Politecnico di Torino

POLITECNICO DI TORINO

Dipartimento di Scienze Matematiche

Corso di Laurea Magistrale in
INGEGNERIA MATEMATICA

Metodo di Galerkin discontinuo semi-lagrangiano per equazioni di Hamilton-Jacobi-Bellman

Relatore:
Prof. Adriano Festa

Candidato:
Carlo De Simone

Anno accademico 2023/2024

Dal figlio, alla Madre
Da Carlo, agli amici e a Torino

Abstract

L'equazione di Hamilton-Jacobi-Bellman è un'equazione differenziale alle derivate parziali non lineare del primo ordine che emerge nel contesto dei problemi di controllo ottimo. Si può considerare come un'espressione differenziale del Principio della Programmazione Dinamica soddisfatto dalla *value function*, ossia una mappa che intercorre tra qualsiasi condizione iniziale e il valore corrispondente al controllo ottimo, non necessariamente unico.
L'equazione di Hamilton-Jacobi-Bellman, a causa della sua non linearità, in generale non ammette una soluzione classica, anche per dati regolari. Tuttavia, negli anni '80 è stata sviluppata una teoria delle soluzioni deboli, note come soluzioni di viscosità, che fornisce un quadro appropriato alla gestione della mancanza di regolarità della *value function*.
Negli ultimi due decenni, la letteratura scientifica ha proposto diversi approcci per risolvere la classe piuttosto generale di equazioni di Hamilton-Jacobi, tra cui Galerkin discontinuo in virtù della sua natura locale, flessibilità e robustezza.
In questo lavoro di tesi, si propone un metodo numerico per risolvere l'equazione di Hamilton-Jacobi-Bellman evolutiva in una dimensione. Esso consiste nella combinazione di uno schema semi-lagrangiano, orientato a ricostruire le linee caratteristiche, con un metodo di Galerkin discontinuo, atto a generare una soluzione approssimata come combinazione lineare di prescritte funzioni di base discontinue e a supporto compatto.
Per dimostrare le prestazioni del modello proposto, sono raccolti e presentati una serie di esperimenti numerici con dati regolari, semplicemente continui e discontinui.

Indice

1 Introduzione

Un problema di controllo ottimo consiste nel determinare l'insieme di azioni e scelte che permettono al sistema dinamico, contemporaneamente, di soddisfare i vincoli fisici e ottimizzare un certo criterio, la funzione obiettivo.
In virtù del Principio della Programmazione Dinamica, sviluppato da Richard Bellman negli anni '50, si definisce la *value function*, una mappa che lega la condizione iniziale con il valore della funzione obiettivo quando è applicato il controllo ottimo. Secondo tale principio, la *value function* associata ad una certa condizione iniziale si può esprimere come la somma della *value function* dello stato futuro e del valore della funzione obiettivo per raggiungere tale stato futuro in modo ottimale. L'equazione di Hamilton-Jacobi-Bellman (HJB) rappresenta in termini differenziali esattamente questo concetto. I risvolti applicativi sono molteplici, si adopera in campi come l'ingegneria aerospaziale, l'industria energetica, la finanza quantitativa e l'elaborazione delle immagini.
L'equazione di HJB appartiene alla classe più generale di equazioni di Hamilton-Jacobi (HJ), si tratta di equazioni alle derivate parziali non lineari del primo ordine studiate, originariamente da Hamilton e successivamente da Jacobi, nel contesto della meccanica classica e calcolo delle variazioni. In generale, questa tipologia di equazioni, non ammette soluzione in senso classico, anche per dati regolari, pertanto negli anni '80 fu sviluppata la teoria delle soluzione di viscosità dagli autori M. G. Crandall, L. C. Evans e P. L. Lions. Tale classe di soluzioni soddisfa l'equazione in senso debole e sotto ipotesi relativamente generali assicura la buona positura dei problemi associati. In particolare, il quadro teorico che ne emerge risulta completo e in grado di fornire la necessaria caratterizzazione della *value function* come unica soluzione di viscosità per equazioni di Hamilton-Jacobi-Bellman.

In questo contesto, si pone una sfida sulla costruzione di un metodo numerico affidabile che sia capace di catturare l'unica soluzione di viscosità. In letteratura, sono stati proposti diversi schemi per risolvere le equzioni di Hamilton-Jacobi. Crandall e Lions in [1] svilupparono un metodo alle Differenze Finite e dimostrarono che uno

schema monotono di questo tipo converge alla soluzione esatta. Dall'adattamento di strategie adoperate per le leggi di conservazione è stata proposta in [2] una classe di schemi *Essentially Non-Oscillatory* (ENO), successivamente anche formulazioni *Weighted* ENO (WENO) sono state introdotte per risolvere le equazioni di HJ [3, 4] così come metodi *central-upwind* [5]. Relativamente agli Elementi Finiti, Hu e Shu in [6] applicarono Galerkin discontinuo (DG) all'equazione di Hamilton-Jacobi, riformulata in una legge di conservazione. Cheng e Shu in [7] svilupparono quindi un metodo diretto per la risoluzione di equazioni di HJ con DG. Successivamente, ne sono nate altre estensioni e varianti [8, 9, 10, 11]. Negli ultimi anni, sono stati impiegati anche approcci semi-lagrangiani (SL) per l'approssimazione delle soluzioni di HJ [12]. In particolare, la combinazione con un metodo WENO è stata sviluppata in [13], invece in [14] è stata proposto un metodo di Galerkin discontinuo e SL per equazioni alle derivate parziali non stazionarie, lineari, del primo e secondo ordine.

L'obiettivo della tesi è di elaborare un nuovo metodo per risolvere e approssimare la soluzione di viscosità delle equazioni di Hamilton-Jacobi-Bellman. Inoltre, si desidera che tale metodo, derivando dallo schema semi-lagrangiano, sia stabile incondizionatamente rispetto al passo temporale. Allo stesso modo si vuole che preservi due tipiche qualità del metodo di Galerkin discontinuo a cui si ispira, l'adattività degli elementi nel raffinamento e grado polinomiale (*hp - adaptivity*), decomposizione del problema generale in parti più semplici e risolvibili separatamente.
Il presente elaborato è così organizzato:

- nel seguito di questa Introduzione, inizialmente si definisce matematicamente un problema di controllo ottimo ad orizzonte finito e si deriva l'equazione evolutiva di Hamilton-Jacobi-Bellman, successivamente si analizza il quadro teorico (definizioni, proprietà, Teoremi) delle soluzioni di viscosità per equazioni differenziali scalari del primo ordine da un punto di vista matematico;

- nel Capitolo 2 si affronta la teoria dei metodi agli Elementi Finiti, inizialmente in relazione alle equazioni ellittiche per cui risultano naturalmente adatti, successivamente per le equazioni iperboliche per le quali Galerkin discontinuo fu sviluppato in principio;

- nel Capitolo 3 si descrive il metodo di Galerkin discontinuo semi-lagrangiano proposto, procedendo gradualmente verso la formulazione e gli aspetti implementativi connessi all'equazione di Hamilton-Jacobi-Bellman evolutiva in una dimensione;

- nel Capitolo 4 si espongono i test effettuati e se ne valuta l'errore di discretizzazione, il quale consente un confronto con altri metodi noti dalla letteratura;

- infine, nel Capitolo 5 si commentano i risultati dei test, evidenziando i punti di forza e debolezza del metodo proposto, quindi si suggeriscono possibili direzioni di ricerca futura.

1.1 Equazione di Hamilton-Jacobi-Bellman

Un sistema di controllo (CS) in dimensione d nell'intervallo temporale $\mathcal{T} \equiv [t, T]$ è governato dalle equazioni della dinamica

$$\begin{cases} \dot{y}(s) = f(s, y(s), u(s)), & 0 \leq t < s < T, \\ y(t) = x, & x \in \mathbb{R}^d, \end{cases} \tag{CS}$$

$y : \mathcal{T} \to \mathbb{R}^d$ rappresenta la funzione dello stato del sistema, $u : \mathcal{T} \to U$ è la funzione di controllo a valori nello spazio $U \subset \mathbb{R}^m$ compatto, $f : \mathcal{T} \times \mathbb{R}^d \times U \to \mathbb{R}^d$ è la mappa della dinamica del sistema. Per ogni funzione di controllo che appartiene all'insieme $\mathcal{U}_{ad} = \{u : \mathcal{T} \to U, \text{ misurabile}\}$, il Teorema di Carathéodory garantisce l'esistenza e l'unicità di una soluzione continua e differenziabile quasi ovunque per (CS) sotto le ipotesi di f continua, lipschitziana in y (uniformemente rispetto a s e u) e misurabile rispetto a s [12, Teorema 8.1].
Per formulare un problema di controllo ottimo, è necessario introdurre con (CS) un funzionale costo che si desiderare minimizzare[1]. Nella formulazione di Bolza, il funzionale dei problemi ad orizzonte finito, $T \in \mathbb{R}$, ha la forma

$$J_{t,x}(u) = \int_t^T \ell(y_x^u(s), u(s)) e^{-\lambda(s-t)}\, ds + L(y_x^u(T)) e^{-\lambda(T-t)},$$

y_x^u è la traiettoria derivante dal controllo $u \in \mathcal{U}_{ad}$ con condizione iniziale x al tempo t, la mappa $\ell : \mathbb{R}^d \times U \to \mathbb{R}$ è il costo corrente (per unità di tempo) ed è assunta lipschitziana in y (uniformemente rispetto a u) e limitata, $L : \mathbb{R}^d \to \mathbb{R}$ è la funzione costo terminale ed è supposta lipschitziana e limitata. $\lambda \geq 0$ è un parametro che rende paragonabili i costi ad istanti temporali diversi, nei contesti di carattere economico è detto fattore di sconto.
Il problema del controllo ottimo può essere quindi descritto in termini della *value function*

$$v(t, x) = \inf_{u \in \mathcal{U}_{ad}} J_{t,x}(u),$$

il cui significato è il costo minimo associato con la condizione iniziale (t, x). In virtù delle ipotesi fatte sulle equazioni della dinamica e sulle funzioni costo, si deduce che $v : \mathcal{T} \times \mathbb{R}^d \to \mathbb{R}$ è limitata e lipschitziana [15, Cap. 3, Proposizione 3.1].
Inoltre, la *value function* soddisfa il Principio della Programmazione Dinamica (DPP), noto altresì come Principio di Ottimalità di Bellman, cioè $\forall \tau \in (t, T]$

$$v(t, x) = \inf_{u \in \mathcal{U}_{ad}} \left\{ \int_t^\tau \ell(y_x^u(s), u(s)) e^{-\lambda(s-t)}\, ds + v(\tau, y_x^u(\tau)) e^{-\lambda(\tau-t)} \right\}, \tag{DPP}$$

la dimostrazione di questo risultato è riportata in [16, sez. 10.3, Teorema 1]. Essenzialmente il problema di partenza è stato scomposto in due parti (figura 1.1), infatti

[1] È indifferente riferirsi ad un problema di massimo poiché $\max_{x \in S} g(x) = -\min_{x \in S}(-g(x))$.

$v(t,x)$ è il valore associato al problema di controllo ottimo nell'intervallo temporale $[t,\tau]$ con costo corrente ℓ e costo terminale $v(\tau, y_x^{u^*}(\tau))$ e quest'ultimo è ancora un problema di controllo ottimo nell'intervallo temporale $[\tau,T]$ con costo corrente ℓ e costo terminale $L(y_x^{u^*}(T))$.

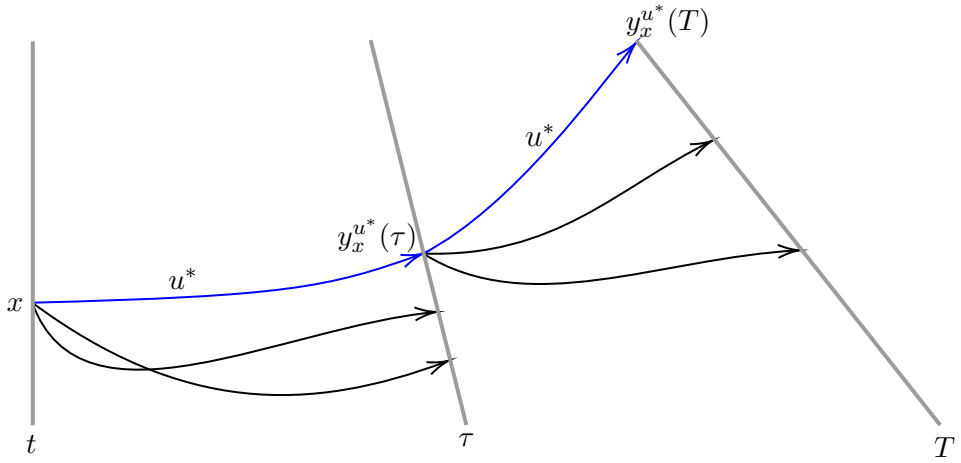

Figura 1.1: Il problema di ottimizzazione viene suddiviso in due parti, rispettivamente, nell'intervallo $[t,\tau]$ e $[\tau,T]$, le scelte ottime sono rappresentate in blu.

Sotto l'ipotesi di *value function* differenziabile con continuità (ipotesi che può essere inseguito indebolita) si può ottenere una versione infinitesimale di (DPP), cioè un'equazione alle derivati parziali che, accoppiata con un'opportuna condizione al bordo, viene risolta da v. Infatti da (CS),

$$\begin{aligned} y_x^u(t+h) &= y_x^u(t) + hf(t,x,u(t)) + o(h), \\ v(t+h, y_x^u(t+h)) &= v(t,x) + h\frac{\partial v}{\partial t}(t,x) + \nabla_x v(t,x)\cdot hf(t,x,u(t)) + o(h), \\ \int_t^{t+h} \ell(y_x^u(s),u(s))e^{-\lambda(s-t)}\,ds &= h\ell(x,u(t)) + o(h), \end{aligned}$$

sostituendo le uguaglianze appena scritte in (DPP) per $\tau = t+h$ e supponendo $\lambda = 0$, si ottiene

$$\begin{aligned} v(t,x) &= \inf_{u\in\mathcal{U}_{ad}} \left\{ h\ell(x,u(t)) + v(t,x) + h\frac{\partial v}{\partial t}(t,x) + \nabla_x v(t,x)\cdot hf(t,x,u(t)) + o(h) \right\}, \\ 0 &= \inf_{u\in\mathcal{U}_{ad}} \left\{ h\ell(x,u(t)) + h\frac{\partial v}{\partial t}(t,x) + \nabla_x v(t,x)\cdot hf(t,x,u(t)) + o(h) \right\}, \\ 0 &= \inf_{u\in\mathcal{U}_{ad}} \left\{ \ell(x,u(t)) + \frac{\partial v}{\partial t}(t,x) + \nabla_x v(t,x)\cdot f(t,x,u(t)) + \frac{o(h)}{h} \right\}. \end{aligned}$$

Passando al limite $h \to 0$, l'infinitesimo di ordine superiore a h tende a 0 e con esso scompaiono anche i termini che coinvolgono $u(s)$ con $s > t$, quindi l'estremo inferiore viene valutato direttamente nello spazio U, in totale si ricava[2]

$$\frac{\partial v}{\partial t}(t,x) = -\inf_{u\in U}\left\{\ell(x,u) + \nabla_x v(t,x)\cdot f(t,x,u)\right\},$$

con $0 \le t < T$, $x \in \mathbb{R}^d$ e condizione al bordo che deriva dalla conoscenza della *value function* in (T,x),

$$v(T,x) = \inf_{u\in\mathcal{U}_{ad}} J_{T,x}(u) = \inf_{u\in\mathcal{U}_{ad}} L(y_x^u(T)) = L(x).$$

In modo equivalente, si può riscrivere per $0 \le t < T$ e $x \in \mathbb{R}^d$

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x) = \sup\limits_{u\in U}\{-\ell(x,u) - \nabla_x v(t,x)\cdot f(t,x,u)\}, \\ v(T,x) = L(x), \end{cases} \tag{HJB}$$

l'equazione alle derivate parziali di (HJB), nel contesto dei problemi di controllo ottimo ad orizzonte finito, si definisce di Hamilton-Jacobi-Bellman perchè rappresenta il principio di Bellman in termini dell'equazione evolutiva di Hamilton-Jacobi, rappresentata in (HJ). Precisamente, per $0 < t \le T$ e $x \in \mathbb{R}^d$, la formulazione

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x) + H(t,x,\nabla_x v(t,x)) = 0, \\ v(0,x) = v_0(x), \end{cases} \tag{HJ}$$

con H operatore Hamiltaniano, equivale a (HJB) dei problemi di controllo ottimo quando $H(t,x,p) = \inf_{u\in U}\{\ell(x,u) + p\cdot f(t,x,u)\}$ e si esegue il cambio di variabile $t \to T - t$.

1.2 Soluzione di Viscosità

In generale, è noto che il problema di Cauchy associato alle equazioni di Hamilton-Jacobi (HJ) non ammette una soluzione classica[3], per esempio quando le curve caratteristiche si incontrano in almeno un punto del dominio. In taluni casi è stato introdotto il concetto di soluzione debole, poiché risolve l'equazione alle derivate parziali quasi ovunque. Tuttavia, questa generalizzazione è risultata inadeguata nel contesto delle equazioni di Hamilton-Jacobi perché non è sempre in grado di assicurare l'unicità della soluzione.

[2]Per λ non nullo, ricordando che $e^{-\lambda h} \approx 1 - \lambda h$ e seguendo gli stessi passaggi, si giunge alla formulazione $\frac{\partial v}{\partial t}(t,x) = \lambda v(t,x) - \inf_{u\in U}\{\ell(x,u) + \nabla_x v(t,x)\cdot f(t,x,u)\}$.

[3]Soluzione globalmente differenziabile con continuità N volte, con N ordine dell'equazione alle derivate parziali che risolve.

Esempio 1.1

Il problema monodimensionale con $x \in \mathbb{R}$ *e* $0 < t \leq T$

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x) + \left|\dfrac{\partial v}{\partial x}(t,x)\right| = 0, \\ v(0,x) = |x|, \end{cases} \tag{1.1}$$

corrisponde a (HJ) *quando* $H(t,x,p) = |p|$ *e* $v_0(x) \equiv 0$. *È facile verificare che sono infinite le soluzioni in senso debole di tale problema, di seguito sono riportate tre funzioni lipschitziane,* $v_a(t,x) = |x| - t$,

$$v_b(t,x) = \begin{cases} |x| - t, & \text{se } |x| > t, \\ t - |x|, & \text{se } |x| \leq t, \end{cases} \qquad v_c(t,x) \begin{cases} |x| - t, & \text{se } |x| > t, \\ 0, & \text{se } |x| \leq t, \end{cases}$$

che soddisfano quasi ovunque (1.1), *eccetto per i punti che giacciono su* $x = 0$ (*casi a e b*) *e su* $x = \pm t$ (*casi b e c*).

Per tale ragione, all'inizio degli anni '80, M. G. Crandall, L. C. Evans e P. L. Lions proposero la nozione di soluzione di viscosità [17] e le relative proprietà [18] per equazioni differenziali alle derivate parziali scalari, classe cui appartengono le equazioni di Hamilton-Jacobi. Tale soluzione, come si vedrà, esiste, è unica ed è stabile, queste condizioni rappresentano la buona positura di un problema.
Al fine di introdurre la teoria alla base di questi risultati, si consideri l'equazione differenziale scalare del primo ordine

$$F(q, v(q), \nabla v(q)) = 0 \quad q \in Q, \tag{DE$_1$}$$

con $Q \subseteq \mathbb{R}^{d+1}$ aperto e $F : Q \times \mathbb{R} \times \mathbb{R}^{d+1} \to \mathbb{R}$ mappa continua sul suo dominio.

Definizione 1.1

Una funzione $v \in C(Q)$ *si definisce subsoluzione di viscosità in* Q *per l'equazione* $F(q, v(q), \nabla v(q)) = 0$ *se,* $\forall \phi \in C^1(Q)$,

$$F(q_M, v(q_M), \nabla \phi(q_M)) \leq 0 \tag{SubS}$$

ad ogni punto $q_M \in Q$ *di massimo locale per* $v - \phi$.

Definizione 1.2

Una funzione $v \in C(Q)$ *si definisce supersoluzione di viscosità in* Q *per l'equazione* $F(q, v(q), \nabla v(q)) = 0$ *se,* $\forall \phi \in C^1(Q)$,

$$F(q_m, v(q_m), \nabla \phi(q_m)) \geq 0 \tag{SuperS}$$

ad ogni punto $q_m \in Q$ *di minimo locale per* $v - \phi$.

Definizione 1.3

Una funzione $v \in C(Q)$ si dice soluzione di viscosità in Q per l'equazione $F(q, v(q), \nabla v(q)) = 0$ se è simultaneamente subsoluzione e supersoluzione.

Le funzioni ϕ appena introdotte nelle definizioni sono dette di test.
In virtù della condizione (SubS) e (SuperS), non è mai restrittivo testare una presunta subsoluzione con delle funzioni ϕ tali che $v(q_M) = \phi(q_M)$ e una presunta supersoluzione con delle funzioni ϕ tali che $v(q_m) = \phi(q_m)$, si veda figura 1.2.

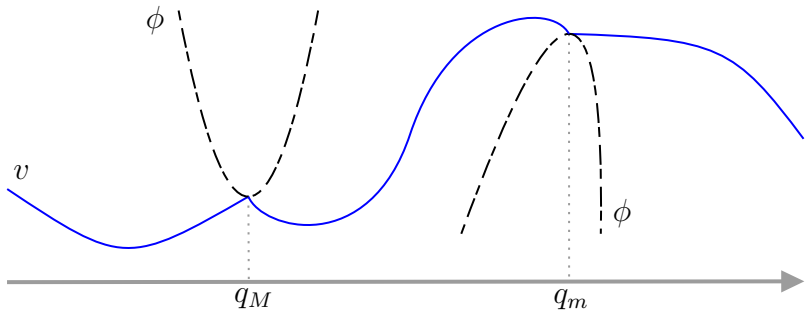

Figura 1.2: Funzione test ϕ per v candidata subsoluzione (in q_M) e supersoluzione (in q_m) di viscosità.

Una proprietà peculiare di questa tipologia di soluzione è che, in generale, non sono preservate da un cambio di segno nell'equazione, in pratica le soluzione di viscosità per $F(q, v(q), \nabla v(q)) = 0$, quasi sempre, non corrispondono a quelle per $-F(q, v(q), \nabla v(q)) = 0$.
Il termine viscosità, che identifica questa classe di soluzioni deboli, deriva dal processo usato inizialmente per ottenerle. Il metodo consiste nel considerare le soluzioni classiche v_ε del problema regolarizzato $F(q, v_\varepsilon(q), \nabla v_\varepsilon(q)) = \varepsilon \Delta v_\varepsilon(q)$ con $q \in Q$ e calcolare v come

$$v = \lim_{\varepsilon \to 0^+} v_\varepsilon,$$

tale procedura si definisce metodo della Viscosità Evanescente poiché, nel contesto della Meccanica dei Fluidi, ε rappresenterebbe la viscosità del fluido.
L'equivalenza tra quest'ultima nozione di soluzione di viscosità e la definizione 1.3 richiede una giustificazione formale, per cui si rimanda all'articolo [18, Teorema 3.1]. Tuttavia, per dare semplicemente un'idea, siano q_M e q_m rispettivamente punti di massimo e minimo locale di $v_\varepsilon - \phi$, allora[4] $\Delta(v_\varepsilon - \phi)(q_M) \leq 0$ e $\Delta(v_\varepsilon - \phi)(q_m) \geq 0$, ricordando che v_ε è soluzione classica del problema regolarizzato, si ha

$$0 = F(q_M, v_\varepsilon(q_M), \nabla v_\varepsilon(q_M)) - \varepsilon \Delta v_\varepsilon(q_M) \geq F(q_M, v_\varepsilon(q_M), \nabla \phi(q_M)) - \varepsilon \Delta \phi(q_M),$$
$$0 = F(q_m, v_\varepsilon(q_m), \nabla v_\varepsilon(q_m)) - \varepsilon \Delta v_\varepsilon(q_m) \leq F(q_m, v_\varepsilon(q_m), \nabla \phi(q_m)) - \varepsilon \Delta \phi(q_m).$$

[4] $\Delta f = \mathrm{Tr}(\nabla(\nabla f))$, x punto di massimo locale per $f \implies \nabla(\nabla f)(x) \preccurlyeq 0 \implies \Delta f(x) \leq 0$, al contrario x punto di minimo locale per $f \implies \nabla(\nabla f)(x) \succcurlyeq 0 \implies \Delta f(x) \geq 0$ ($A \preccurlyeq 0$ e $A \succcurlyeq 0$ significa che A è, rispettivamente, semidefinita negativa e positiva).

passando quindi al limite $\varepsilon \to 0$, si ritrovano rispettivamente le condizioni (SubS) e (SuperS).

Proposizione 1.1

a) Se $v \in C(Q)$ è una soluzione di viscosità in Q per $F(q, v(q), \nabla v(q)) = 0$, allora v è soluzione di viscosità per la stessa equazione in qualsiasi aperto $Q' \subset Q$.

b) Se $v \in C(Q)$ è una soluzione classica in Q di $F(q, v(q), \nabla v(q)) = 0$ (cioè v è differenziabile con continuità e soddisfa l'equazione $\forall x \in Q$), allora v è anche soluzione di viscosità.

c) Se $v \in C^1(Q)$ è soluzione di viscosità in Q per $F(q, v(q), \nabla v(q)) = 0$, allora v è anche una soluzione classica.

Dimostrazione :

a) Se $q_M \in Q'$ è un punto di massimo locale per $v - \phi'$ con $\phi' \in C^1(Q')$, allora q_M è un punto di massimo locale anche per $v - \phi$ con $\phi \in C^1(Q)$ e tale che $\phi \equiv \phi'$ in un intorno circolare chiuso di raggio $r > 0$ centrato in q_M. Quindi, dalla definizione di subsoluzione di viscosità, si ha

$$F(q_M, v(q_M), \nabla \phi'(q_M)) = F(q_M, v(q_M), \nabla \phi(q_M)) \leq 0.$$

Un ragionamento identico si applica per dimostrare che v è anche supersoluzione di viscosità in Q'.

b) Presa qualsiasi $\phi \in C^1(Q)$ e un punto di massimo locale per $v - \phi$, per l'ipotesi di differenziabilità di v, si ha

$$\nabla(u - \phi)(q_M) = 0 \implies \nabla u(q_M) = \nabla \phi(q_M)$$

quindi, dalla definizione di soluzione classica,

$$0 = F(q_M, v(q_M), \nabla v(q_M)) = F(q_M, v(q_M), \nabla \phi(q_M)) \leq 0.$$

Questo risultato prova che v è subsoluzione di viscosità, lo stesso ragionamento si applica per dimostrare che v è anche supersoluzione di viscosità.

c) Per l'ipotesi di $v \in C^1(Q)$, si può scegliere come funzione test $\phi \equiv v$, allora $v - \phi \equiv 0$ e di conseguenza ogni punto del dominio è sia massimo che minimo locale. Dalle condizioni (SubS) e (SuperS) si ha che $\forall q \in Q$

$$0 \geq F(q, v(q), \nabla v(q)) \geq 0 \iff F(q, v(q), \nabla v(q)) = 0.$$

■

È importante sottolineare che il primo enunciato giustifica il carattere locale della nozione di soluzione di viscosità, invece gli altri due evidenziano il legame con il concetto di soluzione classica.

