioni $\hat{\phi}_i$, $i = 1,2,3$, sono definite come gli elementi della base Lagrangiana dello spazio dei polinomi di primo grado $\mathbb{P}^1(D)$

$$\begin{cases} \hat{\phi}_1(\hat{x},\hat{y}) = \hat{x}, \\ \hat{\phi}_2(\hat{x},\hat{y}) = \hat{y}, \\ \hat{\phi}_3(\hat{x},\hat{y}) = 1 - \hat{x} - \hat{y}. \end{cases} \tag{2.2.4}$$

Costruita la triangolazione, è possibile applicare il metodo di Galerkin per discretizzare il problema variazionale (2.1.9), considerando il sottospazio di dimensione finita $V_h \subset V \equiv H_0^1(D)$, definito come lo spazio vettoriale generato dalle funzioni di base Lagrangiane ϕ_j

$$V_h = \mathrm{Span}\{\phi_j, j = 1 \dots N_{\mathrm{dof}}\}, \tag{2.2.5}$$

dove N_{dof} è il numero di gradi di libertà globali della triangolazione (in questo caso, pari al numero di nodi interni).
Ricordando la definizione della mappa $\mathcal{F}$, è possibile mappare le restrizioni delle basi al generico elemento $\phi_j|_e = \phi_{\hat{\jmath}_e}$ sull'elemento di riferimento

$$\phi_{\hat{\jmath}_e}(\mathcal{F}(\hat{x},\hat{y})) = \hat{\phi}_{\hat{\jmath}_e}(\hat{x},\hat{y}), \qquad \forall (\hat{x},\hat{y}) \in \hat{e},$$

dove l'indice $\hat{\jmath}_e \in \{1,\dots,N_e\}$ è riferito agli N_e gradi di libertà locali del singolo elemento della triangolazione.
Per la discretizzazione dei gradienti delle funzioni di base si consideri una riformulazione della mappa $\mathcal{F}$

$$\mathcal{F}(\hat{\mathbf{x}}) = \mathbf{a}_3 + \mathbf{B}\hat{\mathbf{x}},$$

dove la matrice $\mathbf{B}$, che coincide con la matrice Jacobiana $\mathbf{J}_{\mathcal{F}}$ della trasformazione, contiene come colonne i vettori $\mathbf{a}_i - \mathbf{a}_3$, $i = 1,2$. In componenti

$$\mathbf{B} = \begin{bmatrix} x_1 - x_3 & x_2 - x_3 \\ y_1 - y_3 & y_2 - y_3 \end{bmatrix}.$$

Pertanto, applicando la derivazione della funzione composta e trasponendo entrambi i membri, si ha che

$$\begin{aligned} \nabla \hat{\phi}_{\hat{\jmath}_e}(\hat{x},\hat{y}) &= \mathbf{J}_{\mathcal{F}}(\hat{x},\hat{y})^T \nabla \phi_{\hat{\jmath}_e}(\mathcal{F}(\hat{x},\hat{y})) \\ &= \mathbf{B}^T \nabla \phi_{\hat{\jmath}_e}(\mathcal{F}(\hat{x},\hat{y})), \end{aligned} \tag{2.2.6}$$

da cui segue la relazione inversa, utile nella discretizzazione della forma bilineare $a(u,v)$

$$\nabla \phi_{\hat{\jmath}_e}(\mathcal{F}(\hat{x},\hat{y})) = \mathbf{B}^{-T} \nabla \hat{\phi}_{\hat{\jmath}_e}(\hat{x},\hat{y}). \tag{2.2.7}$$

Osservazione 2.2.1. *Nelle espressioni* (2.2.6) *e* (2.2.7)*, il gradiente è calcolato rispetto alle variabili* (x,y) *del generico elemento quando applicato ad una funzione* $\phi(x,y)$*, mentre è calcolato rispetto a* $(\hat{x},\hat{y})$ *nel caso di una funzione definita sull'elemento di riferimento* $\hat{\phi}(\hat{x},\hat{y})$*.*

2.2.2 Strategia *optimize then discretize*

Come anticipato, scegliendo un approccio di tipo *optimize then discretize* la strategia prevede di ottimizzare il problema a livello continuo, per poi procedere alla discretizzazione. Si richiami il sistema di condizioni di ottimalità (2.1.23), cambiando il segno al secondo membro dell'equazione aggiunta (e di conseguenza al secondo termine dell'equazione di Eulero) per convenienza computazionale. Si definisce il problema variazionale seguente.
Trovare $y \in V \equiv H_0^1(D)$, $p \in V \equiv H_0^1(D)$ *e* $u \in \mathcal{U} \equiv L^2(D)$ *tali che*

$$\begin{cases} a(y,\phi) = (f+u,\phi)_{L^2(D)} & \forall \phi \in V, \\ a^*(p,\psi) = -(y-z_d,\psi)_{L^2(D)} & \forall \psi \in V, \\ (\nu u - p, v)_{L^2(D)} = 0 & \forall v \in \mathcal{U}. \end{cases} \tag{2.2.8}$$

Per discretizzare il sistema si introducono due spazi vettoriali finito dimensionali $V_h \subset V$ e $\mathcal{U}_h \subset \mathcal{U}$ aventi basi $\{\phi_j,\ j = 1,\dots,N_v\}$ e $\{\psi_j,\ j = 1,\dots,N_u\}$ dove N_v e N_u sono le rispettive dimensioni finite degli spazi. Esprimendo il sistema di condizioni di ottimalità in virtù dell'approssimazione degli spazi di riferimento si ottiene il seguente problema variazionale discreto.
Trovare $y_h \in V_h$, $p_h \in V_h$ *e* $u_h \in \mathcal{U}_h$ *tali che*

$$\begin{cases} a(y_h,\phi_h) = (f+u_h,\phi_h)_{L^2(\Omega)} & \forall \phi_h \in V_h, \\ a^*(p_h,\psi_h) = -(y_h-z_d,\psi_h)_{L^2(\Omega)} & \forall \psi_h \in V_h, \\ (\nu u_h - p_h, v_h)_{L^2(D)} = 0 & \forall v_h \in \mathcal{U}_h. \end{cases} \tag{2.2.9}$$

Poichè le funzioni incognite del problema sono elementi degli spazi V_h e $\mathcal{U}_h$, possono essere espresse come combinazioni lineari degli elementi delle rispettive basi. Utilizzando quindi la triangolazione e i relativi strumenti computazionali presentati è possibile riformulare il problema variazionale discreto come un sistema lineare di equazioni algebriche.

In particolare, procedendo termine a termine:

- Espansione di y_h, p_h e u_h

$$\begin{aligned} y_h &= \sum_{k=1}^{N_v} y_{h,k}\phi_k = \sum_{e\in\mathcal{T}}\sum_{\hat{k}=1}^{N_e} y_{k_e(\hat{k})}\phi_{\hat{k}}, \\ p_h &= \sum_{k=1}^{N_v} p_{h,k}\phi_k = \sum_{e\in\mathcal{T}}\sum_{\hat{k}=1}^{N_e} p_{k_e(\hat{k})}\phi_{\hat{k}}, \\ u_h &= \sum_{k=1}^{N_u} u_{h,k}\psi_k = \sum_{e\in\mathcal{T}}\sum_{\hat{k}=1}^{N_e} u_{k_e(\hat{k})}\psi_{\hat{k}}. \end{aligned}$$

- Forme bilineari $a(y_h, \phi_j)$, e $a^*(p_h, \phi_j)$

$$\begin{aligned}
a(y_h, \phi_j) &= \int_D \nabla y_h \, \nabla \phi_j \, d\mathbf{x} + \int_D (\mathbf{b} \cdot \nabla y_h) \, \phi_j \, d\mathbf{x} \\
&= \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \Big(\int_e \nabla \phi_{\hat{k}} \, \nabla \phi_{\hat{\jmath}} \, dxdy + \int_e \mathbf{b} \cdot \nabla \phi_{\hat{k}} \, \phi_{\hat{\jmath}} \, dxdy \Big) y_{k_e(\hat{k})} \\
&= \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \Big(2|e| \int_{\hat{e}} \nabla \hat{\phi}_{\hat{k}}(\hat{x}, \hat{y}) \, \mathbf{B}^{-1} \mathbf{B}^{-T} \, \nabla \hat{\phi}_{\hat{\jmath}}(\hat{x}, \hat{y}) \, d\hat{x} d\hat{y} + \\
&\quad + 2|e| \int_{\hat{e}} \mathbf{b}^T (\mathcal{F}(\hat{x}, \hat{y})) \, \mathbf{B}^{-T} \, \nabla \hat{\phi}_{\hat{k}}(\hat{x}, \hat{y}) \, \hat{\phi}_{\hat{\jmath}}(\hat{x}, \hat{y}) d\hat{x} d\hat{y} \Big) y_{k_e(\hat{k})} \\
&\approx \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \Big(2|e| \sum_{q=1}^{N_q} \omega_q \nabla \hat{\phi}_{\hat{k}}(\hat{x}_q, \hat{y}_q) \, \mathbf{B}^{-1} \mathbf{B}^{-T} \, \nabla \hat{\phi}_{\hat{\jmath}}(\hat{x}_q, \hat{y}_q) + \\
&+ 2|e| \sum_{q=1}^{N_q} \omega_q \, \mathbf{b}^T (\mathcal{F}(\hat{x}_q, \hat{y}_q)) \, \mathbf{B}^{-T} \, \nabla \hat{\phi}_{\hat{k}}(\hat{x}_q, \hat{y}_q) \, \hat{\phi}_{\hat{\jmath}}(\hat{x}_q, \hat{y}_q) \Big) y_{k_e(\hat{k})},
\end{aligned}$$

  dove $\mathcal{T}_j = \{e \in \mathcal{T} \,|\, \phi_j \not\equiv 0 \text{ in } e\}$ è la sotto-triangolazione relativa al supporto di ϕ_j. Nell'ultimo passaggio viene applicata una formula di quadratura per approssimare l'integrale sull'elemento di riferimento. Per una generica funzione f

$$\int_{\hat{e}} f(\hat{x}, \hat{y}) \, d\hat{x} d\hat{y} \approx \sum_{q=1}^{N_q} \omega_q f(\hat{x}_q, \hat{y}_q). \tag{2.2.10}$$

  In conclusione, introducendo la matrice di rigidezza $(\mathbb{A})_{kj} = a(\phi_k, \phi_j)$, $j, k = 1, \dots N_v$, il primo termine dell'equazione di stato si scrive in forma discreta come

$$\mathbb{A}\mathbf{y} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} (\mathbb{A})_{\hat{k}\hat{\jmath}} y_{k_e(\hat{k})}. \tag{2.2.11}$$

  Analogamente, il primo termine dell'equazione aggiunta diventa

$$\mathbb{A}^T\mathbf{p} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} (\mathbb{A})^T_{\hat{k}\hat{\jmath}} p_{k_e(\hat{k})}. \tag{2.2.12}$$

- Forme lineari $F(\phi_j) = (f, \phi_j)_{L^2(D)}$ e $Z_d(\phi_j) = (z_d, \phi_j)_{L^2(D)}$, $j = 1, \dots, N_u$

$$\begin{aligned}
(f, \phi_j)_{L^2(D)} &= \int_D f \phi_j \, d\mathbf{x} = \sum_{e \in \mathcal{T}_j} \int_e f \, \phi_{\hat{\jmath}} \, dxdy \\
&= \sum_{e \in \mathcal{T}_j} 2|e| \int_{\hat{e}} f(\mathcal{F}(\hat{x}, \hat{y})) \; \hat{\phi}_{\hat{\jmath}}(\hat{x}, \hat{y}) \, d\hat{x} d\hat{y} \\
&\approx \sum_{e \in \mathcal{T}_j} 2|e| \sum_{q=1}^{N_q} \omega_q \, f(\mathcal{F}(\hat{x}_q, \hat{y}_q)) \; \hat{\phi}_{\hat{\jmath}}(\hat{x}_q, \hat{y}_q).
\end{aligned}$$

Introducendo il vettore $\mathbf{b}_j = (f, \phi_j)_{L^2(D)}$, e, procedendo in maniera analoga per Z_d, il vettore $\mathbf{z_d}_j = (z_d, \phi_j)_{L^2(D)}$ si ottengono i termini

$$\mathbf{f} = \sum_{e \in \mathcal{T}_j} \mathbf{f}_{\hat{\jmath}}, \qquad \mathbf{z_d} = \sum_{e \in \mathcal{T}_j} \mathbf{z_d}_{\hat{\jmath}}. \tag{2.2.13}$$

- Prodotti scalari $(\phi_k, \phi_j)_{L^2(D)}$ e $(\psi_k, \psi_j)_{L^2(D)}$

$$\begin{aligned}(\phi_k, \phi_j)_{L^2(D)} &= \int_D \phi_k \phi_j \, d\mathbf{x} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \int_e \phi_{\hat{k}} \, \phi_{\hat{\jmath}} \, dxdy \\ &= \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} 2|e| \int_{\hat{e}} \hat{\phi}_{\hat{k}}(\hat{x}, \hat{y}) \, \hat{\phi}_{\hat{\jmath}}(\hat{x}, \hat{y}) \, d\hat{x}d\hat{y} \\ &= \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} 2|e| \sum_{q=1}^{N_q} \omega_q \, \hat{\phi}_{\hat{k}}(\hat{x}_q, \hat{y}_q) \, \hat{\phi}_{\hat{\jmath}}(\hat{x}_q, \hat{y}_q).\end{aligned}$$

Indicando la matrice di massa relativa a V_h con $(\mathbb{M})_{kj} = (\phi_k, \phi_j)_{L^2(D)}$, $j, k = 1, \dots N_u$ e con $(\mathbb{N})_{kj} = (\psi_k, \psi_j)_{L^2(D)}$, $j, k = 1, \dots N_v$ quella relativa a $\mathcal{U}_h$, si ottengono le espressioni

$$\mathbb{M} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} (\mathbb{M})_{\hat{k}\hat{\jmath}}, \qquad \mathbb{N} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} (\mathbb{N})_{\hat{k}\hat{\jmath}}. \tag{2.2.14}$$

- Prodotto scalare $(u_h, \phi_j)_{L^2(D)}$

$$\begin{aligned}(u_h, \phi_j)_{L^2(D)} &= \int_D u_h \phi_j \, d\mathbf{x} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \left(\int_e \psi_{\hat{k}} \, \phi_{\hat{\jmath}} \, dxdy \right) u_{k_e(\hat{k})} \\ &= \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \left(2|e| \int_{\hat{e}} \hat{\psi}_{\hat{k}}(\hat{x}, \hat{y}) \, \hat{\phi}_{\hat{\jmath}}(\hat{x}, \hat{y}) \, d\hat{x}d\hat{y} \right) u_{k_e(\hat{k})} \\ &= \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} \left(2|e| \sum_{q=1}^{N_q} \omega_q \, \hat{\psi}_{\hat{k}}(\hat{x}_q, \hat{y}_q) \, \hat{\phi}_{\hat{\jmath}}(\hat{x}_q, \hat{y}_q) \right) u_{k_e(\hat{k})}.\end{aligned}$$

In maniera più compatta, introdotta la matrice di controllo $(\mathbb{B})_{jk} = (\psi_k, \phi_j)_{L^2(D)}$

$$\mathbb{B}\mathbf{u} = \sum_{e \in \mathcal{T}_j} \sum_{\hat{k}=1}^{N_e} (\mathbb{B})_{\hat{\jmath}\hat{k}} u_{k_e(\hat{k})}. \tag{2.2.15}$$

Alla luce della formulazione discreta presentata, è possibile costruire un sistema di equazioni algebriche per risolvere computazionalmente il problema variazionale (2.1.9).

$$\begin{cases} \mathbb{A}\mathbf{y} = \mathbf{f} + \mathbb{B}\mathbf{u}, \\ \mathbb{A}^T\mathbf{p} = -\mathbb{M}(\mathbf{y} - \mathbf{z_d}), \\ \nu\mathbb{N}\mathbf{u} - \mathbb{B}^T\mathbf{p} = 0. \end{cases} \tag{2.2.16}$$

In forma matriciale

$$\begin{bmatrix} \mathbb{M} & 0 & \mathbb{A}^T \\ 0 & \nu\mathbb{N} & -\mathbb{B}^T \\ \mathbb{A} & -\mathbb{B} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbb{M}\,\mathbf{z_d} \\ 0 \\ \mathbf{f} \end{bmatrix}. \tag{2.2.17}$$

Osservazione 2.2.2. *Nel caso particolare in cui il controllo è distribuito sull'intero dominio D, e la discretizzazione operata coinvolge la stessa base per gli spazi V_h e $\mathcal{U}_h$, si osserva come le matrici di massa e quella di controllo siano tutte coincidenti $\mathbb{B} = \mathbb{M} = \mathbb{N}$.*

2.2.3 Strategia *discretize then optimize*

In alternativa, è possibile optare per la strategia *discretize then optimize*, secondo la quale la discretizzazione avviene a livello del problema di controllo (2.1.1), (2.1.2), al fine di costruire direttamente un sistema di condizioni di ottimalità su spazi finito-dimensionali. Si consideri pertanto il problema di minimizzazione discreto

$$\begin{aligned} \min_{u_h} J_h(y_h, u_h) &= \frac{1}{2}\int_D (y_h - z_d)^2 d\mathbf{x} + \frac{\nu}{2}\int_D u_h^2 d\mathbf{x} \\ \text{s.t.} \quad & a(y_h, \phi_h) = (f + u_h, \phi_h)_{L^2(D)}, \quad \forall \phi_h \in V_h, \\ & y_h \in V_h, \quad u_h \in \mathcal{U}_h. \end{aligned} \tag{2.2.18}$$

Applicando la discretizzazione di Galerkin agli elementi finiti presentata, è possibile riformulare il problema in termini matriciali.

$$\begin{aligned} \min_{\mathbf{u}} J_h(\mathbf{y}(\mathbf{u}), \mathbf{u}) &= \frac{1}{2}(\mathbf{y}(\mathbf{u}) - \mathbf{z_d})^T \mathbb{M}(\mathbf{y}(\mathbf{u}) - \mathbf{z_d}) + \frac{\nu}{2}\mathbf{u}^T \mathbb{N}\mathbf{u} \\ \text{s.t.} \quad & \mathbb{A}\mathbf{y} = \mathbf{f} + \mathbb{B}\mathbf{u}, \\ & \mathbf{y} \in \mathbb{R}^{N_v}, \quad \mathbf{u} \in \mathbb{R}^{N_u}. \end{aligned} \tag{2.2.19}$$

Il seguente risultato dimostra che il problema discreto ottenuto è equivalente al sistema algebrico di condizioni di ottimalità (2.1.23).

Teorema 2.2.1. *Siano definite le matrici $\mathbb{A} \in \mathbb{R}^{N_v \times N_v}$ non singolare e $\mathbb{B} \in \mathbb{R}^{N_v \times N_u}$ di rango massimo. Per $\nu > 0$, si consideri il problema di controllo ottimo*

$$\min_{\mathbf{u} \in \mathbb{R}^{N_u}} J_h(\mathbf{y}(\mathbf{u}), \mathbf{u}) = \frac{1}{2}|\mathbf{y}(\mathbf{u}) - \mathbf{z_d}|^2 + \frac{\nu}{2}|\mathbf{u}|^2. \tag{2.2.20}$$

Indicato con $\hat{u}$ il minimo e con $\hat{y} = y(\hat{u})$ il corrispondente stato ottimale, allora $\exists \hat{p} = p(\hat{u})$ tale che è verificato il seguente sistema di condizioni di ottimalità del primo ordine

$$\begin{cases} \mathbb{A}\hat{\mathbf{y}} = \mathbf{f} + \mathbb{B}\hat{\mathbf{u}}, \\ \mathbb{A}^T\hat{\mathbf{p}} = \hat{\mathbf{y}} - \mathbf{z_d}, \\ (\mathbb{B}^T\hat{\mathbf{p}} + \nu\hat{\mathbf{u}}, \mathbf{v} - \hat{\mathbf{u}}) \geq 0, \quad \forall \mathbf{v} \in \mathbb{R}^{N_u}. \end{cases} \tag{2.2.21}$$

Osservazione 2.2.3. *L'unica differenza tra il problema in esame e il teorema enunciato è la scelta della matrice di massa: per ottenere* (2.2.16) *si considera in generale una $\mathbb{M}$ eventualmente diversa da $\mathbb{M} = \mathbb{I}$.*

2.3 Tecniche di risoluzione numerica

In questa terza sezione saranno analizzate alcune delle strategie computazionali applicate per risolvere il problema di ottimizzazione vincolata analizzato. Sotto le opportune ipotesi evidenziate nella sezione precedente, gli approcci presentati conducono allo stesso sistema lineare con struttura di tipo punto-sella. Gli algoritmi di risoluzione si dividono in base all'obiettivo computazionale perseguito, oltre che alla consueta distinzione tra metodi diretti e iterativi. In particolare, è possibile risolvere in maniera diretta, tramite tecniche di precondizionamento, il sistema algebrico, ottenendo contemporaneamente la soluzione del problema di stato, la variabile aggiunta e il controllo ottimo, oppure costruire iterativamente un'approssimazione del controllo ottimo a partire da una stima iniziale. La presente trattazione sarà principalmente incentrata su quest'ultimo obiettivo, proponendo alcuni metodi iterativi noti in letteratura.

2.3.1 Metodi iterativi

Una possibile strategia per ottenere una soluzione approssimata al problema (2.2.16) è l'applicazione di metodi di discesa tipici dell'ottimizzazione numerica non vincolata. Seguendo questo tipo di approccio, il funzionale obiettivo viene minimizzato solamente rispetto alla variabile di controllo, sostituendo l'espressione della variabile di stato $\mathbf{y} = \mathbf{y}(\mathbf{u})$, ottenuta come soluzione dell'equazione differenziale.
Per costruire il generico metodo di discesa, si consideri una stima iniziale della variabile di controllo $\mathbf{u}_0 \in \mathbb{R}^{N_u}$, dalla quale costruire una sequenza $\{\mathbf{u}_k\}$ secondo la formula ricorsiva

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \tau_k \mathbf{d}_k. \tag{2.3.1}$$

La definizione della direzione di discesa $\mathbf{d}_k$ varia a seconda del particolare metodo iterativo scelto: alcune possibili alternative sono di seguito presentate, quali il metodo *steepest descent*, il metodo del *gradiente coniugato* e il metodo *BFGS*. Il parametro τ_k, detto *step-size*, è determinato implementando la tecnica del *backtracking*, che consiste nel ridurre tale parametro finchè la seguente condizione (detta condizione di Armijo) non è soddisfatta.

$$J_h(\mathbf{u}_k + \tau_k \mathbf{d}_k) < J_h(\mathbf{u}_k) - \sigma \tau_k \nabla J_h(\mathbf{u}_k)^T \mathbf{d}_k, \tag{2.3.2}$$

dove $\sigma \in (10^{-5}, 10^{-1})$ è un parametro fissato.
Tuttavia, nel caso particolare in esame, il funzionale da minimizzare è lineare-quadratico e pertanto è possibile determinare analiticamente il valore ottimale della step-size τ_k

$$\begin{aligned} \tau_k =& \arg\min_{\tau>0} J_h(\mathbf{u}_k + \tau \mathbf{d}_k) \\ =& - \frac{\nu(\mathbf{u}_k, \mathbf{d}_k)_{\mathbb{N}} + (\mathbf{y}_k - \mathbf{z}_{\mathbf{d}}, \mathbf{y}(\mathbf{d}_k) - \mathbf{y}(\mathbf{0}))_{\mathbb{M}}}{||\mathbf{y}(\mathbf{d}_k) - \mathbf{y}(\mathbf{0})||^2_{\mathbb{M}} + \nu ||\mathbf{d}_k||^2_{\mathbb{N}}} \\ =& - \frac{\nabla J_h(\mathbf{u}_k)^T \mathbf{d}_k}{\mathbf{d}_k \mathbb{H} \mathbf{d}_k}. \end{aligned} \tag{2.3.3}$$

Metodo *steepest descent*

Per implementare il metodo *steepest descent*, si opta per il gradiente del funzionale J_h cambiato di segno come direzione di discesa nella formula ricorsiva (2.3.1)

$$\mathbf{d}_k = -\nabla J_h(\mathbf{u}_k), \tag{2.3.4}$$

Chiaramente, nel problema in esame, la valutazione del gradiente del funzionale obiettivo richiede ad ogni iterazione del metodo sia la soluzione dell'equazione di stato $\mathbf{y} = \mathbf{y}(\mathbf{u}_k)$ che dell'equazione aggiunta $\mathbf{p} = \mathbf{p}(\mathbf{u}_k)$, in quanto, sostituendo la prima nell'espressione (2.1.23) e derivando si ha

$$\nabla J_h(\mathbf{u}_k) = \nu \mathbb{N}\mathbf{u}_k - \mathbb{B}^T\mathbf{p}(\mathbf{u}_k). \tag{2.3.5}$$

Questo metodo pertanto rischia di essere molto dispendioso dal punto di vista computazionale, soprattutto quando le dimensioni degli spazi vettoriali considerati sono notevoli. Nel seguito è presentato il relativo Algoritmo 1 utilizzato per l'implementazione numerica.

Algoritmo 1 Steepest Descent (SD)

Input: $\mathbf{u}_0$, K_{max}, tol
1: risoluzione di equazione di stato ed equazione aggiunta, ottenendo $\mathbf{y}_0, \mathbf{p}_0$
2: valutazione $J_h(\mathbf{u}_0)$ e $\nabla J_h(\mathbf{u}_0)$
3: $k \leftarrow 0$
4: **while** $k \leq K_{max}$ and $||\nabla J_h(\mathbf{u}_k)|| \geq tol$ **do**
5: calcolo di τ_k tramite la (2.3.3)
6: $\mathbf{u}_{k+1} \leftarrow \mathbf{u}_k - \tau_k \nabla J_h(\mathbf{u}_k)$
7: risoluzione di equazione di stato ed equazione aggiunta, ottenendo $\mathbf{y}_{k+1}, \mathbf{p}_{k+1}$
8: valutazione $J_h(\mathbf{u}_{k+1})$ e $\nabla J_h(\mathbf{u}_{k+1})$
9: $k \leftarrow k + 1$
10: **end while**

Metodo del gradiente coniugato

Il metodo del gradiente coniugato estende l'idea del metodo steepest descent a direzioni *coniugate* rispetto alla matrice Hessiana del sistema $\mathbb{H}$

$$\mathbf{d}_k^T \mathbb{H}(\mathbf{u}_k)\mathbf{d}_{k-1} = 0. \tag{2.3.6}$$

La direzione è determinata quindi dalle seguenti espressioni.

$$\mathbf{d}_k = -\nabla J_h(\mathbf{u}_k) + \lambda_k \mathbf{d}_{k-1}, \qquad \lambda_k = \frac{\nabla J_h(\mathbf{u}_k)^T \mathbb{H}(\mathbf{u}_k)\mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^T \mathbb{H}(\mathbf{u}_k)\mathbf{d}_{k-1}}, \tag{2.3.7}$$

dove il parametro λ_k viene scelto come indicato nel caso di funzionali lineari-quadratici. In generale, quindi, si può costruire l'algoritmo come in Algoritmo 2.

Metodo *BFGS*

Il metodo *BFGS*, così battezzato dai nomi degli studiosi, Broyden, Fletcher, Goldfarb e Shanno, che lo idearono. è classificato tra i metodi *quasi-Newton*, nei quali la matrice Hessiana del sistema (o la sua inversa) è sostituita da una sua approssimazione, calcolata iterativamente in maniera contestuale alla ricerca del controllo ottimo. Questi metodi rappresentano una versione *Jacobian-free* dei metodi di Newton, nei quali la direzione di discesa scelta si ottiene risolvendo un sistema lineare con la matrice Hessiana come matrice dei coefficienti e il gradiente cambiato di segno del funzionale obiettivo come termine noto

$$\mathbb{H}(\mathbf{u}_k)\mathbf{d}_k = -\nabla J_h(\mathbf{u}_k). \tag{2.3.8}$$

Una strategia computazionalmente efficiente per risolvere questo sistema lineare consiste nel definire ricorsivamente un'approssimazione della matrice inversa $\mathbb{B}_k \approx \mathbb{H}^{-1}(\mathbf{u}_k)$, partendo da una stima iniziale $\mathbb{B}_0$ secondo l'espressione

$$\mathbb{B}_{k+1} = (\mathbb{I} - \rho_k \mathbf{g}_k \mathbf{s}_k^T)^T \mathbb{B}_k (\mathbb{I} - \rho_k \mathbf{s}_k \mathbf{g}_k^T) + \rho_k \mathbf{g}_k \mathbf{g}_k^T, \tag{2.3.9}$$

dove sono definiti

$$\mathbf{s}_k = \mathbf{u}_{k+1} - \mathbf{u}_k, \qquad \mathbf{g}_k = \nabla J_h(\mathbf{u}_{k+1}) - \nabla J_h(\mathbf{u}_k), \qquad \rho_k = \frac{1}{\mathbf{g}_k^T \mathbf{s}_k}. \tag{2.3.10}$$

Questo metodo si basa sull'assunzione fondamentale di positività di $\mathbf{s}_k^T \mathbf{g}_k$, detta *condizione di curvatura*. Per assicurare ad ogni iterazione la validità di questo vincolo, è necessario implementare le cosiddette condizioni di Wolfe sul parametro τ_k, aggiungendo alla consueta condizione di Armijo, il vincolo

$$\mathbf{g}_k^T \mathbf{s}_k \geq (c_2 - 1)\tau_k \nabla J_h(\mathbf{u}_k)\mathbf{d}_k, \tag{2.3.11}$$

dove $c_2 = 0.9$ è una costante fissata.
L'Algoritmo 3 risultante è di seguito riportato.
Nel caso di problemi di elevate dimensioni, una modifica computazionalmente efficiente di questo metodo consiste nell'evitare di memorizzare l'intera matrice approssimata $\mathbb{B}_k$, utilizzando cioè tutte le coppie $(\mathbf{s}_i, \mathbf{g}_i)$ disponibili fino a quell'iterata, limitandosi a mantenere solo le m coppie più recenti: questa strategia viene detta *Limited-Memory-BFGS*.

Algoritmo 2 Gradiente Coniugato (CG)

Input: $\mathbf{u}_0, \mathbb{H}, \mathbf{b}$, K_{max}, *tol*

1: calcolo del residuo iniziale $\mathbf{r}_0 = \mathbb{H}\mathbf{u}_0 - \mathbf{b}$ e della direzione iniziale $\mathbf{d}_0 = -\mathbf{r}_0$
2: $k \leftarrow 0$
3: **while** $k \leq K_{max}$ and $||\mathbf{r}_k||/||\mathbf{b}|| \geq tol$ **do**
4: calcolo di τ_k tramite la (2.3.3)
5: $\mathbf{u}_{k+1} \leftarrow \mathbf{u}_k + \tau_k \mathbf{d}_k$
6: $\mathbf{r}_{k+1} \leftarrow \mathbb{H}\mathbf{u}_{k+1} - \mathbf{b}$
7: calcolo di λ_{k+1} tramite la (2.3.7)
8: $\mathbf{d}_{k+1} \leftarrow -\mathbf{r}_{k+1} + \lambda_{k+1}\mathbf{d}_k$
9: $k \leftarrow k + 1$
10: **end while**

Algoritmo 3 BFGS

Input: $\mathbf{u}_0$, $\mathbb{B}_0$, K_{max}, *tol*

1: risoluzione di equazione di stato ed equazione aggiunta, ottenendo $\mathbf{y}_0, \mathbf{p}_0$
2: valutazione di $\nabla J_h(\mathbf{u}_0)$
3: $k \leftarrow 0$
4: **while** $k \leq K_{max}$ and $||\nabla J_h(\mathbf{u}_k)|| \geq tol$ **do**
5: $\mathbf{d}_k \leftarrow -\mathbb{B}_k \nabla J_h(\mathbf{u}_k)$
6: calcolo di τ_k tramite *line-search* con condizioni di Wolfe
7: $\mathbf{u}_{k+1} \leftarrow \mathbf{u}_k + \tau_k \mathbf{d}_k$
8: risoluzione di equazione di stato ed equazione aggiunta, ottenendo $\mathbf{y}_{k+1}, \mathbf{p}_{k+1}$
9: valutazione di $\nabla J_h(\mathbf{u}_{k+1})$
10: $\mathbf{s}_k \leftarrow \mathbf{u}_{k+1} - \mathbf{u}_k$
11: $\mathbf{g}_k \leftarrow \nabla J_h(\mathbf{u}_{k+1}) - \nabla J_h(\mathbf{u}_k)$
12: calcolo di $\mathbb{B}_{k+1}$ tramite la (2.3.9)
13: $k \leftarrow k + 1$
14: **end while**

Capitolo 3

Caso Stocastico

In questo terzo capitolo, il problema di controllo ottimo vincolato da equazioni differenziali alle derivate parziali viene analizzato in presenza di parametri affetti da incertezza. Questa generalizzazione della trattazione richiede, ai fini dell'analisi numerica del problema, di associare una discretizzazione in probabilità a quella spaziale, con l'obiettivo di stimare numericamente determinate quantità di interesse statistiche, quali valore atteso o varianza.

3.1 Preliminari

Prima di addentrarsi nell'analisi del problema dal punto di vista funzionale e numerico, è opportuno presentare alcuni strumenti utili alla descrizione delle funzioni affette da incertezza introdotte e dei loro spazi di definizione.

3.1.1 Random fields ed espansioni di Karhunen-Loève

Si considerino un dominio spaziale $D \in \mathbb{R}^d$ e uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. Una funzione $a : D \times \Omega \to \mathbb{R}$, dipendente da una variabile spaziale $\mathbf{x} \in D$ e da un evento $\omega \in \Omega$, viene definita *random field*. In particolare, si osserva che, fissato $\mathbf{x} \in D$, $a(\mathbf{x}, \cdot)$ è una variabile casuale definita su Ω, mentre, fissato un evento $\omega \in \Omega$, $a(\cdot, \omega)$ viene detta realizzazione del random field. Relativamente al random field è possibile definire alcune quantità statistiche di interesse quali:

- Valore atteso
$$\mathbb{E}[a(\mathbf{x}, \cdot)] \coloneqq \int_\Omega a(\mathbf{x}, \omega) d\mathbb{P}(\omega). \tag{3.1.1}$$

- Funzione di covarianza
$$Cov_a(\mathbf{x}, \mathbf{x}') \coloneqq \mathbb{E}[(a(\mathbf{x}, \cdot) - \mathbb{E}[a(\mathbf{x}, \cdot)])(a(\mathbf{x}', \cdot) - \mathbb{E}[a(\mathbf{x}', \cdot)])]. \tag{3.1.2}$$

- Varianza
$$Var(\mathbf{x}) = Cov_a(\mathbf{x}, \mathbf{x}) = \int_\Omega (a(\mathbf{x}, \omega) - \mathbb{E}[a(\mathbf{x}, \cdot)])^2 \, d\mathbb{P}(\omega). \tag{3.1.3}$$

Ai fini computazionali, per discretizzare i random fields è conveniente separare la dipendenza spaziale da quella in probabilità tramite un set finito di variabili casuali $\xi_k(\omega)$, $k = 1, \dots, N$. Tra i vari metodi di espansione, in questa trattazione è presentata l'approssimazione di Karhunen-Loève, in applicazione a due differenti tipi di random field: quello Gaussiano, in cui fissati alcuni nodi $\mathbf{x}_i, i = 1, \dots, M$ la variabile casuale a valori in $\mathbb{R}^M$ definita come $X(\omega) = a(\mathbf{x}_i, \omega)_{1 \le i \le M}$ ha distribuzione Gaussiana multivariata

$$X \sim \mathcal{N}(\mu, C), \quad \text{con} \quad \mu = \mathbb{E}[a(\mathbf{x}_i, \cdot)]_{1 \le i \le M} \quad \text{e} \quad C_{ij} = Cov_a(\mathbf{x}_i, \mathbf{x}_j), \qquad 1 \le i, j \le M,$$

e quello log-normale, in cui ad avere distribuzione Gaussiana multivariata è il logaritmo del random field.
In generale, l'obiettivo di questa sezione consiste nel fornire una rappresentazione di un random field nella forma seguente

$$a(\mathbf{x}, \omega) = \mathbb{E}[a(\mathbf{x}, \cdot)] + \sum_{k=1}^{\infty} b_k \tilde{\xi_k},$$

dove $\{b_k\}_{k \in \mathbb{N}} \in L^2(D)$, è un set di funzioni ortonormali e $\{\tilde{\xi_k}\}_{k \in \mathbb{N}}$ è un set di variabili casuali scorrelate con medie nulle e varianze σ_k^2. Senza perdita di generalità, è possibile definire $\xi_k = \frac{1}{\sigma_k} \tilde{\xi_k}$, ottenendo quindi

$$a(\mathbf{x}, \omega) = \mathbb{E}[a(\mathbf{x}, \cdot)] + \sum_{k=1}^{\infty} \sigma_k b_k \xi_k. \tag{3.1.4}$$

Random fields Gaussiani

I random fields Gaussiani ammettono la rappresentazione (3.1.4), considerando, in particolare, variabili casuali $\{\xi_k\}_{k \in \mathbb{N}}$ Gaussiane. Vale, pertanto, il seguente teorema [13, 19].

