UNIVERSITÉ PARIS VI — PIERRE et MARIE CURIE

U.F.R de Mathématiques

THÈSE DE DOCTORAT

Spécialité Mathématiques Appliquées

Présentée par

Nadia MEGDICH

Pour obtenir le grade de docteur de l'Université Paris VI

Soutenue le 18 Janvier 2008

Méthodes Anti-dissipatives pour les Equations Hamilton Jacobi Bellman

Jury :

Mr. Frédéric Bonnans	Directeur de thèse
Mr. Olivier Bokanowski	Co-directeur de thèse
Mme. Hasnaa Zidani	Co-directrice de thèse
Mr. Régis Monneau	Rapporteur
Mr. Rémi Munos	Rapporteur
Mr. Bruno Després	Examinateur
Mr. Maurizio Falcone	Examinateur
Mr. Sylvain Sorin	Président du jury

Unité de Mathématiques Appliquées

Cette thèse a été préparée conjointement au Laboratoire de Mathématiques Appliquées (ENSTA) et au Projet COMMANDS (INRIA Futurs CMAP)

Projet COMMANDS

Contents

Remerciements

Je tiens en premier lieu à remercier sincèrement mes directeurs de thèse, Frédéric BONNANS, Olivier BOKANOWSKI et Hasnaa ZIDANI. Je remercie Frédéric de m'avoir acceptée dans son équipe et d'avoir toujours veillé pour que cette thèse se déroule bien. Je suis extrémement reconnaissante envers Hasnaa et Olivier, je les remercie pour le temps qu'ils m'ont consacré, pour leur gentillesse et leur disponibilité.

Je remercie du fond du coeur Régis MONNEAU et Rémi MUNOS d'avoir accepté de rapporter sur cette thèse, de lui avoir consacré de leur temps et de leur énergie. Je les remercie pour leur efficacité, leur rapidité et les remarques précieuses qu'ils m'ont fait.

Je remercie Maurizio FALCONE, Bruno DESPRES et Sylvain SORIN d'avoir accepté de faire partie du jury de thèse, leur présence est un grand honneur pour moi.

Ces années de thèse ont aussi été l'occasion de faire mes débuts dans l'enseignement. Je remercie Hasnaa qui m'a permis de m'initier à l'enseignement à ses côtés à l'ENSTA, je la remercie sincèrement pour son soutien pendant mon ATER à Orsay. Un grand merci également à Frédéric JEAN, à Jean Michel CORON et à Emmanuel TRELAT pour leurs conseils précieux et leur disponibilité pendant ma première année d'ATER. Je n'oublie pas Béatrice LAROCHE, collaborer avec elle fut un vrai plaisir.

Je voudrais également remercier les membres de l'UMA pour leur convivialité, leur disponibilité et leur gentillesse. C'était trés agréable de travailler au sein de leur grande famille. Un grand merci va particulièrement à mes chers amis Grace, Eve Marie, Stefania, Elisabeth, Nicolas, Carlo Maria, Colin, Kamel. Je remercie également mes amis qui ont quitté l'UMA Fabrice, Guillaume, Francois et Bassem. De l'autre côté de la méditerrannée, une pensée particulière va à Fakher et Lobna. Je n'oublie pas mes amis Tunisiens à Paris, en particulier Asma, Hicham, Widad, Farouk, Khalil, Adnane, Anis, Mohamed Ali et mon cher ex-voisin de la maison de Tunisie Lotfi.

Cette thèse est l'aboutissement de longues années de travail, avec des hauts et des bas. Un merci du plus profond du coeur aux deux personnes qui m'ont soutenue aux moments les plus difficiles et jusqu'au bout, tout d'abord à mon amie de toujours Rakia. Je la remercie d'avoir toujours trouvé les mots justes pour me réconforter, de m'avoir bien conseillé et de sa disponibilité. Ma gratitude va ensuite à ma chère amie Grace, un grand merci pour sa complicité, pour sa gentillesse, pour son écoute. Merci Grace d'avoir toujours été là quand j'avais besoin de toi.

Pour finir, à mes chers parents je dédie avec fierté cette thèse. Je les remercie pour leurs sacrifices, d'avoir toujours été à mes côtés et de m'avoir soutenue en toutes circonstances. Je les remercie pour leur amour inconditionnel, ma plus grande joie est de les savoir fiers de moi. Je remercie également mes soeurs Emna et Mariam et mon petit frère Mohamed, leur amour attentionné me fait chaud au coeur. Une pensée particulière va à la mémoire de mes grands pères, j'aurais souhaité qu'ils soient là pour partager ma joie.

Introduction

Ce travail s'inscrit dans le cadre général de la théorie du contrôle. Cette discipline a pour objet l'analyse des systèmes dynamiques sur lesquels on peut agir au moyen d'une commande (ou contrôle). Ainsi on est amené à se poser en premier lieu la question suivante: peut on amener le système d'un état initial donné à un état final souhaité (en temps fini) et au moyen de quels contrôles? c'est la question de la contrôlabilité [22, 65, 31] qu'on n'abordera pas dans ce manuscrit. Une fois la question de la contrôlabilité résolue, il est naturel de vouloir faire le transfert entre l'état initial et l'état final en minimisant un critère (la durée du transfert, l'énergie dépensée,...): c'est l'objet du contrôle optimal, cadre dans lequel s'inscrit ce travail. Plus particulièrement, cette thèse a pour objet l'étude numérique des problèmes de contrôle optimal déterministes en horizon fini. Ces problèmes s'écrivent sous la forme générale suivante

$$\mathcal{P}_{T,x} \qquad \begin{cases} \text{Minimiser } \varphi(y(T)), \\ y(0) = x, \\ \dot{y}(t) = f(y(t), \alpha(t)), \quad \alpha(t) \in \mathcal{A} \ p.p.\ t \geq 0 \\ y(t) \in \mathcal{K} \ \forall t \in [0, T], \end{cases}$$

où x est l'état initial, T est l'instant final et l'on cherche à minimiser le coût final φ.

La résolution de ce genre de problèmes peut s'effectuer par une méthode générale non spécifique aux problèmes de contrôle comme la discrétisation totale (appelée encore programmation non linéaire) [57, Chapitre 15] ou encore l'homotopie (appelée aussi méthode de continuation) [57, Section 11.3]. La première méthode est très intuitive, elle consiste en la discrétisation de l'état et du contrôle ce qui conduit à un problème de programmation non linéaire en dimension finie. Il existe une infinité de variantes qui dépendent de la discrétisation du contrôle et de celle de l'équation différentielle. Les contrôles peuvent être approchés par des fonctions constantes par morceaux, splines,... La discrétisation de l'équation différentielle peut se faire quant à elle par une méthode d'Euler explicite ou implicite, Runge-Kutta,... Dans cette méthode, la prise en compte des contraintes sur l'état peut se faire par une pénalisation par exemple, voir la référence détaillée de Bonnans et al. [23].

La méthode d'homotopie consiste en l'approximation du problème $\mathcal{P}_{T,x}$ par une famille de problèmes $(\mathcal{P}_m)_m$ dépendant d'un paramètre m lié à la physique du problème, et qui soient plus simples à résoudre [65].

Il existe également deux méthodes spécifiques au contrôle optimal , toutes deux basées sur des propriétés intrinsèques du problème, à savoir les méthodes de Tir et l'approche Hamilton-Jacobi-Bellman (HJB).

D'une part Les méthodes de tir, qualifiées par indirectes, se basent sur le principe du maximum de Pontryaguin [59] et consistent en la résolution des conditions nécessaires d'optimalité du premier ordre en l'état et l'état adjoint. Cette démarche permet d'approcher un optimum local en boucle ouverte et nécessite une connaissance préliminaire du problème, comme le nombre et les instants de commutation du contrôle. Elle est très sensible à l'initialisation car utilise la méthode de Newton. En particulier l'initialisation de l'état adjoint reste une étape délicate peu intuitive. De plus, tenir compte de contraintes sur l'état nécessite une étude géométrique préliminaire. C'est cependant la méthode principale utilisée en mécanique

spatiale par exemple, car très précise et moins couteuse en mémoire et en temps de calcul que la plupart des méthodes existantes.

D'autre part, l'approche HJB consiste, en un premier temps, à caractériser la fonction valeur du problème $\mathcal{P}_{T,x}$

$$\vartheta(T,x) := \inf \mathcal{P}_{T,x},$$

(qui vaut l'infimum de $\varphi(y_x(T))$ sur l'ensemble des trajectoires y_x admissibles. Par convention cet infimum vaut $+\infty$ s'il n'existe aucune trajectoire admissible) comme l'unique solution d'une Equation aux Dérivées Partielles, dite équation d'Hamilton Jacobi Bellman. Formellement cette équation est de la forme suivante:

$$\vartheta_t(t,x) + \max_{\alpha\in\mathcal{A}}\{-f(x,\alpha)\cdot\vartheta_x(t,x)\} = 0,\ t\in]0,+\infty[,\ x\in\mathcal{K}, \tag{1}$$

avec la condition initiale

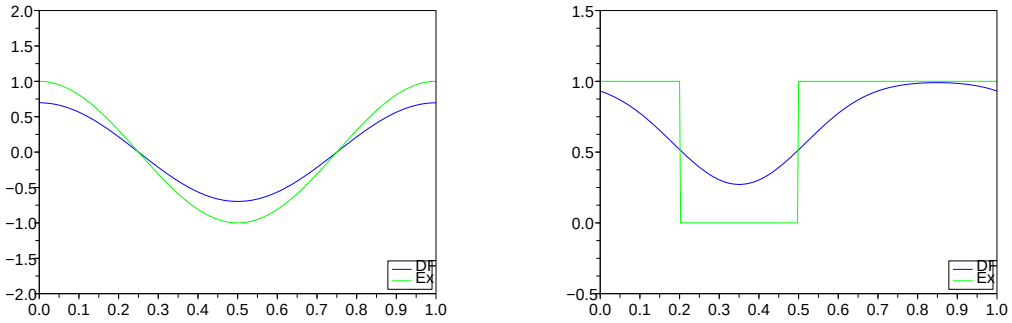
$$\vartheta(0,x) = \varphi(x). \tag{2}$$

Notons au passage que ce type d'EDP peut également modéliser d'autres problèmes comme la propagation de fronts [5, 58], la viabilité et les bassins de capture en biologie [3], la dynamique des dislocations dans les matériaux [2, 47],...

Le premier résultat de caractérisation est dû à Bellman dans le cas où la fonction valeur ϑ est C^1-régulière (mais ceci n'est pas vérifié en général) et remonte aux années 60 [13]. La généralisation de cette caractérisation au cas continu remonte aux années 80 avec le formalisme de la théorie des solutions de viscosité, dû aux travaux de Crandall, Evans et Lions [32, 33]. On se doit également de citer le travail de Soner [63]. Cette théorie s'est encore élargie par la suite pour englober les solutions discontinues dans un cadre non contraint tout d'abord ($\mathcal{K} = \mathbb{R}^n$) étudié par Barron et Jensen [8, 9] et par Frankowska [42]. Puis généralisé au cas contraint [45, 43, 44]. Les excellents livres de Bardi et Capuzzo Dolcetta [4] et de Barles [6] sont deux références incontournables dans ce contexte.
Ceci est l'objet du dernier chapitre qui reprend quelques résultats dûs à Frankowska et Vinter [45] pour lesquels on apporte une extension immédiate.

Suite à cette étape de caractérisation, la résolution de l'équation HJB permet de donner une approximation de la fonction ϑ. Cette résolution pose deux types de difficultés. D'une part celle de choisir un schéma qui soit adapté à l'approximation de la fonction ϑ a priori discontinue. D'autre part la difficulté liée à la grande dimension.

Commençons par le premier point. En effet plusieurs schémas existent pour l'approximation des EDP. Ils sont basés pour la plupart sur des différences finies [50, 34] comme le schéma semi-lagrangien [38, 39, 40] ou encore les schémas d'ordre élevé du type ENO et WENO proposés par Osher et Shu [61, 62]. Ces schémas donnent une bonne approximation de ϑ lorsqu'elle est continue. Cependant, en général, une étape d'interpolation intervient dans ces schémas et cause une perte croissante de précision en temps long, voir la Figure 1 (a). D'autant que pour l'approximation de fonctions discontinues, cette interpolation provoque une diffusion numérique autour des discontinuités, voir Figure 1 (b). Ainsi ces schémas ne semblent pas bien adaptés pour approcher ϑ.



(a) (b)

Figure 1: Transport d'une sinusoide (a) et d'un créneau (b). Diffusion du schéma Décentré: perte du maximum (a) et des discontinuités (b).

Une façon de remédier à cet effet de diffusion est de considérer la classe de schémas avec limiteurs développés pour l'approximation des lois de conservation scalaire. Cette classe de schémas a été unifiée par Sweby [64] sous le formalisme *volumes finis*. On va présenter le formalisme volumes finis dans le cadre de l'équation d'advection de vitesse $a > 0$, bien qu'elle ne soit pas conservative:

$$\begin{aligned} &v_t(t,x) + a(x) \cdot v_x(t,x) = 0, \quad && t \geq 0, x \in \mathbb{R}, \\ &v(0,x) = v_0(x), && x \in \mathbb{R}. \end{aligned}$$

Notons que lorsque l'ensemble de contrôles est un singleton $\mathcal{A} = \{\alpha\}$, l'équation HJB est réduite à une équation d'advection de vitesse $a(x) := -f(x,\alpha)$.

Les schémas aux volumes finis ont la forme générale suivante:

$$\frac{V_j^{n+1} - V_j^n}{\Delta t} + a(x_j)\frac{V_{j+\frac{1}{2}}^n - V_{j-\frac{1}{2}}^n}{\Delta x} = 0, \tag{3}$$

où les $(V_{j+\frac{1}{2}}^n)_{j\in\mathbb{Z}}$ sont les flux aux interfaces des mailles et sont à définir. D'une façon générale, la construction d'un schéma avec limiteur se fait par l'addition d'un schéma d'ordre un avec un flux antidiffusif. Pour plus de détails sur les schémas numériques, une présentation complète est proposée dans les livres de Godlewski et Raviart [49, 48].

Dans cette classe de schémas à volumes finis, on investit en particulier le schéma UltraBee (UB), développé par Després et Lagoutière dans le cadre de la résolution des EDP hyperboliques pour les discontinuités de contact des gaz compressibles [53, 37], et adapté aux équations HJB (UB-HJB) par Bokanowski et Zidani [20]. Ce schéma approche la valeur moyenne de ϑ sur chaque maille de la grille et présente ainsi la particularité de bien localiser les discontinuités au cours du temps. En particulier, dans le cas de l'équation Eikonale en

dimension 1 où c est une constante positive donnée:

$$\begin{array}{ll} \vartheta_t(t,x) + c|\vartheta_x(t,x)| = 0, & t \geq 0,\ x \in \mathbb{R}, \\ \vartheta(0,x) = \varphi(x), & x \in \mathbb{R}, \end{array}$$

et quand la fonction φ a une forme étagée avec des discontinuités "suffisamment" éloignées, le schéma UltraBee calcule exactement la valeur moyenne de ϑ sur les mailles de la grille de discrétisation, à chaque étape de temps. Cette advection exacte est en particulier vraie lorsque φ est associée à une cible $\mathcal{C}$:

$$\varphi(x) = \begin{cases} 0 & \text{sur } \mathcal{C}, \\ 1 & \text{ailleurs}, \end{cases} \tag{4}$$

où $\mathcal{C}$ est un fermé de $\mathbb{R}$.

Dans les chapitres 2, 3 et 4 on s'intéressera exclusivement à des problèmes faisant intervenir la classe (4) de coût final φ. On explicitera en particulier quelques problèmes qui font intervenir un tel coût final, comme le problème cible ou encore le problème de temps minimal.

Notons qu'il existe d'autres méthodes bien adaptées pour approcher l'équation HJB (1) avec une condition initiale de type (4). Les méthodes "Fast Marching" et "Level Set" étudiées par Sethian [60] ont en particulier été développées pour les problèmes de poursuite d'interface ou encore de propagation de fronts.
Soit Γ_0 la position initiale de la discontinuité définie par

$$\Gamma_0 = \partial\mathcal{C}.$$

La méthode Level Set consiste à construire une fonction ψ qui s'annule sur l'interface Γ_t [41]. Ainsi pour traquer l'évolution de l'interface à l'instant t, il suffit de suivre la courbe de niveau 0 de ψ:

$$\Gamma_t := \{x,\ \psi(t,x) = 0\}.$$

La méthode Fast Marching quant à elle consiste à explorer, à partir des noeuds de la grille traversés par le front Γ_t, leurs voisins. Ainsi les meilleurs candidats (qui seront atteints par le front après une étape de temps) sont sélectionnés et admis. L'évolution du front se fait sur une bande mince par exploration de la grille de proche en proche. Cette méthode a été testée sur divers exemples provenant de la théorie du contrôle mais aussi des jeux différentiels et du problème inverse de traitement d'image "shape from shading" [35]. Une généralisation de cette méthode a récemment été étudiée pour l'étendre à des vitesses de signe variable [28].

On étudie dans le premier chapitre la convergence du schéma UltraBee. Dans le contexte de l'étude de convergence, la monotonie du schéma joue un rôle fondamental. En effet, sous des hypothèses génériques de monotonie, de stabilité et de consistance, un résultat général de convergence (vers la solution de viscosité) dû à Barles et Souganidis est prouvé dans [7]. Lorsque la monotonie du schéma n'est pas assurée, une "presque monotonie" permet encore de prouver la convergence du schéma ε-monotone d'Abgrall et Augoula [1]. Le seul résultat de convergence ne faisant intervenir aucune hypothèse de monotonie est à notre connaissance

celui de Lions et Souganidis [55] pour les schémas MUSCL (ce sont les schémas TVD du second ordre) implicites.

La difficulté de cette partie réside dans le fait que le schéma UltraBee est non monotone, de plus il s'agit d'un schéma explicite. En plus de la convergence, on explicite une estimation de l'erreur, c'est à notre connaissance le premier résultat de ce type pour un schéma explicite et non monotone.

Les chapitres 2 et 3 de ce mémoire traitent l'aspect numérique du problème. Il est bien connu que la méthode HJB souffre de la "malédiction de la dimension" ce qui représente en général un handicap à son application à des problèmes de grande dimension (6 ou 7 par exemple dans les problèmes d'aérospatiale). On s'est intéressé pour cela à développer des méthodes de résolution rapide. Ces méthodes se basent sur la forme particulière des fonctions valeur qu'on considère (constantes par morceaux à valeurs dans $\{0,1\}$) d'une part, et sur la bonne localisation des discontinuités au cours du temps dûe à l'UltraBee d'autre part.
On propose deux méthodes d'implémentation rapide.

Tout ce travail de thèse est motivé par l'applicabilité de l'approche HJB aux problèmes industriels. Dans ce contexte, une méthode qui utilise l'approche HJB a été développée par Guilbaud [51]. Son étude a été motivée par l'optimisation de l'architecture de moteurs hybrides thermique-électrique (RENAULT).

Le quatrième chapitre de cette thèse s'inscrit dans le cadre du projet OPALE entre l'INRIA et le CNES. Il concerne l'applicabilité de la méthode sparse que nous proposons à quelques problèmes provenant de l'aérospatiale. On s'intéresse au problème de la rentrée atmosphérique d'une navette spatiale dont on traite quelques variantes.

Ce problème a déjà fait l'objet de plusieurs études antérieures, par diverses méthodes. On peut citer en particulier les multiples travaux de Betts sur l'approximation de la trajectoire optimale [12, 10, 16, 14, 15], mais aussi les contributions de Bonnard, Faubourg et Trélat [25, 26]. Mentionnons également les articles de Bonnans et al [24, 27, 21] et la thèse de Laurent-Varin [54].

Chapitre 1. Convergence d'un schéma non monotone pour les équations HJB
Pour donner une approximation numérique de la fonction valeur, on discrétise l'équation HJB (1)-(2). On investit pour cela le schéma UltraBee (UB). Ce schéma a été étudié par Després et Lagoutière [53, 37] pour l'équation de transport de vitesse constante et étendu par Bokanowski et Zidani [20] dans le cas où la vitesse est variable. Il a également été appliqué à la résolution de l'équation HJB (UB-HJB) sur un maillage structuré [20, 17].
Nous nous intéressons dans le premier chapitre de cette thèse à étudier la convergence du schéma UltraBee en 1D pour l'équation (1)-(2) avec $\mathcal{K} = \mathbb{R}^n$.
On fait les hypothèses suivantes sur la dynamique f et l'ensemble de contrôles $\mathcal{A}$:

$$(H1) \qquad \exists L \geq 0,\ \forall \alpha \in \mathcal{A},\ \forall x, y \in \mathbb{R},\ |f(y,\alpha) - f(x,\alpha)| \leq L|y-x|,$$

Définissons pour tout $x \in \mathbb{R}$ les vitesses maximale et minimale d'évolution des discontinuités,

$$f_m(x) := \min_{\alpha\in\mathcal{A}} -f(x,\alpha) \quad \text{et} \quad f_M(x) := \max_{\alpha\in\mathcal{A}} -f(x,\alpha).$$

On suppose de plus que f_m et f_M vérifient

$(H2)$ f_m et f_M sont chacune de signe constant,
$(H3)$ f_m et f_M sont croissantes.

Alors (1)-(2) se ramène à:

$$\vartheta_t(t,x) + \max\{f_m(x)\ \vartheta_x(t,x)\ ;\ f_M(x)\ \vartheta_x(t,x)\} = 0, \qquad t>0,\ x\in\mathbb{R}, \tag{5a}$$
$$\vartheta(0,x) = v_0(x), \qquad x\in\mathbb{R}, \tag{5b}$$

où la condition initiale $v_0 \equiv \varphi$ est supposée s.c.i.

Soit $\mathcal{G}$ une grille régulière de pas Δx et soit $x_j = j\Delta x,\ j\in\mathbb{Z}$, et Δt le pas de temps. Rappelons que le schéma UltraBee est de la forme volumes finis (3) et qu'on l'initialise avec la valeur moyenne de v_0 sur $\mathcal{G}$:

$$V_j^0 = \frac{1}{\Delta x}\int_{]x_{j-\frac12},x_{j+\frac12}[} v_0(x)dx \quad \text{avec}\ \ x_{j+\frac12} = (j+\frac12)\Delta x.$$

Dans le cas où $\mathcal{A}=\{\alpha\}$ et f_M est constante et positive, le flux UB a l'expression suivante:

$$V^n_{j+\frac12} := \min(\max(V^n_{j+1}, b^+_j), B^+_j)$$

avec

$$b^+_j := \max(V^n_j, V^n_{j-1}) + \tfrac{1}{\nu_j}(V^n_j - \max(V^n_j, V^n_{j-1})),$$
$$B^+_j := \min(V^n_j, V^n_{j-1}) + \tfrac{1}{\nu_j}(V^n_j - \min(V^n_j, V^n_{j-1})),$$

où $\nu_j := f_M(x_j)\frac{\Delta t}{\Delta x}$.
Pour une définition détaillée du schéma UB-HJB quand la dynamique change de signe, on renvoie aux sections 1.2.2 et 1.5 du chapitre 1. Rappelons également que ce schéma est stable (au sens où $\max_j |V^{n+1}_j| \le \max_j |V^n_j|$) sous la condition CFL:

$$\max_{x\in\mathbb{R}} \max(|f_m(x)|\ ,\ |f_M(x)|)\frac{\Delta t}{\Delta x} \le 1. \tag{6}$$

On s'intéresse tout d'abord au cas élémentaire représenté dans la Figure 2 (cas 1) d'une fonction constante par morceaux avec un seul minimum limité par deux discontinuités:

$$v_0(x) := \mathbf{1}_{]-\infty,a[}(x) + \mathbf{1}_{]b,+\infty[}(x),\ x\in\mathbb{R}.$$

On suppose que les deux discontinuités sont séparées initialement par au moins deux mailles entières il suffit pour cela que le pas Δx vérifie:

$$b - a \geq 3\Delta x. \tag{7}$$

Alors on montre sous (H1)-(H3) que le schéma UB-HJB calcule à chaque étape de temps t_n l'évolution exacte de la condition initiale v_0 avec la vitesse constante sur chaque maille:

$$V_j^n = \frac{1}{\Delta x} \int_{]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[} \vartheta^S(t^n, x)dx, \quad \forall j \in \mathbb{Z},$$

avec $\vartheta^S(t,x) := \mathbf{1}_{]-\infty, X_a^{m,S}(t)[}(x) + \mathbf{1}_{]X_b^{M,S}(t), +\infty[}(x)$, et les caractéristiques approchées $X_x^{m,S}$ et $X_x^{M,S}$ définies pour x donné dans $\mathbb{R}$ par:

$$X_x^{m,S}(t) := x + \int_0^t \sum_j f_m(x_j) 1_{]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[}(X_x^{m,S}(\tau))d\tau,$$

$$X_x^{M,S}(t) := x + \int_0^t \sum_j f_M(x_j) 1_{]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[}(X_x^{M,S}(\tau))d\tau.$$

En particulier dans le cas de l'équation Eikonale par exemple:

$$\vartheta_t(t,x) + |\vartheta_x(t,x)| = 0, \ t > 0, \ x \in \mathbb{R}, \tag{8}$$

les vitesses $f_m \equiv -1$ et $f_M \equiv 1$ sont constantes et le schéma calcule exactement la valeur moyenne de ϑ:

$$V_j^n = \frac{1}{\Delta x} \int_{]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[} \vartheta(t^n, x)dx, \ \forall j \in \mathbb{Z}.$$

Figure 2: Cas élémentaires

Le deuxième cas élémentaire est celui d'une fonction constante par morceaux avec un seul maximum représenté dans la figure 2 (cas 2):

$$v_0(x) := \mathbf{1}_{]a,b[}(x), \ x \in \mathbb{R}.$$

On suppose (7) vérifiée, en particulier cela implique que les discontinuités sont séparées au moins par deux mailles entières. Alors on montre que le schéma calcule exactement les valeurs moyennes sur la grille $\mathcal{G}$ de ϑ^S définie par

$$\vartheta^S(t,x) := \mathbf{1}_{]X_a^{M,S}(t),X_b^{m,S}(t)[}(x), \quad \forall t \geq 0,\ x \in \mathbb{R}.$$

La résolution de l'équation Eikonale (8) est encore exacte dans ce cas, tant que les discontinuités sont assez éloignées.
Ce résultat généralise aux équations HJB les résultats de Després et Lagoutière qui montrent que l'UltraBee transporte exactement une classe de fonctions constantes par morceaux, avec une vitesse approchée constante sur chaque maille.

Quand les discontinuités sont "trop proches", le schéma n'est plus capable d'estimer correctement les moyennes de ϑ^S: on effectue une étape de troncature (voir algorithme 2 du chapitre 1). Grâce à l'estimation suivante:

$$X_b^{m,S}(t) - X_a^{M,S}(t) \leq (1+b-a)e^{Lt}\Delta x,$$

tant que $X_b^{m,S}(t) \geq X_a^{M,S}(t)$, l'erreur qu'on commet entre V^n et les moyennes de $\vartheta^S(t_n,\cdot)$ est contrôlable en Δx.

Ce résultat se généralise au cas d'une condition initiale v_0 s.c.i. quelconque à variation totale finie et ayant un nombre fini d'extrema. On suppose que

$$(H4)\quad \begin{array}{l} \text{Il existe q+1 minima de } v_0 \text{ situés en } A_1,\cdots,A_{q+1} \\ \text{et q maxima situés en } B_1,\cdots,B_q, \text{ avec} \\ A_1 = -\infty \leq B_1 \leq A_2 \leq \cdots \leq B_q \leq A_{q+1} = +\infty, \\ \text{avec éventuellement } B_1 = -\infty \text{ ou bien } B_q = +\infty. \end{array}$$

On suppose que le pas Δx vérifie

$$\Delta x < \frac{B_i - A_i}{3} \quad \text{et} \quad \Delta x < \frac{A_{i+1} - B_i}{3} \quad \forall i = 1,\cdots,q, \tag{9}$$

et Δt vérifie (6).

Soit la fonction v_0^P la projection de v_0 sur la grille de pas $3\Delta x$, pour fixer les idées on définit $U_j =]x_{3j-\frac{1}{2}}, x_{3j+\frac{5}{2}}[$ et v_0^P par (10)-(12):

- Si $\overline{U_j} \cap \{(A_k)_{k=2,\ldots,q}, (B_k)_{k=1,\ldots,q}\} = \emptyset$, alors

$$v_0^P(x) := \frac{1}{3\Delta x}\int_{U_j} v_0(y)dy, \quad \forall x \in U_j. \tag{10}$$

- Si $A_k \in \overline{U_j}$ (resp. $B_k \in \overline{U_j}$) alors

$$v_0^P(x) := v_0(A_k)(\text{resp. } v_0(B_k)) \quad \forall x \in U_j. \tag{11}$$

- Etendre v_0^P par semi continuité inférieure:

$$v_0^P(x_{3j-\frac{1}{2}}) = \min\left(v_0^P(x^+_{3j-\frac{1}{2}}),\, v_0^P(x^-_{3j-\frac{1}{2}})\right). \tag{12}$$

L'initialisation du schéma UB-HJB se fait cette fois avec

$$V_j^0 = v_0^P(x_j), \quad \forall j \in \mathbb{Z}.$$

Le but de cette initialisation est de se ramener à une fonction constante par morceaux dont les discontinuités sont séparées par $3\Delta x$ au moins. D'autre part noter que les valeurs des extrema de v_0 sont préservées dans v_0^P, ce sont ces valeurs qui se propagent au cours du temps.

A partir des valeurs $(V_j^n)_{j\in\mathbb{Z}}$ données par le schéma UB-HJB, on définit presque partout sur $\mathbb{R}^+ \times \mathbb{R}$ la fonction V constante par morceaux par:

$$V(t_n, x) := V_j^n, \quad \forall x \in]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[,\ \forall t_n = n\Delta t. \tag{13}$$

Théorème 0.1. *Soit $v_0 : \mathbb{R} \to \mathbb{R}$ une fonction s.c.i. avec un nombre fini d'extrema vérifiant (H4) et à variation totale finie. On suppose (H1)-(H3) et que les pas Δx et Δt vérifient (6) et (9). Soit ϑ la solution de viscosité s.c.i. de (5), alors*

$$||V(t_n,.) - \vartheta(t_n,.)||_{L^1(\mathbb{R})} \leq (1 + Lt_n\, e^{Lt_n})\, TV(v_0)\, \Delta x, \quad \forall n \geq 0, \tag{14}$$

où V est définie par (13) et les $(V_j^n)_j$ sont donnés par l'algorithme UB-HJB initialisé avec $V_j^0 = v_0^P(x_j)$, $\forall j$.

Les hypothèses (H2) et (H3) étant trop restrictives, on étudie également l'erreur du schéma quand ces deux hypothèses sont remplacées par l'une des hypothèses suivantes:

(H5a) $\exists \varepsilon > 0,\ \forall x \in \mathbb{R},\ f_m(x) + \varepsilon \leq f_M(x)$ et $\Delta x < \frac{\varepsilon}{2L}$,
($H5b$) $\forall x \in \mathbb{R},\ f_m(x) \leq 0,\ f_M(x) \geq 0$,
($H5c$) $f_m = f_M$ et est une fonction croissante,

En absence de (H2)-(H3), deux discontinuités dans une zone monotone de v_0 peuvent se rapprocher au cours du temps. Soit $a < b$ et

$$v_0(x) = \mathbf{1}_{]-\infty,a[}(x) + \mathbf{1}_{]-\infty,b[}(x),$$

alors comme illustré par la Figure 3, les discontinuités peuvent devenir très proches. Dans ce cas aussi le schéma UB-HJB fait des erreurs de calcul de la moyenne. Pour remédier à cela, on effectue dans l'algorithme une étape de prédiction qui induit une erreur contrôlable en Δx.

On obtient ainsi les estimations d'erreur suivantes

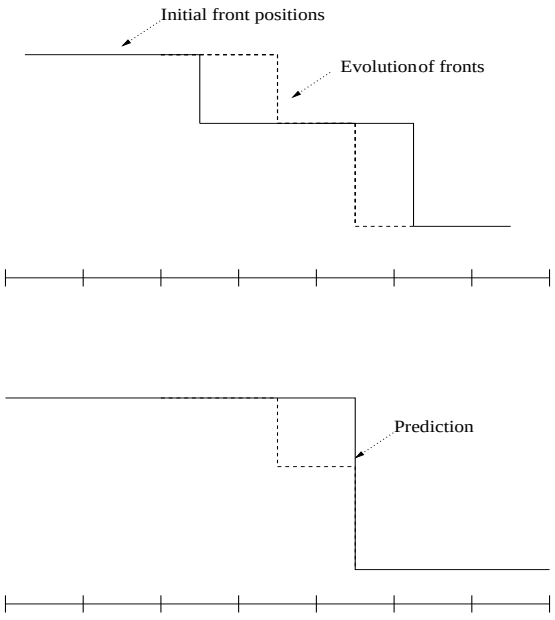


Figure 3: Etape de prediction

Théorème 0.2. *On suppose (H1). Soit v_0 une fonction s.c.i. vérifiant (H4) et telle que $TV(v_0) < \infty$.*

Soient les pas $\Delta x > 0$ et $\Delta t > 0$ vérifiant la condition CFL (6) et (9). Soient V_j^n les valeurs données par le schéma, V définie par (13) et ϑ la solution de viscosité s.c.i. de (5). Alors:
(i) Si v_0 est constante par morceaux avec p discontinuités séparées au moins de $3\Delta x$ et l'une des hypothèses (H5a) ou (H5b) est satisfaite, alors

$$||V(t_n,\cdot) - \vartheta(t_n,\cdot)||_{L^1(\mathbb{R})} \leq (Lt_n + 4p)e^{Lt_n} TV(v_0)\Delta x, \quad \forall n \geq 0. \tag{15}$$

(ii) Si (H5c) est vérifiée alors

$$||V(t_n,\cdot) - \vartheta(t_n,\cdot)||_{L^1(\mathbb{R})} \leq (1 + Lt_n e^{Lt_n}) TV(v_0)\Delta x, \quad \forall n \geq 0. \tag{16}$$

Pour conclure, précisons qu'à notre connaissance ce résultat de convergence (ainsi que l'estimation en norme L^1 qu'on donne) est le seul qui existe pour un schéma non-monotone explicite dans un contexte discontinu, du moins pour les équations HJB. Dans le cadre des équations linéaires, une estimation d'erreur pour les schémas aux volumes finis a été donnée par Després [36] et permet d'étendre le théorème de Lax.

Mise en oeuvre rapide

On se focalise dans cette partie sur l'approximation de la fonction valeur dans le cas où elle prend uniquement les valeurs 0 et 1. Ceci est en particulier le cas d'une propagation de front modélisée non pas avec les techniques classiques de "courbes de niveaux " mais par l'évolution d'une fonction ϑ prenant les valeurs 0 et 1 d'un côté ou de l'autre du front. La forme particulière de cette fonction nécessite uniquement de bien localiser la zone du front à chaque étape de temps t_n. De part et d'autre de la discontinuité, $\vartheta(t_n, \cdot)$ est de valeur connue égale à 0 ou 1: elle prend la valeur 0 aux points admissibles déjà atteints par le front, et la valeur 1 aux points non encore atteints ou non admissibles (sur l'obstacle par exemple).
Pour l'approximation de ces fonctions, l'UltraBee est particulièrement bien adapté. En effet comme nous l'avons déjà expliqué dans le chapitre 1, ce schéma localise bien les discontinuités car il donne une bonne approximation de la valeur moyenne de ϑ sur chaque maille. Dans certains cas (par exemple pour l'equation Eikonale) il est même capable de calculer exactement la valeur moyenne de ϑ sur chaque maille. En dimension supérieure à 1, l'approximation que donne l'UltraBee de la valeur moyenne reste de bonne précision. Ceci est confirmé par des tests numériques effectués sur des exemples académiques en 2D et en 3D (sections 2.4, 3.6 et 3.7).

De ce fait, il est possible de coupler le schéma UltraBee avec une structuration de stockage des données qui permet d'économiser la mémoire et de réaliser un gain en termes de précision et de temps de calcul. Ce gain permettra d'appliquer ces méthodes à des problèmes en dimension supérieure.
On propose en particulier deux méthodes dans la suite.

Chapitre 2. Méthode adaptative Ce chapitre a fait l'objet d'une publication [19], une version plus détaillée est également disponible en rapport de Recherche INRIA [18].
Cette première méthode, consiste à résoudre l'équation sur un maillage adaptatif. L'adaptation de maillage utilise la forme étagée de la fonction valeur. En effet, les mailles sont raffinées au maximum (un niveau de raffinement maximal est choisi pour cela) au voisinage d'une discontinuité, et elles sont grossières sur les paliers où la fonction est constante. Ainsi le raffinement du maillage "traque" l'évolution du front. La structuration de la grille est effectuée par les quadtree linéaires [46], une technique proposée par Gargantini qui permet d'hierarchiser la grille en arbre dont on stocke uniquement les feuilles (ce sont les noeuds terminaux de l'arbre).

L'idée d'adaptation de la grille a déjà été proposée par Grune [50] dans le cadre de l'approximation d'une solution continue de l'équation HJB. Dans [50], l'algorithme de raffinement repose sur un test d'erreur utilisant l'interpolation, donc la continuité de la solution. On peut également citer le travail de Munos et Moore [56] et de Cockburn et Yenikaya [29, 30]. L'apport dans notre travail consiste à proposer un test d'arrêt du raffinement qui soit adapté à la discontinuité de ϑ. On a dû imposer a priori un niveau de raffinement maximal. Ce niveau est atteint uniquement au voisinage du front numérique.

Il est à noter que cette méthode permet d'avoir une solution approchée aussi précise que si on avait effectué les calculs sur une grille régulière, avec un pas minimal.

Cette méthode permet une meilleure gestion de la mémoire. On a pû ainsi améliorer les approximations et atteindre des précisions qu'on ne pouvait pas avoir sur un maillage structuré (à cause de la saturation de l'espace mémoire de la machine). La méthode est validée sur divers exemples académiques en dimension 2.

Chapitre 3. Méthode sparse Dans la deuxième méthode, la résolution est effectuée sur un maillage structuré mais seules les valeurs sur les mailles au voisinage d'une discontinuité sont stockées. En effet lors de l'évolution du front numérique, seules les mailles voisines sont susceptibles de changer de valeur (car le schéma est stable sous la condition CFL). Ainsi en une étape de temps, les mailles éloignées du front gardent la même valeur, et de ce fait il est inutile de réactualiser leurs valeurs et de les stocker.

La structuration du code est organisée de telle sorte que l'extension à des dimensions supérieures ne pose aucune difficulté supplémentaire. On dispose d'un code qui permet d'effectuer la résolution en 2D et en 3D. Cette méthode a été testée sur des problèmes provenant de divers domaines.

L'apport des chapitres numériques 2 et 3 a été l'élaboration et l'implémentation des algorithmes de ces deux méthodes: en 1D et 2D pour la méthode adaptative, en 2D et 3D pour la méthode sparse. On a également pû valider ces codes sur divers exemples académiques.

Chapitre 4. Application à l'aérospatiale L'exploration de l'espace a toujours représenté un grand défi et un moteur pour l'avancée technologique. De part l'importance des satellites artificiels dans l'amélioration de notre vie moderne, les problèmes de lancement de ces corps occupe une place de choix dans les études aérospatiales.
Depuis le premier satellite artificiel Sputnik lancé par les russes dans les années 50, la technique repose toujours sur l'utilisation de lanceurs consommables qui sont détruits dans l'opération de lancement et dont on ne récupère aucune composante.
Pour des raisons de réduction de coût, on s'intéresse de plus en plus aux lanceurs réutilisables dont la mission comprendrait une phase de rentrée pendant laquelle aucune force de poussée n'est appliquée: le véhicule est un planeur commandé uniquement par ses angles d'attitude.

Ce problème est bien connu dans la littérature. En particulier Betts a beaucoup contribué à son étude numérique [12, 10, 11, 16, 14, 15] par les méthodes de Tir et la Discrétisation totale essentiellement. Il s'est également intéressé à approcher la trajectoire optimale. On peut également citer les travaux de Bonnard et Trélat [25, 26] et de Bonnans et Launay [24]. Une référence plus récente est la thèse de Laurent Varin [54].

On se propose dans le chapitre 4 de cette thèse de traiter cette phase de rentrée, en utilisant la méthode sparse proposée dans le chapitre 3. En particulier on s'intéresse à déterminer

le domaine atteignable dans lequel on peut espérer reconstruire les trajectoires optimales. Nous traitons en particulier deux modèles de ce problème en 3D, une contrainte thermique intervient durant la phase atmosphérique.

Chapitre 5. Problème de Rendez Vous avec contraintes sur l'état Nous considérons le problème de contrôle $\mathcal{P}_{T,x}$ où l'ensemble $\mathcal{K}$ est un fermé de $\mathbb{R}^n$ qui définit la contrainte sur l'état et le coût final $\varphi : \mathbb{R}^n \to \mathbb{R}$ est s.c.i. La dynamique du système est définie par la fonction f et contrôlée par $\alpha \in A := L^\infty(\mathbb{R}^+, \mathcal{A})$. On se situe pour l'étude de ce problème sous des hypothèses classiques sur la dynamique f et l'ensemble de contrôles $\mathcal{A}$:

$$(A1)\qquad \begin{array}{l} \mathcal{A} \text{ est un compact convexe de } \mathbb{R}^n. \\ f \text{ est continue.} \\ |f(y,\alpha)| \leq k_1(1+|y|),\ \forall (y,\alpha) \in \mathbb{R}^n \times \mathcal{A}. \\ |f(y_1,\alpha) - f(y_2,\alpha)| \leq k_2|y_1 - y_2|,\ \forall y_1, y_2 \in \mathbb{R}^n,\ \alpha \in \mathcal{A}. \\ f(y,\mathcal{A}) := \{f(y,\alpha),\ \alpha \in \mathcal{A}\} \neq \emptyset \text{ est un convexe de } \mathbb{R}^n,\ \forall y \in \mathbb{R}^n. \end{array}$$

Plusieurs travaux se situent dans le cadre de l'étude du problème $(\mathcal{P}_{T,x})$, l'objectif principal étant de caractériser sa fonction valeur ϑ comme l'unique solution de l'équation HJB évolutive:

$$\vartheta_t(t,x) + \max_{\alpha \in \mathcal{A}} \{-f(x,\alpha) \cdot \vartheta_x(t,x)\} = 0 \quad (t,x) \in]0,+\infty[\times \mathcal{K}, \tag{17a}$$

$$\vartheta(0,x) = \varphi(x) \quad x \in \mathbb{R}^n. \tag{17b}$$

Donnons tout d'abord une vue d'ensemble de la plupart des résultats existants dans la littérature. La fonction ϑ étant peu régulière en général, on se doit de préciser le sens de solution qu'on donne à l'équation (17a) ainsi qu'à la condition intiale (17b). Ce sens dépend essentiellement de la régularité du coût φ et de l'ensemble $\mathcal{K}$.
Les premiers résultats sont dûs à Crandall, Evans et Lions [34, 33] dans le cas où φ est bornée uniformément continue et en absence de contraintes sur l'état: $\mathcal{K} := \mathbb{R}^n$. Ils montrent que ϑ est l'unique solution de viscosité continue bornée de (17).

Lorsque φ est s.c.i., la fonction valeur est caractérisée [8, 9, 42] comme l'unique sursolution bilatérale de de (17).

Quand l'ensemble $\mathcal{K}$ est un fermé, la notion de viscosité (définie initialement sur des ouverts) est à manipuler avec précaution. Lorsque φ est continue et sous l'hypothèse de qualification rentrante (Inward Pointing)

$$\exists \nu > 0,\ \forall x \in \partial\mathcal{K},\ \exists \alpha \in \mathcal{A},\ f(x,\alpha).\eta(x) \leq -\nu, \tag{IP}$$

(où η est la normale sortante en x à $\mathcal{K}$ et $\partial\mathcal{K}$ est supposé régulier), Soner [63] prouve que ϑ est continue et qu'elle est l'unique solution de viscosité contrainte de (17a). Ce résultat est généralisé par Ishii et Koike [52] sous une hypothèse rentrante plus faible appelée contrainte de viabilité (Viability Constraint):

$$\mathcal{A}(x) \neq \emptyset, \quad \forall x \in \partial\mathcal{K}, \tag{VC}$$

où $\mathcal{A}(x) = \{\alpha \in \mathcal{A},\ \exists r > 0,\ y(t) \in \mathcal{K}\ \forall \xi \in \mathcal{B}(x,r) \cap \mathcal{K},\ t \in [0,r],\ \dot{y}(t) = f(y(t),\alpha(t)),\ y(0) = \xi\}$.

La contrainte (VC) signifie que pour tout point x dans $\mathcal{K}$ il existe une trajectoire viable i.e. qui ne quitte jamais $\mathcal{K}$. La contrainte (IP) plus forte indique l'existence d'une trajectoire qui reste à l'intérieur de $\mathcal{K}$ (elle touche éventuellement $\partial\mathcal{K}$ ponctuellement si $x \in \partial\mathcal{K}$).
Une autre contrainte de qualification encore plus forte que (IP) et (VC), appelée contrainte sortante (Outward Pointing)

$$\forall x \in \partial\mathcal{K},\ \exists \alpha \in \mathcal{A},\ f(x,\alpha).\eta(x) > 0, \tag{OP}$$

apparaît également dans la littérature. Elle est utilisée en particulier par Frankowska et ses co-auteurs [43, 45, 44] pour caractériser la fonction ϑ comme étant l'unique solution de viscosité s.c.i de l'équation HJB (17).

La contrainte (OP) signifie qu'en tout point du bord $\partial\mathcal{K}$ il existe une trajectoire qui sort strictement de $\mathcal{K}$ et non pas tangentiellement au bord. Cela signifie par un raisonnement "backward" en temps qu'en tout point du bord arrive une trajectoire qui vient de l'intérieur de $\mathcal{K}$.