Continuazione esempio 1.1

Sia $v_a(t,x)$ che $v_b(t,x)$ non sono soluzione di viscosità del problema (1.1). Infatti, si consideri $v_a(t,x)$ e sia $\phi = x^2 - t$, chiaramente $\phi \in C^1((0,T]\times\mathbb{R})$ e $v_a - \phi$ ha un punto di minimo locale in $(t,0)$ per ogni $t \in (0,T]$, ma

$$\frac{\partial\phi}{\partial t}(t,0) + \left|\frac{\partial\phi}{\partial x}(t,0)\right| = -1 + 0 = -1 \not\geq 0.$$

Si consideri, allora, $v_b(t,x)$ e sia $\phi = t - x^2$, chiaramente $\phi \in C^1((0,T]\times\mathbb{R})$ e $v_b - \phi$ ha un punto di massimo locale in $(t,0)$ per ogni $t \in (0,T]$, ma

$$\frac{\partial\phi}{\partial t}(t,0) + \left|\frac{\partial\phi}{\partial x}(t,0)\right| = 1 + 0 = 1 \not\leq 0.$$

Definizione 1.4

Si definisce superdifferenziale di v in $a \in Q$ l'insieme

$$\nabla^+ v(a) = \left\{ p \in \mathbb{R}^{d+1} : \limsup_{b\to a, b\in Q} \frac{v(b) - v(a) - p\cdot(b-a)}{|b-a|} \leq 0 \right\}.$$

Si definisce subdifferenziale di v in $a \in Q$ l'insieme

$$\nabla^- v(a) = \left\{ p \in \mathbb{R}^{d+1} : \liminf_{b\to a, b\in Q} \frac{v(b) - v(a) - p\cdot(b-a)}{|b-a|} \geq 0 \right\}.$$

Gli insiemi appena definiti sono sottoinsiemi chiusi e convessi di $\mathbb{R}^{d+1}$. Se in un punto $q \in Q$, sia $\nabla^+ v(q)$ che $\nabla^- v(q)$ sono non vuoti, allora entrambi contengono solo l'elemento $\nabla v(q)$, quindi v è differenziabile in q, inoltre vale anche il viceversa.

Lemma 1.2

Sia $v \in C(Q)$, allora

a) $p \in D^+(q) \iff \exists \phi \in C^1(Q)$ tale che $\nabla\phi(q) = p$ e $v - \phi$ ha un massimo locale in q,

b) $p \in D^-(q) \iff \exists \phi \in C^1(Q)$ tale che $\nabla\phi(q) = p$ e $v - \phi$ ha un minimo locale in q.

Sia il Lemma appena riportato, che le proprietà esposte in relazione agli insiemi subdifferenziale e superdifferenziale, sono dimostrate in [15, Cap. 2, Lemma 1.7 e Lemma 1.8]. Come diretta conseguenza, tali enunciati permettono di scrivere una definizione equivalente di soluzione di viscosità.

Definizione 1.5

Una funzione $v \in C(Q)$ si definisce subsoluzione di viscosità in Q per l'equazione $F(q, v(q), \nabla v(q)) = 0$ se, $\forall q \in Q$ e $\forall p \in \nabla^+ v(q)$,

$$F(q, v(q), p) \leq 0.$$

Definizione 1.6

Una funzione $v \in C(Q)$ si definisce supersoluzione di viscosità in Q per l'equazione $F(q, v(q), \nabla v(q)) = 0$ se, $\forall q \in Q$ e $\forall p \in \nabla^- v(q)$,

$$F(q, v(q), p) \geq 0.$$

È conveniente introdurre queste nuove definizioni poiché sono più adatte a dimostrare importanti proprietà di questa classe di soluzioni, tra cui quelle di seguito enunciate.

Proposizione 1.3

a) Se $v \in C(Q)$ è una soluzione di viscosità in Q per $F(q, v(q), \nabla v(q)) = 0$, allora $F(q, v(q), \nabla v(q)) = 0$ in ogni punto $q \in Q$ in cui v è differenziabile.

b) Se v è localmente lipschitziana ed è una soluzione di viscosità in Q per l'equazione $F(q, v(q), \nabla v(q)) = 0$, allora $F(q, v(q), \nabla v(q)) = 0$ quasi ovunque in Q.

Dimostrazione :

a) Sia $q \in Q$ un punto di differenziabilità per v, allora gli insiemi $\{\nabla v(q)\}$, $\nabla^+ v(q)$ e $\nabla^- v(q)$ coincidono, in virtù delle definizioni 1.5 e 1.6 si ricava che

$$0 \geq F(q, v(q), \nabla v(q)) \geq 0 \iff F(q, v(q), \nabla v(q)) = 0.$$

b) Per il Teorema di Rademacher[5] v è differenziabile quasi ovunque in Q, per l'enunciato *a)* della proposizione 1.3 si deduce quindi che $F(q, v(q), \nabla v(q)) = 0$ quasi ovunque in Q. ■

Conclusione esempio 1.1

Si consideri la soluzione debole $v_c(t, x)$ con $(t, x) \in Q = (0, T] \times \mathbb{R}$, gli unici punti

[5] $f : \Omega \to \mathbb{R}^m$ localmente lipschitziana in $\Omega \subseteq \mathbb{R}^n$ aperto $\implies$ f è differenziabile quasi ovunque [16, sez. 5.8, Teorema 6].

di non differenziabilità sono posti in $x = \pm t$. *Allora*

$$\nabla^- v(t,t) = \left\{ p \in \mathbb{R}^2 : \liminf_{\substack{(s,x)\to(t,t),\\(s,x)\in Q}} \frac{v(s,x) - v(t,t) - p\cdot(s-t,x-t)}{|(s-t,x-t)|} \geq 0 \right\} =$$

$$= \left\{ p \in \mathbb{R}^2 : \liminf_{\substack{(s,x)\to(t,t),\\(s,x)\in Q}} \frac{\max(|x|-s,0)}{|(s-t,x-t)|} - |p|\cos\theta \geq 0 \right\},$$

dove θ *indica l'angolo compreso tra il vettore* p *e* $(s-t, x-t)$. *Si denoti l'insieme* $S = \{(-\alpha,\alpha) : \alpha \geq 0\}$, $\forall p \in \mathbb{R}^2 \setminus S$ *si può prendere* $(s,x) \to (t,t)$, *con* $|x| < s$, *tale che* $\cos\theta > 0$, *quindi* $|p|\cos\theta > 0$ *e*

$$\liminf_{\substack{(s,x)\to(t,t),\\(s,x)\in Q}} \frac{\max(|x|-s,0)}{|(s-t,x-t)|} - |p|\cos\theta = \liminf_{\substack{(s,x)\to(t,t),\\(s,x)\in Q}} 0 - |p|\cos\theta \not\geq 0.$$

Per gli elementi di S, *sotto le stesse ipotesi, si ha* $\cos\theta < 0$, *quindi* $|p|\cos\theta < 0$ *e*

$$\liminf_{\substack{(s,x)\to(t,t),\\(s,x)\in Q}} \frac{\max(|x|-s,0)}{|(s-t,x-t)|} - |p|\cos\theta = \liminf_{\substack{(s,x)\to(t,t),\\(s,x)\in Q}} 0 - |p|\cos\theta = 0.$$

Tuttavia, non è ancora conclusa la verifica che gli elementi $p = \alpha(-1,1)$, *con* $\alpha \geq 0$, *appartengono al subdifferenziale di* v *in* (t,t). *Se* $(s,x) \to (t,t)$, *con* $|x| \geq s$, *si ricava*

$$\frac{\max(|x|-s,0) - \alpha(t-s+x-t)}{|(s-t,x-t)|} = \frac{|x|-s-\alpha(x-s)}{|(s-t,x-t)|} = \frac{x-s-\alpha(x-s)}{|(s-t,x-t)|},$$

poiché $x > 0$ *quando* $x \to t$. *Per cui*

$$\frac{x-s-\alpha(x-s)}{|(s-t,x-t)|} = \frac{(1-\alpha)(x-s)}{|(s-t,x-t)|} \geq 0 \iff \alpha \leq 1.$$

Quindi, risulta $\nabla^- v(t,t) = \{(-\alpha,\alpha) : \alpha \in [0,1]\}$ *e identicamente si ottiene anche* $\nabla^- v(t,-t) = \{(\beta,\beta) : \beta \in [-1,0]\}$. *In virtù delle proprietà accennate in relazione agli insiemi subdifferenziale e superdifferenziale, si ha necessariamente* $\nabla^+ v(t,t) = \nabla^+ v(t,-t) = \emptyset$.
Dunque, richiamando la definizione 1.6,

$$\begin{cases} \forall p \in \nabla^- v(t,t), & p_1 + |p_2| = -\alpha + |\alpha| = -\alpha + \alpha = 0, \\ \forall p \in \nabla^- v(-t,t), & p_1 + |p_2| = \beta + |\beta| = \beta - \beta = 0, \end{cases}$$

pertanto $v_c(t,x)$ *risulta essere una soluzione di viscosità dell'equazione* (1.1).

L'equazione (DE_1) descrive anche l'equazione evolutiva di Hamilton-Jacobi,

$$\frac{\partial v}{\partial t}(t,x) + H(t,x,\nabla_x v(t,x)) = 0, \quad (t,x) \in (0,T) \times \mathbb{R}^d,$$

quando $Q = (0,T) \times \mathbb{R}^d \implies q = (t,x)$ e

$$F(q, v(q), \nabla v(q)) = \frac{\partial v}{\partial q_1}(q) + H\left(q, \left[\frac{\partial v}{\partial q_2}(q), \ldots, \frac{\partial v}{\partial q_{d+1}}(q)\right]\right).$$

Si può quindi adattare la definizione di soluzione di viscosità a questa classe di soluzioni. Semplicemente, le condizioni che $v \in C((0,T)\times\mathbb{R}^d)$ deve soddisfare per essere subsoluzione e supersoluzione di viscosità in $(0,T) \times \mathbb{R}^d$ diventano rispettivamente

$$\frac{\partial \phi}{\partial t}(t_M, x_M) + H(t_M, x_M, \nabla_x \phi(t_M, x_M)) \leq 0$$
$$\frac{\partial \phi}{\partial t}(t_m, x_m) + H(t_m, x_m, \nabla_x \phi(t_m, x_m)) \geq 0,$$

con (t_M, x_M) e (t_m, x_m) rispettivamente punto di massimo e minimo locale di $v - \phi$, $\forall \phi \in C^1((0,T) \times \mathbb{R}^d)$.
Il principale risultato che si vuole dimostrare è l'unicità della soluzione di viscosità per questa classe di equazioni evolutive di Hamilton-Jacobi. Per i teoremi sull'esistenza delle soluzioni di viscosità per equazioni differenziali scalari del primo ordine (DE_1), si consiglia l'articolo [19] in cui viene introdotto il metodo di Perron, per i risultati sulla stabilità si consulti il testo [15, Cap. 2, Proposizione 2.2].
Al fine di alleggerire in parte la scrittura, si definiscono le notazioni $Q = (0,T)\times\mathbb{R}^d$, $\widehat{Q} = (0,T] \times \mathbb{R}^d$ e $\overline{Q} = [0,T] \times \mathbb{R}^d$.

Lemma 1.4

Siano $v, \tilde{v} \in C(\widehat{Q})$ rispettivamente subsoluzione e supersoluzione di viscosità in Q dell'equazione

$$\frac{\partial v}{\partial t}(t,x) + H(t,x,\nabla_x v(t,x)) = 0,$$

allora v e $\tilde{v}$ sono anche subsoluzione e supersoluzione di viscosità in $\widehat{Q}$.

La dimostrazione del Lemma 1.4 è riportata in [20, Lemma 2.8].

Teorema 1.5 (*Principio del confronto*)

Sia $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ una mappa che soddisfa le seguenti ipotesi:

- *$\exists m : \mathbb{R}^+ \to \mathbb{R}^+$ continua, crescente, con $m(0) = 0$, tale che $\forall x, y \in \mathbb{R}^d$, $\forall t \in (0,T]$ e $\forall p \in \mathbb{R}^d$ si ha $|H(t,x,p) - H(t,y,p)| \leq m(|x-y|(1+|p|))$;*
- *H è uniformemente continua in $[0,T] \times \mathbb{R}^d \times \overline{B}_r(0)$ $\forall r \geq 0$.*

Siano v e $\tilde{v}$ funzioni limitate e uniformemente continue in $\overline{Q}$, inoltre siano rispettivamente subsoluzione e supersoluzione di viscosità in Q dell'equazione

$$\frac{\partial v}{\partial t}(t,x) + H(t,x,\nabla_x v(t,x)) = 0,$$

allora

$$\sup_{(t,x)\in Q}\{v(t,x)-\tilde{v}(t,x)\} \leq \sup_{x\in\mathbb{R}^d}\{v(0,x)-\tilde{v}(0,x)\}.$$

Dimostrazione :

Sia $C = \sup_{x\in\mathbb{R}^d}\{v(0,x)-\tilde{v}(0,x)\} \in \mathbb{R}$ poiché per ipotesi v e $\tilde{v}$ sono limitate, sia inoltre $M = \sup_{\widehat{Q}}\{v-\tilde{v}\} - C$ e si supponga per assurdo $M > 0$.
Si consideri la funzione ausiliaria $\psi : \overline{Q}\times\overline{Q} \to \mathbb{R}$,

$$\psi(t,x,s,y) = v(t,x) - \tilde{v}(s,y) - \lambda t - \frac{|t-s|^2}{\eta^2} - \frac{|x-y|^2}{\varepsilon^2} - \alpha(|x|^2+|y|^2), \quad (1.2)$$

con $\lambda,\eta,\varepsilon > 0$ e $\alpha \in (0,1]$. Quando η e ε sono sufficientemente piccoli, il massimo di ψ si sposta verso i punti in cui (t,x) coincide con (s,y). Tuttavia, siccome il dominio non è compatto, il termine di penalizzazione con coefficiente α è necessario per garantire che il massimo di ψ sia raggiunto.
Infatti, la funzione ausiliaria è continua nel suo dominio, limitata superiormente e diverge negativamente per $|x|,|y| \to +\infty$, pertanto ammette un massimo globale M_0 in (t_0,x_0,s_0,y_0), con dipendenza dai parametri $\lambda,\eta,\varepsilon,\alpha$.
In particolare, è sempre possibile ottenere $t_0, s_0 > 0$, aggiustando opportunamente i parametri di ψ. Infatti, si possono scegliere λ,α così piccoli da imporre sia vera l'ultima disequazione prima del segno di implicazione,

$$M_0 = \max_{\overline{Q}\times\overline{Q}}\psi \geq \max_{(t,x)\in\overline{Q}}\psi(t,x,t,x) \geq \frac{M}{2} + C \implies$$
$$\implies v(t_0,x_0) - \tilde{v}(s_0,y_0) \geq \psi(t_0,x_0,s_0,y_0) = M_0 \geq \frac{M}{2} + C.$$

Aggiungendo e sottraendo al primo membro gli stessi termini, si ottiene

$$\frac{M}{2} + C \leq v(t_0,x_0) - v(0,x_0) + v(0,x_0) - \tilde{v}(0,x_0) +$$
$$+\tilde{v}(0,x_0) - \tilde{v}(t_0,x_0) + \tilde{v}(t_0,x_0) - \tilde{v}(s_0,y_0),$$

richiamando la definizione di C e denotando con ω e $\tilde{\omega}$, rispettivamente, il modulo di continuità[6] delle funzioni v e $\tilde{v}$, si ha

$$\frac{M}{2} + C \leq \omega(t_0) + C + \tilde{\omega}(t_0) + \tilde{\omega}(|t_0-s_0| + |x_0-y_0|),$$
$$\frac{M}{2} \leq \omega(t_0) + \tilde{\omega}(t_0) + \tilde{\omega}(\sqrt{4R}\eta + \sqrt{4R}\varepsilon).$$

[6]Una funzione $f : \mathbb{R}^+ \to \mathbb{R}^+$ continua, crescente, con $f(0) = 0$.

Si precisa che l'ultima disuguaglianza tiene conto delle relazioni (1.3) che verranno ricavate successivamente. Per concludere questa parte di dimostrazione, prendendo η, ε sufficientemente piccoli, si ottiene

$$\frac{M}{4} \leq \omega(t_0) + \tilde{\omega}(t_0),$$

ciò implica, per le proprietà dei moduli di continuità, che $t_0 > 0$. In modo del tutto analogo si deduce anche $s_0 > 0$.
Le stime di $|t_0 - s_0|$ e $|x_0 - y_0|$ si ricavano attraverso i seguenti passaggi,

$$\begin{aligned} &\psi(t,x,s,y) \leq \psi(t_0,x_0,s_0,y_0) \quad \forall (t,x,y,s) \in \overline{Q} \times \overline{Q} \implies \\ &\implies \psi(0,0,0,0) \leq \psi(t_0,x_0,s_0,y_0) \leq \psi(t_0,x_0,s_0,y_0) + \lambda t_0, \end{aligned}$$

esplicitamente si ha

$$\begin{aligned} &v(0,0) - \tilde{v}(0,0) \leq v(t_0,x_0) - \tilde{v}(s_0,y_0) - \frac{|t_0-s_0|^2}{\eta^2} - \frac{|x_0-y_0|^2}{\varepsilon^2} - \alpha(|x_0|^2 + |y_0|^2), \\ &\frac{|t_0-s_0|^2}{\eta^2} + \frac{|x_0-y_0|^2}{\varepsilon^2} + \alpha(|x_0|^2 + |y_0|^2) \leq v(t_0,x_0) - \tilde{v}(s_0,y_0) - v(0,0) + \tilde{v}(0,0), \\ &\frac{|t_0-s_0|^2}{\eta^2} + \frac{|x_0-y_0|^2}{\varepsilon^2} + \alpha(|x_0|^2 + |y_0|^2) \leq 4R^2, \end{aligned}$$

dove $R^2 = \max(\|v\|_\infty, \|\tilde{v}\|_\infty)$ e $R > 0$. L'ultima disuguaglianza deriva dall'ipotesi che v e $\tilde{v}$ sono limitate in $\overline{Q}$ indipendentemente da $\lambda, \eta, \varepsilon, \alpha$, pertanto

$$|t_0 - s_0| \leq 2R\eta, \quad |x_0 - y_0| \leq 2R\varepsilon, \tag{1.3}$$

$$\alpha|x_0| \leq 2R\sqrt{\alpha}, \quad \alpha|y_0| \leq 2R\sqrt{\alpha}. \tag{1.4}$$

In aggiunta, si può ricavare la stima (1.5) che risulterà utile successivamente,

$$\begin{aligned} &\psi(t,x,s,y) \leq \psi(t_0,x_0,s_0,y_0) \quad \forall (t,x,y,s) \in \overline{Q} \times \overline{Q} \implies \\ &\implies \psi(t_0,x_0,t_0,x_0) = v(t_0,x_0) - \tilde{v}(t_0,x_0) - \lambda t_0 - 2\alpha|x_0|^2 \leq \psi(t_0,x_0,s_0,y_0), \end{aligned}$$

dunque

$$\begin{aligned} &0 \leq \psi(t_0,x_0,s_0,y_0) - v(t_0,x_0) + \tilde{v}(t_0,x_0) + \lambda t_0 + 2\alpha|x_0|^2, \\ &0 \leq \tilde{v}(t_0,x_0) - \tilde{v}(s_0,y_0) - \frac{|t_0-s_0|^2}{\eta^2} - \frac{|x_0-y_0|^2}{\varepsilon^2} + \alpha(|x_0|^2 - |y_0|^2), \end{aligned}$$

quindi

$$\begin{aligned} \frac{|t_0-s_0|^2}{\eta^2} + \frac{|x_0-y_0|^2}{\varepsilon^2} &\leq \tilde{v}(t_0,x_0) - \tilde{v}(s_0,y_0) + \alpha(x_0+y_0)\cdot(x_0-y_0) \leq \\ &\leq \tilde{\omega}(|t_0-s_0| + |x_0-y_0|) + \alpha|x_0+y_0||x_0-y_0| \leq \\ &\leq \tilde{\omega}(2R\eta + 2R\varepsilon) + \alpha(|x_0|+|y_0|)2R\varepsilon \leq \\ &\leq \tilde{\omega}(2R\eta + 2R\varepsilon) + 4R\sqrt{\alpha}2R\varepsilon \leq \\ &\leq \tilde{\omega}(2R\eta + 2R\varepsilon) + 8R^2\varepsilon, \end{aligned} \tag{1.5}$$

la seconda disuguaglianza tiene conto della maggiorazione tramite modulo di continuità di $\tilde{v}$, invece le restanti disuguaglianze coinvolgono le stime (1.3), (1.4). Ricapitolando, finora si è dimostrato che la funzione ausiliaria ψ (1.2) ammette un massimo globale in $(t_0, x_0, s_0, y_0) \in \widehat{Q}$.
Si consideri la funzione $\varphi \in C^1(\widehat{Q})$,

$$\varphi(x,t) = \tilde{v}(s_0,y_0) + \lambda t + \frac{|t-s_0|^2}{\eta^2} + \frac{|x-y_0|^2}{\varepsilon^2} + \alpha(|x|^2 + |y_0|^2),$$

chiaramente $(x_0, t_0) \in \widehat{Q}$ è un punto di massimo di $v - \varphi$, inoltre v è subsoluzione di viscosità in $\widehat{Q}$ (Lemma 1.4), pertanto

$$\begin{aligned} \frac{\partial \varphi}{\partial t}(t_0,x_0) + H(t_0,x_0,\nabla_x \varphi(t_0,x_0)) \leq 0, \\ \lambda + \frac{2(t_0-s_0)}{\eta^2} + H\left(t_0, x_0, \frac{2(x_0-y_0)}{\varepsilon^2} + 2\alpha x_0\right) \leq 0. \end{aligned} \tag{1.6}$$

Analogamente, si consideri la nuova funzione $\varphi \in C^1(\widehat{Q})$,

$$\varphi(s,y) = v(t_0,x_0) - \lambda t_0 - \frac{|t_0-s|^2}{\eta^2} - \frac{|x_0-y|^2}{\varepsilon^2} - \alpha(|x_0|^2 + |y|^2),$$

chiaramente $(s_0, y_0) \in \widehat{Q}$ è un punto di massimo di $\varphi - \tilde{v}$, cioè un punto di minimo di $\tilde{v} - \varphi$, inoltre $\tilde{v}$ è supersoluzione di viscosità in $\widehat{Q}$ (Lemma 1.4), pertanto

$$\begin{aligned} \frac{\partial \varphi}{\partial t}(s_0,y_0) + H(s_0,y_0,\nabla_x \varphi(s_0,y_0)) \geq 0, \\ \frac{2(t_0-s_0)}{\eta^2} + H\left(s_0, y_0, \frac{2(x_0-y_0)}{\varepsilon^2} - 2\alpha y_0\right) \geq 0. \end{aligned} \tag{1.7}$$

Unendo le relazioni (1.6) e (1.7), si ottiene

$$\begin{aligned} \lambda + H\left(t_0, x_0, \frac{2(x_0-y_0)}{\varepsilon^2} + 2\alpha x_0\right) \leq H\left(s_0, y_0, \frac{2(x_0-y_0)}{\varepsilon^2} - 2\alpha y_0\right), \\ \lambda \leq H\left(s_0, y_0, \frac{2(x_0-y_0)}{\varepsilon^2} - 2\alpha y_0\right) - H\left(t_0, x_0, \frac{2(x_0-y_0)}{\varepsilon^2} + 2\alpha x_0\right), \end{aligned}$$

aggiungendo e sottraendo al secondo membro gli stessi termini, si ricava

$$\begin{aligned} \lambda \leq H\left(s_0, y_0, \frac{2(x_0-y_0)}{\varepsilon^2} - 2\alpha y_0\right) - H\left(t_0, x_0, \frac{2(x_0-y_0)}{\varepsilon^2} + 2\alpha x_0\right) + \\ + H\left(t_0, y_0, \frac{2(x_0-y_0)}{\varepsilon^2} + 2\alpha x_0\right) - H\left(t_0, y_0, \frac{2(x_0-y_0)}{\varepsilon^2} + 2\alpha x_0\right), \end{aligned} \tag{1.8}$$

in virtù della prima ipotesi relativa ad H, la disuguaglianza precedente diventa

$$\lambda \leq H(s_0, y_0, p_0 - 2\alpha(x_0 + y_0)) - H(t_0, y_0, p_0) + m(|x_0 - y_0|(1 + |p_0|)),$$

con $p_0 = 2(x_0 - y_0)/\varepsilon^2 + 2\alpha x_0$.
Si consideri l'argomento della funzione m, allora

$$\begin{aligned}|x_0 - y_0|(1 + |p_0|) &\leq |x_0 - y_0| + 2\frac{|x_0 - y_0|^2}{\varepsilon^2} + 2\alpha|x_0||x_0 - y_0| \leq \\ &\leq |x_0 - y_0|(1 + 2\alpha|x_0|) + 2\frac{|x_0 - y_0|^2}{\varepsilon^2} \leq \\ &\leq 2R\varepsilon(1 + 4R\sqrt{\alpha}) + 2\tilde{\omega}(2R\eta + 2R\varepsilon) + 16R^2\varepsilon \leq \\ &\leq 2R\varepsilon(1 + 4R) + 2\tilde{\omega}(2R\eta + 2R\varepsilon) + 16R^2\varepsilon,\end{aligned}$$

si precisa che nel penultimo passaggio si è tenuto conto delle stime (1.3), (1.4), (1.5).
Dunque, per ε e η sufficientemente piccoli risulta $m(|x_0 - y_0|(1 + |p_0|)) \leq \lambda/2$.
Inoltre, si ha

$$|p_0| \leq 2\frac{|x_0 - y_0|}{\varepsilon^2} + 2\alpha|x_0| \leq 2\frac{2R\varepsilon}{\varepsilon^2} + 4R\sqrt{\alpha} \leq \frac{4R}{\varepsilon} + 4R \leq r(\varepsilon),$$

dove nuovamente sono state utilizzate le stime (1.3), (1.4), identicamente si ricava $|p_0 - 2\alpha(x_0 + y_0)| \leq r(\varepsilon)$. Pertanto, attraverso l'ipotesi di continuità uniforme di H si ottiene il seguente risultato,

$$H(s_0, y_0, p_0 - 2\alpha(x_0 + y_0)) - H(t_0, y_0, p_0) \leq m_\varepsilon(|t_0 - s_0| + 2\alpha|x_0 + y_0|),$$

con m_ε[7], modulo di continuità di H in $[0, T] \times \mathbb{R}^d \times \overline{B}_{r(\varepsilon)}(0)$.
In sintesi, per ε e η sufficientemente piccoli, la disequazione (1.8) diventa

$$\begin{aligned}&\lambda \leq m_\varepsilon(|t_0 - s_0| + 2\alpha|x_0 + y_0|) + \frac{\lambda}{2}, \\ &\frac{\lambda}{2} \leq m_\varepsilon(|t_0 - s_0| + 2\alpha(|x_0| + |y_0|)), \\ &\lambda \leq 2m_\varepsilon(2R\eta + 8R\sqrt{\alpha}).\end{aligned}$$

Infine, prendendo il limite $\eta, \alpha \to 0$, si giunge alla contraddizione $0 < \lambda \leq 0$ che conclude la dimostrazione. Infatti, necessariamente deve essere $M \leq 0$, cioè

$$\begin{aligned}&\sup_{(t,x)\in\widehat{Q}} \{v(t, x) - \tilde{v}(t, x)\} - \sup_{x\in\mathbb{R}^d} \{v(0, x) - \tilde{v}(0, x)\} \leq 0, \\ &\sup_{(t,x)\in\widehat{Q}} \{v(t, x) - \tilde{v}(t, x)\} \leq \sup_{x\in\mathbb{R}^d} \{v(0, x) - \tilde{v}(0, x)\}.\end{aligned}$$

■

Si precisa che per la stesura di questa dimostrazione, si è fatto principalmente riferimento alla monografia [20, Teorema 2.4 e Teorema 2.8] e in modo complementare al testo [16, sez. 10.2, Teorema 1].