Teorema 3.1.1. *Siano $D \subset \mathbb{R}^d$ un dominio limitato e $a(\mathbf{x}, \omega)$ un random field con funzione di covarianza $Cov_a : \overline{D} \times \overline{D} \to \mathbb{R}$. Allora $a(\mathbf{x}, \omega)$ ammette l'espansione di Karhunen-Loève*

$$a(\mathbf{x}, \omega) = \mathbb{E}[a(\mathbf{x}, \cdot)] + \sum_{k=1}^{\infty} \sqrt{\lambda_k} b_k \xi_k, \tag{3.1.5}$$

dove la sommatoria è convergente nello spazio di Bochner $L^2_{\mathbb{P}}(\Omega, L^2(D))$. Inoltre, se il random field è Gaussiano si ha che $\{\xi_k\} \sim \mathcal{N}(0,1)$, indipendenti e identicamente distribuite.

Definita la funzione di covarianza come nella (3.1.2), risulta che

$$\begin{aligned} \int_D Cov_a(\mathbf{x}, \mathbf{x}') b_k' d\mathbf{x} &= \int_D \left(\sum_{j=1}^{\infty} \sigma_j^2 b_j b_j' \right) b_k' d\mathbf{x} \\ &= \sum_{j=1}^{\infty} \sigma_j^2 b_j \int_D b_j' b_k' d\mathbf{x} = \sum_{j=1}^{\infty} \sigma_j^2 b_j \delta_{jk} = \sigma_k^2 b_k, \end{aligned}$$

dove nel penultimo passaggio è stata utilizzata l'ipotesi di ortonormalità delle funzioni $\{b_k\}_{k \in \mathbb{N}}$. Questo risultato mostra come le suddette funzioni e le varianze σ_k^2 rappresentino rispettivamente le autofunzioni e gli autovalori relativi alla funzione di covarianza. La legittimità di questo risultato è garantita dal seguente teorema [13, 19].

Teorema 3.1.2 (di Mercer). *Sia definito l'operatore* $\mathcal{C} : L^2(D) \to L^2(D)$ *tale che*

$$\mathcal{C}\psi = \int_D C(\mathbf{x}, \mathbf{x}')\psi(\mathbf{x}')d\mathbf{x}',$$

dove $C : \overline{D} \times \overline{D} \to \mathbb{R}$ *è una funzione continua, simmetrica e non-negativa definita. Allora esiste una sequenza di autofunzioni ortonormali* $\{b_k(\mathbf{x})\}_{k\in\mathbb{N}}$ *di* $\mathcal{C}$ *tale che i corrispondenti autovalori* $\{\lambda_k(\mathbf{x})\}_{k\in\mathbb{N}}$ *siano positivi. Inoltre la serie*

$$C(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} \lambda_k b_k(\mathbf{x}) b_k(\mathbf{x}')$$

converge assolutamente e uniformemente in $\overline{D} \times \overline{D}$.

Osservazione 3.1.1. *Le coppie autovalore-autofunzione dell'espansione di Karhunen-Loève si ottengono applicando il Teorema di Mercer all'operatore* $\mathcal{C}$*, con* $C(\mathbf{x}, \mathbf{x}') = Cov_a(\mathbf{x}, \mathbf{x}')$.

In termini di approssimazione numerica, è pertanto naturale la scelta di un troncamento di tale espansione ad un numero finito K di variabili casuali

$$a_K(\mathbf{x}, \omega) = \mathbb{E}[a(\mathbf{x}, \cdot)] + \sum_{k=1}^{K} \sqrt{\lambda_k} b_k \xi_k. \tag{3.1.6}$$

Tuttavia, i random fields Gaussiani non garantiscono che $a(\mathbf{x}, \omega) > 0$, come sarebbe preferibile a livello applicativo, motivo per cui vengono introdotti i random fields log-normali.

Random fields log-normali

In molte applicazioni, le variabili affette da incertezza devono avere valori positivi. Pertanto i random fields Gaussiani non sembrano essere la migliore scelta per questo tipo di parametri, in quanto associano probabilità positive a valori negativi. Un'alternativa è rappresentata dall'espressione

$$a(\mathbf{x}, \omega) = \exp(\mu(\mathbf{x}) + \sigma(\mathbf{x})\mathcal{Z}(\mathbf{x}, \omega)), \tag{3.1.7}$$

dove $\mathcal{Z}(\mathbf{x}, \omega)$ è un random field Gaussiano con media nulla e varianza unitaria, come descritto nel paragrafo precedente. Media e varianza del campo $a(\mathbf{x}, \omega)$ sono definite come

$$\overline{a}(\mathbf{x}) = \exp\left(\mu(x) + \frac{1}{2}\sigma^2(\mathbf{x})\right), \qquad \mathrm{Var}(a(\mathbf{x})) = e^{2\mu(\mathbf{x})+\sigma^2(\mathbf{x})}\left(e^{\sigma^2(\mathbf{x})} - 1\right),$$

da cui si ricavano tramite formule inverse le espressioni di $\mu(\mathbf{x})$ e $\sigma^2(\mathbf{x})$ che permettono di definire

$$\mu^*(\mathbf{x}) = e^{\mu(\mathbf{x})} \qquad \textit{Parametro di scala,}$$
$$\sigma^*(\mathbf{x}) = e^{\sigma^2(\mathbf{x})} \qquad \textit{Parametro di forma.}$$

Infine, il troncamento dell'espressione (3.1.7), alla luce della formula ricavata nel caso Gaussiano, rappresenta un'approssimazione numerica del random field log-normale

$$a_K(\mathbf{x}, \omega) = \exp\left(\mu(\mathbf{x}) + \sigma(\mathbf{x}) \sum_{k=1}^{K} \sqrt{\lambda_k} b_k(\mathbf{x}) \xi_k(\omega)\right). \tag{3.1.8}$$

3.1.2 Prodotti tensoriali e spazi funzionali

La trattazione dei random fields deve tenere presente la differente struttura funzionale tra dipendenza spaziale e in probabilità. In particolare, siano definiti gli spazi $L^2(D; d\mathbf{x})$, dove $d\mathbf{x}$ rappresenta la misura di Lebesgue su D e $L^2(\Omega; d\mathbb{P}(\omega))$ lo spazio delle funzioni $g : \Omega \to \mathbb{R}$ a modulo quadrato integrabile rispetto alla misura di probabilità $\mathbb{P}$. In generale si ha quindi che $y(\mathbf{x}, \omega) \in L^2(D \times \Omega, d\mathbf{x} \times d\mathbb{P}(\omega))$, dove y è un generico random field. Pertanto una sua rappresentazione conveniente coinvolge il prodotto tensoriale tra gli spazi funzionali di definizione. Siano $\{\phi_i(\mathbf{x})\}_{i\in\mathbb{N}}$ e $\{\psi_j(\omega)\}_{j\in\mathbb{N}}$ le basi ortonormali rispettivamente di $L^2(D; d\mathbf{x})$ e $L^2(\Omega; d\mathbb{P}(\omega))$, sia introdotta la mappa

$$\begin{aligned} U : L^2(D; d\mathbf{x}) \otimes L^2(\Omega; d\mathbb{P}(\omega)) &\to L^2(D \times \Omega, d\mathbf{x} \times d\mathbb{P}(\omega)), \\ (\phi_i(\mathbf{x}) \otimes \psi_j(\omega)) &\mapsto \phi_i(\mathbf{x})\psi_j(\omega). \end{aligned} \tag{3.1.9}$$

Pertanto si ha che, definite due generiche funzioni

$$\phi(\mathbf{x}) = \sum_{i=1}^{\infty} c_i \phi_i(\mathbf{x}) \in L^2(D; d\mathbf{x}), \qquad \psi(\omega) = \sum_{j=1}^{\infty} c'_j \psi_j(\omega) \in L^2(\Omega; d\mathbb{P}(\omega)),$$

vale la seguente espressione

$$U(\phi \otimes \psi) = U\Big(\sum_{i,j=1}^{\infty} c_i c'_j \phi_i(\mathbf{x}) \otimes \psi_j(\omega) \Big) = \sum_{i,j=1}^{\infty} c_i c'_j \phi_i(\mathbf{x})\psi_j(\omega) = \phi\psi. \tag{3.1.10}$$

In conclusione, si dimostra che U è un isomorfismo naturale tra gli spazi $L^2(D; d\mathbf{x}) \otimes L^2(\Omega; d\mathbb{P}(\omega))$ e $L^2(D \times \Omega, d\mathbf{x} \times d\mathbb{P}(\omega))$, quindi si può scrivere

$$L^2(D; d\mathbf{x}) \otimes L^2(\Omega; d\mathbb{P}(\omega)) \cong L^2(D \times \Omega, d\mathbf{x} \times d\mathbb{P}(\omega)). \tag{3.1.11}$$

In maniera più formale, è possibile generalizzare questa trattazione per un qualunque spazio di Hilbert H (caso particolare $H = L^2(D; d\mathbf{x})$), ottenendo che esiste un isomorfismo naturale

$$\begin{aligned} \overline{U} : L^2(\Omega; d\mathbb{P}(\omega)) \otimes H &\to L^2_{\mathbb{P}}(\Omega; H), \\ \sum_{k=1}^{N} g_k(\omega) \otimes \phi_k &\mapsto \sum_{k=1}^{N} g_k(\omega)\phi_k, \end{aligned} \tag{3.1.12}$$

dove g_k sono definiti come coefficienti della seguente rappresentazione di una generica funzione definita sullo spazio di Bochner $g \in L^2_{\mathbb{P}}(\Omega; H)$, dati dal suo prodotto scalare rispetto alla base $\{\phi_i(\mathbf{x})\}_{i\in\mathbb{N}}$ di H

$$g_k(\omega) = \langle g(\omega), \phi_k \rangle_H , \qquad g(\omega) = \lim_{N\to\infty} \sum_{k=1}^{N} g_k(\omega)\phi_k. \tag{3.1.13}$$

Si conclude che $L^2(\Omega; d\mathbb{P}(\omega)) \otimes H \cong L^2_{\mathbb{P}}(\Omega; H)$.

Queste considerazioni sono riassunte dal seguente risultato [19].

Teorema 3.1.3. *Siano definiti gli spazi misurabili $L^2(M_1, d\mu_1)$ e $L^2(M_2, d\mu_2)$. Allora valgono le seguenti implicazioni:*

- $\exists! \, U : L^2(M_1, d\mu_1) \otimes L^2(M_2, d\mu_2) \to L^2(M_1 \times M_2, d\mu_1 \times d\mu_2)$ *isomorfismo tale che, se $\phi \in L^2(M_1, d\mu_1)$ e $\psi \in L^2(M_2, d\mu_2)$ si ha $\phi \otimes \psi \mapsto \phi\psi$.*
- *se $(H, \langle \cdot, \cdot \rangle_H)$ è uno spazio di Hilbert separabile allora $\exists! \, \overline{U} : L^2(M_1, d\mu_1) \otimes H \to L^2(M_1, d\mu_1; H)$ tale che se $g \in H$ si ha $g \otimes \phi \mapsto g\phi$.*

3.2 Analisi del problema

Questa sezione contiene l'analisi variazionale del problema di controllo ottimo, con l'obiettivo di dimostrare buona positura ed esistenza delle soluzioni. L'equazione di Poisson-Laplace è utilizzata come problema modello per l'analisi, considerando affetti da incertezza il coefficiente di diffusione e la forzante.

3.2.1 Equazioni differenziali alle derivate parziali in condizioni di incertezza

Si consideri il seguente problema differenziale ellittico

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla y(\mathbf{x}, \omega)) = \phi(\mathbf{x}, \omega) & \text{in } D \times \Omega, \\ y(\mathbf{x}, \omega) = 0 & \text{su } \partial D \times \Omega. \end{cases} \tag{3.2.1}$$

Si osserva che, nel problema di controllo ottimo associato a questa equazione, la forzante viene ridefinita in termini della variabile di controllo $\phi(\mathbf{x}, \omega) = f(\mathbf{x}, \omega) + u(\mathbf{x})$.
La soluzione dell'equazione differenziale stocastica presentata, secondo la rappresentazione descritta nella sezione precedente, è un random field definito sullo spazio di Bochner $y(\mathbf{x}, \omega) \in L^2_{\mathbb{P}}(\Omega; H^1_0(D))$, su cui è introdotta la norma

$$||y||_{L^2_{\mathbb{P}}(\Omega; H^1_0(D))} = \left(\int_\Omega ||y(\omega)||^2_{H^1_0(D)} d\mathbb{P}(\omega) \right)^{1/2}. \tag{3.2.2}$$

Per scrivere la formulazione variazionale dell'equazione, i termini vengono integrati per parti rispetto alla misura di Lebesgue sul dominio spaziale $d\mathbf{x}$ e poi in media rispetto alla misura di probabilità $d\mathbb{P}(\omega)$. Pertanto si ottiene

$$\int_\Omega \int_D a(\mathbf{x}, \omega) \nabla y \cdot \nabla v \, d\mathbf{x} d\mathbb{P}(\omega) = \int_\Omega \int_D [\phi(\mathbf{x}, \omega)] v \, d\mathbf{x} d\mathbb{P}(\omega), \quad \forall v \in L^2_{\mathbb{P}}(\Omega; H^1_0(D)). \tag{3.2.3}$$

Definite la forma bilineare $A(y, v) : L^2_{\mathbb{P}}(\Omega; H^1_0(D)) \times L^2_{\mathbb{P}}(\Omega; H^1_0(D)) \to \mathbb{R}$, e la forma lineare $F : L^2_{\mathbb{P}}(\Omega; H^1_0(D)) \to \mathbb{R}$, corrispondenti rispettivamente al primo e al secondo termine dell'equazione variazionale, è possibile riscrivere il problema in forma astratta

$$A(y, v) = F(v), \quad \forall v \in L^2_{\mathbb{P}}(\Omega; H^1_0(D)). \tag{3.2.4}$$

Per garantire la validità delle ipotesi del teorema di Lax-Milgram, quali continuità e coercività della forma bilineare e continuità di quella lineare, sono necessarie due assunzioni su a e ϕ:

- $a \in L^{\infty}_{\mathbb{P}}(\Omega; L^{\infty}(D))$ ed esistono $0 < a_{min} \leq a_{max} < +\infty$ tali che

$$\mathbb{P}(\omega \in \Omega \,|\, a(\mathbf{x}, \omega) \in [a_{min}, a_{max}], \forall \mathbf{x} \in \overline{D}) = 1.$$

- $\phi \in L^2_{\mathbb{P}}(\Omega; L^2(D))$.

Pertanto, si può enunciare il seguente teorema [19].

Teorema 3.2.1. *Se valgono le due assunzioni precedenti,* $\exists!\, y(\mathbf{x}, \omega) \in L^2_{\mathbb{P}}(\Omega; H^1_0(D))$ *che risolve il problema variazionale* (3.2.4)*. Inoltre, definita la costante di Poincaré* C_P *come nella* (1.2.30)*, si dimostra che vale la stima*

$$||y||_{L^2_{\mathbb{P}}(\Omega; H^1_0(D))} \leq \frac{C_P}{a_{min}} ||\phi||_{L^2_{\mathbb{P}}(\Omega; L^2(D))}. \tag{3.2.5}$$

Tale disuguaglianza segue dalle ipotesi sul coefficiente $a(\mathbf{x}, \omega)$ e dalla disuguaglianza di Cauchy-Schwarz

$$\begin{aligned} a_{min}||y||^2_{L^2_{\mathbb{P}}(\Omega; H^1_0(D))} &= a_{min} \int_{\Omega} \int_{D} |\nabla y|^2 \, d\mathbf{x} d\mathbb{P}(\omega) \\ &\leq \int_{\Omega} \int_{D} a(\mathbf{x}, \omega) |\nabla y|^2 \, d\mathbf{x} d\mathbb{P}(\omega) = \int_{\Omega} \int_{D} \phi(\mathbf{x}, \omega) y \, d\mathbf{x} d\mathbb{P}(\omega) \\ &\leq ||\phi||_{L^2(D) \otimes L^2_{\mathbb{P}}(\Omega)} ||y||_{L^2_{\mathbb{P}}(\Omega; L^2(D))} \leq C_P ||\phi||_{L^2(D) \otimes L^2_{\mathbb{P}}(\Omega)} ||y||_{L^2_{\mathbb{P}}(\Omega; H^1_0(D))}. \end{aligned}$$

Osservazione 3.2.1. *Si ricorda che, alla luce delle considerazioni della sezione* (3.1.2) *l'analisi può essere analogamente condotta nello spazio isomorfo definito dal prodotto tensoriale* $H^1_0(D) \otimes L^2_{\mathbb{P}}(\Omega)$*, richiedendo per la forzante* $\phi \in L^2(D) \otimes L^2_{\mathbb{P}}(\Omega)$*, oltre all'assunzione su* $a(\cdot, \cdot)$*.*

3.2.2 Funzionale obiettivo e condizioni di ottimalità

Dopo aver introdotto una formulazione variazionale dell'equazione differenziale che rappresenta il vincolo del problema di ottimizzazione in esame, è possibile ricavare il relativo sistema di condizioni di ottimalità, come nel caso deterministico, utilizzando il metodo del problema aggiunto e riducendo il funzionale obiettivo alla sola dipendenza dalla variabile di controllo. Per semplicità notazionale, di seguito è introdotta una riformulazione in senso generalizzato degli spazi funzionali di appartenenza delle variabili coinvolte, definendo $V \equiv H^1_0(D)$ e $\mathcal{U} \equiv L^2(D)$. Nella presente trattazione, il funzionale obiettivo dipende da una certa *misura di rischio*, quale ad esempio il valore atteso o la varianza della distanza in norma tra la soluzione dell'equazione di stato, indicizzata dalla componente relativa all'incertezza, $y = y^{\omega}(\mathbf{x}) \in V$ e una funzione target $z_d(\mathbf{x}) \in \hat{V}$, dove viene definito $\hat{V} \subset V$ come il sottoinsieme di osservazione relativo alla variabile di stato. Una scelta alternativa è rappresentata da misure più complesse quali il *Conditional Value at Risk (CVaR)*, che verrà trattato in dettaglio nel capitolo seguente. In particolare, per questa analisi si consideri il funzionale obiettivo

$$J(y, u) = \frac{1}{2} \mathbb{E}[||Cy^{\omega} - z_d||^2_{\hat{V}}] + \frac{\nu}{2} ||u||^2_{\mathcal{U}}, \tag{3.2.6}$$

dove $C : V \to \hat{V}$ è un operatore tra spazi di Hilbert lineare e continuo, detto *operatore di osservazione*, mentre $\mathcal{U}$ rappresenta lo spazio di Hilbert relativo alla variabile di controllo $u \in \mathcal{U}$. Inoltre, viene definita $y^\omega = y^\omega(u)$ la soluzione del problema differenziale in forma astratta

$$A_\omega(y^\omega, v) = {}_{V'}\langle f^\omega + Bu, v\rangle_V \,, \forall v \in V, \mathbb{P} - q.o.\, \omega \in \Omega, \tag{3.2.7}$$

dove la forma bilineare $A_\omega : V \times V \to V'$ soddisfa

$$A_\omega(y, v) = {}_{V'}\langle -\nabla(a_\omega \nabla y), v\rangle_V \,, \quad \forall v \in V,$$

dove $a_\omega \coloneqq a(\cdot, \omega)$. A secondo membro, oltre alla forzante $f^\omega \in V$, è definito l'*operatore di controllo*, lineare e limitato, $B : \mathcal{U} \to V'$. Analogamente al problema del caso deterministico, è possibile riformulare il funzionale obiettivo esprimendo la variabile di stato in funzione del controllo $u \in \mathcal{U}$, o eventualmente definito su un suo sottospazio chiuso e convesso, ottenendo il funzionale ridotto

$$\hat{J}(u) = \frac{1}{2}\mathbb{E}[||Cy^\omega(u) - z_d||^2_{\hat{V}}] + \frac{\nu}{2}||u||^2_{\mathcal{U}} = \frac{1}{2}\mathbb{E}[||C\hat{S}(f^\omega + Bu) - z_d||^2_{\hat{V}}] + \frac{\nu}{2}||u||^2_{\mathcal{U}}, \tag{3.2.8}$$

Nel secondo passaggio viene introdotto l'operatore controllo-stato $\hat{S}$, definito come

$$\begin{aligned} \hat{S} : V' &\to L^2_{\mathbb{P}}(\Omega; V), \\ h &\mapsto S_\omega h, \end{aligned} \tag{3.2.9}$$

dove $S_\omega : V' \to V$ è l'operatore di risoluzione del problema differenziale, che mappa $(f^\omega + Bu) \in V'$ in $S_\omega(f^\omega + Bu) \in V$ tale che $A_\omega(S_\omega(f^\omega + Bu), v) = {}_{V'}\langle (f^\omega + Bu), v\rangle_V \,, \forall v \in V, \mathbb{P} - q.o.\, \omega \in \Omega$. In definitiva, il problema ridotto può essere riformulato come

$$\min_{u \in \mathcal{U}} \hat{J}(u) = \frac{1}{2}\mathbb{E}[||C\hat{S}(f^\omega + Bu) - z_d||^2_{\hat{V}}] + \frac{\nu}{2}||u||^2_{\mathcal{U}}. \tag{3.2.10}$$

Definito il problema di riferimento, è necessario dimostrare l'esistenza di un controllo ottimo $u^\star \in \mathcal{U}$ tale che

$$\hat{J}(u^\star) \leq \hat{J}(u), \qquad \forall u \in \mathcal{U}. \tag{3.2.11}$$

A tale scopo, è possibile osservare che vale l'enunciato del Teorema 2.1.1 per il problema deterministico, mantenendo la stessa dimostrazione. In ottica di una trattazione dal punto di vista numerico del problema di controllo ottimo, è necessario definire un sistema di condizioni che garantiscano l'ottimalità della soluzione. Per raggiungere questo obiettivo, viene seguito un approccio di tipo Lagrangiano, simile a quello utilizzato nel caso deterministico, con la finalità di ottenere un'espressione per la derivata del funzionale obiettivo ridotto. Pertanto, definita la variabile aggiunta $p^\omega(\mathbf{x}) \in V$ si introduce la funzione Lagrangiana

$$\begin{aligned} \mathcal{L}(y, u, p) =& J(y, u) - \mathbb{E}\Big[A_\omega(y, p) - {}_{V'}\langle f^\omega + Bu, p\rangle_V\Big] \\ =& J(y, u) - \mathbb{E}\Big[\int_D (a(\mathbf{x}, \omega)\nabla y \cdot \nabla p - (f^\omega + Bu)p)\, d\mathbf{x}\Big]. \end{aligned} \tag{3.2.12}$$

Imponendo nulle le derivate parziali della Lagrangiana si ottengono le condizioni di ottimalità richieste, in particolare:

- Derivando rispetto a p si ottiene l'equazione di stato

$$A_\omega(y^\omega, v) = {}_{V'}\langle f^\omega + Bu, v\rangle_V, \qquad \forall v \in V, \mathbb{P}-q.o.\,\omega \in \Omega. \tag{3.2.13}$$

- Derivando rispetto a y si ottiene l'equazione aggiunta, tenendo conto della presenza dell'operatore di osservazione

$$A_\omega(v, p^\omega) = {}_{\hat{V}'}\langle R_{\hat{V}}(Cy^\omega - z_d), Cv\rangle_{\hat{V}}, \qquad \forall v \in V, \mathbb{P}-q.o.\,\omega \in \Omega, \tag{3.2.14}$$

  dove viene definito l'operatore di Riesz $R_{\hat{V}} : \hat{V} \to \hat{V}'$.

- Derivando infine rispetto al controllo u e ricordando la proprietà di differenziabilità secondo Fréchet del funzionale obiettivo in forma ridotta, che segue dalla linearità dell'equazione di stato, si ottiene

$$\hat{J}'(u)q = {}_{\mathcal{U}'}\langle \nu R_U\, u - B^*\mathbb{E}[p^\omega(u)], q\rangle_{\mathcal{U}}, \qquad \forall q \in \mathcal{U}. \tag{3.2.15}$$

  con $R_U : \mathcal{U} \to \mathcal{U}'$ operatore di Riesz riferito allo spazio dei controlli e $B^* : V' \to \mathcal{U}'$ operatore duale di B. In definitiva, introdotta la soluzione del problema di controllo ottimo $u^\star \in \mathcal{U}$, vale il sistema di condizioni di ottimalità (dette di Karush-Kuhn-Tucker), $\forall v \in V, \forall q \in \mathcal{U}, \mathbb{P}-q.o.\,\omega \in \Omega$

$$\begin{cases} \nabla_y \mathcal{L}(y^\omega, u^\star, p^\omega) = A_\omega(y^\omega, v) - {}_{V'}\langle f_\omega + Bu^\star, v\rangle_V = 0, \\ \nabla_p \mathcal{L}(y^\omega, u^\star, p^\omega) = A_\omega(v, p^\omega) - {}_{V'}\langle C^* R_{\hat{V}}(Cy^\omega - z_d), v\rangle_V = 0, \\ \quad {}_{\mathcal{U}'}\langle \nabla_u \mathcal{L}(y^\omega, u^\star, p^\omega), q - u^\star\rangle_{\mathcal{U}} = {}_{\mathcal{U}'}\langle \nu R_U\, u^\star - B^*\mathbb{E}[p^\omega(u^\star)], q - u^\star\rangle_{\mathcal{U}} = 0. \end{cases} \tag{3.2.16}$$

Nelle sezioni seguenti vengono introdotti strumenti per la discretizzazione a elementi finiti del problema, inoltre sono presentati alcuni casi di interesse quali la restrizione del controllo sul bordo del dominio.

3.3 Discretizzazione spaziale

La trattazione computazionale del problema di controllo ottimo in condizioni di incertezza si basa su una discretizzazione agli elementi finiti nella variabile spaziale, come già descritto per il caso deterministico del capitolo precedente. Siano, pertanto, introdotti gli spazi vettoriali finito-dimensionali V_h per la variabile spaziale e $\mathcal{U}_h$ per quella di controllo. Il problema discreto risultante viene costruito considerando la formulazione ridotta del funzionale J_h, in cui la variabile di stato affetta da incertezza y_h^ω viene espressa in funzione del controllo u_h

$$\begin{aligned} \min_{u_h \in \mathcal{U}_h} \hat{J}_h(u_h) &= \frac{1}{2}\mathbb{E}||(y_h^\omega(u_h) - z_d)||^2_{L^2(D)} + \frac{\nu}{2}||u_h||^2_{\mathcal{U}} \\ \text{s.t.} \quad & A_\omega(y_h^\omega(u_h), v_h) = {}_{V'}\langle (f^\omega + Bu_h), v_h\rangle_V, \quad \forall v_h \in V_h, \quad \mathbb{P}-q.o.\,\omega \in \Omega, \\ & y_h^\omega \in V_h. \end{aligned} \tag{3.3.1}$$

In analogia con la trattazione variazionale, è possibile definire le rappresentazioni discrete dell'operatore di risoluzione dell'equazione differenziale

$$S_h^\omega : V' \to V_h, \qquad A_\omega(S_h^\omega(f^\omega + Bu_h), v_h) = {}_{V'}\langle (f^\omega + Bu_h), v_h \rangle_V, \quad \forall v_h \in V_h,$$

e di conseguenza l'operatore controllo-stato discreto

$$\hat{S}_h : V_h' \to L^2_{\mathbb{P}}(\Omega; V_h), \qquad \hat{S}_h(f^\omega + Bu_h)(\omega) = S_h^\omega(f^\omega + Bu_h),$$

dove in entrambe le definizioni viene considerato $f^\omega + Bu_h \in V_h'$. Introducendo l'operatore aggiunto di S_h^ω, indicato come $S_h^{\omega *}$, è possibile definire la variabile aggiunta in forma discreta, rispetto alla soluzione del problema discreto $u_h^\star$

$$p_h^\omega(u_h^\star) = S_h^{\omega *}(C^* R_{\hat{V}}(z_d - C S_h^\omega(f^\omega + Bu_h^\star))). \tag{3.3.2}$$

Alla luce di queste definizioni, vale il seguente risultato.

Lemma 3.3.1. *Il problema di controllo ottimo* (3.3.1) *è ben posto ed il gradiente del funzionale obiettivo in forma ridotta è espresso da*

$$\nabla J_h(u_h^\star) = \nu u_h^\star - \mathbb{E}[p_h^\omega(u_h^\star)]. \tag{3.3.3}$$

Di conseguenza, è possibile riscrivere la condizione di ottimalità in forma discreta

$$\langle \nabla J_h(u_h^\star), v_h - u_h^\star \rangle = 0, \quad \forall v_h \in \mathcal{U}_h. \tag{3.3.4}$$

3.3.1 Stima dell'errore

Prima di completare la discretizzazione del problema occupandosi della variabile casuale, è opportuno quantificare l'errore introdotto sulle variabili di stato, aggiunta e di controllo dalla discretizzazione a elementi finiti. A tal proposito viene seguito un approccio basato sulle stime del caso deterministico dei Teoremi 3.4 and 3.5 di [9]. Siano definite la soluzione del problema continuo $u^\star$ e quella del problema discreto $u_h^\star$. Inoltre sia introdotta

$$\tilde{p}_h^\omega(u_\star) = S_h^{\omega *}(C^* R_{\hat{V}}(z_d - C S_\omega(f + Bu^\star))). \tag{3.3.5}$$

Si dimostra che vale la seguente stima [9, 17].

Lemma 3.3.2. *Si ha che*

$$\begin{aligned} \frac{\nu}{2}||u^\star - u_h^\star||^2_{L^2(D)} + \frac{1}{2}\mathbb{E}[||y^\omega(u^\star) - y_h^\omega(u_h^\star)||^2_{L^2(D)}] \leq \\ \frac{1}{2\nu}\mathbb{E}[||p^\omega(u^\star) - \tilde{p}_h^\omega(u^\star)||^2_{L^2(D)}] + \frac{1}{2}\mathbb{E}[||y^\omega(u^\star) - y_h^\omega(u^\star)||^2_{L^2(D)}]. \end{aligned} \tag{3.3.6}$$

Dimostrazione. Poichè $\nabla J_h(u_h^\star) \in U_h$, la condizione di ottimalità, scelto $v_h = u^\star$ si riduce a

$$_{\mathcal{U}'}\langle \nu(u^\star - u_h^\star) + \mathbb{E}[p^\omega(u^\star)] - \mathbb{E}[p_h^\omega(u_h^\star)], u_h^\star - u^\star \rangle_{\mathcal{U}} = 0.$$

Sommando e sottraendo nel primo membro del prodotto scalare il valore atteso di $\tilde{p}_h^{\omega}(u_\star)$

$$\mathcal{U}' \langle \nu(u^\star - u_h^\star) + \mathbb{E}[p^\omega(u^\star)] + \mathbb{E}[\tilde{p}_h^\omega(u_\star)] - \mathbb{E}[\tilde{p}_h^\omega(u_\star)] - \mathbb{E}[p_h^\omega(u_h^\star)], u_h^\star - u^\star \rangle_{\mathcal{U}} = 0.$$

da cui svolgendo il prodotto scalare, si ottiene

$$\nu ||(u^\star - u_h^\star)||_{\mathcal{U}}^2 = \mathcal{U}' \langle \mathbb{E}[p^\omega(u^\star)] + \mathbb{E}[\tilde{p}_h^\omega(u^\star)] - \mathbb{E}[\tilde{p}_h^\omega(u_\star)] - \mathbb{E}[p_h^\omega(u_h^\star)], u_h^\star - u^\star \rangle_{\mathcal{U}'}.$$

Per un ω fissato si ha l'approssimazione

$$\begin{aligned}
\mathcal{U}' \langle \tilde{p}_h^\omega(u^\star) - p_h^\omega(u_h^\star), u_h^\star - u^\star \rangle_{\mathcal{U}} =& A_\omega(y_h^\omega(u_h^\star) - y_h^\omega(u^\star), \tilde{p}_h^\omega(u^\star) - p_h^\omega(u_h^\star)) \\
=& \int_D (y_h^\omega(u_h^\star) - y_h^\omega(u^\star))(y^\omega(u_h^\star) - y_h^\omega(u_h^\star)) d\mathbf{x} \\
\leq& -\frac{1}{2} ||y^\omega(u^\star) - y_h^\omega(u_h^\star)||_{L^2(D)}^2 + \frac{1}{2} ||y^\omega(u^\star) - y_h^\omega(u^\star)||_{L^2(D)}^2.
\end{aligned}$$

Prendendo la media su tutte le realizzazioni e applicando il teorema di Fubini, si ottiene il risultato

$$\begin{aligned}
&\nu ||(u^\star - u_h^\star)||_{L^2(D)}^2 + \frac{1}{2} ||y(u^\star) - y_h(u_h^\star)||_{L^2(D)}^2 \\
&\quad \leq \mathbb{E}\left[\mathcal{U}' \langle p(u^\star) - \tilde{p}_h(u^\star), u_h^\star - u^\star \rangle_{\mathcal{U}} \right] + \frac{1}{2} \mathbb{E}[||y(u^\star) - y_h(u^\star)||_{L^2(D)}^2] \\
&\quad \leq \frac{1}{2\nu} ||p(u^\star) - \tilde{p}_h(u^\star)||_{L^2(D)}^2 + \frac{\nu}{2} ||(u_h^\star - u^\star)||_{L^2(D)}^2 + \frac{1}{2} \mathbb{E}[||y(u^\star) - y_h(u^\star)||_{L^2(D)}^2].
\end{aligned}$$

□

Questo risultato può essere esteso tramite le ipotesi di coercività e le assunzioni sulla forma bilineare al controllo della norma $H_0^1(D)$ dell'errore di discretizzazione sulla variabile di stato, come stabilito dal seguente lemma [9, 17].

Lemma 3.3.3. *Si dimostra che esiste una costante $C > 0$ indipendente dal passo di discretizzazione h tale che*

$$\begin{aligned}
&||u^\star - u_h^\star||_{L^2(D)}^2 + \mathbb{E}[||y(u^\star) - y_h(u_h^\star)||_{L^2(D)}^2] + h^2 \mathbb{E}[||y(u^\star) - y_h(u_h^\star)||_{H_0^1(D)}^2] \\
&\leq C \{ \mathbb{E}[||p(u^\star) - \tilde{p}_h(u^\star)||_{L^2(D)}^2] + \mathbb{E}[||y(u^\star) - y_h(u^\star)||_{L^2(D)}^2] + h^2 \mathbb{E}[||y(u^\star) - y_h(u^\star)||_{H_0^1(D)}^2] \}.
\end{aligned} \tag{3.3.7}$$

Per concludere questa trattazione, si applica una stima del secondo membro della condizione di quest'ultimo lemma, ottenuta supponendo adeguate ipotesi di differenziabilità delle variabili in esame.

Corollario 3.3.1. *Siano $y(u^\star), p(u^\star) \in L_{\mathbb{P}}^2(\Omega; V)$, con $V = H^{r+1}(D)$. Allora vale la seguente stima*

$$\begin{aligned}
&||u^\star - u_h^\star||_{L^2(D)}^2 + \mathbb{E}[||y(u^\star) - y_h(u_h^\star)||_{L^2(D)}^2] + h^2 \mathbb{E}[||y(u^\star) - y_h(u_h^\star)||_{H_0^1(D)}^2] \\
&\leq C h^{2r+2} \{ \mathbb{E}[|y(u^\star)|_V^2] + \mathbb{E}[|p(u^\star)|_V^2] \}.
\end{aligned} \tag{3.3.8}$$

Nel caso in esame durante le simulazioni numeriche, è opportuno considerare un caso notevole, descritto nella seguente osservazione.

Osservazione 3.3.1. *Se le variabili appartengono allo spazio di Sobolev $H^2(D)$ e lo spazio ad elementi finiti utilizza polinomi continui e affini a tratti, allora si osserva che $r = 1$ e quindi l'esponente del passo di discretizzazione a secondo membro risulta pari a 4. Di conseguenza, il comportamento atteso dell'errore di discretizzazione spaziale $||u^\star - u_h^\star||_{L^2(D)}$ è rappresentato da un decadimento del secondo ordine.*

3.4 Metodi di approssimazione in probabilità

In questa sezione vengono presentate alcune strategie di approssimazione di quantità stocastiche, con particolare riferimento al valore atteso di una generica variabile casuale. In particolare sono analizzati metodi basati sul campionamento, che approssimano il valore atteso sommando una serie di realizzazioni della variabile casuale. Un'alternativa è rappresentata dai metodi di tipo Stochastic Galerkin che discretizzano il problema in probabilità in maniera analoga alla discretizzazione spaziale.

Prima di procedere all'analisi di ogni metodo di discretizzazione in probabilità, è necessaria una fondamentale assunzione per la risoluzione numerica del problema di controllo ottimo.