Dans le chapitre 5 on s'est posé les questions suivantes: Est-il possible d'obtenir un résultat de caractérisation sous des hypothèses plus générales que (OP) et peut on caractériser ϑ sans supposer aucune hypothèse de qualification ?
Pour répondre à la première question, on donne une extension immédiate du résultat d'unicité de Frankowska et Vinter [45] qui permet de traiter certains problèmes d'advection, problèmes pour lesquels la contrainte (OP) n'est pas satisfaite. On suppose que la condition initiale φ vérifie

$$(A0) \qquad \begin{array}{l} \varphi : \mathbb{R}^n \to [0,1] \text{ continue sur } \mathcal{C}, \\ \varphi(x) = 1 \text{ si } x \in \mathbb{R}^n \backslash \mathcal{C} \quad \text{ et } \varphi(x) \in [0,1[\text{ si } x \in \mathcal{C}, \end{array}$$

avec $\mathcal{C}$ la cible à atteindre. On suppose aussi que $\mathcal{K}$ et $\mathcal{C}$ vérifient

$$(A2) \qquad \mathcal{K},\ \mathcal{C} \neq \emptyset \text{ sont des fermés bornés de } \mathbb{R}^n,\ \mathcal{C} \subset \overset{\circ}{\mathcal{K}},\ \overset{\circ}{\mathcal{C}} \neq \emptyset \text{ et } \overline{\overset{\circ}{\mathcal{K}}} = \mathcal{K}.$$

On suppose également que $\mathcal{K}$ est régulier ou régulier par morceaux défini par:

$$(A3) \qquad \mathcal{K} := \cap_{j=1,..,r}\{x,\ h_j(x) \leq 0\},$$

où les $h_j : \mathbb{R}^n \to \mathbb{R}$ sont des fonctions C^1-régulières avec un gradient localement Lipschitz. On notera par

$$I(x) := \{j = 1, \cdots, r,\ h_j(x) = 0\},$$

les indices des contraintes actives.

Théorème 0.3. *On suppose (A0)-(A3) satisfaites, $\mathcal{K}$ vérifie la contrainte de qualification:*

$$\forall x \in \partial\mathcal{K}, \quad \begin{array}{ll} \exists \alpha \in \mathcal{A} \ \ t.q. \ \nabla h_j(x)\cdot f(x,\alpha) > 0, \ \forall j \in I(x), & (OP) \\ ou\ bien \ \forall \alpha \in \mathcal{A}\ \big[\nabla h_j(x)\cdot f(x,\alpha)\cdot \eta(x) < 0, \ \forall j \in I(x)\big] \ ou\ bien\ \big[f(x,\alpha)=0\big]. & (SIP) \end{array}$$

Alors ϑ est l'unique fonction s.c.i. à valeurs dans $[0,1]$ et telle que $\vartheta(t,x) = 1, \ \forall t \geq 0, \ x \in \mathcal{K}^c$, vérifiant
i) Pour tout $(t,x) \in]0,+\infty[\times \overset{\circ}{\mathcal{K}})$, et tout $(p_t,p_x) \in \partial_- \vartheta(t,x)$,

$$p_t + \max_{\alpha\in\mathcal{A}}\{-f(x,\alpha)\cdot p_x\} = 0.$$

ii) Pour tout $(t,x) \in]0,+\infty[\times\partial\mathcal{K})$, et tout $(p_t,p_x) \in \partial_- \vartheta(t,x)$,

$$p_t + \max_{\alpha\in\mathcal{A}}\{-f(x,\alpha)\cdot p_x\} \geq 0.$$

iii) Pour tout $x \in \mathcal{K}$,

$$\liminf_{\substack{t'\to 0^+ \\ x'\to x,\ x'\in\overset{\circ}{\mathcal{K}}}} \vartheta(t',x') = \vartheta(0,x) = \varphi(x).$$

En ce qui concerne la deuxième question, les hypothèses de qualifications semblent être incontournables pour ce type de résultat.
Notre approche a été tout d'abord de modifier la dynamique de telle façon qu'on obtient un problème équivalent sans contraintes sur l'état

$$\widetilde{\mathcal{P}}_{T,x} \qquad \begin{cases} \text{Minimiser } \varphi(y(T)), \\ y(0) = x, \\ \dot{y}(t) = \lambda(t)\cdot f(y(t),\alpha(t)), \quad \alpha(t)\in\mathcal{A},\ \lambda(t)\in\Lambda(y(t))\ p.p.\ t \geq 0, \end{cases}$$

mais avec une dynamique s.c.s. définie par:

$$\Lambda(y) := \begin{cases} \{1\} & \text{si } y \in \overset{\circ}{\mathcal{K}}, \\ [0,1] & \text{si } y \in \partial\mathcal{K}, \\ \{0\} & \text{si } y \in \mathcal{K}^c, \end{cases} \tag{18}$$

et

$$\mathcal{F}(y) := \{\lambda f(y,\alpha),\ \lambda\in\Lambda(y),\ \alpha\in\mathcal{A}\}. \tag{19}$$

On s'intéresse désormais au cas particulier où φ est définie par

$$(A0') \qquad \varphi(x) := \begin{cases} 0 \text{ si } x \in \mathcal{C}, \\ 1 \text{ sinon.} \end{cases}$$

En utilisant les outils classiques en contrôle optimal, notamment ceux proposés par Bardi et Capuzzo Dolcetta [4], on prouve que

Proposition 0.4. *Sous les hypothèses (A0'), (A1)et (A2), ϑ vérifie l'équation suivante*

$$\min\left\{\vartheta_t(t,x)+\max_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}}\{-\lambda f(x,\alpha).\vartheta_x(t,x)\};\vartheta(t,x)-\chi_{\mathcal{K}}(x)\right\}=0,\ t>0,\ x\in\mathbb{R}^n. \tag{20}$$

où la fonction

$$\chi_{\mathcal{K}}(x)=\begin{cases}0 & si\ x\in\mathcal{K},\\ 1 & si\ x\in\mathcal{K}^c.\end{cases}$$

En deuxième lieu on s'intéresse aux outils de l'analyse non lisse et à l'approche de Frankowska et al. par la notion de dérivée contingente. On étend ainsi certains des résultats de Frankowska [42] et de Frankowska et Vinter [45] au cas où la dynamique est uniquement s.c.s. et définie par (19) et (18).

Rappelons la notion d'épidérivée définie pour une fonction $v:\mathbb{R}^n\to\mathbb{R}\cup\{\pm\infty\}$ en $x\in Domv:=\{x\in\mathbb{R}^n,\ v(x)\notin\{\pm\infty\}\}$ dans la direction $u\in\mathbb{R}^n$ par:

$$D_{\uparrow}v(x)(u)=\liminf_{\substack{h\to0^+\\ u'\to u}}\frac{v(x+hu')-v(x)}{h},$$

et introduisons une nouvelle notion d'épidérivée, qu'on appelle $\mathcal{F}$-épidérivée contingente, définie par

$$D^{\mathcal{F}}_{\uparrow}v(t,x)=\sup_{\substack{\alpha\in A,\ \tau>0\\ y(0)=x,\ \dot y(\theta)=f(y(\theta),\alpha(\theta)),\ \theta\in[-\tau,0]}}\liminf_{h\to0^+}\frac{v(t+h,y(-h))-v(t,x)}{h}.$$

On montre grâce à cette nouvelle notion le résultat d'unicité suivant

Théorème 0.5. *Sous les hypothèses (A0'), (A1), (A2), la fonction valeur ϑ est l'unique fonction s.c.i. à valeurs dans $\{0,1\}$ vérifiant pour tout (t,x) dans $]0,+\infty[\times\mathcal{K}$:*

$$\begin{aligned}
&\liminf_{\substack{\tau\to0^+\\ \xi\to x,\ \xi\in\mathring{\mathcal{K}}}}\vartheta(\tau,\xi)=\vartheta(0,x)=\varphi(x), && \forall x\in\mathcal{K},\\
&\inf_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}}D_{\uparrow}\vartheta(t,x)(-1,\lambda f(x,\alpha))\le0, && \forall t\ge0,\ x\in\mathbb{R}^n,\\
&\sup_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\{0,1\}\cap\Lambda(x)}}D_{\uparrow}\vartheta(t,x)(1,-\lambda f(x,\alpha))\le0, && \forall t\ge0,\ x\in\mathring{\mathcal{K}},\\
&D^{\mathcal{F}}_{\uparrow}\vartheta(t,x)\le0, && \forall t\ge0,\ x\in\partial\mathcal{K}.
\end{aligned}$$

Comme on le verra dans le chapitre 5, la dernière inégalité permet de déduire la croissance de ϑ le long des trajectoires admissibles de $\mathcal{P}_{T,x}$ qui touchent le bord $\partial\mathcal{K}$, sans supposer aucune hypothèse de qualification.

Bibliography

[1] R. Abgrall and S. Augoula. High order numerical discretization for hamilton-jacobi equations on triangular meshes. *J. Scientific Computing*, 15(2):197–229, 2000.

[2] O. Alvarez, E. Carlini, Y. LeBouar, and R. Monneau. Dislocation dynamics described by non-local Hamilton Jacobi equations. *Materials science & Engineering*, pages 162–165, 2005.

[3] J.P. Aubin. *Viability theory*. Birkhauser, 1991.

[4] M. I. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and viscosity solutions of Hamilton Jacobi Bellman equations*. Birkhäuser Boston, 1997.

[5] G. Barles. Remarks on a flame propagation model. *Rapport de recherche INRIA RR 451*, pages 1–38, 1985.

[6] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications*. Springer, Paris, 1994.

[7] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[8] E. N. Barron and R. Jensen. Semicontinuous viscosity solutions for Hamilton Jacobi equations with convex hamiltonian. *Comm. Partial Differential equations*, 15:1713–1742, 1990.

[9] E. N. Barron and R. Jensen. Optimal control and semi-continuous viscosity solutions. *Proc. Amer. Math. Soc.*, 113(2):397–402, 1991.

[10] T.P. Bauer, K.P. Zondervan, J.T. Betts, and W.P. Huffman. Solving the optimal control problem using a nonlinear programming technique. part 1: General formulation. *American Institute of aeronautics and astronautics. Astrodynamics conference, Seattle.*, 1984.

[11] T.P. Bauer, K.P. Zondervan, J.T. Betts, and W.P. Huffman. Solving the optimal control problem using a nonlinear programming technique. part 2: Optimal shuttle ascent trajectories. *American Institute of aeronautics and astronautics. Astrodynamics conference, Seattle.*, 1984.

[12] T.P. Bauer, K.P. Zondervan, J.T. Betts, and W.P. Huffman. Solving the optimal control problem using a nonlinear programming technique. part 3: Optimal shuttle reentry trajectories. *Proceedings of the AIAA/AAS Astrodynamics conference, Seattle.*, 1984.

[13] R. Bellman. *Dynamic programming*. Princeton university press, 1961.

[14] J.T. Betts. Survey of numerical methods for trajectory optimization. *Journal of Guidance Control and Dynamics*, 21(2):193–207, 1998.

[15] J.T. Betts. *Practical methods for optimal control using nonlinear programming*. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[16] J.T. Betts, S.K. Eldersveld, P.D. Frank, and J.G. Lewis. An interior point algorithm for large scale optimization. in large scale pde-constrained optimization. *Lect. notes Comput. Sci. Eng.*, 30:184–198, 2003.

[17] O. Bokanowski, S. Martin, R. Munos, and H. Zidani. An anti-diffusive scheme for viability problems. *Applied Numerical Mathematics*, 56:1147–1162, 2006.

[18] O. Bokanowski, N. Megdich, and H. Zidani. An adaptative antidissipative method for optimal control problems. *INRIA Report RR-5770*, 2005.

[19] O. Bokanowski, N. Megdich, and H. Zidani. A method for optimal control problems. an adaptative antidissipative method for optimal control problems. *ARIMA*, 5:256–271, 2006.

[20] O. Bokanowski and H. Zidani. Anti-dissipative schemes for advection and application to Hamilton Jacobi Bellman equations. *J. Sci. Comp.*, 30(1):1–33, 2007.

[21] F. Bonnans, P. Martinon, and E. Trélat. Singular arcs in the generalized goddard's problem. *accepted in J. Optim. Theory Appl.*, 2007.

[22] F. Bonnans and P. Rouchon. *Commande et optimisation de systèmes dynamiques*. Les éditions de l'école polytechnique, 2005.

[23] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimisation numérique*. Springer, 1997.

[24] J.F. Bonnans and G. Launay. Large scale direct optimal control applied to the re-entry problem. *Journal of Guidance Control and Dynamics*, 21(6):996–1000, 1998.

[25] B. Bonnard, L. Faubourg, and E. Trélat. Optimal control of the atmospheric arc of a space shuttle and numerical simulations with multiple shooting method. *Mathematical Models and Methods in Applied Sciences*, 15(1):109–140, 2005.

[26] B. Bonnard and E. Trélat. Une approche géométrique du contrôle optimal de l'arc atmosphérique de la navette spatiale. *ESAIM: Control,Optimization and Calculus of Variations*, 7:179–222, 2002.

[27] N. Bérend, J.F. Bonnans, J. Laurent-Varin, M. Haddou, and C. Talbot. An interior point approach to trajectory optimization. *INRIA report n 5613*, 2005.

[28] E. Carlini, M. Falcone, N. Forcadel, and R. Monneau. Convergence of a generalized fast marching method for a non-convex eikonal equation. *preprint*, 2006.

[29] B. Cockburn and B. Yenikaya. An adaptive method with rigorous error control for the hamilton-jacobi equations. part i: the one-dimensional steady-state case. *App. Num. Math.*, 52:175–195, 2005.

[30] B. Cockburn and B. Yenikaya. An adaptive method with rigorous error control for the hamilton-jacobi equations. part ii: the two-dimensional steady-state case. *J. Comput. ph.*, 209:391–405, 2005.

[31] J. M. Coron. *Control and nonlinearity*. Mathematical surveys and monographs, 2007.

[32] M. G. Crandall and P. L. Lions. Viscosity solutions of Hamilton Jacobi equations. *Tran. Amer. Math. Soc.*, 277:1–42, 1983.

[33] M.G. Crandall, L.C. Evans, and P.-L. Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc*, 282:487–502, 1984.

[34] M.G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of Computation*, 43:1–19, 1984.

[35] E. Cristiani. *Fast Marching and Semi Lagrangian Methods for Hamilton Jacobi Equations with Applications*. PH. D. Thesis, 2006.

[36] B. Després. Lax theorem and finite volume schemes. *Math. Comp.*, 73(247):1203–1234 (electronic), 2004.

[37] B. Després and F. Lagoutière. Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. *J.Sci. Comput.*, 16:479–524, 2001.

[38] M. Falcone. A numerical approach to the infinite horizon problem of deterministic control theory. *Applied Mathematics and Optimization*, 15:1–13, 1987.

[39] M. Falcone and R. Ferretti. Semi-Lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and Godunov methods. *Journal of Computational Physics*, 175:559–575, 2002.

[40] M. Falcone and T. Giorgi. An approximation scheme for evolutive Hamilton Jacobi equations. *Stochastic analysis, Control, Optimization and applications*, pages 289–303, 1999.

[41] M. Falcone, T. Giorgi, and P. Loretti. Level sets of viscosity solutions: some applications to fronts and rendez-vous problems. *SMAI J. Appl. Math.*, 54(5):1335–1354, 1994.

[42] F. Frankowska. Lower semi-continuous solutions of hamilton-jacobi-equations. *SIAM J.Control Optim.*, 31:257–272, 1993.

[43] F. Frankowska and S. Plaskacz. Semicontinuous solutions of hamilton-jacobi equations with state constraints. *Differential inclusions and optimal control, Lecture notes in nonlinear analysis*, 2:145–161, 1998.

[44] F. Frankowska and S. Plaskacz. Semicontinuous solutions of Hamilton Jacobi equations with degenerate state constraints. *JMAA*, pages 818–838, 2000.

[45] F. Frankowska and R. B. Vinter. Existence of neighboring feasible trajectories: applications to dynamic programming for state constrained optimal control problems. *I.Optim.Theory Appl.*, 104:27–40, 2000.

[46] I. Gargantini. An effective way to represent quadtrees. *Communications of the ACM*, 25(12):905–910, 1982.

[47] A. Ghorbel and R. Monneau. Equation d'Hamilton Jacobi non locale modélisant la dynamique des dislocations. *Acte du congrès Tendances des Applications Mathématiques en Tunisie, Algérie et Maroc*, pages 322–328, 2005.

[48] E. Godlewski and P. A. Raviart. *Hyperbolic Systems of Conservation laws.* Mathématiques et applications, Ellipses.

[49] E. Godlewski and P. A. Raviart. *Numerical approximation of Hyperbolic Systems of Conservation laws.* Applied Mathematical Sciences, Springer.

[50] L. Grune. Adaptative grid generation for evolutive Hamilton Jacobi Bellman equations. *Numerical methods for viscosity solutions and applications*, pages 153–172, 2001.

[51] Th. Guilbaud. *Méthodes numériques pour la commande optimale. Thèse de doctorat, Université Paris 6.* 2002.

[52] H. Ishii and S. Koike. A new formulation of state constraint problems for first order pdes. *SIAM J. Control and Optimization*, 34(2):554–571, 1996.

[53] F. Lagoutière. *Modélisation mathématique et résolution numérique de problèmes de fluides compressibles à plusieurs constituants. Thèse de doctorat, Université Paris 6.* 2000.

[54] J. Laurent-Varin. *Calcul de trajectoires optimales de lanceurs spatiaux réutilisables par une méthode de point intérieur.* Thèse de doctorat de l'Ecole Polytechnique, 2005.

[55] P. L. Lions and P. E. Souganidis. Convergence of muscl and filtered schemes for scalar conservation laws and Hamilton Jacobi equations. *Numer. Math.*, 69:441–470, 1995.

[56] R. Munos and A. Moore. Influence and variance of markov chain: Application to adaptive discretization in optimal control. *IEEE conference on decision and control*, 1999.

[57] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer, 1999.

[58] S. Osher and J.A. Sethian. Fronts propagating with curvature dependent speed: algorithms based on Hamilton Jacobi formulations. *Journal of computational Physics*, pages 12–49, 1988.

[59] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes*. Interscience, NewYork, 1962.

[60] J. A. Sethian. *Level Set Methods and Fast Marching Methods. Evolving interfaces in computational geometry , fluid mechanics, computer vision and materials science*. Cambridge university press, 1999.

[61] C. W. Shu and S. Osher. Efficient implementation of essentially non oscillatory shock capturing schemes. *Journal of computational physics*, 77:439–471, 1988.

[62] C. W. Shu and S. Osher. Efficient implementation of essentially non oscillatory shock capturing schemes, ii. *Journal of computational physics*, 83:32–78, 1989.

[63] H. M. Soner. Optimal control with state space constraint. *SIAM Journal of Control and Optimization*, 24(3):552–561, 1986.

[64] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM Journal of Numerical Analysis*, 21(5):995–1011, 1984.

[65] E. Trélat. *Controle optimal: théorie et applications*. Vuibert, collection "mathématiques concrètes", 2005.

CHAPTER 1

Convergence of a non monotone scheme for HJB equations

Abstract. We prove the convergence of a non-monotonous scheme for a one-dimensional first order Hamilton-Jacobi-Bellman equation of the form $v_t + \max_\alpha(f(x,\alpha)v_x) = 0$, $v(0,x) = v_0(x)$. The scheme is related to the HJB-UltraBee scheme suggested in [7]. We show for general discontinuous initial data a first-order convergence of the scheme, in L^1-norm, towards the viscosity solution. We also illustrate the non-diffusive behavior of the scheme on several numerical examples.

1.1 Introduction

We consider the following first order Hamilton-Jacobi-Bellman equation:

$$\vartheta_t(t,x) + \max_{\alpha\in\mathcal{A}}(f(x,\alpha)\ \vartheta_x(t,x)) = 0, \quad t>0,\ x\in\mathbb{R}, \tag{1.1.1a}$$

$$\vartheta(0,x) = v_0(x), \quad x\in\mathbb{R}, \tag{1.1.1b}$$

with discontinuous initial data v_0. In optimal control theory, the solution ϑ of equation (1.1.1) corresponds to the value function of an optimization problem [3, 2]. It is usual that this function, as well as the "final" cost v_0, is discontinuous (for instance for target or Rendez-Vous problems).

In the continuous case (v_0 is continuous), there are several contributions dealing with numerical schemes for the discretization of HJB equations [8, 11, 1, 14]. In [4], Barles and Souganidis give a general framework for the convergence of approximated solutions towards the viscosity solution, under generic monotonicity stability and consistency assumptions. In that case, an L^∞-error bound in $\Delta x^{\frac{\gamma}{2}}$ is exhibited [8, 10], Δx being the mesh size, whenever the function v_0 is bounded γ-Hölder ($\gamma\in(0,1]$).

Nevertheless, when we deal with discontinuous intial data v_0, classical monotone schemes are no more adapted. In fact, if we attempt to use these schemes, we observe an increasing numerical diffusion around discontinuities, and this is due to the fact that monotone schemes use at some level finite differences and/or interpolation technics.

In this paper, we analyse an explicit scheme for the numerical resolution of (1.1.1), closely related to the HJB-UltraBee scheme proposed in [7]. We give a convergence proof, show anti-dissipative properties of the scheme, and give a L^1-error estimate.

The UltraBee scheme has been developed to study compressible gas dynamics [9], and more precisely to solve the transport equation. A generalization to HJB equations and many academic tests have been done to evaluate the behaviour of the scheme when dealing with discontinuities [7]. Its comparison with the viability algorithm [5] [16] was encouraging to study more deeply convergence results.

Let us stress on that this scheme is explicit and non-monotonous (neither ϵ-monotone in the sense of [1, 14]). As far as we know, there are few non-monotone schemes that have been proved to converge for HJ equations. In [13], Lions and Souganidis show the convergence of a TVD second order scheme, but which is implicit.

For a large class of discontinuous initial data v_0, and under some assumptions on the dynamics f (see assumption (H3) in Section 1.2), we obtain a first-order error bound in L^1

norm, of the following form:

$$\|V(t_n,.) - \vartheta(t_n,.)\|_{L^1(\mathbb{R})} \leq C(L,t_n,v_0)\Delta x \quad \forall t_n \geq 0, \tag{1.1.2}$$

where ϑ is the viscosity solution of (1.1.1), V is the numerical approximation and $C(L,t_n,v_0)$ is a positive constant which depends only on t_n, on L (Lipschitz constant of $x \to f(x,\alpha)$, see Section 1.2.1), t_n, and on the total variation of v_0 (see Definition 1.8).

This is the first result of this kind to our knowledge, in the case of discontinuous viscosity solutions. Furthermore, in some particular cases, such as the eikonal equation ($\vartheta_t + |\vartheta_x| = 0$, corresponding to $\mathcal{A} := \{\pm 1\}$ and $f(x,\alpha) = \alpha$), the constant C does not depend of t_n. This shows a "non-diffusive" behavior of the scheme, as identified in [9] for the advection case with constant sign velocity.

The paper is organized as follows: Sections 1.2 and 1.3 are devoted to the proof of (1.1.2) in the case of piece-wise constant initial data. In Section 1.4, we prove (1.1.2) for more general discontinuous initial data. In Section 1.5, we weaken some assumptions made before on the velocities and prove a similar estimate for a modified scheme. We conclude in Section 6 by some numerical illustrations, in particular showing the non-diffusive behavior of the proposed scheme. The appendix contains some useful technical results.

1.2 Preliminaries

1.2.1 Notations and preliminary results

We denote by $v_0 : \mathbb{R} \to \mathbb{R}$ a bounded lower semicontinuous (l.s.c.) function, $\mathcal{A}$ a compact set, and $f : \mathbb{R} \times \mathcal{A} \to \mathbb{R}$ a continuous function satisfying:

$$\exists L \geq 0,\ \forall \alpha \in \mathcal{A},\ \forall x,y \in \mathbb{R},\ |f(y,\alpha) - f(x,\alpha)| \leq L|y-x|. \tag{1.2.1}$$

It is known that, under these assumptions, equation (1.1.1) admits a unique bounded l.s.c. *bilateral viscosity* solution [3, 2]. For convenience of the reader, we recall in the following definition the viscosity notion we use.

Definition 1.1. *A bounded l.s.c. function ϑ is a bilateral viscosity solution of (1.1.1) if,*

i) for any $\phi \in C^1((0,+\infty)\times\mathbb{R})$ and at any local minimum $(t,x) \in]0,+\infty[\times\mathbb{R}$ of $\vartheta - \phi$,

$$\phi_t(t,x) + \max_{\alpha\in\mathcal{A}}(f(x,\alpha)\,\phi_x(t,x)) = 0.$$

ii) $\liminf\limits_{\substack{t\to 0^+ \\ y\to x}} \vartheta(t,y) = v_0(x), \quad \forall x \in \mathbb{R}.$

If we set for all $x \in \mathbb{R}$

$$f_m(x) := \min_{\alpha\in\mathcal{A}} f(x,\alpha) \quad \text{and} \quad f_M(x) := \max_{\alpha\in\mathcal{A}} f(x,\alpha),$$

then equation (1.1.1) can be rewritten in the equivalent form:

$$\vartheta_t(t,x) + \max(f_m(x)\ \vartheta_x(t,x),\ f_M(x)\ \vartheta_x(t,x)) = 0, \qquad t>0,\ x\in\mathbb{R}, \quad (1.2.2a)$$
$$\vartheta(0,x) = v_0(x), \quad x\in\mathbb{R}. \quad (1.2.2b)$$

Notice that by (1.2.1) and the definitions of f_m and f_M, we have $f_m(x) \le f_M(x)$, $\forall x\in\mathbb{R}$, and also

(**H1**) f_m and f_M are L-Lipschitz continuous.

For simplicity of presentation, we first suppose the simplifying additional assumptions:

(**H2**) f_m and f_M are of constant sign,
(**H3**) f_m and f_M are non-decreasing functions on $\mathbb{R}$.

These assumptions will be weakened in Section 1.5.

Remark 1.2. *Assumptions (H1)-(H3) are satisfied in the particular case of the Eikonal equation:* $\vartheta_t + c|\vartheta_x| = 0$, *where* c *is a given constant and* $c\ge 0$ *(taking* $f_M(x) = -f_m(x) = c$*).*

We now define exact and approximated characteristics that will be very useful throughout the paper. Let $x_j := j\,\Delta x$ be a uniform mesh with $\Delta x>0$ and $j\in\mathbb{Z}$, and denote:

$$x_{j+\frac12} := (j+\frac12)\,\Delta x, \quad \text{and} \quad I_j :=]x_{j-\frac12}, x_{j+\frac12}[.$$

As the dynamics f_m and f_M are lipschitz continuous, then for any $x\in\mathbb{R}$ we can define characteristics X_x^M and X_x^m as the solutions of the Cauchy problems:

$$\begin{cases} \dot X_x^M(t) = f_M(X_x^M(t)), \\ X_x^M(0) = x, \end{cases} \quad \text{and} \quad \begin{cases} \dot X_x^m(t) = f_m(X_x^m(t)), \\ X_x^m(0) = x. \end{cases} \quad (1.2.3)$$

We also define approximated piece-wise constant velocity functions f_M^S and f_m^S, such that, for $j\in\mathbb{Z}$:

$$f_M^S(x) = f_M(x_j), \quad \forall x\in I_j,$$
$$f_M^S(x_{j+\frac12}) = \begin{cases} 0 & \text{if } f_M(x_j)\ f_M(x_{j+1}) \le 0, \\ f_M(x_j) & \text{otherwise,} \end{cases}$$

and

$$f_m^S(x) = f_m(x_j), \quad \forall x\in I_j,$$
$$f_m^S(x_{j+\frac12}) = \begin{cases} 0 & \text{if } f_m(x_j)\ f_m(x_{j+1}) \le 0, \\ f_m(x_j) & \text{otherwise.} \end{cases}$$

In general, the differential equation

$$\dot\chi_x(t) = f_M^S(\chi_x(t)),\ a.e.\ t\ge 0, \quad \chi_x(0) = x, \quad (1.2.4)$$

may have more than one absolutely continuous solution. The non-uniqueness comes from the behavior on boundary points $(x_{j+\frac{1}{2}})$ in the case when the velocity vanishes (or changes sign). Throughout this paper, we shall denote by $X_x^{M,S}$ the function defined by:

$$X_x^{M,S} \text{solution of (1.2.4)}, \quad \text{and} \tag{1.2.5a}$$

$$\text{if } \exists t^* \geq 0,\ \exists j \in \mathbb{Z} \text{ s.t. } \begin{cases} X_x^{M,S}(t^*) = x_{j+\frac{1}{2}}, \\ f_M(x_j) f_M(x_{j+1}) \leq 0 \end{cases} \quad \text{then } X_x^{M,S}(t) = x_{j+\frac{1}{2}}\ \forall t \geq t^*. \tag{1.2.5b}$$

We have uniqueness of such solution (see appendix 1.7.1). We construct $X_x^{m,S}$ in a similar way.

Remark 1.3. *Under assumption (H2), $f_S^M(x_j) f_S^M(x_{j+1}) \leq 0$ happens only if $f_S^M(x_j) = 0$ or $f_S^M(x_{j+1}) = 0$. However, in Section 5, we shall use the definition (1.2.5) of $X_x^{M,S}$ also for changing sign velocities.*

Lemma 1.4. *Assume that (H1) holds. Let a, b be in $\mathbb{R}$. The following assertions are satisfied:*
(i) For every $t \geq 0$, we have:

$$\max\left(|X_a^{M,S}(t) - X_a^M(t)|,\ |X_b^{m,S}(t) - X_b^m(t)|\right) \leq \frac{1}{2} L t\ e^{Lt}\ \Delta x,$$

where L is the Lipschitz constant of f_M and f_m (see (H1)).
(ii) Let $s \leq t$ and assume that $X_a^{m,S}(\theta) \geq X_b^{M,S}(\theta)$, for every $\theta \in [s,t]$. Then

$$|X_a^{m,S}(t) - X_b^{M,S}(t)| + \Delta x \leq e^{L(t-s)}(|X_a^{m,S}(s) - X_b^{M,S}(s)| + \Delta x).$$

Moreover, if $X_a^m(\theta) \geq X_b^M(\theta)$, for every $\theta \in [s,t]$, then

$$|X_a^m(t) - X_b^M(t)| \leq e^{L(t-s)}(|X_a^m(s) - X_b^M(s)|).$$

(iii) Assume $(H1)$ and $(H3)$. If $a > b$, then the functions $t \longmapsto X_a^{M,S}(t) - X_b^{m,S}(t)$ and $t \longmapsto X_a^M(t) - X_b^m(t)$ are non-decreasing for $t \geq 0$.
(iv) Assume $(H1) - (H3)$. If $\frac{\Delta x}{2} < a - b$, then the function $t \longmapsto X_a^{M,S}(t) - X_b^m(t)$ is non-decreasing for $t \geq 0$.
(v) If $X_b^{m,S}(\theta) \geq X_a^{m,S}(\theta)$, for every $\theta \in [s,t]$. Then we have

$$|X_b^{m,S}(t) - X_a^{m,S}(t)| + \Delta x \leq e^{L(t-s)}(|X_b^{m,S}(s) - X_a^{m,S}(s)| + \Delta x).$$

Moreover, if $X_a^{M,S}(\theta) \leq X_b^{M,S}(\theta)$, for every $\theta \in [s,t]$, then

$$|X_b^{M,S}(t) - X_a^{M,S}(t)| + \Delta x \leq e^{L(t-s)}(|X_b^{M,S}(s) - X_a^{M,S}(s)| + \Delta x).$$

Proof. (i) Let $x \in \mathbb{R}$ and $j \in \mathbb{Z}$ such that $x \in I_j$. For $y \in \mathbb{R}$, the following inequality holds

$$\begin{aligned} |f_M^S(x) - f_M(y)| &= |f_M(x_j) - f_M(x) + f_M(x) - f_M(y)| \\ &\leq \frac{1}{2}L\Delta x + L|x-y|. \end{aligned}$$

Therefore, for all $t \geq 0$, we get:

$$\begin{aligned} |X_a^{M,S}(t) - X_a^M(t)| &= |\int_0^t \left(f_M^S(X_a^{M,S}(s)) - f_M(X_a^M(s))\right) ds| \\ &\leq \frac{1}{2}Lt\Delta x + L\int_0^t |X_a^{M,S}(s) - X_a^M(s)| ds. \end{aligned}$$

Hence by using Gronwall's Lemma we obtain the desired estimate for $|X_a^{M,S}(t) - X_a^M(t)|$. The bound for $|X_b^{m,S}(t) - X_b^m(t)|$ is obtained in the same way.

(ii) Let $\delta(\theta) := X_a^{m,S}(\theta) - X_b^{M,S}(\theta) + \Delta x$. We have

$$\begin{aligned} \frac{d}{d\theta}\delta(\theta) &= f_m^S(X_a^{m,S}(\theta)) - f_M^S(X_b^{M,S}(\theta)) \\ &\leq f_M^S(X_a^{m,S}(\theta)) - f_M^S(X_b^{M,S}(\theta)) \\ &\leq L\Big(X_a^{m,S}(\theta) - X_b^{M,S}(\theta) + \Delta x\Big) = L\ \delta(\theta). \end{aligned}$$

The result follows by using a Gronwall estimate. The proof for the second estimate is similar.

(iii) Define $\delta(t) := X_a^{M,S}(t) - X_b^{m,S}(t)$, and $t^* := \inf\{t > 0, \delta(t) < 0\}$. As $\delta(0) > 0$, then $t^* > 0$. Then for all $t \in [0, t^*[$, $\delta(t) \geq 0$ and we have:

$$\frac{d}{dt}\delta(t) = f_M^S(X_a^{M,S}(t)) - f_m^S(X_b^{m,S}(t)) \geq f_m^S(X_a^{M,S}(t)) - f_m^S(X_b^{m,S}(t)),$$

which is positive, for $t \in [0, t^*[$, by (H3). We deduce that δ is non-decreasing on $[0, t^*[$. Suppose that t^* is finite, then we get by continuity of $X_a^{M,S}$ and $X_b^{m,S}$ that $\delta(t^*) \geq \delta(0) > 0$. This contradiction shows that $t^* = +\infty$ and δ is non-decreasing for all $t \geq 0$. (The proof is similar for $t \longmapsto X_a^M(t) - X_b^m(t)$.)

(iv) Similar arguments as in (iii)

(v) The proof is obtained as in (ii). □

Lemma 1.5. *Let v_0 be a bounded l.s.c. function on $\mathbb{R}$, and assume that (H1) holds. Then, the unique viscosity solution of (1.2.2) is given by:*

$$\vartheta(t,x) = \min_{y \in [X_x^M(-t), X_x^m(-t)]} v_0(y), \quad \forall t > 0,\ x \in \mathbb{R}. \tag{1.2.6}$$

Proof. Notice that equation (1.2.2a) can be rewritten as follows

$$\vartheta_t(t,x) + \max_{\alpha \in [0,1]} \left\{((1-\alpha)f_m(x) + \alpha f_M(x)) \cdot \vartheta_x(t,x)\right\} = 0,\ t > 0,\ x \in \mathbb{R}. \tag{1.2.7}$$

The unique viscosity solution of equation (1.2.7) satisfying the initial condition (1.2.2b) (see [2]) is given by

$$\vartheta(t,x) = \min_{\alpha\in L^\infty(\mathbb{R}^+,[0,1])} v_0(X_x^\alpha(-t)) = \min_{y\in[X_x^M(-t),X_x^m(-t)]} v_0(y),$$

where X_x^α is the solution of $X_x^\alpha(0) = x$ and $\dot{X}_x^\alpha(t) = (1-\alpha(t))f_m(X_x^\alpha(t)) + \alpha(t)f_M(X_x^\alpha(t))$ for $t \geq 0$ with $\alpha \in L^\infty(\mathbb{R}^+,[0,1])$. □

We also consider the function ϑ^S which is defined in an analogous way as in (1.2.6), but with the approximated characteristics $X_x^{M,S}, X_x^{m,S}$ instead of X_x^M and X_x^m:

$$\vartheta^S(t,x) := \min_{y\in[X_x^{M,S}(-t),X_x^{m,S}(-t)]} v_0(y), \quad \forall t>0,\ x\in\mathbb{R}. \tag{1.2.8}$$

This approximate function will play an important role throughout the paper.

Proposition 1.6. *Under assumption (H1), we have:*

$$\|\vartheta(t,.) - \vartheta^S(t,.)\|_{L^1(\mathbb{R})} \leq Lte^{Lt}\, TV(v_0)\, \Delta x, \tag{1.2.9}$$

where $TV(v_0)$ denotes the total variation *of v_0 (see Definition 1.8).*

Proof. By using Lemma 1.38 (taking $a_x^1 = X_x^M(-t)$, $b_x^1 = X_x^m(-t)$, and $a_x^2 = X_x^{M,S}(-t)$, $b_x^2 = X_x^{m,S}(-t)$) (whose inverse functions are $X_x^M(t)$, $X_x^m(t)$, and $X_x^{M,S}(t)$, $X_x^{m,S}(t)$ respectively) together with Lemma 1.4 (i), we obtain the L^1-norm estimate. □

Using Lemma 1.40, we also obtain

Proposition 1.7. *Under assumption (H1),*

$$TV(\vartheta^S(t,.)) \leq TV(v_0), \quad \forall t\geq 0.$$

For $E\subset\mathbb{R}$ a given set, we shall use in all the sequel the notation $\mathbf{1}_E$ for the function defined by:

$$\mathbf{1}_E(x) := \begin{cases} 1 & \text{if } x\in E,\\ 0 & \text{otherwise.}\end{cases}$$

Throughout this paper, we will also use the following definition.

Definition 1.8. *Let w be a real-valued function. the total variation of w is defined by:*

$$TV(w) := \sup\left\{\sum_{j=1,\dots,k} |w(y_{j+1}) - w(y_j)|;\ k\in\mathbb{N}^*, \text{and } (y_j)_{1\leq j\leq k+1} \text{ non-decreasing sequence}\right\}.$$

1.2.2 The HJB-UltraBee scheme

Let $\Delta t > 0$ be a constant time step, and $t_n := n\Delta t$ for $n \geq 0$. We set the following notation for local "CFL" numbers:

$$\nu_j^m := \frac{\Delta t}{\Delta x} f_m(x_j) \quad \text{and} \quad \nu_j^M := \frac{\Delta t}{\Delta x} f_M(x_j),$$

and $\nu^m = \{\nu_j^m,\ j \in \mathbb{Z}\}$, $\nu^M = \{\nu_j^M,\ j \in \mathbb{Z}\}$.

An adaptation of the UltraBee scheme has been proposed and numerically tested for the HJB equation [7, 5, 6]. Let us recall this formulation. We first introduce the notation for the average values of initial data:

$$V_j^0 := \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v_0(x)dx, \quad j \in \mathbb{Z}. \tag{1.2.10}$$

As we deal with an explicit scheme, we will assume in all the paper that the mesh size satisfies the CFL condition:

$$\max_{x\in\mathbb{R}} \max(|f_m(x)|, |f_M(x)|) \frac{\Delta t}{\Delta x} \leq 1. \tag{1.2.11}$$

Notice that under this condition, the CFL numbers ν^m and ν^M satisfy

$$|\nu_j^m| \in [0,1] \quad \text{and} \quad |\nu_j^M| \in [0,1], \quad \forall j \in \mathbb{Z}.$$

Algorithm 1 :

Initialization: We compute the initial averages $V^0 = (V_j^0)_{j\in\mathbb{Z}}$ as defined in (1.2.10).

Loop: For $n \geq 0$, We compute $V^{n+1} = (V_j^{n+1})_{j\in\mathbb{Z}}$ by:

- Define "fluxes" $V_{j+\frac{1}{2}}^n(\nu)$ for $\nu \in \{\nu^m, \nu^M\}$ as follows:
  If $\nu_j \geq 0$ for every $j \in \mathbb{Z}$, define:

$$V_{j+\frac{1}{2}}^n(\nu) := \begin{cases} \min(\max(V_{j+1}^n, b_j^+(\nu)), B_j^+(\nu)) & \text{if } \nu_j > 0 \\ V_{j+1}^n & \text{if } \nu_j = 0 \text{ and } V_j^n \neq V_{j-1}^n \\ V_j^n & \text{if } \nu_j = 0 \text{ and } V_j^n = V_{j-1}^n, \end{cases}$$

  where

$$\begin{cases} b_j^+(\nu) := \max(V_j^n, V_{j-1}^n) + \frac{1}{\nu_j}(V_j^n - \max(V_j^n, V_{j-1}^n)), \\ B_j^+(\nu) := \min(V_j^n, V_{j-1}^n) + \frac{1}{\nu_j}(V_j^n - \min(V_j^n, V_{j-1}^n)), \end{cases} \tag{1.2.12}$$

  If $\nu_j \leq 0$ for every $j \in \mathbb{Z}$, define:

$$V_{j-\frac{1}{2}}^n(\nu) := \begin{cases} \min(\max(V_{j-1}^n, b_j^-(\nu)), B_j^-(\nu)) & \text{if } \nu_j < 0 \\ V_{j-1}^n & \text{if } \nu_j = 0 \text{ and } V_j^n \neq V_{j+1}^n \\ V_j^n & \text{if } \nu_j = 0 \text{ and } V_j^n = V_{j+1}^n, \end{cases}$$

where

$$\begin{cases} b_j^-(\nu) := \max(V_j^n, V_{j+1}^n) + \frac{1}{|\nu_j|}(V_j^n - \max(V_j^n, V_{j+1}^n)), \\ B_j^-(\nu) := \min(V_j^n, V_{j+1}^n) + \frac{1}{|\nu_j|}(V_j^n - \min(V_j^n, V_{j+1}^n)). \end{cases} \tag{1.2.13}$$

- For $\nu \in \{\nu^m, \nu^M\}$, let

$$V_j^{n+1}(\nu) := V_j^n - \nu_j \left(V_{j+\frac{1}{2}}^n(\nu) - V_{j-\frac{1}{2}}^n(\nu) \right).$$

- Finally, set $V_j^{n+1} := \min\left(V_j^{n+1}(\nu^m), V_j^{n+1}(\nu^M) \right)$ for every $j \in \mathbb{Z}$.

In all the sequel, we shall use the notation:

$$\mathcal{S}_{UB}(V^n) := \left(\min\left(V_j^{n+1}(\nu^m), V_j^{n+1}(\nu^M) \right) \right)_{j \in \mathbb{Z}}.$$

Under assumption (H2), we notice that the resulting scheme is well defined. We associate to the scheme values $(V_j^n)_j$, the l.s.c. step function $V(t_n, .)$ defined for every $t_n \geq 0$, $x \in \mathbb{R}$ by

$$V(t_n, x) := \begin{cases} V_j^n & \text{if } x \in]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[, \\ \min(V_j^n, V_{j+1}^n) & \text{if } x = x_{j+\frac{1}{2}}. \end{cases} \tag{1.2.14}$$

1.2.3 A first case where fronts do not meet

We consider here the case of an initial data of the form

$$v_0(x) := \mathbf{1}_{]-\infty, b[}(x) + \mathbf{1}_{]a, +\infty[}(x), \tag{1.2.15}$$

where $a, b \in \mathbb{R} \cup \{\pm\infty\}$ and $a > b$. Our aim in this section is to study in this simple case the relationship between the viscosity solution at time t_n, $\vartheta(t_n, .)$ and the values V^n computed by the HJB-UB algorithm. We show, under the CFL condition, an L^1-error estimate in Δx stated in the following theorem:

Theorem 1.9. *We assume that $(H1) - (H3)$ and the CFL condition (1.2.11) are satisfied. Let v_0 be defined by (1.2.15), Δx be such that $a \geq b + 3\Delta x$, and ϑ be the viscosity solution of (1.2.2). Then*

$$\|V(t_n, .) - \vartheta(t_n, .)\|_{L^1(\mathbb{R})} \leq (1 + Lt_n e^{Lt_n}) TV(v_0)\, \Delta x, \quad \forall n \geq 0. \tag{1.2.16}$$

Before dealing with the proof of Theorem 1.9, it will be useful to have the analytic expression of the viscosity solution ϑ (known in this case, since v_0 has a simple form):

Remark 1.10. *Assume that (H1) is satisfied, $a > b$, and v_0 as in (1.2.15). Then using Lemma 1.5 and by a direct calculation we obtain that the l.s.c. viscosity solution of (1.2.2) is given by :*

$$\vartheta(t,x) := \mathbf{1}_{]-\infty, X_b^m(t)[}(x) + \mathbf{1}_{]X_a^M(t),+\infty[}(x). \quad \forall t \geq 0,\ x \in \mathbb{R}.$$

Also, the function ϑ^S defined by (1.2.8) satisfies:

$$\vartheta^S(t,x) = \mathbf{1}_{]-\infty, X_b^{m,S}(t)[}(x) + \mathbf{1}_{]X_a^{M,S}(t),+\infty[}(x), \quad \forall t \geq 0,\ x \in \mathbb{R}.$$

In the following we denote by $\overline{\vartheta}^{S,n}$ the cell averages of $\vartheta^S(t_n,\cdot)$, defined by

$$\overline{\vartheta}_j^{S,n} := \frac{1}{\Delta x}\int_{I_j} \vartheta^S(t_n,x)\ dx \qquad \text{for } j \in \mathbb{Z},\ n \in \mathbb{N}. \tag{1.2.17}$$

Lemma 1.11. *Assume that $(H1)-(H3)$ hold, v_0 is defined by (1.2.15), $a \geq b + 3\Delta x$ and the CFL condition (1.2.11) satisfied. Then the values V_j^n computed by Algorithm 1 satisfy:*

$$V_j^n = \overline{\vartheta}_j^{S,n}, \quad \forall n \geq 0,\ \forall j \in \mathbb{Z}.$$

Proof. *Case 1.* We consider the case of $v_0(x) := \mathbf{1}_{]a,\infty[}(x)$, and proceed by recursion on $n \geq 0$ (the case $v_0(x) = \mathbf{1}_{]-\infty,b[}(x)$ may be treated in a similar way). We suppose that $V_j^n = \overline{\vartheta}_j^{S,n}$ for all $j \in \mathbb{Z}$.
Let $j \in \mathbb{Z}$ be such that the discontinuity position $x_n := X_a^{M,S}(t_n)$ lies in $]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. Then we have $V_k^n = 0$ for $k < j$, $V_j^n \in [0,1[$, and $V_k^n = 1$ for $k > j$. By straightforward calculations, if $\nu^M \geq 0$, we can verify that

$$\begin{array}{rcl}
\overline{\vartheta}_{j-1}^{S,n+1} = V_{j-1}^{n+1} & = & 0, \\
\overline{\vartheta}_j^{S,n+1} = V_j^{n+1} & = & \max(0, V_j^n - \nu_j^M), \\
\overline{\vartheta}_{j+1}^{S,n+1} = V_{j+1}^{n+1} & = & \left\{ \begin{array}{l} 1 - \frac{\nu_{j+1}^M}{\nu_j^M}\ \max(0,\ \nu_j^M - V_j^n) \quad \text{if } \nu_j^M > 0 \\ 1 \quad \text{if } \nu_j^M = 0 \end{array} \right.
\end{array}$$

and, if $\nu^M \leq 0$,

$$\begin{array}{rcl}
\overline{\vartheta}_{j-1}^{S,n+1} = V_{j-1}^{n+1} & = & \left\{ \begin{array}{l} \frac{|\nu_{j-1}^M|}{|\nu_j^M|}\ \big(\max(1, V_j^n + |\nu_j^M|) - 1\big) \quad \text{if } \nu_j^M < 0, \\ 0 \quad \text{if } \nu_j^M = 0 \end{array} \right. \\
\overline{\vartheta}_j^{S,n+1} = V_j^{n+1} & = & \min(1, V_j^n + |\nu_j^M|), \\
\overline{\vartheta}_{j+1}^{S,n+1} = V_{j+1}^{n+1} & = & 1,
\end{array}$$

and in all cases, $\overline{\vartheta}_k^{S,n+1} = V_k^{n+1} = 0$, $\forall k \leq j-2$, and $\overline{\vartheta}_k^{S,n+1} = V_k^{n+1} = 1$, $\forall k \geq j+2$. This means that the HJB-UltraBee scheme computes exactly $V^{n+1}(\nu^M)$ from V^n for the

advection with velocity f_M^S. Notice in particular that under the CFL condition (1.2.11) the discontinuity will not move more than one cell during a step of time. In the same way we obtain that $V^{n+1}(\nu^m)$ computes exactly the average values from V^n for the advection with velocity f_m^S. Hence we have, for every $k \in \mathbb{Z}$,

$$V_k^{n+1}(\nu^M) = \frac{1}{\Delta x} \int_{I_k} \mathbf{1}_{]X_{x_n}^{M,S}(\Delta t),\infty[}(x)dx, \tag{1.2.18}$$

$$V_k^{n+1}(\nu^m) = \frac{1}{\Delta x} \int_{I_k} \mathbf{1}_{]X_{x_n}^{m,S}(\Delta t),\infty[}(x)dx. \tag{1.2.19}$$

Since $X_{x_n}^{m,S}(\Delta t) \leq X_{x_n}^{M,S}(\Delta t)$ we deduce that $V_k^{n+1}(\nu^M) \leq V_k^{n+1}(\nu^m)$, and, for all $k \in \mathbb{Z}$,

$$V_k^{n+1} = V_k^{n+1}(\nu^M).$$

This concludes the proof of $\overline{\vartheta}^{S,n+1} = V^{n+1}$.