[7] Si veda la nota 6 al fondo di pagina 13.

Corollario 1.6

Sotto le ipotesi del Teorema 1.5 relative alla mappa H, il problema (HJ) *ammette al più una soluzione di viscosità limitata e uniformemente continua.*

Dimostrazione :

Si supponga esistano v_1 e v_2 soluzioni di viscosità del problema (HJ) limitate e uniformemente continue. Per ipotesi $v_1(0,x) \equiv v_2(0,x) \equiv v_0(x)\ \forall x \in \mathbb{R}^d$. Siccome v_1 e v_2 sono anche rispettivamente subsoluzione e supersoluzione, dal Teorema 1.5 si ottiene

$$\sup_{(t,x)\in\widehat{Q}} \{v_1(t,x) - v_2(t,x)\} \leq \sup_{x\in\mathbb{R}^d} \{v_1(0,x) - v_2(0,x)\} = 0 \implies v_1 \leq v_2 \text{ in } \widehat{Q},$$

invertendo il ruolo di v_1 e v_2 si giunge invece al risultato opposto, cioè $v_2 \leq v_1$ in $\widehat{Q}$. Pertanto, l'unica possibilità è che $v_1 \equiv v_2$ in $\widehat{Q}$. ■

2 Il metodo di Galerkin continuo e discontinuo

Il metodo agli Elementi Finiti risolve numericamente un problema differenziale riscritto in forma variazionale attraverso la discretizzazione del dominio Ω e l'approssimazione dello spazio funzionale V in cui vive la soluzione v del problema differenziale. Dunque, il problema variazionale di dimensione infinita produce un sistema di equazioni con un numero finito di incognite che, una volta risolto, caratterizzano la soluzione numerica.

2.1 Il problema ellittico

Si consideri il problema classico di un'equazione di diffusione-convezione-reazione nella variabile v e dominio $\Omega \subset \mathbb{R}^d$ limitato,

$$\begin{cases} -\nabla \cdot (\varepsilon \nabla v) + \beta \cdot \nabla v + \sigma v = f & \text{in } \Omega, \\ v = g_D & \text{su } \Gamma_D \subseteq \partial\Omega, \\ \varepsilon \dfrac{\partial v}{\partial n} = g_N & \text{su } \Gamma_N = \partial\Omega \setminus \Gamma_D. \end{cases} \tag{SF}$$

I termini g_D e g_N rappresentano rispettivamente dei dati al bordo di Dirichlet (Γ_D) e Neumann (Γ_N), n è il versore normale uscente dal bordo Γ_N, mentre $\varepsilon, \beta, \sigma, f$ sono funzioni scalari che caratterizzano il problema differenziale. Tuttavia, quest'ultimo non ammette sempre l'esistenza di una soluzione nel senso classico[1], rendendo pertanto intrattabili alcuni casi fisicamente significativi. Per tale ragione, si introduce una formulazione che indebolisce le condizioni che v deve soddisfare.
Si moltiplichino ambo i membri dell'equazione alle derivate parziali di (SF) per una

[1] Si veda la nota 3 al fondo di pagina 5.

generica $w \in H^1_{0,\Gamma_D}(\Omega)$ con $H^1_{0,\Gamma_D}(\Omega) = \{w \in H^1(\Omega) : w_{|\Gamma_D} \equiv 0\}$ e si integri su Ω,

$$\begin{aligned}
&\int_\Omega -\nabla\cdot(\varepsilon\nabla v)w\,d\Omega + \int_\Omega (\beta\cdot\nabla v)w\,d\Omega + \int_\Omega \sigma v w\,d\Omega = \int_\Omega f w\,d\Omega,\\
&\int_\Omega \varepsilon\nabla v\cdot\nabla w\,d\Omega - \int_{\partial\Omega} \varepsilon\frac{\partial v}{\partial n} w\,d\Gamma + \int_\Omega (\beta\cdot\nabla v)w\,d\Omega + \int_\Omega \sigma v w\,d\Omega = \int_\Omega f w\,d\Omega,\\
&\int_\Omega \varepsilon\nabla v\cdot\nabla w\,d\Omega - \int_{\Gamma_N} g_N w\,d\Gamma + \int_\Omega (\beta\cdot\nabla v)w\,d\Omega + \int_\Omega \sigma v w\,d\Omega = \int_\Omega f w\,d\Omega,\\
&\int_\Omega \varepsilon\nabla v\cdot\nabla w\,d\Omega + \int_\Omega (\beta\cdot\nabla v)w\,d\Omega + \int_\Omega \sigma v w\,d\Omega = \int_\Omega f w\,d\Omega + \int_{\Gamma_N} g_N w\,d\Gamma. \quad (2.1)
\end{aligned}$$

Nel primo passaggio, al primo termine, è stata applicata l'identità seguente che deriva dal Teorema della Divergenza e dalla regola di derivazione del prodotto,

$$\begin{aligned}
\int_{\partial\Omega} \varepsilon\frac{\partial v}{\partial n} w\,d\Gamma &= \int_{\partial\Omega} \varepsilon(\nabla v\cdot n)w\,d\Gamma = \int_\Omega \nabla\cdot((\varepsilon\nabla v)w)\,d\Omega =\\
&= \int_\Omega \nabla\cdot(\varepsilon\nabla v)w\,d\Omega + \int_\Omega \varepsilon\nabla v\cdot\nabla w\,d\Omega,
\end{aligned}$$

mentre il secondo passaggio deriva dai dati al bordo di (SF) e da $w_{|\Gamma_D} \equiv 0$.
Ritornando all'equazione (2.1), affinché gli integrali risultino ben definiti, si supponga $v \in H^1(\Omega)$, $\varepsilon, \sigma \in L^\infty(\Omega)$, $\beta \in (L^\infty(\Omega))^d$, $f \in L^2(\Omega)$ e $g_N \in L^2(\Gamma_N)$, avendo ricordato che per costruzione $w \in H^1_{0,\Gamma_D}(\Omega)$.
Più precisamente, effettuando la decomposizione $v = v_0 + \mathcal{R}_{g_D}$, si ottiene come nuova incognita $v_0 \in H^1_{0,\Gamma_D}(\Omega)$, dove $\mathcal{R}_{g_D} \in H^1_{g_D,\Gamma_D}(\Omega)$ è un rilevamento di v e $H^1_{g_D,\Gamma_D}(\Omega)$ è lo spazio delle funzioni $H^1(\Omega)$ con g_D sul bordo di Dirichlet. Di conseguenza, le funzioni w vivono nello stesso spazio funzionale in cui si ricerca la soluzione e tale spazio funzionale è effettivamente un sottospazio chiuso di $H^1(\Omega)$, piuttosto che una varietà affine.
Si introducono allora la forma bilineare $a : H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}$ e la forma lineare $F : H^1(\Omega) \to \mathbb{R}$,

$$a(v,w) = \int_\Omega \varepsilon\nabla v\cdot\nabla w\,d\Omega + \int_\Omega (\beta\cdot\nabla v)w\,d\Omega + \int_\Omega \sigma v w\,d\Omega, \quad (2.2)$$

$$F(w) = \int_\Omega f w\,d\Omega + \int_{\Gamma_N} g_N w\,d\Gamma, \quad (2.3)$$

che, tramite il passaggio $F(w) = a(v,w) = a(v_0 + \mathcal{R}_{g_D}, w) = a(v_0,w) + a(\mathcal{R}_{g_D}, w)$, permettono di scrivere il problema differenziale in forma debole, o variazionale,

$$\begin{cases} \text{trovare } v_0 \in H^1_{0,\Gamma_D}(\Omega) \text{ tale che} \\ a(v_0,w) = F(w) - a(\mathcal{R}_{g_D}, w) \quad \forall w \in H^1_{0,\Gamma_D}(\Omega), \end{cases} \quad \text{(WF)}$$

in contrapposizione alla formulazione forte (SF). In questo modo, viene rilassata l'ipotesi di differenziabilità della soluzione fino all'ordine due e il significato delle

derivate viene sostituito dalla più generale versione distribuzionale.
Tuttavia, è necessario sancire delle condizioni sufficienti che rendano la formulazione (WF) ben posta nel senso di Hadamard (esistenza, unicità e dipendenza con continuità dai dati). A tal fine, si richiamano brevemente le seguenti definizioni,

- a si dice *continua* in V se $\exists C_a > 0$ tale che $|a(v,w)| \leq C_a \|v\|_V \|w\|_V \ \forall v, w \in V$,
- a si dice *coerciva* in V se $\exists \alpha > 0$ tale che $\alpha \|v\|_V^2 \leq a(v,v) \ \forall v \in V$,
- F si dice *continua* in V se $\exists C_F > 0$ tale che $|F(w)| \leq C_F \|w\|_V \ \forall w \in V$,

la più piccola costante C_F per cui è vera la precedente disuguaglianza è

$$\|F\|_{V'} = \sup_{w \in V \setminus \{0\}} \frac{F(w)}{\|w\|_V},$$

con V' spazio duale di V.

Teorema 2.1 (*Lax-Milgram*)

Sia V uno spazio di Hilbert, $a : V \times V \to \mathbb{R}$ una forma bilineare continua e coerciva in V e $F : V \to \mathbb{R}$ una forma lineare continua in V. Allora esiste ed è unica la soluzione del problema

$$\begin{cases} \text{trovare } v \in V \text{ tale che} \\ a(v,w) = F(w) \quad \forall w \in V. \end{cases} \tag{VP}$$

Corollario 2.2

La soluzione del problema (VP) *soddisfa*

$$\|v\|_V \leq \frac{1}{\alpha} \|F\|_{V'},$$

con α costante di coercività della forma bilineare a.

Dimostrazione :

Ricordando la definizione di coercività di una forma bilineare, scegliendo $w = v$ nell'equazione del problema (VP) e richiamando il significato di continuità di una forma lineare, si ha

$$\left.\begin{aligned} \alpha\|v\|_V^2 &\leq a(v,v) \\ a(v,v) &= F(v) \\ F(v) &\leq \|F\|_{V'}\|v\|_V \end{aligned}\right\} \implies \alpha\|v\|_V^2 \leq \|F\|_{V'}\|v\|_V \implies \|v\|_V \leq \frac{1}{\alpha}\|F\|_{V'}.$$

■

Dal corollario del Teorema 2.1, la cui dimostrazione è in [21, Teorema 5.1.1], si deduce un'ultima importante proprietà. Siano v_1 e v_2 soluzione del problema (VP) con dati rispettivamente F_1 e F_2, allora $\forall w \in V$

$$\left.\begin{aligned} a(v_1, w) &= F_1(w) \\ a(v_2, w) &= F_2(w) \end{aligned}\right\} \implies a(v_1 - v_2, w) = (F_1 - F_2)(w) \implies \|v_1 - v_2\|_V \leq \frac{1}{\alpha}\|F_1 - F_2\|_{V'}.$$

In pratica, una piccola perturbazione dei dati genera una piccola variazione della soluzione in norma V.
Ritornando quindi alla formulazione (WF), la risolvibilità è garantita se, denotando con V lo spazio $H^1_{0,\Gamma_D}(\Omega)$, si caratterizza la forma bilineare a dell'equazione (2.2) come continua e coerciva in V e la forma lineare $\tilde{F}$ come continua in V, dove $\tilde{F}(w) = F(w) - a(\mathcal{R}_{g_D}, w)$ e F coincide con l'espressione (2.3).
Prima di procedere, conviene precisare che lo spazio $H^1_{0,\Gamma_D}(\Omega)$ eredita la norma di $H^1(\Omega)$ in quanto suo sottospazio, cioè

$$\|w\|^2_{H^1_{0,\Gamma_D}(\Omega)} = \|w\|^2_{H^1(\Omega)} = \|w\|^2_{L^2(\Omega)} + \|\nabla w\|^2_{(L^2(\Omega))^d},$$

inoltre si deduce immediatamente che

$$\|w\|_{L^2(\Omega)} \leq \|w\|_{H^1(\Omega)}, \quad \|\nabla w\|_{(L^2(\Omega))^d} \leq \|w\|_{H^1(\Omega)}.$$

Dunque, applicando ad ogni addendo della forma bilineare a la disuguaglianza di Hölder estesa al prodotto di tre funzioni[2], si ha che $\forall v, w \in V$

$$\begin{aligned} \left|\int_\Omega \varepsilon \nabla v \cdot \nabla w \, d\Omega\right| &\leq \|\varepsilon\|_{L^\infty(\Omega)} \|\nabla v\|_{(L^2(\Omega))^d} \|\nabla w\|_{(L^2(\Omega))^d} \leq \|\varepsilon\|_{L^\infty(\Omega)} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}, \\ \left|\int_\Omega (\beta \cdot \nabla v) w \, d\Omega\right| &\leq \|\beta\|_{(L^\infty(\Omega))^d} \|\nabla v\|_{(L^2(\Omega))^d} \|w\|_{L^2(\Omega)} \leq \|\beta\|_{(L^\infty(\Omega))^d} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}, \\ \left|\int_\Omega \sigma v w \, d\Omega\right| &\leq \|\sigma\|_{L^\infty(\Omega)} \|v\|_{L^2(\Omega)} \|w\|_{L^2(\Omega)} \leq \|\sigma\|_{L^\infty(\Omega)} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}, \end{aligned}$$

da cui si ottiene immediatamente la continuità,

$$|a(v, w)| \leq (\|\varepsilon\|_{L^\infty(\Omega)} + \|\beta\|_{(L^\infty(\Omega))^d} + \|\sigma\|_{L^\infty(\Omega)}) \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}.$$

Per stabilire la coercività di a, si considerino i seguenti passaggi che derivano dalla conoscenza che $v \in H^1_{0,\Gamma_D}(\Omega)$, dal Teorema della Divergenza e dalla regola di

[2]Siano $f \in L^p$, $g \in L^q$ e $h \in L^r$ con $p, q, r \in [1, \infty]$ tali che $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, allora $fgh \in L^1$ e $\int |fgh| \leq \|f\|_{L^p} \|f\|_{L^q} \|f\|_{L^r}$.

derivazione del prodotto,

$$\begin{aligned}\int_{\Gamma_N}(\beta\cdot n)v^2\,d\Gamma &= \int_{\partial\Omega}(\beta\cdot n)v^2\,d\Gamma = \int_\Omega \nabla\cdot(\beta v^2)\,d\Omega = \\ &= \int_\Omega(\nabla\cdot\beta)v^2\,d\Omega + \int_\Omega \beta\cdot\nabla(v^2)\,d\Omega = \\ &= \int_\Omega(\nabla\cdot\beta)v^2\,d\Omega + 2\int_\Omega(\beta\cdot\nabla v)v\,d\Omega,\end{aligned}$$

dove si è supposto in aggiunta $\nabla\cdot\beta\in L^\infty(\Omega)$. Allora,

$$\begin{aligned}a(v,v) &= \int_\Omega \varepsilon|\nabla v|^2\,d\Omega + \int_\Omega(\beta\cdot\nabla v)v\,d\Omega + \int_\Omega \sigma v^2\,d\Omega, = \\ &= \int_\Omega \varepsilon|\nabla v|^2\,d\Omega + \frac{1}{2}\int_{\Gamma_N}(\beta\cdot n)v^2\,d\Gamma - \frac{1}{2}\int_\Omega(\nabla\cdot\beta)v^2\,d\Omega + \int_\Omega\sigma v^2\,d\Omega = \\ &= \int_\Omega \varepsilon|\nabla v|^2\,d\Omega + \int_\Omega\left(\sigma - \frac{1}{2}\nabla\cdot\beta\right)v^2\,d\Omega + \frac{1}{2}\int_{\Gamma_N}(\beta\cdot n)v^2\,d\Gamma,\end{aligned}$$

pertanto, assumendo che in Ω siano verificate

$$\varepsilon\geq\varepsilon_m>0,\quad \sigma-\frac{1}{2}\nabla\cdot\beta\geq\sigma_m>0$$

e lungo il bordo Γ_N valga $\beta\cdot n\geq 0$, si ottiene

$$a(v,v)\geq\varepsilon_m\int_\Omega|\nabla v|^2\,d\Omega+\sigma_m\int_\Omega v^2\,d\Omega\geq\min(\varepsilon_m,\sigma_m)\|v\|^2_{H^1(\Omega)}.$$

Infine, estendendo la continuità di a alle funzioni dello spazio $H^1(\Omega)$, si ha

$$|a(\mathcal{R}_{g_D},w)|\leq C_a\|\mathcal{R}_{g_D}\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)},$$

dove $C_a=\|\varepsilon\|_{L^\infty(\Omega)}+\|\beta\|_{(L^\infty(\Omega))^d}+\|\sigma\|_{L^\infty(\Omega)}$, per cui

$$|\tilde{F}(w)|\leq|F(w)|+|a(\mathcal{R}_{g_D},w)|\leq|F(w)|+C_a\|\mathcal{R}_{g_D}\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}.$$

La continuità della forma $\tilde{F}$ si ottiene quindi dalla disequazione precedente applicando la Disuguaglianza di Cauchy-Schwarz e il Teorema della Traccia per gli spazi di Sobolev[3] [22, Teorema 2.3] al termine $|F(w)|$, cioè

$$\begin{aligned}|F(w)| &\leq \|f\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}+\|g_N\|_{L^2(\Gamma_N)}\|Tw\|_{L^2(\Gamma_N)}\leq \\ &\leq \|f\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}+\|g_N\|_{L^2(\Gamma_N)}(C_T\|w\|_{H^1(\Omega)})\leq \\ &\leq (\|f\|_{L^2(\Omega)}+C_T\|g_N\|_{L^2(\Gamma_N)})\|w\|_{H^1(\Omega)}.\end{aligned}$$

[3]Sia $\Omega\subset\mathbb{R}^N$ un aperto limitato e sia $\partial\Omega$ la sua frontiera, se quest'ultima è sufficientemente regolare, $\exists!T: H^1(\Omega)\to L^2(\Omega)$, applicazione lineare e continua, tale che $Tf=f|_{\partial\Omega}$ per ogni $f\in H^1(\Omega)\cap C(\overline{\Omega})$, inoltre $\exists C_T>0$ tale che $\|Tf\|_{L^2(\partial\Omega)}\leq C_T\|f\|_{H^1(\Omega)}$.

2.1.1 Il metodo di Galerkin

Sia $V_h \subset V$ sottospazio di dimensione finita N_h, pertanto chiuso, con h un parametro. Tale spazio permette di definire il problema approssimato, o discreto,

$$\begin{cases} \text{trovare } v_h \in V_h \text{ tale che} \\ a(v_h, w_h) = F(w_h) \quad \forall w_h \in V_h. \end{cases} \tag{DVP}$$

Le ipotesi del Teorema 2.1 sono automaticamente soddisfatte se valgono per il problema esatto (VP), poichè la struttura dell'equazione differenziale è la medesima con gli operatori semplicemente ristretti al sottospazio chiuso V_h. Quindi esiste ed è unica la soluzione che risolve (DVP).
Indicando con $\{\varphi_j, j = 1, \dots N_h\}$ una base di V_h, in virtù delle proprietà di linearità delle forme a e F, (DVP) è equivalente a

$$\begin{cases} \text{trovare } v_h \in V_h \text{ tale che} \\ a(v_h, \varphi_j) = F(\varphi_j) \quad \forall j = 1, \dots, N_h, \end{cases}$$

inoltre $v_h \in V_h$ quindi si può esprimere come combinazione lineare di una sua base, cioè $v_h = \sum_{k=1}^{N_h} v_k \varphi_k$, con v_k coefficienti incogniti che caratterizzano la soluzione. Dunque si ottiene la formulazione equivalente

$$\begin{cases} \text{trovare } \{v_k\}_{k=1}^{N_h} \text{ tale che} \\ v_k a(\varphi_k, \varphi_j) = F(\varphi_j) \quad \forall j = 1, \dots, N_h, \end{cases} \iff \begin{cases} \text{trovare } \mathsf{v} \text{ tale che} \\ \mathsf{Av} = \mathsf{f}, \end{cases}$$

$\mathsf{A} \in \mathbb{R}^{N_h \times N_h}$ e $\mathsf{A_{jk}} = a(\varphi_k, \varphi_j)$, $\mathsf{v} \in \mathbb{R}^{N_h}$ e $\mathsf{v_k} = v_k$, $\mathsf{f} \in \mathbb{R}^{N_h}$ e $\mathsf{f_j} = F(\varphi_j)$.
A è nota come matrice di rigidezza ed è definita positiva, tale proprietà discende direttamente dalla coercività della forma bilineare a, infatti $\forall \mathsf{v} \in \mathbb{R}^{N_h}$ si ha

$$\mathsf{v}^\mathsf{T}\mathsf{Av} = \sum_{\mathsf{j,k}=1}^{N_h} \mathsf{v_j A_{jk} v_k} = \sum_{j=1}^{N_h} v_j \sum_{k=1}^{N_h} a(\varphi_k, \varphi_j) v_k = \sum_{j=1}^{N_h} v_j a(v_h, \varphi_j) = a(v_h, v_h) \geq \alpha \|v_h\|_V^2.$$

2.1.2 Elemento Finito

Il metodo degli Elementi Finiti non è altro che una particolare classe di possibili scelte per lo spazio V_h. Esso corrisponde a decomporre preventivamente il dominio Ω in parti geometriche, ognuna delle quali è genericamente indicata con E. Tipicamente, nel caso monodimensionale gli elementi della decomposizione sono degli intervalli e nel caso bidimensionale sono dei triangoli.
In ogni E, si sceglie una spazio di funzioni finito dimensionale, denotato con $V_{E,h}$ di dimensione N_E, grazie al quale è possibile costruire le funzioni di forma globali dello spazio V_h. Una scelta classica per lo spazio $V_{E,h}$ coincide con $\mathbb{P}_r(E)$, lo spazio dei polinomi di grado minore o uguale a $r \in \mathbb{N}$ definiti sull'elemento E.

Ogni funzioni di $V_{E,h}$ si può identificare assegnando N_E condizioni indipendenti, o gradi di libertà. Una scelta semplice e conveniente consiste nel prescrivere tali condizioni mediante una base dello spazio duale di $V_{E,h}$.
Dai concetti appena illustrati, si giunge alla seguente formalizzazione matematica.

Definizione 2.1

In $\mathbb{R}^d$, si definisce Elemento Finito la terna $(E, V_{E,h}, \mathcal{L}_E)$,

- *$E \subset \mathbb{R}^d$ è un insieme compatto, connesso e non vuoto, tale che il suo bordo è lipschitziano e $E = \overline{\mathring{E}}$;*
- *$V_{E,h}$ è uno spazio lineare di funzioni scalari definite su E, di dimensione finita N_E;*
- *$\mathcal{L}_E = \{l_j : j \in \mathbb{N} \wedge 1 \leq j \leq N_E\}$ è un insieme di forme lineari $l_j : V_{E,h} \to \mathbb{R}$, unisolvente[4] per $V_{E,h}$.*

Una volta fissato l'Elemento Finito, si individua la base canonica di $V_{E,h}$ che permette di scrivere ogni funzione di questo spazio in termini dei gradi di libertà, cioè

$$v_h \in V_{E,h} \implies v_h = \sum_{k=1}^{N_E} l_k(v_h)\phi_k,$$

con ϕ_k tale che

$$l_j(\phi_k) = \delta_{jk} \quad \forall j, k \in \{1, \ldots, N_E\}.$$

Conseguentemente, si consideri lo spazio V_E contenente $V_{E,h}$ e tale che le forme lineari dell'insieme $\mathcal{L}_E$ restino ben definite, allora si può introdurre l'operatore di proiezione locale $\Pi_E : V_E \to V_{E,h}$,

$$v \in V_E \implies \Pi_E v = \sum_{k=1}^{N_E} l_k(v)\phi_k.$$

È importante sottolineare che nella scelta arbitraria dell'Elemento Finito risiede la possibilità di ottenere automaticamente una regolarità globale della soluzione approssimata, come la continuità.
Si denoti con $\mathcal{T}_h$ la decomposizione del dominio Ω in un numero finito di elementi E_i, tale partizionamento (o *mesh*) risulta essere conforme se possiede le seguenti proprietà,

- $\cup_i E_i = \overline{\Omega}$;
- $E_i \neq \emptyset \ \forall i$;

[4] $\mathcal{F} = \{f_j\}_{j=1}^n$ è unisolvente per U se $\forall x \in \mathbb{R}^n$, $\exists! u \in U : f_j(u) = x_j$ per ogni $j \in \{1, \ldots, n\}$.

- $\mathring{E}_m \cap \mathring{E}_n = \emptyset \ \forall m \neq n$;
- $\forall m \neq n$, se $E_m \cap E_n \neq \emptyset$, l'intersezione è un vertice o una faccia massimale[5].