Assunzione 3.4.1. *Si assume che la stocasticità inclusa nell'equazione differenziale alle derivate parziali dipenda da un numero finito di variabili casuali scorrelate a valori reali. Pertanto si ha*

$$\xi(\omega) = (\xi_1(\omega), \dots, \xi_N(\omega)), \tag{3.4.1}$$

dove $\xi_i(\omega) : \Omega \to \mathbb{R}, \quad i = 1, \dots, N$.

Quindi è possibile riscrivere il problema di controllo ottimo nella forma semi-discreta, esplicitando l'introduzione della stima del valore atteso, indicata con $\hat{\mathbb{E}}[X] = \sum_{i=1}^{N} z_i X(\zeta_i)$, $i = 1, \dots, N$, dove ζ_i sono nodi di quadratura rappresentati da N realizzazioni della variabile casuale. Alcune possibili scelte per tale stima sono presentate nel seguito.
Il problema di controllo ottimo viene riscritto applicando come segue la discretizzazione in probabilità

$$\begin{aligned} \min_{u \in \mathcal{U}} J(u) &= \frac{1}{2}\hat{\mathbb{E}}||(y(u,\zeta_i) - z_d)||^2_{L^2(D)} + \frac{\nu}{2}||u||^2_{\mathcal{U}} \\ \text{s.t.} \quad A(y(u,\zeta_i), v) &= {}_{V'}\left\langle (f(\zeta_i) + Bu), v \right\rangle_V, \quad \forall v \in V. \end{aligned} \tag{3.4.2}$$

3.4.1 Metodo Monte Carlo

I metodi di tipo Monte Carlo si propongono di approssimare il valore atteso della variabile casuale

$$X : \Omega \to \mathbb{R}, \qquad \mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega), \tag{3.4.3}$$
$$\omega \mapsto X(\omega)$$

tramite N realizzazioni scelte in maniera casuale, indicate con ζ_i, ognuna associata allo stesso peso $1/N$. Valutando X su questi nodi si imposta una formula di quadratura per approssimare l'integrale del valore atteso

$$\mathbb{E}[X] \approx \hat{\mathbb{E}}[X] = \frac{1}{N}\sum_{i=1}^{N} X(\zeta_i). \tag{3.4.4}$$

Stima dell'errore

Relativamente a questa stima del valore atteso della variabile casuale, è necessario fornire una misura dell'errore introdotto sulla variabile di controllo. Sia definita la soluzione del problema semi-discreto $\hat{u}^\star$, si dimostra che la stima ottenuta sull'errore per il metodo Monte Carlo viene espressa dal seguente lemma [17, 18].

Lemma 3.4.1. *Definita, oltre ad $\hat{u}^\star$, la soluzione del problema continuo $u^\star$ si ha*

$$\frac{\nu}{2}\mathbb{E}[||\hat{u}^\star - u^\star||^2_{L^2(D)}] + \mathbb{E}[\hat{\mathbb{E}}[||y(u^\star) - y(\hat{u}^\star)||^2_{L^2(D)}]] \leq \frac{1}{N}\frac{1}{2\nu}\mathbb{E}[||p(u^\star)||^2_{L^2(D)}]. \tag{3.4.5}$$

Dimostrazione. Sia $p^\omega(\hat{u}^\star)$ la soluzione del problema differenziale

$$A_\omega(v, p^\omega(\hat{u}^\star)) = \langle v, y^\omega(\hat{u}^\star) - z_d\rangle\,, \quad \forall v \in V, \quad \mathbb{P} - q.o.\,\omega \in \Omega.$$

Scrivendo la condizione di ottimalità sul gradiente per il valore atteso approssimato

$$\left\langle \nabla \hat{J}(\hat{u}^\star), v - \hat{u}^\star \right\rangle = 0,\ \forall v \in \mathcal{U}, \qquad \nabla \hat{J}(\hat{u}^\star) = \nu\hat{u}^\star + \hat{\mathbb{E}}[p(\hat{u}^\star)],$$

e sottraendo termine a termine tale espressione con quella analoga per il problema continuo, si ottiene

$$\nu||u^\star - \hat{u}^\star||^2_{L^2(D)} = \left\langle \mathbb{E}[p(u^\star)] - \hat{\mathbb{E}}[p(u^\star)], \hat{u}^\star - u^\star \right\rangle + \left\langle \hat{\mathbb{E}}[p(u^\star)] - \hat{\mathbb{E}}[p(\hat{u}^\star)], \hat{u}^\star - u^\star \right\rangle. \tag{3.4.6}$$

Per quanto riguarda il primo termine del secondo membro, vale la stima

$$\left\langle \mathbb{E}[p(u^\star)] - \hat{\mathbb{E}}[p(u^\star)], \hat{u}^\star - u^\star \right\rangle \leq \frac{1}{2\nu}||\mathbb{E}[p(u^\star)] - \hat{\mathbb{E}}[p(u^\star)]||^2_{L^2(D)} + \frac{\nu}{2}||\hat{u}^\star - u^\star||^2_{L^2(D)}.$$

Per il secondo termine si osserva che

$$\begin{aligned}\langle \hat{u}^\star - u^\star, p^{\omega_i}(u^\star) - p^{\omega_i}(\hat{u}^\star)\rangle =& A_{\omega_i}(y^{\omega_i}(\hat{u}^\star) - y^{\omega_i}(u^\star), p^{\omega_i}(u^\star) - p^{\omega_i}(\hat{u}^\star)) \\ =& \langle y^{\omega_i}(u^\star) - y^{\omega_i}(\hat{u}^\star), y^{\omega_i}(\hat{u}^\star) - y^{\omega_i}(u^\star)\rangle \\ =& -||y^{\omega_i}(u^\star) - y^{\omega_i}(\hat{u}^\star)||^2_{L^2(D)},\end{aligned}$$

pertanto risulta, applicando il valore atteso

$$\left\langle \hat{u}^\star - u^\star, \hat{\mathbb{E}}[p(u^\star)] - \hat{\mathbb{E}}[p(\hat{u}^\star)] \right\rangle = -\hat{\mathbb{E}}[||y(u^\star) - y(\hat{u}^\star)||^2_{L^2(D)}].$$

In definitiva, considerando il valore atteso della (3.4.6), oltre alla proprietà di non distorsione dello stimatore Monte Carlo, si ha

$$\begin{aligned}\frac{\nu}{2}\mathbb{E}[||\hat{u}^\star - u^\star||^2_{L^2(D)}] + \mathbb{E}[\hat{\mathbb{E}}[||y(u^\star) - y(\hat{u}^\star)||^2_{L^2(D)}]] \leq & \frac{1}{2\nu}\mathbb{E}[||\mathbb{E}[p(u^\star)] - \hat{\mathbb{E}}[p(u^\star)]||^2_{L^2(D)}] \\ = & \frac{1}{2\nu}\mathbb{E}\left[||\frac{1}{N}\sum_{i=1}^{N} p^{\omega_i}(u^\star) - \mathbb{E}[p(u^\star)]||^2_{L^2(D)}\right] \\ = & \frac{1}{2\nu}\mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N} ||p^{\omega_i}(u^\star) - \mathbb{E}[p(u^\star)]||^2_{L^2(D)}\right] \\ = & \frac{1}{2\nu}\frac{1}{N}\mathbb{E}[||p(u^\star) - \mathbb{E}[p(u^\star)]||^2_{L^2(D)}] \\ = & \frac{1}{2\nu}\frac{1}{N}\mathbb{E}[||p(u^\star)||^2_{L^2(D)}].\end{aligned}$$

□

3.4.2 Metodi Stochastic Collocation

Un'alternativa alla scelta casuale dei nodi di quadratura consiste nel fissare a priori l'insieme delle realizzazioni ω_i, definendo i nodi di Gauss-Legendre su un intervallo e costruendo una griglia tramite prodotto tensoriale nel caso ogni singola osservazione sia rappresentata da un vettore.

Caso univariato

Nel caso più semplice, la variabile casuale è definita su un intervallo quale $\Omega = [-1,1]$ su cui si definisce un set di campioni ω_i, $i = 1, \ldots, N+1$ tali che $-1 \leq \omega_0 \leq \omega_1 \leq \cdots \leq \omega_N \leq 1$. Una scelta efficace è rappresentata dai nodi di quadratura di Gauss-Legendre, definiti come gli zeri dei polinomi ortogonali di Legendre, ottenuti applicando il procedimento di ortogonalizzazione di Gram-Schmidt alle funzioni monomie $x^n, n = 0,1,2,\ldots$ ed espressi in forma normalizzata

$$\mathcal{P}_n(x) = \sqrt{\frac{2n+1}{2}}P_n(x), \qquad P_n(x) = \frac{1}{2^n n!}\frac{d^n(x^2-1)^n}{dx^n}, \qquad n = 0,1,2,\ldots. \tag{3.4.7}$$

Inoltre vengono indicati con $L_n(\omega)$ i polinomi di Lagrange di grado N

$$L_n(\omega) = \prod_{n' \neq n} \frac{\omega - \omega_{n'}}{\omega_n - \omega_{n'}}, \qquad 0 \leq n \leq N. \tag{3.4.8}$$

Si definisce pertanto un operatore di interpolazione basato su tali polinomi per una generica realizzazione $\omega \in \Omega$

$$\mathcal{L}(X(\omega)) = \sum_{n=0}^{N} X(\omega_n)L_n(\omega). \tag{3.4.9}$$

Pertanto il valore atteso della variabile casuale è approssimato come

$$\mathbb{E}[X] \approx \mathbb{E}[\mathcal{L}(X(\omega))] = \sum_{n=0}^{N} \left(\int_{\Omega} L_n(\omega) d\mathbb{P}(\omega) \right) X(\omega_n) = \sum_{n=0}^{N} z_n X(\omega_n), \tag{3.4.10}$$

dove sono definiti i pesi $z_n = \int_{\Omega} L_n(\omega) d\mathbb{P}(\omega),\ n = 0, \ldots, N$.

Caso multivariato

Sia $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_M\} \in \Omega = [-1,1]^M$ una generica realizzazione della variabile casuale X di dimensione $M \geq 1$. Definito un indice $k = 1, \ldots, M$ per la generica dimensione, la formula di interpolazione si scrive come

$$\mathcal{L}_k(X(\omega^{n_k})) = \sum_{n_k=0}^{N_k} X(\omega_k^{n_k}) L_k^{n_k}(\omega_k), \qquad \omega_k^{n_k} \in \Omega_k. \tag{3.4.11}$$

Quindi l'operatore nel caso multivariato è ottenuto calcolando il prodotto tensoriale degli M operatori univariati

$$(\mathcal{L}_1 \otimes \cdots \otimes \mathcal{L}_M)(X(\omega)) = \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \cdots \sum_{n_M=0}^{N_M} X(\omega_1^{n_1}, \ldots, \omega_M^{n_M})(L_1^{n_1}(\omega_1) \otimes \cdots \otimes L_M^{n_M}(\omega_M)). \tag{3.4.12}$$

In maniera analoga al caso univariato, si definisce la formula di quadratura per il valore atteso, scegliendo nodi e pesi di Gauss-Legendre in ognuna delle M dimensioni

$$\mathbb{E}[X] \approx \mathbb{E}[(\mathcal{L}_1 \otimes \cdots \otimes \mathcal{L}_M)(X(\omega))] = \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \cdots \sum_{n_M=0}^{N_M} X(\omega_1^{n_1}, \ldots, \omega_M^{n_M}) \prod_{k=1}^{M} z_k^{n_k}. \tag{3.4.13}$$

Stima dell'errore

L'approssimazione del valore atteso tramite nodi ottenuti dal prodotto tensoriale di un set di nodi notevoli fissati a priori, può garantire una velocità di convergenza di tipo esponenziale, secondo la seguente stima, per la soluzione del problema approssimato $\hat{u}^\star$

$$||\hat{u}^\star - u^\star||_{L^2(D)} \leq C \sum_{k=1}^{M} e^{-r_k \beta_k}, \tag{3.4.14}$$

dove sono indicati con β_k il numero di nodi di quadratura relativi ad ogni realizzazione ω_k e con $\mathbf{r} = \{r_k\}_{k=1}^M$ un set di coefficienti dipendenti dalla regione olomorfa della mappa $\omega \in \Omega \to p(\omega) \in V$ sul piano complesso.
Una doverosa precisazione, tuttavia, riguarda la specifica scelta dei nodi di collocazione. In particolare la scelta del prodotto tensoriale uniforme rispetto alle dimensionalità del problema, comporta uno svantaggio relativo alla quantità di nodi necessari per l'approssimazione, che è determinata dall'espressione $N_M \coloneqq \prod_{k=1}^{M} \beta_k$. Per questo motivo si osserva che questa approssimazione, pur essendo maggiormente accurata rispetto al metodo

Monte Carlo, rischia di soffrire un'eventuale elevata dimensionalità del problema (*curse of dimensionality*). Esistono varie strategie di campionamento nell'ambito dei metodi Stochastic Collocation con l'obiettivo di ridurre l'elevato numero di nodi di quadratura richiesti dalla griglia regolare per ogni realizzazione. Un comune approccio modellistico consiste nel definire griglie sparse differenti per ogni realizzazione, al fine di avere un numero maggiore di punti per le variabili più influenti nel metodo. Questo tipo di campionamento prende il nome di *griglia di Smolyak* e viene ad esempio usato nella trattazione di [19].

3.4.3 Metodi Stochastic Galerkin

Come anticipato, i metodi di tipo Galerkin applicati al problema stocastico richiedono una discretizzazione in probabilità analoga a quella spaziale. Pertanto vengono definiti i sottospazi finito-dimensionali $V_h \subset V$ per la parte spaziale e $P_K \subset L^2(\Omega)$ per quella in probabilità con le relative basi $\{\phi_j(\mathbf{x})\}_{j=1,\dots,J}$ e $\{\psi_k(\omega)\}_{k=1,\dots,K}$. In generale, utilizzando la notazione tensoriale, la soluzione del problema differenziale sarà definita come

$$y_{SG}(\mathbf{x},\omega) = \sum_{j=1}^{J}\sum_{k=1}^{K} c_{jk}\phi_j(\mathbf{x})\psi_k(\omega), \qquad y_{SG}(\mathbf{x},\omega) \in V_h \times P_K. \tag{3.4.15}$$

Tale soluzione, con i relativi coefficienti c_{jk} viene ottenuta riscrivendo il problema differenziale nella forma debole accoppiata (3.2.3) e discretizzando rispetto alle basi introdotte

$$\begin{aligned}&\int_\Omega \int_D a(\mathbf{x},\omega)\nabla\Big(\sum_{j=1}^{J}\sum_{k=1}^{K} c_{jk}\phi_j(\mathbf{x})\psi_k(\omega)\Big)\cdot\nabla(\phi_{j'}(\mathbf{x}))\psi_{k'}(\omega)\,d\mathbf{x}d\mathbb{P}(\omega) = \\ &\int_\Omega \int_D f(\mathbf{x},\omega)\phi_j(\mathbf{x})\psi_k(\omega)\,d\mathbf{x}d\mathbb{P}(\omega), \qquad j' \in 1,\dots,J, \qquad k' \in 1,\dots,K.\end{aligned} \tag{3.4.16}$$

L'integrazione è approssimata numericamente tramite formule di quadratura, tenendo presente che non è obbligatorio utilizzare stessi nodi e pesi per gli integrali rispetto alla misura spaziale e per quelli in probabilità.

3.5 Trattazione numerica del problema discreto

In questa sezione viene presentata la forma completamente discreta del problema in esame, unendo la discretizzazione spaziale con quella in probabilità, ottenuta tramite una formula di quadratura generica $\hat{\mathbb{E}}$. Il problema risultante è

$$\begin{aligned}\min_{u_h\in\mathcal{U}_h} J_h(u_h) &= \frac{1}{2}\hat{\mathbb{E}}[||(y_h^\omega(u_h) - z_d)||^2_{L^2(D)}] + \frac{\nu}{2}||u_h||^2_{\mathcal{U}} \\ \text{s.t.} \quad A_\omega(y_h^\omega(u_h), v_h) &= (f^\omega + Bu_h, v_h)_{L^2(D)}, \quad \forall v_h \in V_h.\end{aligned} \tag{3.5.1}$$

La risoluzione numerica di questo problema richiede l'applicazione di un metodo iterativo quale lo *steepest descent* o il *metodo di Newton* rispetto al sistema di condizioni di ottimalità ricavate in forma discreta

$$\begin{cases} A_{\omega_i}(y_h^{\omega_i}(u_h^\star), \phi_h) = \langle (f^\omega + Bu_h^\star), \phi_h\rangle, & \forall \phi_h \in V_h, \quad i = 1,\dots,N, \\ A_{\omega_i}(\phi_h, p_h^{\omega_i}(u_h^\star)) = \langle \phi_h, y_h^{\omega_i}(u_h^\star) - z_d\rangle, & \forall \phi_h \in V_h, \quad i = 1,\dots,N, \\ \nabla J_h(u_h^\star) = \nu u_h^\star - \hat{\mathbb{E}}[p_h^{\omega_i}(u_h^\star)] = 0, \end{cases} \tag{3.5.2}$$

con soluzione generata dal metodo $u_h^\star$.
I metodi iterativi sopracitati vengono implementati analogamente agli algoritmi del capitolo 1, tenendo presente la discretizzazione in probabilità del valore atteso. L'aggiunta di incertezza provoca un considerevole aumento del costo computazionale dei metodi, in quanto aumenta il numero di equazioni differenziali alle derivate parziali da risolvere, proporzionalmente alla dimensione del campione. La costruzione del sistema lineare di equazioni algebriche, basato sulle condizioni di ottimalità, al quale applicare i metodi iterativi, segue dalla rappresentazione del sistema (3.5.2) in forma espansa, rispetto alle funzioni base selezionate, come già mostrato nel caso deterministico. Supponendo pari a N la grandezza del campione considerato, sia l'equazione di stato che l'equazione aggiunta devono essere risolte N volte, una per ciascuna realizzazione della variabile casuale. Di conseguenza, vengono costruite N matrici di rigidezza associate alla forma bilineare $A_{\omega_i}(\cdot,\cdot)$, oltre a corrispondenti N termini noti associati alla forzante f^{ω_i}. Applicando la discretizzazione ad elementi finiti, si ottiene la seguente rappresentazione, per ciascuno degli N campioni

$$\begin{aligned}
&\text{Equazione di stato} && \mathbb{K}_i \mathbf{y}_i = \mathbf{f}_i + \mathbb{M}\mathbf{u}, && i = 1,\dots,N,\\
&\text{Equazione aggiunta} && \mathbb{K}_i^\top \mathbf{p}_i = -\mathbb{M}(\mathbf{y}_i - \mathbf{z_d}), && i = 1,\dots,N,\\
&\text{Condizione di ottimalità} && \nu\mathbb{M}\mathbf{u} - \mathbb{M}^\top \mathbb{E}[\mathbf{p}] = 0.
\end{aligned}$$

Raggruppando le componenti algebriche del problema in forma matriciale si ottiene il seguente sistema definito a blocchi

$$\begin{bmatrix}
\mathbb{M} & & & & & \mathbb{K}_1^\top & & \\
& \ddots & & & & & \ddots & \\
& & \mathbb{M} & & & & & \mathbb{K}_N^\top \\
& & & \nu\mathbb{M} & -\mathbb{M} & \dots & -\mathbb{M} & \\
\mathbb{K}_1 & & & -\mathbb{M} & & & & \\
& \ddots & & \vdots & & & & \\
& & \mathbb{K}_N & -\mathbb{M} & & & &
\end{bmatrix}
\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \\ \mathbf{u} \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_N \end{bmatrix}
=
\begin{bmatrix} \mathbf{z_d} \\ \vdots \\ \mathbf{z_d} \\ 0 \\ \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_N \end{bmatrix}. \tag{3.5.3}$$

Una strategia di risoluzione consiste nell'eliminare la variabile di stato e l'aggiunta risolvendo $2N$ equazioni differenziali alle derivate parziali, quindi, definite le matrici diagonali a blocchi

$$\mathcal{K} = diag(\mathbb{K}_1,\dots,\mathbb{K}_N), \quad \mathcal{M} = diag(\mathbb{M},\dots,\mathbb{M}), \quad n = 1,\dots,N, \tag{3.5.4}$$

con dimensioni $\mathcal{K},\mathcal{M} \in \mathbb{R}^{N_v\cdot N \times N_v\cdot N}$ (si ricorda che N_v è la dimensione dello spazio finito-dimensionale V_h), oltre al vettore di matrici

$$M = (\mathbb{M},\dots,\mathbb{M}), \tag{3.5.5}$$

con dimensione $M \in \mathbb{R}^{N_v \times N_v\cdot N}$ è possibile riscrivere il sistema algebrico in termini della sola variabile di controllo

$$\mathcal{G}\mathbf{u} = \mathbf{d}, \tag{3.5.6}$$

dove sono definiti

$$\begin{aligned}\mathcal{G} =&\nu N\mathbb{M} + M^T\mathcal{K}^{-T}\mathcal{M}\mathcal{K}^{-1}M,\\ \mathbf{d} =&M^T\mathcal{K}^{-T}(\mathbf{z}_d - \mathcal{M}\mathcal{K}^{-1}\mathbf{f}).\end{aligned} \tag{3.5.7}$$

Il sistema viene quindi risolto utilizzanto metodi *gradient-based*, richiedendo un costo computazionale di $2N$ PDE risolte ad ogni iterazione. L'algoritmo implementato per l'utilizzo del metodo iterativo del gradiente per la risoluzione del sistema lineare algebrico è presentato in Algoritmo 4.

Algoritmo 4 Metodo del gradiente per il problema stocastico

Input: $\mathbf{u}_0$, $\mathbf{u}_{ref}$, $\mathbf{z_d}$, K_{max}, tol, ν, N

1: Definizione di N realizzazioni del coefficiente di diffusione ed eventualmente della forzante
2: Costruzione delle matrici ad elementi finiti del sistema $\mathbb{M}, \mathbb{K}_n$, oltre ai termini noti $\mathbf{f}_n$
3: $k \leftarrow 0$
4: **while** $k \leq K_{max}$ and $||\nabla J_h(\mathbf{u}_k)|| \geq tol$ **do**
5: **for** $n = 1, \dots, N$ **do**
6: Risoluzione di $\mathbb{K}_n\mathbf{y}_n = \mathbf{f}_n + \mathbb{M}\mathbf{u}_k$
7: Risoluzione di $\mathbb{K}_n^\top\mathbf{p}_n = -\mathbb{M}(\mathbf{y}_n - \mathbf{z_d})$
8: Aggiornamento $\nabla J_h(\mathbf{u}_k) += \mathbb{M}^\top\mathbf{p}_n$
9: **end for**
10: $\nabla J_h(\mathbf{u}_k) \leftarrow -\frac{1}{N}\nabla J_h(\mathbf{u}_k) + \nu\mathbf{u}_k$
11: Calcolo della direzione $\mathbf{d}_k = -\nabla J_h(\mathbf{u}_k)$
12: Calcolo di τ_k tramite *line-search*
13: $\mathbf{u}_{k+1} \leftarrow \mathbf{u}_k + \tau_k\mathbf{d}_k$
14: $k \leftarrow k + 1$
15: **end while**
16: Valutazione dell'errore rispetto ad una soluzione di riferimento $\mathbf{u}_{ref}$

3.5.1 Stima dell'errore complessivo

A conclusione della presentazione del problema stocastico, viene fornita una stima completa dell'errore di discretizzazione, dovuto sia all'approssimazione introdotta dal metodo iterativo utilizzato, sia dalle discretizzazioni in spazio e in probabilità. Considerando un funzionale obiettivo fortemente convesso e Lipschitziano, è possibile dimostrare il seguente risultato [17], propedeutico alla stima dell'errore, considerando ad esempio il metodo iterativo del gradiente per la risoluzione del sistema algebrico.

Lemma 3.5.1. *Sia $u^\star$ la soluzione esatta del problema differenziale continuo. Definite le iterazioni del metodo $\{u_j\}_{j\in\mathbb{N}}$, si ha che, introdotte $l = 2\nu$ e la costante di Lipschitz L, per $0 \leq \tau \leq l/L^2$*

$$||u_{j+1}-u^\star||^2_{L^2(D)} \leq (1-\tau l+\tau^2L^2)||u_j-u^\star||^2_{L^2(D)} \leq (1-\tau l+\tau^2L^2)^{j+1}||u_0-u^\star||^2_{L^2(D)}. \tag{3.5.8}$$

Inoltre $||u_j - u^\star||^2_{L^2(D)} \to 0$, per $j \to \infty$.

Dimostrazione. Applicando la formula ricorsiva e ricordando che $\nabla J(u^\star) = 0$, si ha

$$u_{j+1} - u^\star = u_j - u^\star - \tau \mathbb{E}[\nabla\psi(u_j, \omega) - \nabla\psi(u^\star, \omega)],$$

dove $\psi(u, \omega) = 1/2||y^\omega(u) - z_d||^2_{L^2(D)} + \nu/2||u||^2_{L^2(D)}$.
Di conseguenza si ha

$$\begin{aligned} ||u_{j+1} - u^\star||^2_{L^2(D)} =& ||u_j - u^\star||^2_{L^2(D)} + \tau^2 ||\mathbb{E}[\nabla\psi(u_j, \omega) - \nabla\psi(u^\star, \omega)]||^2_{L^2(D)} \\ & - 2\tau \langle u_j - u^\star, \mathbb{E}[\nabla\psi(u_j, \omega) - \nabla\psi(u^\star, \omega)]\rangle \\ \leq& (1 - \tau l + \tau^2 L^2)||u_j - u^\star||^2_{L^2(D)}. \end{aligned}$$

In conclusione, applicando la disuguaglianza fino ad u_0 e osservando che la condizione imposta su τ implica che $0 \leq 1 - \tau l + \tau^2 L^2 \leq 1$ si ha la tesi, che vale per $\mathbb{P}$-q.o. $\omega \in \Omega$. □

A partire da questo lemma, nel teorema seguente [17, 18] viene fornita in maniera dettagliata una stima dell'errore di discretizzazione.

Teorema 3.5.1. *Sia definita la soluzione approssimata ottenuta dalle iterazioni del metodo iterativo $\hat{u}^j_h$. Allora $\exists\, C_1, C_2, C_3 > 0$, indipendenti dal passo di discretizzazione h e dal numero di campioni N tali che*

$$\mathbb{E}[||\hat{u}^j_h - u^\star||^2_{L^2(D)}] \leq C_1 e^{-\rho j} + \frac{C_2}{N} + C_3 h^{2r+2}, \tag{3.5.9}$$

dove $\rho = -2\log(1 - \tau l + \tau^2 L^2)$ è una costante.

Dimostrazione. La stima dell'errore di discretizzazione può essere decomposta in 3 contributi, relativi rispettivamente al metodo iterativo di risoluzione del sistema algebrico, alla discretizzazione spaziale e all'approssimazione in probabilità. Pertanto

$$\mathbb{E}[||\hat{u}^j_h - u^\star||^2_{L^2(D)}] \leq 3\mathbb{E}[||\hat{u}^j_h - \hat{u}^\star_h||^2_{L^2(D)}] + 3\mathbb{E}[||\hat{u}^\star_h - u^\star_h||^2_{L^2(D)}] + 3\mathbb{E}[||u^\star_h - u^\star||^2_{L^2(D)}].$$

Riprendendo i risultati ottenuti nelle sezioni precedenti, è stato già dimostrato il limite relativo al secondo termine, controllato da $\frac{C_2}{N}$, per il metodo Monte Carlo, e al terzo, limitato da $C_3 h^{2r+2}$, riferito alla convergenza rispetto al passo spaziale. La dimostrazione si conclude osservando che, per il metodo di Newton, si ha, grazie al lemma precedente

$$||\hat{u}^j_h - \hat{u}^\star_h||^2_{L^2(D)} \leq M^{2j}||\hat{u}^0_h - \hat{u}^\star_h||^4_{L^2(D)} = e^{-\rho j}||\hat{u}^0_h - \hat{u}^\star_h||^4_{L^2(D)},$$

opportunamente definita la costante ρ.
Considerando il valore atteso di entrambi i membri di tale disuguaglianza, segue il risultato richiesto

$$\mathbb{E}[||\hat{u}^j_h - \hat{u}^\star_h||^2_{L^2(D)}] \leq e^{-\rho j}\mathbb{E}[||\hat{u}^0_h - \hat{u}^\star_h||^2_{L^2(D)}] = C_1 e^{-\rho j}.$$

□

3.6 Problemi di controllo al bordo

Un'interessante applicazione alternativa della teoria del controllo ottimo in condizioni di incertezza è rappresentata dal caso particolare in cui la variabile di controllo è definita solamente sul bordo del dominio, o su parte di esso. Specificatamente, in questa analisi il controllo agisce sul sistema per mezzo di una condizione di Neumann applicata all'equazione differenziale alle derivate parziali ellittica in esame.
Siano pertanto definiti la variabili di stato $y(\mathbf{x},\omega) \in L^2_{\mathbb{P}}(\Omega; H^1_d(D))$, analogamente al caso distribuito, e quella di controllo $u \in L^2(\partial D_n)$ sul bordo di Neumann $\partial D_n \subset \partial D$. Mantenendo le definizioni e proprietà di $a(\cdot,\cdot)$ ed $f(\cdot)$ introdotte nella (3.2.1), si ottiene il problema al contorno

$$\begin{cases} -\nabla \cdot (a(\mathbf{x},\omega)\nabla y(\mathbf{x},\omega)) = f(\mathbf{x},\omega) & \text{in } D \times \Omega, \\ y(\mathbf{x},\omega) = 0 & \text{su } \partial D_d \times \Omega, \\ \dfrac{\partial y(\mathbf{x},\omega)}{\partial n} = u(\mathbf{x}) & \text{su } \partial D_n \times \Omega. \end{cases} \tag{3.6.1}$$

Definite le funzioni test $v \in L^2_{\mathbb{P}}(\Omega; H^1_d(D))$ è possibile scrivere la forma variazionale del problema

$$\int_\Omega \int_D a(\mathbf{x},\omega)\nabla y \nabla v \, d\mathbf{x} d\mathbb{P}(\omega) = \int_\Omega \int_D f(\mathbf{x},\omega) v \, d\mathbf{x} d\mathbb{P}(\omega) + \int_\Omega \int_{\partial D_n} u(\mathbf{x}) v \, d\mathbf{x} d\mathbb{P}(\omega). \tag{3.6.2}$$

La corrispondente forma astratta rimane analoga alla (3.2.3), ridefinendo il funzionale $F(v) : L^2_{\mathbb{P}}(\Omega; H^1_d(D)) \to \mathbb{R}$ in modo tale che corrisponda al secondo membro dell'equazione variazionale. Come nel caso con controllo distribuito, il problema di ottimizzazione mantiene la formulazione in termini operatoriali data dalla (3.2.8), tuttavia è necessario specificare come varia la natura dell'operatore di controllo, relativamente al suo spazio vettoriale di definizione. Infatti, mentre nel caso distribuito la matrice associata all'operatore di controllo corrisponde alla matrice di massa relativa alla PDE, quando il controllo è localizzato sul bordo tale matrice cessa di essere quadrata e simmetrica, dovendo proiettare il vettore associato alla variabile di controllo, con lunghezza pari al numero di nodi del bordo di Neumann N_n, sullo spazio vettoriale finito-dimensionale di riferimento della variabile di stato, che comprende, oltre ai nodi sul bordo, anche i gradi di libertà interni. Concretamente, definita la matrice di massa relativa ad una discretizzazione 1D del bordo di Neumann $B_0 \in \mathbb{R}^{N_n \times N_n}$, è possibile considerare la matrice di controllo

$$\mathbb{B} = \begin{bmatrix} \mathbb{O} \\ B_0 \end{bmatrix}, \qquad \mathbb{B} \in \mathbb{R}^{(N_v+N_n)\times N_n}, \qquad \mathbb{O} \in \mathbb{R}^{N_v \times N_n}, \tag{3.6.3}$$

dove N_v è il numero di gradi di libertà interni mentre $\mathbb{O}$ è la matrice identicamente nulla. Pertanto si ottiene il set di condizioni di ottimalità

$$\begin{aligned} &\text{Equazione di stato} && \mathbb{K}_i \mathbf{y}_i = \mathbf{f}_i + \mathbb{B}\mathbf{u}, && i = 1, \dots, N, \\ &\text{Equazione aggiunta} && \mathbb{K}_i^\top \mathbf{p}_i = -\mathbb{M}(\mathbf{y}_i - \mathbf{z_d}), && i = 1, \dots, N, \\ &\text{Condizione di ottimalità} && \nu B_0 \mathbf{u} - \mathbb{B}^\top \mathbb{E}[\mathbf{p}] = 0. \end{aligned}$$

Di conseguenza, è possibile scrivere queste condizioni tramite un sistema lineare algebrico completo

$$\begin{bmatrix} \mathbb{M} & & & & \mathbb{A}_1^\top & & \\ & \ddots & & & & \ddots & \\ & & \mathbb{M} & & & & \mathbb{A}_N^\top \\ & & & \nu B_0 & -\mathbb{B}^\top & \dots & -\mathbb{B}^\top \\ \mathbb{A}_1 & & & -\mathbb{B} & & & \\ & \ddots & & \vdots & & & \\ & & \mathbb{A}_N & -\mathbb{B} & & & \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \\ \mathbf{u} \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_N \end{bmatrix} = \begin{bmatrix} \mathbf{z_d} \\ \vdots \\ \mathbf{z_d} \\ 0 \\ \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_N \end{bmatrix}. \tag{3.6.4}$$

Ancora una volta, costruite le suddette matrici, è necessario implementare un metodo iterativo quale steepest descent, o il metodo di Newton, per determinare il controllo ottimo e il relativo stato.
L'analisi della convergenza e la stima dell'errore complessivo seguono in maniera analoga al caso distribuito, con le opportune modifiche legate alla diversa natura dell'operatore di controllo.

Capitolo 4

Problemi di controllo *Risk-Averse*

Questo capitolo è dedicato all'analisi di particolari problemi di controllo ottimo in condizioni di incertezza, con l'obiettivo di minimizzare il rischio associato al verificarsi di eventi rari, matematicamente rappresentati da realizzazioni della variabile casuale lontane dalla media campionaria. In letteratura esistono varie scelte in termini di funzionale obiettivo o di vincoli per modellizzare questo comportamento. Un esempio significativo, studiato da D. Kouri e T. Surowiec [11] è rappresentato dal *Conditional Value-at-Risk (CVaR)*, ossia il valore atteso di una variabile aleatoria calcolata sulle realizzazioni appartenenti ad una coda della distribuzione di probabilità. Il problema in esame in [11] consiste nella minimizzazione di un funzionale obiettivo composto dalla sopracitata misura di rischio rispetto ad una funzione della variabile di stato sommato al consueto termine di regolarizzazione sulla norma della variabile di controllo, già utilizzato nei capitoli precedenti. L'obiettivo del presente lavoro sarà l'analisi di una versione modificata di questo problema, in cui il funzionale da minimizzare è quello definito per il problema stocastico di controllo robusto, oppure un funzionale risk-averse. Inoltre la misura CVaR viene introdotta come vincolo, imponendovi una soglia limite. Due differenti approcci saranno quindi discussi per la risoluzione numerica del problema, sempre basandosi sulle discretizzazioni in spazio e in probabilità descritte precedentemente: nel primo si implementa una riformulazione dei vincoli tramite un set di variabili ausiliarie, mentre nel secondo viene implementato un metodo primal-dual interior point, ricavando le condizioni KKT e trattando la variabile aggiuntiva introdotta dal CVaR come una funzione implicita della variabile di controllo.

4.1 Il Conditional Value-at-Risk

Prima di addentrarsi nell'analisi del problema di controllo ottimo risk-averse è necessario definire il concetto matematico di *Conditional Value-at-Risk (CVaR)*, oltre a darne un'interpretazione numerica in forma discreta. A tal proposito, verrà seguito l'approccio proposto da T. Rockafellar e S. Uryasev [24, 25], introdotto per lo studio di problemi relativi all'ottimizzazione di un portafoglio di investimenti finanziari con l'obiettivo di

ridurre il rischio di gravi perdite.