Case 2. Consider the case of $v_0(x) := \mathbf{1}_{]-\infty,b[}(x) + \mathbf{1}_{]a,\infty[}(x)$. As $a - b \geq 3\Delta x$, by Lemma 1.4 *(iii)* and (H3), we get $X_a^{M,S}(t) \geq 3\Delta x + X_b^{m,S}(t)$ for $t \geq 0$. This means that there are at least two successive cells with value $V_j^n = V_{j+1}^n = 0$ separating $X_b^{m,S}(t_n)$ and $X_a^{M,S}(t_n)$. Then as in Case 1 we obtain for $k \geq j+1$ that (1.2.18) and (1.2.19) are also valid, and thus

$$\text{for } k \geq j+1, \quad V_k^{n+1} = V_k^{n+1}(\nu^M) = \overline{\vartheta}_k^{S,n+1}$$

(i.e, an exact evolution following the discontinuity position $X_a^{M,S}(t_{n+1})$). Also, in the same way as in Case 1, we obtain

$$\text{for } k \leq j, \quad V_k^{n+1} = V_k^{n+1}(\nu^m) = \overline{\vartheta}_k^{S,n+1}$$

(i.e, an exact evolution following the discontinuity position $X_b^{m,S}(t_{n+1})$). This concludes to $V_k^{n+1} = \overline{\vartheta}_k^{S,n+1}$ for all $k \in \mathbb{Z}$. □

Proof of Theorem 1.9: Since $V_j^n = \overline{\vartheta}_j^{S,n}$, for all $j \in \mathbb{Z}$ and $n \geq 0$, we obtain:

$$\|\vartheta^S(t_n,.) - V(t_n,.)\|_{L^1(\mathbb{R})} \leq \Delta x\, TV(\vartheta^S(t_n,.)) = 2\Delta x = TV(v_0)\ \Delta x. \tag{1.2.20}$$

Inequalities (1.2.20) and (1.2.9) lead to the desired result (1.2.16). □

Remark 1.12. *Lemma 1.11 shows the behaviour of the HJB-UltraBee scheme: when two discontinuities are far from each other, the algorithm is able to recover, from the average values, their exact positions $X_a^{M,S}(t)$ and $X_b^{m,S}(t)$. Then the scheme makes these discontinuities evolve with the piece-wise constant velocities f_M^S and f_m^S. This is due to the fact that, as long as the discontinuities are separated by more than $3\Delta x$ from each other, the extrema of ϑ^S can be identified by the scheme.*

This interpretation of the scheme extends some results of [12] (for the advection case) to HJB equations. We will see in the next section that this exact computation of the averages of ϑ^S is no more possible when two discontinuities lie in two successive cells, or in the same cell.

1.3 Case of piece-wise constant initial data

Let $(y_i)_{i=0,\cdots,p+1}$ be an increasing sequence of real values with $y_0 = -\infty$, $y_{p+1} = +\infty$, and let $(\gamma_i)_{i=0,\cdots,p}$ be real values. Assume that the function v_0 is of the following form:

$$v_0(x) = \begin{cases} \sum_{i=0,\ldots,p} \gamma_i \mathbf{1}_{]y_i,y_{i+1}[}(x) & \text{for } x \in \mathbb{R}\backslash\{y_1,\ldots,y_p\}, \\ \min(\gamma_{i-1},\gamma_i) & \text{for } x = y_i \text{ and for } i = 1,\ldots,p. \end{cases} \tag{1.3.1}$$

With this definition, v_0 is a l.s.c. piece-wise constant function. We also assume that the mesh size Δx satisfies:

$$\Delta x \leq \frac{1}{3} \min_{1\leq i\leq p-1} (y_{i+1} - y_i). \tag{1.3.2}$$

We have derived in the previous section an error estimate when the discontinuities keep far enough from each other (more precisely, when they are separated by at least two entire cell intervals). In general, two discontinuities may become very close. Two critical cases may happen, see Figure 1.1.

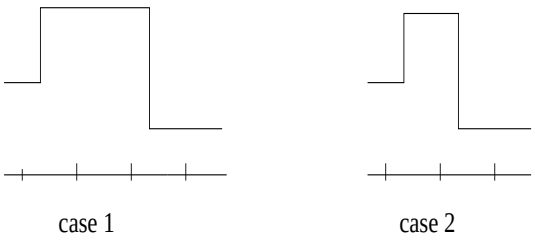


Figure 1.1: Critical cases of truncation

In these cases, one time step of the UltraBee scheme given in algorithm 1 may not compute the average values exactly. In fact, it would do a false interpretation of the maximum value and of the discontinuity localization. The idea is to anticipate this critical situation. Hence when two discontinuities are too close, a truncation is done such that just one discontinuity remains in its right location (see Fig. 1.2). Here we modify the HJB-UltraBee scheme around maxima when one of the two critical cases of Fig. 1.1 occur.

Algorithm 2

Initialization: We compute the averages $(V_j^0)_{j\in\mathbb{Z}}$ by (1.2.10)

Loop: For $n \geq 0$:
A) Compute $W := \mathcal{S}_{UB}(V^n)$ (HJB-UltraBee step).

B) (Truncation step)

• For all indexes j such that

$$\left\{ \begin{array}{c} W_j > \max(W_{j-1}, W_{j+1}), \text{ and } W_j = V_j^n \\ \text{or} \\ W_j > W_{j-1},\ W_{j+1} > W_{j+2}, \text{ and } W_j < V_j^n \end{array} \right\} \ \underline{\text{and}} \quad V_j^n - W_{j-2} < V_j^n - W_{j+2},$$

set

$$\begin{array}{l} V_{j-1}^{n+1} := W_{j-2}, \quad V_j^{n+1} := W_{j-2}, \\ \text{and} \quad V_{j+1}^{n+1} := W_{j+2} + \dfrac{W_{j-2} - W_{j+2}}{V_j^n - W_{j+2}}(W_{j+1} - W_{j+2}), \end{array}$$

• For all indexes j such that

$$\left\{ \begin{array}{c} W_j > \max(W_{j-1}, W_{j+1}), \text{ and } W_j = V_j^n \\ \text{or} \\ W_{j-1} > W_{j-2},\ W_j > W_{j+1}, \text{ and } W_j < V_j^n \end{array} \right\} \ \underline{\text{and}} \quad V_j^n - W_{j-2} \geq V_j^n - W_{j+2},$$

set

$$\begin{array}{l} V_{j+1}^{n+1} := W_{j+2}, \quad V_j^{n+1} := W_{j+2}, \\ \text{and} \quad V_{j-1}^{n+1} := W_{j-2} + \dfrac{W_{j+2} - W_{j-2}}{V_j^n - W_{j-2}}(W_{j-1} - W_{j-2}). \end{array}$$

• Otherwise set $V_j^{n+1} := W_j$.

Hereafter the truncation step will be denoted by

$$V^{n+1} := T_{V^n}(W).$$

Remark 1.13. *The truncation step modifies values only near local strict maxima of* $(V_j^n)_{j\in\mathbb{Z}}$.

The test $V_j^n - W_{j-2} < V_j^n - W_{j+2}$ *(resp.* $V_j^n - W_{j-2} \geq V_j^n - W_{j+2}$*) allows to check if the left jump is strictly smaller (resp. equal or greater) than the right jump near a local maxima, see Figure 1.2. In this case, the truncation corresponds to a recomputation of the average values such that the right (resp. left) discontinuity position is correctly coded. Hence the truncation allows to get rid of the left discontinuity and to keep the right one in its correct location. This truncation step aims to improve the treatment by the UltraBee scheme and to prevent the presence of two discontinuities in the same cell or in adjacent*[1] *cells.*

The main result of the section is the following.

[1] We mean by adjacent here the neighboring cells from the left and from the right

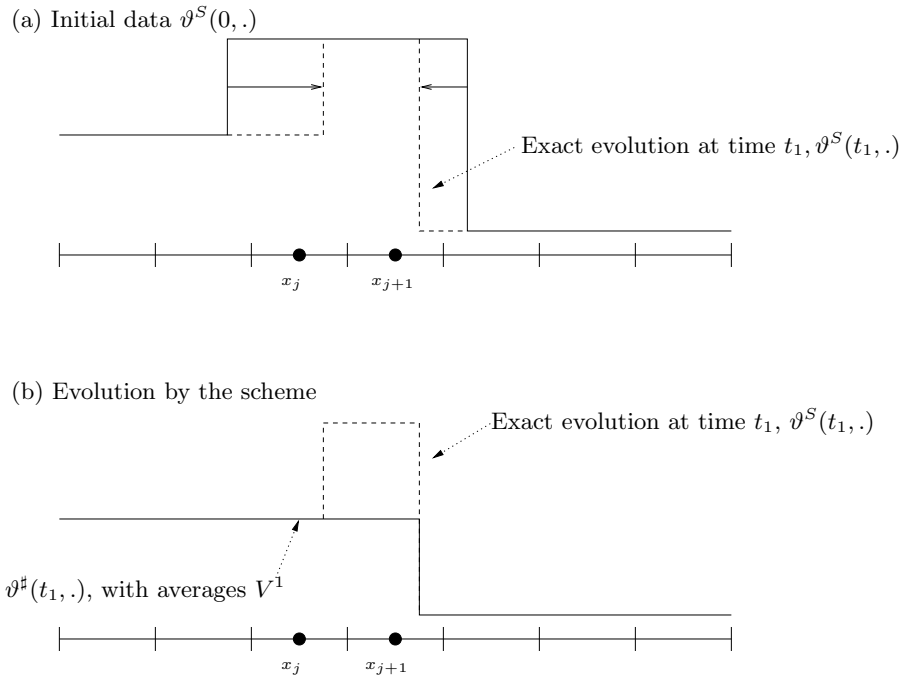


Figure 1.2: Truncation step for critical case 2

Theorem 1.14. *Assume that $(H1)-(H3)$ are satisfied. Let v_0 be a piece-wise constant function as in (1.3.1). We assume that the mesh size satisfies the CFL condition (1.2.11) and (1.3.2). Let ϑ be the viscosity solution of (1.2.2), and let V be defined by (1.2.14) and algorithm 2. We have*

$$\|V(t_n,.) - \vartheta(t_n,.)\|_{L^1(\mathbb{R})} \leq (Lt_n + 4)e^{Lt_n}\, TV(v_0)\, \Delta x, \quad \forall n \geq 0.$$

Notice that the total variation of v_0 given by (1.3.1) is $TV(v_0) := \sum_{i=1,\dots,p-1} |\gamma_{i+1} - \gamma_i|$.

Remark 1.15. *We shall also see in the proof, for instance in the case of the eikonal equation, that $\overline{\vartheta}_j^{S,n} = V_j^n$ as long as the discontinuities are far enough one from each other.*

1.3.1 A first simple case when two fronts may meet

Let a, b be in $\mathbb{R}$, with $b \geq a + 3\Delta x$. Consider the following initial data:

$$v_0(x) = \mathbf{1}_{]a,b[}(x), \quad a.e.\ x \in \mathbb{R}. \tag{1.3.3}$$

Remark 1.16. *Thanks to (1.2.6), under assumption (H1), the unique l.s.c. viscosity solution of (1.2.2) is given by (for $t \geq 0$, $x \in \mathbb{R}$):*

$$\vartheta(t,x) := \begin{cases} \mathbf{1}_{]X_a^M(t), X_b^m(t)[}(x), & \text{if } X_a^M(t) < X_b^m(t), \\ 0 & \text{otherwise.} \end{cases} \tag{1.3.4}$$

Also, by definition of ϑ^S (see (1.2.8)), for $t \geq 0$ and $x \in \mathbb{R}$ we get

$$\vartheta^S(t,x) := \begin{cases} \mathbf{1}_{]X_a^{M,S}(t), X_b^{m,S}(t)[}(x), & \text{if } X_a^{M,S}(t) < X_b^{m,S}(t), \\ 0 & \text{otherwise} \end{cases} \tag{1.3.5}$$

(the approximated characteristics $X_a^{M,S}$ and $X_b^{m,S}$ are defined as in (1.2.5)).

As in (1.2.17) we denote by $\overline{\vartheta}^{S,n}$ the average cell values of $\vartheta^S(t_n, \cdot)$.

Lemma 1.17. *Assume that $(H1)-(H3)$ are satisfied. Let v_0 be as in (1.3.3), and the mesh size satisfy (1.2.11) and $\Delta x \leq \dfrac{b-a}{3}$. For every $n \geq 0$, we have*

$$\|V(t_n,.) - \vartheta^S(t_n,.)\|_{L^1(\mathbb{R})} \leq 4\Delta x\, e^{Lt_n}. \tag{1.3.6}$$

Proof. For $n \in \mathbb{N}$, let j_n and ℓ_n be two integers such that $X_a^{M,S}(t_n) \in]x_{j_n-\frac{1}{2}}, x_{j_n+\frac{1}{2}}]$ and $X_b^{m,S}(t_n) \in]x_{\ell_n-\frac{1}{2}}, x_{\ell_n+\frac{1}{2}}]$.
Two cases may occur:

(i) $\ell_n \geq j_n + 3\ \forall n \geq 0$.
(ii) There exists a first index $n \geq 1$ such that $\ell_n < j_n + 3$.

Assume (i). By the results of section 1.2.3 we know that the scheme computes the exact averages of ϑ^S as long as the fronts are separated by at least two cells, that is, in the case for all $k \leq n$, $\ell_k \geq j_k + 3$. In particular, no truncation step has occured in this case. We then have $V_j^n = \overline{\vartheta}_j^{S,n}$ and can conclude as in (1.2.20) to

$$\|\vartheta^S(t_n,.) - V(t_n,.)\|_{L^1(\mathbb{R})} \leq \Delta x\, TV(\vartheta^S(t_n,.)) = 2\Delta x = TV(v_0)\ \Delta x. \tag{1.3.7}$$

Assume now (ii). For $k < n$, the estimate (1.3.7) holds (no truncation done yet). For $k \geq n$, a truncation has occured at step n and thus $V^k = 0$. For $k \geq n$, we then have (using Lemma 1.4(ii)) $||V(t_k,.) - \vartheta^S(t_k,.)||_{L^1} = ||\vartheta^S(t_k,.)||_{L^1} = \max\left(0, X_b^{m,S}(t_k) - X_a^{M,S}(t_k)\right) \leq e^{L(t_k-t_n)}(X_b^{m,S}(t_n) - X_a^{M,S}(t_n) + \Delta x) \leq e^{Lt_k} 4\Delta x$. This gives the desired bound. □

Proof of theorem 1.14 in the case v_0 is given by (1.3.3): It is now a simple consequence of (1.3.6) and of Proposition 1.6.

1.3.2 Proof of Theorem 1.14 in the general case

We notice that since v_0 is a step function, $\vartheta^S(t,.)$ is also a step function. We need to define a truncation $\vartheta^\sharp$ of ϑ^S that will be connected to the scheme values.

For a given piece-wise constant l.s.c. function w (of the form (1.3.1)), we define the truncation function $Trunc(w)$ as follows. For $x \in \mathbb{R}$, set

$$z_1^x := \sup\{z,\ z \le x,\ w(z) \neq w(x)\} \in\ [-\infty,\ \infty[,$$
$$z_2^x := \inf\{z,\ z \ge x,\ w(z) \neq w(x)\} \in\ ]-\infty,\ \infty]$$

(i.e. the closest left and right discontinuities of w to x). Let j_1 be such that $z_1^x \in]x_{j_1-\frac{1}{2}}, x_{j_1+\frac{1}{2}}]$ and j_2 be such that $z_2^x \in [x_{j_2-\frac{1}{2}}, x_{j_2+\frac{1}{2}}[$. Then set (see Fig. 1.2(b)):

$$Trunc(w)(x) := \begin{cases} \max(w(z_1^x), w(z_2^x)) & \text{if } j_2 \in \{j_1+1,\ j_1+2\} \\ & \text{and } w(x) > \max(w(z_1^x), w(z_2^x)), \\ w(x) & \text{otherwise.} \end{cases}$$

We now define the function $\vartheta^\sharp$ by:

- $\vartheta^\sharp(0,.) := v_0$, and
- $\forall n \ge 0$, $\vartheta^\sharp(t_{n+1},.) = Trunc(w^{n+1})$ where

$$w^{n+1}(x) := \min_{y \in [X_x^{M,S}(-\Delta t), X_x^{m,S}(-\Delta t)]} \vartheta^\sharp(t_n, y). \tag{1.3.8}$$

In the next result we derive an L^1-error bound for $\vartheta^\sharp - \vartheta^S$. We also prove that the cell averages of $\vartheta^\sharp$ are exactly the values given by algorithm 2.

Lemma 1.18. *Assume (H1)-(H3).*
(i) $\forall n \ge 0, \quad \|\vartheta^S(t_n,.) - \vartheta^\sharp(t_n,.)\|_{L^1(\mathbb{R})} \le (4e^{Lt_n} - 1)TV(v_0)\ \Delta x.$
(ii) $\forall j \in \mathbb{Z}$, $\forall n \ge 0$, *we have*

$$\frac{1}{\Delta x}\int_{I_j} \vartheta^\sharp(t_n, x)\ dx = V_j^n.$$

Proof. *(i)* First we notice that for $t \ge 0$, $\vartheta^\sharp(t,.)$ is a piece-wise constant l.s.c function (this can be proved, as for ϑ^S, by using a recursion argument).

At a given time t_n, let $(]\alpha_i, \beta_i[)_{i=1,\dots,p}$ be the local maxima intervals of $\vartheta^S(t_n,.)$ (i.e. $\vartheta^S(t_n,x) \equiv const = \mu_i$ on $]\alpha_i,\beta_i[$ and $\mu_i > \max(\vartheta^S(t_n,\alpha_i), \vartheta^S(t_n,\beta_i))$).

Then we obtain

(a) $\vartheta^\sharp(t_n,.)$ and $\vartheta^S(t_n,.)$ can only differ on the intervals $\cup_{i=1,\dots,p}]\alpha_i,\beta_i[$,
(b) $\forall i$, if $\vartheta^\sharp(t_n,.)$ and $\vartheta^S(t_n,.)$ differ on $]\alpha_i,\beta_i[$, then $\forall x \in]\alpha_i,\beta_i[$, $\vartheta^\sharp(t_n,x) = \max(\vartheta^S(t_n,\alpha_i), \vartheta^S(t_n,\beta_i))$.

Indeed, by Lemma 1.4(iii), the length of the minima regions of ϑ^S can only increase, so they will not deasappear and create new local extrema; Also by (H3) the fronts of an increasing region of ϑ^S cannot get closer, as well as the fronts of a decreasing region of ϑ^S.

Now if $\vartheta^S(t_n,.)$ and $\vartheta^\sharp(t_n,.)$ differ on some interval $]\alpha_i,\beta_i[$, we have for $x\in]\alpha_i,\beta_i[$:

$$\begin{aligned}\|\vartheta^S(t_n,.)-\vartheta^\sharp(t_n,.)\|_{L^1([\alpha_i,\beta_i])} &= \Big(\vartheta^S(t_n,x)-\max(\vartheta^S(t_n,\alpha_i),\vartheta^S(t_n,\beta_i))\Big)\ |\beta_i-\alpha_i|\\ &\le TV(\vartheta^S(t_n,.);]\alpha_i,\beta_i])\ |\beta_i-\alpha_i|.\end{aligned}$$

Also we know that a truncation has occured at some time $t_k<t_n$. At time t_k, the discontinuity positions corresponding to α_i and β_i are located in $\alpha_i^0:=X^{M,S}_{\alpha_i}(-(t_n-t_k))$ and $\beta_i^0:=X^{m,S}_{\beta_i}(-(t_n-t_k))$ respectively. Since the truncation has occured, α_i^0 and β_i^0 are separated by less than two cell intervals, and $|\beta_i^0-\alpha_i^0|\le 3\Delta x$. By Lemma 1.4($ii$), we thus have

$$\begin{aligned}|\beta_i-\alpha_i| &\le e^{L(t_n-t_k)}(|\beta_i^0-\alpha_i^0|+\Delta x)-\Delta x\\ &\le (4e^{Lt_n}-1)\Delta x.\end{aligned}$$

Summing these bounds for all local maxima intervals $[\alpha_i,\beta_i]$ we obtain $\|\vartheta^S(t_n,.)-\vartheta^\sharp(t_n,.)\|_{L^1(\mathbb{R})}\le (4e^{Lt_n}-1)\Delta x TV(\vartheta^S(t_n,.))$. Then we conclude the proof of (i) using Proposition. 1.7.

(ii) Now, we prove recursively that for all $n\ge 0$:

$$(P_n)\quad V_j^n=\frac{1}{\Delta x}\int_{I_j}\vartheta^\sharp(t_n,x)dx,\ \forall j\in\mathbb{Z}$$

and

$$(Q_n)\quad \begin{cases}\text{Any two successive discontinuity positions in } \vartheta^\sharp(t_n,.)\\ \text{are separated by at least two cell intervals}\end{cases}$$

((Q_n) amounts to have $\vartheta^\sharp(t_n,.)=\sum_i \mu_i^n\mathbf{1}_{]y_i,y_{i+1}[}$ a.e. , with $y_i\in]x_{j_1^i-\frac12},x_{j_1^i+\frac12}]$, $y_{i+1}\in[x_{j_2^i-\frac12},x_{j_2^i+\frac12}[$ and $j_2^i\ge j_1^i+3$.)

For $n=0$, $\vartheta^\sharp(0,.)=v_0=\vartheta(0,.)$ and (P_0) follows, also (Q_0) is true by the definition of v_0 and of the mesh step Δx.

Now let us suppose (P_n) and (Q_n). Let $W:=\mathcal{S}_{UB}(V^n)$, and define w^{n+1} as in (1.3.8). Since V^n codes the average values of $\vartheta^\sharp(t_n,.)$ where the discontinuity positions are separated by at least two cells, the Ultra-Bee scheme computes the correct averages for one time-step evolution, and we will have

$$W_j=\frac{1}{\Delta x}\int_{I_j}w^{n+1}(x)dx$$

(this uses the same arguments as in the proof of Lemma 1.11).

As long as the discontinuity positions in w^{n+1} are separated by at least two entire cells, $\vartheta^\sharp(t_{n+1},.) = Trunc(w^{n+1}) = w^{n+1}$, and we have (Q_{n+1}). Also V_j^{n+1} values in step B) of algorithm 2 are unchanged. Hence $V_j^{n+1} = W_j$ which proves (P_{n+1}).

Because of assumption (H3), the only case the discontinuity positions of w^{n+1} may be separated by less than two entire cells is around the local maxima. (Recall that by (H3), discontinuity positions in a monotonous part of $\vartheta^\sharp(t_n,.)$ can only get far from each other. This is also true for discontinuity positions around a local minimum, by Lemma 1.4(iii)).

Now let us assume that $\vartheta^\sharp(t_{n+1},x) := Trunc(w^{n+1})(x) \neq w^{n+1}(x)$ for some $x \in \mathbb{R}$. This means that x is surrounded by two discontinuities of w^{n+1}, denoted z_1^x and z_2^x, which are separated by one cell (critical case 1), or which lie in two successive cells (critical case 2), see Fig. 1.1. Let us consider for instance that we have a critical case 2 (critical case 1 being similar). We may also assume that $V_j^n - W_{j-2} < V_j^n - W_{j+2}$ as illustrated in Fig. 1.2.

In this case we see that the average values in cells I_{j-1} and I_j can be set to W_{j-2} (hence the redefinition $V_{j-1}^{n+1} = V_j^{n+1} := W_{j-2}$ in algorithm 2). Then the remaining discontinuity position $\bar{x} \in I_{j+1}$ can be computed by two different ways: first, by

$$\frac{\bar{x} - x_{j+\frac{1}{2}}}{\Delta x} = \frac{W_{j+1} - W_{j+2}}{V_j^n - W_{j+2}},$$

and second, if V_{j+1}^{n+1} codes the correct average value on I_{j+1} after truncation,

$$\frac{\bar{x} - x_{j+\frac{1}{2}}}{\Delta x} = \frac{V_{j+1}^{n+1} - W_{j+2}}{W_{j-2} - W_{j+2}}$$

(recall that the scheme values should code the average of an exact piece-wise function). Thus we obtain the desired definition of V_j^{n+1} in terms of the (W_k) as in algorithm 2 B), which proves (P_{n+1}).

On the other hand, after the truncation step, there are no more critical cases in $\vartheta^\sharp(t_{n+1},.)$ and thus we have (Q_{n+1}). □

Proof of theorem 1.14 in the general piece-wise constant case: Combining Proposition 1.6 and Lemma 1.18(i) we obtain

$$\begin{aligned} \|\vartheta(t_n,.) - \vartheta^\sharp(t_n,.)\|_{L^1(\mathbb{R})} &\leq \|\vartheta(t_n,.) - \vartheta^S(t_n,.)\|_{L^1(\mathbb{R})} + \|\vartheta^S(t_n,.) - \vartheta^\sharp(t_n,.)\|_{L^1(\mathbb{R})} \\ &\leq (Lt_n + 4)e^{Lt_n} - 1)TV(v_0)\ \Delta x. \end{aligned} \tag{1.3.9}$$

Then, using Lemma 1.18(ii), we obtain

$$\|\vartheta^\sharp(t_n,.) - V(t_n,.)\|_{L^1(\mathbb{R})} \leq TV(\vartheta^\sharp(t_n,.))\Delta x. \tag{1.3.10}$$

By construction of $\vartheta^\sharp$ (and using Lemma 1.40), we also have $TV(\vartheta^\sharp(t,.)) \leq TV(\vartheta^S(t,.)) \leq TV(v_0)$. Together with (1.3.9), (1.3.10) we obtain the desired bound of Theorem 1.14. □

1.4 Case of a general discontinuous initial data

In this section we generalize Theorem 1.14 (where we supposed that v_0 was a piece-wise constant function) to more general discontinuous initial data v_0.

We assume that $v_0 : \mathbb{R} \to \mathbb{R}$ is a l.s.c. function, with $TV(v_0) < \infty$, and that v_0 has a finite number of extrema. More precisely we consider the following assumption :

$$\textbf{(H4)} \quad \begin{cases} \text{There exist } A_1, \dots, A_{q+1} \text{ and } B_1, \dots, B_q \text{ real numbers with} \\ \qquad A_1 = -\infty \le B_1 < A_2 < \dots < B_q \le A_{q+1} = +\infty, \\ \text{(with possibly } B_1 = -\infty \text{ or } B_q = +\infty\text{), such that } v_0 \nearrow \text{ on each } [A_i, B_i[, \\ v_0 \searrow \text{ on each }]B_i, A_{i+1}], \text{ and } v_0(B_i) = \min(v_0(B_i^-), v_0(B_i^+)). \end{cases}$$

In particular A_i are local minima of v_0, and B_i are local maxima of v_0.

We also consider Δx small enough such that the minima and maxima of v_0 are separated at least by $3\Delta x$:

$$\Delta x < \frac{1}{3} \min_{i=1,\dots,q} \min\left(B_i - A_i,\ A_{i+1} - B_i\right). \tag{1.4.1}$$

Let a regular grid with mesh size Δx, and let $\Delta t > 0$.

We denote by v_0^P the l.s.c. function associated to v_0 as follows. v_0^P is the projection of v_0 on the grid with mesh size $3\Delta x$, with modified extrema. We set $U_j :=]x_{3j-\frac{1}{2}}, x_{3j+\frac{5}{2}}[$, for all $j \in \mathbb{Z}$,

- If $\overline{U_j} \cap \{(A_k)_{k=2,\dots,q}, (B_k)_{k=1,\dots,q}\} = \emptyset$, set
$$v_0^P(x) := \frac{1}{3\Delta x} \int_{U_j} v_0(y)dy, \quad \forall x \in U_j. \tag{1.4.2a}$$
- otherwise if $A_k \in \overline{U_j}$ (resp. $B_k \in \overline{U_j}$) set
$$v_0^P(x) := v_0(A_k) (\text{resp. } v_0(B_k)) \quad \forall x \in U_j. \tag{1.4.2b}$$
- Extend v_0^P by lower semi-continuity :
$$v_0^P(x_{3j-\frac{1}{2}}) = \min\left(v_0^P(x^+_{3j-\frac{1}{2}}),\ v_0^P(x^-_{3j-\frac{1}{2}})\right). \tag{1.4.2c}$$

Then the scheme values are initialized as usual but starting from v_0^P, i.e.

$$V_j^0 := \frac{1}{\Delta x} \int_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]} v_0^P(y)dy = v_0^P(x_j) \tag{1.4.3}$$

This initialization ensures that the initial step function v_0^P will satisfy (1.3.1)-(1.3.2), and that the coded discontinuities are separated by at least $3\Delta x$. The aim of step (1.4.2b) is also to keep the correct (local) extremal value of v_0. This is motivated by the fact that the exact minima and maxima values should propagate. It allows to obtain better long-time approximations.

The general convergence result is the following.

Theorem 1.19. *We assume (H1)-(H3). Consider a l.s.c. function v_0 satisfying (H4), and such that $TV(v_0) < \infty$. We also assume that mesh steps Δx and Δt satisfy the CFL condition (1.2.11) and (1.4.1). Let ϑ be the unique viscosity solution of (1.2.2a)-(1.2.2b). We consider (V_j^0) defined by (1.4.2)-(1.4.3) and $(V^n)_{n\geq 1}$ given by algorithm 2. Let V be defined by (1.2.14). Then*

$$\|V(t_n,.) - \vartheta(t_n,.)\|_{L^1(\mathbb{R})} \leq (Lt_n + 7)e^{Lt_n}\, TV(v_0)\, \Delta x, \quad \forall n \geq 0.$$

Remark 1.20. *We can treat a discontinuous initial data of the form $v_0 := \mathbf{1}_{\mathbb{R}\setminus\{0\}}$. In this case we have $v_0^P = \mathbf{1}_{\mathbb{R}\setminus[-\frac{\Delta x}{2},5\frac{\Delta x}{2}]}$.*

We shall need preliminary estimates

Lemma 1.21. *We have*

$$\|v_0 - v_0^P\|_{L^1(\mathbb{R})} \leq 3\Delta x\, TV(v_0).$$

Proof. The result is immediate by the definition of v_0^P. □

Now let ϑ^P be the l.s.c. viscosity solution of (1.2.2a) with initial data v_0^P, i.e.,

$$\vartheta^P(0,x) := v_0^P(x), \quad \forall x \in \mathbb{R}.$$

The following estimate is essential in the analysis (the proof is postponed to the end of the section).

Proposition 1.22. *We assume $(H1)$. Let u_0 and v_0 be two l.s.c. functions, such that $v_0 - u_0 \in L^1(\mathbb{R})$. We suppose furthermore that*
(i) u_0 satisfies assumption $(H4)$ (with $(A_i)_{i=2,\dots,q}$ local minima);
(ii) for all interval $I \subset \mathbb{R}$,

$$\begin{cases} u_0 \nearrow \text{ on } I \Rightarrow v_0 \nearrow \text{ on } I, \\ u_0 \searrow \text{ on } I \Rightarrow v_0 \searrow \text{ on } I; \end{cases} \tag{1.4.4}$$

(iii) for any local minima A_i of u_0 $(i = 2,\dots,q)$, $u_0(A_i) = v_0(A_i)$.
Let u and v be defined by

$$u(t,x) := \min_{y\in[X_x^M(-t),X_x^m(-t)]} u_0(y), \quad \text{and} \quad v(t,x) := \min_{y\in[X_x^M(-t),X_x^m(-t)]} v_0(y).$$

We have

$$\|v(t,.) - u(t,.)\|_{L^1(\mathbb{R})} \leq e^{Lt}\, \|v_0 - u_0\|_{L^1(\mathbb{R})} \quad \forall t \geq 0. \tag{1.4.5}$$

Proof of Theorem 1.19. By construction of v_0^P, and under assumption $(H4)$, the function v_0^P is increasing on intervals where v_0 is increasing, and decreasing on intervals where v_0 is decreasing, and the local minima of v_0^P are the same as the ones of v_0. Hence we can apply Proposition 1.22 to compare ϑ and ϑ^P and obtain together with Lemma 1.21:

$$\|\vartheta(t,\cdot) - \vartheta^P(t,\cdot)\|_{L^1(\mathbb{R})} \leq e^{Lt}(3\Delta x TV(v_0)), \quad \forall t \geq 0. \tag{1.4.6}$$

Now by Theorem 1.14 we also have

$$\|V(t_n,.) - \vartheta^P(t_n,.)\|_{L^1(\mathbb{R})} \leq (Lt_n + 4)e^{Lt_n} TV(v_0^P)\Delta x, \quad \forall n \geq 0.$$

Furthermore, it is easy to see that $TV(v_0^P) \leq TV(v_0)$. Together with (1.4.6) this concludes the proof of Theorem 1.19. □

We now conclude the section with the proof of Proposition 1.22.

Proof of Proposition 1.22: *Step 1.* We first study the case when u_0 is a monotonous increasing function. In this case v_0 is also increasing. $u(t,x) = u_0(X_x^M(-t))$ and $v(t,x) = v_0(X_x^M(-t))$. Hence, using the change of variable $y = X_x^M(t)$, we get

$$\begin{aligned} \|u(t,.) - v(t,.)\|_{L^1(\mathbb{R})} &= \int_{\mathbb{R}} |u(t,y) - v(t,y)|dy \\ &= \int_{\mathbb{R}} |u(t, X_x^M(t)) - v(t, X_x^M(t))| \left|\frac{dy}{dx}(t)\right| dx. \end{aligned}$$

Here, as f_M is Lipschitz then, by the Rademacher theorem, it is almost everywhere differentiable and we get: $\frac{dy}{dx}(t) = exp\big(\int_0^t f_M'(X_x^M(s))ds\big)$. In particular, $|\frac{dy}{dx}(t)| \leq e^{Lt}$ for all $t \geq 0$, and we obtain the bound

$$\|u(t,.) - v(t,.)\|_{L^1(\mathbb{R})} \leq e^{Lt}\|u_0 - v_0\|_{L^1(\mathbb{R})}.$$

The proof is similar when u_0 is a decreasing function.
Step 2. Now we study the case when u_0 has essentially only one local maximum located in B_1. More precisely, we suppose that $u_0 \nearrow$ on $(-\infty, B_1)$ and $u_0 \searrow$ on (B_1, ∞). The representation Lemma 1.43 allows to write u_0 as:

$$u_0 = \min(u_{01}, u_{02}), \quad \text{with } u_{01} \nearrow \text{ and } \quad u_{02} \searrow$$

where u_{01} and u_{02} are defined as in Lemma 1.43. Then the viscosity solution is given by:

$$u(t,x) = \min\left(u_{01}(X_x^M(-t)), u_{02}(X_x^m(-t))\right), \quad t \geq 0,\ x \in \mathbb{R}.$$

By assumption (ii), v_0 also satisfies $v_0 \nearrow$ on $(-\infty, B_1)$ and $v_0 \searrow$ on (B_1, ∞). We can also write $v_0 = \min(v_{01}, v_{02})$ where v_{01} and v_{02} are also defined as in Lemma 1.43, and we have

$$v(t,x) = \min\left(v_{01}(X_x^M(-t)),\ v_{02}(X_x^m(-t))\right) \quad t \geq 0,\ x \in \mathbb{R}. \tag{1.4.7}$$

Next, we define b as a meeting point for the following two curves at a given time t:

$$x \to v_{01}(X_x^M(-t)) \quad \text{and} \quad x \to v_{02}(X_x^m(-t)).$$

Then

$$\|u(t,.) - v(t,.)\|_{L^1(\mathbb{R})} = \int_{-\infty}^{b} |u(t,x) - v(t,x)|dx + \int_{b}^{+\infty} |u(t,x) - v(t,x)|dx.$$

We notice that for $x \leq b$, $u(t,x) = u_{01}(X_x^M(-t))$ and $v(t,x) = v_{01}(X_x^M(-t))$. We thus get

$$\begin{aligned}
\int_{-\infty}^{b} |u(t,x) - v(t,x)|\ dx &= \int_{-\infty}^{b} |u_{01}(X_x^M(-t)) - v_{01}(X_x^M(-t))|\ dx \\
&= \int_{-\infty}^{X_b^M(-t)} |u_{01}(x) - v_{01}(x)|\ \exp\left(\int_0^t f_M'(X_x^M(s))ds\right)\ dx \\
&\leq e^{LT}\|u_{01} - v_{01}\|_{L^1(-\infty, X_b^M(-t))}, \qquad (1.4.8)
\end{aligned}$$

In the same way:

$$\int_{(b,+\infty)} |u(t,x) - v(t,x)|\ dx \leq e^{LT}||u_{02} - v_{02}||_{L^1(X_b^m(-t),+\infty)} \qquad (1.4.9)$$

Since $f_M \geq f_m$, then $X_b^M(-t) \leq X_b^m(-t)$. Hence combining (1.4.8) and (1.4.9) we obtain (1.4.5).

Notice that for Step 1 and Step 2 there is no need of assumption (iii).

Step 3. We turn now to the proof in the general case (u_0 as in (H4)). Using assumptions (1.3.1)-(1.3.2) we can decompose u_0 into monotonous parts: there exist an integer $q \geq 1$ and real numbers $A_1, \ldots, A_{q+1}$, $B_1, \ldots, B_q$ as in (H4) (with possibly $B_1 = -\infty$ or $B_q = +\infty$), and such that $u_0 \nearrow$ on each $[A_i, B_i[$ and $\searrow$ on each $]B_i, A_{i+1}]$.

We first consider the time interval $[0, \tau_1[$ such that the number of local maxima q of u keeps constant.[2] We note that around a local minima A_i, the solutions u and v will stay constant in the following sense: if we set $I_i^t := [X_{A_i}^m(t), X_{A_i}^M(t)]$ for $t \in [0, \tau_1]$, we have

$$u(t,x) = u_0(A_i) \text{ and } v(t,x) = u_0(A_i), \quad \forall x \in I_i^t.$$

Hence, in order to bound $||u(t,.) - v(t,.)||_{L^1(\mathbb{R})}$, we have to estimate the difference on the remaining intervals $J_i^t := [X_{A_i}^M(t), X_{A_{i+1}}^m(t)]$:

$$\|u(t,.) - v(t,.)\|_{L^1(\mathbb{R})} = \sum_{i=1}^{q} \|u(t,.) - v(t,.)\|_{L^1(J_i^t)},$$

[2] For instance we consider $\tau_1 > 0$ to be the first time such that $\min_{i=1,\ldots,q} X_{a_{i+1}}^m(t) - X_{b_i}^M(t)$ vanishes, with $a_i := \inf\{x, u_0(y) = u_0(A_i)\ \forall x \leq y \leq A_i\}$ and $b_i := \sup\{x, u_0(y) = u_0(A_i)\ \forall A_i \leq y \leq x\}$.

(the J_i^t are disjoint as long as $t < \tau_1$). Notice that $u(t,.)$ and $v(t,.)$ admit only one maximum on J_i^t. Then as in Step 2 we can show that:

$$\|u(t,.) - v(t,.)\|_{L^1(J_i^t)} \leq e^{Lt} \|u_0 - v_0\|_{L^1(J_i^0)} = e^{Lt} \|u_0 - v_0\|_{L^1([A_i, A_{i+1}])}.$$

Summing the previous bounds we obtain for $t \in [0, \tau_1]$:

$$\begin{aligned} \|u(t,.) - v(t,.)\|_{L^1(\mathbb{R})} &\leq e^{Lt} \sum_{i=1}^{q} \|u_0 - v_0\|_{L^1([A_i, A_{i+1}])} \\ &\leq e^{Lt} \|u_0 - v_0\|_{L^1(\mathbb{R})}. \end{aligned} \tag{1.4.10}$$

Now we consider the case when the number of maxima q may lower, and proceed recursively on q. We obtain on a time interval $[\tau_1, \tau_2[$, where the number of maxima is constant and equal to $q-1$, the similar bound:

$$\|u(t,.) - v(t,.)\|_{L^1(\mathbb{R})} \leq e^{L(t-\tau_1)} \|u(\tau_1,.) - v(\tau_1,.)\|_{L^1(\mathbb{R})}, \quad t \in [\tau_1, \tau_2[.$$

Then with (1.4.10) we obtain for all $t \leq \tau_2$ the desired bound, and so on. This concludes the proof of Proposition 1.22. □

1.5 Case of changing sign velocities

We explain in this section how to get a general error estimate using mainly assumption $(H1)$. Instead of assumptions $(H2)$ and $(H3)$ we shall use less restrictive hypothesis that will be detailled in Theorem 1.23 below. Let $v_0 : \mathbb{R} \to \mathbb{R}$ be a l.s.c. function such that $TV(v_0) < \infty$. We consider a regular grid (x_j) as in Section 1.2, and define f_M^S and f_m^S as in Section 1.2.1.

The following algorithm introduces two modifications to algorithm 2: *left and right fluxes* (denoted by $V^{n,L}$ and $V^{n,R}$), for computing the UltraBee estimates, and a *prediction step*. We will explain later on the relevance of these modifications.

Algorithm 3

Initialization: We compute the initial averages (V_j^0) as in (1.4.2)-(1.4.3).
Loop: For $n \geq 0$, we compute V^{n+1} from V^n in three steps:

A) Evolution by a modified HJB-UltraBee scheme:

- Define "fluxes" $V_{j+\frac{1}{2}}^{n,L}(\nu)$, $V_{j+\frac{1}{2}}^{n,R}(\nu)$ for $\nu \in \{\nu^m, \nu^M\}$ as follows:

  If $\nu_j \geq 0$, set

$$V_{j+\frac{1}{2}}^{n,L}(\nu) := \begin{cases} \min(\max(V_{j+1}^n, b_j^+(\nu)), B_j^+(\nu)) & \text{if } \nu_j > 0 \\ V_{j+1}^n & \text{if } \nu_j = 0 \text{ and } V_j^n \neq V_{j-1}^n \\ V_j^n & \text{if } \nu_j = 0 \text{ and } V_j^n = V_{j-1}^n, \end{cases}$$

If $\nu_j \leq 0$, set

$$V^{n,R}_{j-1/2}(\nu) := \begin{cases} \min(\max(V^n_{j-1}, b^-_j(\nu)), B^-_j(\nu)) & \text{if } \nu_j < 0 \\ V^n_{j-1} & \text{if } \nu_j = 0 \text{ and } V^n_j \neq V^n_{j+1} \\ V^n_j & \text{if } \nu_j = 0 \text{ and } V^n_j = V^n_{j+1}, \end{cases}$$

(where b^+_j, b^-_j, B^+_j and B^-_j are defined by (1.2.12)-(1.2.13)).

If $\nu_j \geq 0$ and $\nu_{j+1} > 0$, set $V^{n,R}_{j+\frac{1}{2}}(\nu) := V^{n,L}_{j+\frac{1}{2}}(\nu)$.

If $\nu_{j+1} \leq 0$ and $\nu_j < 0$, set $V^{n,L}_{j+\frac{1}{2}}(\nu) := V^{n,R}_{j+\frac{1}{2}}(\nu)$.