Per ogni elemento $E \in \mathcal{T}_h$, sia h_E la lunghezza del segmento più lungo contenuto in E, allora si può caratterizzare il parametro h, introdotto inizialmente, come

$$h = \max_{E \in \mathcal{T}_h} h_E.$$

Tuttavia, per evitare che alcuni elementi non abbiano una forma sufficientemente regolare, si impone che $\forall E \in \mathcal{T}_h$, il rapporto h_E/ρ_E sia limitato da una costante positiva fissata, dove ρ_E è il raggio della palla inscritta in E.
Per imporre un certo livello di regolarità globale, una possibilità consiste nel considerare dei gradi di libertà, incidenti sull'interfaccia tra elementi contigui, che siano sufficienti a stabilire questa proprietà. Per essere più precisi, si considerino due elementi adiacenti, E_m e E_n, sia $\Gamma = E_m \cap E_n$ l'interfaccia che li separa e siano $v_m \in V_{E_m,h}$ e $v_n \in V_{E_n,h}$. Si denoti con v la funzione definita su $E_m \cup E_n$ tale che $v_{|E_m} = v_m$ e $v_{|E_n} = v_n$, se $V_{E_m,h} \subset C^0(E_m)$ e $V_{E_n,h} \subset C^0(E_n)$ allora

$$v \in C^0(E_m \cup E_n) \iff v_{m|\Gamma} = v_{n|\Gamma},$$

cioè quando i gradi di libertà che incidono sull'interfaccia determinano $v_{|\Gamma}$. Analogamente, se $V_{E_m,h} \subset C^1(E_m)$ e $V_{E_n,h} \subset C^1(E_n)$ allora

$$v \in C^1(E_m \cup E_n) \iff \begin{cases} v_{m|\Gamma} = v_{n|\Gamma}, \\ (\nabla v_m)_{|\Gamma} = (\nabla v_n)_{|\Gamma}. \end{cases}$$

cioè quando i gradi di libertà che incidono sull'interfaccia determinano $v_{|\Gamma}$ e $(\nabla v)_{|\Gamma}$. Dunque, definendo V_h come

$$V_h = \{v_h \in V : v_{h|E} \in V_{E,h} \quad \forall E \in \mathcal{T}_h\},$$

se si è implementato un certo livello di regolarità globale, si possono caratterizzare i gradi di libertà globali a partire da quelli locali. Infatti, estendendo il dominio di tutte le forme lineari l_j da $V_{E,h}$ a V_h per ogni elemento $E \in \mathcal{T}_h$, unendo inoltre le forme l_j di elementi adiacenti che incidono sulla medesima interfaccia, si possono definire l'insieme delle funzioni di base composite $\varphi_k \in V_h$, tale che

$$l_j(\varphi_k) = \delta_{jk} \quad \forall j, k \in \{1, \dots N_h\}.$$

Per cui,

$$v_h \in V_h \implies v_h = \sum_{k=1}^{N_h} l_k(v_h)\varphi_k$$

[5]In geometria, la faccia massimale per un politopo è l'analogo della faccia per un poliedro e del lato per un poligono.

e l'operatore di proiezione globale $\Pi : V \to V_h$ si definisce come

$$v \in V \implies \Pi v = \sum_{k=1}^{N_h} l_k(v)\varphi_k. \tag{2.4}$$

2.1.3 Analisi dell'errore a priori

Si considerino le formulazione (VP) e (DVP), le relative soluzioni soddisfano

$$a(v - v_h, w_h) = a(v, w_h) - a(v_h, w_h) = F(w_h) - F(w_h) = 0, \tag{2.5}$$

poichè $w_h \in V_h \subset V$. La proprietà appena dedotta è nota come ortogonalità di Galerkin e vale $\forall w_h \in V_h \subset V$ e per la quantità $v - v_h$, che rappresenta l'errore di discretizzazione dello schema.

Lemma 2.3 (*di Céa*)

L'errore di discretizzazione del metodo di Galerkin applicato a (VP) *soddisfa la maggiorazione*

$$\|v - v_h\|_V \leq \frac{C_a}{\alpha} \inf_{w_h \in V_h} \|v - w_h\|_V,$$

con C_a e α costanti legate rispettivamente alla continuità e coercività della forma bilineare a.

Dimostrazione :

Siano v e v_h, soluzioni rispettivamente di (VP) e (DVP), allora

$$\begin{aligned}\alpha\|v - v_h\|_V^2 &\leq a(v - v_h, v - v_h) = a(v - v_h, v - w_h + w_h - v_h) = \\ &= a(v - v_h, v - w_h) + a(v - v_h, w_h - v_h) = a(v - v_h, v - w_h) \leq \\ &\leq |a(v - v_h, v - w_h)| \leq C_a\|v - v_h\|_V\|v - w_h\|_V,\end{aligned}$$

le maggiorazioni derivano, nell'ordine, dalla coercività e continuità di a, il termine $a(v-v_h, w_h-v_h)$ viene eliso poichè $w_h \in V_h$, quindi $w_h-v_h \in V_h$ e vale l'ortogonalità di Galerkin (2.5). Dunque, $\forall w_h \in V_h$ risulta

$$\alpha\|v - v_h\|_V^2 \leq C_a\|v - v_h\|_V\|v - w_h\|_V \implies \|v - v_h\|_V \leq \frac{C_a}{\alpha}\|v - w_h\|_V,$$

tale disuguaglianza vale ancora prendendo, tra le funzioni $w_h \in V_h$, l'estremo inferiore del termine che maggiora. ■

In virtù di tale lemma, considerando l'approssimazione generata dall'operatore di proiezione globale (2.4), si ottiene

$$\|v - v_h\|_V \leq \frac{C_a}{\alpha}\|v - \Pi v\|_V,$$

sotto opportune ipotesi è possibile maggiorare il termine a destra con un termine funzione del grado di approssimazione della soluzione e del raffinamento della discretizzazione.

Teorema 2.4

Sia $\{\mathcal{T}_h\}_{h>0}$ *una famiglia di decomposizioni conformi e dalla forma sufficientemente regolare*[6] *del dominio* Ω. *Sia* $v \in V \subset H^1(\Omega)$ *la soluzione esatta di* (VP) *e* $v_h \in V_h = \{v_h \in C^0(\Omega) : v_{h|E} \in \mathbb{P}_r(E) \quad \forall E \in \mathcal{T}_h\}$ *la soluzione approssimata che risolve* (DVP). *Se* $v \in H^{r+1}(\Omega)$, *allora vale la seguente stima a priori dell'errore di discretizzazione,*

$$\|v - v_h\|_{H^1(\Omega)} \leq K \frac{C_a}{\alpha} h^r |v|_{H^{r+1}(\Omega)},$$

con $K > 0$ *costante indipendente da* h *e* v.

Per la dimostrazione e per stime più generali si consulti il testo [22, sottosez. 4.5.3 e 4.5.4]. Questo teorema evidenzia che, sfruttando gli Elementi Finiti, quando la *mesh* viene raffinata, cioè $h \to 0$, la soluzione approssimata v_h del metodo di Galerkin converge alla soluzione esatta v in norma $H^1(\Omega)$.

2.1.4 La difficoltà della convezione dominante

Si consideri il problema (SF) caratterizzato in particolare da $\varepsilon \equiv \varepsilon_m > 0$, $\sigma \equiv 0$, $\Gamma_N = \emptyset$, $g_D \equiv 0$, cioè

$$\begin{cases} -\varepsilon_m \Delta v + \beta \cdot \nabla v = f & \text{in } \Omega, \\ v = 0 & \text{su } \partial\Omega. \end{cases} \tag{2.6}$$

Scrivendo tale problema in forma debole[7], si ottiene la formulazione

$$\begin{cases} \text{trovare } v \in H^1_{0,\partial\Omega}(\Omega) \text{ tale che} \\ a(v, w) = F(w) \quad \forall w \in H^1_{0,\partial\Omega}(\Omega), \end{cases}$$

con

$$a(v, w) = \varepsilon_m \int_\Omega \nabla v \cdot \nabla w \, d\Omega + \int_\Omega (\beta \cdot \nabla v) w \, d\Omega, \quad F(w) = \int_\Omega f w \, d\Omega.$$

Dunque, analogamente a quanto visto alla fine della sezione 2.1 per la forma bilineare a, la costante di continuità C_a si ottiene considerando la seguente disequazione,

$$|a(v, w)| \leq (e_m + \|\beta\|_{(L^\infty(\Omega))^d}) \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} = C_a \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}$$

[6] Si veda la sezione 2.1.2.

[7] Si seguono gli stessi passaggi effettuati per ricavare (WF).

e la costante di coercività α si ricava dai passaggi successivi,

$$\begin{aligned} a(v,v) &\geq \varepsilon_m \int_\Omega |\nabla v|^2 \, d\Omega - \frac{1}{2}\int_\Omega (\nabla \cdot \beta) v^2 \, d\Omega \geq \\ &\geq \varepsilon_m \int_\Omega |\nabla v|^2 \, d\Omega = \varepsilon_m \|\nabla v\|^2_{(L^2(\Omega))^d} \geq \\ &\geq \frac{\varepsilon_m}{1 + C_P^2(\Omega)} \|v\|^2_{H^1(\Omega)} = \alpha \|v\|^2_{H^1(\Omega)}, \end{aligned}$$

Nella seconda maggiorazione si è supposto $\nabla \cdot \beta \leq 0$ e nella terza si è applicato il risultato

$$\|v\|^2_{H^1(\Omega)} = \|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{(L^2(\Omega))^d} \leq (1 + C_P^2(\Omega)) \|\nabla v\|^2_{(L^2(\Omega))^d}$$

che deriva dalla disuguaglianza di Poincaré.

Proposizione 2.5 (*Disuguaglianza di Poincaré*)

Sia $\Omega \subset \mathbb{R}^d$ un aperto limitato con bordo lipschitziano, allora $\exists C_P > 0$ dipendente solo da Ω tale che

$$\|v\|_{L^2(\Omega)} \leq C_P(\Omega) \|\nabla v\|_{(L^2(\Omega))^d}, \quad \forall v \in H^1_{0,\partial\Omega}(\Omega).$$

Il fattore C_P si definisce costante di Poincaré per il dominio Ω.

Quindi,

$$\frac{C_a}{\alpha} = \frac{1 + C_P^2(\Omega)}{\varepsilon_m} \left(e_m + \sqrt{\sum_{i=1}^d \|\beta_i\|^2_{L^\infty(\Omega)}} \right) = (1 + C_P^2(\Omega))(1 + \mathcal{P}e),$$

con $\mathcal{P}e$, numero di Péclet, che rappresenta il rapporto tra il coefficiente convettivo e diffusivo dell'equazione alle derivate parziali di (2.6). In riferimento al Teorema di Lax-Milgram 2.1, quando il termine convettivo è dominante, cioè $\mathcal{P}e \gg 1$, la stima dell'errore di discretizzazione a priori diventa estremamente grande, in questo modo il metodo di Galerkin non è in grado di raggiungere una soluzione approssimata soddisfacente a meno, ad esempio, di raffinamenti elevati della *mesh*. Non è insolito, infatti, verificare in tali casi una soluzione discreta che presenta ripide oscillazioni non fisiche.

2.2 Il problema iperbolico

Si consideri un generico problema iperbolico di convezione-reazione nella variabile v e dominio $\Omega \subset \mathbb{R}^d$ limitato,

$$\begin{cases} \beta \cdot \nabla v + \sigma v = f & \text{in } \Omega, \\ v = g_{in} & \text{su } \Gamma_{in} \subseteq \partial\Omega, \end{cases} \tag{SF}$$

con g_{in} dato di Dirichlet legato alla porzione della frontiera di *inflow* del dominio, cioè $\Gamma_{in} = \{x \in \partial\Omega : \beta(x) \cdot n(x) < 0\}$ e n è il versore normale uscente da $\partial\Omega$, mentre β, σ, f sono funzioni scalari che caratterizzano il problema differenziale.
Per derivare una formulazione debole di (SF) che includa le condizioni al bordo al suo interno, si moltiplichino ambo i membri dell'equazione alle derivate parziali per una generica $w \in H^{1,\beta}(\Omega) = \{v \in L^2(\Omega) | \beta \cdot \nabla v \in L^2(\Omega)\}$[8] e si integri dunque su Ω,

$$\int_\Omega (\beta \cdot \nabla v) w \, d\Omega + \int_\Omega \sigma v w \, d\Omega = \int_\Omega f w \, d\Omega. \tag{2.7}$$

Il primo integrale si può riscrivere tenendo conto della seguente identità che deriva dal Teorema della Divergenza e dalla regola di derivazione del prodotto,

$$\int_{\partial\Omega} (\beta \cdot n) v w \, d\Gamma = \int_\Omega \nabla \cdot (\beta v w) \, d\Omega = \int_\Omega \nabla \cdot (\beta w) v \, d\Omega + \int_\Omega (\beta \cdot \nabla v) w \, d\Omega,$$

infatti

$$\begin{aligned}
\int_\Omega (\beta \cdot \nabla v) w \, d\Omega &= \int_{\partial\Omega} (\beta \cdot n) v w \, d\Gamma - \int_\Omega \nabla \cdot (\beta w) v \, d\Omega = \\
&= \int_{\Gamma_{in}} (\beta \cdot n) g_{in} w \, d\Gamma + \int_{\partial\Omega \setminus \Gamma_{in}} (\beta \cdot n) v w \, d\Gamma - \int_\Omega \nabla \cdot (\beta w) v \, d\Omega = \\
&= \int_{\Gamma_{in}} (\beta \cdot n) g_{in} w \, d\Gamma - \int_{\Gamma_{in}} (\beta \cdot n) v w \, d\Gamma + \int_\Omega (\beta \cdot \nabla v) w \, d\Omega,
\end{aligned}$$

dove si è richiamata la scomposizione del bordo di Ω e che $v = g_{in}$ su Γ_{in}. Quindi, l'equazione (2.7) diventa

$$\int_\Omega (\beta \cdot \nabla v) w \, d\Omega + \int_\Omega \sigma v w \, d\Omega - \int_{\Gamma_{in}} (\beta \cdot n) v w \, d\Gamma = \int_\Omega f w \, d\Omega - \int_{\Gamma_{in}} (\beta \cdot n) g_{in} w \, d\Gamma.$$

Affinché gli integrali introdotti risultino ben definiti, si supponga $v \in H^{1,\beta}(\Omega)$, $\beta \in (L^\infty(\Omega))^d$, $\sigma \in L^\infty(\Omega)$, $f \in L^2(\Omega)$ e $g_{in} \in L^2(\Gamma_{in})$. Si definiscono allora la forma bilineare $b : H^{1,\beta}(\Omega) \times H^{1,\beta}(\Omega) \to \mathbb{R}$ e la forma lineare $F : H^{1,\beta}(\Omega) \to \mathbb{R}$,

$$b(v, w) = \int_\Omega (\beta \cdot \nabla v) w \, d\Omega + \int_\Omega \sigma v w \, d\Omega - \int_{\Gamma_{in}} (\beta \cdot n) v w \, d\Gamma, \tag{2.8}$$

$$F(w) = \int_\Omega f w \, d\Omega - \int_{\Gamma_{in}} (\beta \cdot n) g_{in} w \, d\Gamma, \tag{2.9}$$

che permettono di scrivere il problema differenziale in forma debole, o variazionale,

$$\begin{cases} \text{trovare } v \in H^{1,\beta}(\Omega) \text{ tale che} \\ b(v, w) = F(w) \quad \forall w \in H^{1,\beta}(\Omega), \end{cases} \tag{WF}$$

in contrapposizione alla formulazione forte (SF).

[8] Lo spazio $H^{1,\beta}(\Omega) = \{v \in L^2(\Omega) | \beta \cdot \nabla v \in L^2(\Omega)\}$ è di Hilbert [23, Proposizione 2.2].

Teorema 2.6

Si supponga che $\nabla \cdot \beta \in L^\infty(\Omega)$ *e che quasi ovunque in* Ω *valga la relazione* $\sigma - (\nabla \cdot \beta)/2 \geq \sigma_m > 0$, *se*

$$v, g \in L^2(|\beta \cdot n|; \partial\Omega) = \left\{ v \text{ è misurabile su } \partial\Omega \,\middle|\, \int_{\partial\Omega} |\beta \cdot n| v^2 \, d\Omega < \infty \right\},$$

allora la formulazione (WF) *è ben posta nel senso di Hadamard, quindi la soluzione esiste, è unica e dipende con continuità dai dati.*

La dimostrazione di questo risultato è presentata nel testo [23, Lemma 2.5, Lemma 2.11 e Teorema 2.12] e si fonda sul Teorema di Banach–Nečas–Babuška.

2.2.1 Approssimazione discontinua

Nel 1973, venne introdotto da Reed e Hill [24], un primo metodo di Galerkin discontinuo come tecnica per risolvere l'equazione stazionaria del trasporto di neutroni, cioè un'equazione iperbolica lineare e indipendente dal tempo.
In riferimento al metodo di Galerkin classico (DVP), si rilassa la restrizione per cui lo spazio di dimensione finita V_h è contenuto in V, ma si richiede che V_h sia parte di uno spazio più grande W tale che $V \subset W$. Nella letteratura, i metodi agli Elementi Finiti in cui $V_h \not\subset V$ sono definiti non conformi.
Sia $\mathcal{T}_h$ una decomposizione conforme e dalla forma sufficientemente regolare[9] di Ω in un numero finito di elementi E_i e sia, con un leggero abuso di notazione,

$$\mathbb{P}_r(\mathcal{T}_h) = \{v \in L^2(\Omega) : v_{|E} \in \mathbb{P}_r(E) \quad \forall E \in \mathcal{T}_h\},$$

lo spazio dei polinomi di grado minore o uguale a $r \in \mathbb{N}$, discontinui lungo le interfacce che costituiscono la *mesh*. Inoltre, sia $\partial E_{in} = \{x \in \partial E : \beta(x) \cdot n(x) < 0\}$, la porzione di bordo di *inflow* di un generico elemento E, con n versore normale uscente da ∂E.
Si definisca la funzione residuo $\mathcal{R}_h : \Omega \to \mathbb{R}$ come

$$\mathcal{R}_h = \beta \cdot \nabla v_h + \sigma v_h - f,$$

con $v_h \in \mathbb{P}_r(\mathcal{T}_h)$, allora si vuole che tale quantità sia ortogonale ad ogni funzione $w_h \in \mathbb{P}_r(\mathcal{T}_h)$ in Ω, cioè

$$\int_\Omega \mathcal{R}_h w_h \, d\Omega = \sum_{E \in \mathcal{T}_h} \int_E \mathcal{R}_h w_h \, d\Omega = \sum_{E \in \mathcal{T}_h} \int_E (\beta \cdot \nabla v_h + \sigma v_h - f) w_h \, d\Omega = 0.$$

In virtù dell'arbitrarietà delle funzioni w_h, il problema è equivalente a imporre

$$\int_E (\beta \cdot \nabla v_h) w_h \, d\Omega + \int_E \sigma v_h w_h \, d\Omega = \int_E f w_h \, d\Omega, \quad \forall E \in \mathcal{T}_h.$$

[9] Si veda la sezione 2.1.2.

Seguendo gli stessi passaggi descritti nella sezione precedente, si ottiene

$$\int_E (\beta\cdot\nabla v_h) w_h \, d\Omega + \int_E \sigma v_h w_h \, d\Omega - \int_{\partial E_{in}} (\beta\cdot n) v_h^+ w_h^+ \, d\Gamma = \int_E f w_h \, d\Omega - \int_{\partial E_{in}} (\beta\cdot n) v_h^- w_h^+ \, d\Gamma,$$

con v_h^+ si intende il valore della funzione, ipoteticamente discontinua all'interfaccia con gli elementi adiacenti, presa dall'interno dell'elemento E, viceversa con v_h^- si considera il valore della funzione presa dall'interno dell'elemento che condivide l'interfaccia con E e lungo il quale si sta calcolando l'integrale. Inoltre, se $\partial E_{in} \cap \Gamma_{in} \neq \emptyset$, allora si pone $v_h^- = g_{in}$. Si noti che il significato di tali integrali di bordo non cambia adottando la notazione dei valori *upwind* di v_h, cioè $v_h^{\pm}(x) = \lim_{t\to 0^+} v_h(x \pm \beta t)$.
Si consideri l'equazione precedente e le seguenti equivalenze

$$\int_{\partial E_{in}} (\beta \cdot n) v_h^+ w_h^+ \, d\Gamma = \int_{\partial E_{in}\cap\Gamma_{in}} (\beta \cdot n) v_h^+ w_h^+ \, d\Gamma + \int_{\partial E_{in}\setminus\Gamma_{in}} (\beta \cdot n) v_h^+ w_h^+ \, d\Gamma,$$
$$\int_{\partial E_{in}} (\beta \cdot n) v_h^- w_h^+ \, d\Gamma = \int_{\partial E_{in}\cap\Gamma_{in}} (\beta \cdot n) g_{in} w_h^+ \, d\Gamma + \int_{\partial E_{in}\setminus\Gamma_{in}} (\beta \cdot n) v_h^- w_h^+ \, d\Gamma,$$

portando i termini che includono v_h a destra dell'uguale, i restanti a sinistra e sommando per ogni elemento del partizionamento $E \in \mathcal{T}_h$, si ottiene al primo membro

$$b_h(v_h, w_h) = \sum_{E\in\mathcal{T}_h} \int_E (\beta \cdot \nabla v_h) w_h \, d\Omega + \sum_{E\in\mathcal{T}_h} \int_E \sigma v_h w_h \, d\Omega -$$
$$- \sum_{E\in\mathcal{T}_h} \int_{\partial E_{in}\cap\Gamma_{in}} (\beta \cdot n) v_h^+ w_h^+ \, d\Gamma - \sum_{E\in\mathcal{T}_h} \int_{\partial E_{in}\setminus\Gamma_{in}} (\beta \cdot n) [\![v_h]\!] w_h^+ \, d\Gamma,$$

dove $[\![v_h]\!] = v_h^+ - v_h^-$, al secondo membro

$$F_h(w_h) = \sum_{E\in\mathcal{T}_h} \int_E f w_h \, d\Omega - \sum_{E\in\mathcal{T}_h} \int_{\partial E_{in}\cap\Gamma_{in}} (\beta \cdot n) g_{in} w_h^+ \, d\Gamma.$$

Allora il metodo di Galerkin discontinuo per il problema iperbolico lineare (SF) consiste nel

$$\begin{cases} \text{trovare } v_h \in \mathbb{P}_r(\mathcal{T}_h) \text{ tale che} \\ b_h(v_h, w_h) = F_h(w_h) \quad \forall w_h \in \mathbb{P}_r(\mathcal{T}_h). \end{cases}$$

Nell'articolo di conferenza [25, sez 5.1] è presentata, sotto opportune ipotesi, una stima dell'errore a priori relativa a tale approccio, così come in [26].

2.2.2 Il caso non stazionario

Storicamente, il metodo di Galerkin discontinuo (DG) è stato ampiamente studiato ed adoperato nel campo delle leggi di conservazione iperboliche. Tale metodo ha

suscitato interesse poiché capace di risolvere efficacemente equazioni iperboliche lineari scalari, oltre ad incorporare naturalmente l'idea dei flussi numerici del metodo ai Volumi Finiti (già molto sviluppato per le leggi di conservazione) [27].
Inoltre, è noto che le soluzioni dei problemi iperbolici, sviluppano discontinuità nella soluzione o nella derivata, anche per dati molto regolari, pertanto DG è potenzialmente adatto a catturare i salti fisicamente rilevanti della soluzione senza produrre oscillazioni spurie.
Si consideri la legge di conservazione scalare non lineare

$$\begin{cases} \dfrac{\partial v}{\partial t} + \nabla \cdot F(v) = 0, & \text{in } (0,T] \times \Omega, \\ v = g_{in}, & \text{su } (0,T] \times \Gamma_{in}, \\ v = v_0, & \text{in } \{0\} \times \Omega, \end{cases} \tag{CL}$$

$\Omega \subset \mathbb{R}^d$, $v : [0,T] \times \Omega \to \mathbb{R}$ è la quantità conservata incognita, $F : \mathbb{R} \to \mathbb{R}^d$ è la funzione flusso e Γ_{in} è il bordo di *inflow*, cioè la porzione di $\partial\Omega$ in cui $\partial F/\partial v \cdot n < 0$, con n versore normale uscente da $\partial\Omega$.
Nel caso in cui $F(v) = \beta v$, il problema a cui si giunge è la versione non stazionaria di (SF) con $\sigma = \nabla \cdot \beta$ e $f \equiv 0$.
Dunque, richiamando la trattazione della sezione precedente, si moltiplichi l'equazione di (CL), calcolata in $v_h \in \mathbb{P}_r(\mathcal{T}_h)$, per una generica $w_h \in \mathbb{P}_r(\mathcal{T}_h)$ e si integri sul dominio spaziale Ω,

$$\int_\Omega \frac{\partial v_h}{\partial t} w_h \, d\Omega + \int_\Omega (\nabla \cdot F(v_h)) w_h \, d\Omega = \sum_{E \in \mathcal{T}_h} \frac{\partial}{\partial t} \int_E v_h w_h \, d\Omega + \sum_{E \in \mathcal{T}_h} \int_E (\nabla \cdot F(v_h)) w_h \, d\Omega = 0.$$

L'equazione appena scritta, in virtù della seguente identità che deriva dal Teorema della Divergenza e dalla regola di derivazione del prodotto,

$$\int_{\partial E} (F(v_h) \cdot n) w_h \, d\Gamma = \int_E \nabla \cdot (F(v_h) w_h) \, d\Omega = \int_E (\nabla \cdot F(v_h)) w_h \, d\Omega + \int_E F(v_h) \cdot \nabla w_h \, d\Omega,$$

diventa

$$\sum_{E \in \mathcal{T}_h} \frac{\partial}{\partial t} \int_E v_h w_h \, d\Omega - \sum_{E \in \mathcal{T}_h} \int_E F(v_h) \cdot \nabla w_h \, d\Omega + \sum_{E \in \mathcal{T}_h} \int_{\partial E} (F(v_h) \cdot n) w_h \, d\Gamma = 0.$$

Gli integrali relativi alla frontiera degli elementi di $\mathcal{T}_h$ permettono di introdurre debolmente la condizione sul bordo di *inflow*, infatti vale la seguente uguaglianza,

$$\int_{\partial E} (F(v_h) \cdot n) w_h \, d\Gamma = \int_{\partial E \setminus \Gamma_{in}} (F(v_h) \cdot n) w_h \, d\Gamma + \int_{\partial E \cap \Gamma_{in}} (F(g_{in}) \cdot n) w_h \, d\Gamma.$$

Tuttavia, nel metodo di Galerkin discontinuo è ammesso che la soluzione numerica sia discontinua lungo le interfacce interne della *mesh*, pertanto è necessario formulare un flusso numerico $H : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^d$, tale che

$$\int_{\partial E \setminus \Gamma_{in}} (F(v_h) \cdot n) w_h \, d\Gamma \approx \int_{\partial E \setminus \Gamma_{in}} (H(v_h^+, v_h^-) \cdot n) w_h \, d\Gamma,$$

dove v_h^+ e v_h^-, rappresentano v_h, rispettivamente, dall'interno e dall'esterno di E. Sui bordi che costituiscono $\partial\Omega \setminus \Gamma_{in}$, in cui non c'è ambiguità sul valore della soluzione numerica, si ha $H(v_h^+, v_h^-) = F(v_h^+)$.
Come riportato in [28], è importante che l'espressione di H scelta possieda le seguenti proprietà:

- consistenza, quindi $H(v, v) = F(v)$;
- monotonia, cioè $H(a, b)$ crescente in a e decrescente in b.