Definizione

Sia definita una funzione obiettivo $\mathcal{F} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, che dipende dal vettore di controllo $\mathbf{u} \in \mathbb{R}^n$ e da un vettore casuale $\boldsymbol{\omega} \in \mathbb{R}^m$, con densità di probabilità $p_{\mathbf{u}}(\boldsymbol{\omega})$. Si definisce pertanto la funzione di probabilità cumulata rispetto ad una soglia t

$$\Psi(\mathbf{u}, t) = \int_{\mathcal{F}(\mathbf{u},\boldsymbol{\omega})\leq t} p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{4.1.1}$$

Pertanto, è possibile definire il β-quantile della distribuzione, che assume l'equivalente definizione di *Value-at-Risk (VaR)*, tramite la funzione $t(\mathbf{u}, \beta)$

$$\mathrm{VaR}_\beta(\mathcal{F}(\mathbf{u}, \cdot)) = t(\mathbf{u}, \beta) = \min\{t \in \mathbb{R} \mid \Psi(\mathbf{u}, t) \geq \beta\}. \tag{4.1.2}$$

Infine, il Conditional Value-at-Risk viene definito come il valore atteso della funzione obiettivo, condizionato al superamento della soglia definita dal Value-at-Risk

$$\begin{aligned} \mathrm{CVaR}_\beta(\mathcal{F}(\mathbf{u}, \cdot)) = \mathbb{E}[\mathcal{F}(\mathbf{u}, \cdot) | \mathcal{F}(\mathbf{u}, \cdot) \geq \mathrm{VaR}_\beta(\mathcal{F}(\mathbf{u}, \cdot))] = \\ \frac{1}{1-\beta} \int_{\mathcal{F}(\mathbf{u},\boldsymbol{\omega})\geq t(\mathbf{u},\beta)} \mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \end{aligned} \tag{4.1.3}$$

Una difficoltà nella caratterizzazione numerica di queste quantità consiste nella non-differenziabilità del β-quantile, oltre alla complessità concettuale della sua definizione, pertanto risulta necessario considerare una funzione, dipendente sia dal vettore di controllo $\mathbf{u}$ che dal parametro α, che sia, almeno sotto determinate ipotesi, differenziabile e la cui minimizzazione sia equivalente al CVaR in termini di significato matematico

$$F(\mathbf{u}, t) = t + \frac{1}{1-\beta} \int_{\mathcal{F}(\mathbf{u},\boldsymbol{\omega})\geq t} (\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) - t) p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{4.1.4}$$

Equivalentemente, definita la parte positiva come $(\cdot)^+ = \max(\cdot, 0)$, è possibile riscrivere tale funzione come

$$F(\mathbf{u}, t) = t + \frac{1}{1-\beta} \int_{\boldsymbol{\omega}\in\mathbb{R}^m} (\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) - t)^+ p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{4.1.5}$$

Si dimostra che la funzione F così definita risulta differenziabile con continuità e convessa sia rispetto a t che a $\mathbf{u}$, sotto l'ipotesi di convessità del funzionale obiettivo $\mathcal{F}$ e supponendo che la densità p dipenda solamente dalla variabile $\boldsymbol{\omega}$ [28]. Per dimostrare l'equivalenza tra la minimizzazione del CVaR e quella della funzione F, sono utili i due seguenti risultati [24, 25].

Teorema 4.1.1. *Sia $\Psi(\mathbf{u}, t)$ continua rispetto a t per qualunque $\mathbf{u} \in U \subset \mathbb{R}^n$, dove U è il feasible set. Sia definito l'insieme dei valori di t per cui la distribuzione cumulata assume valore pari al quantile*

$$\Theta(\mathbf{u}, \beta) = \{t \mid \Psi(\mathbf{u}, t) = \beta\}. \tag{4.1.6}$$

Poichè $F(\mathbf{u}, t)$ è differenziabile rispetto ad α $\forall \mathbf{u} \in U \subset \mathbb{R}^n$, allora si ha

$$\begin{aligned} \Theta(\mathbf{u}, \beta) &= \{t \mid F(\mathbf{u}, t) = \min_{t \in \mathbb{R}} \{F(\mathbf{u}, t)\}\}, \\ t(\mathbf{u}, \beta) &= \min_{t \in \mathbb{R}} \{t \mid t \in \Theta(\mathbf{u}, \beta)\}. \end{aligned} \tag{4.1.7}$$

Dimostrazione. L'ipotesi di continuità rispetto a t della funzione di probabilità cumulata Ψ è necessaria per escludere la presenza di *atomi*, cioè di singoli valori con probabilità positiva, poichè implica

$$\int_{\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) = t} p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega} = 0. \tag{4.1.8}$$

Poichè F è convessa e differenziabile rispetto ad α, i suoi minimi possono essere determinati calcolando la derivata parziale corrispondente

$$\begin{aligned} \nabla_t F(\mathbf{u}, t) =& \nabla_t \left(t + \frac{1}{1-\beta} \int_{\boldsymbol{\omega} \in \mathbb{R}^m} (\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) - t)^+ p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega} \right) \\ =& 1 + \frac{1}{1-\beta} \int_{\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) \geq t} \nabla_t (\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) - t)^+ p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ =& \frac{1}{1-\beta} ((1-\beta) - (1 - \Psi(\mathbf{u}, t))) = \frac{1}{1-\beta} (\Psi(\mathbf{u}, t) - \beta). \end{aligned} \tag{4.1.9}$$

Si conclude che i minimi rispetto a t di F sono individuati dall'uguaglianza $\Psi(\mathbf{u}, t) = \beta$, che ha almeno una soluzione poichè Ψ è continua e non decrescente rispetto a t. Quindi, per definizione, il quantile è il valore minore che soddisfa tale equazione, da cui segue la seconda espressione della (4.1.7). □

Teorema 4.1.2. *Sotto le ipotesi del teorema precedente, si ha che $F(\mathbf{u}, t)$ è costante rispetto a t, quando $t \in \Theta(\mathbf{u}, \beta)$ ed inoltre*

$$F(\mathbf{u}, t) = \mathrm{CVaR}_{\beta}(\mathcal{F}(\mathbf{u}, \cdot)). \tag{4.1.10}$$

Dimostrazione. Per concludere la dimostrazione, alla luce del risultato precedente, è sufficiente stabilire che

$$\mathrm{CVaR}_{\beta}(\mathcal{F}(\mathbf{u}, \cdot)) = F(\mathbf{u}, t(\mathbf{u}, \beta)).$$

Il secondo membro di questa espressione si può scrivere come

$$F(\mathbf{u}, t(\mathbf{u}, \beta)) = t(\mathbf{u}, \beta) + \frac{1}{1-\beta} \int_{\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) \geq t(\mathbf{u}, \beta)} (\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) - t(\mathbf{u}, \beta)) p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{4.1.11}$$

Per definizione di Θ, sia ha inoltre che $\beta = \Psi(\mathbf{u}, t(\mathbf{u}, \beta))$, quindi il secondo termine della precedente equazione si può riscrivere con la seguente riformulazione del quantile

$$\begin{aligned} t(\mathbf{u}, \beta) =& t(\mathbf{u}, \beta) \frac{1}{1-\beta} (1 - \Psi(\mathbf{u}, t(\mathbf{u}, \beta))) \\ =& t(\mathbf{u}, \beta) \frac{1}{1-\beta} \int_{\mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) \geq t(\mathbf{u}, \beta)} p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \end{aligned} \tag{4.1.12}$$

Osservazione 4.1.1. *Questo artificio matematico, utile ai fini della dimostrazione, è giustificato dal fatto che* $\beta = \Psi(\mathbf{u}, t(\mathbf{u}, \beta))$*, quindi* $1-\beta = 1-\Psi(\mathbf{u}, t(\mathbf{u}, \beta))$ *ed in conclusione si ha che* $\frac{1}{1-\beta}(1 - \Psi(\mathbf{u}, t(\mathbf{u}, \beta))) = 1$*, mostrando l'equivalenza dei termini nella prima uguaglianza della* (4.1.12)*.*

Unendo questo risultato all'espressione (4.1.11) risulta infine

$$F(\mathbf{u}, t(\mathbf{u}, \beta)) = \frac{1}{1-\beta} \int_{\mathcal{F}(\mathbf{u},\boldsymbol{\omega}) \geq t(\mathbf{u},\beta)} \mathcal{F}(\mathbf{u}, \boldsymbol{\omega}) p_{\mathbf{u}}(\boldsymbol{\omega}) d\boldsymbol{\omega} = \mathrm{CVaR}_{\beta}(\mathcal{F}(\mathbf{u}, \cdot)). \tag{4.1.13}$$

□

In conclusione, è necessario mostrare che i teoremi enunciati implicano la possibilità di minimizzare $F(\mathbf{u}, t)$ in maniera equivalente al Conditional Value-at-Risk, rispetto ad un feasible set U

$$\min_{\mathbf{u}\in U, t\in\mathbb{R}} F(\mathbf{u}, t) = \min_{\mathbf{u}\in U} \min_{t\in\mathbb{R}} F(\mathbf{u}, t) = \min_{\mathbf{u}\in U} F(\mathbf{u}, t(\mathbf{u}, \beta)) = \min_{\mathbf{u}\in U} \mathrm{CVaR}_{\beta}(\mathcal{F}(\mathbf{u}, \cdot)). \tag{4.1.14}$$

Sia ora $(\mathbf{u}^\star, t^\star)$ la soluzione ottima determinata con un metodo di ottimizzazione, allora si ha che $F(\mathbf{u}^\star, t^\star) = \min_{\mathbf{u}\in U, t\in\mathbb{R}} F(\mathbf{u}, t)$. Di conseguenza $\mathbf{u}$ minimizza il CVaR su U ed il quantile ottimale è rappresentato da $t^\star$.

Discretizzazione e regolarizzazione

Nell'ottica di procedere alla risoluzione numerica del problema, è necessario fornire una rappresentazione discreta della funzione F, in riferimento ad un set di realizzazioni della variabile casuale $\boldsymbol{\omega}$. A questo proposito, definiti tali elementi come $\omega_j,\ j = 1, \ldots, N$, l'integrale dell'espressione (4.1.5) viene approssimato da una formula di quadratura di tipo Monte Carlo, ottenendo l'espressione discreta

$$\hat{F}(\mathbf{u}, t) = t + \frac{1}{1-\beta} \frac{1}{N} \sum_{j=1}^{N} (\mathcal{F}(\mathbf{u}, \omega_j) - t)^+. \tag{4.1.15}$$

Oltre all'applicazione dell'approssimazione basata sulla media campionaria per la discretizzazione in probabilità, è utile per la trattazione fornire un'espressione regolarizzata della funzione $\hat{F}$, in quanto la parte positiva $(\cdot)^+$ non risulta differenziabile. La regolarizzazione, basata sull'approccio presentato in [11], consiste nel definire un parametro sufficientemente piccolo $\epsilon > 0$ ed una funzione g_ϵ tale che

$$g_\epsilon(\cdot) \to (\cdot)^+, \qquad \epsilon \to 0.$$

Questa strategia permette di calcolare gradienti e matrici Hessiane relative alla funzione $\hat{F}$, necessarie per l'implementazione di algoritmi di ottimizzazione. La scelta della specifica funzione g_ϵ e l'analisi delle sue proprietà sono argomentate nella sezione riguardante il relativo metodo di ottimizzazione. Pertanto, la forma regolarizzata viene scritta come

$$\hat{F}_\epsilon(\mathbf{u}, t) = t + \frac{1}{1-\beta} \frac{1}{N} \sum_{j=1}^{N} g_\epsilon(\mathcal{F}(\mathbf{u}, \omega_j) - t). \tag{4.1.16}$$

4.2 Definizione del problema e trattazione numerica

In questa sezione vengono presentati nel dettaglio due possibili approcci per la risoluzione numerica del problema di controllo ottimo di un funzionale obiettivo vincolato da un'equazione differenziale alle derivate parziali e da una restrizione sul valore della misura di rischio in esame (CVaR). La prima strategia risolutiva è basata su un metodo di tipo *primal interior-point*, che trasforma il problema vincolato in un insieme di problemi non-vincolati sommando al funzionale obiettivo i vincoli per mezzo di una *funzione barriera*, regolata da un parametro μ. Ogni problema non-vincolato viene quindi risolto utilizzando il metodo di Newton fino al raggiungimento di un'adeguata accuratezza computazionale. Un'alternativa a questo metodo è rappresentata dall'approccio *primal-dual interior-point*, in cui viene considerata una trasformazione dei vincoli in sole uguaglianze tramite variabili ausiliarie. Il metodo di Newton viene quindi applicato al sistema di condizioni KKT ottenute dalla funzione Lagrangiana, introducendo opportuni moltiplicatori di Lagrange associati ai vincoli.
Il problema di controllo ottimo in esame viene scritto in forma generale come

$$\begin{aligned} \min_{u\in\mathcal{U},y\in V} \; & J(y,u) = \frac{1}{2}\mathbb{E}[||y_\omega - z_d||_V^2] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ \text{s.t. } & A_\omega(y_\omega, v) =_{V'} \langle f_\omega + u, v\rangle_V \,, \; \forall v \in V, \; \mathbb{P}-q.o.\; \omega \in \Omega, \\ & \mathrm{CVaR}_\beta(\theta(y_\omega)) \leq \tau, \end{aligned} \tag{4.2.1}$$

dove $\theta(\cdot)$ è una funzione della variabile di stato, quale ad esempio $\theta(y_\omega) = 1/2||y_\omega||_V^2$. Nella trattazione, sarà tuttavia considerata la corrispondente forma ridotta, ottenuta risolvendo preventivamente l'equazione di stato

$$\begin{aligned} \min_{u\in\mathcal{U}} \; & \hat{J}(u) = \frac{1}{2}\mathbb{E}[||S_\omega(f+Bu) - z_d||_V^2] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ \text{s.t. } & \mathrm{CVaR}_\beta(\theta(S_\omega(f+Bu))) \leq \tau. \end{aligned} \tag{4.2.2}$$

4.2.1 Esistenza delle soluzioni del problema vincolato

Prima di addentrarsi nell'analisi dei metodi numerici implementati per la risoluzione del problema di controllo ottimo vincolato, risulta necessario mostrare la buona positura del problema, garantendo l'esistenza di soluzioni, ed eventualmente l'unicità per funzionali obiettivo strettamente convessi. A tale scopo, è possibile enunciare il seguente lemma.

Lemma 4.2.1. *Le due formulazioni generalizzate del problema vincolato*

$$\begin{aligned} &\min_u \hat{J}(u) \\ &s.t. \min_t \{g(u,t)\} \leq \tau, \end{aligned} \qquad\qquad \begin{aligned} &\min_{u,t} \hat{J}(u) \\ &s.t.\; g(u,t) \leq \tau, \end{aligned}$$

ammettono rispettivamente soluzioni ottime $(u_1^\star, t_1^\star)$ *e* $(u_2^\star, t_2^\star)$.
Supponendo la stretta convessità, oltre che del funzionale obiettivo, anche di $g(u,t)$*, in generale si ha che* $u_1^\star = u_2^\star$*. Inoltre, se il vincolo è attivo le soluzioni ottime sono equivalenti.*

Dimostrazione. Come osservazione preliminare, si nota che, date le citate proprietà di stretta convessità, esiste un'unica soluzione per entrambi i problemi. Siano definiti i feasible set per ogni formulazione

$$\mathcal{F}_1 = \{u \mid \min_t g(u,t) \leq \tau\},$$
$$\mathcal{F}_2 = \{(u,t) \mid g(u,t) \leq \tau\}.$$

Per dimostrare l'equivalenza delle soluzioni rispetto alla variabile u, è necessario mostrare che $\mathcal{F}_1 = \Pi_u(\mathcal{F}_2)$, dove Π_u è l'operatore di proiezione definito da

$$\Pi_u(\mathcal{F}_2) = \{u \mid \exists t \text{ t.c. } g(u,t) \leq \tau\}.$$

- $\boxed{\mathcal{F}_1 \subseteq \Pi_u(\mathcal{F}_2)}$

  Sia $u \in \mathcal{F}_1$. Allora si ha che $\min_t g(u,t) \leq \tau$, quindi si può concludere che $\exists t_1$ t.c. $g(u,t_1) \leq \tau$ e pertanto $u \in \Pi_u(\mathcal{F}_2)$.

- $\boxed{\Pi_u(\mathcal{F}_2) \subseteq \mathcal{F}_1}$

  Sia ora $u \in \Pi_u(\mathcal{F}_2)$. Dalla definizione del feasible set si ha che $\exists t_2$ t.c. $g(u,t_2) \leq \tau$. Ma poichè $\min_t g(u,t) \leq g(u,t_2) \leq \tau$ si ha che $u \in \mathcal{F}_1$.

In definitiva, poichè i feasible set, rispetto alla variabile u, coincidono, è possibile concludere che $u_1^\star = u_2^\star := u^\star$.
Rimane da mostrare che, quando il vincolo è attivo, anche le soluzioni ottime rispetto alla variabile t sono coincidenti. Infatti, se supponiamo che $\min_t g(u^\star,t) = \tau$, allora per convessità di $g(u,t)$, $\exists t^\star$ t.c. $g(u^\star,t^\star) = \tau$, soluzione ottima per il primo problema. Poichè il funzionale obiettivo non dipende da t, è quindi immediato concludere che tale $t^\star$ sia ancora una soluzione ottimale anche per il secondo problema. □

Osservazione 4.2.1. *L'ipotesi che il vincolo su $g(u,t)$ sia attivo è strettamente necessaria per concludere che i due suddetti problemi siano equivalenti. Infatti, nel caso in cui il vincolo non sia attivo, quindi $g(u,t) < \tau$, non è garantito che il secondo problema restituisca il valore ottimale di t. Questo comportamento si spiega constatando l'indipendenza del funzionale obiettivo da t. Poichè questa variabile non influenza direttamente la funzione da minimizzare, un generico algoritmo risolutivo tenderà a scegliere un qualunque t per cui $g(u,t) < \tau$, ma non specificamente quello ottimale. In definitiva, il risultato numerico ottenuto per la variabile t avrà, nel caso dell'ottimizzazione vincolata da CVaR, il significato di quantile del funzionale di interesse $\theta_\omega(u)$ solo nel caso in cui il vincolo sia attivo.*

4.2.2 I metodi interior-point

I metodi interior-point rappresentano una efficace strategia risolutiva per problemi di ottimizzazione vincolata. La caratteristica principale di questa classe di metodi, che li differenzia da approcci alternativi, quali ad esempio i metodi *active-set*, è la stretta

ammissibilità, rispetto ai vincoli considerati, di ogni iterazione intermedia. In altre parole, il percorso seguito dal metodo, a partire da una stima iniziale *feasible*, in direzione della soluzione ottima, si sviluppa interamente all'interno del *feasible set*. A tale scopo, viene definito il parametro di barriera μ, che viene progressivamente ridotto garantendo una crescente accuratezza della soluzione, avvicinandosi al bordo del feasible set.
Al fine di chiarire la strategia risolutiva adottata da questo metodo di ottimizzazione, è opportuno studiarne l'applicazione ad un problema modello, che rappresenta una generalizzazione di quello in esame, come ad esempio

$$\begin{aligned} &\min_x \ f(x) \\ &\text{s.t. } g_i(x) \leq 0, \qquad i = 1, \ldots, N. \end{aligned} \tag{4.2.3}$$

La trattazione numerica di questo problema, può essere condotta sia in maniera diretta, seguendo l'approccio *primal interior-point*, oppure tramite l'aggiunta di variabili fittizie per trasformare i vincoli di disuguaglianza in uguaglianze e la definizione del problema duale (approccio *primal-dual interior-point*).

Il metodo primal interior-point Il metodo di risoluzione diretta del problema primale rappresenta un primo tentativo di trattare l'ottimizzazione non-lineare vincolata, introdotto inizialmente da Frisch [7], quindi perfezionato da Fiacco e McCormick [6] negli anni '60. La strategia risolutiva consiste nel riformulare il problema vincolato in una famiglia di problemi non-vincolati, indicizzata dal parametro di barriera μ, includendo il vincolo nel funzionale obiettivo, tramite l'utilizzo di una *funzione barriera* di tipo logaritmico

$$\phi(x) = -\sum_{i=1}^{N} \log(-g_i(x)). \tag{4.2.4}$$

Il set di problemi non-vincolati risultante viene pertanto definito come

$$\min_x \ \{f_\mu(x) = f(x) + \mu\phi(x)\}. \tag{4.2.5}$$

La trattazione numerica proposta per questi problemi di ottimizzazione prevede l'applicazione di un metodo iterativo di Newton per una sequenza decrescente di valori del parametro di barriera. Per costruire tale metodo, è necessario calcolare gradiente e matrice Hessiana del funzionale obiettivo modificato $f_\mu(x)$

$$\nabla f_\mu(x) = \nabla f(x) - \mu \sum_{i=1}^{N} \left(\frac{1}{g_i(x)} \nabla g_i(x) \right), \tag{4.2.6}$$

$$\nabla^2 f_\mu(x) = \nabla^2 f(x) + \mu \sum_{i=1}^{N} \left(\frac{1}{g_i(x)} \nabla g_i(x) \nabla g_i(x)^T - \frac{1}{g_i^2(x)} \nabla^2 g_i(x) \right). \tag{4.2.7}$$

Per ogni valore di $\mu > 0$, pertanto, si ottiene dal metodo iterativo una soluzione $x^\star(\mu)$. Idealmente, al decrescere del parametro di barriera, la successione di tali soluzioni tende

alla soluzione ottima del problema, rimanendo all'interno del feasible set. L'insieme delle soluzioni intermedie ottenute dal metodo $\{x^\star(\mu), \mu > 0\}$ costituisce il cosiddetto *central path*.
A conclusione della descrizione di questo metodo, è necessario fornire una misura della subottimalità delle soluzioni intermedie generate dal metodo, mostrando la loro convergenza alla soluzione ottima, al diminuire di μ. A tale scopo, si utilizza una importante proprietà del central path, ossia l'esistenza di elementi duali ad ogni $x^\star(\mu)$, definiti come

$$\lambda_i^\star(\mu) = \frac{\mu}{g_i(x^\star(\mu))}, \qquad i = 1, \dots, N. \tag{4.2.8}$$

Si introduce quindi la funzione Lagrangiana, utilizzando i $\lambda_i = \lambda_i^\star(\mu)$ come moltiplicatori di Lagrange associati ai vincoli

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{N} \lambda_i g_i(x), \tag{4.2.9}$$

da cui si ricava la funzione duale

$$p(\lambda^\star(\mu)) = f(x^\star(\mu)) + \sum_{i=1}^{N} \lambda_i^\star(\mu) g_i(x^\star(\mu)) \tag{4.2.10}$$

$$= f(x^*(\mu)) + N\mu. \tag{4.2.11}$$

In definitiva, detta $x^\star$ la soluzione ottima del problema, poichè per il teorema della dualità debole si ha che $p(\lambda) \leq f(x)$ per ogni x ammissibile, si ottiene la seguente disuguaglianza

$$f(x^\star(\mu)) - f(x^\star) \leq N\mu, \tag{4.2.12}$$

dove la quantità $N\mu$ assume la denominazione di *duality gap*. Questo risultato mostra come al decrescere di μ, la successione dei valori del funzionale obiettivo ottenuta dal metodo di Newton, converga alla soluzione ottima $f(x^\star)$. Nella trattazione computazionale, viene posta una soglia ϵ sul duality gap, al fine di accettare, quando sufficientemente accurata, una soluzione intermedia data dal metodo iterativo.

I metodi primal-dual interior-point Questi metodi computazionali vennero sviluppati come evoluzione dei metodi primal, al fine di superare alcuni limiti numerici del precedente approccio, come ad esempio la difficoltà nella ricerca di una soluzione quando il parametro di barriera è molto piccolo, per via della forte non-linearità del funzionale obiettivo. In particolare, la strategia computazionale adottata, prevede di riformulare il problema trasformando i vincoli in sole uguaglianze, tramite un insieme di variabili ausiliarie, quindi di risolvere un sistema lineare ricavato dalle condizioni KKT calcolate con l'approccio duale. Considerando il problema generalizzato in esame (4.2.3), si applica la suddetta riformulazione, utilizzando una funzione barriera logaritmica

$$\begin{aligned} &\min_{x,s} \; f(x) - \mu \sum_{i=1}^{N} \log s_i \\ &\text{s.t. } g_i(x) + s_i = 0, \qquad i = 1, \dots, N. \end{aligned} \tag{4.2.13}$$

Pertanto, per ottenere un sistema lineare algebrico, è necessario calcolare le condizioni KKT relative a questo problema, definendo la funzione Lagrangiana associata al sistema

$$\mathcal{L}(x, s, \lambda) = f(x) - \mu \sum_{i=1}^{N} \log s_i + \sum_{i=1}^{N} \lambda_i (g_i(x) + s_i). \tag{4.2.14}$$

Derivando rispetto a x e s si ottiene il sistema

$$\begin{cases} \nabla f(x) \ + G^T \lambda = 0, \\ -\mu / s_i + \lambda_i = 0, \qquad i = 1, \dots, N, \\ g_i(x) + s_i = 0, \qquad i = 1, \dots, N, \end{cases} \tag{4.2.15}$$

dove G è la matrice Jacobiana relativa alle funzioni $g_i(x),\ i = 1, \dots, N$.
Supponendo $x \in \mathbb{R}^n$ si può riscrivere il sistema in forma compatta tramite la funzione

$$F(x, s, \lambda) = \begin{bmatrix} \nabla f(x) \ + G^T \lambda \\ s_i \lambda_i - \mu \\ g_i(x) + s_i \end{bmatrix} \in \mathbb{R}^{n+2N}. \tag{4.2.16}$$

Per costruire il metodo di Newton, è necessario infine calcolare la matrice Jacobiana J_F, al fine di impostare il sistema

$$\begin{bmatrix} \nabla^2_{xx} \mathcal{L} & 0 & G^T \\ 0 & S & \Lambda \\ G & \mathbb{I} & 0 \end{bmatrix} \begin{bmatrix} p_x \\ p_s \\ p_\lambda \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \ + G^T \lambda \\ S\Lambda e - \mu e \\ g(x) + s \end{bmatrix}, \tag{4.2.17}$$

dove sono definite le matrici diagonali S e Λ formate dai termini s_i e λ_i, mentre il simbolo 0 rappresenta una matrice nulla di dimensioni opportune.

4.3 Applicazione al problema di controllo vincolato da CVaR

Definiti i metodi numerici per la risoluzione di un problema di ottimizzazione modello, è necessario riformulare opportunamente, utilizzando i risultati di Rockafellar e Uryasev, il vincolo sul CVaR. Nella seguente trattazione, sono confrontate due strategie, basate sui due differenti approcci interior-point descritti. Inoltre, esse differiscono per il metodo utilizzato nell'approssimazione della funzione $(\cdot)^+$ contenuta nel CVaR, tramite aggiunta di variabili ausiliarie o per mezzo di una funzione di regolarizzazione.

4.3.1 Approccio interior-point con riformulazione epigrafica

La strategia risolutiva adottata da questo metodo consiste nell'includere l'ultimo vincolo, quello di disuguaglianza sul CVaR, nel funzionale obiettivo del problema di ottimizzazione in forma ridotta, tramite l'utilizzo della funzione barriera di tipo logaritmico

$$\phi(y(u)) = -\log(-(\mathrm{CVaR}_\beta(\theta(y_\omega(u))) - \tau)). \tag{4.3.1}$$

Il problema risultante sarà pertanto non-vincolato

$$\min_{u\in\mathcal{U}} \hat{J}(y(u),u) + \mu\phi(y(u)), \tag{4.3.2}$$

indicizzato dal parametro di barriera μ, che assumerà valori decrescenti all'avvicinarsi della soluzione.
Tuttavia, per poter trattare numericamente questo problema, è opportuno riformulare il vincolo sul CVaR della (4.2.2), introducendo un set di variabili ausiliarie $\{z_{\omega_j}\}_{j=1}^N$ in numero pari alla grandezza del campione considerato per la variabile casuale ω. Il problema ottenuto, in forma ridotta, contiene quindi una *riformulazione epigrafica* del vincolo

$$\begin{aligned}\min_{u\in\mathcal{U},t\in\mathbb{R},\mathbf{z}\in\mathbb{R}^N}\ & \hat{J}(u)\\ \text{s.t. } & t + \frac{1}{1-\beta}\frac{1}{N}\sum_{j=1}^{N} z_{\omega_j} \le \tau,\\ & \theta(y_{\omega_j}(u)) - t \le z_{\omega_j}, \qquad j=1,\dots,N,\\ & z_{\omega_j} \ge 0, \qquad j=1,\dots,N.\end{aligned} \tag{4.3.3}$$

Si osserva che, in totale, nella riformulazione presentata, il vincolo relativo al CVaR viene spezzato in $2N+1$ vincoli rispetto alle variabili ausiliarie $\{z_{\omega_j}\}_{j=1}^N$. In particolare, oltre agli N vincoli di non-negatività delle variabili ausiliarie si hanno N ulteriori vincoli riferiti alla parte positiva all'interno del CVaR e un vincolo per la limitazione del valore atteso calcolato sulle variabili ausiliarie tramite un metodo Monte Carlo. Idealmente, il comportamento atteso è la riduzione fino a valori trascurabili delle componenti del vettore $\mathbf{z}$ relative ai valori della funzione θ, calcolati per mezzo delle realizzazioni della variabile casuale, che risultano inferiori al quantile di riferimento t. Un risultato accettabile del metodo, consiste nel determinare un controllo ottimo cui è associato, tramite la funzione θ, il relativo quantile t ed il vettore $\mathbf{z}$ con componenti non nulle in corrispondenza dei valori che eccedono il suddetto quantile.
Definito l'approccio generale, si introducano le funzioni $\{h_j\}_{j=1}^{2N+1}$.

$$\begin{aligned}\min_{u\in\mathcal{U},t\in\mathbb{R},\mathbf{z}\in\mathbb{R}^N}\ & \hat{J}(u)\\ \text{s.t. } & h_j(u,t,\mathbf{z}) \le 0, \qquad j=1,\dots,2N+1,\end{aligned} \tag{4.3.4}$$

dove

$$\begin{aligned}& h_j(u,t,\mathbf{z}) = \theta(y_{\omega_j}(u)) - t - z_{\omega_j}, \qquad j=1,\dots,N,\\ & h_{j+N}(u,t,\mathbf{z}) = z_{\omega_j}, \qquad j=1,\dots,N,\\ & h_{2N+1}(u,t,\mathbf{z}) = t + \frac{1}{1-\beta}\frac{1}{N}\sum_{i=1}^{N} z_{\omega_i} - \tau.\end{aligned} \tag{4.3.5}$$

Pertanto applicando il metodo della barriera si ottiene il problema di ottimizzazione non vincolata parametrizzato da μ

$$\min_{u\in\mathcal{U},t\in\mathbb{R},\mathbf{z}\in\mathbb{R}^N} \hat{J}(u) - \frac{\mu}{2N+1}\sum_{j=1}^{2N+1}\log(-h_j(u,t,\mathbf{z})). \tag{4.3.6}$$

Questo problema, fissato un valore del parametro μ, può essere risolto utilizzando tecniche numeriche standard dell'ottimizzazione non vincolata, quale ad esempio il metodo di Newton.

Costruzione del metodo di Newton

L'applicazione del metodo di Newton richiede, oltre al calcolo del gradiente del funzionale obiettivo completo del termine della funzione barriera, quello della corrispondente matrice Hessiana, rispetto alle variabili del sistema, il controllo u, la variabile t e il vettore di variabili ausiliarie $\mathbf{z}$. Sia definito pertanto il funzionale obiettivo

$$\hat{J}_B(u,t,\mathbf{z}) = \hat{J}(u) - \frac{\mu}{2N+1} \sum_{j=1}^{2N+1} \log(-h_j(u,t,\mathbf{z})). \tag{4.3.7}$$

Come primo step si calcolano gradiente ed Hessiana mantenendo indicate le funzioni h_j

$$\begin{aligned}
\nabla \hat{J}_B(u,t,\mathbf{z}) =& \nabla \hat{J}(u) - \frac{\mu}{2N+1} \sum_{j=1}^{2N+1} \frac{\nabla h_j(u,t,\mathbf{z})}{h_j(u,t,\mathbf{z})}, \\
\nabla^2 \hat{J}_B(u,t,\mathbf{z}) =& \nabla^2 \hat{J}(u) + \frac{\mu}{2N+1} \sum_{j=1}^{2N+1} \left(\frac{\nabla h_j(u,t,\mathbf{z}) \nabla h_j(u,t,\mathbf{z})^T}{h_j^2(u,t,\mathbf{z})} - \frac{\nabla^2 h_j(u,t,\mathbf{z})}{h_j(u,t,\mathbf{z})} \right).
\end{aligned}$$

Quindi è necessario esplicitare le derivate rispetto ad ogni funzione h_j per costruire il sistema lineare algebrico. A tal proposito viene utilizzata la discretizzazione agli elementi finiti in uso nei capitoli precedenti. Si ricorda la dimensione dello spazio vettoriale finito-dimensionale V_h pari a N_V, oltre alla definizione dell'operatore di controllo-stato $\hat{S}$, espressa dalla (3.2.9). Inoltre, per questa analisi, viene considerata come funzione θ la norma quadrata della variabile di stato. Quindi, applicando la discretizzazione spaziale

$$\theta(y_h^\omega(u_h)) = \frac{1}{2}||y_h^\omega(u_h)||_V^2 = \frac{1}{2}||\hat{S}_h^\omega(f + Bu_h)||_V^2. \tag{4.3.8}$$

Ad esempio, considerando la PDE in forma operatoriale $A_\omega y_\omega = f_\omega + Bu$, con il relativo operatore di controllo stato, è possibile calcolare i gradienti delle h_j, ciascuno di dimensione $N_V + 1 + N$,

$$\nabla h_j(u_h,t,\mathbf{z}) = \begin{bmatrix} (A_{\omega_j}^{-1}B)^T M (A_{\omega_j}^{-1}(f_{\omega_j} + Bu_h)) \\ -1 \\ -e_j \end{bmatrix}, \qquad j = 1,\dots,N,$$

$$\nabla h_j(u_h,t,\mathbf{z}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ e_j \end{bmatrix}, \qquad j = N+1,\dots,2N,$$

$$\nabla h_{2N+1}(u_h, t, \mathbf{z}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \frac{1}{1-\beta}\frac{1}{N} \\ \vdots \\ \frac{1}{1-\beta}\frac{1}{N} \end{bmatrix},$$

dove gli e_j sono i vettori della base canonica di $\mathbb{R}^N$.
Per quanto riguarda le matrici Hessiane relative alle funzioni h_j, le uniche non identicamente nulle sono quelle relative ai vincoli contenenti la norma della variabile di stato, che risultano pari a

$$\nabla^2 h_j(u_h, t, \mathbf{z}) = \begin{bmatrix} (A_{\omega_j}^{-1}B)^T M (A_{\omega_j}^{-1}B) \\ \mathbb{O}^{(N+1)\times(N+1)} \end{bmatrix}, \qquad j = 1, \dots, N.$$

Completata la costruzione di matrici e gradienti, il metodo ricorsivo di Newton viene applicato alla variabile $\boldsymbol{\eta} = (u_h, t, \mathbf{z})$ tramite l'espressione

$$\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k - t(\nabla^2 \hat{J}_B)^{-1} \nabla \hat{J}_B, \tag{4.3.9}$$

partendo da una stima iniziale strettamente feasibile $\boldsymbol{\eta}_0$ e applicando opportunamente le regole di backtracking per ricavare la step-size t. Nell'Algoritmo 5 viene schematizzato il procedimento descritto per questo metodo.

Algoritmo 5 Metodo della barriera con approccio primal interior point

Input: $\boldsymbol{\eta}_0,\ \mu_0,\ K_{max},\ tol,\ \tau,\ \kappa$

while Criterio di arresto su μ non soddisfatto **do**
  $j = 1$
  while Criterio di arresto per Newton non soddisfatto **do**
    $k = 0$
    Costruzione di $\nabla^2 \hat{J}_B, \nabla \hat{J}_B$, tramite risoluzione PDE di stato e aggiunta
    Step-size iniziale $\alpha = 1$
    $\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k - t(\nabla^2 \hat{J}_B)^{-1} \nabla \hat{J}_B$
    Backtracking su α tramite condizione di Armijo
    Aggiornamento di $\boldsymbol{\eta}^k \leftarrow \boldsymbol{\eta}^{k+1}$
    $k \leftarrow k + 1$
  end while
  $\mu_{j+1} \leftarrow \kappa \mu_j$
  $j \leftarrow j + 1$
end while

4.3.2 Approccio smoothing-splitting

Come anticipato, nel secondo approccio proposto, il problema (4.2.1) viene trattato applicando una regolarizzazione della funzione $(\cdot)_+$ contenuta nel vincolo relativo al CVaR.

Sono pertanto analizzate proprietà e caratteristiche di alcune funzioni utili a tale scopo. Quindi, il problema di ottimizzazione è risolto applicando il metodo di Newton seguendo una strategia di splitting delle variabili, per evitare la degenerazione della matrice Hessiana. Nel seguito, viene presentata la costruzione passo-passo del metodo e l'analisi delle sue caratteristiche.

Funzioni di smoothing

Una strategia efficace per rendere differenziabile il vincolo relativo al CVaR consiste nell'approssimazione diretta della funzione $(\cdot)^+$ tramite una famiglia di funzioni, parametrizzata da $\epsilon > 0$, definite come la convoluzione tra la funzione parte positiva e una densità di probabilità $\rho : \mathbb{R} \to \mathbb{R}$. Questo approccio è basato sul lavoro di Chen e Mangasarian [4], ed è stato adattato all'ambito dell'ottimizzazione vincolata da PDE *risk-averse* in [11, 15]. Pertanto si definisce tale famiglia di funzioni come

$$g_\epsilon(x) = \frac{1}{\epsilon}\int_{-\infty}^{\infty}(x-t)^+\,\rho\left(\frac{t}{\epsilon}\right)dt = \int_{-\infty}^{x/\epsilon}(x-\epsilon t)^+\,\rho(t)dt, \tag{4.3.10}$$

dove per la seconda uguaglianza è stato applicato il cambio di variabile $t \to t\epsilon$.
Si richiede inoltre che ρ soddisfi le seguenti assunzioni, oltre alle proprietà legate alla sua natura di densità di probabilità ($\rho(x) \geq 0\,\forall x \in \mathbb{R}$ e $\int_{-\infty}^{\infty}\rho(x)dx = 1$):

- $\exists M \in \mathbb{R}$ t.c. $\rho(x) \leq M$, $\forall x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty}\rho(x)|x|dx < \infty$.