If $\nu_j < 0$ and $\nu_{j+1} > 0$, then set

$$V^{n,R}_{j+\frac{1}{2}}(\nu) := \begin{cases} V^n_{j+1} & \text{if } V^n_{j+1} = V^n_{j+2} \\ V^n_j & \text{otherwise} \end{cases} \quad \text{and} \quad V^{n,L}_{j+\frac{1}{2}}(\nu) := \begin{cases} V^n_j & \text{if } V^n_j = V^n_{j-1} \\ V^n_{j+1} & \text{otherwise.} \end{cases} \tag{1.5.1}$$

- For $\nu \in \{\nu^m, \nu^M\}$, let $\quad V^{n+1}_j(\nu) := V^n_j - \nu_j \left(V^{n,L}_{j+\frac{1}{2}}(\nu) - V^{n,R}_{j-\frac{1}{2}}(\nu) \right)$.

- Set $V^{n+1,1} := \min\left(V^{n+1}(\nu^m), V^{n+1}(\nu^M) \right)$.

B) Truncation: Set $V^{n+1,2} := T_{V^n}(V^{n+1,1})$ as in algorithm 2.

C) Prediction: Set $W := V^{n+1,2}$.

- (decreasing critical cases)

  Let $J^- := \left\{ \begin{array}{ll} j \in \mathbb{Z}, & \left[W_{j-1} > W_j > W_{j+1} > W_{j+2} \text{ and } V^n_j < W_j \right] \\ & \text{or } \left[W_{j-2} > W_{j-1} > W_j > W_{j+1} > W_{j+2} \text{ and } V^n_j = W_j \right] \end{array} \right\}$

  and $J^{-,*} := J^- \setminus \{ j \in J^-, \ s.t. \ j+2 \in J^- \}$.

  For $j \in J^{-,*}$, set

$$\begin{aligned} &V^{n+1}_{j-1} := W_{j-2}, \quad V^{n+1}_j := W_{j-2}, \text{ and} \\ &V^{n+1}_{j+1} := W_{j+2} + \frac{W_{j+1} - W_{j+2}}{V^n_j - W_{j+2}} (W_{j-2} - W_{j+2}). \end{aligned} \tag{1.5.2}$$

- (increasing critical cases)

  Let $J^+ := \left\{ \begin{array}{ll} j \in \mathbb{Z}, & \left[W_{j+1} > W_j > W_{j-1} > W_{j-2} \text{ and } V^n_j < W_j \right] \\ & \text{or } \left[W_{j+2} > W_{j+1} > W_j > W_{j-1} > W_{j-2} \text{ and } V^n_j = W_j \right] \end{array} \right\}$

  and $J^{+,*} := J^+ \setminus \{ j \in J^+, \ s.t. \ j-2 \in J^+ \}$.

For $j \in J^{+,*}$, set

$$\begin{aligned} &V_{j+1}^{n+1} := W_{j+2}, \quad V_j^{n+1} := W_{j+2}, \\ &\text{and} \quad V_{j-1}^{n+1} := W_{j-2} + \frac{W_{j-1} - W_{j-2}}{V_j^n - W_{j-2}}(W_{j+2} - W_{j-2}). \end{aligned} \tag{1.5.3}$$

- otherwise set $V_j^{n+1} = W_j$ (i.e. for j such that $\{j-1, j, j+1\} \cap (J^{-,*} \cup J^{+,*}) = \emptyset$).

Then we have the following general error estimate:

Theorem 1.23. *We suppose (H1). Let v_0 be a l.s.c. function satisfying (H4) and such that $TV(v_0) < \infty$ (in particular v_0 has a finite number of extrema).*
Let $\Delta x > 0$ and $\Delta t > 0$ satisfy the CFL condition (1.2.11) and (1.4.1). Let V_j^n be defined by algorithm 3, V as in (1.2.14), and ϑ be the viscosity solution of (1.2.2).
(i) If v_0 is piece-wise constant as in (1.3.1), with p discontinuities, and if

$$\textbf{(H5a)} \quad \exists \varepsilon > 0,\ \forall x \in \mathbb{R},\ f_m(x) + \varepsilon \leq f_M(x) \quad \textit{and} \quad \Delta x < \frac{\varepsilon}{2L},$$

or

$$\textbf{(H5b)} \quad \forall x \in \mathbb{R},\ f_m(x) \leq 0,\ f_M(x) \geq 0,$$

then

$$\|V(t_n, \cdot) - \vartheta(t_n, \cdot)\|_{L^1(\mathbb{R})} \leq (Lt_n + 4p)e^{Lt_n} TV(v_0)\Delta x, \quad \forall n \geq 0. \tag{1.5.4}$$

(ii) If

$$\textbf{(H5c)} \quad f_m = f_M \quad \textit{and is an non-decreasing function,}$$

then

$$\|V(t_n, \cdot) - \vartheta(t_n, \cdot)\|_{L^1(\mathbb{R})} \leq (1 + Lt_n e^{Lt_n}) TV(v_0)\Delta x, \quad \forall n \geq 0. \tag{1.5.5}$$

Remark 1.24. *Assumption (H5b) is satisfied by the eikonal equation $\vartheta_t + c(x)|\vartheta_x| = 0$, where c is a L-lipschitz positive function.*

Remark 1.25. *As in [7], when f_M (resp. f_m) changes sign, it is important to use two fluxes $V_{j+\frac{1}{2}}^{n,L}$ and $V_{j+\frac{1}{2}}^{n,R}$, which may be different on the cell's interface containig the zero of f_M (resp. f_m). The choice made for these fluxes insures the stability, consistency and TVD[3] properties, see [7, Remark 2.1].*

[3] Total variation diminishing

Remark 1.26. *When assumptions* $(H1)-(H3)$ *hold, the discontinuities around a minimum could only get far from each other. However, if we assume only* $(H1)$*, the discontinuities may become closer. A truncation in this feature would produce an error which is not always controlled by the mesh size* Δx*. Hence we assume one of the* $(H5)$ *assumptions in order to avoid this truncation. Indeed, with (H5) two discontinuities around a local minimum cannot become closer than* $2\Delta x$ *(see Lemma 1.32).*

Remark 1.27. *When removing assumption* $(H3)$*, two successive discontinuities in a monotone zone of* v_0 *can become very close. This motivates the* prediction step*: if the discontinuities are too close, we keep only one of them as shown in Figure 1.3. The algorithm then codes correctly the exact localization of the remaining discontinuity. This prediction step is handled only in the two critical cases explicited in figure 1.4, i.e. when the discontinuities are separated by less than two entire cells.*

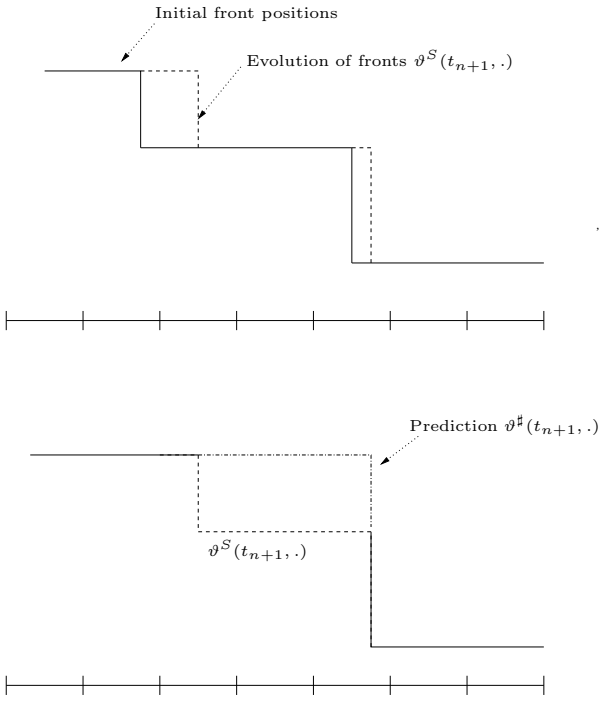


Figure 1.3: Prediction step

1.5.1 Preliminaries

We first prove in the following Lemma that the scheme computes the exact average values of ϑ^S in some elementary cases, where truncation and prediction steps are not used.

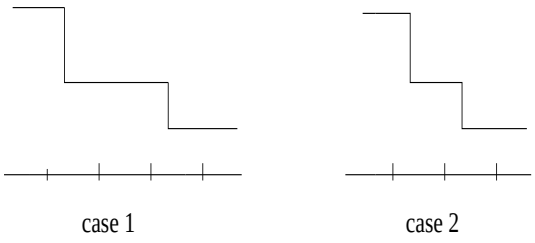


Figure 1.4: Critical cases of prediction

Lemma 1.28. *We assume* $(H1)$ *and (1.2.11). Let* a, b, α, β *be real numbers, with* $\beta \geq 0$. *Let* $v_0(x) := \alpha + \beta \mathbf{1}_{]a,+\infty[}(x)$ *(resp.* $v_0(x) := \alpha + \beta \mathbf{1}_{]-\infty,b[}(x)$*).*
(i) *We have* $\vartheta^S(t,x) := \alpha + \beta \mathbf{1}_{]X_a^{M,S}(t),+\infty[}(x)$ *(resp.* $\vartheta^S(t,x) := \alpha + \beta \mathbf{1}_{]-\infty,X_b^{m,S}(t)[}(x)$*),*
(ii) $\forall n \geq 0$, $\forall j \in \mathbb{Z}$ *:*

$$V_j^n = \frac{1}{\Delta x} \int_{I_j} \vartheta^S(t_n, x) dx \tag{1.5.6}$$

Proof. Part (i) is obtained by direct verifications using (1.2.8). For (ii), we shall treat the case of $v_0(x) = \mathbf{1}_{]a,+\infty[}(x)$ (the other cases beeing similar). We proceed as in Lemma 1.11. We prove the statement by recursion. Let us denote by $x_n := X_a^{M,S}(t_n)$ the discontinuity position at time t_n. Note that it suffices to prove that

$$V_j^{n+1}(\nu^M) = \frac{1}{\Delta x} \int_{I_j} \mathbf{1}_{]X_{x_n}^{M,S}(\Delta t),+\infty[}(x) dx \quad \forall j \in \mathbb{Z}. \tag{1.5.7}$$

Indeed we have in the same way:

$$V_j^{n+1}(\nu^m) = \frac{1}{\Delta x} \int_{I_j} \mathbf{1}_{]X_{x_n}^{m,S}(\Delta t),+\infty[}(x) dx \quad \forall j \in \mathbb{Z}.$$

And this will prove that $V_j^{n+1} = V_j^{n+1}(\nu^M)$ as desired. In the case when for all $j \in \mathbb{Z}$, $\nu_j^M \geq 0$, (or if for all $j \in \mathbb{Z}$, $\nu_j^M \leq 0$), the result comes from Lemma 1.11. It remains to treat the case when (ν_j^M) changes signs. We denote $\nu := \nu^M$ for simplicity.

Assume that $x_n \in]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ (i.e., $V_k^n = 0$ for $k < j$, $V_j^n \in [0,1[$, and $V_k^n = 1$ for $k > j$). We furthermore assume that $x_n \neq x_{j+\frac{1}{2}}$ (the case of $x_n = x_{j+\frac{1}{2}}$ can be treated in a similar way). In particular we have $V_{j-1}^n = V_{j-2}^n$ and $V_{j+1}^n = V_{j+2}^n$. Let us show that the values $(V_k^{n+1}(\nu))_{k=j-1,j,j+1}$ are correctly computed by the modified UltraBee scheme. We first state the following Lemma:

Lemma 1.29. (i) $V_{j+\frac{1}{2}}^{n,L/R} \in [\min(V_j^n, V_{j+1}^n),\ \max(V_j^n, V_{j+1}^n)]$ *(Consistency).*
(ii) *(Stability)* $V_j^n = V_{j-1}^n$ *and* $\nu_j \geq 0 \Rightarrow V_j^{n+1} = V_j^n$.
(iii) *(Stability)* $V_j^n = V_{j+1}^n$ *and* $\nu_j \leq 0 \Rightarrow V_j^{n+1} = V_j^n$.

Proof of Lemma 1.29 Assertion (i) can be obtained as in [7].

(ii) If $\nu_j = 0$, is immediate. If $\nu_j > 0$, we notice that $V^{n,L}_{j+\frac{1}{2}} = V^n_j$ because $b^+_j = B^+_j = V^n_j$. On the other hand, if $\nu_{j-1} > 0$, then $V^{n,R}_{j+\frac{1}{2}} = V^{n,L}_{j-\frac{1}{2}} = V^n_j$ (using $V^n_{j-1} = V^n_j$ and the consistency property). If otherwise $\nu_{j-1} \leq 0$, then $V^{n,R}_{j+\frac{1}{2}} \in \{V^n_j, V^n_{j-1}\}$ (see definition), hence $V^{n,R}_{j+\frac{1}{2}} = V^n_j$.

(iii) the proof of this assertion is similar to (ii). □

We come back to the proof of Lemma 1.28. In the case $\nu_j = 0$, we have $V^{n+1}_j(\nu) = V^n_j$, which is the correct value (the characteristics do not evolve in I_j). If $\nu_{j+1} > 0$, we already know from Section 2 that $V^{n+1}_{j+1}(\nu)$ is correctly computed (we are in a subcase of $\nu_j \geq 0, \nu_{j+1} \geq 0$). The case when $\nu_{j+1} < 0$ is similar. We can use the same arguments to see that $V^{n+1}_{j-1}(\nu)$ is also correctly computed.

Now we assume that $\nu_j > 0$ (the case when $\nu_j < 0$ can be treated in a similar way), and study the different remaining cases.

Case 1: when $\nu_{j-1} > 0$. We can also suppose that $\nu_{j+1} \leq 0$ (otherwise, $(\nu_k)_{k=j-1,j,j+1} \geq 0$ and this has already been treated).

We first have $V^{n+1}_{j-1} = V^n_{j-1}$ (using Lemma 1.29(i)), the correct expected value.

Also $V^{n+1}_j = V^n_j - \nu_j(V^{n,L}_{j+\frac{1}{2}} - V^{n,R}_{j-\frac{1}{2}})$, where $V^{n,L}_{j+\frac{1}{2}}$ is computed as in Algorithm 1 (Section 2) for positive velocities, and where $V^{n,R}_{j-\frac{1}{2}} = V^{n,L}_{j-\frac{1}{2}} = V^n_{j-1}$. Hence the computation of V^{n+1}_j is as in the case of positive velocities, and gives the correct expected value. Finally, $V^{n+1}_{j+1} = V^n_{j+1}$ using Lemma 1.29 (iii).

Case 2: $\nu_{j-1} < 0$. First, $V^{n+1}_{j-1} = V^n_{j-1} - \nu_{j-1}(V^{n,L}_{j-\frac{1}{2}} - V^{n,R}_{j-\frac{3}{2}})$, where $V^{n,L}_{j-\frac{1}{2}} = V^n_{j-1}$ (by definition in Algorithm 3), $V^{n,R}_{j-\frac{3}{2}} \in [V^n_{j-1}; V^n_{j-2}] = \{V^n_{j-1}\}$ (by consistency), hence $V^{n+1}_{j-1} = V^n_{j-1}$. Then, $V^{n+1}_j = V^n_j - \nu_j(V^{n,L}_{j+\frac{1}{2}} - V^{n,R}_{j-\frac{1}{2}})$ where $V^{n,L}_{j+\frac{1}{2}}$ has the flux definition with $\nu_j > 0$, and $V^{n,R}_{j-\frac{1}{2}}$ takes the value V^n_{j-1} (since $V^n_j \neq V^n_{j+1}$ using the fact that $x_n \neq x_{j-\frac{1}{2}}$). Hence in the interval I_j, the estimate of the fluxes are the same as in the case of positive velocities, and are thus correct (exact evolution of the average values locating the discontinuity position in I_j or I_{j+1}. Finally, we have only to check the value of V^{n+1}_{j+1} in the case $\nu_{j+1} < 0$. In this case, $V^{n,R}_{j+\frac{1}{2}} = V^n_{j+1}$ (by Lemma 1.29(ii)), and $V^{n,L}_{j+\frac{3}{2}} = V^n_{j+1}$ (by Lemma 1.29(i)), hence $V^{n+1}_{j+1} = V^n_{j+1}$.

Case 3: $\nu_{j-1} = 0$. We first obtain $V^{n+1}_{j-1} = V^n_{j-1}$. Then V^{n+1}_j and V^{n+1}_{j+1} are computed as in the case $\nu_{j-1} < 0$ (proof is left to the reader). This concludes the proof of Lemma 1.28 □

Remark 1.30. *Indeed we have proved a more precise result: if at some time t_n we have that $\vartheta^S(t_n, .)$ is a piece-wise constant fonction with all successive discontinuities separated by at least two entire intervals, and if (1.5.6) holds for $j \in \mathbb{Z}$, then we have*

$$V^{n+1}_j = \frac{1}{\Delta x}\int_{I_j} \vartheta^S(t_{n+1}, x)dx \qquad \forall j \in \mathbb{Z}.$$

Remark 1.31. *Several estimates of Section 2 can be extended here. In particular, under assumption (H1) only, the estimates of Lemma 1.4(i),(ii) and (v) hold true (for changing sign velocities).*

Lemma 1.32. *We assume* $(H1)$ *and one of the* $(H5)$ *assumptions. Let* $a, b \in \mathbb{R}$ *be such that* $b - a \geq 2\Delta x$. *Then for all* $t \geq 0$, $X_b^{M,S}(t) - X_a^{m,S}(t) \geq 2\Delta x$.

Proof. If we assume $(H5b)$ or $(H5c)$ it is easy to see that $\frac{d}{dt}(X_b^{M,S}(t) - X_a^{m,S}(t)) \geq 0$, for a.e. $t \geq 0$, hence the result. Now assume $(H5a)$. Suppose there exists $\theta \geq 0$ such that $X_b^{M,S}(\theta) - X_a^{m,S}(\theta) < 2\Delta x$. By continuity there exists $\tau \geq 0$ such that $X_b^{M,S}(\tau) - X_a^{m,S}(\tau) = 2\Delta x$ and

$$X_b^{M,S}(t) - X_a^{m,S}(t) < 2\Delta x, \quad \text{for } t \text{ in a neighborhood } \mathcal{V}(\tau^+) \text{ of } \tau^+.$$

Case 1. We first suppose that $X_a^{m,S}(\tau) \in I_{j-1}$ for some $j \in \mathbb{Z}$ (i.e., $X_a^{m,S}(\tau)$ belongs to the interior of a mesh interval) and thus also $X_b^{M,S}(\tau) \in I_{j+1}$. In particular $\delta : t \to X_b^{M,S}(t) - X_a^{m,S}(t)$ is differentiable at $t = \tau$ and necessarily we have $\dot{\delta}(\tau) \leq 0$. Hence $f_m(x_{j-1}) \geq f_M(x_{j+1})$. Then

$$f_M(x_{j-1}) - 2L\Delta x \leq f_M(x_{j+1}) \leq f_m(x_{j-1}) \leq f_M(x_{j-1}) - \varepsilon,$$

and we get $2L\Delta x \geq \varepsilon$ which contradicts $(H5a)$. Thus this case cannot occur.

Case 2. Now we suppose that $X_a^{m,S}(\tau) = x_{j-\frac{1}{2}}$ for some $j \in \mathbb{Z}$ (and thus $X_b^{M,S}(\tau) = x_{j+1+\frac{1}{2}}$). If the two characteristics $X_a^{m,S}(\theta)$ and $X_b^{M,S}(\theta)$ move such that $X_a^{m,S}(\theta) \in I_j$ and $X_b^{M,S}(\theta) \in I_{j+1}$ for $\theta \in \mathcal{V}(\tau^+)$, then using $(H5a)$ we get

$$f_M(x_{j+1}) \leq f_m(x_j) \leq f_M(x_j) - \varepsilon.$$

Since we also have $f_M(x_{j+1}) \geq f_M(x_j) - L\Delta x$, we obtain

$$f_M(x_j) - L\Delta x \leq f_M(x_j) - \varepsilon \leq f_M(x_j) - 2L\Delta x$$

which leads to $2L \leq L$, a contradiction. Otherwise, if the two characteristics move in the same direction, we obtain a contradiction in the same way as in Case 1. □

1.5.2 Proof of Theorem 1.23

We first define a *prediction operator* as follows. Let w be a real piece-wise constant and l.s.c. function. For $x \in \mathbb{R}$, set

$$z_1^x := \sup\{z,\ z \leq x,\ w(z) \neq w(x)\} \in\ [-\infty,\ \infty[,$$
$$z_2^x := \inf\{z,\ z \geq x,\ w(z) \neq w(x)\} \in\ ]-\infty,\ \infty]$$

i.e. the closest left and right discontinuities of w to x. Let j and k be such that

$$z_1^x \in]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \quad \text{and} \quad z_2^x \in [x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}[.$$

Let us denote $w(x^-) := \lim_{y\to x,\ y<x} w(y)$ and $w(x^+) := \lim_{y\to x,\ y>x} w(y)$. Then set

$$Pred(w)(x) := \begin{cases} w((z_1^x)^-) & \text{if } w((z_1^x)^-) > w(x) > w((z_2^x)^+),\ k \in \{j+1, j+2\}, \\ & \quad \text{and } z_2^{z_2^x} \notin]x_{k+2-\frac{1}{2}}, x_{k+2+\frac{1}{2}}[\\ w((z_2^x)^+) & \text{if } w((z_1^x)^-) < w(x) < w((z_2^x)^+),\ k \in \{j+1, j+2\}, \\ & \quad \text{and } z_1^{z_1^x} \notin]x_{j-2-\frac{1}{2}}, x_{j-2+\frac{1}{2}}[\\ w(x) & \text{otherwise.} \end{cases} \tag{1.5.8}$$

Remark 1.33. *The above definition means that we set $Pred(w)(x) = w((z_1^x)^-)$ in the case x belongs to a decreasing zone of w (for instance), surrounded by two discontinuities of w that are separated at most by one cell, but also such that the next right discontinuity $z_2^{z_2^x}$ is at least separated by two cells from the right discontinuity z_2^x (see Fig. 1.3).*

Then we define $\vartheta^\sharp(t_n, .)$ for all $n \geq 0$ by :

- $\vartheta^\sharp(0, .) := v_0^P$,
- $\forall n \geq 0$, $\vartheta^\sharp(t_{n+1}, .) = Pred(Trunc(w_{n+1}))$ where w_{n+1} is as in (1.3.8).

We also define ϑ^S as in (1.2.8) but starting from v_0^P, i.e.:

$$\vartheta^S(t, x) := \min_{y\in[X_x^{M,S}(-t), X_x^{m,S}(-t)]} v_0^P(y), \quad \forall t > 0,\ x \in \mathbb{R}. \tag{1.5.9}$$

Now, we focus on the proof of Theorem 1.23 (i). We follow the proofs of Theorem 1.14 and 1.19. For a given piece-wise constant initial data v_0 (with discontinuities separated by at least $3\Delta x$), it remains to prove that the following still holds - in a similar way as in Lemma 1.18.

Lemma 1.34. *Under the assumptions of Theorem 1.23(i), we have*
(i)

$$\forall n \geq 0, \quad \|\vartheta^S(t_n, .) - \vartheta^\sharp(t_n, .)\|_{L^1(\mathbb{R})} \leq p(4e^{Lt_n} - 1)TV(v_0)\ \Delta x \tag{1.5.10}$$

(ii)

$$\forall j \in \mathbb{Z},\ \forall n \geq 0, \quad V_j^n = \frac{1}{\Delta x}\int_{I_j} \vartheta^\sharp(t_n, x)\ dx. \tag{1.5.11}$$

Proof of Theorem 1.23. The proof of (i) follows by using (1.5.10) and previous estimates :

$$\begin{aligned} &\|V(t_n, .) - \vartheta(t_n, .)\|_{L^1(\mathbb{R})} \\ &\leq \|V(t_n, .) - \vartheta^\sharp(t_n, .)\|_{L^1(\mathbb{R})} + \|\vartheta^\sharp(t_n, .) - \vartheta^S(t_n, .)\|_{L^1(\mathbb{R})} + \|\vartheta^S(t_n, .) - \vartheta(t_n, .)\|_{L^1(\mathbb{R})} \\ &\leq TV(v_0)\Delta x + p(4e^{Lt_n} - 1)TV(v_0)\Delta x + Lt_n e^{Lt_n} TV(v_0)\Delta x \\ &\leq (Lt_n + 4p)e^{Lt_n}\ TV(v_0)\Delta x. \end{aligned}$$

Now to prove (ii), notice that under assumption (H5c), equation (1.2.2a) is reduced to an advection equation. In particular, as f_M is non-decreasing then the discontinuities never meet: we do not need any truncation or prediction step in the scheme. The proof of Theorem 1.23(ii) is similar to the proof of Theorem 1.9.

□

In order to prove Lemma 1.34, we first establish the following.

Lemma 1.35. *Assume that v_0 satisfies assumption (H4) (with $(A_i)_{i=2,\dots,q}$ local minima).*
(i) for all interval $I \subset \mathbb{R}$,

$$\begin{cases} \vartheta^S \nearrow \text{ on } I \Rightarrow \vartheta^\sharp \nearrow \text{ on } I, \\ \vartheta^S \searrow \text{ on } I \Rightarrow \vartheta^\sharp \searrow \text{ on } I, \end{cases}$$

(ii) If $X^{M,S}_{A_i}(t_n)$ is a local minima of $\vartheta^S(t_n,.)$ $(i = 2,\dots,q)$, then it is also a local minima of $\vartheta^\sharp(t_n,.)$, and we have $\vartheta^\sharp(t_n,y) = \vartheta^S(t_n,y)$ $(= \vartheta^S(t_n, X^{M,S}_{A_i}(t_n)))$ for all $y \in [X^{m,S}_{A_i}(t_n), X^{M,S}_{A_i}(t_n)]$.

Proof. By Lemma 1.32, two discontinuities positions a_i and b_i limiting a given minimum A_i in v_0, initially separated by at least $2\Delta x$, stay at more than $2\Delta x$ in $\vartheta^S(t_n,.)$ and thus stay separated at least by one cell interval. In particular they cannot meet under our assumption.

Also since the operators $Pred$ and $Trunc$ do not modify the monotonicity and keep the minima values, we obtain that $\vartheta^\sharp(t_n,.)$ will keep the monotonicity regions of $\vartheta^S(t_n,.)$ (as well as the minima regions of $\vartheta^S(t_n,.)$). This proves both (i) and (ii). □

Remark 1.36. *Note that a minima zone of $\vartheta^S(t_n,.)$ can disappear. This happens only in the case the neighboring left or right maxima zones disappear. By the previous Lemma $\vartheta^\sharp(t_n,.)$ will have a similar property.*

Proof of Lemma 1.34(i). Let $J_i := [X^{M,S}_{A_i}(t_n), X^{m,S}_{A_{i+1}}(t_n)]$. To show (1.5.10), by using Lemma 1.35(ii), it is sufficient to obtain the following bound

$$\|\vartheta^S(t_n,.) - \vartheta^\sharp(t_n,.)\|_{L^1(J_i)} \le p(4e^{Lt_n} - 1)TV(v_0, [A_i, A_{i+1}])\ \Delta x \tag{1.5.12}$$

(the result will then follow by summation on i). It means that we need to show (1.5.10) in the particular case when $v_0 \nearrow$ on $[A_1, B_1]$ and $v_0 \searrow$ on $[B_1, A_2]$. We can assume $A_1 = -\infty$ and $A_2 = \infty$ to simplify.

Set $\vartheta^{\sharp,0} := \vartheta^S$. We define recursively the function $\vartheta^{\sharp,k}$ for $k \ge 0$ as follows. Let t_{n_k} be the first time where a prediction should be performed in the function $\vartheta^{\sharp,k-1}$ (i.e., two successive discontinuities in a decreasing or in an increasing region of $\vartheta^{\sharp,k-1}(t_{n_k},.)$ are separated by less than two entire cell intervals). Then set (for $k \ge 0$)

$$\begin{aligned} &\vartheta^{\sharp,k}(t,.) := \vartheta^{\sharp,k-1}(t,.) \quad \text{for } t < t_{n_k}, \\ &\vartheta^{\sharp,k}(t_{n_k},.) := Pred(\vartheta^{\sharp,k-1}(t^-_{n_k},.)), \\ &\vartheta^{\sharp,k}(t,x) := \min_{y \in [X^{M,S}_x(t_{n_k}-t), X^{m,S}_x(t_{n_k}-t)]} \vartheta^{\sharp,k}(t_{n_k},y) \quad \text{for } t \ge t_{n_k} \text{ and } x \in \mathbb{R}. \end{aligned}$$

This means that $\vartheta^{\sharp,1}$ is the function where only the first occuring prediction in $\vartheta^{\sharp,0} = \vartheta^S$ is taken into account, $\vartheta^{\sharp,2}$ is the function where the two first prediction in ϑ^S are taken into account, etc.

We obtain for $k = 1, \dots, p$,

$$\|\vartheta^{\sharp,k}(t_n,.) - \vartheta^{\sharp,k-1}(t_n,.)\|_{L^1} \leq (4e^{Lt_n} - 1)TV(v_0)\ \Delta x \tag{1.5.13}$$

Indeed, if $t_n < t_{n_k}$, $||\vartheta^{\sharp,k}(t_n,.) - \vartheta^{\sharp,k-1}(t_n,.)||_{L^1} = 0$, and if $t_n \geq t_{n_k}$, we obtain the bound (1.5.13) by using similar arguments as in the proof of Lemma 1.18(i).

Also we note that $\vartheta^{\sharp,p}(t_n,.) \equiv \vartheta^{\sharp}(t_n,.)$, because there are at most p prediction steps that can be done in $\vartheta^{\sharp}(t_n,.)$. Summing these bounds for $k = 1, \dots, p$ and by a triangular inequality this proves (1.5.12). □

Proof of Lemma 1.34(ii)**.** The proof is obtained by a recursion argument on $n \geq 0$.

Lemma 1.28 shows that an isolated discontinuity of $\vartheta^S(t_n,.)$, as long as it keeps separated from other discontinuity positions by at least two cell intervals, has its evolution correctly coded by algorithm 3 for one time step. Also Lemma 1.32 implies that two discontinuities positions a_i and b_i limiting a given minimum A_i in v_0 (and evolving as $X_{a_i}^{m,S}(t)$ and $X_{b_i}^{M,S}(t)$) stay at more than $2\Delta x$, and thus are separated at least by two cell intervals. Hence the problem of having discontinuity positions in ϑ^S no more separated by two cell intervals can only come from maxima regions or monotonous regions of ϑ^S.

Then as in Lemma 1.18(ii) we can show that the cell averages of $Trunc(w_{n+1})$ are well coded by the scheme values $V_j^{n+1,2}$ after Step A) and Step B).

Finally it remains to prove that the prediction step C) corresponds to the application of prediction operator $Pred$ on the function $Trunc(w_{n+1})$. The proof is very similar to the proof of Lemma 1.18(ii) for the truncation step (i.e. a discontinuity position vanishes and the remaining discontinuity leads to a recomputation of average values). □

1.6 Numerical tests

In this section, we apply algorithm 2 and 3 to some examples. These tests show the numerical relevance of the method especially for the truncation step. The L^1-error is computed by the formula:

$$\text{error} \quad \equiv \quad \sum_j \Delta x \left| V_j^n - \frac{1}{\Delta x} \int_{I_j} \vartheta(t^n, x) dx \right|,$$

where (V_j^n) are the numerical values and ϑ is the exact viscosity solution of (1.2.2).

Example 1: piece-wise constant initial data. We consider the following Eikonal equation:

$$\vartheta_t(t,x) + |\vartheta_x(t,x)| = 0, \quad t \geq 0, \tag{1.6.1}$$

for $x \in (-2,2)$ and with periodic boundary conditions. Notice that we can take $f_M = 1$ and $f_m = -1$. The CFL condition is $\frac{\Delta t}{\Delta x} \leq 1$ (here we have chosen the CFL number to be $\frac{\Delta t}{\Delta x} = 0.9$). The initial condition is defined as follows:

$$\vartheta(0,x) = \begin{cases} 2 & \text{if } x \in]-1.6,-1[, \\ 1.7 & \text{if } x \in [-1,0.1[, \\ 0.2 & \text{if } x \in [0.1,0.6], \\ 1.2 & \text{if } x \in]0.6,1.2[, \\ 0.7 & \text{if } x \in [1.2,1.6[, \\ 0 & \text{otherwise.} \end{cases}$$

In Figure 1.5, we show the exact solution (black line) and the numerical values of algorithm 2 (cross) at different times.

We follow the evolution of two local maxima until they disappear, in particular in the last two time steps: just before they disappear (c) and then just after (d). Notice that the scheme computes exactly the mean value of the solution far from critical situations. In fact the dynamics f_m and f_M are constant here, and the approximated characteristics $X_x^{M,S}$ and $X_x^{m,S}$ coincide with X_x^M and X_x^m respectively. This leads to an exact computation of the average values of $\vartheta \equiv \vartheta^S$ by the scheme (except when discontinuities are too close)

Example 2: piece-wise continuous initial data.
We still consider equation (1.6.1) with the initial condition given by

$$\vartheta(0,x) = \begin{cases} -x^2 + 1.5 & \text{if } x \in [-1,1], \\ 0 & \text{otherwise.} \end{cases}$$

The results are shown in Figure 1.6.

Example 3: two local maxima.
Here we consider a piece-wise continuous initial data with two local maxima :

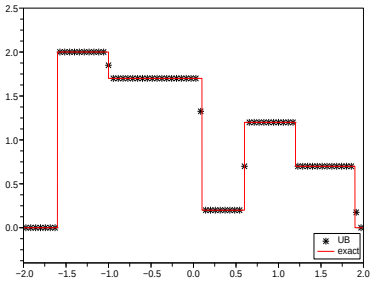
$$\vartheta(0,x) = \begin{cases} x^2 + 0.7 & \text{if } x \in [-1,1], \\ 0 & \text{otherwise.} \end{cases}$$

The results are shown in Figure 1.7 (we still consider the equation (1.6.1)).

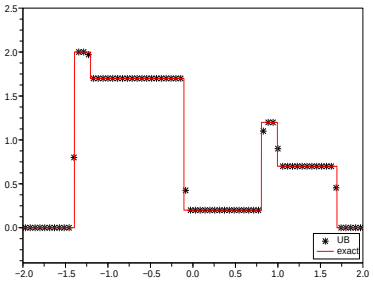
Remark 1.37. *In pratice, numerical tests show that when the initialization of the algorithm is done with a projection with mesh size Δx (instead of $3\Delta x$), the error gets smaller (up to 10 times smaller).*

Example 4: We consider the eikonal equation (1.6.1) on $(-1,1)$ but now with an initial condition composed of two discontinuities, as follows:

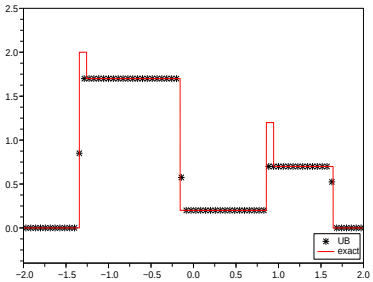
$$\vartheta(0,x) = \begin{cases} 1 & \text{if } x \in]-0.7,0.9[, \\ 0 & \text{otherwise.} \end{cases}$$



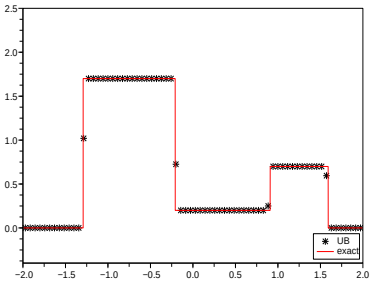
(a) error=0, $t = 0$



(b) error=0, $t = 0.20$

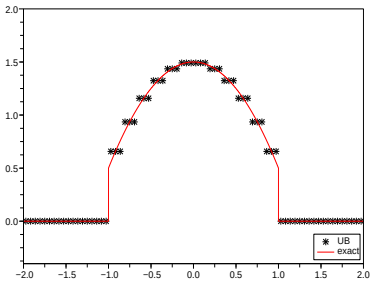


(c) error=0.06, $t = 0.26$

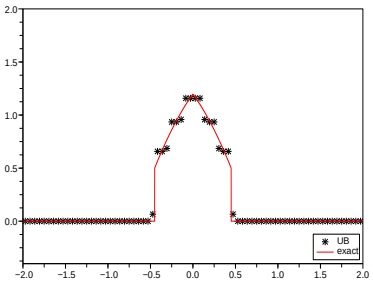


(d) error=0, $t = 0.3$

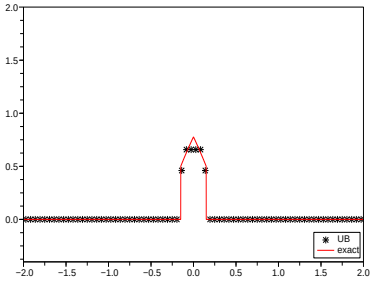
Figure 1.5: (Example 1) Piece-wise constant function, #cells=70.



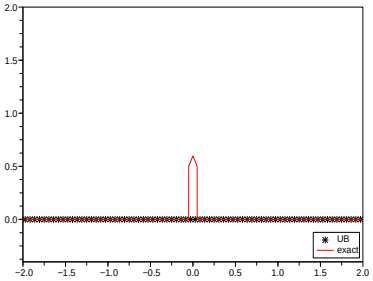
(a) error=0.076, $t = 0$



(b) error=0.047, $t = 0.55$

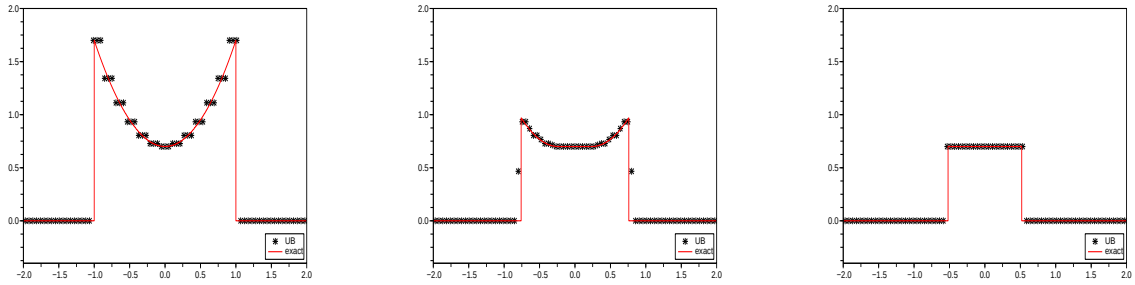


(c) error=0.020, $t = 0.85$



(d) error=0.055, $t = 0.95$

Figure 1.6: (Example 2) Evolution of a piece wise continuous function with one maximum, # cells=72, initialization of the algorithm using v_0^P.



(a) error=0.22, $t = 0$ (b) error=0.08, $t = 0.24$ (c) error=0.05, $t = 0.48$

Figure 1.7: (Example 3) Piece wise continuous function with two maxima, # cells=75.

Δx	error	Truncation Time	# cells
0.1	0.16	0.72	20
0.05	0.07	0.76	40
0.025	0.025	0.79	80
0.0125	0.0025	0.8	160

Table 1.1: (Example 4) Evolution of the error with the mesh size Δx, $CFL = 0.9$

We see in Table 1.1 that the error is bounded by $2\Delta x$.

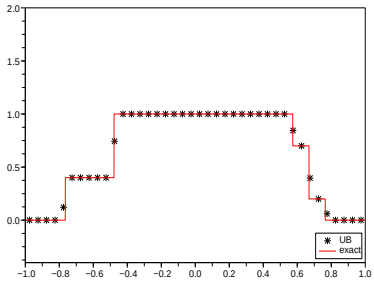
Example 5: Eikonal equation with varying velocity
In this example we deal with algorithm 3 in order to illustrate the prediction step. We consider the following Eikonal equation:

$$\vartheta_t(t,x) + |x| \cdot |\vartheta_x(t,x)| = 0, \quad x \in (-1,1),$$

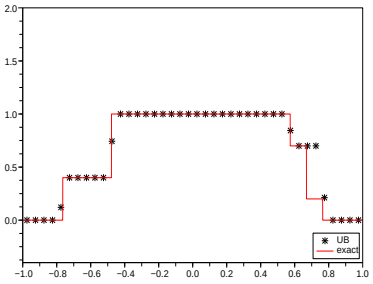
with periodic boundary conditions. Here we take $f_M(x) = |x|$ and $f_m(x) = -|x|$. The initial condition is given by

$$\vartheta(0,x) = \begin{cases} 0.4 & \text{if } x \in]-0.8,-0.5], \\ 1 & \text{if } x \in [-0.5,0.6], \\ 0.7 & \text{if } x \in [0.6,0.7[, \\ 0.2 & \text{if } x \in [0.7,0.8[, \\ 0 & \text{otherwise.} \end{cases}$$

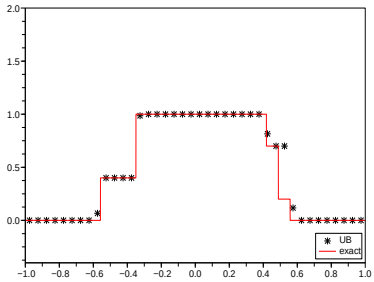
Since f_m and f_M are not constant, ϑ and ϑ^S will differ. Results are shown in Figure 1.8. In Fig.1.8(a), two critical cases of prediction appear simultaneously in the decreasing zone. The algorithm handles only a prediction for the right discontinuity (Fig.1.8(b)). In Fig.1.8(c), a prediction step is again needed, and so on.



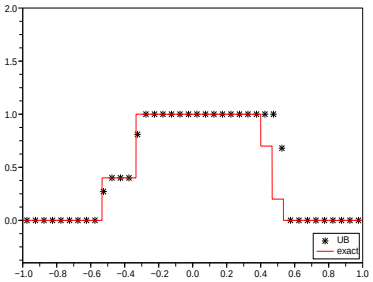
(a) error=0.095, $t = 0.045$
(before prediction step)



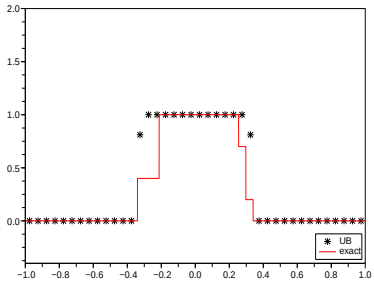
(b) error=0.0487, $t = 0.045$
(after prediction step)



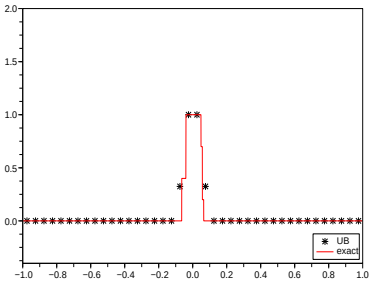
(c) error=0.035, $t = 0.4$
(before prediction step)



(d) error=0.074, $t = 0.4$
(after prediction step)



(e) error=0.12, $t = 0.855$



(f) error=0.027, $t = 2.52$

Figure 1.8: (Example 5) Prediction steps in Algorithm 3, # cells=40

1.7 Appendix

1.7.1 Definition of the approximated characteristics

We prove existence and uniqueness of the absolutely continuous solution $X_a^{M,S}$ of the differential equation:

$$\dot{X}_a^{M,S}(t) = f_M^S(X_a^{M,S}(t)) \text{ a.e } t \geq 0, \qquad X_a^{M,S}(0) = a, \tag{1.7.1a}$$

$$\text{if } \exists t^* \geq 0 \text{ s.t. } \begin{cases} X_a^{M,S}(t^*) = x_{j+\frac{1}{2}}, \\ \text{and } f_M(x_j) f_M(x_{j+1}) \leq 0 \end{cases} \text{ then } X_a^{M,S}(t) = x_{j+\frac{1}{2}} \ \forall t \geq t^* \tag{1.7.1b}$$

We define recursively the characteristic as follows. Assume that $a \in]x_{j_0-\frac{1}{2}}, x_{j_0+\frac{1}{2}}]$ for some index j_0. We consider the case $f_M(x_{j_0}) \geq 0$ (the case $f_M(x_{j_0}) < 0$ being similar). We define $\tau_0 := 0$,

$$\tau_1 := \tau_0 + \frac{x_{j_0+\frac{1}{2}} - a}{f_M(x_{j_0})}, \quad \text{if } f_M(x_{j_0}) > 0,$$

and for $k \geq 1$,

$$\tau_{k+1} := \tau_k + \frac{\Delta x}{f_M(x_{j_0+k})}, \quad \text{if } f_M(x_{j_0+k}) > 0$$

(i.e., $\frac{\Delta x}{f_M(x_{j_0+k})}$ is the time needed for a characteristic to cross the interval I_{j_0+k}). Otherwise, if there exists a first index $k^* \geq 0$ such that $f_M(x_{j_0+k^*}) \leq 0$, then we define $\tau_{k^*+1} := +\infty$ and stop the iterations. Note that since f_M is Lipschitz, we have either $\lim_{k\to\infty} \tau_k = +\infty$ (in this case set $k^* = +\infty$), or there exists $k^* < +\infty$ such that $\tau_{k^*+1} := +\infty$.

Now, $t \in [\tau_k, \tau_{k+1}[$ and $k < k^*$, we set

$$\chi(t) := \chi(\tau_k) + (t - \tau_k) f_M(x_{j_0+k})$$

(where $\chi(\tau_0) = a$ and $\chi(\tau_k) = x_{j_0+k-\frac{1}{2}}$ for $k \geq 1$), and if $t \geq \tau_{k^*}$, we set

$$\chi(t) = \chi(\tau_{k^*}).$$

Then $\chi(t)$ is a solution of (1.7.1).

In order to show the uniqueness of the solution of (1.7.1), we first notice that the first time t^* when two solutions of (1.7.1a) may differ must be such that $\chi(t^*)$ is on an interface: $\exists j \in \mathbb{Z}$, $\chi(t^*) = x_{j+\frac{1}{2}}$. In the case $f_M(x_j) f_M(x_{j+1}) \leq 0$, by definition we have $\chi(t) = x_{j+\frac{1}{2}}$ for all $t \geq t^*$, and uniqueness. Otherwise, in the case $f_M(x_j) f_M(x_{j+1}) > 0$, we have necessarily $f_M(x_j) > 0$ (we assume here that $f_M(x_{j_0}) > 0$). Then the only solution for $t > t^*$ in a neighborhood of t^*, is given by

$$\chi(t) = \chi(t^*) + (t - t^*) f_M(x_{j+1}).$$

This shows uniqueness.