Un esempio è il flusso di Lax-Friedrichs locale

$$H(a, b) = \frac{F(a) + F(b)}{2} + \frac{C}{2}(a - b)n,$$

con $C = \max_{v \in I(a,b)} |\partial F/\partial v \cdot n|$ e $I(a, b)$ intervallo di estremi a e b.
In totale, si ottiene quindi la discretizzazione spaziale del problema (CL) secondo un metodo di Galerkin discontinuo,

$$\sum_{E \in \mathcal{T}_h} \frac{\partial}{\partial t} \int_E v_h w_h \, d\Omega - \sum_{E \in \mathcal{T}_h} \int_E F(v_h) \cdot \nabla w_h \, d\Omega + \\ + \sum_{E \in \mathcal{T}_h} \int_{\partial E \setminus \Gamma_{in}} (H(v_h^+, v_h^-) \cdot n) w_h \, d\Gamma = - \sum_{E \in \mathcal{T}_h} \int_{\partial E \cap \Gamma_{in}} (F(g_{in}) \cdot n) w_h \, d\Gamma.$$

Per gli schemi di avanzamento in tempo sviluppati in questo specifico contesto si consulti il testo [23, sottosez. 3.2.3].

3 Metodo DGSL per l'equazione di HJB

Nella risoluzione dei problemi iperbolici, talvolta si richiede da parte dello schema numerico un margine di stabilità in relazione al passo temporale, in particolare nelle simulazioni in cui si considerano tempi lunghi. Un metodo semi-lagrangiano garantisce la stabilità necessaria senza vincoli sul passo di discretizzazione temporale e spaziale [12, sottosez. 5.1.3, Stabilità]. Insieme a questa proprietà, si desidera ottenere, in unico schema, anche i vantaggi offerti da una ricostruzione della soluzione con Elementi Finiti discontinui.

3.1 Lo schema semi-lagrangiano

In generale, i metodi semi-lagrangiani (SL) si definiscono tali poichè, ad ogni istante temporale, la soluzione viene individuata seguendo le traiettorie delle singole parti, cioè il punto di vista lagrangiano. Per le equazioni iperboliche, questo si traduce nell'analizzare l'evoluzione del sistema tramite il metodo delle caratteristiche.
Si consideri un dominio spaziale di dimensione $d = 1$ e il problema di trasporto con sorgente

$$\begin{cases} v_t(t,x) + f(t,x)v_x(t,x) = g(t,x), & (t,x) \in (0,T] \times \mathbb{R}, \\ v(0,x) = v_0(x), & x \in \mathbb{R}, \end{cases} \tag{3.1}$$

$f : (0,T] \times \mathbb{R} \to \mathbb{R}$ è il termine di convezione e $g : (0,T] \times \mathbb{R} \to \mathbb{R}$ rappresenta la forzante.
L'idea del metodo delle caratteristiche è di individuare, per ogni $x \in \mathbb{R}$, una curva nella porzione di piano $[0,t] \times \mathbb{R}$ che congiunga (t,x) con l'asse delle ascisse e lungo la quale il problema di partenza possa essere riscritto come un'equazione differenziale ordinaria del primo ordine.
Dunque, considerando l'equazione del problema (3.1), si deriva la soluzione v[1] nel

[1] Si Supponga che esista e sia sufficientemente regolare.

tempo lungo una curva della forma $(s, y_x(s))$,

$$\begin{aligned}\frac{dv}{dt}(t,x) &= \frac{dv}{dt}(t, y_x(t)) = v_t(t, y_x(t)) + \dot{y}_x(t) v_x(t, y_x(t)) = \\ &= v_t(t,x) + \dot{y}_x(t) v_x(t,x) = v_t(t,x) + f(t,x) v_x(t,x) = g(t,x),\end{aligned}$$

dove la penultima uguaglianza deriva dall'imposizione che $y_x(s)$ risolva il seguente sistema dinamico,

$$\begin{cases} \dot{y}(s) = f(s, y(s)), & 0 < s < t \leq T, \\ y(t) = x, & x \in \mathbb{R}. \end{cases} \tag{3.2}$$

Allora, la soluzione v si calcola in modo esplicito come

$$v(t,x) = v_0(y_x(0)) + \int_0^t g(s, y_x(s))\, ds \tag{3.3}$$

e $(s, y_x(s))$ si definisce curva caratteristica del problema (3.1). In particolare, quando il termine di sorgente $g \equiv 0$, si ottiene

$$\frac{dv}{dt}(t,x) = v_t(t,x) + \dot{y}_x(t) v_x(t,x) = 0 \implies v(t,x) = v_0(y_x(0)),$$

dunque nota la condizione iniziale v_0 e le curve caratteristiche, si ha completa conoscenza di $v(t,x)$ quasi ovunque[2] in $(0,T] \times \mathbb{R}$ a meno che f, g e v_0 non siano sufficientemente regolari. Infatti dal testo [12, Teorema 1.1] si riporta quanto segue.

Teorema 3.1

In riferimento al problema (3.1)*, siano* $f, g \in C^1((0,T) \times \mathbb{R})$ *e sia* $v_0 \in C^1(\mathbb{R})$*, allora la soluzione esiste, è unica e* $v \in C^1((0,T) \times \mathbb{R})$*.*

Una rappresentazione grafica del metodo delle caratteristiche, nel caso $g \equiv 0$ è rappresentata in figura 3.1. Lo schema semi-lagrangiano nasce dalla valutazione della formula (3.3), ottenuta attraverso la teoria del metodo delle caratteristiche, in un numero finito di punti del dominio,

$$v(t_{n+1}, x_i) = v(t_n, y_{x_i}(t_n)) + \int_{t_n}^{t_{n+1}} g(s, y_{x_i}(s))\, ds, \tag{SL}$$

si tratta di uno schema di avanzamento in tempo il cui obiettivo è di calcolare la soluzione nei nodi assegnati del dominio spaziale $x_i \in \mathbb{R}$ ad ogni istante temporale scelto, ad esempio $t_n = n\Delta t \in (0,T]$.
Tuttavia, quando non è nota la soluzione del sistema dinamico (3.2), è necessario

[2] L'intersezione delle caratteristiche forma un'onda di shock nella soluzione, una discontinuità che evolve nel tempo, al contrario l'allontanamento delle caratteristiche genera un'onda di rarefazione (espansione) nella soluzione, una zona intermedia che si espande nel tempo [29, sez. 2.7].

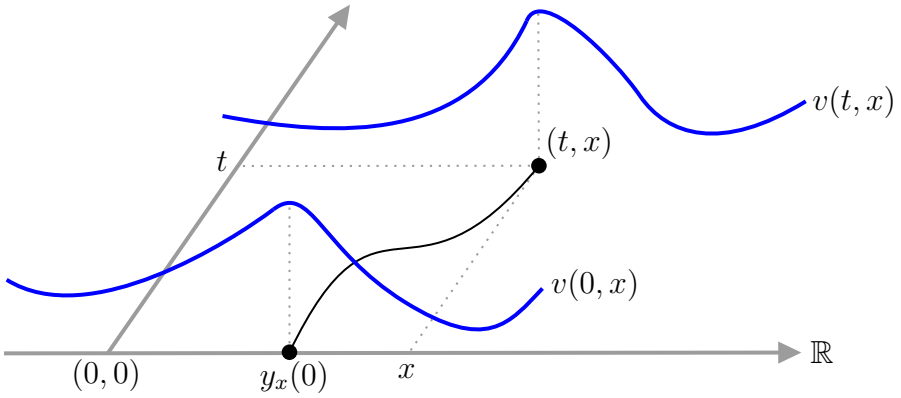

Figura 3.1: Metodo delle caratteristiche per l'equazione $v_t + f v_x = 0$ con f non costante e condizione iniziale nota, in nero è rappresentata la caratteristica $(s, y_x(s))$, $s \in [0,t]$, associata al punto (t,x).

introdurre delle approssimazioni, una scelta possibile e molto semplice consiste nel metodo di Eulero esplicito, quindi

$$x_i = y_{x_i}(t_{n+1}) \approx y_{x_i}(t_n) + f(t_{n+1}, x_i)\Delta t \implies y_{x_i}(t_n) \approx x_i - f(t_{n+1}, x_i)\Delta t.$$

Per quanto riguarda il termine integrale, la più elementare discretizzazione in questo contesto è data da

$$\int_{t_n}^{t_{n+1}} g(s, y_{x_i}(s))\, ds \approx g(t_{n+1}, y_{x_i}(t_{n+1}))\Delta t = g(t_{n+1}, x_i)\Delta t.$$

Inserendo le approssimazioni appena descritte nell'equazione (SL), si ottiene lo schema che prende il nome dai suoi autori, Courant-Isaacson-Rees [30],

$$\begin{cases} \tilde{v}(t_{n+1}, x_i) = \tilde{v}(t_n, x_i - f(t_{n+1}, x_i)\Delta t) + g(t_{n+1}, x_i)\Delta t, \\ \tilde{v}(0, x_i) = v_0(x_i). \end{cases} \tag{CIR}$$

La notazione $\tilde{v}$ non solo fa riferimento alle approssimazioni introdotte nel metodo, ma anche alla ricostruzione numerica della soluzione in tutto il dominio spaziale.
La caratteristica fondamentale dello schema (CIR), come viene riportato in [12, sottosez. 5.1.3] trattando la consistenza, stabilità e convergenza del metodo, è la libertà di considerare una distanza tra nodi x_i anche molto minore dello spazio percorso dall'informazione sul valore v in un tempo Δt.

3.2 Approssimazione di Galerkin per SL

Si discretizzi il dominio spaziale di interesse $[a,b]$ in N intervalli disgiunti, cioè si consideri il partizionamento $\mathcal{T}_h = \{I_i : i \in \mathbb{N} \wedge 1 \leq i \leq N\}$ con $I_i = [x_{i-1}, x_i]$ tale che

- $[a,b] = \cup_{i=1}^{N} I_i$;
- $a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b$;
- $x_i - x_{i-1} = h_i \neq 0$ per ogni $i \in \{1, \ldots, N\}$;
- $h = \max_{i \in \{1,\ldots,N\}} h_i$.

Tale decomposizione risulta essere conforme secondo la definizione riportata nella sezione 2.1.2. Inoltre, per semplicità, si consideri la griglia equispaziata, in questo modo, si può indicare $x_i = a + ih$ con $h = (b-a)/N$.
Sia $V_h \subset W$ lo spazio di dimensione finita, attraverso il quale si vuole ottenere un'approssimazione della soluzione $v \in V \subset W$, senza necessariamente imporre $V_h \subset V$. Ad esempio, nel caso degli Elementi Finiti continui, si potrebbe avere $V = H^1([a,b]) \subset L^2([a,b]) = W$ e

$$V_h = \{v \in C^0([a,b]) : v_{|I_i} \in \mathbb{P}_r(I_i) \quad \forall I_i \in \mathcal{T}_h\}.$$

Dunque, a partire dall'equazione (SL), la formulazione debole si ricava moltiplicando ambo i membri per una generica $w \in W$ e integrando sull'insieme $[a,b]$, cioè

$$\int_a^b v(t_{n+1},x)w(x)\,dx = \int_a^b v(t_n, y_x(t_n))w(x)\,dx + \int_a^b \int_{t_n}^{t_{n+1}} g(s, y_x(s))w(x)\,dsdx,$$

da cui si ottiene immediatamente il problema approssimato di trovare $v_h \in V_h$ tale che $\forall w_h \in V_h$

$$\int_a^b v_h^{n+1}(x)w_h(x)\,dx = \int_a^b v_h^n(y_x(t_n))w_h(x)\,dx + \int_a^b G(t_n, t_{n+1}, x)w_h(x)\,dx,$$

dove, al fine di alleggerire la notazione,

$$v_h^n(x) = v_h(t_n, x), \quad G(t_n, t_{n+1}, x) = \int_{t_n}^{t_{n+1}} g(s, y_x(s))\,ds.$$

Ogni elemento dello spazio V_h, di dimensione finita N_h, può essere espresso come combinazione lineare di una sua base,

$$v_h^n(x) = \sum_{k=1}^{N_h} v_k^n \varphi_k(x),$$

con $\varphi_k : [a,b] \to \mathbb{R}$ per ogni k.
Ritornando alla formulazione debole, essa risulta essere equivalente, in virtù della proprietà di linearità dell'integrale rispetto alla somma, al sistema lineare con incognite $v_1^{n+1}, \ldots, v_{N_h}^{n+1}$ di seguito riportato, cioè $\forall j \in \{1, \ldots, N_h\}$ deve risultare

$$\sum_{k=1}^{N_h} v_k^{n+1} \int_a^b \varphi_k(x)\varphi_j(x)\,dx = \sum_{k=1}^{N_h} v_k^n \int_a^b \varphi_k(y_x(t_n))\varphi_j(x)\,dx + \int_a^b G(t_n, t_{n+1}, x)\varphi_j(x)\,dx.$$

In forma matriciale, esso corrisponde a risolvere

$$\mathsf{M}\mathsf{v}^{\mathsf{n}+1} = \mathsf{C}^{\mathsf{n}}\mathsf{v}^{\mathsf{n}} + \mathsf{G}^{\mathsf{n}} \text{ con } \begin{cases} \mathsf{M} \in \mathbb{R}^{N_h \times N_h}, & \mathsf{M}_{\mathsf{jk}} = \displaystyle\int_a^b \varphi_k(x)\varphi_j(x)\,dx, \\ \mathsf{v}^{\mathsf{n}} \in \mathbb{R}^{N_h}, & \mathsf{v}^{\mathsf{n}}_{\mathsf{k}} = v_k^n, \\ \mathsf{C}^{\mathsf{n}} \in \mathbb{R}^{N_h \times N_h}, & \mathsf{C}^{\mathsf{n}}_{\mathsf{jk}} = \displaystyle\int_a^b \varphi_k(y_x(t_n))\varphi_j(x)\,dx, \\ \mathsf{G}^{\mathsf{n}} \in \mathbb{R}^{N_h}, & \mathsf{G}^{\mathsf{n}}_{\mathsf{j}} = \displaystyle\int_a^b G(t_n, t_{n+1}, x)\varphi_j(x)\,dx, \end{cases} \tag{3.4}$$

Tuttavia, questa scrittura generale non evidenzia le potenzialità e le differenze del metodo di Galerkin discontinuo rispetto alla sua versione più classica.

3.2.1 Implementazione DGSL - $\mathbb{P}_r$

Si consideri lo spazio delle funzioni polinomiali discontinue lungo le interfacce che costituiscono la *mesh*,

$$V_h = \mathbb{P}_r(\mathcal{T}_h) = \{v \in L^2([a,b]) : v_{|I_i} \in \mathbb{P}_r(I_i) \quad \forall I_i \in \mathcal{T}_h\},$$

insieme all'Elemento Finito $(I, \mathbb{P}_r(I), \mathcal{L}_{Lagr})$ in ogni intervallo I della *mesh*. In riferimento alla definizione 2.1, si intende che $\forall i \in \{1, \dots, N\}$

- $E = I_i$;
- $V_{E,h} = \mathbb{P}_r(I_i)$ e $N_E = r + 1$;
- $\mathcal{L}_E = \mathcal{L}_{Lagr} = \{l_j : l_j(v) = v(\xi_j)\}$;

dove $x_{i-1} = \xi_1 < \xi_2 < \ldots < \xi_r < \xi_{r+1} = x_i$ e l'insieme $\mathcal{L}_{Lagr}$ soddisfa la proprietà di unisolvenza per lo spazio $\mathbb{P}_r(I)$[3]. Pertanto la base canonica $\{\phi_1, \dots, \phi_{r+1}\}$ nell'intervallo I_i si identifica richiedendo

$$l_j(\phi_k) = \phi_k(\xi_j) = \delta_{jk} \quad \forall j, k \in \{1, \dots, r+1\}.$$

L'unione degli insiemi costituiti dalle basi di tutti gli intervalli $I \in \mathcal{T}_h$, una volta esteso opportunamente il dominio su cui sono definite, forma una base dello spazio V_h. Esplicitamente, $N_h = \dim(V_h) = N(r+1)$, imponendo

$$\varphi_k(x) = \begin{cases} \phi_j(x) & \text{se } x \in I_i, \\ 0 & \text{altrimenti,} \end{cases}$$

[3]Sia $f = a_0 + a_1 x + \ldots a_k x^k \in \mathbb{P}_k$, valutando f in $k+1$ punti si ottiene un sistema lineare di $k+1$ equazioni in altrettante incognite, $\mathsf{Va} = \mathsf{f}$, V è la matrice di Vandermonde e risulta invertibile se $\det(\mathsf{V}) = \prod_{0 \le i < j \le k}(x_j - x_i) \neq 0$, cioè se i $k+1$ punti sono distinti.

tramite la relazione $k = (i-1)(r+1)+j$, si ottiene

$$v_h^n(x) = \sum_{k=1}^{N_h} v_k^n \varphi_k(x) = \sum_{i=1}^{N}\sum_{j=1}^{r+1} v_k^n \phi_j(x) = \sum_{i=1}^{N}\sum_{j=1}^{r+1} v_h^n|_{I_i}(\xi_j)\phi_j(x),$$

dove si sottintende la dipendenza della base ϕ_j dall'intervallo I_i a cui si riferisce. In figura 3.1 è riportata una rappresentazione grafica delle funzioni di base globali del metodo proposto quando $r = 1$ e una eventuale soluzione numerica che si può ottenere.

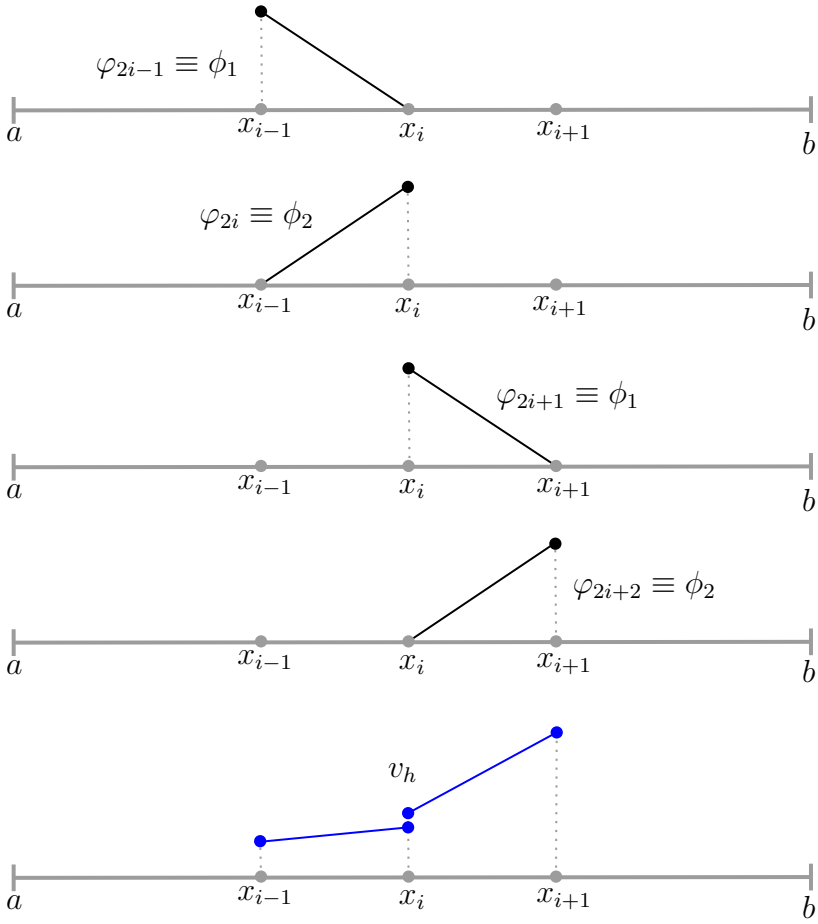

Figura 3.2: Rappresentazione delle funzioni di base globali (in nero) di $(I_i, \mathbb{P}_r(I_i), \mathcal{L}_{Lagr})$ e $(I_{i+1}, \mathbb{P}_r(I_{i+1}), \mathcal{L}_{Lagr})$ per gli elementi finiti discontinui e una possibile approssimazione v_h (in blu).

Dunque, relativamente alla formulazione matriciale descritta nell'equazione (3.4), si ha

$$\mathsf{M} = \begin{bmatrix} \boxed{\mathsf{M}_1} & & 0 \\ & \ddots & \\ 0 & & \boxed{\mathsf{M}_N} \end{bmatrix} \quad \text{con } (\mathsf{M}_i)_{jk} = \int_{I_i} \phi_k(x)\phi_j(x)dx,$$

poiché l'integrale su tutto lo spazio $[a, b]$ di due generiche funzioni di basi, ϕ_k e ϕ_j, è pari a 0 se l'intervallo a cui fanno riferimento è diverso, altrimenti coincide con l'integrale sull'intervallo I_i a cui appartengono. Inoltre, grazie alle ipotesi di griglia $\mathcal{T}_h$ equispaziata e di stesso Elemento Finito per ogni intervallo della *mesh*, tutti i blocchi non nulli di M sono uguali tra loro, cioè $\mathsf{M_i} = \mathsf{S_M}$ per ogni i.
Per quanto concerne le altre matrici dell'equazione (3.4), l'unica semplificazione è relativa al dominio di integrazione che diventa l'intervallo I_i in cui la funzione di base $\varphi_j|_{I_i} = \phi_j$ è non nulla.
Quindi, rimodellando il vettore colonna $\mathsf{v^n}$ in una struttura dati matriciale di $r+1$ righe (come il numero di gradi di libertà di $\mathcal{L}_{Lagr}$) e N colonne (come il numero degli intervalli) si ricava il metodo di Galerkin discontinuo semi-lagrangiano (DGSL) proposto, cioè

$$\mathsf{S_M v^{n+1}} = \mathsf{I_C^n}(\mathsf{v^n}) + \mathsf{G^n} \text{ con } \begin{cases} \mathsf{S_M} \in \mathbb{R}^{(r+1)\times(r+1)}, & (\mathsf{S_M})_{\mathsf{jk}} = \displaystyle\int_I \phi_k(x)\phi_j(x)\,dx, \\ \mathsf{v^n} \in \mathbb{R}^{(r+1)\times N}, & \mathsf{v^n_{jk}} = v_h^n|_{I_k}(\xi_j), \\ \mathsf{I_C^n}(\mathsf{v^n}) \in \mathbb{R}^{(r+1)\times N}, & (\mathsf{I_C^n}(\mathsf{v^n}))_{\mathsf{jk}} = \displaystyle\int_{I_k} v_h^n(y_x(t_n))\phi_j(x)\,dx, \\ \mathsf{G^n} \in \mathbb{R}^{(r+1)\times N}, & \mathsf{G^n_{jk}} = \displaystyle\int_{I_k} G(t_n, t_{n+1}, x)\phi_j(x)\,dx, \end{cases}$$

dove si vuole rappresentare con $\mathsf{I_C^n} : \mathbb{R}^{(r+1)\times N} \to \mathbb{R}^{(r+1)\times N}$ una funzione che opera sugli elementi delle matrici. Si preferisce tale scrittura perché, a differenza di quanto descritto per la matrice M, non ci sono dei pattern che semplificano la struttura di $\mathsf{C^n}$ dell'equazione (3.4), il motivo risiede nella caratteristica $y_x(t_n)$ che genera una traslazione della funzione di base non nota a priori, infatti in generale

$$\int_{I_i} \phi_k(y_x(t_n))\phi_j(x)\,dx \neq 0,$$

per ogni scelta di j, k.
In questo lavoro di tesi, si è scelto di calcolare gli integrali $(\mathsf{I_C^n}(\mathsf{v^n}))_{\mathsf{jk}}$ direttamente, poiché in numero $N(r+1)$ invece di $N^2(r+1)^2$, seguendo un ordine specifico che permetta di riutilizzare dei conti già effettuati. Si consideri una generica formula di quadratura[4], si ha dunque

$$\begin{aligned} \int_{I_k} v_h^n(y_x(t_n))\phi_j(x)\,dx &\approx \sum_{q=1}^{N_q} w_q v_h^n(y_{x_q}(t_n))\phi_j(x_q) = \\ &= \sum_{q=1}^{N_q} w_q \left[\sum_{\ell=1}^{r+1} v_h^n|_{I_{q'}}(\xi_\ell)\phi_\ell(y_{x_q}(t_n))\right]\phi_j(x_q), \end{aligned} \tag{3.5}$$

[4]Sia $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$, una formula di quadratura consiste di nodi ($x_q \in \Omega$) e relativi pesi (ω_q) di numero N_q affinché $\int_\Omega f(x)d\Omega \approx \sum_{q=1}^{N_q} \omega_q f(x_q)$.

con q' indice associato all'intervallo in cui ricade il punto $y_x(t_n)$, formalmente

$$q'(y_{x_q}(t_n)) = \{i \in \{1, \dots, N\} : y_{x_q}(t_n) \in I_i\}.$$

Si sottolinea che il termine tra parentesi quadre dell'equazione (3.5) è quella quantità che può non essere ricalcolata per ogni ϕ_j dell'intervallo I_k, perché comune a tutti. Infine, la nuova formulazione mette in risalto il carattere locale dell'approssimazione con un metodo agli Elementi Finiti discontinui. Infatti già dalla struttura matriciale di M, diagonale a blocchi, arriva il suggerimento che la soluzione numerica in un determinato intervallo della discretizzazione non è direttamente legata a quello che accade agli altri. In tal senso, la risoluzione dell'equazione alle derivate parziali viene divisa in problemi più piccoli che vengono risolti separatamente, a differenza degli elementi finiti continui.