Di seguito sono riportati alcuni esempi comuni di funzioni parametrizzate g_ϵ utilizzate per la regolarizzazione

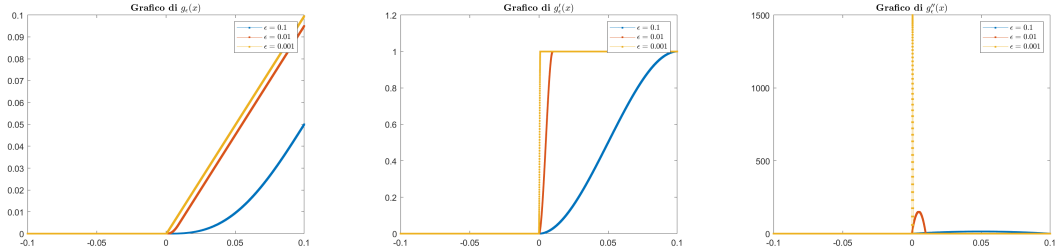
$$g_{\epsilon,1}(x) = x + \epsilon\log\left(1+\exp\left(-\frac{x}{\epsilon}\right)\right), \tag{4.3.11}$$

$$g_{\epsilon,2}(x) = \begin{cases} 0 & \text{se } x \leq 0, \\ \frac{x^3}{\epsilon^2} - \frac{x^4}{2\epsilon^3} & \text{se } 0 \leq x \leq \epsilon, \\ x - \frac{\epsilon}{2} & \text{se } x \geq \epsilon, \end{cases} \tag{4.3.12}$$

$$g_{\epsilon,3}(x) = \begin{cases} 0 & \text{se } x \leq -\frac{\epsilon}{2}, \\ \frac{\left(x-\frac{\epsilon}{2}\right)^3}{\epsilon^2} - \frac{\left(x-\frac{\epsilon}{2}\right)^4}{2\epsilon^3} & \text{se } -\frac{\epsilon}{2} \leq x \leq \frac{\epsilon}{2}, \\ x & \text{se } x \geq \frac{\epsilon}{2}. \end{cases} \tag{4.3.13}$$

Osservazione 4.3.1. *Si nota che la funzione $g_{\epsilon,1}$ è infinitamente differenziabile, mentre le altre sono differenziabili con continuità solo due volte.*

La Figura 4.1 mostra la rappresentazione grafica della funzione $g_{\epsilon,2}(x)$, selezionata per la regolarizzazione durante la trattazione successiva, per alcuni valori del parametro ϵ. Si dimostra che vale la seguente proposizione [15].



(a) (b) (c)

Figura 4.1: Funzione di smoothing $g_{\epsilon,2}$ e sue derivate, nell'intervallo $[-0.1,0.1]$, per alcuni valori di ϵ

Proposizione 4.3.1. *Sotto le ipotesi considerate, la funzione di smoothing $g_\epsilon(x)$ è convessa, non-decrescente e (almeno) due volte differenziabile con continuità. Inoltre si ha che $0 \leq g'_\epsilon(x) \leq 1$ e $0 \leq g''_\epsilon(x) \leq M/\epsilon$, $\forall x \in \mathbb{R}$. Infine è possibile dimostrare che $|g_\epsilon(\cdot) - (\cdot)_+| < c\epsilon$, per un'opportuna costante $c > 0$.*

Sia pertanto definita la funzione CVaR regolarizzata

$$G^\beta_\epsilon(\theta_\omega(u), t) = \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta_\omega(u) - t)] \right\}, \tag{4.3.14}$$

dove θ_ω rappresenta una quantità di interesse su cui è calcolato il CVaR, ad esempio $\theta_\omega(u) = 1/2||y_\omega(u)||^2_{L^2(D)}$. Il seguente risultato [11] fornisce una stima dell'errore commesso con l'approssimazione tramite la funzione g_ϵ.

Lemma 4.3.1. *Considerate le assunzioni supposte per ρ, si ha che*

$$|G^\beta_\epsilon(\theta_\omega(u), t) - CVaR_\beta(\theta_\omega(u))| \leq \frac{c}{1-\beta}\epsilon. \tag{4.3.15}$$

Dimostrazione. Supponendo esista $t^\star = \mathrm{VaR}_\beta(\theta_\omega(u))$, *minimizer* del CVaR, è possibile provare, sfruttando la limitatezza dei level set $t \to G^\beta_\epsilon(\theta_\omega(u), t)$, che esiste $t^\star_\epsilon$, *minimizer* della funzione regolarizzata. Supponendo che $g_\epsilon(\cdot) \geq (\cdot)^+$, si ha

$$\begin{aligned} t^\star + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta_\omega(u) - t^*)] &\geq t^\star_\epsilon + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta_\omega(u) - t^*_\epsilon)] = G^\beta_\epsilon(\theta_\omega(u), t) \\ &\geq t^\star_\epsilon + \frac{1}{1-\beta}\mathbb{E}[(\theta_\omega(u) - t^\star_\epsilon)^+] \\ &\geq t^\star + \frac{1}{1-\beta}\mathbb{E}[(\theta_\omega(u) - t^\star)^+] = \mathrm{CVaR}_\beta(\theta_\omega(u)). \end{aligned}$$

Viceversa, nel caso in cui $g_\epsilon(\cdot) \leq (\cdot)_+$, risulta che

$$\begin{aligned} t^\star_\epsilon + \frac{1}{1-\beta}\mathbb{E}[(\theta_\omega(u) - t^\star_\epsilon)^+] &\geq t^\star + \frac{1}{1-\beta}\mathbb{E}[(\theta_\omega(u) - t^\star)^+] = \mathrm{CVaR}_\beta(\theta_\omega(u)) \\ &\geq t^\star + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta_\omega(u) - t^\star)] \\ &\geq t^\star_\epsilon + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta_\omega(u) - t^\star_\epsilon)] = G^\beta_\epsilon(\theta_\omega(u), t). \end{aligned}$$

Infine, applicando la stima per $g_\epsilon(\cdot) - (\cdot)_+$ segue il risultato

$$\left| \left(t + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta_\omega(u) - t)] \right) - \left(t + \frac{1}{1-\beta}\mathbb{E}[(\theta_\omega(u) - t)^+] \right) \right| < \frac{c}{1-\beta}\epsilon.$$

□

Infine, è interessante valutare la velocità di convergenza delle soluzioni ottime approssimate $(u_\epsilon^\star, t_\epsilon^\star)$ alla soluzione ottima $(u^\star, t^\star)$. Per raggiungere tale obiettivo, è tuttavia necessario mostrare che vale il seguente teorema [15].

Teorema 4.3.1. *Sia indicato con $\hat{J}_\epsilon$ il funzionale obiettivo relativo al problema riformulato con la funzione di smoothing. Supponendo esistano C_1, C_2 t.c.*

$$\begin{aligned} -C_1\epsilon \leq & \hat{J}(u^\star, t^\star) - \hat{J}_\epsilon(u^\star, t^\star) \\ & \hat{J}(u_\epsilon^\star, t_\epsilon^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) \leq C_2\epsilon, \end{aligned} \tag{4.3.16}$$

e che inoltre esista $\alpha > 0$ t.c.

$$\left\langle \nabla \hat{J}_\epsilon(u,t) - \nabla \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star), (u,t) - (u_\epsilon^\star, t_\epsilon^\star) \right\rangle \geq \alpha ||(u,t) - (u_\epsilon^\star, t_\epsilon^\star)||^2, \tag{4.3.17}$$

per un generico $(u,t) = (u_\epsilon^\star, t_\epsilon^\star) + s((u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)), \quad s \in (0,1)$.
Allora si conclude che

$$||(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)||^2 \leq \frac{2}{\alpha}(C_1 + C_2)\epsilon. \tag{4.3.18}$$

Dimostrazione. Si osserva che

$$\begin{aligned} \hat{J}(u^\star, t^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) &= \hat{J}(u^\star, t^\star) - \hat{J}_\epsilon(u^\star, t^\star) + \hat{J}_\epsilon(u^\star, t^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) \\ &\geq -C_1\epsilon + \int_0^1 \left\langle \nabla \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star + s(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)), (u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star) \right\rangle ds. \end{aligned}$$

Utilizzando l'ipotesi (4.3.17), si ha che il termine integrando diventa

$$\begin{aligned} & \left\langle \nabla \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star + s(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)), (u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star) \right\rangle \\ &= \frac{1}{s} \left\langle \nabla \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star + s(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)) - \nabla \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star), s((u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)) \right\rangle \\ &\geq \frac{\alpha}{s} ||s(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)||^2 = s\alpha ||(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)||^2. \end{aligned}$$

Sostituendo questo termine nell'integrale rispetto a s, si ottiene la disuguaglianza

$$\hat{J}(u^\star, t^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) + C_1\epsilon \geq \frac{\alpha}{2} ||(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)||^2.$$

Quindi, poiché

$$\begin{aligned} \hat{J}(u^\star, t^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) &= \hat{J}(u^\star, t^\star) - \hat{J}_\epsilon(u^\star, t^\star) + \hat{J}_\epsilon(u^\star, t^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) \\ &\leq \hat{J}_\epsilon(u^\star, t^\star) - \hat{J}_\epsilon(u_\epsilon^\star, t_\epsilon^\star) \leq C_2\epsilon, \end{aligned}$$

è possibile concludere che

$$||(u^\star, t^\star) - (u_\epsilon^\star, t_\epsilon^\star)||^2 \leq \frac{2}{\alpha}(C_1 + C_2)\epsilon.$$

□

Splitting delle variabili e costruzione del metodo primale

Il problema di controllo ottimo vincolato da risolvere viene espresso, nella forma ridotta, come

$$\begin{aligned}\min_{u\in\mathcal{U},t\in\mathbb{R}} \hat{J}(u) &= \frac{1}{2}\mathbb{E}[||S_\omega(f+Bu)-z_d||_V^2] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ \text{s.t. } t + \frac{1}{1-\beta}&\mathbb{E}[g_\epsilon(\theta(S_\omega(f+Bu))-t)] \leq \tau,\end{aligned} \tag{4.3.19}$$

Una interessante strategia computazionale per trattare il vincolo consiste nel riformulare il problema esprimendo la variabile $t \in \mathbb{R}$ in termini della variabile di controllo, utilizzando una *funzione implicita* $t = h(u)$. Sia pertanto introdotto il funzionale

$$F(u,t) = t + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta(S_\omega(f+Bu))-t)]. \tag{4.3.20}$$

Il valore ottimo di t è quindi definito come l'argomento che minimizza tale funzionale

$$t = h(u) = \arg\min_{t\in\mathbb{R}} F(u,t). \tag{4.3.21}$$

Con questa rappresentazione, è possibile sostituire tale espressione nel funzionale, riducendo l'insieme delle variabili indipendenti al solo controllo $F(u, h(u))$. Pertanto, il calcolo del gradiente e della matrice Hessiana associata al funzionale, dovranno tenere presente la struttura della funzione implicita, ricordando tuttavia che, per definizione di t

$$\frac{\partial F}{\partial t}(u,h(u)) = 0. \tag{4.3.22}$$

Sotto l'assunzione di unicità di $t^\star$ tale per cui il funzionale F è minimizzato rispetto a t, oltre che di forte convessità e Lipschitzianità della matrice Hessiana associata alla quantità di interesse $\theta(u)$, è possibile dimostrare che anche il funzionale F è fortemente convesso, con gradienti Lipschitz-continui.

Per la costruzione del metodo, è necessario dare un'espressione esplicita per gradiente e matrice Hessiana di F. Per compattezza notazionale, è opportuno definire

$$\Theta_\omega(u) \coloneqq \theta(S_\omega(f+Bu)), \tag{4.3.23}$$

con le relative derivate Θ'_ω e Θ''_ω. Si osserva che, vista la condizione (4.3.22)

$$\begin{aligned}\nabla F(u,h(u)) &= \frac{\partial F}{\partial u}(u,h(u)) + \frac{\partial F}{\partial t}(u,h(u))h'(u) = \frac{\partial F}{\partial u}(u,h(u)) \\ &= \frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_\omega(u)-h(u))\Theta'_\omega(u)].\end{aligned} \tag{4.3.24}$$

Inoltre si ha che

$$\begin{aligned}
H_{uu}F(u,h(u)) &= \frac{\partial^2 F}{\partial u^2}(u,h(u)) = \\
&= \frac{1}{1-\beta}\mathbb{E}[g''_\epsilon(\Theta_\omega(u)-h(u))\,(\Theta'_\omega(u))^2 + g'_\epsilon(\Theta_\omega(u)-h(u))\,\Theta''_\omega(u)], \\
H_{ut}F(u,h(u)) &= \frac{\partial^2 F}{\partial u \partial t}(u,h(u)) = -\frac{1}{1-\beta}\mathbb{E}[g''_\epsilon(\Theta_\omega(u)-h(u))\Theta'_\omega(u)], \\
H_{tt}F(u,h(u)) &= \frac{\partial^2 F}{\partial t^2}(u,h(u)) = \frac{1}{1-\beta}\mathbb{E}[g''_\epsilon(\Theta_\omega(u)-h(u))].
\end{aligned} \tag{4.3.25}$$

Costruita l'Hessiana nelle due variabili (u,t), si ha che è possibile ridurre tale matrice alla sola dipendenza dal controllo utilizzando il complemento di Schur

$$H_F(u,h(u)) = H_{uu}(u,h(u)) - H_{ut}(u,h(u))H_{tt}(u,h(u))^{-1}H_{ut}(u,h(u))^\top. \tag{4.3.26}$$

Sostituendo e svolgendo il calcolo, si conclude che

$$H_F(u,h(u)) = \frac{1}{1-\beta}\mathbb{E}[g''_\epsilon(\Theta_\omega(u)-h(u))\,(\Theta'_\omega(u))^2 + g'_\epsilon(\Theta_\omega(u)-h(u))\,\Theta''_\omega(u)] - \tag{4.3.27}$$

$$\frac{1}{1-\beta}\frac{\mathbb{E}[g''_\epsilon(\Theta_\omega(u)-h(u))\Theta'_\omega(u)]^2}{\mathbb{E}[g''_\epsilon(\Theta_\omega(u)-h(u))]}. \tag{4.3.28}$$

A questo punto, come nel caso precedente, si risolve il problema vincolato utilizzando una funzione barriera logaritmica, ottenendo il funzionale

$$\hat{J}_B(u) = \hat{J}(u) - \mu\log(-(F(u,h(u))-\tau)). \tag{4.3.29}$$

Quindi, utilizzando i risultati dei calcoli precedenti, è possibile costruire il sistema di Newton

$$\nabla^2\hat{J}_B(u)\,\delta u = -\nabla\hat{J}_B(u), \tag{4.3.30}$$

dove si ha che

$$\begin{aligned}
\nabla\hat{J}_B(u) &= \nabla\hat{J}(u) - \mu\frac{\nabla F(u,h(u))}{F(u,h(u))-\tau}, \\
\nabla^2\hat{J}_B(u) &= \nabla^2\hat{J}(u) - \mu\left(\frac{\nabla F(u,h(u))\nabla F(u,h(u))'}{(F(u,h(u))-\tau)^2} - \frac{H_F(u,h(u))}{F(u,h(u))-\tau}\right).
\end{aligned}$$

Risoluzione con il metodo primal-dual interior point

Una valida alternativa al metodo primale, come descritto nell'introduzione teorica, è rappresentata dal metodo primal-dual interior point, che prevede la trasformazione dei vincoli di disuguaglianza in sole uguaglianze tramite l'introduzione di variabili ausiliarie e la risoluzione di un sistema di Newton basato sulle condizioni KKT. Risulta quindi necessario riformulare il vincolo sul CVaR, cui viene applicato lo smoothing tramite la

funzione g_ϵ introducendo la *slack variable* $s \in \mathbb{R}^+$ e utilizzando una funzione barriera logaritmica

$$\begin{aligned} &\min_{u\in\mathcal{U}} \hat{J}_B(u) = \hat{J}(u) - \mu \log(s) \\ &\text{s.t. } t + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta(S_\omega(f+Bu)) - t)] + s = \tau. \end{aligned} \tag{4.3.31}$$

Per costruire il sistema di Newton, bisogna ricavare le condizioni KKT del problema, ottenute derivando la Lagrangiana. Per scriverne l'espressione, si esprime la variabile ausiliaria t in termini della variabile di controllo tramite una funzione implicita $t = h(u)$.

$$\mathcal{L}(u, s, \lambda) = \hat{J}(u) - \mu \log(s) + \lambda \left(h(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\theta(S_\omega(f+Bu)) - h(u))] + s - \tau \right). \tag{4.3.32}$$

Si ottiene pertanto il seguente sistema di condizioni di ottimalità, ricordando la definizione di $\Theta_\omega(u)$ e delle relative derivate

$$\begin{cases} \nabla \hat{J}(u) + \lambda \frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_\omega(u) - h(u))\Theta'_\omega(u)] = 0, \\ -\mu/s + \lambda = 0, \\ h(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_\omega(u) - h(u))] + s - \tau = 0. \end{cases} \tag{4.3.33}$$

Supponendo $u \in \mathbb{R}^n$ si può riscrivere il sistema in forma compatta tramite la funzione

$$F_{\text{KKT}}(u, s, \lambda) = \begin{bmatrix} \nabla \hat{J}(u) + \lambda \frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_\omega(u) - h(u))\Theta'_\omega(u)] = 0 \\ s\lambda - \mu \\ h(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_\omega(u) - h(u))] + s - \tau \end{bmatrix} \in \mathbb{R}^{n+2}. \tag{4.3.34}$$

Per applicare il metodo di Newton, è necessario infine calcolare la matrice Jacobiana J_F. A tal proposito, si ricava la matrice Hessiana $H_F(u, h(u))$, tenendo presenti i termini aggiuntivi introdotti dalla funzione implicita, come nella (4.3.26).
Definito inoltre $G_u = \frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_\omega(u) - h(u))\Theta'_\omega(u)]$, si costruisce il sistema

$$\begin{bmatrix} H_F & 0 & G_u^\top \\ 0 & s & \lambda \\ G_u & 1 & 0 \end{bmatrix} \begin{bmatrix} p_u \\ p_s \\ p_\lambda \end{bmatrix} = - \begin{bmatrix} \nabla \hat{J}(u) + \lambda G_u \\ s\lambda - \mu \\ h(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_\omega(u) - h(u))] + s - \tau \end{bmatrix}. \tag{4.3.35}$$

L'implementazione numerica del metodo viene condotta come descritto nell'Algoritmo 6.

4.3.3 Criteri di arresto

Un passo cruciale nella definizione del metodo è rappresentato dalla scelta di opportuni criteri di arresto sia per quanto riguarda le iterazioni interne del metodo di Newton, sia per quelle esterne. In particolare, fissato un numero massimo di iterazioni oltre il quale si ritiene fallita la convergenza del metodo, si possono considerare alcune strategie di arresto comuni per i metodi iterativi di ottimizzazione, quali

Algoritmo 6 Metodo della barriera con approccio primal-dual interior point

Input: u_0, μ_0, K_{max}, tol, τ, κ

while Criterio di arresto su μ non soddisfatto **do**
  $j = 1$
  while Criterio di arresto per Newton non soddisfatto **do**
    $k = 0$
    Risoluzione PDE di stato e aggiunta
    Calcolo di $t^k = \min_{t \in \mathbb{R}} F(u, t)$
    Definizione del sistema KKT $F_{\text{KKT}}(u, s, \lambda)$
    Assemblaggio della matrice Jacobiana
    Step-size iniziale $\alpha = 1$
    $u^{k+1} = u^k - \alpha(\nabla^2 \hat{J}_B)^{-1} \nabla \hat{J}_B$
    Backtracking su α tramite condizione di Armijo
    Aggiornamento di $u^k \leftarrow u^{k+1}$
    $k \leftarrow k + 1$
  end while
  $\mu_{j+1} \leftarrow \kappa \mu_j$
  $j \leftarrow j + 1$
end while

- controllare la norma del gradiente del funzionale obiettivo: quando quest'ultima scende sotto una soglia predefinita ϵ, la soluzione trovata viene considerata un'approssimazione accettabile di quella reale

$$||\nabla \hat{J}(u^k)|| < \epsilon. \tag{4.3.36}$$

- stabilire uno scarto sufficientemente piccolo tra due valori del funzionale obiettivo ottenuti da iterate consecutive

$$|\hat{J}(u^k) - \hat{J}(u^{k-1})| < \epsilon. \tag{4.3.37}$$

Oltre a questi criteri classici, nella presente trattazione viene implementata una strategia di arresto specifica del metodo di Newton, basata sull'approssimazione quadratica del funzionale obiettivo

$$\hat{J}(u + \delta u) \approx \hat{J}(u) + \nabla \hat{J}(u)^T \delta u + \frac{1}{2} \delta u^T \nabla^2 \hat{J}(u) \delta u. \tag{4.3.38}$$

Il metodo minimizza la funzione a secondo membro rispetto a δu, che soddisfa

$$\nabla \hat{J}(u) + \nabla^2 \hat{J}(u) \delta u = 0, \tag{4.3.39}$$

da cui si ottiene l'espressione per la direzione di discesa $\delta u = -\nabla^2 \hat{J}(u)^{-1} \nabla \hat{J}(u)$. Sostituendo tale risultato nella (4.3.38) si ha

$$\begin{aligned}\hat{J}(u) - \hat{J}(u+\delta u) \approx & -\nabla \hat{J}(u)^T(-\nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u))+ \\ & -\frac{1}{2}(\nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u))^T \nabla^2 \hat{J}(u)(-\nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u)) \\ & = \nabla \hat{J}(u)^T \nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u) - \frac{1}{2}\nabla \hat{J}(u))^T \nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u) \\ & = x\frac{1}{2}\nabla \hat{J}(u)^T \nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u).\end{aligned}$$

Si definisce pertanto il *decremento di Newton*, come la funzione

$$\lambda = \sqrt{\nabla \hat{J}(u)^T \nabla^2 \hat{J}(u)^{-1}\nabla \hat{J}(u)}. \tag{4.3.40}$$

In definitiva, pertanto, una misura efficace dell'errore di approssimazione introdotto dal metodo consiste nel valutare $\lambda^2/2$, interrompendo il metodo quando tale valore raggiunge un parametro piccolo ϵ

$$\frac{\lambda^2}{2} < \epsilon. \tag{4.3.41}$$

Per quanto riguarda invece il metodo della barriera, una strategia comune per stabilire un criterio di arresto è rappresentata dall'analisi di una quantità caratteristica del problema di ottimizzazione detta *duality gap*, già introdotta nella trattazione generica dei metodi interior-point. Nel caso del metodo primale, ad esempio, tale quantità è rappresentata dal prodotto tra il parametro di barriera μ e il numero di vincoli, indicato con m. Pertanto, la condizione di arresto si scrive come

$$\frac{\lambda^2}{2} < \eta m \mu, \tag{4.3.42}$$

con $\eta > 0$ costante (ad esempio $\eta = 0.1$). In alternativa, se invece di sommare tutti i termini barriera si opta per una media sul numero totale di vincoli, il criterio diventa

$$\frac{\lambda^2}{2} < \eta \mu. \tag{4.3.43}$$

4.4 Problemi di minimizzazione vincolata con funzionale obiettivo *risk-averse*

In questa sezione viene presentato un problema di controllo ottimo vincolato da PDE in cui viene richiesta la minimizzazione di un funzionale obiettivo dipendente dal CVaR di una certa quantità di interesse, quale la distanza della soluzione della PDE da una funzione target, e contemporaneamente che sia soddisfatto un vincolo sul CVaR di un altro funzionale di interesse, come ad esempio l'energia della soluzione. Considerando la forma ridotta, questo problema si scrive come

$$\begin{aligned}&\min_{u \in \mathcal{U}} \hat{J}(u) = \mathrm{CVaR}_\beta(\theta_J(S_\omega(f+Bu))) + \frac{\nu}{2}||u||^2_{\mathcal{U}} \\ &\text{s.t. } \mathrm{CVaR}_\beta(\theta_G(S_\omega(f+Bu))) \leq \tau,\end{aligned} \tag{4.4.1}$$

dove vengono definite le quantità di interesse

$$\theta_J(S_\omega(f+Bu)) = \frac{1}{2}||S_\omega(f+Bu) - z_d||_V^2, \tag{4.4.2}$$

$$\theta_G(S_\omega(f+Bu)) = \frac{1}{2}||S_\omega(f+Bu)||_V^2. \tag{4.4.3}$$

La risoluzione numerica del problema segue le due tecniche presentate per il problema con funzionale *risk-neutral*, tenendo in considerazione la presenza delle due misure di rischio CVaR differenti. A tale scopo, è necessario esplicitare il CVaR tramite la rappresentazione di Rockafellar e Uryasev, introducendo opportunamente le variabili ausiliarie t_J, t_G, z_J, z_G, i cui pedici indicano rispettivamente il riferimento al CVaR nell'obiettivo (J) o a quella nel vincolo (G).

$$\begin{aligned} &\min_{u\in\mathcal{U}} \hat{J}(u) = t_J + \frac{1}{1-\beta}\mathbb{E}[(\theta_J(S_\omega(f+Bu)) - t_J)^+] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ &\text{s.t. } t_G + \frac{1}{1-\beta}\mathbb{E}[(\theta_G(S_\omega(f+Bu)) - t_G)^+] \leq \tau. \end{aligned} \tag{4.4.4}$$

Riformulazione epigrafica

Seguendo il primo approccio, basato sul metodo primal interior point e la riformulazione epigrafica si ottiene il problema

$$\begin{aligned} \min_{u\in\mathcal{U},t_J,t_G\in\mathbb{R},\mathbf{z}_J,\mathbf{z}_G\in\mathbb{R}^N} \; & t_J + \frac{1}{1-\beta}\frac{1}{N}\sum_{k=1}^{N} z_{\omega_j}^J + \frac{\nu}{2}||u||_{\mathcal{U}}^2 =: \hat{J}_{\text{CVaR}}(u) \\ \text{s.t. } & t_G + \frac{1}{1-\beta}\frac{1}{N}\sum_{j=1}^{N} z_{\omega_k}^G \leq \tau, \\ & \theta_J(y_{\omega_j}(u)) - t_J \leq z_{\omega_j}^J, \qquad j = 1,\dots,N, \\ & \theta_G(y_{\omega_k}(u)) - t_G \leq z_{\omega_k}^G, \qquad k = 1,\dots,N, \\ & z_{\omega_j}^J \geq 0, \qquad j = 1,\dots,N, \\ & z_{\omega_k}^G \geq 0, \qquad k = 1,\dots,N, \end{aligned} \tag{4.4.5}$$

dove sono definite le variabili ausiliarie $\{z_{\omega_j}\}_{j=1}^N$ relative al CVaR nel funzionale obiettivo, e $\{z_{\omega_k}\}_{k=1}^N$ relative a quella nel vincolo. Definendo opportunamente le funzioni $h_j, j = 1,\dots,4N+1$ che rappresentano i vincoli, è possibile risolvere il problema applicando un metodo della barriera, costruito in maniera analoga al caso risk-neutral

$$\min_{u\in\mathcal{U},t_J,t_G\in\mathbb{R},\mathbf{z}_J,\mathbf{z}_G\in\mathbb{R}^N} \hat{J}_B(u) = \hat{J}_{\text{CVaR}}(u) - \frac{\mu}{4N+1}\sum_{j=1}^{4N+1} \log(-h_j(u,t_J,t_G,z_J,z_G)). \tag{4.4.6}$$

La risoluzione numerica viene completata calcolando gradiente e matrice Hessiana, quindi applicando il metodo di Newton.

Smoothing-splitting

Infine, viene presentata l'applicazione dell'approccio smoothing-splitting al problema di minimizzazione (4.4.1). La formulazione del problema ottenuta approssimando la parte positiva con la funzione g_ϵ è espressa dalla seguente

$$\begin{aligned}\min_{u\in\mathcal{U},t_J\in\mathbb{R},t_G\in\mathbb{R}} \hat{J}(u) = t_J + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_J(u)-t_J)] + \frac{\nu}{2}||u||^2_{\mathcal{U}} \\ \text{s.t. } t_G + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_G(u)-t_G)] \leq \tau,\end{aligned} \tag{4.4.7}$$

dove sono definite Θ_J e Θ_G applicando la notazione (4.3.23) alle espressioni (4.4.2)-(4.4.3). Ad ogni iterazione interna, i quantili t_J, t_G vengono aggiornati utilizzando un metodo di ottimizzazione 1D, come il metodo di bisezione, tramite la definizione delle funzioni implicite

$$t_J = h_J(u) = \arg\min_{t_J\in\mathbb{R}} F_J(u,t_J) \coloneqq t_J + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_J(u)-t_J)], \tag{4.4.8}$$

$$t_G = h_G(u) = \arg\min_{t_G\in\mathbb{R}} F_G(u,t_G) \coloneqq t_G + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_G(u)-t_G)]. \tag{4.4.9}$$

Il problema viene riformulato per costruire il metodo primal-dual interior point, aggiungendo una slack variable $s \in \mathbb{R}$ similmente alla (4.3.31) e scrivendo la funzione Lagrangiana

$$\mathcal{L}(u,s,\lambda) = h_J(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_J(u)-h_J(u))] + \frac{\nu}{2}||u||^2_{\mathcal{U}} - \mu\log(s) + \tag{4.4.10}$$

$$\lambda(h_G(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_G(u)-h_G(u))] + s - \tau). \tag{4.4.11}$$

Dalla derivazione della Lagrangiana si ottiene il sistema di condizioni KKT

$$F_{\text{KKT}}(u,s,\lambda) = \begin{bmatrix}\nabla\tilde{J}(u) + \lambda\frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_G(u)-h_G(u))\Theta'_G(u)] \\ s\lambda - \mu \\ h_G(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_G(u)-h_G(u))] + s - \tau\end{bmatrix} \in \mathbb{R}^{n+2}, \tag{4.4.12}$$

dove si definisce

$$\nabla\tilde{J}(u) = \nabla\hat{J}(u,h_J(u)) = \nu u + \frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_J(u)-h_J(u))\Theta'_J(u)].$$

Infine, per risolvere con un metodo di Newton è necessario determinare la matrice Jacobiana associata a $F_{\text{KKT}}(u,s,\lambda)$. Analogamente al caso con funzionale risk-averse, i contributi relativi ai quantili vengono gestiti tramite il complemento di Schur, per ridurre il sistema alla sola dipendenza dalla variabile di controllo, oltre che da variabile ausiliaria e relativo moltiplicatore. Questa strategia viene applicata sia all'Hessiana relativa al CVaR nel funzionale obiettivo, sia a quella nel vincolo, come segue

$$H_J(u,h_J(u)) = H_{J_{uu}}(u,h_J(u)) - H_{J_{ut}}(u,h_J(u))H_{J_{tt}}(u,h_J(u))^{-1}H_{J_{ut}}(u,h_J(u))^\top, \tag{4.4.13}$$

$$H_G(u,h_G(u)) = H_{G_{uu}}(u,h_G(u)) - H_{G_{ut}}(u,h_G(u))H_{G_{tt}}(u,h_G(u))^{-1}H_{G_{ut}}(u,h_G(u))^\top. \tag{4.4.14}$$

I singoli termini presenti in queste espressioni sono ricavati in maniera analoga alla (4.3.25). In definitiva, risulta possibile costruire il sistema

$$\begin{bmatrix} H_J + \nu I + \lambda H_G & 0 & G_u^\top \\ 0 & s & \lambda \\ G_u & 1 & 0 \end{bmatrix} \begin{bmatrix} p_u \\ p_s \\ p_\lambda \end{bmatrix} = - \begin{bmatrix} \nabla \tilde{J}(u) + \lambda G_u \\ s\lambda - \mu \\ h_G(u) + \frac{1}{1-\beta}\mathbb{E}[g_\epsilon(\Theta_G(u) - h_G(u))] + s - \tau \end{bmatrix}, \tag{4.4.15}$$

dove I rappresenta la matrice identità, mentre $G_u = \frac{1}{1-\beta}\mathbb{E}[g'_\epsilon(\Theta_G(u) - h_G(u))\Theta'_G(u)]$. Il sistema così costruito viene risolto iterativamente con un metodo di Newton adattato per il problema primal-dual interior-point.

Capitolo 5

Simulazioni numeriche

Il capitolo conclusivo di questo elaborato è dedicato alla presentazione di alcuni risultati numerici ottenuti implementando i metodi descritti, relativamente al controllo ottimo nei casi deterministico e stocastico, oltre ai problemi di controllo risk-averse. Inizialmente vengono confrontati alcuni metodi numerici di risoluzione del problema deterministico descritti nel primo capitolo. Quindi viene condotta un'analisi dell'errore rispetto alle discretizzazioni spaziali e in probabilità per il problema di controllo ottimo stocastico, al fine di osservare gli ordini di convergenza dimostrati nel Capitolo 2. A tale scopo sono utilizzati alcuni problemi modello con controllo distribuito oppure localizzato su un bordo di Neumann del dominio. Infine vengono confrontati i risultati ottenuti applicando le due strategie computazionali per la trattazione di problemi vincolati dal Conditional Value-at-Risk, sia nel caso di funzionali risk-neutral che risk-averse.

5.1 Problema di controllo ottimo non vincolato

In questa prima sezione sono presentati alcuni risultati relativi a problemi di controllo ottimo, sia nel caso deterministico che stocastico. In particolare, viene introdotta una variabile casuale nell'equazione differenziale di stato, con l'obiettivo di studiare il comportamento del sistema rispetto ad una serie di sue realizzazioni. Il caso particolare deterministico viene analizzato fissando una singola realizzazione. I risultati presentati sono relativi a due problemi di riferimento, uno con variabile di controllo definita sull'intero dominio, uno con variabile introdotta come condizione di Neumann su uno dei bordi. Lo scopo dell'analisi è la verifica dei risultati di convergenza teorici, rispetto alle discretizzazioni spaziale e in probabilità (per il problema stocastico).

5.1.1 Definizione dei problemi modello e della geometria

Al fine di condurre lo studio di convergenza, vengono definiti i due seguenti modelli, che si differenziano per il dominio di definizione della variabile di controllo (distribuita o al bordo).