1.7.2 TV bounds

Lemma 1.38. *Let v_0 be an l.s.c. function such that $TV(v_0) < \infty$. For $j = 1, 2$, let $x \to a_x^j$ and $x \to b_x^j$ be two functions from $\mathbb{R}$ onto $\mathbb{R}$. We assume that*
(i) $x \to a_x^j$ and $x \to b_x^j$ are non-decreasing, (for $j = 1, 2$).
(ii) $a_x^j \leq b_x^j$ for $j = 1, 2$ and $x \in \mathbb{R}$.
(iii) there exists $\delta \geq 0$, such that,

$$\forall x \in \mathbb{R}, \quad \max(|(a^2)^{-1}(x) - (a^1)^{-1}(x)|, |(b^2)^{-1}(x) - (b^1)^{-1}(x)|) \leq \delta.$$

(where $(a^j)^{-1}$ and $(b^j)^{-1}$ denote the reciprocal functions of a^j and b^j resp.). Set

$$v_j(x) := \min_{y \in [a_x^j, b_x^j]} v_0(y).$$

Then

$$||v_1 - v_2||_{L^1(\mathbb{R})} \leq 2\delta \; TV(v_0). \tag{1.7.2}$$

Proof. We first assume that $a_x^2 = a_x^1$, for all $x \in \mathbb{R}$. Then we claim that:

$$|v_2(x) - v_1(x)| \leq TV(v_0; [b_x^1; b_x^2]) \tag{1.7.3}$$

where $[\alpha; \beta]$ denotes the interval $[\min(\alpha, \beta), \max(\alpha, \beta)]$. To prove (1.7.3), assume for instance the case $b_x^1 \leq b_x^2$, and consider $z_1 \in [a_x^1, b_x^1]$ such that $v_1(x) = v_0(z_1)$ and $z_2 \in [a_x^1, b_x^2]$ such that $v_2(x) = v_0(z_2)$. If the minimum for v_2 can be reached in $[a_x^1, b_x^1]$ then $v_1(x) = v_2(x)$. Otherwise we have $z_2 \in [b_x^1, b_x^2]$ and $v_2(x) = v_0(z_2) < v_0(z_1) = v_1(x)$. Hence

$$\begin{aligned} |v_2(x) - v_1(x)| &= v_0(z_1) - v_0(z_2) \\ &= v_0(z_1) - v_0(b_x^1) + v_0(b_x^1) - v_0(z_2) \\ &\leq v_0(b_x^1) - v_0(z_2) \leq TV(v_0; [b_x^1, b_x^2]). \end{aligned}$$

The case when $b_x^2 \leq b_x^1$ is similar.

Now we establish the following.

Lemma 1.39. *There exists a positive real measure μ such that $\mu(\mathbb{R}) = TV(v_0)$ and $TV(v_0;]-\infty, \beta[) = \int_{y \in]-\infty, \beta[} d\mu(y)$, for a.e $\beta \in \mathbb{R}$.*

Proof. Let $\tilde{v}_0$ be defined by $\tilde{v}_0(x) := \lim_{y \to x, y < x} v_0(y)$ for every $x \in \mathbb{R}$ (i.e, $\tilde{v}_0(x) = v_0(x^-)$. Then $\tilde{v}_0$ is left-continuous, and there exists a positive measure $\tilde{\mu}$ such that $TV(\tilde{v}_0,]-\infty, x[) = \tilde{\mu}(]-\infty, x[)$ for every $x \in \mathbb{R}$. (This can be deduced for instance from [15]). We have also $\tilde{\mu}(\{x\}) = |v_0(x^-) - v_0(x^+)|$, for every $x \in \mathbb{R}$.

Also, since $TV(v_0) < \infty$, v_0 admits at most a countable set of discontinuity points denoted $(a_n)_{n \geq 0}$. Let

$$q_n := |v_0(a_n^-) - v_0(a_n)| + |v_0(a_n) - v_0(a_n^+)| - |v_0(a_n^-) - v_0(a_n^+)|$$

$(q_n \geq 0)$ and

$$\mu := \tilde{\mu} + \sum_{n\in\mathbb{N}} q_n \delta_{x=a_n}$$

where $\delta_{x=a_n}$ is the dirac measure centered in a_n. Then for $x \notin \{a_n, n \in \mathbb{N}\}$, we have $\mu(]-\infty, x[) = \tilde{\mu}(]-\infty, x[) + \sum_{a_n<x} q_n = TV(v_0,]-\infty, x[)$. Passing to the limit $x \to \infty$ we obtain $\mu(\mathbb{R}) = TV(v_0)$. □

We now come back to the proof of Lemma 1.38 In particular we obtain from (1.7.3)

$$|v_2(x) - v_1(x)| \leq \mu([b_x^1; b_x^2[), \quad \text{a.e. } x \in \mathbb{R}.$$

Then we have, using the Fubini Theorem :

$$\begin{aligned}
\int_{\mathbb{R}} |v_2(x) - v_1(x)| dx &\leq \int_{x\in\mathbb{R}} \left(\int_{y\in\mathbb{R}} 1_{y\in[b_x^1; b_x^2[} d\mu(y) \right) dx \\
&= \int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} 1_{y\in[b_x^1; b_x^2[} dx) \right) d\mu(y) \\
&= \int_{y\in\mathbb{R}} |(b^2)^{-1}(y) - (b^1)^{-1}(y)| d\mu(y) \\
&\leq \int_{y\in\mathbb{R}} \delta \, d\mu(y) \leq \delta \, TV(v_0).
\end{aligned}$$

Now in the general case when $a_x^2 \neq a_x^1$, with the same arguments to (1.7.3)), we have

$$|v_2(x) - v_1(x)| \leq TV(v_0; [a_x^1; a_x^2]) + TV(v_0; [b_x^1; b_x^2]) \tag{1.7.4}$$

Then both parts of the R.H.S. of (1.7.4) can be handled as before and we deduce (1.7.2). □

Lemma 1.40. *Let v_0 be an l.s.c. function with $TV(v_0) < \infty$. We assume that for all $x \in \mathbb{R}$, $a_x \leq b_x$, and $x \to a_x$, $x \to b_x$ are non-decreasing functions, and consider*

$$w(x) := \min_{y\in[a_x, b_x]} v_0(y).$$

We have

$$TV(w) \leq TV(v_0).$$

Proof. Let $x_0 < x_1, \cdots < x_p$ be real numbers. We want to estimate

$$\delta := \sum_{j=1,\ldots,p} |w(x_j) - w(x_{j-1})|$$

For all j, we define y_j as the smallest real number of $[a_{x_j}, b_{x_j}]$ such that $w(x_j) := \min_{[a_{x_j}, b_{x_j}]} v_0(y) = v_0(y_j)$. Let us prove that (y_j) is a non-decreasing sequence. It suffices to check that $x_0 \leq x_1$

$\Rightarrow y_0 \leq y_1$.
a) In the case $b_{x_0} \leq a_{x_1}$, we obtain $y_0 \leq y_1$ trivially.
b) Otherwise, if $\min_{[a_{x_0},b_{x_0}]} v_0 = \min_{[a_{x_0},a_{x_1}]} v_0$, then $y_0 \in [a_{x_0}, a_{x_1}]$ and thus $y_0 \leq y_1$.
c) Otherwise, we have $\min_{[a_{x_0},b_{x_0}]} v_0 = \min_{[a_{x_1},b_{x_0}]} v_0$. If the minimum $w(y_1) = \min_{[a_{x_1},b_{x_1}]} v_0$ is reached on $[b_{x_0}, b_{x_1}]$, then $y_0 \leq b_{x_0} \leq y_1$. If $w(y_1)$ is reached on $[a_{x_1}, b_{x_0}]$, then $y_0 = y_1$. Hence we have proved that $y_0 \leq y_1$ in all cases.

Then, by definition of $TV(v_0)$,

$$\delta = \sum_{j=1,\dots,p} |v_0(y_j) - v_0(y_{j-1})| \leq TV(v_0).$$

Taking the supremum over all non-decreasing sequences (x_j), we obtain the desired result. □

Lemma 1.41. *Let v_0 be a real valued function such that $TV(v_0) < \infty$, then*

$$TV(v_0^P) \leq TV(v_0),$$

where v_0^P is defined in (1.4.2).

Proof. : The number of discontinuity points of v_0 is of zero measure as $TV(v_0) < \infty$. It is clear that $|v_0(x + \Delta x) - v_0(x)| \leq TV(v_0; [x, x + \Delta x[)$ for almost every $x \in \mathbb{R}$. Then we can write

$$\begin{aligned}
TV(v_0^P) &= \sum_{j\in\mathbb{Z}} |V_{j+1}^0 - V_j^0| \\
&\leq \frac{1}{\Delta x} \sum_{j\in\mathbb{Z}} \int_{I_j} |v_0(x + \Delta x) - v_0(x)| \, dx \\
&\leq \frac{1}{\Delta x} \sum_{j\in\mathbb{Z}} \int_{I_j} TV(v_0, [x, x + \Delta x[) \, dx \\
&= \frac{1}{\Delta x} \sum_{j\in\mathbb{Z}} \int_{[-\frac{1}{2},\frac{1}{2}[} TV(v_0, [y\Delta x + x_j, y\Delta x + x_{j+1}[)\Delta x dy \\
&= \int_{[-\frac{1}{2},\frac{1}{2}[} TV(v_0) \, dy = TV(v_0)
\end{aligned}$$

To invert the sum and the integral, we have used that for almost every $x \in \mathbb{R}$, v_0 is continuous at the points $(x + x_j)_{j\in\mathbb{Z}}$. □

1.7.3 Representation Lemma

In this section we denote by ϑ^* the upper semi continuous (u.s.c.) envelope of a real valued function ϑ. We also denote

$$\vartheta(B^-) := \lim_{y\to B,\ y<B} \vartheta(y), \quad \vartheta(B^+) := \lim_{y\to B,\ y>B} \vartheta(y).$$

We start with an elementary result related to the minimum of two viscosity solutions.

Lemma 1.42. *Let v_0, v_{01} and $v_{02} : \mathbb{R} \to \mathbb{R}$ be l.s.c. functions. Let ϑ, ϑ_1 and ϑ_2 be the l.s.c solutions of (1.2.2a) with initial data v_0, v_{01} and v_{02} respectively. If $v_0 = \min(v_{01}, v_{02})$ then for all $t \geq 0$, $x \in \mathbb{R}$, $\vartheta(t,x) = \min(\vartheta_1(t,x), \vartheta_2(t,x))$.*

Proof. Let $w = \min(\vartheta_1, \vartheta_2)$. It is easy to check that the function w is l.s.c, and satisfies the initial condition $w(0,x) = v_0(x)$ in the sense of Definition 1.1 *ii*). Let ϕ be C^1-regular and $(t,x) \in \mathbb{R}^+ \times \mathbb{R}$ a minimum of $w - \phi$ with $w(t,x) = \vartheta_1(t,x)$ (the case $w(t,x) = \vartheta_2(t,x)$ being similar). Then $\vartheta_1 - \phi$ has also a minimum at (t,x). Since ϑ_1 is a viscosity solution, we obtain that ϕ satisfies (1.2.2a). Hence w is a l.s.c. viscosity solution of (1.2.2a). By uniqueness of ϑ, we obtain $w = \vartheta$. □

Notice that the same arguments would not work for the maximum of two viscosity solutions instead of the minimum.

We now give a representation formula for a viscosity solution in a particular case.

Lemma 1.43. *Let $v_0 : \mathbb{R} \to \mathbb{R}$ be a l.s.c function. We assume that there exists B_1 such that $v_0 \nearrow$ for $x < B_1$, and $v_0 \searrow$ for $x > B_1$, and $v_0(B_1) = \min(v_0(B_1^-), v_0(B_1^+))$. We define the functions:*

$$v_{01}(x) := \begin{cases} v_0(x) \text{ if } x < B_1, \\ v_0(B_1^-) \text{ if } x = B_1, \\ v_0^*(B_1) \text{ if } x > B_1, \end{cases} \quad \text{and} \quad v_{02}(x) := \begin{cases} v_0^*(B_1) \text{ if } x < B_1, \\ v_0(B_1^+) \text{ if } x = B_1, \\ v_0(x) \text{ if } x > B_1. \end{cases}$$

Let ϑ_1 (respectively ϑ_2) be the l.s.c. viscosity solution of (1.2.2a) with initial condition v_{01} (respectively v_{02}). Then $v_0 = \min(v_{01}, v_{02})$ and

$$\vartheta(t,x) := \min(\vartheta_1(t,x), \vartheta_2(t,x)), \quad t \geq 0, x \in \mathbb{R}$$

is the l.s.c. viscosity solution of (1.2.2a)-(1.2.2b). In particular, we have

$$\vartheta(t,x) = \min\left(v_{01}(X_x^M(-t)), v_{02}(X_x^m(-t))\right).$$

Proof. : It is easy to check that $v_0 = \min(v_{01}, v_{02})$. Then we apply Lemma 1.42 to obtain $\vartheta = min(\vartheta_1, \vartheta_2)$. Let $I_x^t := [X_x^M(-t), X_x^m(-t)]$. We also have $\vartheta_1(t,x) = \inf_{y \in I_x^t} v_{01}(y) = v_{01}(X_x^M(-t))$ as v_{01} is an increasing function, and similarily $\vartheta_2(t,x) = v_{02}(X_x^m(-t))$ as v_{02} is decreasing. □

Bibliography

[1] R. Agbrall and S. Augoula. High order numerical discretization for hamilton-jacobi equations on triangular meshes. *J. Scientific Computing*, 15(2):197–229, 2000.

[2] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations.* Systems and Control: Foundations and Applications. Birkhäuser, Boston, 1997.

[3] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques & Applications (Berlin)*. Springer-Verlag, Paris, 1994.

[4] G. Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[5] 0. Bokanowski, S. Martin, R. Munos, and H. Zidani. An anti-diffusive scheme for viability problems. *Applied Numerical Mathematics*, 56(9):1135–1254, 2006.

[6] O. Bokanowski, N. Megdich, and H. Zidani. An adaptative antidissipative method for optimal control problems. *Arima*, 5:256–271, 2006.

[7] O. Bokanowski and H. Zidani. Anti-diffusive schemes for linear advection and application to Hamilton-Jacobi-Bellman equations. *J. Sci. Computing*, 30(1):1–33, 2007.

[8] M.G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton Jacobi equations. *Math. Comp.*, 43:1–19, 1984.

[9] B. Després and F. Lagoutière. Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. *J. Sci. Comput.*, 16:479–524, 2001.

[10] M. Falcone. A numerical approach to the infinite horizon problem. *Appl. Math. Optim.*, 15(13):213–214, 1987, and **23,** 1991.

[11] M. Falcone and R. Ferretti. Semi-lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and godunov methods. *J. Computational Physics*, 175:559–575, 2002.

[12] F. Lagoutiere. A non-dissipative entropic scheme for convex scalar equations via discontinuous cell reconstruction. *C. R. Acad. Sci.*, 338(7):549–554, 2004.

[13] P.L. Lions and P.E. Souganidis. Convergence of muscl and filtered schemes for scalar conservation laws and hamilton jacobi equations. *Numerische Mathematik*, (69):441–470, 1995.

[14] S. Osher and C-W. Shu. High essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4):907–922, 1991.

[15] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[16] P. Saint-Pierre. Approximation of viability kernel. *Appl. Math. Optim.*, 29:187–209, 1994.

CHAPTER 2

An adaptative antidissipative method for optimal control problems

[1]

[1] Joint work with O.Bokanowski and H.Zidani, published in ARIMA, volume 5, 2006. An extended version is also available under INRIA report form RR 5770 november 2005.

résumé On étudie une méthode numérique pour les équations HJB provenant des problèmes de contrôle optimal avec contraintes sur l'état. Plus précisément on présente un schéma antidissipatif sur une grille adaptative. La grille adaptative est générée en utilisant la structure des quadtree linéaires. Cette technique facilite le stockage et la maniabilité des mailles.

abstract We deal with a numerical method for HJB equations coming from optimal control problems with state constraints.
More precisely, we present here an antidissipative scheme applied on an adaptative grid. The adaptative grid is generated using linear quadtree structure. This technique of adaptation facilitates stocking data and dealing with large numerical systems.

mots clées problèmes de contrôle optimal, équations HJB, schéma antidissipatif, quadtree linéaire.

key words optimal control problems, HJB equations, antidissipative scheme, linear quadtree.

2.1 Introduction

In this paper, we deal with an optimal control problem $(\mathcal{P}_{s,x})$ with state constraint:

$$\begin{cases} \min \varphi(y_{x,s}(T)), \\ \dot{y}_{x,s}(t) = f(y_{x,s}(t), a(t)) \ \forall t \in [s,T] \\ y_{x,s}(s) = x, \ x \in \mathbb{R}^n, \\ a(t) \in \mathcal{A} \ \ \forall t \in [s,T], \\ y_{x,s}(t) \in \mathcal{K} \ \ \forall t \in [s,T]. \end{cases} \tag{2.1.1}$$

The set of controls $\mathcal{A}$ is a compact of $\mathbb{R}^m$, $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi continuous (l.s.c), and T is a fixed final time. The set $\mathcal{K} \neq \emptyset$ is a compact convex set of $\mathbb{R}^n$. The dynamics $f : \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}^n$ is assumed to be Lipschitz and bounded.

Let $v : [0,T] \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the value function defined by $v(s,x) = \inf(\mathcal{P}_{s,x})$. For every $s \in [0,T]$ and $x \notin \mathcal{K}$, $v(s,x) = +\infty$ and for $x \in \mathcal{K}$, $v(T,x) = \varphi(x)$.
It is known that the value function v satisfies the *Dynamic Programming Principle* (DPP):

$$v(s,x) = \inf_{a(.) \in A(s,\tau;x)} v(\tau, y_{x,s}(\tau)), \qquad \forall \tau \in]s,T], \ \forall x \in \mathcal{K}, \tag{2.1.2}$$

where $A(s,\tau;x) := \{a : [0,+\infty[\to \mathcal{A} \text{ measurable}, \ y_{x,s}(t) \in \mathcal{K}, \ \forall t \in [s,\tau]\}$. In the case when the final cost function φ is continuous and $\mathcal{K} = \mathbb{R}^n$, the value function is the unique continuous "viscosity" solution [1, 2, 6] of the *Hamilton-Jacobi-Bellman* (HJB) equation:

$$\begin{cases} -v_t(t,x) - \min_{a \in \mathcal{A}} \{f(x,a).v_x(t,x)\} = 0, & (t,x) \in [0,T] \times \mathcal{K}, \\ v(T,x) = \varphi(x), & x \in \mathcal{K}. \end{cases} \tag{2.1.3}$$

Here, we are interested in the case when φ is given by

$$\varphi(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ +\infty & \text{otherwise,} \end{cases} \tag{2.1.4}$$

where $\mathcal{C} \neq \emptyset$ is a compact convex set of $\mathbb{R}^n$, $\mathcal{C} \subset \mathcal{K}$ and $\mathcal{K} \neq \mathbb{R}^n$. In section 4, we will see that this case modelizes several control problems (target problem, Rendez-Vous problem, viability kernels,...). Here, the value function v may clearly be discontinuous and takes its values in $\{0, +\infty\}$. It still satisfies equation (2.1.3) in a sense given by Frankowska and her co-authors, see [14, 13] and the references therein for all the details.

Several numerical schemes have been studied for discretizing (2.1.3). The most popular are the Semi-Lagrangian schemes [11, 12, 16] and the finite differences schemes [18, 9]. These schemes provide a good approximation for a continuous value function. However, they all use interpolation techniques at some level and are no more suitable for the approximation of discontinuous value functions. Indeed, the interpolation steps produce more or less numerical diffusion, which causes an increasing loss of precision mainly around the discontinuities. To our knowledge, the only scheme which doesn't use any interpolation technique is the one based on the viability algorithm developed by P. Saint Pierre and his co-authors [17]. But, as already shown in [3] this scheme still diffuses.

The approximation method we study here is a mixture of the antidissipative *UltraBee* (UB) scheme [10, 5] and of an adaptative gridding technique. The UltraBee scheme has been studied by B. Désprès and F. Lagoutière [10] for solving the transport equation with positive constant velocity. It has been extended by O. Bokanowski and H. Zidani [5] for the transport equation with a changing sign velocity and applied for the resolution of Hamilton Jacobi equations (2.1.3) on a regular grid.
In our case, the value function takes only values 0 and 1 (the value 1 coding in fact the $+\infty$ value). In this special situation, the UltraBee scheme has a nice property: it is able to localize accurately the discontinuity of v corresponding to the interface Γ_t separating the region where $v(t,.)$ takes the value 1 from the region where it takes the value 0. This property allows to design a simple method for adaptative gridding. In fact, the real calculations at every time $t^n = n\Delta t$ (Δt being the time step) have only to be done on a small neighborhood of the interface Γ_{t^n}. Hence adaptative gridding is particularly interesting in our case. Moreover, we use linear quadtrees which provide a good way to handle easily adaptative grids and to achieve a significant save of memory.

Adaptative gridding for solving HJB equations has already been studied in the case of a continuous value function [7, 8, 16]. In [16] for example, L.Grune has handled the Semi-Lagrangian scheme to solve (2.1.3) and explained the criteria he used for the refinement and coarsening steps. These criteria are based on a fixed tolerance for the interpolation error. The presence of discontinuities in our case makes these criteria no more suitable.

The paper is organized as follows. In section 2 we give the formulation of the UltraBee scheme and some of its properties. In section 3 we present the adaptative technique that we use and explain the steps of the proposed method. Finally in section 4, we give several

numerical simulations in 2 dimensions coming from control problems and propagating front problems.

2.2 The UltraBee scheme

Notice that when we deal with only one control, the HJB equation (2.1.3) becomes a transport equation. Hence we will first present the UltraBee scheme in this simple case in one space dimension ($n = 1$).

2.2.1 Transport equation

Let $f : \mathbb{R} \to \mathbb{R}$ be Lipschitz bounded and $u_0 : \mathbb{R} \to \mathbb{R}$ be lower semi continuous. We consider the transport problem:

$$\begin{cases} u_t(t,x) + f(x)u_x(t,x) = 0, \ x \in \mathbb{R}, \ t \geq 0, \\ u(0,x) = u_0(x), \ x \in \mathbb{R}. \end{cases} \tag{2.2.1}$$

In all the sequel, we will use the following notations: Δt denotes the time step, Δx is the space step of a regular grid $\mathcal{G}$ of $\mathbb{R}$ and ν_j is the local CFL number at cell M_j defined by:

$$\nu_j := \frac{f(x_j)\Delta t}{\Delta x}. \tag{2.2.2}$$

Consider the following scheme of finite volumes type:

$$\begin{cases} \dfrac{U_j^{n+1} - U_j^n}{\Delta t} + f(x_j)\dfrac{U_{j+\frac{1}{2}}^{n,L} - U_{j-\frac{1}{2}}^{n,R}}{\Delta x} = 0, \ \forall j \in \mathbb{Z}, \ \forall n \in \mathbb{N}, \\ U_j^0 = \dfrac{1}{\Delta x}\displaystyle\int_{M_j} u_0(x)dx, \ \forall j \in \mathbb{Z}. \end{cases} \tag{2.2.3}$$

where x_j is the middle point of cell $M_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, U_j^n is an approximation of the mean value $\dfrac{1}{\Delta x}\displaystyle\int_{M_j} u(t^n, x)dx$ of u on cell M_j at t^n, and $U_{j+\frac{1}{2}}^{n,L}$, $U_{j+\frac{1}{2}}^{n,R}$ are fluxes respectively on the left and on the right of the interface of cells M_j and M_{j+1} at time t^n.
For the UltraBee scheme, these fluxes are defined in the following way.

- In the case when the velocity $f(.) \equiv f$ is a positive constant, the fluxes $U_{j+\frac{1}{2}}^{n,L}$ and $U_{j+\frac{1}{2}}^{n,R}$ coincide and we have $U_{j+\frac{1}{2}}^{n,L} = U_{j+\frac{1}{2}}^{n,R} =: U_{j+\frac{1}{2}}^n$. The scheme becomes:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + f(x_j)\frac{U_{j+\frac{1}{2}}^n - U_{j-\frac{1}{2}}^n}{\Delta x} = 0.$$

  The UltraBee scheme, as defined in [10], is a downwind choice of the fluxes under some stability conditions. This choice replaces the classical Upwind flux which is stable but

dissipative. More precisely, the flux $U^n_{j+\frac{1}{2}}$ is given by solving the minimization problem: $\min_{b^{n,+}_j \le U \le B^{n,+}_j} |U - U^n_{j+1}|$ where $b^{n,+}_j$ and $B^{n,+}_j$ are defined by:

$$b^{n,+}_j = \frac{1}{\nu_j}(U^n_j - \max(U^n_j, U^n_{j-1})) + \max(U^n_j, U^n_{j-1}),$$

$$B^{n,+}_j = \frac{1}{\nu_j}(U^n_j - \min(U^n_j, U^n_{j-1})) + \min(U^n_j, U^n_{j-1}).$$

It follows that

$$U^n_{j+\frac{1}{2}} = \min(\max(U^n_{j+1}, b^{n,+}_j), B^{n,+}_j). \tag{2.2.4}$$

- In the case when f is of changing sign, the *UltraBee generalized scheme* (UB-G) [5] is defined as follows

  (a) if $\nu_j > 0$, $U^{n,L}_{j+\frac{1}{2}} = \min(\max(U^n_{j+1}, b^{n,+}_j), B^{n,+}_j)$ as proposed in (2.2.4).

  (b) if $\nu_j < 0$, we define symetrically $U^{n,R}_{j-\frac{1}{2}} = \min(\max(U^n_{j-1}, b^{n,-}_j), B^{n,-}_j)$ with $b^{n,-}_j = \frac{1}{|\nu_j|}(U^n_j - \max(U^n_j, U^n_{j+1})) + \max(U^n_j, U^n_{j+1})$, and $B^{n,-}_j = \frac{1}{|\nu_j|}(U^n_j - \min(U^n_j, U^n_{j+1})) + \min(U^n_j, U^n_{j+1})$.

  (c) if $\nu_j \le 0$ and $\nu_{j+1} \ge 0$, $U^{n,L}_{j+\frac{1}{2}} = U^n_j$, $U^{n,R}_{j+\frac{1}{2}} = U^n_{j+1}$.

  (d) if $\nu_j \nu_{j+1} > 0$, $U^{n,R}_{j+\frac{1}{2}} = U^{n,L}_{j+\frac{1}{2}}$ (if $\nu_j > 0$) or $U^{n,L}_{j+\frac{1}{2}} = U^{n,R}_{j+\frac{1}{2}}$ (if $\nu_{j+1} < 0$).

When the velocity is constant, and under the CFL condition,

$$|\nu_j| \le 1 \ \forall j \in \mathbb{Z}, \tag{2.2.5}$$

one interesting property of the UltraBee scheme is an exact advection [10, Theorem 3] for a class of step functions defined by: $\exists k^0 \in [0, 1[$ such that $\forall j \in \mathbb{Z}$,

$$U^0_{3j+1} = U^0_{3j}, \quad U^0_{3j+2} = k^0 U^0_{3j+1} + (1 - k^0) U^0_{3j+3}. \tag{2.2.6}$$

Exact advection means that the computed value U^n_j is the exact mean value,

$$U^n_j = \frac{1}{\Delta x} \int_{M_j} u(t^n, x) dx,$$

where u is the exact solution of the advection problem. For the convergence proofs of the UltraBee scheme, we refer to [10, 5].

2.2.2 HJB equation

Here, we are still in dimension 1. First, we consider the simple change of variable,

$$\hat{v}(t,x) = v(T-t,x),\ \forall t \in [0,T],\ \forall x \in \mathbb{R}.$$

Then the function $\hat{v}$ satisfies

$$\begin{cases} \hat{v}_t(t,x) - \min_{a\in\mathcal{A}}\{f(x,a).\hat{v}_x(t,x)\} = 0,\ \forall (t,x) \in [0,T]\times\mathcal{K}, \\ \hat{v}(0,x) = \varphi(x),\ \forall x \in \mathcal{K}. \end{cases} \tag{2.2.7}$$

The application of UB-G to the HJB equation (2.2.7) consists, on a regular grid $\mathcal{G}$ of $\mathcal{K}$, in the following steps (UB-HJB):

- Step 1: We compute the discrete initial condition

$$V_j^0 = \frac{1}{\Delta x}\int_{M_j} \varphi(x)dx,\ \forall j \in J := \{j \in \mathbb{Z},\ M_j \in \mathcal{G}\}.$$

- Step 2: We discretize the set of controls $\mathcal{A}$ into N_a controls, $a_1, a_2, ..., a_{N_a}$.

- Step 3: For $n \geq 1$, knowing the approximation $(V_j^n)_{j\in J}$ of $\hat{v}(t^n,.)$

  - We compute, for each $i = 1...N_a$, $(U_j^{n+1}(a_i))_{j\in J}$ given by the UB-G scheme:

$$\begin{cases} U_j^{n+1}(a_i) = U_j^n(a_i) - \dfrac{f(x_j,a_i)\Delta t}{\Delta x}(U_{j+\frac{1}{2}}^{n,L}(a_i) - U_{j-\frac{1}{2}}^{n,R}(a_i)), \\ U_j^n(a_i) = V_j^n, \quad \forall j \in J. \end{cases}$$

  - We take $V_j^{n+1} := \min_{i=1..N_a} U_j^{n+1}(a_i),\ \forall j \in J$. This defines the numerical approximation of $\hat{v}$ at t^{n+1}.

In dimension 2, we apply the UB-HJB using the classical Trotter splitting: the numerical solution evolves during a time step in the x^1-direction and then during another time step in the x^2-direction. The resolution using this splitting technique is stable under the CFL condition,

$$\max(|\frac{f_1(x_j,a_i)\Delta t}{\Delta x^1}|, |\frac{f_2(x_j,a_i)\Delta t}{\Delta x^2}|) \leq 1,\ \forall j \in J,\ \forall i = 1,..,N_a. \tag{2.2.8}$$

Here, the dynamics f is defined by $f := (f_1, f_2)$, Δx^1 and Δx^2 are the space steps respectively in the x^1-direction and in the x^2-direction, and $x_j \in \mathbb{R}^2$ is the center of the cell M_j.

In [4], under some suitable assumptions, we prove in one space dimension the convergence of the UB-HJB scheme towards the value function for any initial condition φ which is C^1-piecewise regular with compact support.

Notice that, at the first step of UB-HJB scheme, when we compute the average values $(V_j^0)_{j\in J}$, the only components whose values are strictly between 0 and 1 are those corresponding to the cells containing the front Γ_0 (we recall that Γ_0 is the interface separating 0-values of $\hat{v}(t=0,.)$ and its 1-values). In dimension 1, we prove in [4] that, for every $n \geq 0$, the interface Γ_{t^n} is localized on no more than one cell. In dimension 2, we shall verify numerically that Γ_{t^n} is still well localized by the UB-HJB scheme, but we don't have yet any precise theoretical result to claim.

2.3 The adaptative method

We explain in this section the details of the method that we propose. For sake of simplicity, we take $n = 2$ and $\mathcal{K} = [X^1_{\min}, X^1_{\max}] \times [X^2_{\min}, X^2_{\max}]$. Before dealing with details, we start by presenting the linear quadtrees technique that we use for stocking data.

2.3.1 Linear quadtrees

As we deal with adaptative grids, we look for a technique that facilitates stocking and finding data relative to each cell of the grid. This technique is explained by I. Gargantini in [15] and uses the notion of *linear quadtree*. If we represent our final adapted grid by a tree, each cell is a leaf (final node of the tree) and the initial quadrant (all the domain $\mathcal{K}$ before adaptation) is the root of the tree. The method for stocking data using quadtrees is based on coding each leaf of the tree with a quaternary function. This code representation is implicitly the path from the root to the concerned leaf.
Every code is composed of 0, 1, 2, 3. When dividing a cell into four subcells, the NW quadrant is indexed by 0, the NE by 1, the SW by 2 and the SE by 3. The code of each subcell is the concatenation of the code of the mother cell with the index of the subcell (as shown in figure 2.1). Here cells 20, 21, 22 and 23 are sisters and 2 is the mother cell.

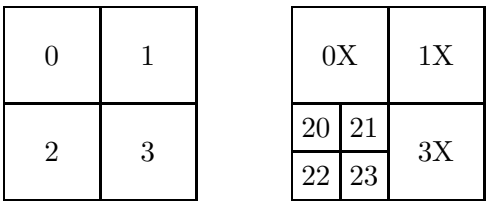


Figure 2.1: Refinement of a cell by quadtrees

Notice that when coding the grid using a tree, all intermediate cells have to be memorized, for example we memorize cells 2, 20, 21, 22, 23. However, in a linear quadtree, we have to stock only final cells of the grid, i.e. cells 20, 21, 22, 23. Then, to find the intermediate cells, we have just to truncate the codes. Furthermore the use of linear quadtrees allows to manage efficiently the adapted grid. In fact, thanks to fast algorithms, operations like encoding a cell into its quaternary code and finding adjacencies of a cell are run in logarithmic time [15].

2.3.2 Algorithm of the method

Our contribution consists in finding a suitable criterion to adapt the computational domain. This criterion must be compatible with the fact that we deal with mean values on each cell and that our value function is discontinuous. Let $L_{\max}$ be a fixed integer that corresponds to the maximal level of refinement. We set

$$\Delta X^1_{\min} = \frac{|X^1_{\max} - X^1_{\min}|}{2^{L_{\max}}} \quad \text{and} \quad \Delta X^2_{\min} = \frac{|X^2_{\max} - X^2_{\min}|}{2^{L_{\max}}}.$$

Then $(\Delta X^1_{\min}, \Delta X^2_{\min})$ is the minimal cell size. The maximal level $L_{\max}$ is chosen such that the following CFL condition holds:

$$\max(|\frac{f_1(x_j, a_i)\Delta t}{\Delta X^1_{\min}}|, |\frac{f_2(x_j, a_i)\Delta t}{\Delta X^2_{\min}}|) \leq 1, \ \forall j \in J, \ \forall i = 1, .., N_a. \tag{2.3.1}$$

In all the sequel, for every $n \geq 0$, we denote by $\mathcal{G}^n$ the adaptative grid at time $t_n = n\Delta t$. We also use the notation $\mathcal{G}_l$, $l = 1, \cdots, L_{\max}$, for the regular grid with mesh steps:

$$\Delta_l X^1 = \frac{|X^1_{\max} - X^1_{\min}|}{2^l}, \qquad \Delta_l X^2 = \frac{|X^2_{\max} - X^2_{\min}|}{2^l}.$$

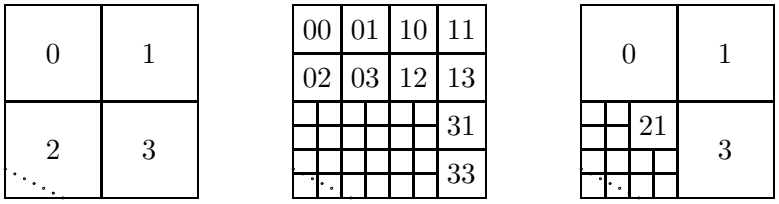
We give now the algorithm of the method which begins by an initialization step. First, we handle refinement steps in order to localize the discontinuity. At every refinement level, a cell which is surrounded by cells not having the same mean value is refined. After these refinement steps, the grid we obtain may contain sister cells having the same mean value (0 or 1) as their immediate neighboring cells. These cells do not contain any discontinuity and have to be coarsened: this is the coarsening step. Finally, we get the grid $\mathcal{G}^0$ where a cell of minimal size either contains a discontinuity or is a neighbor of a cell containing a discontinuity or is a sister of such a cell.

Step 1: The construction of the grid $\mathcal{G}^0$.

- Step 1.1: Take $\mathcal{G}^{0,0} = \mathcal{K}$ (one cell), and define $\mathcal{G}^{0,1}$ as the domain $\mathcal{K}$ splitted into four cells. Set $l = 1$.
- Step 1.2: For $1 \leq l \leq L_{\max} - 1$. For all cells $M_j \in \mathcal{G}^{0,l} \setminus \mathcal{G}^{0,l-1}$, compare the value on M_j with its neighboring values. If the values are different, then refine cell M_j. We obtain a new grid denoted $\mathcal{G}^{0,l+1}$. Set $l = l + 1$ and go to Step 1.2.
  Otherwise, set $l = L_{\max}$ and go to Step 1.3.
- Step 1.3: Set $\widetilde{\mathcal{G}}^{0,L_{\max}} = \mathcal{G}^{0,L_{\max}}$.
- Step 1.4: For $l = L_{\max}, \cdots, 2$, for every cell $M_j \in \widetilde{\mathcal{G}}^{0,l} \cap \mathcal{G}_l$, if the sisters of M_j have the same mean value (0 or 1) and if the neighboring cells of the four sisters have also the same value, then coarsen M_j with its sisters. We obtain a new grid denoted $\widetilde{\mathcal{G}}^{0,l-1}$. Set $l = l - 1$, and go to Step 1.4. Otherwise, set $l = 1$, and go to Step 1.5.

- Step 1.5: Set $\mathcal{G}^0 := \widetilde{\mathcal{G}}^{0,1}$ and define V^0 on $\mathcal{G}^0$.

For example, the construction of the adapted grid $\mathcal{G}^0$ follows the refinement steps explicited in figure 2.2 for $L_{max} = 3$. The value function here takes value 1 below the discontinuity and value 0 beyond. At the second level of refinement, cell 2 is refined as its mean value is in $]0, 1[$. Cells 0, 1, 3 are refined too because their value 0 differs from the value of their neighbor cell 2. Notice that, at level 3, $\mathcal{G}^{0,3}$ is such that all cells containing the discontinuity are of minimal size as well as their neighboring cells. After refinement, we carry out a coarsening step. Following the test of the algorithm, we coarsen subcells of 21, 30, 32 and then subcells of 0, 1 and 3.



$\mathcal{G}^{0,1}$ $\mathcal{G}^{0,3}$ $\mathcal{G}^0$

Figure 2.2: Construction of $\mathcal{G}^0$, the discontinuity is plotted with dots.

Now, for $n \geq 0$, we have the adapted grid $\mathcal{G}^n$ (at time t^n) and the numerical solution V^n on $\mathcal{G}^n$. The discontinuity at t^n lies in the region of $\mathcal{G}^n$ where the cells are of minimal size. Because of the CFL condition (2.3.1), we know that the discontinuity is still in this region at t^{n+1} (as already explained the discontinuity does not evolve of more than one cell of size $(\Delta X^1_{\min}, \Delta X^2_{\min})$ during a time step). We conclude that $V_j^{n+1} = V_j^n$ (=0 or 1) whenever $M_j \in \mathcal{G}^n$ is not of minimal size, and the only computations which remain to be done correspond to the cells of minimal size.

Step 2: UB-HJB computation and construction of $\mathcal{G}^{n+1}$.

- Step 2.0: Do an iteration of UB-HJB scheme on the cells of minimal size $(\Delta X^1_{\min}, \Delta X^2_{\min})$ of $\mathcal{G}^n$. We obtain V^{n+1} on $\mathcal{G}^n$.
- Step 2.1: Define $\mathcal{G}^{n+1,0} := \mathcal{G}^n$, and $V^{n+1,0} := V^{n+1}$ on $\mathcal{G}^n$.
- Step 2.2: For $1 \leq l \leq L_{\max} - 1$, for all cells $M_j \in \mathcal{G}^{n+1,l} \cap \mathcal{G}_l$, compare the value[2] $V_j^{n+1,l}$ on M_j with its neighboring values. If the values are different, then refine cell M_j, attribute to the daugther cells of M_j the same value $V_j^{n+1,l}$. This defines a new grid $\mathcal{G}^{n+1,l+1}$. Set $l = l + 1$, and go to Step 2.2.
  Otherwise, set $l = L_{max}$ and go to step 2.3.
- Step 2.3: Set $\widetilde{\mathcal{G}}^{n+1,L_{\max}} := \mathcal{G}^{n+1,L_{\max}}$ and $\widetilde{V}^{n+1,L_{\max}} := V^{n+1,L_{\max}}$.

[2]Recall that for every $M_j \in \mathcal{G}^{n+1,l} \cap \mathcal{G}_l$, with $l < L_{\max}$, the mean value $V_j^{n+1,l}$ is equal to 0 or 1.

- Step 2.4: For $l = L_{\max}, \cdots, 2$, do a corsening step following the same idea as in Step 1.4. If there is no coarsening to do, then set $l = 1$, and go to Step 2.5.
- Step 2.5: Set $\mathcal{G}^{n+1} := \widetilde{\mathcal{G}}^{n+1,1}$, and $V^{n+1} := \widetilde{V}^{n+1,1}$. This corresponds to the approximation on $\mathcal{G}^{n+1}$ of the solution $\hat{v}$ of (2.2.7) at $t = t^{n+1}$.

By construction, we have the following equivalence result:
Theorem: Let $L_{\max}$ be a fixed integer. Under the CFL condition (2.3.1), the approximation of (2.2.7) using the UB-HJB scheme on an adaptative grid gives the same numerical solution as the resolution using the UB-HJB scheme on a regular grid $\mathcal{G}_{L_{\max}}$.

2.4 Numerical simulations

In the graphics through all this section, we use the black color for cells with mean value strictly between 0 and 1, white for cells with value 0 and light gray for cells with value 1. We also use the notation $\mathcal{B}(c_0, r)$ for the ball centered in c_0 and with radius r.

Example 1: A propagating front problem
We first start with a propagating fronts problem. The initial condition is here two sources from which a fire spreads. Let φ be a function that modelizes the burnt region at $t = 0$, and defined as :

$$\varphi(x) = \begin{cases} 0 & \text{if } x \in \mathcal{B}(c_1, 0.1) \cup \mathcal{B}(c_2, 0.1), \\ 1 & \text{otherwise,} \end{cases}$$

with $c_1 = (0.4, 0.4)$ and $c_2 = (0.6, 0.6)$. Let $\mathcal{K}$ denote the domain $[-0.5, 2.5] \times [-1.5, 1.5]$. We associate to this problem the function $\hat{v}$ which takes value 0 in $(t, x) \in [0, T] \times \mathcal{K}$ if the flame front has already reached x at t, and 1 otherwise. In fact, $\hat{v}$ satisfies the Eikonal equation,

$$\begin{cases} \hat{v}_t(t,x) + ||\nabla \hat{v}(t,x)|| + (-x_2, x_1)^t . \nabla \hat{v}(t,x) = 0, \ \forall t \in [0,T], \ \forall x = (x_1, x_2) \in \mathcal{K}, \\ \hat{v}(0,x) = \varphi(x), \ \forall x \in \mathcal{K}. \end{cases} \tag{2.4.1}$$

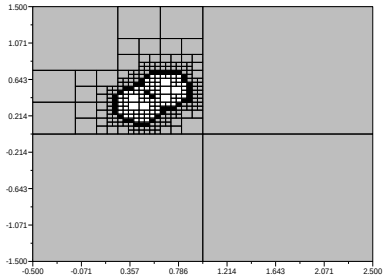
The discontinuity of $\hat{v}$ at time t is the position of the flame front at time t. Hence the set $\{x \in \mathcal{K}, \hat{v}(t,x) = 0\}$ represents the burnt zone at time t.
Although this problem comes from front propagation, it takes place in the formalism we study. Indeed, the Eikonal equation (2.4.1) can be written as an HJB equation:

$$\begin{cases} \hat{v}_t(t,x) - \min_{a \in \mathcal{A}} f(x,a) . \nabla \hat{v}(t,x) = 0, \ \forall t \in [0,T], \ \forall x \in \mathcal{K}, \\ \hat{v}(0,x) = \varphi(x), \ \forall x \in \mathcal{K}, \end{cases}$$

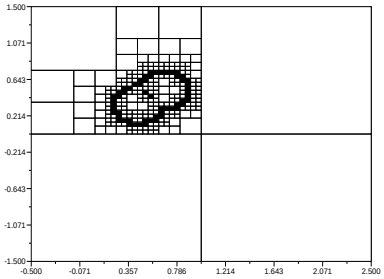
where the set of controls is $\mathcal{A} = [0, 2\pi]$, and the dynamics is given by:

$$f(x,a) = (x_2 - \cos(a), -x_1 - \sin(a))^t, \ \forall x \in \mathbb{R}^2, \ \forall a \in \mathcal{A}.$$

In the numerical tests, we discretize $\mathcal{A}$ into $N_a = 8$ controls, and we choose $L_{max} = 6$ as maximal level of refinement. We visualize the computed solution and the error which is

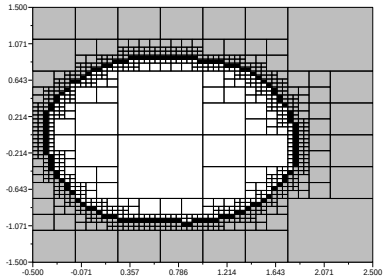


(a) computed solution

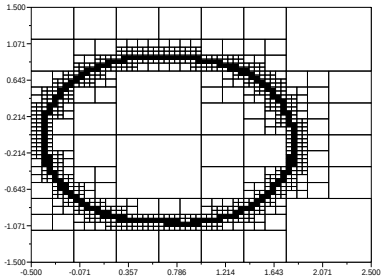


(b) error $(\varepsilon_j^n)_{j\in J}$

Figure 2.3: Computed solution and error at T=0.11, # cells=244, $L_{max} = 6$.



(a) computed solution



(b) error $(\varepsilon_j^n)_{j\in J}$

Figure 2.4: Computed solution and error at T=0.87, # cells=817, $L_{max} = 6$.

defined on each cell M_j, for $j \in J$, by $\varepsilon_j^n = |V_j^n - \widetilde{V}_j^n|$. Here $\widetilde{V}_j^n$ is the average value of the exact solution $\hat{v}$ on cell M_j at time t^n.

We display graphics at $T = 0.11$ (figure 2.3) when the two fronts meet, and then at $T = 0.87$ (figure 2.4) when we get only one front which is already far from the sources of fire. Notice that the error is localized in a thin region around the discontinuity of bandwidth of no more than twice the size of a minimal cell. This is the antidissipative behavior of the scheme. Notice also that the approximation quality isn't distorted when the discontinuity evolves in time. This is another feature of the antidissipative behavior.