3.2.2 Le condizioni al bordo

Nella teoria e nel metodo finora descritti, non si è menzionata la presenza delle condizioni al bordo che nascono spontaneamente quando si definisce un problema su un dominio limitato.
In generale, quando si valuta la curva caratteristica è possibile che quest'ultima superi i bordi del dominio di interesse, $y_x(t_n) \notin [a, b]$, creando quindi ambiguità in assenza di informazioni, le condizioni al bordo appunto.
Il caso con condizioni periodiche, che per il problema (3.1) si traduce formalmente in $v(t, a) = v(t, b)$ per ogni $t \in (0, T]$, non necessita di correggere la formulazione del metodo DGSL - $\mathbb{P}_r$ descritta precedentemente. L'unico accorgimento è che il valore assunto dall'approssimazione v_h in un qualsiasi punto esterno al dominio di interesse si identifica con un punto interno secondo la relazione

$$v_h^n(y_x(t_n)) = v_h^n(y_x(t_n) \pm m(b - a)), \quad \forall m \in \mathbb{N}.$$

Invece, nel caso di condizioni di Dirichlet, in generale non omogenee, è necessario avere informazioni sul bordo mediante delle funzioni $B(t)$, ad esempio $B_a(t)$ per il bordo di sinistra ($x = a$) e $B_b(t)$ per il bordo di destra ($x = b$) con $t \in (0, T]$. Allora per una caratteristica che proviene dall'esterno del dominio spaziale, $y_x(t_n) \notin [a, b]$, si indica con

$$\theta_x = \sup\{s \in (t_n, t_{n+1}] : y_x(s) \in \mathbb{R} \setminus (a, b)\}$$

il tempo in cui si ha la prima intersezione con il bordo. Se nell'intervallo $(t_n, t_{n+1}]$ non si hanno intersezioni, per convenzione si pone $\theta_x = \sup\{\emptyset\} = -\infty$.
Dunque, si consideri lo schema semi-lagrangiano esatto (SL), tenendo conto delle condizioni al bordo di Dirichlet si ha

$$v(t_{n+1}, x_i) = \begin{cases} v(t_n, y_{x_i}(t_n)) + \displaystyle\int_{t_n}^{t_{n+1}} g(s, y_{x_i}(s))\, ds, & \text{se } \theta_{x_i} \le t_n, \\ B(\theta_{x_i}) + \displaystyle\int_{\theta_{x_i}}^{t_{n+1}} g(s, y_{x_i}(s))\, ds, & \text{se } \theta_{x_i} > t_n. \end{cases}$$

Analogamente, in riferimento al metodo DGSL - $\mathbb{P}_r$, nell'equazione (3.5) si ha la seguente casistica,

$$v_h^n(y_{x_q}(t_n)) = \begin{cases} \sum_{\ell=1}^{r+1} v_h^n|_{I_{q'}}(\xi_\ell)\phi_\ell(y_{x_q}(t_n)), & \text{se } \theta_{x_q} \le t_n, \\ B(\theta_{x_q}), & \text{se } \theta_{x_q} > t_n, \end{cases}$$

inoltre, il termine $\mathsf{G}_{\mathsf{jk}}^{\mathsf{n}}$ diventa

$$\mathsf{G}_{\mathsf{jk}}^{\mathsf{n}} = \int_{I_k} \int_{\max\{t_n,\theta_x\}}^{t_{n+1}} g(s, y_x(s))\phi_j(x)\, ds\, dx.$$

3.3 Applicazione al caso di Hamilton-Jacobi-Bellman

Si richiamano dalla sezione 1.1 l'equazione di Hamilton-Jacobi-Bellman e la sua condizione al bordo, derivanti dal problema di controllo ottimo con dinamica governata da f, costo corrente ℓ, costo terminale L, spazio dei valori del controllo U, $\lambda = 0$ e dominio $[0, T) \times \mathbb{R}$,

$$\begin{cases} v_t(t,x) = \sup_{u\in U} \{-\ell(x,u) - f(t,x,u)v_x(t,x)\}, \\ v(T,x) = L(x). \end{cases} \tag{HJB}$$

La teoria delle soluzioni di viscosità si applica a questa classe di equazione che appartiene al più generale insieme dei problemi di Hamilton-Jacobi (HJ).

Teorema 3.2

La value function $v(t,x)$ è l'unica soluzione di viscosità dell'equazione di (HJB).

Dimostrazione :

In primo luogo, si riscriva l'equazione di (HJB) nella forma più adatta alla teoria della viscosità, cioè $v_t(t,x) + H(t,x,v_x) = 0$ con $(t,x) \in (0,T] \times \mathbb{R}$. Dunque, si consideri il cambio di variabile $t \to T - t$, per cui $v_t \to -v_t$ e

$$\begin{cases} -v_t(T-t,x) = \sup_{u\in U} \{-\ell(x,u) - f(T-t,x,u)v_x(T-t,x)\}, \\ v(T,x) = L(x), \end{cases}$$

quindi, ponendo $v(T-t,x) = V(t,x)$ e $f(T-t,x,u) = F(t,x,u)$, si ha

$$\begin{aligned} -V_t(t,x) &= \sup_{u\in U} \{-\ell(x,u) - F(t,x,u)V_x(t,x)\}, \\ -V_t(t,x) &+ \inf_{u\in U} \{\ell(x,u) + F(t,x,u)V_x(t,x)\} = 0, \\ V_t(t,x) &- \inf_{u\in U} \{\ell(x,u) + F(t,x,u)V_x(t,x)\} = 0. \end{aligned}$$

Sintetizzando, (HJB) risulta essere equivalente a

$$\begin{cases} V_t(t,x) - \inf\limits_{u\in U}\{\ell(x,u) + F(t,x,u)V_x(t,x)\} = 0, \\ V(0,x) = L(x), \end{cases}$$

con $(t,x) \in (0,T] \times \mathbb{R}$.
Al fine di verificare che $V \in C((0,T]\times\mathbb{R})$[5] è subsoluzione di viscosità per l'equazione del problema sopra riportato, si supponga per assurdo che esista un punto (t_M, x_M) di massimo locale di $V - \phi$, con $\phi \in C^1((0,T] \times \mathbb{R})$, in cui non sia soddisfatta la condizione (SubS). Allora

$$\phi_t(t_M,x_M) - \inf_{u\in U}\{\ell(x_M,u) + F(t_M,x_M,u)\phi_x(t_M,x_M)\} > 0,$$

per semplicità si supponga che esista il controllo u^* che minimizzi la quantità tra parentesi graffe, quindi

$$\phi_t(t_M,x_M) - \ell(x_M,u^*) - F(t_M,x_M,u^*)\phi_x(t_M,x_M) > 0.$$

Prendendo (t_M, x_M) come condizione iniziale, si consideri la traiettoria risultante dall'applicazione di un controllo costante $u(s) \equiv u^*$ per $s \in [t_M - \Delta t, t_M]$, cioè

$$\begin{cases} \dot{y}(s) = -F(s,y(s),u(s)), & t_M - \Delta t < s < t_M, \\ y(t_M) = x_M, & x \in \mathbb{R}, \end{cases}$$

il segno meno deriva dai medesimi motivi per cui $v_t \to -v_t$.
Indicando con $y_{x_M}(t)$ la traiettoria risultante da tale dinamica, Δt deve essere preso sufficientemente piccolo da mantenere vere le seguenti disequazioni nell'intervallo temporale $[t_M - \Delta t, t_M]$,

$$\begin{gathered} \phi_t(t,y_{x_M}(t)) - \ell(y_{x_M}(t),u^*) - F(t,y_{x_M}(t),u^*)\phi_x(t,y_{x_M}(t)) > 0, \\ V(t_M,x_M) - \phi(t_M,x_M) \geq V(t,y_{x_M}(t)) - \phi(t,y_{x_M}(t)), \end{gathered}$$

dove la seconda è una diretta conseguenza dell'ipotesi che $V - \phi$ abbia un massimo locale in (t_M, x_M) e sia una funzione continua. Dunque,

$$V(t_M,x_M) - V(t,y_{x_M}(t)) \geq \phi(t_M,x_M) - \phi(t,y_{x_M}(t)) = \int_t^{t_M} \frac{d\phi}{dt}(s,y_{x_M}(s))\,ds =$$
$$= \int_t^{t_M} \phi_t(s,y_{x_M}(s)) - F(s,y_{x_M}(s),u^*)\phi_x(s,y_{x_M}(s))\,ds > \int_t^{t_M} \ell(y_{x_M}(s),u^*)\,ds.$$

In $t = t_M - \Delta t = t_M^-$ si ha

$$V(t_M,x_M) > \int_{t_M^-}^{t_M} \ell(y_{x_M}(s),u^*)\,ds + V(t_M^-, y_{x_M}(t_M^-)),$$
$$v(T-t_M,x_M) > \int_{T-t_M^-}^{T-t_M} \ell(y_{x_M}(T-s),u^*)\,(-ds) + v(T-t_M^-, y_{x_M}(T-t_M^-)),$$
$$v(T-t_M,x_M) > \int_{T-t_M}^{T-t_M^-} \ell(y_{x_M}(T-s),u^*)\,ds + v(T-t_M^-, y_{x_M}(T-t_M^-)), \quad (3.6)$$

[5]Nella sezione 1.1, si riporta una Proposizione che sancisce v limitata e lipschitziana, quindi continua e di conseguenza anche V, poichè le relazioni che le legano preservano tale proprietà.

dove nel secondo passaggio si è ritornati alla notazione con v e nell'integrale si è effettuato il cambio di variabile $s \to T - s$.
L'equazione (3.6) contraddice il principio della programmazione dinamica (DPP) e pertanto V è subsoluzione di viscosità. In modo informale, il costo ottimo da $(T - t_M, x_M)$ è in realtà maggiore del costo che si ottiene applicando un controllo costante nell'intervallo $[T - t_M, T - t_M + \Delta t]$ e il controllo ottimo nel tempo rimanente.
Analogamente si dimostra anche che V è supersoluzione, quindi in totale V, ma specialmente la *value function* v, è una soluzione di viscosità dell'equazione di (HJB). Relativamente all'unicità, la funzione $H(t, x, p) = -\inf_{u \in U}\{\ell(x, u) + F(t, x, u)p\}$ soddisfa le ipotesi del Teorema 1.5, in realtà una versione dello stesso Teorema con delle ipotesi più rigide, come si vede nel testo [16, sez. 10.3, Teorema 2, Nota 1]. ■

La dimostrazione riportata è un adattamento della versione contenuta nel testo [31, Teorema 8.7.1].
Si consideri nuovamente, la formulazione equivalente di (HJB), ottenuta dalla riflessione e traslazione dell'asse dei tempi ($t \to T - t$),

$$\begin{cases} V_t(t, x) - \inf\limits_{u \in U} \{\ell(x, u) + F(t, x, u)V_x(t, x)\} = 0, \\ V(0, x) = L(x), \end{cases} \tag{HJB$^{\text{back}}_{\text{ward}}$}$$

dove $V(t, x) = v(T - t, x)$ e $F(t, x, u) = f(T - t, x, u)$. Anche la funzione V verifica la definizione, opportunamente adattata, di costo minimo del problema di controllo ottimo,

$$V(t, x) = \inf_{u \in \mathcal{U}_{ad}} \left\{ L(y^u_x(0)) + \int_0^t \ell(y^u_x(s), u(s))\, ds \right\}, \tag{3.7}$$

con $\mathcal{U}_{ad} = \{u : [0, t] \to U, \text{ misurabile}\}$ e $y_x(s)$ soluzione del sistema dinamico

$$\begin{cases} \dot{y}(s) = -F(s, y(s), u(s)), & 0 < s < t \le T, \\ y(t) = x, & x \in \mathbb{R}. \end{cases}$$

Il problema (HJB$^{\text{back}}_{\text{ward}}$) e l'equazione (3.7) ricordano, a meno dell'estremo inferiore, un problema di trasporto con sorgente (3.1) e la sua soluzione esplicita (3.3). In tale ottica, le caratteristiche dell'equazione differenziale di (HJB$^{\text{back}}_{\text{ward}}$) sono proprio le traiettorie ottime associate al problema di controllo e lo schema semi-lagrangiano che ne deriva,

$$V(t_{n+1}, x) = \inf_{u \in \mathcal{U}^{\Delta t}_{ad}} \left\{ V(t_n, y^u_x(t_n)) + \int_{t_n}^{t_{n+1}} \ell(y^u_x(s), u(s))\, ds \right\}, \tag{3.8}$$

rappresenta semplicemente il principio della programmazione dinamica (DPP) in forma discreta.
In riferimento all'equazione precedente, $\mathcal{U}^{\Delta t}_{ad} = \{u : [t_n, t_{n+1}] \to U, \text{ misurabile}\}$.

3.3.1 Aspetti implementativi dell'approssimazione DG

Avendo formalizzato e giustificato i passaggi con cui si è giunti all'equazione (3.8), è finalmente possibile addentrarsi nel metodo proposto.
Sia $[a, b]$ il dominio di interesse e si consideri la stessa *mesh* $\mathcal{T}_h$ della sezione 3.2, sia inoltre $V_h = \mathbb{P}_r(\mathcal{T}_h)$, allora nell'intervallo I_k, la formulazione debole descritta nella sottosezione 3.2.1 impone

$$\int_{I_k} V_h^{n+1}(x)\phi_j(x)\,dx = \int_{I_k} \inf_{u\in\mathcal{U}_{ad}^{\Delta t}} \left\{ V_h^n(y_x^u(t_n)) + \int_{t_n}^{t_{n+1}} \ell(y_x^u(s), u(s))\,ds \right\} \phi_j(x)\,dx,$$

$\{\phi_1, \ldots, \phi_{r+1}\}$ rappresenta la base canonica dell'Elemento Finito $(I_k, \mathbb{P}_r(I_k), \mathcal{L}_{Lagr})$. Per calcolare l'integrale a secondo membro si introduce direttamente una formula di quadratura numerica, quindi

$$\begin{aligned}
&\int_{I_k} \inf_{u\in\mathcal{U}_{ad}^{\Delta t}} \left\{ V_h^n(y_x^u(t_n)) + \int_{t_n}^{t_{n+1}} \ell(y_x^u(s), u(s))\,ds \right\} \phi_j(x)\,dx = \\
&\quad = \sum_{q=1}^{N_q} w_q \inf_{u\in\mathcal{U}_{ad}^{\Delta t}} \left\{ V_h^n(y_{x_q}^u(t_n)) + \int_{t_n}^{t_{n+1}} \ell(y_{x_q}^u(s), u(s))\,ds \right\} \phi_j(x_q) = \\
&\quad\quad = \sum_{q=1}^{N_q} w_q \inf_{u\in U} \{ V_h^n(x_q + F(t_{n+1}, x_q, u)\Delta t) + \ell(x_q, u)\Delta t \} \phi_j(x_q),
\end{aligned}$$

dove nell'ultimo passaggio sono state impiegate le approssimazioni dello schema (CIR), cioè

$$\begin{aligned}
x = y_x^u(t_{n+1}) \approx y_x(t_n) - F(t_{n+1}, x, u)\Delta t \implies y_x(t_n) \approx x + F(t_{n+1}, x, u)\Delta t, \\
\int_{t_n}^{t_{n+1}} \ell(y_x^u(s), u(s))\,ds \approx \ell(y_x^u(t_{n+1}), u)\Delta t = \ell(x, u)\Delta t,
\end{aligned}$$

insieme alla semplificazione di un controllo costante a tratti,

$$\mathcal{U}_{ad}^{\Delta t} \approx \{u : [t_n, t_{n+1}] \to U, \text{ costante}\}.$$

Nell'algoritmo proposto, il problema di ottimizzazione viene risolto numericamente scegliendo N_a controlli da U, $\{u_a\}_{a=1}^{N_a}$, e calcolando il minimo tra questi valori, esplicitamente

$$\inf_{u\in U}\{f_{obj}(u)\} \approx \min_{a\in\{1,\ldots,N_a\}}\{f_{obj}(u_a)\} = f_{obj}(u_a^*).$$

Dunque,

$$\int_{I_k} \inf_{u\in\mathcal{U}_{ad}^{\Delta t}} \left\{ V_h^n(y_x^u(t_n)) + \int_{t_n}^{t_n+1} \ell(y_x^u(s),u(s))\,ds \right\} \phi_j(x)\,dx \approx$$

$$\approx \sum_{q=1}^{N_q} w_q \min_{a\in\{1,\dots,N_a\}} \{V_h^n(x_q + F(t_{n+1},x_q,u_a)\Delta t) + \ell(x_q,u_a)\Delta t\}\phi_j(x_q) =$$

$$= \sum_{q=1}^{N_q} w_q \left[\sum_{\ell=1}^{r+1} V_h^n|_{I_{q'}}(\xi_\ell)\phi_\ell(x_q + F(t_{n+1},x_q,u_a^*)\Delta t) + \ell(x_q,u_a^*)\Delta t \right] \phi_j(x_q),$$

nell'ultimo passaggio è stata sottintesa la decomposizione della soluzione V_h^n come combinazione lineare delle $r+1$ funzioni di base dell'intervallo $I_{q'} \in \mathcal{T}_h$, cioè

$$V_h^n(x_q + F(t_{n+1},x_q,u)\Delta t) = \sum_{\ell=1}^{r+1} V_h^n|_{I_{q'}}(\xi_\ell)\phi_\ell(x_q + F(t_{n+1},x_q,u)\Delta t),$$

con $q'(y_{x_q}(t_n)) = \{i \in \{1,\dots,N\} : y_{x_q}(t_n) \in I_i\}$.
Ricapitolando, per risolvere le equazioni di ($\mathrm{HJB}_{\mathrm{ward}}^{\mathrm{back}}$), il metodo DGSL - $\mathbb{P}_r$ proposto ha la forma

$$\mathsf{S_M v^{n+1}} = \mathsf{I_C^n(v^n,u)} \text{ con } \begin{cases} \mathsf{S_M} \in \mathbb{R}^{(r+1)\times(r+1)}, & (\mathsf{S_M})_{jk} = \displaystyle\int_{I_k} \phi_k(x)\phi_j(x)\,dx, \\ \mathsf{v^n} \in \mathbb{R}^{(r+1)\times N}, & \mathsf{v_{jk}^n} = V_h^n|_{I_k}(\xi_j), \\ \mathsf{u} \in \mathbb{R}^{N_a}, & \mathsf{u_a} = u_a, \\ \mathsf{I_C^n(v^n)} \in \mathbb{R}^{(r+1)\times N}, & (\mathsf{I_C^n(v^n,u)})_{jk} = \displaystyle\sum_{q=1}^{N_q} w_q[\dots]\phi_j(x_q), \end{cases}$$

dove

$$[\dots] = \min_{a\in\{1,\dots,N_a\}} \left\{ \sum_{\ell=1}^{r+1} V_h^n|_{I_{q'}}(\xi_\ell)\phi_\ell(x_q + F(t_{n+1},x_q,u_a)\Delta t) + \ell(x_q,u_a)\Delta t \right\}.$$

4 Simulazioni numeriche

Al fine di valutare l'efficacia e l'accuratezza del metodo DGSL - $\mathbb{P}_r$ nella risoluzione delle equazioni di Hamilton-Jacobi-Bellman, sono stati scelti alcuni esempi modello noti dalla letteratura scientifica. Si tratta di problemi specifici in cui la soluzione di viscosità ha un'espressione che si conosce esattamente, ma risulta numericamente difficile da catturare. Infine, si è sperimentato il metodo proposto su un problema di controllo ottimo, precisamente di tempo minimo.

4.1 Impostazione delle simulazioni

Gli esperimenti numerici sono stati condotti attraverso il software MATLAB [32], implementando l'algoritmo DGSL - $\mathbb{P}_r$ per risolvere l'equazione ($\mathrm{HJB}^{\mathrm{back}}_{\mathrm{ward}}$).
In riferimento a tale equazione, che si richiama di seguito specificandone il dominio,

$$\begin{cases} V_t(t,x) - \inf\limits_{u\in U}\{\ell(x,u) + F(t,x,u)V_x(t,x)\} = 0, & (t,x) \in (0,T]\times[a,b], \\ V(0,x) = L(x), & x \in [a,b], \end{cases}$$

il problema viene caratterizzato prescrivendo gli estremi del dominio a e b, il tempo finale T, le funzioni $\ell(x,u)$ e $F(t,x,u)$, la condizione iniziale $L(x)$ e gli estremi dello spazio dei valori del controllo $U = [u_m, u_M]$. Quando le condizioni al bordo non sono periodiche ma di Dirichlet, è necessario specificare anche tale informazione così come le funzioni di bordo $B_a(t)$ e $B_b(t)$.
Per procedere dunque con la simulazione, si indicano

- il numero N di intervalli $\{I_i\}_{i=1}^N$ disgiunti ed equispaziati in cui suddividere il dominio spaziale;
- il numero M di passi temporali uniformi in cui suddividere l'intervallo $(0,T]$;
- il grado massimo r dei polinomi definiti su ogni intervallo della *mesh* per ricostruire la soluzione numerica;

- il numero N_a di controlli equidistanti e compresi tra u_m e u_M inclusi, con cui approssimare lo spazio U.

In ogni test effettuato, si è considerato $r = 1$, per cui i gradi di libertà del generico intervallo $I_i = [x_{i-1}, x_i]$ sono $v_h(x_{i-1})$ e $v_h(x_i)$. Inoltre, se non diversamente specificato, si è posto $N_a = 21$.
Relativamente al calcolo degli integrali richiesti dal metodo DGSL - $\mathbb{P}_r$, si è scelto il metodo di Gauss-Legendre, poiché a parità di numero di nodi, garantisce la maggiore accuratezza e il minor costo computazionale. In particolare, la formula di quadratura considerata ha $N_q = 5$ nodi, quindi esattezza polinomiale fino al grado $2N_q + 1 = 11$. In tabella 4.1 sono riportati i valori dei nodi x_q e dei pesi w_q sull'intervallo di riferimento $[-1, 1]$.

q	1	2	3	4	5
x_q	$-\frac{\sqrt{5+2\sqrt{\frac{10}{7}}}}{3}$	$-\frac{\sqrt{5-2\sqrt{\frac{10}{7}}}}{3}$	0	$\frac{\sqrt{5-2\sqrt{\frac{10}{7}}}}{3}$	$\frac{\sqrt{5+2\sqrt{\frac{10}{7}}}}{3}$
ω_q	$\frac{322-13\sqrt{70}}{900}$	$\frac{322+13\sqrt{70}}{900}$	$\frac{128}{225}$	$\frac{322+13\sqrt{70}}{900}$	$\frac{322-13\sqrt{70}}{900}$

Tabella 4.1: coordinate dei nodi e relativi pesi della quadratura di Gauss-Legendre a 5 punti sull'intervallo di riferimento $[-1, 1]$.

Le prestazioni del metodo numerico proposto sono state valutate analizzando l'errore di discretizzazione (spaziale) e_h, che rappresenta lo scostamento, ad un certo istante temporale, della soluzione approssimata V_h dalla soluzione esatta V. Tale quantità, misurata nelle norme L^2, L^1 e L^∞, si calcola nel modo seguente, rispettivamente

$$\|e_h\|_{L^2} = \|V - V_h\|_{L^2([a,b])} = \sqrt{\int_a^b (V(x) - V_h(x))^2\,dx} = \sqrt{\sum_{i=1}^{N} \int_{I_i} (V(x) - V_h(x))^2\,dx},$$

$$\|e_h\|_{L^1} = \|V - V_h\|_{L^1([a,b])} = \int_a^b |V(x) - V_h(x)|\,dx = \sum_{i=1}^{N} \int_{I_i} |V(x) - V_h(x)|\,dx,$$

$$\|e_h\|_{L^\infty} = \|V - V_h\|_{L^\infty([a,b])} = \sup_{x\in[a,b]} |V(x) - V_h(x)| = \sup_{i\in\{1,\dots,N\}} \sup_{x\in I_i} |V(x) - V_h(x)|.$$

Inoltre, nelle sezioni successive, l'errore commesso dallo schema in oggetto viene confrontato con alcuni metodi di Galerkin discontinuo proposti dalla letteratura, in particolare rispetto al numero di Courant-Friedrichs-Lewy (CFL). Si tratta di un parametro adimensionale che sancisce una condizione necessaria per la stabilità di uno schema esplicito nel contesto dei problemi iperbolici con condizione iniziale [22, Osservazione 13.2], il requisito per le Differenze Finite è

$$\mathrm{CFL} = \frac{\|F\|_{L^\infty}\Delta t}{\Delta x} \leq 1. \tag{4.1}$$

4.2 Test 1 - H quadratico, dato C^∞

Il primo test deriva dall'articolo [8, esempio 4.4],

$$\begin{cases} V_t(t,x) + \dfrac{(V_x(t,x))^2}{2} = 0, & (t,x) \in (0,1.5] \times [0,2\pi], \\ V(0,x) = -\cos(x), & x \in [0,2\pi]. \end{cases} \tag{4.2}$$

Si tratta di un'equazione di Hamilton-Jacobi evolutiva (HJ) con $H(t,x,p) = p^2/2$, condizione iniziale di classe C^∞ e condizioni al bordo periodiche, $V(t,0) = V(t,2\pi)$ per ogni $t \in (0,1.5]$.
Il test (4.2) è equivalente a un problema ($\mathrm{HJB}^{\mathrm{back}}_{\mathrm{ward}}$), infatti siano $\ell(x,u) = u^2/2$, $F(t,x,u) = u$ e $L(x) = -\cos(x)$,

$$\begin{cases} V_t(t,x) - \inf\limits_{u \in U} \left\{ \dfrac{u^2}{2} + uV_x(t,x) \right\} = 0, & (t,x) \in (0,1.5] \times [0,2\pi], \\ V(0,x) = -\cos(x), & x \in [0,2\pi]. \end{cases} \tag{T1}$$

Fissando (t,x), la funzione da ottimizzare è una parabola convessa in u e il dominio U è compatto, pertanto il minimo esiste,

$$f_{obj}(u) = \frac{u^2}{2} + uV_x(t,x), \quad f'_{obj}(u) = u + V_x(t,x) = 0 \implies u^* = -V_x(t,x),$$
$$f_{obj}(u^*) = \frac{(-V_x(t,x))^2}{2} + (-V_x(t,x))V_x(t,x) = -\frac{(V_x(t,x))^2}{2},$$

allora

$$V_t(t,x) - \inf_{u \in U} \left\{ \frac{u^2}{2} + uV_x(t,x) \right\} = V_t(t,x) - f_{obj}(u^*) = V_t(t,x) + \frac{(V_x(t,x))^2}{2}.$$

Tuttavia, affinché l'equivalenza tra le due formulazioni valga, è importante verificare che $V_x(t,x) \in U$ per ogni $(t,x) \in (0,1.5] \times [0,2\pi]$, in questo caso è sufficiente avere $U = [-1,1]$.
Per $t \in (0,1]$, la soluzione di viscosità si ottiene calcolando nell'ordine

$$y + \cos(y)t = x - \frac{\pi}{2}, \quad V(t,x) = \sin(y) + \frac{\cos^2(y)}{2}t,$$

in $t = 1$ si forma una singolarità nella derivata a causa di un'onda di shock[1], che rende la soluzione esatta semplicemente continua.
In figura 4.1 si riporta il confronto, in $t = 1$, tra la soluzione esatta (in nero) e la sua approssimazione (in blu) ottenuta con il metodo DGSL - $\mathbb{P}_1$. La *mesh* considerata è composta da $N = 320$ intervalli disgiunti ed equispaziati e si sono eseguiti $M = 13$ passi temporali uniformi, dunque CFL ≈ 4.

[1] Si veda la nota 2 al fondo di pagina 36.