- *Problema 1 - Controllo distribuito*
  Il problema di minimizzazione del funzionale risk-neutral, in forma ridotta, viene scritto come

$$\begin{aligned} &\min_{u\in\mathcal{U}} J(u) = \frac{1}{2}\mathbb{E}[||y_\omega(u) - z_d||_V^2] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ &\text{s.t. } a(y_\omega(u), v) = \langle f_\omega + u, v\rangle, \qquad \forall v \in V. \end{aligned} \tag{5.1.1}$$

  Il dominio è rappresentato da un quadrato $D = (0,1)^2$, su cui sono definite sia la variabile di stato $y(u) \in V \equiv H_0^1(D)$ sia quella di controllo $u \in \mathcal{U} \equiv L^2(D)$. L'equazione differenziale alle derivate parziali ellittica in esame è l'equazione di Laplace con coefficiente di diffusione, ed eventualmente termine forzante, stocastici, con condizioni al bordo di Dirichlet omogenee

$$\begin{cases} -\nabla\cdot(\epsilon(x_1,x_2,\xi)\nabla y) = f(x_1,x_2,\xi) + u & \text{in D,} \\ y = 0 & \text{su } \partial\text{D.} \end{cases} \tag{5.1.2}$$

  Sia definito un vettore di variabili casuali $\xi = (\xi_1,\xi_2,\xi_3,\xi_4)$, tra loro indipendenti, tutte con distribuzione uniforme $\xi_i \overset{\text{iid}}{\sim} U(-1,1)$. Allora il coefficiente di diffusione è definito come

$$\begin{aligned} \epsilon(x_1,x_2,\xi) = 1 + \exp(\sigma^2(&\xi_1\cos(\pi x_1)\sin(\pi x_2) + \xi_2\cos(2\pi x_1)\sin(\pi x_2) + \\ &+\xi_3\cos(\pi x_1)\sin(2\pi x_2) + \xi_4\cos(2\pi x_1)\sin(2\pi x_2))), \end{aligned} \tag{5.1.3}$$

  dove viene introdotto $\sigma^2 = \exp(-1.125)$. Infine, vengono fissati $\nu = 10^{-4}$ e la soluzione di riferimento $z_d(x_1,x_2) = \sin(2\pi x_1)\sin(2\pi x_2)$. Per quanto riguarda il termine forzante, vengono confrontate alcune differenti scelte, qui ordinate in maniera decrescente rispetto alla distanza dalla soluzione target, differenziate inoltre dalla eventuale dipendenza dalle variabili casuali.

$$\begin{aligned} f_1(x_1,x_2,&\xi) = 4\pi^2\epsilon(x_1,x_2,\xi)z_d(x_1,x_2), \\ f_2(x_1,x_2) &= 4\pi^2 z_d(x_1,x_2), \\ f_3(x_1,x_2) &= 1. \end{aligned}$$

- *Problema 2 - Controllo al bordo di Neumann*
  Il problema di minimizzazione del funzionale risk-neutral, assume la forma

$$\begin{aligned} &\min_{u\in\mathcal{U}} J(u) = \frac{1}{2}\mathbb{E}[||y_\omega(u) - z_d||_V^2] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ &\text{s.t. } a(y_\omega(u), v) = \langle f_\omega, v\rangle + \int_{\partial D_n} uv\, ds, \qquad \forall v \in V. \end{aligned} \tag{5.1.4}$$

  Il dominio è sempre rappresentato da un quadrato $D = (0,1)^2$, con bordo $\partial D = \partial D_d \cup \partial D_n$, su cui sono definite la variabile di stato $y(u) \in V \equiv H_0^1(D)$ e quella di controllo, ora solo sul bordo di Neumann del dominio $u \in \mathcal{U} \equiv L^2(\partial D_n)$. L'equazione differenziale alle derivate parziali ellittica in esame è l'equazione di Laplace con

coefficiente di diffusione, ed eventualmente termine forzante, stocastici, con condizioni al bordo in parte di Dirichlet omogenee e in parte di Neumann

$$\begin{cases} -\nabla \cdot (\epsilon(x_1, x_2, \xi)\nabla y) = f(x_1, x_2, \xi) & \text{in } D, \\ y = 0 & \text{su } \partial D_d, \\ \frac{\partial y}{\partial n} = u & \text{su } \partial D_n. \end{cases} \tag{5.1.5}$$

Fissata la soluzione target

$$z_d(x_1, x_2) = \exp(x_1 + x_2)\sin(2\pi x_1)\sin(2\pi x_2)\, I_{[0.5,1;0,1]}(x_1, x_2), \tag{5.1.6}$$

viene mantenuta l'espressione (5.1.3) per il coefficiente di diffusione stocastico ed inoltre viene scelto $\nu = 10^{-4}$. Anche in questo caso vengono testate alcune espressioni della forzante f, in particolare

$$\begin{aligned} f_1(x_1, x_2,\xi) &= 4\pi^2 \epsilon(x_1, x_2, \xi) z_d(x_1, x_2), \\ f_2(x_1, x_2) &= 4\pi^2 z_d(x_1, x_2), \\ f_3(x_1, x_2) &= 1. \end{aligned}$$

Geometria del problema

La discretizzazione spaziale agli elementi finiti in 2D viene costruita dal triangolatore *bbtr30* [2], sviluppato al Politecnico di Torino da A.M.A. Barbera e S. Berrone. Questo codice permette di generare mesh uniformi sul dominio scelto fissando alcuni parametri di riferimento, quali l'area massima o l'angolo minimo consentiti per i triangoli. In aggiunta, per il problema di controllo al bordo, è possibile selezionare una regione adiacente al bordo di Neumann in cui considerare un raffinamento ulteriore della mesh, per garantire maggiore accuratezza nella risoluzione. Inoltre, il codice permette di costruire geometrie più complesse aggiungendo segmenti, partizioni o buchi all'interno del dominio.

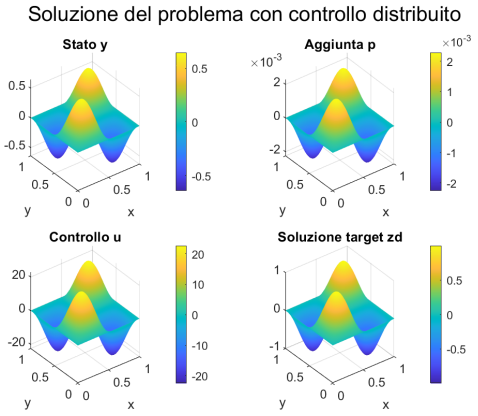
5.1.2 Analisi del problema distribuito

Si consideri in questa sezione il problema (5.1.1) introdotto precedentemente.
Al fine di confrontare i risultati ottenuti variando numero di campioni o accuratezza della mesh, è necessario fissare una soluzione di riferimento, relativamente al controllo u. A tale scopo viene costruita una mesh dal triangolatore *bbtr30*, fissando come massimo valore per l'area dei triangoli $h = 2^{-11}$. Inoltre vengono campionate 2000 realizzazioni della variabile casuale vettoriale ξ, costruendo le relative matrici di rigidezza e i termini noti relativi alle espressioni fornite per la forzante. Il sistema è quindi risolto con il metodo di Newton, fissando una tolleranza sulla norma del gradiente $||\nabla J(u)|| < tol = 10^{-9}$. La Figura 5.1 (a-c) riassume i risultati ottenuti, con $\nu = 10^{-4}$, per la variabile di stato, l'aggiunta e quella di controllo, oltre a fornire un confronto con la soluzione target z_d.

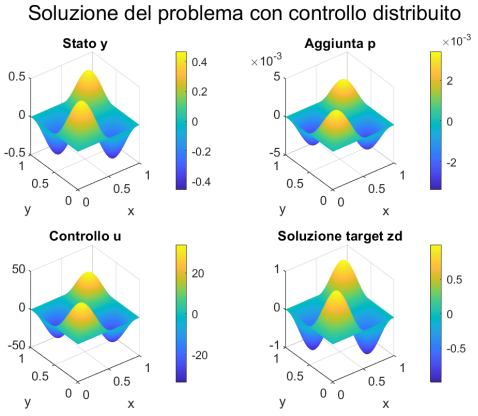
Caso deterministico

Per l'analisi del caso deterministico, viene fissata la singola realizzazione generata dai nodi di quadratura di Gauss-Legendre sull'intervallo [−1,1],

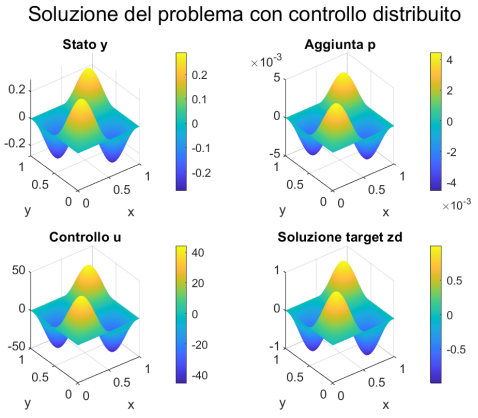
$$\xi = [-0.8611, -0.3400, 0.3400, 0.8611].$$



(a)



(b)



(c)

Figura 5.1: Soluzioni di riferimento relative alle forzanti f_1 (a), f_2 (b), f_3 (c)

Il problema viene risolto numericamente applicando alcune strategie computazionali, in particolare

- Metodi diretti, quali il calcolo del sistema completo attraverso l'operatore *backslash* di MATLAB, o tramite il calcolo del complemento di Schur della matrice $\mathbb{H}$.
- Metodi iterativi per il sistema ridotto, quali Steepest Descent o Newton, oppure gmres per il sistema completo.

La Tabella 5.1 riassume i tempi di esecuzione di questi metodi e il valore finale del funzionale obiettivo calcolato per $\nu = 10^{-4}$.

Metodo	**Tempo**	**Funzionale**
One-shot (2.2.16)	0.02817 s	0.02235
Schur su (2.2.16)	0.14235 s	0.02235
gmres su (2.2.16)	1.57883 s	0.02235
SD (ridotto) Alg. 1	0.85960 s	0.02235
Newton (ridotto) (2.3.8)	1.21104 s	0.02235

Tabella 5.1: Tempi di esecuzione e funzionale obiettivo

Si osserva pertanto come sembri conveniente utilizzare metodi di tipo diretto per la risoluzione del problema deterministico. Questo comportamento è dovuto alle contenute dimensioni del problema in esame e non è percorribile nel caso di multiple realizzazioni della variabile casuale, almeno senza utilizzare precondizionatori efficaci, poichè ognuna di esse aggiungerebbe nuove matrici di rigidezza e termini noti, aumentando enormemente le dimensioni del sistema lineare completo. Un metodo efficace per aggirare tale difficoltà consiste nel considerare il problema ridotto, cioè nel risolvere l'equazione di stato e l'aggiunta per l'espressione corrente del controllo, quindi aggiornare la variabile di controllo, costruendo un metodo iterativo.
Nella Figura 5.2, si riporta l'andamento del valore del funzionale obiettivo durante l'esecuzione delle iterazioni di Steepest Descent e Newton, da cui si nota che quest'ultimo sembra essere molto veloce a raggiungere la convergenza, grazie alla struttura lineare-quadratica del problema. Tuttavia la costruzione esplicita della matrice Hessiana

$$\mathbb{H} = \nu\mathbb{M} + (\mathbb{K}^{-1}\mathbb{B})^{\top}\mathbb{M}(\mathbb{K}^{-1}\mathbb{B}), \tag{5.1.7}$$

con $\mathbb{B} \equiv \mathbb{M}$ nel caso distribuito, richiede la risoluzione di un sistema lineare e prodotti con matrici dense, risultando molto onerosa dal punto di vista computazionale. Una soluzione è rappresentata dall'uso del metodo del gradiente coniugato matrix-free anche per il sistema ridotto, che non richiede $\mathbb{H}$ esplicitamente ma come prodotto matrice-vettore $\mathbb{H}(v)$.

Risulta infine interessante considerare l'evoluzione dell'errore rispetto ad una configurazione di riferimento deterministica, al variare della lunghezza caratteristica della mesh o al numero di gradi di libertà. Gli ordini di convergenza attesi, dimostrati nel Capitolo 3, sono rispettati dall'evidenza sperimentale. La Figura 5.3 riassume tale analisi per la forzante f_2.

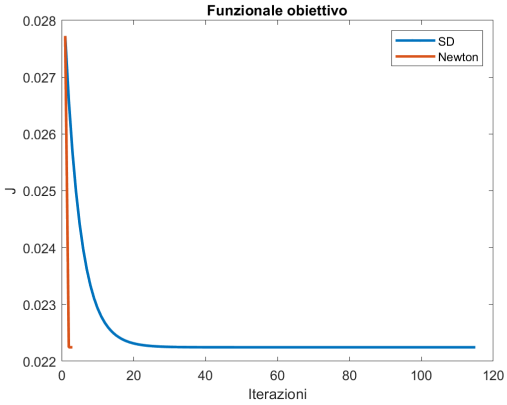
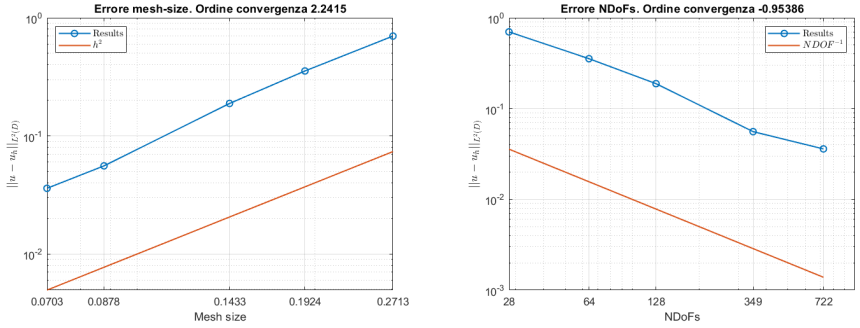


Figura 5.2: Confronto Steepest Descent vs Newton



(a) (b)

Figura 5.3: Errori di discretizzazione spaziale

Caso stocastico

Dopo aver discusso il caso particolare deterministico, si propone un'analisi del problema in condizioni di incertezza, basata sulla tecnica di Sample Average Approximation (SAA), che consiste nel risolvere con un metodo iterativo deterministico il problema mediato sulle realizzazioni della variabile casuale. Per la discretizzazione in probabilità viene utilizzato un metodo Monte-Carlo su 2000 campioni, durante il calcolo della soluzione di riferimento. Relativamente alle espressioni introdotte per la forzante f, viene quindi risolto il problema di controllo ottimo fissando valori decrescenti del parametro di regolarizzazione sulla norma della variabile di controllo $\nu = 10^{-1}, \ldots, 10^{-6}$, per osservare il comportamento del funzionale obiettivo, anche in maniera separata tra i due termini che lo compongono, quello relativo alla distanza tra stato e target e la regolarizzazione sul controllo. L'obiettivo di questa analisi è mostrare la progressiva riduzione del funzionale obiettivo, e l'avvicinamento alla soluzione target della variabile di stato, al diminuire del parametro ν. Tuttavia si osserva come ridurre eccessivamente tale valore, permetterebbe una forte crescita della variabile di controllo, consentendole di assumere valori estremi in alcune zone del dominio,

al fine di spingere la variabile di stato verso il target. Questo rischia di introdurre instabilità numerica nel problema, oltre a rappresentare un limite nel caso sia difficile riprodurre valori così estremi del controllo. Pertanto risulta necessario trovare una sintesi tra l'accuratezza della soluzione e l'energia del controllo, scegliendo in maniera opportuna il valore di ν. Una efficace strategia per visualizzare questo tradeoff consiste nel rappresentare sul piano alcuni punti (tanti quanti i valori di ν studiati) aventi come coordinate la distanza in norma tra y e z_d e la norma della variabile di controllo. Tali punti formano una curva sul piano, detta *curva di Pareto*. Nella Figura 5.4 vengono confrontate le curve di Pareto relative alle espressioni fornite della forzante.

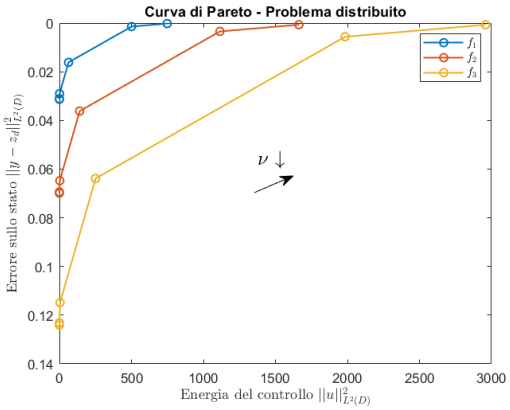


Figura 5.4: Curva di Pareto per il problema (5.1.1)

Per procedere con l'analisi degli ordini di convergenza, nelle Figure 5.6-5.8 vengono riportate le configurazioni di riferimento per $\nu = 10^{-4}$, riferite alle forzanti f_1,f_2,f_3. In particolare, dopo un'immagine della soluzione target z_d, sono raffigurati i risultati per la variabile di stato e per quella di controllo.

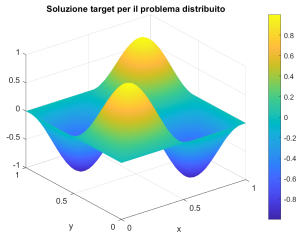
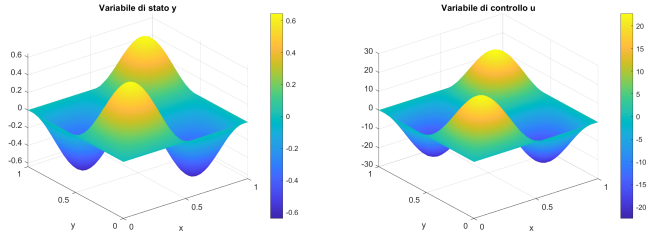


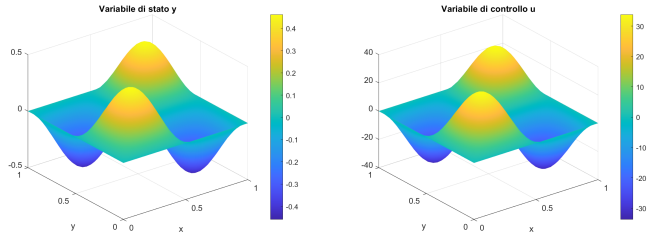
Figura 5.5: Soluzione target z_d per il problema (5.1.1)

Come per il caso deterministico, risulta interessante studiare l'andamento dell'errore relativo alla discretizzazione spaziale, ottenuto riducendo in maniera progressiva la dimensione caratteristica della mesh. A tale scopo, poichè il triangolatore permette di raffinare la mesh solo tramite i parametri di area massima e angolo minimo, risulta necessario calcolare



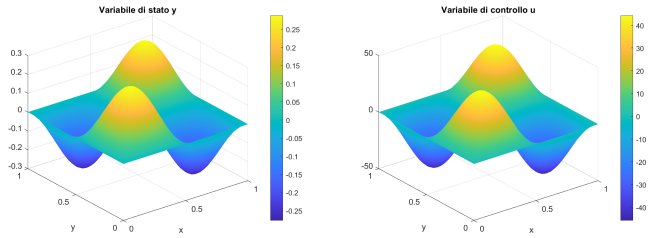
(a) (b)

Figura 5.6: Soluzioni di riferimento relative alla forzante f_1 per (5.1.1)



(a) (b)

Figura 5.7: Soluzioni di riferimento relative alla forzante f_2 per (5.1.1)



(a) (b)

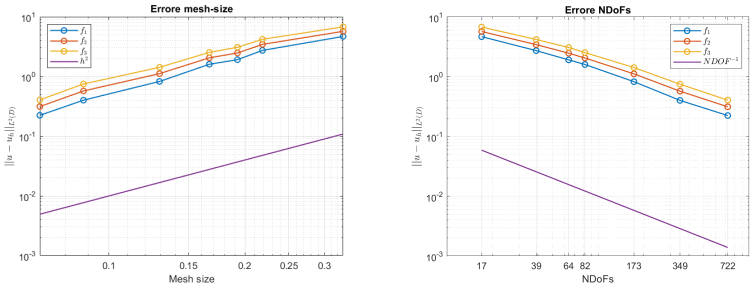
Figura 5.8: Soluzioni di riferimento relative alla forzante f_3 per (5.1.1)

per ogni geometria la quantità

$$h = \max_{e \in \mathcal{T}} h_e, \tag{5.1.8}$$

che rappresenta il lato di lunghezza massima della triangolazione. Inoltre, viene fornita una misura dell'errore anche rispetto al numero di gradi di libertà del sistema, per la quale è logico attendersi un ordine di convergenza dimezzato rispetto a quello relativo alla lunghezza. Infatti si ha che il numero dei gradi di libertà del sistema (cioè dei nodi) è simile all'area del quadratino costruito sul lato di area massima $N_{DoF} \sim \frac{1}{h^2}$.

Oltre all'errore rispetto alla discretizzazione spaziale, viene fornita una conferma sperimentale dell'ordine di convergenza dimostrato per la discretizzazione in probabilità tramite approssimazione Monte Carlo. La soluzione di riferimento fissata, con 2000



(a) (b)

Figura 5.9: Errori sulla discretizzazione spaziale rispetto ad h ed N_{DoF} per (5.1.1)

realizzazioni, viene confrontata con le espressioni ottenute considerando alcuni sottoinsiemi del campione, con dimensione crescente $N = 2^1, \ldots, 2^8$. L'analisi viene condotta in riferimento alla forzante f_2, ma l'output risulta analogo anche per le altre sue espressioni. I risultati delle analisi di convergenza sono presentati nelle Figure 5.9-5.10, mentre nella

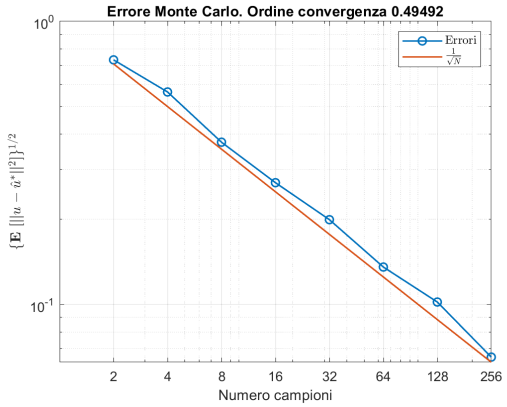


Figura 5.10: Errore di discretizzazione in probabilità per (5.1.1)

Tabella 5.2 sono riassunti gli ordini osservati per la discretizzazione spaziale. Gli ordini di convergenza sono presentati in valore assoluto, tuttavia si ricorda la diversa variazione delle misure di riferimento scelte (il passo h decresce, mentre il numero di gradi di libertà aumenta).

Forzante	**Ordine vs h**	**Ordine vs Ndof**
f_1	2.1552	0.8793
f_2	2.0486	0.8363
f_3	1.9916	0.8131

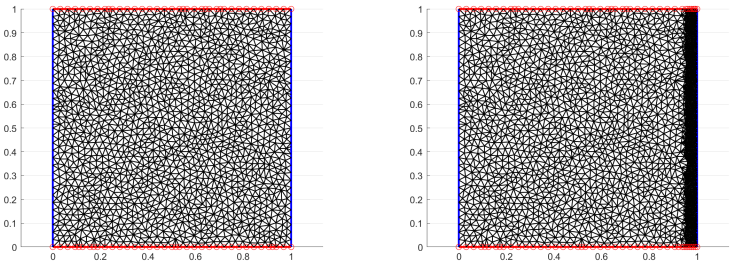
Tabella 5.2: Ordini di convergenza relativi alle Figure 5.9-5.10

5.1.3 Analisi del problema al bordo di Neumann

Si consideri ora il problema (5.1.4), con la variabile di controllo definita sul bordo di Neumann del dominio, in questo caso specifico rappresentato dai lati destro e sinistro del quadrato. In questa analisi, vista l'espressione della soluzione target, verrà maggiormente posta l'attenzione sul bordo destro del quadrato [0,1], dove è concentrata l'energia della soluzione. Per condurre la simulazione, è necessario costruire opportunamente la matrice di massa sui bordi di Neumann in esame B_0 e la relativa matrice $\mathbb{B}$ che mappa il controllo al bordo sullo spazio delle variabili di stato e aggiunta, come descritto nella sezione finale del capitolo 2. Come per il caso distribuito, verranno discussi i risultati ottenuti sia per il problema puramente deterministico, sia per quello stocastico.

Problema deterministico

Prima di addentrarsi nell'analisi dei vari metodi risolutivi, risulta opportuno, in vista dell'analisi stocastica e dell'obiettivo di risolvere problemi vincolati con misure di rischio quali il CVaR, considerare una differente strategia per la discretizzazione spaziale del dominio quadrato. In particolare, è possibile confrontare due differenti approcci: la triangolazione uniforme, analoga a quella utilizzata per il problema distribuito, ed una mesh adattiva, opportunamente raffinata in una piccola regione adiacente al dominio di interesse, rappresentato nel caso in esame dal lato destro del dominio. In questa maniera, è possibile aumentare il numero di nodi sul bordo di Neumann, incrementando l'accuratezza della soluzione per la variabile di controllo. Inoltre, i risultati ottenuti riguardo gli errori di discretizzazione non sembrano essere degradati dal cambio di mesh, pertanto questa scelta è giustificata per l'analisi del caso stocastico e per i problemi *risk-averse*. Nelle figure sottostanti vengono rappresentate le discretizzazioni presentate, fissato come parametro per l'area massima $h = 2^{-11}$ sul dominio e $h = 2^{-15}$ nella zona raffinata vicino al bordo di Neumann.



(a) (b)

Figura 5.11: Mesh uniforme e raffinata

La Tabella 5.3 riassume i risultati ottenuti, con alcuni metodi di ottimizzazione, per la discretizzazione spaziale non uniforme appena introdotta, confrontando i tempi di calcolo e il risultato ottenuto in termini di funzionale obiettivo

Metodo	Tempo	Funzionale
One-shot (3.6.4)	1.5868 s	0.5843
SD (ridotto) Alg. 1	174.3395 s	0.5843
Newton (ridotto) (2.3.8)	3.6495 s	0.5843

Tabella 5.3: Tempi di esecuzione e funzionale obiettivo

Si osserva come il metodo diretto sia preferibile per il caso deterministico, mentre il metodo steepest descent risulta molto lento in virtù dell'elevato numero di iterazioni. Tuttavia, aumentando la dimensione del campione o raffinando ulteriormente la mesh, il calcolo esplicito della matrice Hessiana influenzerà negativamente il metodo di Newton, equilibrando il costo computazionale rispetto a Steepest Descent. Come per il caso deterministico, sono confrontati i metodi di Newton e Steepest Descent rappresentando la progressione dei valori del funzionale obiettivo, nella Figura 5.12.

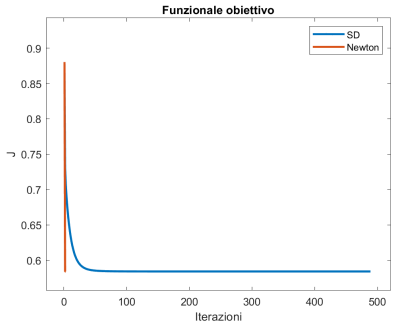
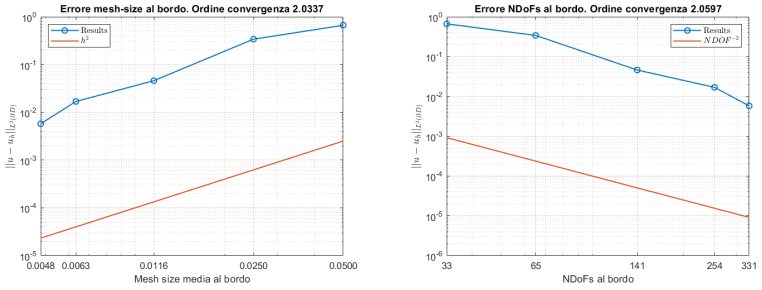


Figura 5.12: Confronto Steepest Descent vs Newton

Infine, a dimostrazione del fatto che il cambio di triangolazione non influisce sull'ordine di convergenza dell'errore di discretizzazione rispetto ad una configurazione di riferimento, nelle Figure 5.13-5.14 sono rappresentati i risultati ottenuti variando la meshsize e di conseguenza il numero di gradi di libertà, sia per la variabile di controllo che per quella di stato. Nel caso con controllo al bordo, l'ordine di convergenza atteso è quadratico sia rispetto alla lunghezza caratteristica, che in questo caso viene fissata pari al segmento medio sul bordo di Neumann, sia rispetto al numero di gradi di libertà, cioè il numero di nodi sul bordo. Questo segue dal fatto che, in un dominio unidimensionale, il numero dei gradi di libertà è inversamente proporzionale al passo di discretizzazione.

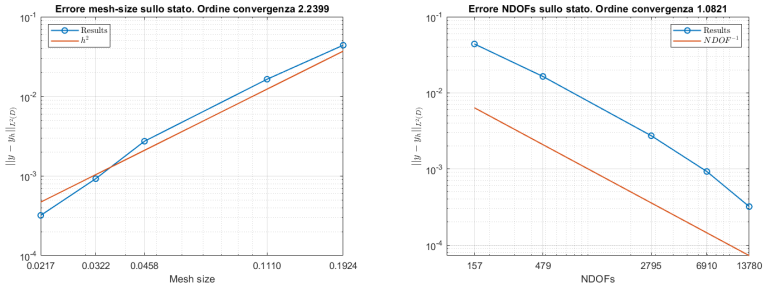
Problema stocastico

A conclusione della sezione dedicata ai problemi di controllo ottimo non vincolati, vengono presentati i risultati relativi al problema stocastico con controllo al bordo (5.1.4).
La soluzione target viene definita solo in una porzione del dominio, tramite la funzione indicatrice, come nell'espressione (5.1.6), e rappresentata nella Figura 5.15.



(a) (b)

Figura 5.13: Errori di discretizzazione con la mesh adattiva sul controllo



(a) (b)

Figura 5.14: Errori di discretizzazione con la mesh adattiva sullo stato

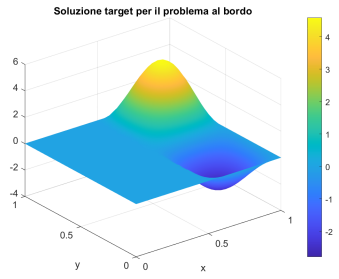


Figura 5.15: Soluzione target z_d per il problema (5.1.4)

Viene riproposta in Figura 5.16 l'analisi relativa al parametro di regolarizzazione ν e la raffigurazione della curva di Pareto per valutare il trade-off tra accuratezza nella soluzione della PDE di stato ed energia del controllo richiesta.

Nelle Figure 5.17–5.19 sono presentate le soluzioni di riferimento relative alle espressioni della forzante f indicate nel problema (5.1.4).

Analogamente al caso distribuito, risulta necessario studiare l'andamento della soluzione rispetto ad una configurazione di riferimento variando la dimensione della mesh o il numero di campioni. Per quanto riguarda la discretizzazione spaziale, le Figure 5.20-5.21 mostrano

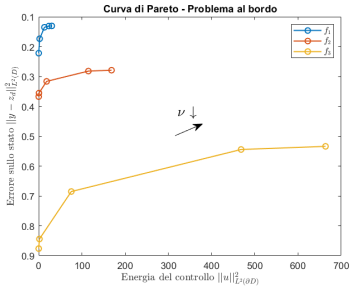
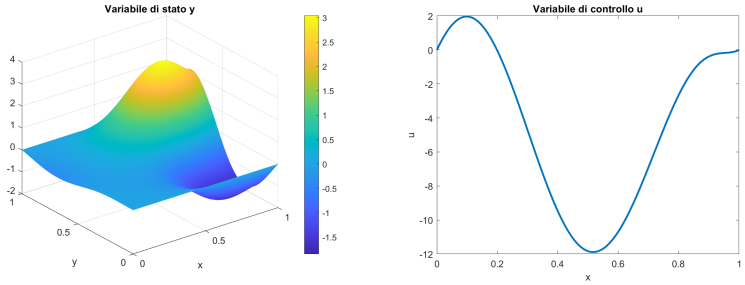
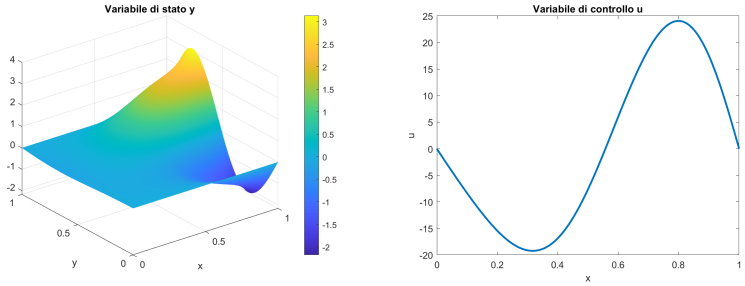


Figura 5.16: Curva di Pareto per il problema (5.1.4)



(a) (b)

Figura 5.17: Soluzioni di riferimento relative alla forzante f_1 per (5.1.4)

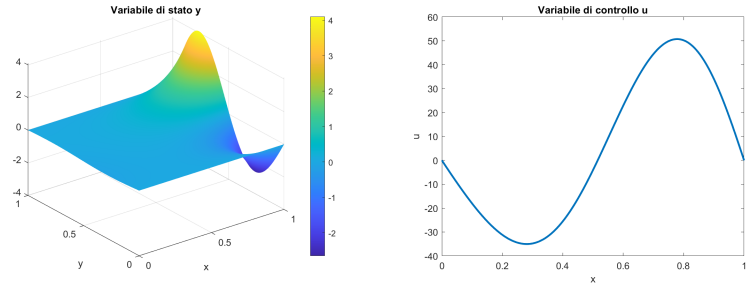


(a) (b)

Figura 5.18: Soluzioni di riferimento relative alla forzante f_2 per (5.1.4)

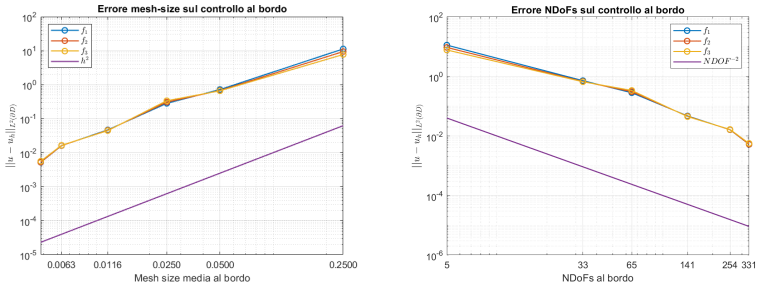
gli ordini di convergenza sia dell'errore sulla variabile di controllo, definita sul bordo di Neumann, (confrontato quindi rispetto al numero di nodi sul bordo e al relativo passo di lunghezza media) e sulla variabile di stato, definita sull'intero dominio (confrontata con lunghezza massima del passo e N_{DoF} totali).

Infine, fissata f_2 come forzante in esame viene riportata l'analisi della convergenza del metodo Monte Carlo, in Figura 5.22.



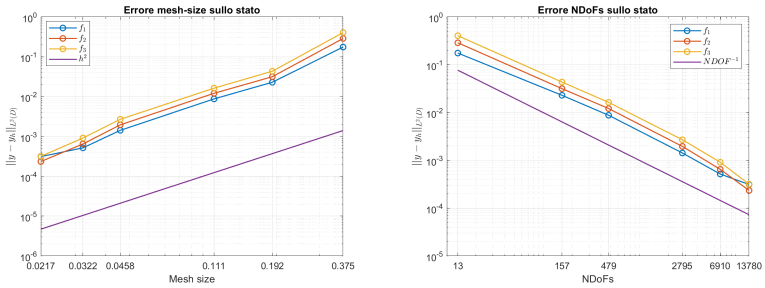
(a) (b)

Figura 5.19: Soluzioni di riferimento relative alla forzante f_3 per (5.1.4)



(a) (b)

Figura 5.20: Errore di discretizzazione spaziale sul controllo per (5.1.4)



(a) (b)

Figura 5.21: Errore di discretizzazione spaziale sullo stato per (5.1.4)

Nella Tabella 5.4 sono riassunti i risultati di convergenza per la discretizzazione spaziale, riferiti alla variabile di controllo e a quella di stato, rispetto a dimensione caratteristica e numero di gradi di libertà.

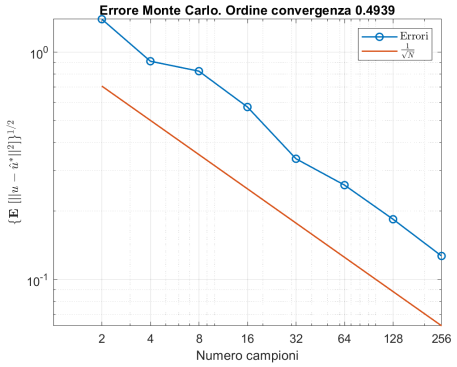


Figura 5.22: Errore di discretizzazione in probabilità per (5.1.4)

	Controllo		Stato	
Forzante	**Ordine vs h**	**Ordine vs Ndof**	**Ordine vs h**	**Ordine vs Ndof**
f_1	2.0677	2.0937	2.0406	0.9868
f_2	2.0563	2.0823	2.2451	1.0848
f_3	2.0441	2.0702	2.2416	1.0829

Tabella 5.4: Ordini di convergenza relativi alle Figure 5.20–5.22

5.2 Problemi di controllo ottimo vincolati

Questa sezione contiene i risultati numerici relativi ai problemi vincolati di tipo *risk-averse* descritti nel quarto capitolo.

L'analisi viene condotta con l'obiettivo di confrontare i due approcci risolutivi presentati

- Riformulazione epigrafica e soluzione tramite metodo interior-point primale;
- Smoothing-splitting e soluzione tramite metodo primal-dual interior-point.

In particolare, questi metodi saranno applicati a due problemi modello vincolati dal CVaR, uno con l'obiettivo di minimizzare un funzionale risk-neutral e l'altro un funzionale risk-averse, anch'esso contenente la medesima misura di rischio, calcolata su una diversa quantità di interesse.

5.2.1 Problemi modello e scelte computazionali

I problemi considerati sono i seguenti

- *Problema 1 - Funzionale risk-neutral*

$$\begin{aligned} &\min_{u\in\mathcal{U}} \hat{J}(u) = \frac{1}{2}\mathbb{E}[||S_\omega(f+Bu)-z_d||_V^2] + \frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ &\text{s.t. } \mathrm{CVaR}_\beta\left(\frac{1}{2}||S_\omega(f+Bu)||_V^2\right) \leq \tau. \end{aligned} \tag{5.2.1}$$

- *Problema 2 - Funzionale risk-averse*

$$\begin{aligned} &\min_{u\in\mathcal{U}} \hat{J}(u) = \mathrm{CVaR}_\beta\left(\frac{1}{2}||S_\omega(f+Bu)-z_d||_V^2\right)+\frac{\nu}{2}||u||_{\mathcal{U}}^2 \\ &\text{s.t. } \mathrm{CVaR}_\beta\left(\frac{1}{2}||S_\omega(f+Bu)||_V^2\right)\leq\tau. \end{aligned} \tag{5.2.2}$$

Il dominio di riferimento è sempre rappresentato dal quadrato $[0,1]^2$, discretizzato per mezzo di una mesh adattiva generata dal triangolatore *bbtr30*, come nella Figura 5.11. L'equazione differenziale alle derivate parziali scelta rimane quella di Laplace, con controllo definito sul bordo di Neumann, analogamente a quella presentata nel problema non vincolato espressa dalla (5.1.5). Inoltre sono fissati il parametro di regolarizzazione $\mu = 10^{-4}$, il livello di confidenza $\beta = 0.95$ e la soglia τ, pari alla metà del valore di riferimento del CVaR calcolata utilizzando la soluzione del corrispondente problema *unconstrained*. La tolleranza sulla riduzione del parametro di barriera μ è fissata pari a $tol_\mu = 10^{-9}$, inoltre viene fissato, per ogni valore di μ il seguente criterio di arresto per le iterazioni interne

$$\frac{\lambda^2}{2} < \eta\mu, \qquad \eta = 0.1. \tag{5.2.3}$$

Nei paragrafi seguenti sono analizzati i risultati forniti da ogni metodo, infine viene condotta una comparazione tra i risultati ottenuti per stabilire vantaggi e svantaggi computazionali di ogni metodo e fornire un'interpretazione delle soluzioni ottenute rispetto ai diversi funzionali obiettivo considerati.