Grid	L_{max}	L^1 error	# cells	gain
adapt	6	8.8E-3	244	16.78
reg	6	8.8E-3	4096	-
adapt	7	5.67E-3	547	29.95
reg	7	5.67E-3	16384	-
adapt	8	4E-3	1108	59.14
reg	8	4E-3	65536	-
adapt	9	3E-3	1939	135.195
reg	9	3E-3	262144	-
adapt	10	2.46E-3	3940	266.14
reg	10	-	1048576	-

Table 2.1: Gain relative to each refinement level at T=0.11.

Table 2.1 summerizes the gain we obtain in terms of number of cells when we apply adaptation by comparison to an equivalent regular grid. We can notice that, as expected, we obtain exactly the same error on an adaptative grid and on a regular one. Notice that when we increase the refinement level by 1, the gain is multiplied by 2 and the error is multiplied by $\frac{2}{3}$. This reflects optimization in the management of cells. We can also notice at $T = 0.11$ that when we fix $L_{max} = 10$ we handle almost 4000 cells on an adaptative grid, as much cells as if we did calculations on a regular grid corresponding to $L_{max} = 6$. Hence when we adapt the grid we improve the precision 3 times without spending any additional memory cost. For refinement levels bigger than 10, it is no more possible to handle calculations on a regular grid. Hence we can not have better precision on a regular grid: this reflects the gain of precision achieved by the use of the adaptative algorithm.

Example 2: A capture basin problem (Zermelo problem)
Let $\mathcal{K} := [-6, 2] \times [-2, 2]$ and $\mathcal{C} := \mathcal{B}(c_0, r)$ with $c_0 = (0, 0)$ and $r = 0.44$. We define the dynamics $f : \mathbb{R}^2 \times \mathcal{A} \to \mathbb{R}^2$,

$$f(x, a, \theta) = (1 - \beta x_2^2 + a\cos(\theta), a\sin(\theta)),$$

where the constant $\beta = 0.1$, and $\mathcal{A}$ denotes the set $[0, 0.44] \times [0, 2\pi[$.

Our aim is to approximate the capture basin of $\mathcal{C}$ which is the subset of initial states $x \in \mathcal{K}$ for which exists an admissible control $(a,\theta) \in L^\infty([0,+\infty[;\mathcal{A})$ and a finite time $t \geq 0$ such that the trajectory $y_{x,0}(.)$ evolving with the dynamics f under (a,θ) lives in $\mathcal{K}$ and reaches $\mathcal{C}$ at time t :

$$\mathrm{Capt}_f(\mathcal{C}) := \{x \in \mathcal{K},\ \exists t \geq 0,\ \exists (a,\theta) \in L^\infty(\mathbb{R}^+;\mathcal{A}),\ y_{x,0}(\tau) \in \mathcal{K}\ \forall \tau \in [0,t],\ y_{x,0}(t) \in \mathcal{C}\}.$$

We consider the capture basin of $\mathcal{C}$ before time T:

$$\mathrm{Capt}_f(T,\mathcal{C}) := \{x \in \mathcal{K},\ \exists t \in [0,T],\ \exists (a,\theta) \in L^\infty([0,T];\mathcal{A}),\ y_{x,0}(.) \in \mathcal{K},\ y_{x,0}(t) \in \mathcal{C}\}.$$

It is clear that $T \mapsto \mathrm{Capt}_f(T,\mathcal{C})$ is increasing for inclusion. Moreover, we can prove [3] that $\lim_{T\to+\infty} \mathrm{Capt}_f(T,\mathcal{C}) = \mathrm{Capt}_f(\mathcal{C})$. Let us set

$$\varphi(x) = 0 \text{ if } x \in \mathcal{C}, \quad \text{and 1 otherwise,}$$

and consider the set-valued map defined by

$$\Lambda(x) = \begin{cases} 0 & \text{if } x \in \overset{\circ}{\mathcal{C}}, \\ [0,1] & \text{if } x \in \partial\mathcal{C}, \\ \{1\} & \text{if } x \in \mathcal{K} \setminus \mathcal{C}. \end{cases}$$

Let v_T be the value function of the following control problem:

$$\begin{aligned} &\min\{\varphi(y_{x,s}(T)), \\ &\dot{y}_{x,s}(t) = \lambda(t) f(y_{x,s}(t), a(t), \theta(t)), \quad y_{x,s}(s) = x, \\ &(a(t),\theta(t)) \in \mathcal{A} \quad \& \quad \lambda(t) \in \Lambda(y_{x,s}(t)) \quad \text{for a.e. } t \in (0,T), \\ &y_{x,s}(t) \in \mathcal{K}\ \forall t \in [0,T]. \end{aligned}$$

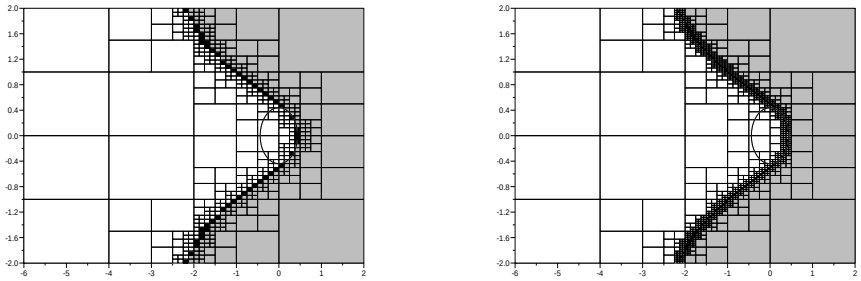
We use the classical change of variable: $\hat{v}(t,x) = v_T(T-t,x),\ \forall t \in [0,T],\ \forall x \in \mathcal{K}$. Then, following [3], we have:

$$\mathrm{Capt}_f(T,\mathcal{C}) = \{x \in \mathcal{K},\ v_T(0,x) = 0\} = \{x \in \mathcal{K},\ \hat{v}(T,x) = 0\}.$$

Then as in [3], in order to approximate $\mathrm{Capt}_f(\mathcal{C})$, we compute an approximation V^n of $\hat{v}(t^n,\cdot)$, with $t^n := n\Delta t$, for n large enough and satisfying the stopping test

$$||V^n - V^{n-1}||_{L^1} := \sum_j \Delta x^1_{min} \Delta x^2_{min} |V^n_j - V^{n-1}_j| \leq 10^{-4}. \tag{2.4.2}$$

In figure 2.5, we show the graphics that we obtain for maximal level of refinement $L_{max} = 6$ (figure 2.5 (a)) and $L_{max} = 7$ (figure 2.5 (b)). In the graphics, the black circle is the border of the target $\mathcal{C}$. We give in the following table the gain obtained for each refinement level, the stopping time (i.e time for which the stopping test (2.4.2) is fulfilled) and the value of



(a) L_{max}=6, $T = 12.96$, # cells=478 (b) L_{max}=7, $T = 7.18$, # cells=952

Figure 2.5: Capture basin of a Zermelo problem.

Grid	L_{max}	# cells	gain	stopping time	$\|\|V^n - V^{n-1}\|\|_{L^1}$
adapt	5	232	4.41	26.125	3.89 E-7
reg	5	1024	-	26.125	3.89 E-7
adapt	6	478	8.56	12.96	1.82 E-8
reg	6	4096	-	12.96	1.82 E-8
adapt	7	952	17.21	7.187	2.38 E-5
reg	7	16384	-	7.187	2.38 E-5

Table 2.2: Gain relative to each refinement level with the value of the residual and the stopping time.

the residual $||V^n - V^{n-1}||_{L^1}$. Notice that when we increase the refinement level L_{max}, the precision required in the stopping test is reached faster. For example, with $L_{max} = 7$ we obtain a good solution on the adaptative grid at $T = 7.187$ using 952 cells, whereas with $L_{max} = 5$, the stopping test is fulfilled only after $T = 26.125$ and the corresponding regular grid contains 1024 cells. Hence adaptative gridding allows not only to have a better precision using almost the same number of cells but also to handle less calculations and to save time.

Example3: A viability kernel problem (consumption problem)
Let $\mathcal{K} = [0, 2] \times [0, 3]$, $\mathcal{A} = [-\frac{1}{2}, \frac{1}{2}]$ and $f(x, a) = (x_1 - x_2, a)$, $\forall x \in \mathbb{R}^2$, $\forall a \in \mathcal{A}$.

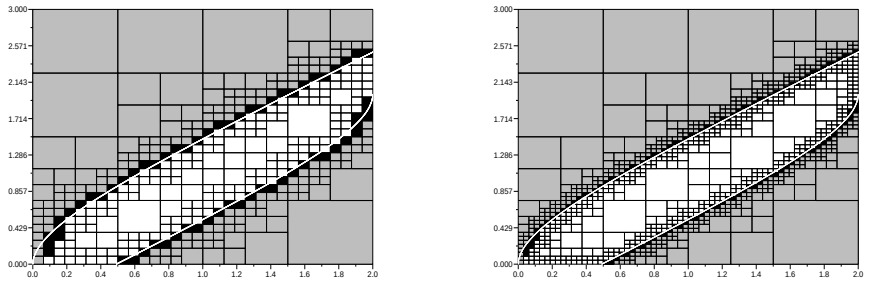
We define the viability kernel associated to $\mathcal{K}$ as the set of initial states $x \in \mathcal{K}$ such that exists a control $a \in L^\infty(\mathbb{R}^+; \mathcal{A})$ and a trajectory $y_{x,0}(.)$ (evolving under the control a) which never leaves $\mathcal{K}$,

$$\text{Viab}(\mathcal{K}) := \{x \in \mathcal{K}, \exists a \in L^\infty(\mathbb{R}^+; \mathcal{A}),\ y_{x,0}(t) \in \mathcal{K}\ \forall t \geq 0\}.$$

We also define $\text{Viab}(T, \mathcal{K}) := \{x \in \mathcal{K}, \exists a \in L^\infty([0, T], \mathcal{A}),\ y_{x,0}(t) \in \mathcal{K}\ \forall t \in [0, T]\}$. Let $\varphi(x) = 0$ if $x \in \mathcal{K}$, and 1 otherwise. Consider the value function v_T of the control problem (2.1.1) associated to the dynamics f, the final cost φ and the set $\mathcal{K}$.

Let $\hat{v}(t, x) = v_T(T - t, x)$. From [3], we have: $\text{Viab}(T, \mathcal{K}) = \{x \in \mathcal{K},\ \hat{v}(T, x) = 0\}$, and $\text{Viab}(T, \mathcal{K}) \longrightarrow_{T \to +\infty} \text{Viab}(\mathcal{K})$. As in the previous example, in order to approximate $\text{Viab}(\mathcal{K})$, we compute an approximation V^n of $\hat{v}(t^n, \cdot)$ for n satisfying the same stopping test (2.4.2).

We show the graphics we obtain for maximal level $L_{max} = 5$ (figure 2.6 (a)) and $L_{max} = 6$ (figure 2.6 (b)). In these figures, the white line is the border of the exact viability kernel. Here, the set of controls $\mathcal{A}$ is discretized into $N_a = 2$ controls ($a \in \{-\frac{1}{2}, \frac{1}{2}\}$).



(a)L_{max}=5, $T \approx 256$, # cells=421 (b)L_{max}=6, $T \approx 5$, # cells=916

Figure 2.6: Viability kernel of the consumption problem.

In a previous work [3], the UB-HJB scheme has been compared to the viability algorithm [17]. Many numerical examples have been handled on a regular grid (among others the Zermelo problem and the consumption problem) to prove the relevance of UB-HJB in this

kind of problems. In fact, UB-HJB provides a much better approximation of these sets (viability kernels, capture basins). Here we continue in the same direction and improve results given by UB-HJB scheme. The use of the adaptative method allows us to reach a better precision by optimizing the management of the memory.

Bibliography

[1] M. I. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Birkhäuser Boston, 1997.

[2] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications*. Springer, Paris, 1994.

[3] O. Bokanowski, S. Martin, R. Munos, and H. Zidani. An anti-diffusive scheme for viability problems. *Applied Numerical Mathematics*, 56:1147–1162, 2006.

[4] O. Bokanowski, N. Megdich, and H. Zidani. Convergence of a non monotone scheme for Hamilton Jacobi Bellman equations with discontinuous initial data. *in preparation*, 2007.

[5] O. Bokanowski and H. Zidani. Anti-dissipative schemes for advection and application to Hamilton-Jacobi-Bellman equations. *J. Sci. Comp.*, 30(1):1–33, 2007.

[6] I. Capuzzo-Dolcetta and P. L. Lions. Hamilton Jacobi equations with state constraints. *Tran. Amer. Math. Soc.*, 318(2), 1990.

[7] B. Cockburn and B. Yenikaya. An adaptive method with rigorous error control for the Hamilton-Jacobi equations. part i: the one-dimensional steady-state case. *App. Num. Math.*, 52:175–195, 2005.

[8] B. Cockburn and B. Yenikaya. An adaptive method with rigorous error control for the Hamilton-Jacobi equations. part ii: the two-dimensional steady-state case. *J. Comput. ph.*, 209:391–405, 2005.

[9] M. G. Crandall and P. L. Lions. Viscosity solutions of Hamilton Jacobi equations. *Tran. Amer. Math. Soc.*, 277:1–42, 1983.

[10] B. Després and F. Lagoutière. Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. *J.Sci. Comput.*, 16:479–524, 2001.

[11] M. Falcone and R. Ferretti. Semi-Lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and Godunov methods. *Journal of Computational Physics*, 175:559–575, 2002.

[12] M. Falcone and T. Giorgi. An approximation scheme for evolutive Hamilton-Jacobi equations. *Stochastic Analysis, Control, Optimization and Applications*, pages 289–1303, 1999.

[13] F. Frankowska and S. Plaskacz. Semicontinuous solutions of Hamilton-Jacobi equations with degenerate state constraints. *JMAA*, pages 818–838, 2000.

[14] F. Frankowska and R. B. Vinter. Existence of neighboring feasible trajectories: applications to dynamic programming for state constrained optimal control problems. *I.Optim.Theory Appl.*, 104:27–40, 2000.

[15] I. Gargantini. An effective way to represent quadtrees. *Communications of the ACM*, 25(12):905–910, 1982.

[16] L. Grune. Adaptative grid generation for evolutive Hamilton-Jacobi-Bellman equations. *Numerical methods for viscosity solutions and applications*, pages 153–172, 2001.

[17] P. Saint-Pierre. Approximation of viability kernel. *Appl. Math. Optim.*, 29:187–209, 1994.

[18] P. E. Souganidis. Approximation schemes for viscosity solutions of Hamilton-Jacobi equations. *Journal of differential equations*, 59:1–43, 1985.

CHAPTER 3

A fast anti-dissipative sparse implementation method

3.1 Introduction

We deal in this chapter with the numerical approximation of the first order Hamilton Jacobi Bellman (HJB) equation arising in some deterministic optimal control problems with state constraints.
We propose a fast implementation method relying on the HJB-UltraBee scheme and on a sparse technique of storage. This method uses in particular the special shape of the function we are interested in: we deal with piece wise constant initial data which take values 0 and 1. The corresponding solution is also piece-wise constant with values in $\{0,1\}$. We want in particular to localise accurately its discontinuity (the interface between 0 values and 1 values).

We have already coupled in the previous chapter the UltraBee scheme with an adaptative gridding technic leading to better precision. The method that we propose here achieves furthermore a gain of time and a better management of the storage capacity.
For the clarity of presentation, we detail the method in the two dimensional case. We also propose a trajectory reconstruction algorithm and validate the method on several academic tests.

Extension of the method to higher dimension does not involve any additional difficulty. We give an example in 3D. This method will also be applied to the atmospheric reentry model in the next chapter.

3.2 Preliminaries

We are interested in the numerical approximation of the following Hamilton-Jacobi-Bellman (HJB) equation: for all $t \in]0,T[$ and $x \in \mathbb{R}^n$

$$\min\Big(\vartheta_t(t,x) + \max_{\substack{\alpha\in\mathcal{A},\\ \lambda\in\Lambda(x)}} \{-\lambda f(x,\alpha).\vartheta_x(t,x)\};\ \vartheta(t,x) - \chi_{\mathcal{K}}(x)\Big) = 0, \tag{3.2.1a}$$

$$\vartheta(0,x) = \varphi(x), \tag{3.2.1b}$$

where the control set $\mathcal{A}$ is a compact of $\mathbb{R}^m$, Λ is a set valued function defined on $\mathbb{R}^n$ by

$$\Lambda(x) := \begin{cases} [0,1] & \text{if } x \in \mathcal{C}, \\ \{1\} & \text{otherwise.} \end{cases}$$

We assume that the dynamics $f : \mathbb{R}^n \mapsto \mathcal{A} \times \mathbb{R}^n$ is Lipschitz continuous. In the sequel, $\mathcal{K}$ and $\mathcal{C}$ will denote two closed sets of $\mathbb{R}^n$ such that $\mathcal{C} \subset \mathcal{K}$ and $\overset{\circ}{\mathcal{C}} \neq \emptyset$, $\int k \neq \emptyset$. The functions φ and $\chi_{\mathcal{K}}$ are defined on $\mathbb{R}^n$ by:

$$\varphi(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ 1 & \text{otherwise,} \end{cases} \quad \text{and} \quad \chi_{\mathcal{K}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{K}, \\ 1 & \text{otherwise,} \end{cases} \tag{3.2.2}$$

and are l.s.c. under our assumptions.

In order to be precise, the meaning we give to a discontinuous solution of (3.2.1) is the viscosity sense [10, 11] that we recall:
Let v be a l.s.c. function defined on $\mathbb{R}^+ \times \mathbb{R}^n$. v is a l.s.c. viscosity solution of (3.2.1) if for all test function $\phi \in C^1(\mathbb{R}^+ \times \mathbb{R}^n)$ and all minimum $(t,x) \in]0,+\infty[\times\mathbb{R}^n$ of $v-\phi$,

$$\min\left(\phi_t(t,x) + \max_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}} \{-\lambda f(x,\alpha)\cdot\phi_x(t,x)\}; v(x)-\chi_{\mathcal{K}}(x)\right) = 0, \tag{3.2.3}$$

and for all $x \in \mathcal{K}$,

$$\liminf_{\substack{\tau>0,\ \tau\to 0\\ \xi\in\overset{\circ}{\mathcal{K}},\ \xi\to x}} v(\tau,\xi) = \varphi(x), \tag{3.2.4}$$

where $\overset{\circ}{\mathcal{K}}$ denotes the interior of the set $\mathcal{K}$. For more details about this sense of solution we refer to chapter 5 of this manuscript and more generally to the relevant books of Bardi and Capuzzo Dolcetta [1] and Barles [2].
Eventhough we do not have any uniqueness result for equation (3.2.1), we are interested here in approaching numerically its solution.

This equation comes from some optimal control problems with state constraints as the Rendez-Vous problem dealt with in chapter 5. Under classical arguments [15], the value function of this problem is proved to be l.s.c.

We will combine for this method, as mentioned in the introduction, on one hand the anti-dissipative UltraBee scheme which will insure a good localization of the discontinuity, and on the other hand a sparse storage technique motivated by the piece wise constant shape of the solution.

We shall recall here the Ultrabee scheme for equation (3.2.1) in the 2-dimensional space through three steps. We first recall the UB scheme for linear advection in 1d [9, 12, 4] already detailed in chapter 2, then its extension to the two dimensional case using *Trotter splitting* device. Finally we give the way of applying the scheme to solve the HJB equation (3.2.1).

We introduce the following notation for the control $(\alpha,\lambda) \in \mathcal{A}\times[0,1]$,

$$u := (\alpha,\lambda).$$

3.3 The Ultrabee scheme

3.3.1 UB scheme for 1d linear advection

We first consider the discretization of the transport equation:

$$\begin{cases} u_t(t,x) + f(x)u_x(t,x) = 0,\ x\in\mathbb{R},\ t\geq 0,\\ u(0,x) = u_0(x),\ x\in\mathbb{R}. \end{cases} \tag{3.3.1}$$

with the initial condition u_0 assumed in $L^1_{loc}(\mathbb{R})$. Let $(x_j)_{j\in\mathbb{Z}}$ define a regular grid such that $x_{j+1}-x_j = \Delta x$, and $t_n = n\Delta t$ be a uniform time discretization. Here Δx and Δt are the

mesh sizes.
Let U_j^n denote a numerical approximation of the solution $u(t_n, x_j)$.
Then the UB scheme for (3.3.1) takes the finite volume form (2.2.3) of chapter 2.

We recall that $x_{j+\frac{1}{2}} = x_j + \frac{\Delta x}{2}$ and that we take as initial approximation of $u_0(x_j)$ the average value of u_0 on cell $]x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[=: M_j$.

This initialization has the particularity of coding exactly the discontinuities of u_0. In fact if u_0 is a step function with values in $\{0, 1\}$, and if a cell M_j contains only one discontinuity of u_0, then U_j^0 defines in a unique way this discontinuity localization (this is valid if furthermore u_0 is constant on M_{j-1} and M_{j+1}).
The numerical fluxes $U_{j+\frac{1}{2}}^{n,L}$ and $U_{j+\frac{1}{2}}^{n,R}$ are defined as in page 80. Notice that we can write (2.2.3) in the equivalent form:

$$U_j^{n+1} = U_j^n - \nu_j \left(U_{j+\frac{1}{2}}^{n,L} - U_{j-\frac{1}{2}}^{n,R} \right), \tag{3.3.2}$$

where the "local CFL" number $\nu_j := \frac{f(x_j)\Delta t}{\Delta x}$. As we are considering an explicit scheme, we need a CFL type condition:

$$|\nu_j| \leq 1 \quad \forall j \in \mathbb{Z}. \tag{3.3.3}$$

It is easy to check that under the CFL condition (3.3.3), we have

$$\min(U_{j-1}^n, U_j^n, U_{j+1}^n) \leq U_j^{n+1} \leq \max(U_{j-1}^n, U_j^n, U_{j+1}^n).$$

For other stability and convergence properties of this scheme, we refer to [9, 4].

Finally, as in [3] we use the notation F^L and F^R in this one-dimensional case:

$$U_{j+\frac{1}{2}}^{n,L} = F^L(U^n)_{j+\frac{1}{2}} \quad \text{for } j \in \mathbb{Z},$$

and

$$U_{j+\frac{1}{2}}^{n,R} = F^R(U^n)_{j+\frac{1}{2}} \quad \text{for } j \in \mathbb{Z}.$$

This notation will be usefull for the two dimensional presentation.

3.3.2 UB scheme for 2d linear advection

Now we consider the transport equation in 2D:

$$\begin{cases} u_t(t,x) + f_1(x)u_{x_1}(t,x) + f_2(x)u_{x_2}(t,x) = 0, \\ u(0,x) = u_0(x), \end{cases} \tag{3.3.4}$$

where $x := (x_1, x_2)$ belongs to a square box $\mathcal{D}$ of $\mathbb{R}^2$. We consider a Cartesian mesh $(x_{1,i}, x_{2,j})$ $(i \in \{1, \ldots, P_{x_1}\},\ j \in \{1, \ldots, P_{x_2}\})$, with constant mesh sizes $x_{1,i+1} - x_{1,i} = \Delta x_1$ and $x_{2,j+1} - x_{2,j} = \Delta x_2$, and assume the CFL condition

$$\max_{i,j} \left(\max \left(\frac{\Delta t}{\Delta x_1} |f_1(x_{1,i}, x_{2,j})|,\ \frac{\Delta t}{\Delta x_2} |f_2(x_{1,i}, x_{2,j})| \right) \right) \leq 1. \tag{3.3.5}$$

The UB scheme (3.3.2) may be extended to (3.3.4) as proposed in [12], i.e. by using simply a Trotter splitting or "alternate direction method", see [8]. We recall the algorithm explicited in [3].
The initialization is done, as in 1d, with the average value of u_0 on each cell:

$$U^0_{i,j} = \frac{1}{\Delta x_1 \Delta x_2} \int_{I_i \times J_j} u_0(x)\ dx, \tag{3.3.6}$$

where $I_i =]x_{1,i} - \frac{\Delta x_1}{2};\ x_{1,i} + \frac{\Delta x_1}{2}[$ and $J_j =]x_{2,j} - \frac{\Delta x_2}{2};\ x_{2,j} + \frac{\Delta x_2}{2}[$. But unlike the one dimensional case, the mean value $U^0_{i,j}$ is not sufficient here (and more generally for dimension $d \geq 2$) to recover the discontinuity localization.

We first make an evolution in the x_1-direction:

$$U^{n,1}_{i,j} = U^n_{i,j} - \frac{\Delta t}{\Delta x_1} f_1(x_{1,i}, x_{2,j}) \left(F^L(U^n_{.,j})_{i+\frac{1}{2}} - F^R(U^n_{.,j})_{i-\frac{1}{2}} \right) \tag{3.3.7}$$

where $U^n_{.,j} = (U^n_{i,j})_{i=1,\dots,P_{x_1}}$. Then we evolve in the x_2-direction:

$$U^{n+1}_{i,j} = U^{n,1}_{i,j} - \frac{\Delta t}{\Delta x_2} f_2(x_{1,i}, x_{2,j}) \left(F^L(U^{n,1}_{i,.})_{j+\frac{1}{2}} - F^R(U^{n,1}_{i,.})_{j-\frac{1}{2}} \right) \tag{3.3.8}$$

where $U^{n,1}_{i,.} = (U^{n,1}_{i,j})_{j=1,\dots,P_{x_2}}$.

The CFL condition (3.3.5) is natural here as it ensures the stability of the two splitting steps (3.3.7) and (3.3.8).

3.3.3 UB-HJB scheme

We now consider the discretization of the HJB equation (3.2.1), and assume the following CFL condition satisfied:

$$\max_{i,j,\alpha} \left(\max \left(\frac{\Delta t}{\Delta x_1} |f_1((x_{1,i}, x_{2,j}), \alpha)|,\ \frac{\Delta t}{\Delta x_2} |f_2((x_{1,i}, x_{2,j}), \alpha)| \right) \right) \leq 1. \tag{3.3.9}$$

The initialization of the scheme is done by the average values of φ:

$$V^0_{i,j} := \frac{1}{\Delta x_1 \Delta x_2} \int_{I_i \times J_j} \varphi(x) dx. \tag{3.3.10}$$

We also compute the average values of $\chi_{\mathcal{K}}$ on the grid

$$W_{i,j} := \frac{1}{\Delta x_1 \Delta x_2} \int_{I_i \times J_j} \chi_{\mathcal{K}}(x) dx. \tag{3.3.11}$$

Let $N_\alpha \geq 1$ be an integer and $(\alpha_k)_{k=1,\dots,N_\alpha}$ be a given discretization of the control set $\mathcal{A}$, We also consider the following approximation $\mathcal{W}$ of Λ:

$$\mathcal{W}(x) = \begin{cases} \{1\} & \text{if } x \notin \mathcal{C}, \\ \{0,1\} & \text{if } x \in \mathcal{C}. \end{cases} \tag{3.3.12}$$

We denote by $\left(V_{i,j}^{n+1,UB}(\alpha,\lambda)\right)_{i,j}$ the UB scheme approximation obtained from $\left(V_{i,j}^{n}\right)_{i,j}$ by using the advection $-\lambda f(\cdot,\alpha)$, i.e. one time step of the UB scheme for

$$u_t(t,x) - \lambda f(x,\alpha)\cdot u_x(t,x) = 0.$$

Then the UB-HJB scheme is given by

$$V_{i,j}^{n+1} = \max\left(W_{i,j}, \min_{\substack{k=1,\cdots,N_\alpha;\\ \lambda\in\mathcal{W}(x_{1,i},x_{2,j})}}\left(V_{i,j}^{n+1,UB}(\alpha_k,\lambda)\right)\right), \tag{3.3.13}$$

that we denote by : $\quad V^{n+1} := S_{UB}(V^n)$.

Remark 3.1. *Notice that, at the first step of UB-HJB scheme, when we compute the average values V^0, the only components which values are strictly between 0 and 1 are those corresponding to the cells containing the front Γ_0 [1] of the initial condition.*
In dimension 1, we proved in chapter 1 that, for every $n \geq 0$, the interface Γ_{t^n} [2] keeps well localized on a few cells. In dimension 2, we observe numerically, on several examples, that:

- *Γ_{t^n} is still well localized by the UB-HJB scheme*
- *for every $t_n \geq 0, i \in \{1,\cdots,P_{x1}\}, j \in \{1,\cdots,P_{x2}\}$, the computed value $V_{i,j}^n$ is a good approximation of the average $\dfrac{1}{\Delta x_1 \Delta x_2}\displaystyle\int_{I_i\times J_j} \vartheta(t_n,(x_1,x_2))\,dx_1dx_2$.*

However, we don't have yet any precise theoretical result to claim (for dimension $d \geq 2$). Nevertheless the good localization of the front allows to use in a relevant way a sparse storage technique that we present in the next section.

3.4 Sparse fast implementation

Let

$$\mathcal{D} := [X_{1\min}, X_{1\max}] \times [X_{2\min}, X_{2\max}], \qquad \text{with } \mathcal{K} \subset \mathcal{D},$$

denote the domain of calculations and consider the grid $\mathcal{G}$ with Cartesian mesh $(x_{1,i}, x_{2,j})$, $i \in \{1,...,P_{x_1}\}$, $j \in \{1,...,P_{x_2}\}$, and constant mesh sizes

$$\Delta x_1 := \frac{X_{1\,\max} - X_{1\,\min}}{P_{x_1}} \quad \text{and} \quad \Delta x_2 := \frac{X_{2\,\max} - X_{2\,\min}}{P_{x_2}}.$$

We will denote by V_I^n the numerical value on cell $M_I \in \mathcal{G}$, where $I := (i,j),\ i \in \{1,\cdots,P_{x_1}\},\ j \in \{1,\cdots,P_{x_2}\}$, and $M_I := I_i \times J_j$ with

$$I_i :=]X_{1min} + (i-1)\Delta x_1; X_{1min} + i\Delta x_1[\text{ and } J_j :=]X_{2min} + (j-1)\Delta x_2; X_{2min} + j\Delta x_2[.$$

[1] Γ_0 is the interface separating 0-values of $\vartheta(t=0,.)$ and its 1-values
[2] Γ_{t_n} is the interface between 0-values and 1-values of $\vartheta(t_n,.)$.

We expose in the sequel the details of how to make calculations more quickly in the case of approximating the solution of (3.2.1) with an initial condition φ and an obstacle function $\chi_{\mathcal{K}}$ defined by (3.2.2).

Initialization
Since the function φ is equal to 0 on $\mathcal{C}$ and 1 outside, it is not necessary to compute the averages $V^0_{i,j}$ (defined in (3.3.10)) for all $(i,j) \in \{1,\cdots,P_{x_1}\} \times \{1,\cdots,P_{x_2}\}$. In fact, it suffices to calculate $V^0_{i,j}$ for the cells $M_{i,j} := I_i \times J_j$ crossed by $\partial\mathcal{C}$ (the boundary of $\mathcal{C}$), and their neighboring[3] cells.

The remaining cells are stored directly in sparse type and their values can be deduced from the computed values. In fact, for a non stored cell M, its mean value is the same as the nearest stored value equal to 0 or 1.

The same remark is valid for the computation of $W_{i,j}$ the average values of $\chi_{\mathcal{K}}$ (as defined in (3.3.11)).

These two remarks allow to reduce the initialization step on the grid to a narrow band around $\partial\mathcal{C}$ and $\partial\mathcal{K}$.

Remark 3.2. *Numerically we did initialization calculations for cells $M_{i,j}$ such that*

$$d(x_{1,i},\mathcal{C}) \leq 2\Delta x_1 \text{ and } d(x_{2,j},\mathcal{C}) \leq 2\Delta x_2,$$

with $x_{1,i} := X_{1min} + (i-\frac{1}{2})\Delta x_1$, $x_{2,j} := X_{2min} + (j-\frac{1}{2})\Delta x_2$ and d is defined by:

$$d(x_{1,i},\mathcal{C}) = \min\{d(x_{1,i},x_1),\ (x_1,x_2) \in \mathcal{C}\}, \quad d(x_{2,j},\mathcal{C}) = \min\{d(x_{2,j},x_2),\ (x_1,x_2) \in \mathcal{C}\}.$$

Boundary conditions
As we suppose $\mathcal{K} \subset \mathcal{D}$, we know that outside $\mathcal{D}$ the value function ϑ takes the value 1. Hence we involve in our implementation a value on the border of the domain $\mathcal{D}$ fixed to 1. This means simply that the exterior of the grid is forbidden for the trajectories.

The algorithm
Recall that, for a given time $t \geq 0$, the domain $\mathcal{D}$ is partitionned into a region where the value function $\vartheta(t,\cdot)$ takes 0 and another region where it takes 1. These two regions are separated by an interface Γ_t which represents the discontinuity line of $\vartheta(t,\cdot)$.

One one hand, we assume that the UB-HJB scheme gives a good approximation of Γ_{t^n} (see Remark 3.1). On the other hand, under the CFL condition, the discontinuities can not jump more than one cell at each time step. Therefore, it is possible to do the computations only in the narrow band around the interface Γ_{t_n}. In other words, we reduce the approximation of $\vartheta(t_n,\cdot)$ to the determination of the evolution of Γ_{t_n}: we treat the problem as a front propagation. Outside this narrow band the cells values do not change. Hence these remaining cells should be non stored or stored in sparse type.

[3] in dimension 2, we mean by neighboring cells of a given cell $M_{i,j}$ the eight cells sharing an edge or a vertex with $M_{i,j}$

Remark 3.3. *i) In the 2D scilab code, numerical values are stored in a column vector V such that to cell M_I, $I=(i,j)$, $1\leq i\leq Px_1$, $1\leq j\leq Px_2$, corresponds the value $V((j-1)Px_1+i)$. In fact, we cover the grid for increasing x_2 coordinates and for each fixed x_2, we cover the x_1 coordinates in an increasing order. Notice that this storage technique can easily be extended for higher dimensions, it suffices to choose an order in which the grid is stored into a vector. ii) In scilab code, the sparse type takes value "0". Hence we shift the numerical values in V (which are in $[0,1]$) by one, then they become in $[1,2]$. In this way we can reserve the value 0 for the sparse type, and the final vector takes values in $\{0\}\cup[1,2]$.*

Let M be a cell of the grid $\mathcal{G}$. We introduce the notations $|M|$ for the area of M (in 2D, $|M|=\Delta x_1\Delta x_2$), and $\mathcal{N}(M)$ for the neighboring cells of M. The neighbors $\mathcal{N}(M)$ are cells of $\mathcal{G}$ sharing an edge or a vertex with M. In particular, in 2D a cell M may have at most eight neighboring cells in $\mathcal{G}$.
We also denote a non stored value associated to a cell M_I in the numerical vector V^n by

$$V_I^n=\emptyset.$$

Sparse algorithm

Initialization

Projection step: For each cell M_I of the grid $\mathcal{G}$, compute:

$$V_I^0=\frac{1}{|M_I|}\int_{M_I}\varphi(x)dx \quad\text{and}\quad W_I=\frac{1}{|M_I|}\int_{M_I}\chi_{\mathcal{K}}(x)dx,$$

and set $n=0$.

Elimination step: For each $M_I\in\mathcal{G}$ such that $V_I^n\in\{0,1\}$,

$$\text{if } \left(\forall M_J\in\mathcal{N}(M_I),\ V_J^n=V_I^n\right) \text{ set } V_I^n=\emptyset.$$

Evolution loop

Extension step: Construct the extended vector $\widetilde{V}^n$ initialized by:

$$\widetilde{V}^n:=V^n.$$

For each $M_I\in\mathcal{G}$ such that $V_I^n\in\{0,1\}$ (see Figure 3.2),

$$\forall M_J\in\mathcal{N}(M_I) \text{ such that } V_J^n=\emptyset, \text{ set } \widetilde{V}_J^n=V_I^n.$$

Evolution step: For each $M_I\in\mathcal{G}$ such that $V_I^n\in[0,1]$, apply the UB-HJB scheme (3.3.13) as described in section 3.3.3,

$$V_I^{n+1}:=S_{UB}(V^n)_I,$$

Prevision step: For each $M_I\in\mathcal{G}$,

$$\text{if } \widetilde{V}_I^n\in\{0,1\} \text{ and } V_I^{n+1}=\emptyset \text{ set } V_I^{n+1}=\widetilde{V}_I^n.$$

Elimination step: Apply for V^{n+1} the previous elimination step.

Hence we get the numerical approximation V^{n+1} of the average values of $\vartheta(t_{n+1},.)$ on $\mathcal{G}$.

Remark 3.4. *Recall that after the elimination step done in the initialization, any non stored value can be recovered from the nearest stored neighbor value.*

Remark 3.5. *For sake of clarity, we explain the relevance of the extension and prevision steps in 1D. After the evolution step is run, let a cell M_I be such that*

$$V_I^n \in]0,1[, \quad V_{I+1}^n = 1 \text{ and } V_{I+2}^n = \emptyset,$$

then at t_{n+1} we may obtain

$$V_I^{n+1} = 0, \quad V_{I+1}^{n+1} \in]0,1[\text{ and } V_{I+2}^{n+1} = \emptyset$$

In this special case, we do not know any more whether $V_{I+2}^{n+1} = \emptyset$ corresponds to value 0 or 1 (both values may be non stored) as its immediate neighboring value is neither 0 nor 1. Thanks to the extension step, the extended vector $\widetilde{V}^n$ allows to define such values without any ambiguity:

$$V_{I+2}^{n+1} = \widetilde{V}_{I+2}^n = 1 \text{ as } \widetilde{V}_{I+2}^n = V_{I+1}^n = 1.$$

This is illustrated in Figure 3.1.

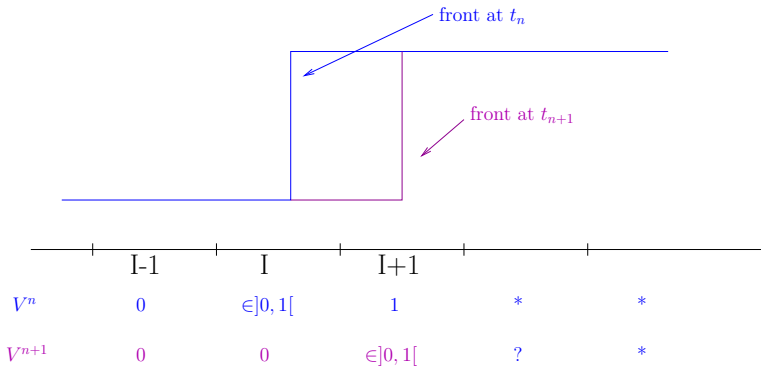
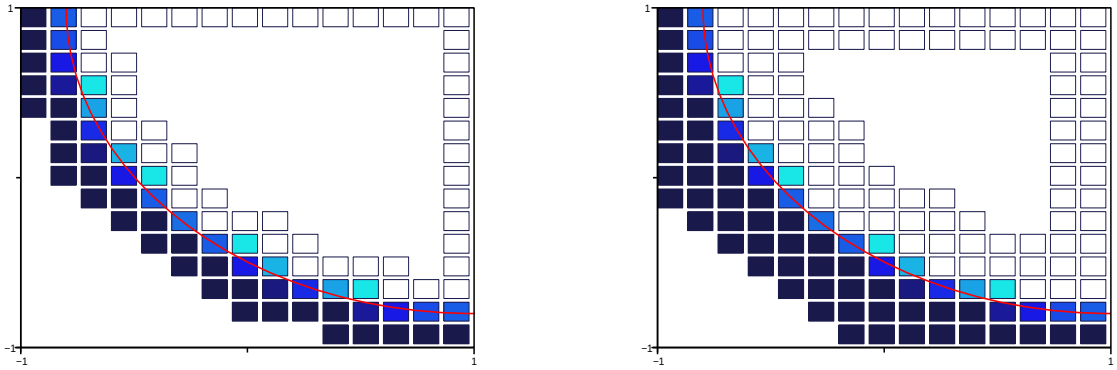


Figure 3.1: Relevance of the Prevision-Extension steps.

Evolution

As precised in section 3.3, the UB scheme is defined in 1D and extended to 2D by using the splitting device. In our code, we first define an evolution function in 1D which applies the UB algorithm in one direction (the algorithm of section 3.3.1):

$$V^{n+1,1} = F_{UB}(V^n, x_1),$$



V $\widetilde{V}$

Figure 3.2: the stored values V and the extended values $\widetilde{V}$

then we use the same function in the "x_2"-direction to get a 2D evolution.
We only have to reorganize suitably the vector $V^{n+1,1}$

$$V^{n+1} = F_{UB}(V^{n+1,1}, x_2).$$

For 3D evolution, we iterate the use of the same function three times

$$V^{n+1,1} = F_{UB}(V^n, x_1),\ V^{n+1,2} = F_{UB}(V^{n+1,1}, x_2),\ V^{n+1} = F_{UB}(V^{n+1,2}, x_3).$$

Notice that this coding structure facilitates the extension to higher dimension.

3.5 Optimal trajectories reconstruction

The question of approximating optimal trajectories is of considerable practical interest in control theory. A scope on the subject is in particular proposed in [1, chapter VI and appendix A] [4] for a Lipschitz continuous initial function ψ.
Let k denote the space mesh size and $N \in \mathbb{N}^*$, $h := \Delta t = \frac{T}{N}$. Let the numerical approximated function be denoted by v_h^k (obtained by a linear interpolation of the numerical values on the grid nodes with the Semi Lagrangian scheme) and let α^* be defined by

$$v_h^k(t,x) = v_h^k(t - \Delta t, x + \Delta t f(x, \alpha^*(x))), \tag{3.5.1}$$

and y_m^* by:

$$\begin{aligned} y_0^* &= x, \\ y_{m+1}^* &= y_m^* + \Delta t f(y_m^*, \alpha^*(y_m^*)). \end{aligned} \tag{3.5.2}$$

[4] The results summarized in [1] are given for the infinite horizon problem, we transpose them to our problem.

Then under the additional assumption that α^*, defined by (3.5.1) is single valued, we can prove [1, Appendix A] (see also [6]) that

$$\psi(y_N^*) \to v(T,x),$$

when $h \to 0^+,\ k \to 0^+$ and $k/h \to 0^+$. Here v is the unique viscosity solution of the HJB equation (without any state constraint) with initial data ψ.

More recently a paper by Rowland and Vinter [14] treats the semi discrete problem (in time) and proposes an algorithm that constructs an approximated optimal trajectory.

The constructed piece wise continuous trajectory admits cluster points with respect to the topology of uniform convergence, and to each cluster point is associated an optimal control for the problem [14, Theorem 3.2].

We explain now our algorithm in order to recover an "approximated time optimal" trajectory, starting from a given initial position x in $\mathcal{K}$.

We suppose that for each time step $t_n = n\Delta t$ the numerical values V^n (obtained by the UB-HJB scheme) give a good approximation of the average values of $\vartheta(t_n,\cdot)$ on the grid. This good approximation is proved in dimension 1 and observed on several examples in dimension 2. In addition to values V^n we also need to save a selection of optimal controls u^n that allow to get V^n from $V^{n,UB}$. Hence using expression (3.3.13), the control $u^n := (\alpha^n, \lambda^n)$ satisfies:

$$\min_{\substack{k=1,\cdots,N_\alpha \\ \lambda\in\mathcal{W}(M_I)}} (V_I^{n,UB}(\alpha_k,\lambda)) = V_I^{n,UB}(\alpha_I^n,\lambda_I^n), \quad \forall M_I \in \mathcal{G}. \tag{3.5.3}$$

If many controls satisfy (3.5.3), we choose one of them.

Let the starting point x be a reachable point i.e. such that

$$\exists I_0 \text{ and } n\in\mathbb{N} \text{ with } x\in M_{I_0} \text{ and } V_{I_0}^n \in [0,1[. \tag{3.5.4}$$

We have to detect the first time t_n such that the numerical front zone reaches cell M_{I_0},

$$\tau^0(x) := \inf\{t_n \text{ such that } V_{I_0}^n \in [0,1[\text{ and } V_{I_0}^{n-1} = 1\}. \tag{3.5.5}$$

Then we assume that $\tau^0(x)$ is an approximation of $\mathcal{T}(x)$ the minimum time for x to reach $\mathcal{C}$,

$$\mathcal{T}(x) := \inf\{t\geq 0,\ \vartheta(t,x) = 0\}.$$

Notice that under our assumption, $\tau^0(x)$ is also an approximation of $\mathcal{T}(z)$ for all $z\in M_{I_0}$.

We suppose by now that $\tau^0(x) < \infty$ and $\tau^0(x) = t_n$.

Recall that the UB-HJB algorithm makes the numerical discontinuity evolve with a piece wise constant dynamics (as shown in chapter 1). We use this approached dynamics in our reconstruction algorithm, we will denote it by f^S and define it almost every where by:

$$f^S(x,\alpha) := f(x_I,\alpha),\ \forall x\in M_I, \tag{3.5.6}$$

where x_I is the center of cell M_I.

Let $y^0 := x$, then we can reconstruct an "approximation" of the position $y_x^*(\Delta t)$ (here y_x^* denotes one of the optimal trajectories starting at x), that we denote by y^1, using the control $u_{I_0}^n := (\alpha_{I_0}^n, \lambda_{I_0}^n)$,

$$y^1 := x + \Delta t\ \lambda_{I_0}^n f^S(x, \alpha_{I_0}^n).$$

Now we assume the following: the cell M_{I_1} containing y^1 is such that

$$V_{I_1}^{n-1} \in [0, 1[. \tag{3.5.7}$$

If (3.5.7) is true, this means that y^1 follows correctly the numerical front position at t_{n-1} which crosses the cell M_{I_1}.
Hence, we can iterate the operation on y^1 knowing the control $u_{I_1}^{n-1}$ (we construct then an "approximation" of $y_{y^1}^*(\Delta t)$ with $y_{y^1}^*$ an optimal trajectory starting at y^1) and so on, we construct a sequence $(y^k)_{0\le k\le n}$, assuming that $M_{I_k} \ni y^k$ is such that

$$(H) \qquad V_{I_k}^{n-k} \in [0, 1[, \quad \forall 0 \le k \le n,$$

such that y^k "approximates" $y_{y^{k-1}}^*(\Delta t)$, $1 \le k \le n$, see Figure 3.3.