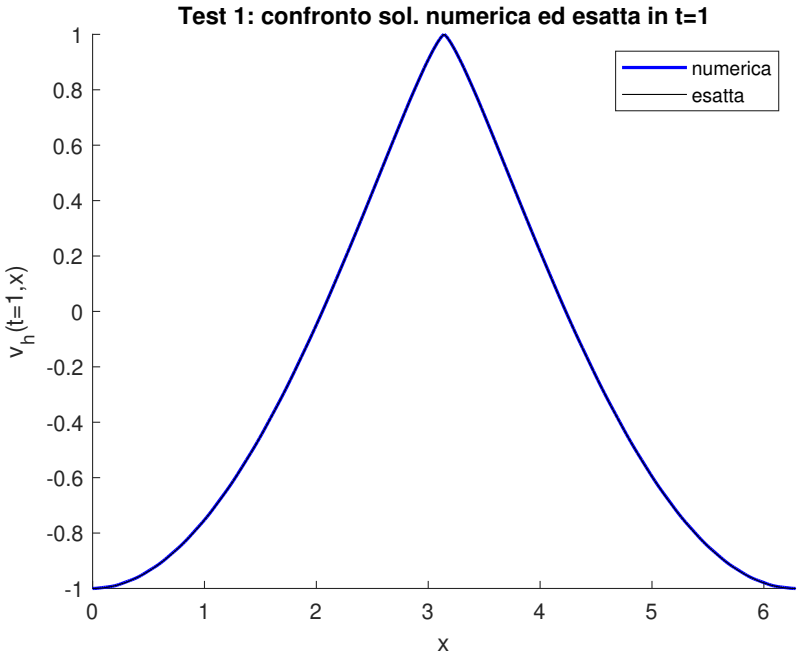

Figura 4.1: Rappresentazione della soluzione numerica DGSL - $\mathbb{P}_1$ ($N = 320$, $M = 13$) ed esatta del problema (T1).

Graficamente, lo schema proposto sembra aver risolto correttamente il problema, nonché il punto angoloso in $x = \pi$.
In tabella 4.2 è riportato l'errore $e_h = V - V_h$ commesso in $t = 1$, con CFL ≈ 0.45 e in tutto il dominio spaziale escluso l'aperto $(3, 2\pi - 3)$, così da confrontare i risultati con [8, tabella 4], a parità di raffinamento spaziale e temporale.

N	10	20	40	80	160	ordine
$\|\|e_h\|\|_{L^2}$	4.7013e-02	2.0098e-02	9.2061e-03	4.1933e-03	2.3776e-03	1.0872
$\|\|e_h\|\|_{L^1}$	8.8735e-02	3.2645e-02	1.4903e-02	6.3181e-03	3.8578e-03	1.1417

Tabella 4.2: Errore commesso nella regione $[0, 2\pi] \setminus (3, 2\pi - 3)$ dal metodo DGSL - $\mathbb{P}_1$ per il test (T1) in $t = 1$, ponendo CFL ≈ 0.45.

Lo schema proposto risulta meno accurato del metodo *Central Discontinuous Galerkin* sviluppato dagli autori Li e Yakovlev [8] e mostra una velocità di convergenza dell'errore inferiore in entrambe le norme considerate, L^2 e L^1. Tuttavia i metodi semi-lagrangiani, tra cui quello proposto, sono noti per essere stabili indipendentemente dalla condizione (4.1). Infatti con $N = 160$ e CFL ≈ 2, *Central Discontinuous Galerkin* risulta completamente instabile e la soluzione calcolata diverge, mentre DGSL - $\mathbb{P}_1$ converge e i dati sull'errore e_h sono raccolti in tabella 4.3, dove inoltre

si è imposto $N_a = 41$.

N	10	20	40	80	160	ordine
$\|\|e_h\|\|_{L^2}$	5.4092e-02	1.2419e-02	3.2347e-03	9.6385e-04	4.2388e-04	1.7679
$\|\|e_h\|\|_{L^1}$	1.0369e-01	2.4048e-02	6.5974e-03	1.9384e-03	8.6166e-04	1.7455

Tabella 4.3: Errore commesso nella regione $[0, 2\pi] \setminus (3, 2\pi - 3)$ dal metodo DGSL - $\mathbb{P}_1$ per il test (T1) in $t = 1$, con CFL ≈ 2 e $N_a = 41$.

Nelle figure 4.2 e 4.3 vengono sinteticamente rappresentati, all'interno dello stesso grafico, i dati dello schema proposto a confronto con il metodo di Li e Yakovlev, sull'errore commesso nel risolvere il problema (4.2), rispettivamente, in norma L^2 e L^1. In particolare, si evince che, regolando opportunamente il valore CFL e aumentando il numero N_a di controlli da valutare, il metodo DGSL - $\mathbb{P}_1$ diventa paragonabile al *Central Discontinuous Galerkin* sia in termini di accuratezza che di velocità di convergenza.

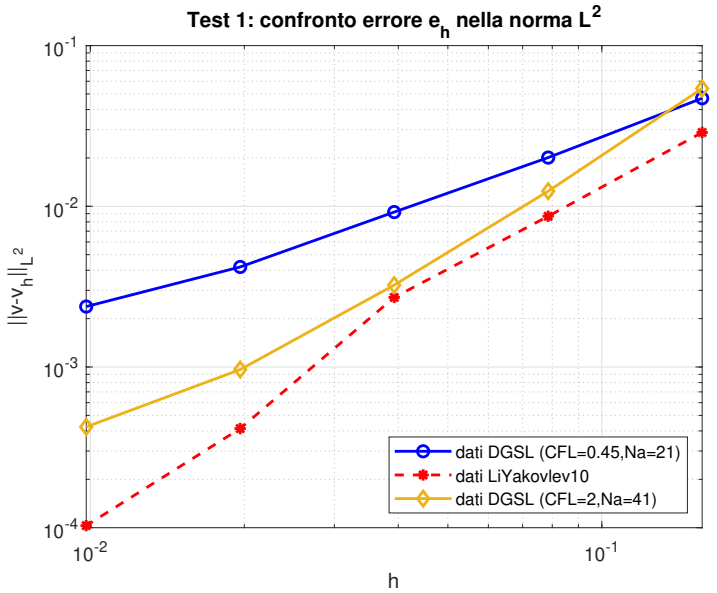

Figura 4.2: Andamento dell'errore in norma L^2 al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T1) a confronto con i risultati dell'articolo [8, tabella 4] (in rosso).

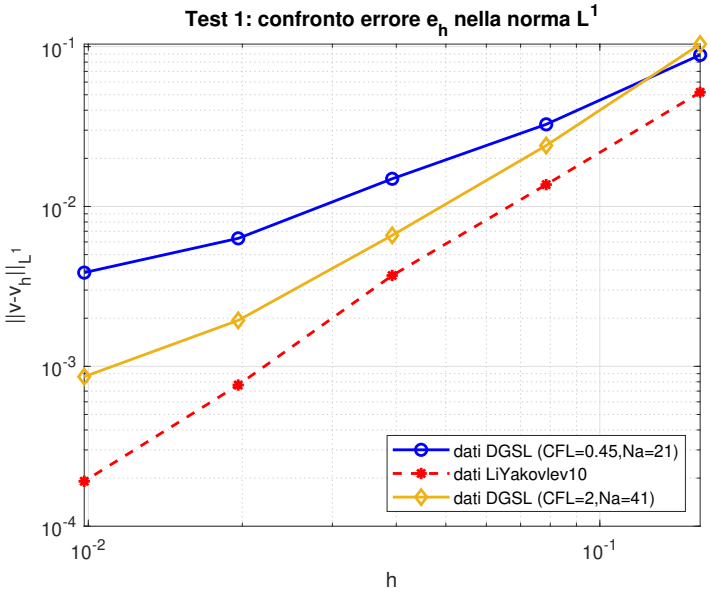

Figura 4.3: Andamento dell'errore in norma L^1 al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T1) a confronto con i risultati dell'articolo [8, tabella 4] (in rosso).

4.3 Test 2 - H lineare non derivabile, dato C^∞

Il secondo test deriva dall'articolo [7, esempio 4.2.1],

$$\begin{cases} V_t(t,x) + \text{sign}(\cos(x))V_x(t,x) = 0, & (t,x) \in (0,1] \times [0,2\pi], \\ V(0,x) = \sin(x), & x \in [0,2\pi]. \end{cases} \tag{4.3}$$

Si tratta di un'equazione di Hamilton-Jacobi evolutiva con $H(t,x,p) = \text{sign}(\cos(x))p$, quindi lineare in p ma con un coefficiente che presenta delle discontinuità di tipo salto in $x = \pi/2$ e $x = 3\pi/2$ che causano, rispettivamente, la formazione di un'onda di shock e un'onda di rarefazione[2] per V_x. La condizione iniziale è di classe C^∞ e si hanno condizioni al bordo periodiche, $V(t,0) = V(t,2\pi)$ per ogni $t \in (0,1]$.
Come viene riportato nell'articolo [11, esempio 4.4], la soluzione di tale problema coincide con quella dell'equazione eikonale evolutiva e omogenea, cioè

$$\begin{cases} V_t(t,x) + |V_x(t,x)| = 0, & (t,x) \in (0,1] \times [0,2\pi], \\ V(0,x) = \sin(x), & x \in [0,2\pi]. \end{cases}$$

[2]Si veda la nota 2 al fondo di pagina 36.

Inoltre, tale formulazione è equivalente ad un problema ($\mathrm{HJB}^{\mathrm{back}}_{\mathrm{ward}}$), infatti siano $\ell(x,u)=0$, $F(t,x,u)=u$, $U=[-1,1]$ e $L(x)=\sin(x)$,

$$\begin{cases} V_t(t,x) - \inf\limits_{u\in U}\{uV_x(t,x)\} = 0, & (t,x)\in(0,1]\times[0,2\pi], \\ V(0,x) = \sin(x), & x\in[0,2\pi]. \end{cases} \tag{T2}$$

Fissando (t,x), la funzione da ottimizzare è una retta in u e il dominio U è compatto, pertanto il minimo esiste. Sia $f_{obj}(u)=uV_x(t,x)$, allora si presentano due casi,

$$\begin{cases} V_x(t,x) > 0 \implies u^* = -1 \implies f_{obj}(u^*) = -V_x(t,x), \\ V_x(t,x) \le 0 \implies u^* = 1 \implies f_{obj}(u^*) = V_x(t,x), \end{cases}$$

sinteticamente

$$u^* = -\operatorname{sign}(V_x(t,x)) \implies f_{obj}(u^*) = -|V_x(t,x)|,$$

allora

$$V_t(t,x) - \inf_{u\in U}\{uV_x(t,x)\} = V_t(t,x) - f_{obj}(u^*) = V_t(t,x) + |V_x(t,x)|.$$

La soluzione di viscosità del problema in esame, per $t\in(0,\pi/2]$, è

$$V(t,x) = \begin{cases} \sin(x-t), & x\in[0,\pi/2], \\ \sin(x+t), & x\in(\pi/2,3\pi/2-t], \\ -1, & x\in(3\pi/2-t,3\pi/2+t], \\ \sin(x-t), & x\in(3\pi/2+t,2\pi]. \end{cases}$$

In figura 4.4 si riporta il confronto, in $t=1$, tra la soluzione esatta (in nero) e la sua approssimazione (in blu) calcolata con il metodo DGSL - $\mathbb{P}_1$. La *mesh* considerata è composta da $N=640$ intervalli disgiunti ed equispaziati e si sono eseguiti $M=25$ passi temporali uniformi, dunque CFL ≈ 4.
Graficamente, l'approssimazione fornita dallo schema proposto per la risoluzione del problema appare corretta, in particolare nell'intorno del punto angoloso $x=\pi/2$ e nella regione di rarefazione.
In tabella 4.4 è riportato l'errore $e_h = V - V_h$ commesso al tempo $t=1$ e considerando CFL ≈ 0.1, così da confrontare i risultati con [7, tabella 4.7], a parità di raffinamento spaziale e temporale.
Lo schema proposto risulta meno accurato del metodo sviluppato da Cheng e Shu in [7] e l'ordine di convergenza dell'errore è leggermente inferiore in tutte le norme considerate, L^2, L^1 e L^∞. Tuttavia, si precisa che, in questo caso, l'approccio sviluppato da Cheng e Shu richiede una procedura aggiuntiva per correggere e reindirizzare la soluzione numerica verso quella di viscosità. Inoltre, il parametro CFL gioca un ruolo fondamentale nella stabilità di tale metodo numerico e lo si può verificare, ad

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	7.0108e-02	2.2709e-02	6.3778e-03	1.7975e-03	5.7754e-04	1.7506
$\|\|e_h\|\|_{L^1}$	1.1026e-01	3.7218e-02	1.1577e-02	3.7071e-03	1.3023e-03	1.6135
$\|\|e_h\|\|_{L^\infty}$	6.1141e-02	1.8322e-02	4.6927e-03	1.1741e-03	3.2572e-04	1.9069

Tabella 4.4: Errore commesso dal metodo DGSL - $\mathbb{P}_1$ per il test (T2) in $t = 1$, ponendo CFL ≈ 0.1.

esempio, provando a risolvere il problema (4.3) imponendo $N = 160$ e CFL > 0.6, la soluzione che si ottiene è fortemente divergente.
Al contrario, il parametro CFL non limita la stabilità del metodo proposto e può migliorarne l'accuratezza, infatti in tabella 4.5, si riportano i dati relativi all'approssimazione ottenuta dal metodo DGSL - $\mathbb{P}_1$ con CFL ≈ 3.

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	3.0542e-03	6.8869e-04	2.1180e-04	4.6041e-05	1.1562e-05	1.9993
$\|\|e_h\|\|_{L^1}$	5.7478e-03	1.4008e-03	4.2445e-04	8.1865e-05	2.0823e-05	2.0314
$\|\|e_h\|\|_{L^\infty}$	2.9984e-03	6.3493e-04	2.3883e-04	4.7233e-05	1.1253e-05	1.9864

Tabella 4.5: Errore commesso dal metodo DGSL - $\mathbb{P}_1$ per il test (T2) in $t = 1$, ponendo CFL ≈ 3.

Nelle figure 4.5, 4.6 e 4.7 vengono sinteticamente rappresentati, all'interno dello stesso grafico, i dati a confronto tra lo schema proposto e il metodo di Cheng e Shu, sull'errore commesso nel risolvere l'esempio in esame, rispettivamente, in norma L^2, L^1 e L^∞. Dalle immagini, si evince che, regolando opportunamente il valore CFL, lo schema DGSL - $\mathbb{P}_1$ diventa paragonabile al metodo di Cheng e Shu sia in termini di accuratezza che di velocità di convergenza, o leggermente superiore in riferimento alla norma L^∞ dell'errore.

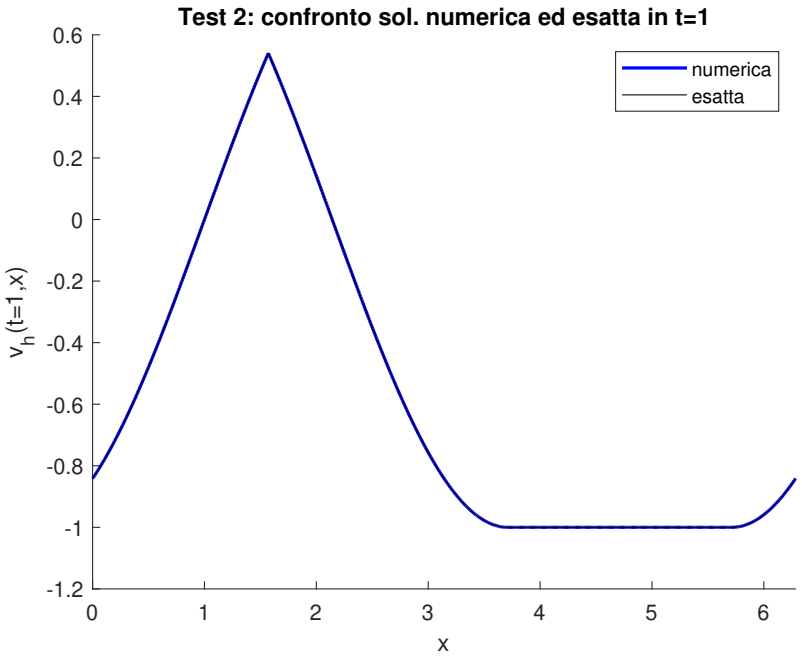

Figura 4.4: Rappresentazione della soluzione numerica DGSL - $\mathbb{P}_1$ ($N = 640$, $M = 25$) ed esatta del problema (T2).

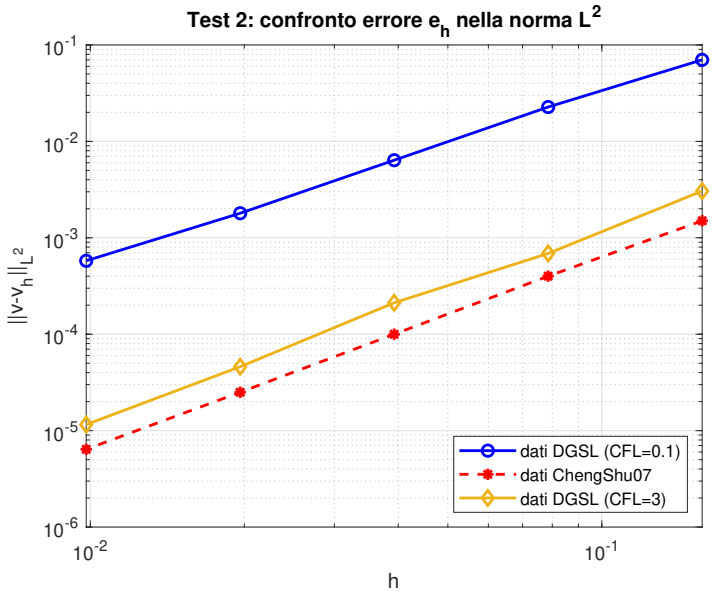

Figura 4.5: Andamento dell'errore in norma L^2 al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T2) a confronto con i risultati dell'articolo [7, tabella 4.7] (in rosso).

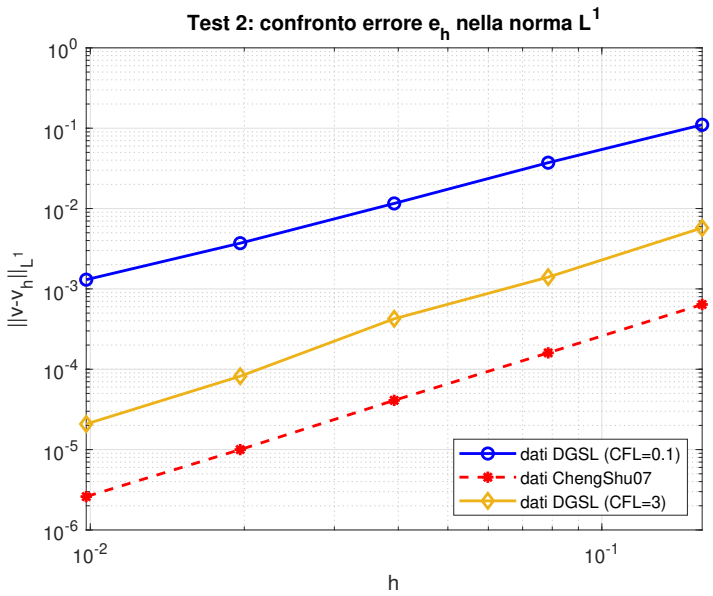

Figura 4.6: Andamento dell'errore in norma L^1 al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T2) a confronto con i risultati dell'articolo [7, tabella 4.7] (in rosso).

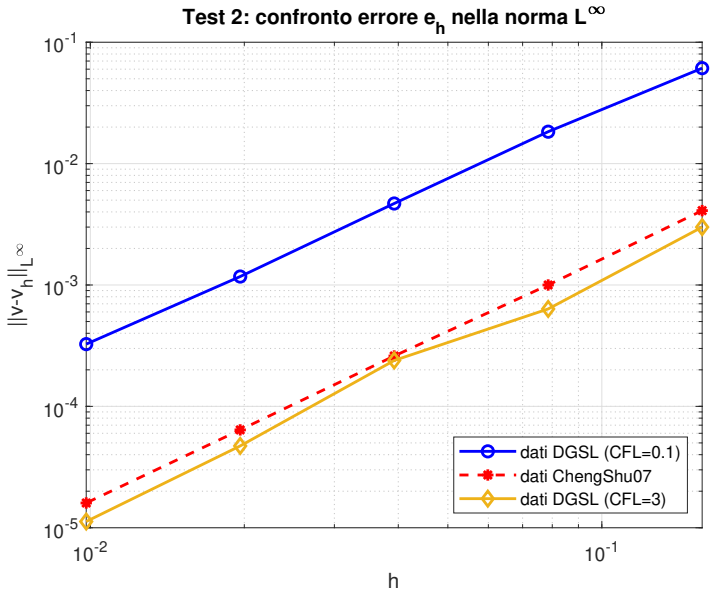

Figura 4.7: Andamento dell'errore in norma L^∞ al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T2) a confronto con i risultati dell'articolo [7, tabella 4.7] (in rosso).

4.4 Test 3 - H quadratico, dato non derivabile

Il terzo test deriva dall'articolo [11, esempio 4.3],

$$\begin{cases} V_t(t,x) + \dfrac{(V_x(t,x))^2}{2} = 0, & (t,x) \in (0,1] \times [0,2\pi], \\ V(0,x) = |x-\pi|, & x \in [0,2\pi]. \end{cases} \tag{4.4}$$

Si tratta di un'equazione di Hamilton-Jacobi evolutiva (HJ) con $H(t,x,p) = p^2/2$, condizione iniziale continua, non derivabile in $x = \pi$ e che causa la formazione di un'onda di rarefazione[3] per V_x. Si hanno inoltre condizioni al bordo periodiche, $V(t,0) = V(t,2\pi)$ per ogni $t \in (0,1]$.
Come visto per il primo test (sezione 4.2), tale problema può essere riscritto nella forma ($\text{HJB}^{\text{back}}_{\text{ward}}$), considerando $\ell(x,u) = u^2/2$, $F(t,x,u) = u$, $L(x) = |x-\pi|$ e $U = [-1,1]$,

$$\begin{cases} V_t(t,x) - \inf\limits_{u \in U} \left\{ \dfrac{u^2}{2} + uV_x(t,x) \right\} = 0, & (t,x) \in (0,1] \times [0,2\pi], \\ V(0,x) = |x-\pi|, & x \in [0,2\pi]. \end{cases} \tag{T3}$$

La soluzione di viscosità del problema in esame, per $t \in (0,\pi]$, è

$$V(t,x) = \begin{cases} \pi - x - \dfrac{t}{2}, & x \in [0, \pi - t], \\ x - \pi - \dfrac{t}{2}, & x \in [\pi + t, 2\pi], \\ \dfrac{(x-\pi)^2}{2t}, & x \in (\pi - t, \pi + t). \end{cases}$$

In figura 4.8 si riporta il confronto, in $t = 1$, tra la soluzione esatta (in nero) e la sua approssimazione (in blu) calcolata con il metodo DGSL - $\mathbb{P}_1$. La *mesh* considerata è composta da $N = 160$ intervalli disgiunti ed equispaziati e si sono eseguiti $M = 25$ passi temporali uniformi, dunque CFL ≈ 1.
Graficamente, l'approssimazione ottenuta attraverso il metodo proposto appare corretta, in particolare nell'intorno del punto $x = \pi$ in cui si ha la formazione della regione di rarefazione.
In tabella 4.6 è riportato l'errore $e_h = V - V_h$ commesso in $t = 1$, con CFL ≈ 0.1, così da confrontare i risultati con [11, tabella 4, $\mathbb{P}_1$], a parità di raffinamento spaziale e temporale.
Lo schema proposto risulta, ancora una volta, meno accurato della formulazione alternativa di Galerkin discontinuo sviluppata da Ke e Guo in [11] e mostra una velocità di convergenza dell'errore inferiore, sia in norma L^2 che L^∞. Tuttavia, il metodo DGSL - $\mathbb{P}_1$, in qualità di schema semi-lagrangiano, è in grado di risolvere il

[3] Si veda la nota 2 al fondo di pagina 36.

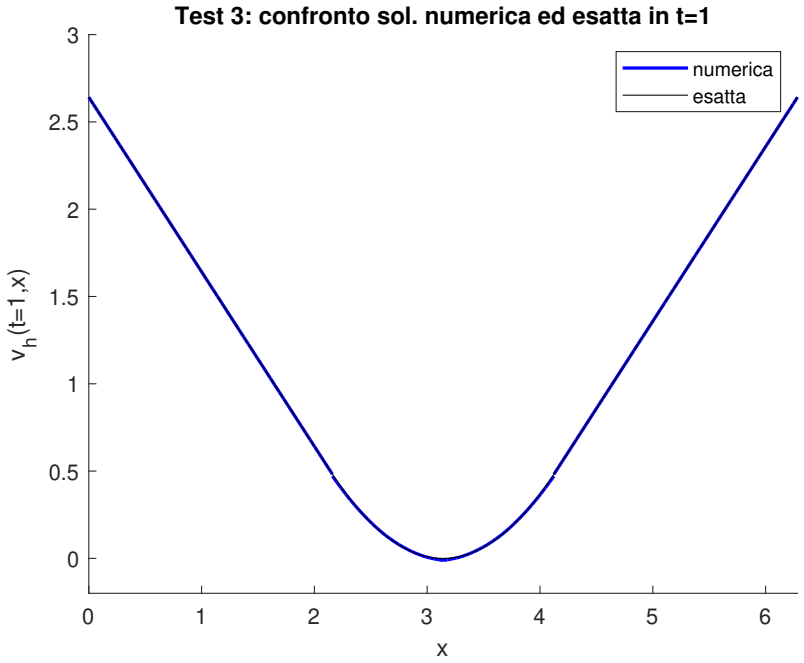

Figura 4.8: Rappresentazione della soluzione numerica DGSL - $\mathbb{P}_1$ ($N = 160$, $M = 25$) ed esatta del problema (T3).

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	1.6680e-01	1.3354e-01	1.0259e-01	7.7996e-02	6.0952e-02	0.3680
$\|\|e_h\|\|_{L^\infty}$	2.7742e-01	2.2845e-01	1.8333e-01	1.4767e-01	1.2452e-01	0.2941

Tabella 4.6: Errore commesso dal metodo DGSL - $\mathbb{P}_1$ per il test (T3) in $t = 1$, ponendo CFL ≈ 0.1.

problema in esame considerando valori CFL che violano la condizione (4.1). Infatti in tabella 4.7, si riportano i dati relativi all'approssimazione del problema (T3) ottenuta con CFL ≈ 4.
Nelle figure 4.9 e 4.10 vengono sinteticamente rappresentati, all'interno dello stesso grafico, il confronto tra i dati dello schema proposto e il metodo di Ke e Guo, sull'errore commesso nel risolvere l'esempio in esame, rispettivamente, in norma L^2 e L^∞. Dalle immagini, si evince che, la sola regolazione opportuna del valore CFL produce un miglioramento delle prestazioni dello schema DGSL - $\mathbb{P}_1$, grazie al quale supera il metodo di Ke e Guo [11] in termini di accuratezza e di velocità di convergenza dell'errore.

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	1.6487e-02	9.4490e-03	5.0044e-03	2.5007e-03	1.8293e-03	0.8262
$\|\|e_h\|\|_{L^\infty}$	1.7727e-02	1.0651e-02	5.8875e-03	3.0801e-03	2.1743e-03	0.7845

Tabella 4.7: Errore commesso dal metodo DGSL - $\mathbb{P}_1$ per il test (T3) in $t = 1$, ponendo CFL ≈ 4.