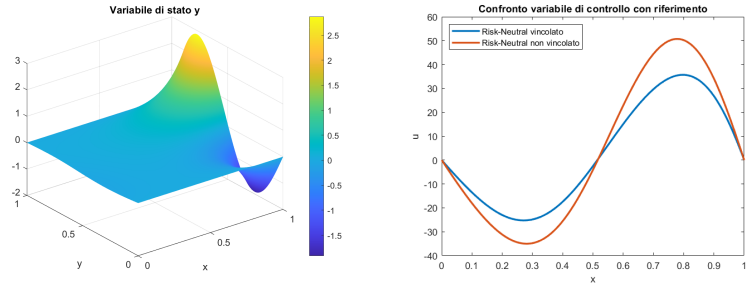
5.2.2 Minimizzazione risk-neutral

In questa sezione viene analizzato il problema (5.2.1) presentato in precedenza, fornendo alcuni risultati e rappresentazioni significative relativamente all'applicazione dei metodi descritti nella parte teorica.

Riformulazione epigrafica

Il primo algoritmo testato prevede una riformulazione del vincolo sul CVaR di tipo epigrafico, con l'introduzione, oltre di una variabile per il quantile, di un vettore di variabili ausiliarie $\mathbf{z}$, che risultano convergenti a 0 per quei campioni i cui valori del funzionale di interesse sono minori del quantile ed assumono un valore positivo per i valori che eccedono il quantile, misurandone la distanza da quest'ultimo.
I grafici della Figura 5.23 rappresentano la soluzione ottenuta per la variabile di stato e per quella di controllo, sul bordo destro del dominio, evidenziandone la differenza con la soluzione ottima del problema non vincolato.

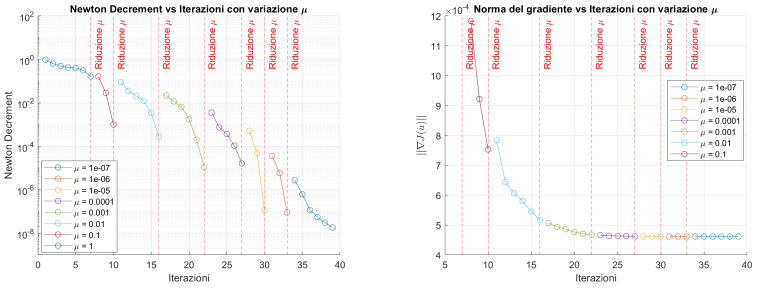
Si osserva una riduzione dell'energia della variabile di controllo, che si riflette su quella di stato, come richiesto dal vincolo imposto. Infatti, il limite sul Conditional Value-at-Risk ha l'effetto di ridurre la variabilità del controllo, agendo sugli scenari peggiori, rappresentati dalla coda della distribuzione, rendendolo più robusto anche in condizioni sfavorevoli o estreme. Questa strategia più conservativa nella scelta del controllo ottimo, richiede



(a) (b)

Figura 5.23: Soluzione per variabile di stato (a) e di controllo (b)

tuttavia un costo aggiuntivo in termini di funzionale obiettivo, che raggiunge un valore ottimale maggiore rispetto al caso non vincolato. L'analisi comparativa dei valori finali dell'obiettivo sarà oggetto di studio nella sezione conclusiva.
Per monitorare l'andamento del metodo, al decrescere del parametro di barriera μ, è opportuno rappresentare il comportamento di alcune quantità notevoli relative al sistema, quali il decremento di Newton e la norma del gradiente, come in Figura 5.24.

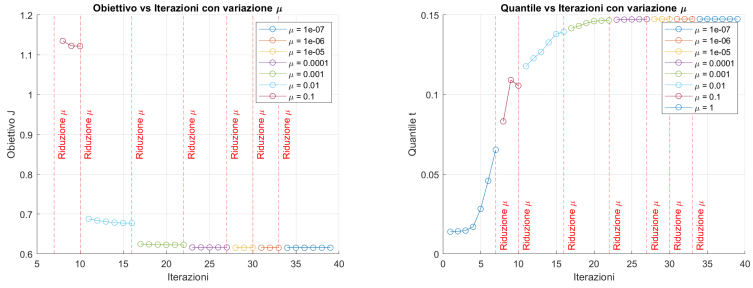


(a) (b)

Figura 5.24: Andamento del decremento di Newton (a) e della norma del gradiente (b)

Il decremento di Newton si riduce progressivamente avvicinandosi alla soluzione ottima, osservando tuttavia alcuni "salti" nel valore dovuti alla riduzione del parametro di barriera. La convergenza del gradiente è limitata dalla presenza del vincolo sul CVaR, che non permette il raggiungimento del valore ottimo effettivo per il funzionale obiettivo in esame. Risulta inoltre interessante descrivere il comportamento, oltre che del funzionale obiettivo, della variabile ausiliaria t, il cui valore ottimo fornito dal metodo rappresenta un'approssimazione del Value-at-Risk, ovvero dell'α-quantile della distribuzione. Tali risultati sono riassunti nella Figura 5.25, osservando per entrambi la convergenza ad un valore ottimale.

Infine, risulta opportuno mostrare l'avvenuta convergenza delle variabili ausiliarie $\mathbf{z}$. La Figura 5.26 evidenzia il numero di valori non nulli ottenuti, confermando l'ipotesi teorica relativa alla loro proporzione all'interno del campione.



(a) (b)

Figura 5.25: Andamento del funzionale obiettivo (a) e di t (b)

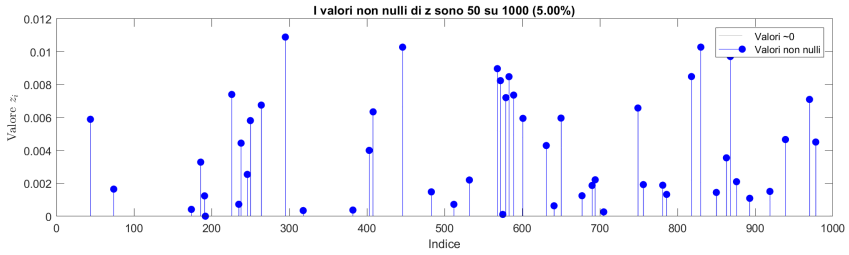
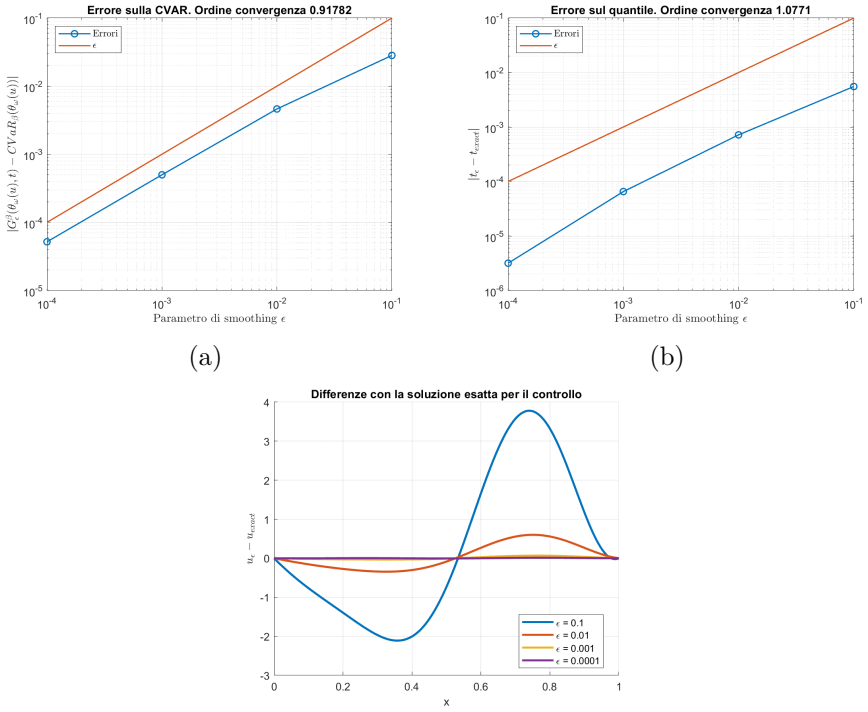


Figura 5.26: Valori della variabile ausiliaria $\mathbf{z}$

Smoothing-splitting

La seconda strategia risolutiva per il problema di controllo ottimo vincolato da CVaR consiste nel regolarizzare (*smoothing*) la misura di rischio tramite una funzione g_ϵ, risolvendo poi, a differenza della riformulazione epigrafica, una versione approssimata del problema originale. La variabile t relativa al VaR viene introdotta nel vincolo come funzione implicita del controllo. Il suo valore viene determinato ad ogni iterazione in maniera separata dall'ottimizzazione del controllo (*splitting*), applicando un metodo di bisezione sui valori correnti del funzionale di interesse.
Oltre ai risultati relativi alla convergenza del metodo, simili a quelli della riformulazione epigrafica, risulta interessante valutare l'errore introdotto dall'approssimazione, al variare del parametro di smoothing ϵ.

La Figura 5.27 mostra l'andamento delle quantità statistiche in esame, quali il CVaR e il VaR ottenute dall'applicazione del metodo con valori decrescenti del parametro $\epsilon = 10^{-1}, \ldots, 10^{-4}$, assumendo come soluzione di riferimento quella ottenuta dal metodo precedente. Si osserva come venga rispettato l'ordine di convergenza lineare rispetto ad ϵ per CVaR e quantile, nella Figura 5.27 (a-b), a conferma del risultato teorico di convergenza espresso dal Teorema 4.3.1. Infine, per dare una misura dell'accuratezza della soluzione, sulla variabile di controllo, viene rappresentata, nel grafico Figura 5.27 (c), la differenza tra la soluzione ottima ricavata per ogni valore di ϵ testato e il controllo risultante dal metodo esatto epigrafico.

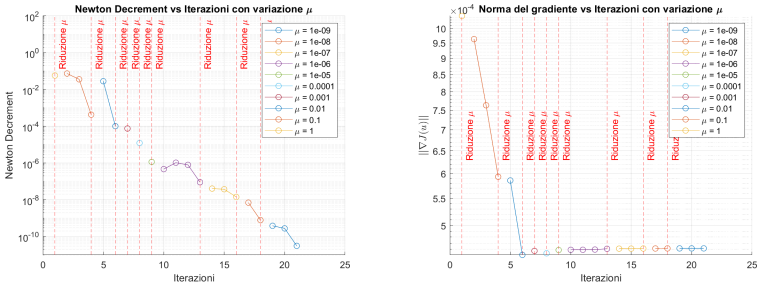


(a) (b)

(c)

Figura 5.27: Andamento rispetto a ϵ di CVaR (a) e quantile (b), differenze sulle soluzioni per la variabile di controllo (c)

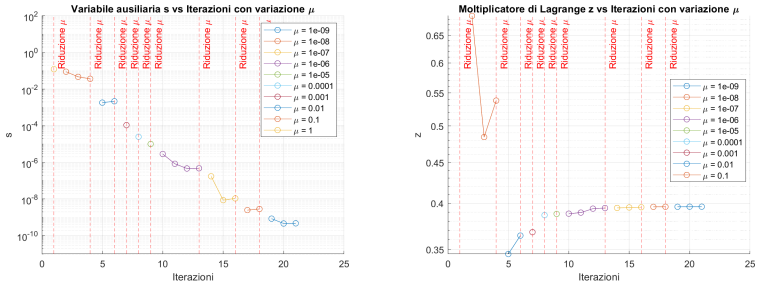
La soluzione ottenuta per le variabili di stato e di controllo, fissato $\epsilon = 10^{-4}$, risulta essere molto simile a quella epigrafica, pertanto è possibile mantenere la rappresentazione della Figura 5.23 per la sua visualizzazione. Risulta invece opportuno mostrare, in Figura 5.28, l'andamento delle misure caratteristiche del metodo di Newton quali il decremento e la norma del gradiente, utili a monitorare la convergenza.



(a) (b)

Figura 5.28: Andamento del decremento di Newton (a) e della norma del gradiente (b)

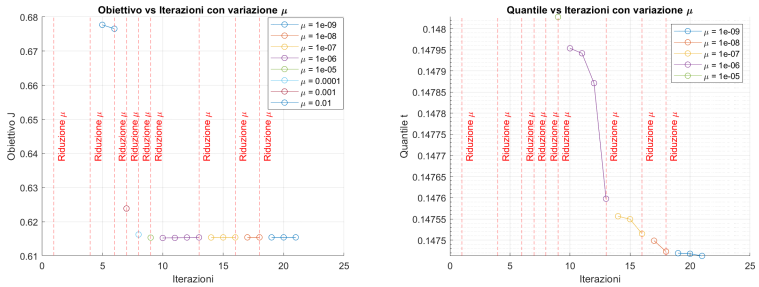
A differenza del caso con riformulazione epigrafica, non è più necessario che il vincolo sul CVaR sia strettamente soddisfatto ad ogni iterazione. Nel metodo primal-dual interior point viene introdotta una variabile ausiliaria s che trasforma il vincolo in un'uguaglianza. Tale variabile, invece, deve necessariamente rispettare un vincolo di non-negatività, come mostrato nella Figura 5.29 (a). Nella Figura 5.29 (b) viene invece rappresentato l'andamento del moltiplicatore di Lagrange associato al vincolo di uguaglianza.



(a) (b)

Figura 5.29: Andamento della variabile ausiliaria s (a) e del moltiplicatore di Lagrange ζ (b)

Nella Figura 5.30 viene infine mostrato il comportamento del funzionale obiettivo e del quantile, nelle ultime iterazioni. Come si può osservare, il funzionale obiettivo sembra ridursi fino ad oltrepassare la soluzione ottima del problema vincolato e contemporaneamente il quantile cresce sopra il suo valore ottimale, per poi stabilizzarsi entrambi al decrescere del parametro di barriera μ. Questo comportamento è permesso dal metodo primal-dual implementato, in quanto il vincolo non deve più essere sempre strettamente soddisfatto ad ogni iterazione, essendo stato trasformato in un vincolo di uguaglianza. Solo la variabile ausiliaria s è ora vincolata ad essere non negativa, come già osservato nell'immagine precedente.



(a) (b)

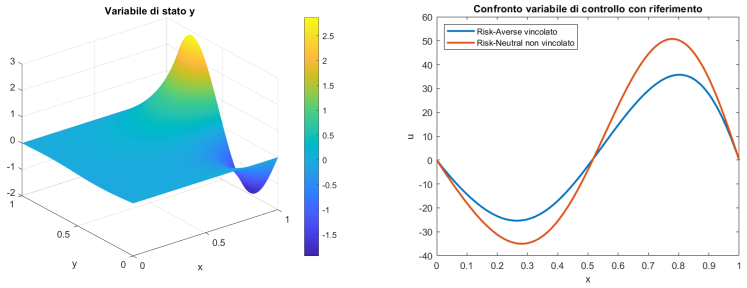
Figura 5.30: Andamento del funzionale obiettivo (a) e del quantile (b)

5.2.3 Minimizzazione risk-averse

In questa sezione viene invece analizzato il problema (5.2.2), in cui il funzionale obiettivo contiene, oltre alla regolarizzazione del controllo, un termine di tipo risk-averse, rappresentato dal CVaR del funzionale di interesse dato dalla distanza in norma tra la variabile di stato y e la soluzione target. Anche per questo problema, come per quello con funzionale risk-neutral vengono analizzati due metodi risolutivi, la riformulazione epigrafica e il metodo smoothing-splitting.

Riformulazione epigrafica

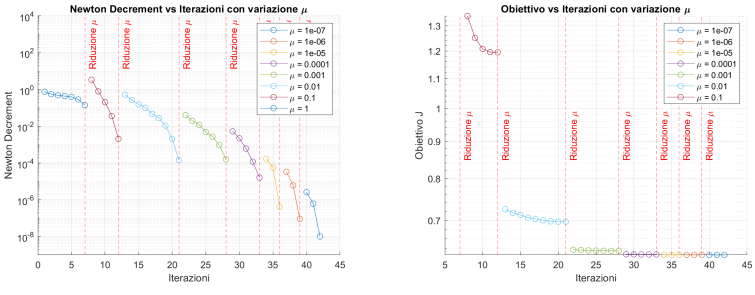
Per ottenere una riformulazione epigrafica del problema di controllo ottimo risk-averse vincolato da CVaR, sono necessari, oltre ai $2N+1$ vincoli aggiuntivi utilizzati nel problema risk-neutral, ulteriori $2N$ vincoli per riformulare il CVaR del funzionale obiettivo.
In Figura 5.31 sono rappresentate le configurazioni ottimali della variabile di stato e di quella di controllo.



(a) (b)

Figura 5.31: Soluzione per variabile di stato (a) e di controllo (b)

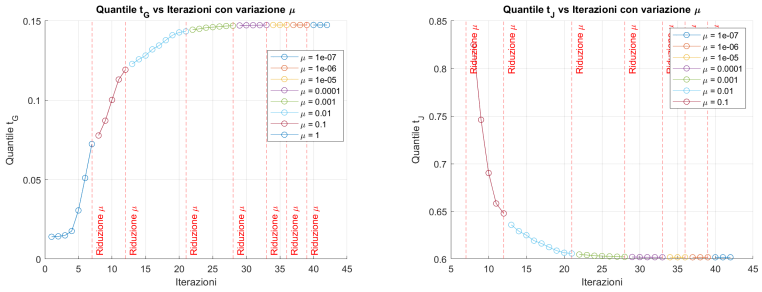
Per certificare l'avvenuta convergenza del metodo, è necessario visualizzare, come di consueto, l'andamento del decremento di Newton, rappresentato in Figura 5.32 (a).



(a) (b)

Figura 5.32: Andamento del decremento di Newton (a) e del funzionale obiettivo (b)

Anche in questo caso si osserva la diminuzione del decremento di Newton, con alcuni salti dovuti alla variazione del parametro di barriera.
La Figura 5.32 (b) riassume invece la convergenza del valore del funzionale obiettivo verso la soluzione ottima.
L'analisi dei risultati ottenuti dal metodo viene quindi condotta verificando il comportamento delle variabili ausiliarie introdotte per gestire contemporaneamente vincolo CVaR e misura di rischio nel funzionale. A tale scopo, saranno indicizzate tramite la lettera G quelle relative ai vincoli derivanti dalla riformulazione epigrafica del CVaR nel vincolo, mentre con la lettera J per quelle riferite al CVaR dell'obiettivo. In prima analisi, è utile alla comprensione del metodo valutare l'andamento dei quantili t_G e t_J, osservandone la convergenza asintotica al procedere delle iterazioni, come in Figura 5.33.



(a) (b)

Figura 5.33: Andamento dei quantili t_G (a) e t_J (b)

Infine, nelle Figure 5.34-5.35 sono riportati i valori non nulli delle variabili ausiliarie $\mathbf{z}_G$, $\mathbf{z}_J$, che rappresentano gli elementi delle code delle distribuzioni individuate dal CVaR.

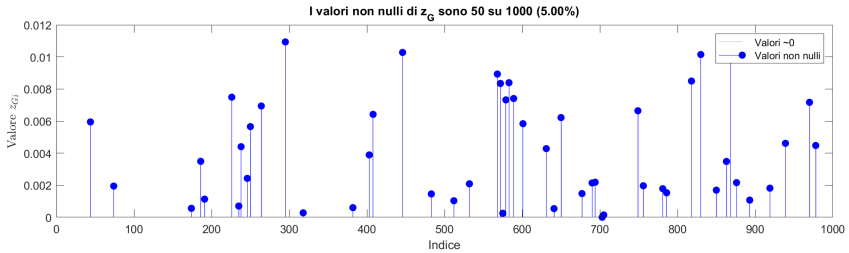


Figura 5.34: Valori della variabile ausiliaria $\mathbf{z}_G$

Osservazione 5.2.1. *Dai risultati riportati, si osserva una variazione molto attenuata nell'andamento del funzionale obiettivo durante il passaggio da risk-neutral a risk-averse. Questo comportamento potrebbe essere dovuto alla presenza del vincolo, che penalizza già abbastanza valori estremi della variabile di controllo (e quindi di quella di stato), oppure al*

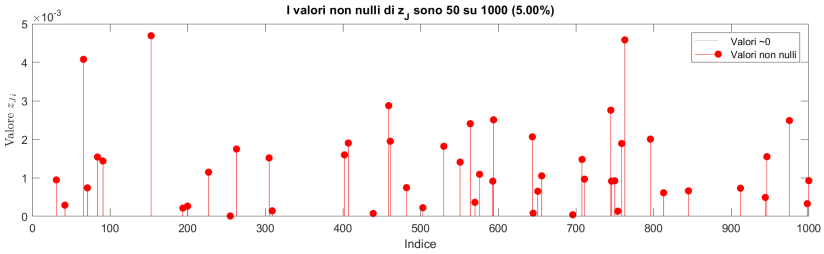


Figura 5.35: Valori della variabile ausiliaria $\mathbf{z}_J$

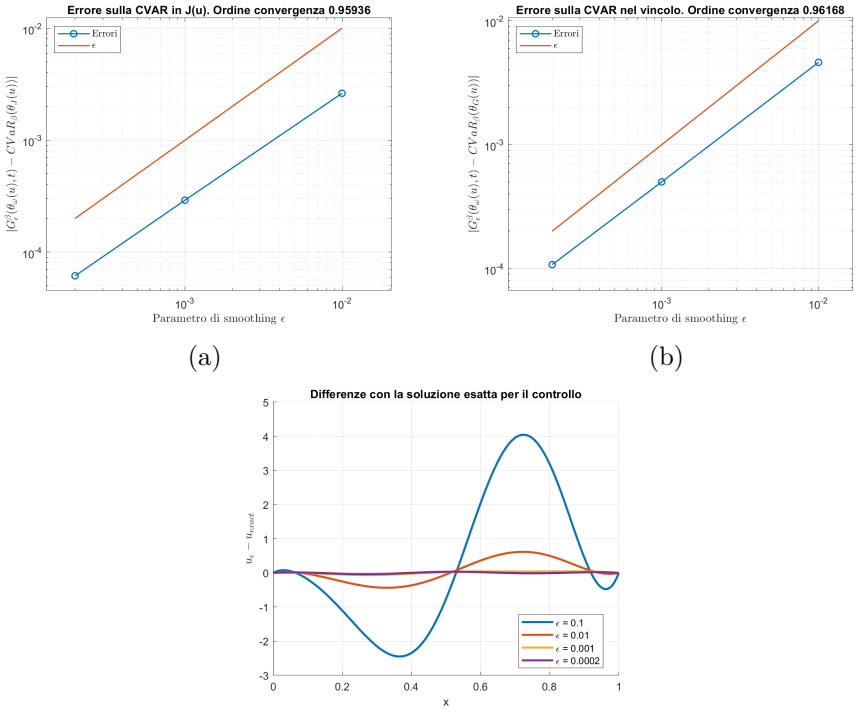
valore basso del parametro di regolarizzazione ν, che avvicina molto lo stato alla soluzione target, rendendo piccole le distanze in norma $L^2(D)$ tra queste variabili.

Smoothing-splitting

Infine, in quest'ultimo paragrafo, saranno riportati i risultati relativi all'implementazione del metodo smoothing-splitting applicato al problema con funzionale risk-averse. Si osserva che, in questo caso, risulta necessario applicare lo splitting sia alla variabile introdotta dal CVaR nell'obiettivo, sia da quella nel vincolo, per ricavare quindi le condizioni KKT e risolvere il problema con un metodo primal-dual interior point.
La risoluzione di questo problema è fortemente condizionata dalla scelta del parametro di smoothing ϵ. Come per il caso risk-neutral, nelle immagini seguenti vengono analizzati gli errori commessi sul CVaR e sul quantile relativi al vincolo per alcuni valori di tale parametro, oltre ad una rappresentazione dell'andamento della distanza tra le soluzioni ottenute e il riferimento calcolato tramite il metodo epigrafico.

Si osserva ancora una volta, come nel caso con funzionale risk-neutral, la convergenza delle soluzioni per la variabile di controllo, che si avvicinano alla soluzione esatta riducendo il parametro di smoothing. Inoltre l'ordine di convergenza lineare rispetto ad ϵ viene preservato per le misure di rischio stimate tramite smoothing, quali i CVaR del vincolo e dell'obiettivo, rappresentate in Figura 5.36, e i relativi quantili. La convergenza del metodo è assicurata dalla riduzione fino ad una soglia fissata $tol = 10^{-10}$ del decremento di Newton, come si osserva nella Figura 5.37 (a). In aggiunta, nella Figura 5.37 (b) viene monitorato l'andamento del gradiente del funzionale obiettivo originale, che rimane non nullo, convergendo ad un valore fissato, per la discrepanza tra l'ottimo del problema non vincolato e la soluzione ottenuta.

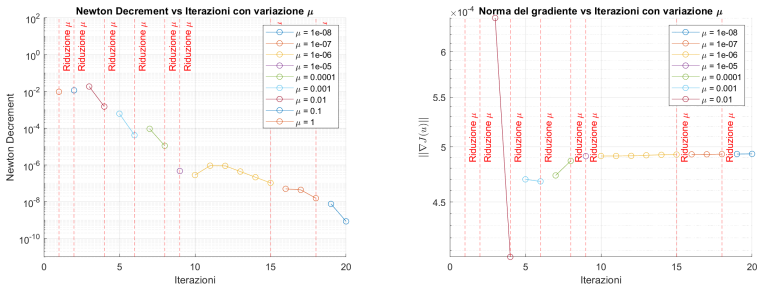
I grafici della Figura 5.38 riportano l'andamento del CVaR calcolata dal metodo per il termine nel funzionale obiettivo e quello nel vincolo, osservando una stabilizzazione verso valori fissati: l'ottimo per il funzionale obiettivo ed un valore vicino alla soglia per il vincolo, sintomo della sua attivazione in prossimità della soluzione ottima.



(a) (b)

(c)

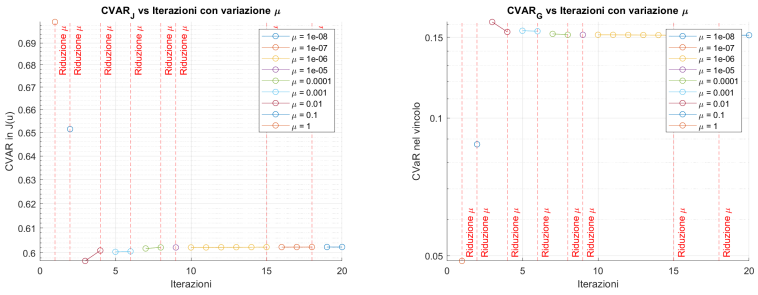
Figura 5.36: Andamento rispetto a ϵ di CVaR (a) e quantile (b), differenze sulle soluzioni per la variabile di controllo (c)



(a) (b)

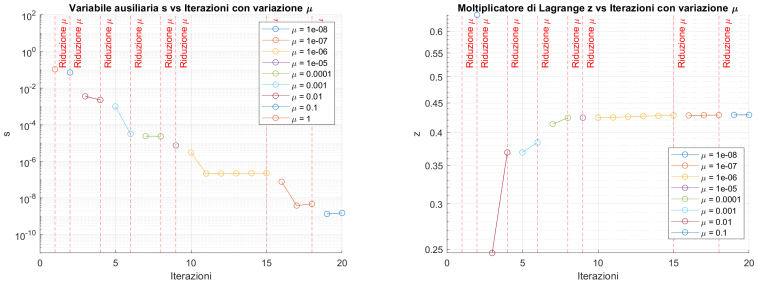
Figura 5.37: Andamento del decremento di Newton (a) e della norma del gradiente (b)

Infine, è opportuno riportare, in Figura 5.39, l'andamento della slack variable s, la cui positività certifica il rispetto del vincolo, che si riduce progressivamente durante le iterazioni, e quello del moltiplicatore di Lagrange ζ, che a sua volta sembra convegere verso un valore fissato.



(a) (b)

Figura 5.38: Andamento della CVaR nel funzionale obiettivo (a) e di quella nel vincolo (b)



(a) (b)

Figura 5.39: Andamento della variabile ausiliaria s (a) e del moltiplicatore di Lagrange ζ (b)

5.2.4 Confronto dei risultati

Infine, quest'ultimo paragrafo è dedicato ad un'analisi comparativa dei metodi numerici implementati, per evidenziarne pregi e difetti in termini di accuratezza computazionale e costo sostenuto per l'implementazione. Le Tabelle 5.5-5.6 riassumono i valori di alcune quantità notevoli ottenute dall'applicazione dei metodi per problemi vincolati da CVaR, mettendole a confronto con i valori ottimali ricavati per il problema unconstrained con funzionale risk-neutral. I problemi sono indicati con una sigla identificativa costituita dal tipo di funzionale obiettivo, (RA) per risk-averse e (RN) per risk-neutral, e dal metodo utilizzato, (E) per la riformulazione epigrafica e (S) per il metodo smoothing-splitting. La lettera (J) indica il risultato ottenuto rispetto al funzionale di interesse relativo al termine nel funzionale obiettivo, mentre (G) indica il valore relativo al vincolo.

Analizzando la prima tabella, relativa ai risultati ottenuti per le misure di rischio in esame, si osserva come tutti i metodi testati controllino i valori del CVaR, e di conseguenza del quantile associato, contenuta nel vincolo imposto, adattandoli alla soglia fissata τ, pari a metà del valore relativo al problema non vincolato, al fine di attivare il vincolo. Per quanto riguarda invece il CVaR del funzionale di interesse nell'obiettivo, dai risultati si deduce un fisiologico aumento rispetto al dato unconstrained causato dall'imposizione del

Problema	**Obiettivo**	**CVaR (J)**	**CVaR (G)**	**VaR (J)**	**VaR (G)**
RN-Unconstrained	0.5914	0.5567	0.3032	0.5545	0.2948
RN-E	0.6155	0.6033	0.1516	0.6020	0.1475
RN-S	0.6153	0.6030	0.1516	0.6017	0.1474
RA-E	0.6270	0.6031	0.1516	0.6018	0.1474
RA-S	0.6263	0.6023	0.1516	0.6010	0.1474

Tabella 5.5: Valori dell'obiettivo, della CVaR e dei quantili

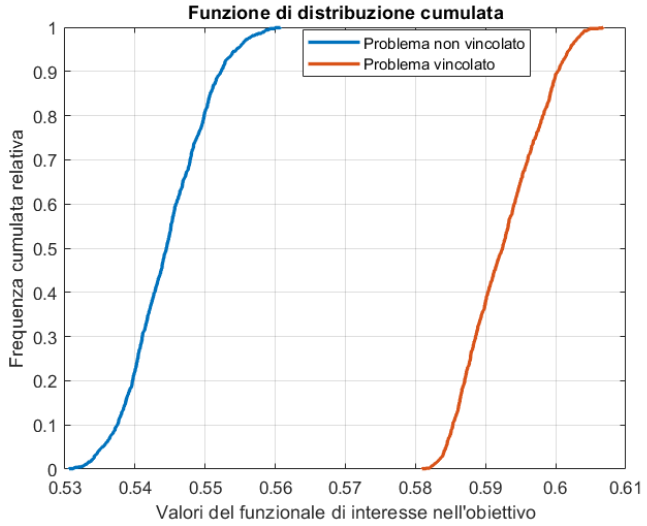
Problema	**CVaR (J)**	**CVaR (G)**	**VaR (J)**	**VaR (G)**
RN-E	-	2.0056e-07	-	3.0077e-07
RN-S	-	4.9516e-04	-	5.5352e-04
RA-E	1.9979e-08	4.6407e-08	5.0626e-08	5.1634e-08
RA-S	5.4733e-04	5.1068e-04	5.4266e-04	4.9733e-04

Tabella 5.6: Errori su CVaR e quantile

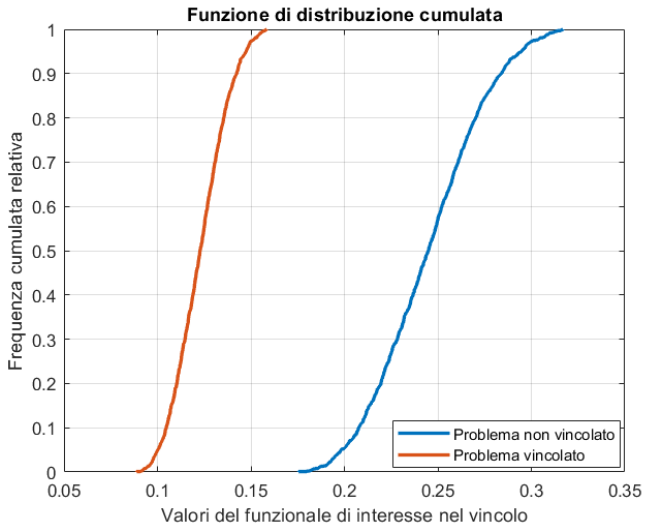
vincolo, che peggiora la soluzione ottima, causando di conseguenza la crescita del funzionale obiettivo. Inoltre, confrontando i dati relativi ai due problemi vincolati considerati (RN vs RA), si osserva la differenza tra la minimizzazione del CVaR e quella della media delle realizzazioni: nel primo caso il valore in tabella risulta inferiore, seppure in minima parte, comportamento dovuto alla bassa varianza dei dati, rispetto al corrispondente risultato per il problema risk-neutral. Come atteso, il funzionale obiettivo risulta maggiore nel caso risk-averse, in quanto la minimizzazione del CVaR genera soluzioni più conservative rispetto alla variabile di controllo. Osservando invece la tabella relativa agli errori di approssimazione tra il risultato ottenuto dai metodi per CVaR e quantili ed il loro valore reale calcolato sulla soluzione per la variabile di controllo, si osserva una discrepanza tra il metodo epigrafico e il metodo smoothing-splitting. In particolare, quest'ultimo metodo introduce un'approssimazione del CVaR tramite una funzione di smoothing, regolata dal parametro ϵ, che risulta limitare l'accuratezza della soluzione ottenuta, per via del risultato teorico dato dal Teorema 4.3.1. Infatti, fissato per l'analisi $\epsilon = 10^{-3}$, si osserva che tutti gli errori relativi a metodi di tipo smoothing-splitting risultano limitati a circa metà del valore di tale parametro. Invece, il metodo epigrafico non approssima la misura di rischio, mantenendo il vincolo originale, solamente riformulato per essere computazionalmente trattabile. Questa strategia risulta pertanto molto più accurata nella stima delle misure di interesse, come dimostrato dagli errori ottenuti rispetto ai valori reali.

Nella Figura 5.40 è possibile osservare l'effetto sui funzionali di interesse nell'obiettivo (4.4.2), in 5.40 (a), e nel vincolo (4.4.3), in 5.40 (b), dovuto alla presenza del vincolo sul CVaR, confrontando il caso RN-Unconstrained con RN-E. Nel caso vincolato, la saturazione della distribuzione cumulata di probabilità del funzionale (4.4.3) avviene in corrispondenza di un valore più basso rispetto al caso non vincolato, in quanto il relativo CVaR è limitato dalla soglia τ. Viceversa, per il funzionale (4.4.2), si osserva che il rispetto del vincolo sul CVaR comporta valori del funzionale obiettivo più elevati rispetto al caso non vincolato,

coerentemente con quanto atteso.



(a) Funzionale di interesse nell'obiettivo (4.4.2)



(b) Funzionale di interesse nel vincolo (4.4.3)

Figura 5.40: Confronto dei funzionali di interesse tra problema non vincolato e problema vincolato da CVAR

Infine, a conclusione di questa analisi, risulta utile confrontare, in Tabella 5.7, alcuni parametri relativi al costo computazionale, in termini di tempo richiesto e numero di PDE risolte, sostenuto da ogni metodo implementato.

Questi dati confermano l'ipotesi di maggiore complessità computazionale richiesta dai metodi basati sulla riformulazione epigrafica del funzionale obiettivo. L'introduzione di

Problema	**Tempo**	**Tempo/iters**	**PDE**	**PDE/iters**
RN-E	5392.17 s	138.26 s	245000	6282
RN-S	2895.38 s	137.88 s	94000	4476
RA-E	8309.38 s	361.27 s	277000	12043
RA-S	2716.65 s	133.92 s	92000	4000

Tabella 5.7: Tempi e numero di PDE risolte

un elevato numero di vincoli aggiuntivi richiede maggiore onere computazionale nella risoluzione del sistema lineare, oltre ad un aumento del numero di iterazioni necessarie per la convergenza. Il diverso numero di equazioni differenziali alle derivate parziali richiesto può essere spiegato da un maggior numero di iterazioni di backtracking effettuate per il caso epigrafico.