We can give by now the algorithm of this reconstruction
The reconstruction algorithm:

Initialization. Run the UB-HJB algorithm for a sufficiently long time t_N and save values V^k and optimal controls u^k satisfying (3.5.3) for $0 \le k \le N$. Let $x \in M_{I_0}$ be a reachable position satisfying (3.5.4). Then define by (3.5.5) the approximated minimum time $\tau^0(x)$. We suppose that $\tau^0(x) = t_n$ and set $y^0 := x$.

Loop. For $k = n, \cdots, 1$, let $M_{I_{n-k}}$ be the cell in $\mathcal{G}$ such that $M_{I_{n-k}} \ni y^{n-k}$. Check that $V_{I_{n-k}}^k \in [0, 1[$. Knowing $y^{n-k} \in M_{I_{n-k}}$ and u^k, set

$$y^{n-k+1} := y^{n-k} + \lambda_{I_{n-k}}^k\ f^S(y^{n-k}, \alpha_{I_{n-k}}^k)\Delta t. \tag{3.5.8}$$

In practice:
The hypothesis (H) is almost never satisfied numerically when applying the previous reconstruction algorithm. Hence we have to handle some corrections at each step of time in order to follow correctly the numerical front. We give some implementation details that we used in our program

- At each reconstruction step, we evolve at most during 3 step of time Δt: when we get y^{n-k+1} in (3.5.8), if we obtain $V_{I_{n-k+1}}^{k-1} = 1$ then we iterate (3.5.8) with $2\Delta t$ (and then $3\Delta t$) instead of Δt.

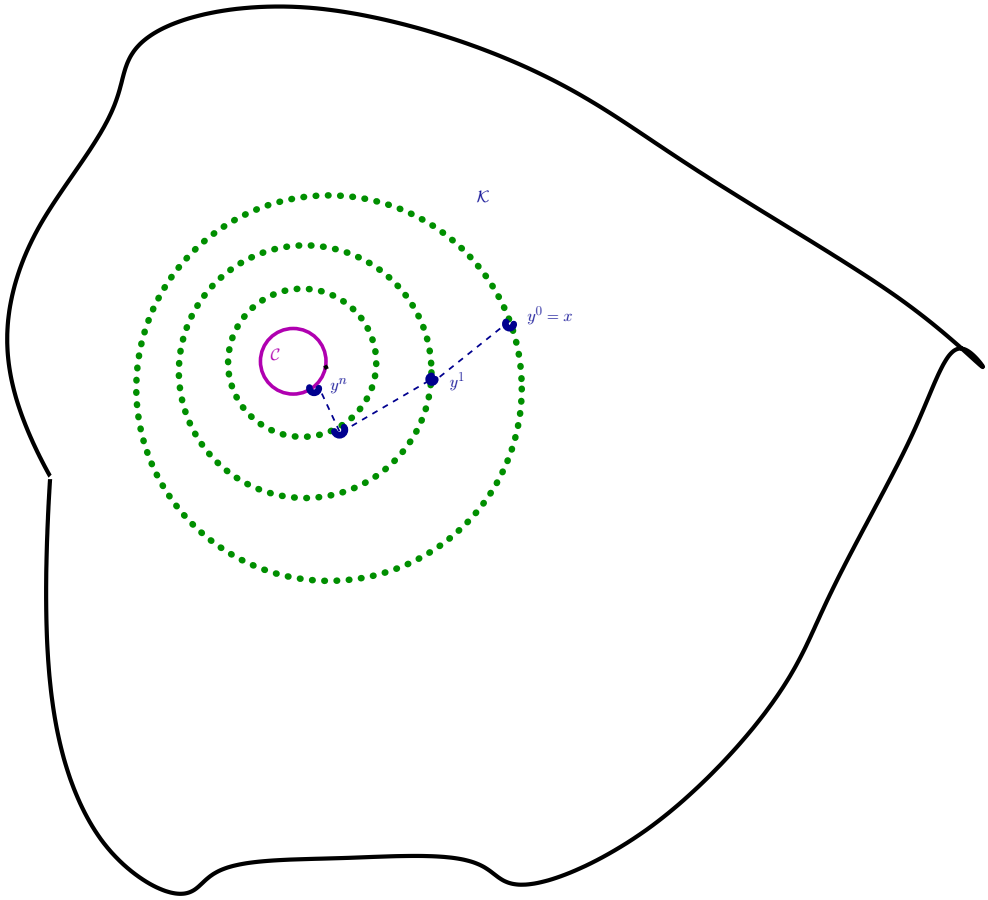


Figure 3.3: Optimal trajectory reconstruction

- When we get $V^{k-1}_{I_{n-k+1}} = 0$ and $V^{k-1}_{I_{n-k}} \in [0,1[$ we go back i.e. set

$$y^{n-k+1} \equiv y^{n-k}.$$

Remark 3.6. *We may store all optimal controls for each cell M_I of the grid and at each step of time t_n,*

$$\mathcal{U}^n_I := \{u^n_I := (\alpha^n_I, \lambda^n_I) \text{ satisfying } (3.5.3)\}.$$

This may facilitate the tracking of the numerical front but will require much more memory allocation.

3.6 Some 2D simulations

In the following simulations, we approximate the solution ϑ of (3.2.1). The dark blue cells in the figures are those with value 1 (prohibited area or not yet reached cells), light blue cells are those with numerical value in $]0,1[$ (area containing the discontinuities of ϑ), and white cells have value 0 (already reached area). We also represent the target in green and the obstacle in red which are specified in each example.

Stopping criteria. By using the UB-HJB algorithm, we evolve in time and compute, for $n \geq 0$, an approximation V^n of $\vartheta(t_n, \cdot)$ (with a time step $\Delta t > 0$ satisfying the CFL condition (3.3.9)). Then we decide to stop the scheme when the numerical front Γ_{t_n} does not evolve any more. In practice, the UB-HJB scheme is stopped when

$$||V^n - V^{n-1}||_{L^1} := \Delta x_1 \Delta x_2 \sum_{i,j} |V^n_{i,j} - V^{n-1}_{i,j}| \leq tol, \tag{3.6.1}$$

where *tol* is a fixed tolerance. This tolerance may be in relation with the mesh size and will be specified in each example. If not specified it is equal to 10^{-4}.

Examples 3.6.4, 3.6.5 and 3.6.6 that we are presenting here are also treated in the PH.D. thesis of Cristiani [7] using a Fast Marching-Semi Lagrangian method in the context of minimum time problem. Notice in particular that the front positions that we obtain numerically in Figures 3.10, 3.11 and 3.12 are comparable to the level sets obtained in [7].

3.6.1 Zermelo navigation problem

This problem is classical in optimal control. Consider a boat moving towards a target $\mathcal{C}$ with a constant magnitude equal to a, and submitted to a stream. The dynamics of the system $f : \mathbb{R}^2 \times \mathcal{A} \to \mathbb{R}^2$ is defined by

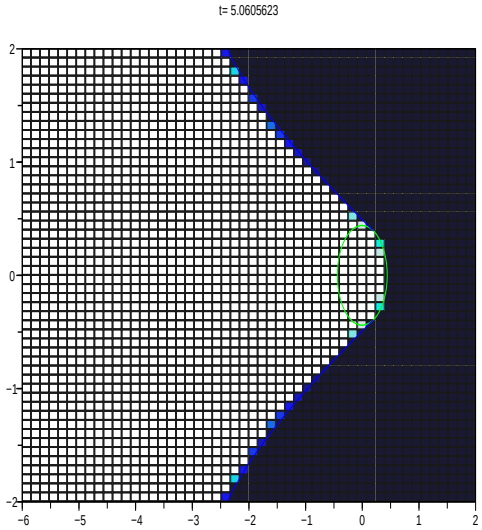
$$f((x_1, x_2), (a, \theta)) = (1 - \beta x_2^2 + a\cos(\theta), a\sin(\theta)),$$

where the constant $\beta = 0.1$, and $\mathcal{A} := [0, 0.44] \times [0, 2\pi]$ is the set of controls.

Let $\mathcal{D} = \mathcal{K} := [-6, 2] \times [-2, 2]$, and $\mathcal{C}$ be the disk with center $(0, 0)$ and with radius $r_{\mathcal{C}} = 0.44$.

In the numerical tests, we consider a Cartesian grid with $P_{x1} \times P_{x2}$ cells. We use N_α admissible control values to approximate the set $\mathcal{A}$:

$$\mathcal{A} \sim \left\{0; 0.44\right\} \times \left\{4\ell\pi/N_\alpha \mid \ell = 0, \cdots, \frac{N_\alpha}{2} - 1\right\}.$$



CFL=0.99, $T \sim 5$, $Px_1 = Px_2 = 25$, and $N_\alpha = 2 * 10$

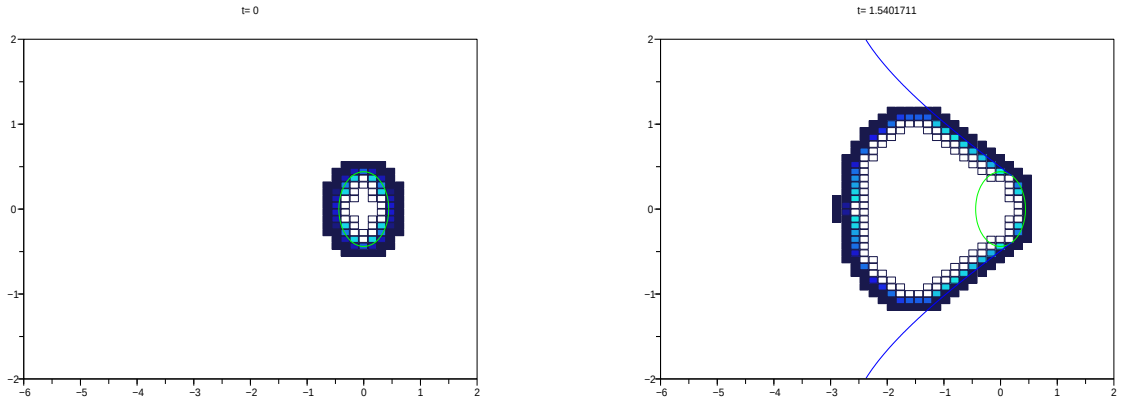
Figure 3.4: Zermelo navigation problem

In Figure 3.4, we show the approximation V^n of the function ϑ at $T = t_n$, with n large enough such that the stopping test (3.6.1) holds for $tol = 10^{-4}$. We also represent the exact interface Γ_T separating the 0-values of $\vartheta(T, \cdot)$ and its 1-values. Here the computations are done, with a $CFL = 0.99$, in a "full" grid of $P_{x1} \times P_{x2} = 25 \times 25$ cells. The number of admissible control values considered is $N_\alpha := 2 * 10$. Let us stress that, even in that coarse grid, the front Γ_T is well localized.

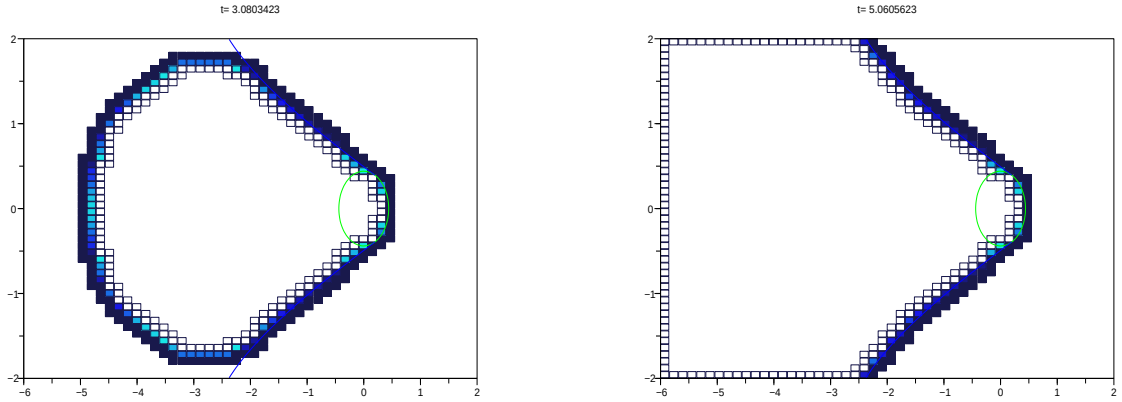
In Figure 3.5, we represent at different times, the sparse grid (i.e. only cells where we made the computations). We give in Table 3.1. the cpu-times for the computations done with scilab code on regular grids and sparse grids.We considered in this table the stopping test

$$||V^n - V^{n-1}||_{L^1} \leq tol \text{ or } t^n \geq 6,$$

with $tol = 0.01\Delta x_1 \Delta x_2$.



$T = 0$ $T = 1.5$



$T = 3$ $T \sim 5$

Figure 3.5: Zermelo problem with CFL=0.99, $P_{x1} = P_{x2} = 50$, and $N_\alpha = 2 \times 10$

Regular grid		
$P_{x1} = P_{x2} \,/\, N_\alpha$	2*5	2*10
25	0.92s	0.92s
50	17.62s	18.19s
100	495.33	496.72

Sparse grid		
$P_{x1} = P_{x2} \,/\, N_\alpha$	2*5	2*10
25	0.58s	0.57s
50	4.76s	4.87s
100	62.92s	62.23

Table 3.1: CPU-time, Scilab code (Zermelo problem)

These calculations have been run with Scilab 4 on a computer equipped with AMD Opteron 2GHz processor, 8GB RAM.

Notice that the computational time is twice bigger on a regular grid for the resolution 25×25 and it is about four times bigger for the resolution 50×50. For a grid of 100×100 the gain is about seven.

We consider by now obstacles (the triangle and the rectangle in red), and we are interested in some optimal trajectories reconstruction, see Figure 3.6. The left figure shows the maximal reachable area, the right one shows some time optimal trajectories starting from admissible points (i.e. points in the reachable area).

3.6.2 An example with constant dynamics

In this example, the domain of calculations is $\mathcal{D} = [-1, 1] \times [-1, 1]$. The dynamics is given by

$$f(x; \alpha) = (-1; -1), \quad \forall x \in \mathcal{D}.$$

We suppose that the target $\mathcal{C}$ is the disk centered in $(x_{1\mathcal{C}} = 0.5; x_{2\mathcal{C}} = 0.5)$ with radius $r_{\mathcal{C}} = 0.3$. We also consider an obstacle $\mathcal{O}$ which is the square defined by

$$\mathcal{O} := \{(x_1, x_2),\ \max(|x_1 - x_{1\mathcal{O}}|, |x_2 - x_{2\mathcal{O}}|) < r_{\mathcal{O}}\},$$

with $r_{\mathcal{O}} = 0.1,\ x_{1\mathcal{O}} = x_{2\mathcal{O}} = -0.1$. The admissible set $\mathcal{K}$ is given by:

$$\mathcal{K} := \mathcal{D} \setminus \mathcal{O}.$$

We can remark (see figure 3.7) that in the coarse simulation ($P_{x_1} = P_{x_2} = 51$), the straight lines (starting at the corners of the obstacle and delimiting the exact reachable set) cross cells with numerical values in $]0, 1[$. This is due to the antidissipative behaviour of the scheme which gives good quality approximation even for long times.

3.6.3 A thin target problem

The example we are treating now is very well known in optimal control, an analysis of necessary conditions on the optimal control is proposed in [5, Example1.29, p198].

We consider a thin target $\mathcal{C}$ located in $\{(0,0)\}$. In this case, we approximate numerically the target by the cell containing $\mathcal{C}$. When $\mathcal{C}$ coincides with a node of the grid, the numerical target is:

$$\{x := (x_1, x_2),\ |x_1| \leq \frac{\Delta x_1}{2},\ |x_2| \leq \frac{\Delta x_2}{2}\}.$$

The trajectories of the problem are governed by the dynamics

$$f((x_1, x_2), \alpha) := (x_2;\ \alpha), \quad \forall (x_1, x_2) \in \mathcal{D},\ \alpha \in \mathcal{A},$$

where the control α takes values in $\mathcal{A} := [-1, 1]$. We set $\mathcal{K} = \mathcal{D} := [-1, 1]^2$.

In figure 3.8, we show the numerical solutions computed with $P_{x_1} = P_{x_2} = 81$, and $N_\alpha = 3$, the set $\mathcal{A}$ being discretized into $\{-1, 0, 1\}$. We also represent, in figure 3.9, some optimal trajectories reaching the target in a time less than $T := 3.07$.

3.6.4 An example with two obstacles

In this example, we set:

$$\mathcal{D} := [-2; 2]^2,$$

the thin target set $\mathcal{C} := \{(-1, -1)\}$ is approximated numerically by the cell containing it. We suppose that we have two obstacles located in

$$\mathcal{O} := \big\{[0; 0.5] \times [-2; 1.5]\big\} \cup \big\{[1; 1.5] \times [-1.5; 2]\big\}.$$

The admissible set is: $\mathcal{K} := \mathcal{D} \setminus \mathcal{O}$, and the dynamics f is defined by:

$$f(x, \theta) := (\cos\theta, \sin\theta), \quad \text{for } \ \theta \in [0, 2\pi].$$

The control set $\mathcal{A} := [0, 2\pi]$ is discretized into $N_\alpha = 15$ controls,

$$\mathcal{A} \sim \{\frac{i}{N_\alpha} 2\pi,\ i = 0, \cdots, N_\alpha - 1\}.$$

We represent in figure 3.10 the fronts Γ_t, for each ten steps of time with $\Delta t = 0.036$. The calculations are stopped at $T \sim 10.3s$ with the stopping test (3.6.1) and the tolerance $tol = 0.5\Delta x_1\ \Delta x_2$. We show some trajectories reconstruction.

we compare in Table 3.2 the time T_F necessary for the front to reach the starting points and the time T_R needed by the reconstructed trajectories to reach the target $\mathcal{C}$. The starting points are given by

$$X_1 = (1.7; 1.5); \quad X_2 = (-0.67; 2); \quad X_3 = (-1.8; 0.67).$$

starting points	T_F	T_R
X_1	9.84 s	4.153s
X_2	2.88s	1.518s
X_3	1.8s	0.858s

Table 3.2: Optimal trajectories reconstruction: CPU-time.

3.6.5 Poincaré model

In this test the target is again thin $\mathcal{C} := (-0.65, -0.65)$, and approximated by the cell containing it. The domain of calculations is

$$\mathcal{D} := [-1, 1]^2,$$

and the dynamics is discontinuous and defined by

$$f((x_1, x_2), \alpha) := \begin{cases} (\cos(\alpha), \sin(\alpha))(1 - (x_1^2 + x_2^2)) & \text{if } x_1^2 + x_2^2 < 1, \\ (0, 0) & \text{otherwise,} \end{cases}$$

The set of controls $\mathcal{A} := [0, 2\pi]$ is discretized into $N_\alpha = 151$

$$\mathcal{A} \sim \{\frac{i}{N_\alpha} 2\pi,\ i = 0, \cdots, N_\alpha - 1\}.$$

We show in Figure 3.11 the numerical front position after each ten step of time $10 * \Delta t$, $\Delta t = 0.01$, until the required tolerance is reached $tol = 10^{-3}\Delta x_1 \Delta x_2$.

3.6.6 An Eikonal example

Here the target is the cell centered in $(0, 0)$. The domain is

$$\mathcal{D} := [-2, 2]^2,$$

and the dynamics is defined by

$$f((x_1, x_2), \alpha) := (\cos(\alpha), \sin(\alpha))|x_1 + x_2|.$$

The set of controls $\mathcal{A} := [0, 2\pi]$ is discretized into $N_\alpha = 150$. We show in Figure 3.12 the numerical front position after each $15 * \Delta t$, $\Delta t = 0.01$, until the required tolerance is reached $tol = 10^{-2}\Delta x_1 \Delta x_2$.

3.7 A 3D simulation

We also implement the sparse algorithm in 3D, We will see in particular in the next chapter the application of this algorithm to the atmospheric re-entry problem. We first give here a preliminary academic example inorder to validate the code.

We consider a thin target example, the analogous of Example 3.6.3 in 3D. For the theoretical study of this example, see Lee and Markus [13, Chapter 2].

The thin target is located in

$$\mathcal{C} := \{(0,0,0)\}.$$

that we approximate numerically by

$$\{x := (x_1, x_2, x_3),\ |x_1| \leq \frac{3\Delta x_1}{2},\ |x_2| \leq \frac{3\Delta x_2}{2},\ |x_3| \leq \frac{3\Delta x_3}{3}\}.$$

The domain of calculations is

$$\mathcal{D} := [-1,1]^3,$$

discretized into 71*71*71 cells.
The dynamics of the problem is as follows:

$$f((x_1,x_2,x_3),\alpha) := (x_2,x_3,\alpha),$$

with the control α taking values in $\mathcal{A} := [-1,1]$ and discretized into

$$\mathcal{A} \sim \{-1,0,1\}.$$

We show in Figure 3.14 the reachable sets at $T = 0.5s$ and $T = 0.9s$. We also propose in Figure 3.15 some optimal trajectories reconstruction (with lower resolution $Px1 = Px_2 = Px_3 = 31$). The target is the cell in light blue.

Bibliography

[1] M. I. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and viscosity solutions of Hamilton Jacobi Bellman equations*. Birkhäuser Boston, 1997.

[2] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications*. Springer, Paris, 1994.

[3] O. Bokanowski, S. Martin, R. Munos, and H. Zidani. An anti-diffusive scheme for viability problems. *Applied Numerical Mathematics*, 56:1147–1162, 2006.

[4] O. Bokanowski and H. Zidani. Anti-dissipative schemes for advection and application to Hamilton Jacobi Bellman equations. *J. Sci. Comp.*, 30(1):1–33, 2007.

[5] F. Bonnans and P. Rouchon. *Commande et optimisation de systèmes dynamiques*. Les éditions de l'école polytechnique, 2005.

[6] I. Capuzzo Dolcetta and H. Ishii. Approximate solutions of the bellman equation of deterministic control theory. *Appl. Math. Optim.*, 11:161–181, 1984.

[7] E. Cristiani. *Fast Marching and Semi Lagrangian Methods for Hamilton Jacobi Equations with Applications.* PH. D. Thesis, 2006.

[8] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 4: Integral equations and numerical methods.* Springer-Verlag, Berlin, 1990.

[9] B. Després and F. Lagoutière. Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. *J.Sci. Comput.*, 16:479–524, 2001.

[10] F. Frankowska. Lower semi-continuous solutions of hamilton-jacobi-equations. *SIAM J.Control Optim.*, 31:257–272, 1993.

[11] F. Frankowska and R. B. Vinter. Existence of neighboring feasible trajectories: applications to dynamic programming for state constrained optimal control problems. *I.Optim.Theory Appl.*, 104:27–40, 2000.

[12] F. Lagoutière. *Modélisation mathématique et résolution numérique de problèmes de fluides compressibles à plusieurs constituants. Thèse de doctorat, Université Paris 6.* 2000.

[13] E.B. Lee and L. Markus. *Foundations of optimal control theory.* John Wiley, New York, 1967.

[14] J.D.L. Rowland and R.B. Vinter. Construction of optimal feedback controls. *Systems and control letters*, 16:357–367, 1991.

[15] J. Warga. Relaxed variational problems. *J. Mathematical Analysis and Applications*, 4:111–128, 1962.

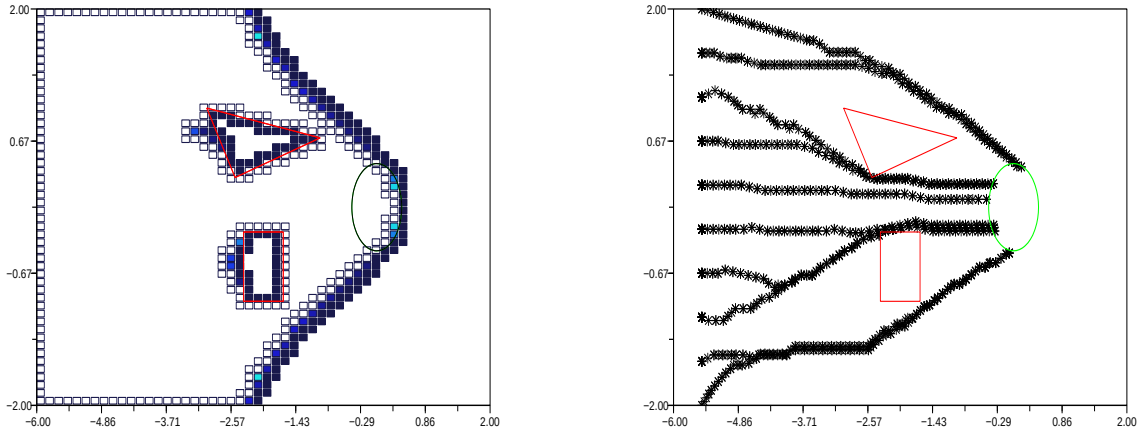
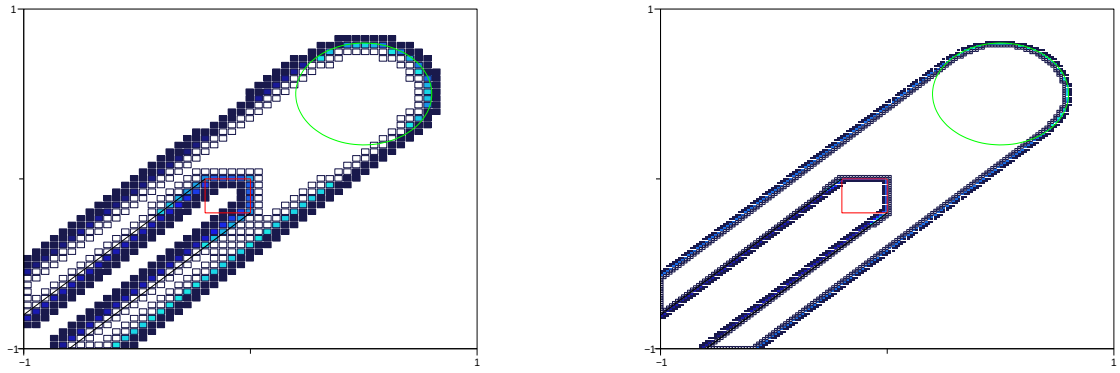
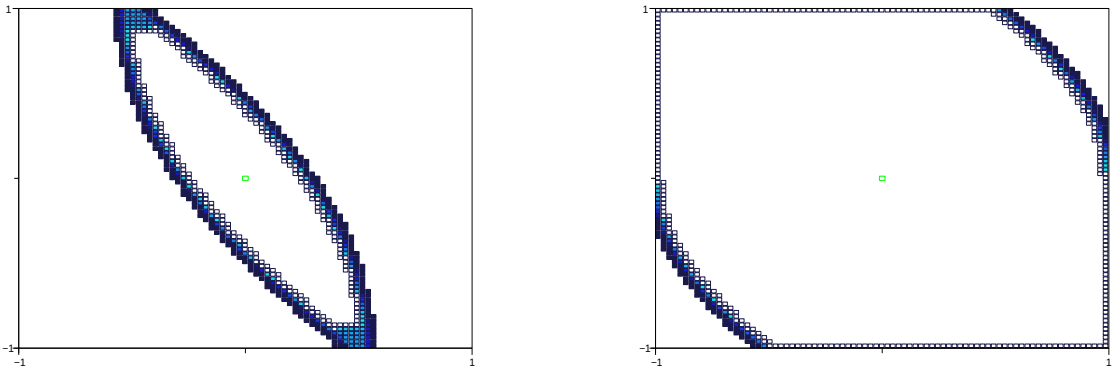


Figure 3.6: Zermelo problem: Capture basin and optimal trajectories reconstruction, $T = 8.28$, $P_{x_1} = P_{x_2} = 50$, CFL=0.9



$P_{x_1} = P_{x_2} = 51$ $(T = 1.38)$ $P_{x1} = P_{x2} = 151$ $(T = 1.37)$

Figure 3.7: Example 2 with constant dynamics, CFL=0.9



$T = 1.02,\ P_{x_1} = P_{x_2} = 81$ $T = 3.02,\ P_{x_1} = P_{x_2} = 81$

Figure 3.8: Thin target problem, CFL=0.9

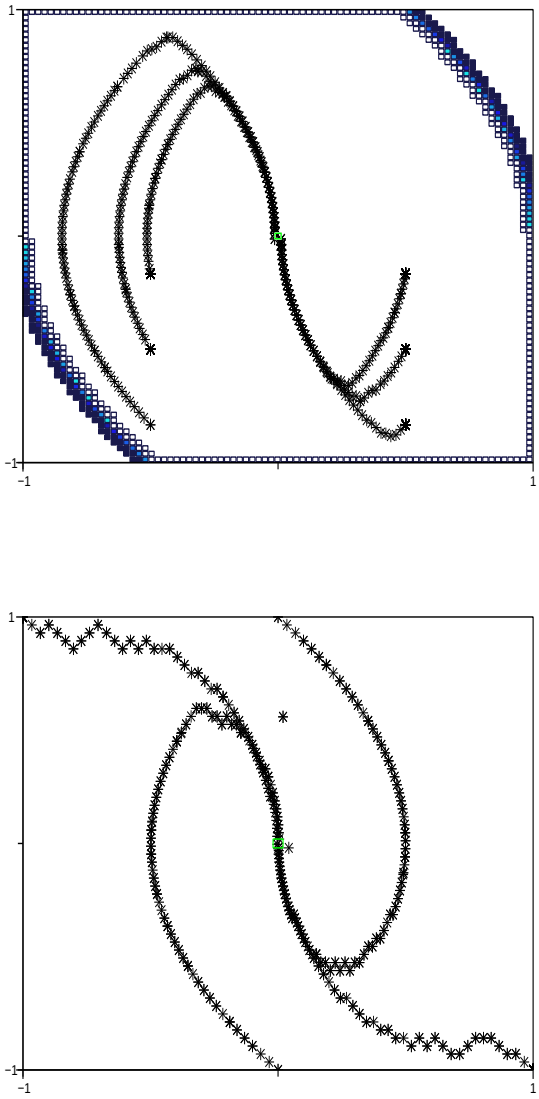
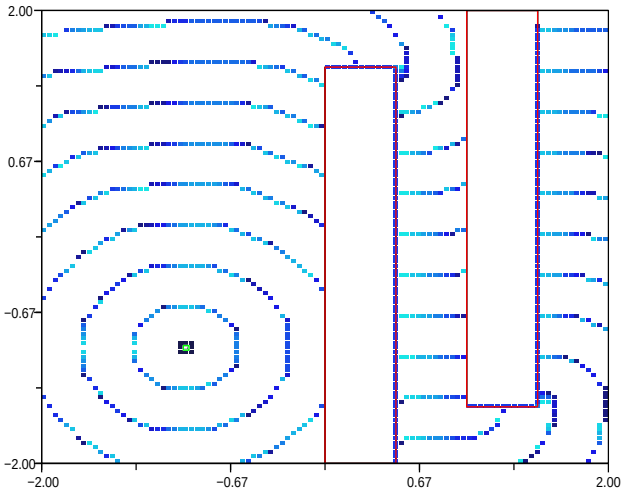
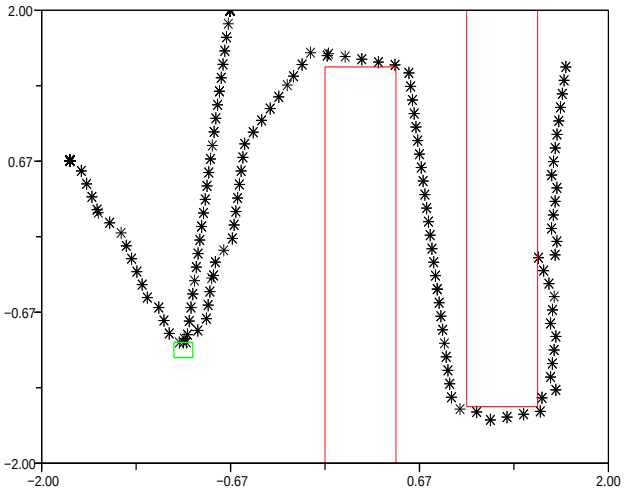


Figure 3.9: Thin target problem: some optimal trajectories, $T = 3.07$

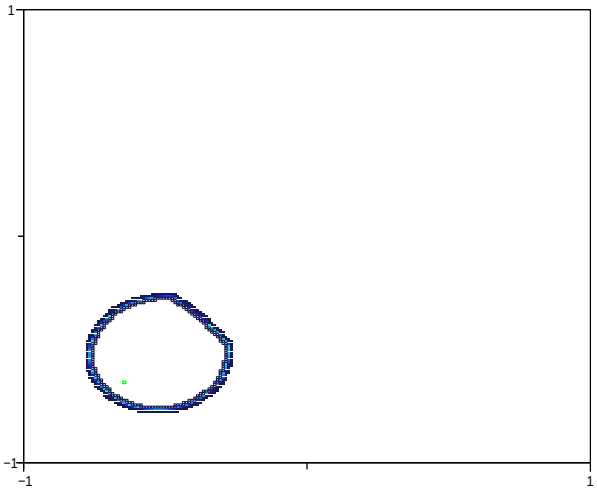


(a)

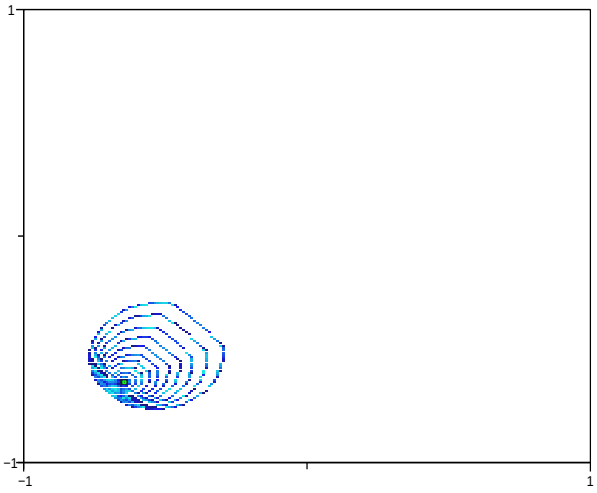


(b)

Figure 3.10: Two obstacles example: front positions every $10\Delta t$, $\Delta t = 0.036$, $P_{x_1} = P_{x_2} = 100$, $N_\alpha = 15$, cpu-time=201s (scilab)(a), some trajectories reconstruction $P_{x_1} = P_{x_2} = 30$(b) , CFL=0.9,

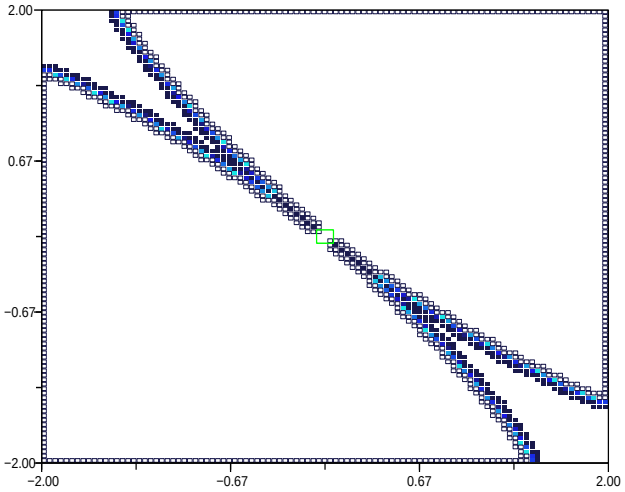


(a)

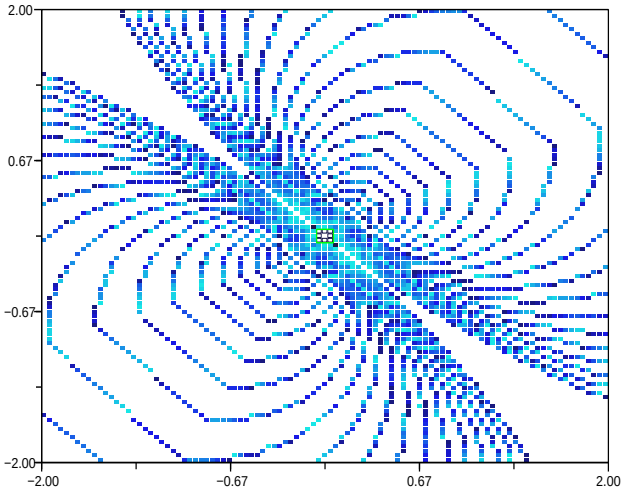


(b)

Figure 3.11: Poincaré model: maximal reachable set (a) and front position every $10\Delta t$, $\Delta t = 0.01$ (b), $T = 0.94s$ $P_{x_1} = P_{x_2} = P_{x3} = 200$, $N_\alpha = 151$, CFL=0.9, CPU time=97.8s



(a)



(b)

Figure 3.12: Eikonal example: maximal reachable set (a) and front position every $15\Delta t$, $\Delta t = 0.01$ (b), $T = 3.7s$ $P_{x1} = P_{x2} = P_{x3} = 101$, $N_\alpha = 150$, CFL=0.9, CPU time=221.97s

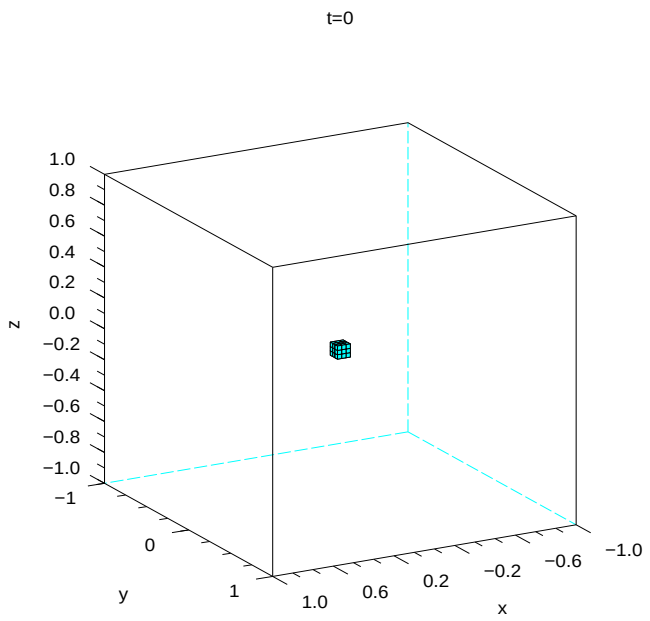


Figure 3.13: 3D example: The target

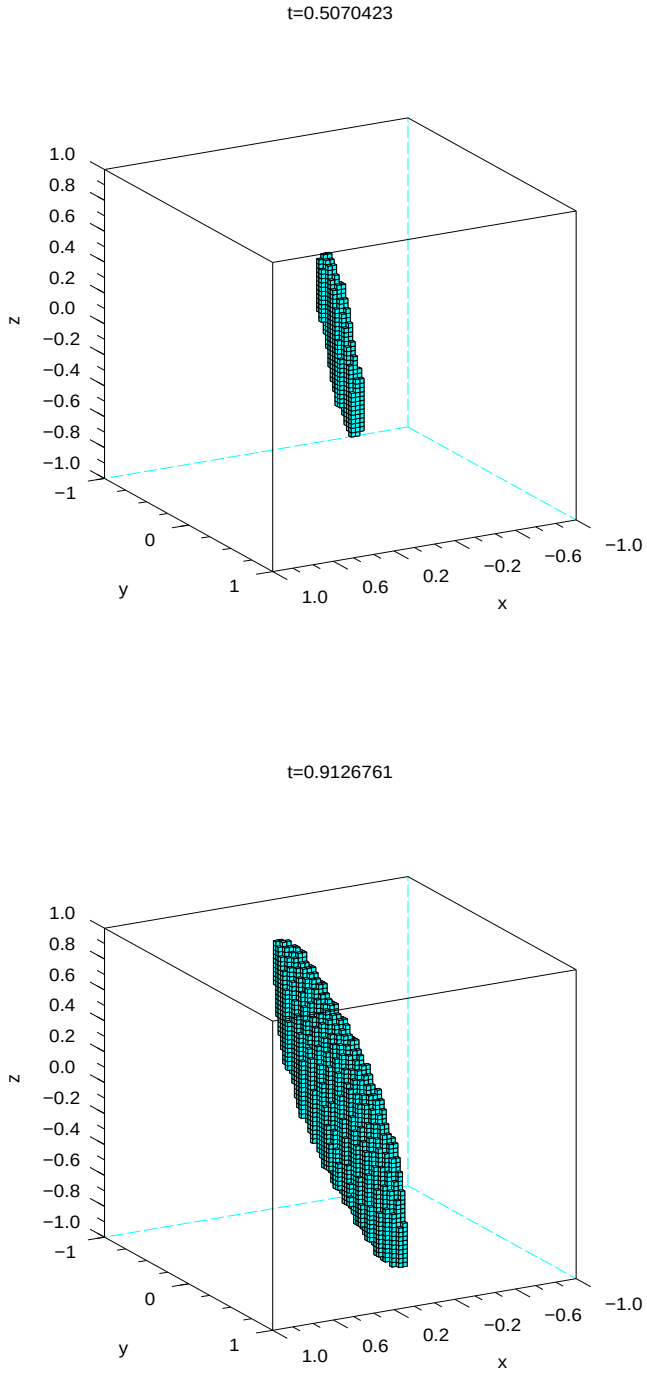


Figure 3.14: 3D example: Reachable set at $T = 0.5s$ and $T = 0.9s$, $P_{x1} = P_{x2} = P_{x3} = 71$, $N_\alpha = 3$, CFL=0.9

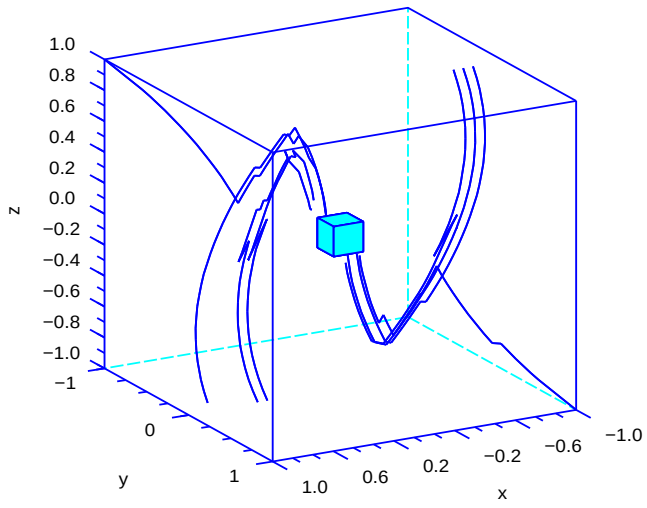


Figure 3.15: 3D example: Some optimal trajectories reconstruction, $P_{x1} = P_{x2} = P_{x3} = 31$, $N_\alpha = 3$, CFL=0.9

CHAPTER 4

Application to Atmospheric re-entry

4.1 Introduction

The atmospheric reentry problem is of considerable practical interest. It has already been studied by many authors and with different approaches. Early results are due to Bulirsh [12] and Dickmanns [14]. Since there, many works have dealt with different versions of the problem. In particular, the contribution of Betts [6, 3, 5, 7, 1, 2]is doubtless one of the most rich. His works dealt with the resolution of problems from aeronautics using different methods, especially total discretization and shooting methods. He was also interested in trajectory optimization [5].

Always in the context of trajectory optimization, Bonnard et al. aim in [10, 11] to stabilize the shuttle on the optimal trajectory that they construct using the maximum Pontryagin principle. The optimization problem that they consider consists in minimizing the total thermal flux.

In [13], Bonnans and his co-authors propose a trajectory optimization procedure based on an interior point approach. Some variants of the atmospheric reentry problem are handled as an application.

The thermal flux criterion is also considered in [9] where the authors study a total discretization of the control problem and the resolution of the resulting nonlinear programming system.

In [6], the considered criterion is the maximization of the final latitude, the final time being free.

We are interested here in minimizing the final time, i.e the time needed by the shuttle to reach a given target on the earth, with respect to some state constraints.

4.1.1 The complete 6D model

The atmospheric re-entry problem consists in optimizing the trajectory of a shuttle during the atmospheric phase. Hence, we aim to steer the vehicle from an initial position in space to a final target on the earth. During this phase, the engine speed is reduced by friction with the atmosphere and we take into account essentially three state constraints: a thermal constraint (as there are passengers inside the engine), a constraint on the normal acceleration (which is related to the flight comfort) and a constraint on the dynamic pressure (it is a technical constraint related to the structure of the engine). We suppose that the shuttle is a glider during the atmospheric arc, which means that no thrust is applied.

The modelization of the re-entry shuttle motion leads to a 6D model that we present in the sequel. We will use the same notations of [10, 11]:

State variables	
r:	the distance between the shuttle and the center of the earth
v:	modulus of the shuttle velocity
ℓ:	longitude (angle)
L:	latitude (angle)
γ:	path inclination of the shuttle (angle)
χ:	Azimuth (angle)

and

Control variables	
α:	angle of attack
μ:	bank angle

These variables describe totally the position of the shuttle (with r, ℓ, L) and its velocity (with v, γ, χ).

The physical problem: State equation

In order to modelize the problem, we first define a suitable frame to describe the state variables. Let O denote the center of the earth and G the mass center of the shuttle. $E = (e_1, e_2, e_3)$ is an inertial frame centered at O. The spherical coordinates of G are (r, ℓ, L), with r the distance to O, ℓ the longitude and L the latitude. The moving frame $R_1' = (e_r, e_\ell, e_L)$ is centered at G and defined such that e_r is the local vertical direction and (e_ℓ, e_L) is the local horizontal plane with e_L pointing to the north, see Figure 4.1 (a).

For conveniency, we parametrize the relative velocity by its modulus v and two angles: the path inclination γ which is the angle between v and the horizontal plane (e_ℓ, e_L), the azimuth χ which is the angle between the projection of v on (e_ℓ, e_L) and e_L, see Figure 4.1 (b). We also introduce the orthonormal frame (i, j, k) defined such that i has the same direction as the velocity v, j is in the plane (i, e_r) and satisfies $j.e_r > 0$ and $k = i \wedge j$.