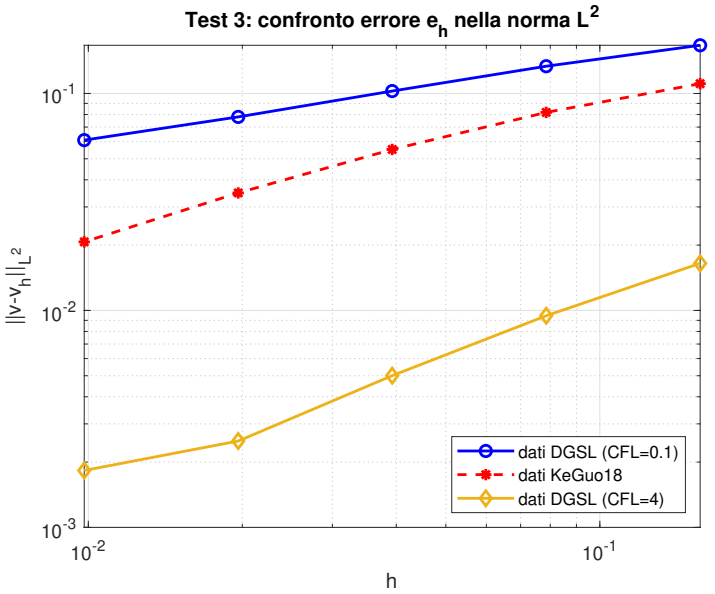

Figura 4.9: Andamento dell'errore in norma L^2 al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T3) a confronto con i risultati dell'articolo [11, tabella 4, $\mathbb{P}_1$] (in rosso).

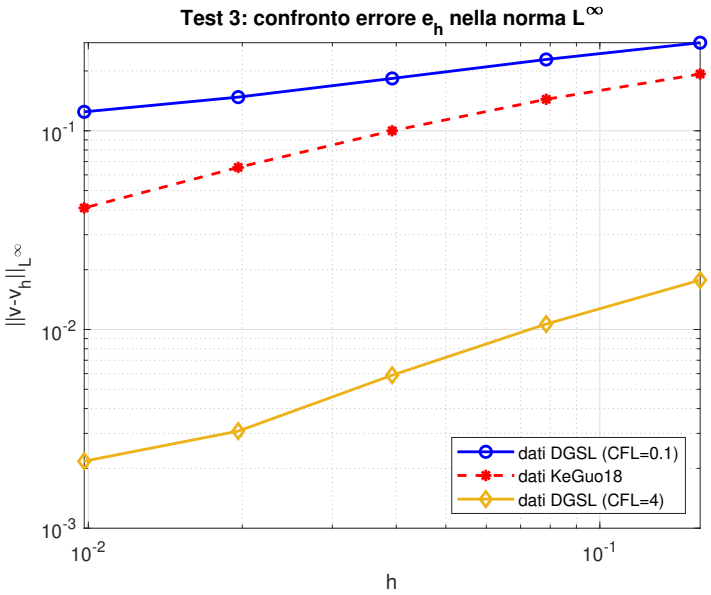

Figura 4.10: Andamento dell'errore in norma L^∞ al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T3) a confronto con i risultati dell'articolo [11, tabella 4, $\mathbb{P}_1$] (in rosso).

4.5 Test 4 - Tempo minimo di evasione

Infine, per il quarto test, si è considerato un problema di controllo ottimo, la cui dinamica è governata dall'equazione

$$\begin{cases} \dot{y}(s) = u(s), & 0 \le t < s \le 4, \\ y(t) = x, & x \in [0, 10], \end{cases}$$

$y : [t, 4] \to [0, 10]$ rappresenta la funzione dello stato del sistema e la funzione di controllo è denotata con $u \in \mathcal{U}_{ad} = \{u : [t, 4] \to [-1, 1], \text{ misurabile}\}$. Il funzionale costo da minimizzare è

$$J_{t,x}(u) = \int_t^{t_f} 1 \, ds + (10 - y_x^u(t_f)) y_x^u(t_f), \tag{4.5}$$

y_x^u è la traiettoria derivante dal controllo $u \in \mathcal{U}_{ad}$ con condizione iniziale x al tempo t, $t_f = \min(4, \theta_{t,x}(u))$ e $\theta_{t,x}(u) = \inf\{s \in (t, 4] : y_x^u(s) = 0 \vee y_x^u(s) = 10\}$ rappresentano, rispettivamente, l'orizzonte finito e il tempo di raggiungimento del bordo del dominio spaziale. Differisce dalla formulazione di Bolza della sezione 1.1 esclusivamente per l'orizzonte finito, che in questo test non è fissato dipendendo dallo stato del sistema e potrebbe essere minore di $T = 4$.
Il funzionale (4.5) è descritto dalla somma di due contributi, il primo è il tempo

impiegato prima di raggiungere uno degli estremi del dominio spaziale nella finestra temporale $[t, 4]$, cioè $t_f - t$, il secondo è un costo associato alla distanza tra il bordo spaziale $\{0, 10\}$ e la posizione raggiunta dal sistema al tempo t_f.
Nella sezione 1.1, si è introdotta la nozione di *value function*

$$v(t,x) = \inf_{u \in \mathcal{U}_{ad}} J_{t,x}(u),$$

che, in virtù del Teorema 3.2 e della teoria delle soluzioni di viscosità, risolve

$$\begin{cases} v_t(t,x) = \sup_{u \in U} \{-1 - u v_x(t,x)\}, & (t,x) \in [0,t_f) \times [0,10], \\ v(t_f,x) = (10-x)x, & x \in [0,10]. \end{cases}$$

con $U = [-1,1]$. In particolare, quando $x = 0$ o $x = 10$ si ha $t_f = \min(4,t) = t$ per ogni $t \in [0,4)$, per cui $v(t,0) = v(t,10) = 0$. Pertanto, in virtù del principio della programmazione dinamica (DPP), si ottiene la formulazione equivalente e classica di un problema differenziale con condizione iniziale e bordo di Dirichlet non omogeneo,

$$\begin{cases} v_t(t,x) = \sup_{u \in U} \{-1 - u v_x(t,x)\}, & (t,x) \in [0,4) \times [0,10], \\ v(T,x) = (10-x)x, & x \in [0,10], \\ v(t,0) = v(t,10) = 0, & t \in [0,4). \end{cases}$$

Come visto nella dimostrazione del Teorema 3.2, dal cambio di variabile $t \to T - t$, si ricava

$$\begin{cases} V_t(t,x) - \inf_{u \in U} \{1 + u V_x(t,x)\} = 0, & (t,x) \in (0,4] \times [0,10], \\ V(0,x) = (10-x)x, & x \in [0,10], \\ V(t,0) = V(t,10) = 0, & t \in (0,4], \end{cases} \tag{T4}$$

che rappresenta un problema ($\mathrm{HJB}^{\mathrm{back}}_{\mathrm{ward}}$) caratterizzato da $\ell(x,u) = 1$, $F(t,x,u) = u$, $L(x) = (10-x)x$, $U = [-1,1]$ e $B_a(t) = B_b(t) \equiv 0$.
La soluzione di viscosità del problema in esame, per $t \in (0,4]$ e $x \in [0,5]$, è

$$V(t,x) = \begin{cases} t, & x \in [0,t], \\ t + (x-t)(10-x+t), & x \in (t,5] \end{cases}$$

ed è simmetrica rispetto all'asse $x = 5$, infatti si ha $V(t,x) = V(t,10-x)$ per $x \in (5,10]$. L'espressione esplicita di $V(t,x)$ è stata ricavata considerando che si vuole minimizzare il tempo di evasione dal dominio spaziale, poiché se non si raggiunge uno dei bordi si paga un costo terminale positivo oltre al tempo impiegato. Quindi è necessario orientare lo stato del sistema verso il bordo più vicino e alla massima velocità, cioè se $x \gtrless 5$ allora $u^* = \pm 1$, in $x = 5$ entrambe le scelte sono ottime. Conoscendo il controllo ottimo u^*, si calcola la *value function* che è legata

a V dalla relazione $v(T-t,x) = V(t,x)$.
In figura 4.11 si riporta il confronto, in $t = 4$, tra la soluzione esatta (in nero) del problema (T4) e la sua approssimazione (in blu) ottenuta dal risolutore DGSL - $\mathbb{P}_1$. La *mesh* considerata è composta da $N = 160$ intervalli disgiunti ed equispaziati e si sono eseguiti $M = 64$ passi temporali uniformi, dunque CFL $= 1$.

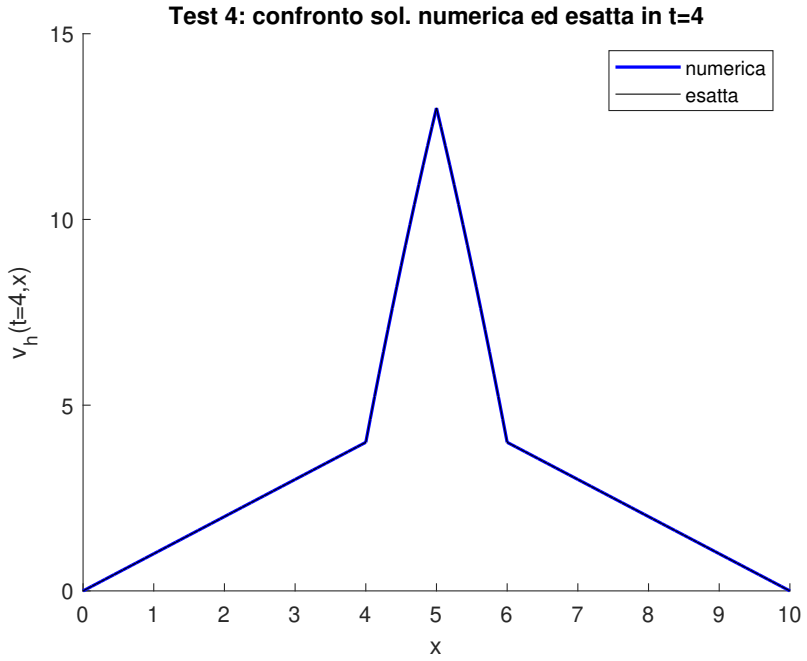

Figura 4.11: Rappresentazione della soluzione numerica DGSL - $\mathbb{P}_1$ ($N = 160$, $M = 64$) ed esatta del problema (T4).

Graficamente, l'approssimazione calcolata dal metodo proposto appare corretta, in particolare nell'intorno dei punti $x = t = 4$, $x = 10 - t = 6$ in cui si hanno delle discontinuità nella derivata causate dalla non linearità del funzionale (4.5). Anche in $x = 5$ la *value function* non è differenziabile poiché la dinamica, attraverso il controllo, non è continua nell'intorno di tale punto, tuttavia l'approssimazione appare senza difetti.
Sulla scia dei test precedentemente effettuati, si è calcolato l'errore $e_h = V - V_h$ commesso in $t = 4$, con CFL $= 0.5$, dallo schema proposto e da un metodo di Galerkin discontinuo (DG) tradizionale con Elemento Finito $(I, \mathbb{P}_1(I), \mathcal{L}_{Lagr})$ per tutti gli intervalli della *mesh* e avanzamento in tempo di Crank-Nicolson[4] (CN).
Infatti, in riferimento alla sezione 2.2, l'equazione iperbolica lineare non stazionaria caratterizzata da $\beta(x) = -\operatorname{sign}(x-5)$, $\sigma \equiv 0$, $f \equiv 1$ e $g_{in} \equiv 0$ coincide con l'equazione alle derivate parziali di (T4), una volta calcolato il controllo ottimo[5].

[4]Sia $\dot{y}(t) = f(t,y)$, allora $(y^{n+1} - y^n)/\Delta t = [f(t^n, y^n) + f(t^{n+1}, y^{n+1})]/2$ è lo schema di Crank-Nicolson che calcola $y^n = y(t^n)$ con $t^n = n\Delta t$, sapendo che $y^0 = y(0)$, l'accuratezza è $\mathcal{O}(\Delta t^2)$.

[5]I passaggi sono identici a quelli visti per il secondo test, sezione 4.3.

I dati sono riportati, rispettivamente, nelle tabelle 4.8 e 4.9.

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	5.0465e-01	2.8166e-01	1.7012e-01	1.1125e-01	7.9707e-02	0.6665
$\|\|e_h\|\|_{L^\infty}$	6.0525e-01	4.1332e-01	3.3681e-01	2.6289e-01	2.1548e-01	0.3633

Tabella 4.8: Errore commesso dal metodo DGSL - $\mathbb{P}_1$ per il test (T4) in $t = 4$, ponendo CFL $= 0.5$.

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	3.7715e-01	1.9267e-01	9.8753e-02	5.0724e-02	2.6073e-02	0.9634
$\|\|e_h\|\|_{L^\infty}$	5.6317e-01	3.5536e-01	2.2379e-01	1.4107e-01	8.9045e-02	0.6655

Tabella 4.9: Errore commesso dal metodo DG ($\mathbb{P}_1$, $\mathcal{L}_{Lagr}$) e avanzamento in tempo CN per il test (T4) in $t = 4$, ponendo CFL $= 0.5$.

A parità di raffinamento spaziale e temporale, lo schema proposto risulta meno preciso del metodo di Galerkin discontinuo tradizionale e l'ordine di convergenza dell'errore è inferiore in tutte le norme considerate, L^2 e L^∞. Tuttavia, per valori CFL che violano la condizione (4.1), DG tradizionale presenta delle oscillazioni spurie che deteriorano la soluzione numerica peggiorando l'errore e_h. Al contrario, il metodo DGSL - $\mathbb{P}_1$ non presenta questa irregolarità e anzi, con CFL $= 2$, i dati riportati in tabella 4.10 sull'errore commesso sono notevolmente inferiori.

N	40	80	160	320	640	ordine
$\|\|e_h\|\|_{L^2}$	1.6137e-02	4.0344e-03	1.0086e-03	2.5215e-04	6.3037e-05	2.0000
$\|\|e_h\|\|_{L^\infty}$	1.5625e-02	3.9063e-03	9.7656e-04	2.4414e-04	6.1035e-05	2.0000

Tabella 4.10: Errore commesso dal metodo DGSL - $\mathbb{P}_1$ per il test (T4) in $t = 4$, ponendo CFL $= 2$.

Nelle figure 4.12 e 4.13 vengono sinteticamente rappresentati, all'interno dello stesso grafico, il confronto tra i dati dello schema proposto e un metodo di Galerkin discontinuo tradizionale, sull'errore commesso nel risolvere il problema di controllo ottimo del tempo minimo di evasione, rispettivamente, in norma L^2 e L^∞.

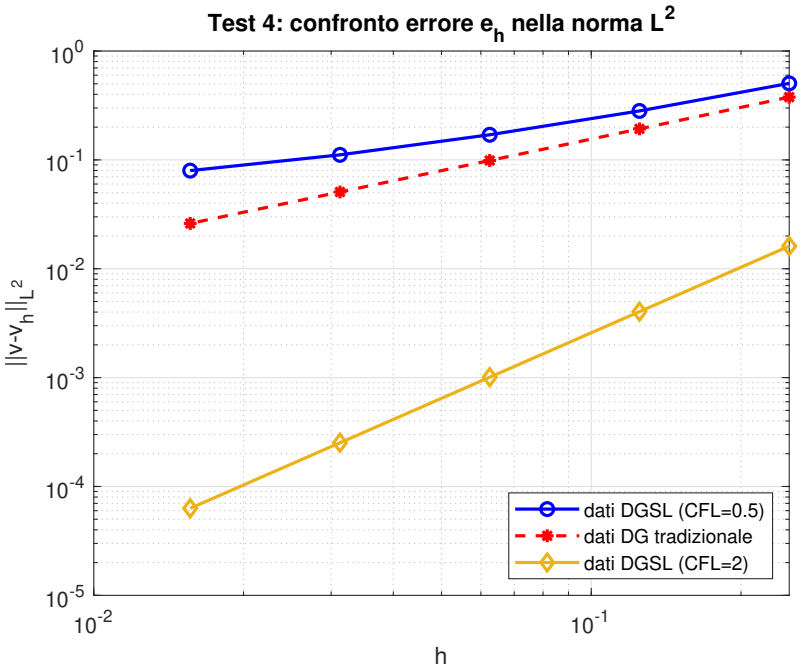

Figura 4.12: Andamento dell'errore in norma L^2 al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T4) a confronto con i risultati dell'approssimazione DG tradizionale (in rosso).

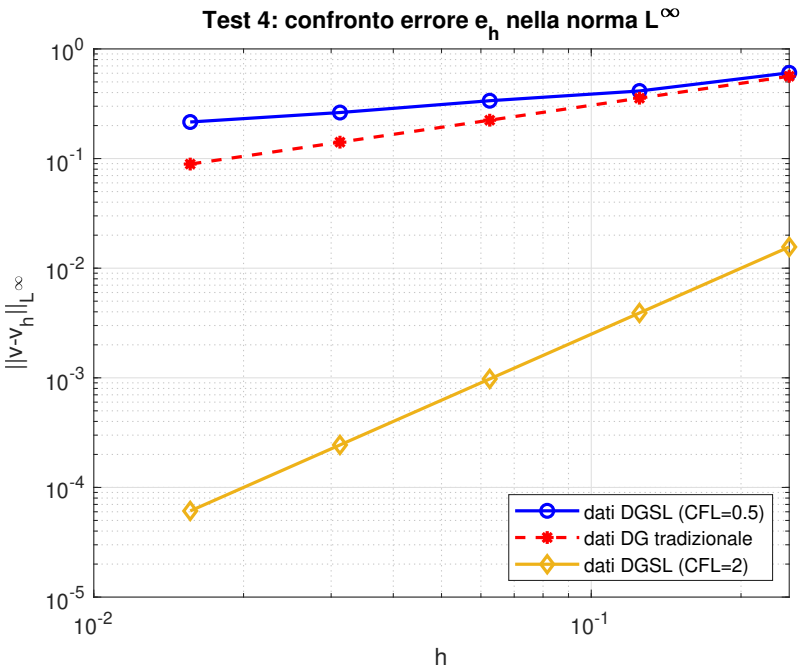

Figura 4.13: Andamento dell'errore in norma L^∞ al crescere di h per la soluzione numerica DGSL - $\mathbb{P}_1$ (in blu e in giallo) del problema (T4) a confronto con i risultati dell'approssimazione DG tradizionale (in rosso).

Dalle immagini, si evince che il metodo DGSL - $\mathbb{P}_1$ si comporta come un metodo di Galerkin discontinuo tradizionale con funzioni di base $\mathbb{P}_1$ e avanzamento in tempo di Crank-Nicolson per bassi valori del parametro CFL. Inoltre, le prestazioni dello schema proposto, al contrario del metodo DG tradizionale, migliorano considerevolmente per valori CFL che violano la condizione (4.1).

5 Conclusioni e sviluppi futuri

Nel presente lavoro di tesi, è stato proposto un metodo numerico, DGSL - $\mathbb{P}_r$, per risolvere equazioni di Hamilton-Jacobi-Bellman evolutive in una dimensione.
Alla luce dei testi effettuati, l'approccio basato sulla formula di rappresentazione semi-lagrangiana e sulla ricostruzione della funzione con Elementi Finiti discontinui è risultato efficace nella cattura della ricca struttura (generata dalle discontinuità nella derivata e dai dati non derivabili ovunque) delle soluzioni di viscosità.
In riferimento agli obiettivi prefissati nell'Introduzione, con gli esperimenti effettuati emerge la stabilità del metodo DGSL - $\mathbb{P}_r$ indipendentemente dal valore del passo temporale e spaziale scelto. Inoltre, come auspicato, la struttura dello schema proposto garantisce sia l'adattività *hp* che il disaccoppiamento del sistema lineare di equazioni in parti più piccole, risolvibili separatamente.
Un aspetto peculiare, che si è delineato con gli esperimenti numerici, riguarda l'accuratezza e l'ordine di convergenza. Infatti, dal confronto con altri metodi noti dalla letteratura, lo schema proposto è risultato meno efficace per valori CFL inferiori all'unità, diventando invece paragonabile o anche superiore per CFL più grandi.

A partire dal lavoro discusso in questa tesi, si aprono diversi percorsi di ricerca futura. In primo luogo, si potrebbe estendere lo studio al caso bidimensionale (o multidimensionale) ponendo attenzione alla ricostruzione della caratteristica nella fase semi-lagrangiana del metodo. Inoltre, per ridurre i tempi computazionali, può essere ragionevole implementare per ogni elemento le funzioni di basi di Legendre, non essendo richiesta la continuità all'interfaccia tra gli elementi.
In aggiunta, il metodo DGSL - $\mathbb{P}_r$ potrebbe essere indirizzato verso applicazioni, sotto le opportune correzioni, che coinvolgono soluzioni di viscosità che sviluppano in tempo finito discontinuità di tipo salto. Infine, potrebbe essere interessante studiare, anche empiricamente, il comportamento dell'errore in relazione al numero CFL e derivare stime teoriche sull'errore di discretizzazione e sull'ordine di convergenza del metodo nelle regioni del dominio in cui la soluzione è regolare.

Bibliografia

[1] Michael G. Crandall and Pierre-Louis Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comp.*, 43(167):1–19, 1984. (Cited on page 1.)

[2] Stanley Osher and Chi-Wang Shu. High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4):907–922, 1991. (Cited on page 2.)

[3] Guang-Shan Jiang and Danping Peng. Weighted ENO schemes for Hamilton-Jacobi equations. *SIAM J. Sci. Comput.*, 21(6):2126–2143, 2000. (Cited on page 2.)

[4] Steve Bryson and Doron Levy. High-order semi-discrete central-upwind schemes for multi-dimensional Hamilton-Jacobi equations. *J. Comput. Phys.*, 189(1):63–87, 2003. (Cited on page 2.)

[5] Alexander Kurganov, Sebastian Noelle, and Guergana Petrova. Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations. *SIAM J. Sci. Comput.*, 23(3):707–740, 2001. (Cited on page 2.)

[6] Changqing Hu and Chi-Wang Shu. A discontinuous Galerkin finite element method for Hamilton-Jacobi equations. *SIAM J. Sci. Comput.*, 21(2):666–690, 1999. (Cited on page 2.)

[7] Yingda Cheng and Chi-Wang Shu. A discontinuous Galerkin finite element method for directly solving the Hamilton-Jacobi equations. *J. Comput. Phys.*, 223(1):398–415, 2007. (Cited on pages 2, 54, 55, 57, and 58.)

[8] Fengyan Li and Sergey Yakovlev. A central discontinuous Galerkin method for Hamilton-Jacobi equations. *J. Sci. Comput.*, 45(1-3):404–428, 2010. (Cited on pages 2, 51, 52, 53, and 54.)

[9] Jue Yan and Stanley Osher. A local discontinuous Galerkin method for directly solving Hamilton-Jacobi equations. *J. Comput. Phys.*, 230(1):232–244, 2011. (Cited on page 2.)

[10] Yingda Cheng and Zixuan Wang. A new discontinuous Galerkin finite element method for directly solving the Hamilton-Jacobi equations. *J. Comput. Phys.*, 268:134–153, 2014. (Cited on page 2.)

[11] Guoyi Ke and Wei Guo. An alternative formulation of discontinous Galerkin schemes for solving Hamilton-Jacobi equations. *J. Sci. Comput.*, 78(2):1023–1044, 2019. (Cited on pages 2, 54, 59, 60, 61, and 62.)

[12] Maurizio Falcone and Roberto Ferretti. *Semi-Lagrangian Approximation Schemes for Linear and Hamilton—Jacobi Equations.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013. (Cited on pages 2, 3, 35, 36, and 37.)

[13] Elisabetta Carlini, Roberto Ferretti, and Giovanni Russo. A weighted essentially nonoscillatory, large time-step scheme for Hamilton-Jacobi equations. *SIAM J. Sci. Comput.*, 27(3):1071–1091, 2005. (Cited on page 2.)

[14] Olivier Bokanowski and Giorevinus Simarmata. Semi-Lagrangian discontinuous Galerkin schemes for some first- and second-order partial differential equations. *ESAIM Math. Model. Numer. Anal.*, 50(6):1699–1730, 2016. (Cited on page 2.)

[15] Martino Bardi and Italo Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations.* Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 1997. With appendices by Maurizio Falcone and Pierpaolo Soravia. (Cited on pages 3, 9, and 12.)

[16] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, second edition, 2010. (Cited on pages 3, 10, 16, and 45.)

[17] Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 277(1):1–42, 1983. (Cited on page 6.)

[18] Michael G. Crandall, Lawrence C. Evans, and Pierre-Louis Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 282(2):487–502, 1984. (Cited on pages 6 and 7.)

[19] Hitoshi Ishii. Perron's method for Hamilton-Jacobi equations. *Duke Math. J.*, 55(2):369–384, 1987. (Cited on page 12.)

[20] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques & Applications (Berlin) [Mathematics & Applications].* Springer-Verlag, Paris, 1994. (Cited on pages 12 and 16.)

[21] Alfio Quarteroni and Alberto Valli. *Numerical approximation of partial differential equations*, volume 23 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin, 1994. (Cited on page 22.)

[22] Alfio Quarteroni. *Modellistica Numerica per Problemi Differenziali*, volume 100 of *UNITEXT – La Matematica per il 3+2.* Springer-Verlag Italia, Milan, sixth edition, 2016. (Cited on pages 23, 28, and 50.)

[23] Daniele A. Di Pietro and Alexandre Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012. (Cited on pages 30, 31, and 34.)

[24] William H. Reed and Thomas R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, New Mexico, USA, October 1973. (Cited on page 31.)

[25] Emmanuil H. Georgoulis. Discontinuous Galerkin methods for linear problems: an introduction. In *Approximation algorithms for complex systems*, volume 3 of *Springer Proc. Math.*, pages 91–126. Springer, Heidelberg, 2011. (Cited on page 32.)

[26] Claes. Johnson and Juhani Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46(173):1–26, 1986. (Cited on page 32.)

[27] Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu. The development of discontinuous Galerkin methods. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 3–50. Springer, Berlin, 2000. (Cited on page 33.)

[28] Jan S. Hesthaven and Tim Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications. (Cited on page 34.)

[29] Yehuda Pinchover and Jacob Rubinstein. *An introduction to partial differential equations*. Cambridge University Press, Cambridge, 2005. (Cited on page 36.)

[30] Richard Courant, Eugene Isaacson, and Mina Rees. On the solution of nonlinear hyperbolic differential equations by finite differences. *Comm. Pure Appl. Math.*, 5:243–255, 1952. (Cited on page 37.)

[31] Alberto Bressan and Benedetto Piccoli. *Introduction to the mathematical theory of control*, volume 2 of *AIMS Series on Applied Mathematics*. American Institute of Mathematical Sciences (AIMS), Springfield, MO, 2007. (Cited on page 45.)

[32] The MathWorks Inc. *MATLAB version 23.2.0 (R2023b)*. The MathWorks Inc., Natick, Massachusetts, United States, 2023. (Cited on page 49.)