Conclusioni

In questo lavoro sono stati studiati problemi di controllo ottimo vincolati da equazioni differenziali alle derivate parziali, caratterizzati dalla presenza di incertezza su alcuni parametri modellistici, con l'obiettivo di fornire una formulazione matematica rigorosa del problema e di sviluppare strategie numeriche idonee alla sua risoluzione, al fine di validare sperimentalmente le osservazioni teoriche.
In prima analisi, è stato considerato, come modello introduttivo, il problema deterministico. Dopo aver definito una opportuna discretizzazione agli elementi finiti sul dominio per risolvere la PDE associata al problema, sono state derivate le condizioni di ottimalità e confrontati diversi metodi numerici per una risoluzione efficiente del problema. Successivamente, l'analisi è stata estesa al caso stocastico, introducendo opportune variabili aleatorie per descrivere l'incertezza sulla PDE. Oltre alla discretizzazione spaziale presentata nel caso deterministico, è stata definita una discretizzazione in probabilità, per stimare il valore atteso contenuto nell'espressione del funzionale obiettivo, in questo caso di tipo risk-neutral. Sono state quindi proposte strategie numeriche risolutive basate su metodi del gradiente (Steepest Descent) e metodi Newton-based. Vengono quindi presentate analisi di convergenza rispetto ad alcuni parametri caratteristici del modello, quali la dimensione della mesh e il numero di campioni considerati. Inoltre, è stata proposta un'estensione del problema classico con controllo distribuito al caso particolare in cui la variabile di controllo viene definita solamente su uno o più bordi del dominio in modo da influenzare la PDE come condizione al contorno di Neumann. Anche per questo caso sono stati osservati sperimentalmente gli ordini di convergenza dimostrati, rispetto alle discretizzazioni spaziale e in probabilità. L'approssimazione del valore atteso viene ottenuta tramite un approccio di tipo Monte Carlo. Alcune possibili alternative, quali i metodi Stochastic Collocation e Stochastic Galerkin, sono citate nella relativa sezione.
Nella seconda parte dell'elaborato, viene introdotta una misura di rischio nota nell'ambito dell'ottimizzazione in condizioni di incertezza, il Conditional Value-at-Risk (CVaR). Tale misura consente di controllare le code della distribuzione di probabilità, fornendo una formulazione più robusta nei confronti degli eventi rari, rispetto al caso risk-neutral in cui si ottimizza unicamente il valore atteso. In particolare, viene considerato un vincolo aggiuntivo al problema stocastico con controllo al bordo, basato su questa misura di rischio, imponendo una soglia massima per il CVaR di un funzionale di interesse. Il problema di ottimizzazione vincolata viene quindi risolto utilizzando una strategia interior-point, per trattare il vincolo di disuguaglianza, unitamente ad un metodo di ottimizzazione iterativo, quale il metodo di Newton. L'obiettivo di questa analisi consiste nel confronto di due

tecniche di risoluzione numerica alternative: una riformulazione di tipo epigrafico del problema, che comporta l'introduzione di un set di variabili aggiuntive, pur garantendo una maggiore accuratezza nella soluzione, ed una approssimazione del CVaR tramite una funzione di regolarizzazione, che garantisce maggiore efficienza computazionale ma è limitata relativamente alla precisione nella soluzione, in quanto il funzionale considerato non è quello originale, ma una sua versione approssimata.
Infine, viene considerato un ultimo problema, in cui oltre al vincolo sul CVaR, viene introdotta la medesima misura di rischio nel funzionale obiettivo, passando dalla formulazione risk-neutral ad una di tipo risk-averse. Gli algoritmi descritti vengono quindi applicati a questo nuovo problema, evidenziandone differenze e proprietà computazionali. Questo lavoro apre varie possibili direzioni di ulteriore sviluppo, relative a numerosi aspetti modellistici e numerici. Una possibile estensione riguarda l'analisi di tipologie differenti di equazioni differenziali alle derivate parziali, ad esempio paraboliche o iperboliche, in cui la presenza di dinamiche più complesse può amplificare gli effetti dell'incertezza e rendere ancora più rilevante il controllo del rischio. Come già osservato, un altro sviluppo è caratterizzato dall'applicazione di differenti strategie di discretizzazione in probabilità, basate ad esempio su griglie sparse anisotrope (Smolyak). Risulta inoltre interessante analizzare tecniche di campionamento adattive per una maggiore accuratezza nella descrizione del CVaR, concentrando la presenza dei campioni nella coda della distribuzione.
Infine, una ulteriore possibile estensione è rappresentata dalla scelta di misure di rischio alternative come l'*entropic risk measure* o la *mean-upper semi-deviation*, al fine di confrontarle con il CVaR in termini di robustezza, interpretabilità e complessità numerica.

Appendice A

Function MATLAB

In questa appendice sono descritte nel dettaglio le *function* MATLAB utilizzate per le simulazioni numeriche.

```
% Questo codice risolve un problema di controllo ottimo stocastico
% vincolato da CVaR. Devono essere forniti in input la geometria del
% problema contenente i dati sulla partizione ad elementi finiti del
% dominio, una struttura FEM_mat che raccoglie i dati del problema ottenuti
% dal campionamento e dall'assemblaggio di matrici e termini noti, oltre
% alla soluzione del problema non vincolato u_ref_CVAR.
% Il metodo utilizzato per la risoluzione e' basato su una riformulazione in
% senso epigrafico del vincolo sul CVaR, introducendo variabili le ausiliarie
% t per il quantile e z per la distanza tra esso e la singola
% realizzazione del funzionale di interesse.
% Vengono quindi costruite delle strutture contenenti i risultati ottenuti
% per le variabili e per alcune misure della convergenza del metodo quali
% il decremento di Newton o la norma del gradiente.

clear all
clc
% Percorsi con funzioni utili
fullFileName = matlab.desktop.editor.getActiveFilename;
disp(['Live script path: ', fullFileName]);
scriptFolder = fileparts(fullFileName);
disp(['Cartella script: ', scriptFolder]);
tmpFolder = fileparts(scriptFolder);
disp(['Cartella tmp: ', tmpFolder]);
```

```
TRIpath = fullfile(tmpFolder, 'Triangolatore/Long/bbtr30_long/
    bbtr30/');
UTpath = fullfile(tmpFolder, 'Utility/');
MGpath = fullfile(tmpFolder, 'Matrici_geometria/');
RESpath = fullfile(tmpFolder, 'Alcuni Risultati/');
addpath(TRIpath);
addpath(UTpath);
addpath(MGpath);
addpath(RESpath);
global geometry params gamma;
% Geometria
dim=9;
s=load('geom_boundary_seg.mat',sprintf('geom_%d', dim));
geom=s.(sprintf('geom_%d', dim));
clear s;
geometry.Ndof=max(geom.pivot.pivot);
geometry.Ne=2+length(geom.pivot.Ne(:,1));
geometry.coords=geom.elements.coordinates;

% Funzioni del problema
epsilon=@(x,y,xi) 1+exp(exp(-1.125).*(xi(1).*cos(pi*x).*sin(pi*y)+
    xi(2).*cos(2*pi*x).*sin(pi*y)+xi(3).*cos(pi*x).*sin(2*pi*y)+xi
    (4).*cos(2*pi*x) .*sin(2*pi*y)));
z_d=@(x,y) (exp(x+y).*sin(2*pi*x).*sin(2*pi*y)).*(x>=.5);
f=@(x,y,xi) 1;

% Download dei campioni e della soluzione unconstrained
load('u_ref1000_boundary.mat')
load('campioni_1000_boundary.mat')

% Inizializzazioni
test_fun.z_d=z_d;
test_fun.epsilon=epsilon;
test_fun.f=f;
gamma=1e-4;
params.Nsamples=1000;
params.c1=1e-4;
params.lambda=.95;
params.btmax=50;
params.iters=1;
params.Kmax=1000;
params.tol=1e-10;
params.outertol=1e-9;
params.rho=.5;
params.mu=1;

y_init=zeros(geometry.Ndof,params.Nsamples);
y_refnew=zeros(geometry.Ndof,params.Nsamples);
y=zeros(geometry.Ndof,params.Nsamples);
```

```
70  p=zeros(geometry.Ndof,params.Nsamples);
71  valsref=zeros(params.Nsamples,1);
72  valsinit=zeros(params.Nsamples,1);
73
74  u0=u_ref_CVAR/5;
75
76  parfor jjj=1:params.Nsamples
77  y_init(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*u0)
        ;
78  valsinit(jjj)=.5*(y_init(:,jjj))'*FEM_mat.M*(y_init(:,jjj));
79  y_refnew(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*
        u_ref_CVAR);
80  valsref(jjj)=.5*(y_refnew(:,jjj))'*FEM_mat.M*(y_refnew(:,jjj));
81  end
82
83  quantile_ref=quantile(valsref,params.lambda);
84  CVAR_ref=quantile_ref+(1/(1-params.lambda))*(mean(max(valsref-
        quantile_ref,0)));
85
86  quantile_init=quantile(valsinit,params.lambda);
87  CVAR_init=quantile_init+(1/(1-params.lambda))*(mean(max(valsinit-
        quantile_init,0)));
88
89  tau=CVAR_ref/2;
90
91  t0=quantile_init;
92  g1=2;
93  z0=.005*ones(params.Nsamples,1);
94
95  zk=z0;
96  m=params.Nsamples+geometry.Ne+1;
97  mu=params.mu;
98
99  uk=u0;
100 gradJ_u0=1;
101 grad_barr=1;
102 e=@(k,n) [zeros(k-1,1);1;zeros(n-k,1)];
103 NewHess=sparse([FEM_mat.Hess,zeros(geometry.Ne,1+params.Nsamples);
        zeros(1+params.Nsamples,m)]);
104 dknorms=cell(15,1);
105 NewtonDecr=cell(15,1);
106 gradJnorms=cell(15,1);
107 fullgradnorms=cell(15,1);
108 z_values=cell(15,1);
109 u_values=cell(15,1);
110 t_values=cell(15,1);
111 CVAR_vals=cell(15,1);
112 J=cell(15,1);
113 kk=1;
```

```
g2=zeros(params.Nsamples,1);
gradg2=zeros(m,params.Nsamples);
gg2=zeros(params.Nsamples,1);
g3=zeros(params.Nsamples,1);
gradg3=zeros(m,params.Nsamples);
tic
pde_counter=0;
while mu>params.outertol && abs(g1)>params.tol
NewtDecr=1e6;
k=3;
dk=1e5;

dknorms_inner=[];
NewtonDecr_inner=[];
gradJnorms_inner=[];
fullgradnorms_inner=[];
z_values_inner=[];
u_values_inner=[];
t_values_inner=[];
CVAR_vals_inner=[];
J_inner=[];

while k<params.Kmax && (.5*NewtDecr)>.1*mu

grad_barr2=zeros(m,1);
grad_barr3=zeros(m,1);
Hess_barr2=sparse(m,m);
Hess_barr3=sparse(m,m);

% Soluzione equazioni di stato ed aggiunta e assemblaggio sistema
parfor jjj=1:params.Nsamples
y(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*u0);
p(:,jjj)=(FEM_mat.Kcell{jjj})'\(-FEM_mat.M*(y(:,jjj)-FEM_mat.zd));
g2(jjj)=.5*(y(:,jjj))'*FEM_mat.M*(y(:,jjj))-t0-z0(jjj);
g3(jjj)=-z0(jjj);
end

for jjj=1:params.Nsamples
Kmat_Util=FEM_mat.Kcell{jjj}\FEM_mat.B;

gradg2(:,jjj)=[((y(:,jjj)'*FEM_mat.M*(Kmat_Util)))';-1;-e(jjj,
    params.Nsamples)];
gradg3(:,jjj)=[zeros(geometry.Ne+1,1);-e(jjj,params.Nsamples)];

grad_barr2=grad_barr2+(gradg2(:,jjj)*(1/g2(jjj)));
grad_barr3=grad_barr3+(gradg3(:,jjj)*(1/g3(jjj)));

Hess_barr2=sparse(Hess_barr2+(gradg2(:,jjj)*gradg2(:,jjj)'*(1/(g2(
    jjj)^2)))+...
```

```
161 -[(((Kmat_Util)'*FEM_mat.M*(Kmat_Util))/g2(jjj)),zeros(geometry.Ne
        ,1+params.Nsamples);zeros(1+params.Nsamples,m)]);
162 Hess_barr3=sparse(Hess_barr3+(gradg3(:,jjj)*gradg3(:,jjj)'*(1/(g3(
        jjj)^2))));
163 end
164 pde_counter=pde_counter+3000;
165
166 gradJ_u0=gamma*FEM_mat.B_mass*u0-FEM_mat.B'*mean(p,2);
167 g1=t0+(1/(1-params.lambda))*(1/params.Nsamples)*(sum(z0))-tau;
168 gradg1=[zeros(geometry.Ne,1);1;repmat((1/(1-params.lambda))*(1/
        params.Nsamples),params.Nsamples,1)];
169
170 J_val=.5*(mean(y,2)-FEM_mat.zd)'*FEM_mat.M*(mean(y,2)-FEM_mat.zd)+
        gamma*.5*u0'*FEM_mat.B_mass*u0+...
171 -1/(2*params.Nsamples+1)*(mu*(log(-g1))+mu*sum(log(-g2))+mu*sum(
        log(-g3)));
172 grad_barr=sparse((1/(2*params.Nsamples+1))*(mu*(gradg1/g1)+mu*
        grad_barr2+mu*grad_barr3));
173 Hess_barr=sparse((1/(2*params.Nsamples+1))*(mu*(gradg1*gradg1')
        *(1/(g1^2))+mu*Hess_barr2+mu*Hess_barr3));
174
175 grad_tot=[gradJ_u0;zeros(params.Nsamples+1,1)]-grad_barr;
176 Hess_tot=sparse(NewHess+Hess_barr);
177
178 tauk=.1;
179 dk=-Hess_tot\grad_tot;
180 NewtDecr=(grad_tot'*(-dk));
181
182 % Backtracking
183
184 bb=0;
185 J_new=0+1i;
186 while J_new~=real(J_new) && bb<params.btmax
187 uk=u0+tauk*dk(1:geometry.Ne);
188 tk=t0+tauk*dk(geometry.Ne+1);
189 zk=z0+tauk*dk(geometry.Ne+2:end);
190
191 parfor jjj=1:params.Nsamples
192 y(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*uk);
193 g2(jjj)=.5*(y(:,jjj))'*FEM_mat.M*(y(:,jjj))-tk-zk(jjj);
194 end
195
196 pde_counter=pde_counter+1000;
197 g1=tk+(1/(1-params.lambda))*(1/params.Nsamples)*(sum(zk))-tau;
198 g3=-zk;
199 J_new=.5*(mean(y,2)-FEM_mat.zd)'*FEM_mat.M*(mean(y,2)-FEM_mat.zd)+
        gamma*.5*uk'*FEM_mat.B_mass*uk+...
200 -1/(2*params.Nsamples+1)*(mu*(log(-g1))+mu*sum(log(-g2))+mu*sum(
        log(-g3)));
```

```
201
202 if J_new==real(J_new) && J_new<J_val-params.c1*tauk*NewtDecr
203 break;
204 end
205
206 tauk=params.rho*tauk;
207 bb=bb+1;
208 end
209
210 %Aggiornamento interno
211 J_val=J_new;
212 u0=uk;
213 t0=tk;
214 z0=zk;
215 k=k+1;
216 dknorms_inner=[dknorms_inner,norm(dk)];
217 NewtonDecr_inner=[NewtonDecr_inner,NewtDecr];
218 gradJnorms_inner=[gradJnorms_inner,norm(gradJ_u0)];
219 fullgradnorms_inner=[fullgradnorms_inner,norm(grad_barr)];
220 z_values_inner=[z_values_inner,zk];
221 u_values_inner=[u_values_inner,uk];
222 t_values_inner=[t_values_inner,tk];
223 CVAR_vals_inner=[CVAR_vals_inner,g1+tau];
224 J_inner=[J_inner;J_new];
225 fprintf('NewtDecr= %.5f\n tk=%.5f\n J=%.5f\n bb=%.5f\n', NewtDecr,
      tk,J_val,bb);
226 disp('-----------------------------------------------------------------
    ')
227 end
228 % Aggiornamento esterno
229 dknorms{kk}=dknorms_inner;
230 NewtonDecr{kk}=NewtonDecr_inner;
231 gradJnorms{kk}=gradJnorms_inner;
232 fullgradnorms{kk}=fullgradnorms_inner;
233 z_values{kk}=z_values_inner;
234 u_values{kk}=u_values_inner;
235 t_values{kk}=t_values_inner;
236 CVAR_vals{kk}=CVAR_vals_inner;
237 J{kk}=J_inner;
238 disp('-----------------------------------------------------------------
    ')
239 disp('-----------------------------------------------------------------
    ')
240 disp('-----------------------------------------------------------------
    ')
241 mu=.1*mu;
242 kk=kk+1;
243 end
244 time=toc;
```

```
245
246 % Plot della soluzione e calcolo degli errori
247 uplot=uk;
248 yplot=assembla_soluzione(mean(y,2),geom);
249 pplot=assembla_soluzione(mean(p,2),geom);
250 zplot=test_fun.z_d(geometry.coords(:,1),geometry.coords(:,2));
251 [~,~,~,Vert_lato2]=assembla_B_2lati(geom);
252 Plotting_loc(yplot,pplot,uplot,zplot,geom,'al bordo');
253 figure
254 plot(Vert_lato2,uk(end+1-length(Vert_lato2):end))
255 hold on;
256 plot(Vert_lato2,u_ref_CVAR(end+1-length(Vert_lato2):end))
257 valsvals=zeros(params.Nsamples,1);
258 y_vals=zeros(size(y));
259 parfor jjj=1:params.Nsamples
260 y_vals(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*uk)
    ;
261 valsvals(jjj)=.5*(y_vals(:,jjj))'*FEM_mat.M*(y_vals(:,jjj));
262 end
263 exact_quantile=quantile(valsvals,params.lambda);
264 exact_CVAR=mean(valsvals(valsvals>exact_quantile));
265 confronta_errori=[zk,valsvals,max(valsvals-tk,0),max(valsvals-
    quantile(valsvals,params.lambda),0)];
266 err_quantile=abs(quantile(valsvals,params.lambda)-tk);
267 CVAR_error=abs(exact_CVAR-(tk+(1/(1-params.lambda)*1/params.
    Nsamples)*sum(zk)));
```

```
1         % Questo codice risolve un problema di controllo ottimo
              stocastico
2         % vincolato da CVaR. Devono essere forniti in input la
              geometria del
3         % problema contenente i dati sulla partizione ad elementi
              finiti del
4         % dominio, una struttura FEM_mat che raccoglie i dati del
              problema ottenuti
5         % dal campionamento e dall'assemblaggio di matrici e
              termini noti, oltre
6         % alla soluzione del problema non vincolato u_ref_CVAR.
7         % Il metodo utilizzato per la risoluzione e' basato su un
              approccio di tipo
8         % smoothing-splitting primal-dual interior point. Tale
              strategia consiste
9         % nell'approssimare il CVaR con una funzione
              regolarizzante, quindi trasformare
10        % il vincolo in uguaglianza tramite una variabile
              ausiliaria s e il
11        % relativo moltiplicatore di Lagrange zeta.
12        % La variabile t viene eliminata (smoothing) introducendo
              una funzione
```

```
13          % implicita h(u) e viene determinata utilizzando il metodo
                di bisezione.
14          % Si ottiene quindi un sistema lineare nelle variabili (u,
                s,zeta)
15          % calcolando il complemento di Schur della matrice
                Hessiana relativamente
16          % al contributo di t=h(u).
17          % Vengono quindi costruite delle strutture contenenti i
                risultati ottenuti
18          % per le variabili e per alcune misure della convergenza
                del metodo quali
19          % il decremento di Newton o la norma del gradiente.
20
21          clear all
22          clc
23          % Percorsi con funzioni utili
24          fullFileName = matlab.desktop.editor.getActiveFilename;
25          disp(['Live script path: ', fullFileName]);
26          scriptFolder = fileparts(fullFileName);
27          disp(['Cartella script: ', scriptFolder]);
28          tmpFolder = fileparts(scriptFolder);
29          disp(['Cartella tmp: ', tmpFolder]);
30          TRIpath = fullfile(tmpFolder, 'Triangolatore/Long/
                bbtr30_long/bbtr30/');
31          UTpath = fullfile(tmpFolder, 'Utility/');
32          MGpath = fullfile(tmpFolder, 'Matrici_geometria/');
33          RESpath = fullfile(tmpFolder, 'Alcuni Risultati/');
34          addpath(TRIpath);
35          addpath(UTpath);
36          addpath(MGpath);
37          addpath(RESpath);
38          global geometry geom params;
39          % Geometria
40          dim=9;
41          s=load('geom_boundary_seg.mat',sprintf('geom_%d', dim));
42          geom=s.(sprintf('geom_%d', dim));
43          clear s;
44          geometry.Ndof=max(geom.pivot.pivot);
45          geometry.coords=geom.elements.coordinates;
46
47          % Funzioni del problema
48          epsilon=@(x,y,xi) 1+exp(exp(-1.125).*(xi(1).*cos(pi*x).*
                sin(pi*y)+xi(2).*cos(2*pi*x).*sin(pi*y)+xi(3).*cos(pi*
                x).*sin(2*pi*y)+xi(4).*cos(2*pi*x) .*sin(2*pi*y)));
49          z_d=@(x,y) (exp(x+y).*sin(2*pi*x).*sin(2*pi*y)).*(x>=.5);
50          f=@(x,y,xi) 1;
51
52          % Download dei campioni e della soluzione unconstrained
53          load('u_ref1000_boundary.mat')
```

```
54          load('campioni_1000_boundary.mat')
55
56          % Inizializzazioni
57          test_fun.z_d=z_d;
58          test_fun.epsilon=epsilon;
59          test_fun.f=f;
60          gamma=1e-4;
61          params.Nsamples=1000;
62          params.c1=1e-4;
63          params.lambda=.95;
64          params.btmax=50;
65          params.iters=1;
66          params.Kmax=1000;
67          params.tol=1e-10;
68          params.outertol=1e-9;
69          params.rho=.5;
70          params.mu=1;
71          params.epsilon=.001;
72          coef=1/(1-params.lambda);
73          sig=1;
74
75          y_init=zeros(geometry.Ndof,params.Nsamples);
76          y_refnew=zeros(geometry.Ndof,params.Nsamples);
77          y=zeros(geometry.Ndof,params.Nsamples);
78          p=zeros(geometry.Ndof,params.Nsamples);
79          valsref=zeros(params.Nsamples,1);
80          valsinit=zeros(params.Nsamples,1);
81          u0=u_ref_CVAR/5;
82
83          parfor jjj=1:params.Nsamples
84          y_init(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+
               FEM_mat.B*u0);
85          valsinit(jjj)=.5*(y_init(:,jjj))'*FEM_mat.M*(y_init(:,jjj)
               );
86          y_refnew(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+
               FEM_mat.B*u_ref_CVAR);
87          valsref(jjj)=.5*(y_refnew(:,jjj))'*FEM_mat.M*(y_refnew(:,
               jjj));
88          end
89
90          quantile_ref=quantile(valsref,params.lambda);
91          CVAR_ref=quantile_ref+(1/(1-params.lambda))*(mean(max(
               valsref-quantile_ref,0)));
92
93          quantile_init=quantile(valsinit,params.lambda);
94          CVAR_init=quantile_init+(1/(1-params.lambda))*(mean(max(
               valsinit-quantile_init,0)));
95
96          tau=CVAR_ref/2;
```

```
97
98          m=params.Nsamples+geometry.Ndof+1;
99          mu=params.mu;
100
101         t0=quantile_init;
102         s0=tau-(t0+coef*mean(g_eps(valsinit-t0,params.epsilon)));
103         zeta0=(sig*mu)/s0;
104         vals=valsinit;
105         e=@(k,n) [zeros(k-1,1);1;zeros(n-k,1)];
106
107         dknorms=cell(12,1);
108         NewtonDecr=cell(12,1);
109         gradJnorms=cell(12,1);
110         fullgradnorms=cell(12,1);
111         zeta_values=cell(12,1);
112         u_values=cell(12,1);
113         t_values=cell(12,1);
114         s_values=cell(12,1);
115         CVAR_vals=cell(12,1);
116         J=cell(12,1);
117
118         dk=1;
119         k=1;
120         g1=2;
121         pde_counter=0;
122         tic
123         while mu>params.outertol
124         kk=1;
125         dknorms_inner=[];
126         NewtonDecr_inner=[];
127         gradJnorms_inner=[];
128         fullgradnorms_inner=[];
129         zeta_values_inner=[];
130         u_values_inner=[];
131         t_values_inner=[];
132         s_values_inner=[];
133         CVAR_vals_inner=[];
134         J_inner=[];
135         NewtDcr_u=10;
136         while kk<params.Kmax && .5*NewtDcr_u>.1*mu
137         g_u=zeros(length(u0),1);
138         dFdu=zeros(length(u0),1);
139         Hess_uu=sparse(zeros(length(u0)));
140         % Calcolo di t tramite bisezione (alla prima iterazione
              viene
141         % fissata al suo valore iniziale)
142         if kk==1 && mu==1
143         tk=t0;
144         else
```

```
145         tk=solve_F_tk(vals,t0);
146         end
147
148         % Soluzione equazioni di stato ed aggiunta e assemblaggio
                sistema
149         parfor jjj=1:params.Nsamples
150         y(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*
                u0);
151         p(:,jjj)=(FEM_mat.Kcell{jjj})'\(-FEM_mat.M*(y(:,jjj)-
                FEM_mat.zd));
152
153         vals(jjj)=.5*y(:,jjj)'*FEM_mat.M*y(:,jjj);
154         end
155
156         for jjj=1:params.Nsamples
157         Util_Mat=FEM_mat.Kcell{jjj}\(FEM_mat.B);
158         dFdu=dFdu+g_eps_der_der(vals(jjj)-tk,params.epsilon)*
                Util_Mat'*FEM_mat.M*y(:,jjj);
159         g_u=g_u+g_eps_der(vals(jjj)-tk,params.epsilon)*Util_Mat'*
                FEM_mat.M*y(:,jjj);
160         Hess_uu=sparse(Hess_uu+g_eps_der_der(vals(jjj)-tk,params.
                epsilon)*(Util_Mat'*FEM_mat.M*y(:,jjj))*(Util_Mat'*
                FEM_mat.M*y(:,jjj))'+...
161         g_eps_der(vals(jjj)-tk,params.epsilon)*(Util_Mat'*FEM_mat.
                M*Util_Mat));
162         end
163
164         pde_counter=pde_counter+3000;
165         dFdt=coef*mean(g_eps_der_der(vals-tk,params.epsilon));
166         dFdu=coef*dFdu/params.Nsamples;
167         g1=tk+coef*(mean(g_eps(vals-tk,params.epsilon)))-tau+s0;
168         J_val=.5*(mean(y,2)-FEM_mat.zd)'*FEM_mat.M*(mean(y,2)-
                FEM_mat.zd)+.5*gamma*u0'*FEM_mat.B_mass*u0-mu*log(s0);
169         gradJ_u0=gamma*FEM_mat.B_mass*u0-FEM_mat.B'*mean(p,2);
170         g_u=coef*(g_u/params.Nsamples);
171         Hess_uu=coef*(Hess_uu/params.Nsamples);
172
173         grad_tot=[gradJ_u0+zeta0*g_u; -(sig*mu)/s0+zeta0; g1];
174         HH=Hess_uu-(dFdu*dFdu')/dFdt;
175         Hess_tot=sparse([FEM_mat.Hess+zeta0*HH, zeros(length(u0)
                ,1), g_u ;
176         zeros(length(u0),1)', (sig*mu)/(s0^2), 1;
177         g_u', 1,0]);
178         dk=-Hess_tot\grad_tot;
179         NewtDcr_u=dk(1:end-2)'*(FEM_mat.Hess+zeta0*HH)*dk(1:end-2)
                ;
180         J_new=0+1i;
181
182         % Backtracking
```

```
183         etak=.95;
184         tauk=max(etak*min([.7,-s0/dk(end-1)]),.3);
185         bb=0;
186         while  J_new~=real(J_new) && bb<params.btmax
187
188         uk=u0+tauk*dk(1:end-2);
189         sk=s0+tauk*dk(end-1);
190         zetak=zeta0+tauk*dk(end);
191
192         parfor jjj=1:params.Nsamples
193         y(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+FEM_mat.B*
                uk);
194         end
195         pde_counter=pde_counter+1000;
196
197         J_new=.5*(mean(y,2)-FEM_mat.zd)'*FEM_mat.M*(mean(y,2)-
                FEM_mat.zd)+.5*gamma*uk'*FEM_mat.B_mass*uk-mu*log(sk);
198
199         if J_new==real(J_new)
200         break;
201         end
202
203         tauk=params.rho*tauk;
204         bb=bb+1;
205         end
206         % Aggiornamento interno
207         newvals=compute_vals(uk,FEM_mat);
208         kk=kk+1;
209         u0=uk;
210         t0=tk;
211         s0=sk;
212         zeta0=zetak;
213         vals=newvals;
214         fprintf('NewtDcr_u= %.5f\n tk=%.5f\n sk=%.5f\n zetak=%.5f\
                n J=%.5f\n bb=%.5f\n', NewtDcr_u, tk,sk,zetak,J_val,
                bb);
215         disp('
                ------------------------------------------------------------
                ')
216
217         NewtonDecr_inner=[NewtonDecr_inner,NewtDcr_u];
218         gradJnorms_inner=[gradJnorms_inner,norm(gradJ_u0)];
219         dkorms_inner=[dknorms_inner,norm(dk)];
220         fullgradnorms_inner=[fullgradnorms_inner,norm(grad_tot)];
221         zeta_values_inner=[zeta_values_inner,zetak];
222         s_values_inner=[s_values_inner,sk];
223         u_values_inner=[u_values_inner,uk];
224         t_values_inner=[t_values_inner,tk];
225         CVAR_vals_inner=[CVAR_vals_inner,g1+tau];
```

```
226         J_inner=[J_inner;J_new];
227         end
228         % Aggiornamento esterno
229         dknorms{k}=dknorms_inner;
230         NewtonDecr{k}=NewtonDecr_inner;
231         gradJnorms{k}=gradJnorms_inner;
232         fullgradnorms{k}=fullgradnorms_inner;
233         zeta_values{k}=zeta_values_inner;
234         u_values{k}=u_values_inner;
235         s_values{k}=s_values_inner;
236         t_values{k}=t_values_inner;
237         CVAR_vals{k}=CVAR_vals_inner;
238         J{k}=J_inner;
239         disp('
              ------------------------------------------------------------
              ')
240         disp('
              ------------------------------------------------------------
              ')
241         disp('
              ------------------------------------------------------------
              ')
242         mu=mu*.1;
243         k=k+1;
244         end
245         time=toc;
246
247         % Plot della soluzione e calcolo degli errori
248         uu=u_values{k-1};
249         yplot=assembla_soluzione(mean(y,2),geom);
250         pplot=assembla_soluzione(mean(p,2),geom);
251         zplot=test_fun.z_d(geometry.coords(:,1),geometry.coords
              (:,2));
252         Plotting_loc(yplot,pplot,uu(:,size(uu,2)),zplot,geom,'al
              bordo');
253         [~,~,~,Vert_lato2]=assembla_B_2lati(geom);
254         figure
255         plot(Vert_lato2,uu(end-length(Vert_lato2)+1:end,size(uu,2)
              ))
256         hold on;
257         plot(Vert_lato2,u_ref_CVAR(end-length(Vert_lato2)+1:end))
258         valsvals=zeros(params.Nsamples,1);
259         y_vals=zeros(size(y));
260         parfor jjj=1:params.Nsamples
261         y_vals(:,jjj)=FEM_mat.Kcell{jjj}\(FEM_mat.rhs(:,jjj)+
              FEM_mat.B*uk);
262         valsvals(jjj)=.5*(y_vals(:,jjj))'*FEM_mat.M*(y_vals(:,jjj)
              );
263         end
```

```
264         exact_quantile=quantile(vals,params.lambda);
265         exact_CVAR=mean(vals(vals>exact_quantile));
266         err_quantile=abs(exact_quantile-tk);
267         CVAR_error=abs(exact_CVAR-(tk+coef*(mean(g_eps(vals-tk,
               params.epsilon)))));
```

Bibliografia

[1] Babuška I., Nobile F., Tempone R. (2007). A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45:1005–1034.

[2] Barbera A. M. A., Berrone S. (2008). BBTR: An unstructured triangular mesh generator. *Quaderni del Dipartimento di Matematica*.

[3] Boyd S., Vandenberghe L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

[4] Chen C. H., Mangasarian O. L. (1995). Smoothing methods for convex inequalities and linear complementarity problems. *Mathematical Programming*, 71:51–69.

[5] Ciaramella G., Nobile F., Vanzan T. (2024). A multigrid solver for PDE-onstrained optimization with uncertain inputs. *Journal of Scientific Computing*, 101(3):Article 13.

[6] Fiacco A. V., McCormick G. P. (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley & Sons, New York, N.Y.

[7] Frisch K. R. (1955). The logarithmic potential method of convex programming. Technical report, University Institute of Economics, Oslo, Norway.

[8] Gunzburger M. D., Lee Hyung-Chun, Lee Jangwoon (2011). Error estimates of stochastic optimal neumann boundary control problems. *SIAM Journal on Numerical Analysis*, 49(4):1532–1552.

[9] Hinze M., Pinnau R., Ulbrich M., Ulbrich S. (2009). *Optimization with PDE Constraints*. Springer, Berlin, first edition.

[10] Jacod J., Protter P. (2004). *Probability Essentials*. Universitext. Springer-Verlag, Berlin.

[11] Kouri D. P., Surowiec T. M. (2016). Risk-averse pde-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization*, 26(1):365–396.

[12] Lions J. L. (1971). *Optimal Control of Systems Governed by Partial Differential Equations*. Springer, Berlin.

[13] Lord G. J., Powell C. E., Shardlow T. (2014). *An Introduction to Computational Stochastic PDEs*, volume 50 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press.

[14] Manzoni A., Quarteroni A., Salsa S. (2021). *Optimal Control of Partial Differential Equations*. Springer, Cham, Switzerland, first edition.

[15] Markowski M. (2019). Efficient solution of smoothed risk-averse pde-constrained optimization problems. Master's thesis, Rice University.

[16] Markowski M. (2022). *Newton-Based Methods for Smoothed Risk-Averse PDE-Constrained Optimization Problems*. PhD thesis, Rice University.

[17] Martin M. (2019). *Stochastic Approximation Methods for PDE-Constrained Optimal Control Problems with Uncertain Parameters*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL).

[18] Martin M., Krumscheid S., Nobile F. (2021). Complexity analysis of stochastic gradient methods for pde-constrained optimal control problems with uncertain parameters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(4):1599–1633.

[19] Martínez-Frutos J., Periago Esparza F. (2018). *Optimal Control of PDEs under Uncertainty: An Introduction with Application to Optimal Shape Design of Structures*. SpringerBriefs in Mathematics, Springer, Cham, Switzerland. Bilbao: BCAM – Basque Center for Applied Mathematics.

[20] Mateos M. (2018). Optimization methods for dirichlet control problems. *Optimization*, 67(5):585–617.

[21] Nobile F., Vanzan T. (2024). Multilevel quadrature formulae for the optimal control of random pdes. *ArXiv preprint*.

[22] Nocedal J., Wright S. J. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition.

[23] Pieraccini S., Vanzan T. (2025). An adaptive importance sampling algorithm for risk-averse optimization: A survey. *ArXiv preprint*.

[24] Rockafellar R. T., Uryasev S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41.

[25] Rockafellar R. T., Uryasev S. (2001). Conditional value-at-risk: Optimization approach. In *Stochastic Optimization: Algorithms and Applications*, volume 54 of *Applied Optimization*, pages 411–435. Springer, Boston, MA.

[26] Royset J. O. (2025). Risk-adaptive approaches to stochastic optimization: a survey. *SIAM Review*, 67(1):3–70.

[27] Shapiro A., Dentcheva D., Ruszczyński A. (2014). *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics; Mathematical Programming Society, Philadelphia, PA, second edition.

[28] Shapiro A., Wardi Y. (1994). Nondifferentiability of the steady-state function in discrete event dynamic systems. *IEEE Transactions on Automatic Control*.

[29] Tiesler H., Kirby R. M. , Xiu D., Preusser T. (2012). Stochastic collocation for optimal control problems with stochastic pde constraints. *SIAM Journal on Control and Optimization*, 50(5):2659–2682.

[30] Tröltzsch F. (2010). *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

[31] Wright S. J. (2001). On the convergence of the newton/log-barrier method. *Mathematical Programming*, 90:71–100.