The forces acting on the shuttle are the gravitational force and the aerodynamic force. This latest one consists of a Drag component opposite to the velocity sense:

$$F_D = -\frac{1}{2}\rho(r)SC_D(v, \alpha)v^2 i,$$

and of a Lift component perpendicular to the velocity:

$$F_L = \left(\frac{1}{2}\rho(r)SC_L(v, \alpha)v^2\right)(\cos\mu j + \sin\mu k),$$

S being a given constant called the reference surface, and C_D, C_L are called respectively drag and lift coefficients.

The vehicle is controlled by the angle of attack α and the bank angle μ. We recall that the angle of attack α is the one between the velocity and the axis of the shuttle. The bank angle is the one between the local vertical e_r and the vertical axis to the plan of the shuttle, see Figure 4.2.

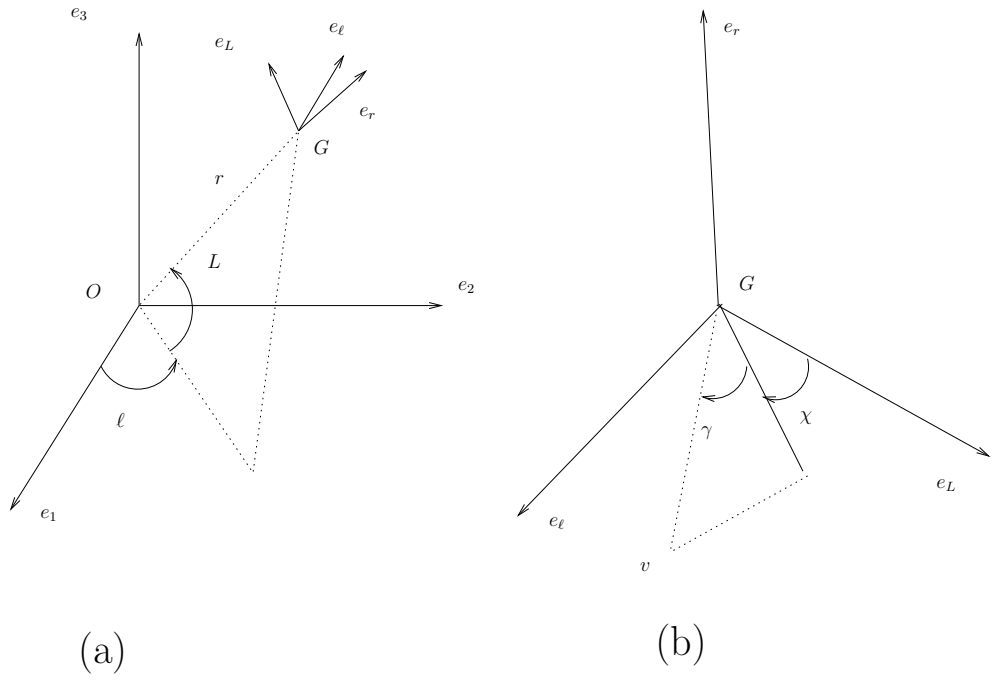


Figure 4.1: setting of the frame

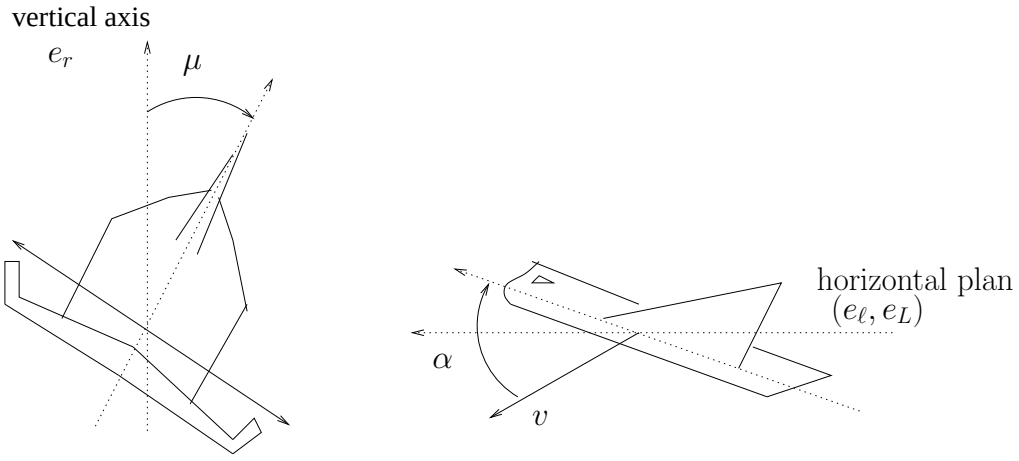


Figure 4.2: The two controls of the model

As we describe the system in a non inertial frame, the shuttle is also submitted to the Coriolis force (orthogonal to the motion of the shuttle and to the rotation axis of the earth) and to the centripetal force (which is orthogonal to the rotation axis of the earth and centripetal). These two forces are due to the rotation of the earth around its axis, they will be neglected during the atmospheric phase.
Then the motion of the shuttle is described for the six state variables by the following dynamic system:

$$\dot{r} = v \sin\gamma, \tag{4.1.1a}$$

$$\dot{v} = -g(r)\sin\gamma - \frac{S}{2}\frac{C_D(v,\alpha)}{m}\rho(r)v^2, \tag{4.1.1b}$$

$$\dot{\gamma} = \cos\gamma\left(-\frac{g(r)}{v} + \frac{v}{r}\right) + \frac{S}{2}\frac{C_L(v,\alpha)}{m}\rho(r)v\cos\mu, \tag{4.1.1c}$$

$$\dot{L} = \frac{v}{r}\cos\gamma\cos\chi, \tag{4.1.1d}$$

$$\dot{\ell} = \frac{v}{r}\frac{\cos\gamma\sin\chi}{\cos L}, \tag{4.1.1e}$$

$$\dot{\chi} = \frac{S}{2}\rho\frac{C_L(v,\alpha)}{m}\frac{v}{\cos\gamma}\sin\mu + \frac{v}{r}\cos\gamma\tan L\sin\chi, \tag{4.1.1f}$$

where C_D and C_L are respectively the Drag and Lift coefficients. They depend on the velocity v and the angle of attack α. They will be defined later.

For more details about the physical variables we refer to the book of Betts [6] and to the PH.D thesis of Laurent-Varin [15].

Spherical atmosphere

We consider a spherical model for the earth, the atmospheric density ρ is given by an exponential model:

$$\rho(r) = \rho_0 exp(-\frac{r - r_T}{h_s}). \tag{4.1.2}$$

On the other hand, during the atmospheric arc, r does not vary a lot. Hence, we may often assume that the potential g is constant during this phase. However, in the present work, we would rather consider that the dependence of g on r follows the following general rule:

$$g(r) := \frac{\mu_T}{r^2}. \tag{4.1.3}$$

Remark 4.1. *Notice that the longitude ℓ does not appear in the dynamics of the five other state variables. This is a natural consequence of the choice of a spherical potential model. Consequently, we can get rid of this variable [9, 6] or fix its value $\ell = 0$ [6]. Hence this simplification allows to handle only five state variables.*

The state constraints

As mentioned at the beginning of this section, the re-entry problem is subject to three state constraints:

- A constraint on the thermal flux:

$$Q := C_q\sqrt{\rho(r)}v^3 \leq Q_{\max}, \tag{4.1.4}$$

  where C_q and $Q_{\max}$ are positive constants that we define below.

- A constraint on the normal acceleration:

$$\gamma_n := \gamma_{n0}(\alpha)\rho v^2 \leq \gamma_n^{max}, \tag{4.1.5}$$

  where as in [9], $\gamma_{n0}(\alpha) := \dfrac{S}{2mg}(C_D(\alpha)\sin\alpha + C_L(\alpha)\cos\alpha)$ and $\gamma_n^{max} = 2.5$.

- A constraint on the dynamic pressure:

$$P := \frac{1}{2}\rho(r)v^2 \leq P^{max}. \tag{4.1.6}$$

Constraints on the control variables

It is also natural to consider that the control variables lie in a bounded subset. For instance, in [6], the controls α and μ satisfy:

$$\frac{-\pi}{2} \leq \alpha \leq \frac{\pi}{2}, \qquad \frac{-89}{90}\frac{\pi}{2} \leq \mu \leq \frac{1}{90}\frac{\pi}{2}. \tag{4.1.7}$$

Constants of the problem

We give in the sequel the constants of the problem as described in [10]:

$m = 7169.602\ (kg)$,	mass of the shuttle
$S = 15.05\ (m^2)$,	reference surface
$Q_{max} = 717300\ (W.m^{-2})$,	maximal thermal flux
$r_T = 6378139\ (m)$,	radius of the earth
$C_q = 1.705\ 10^{-4}\ (S.I)$,	
$\mu_T = 39.86\ 10^{13}\ (m^3.s^{-2})$,	gravitational constant of the earth
$\rho_0 = 1.225\ (kg.m^{-3})$,	
$h_s = 7143\ (m)$.	

4.1.2 A simplified 3D version

Notice that in the system (4.1.1), the three first state variables do not depend on the other variables. Hence we could treat, as a first step, a simplified 3D version of atmospheric re-entry problem where we handle only the state variables r, v and γ, and assume that the longitude, latitude and azimuth are fixed.

We also restrict the state constraints to the thermal one (4.1.4) on the energy of the shuttle. This constraint is the most important physically.

Let us precise that this simplified 3D version is the one proposed in [11].

The dynamics of the model is then reduced to:

$$\dot{r} = vsin\gamma, \tag{4.1.8a}$$

$$\dot{v} = -g(r)sin\gamma - \frac{S}{2m}C_D(v,\alpha)\rho(r)v^2, \tag{4.1.8b}$$

$$\dot{\gamma} = cos\gamma(-\frac{g(r)}{v} + \frac{v}{r}) + \frac{S}{2m}C_L(v,\alpha)\rho(r)v\cos\mu \tag{4.1.8c}$$

As the order of magnitude of the distance r and the velocity v has large variations, during the atmospheric arc, we rather consider the variables:

$$h := \rho(r); \quad \theta := \ln(v); \quad \gamma := \gamma.$$

Then the dynamics of the new variables is:

$$\dot{h} = -\frac{1}{h_s}he^{\theta}\ \sin\gamma, \tag{4.1.9a}$$

$$\dot{\theta} = -g(h)\sin\gamma e^{-\theta} - \frac{S}{2m}C_D(\theta,\alpha)he^{\theta}, \tag{4.1.9b}$$

$$\dot{\gamma} = -\cos\gamma\left(g(h)e^{-\theta} + \frac{e^{\theta}}{h_s\ln(h/\rho_0) + r_T}\right) + \frac{S}{2m}C_L(\theta,\alpha)he^{\theta}\cos\mu, \tag{4.1.9c}$$

with $g(h) := \dfrac{\mu_T}{(r_T - h_s\ln(h/\rho_0))^2}$.

The state constraint (4.1.4) becomes:

$$C_q\sqrt{h}\exp(3\theta) \leq Q_{\max}. \tag{4.1.10}$$

Finally, the control problem is the following

$$\boxed{\begin{array}{ll}\text{Minimize } T & \\ \text{subject to:} & T>0, \\ & (h,\theta,\gamma,\alpha,\mu) \text{ satisfies } (4.1.9)-(4.1.10), \\ & (\alpha(t),\mu(t))\in U \quad \text{a.e. } t\in(0,T), \\ & (h(t),\theta(t),\gamma(t))\in\mathcal{K} \quad \forall t\in(0,T), \\ & (h(T),\theta(T),\gamma(T))\in\mathcal{C}. \end{array}} \tag{4.1.11}$$

We consider the following domain for state and control variables:

$$\begin{aligned} U &= [0,\frac{40}{90}\frac{\pi}{2}]\times[0,\frac{\pi}{2}]; \\ \mathcal{K} &= [\rho(125.10^3+r_T),\rho(7.10^3+r_T)]\times[\ln(400),\ln(8000)]\times[-\pi/2,0]. \end{aligned}$$

In general, the flight path angle is constrained to take only negative values. This allows to avoid rebounds in the optimal trajectories [9].

The control set U requires some explanations compared to the bounds of (4.1.7). In fact the angle of bank μ could take positive or negative values, it does not matter as it intervenes as $\cos\mu$ in the dynamics (4.1.9).

The target that we consider for simulations in this frame is:

$$\mathcal{C} := \{(h_\mathcal{C},\theta_\mathcal{C},\gamma_\mathcal{C})\} = \{(\rho(15.10^3+r_T);\ln(445);-\pi/4)\}.$$

4.2 The HJB approach

4.2.1 The minimum time problem

The model (4.1.11) takes place in a general setting of time optimal control problems of the form:

$$(P_x)\quad \begin{cases} \mathcal{T}(x) := \text{Minimize } T, \\ \text{with: } \begin{cases} \dot{y}(t)=f(y(t),\alpha(t)) \quad t\in(0,T), \\ y(0)=x, \end{cases} \\ \qquad T\geq 0, \\ \qquad \alpha(t)\in\mathcal{A} \quad a.e.\ t\geq 0, \\ \qquad y(T)\in\mathcal{C}, \\ \qquad y(t)\in\mathcal{K} \quad \forall t\in[0,T], \end{cases}$$

where $\mathcal{C} \subset \mathcal{K} \subset \mathbb{R}^n$ (with $n \geq 1$) and $\mathcal{A} \subset \mathbb{R}^m$. The set of control values $\mathcal{A} \neq \emptyset$ is a compact of $\mathbb{R}^m$ with $m \geq 1$, $\mathcal{K}$ and $\mathcal{C}$ are non empty closed sets of $\mathbb{R}^n$. The set $\mathcal{C}$ is the target that we aim to reach in minimum time, and $\mathcal{K}$ is the domain in which the trajectories are allowed to move until they reach $\mathcal{C}$.

In order to solve problem (P_x) on $\mathbb{R}^n$ (i.e to determine the function $\mathcal{T}$ on $\mathbb{R}^n$), we will use the HJB approach which consists in proving that $\mathcal{T}$ is a solution of the nonlinear Hamilton Jacobi Bellman PDE on $]0, +\infty[\times\mathbb{R}^n$. Another difficulty in the calculation of $\mathcal{T}$ lies in the fact that this function is discontinuous. In fact, the minimal time is finite on the area of starting points from which there exists an optimal trajectory reaching the target, and $\mathcal{T}$ is infinite in the area of starting points not able to reach the target.

To avoid these difficulties, we suggest:

- First, we will reduce the calculation of the function $\mathcal{T}$ to the calculation of another function ϑ which takes only values 0 or 1, and which is also the value function of a control problem equivalent to (P_x). At every time $t \geq 0$, the function $\vartheta(t, \cdot)$ takes the value 0 on the area $\Omega_t \subset \mathcal{K}$ corresponding to initial conditions x being able to reach the target, before time t, following an admissible trajectory. And $\vartheta(t, \cdot)$ takes value 1 on $\mathbb{R}^n \setminus \Omega_t$.

- At every $t \geq 0$, the determination of $\vartheta(t, \cdot)$ corresponds to the determination of the front $\Gamma_t = \partial\Omega_t$. This evolution (also described by an HJB equation) requires only local calculations. This simple idea combined with an anti-diffusive numerical scheme makes it possible to design a fast algorithm as the ones proposed in chapters 2 and 3.

We assume the following classical assumptions on the dynamics f:

$$(H1) \qquad \begin{array}{ll} (a) & f \text{ is continuous on } \mathbb{R}^n \times \mathbb{R}^m. \\ (b) & |f(y,\alpha)| \leq k_1(1+|y|),\ \forall (y,\alpha) \in \mathbb{R}^n \times \mathcal{A}. \\ (c) & |f(y_1,\alpha) - f(y_2,\alpha)| \leq k_2|y_1 - y_2|,\ \forall y_1, y_2 \in \mathbb{R}^n,\ \alpha \in \mathcal{A}. \\ (d) & \{f(y,\alpha),\ \alpha \in \mathcal{A}\} \neq \emptyset \text{ is a convex set of } \mathbb{R}^n,\ \forall y \in \mathbb{R}^n. \end{array}$$

Some properties of the function $\mathcal{T}$.

For every $x \in \mathcal{K}$, the value $\mathcal{T}(x)$ is finite if and only if there exists $T \in \mathbb{R}^+$ and a trajectory that starts at x and reaches $\mathcal{C}$ at time T without leaving $\mathcal{K}$ on $[0, T]$, this trajectory may leave $\mathcal{K}$ after time T.

On the other hand, for $x \in \mathbb{R}^n$, we can notice that if x is in $\mathcal{K}^c := \mathbb{R}^n \setminus \mathcal{K}$, then there is no admissible trajectory for (P_x) starting at x and staying in $\mathcal{K}$ until reaching $\mathcal{C}$, so $\mathcal{T}(x) = +\infty$.

In the same way, if x is in $\mathcal{C}$, then all trajectories starting at x reach the target at $T = 0$, and $\mathcal{T}(x) = 0$.

The only case that deserves to be studied is the remaining one: $x \in \mathcal{K} \setminus \mathcal{C}$. If there exists an admissible[1] trajectory of (P_x) that reaches $\mathcal{C}$ in finite time, then we can deduce that

[1] we mean by admissible trajectory a trajectory y that satisfies the state constraint i.e. $y(t) \in \mathcal{K}$ for all $t \in [0, \mathcal{T}(x)]$.

$\mathcal{T}(x) < \infty$. Otherwise if all trajectories leave $\mathcal{K}$ before reaching $\mathcal{C}$, or keep moving inside $\mathcal{K}$ without reaching $\mathcal{C}$, then $\mathcal{T}(x) = +\infty$. Moreover, when a trajectory leaves $\mathcal{K}$ before reaching $\mathcal{C}$, it is no more interesting to follow its evolution as we know that either $\mathcal{T}(x) = +\infty$ or there exists another admissible trajectory that reaches $\mathcal{C}$ in finite time and we would rather follow this latter trajectory. Notice also that, when a trajectory reaches $\mathcal{C}$ while staying in $\mathcal{K}$, its motion after this time does not change $\mathcal{T}(x)$ even if it leaves $\mathcal{K}$.

4.2.2 The link with a Rendez-Vous problem

For $x \in \mathbb{R}^n$ and $T > 0$, let us introduce the finite horizon Rendez-Vous problem $(\mathcal{P}_{T,x})$:

$$(\mathcal{P}_{T,x}) \quad \left\{ \begin{array}{l} \text{Minimize } \varphi(y(T)), \\ \text{with: } \left\{ \begin{array}{l} \dot{y}(t) = \lambda(t).f(y(t),\alpha(t)) \quad \text{a.e } t \in (0,T), \\ y(0) = x, \end{array} \right. \\ (\alpha(t),\lambda(t)) \in \mathcal{A} \times \Lambda(y(t)) \quad \text{a.e } t \in [0,T], \\ y(t) \in \mathcal{K} \quad \forall t \in [0,T], \end{array} \right.$$

where the final cost function φ is given by,

$$\varphi(y) := \left\{ \begin{array}{l} 0 \text{ if } y \in \mathcal{C}, \\ 1 \text{ otherwise,} \end{array} \right.$$

and the set-valued map Λ is defined on $\mathbb{R}^n$ by:

$$\Lambda(y) := \left\{ \begin{array}{ll} \{1\} & \text{if } y \in \mathbb{R}^n \setminus \mathcal{C}, \\ [0,1] & \text{if } y \in \mathcal{C}. \end{array} \right. \tag{4.2.1}$$

Let $\vartheta : [0,+\infty[\times\mathbb{R}^n \longmapsto \mathbb{R}$ denote the value function of the control problem $(\mathcal{P}_{T,x})$:

$$\vartheta(T,x) := \inf(\mathcal{P}_{T,x}).$$

It is obvious that ϑ takes only values 0 and 1. Moreover this function is linked to the function $\mathcal{T}$ and problem (P_x), as shown in the following theorem.

Theorem 4.2. *For every $x \in \mathbb{R}^n$, we have:*
i) $\mathcal{T}(x) = \inf\{T,\ \vartheta(T,x) = 0\}$.
ii) The function $T \mapsto \vartheta(T,x)$ is decreasing for $T \geq 0$.
iii) $\mathcal{T}(x) = +\infty \Longleftrightarrow \vartheta(T,x) = 1 \quad \forall T \geq 0$.

Proof. the proof of the above theorem is straightforward, it uses the same arguments of [8]. □

From the above theorem, it appears that a good way to determine, for a given initial position $x \in \mathcal{K}$, the minimum time $\mathcal{T}(x)$ to reach $\mathcal{C}$, is to define $\vartheta(T,x)$ for increasing $T \geq 0$.

As long as $\vartheta(T,x)=1$ the target is not yet reached by any admissible trajectory of $(\mathcal{P}_{T,x})$ and we can deduce that $\mathcal{T}(x)>T$.

When we get $\vartheta(T,x)=0$ for the first time $T\geq 0$, then we know that the horizon T is sufficient to find a trajectory that starts at x and reaches $\mathcal{C}$ while keeping in $\mathcal{K}$ on $[0,T]$. Consequently we deduce that $\mathcal{T}(x)=T$.

It is known that the value function ϑ satisfies the *Dynamic Programming Principle* or Bellman's principle [4],

$$\vartheta(T,x)=\inf\Big\{\begin{array}{l}\vartheta(T-\tau,y(\tau)),\\ y(0)=x,\\ \dot{y}(t)=\lambda(t)f(y(t),\alpha(t)),\ (\alpha(t),\lambda(t))\in\mathcal{A}\times\Lambda(y(t)),\ a.a.\ t\in[0,\tau]\\ y(t)\in\mathcal{K}\ a.a.\ t\in[0,\tau]\Big\}.\end{array}\tag{4.2.2}$$

From this principle, we prove that the value function ϑ is a *l.s.c. viscosity solution* of the HJB equation as shown in chapter 5:

$$\min\Big(\vartheta_t(t,x)+\mathcal{H}(\vartheta)(t,x)\ ;\ \vartheta(t,x)-\chi_{\mathcal{K}}(x)\Big)=0,\ t\in]0,T[,\ x\in\mathbb{R}^n,\tag{4.2.3a}$$

$$\vartheta(0,x)=\varphi(x),\ x\in\mathbb{R}^n,\tag{4.2.3b}$$

where the Hamiltonian

$$\mathcal{H}(\vartheta)(t,x):=\max_{\substack{\alpha\in\mathcal{A},\\ \lambda\in\Lambda(x)}}\{-\lambda f(x,\alpha).\vartheta_x(t,x)\}$$

and $\chi_{\mathcal{K}}(x)=0$ if $x\in\mathcal{K}$ and 1 otherwise.

In virtue of theorem 4.2, we will be only interested in the value function ϑ. In the sequel, we deal with the discretization of equation (4.2.3) (eventhough we do not have any uniqueness result) using the fast sparse method of chapter 3. We show in particular the reachable set for the 3D model of atmospheric reentry.

4.3 Numerical simulations in 3D

Let us first precise that all the variables and constants are expressed in the International Units System (S.I). In particular the angles are expressed in radian.

We deal here with the atmospheric reentry model in 3D. We aim to define the initial positions from which the shuttle is able to reach a given target on the earth without violating the thermal constraint.

The target in the tests below is thin defined by

$$\mathcal{C}:=\{(\rho(15.10^3+r_T);ln(445);-\frac{\pi}{4})\},$$

and is approached numerically by the cell containing it.
The dynamics is defined by (4.1.9) and the constraint by (4.1.10).

In order to obtain well balanced figures, we scaled the θ variable by 0.1 and the γ variable by 0.3.

4.3.1 With one control

We suppose here as in [11] that the only control is the bank angle μ during the atmospheric arc. In fact the angle of attack is expressed as a function of v in this model. Hence, coefficients C_D and C_L depend only on the velocity v. In the model of [11], the dynamic coefficients C_D and C_L follow piece wise affine rules that we transpose in our variables system:

$$C_D(\theta) = \begin{cases} 0.585 & \text{if } \exp\theta > 3000, \\ 1.7\ 10^{-4}\exp\theta + 0.075 & \text{if } 1000 < \exp\theta \leq 3000, \\ 0.245 & \text{if } \exp\theta \leq 1000. \end{cases} \tag{4.3.1a}$$

$$C_L(\theta) = \begin{cases} 0.55 & \text{if } \exp\theta > 3000, \\ 1.256\ 10^{-4}\exp\theta + 0.1732 & \text{if } \exp\theta \leq 3000. \end{cases} \tag{4.3.1b}$$

The domain of calculations

$$\mathcal{D} := [3.07\ 10^{-8}; 0.46] \times [5.9; 9] \times [-\frac{\pi}{2}, 0],$$

is descretized into 50*50*50 cells. The control μ takes values in

$$U := [0; \frac{\pi}{2}],$$

discretized into 9 controls

$$U \sim \{\frac{i}{9}\frac{\pi}{2},\ i = 0, \cdots, 8\}.$$

The target and the obstacle, as well as the reachable set at $T = 90s$ are exhibited in Figure 4.4.

4.3.2 With the two controls

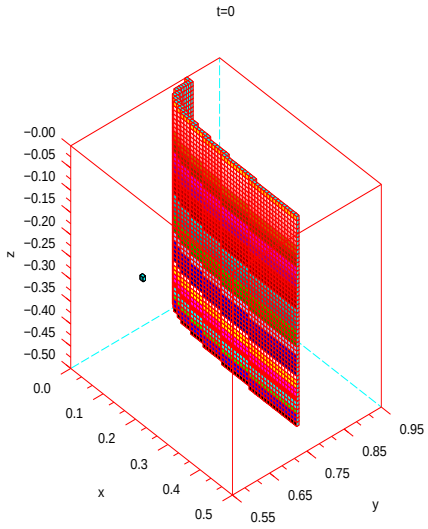
Now, the two controls α and μ are considered. We assume as in [6, 9] that the coefficients C_D and C_L depend only on the angle of attack α (in radian),

$$C_L(\alpha) = a_0 + a_1\frac{180\alpha}{\pi}, \tag{4.3.2a}$$

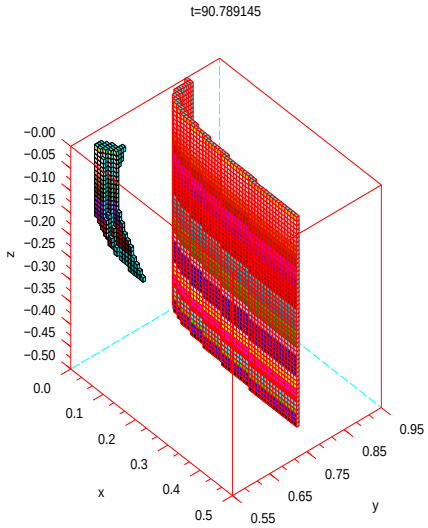
$$C_D(\alpha) = b_0 + b_1\frac{180\alpha}{\pi} + b_2(\frac{180\alpha}{\pi})^2. \tag{4.3.2b}$$

In the model exposed in [15], the constants $a_0,\ b_0,\ a_1,\ b_1,\ b_2$ take the values:

$$a_0 = -0.05;\ a_1 = 1.67\ 10^{-2};\ b_0 = 0.05;\ b_1 = 1.67\ 10^{-4};\ b_2 = 3.67\ 10^{-4}.$$



Target and obstacle



Reachable set $T = 90s$

Figure 4.3: Atmospheric re-entry with one control

We show in Figures 4.5 and 4.6 the evolution of the reachable set at different steps of time. We also precise the target and the thermal obstacle in Figure 4.4. The maximal reachable

set in Figure 4.6 is obtained using the stopping test already used in chapter 3,

$$||V^n - V^{n-1}||_{L^1} \leq tol,$$

with the tolerance $tol = 10^{-8}$.

The domain

$$\mathcal{D} := [3.07\ 10^{-8},\ 0.46] \times [5.9,\ 9] \times [-\frac{\pi}{2},\ 0],$$

is discretized into $P_\rho = 30 \times P_\theta = 30 \times P_\gamma = 10$ cells. The controls α and μ are both discretized into 5 controls. Hence the control set

$$U := [0; \frac{40}{90}\frac{\pi}{2}] \times [0, \frac{\pi}{2}].$$

is discretized into

$$\alpha \in \{\frac{i}{4}\frac{40}{90}\frac{\pi}{2},\ i = 0 \cdots 4\} \text{ and } \mu \in \{\frac{i}{4}\frac{\pi}{2},\ i = 0 \cdots 4\}.$$

This coarse discretization is done inorder to have a first idea about the maximal reachable set.

To give an idea about the duration of calculations, we present the CPU times in Table 4.1.

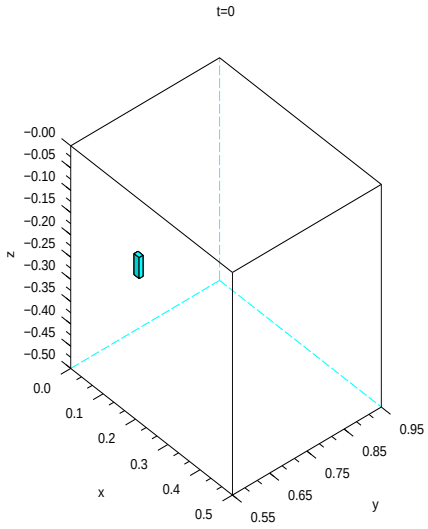
T (s)	CPU time (s)	# step of time
0	15.74	0
26.04	10617.58	1000
52.08	35870.86	2000
78.11	66166.45	3000
104.15	96019.64	4000
130.19	125219.24	5000
145.81	142698.10	5600

Table 4.1: CPU-time, Scilab code

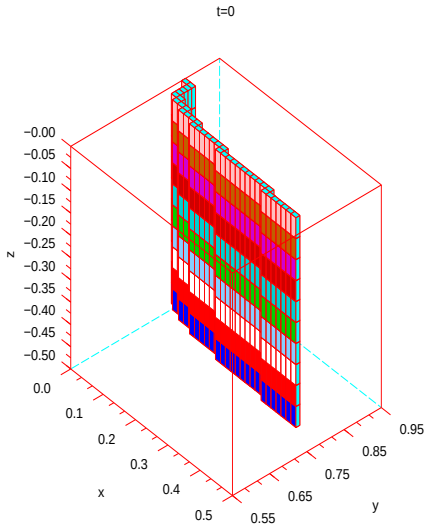
These calculations have been run with Scilab 4 on a computer equipped with AMD 2GHz processor, 4058 MB RAM.

Bibliography

[1] T.P. Bauer, K.P. Zondervan, J.T. Betts, and W.P. Huffman. Solving the optimal control problem using a nonlinear programming technique. part 1: General formulation. *American Institute of aeronautics and astronautics. Astrodynamics conference, Seattle.*, 1984.



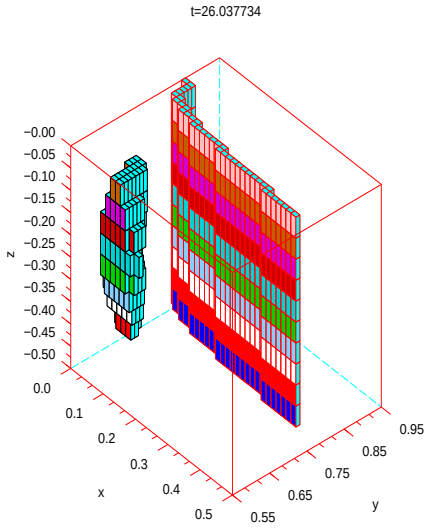
The target



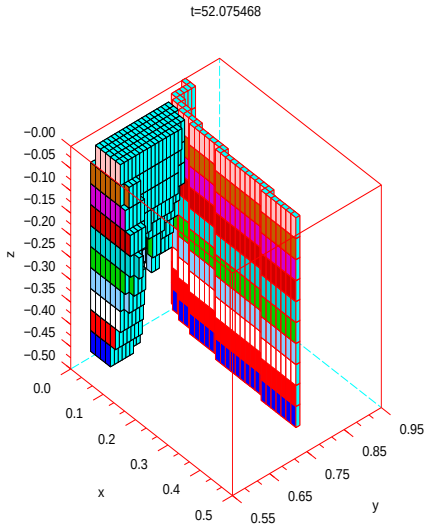
The obstacle

Figure 4.4: The target and the thermal obstacle

[2] T.P. Bauer, K.P. Zondervan, J.T. Betts, and W.P. Huffman. Solving the optimal control problem using a nonlinear programming technique. part 2: Optimal shuttle ascent tra-



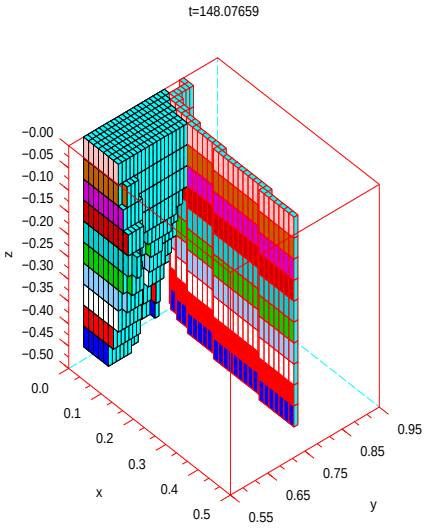
$T = 26s$



$T = 52s$

Figure 4.5: The reachable set after 1000 and 2000 step of time

jectories. *American Institute of aeronautics and astronautics. Astrodynamics conference, Seattle.*, 1984.



$T = 148s$

Figure 4.6: Maximal reachable set (after 5687 step of time)

[3] T.P. Bauer, K.P. Zondervan, J.T. Betts, and W.P. Huffman. Solving the optimal control problem using a nonlinear programming technique. part 3: Optimal shuttle reentry trajectories. *Proceedings of the AIAA/AAS Astrodynamics conference, Seattle.*, 1984.

[4] R. Bellman. *Dynamic programming*. Princeton university press, 1961.

[5] J.T. Betts. Survey of numerical methods for trajectory optimization. *Journal of Guidance Control and Dynamics*, 21(2):193–207, 1998.

[6] J.T. Betts. *Practical methods for optimal control using nonlinear programming*. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[7] J.T. Betts, S.K. Eldersveld, P.D. Frank, and J.G. Lewis. An interior point algorithm for large scale optimization. in large scale pde-constrained optimization. *Lect. notes Comput. Sci. Eng.*, 30:184–198, 2003.

[8] O. Bokanowski, S. Martin, R. Munos, and H. Zidani. An anti-diffusive scheme for viability problems. *Applied Numerical Mathematics*, 56:1147–1162, 2006.

[9] J.F. Bonnans and G. Launay. Large scale direct optimal control applied to the re-entry problem. *Journal of Guidance Control and Dynamics*, 21(6):996–1000, 1998.

[10] B. Bonnard, L. Faubourg, and E. Trélat. Optimal control of the atmospheric arc of a space shuttle and numerical simulations with multiple shooting method. *Mathematical Models and Methods in Applied Sciences*, 15(1):109–140, 2005.

[11] B. Bonnard and E. Trélat. Une approche géométrique du contrôle optimal de l'arc atmosphérique de la navette spatiale. *ESAIM: Control,Optimization and Calculus of Variations*, 7:179–222, 2002.

[12] R. Bulirsch. Die mehrzielmethode zur numerischen losung von nichtlinearen randwertproblemen und aufgaben der optimalen steuerung. *Proceedings of the AIAA/AAS Astrodynamics conference, Seattle.*, 1971.

[13] N. Bérend, J.F. Bonnans, J. Laurent-Varin, M. Haddou, and C. Talbot. An interior point approach to trajectory optimization. *INRIA report n 5613*, 2005.

[14] E.D. Dickmanns. Maximum range three-dimensional lifting planetary entry. *Technical Report TR R-387, National Aeronautics and space administration*, 1972.

[15] J. Laurent-Varin. *Calcul de trajectoires optimales de lanceurs spatiaux réutilisables par une méthode de point intérieur*. Thèse de doctorat de l'Ecole Polytechnique, 2005.

CHAPTER 5

Rendez-vous problem with state constraints

5.1 Introduction

We deal in this chapter with the Rendez-Vous (RDV) control problem subject to state constraints. This problem has already been studied by Frankowska and her co-authors [9, 10, 11] as well as by Ishii and Koike [12] and Soner [13]. A similar study has been carried out by Barles and Perthame [5, 6, 4] on the akin exit time and stopping time problems.
In all these works. the value function of the control problem is characterized, under some controllability hypotheses, as the unique viscosity solution of the Hamilton Jacobi Bellman equation (a suitable sense of solution is proposed in each case [9, 10, 11, 13, 12]).

Our aim here is to give an extension to the result of [11] (involving mixed inward-outward boundary conditions) that allows to treat particular transport problems. We also investigate the impact of some qualification constraints on the trajectories of the problem.

In the last section, we propose a characterization result in terms of epiderivatives, which involves no controllability assumptions.

5.2 The control problem

We deal with the finite horizon Mayer problem $(\widetilde{\mathcal{P}}_{T,x})$.

$$\widetilde{\mathcal{P}}_{T,x} \qquad \begin{cases} \text{Minimize } \varphi(y_x(T)),\ \alpha \in A, \\ \dot{y}_x(t) = f(y_x(t), \alpha(t)),\ a.a.\ t \in [0,T], \\ y_x(0) = x, \\ y_x(t) \in \mathcal{K}\ \forall t \in [0,T], \end{cases}$$

with the final cost function φ supposed lower semi continuous, and satisfying:

$$(H0) \qquad \begin{array}{l} \varphi : \mathbb{R}^n \to [0,1] \text{ continuous on } \mathcal{C}, \\ \varphi(x) = 1 \text{ if } x \in \mathbb{R}^n \backslash \mathcal{C} \quad \text{and } \varphi(x) \in [0,1[\text{ if } x \in \mathcal{C}. \end{array}$$

The set $\mathcal{C}$ is the target that we aim to reach at time T and $\mathcal{K}$ is the domain in which the trajectories are allowed to move until they reach $\mathcal{C}$. The set of controls is

$$A := L^\infty(\mathbb{R}^+, \mathcal{A}).$$

Remark 5.1. *i) This study is not restricted to the Mayer problem. In fact, as usual in control theory, it covers also the general Bolza problem, for more details see [3].*
ii) The function φ is in particular bounded here. When the terminal cost φ is unbounded we can modify the problem such that we recover the setting of $\widetilde{\mathcal{P}}_{T,x}$ without changing the optimal solutions.

Let

$$\widetilde{\vartheta}(T,x) := \inf(\widetilde{\mathcal{P}}_{T,x}),$$

denote the value function of the Mayer problem above.

We will consider in the sequel the following classical assumptions in control theory on the sets $\mathcal{K}$ and $\mathcal{C}$, and on the dynamics f:

$(H1)$
- $(a1)$ $\mathcal{K}$, $\mathcal{C} \neq \emptyset$ are closed bounded sets of $\mathbb{R}^n$, $\mathcal{C} \subset \overset{\circ}{\mathcal{K}}$, $\overset{\circ}{\mathcal{C}} \neq \emptyset$ and $\overline{\overset{\circ}{\mathcal{K}}} = \mathcal{K}$.
- $(b1)$ $\mathcal{A}$ is a compact convex set of $\mathbb{R}^n$.

$(H2)$
- $(a2)$ f is continuous.
- $(b2)$ $|f(y,\alpha)| \leq k_1(1+|y|),\ \forall (y,\alpha) \in \mathbb{R}^n \times \mathcal{A}$.
- $(c2)$ $|f(y_1,\alpha) - f(y_2,\alpha)| \leq k_2|y_1 - y_2|,\ \forall y_1, y_2 \in \mathbb{R}^n,\ \alpha \in \mathcal{A}$.
- $(d2)$ $f(y,\mathcal{A}) := \{f(y,\alpha),\ \alpha \in \mathcal{A}\} \neq \emptyset$ is a convex set of $\mathbb{R}^n,\ \forall y \in \mathbb{R}^n$.

We assume in particular that $\mathcal{K}$ is given by:

$$(H3) \qquad \mathcal{K} := \cap_{j=1,..,r}\{x,\ h_j(x) \leq 0\},$$

where the $h_j : \mathbb{R}^n \to \mathbb{R}$ are C^1-regular functions with locally Lipschitz continuous gradients. We will denote for $x \in \partial\mathcal{K}$ the index active set by

$$I(x) := \{j = 1, \cdots, r,\ h_j(x) = 0\}.$$

Remark 5.2. *i) The set of controls $\mathcal{A}$ being compact and convex, hypothesis (H2)(d2) is satisfied in particular if there exist two functions g_1 and g_2 such that $f(y,\alpha) = g_1(y)\alpha + g_2(y)$ for all y, α, i.e. if the dynamics is affine w.r.t. the control.*
ii) As the set $\mathcal{A}$ is compact and f is continuous, $f(y,\mathcal{A})$ is a compact set for all y in $\mathbb{R}^n$.

Mixed qualification constraints We are looking here for some uniqueness result that extends the work of Frankowska et al. [11, 10], in particular these works deal with the Outward Pointing (OP) qualification constraint

$$\forall x \in \partial\mathcal{K},\ \exists \alpha \in \mathcal{A},\ \nabla h_j(x) \cdot f(x,\alpha) > 0,\ \forall j \in I(x), \qquad (OP)$$

Let us recall that (OP) is used [11, Lemma 4.1] in order to approach an admissible trajectory

$$y_x(t) \in \mathcal{K}\ \forall t \in [0,T],$$

by a sequence of interior admissible trajectories

$$y^k_{x_k} \to y_x \ \text{ and } \ y^k_{x_k}(t) \in \overset{\circ}{\mathcal{K}}\ \forall t \in [0,T].$$

On the other hand, another qualification constraint, called Inward Pointing (IP) has been introduced by Soner [13]

$$\forall x \in \partial\mathcal{K},\ \exists \alpha \in \mathcal{A},\ \nabla h_j(x) \cdot f(x,\alpha) \cdot < 0,\ \forall j \in I(x) \qquad (IP)$$

to show continuity and uniqueness of the viscosity solution of the HJB equation. Another type of inward pointing hypothesis is also used by Barles and Perthame [6]:

$$\forall x \in \partial\Omega, \quad \text{if } \exists \alpha \in \mathcal{A},\ f(x,\alpha)\cdot\eta(x) \leq 0, \quad \text{then } \exists \alpha' \in \mathcal{A},\ f(x,\alpha')\cdot\eta(x) < 0, \tag{5.2.1}$$

(with $\eta(x)$ the outward normal to Ω at x, and Ω regular) in the context of exit time problems (from an open bounded subset Ω) to prove a comparison principle and deduce the continuity of the exit time value function.

Our idea here is to compare the impact of (IP) and (OP) on the trajectories of the problem, and to propose a uniqueness result with mixed boundary controllability conditions. This result is an immediate consequence of [8] and [11].

Let $\mathcal{U} : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be an extended function, then the (Fréchet) subdifferential of $\mathcal{U}$ at $x \in Dom\,\mathcal{U}$ is defined by

$$\partial_-\mathcal{U}(x) = \{p \in \mathbb{R}^n,\ \liminf_{y\to x} \frac{\mathcal{U}(y) - \mathcal{U}(x) - < p, y-x >}{||y-x||} \geq 0\}.$$

Theorem 5.3. *Assume (H0)-(H3) and let the dynamics satisfy the following qualification constraint:*

$$\text{For all } x \in \partial\mathcal{K}, \quad \begin{array}{ll} \exists \alpha \in \mathcal{A} \ \ s.t.\ \nabla h_j(x)\cdot f(x,\alpha) > 0,\ \forall j \in I(x), & (OP) \\ or\ \ \forall \alpha \in \mathcal{A}\ \big[\nabla h_j(x)\cdot f(x,\alpha)\cdot\eta(x) < 0,\ \forall j \in I(x)\big]\ \ or\ \ \big[f(x,\alpha) = 0\big]. & (SIP) \end{array}$$

Then $\widetilde{\vartheta}$ is the unique l.s.c. function with values in $[0,1]$ and such that $\widetilde{\vartheta}(t,x) = 1$ for all $t \geq 0,\ x \in \mathcal{K}^c$, satisfying

i) For all $(t,x) \in]0,+\infty[\times \overset{\circ}{\mathcal{K}})$, and all $(p_t,p_x) \in \partial_-\widetilde{\vartheta}(t,x)$,

$$p_t + \max_{\alpha\in\mathcal{A}}\{-f(x,\alpha)\cdot p_x\} = 0.$$

ii) For all $(t,x) \in]0,+\infty[\times\partial\mathcal{K})$, and all $(p_t,p_x) \in \partial_-\widetilde{\vartheta}(t,x)$,

$$p_t + \max_{\alpha\in\mathcal{A}}\{-f(x,\alpha)\cdot p_x\} \geq 0.$$

iii) For all $x \in \mathcal{K}$,

$$\liminf_{\substack{t'\to 0+ \\ x'\to x,\ x'\in\overset{\circ}{\mathcal{K}}}} \widetilde{\vartheta}(t',x') = \widetilde{\vartheta}(0,x) = \varphi(x).$$

The following Lemma will be usefull for the proof

Lemma 5.4. *Under the assumptions of Theorem 5.3, let $T > 0$ and $z \in \partial\mathcal{K}$ be such that (SIP) is satisfied. If there exists an admissible trajectory $y_x(t),\ t \in [0,T]$ such that $y_x(T) = z$ then*

$$y_x(t) \equiv z,\ \forall t \in [0,T].$$

Proof. As y_x is admissible then for all $j \in I(z)$ we have

$$0 \geq h_j(y_x(T-\tau)) - h_j(z) = \nabla h_j(z)(y_x(T-\tau) - z) + o(\tau)$$

for $\tau > 0$ in a neighborhood of 0. We divide by τ and let $\tau \to 0$ we obtain $0 \geq \nabla h_j(z) \cdot (-f(z,\alpha))$ with $\alpha \in \mathcal{A}$. Then either $0 > \nabla h_j(z) \cdot (-f(z,\alpha))$ which contradicts (SIP), or $\nabla h_j(z) \cdot f(z,\alpha) = 0$ which implies $f(z,\alpha) = 0$ by (SIP). And then y_x is stationnary. □

Proof of Theorem 5.3. The proof of this theorem is a slight modification of the one proposed in [11, Theorem 2.1]. We will only sketch the steps where the qualification constraint intervenes.

Using classical arguments, we can prove that the value function $\widetilde{\vartheta}$ is l.s.c. and satisfies the assertions (i)-(iii) of the theorem.

Now let V be a l.s.c. function with values in $[0,1]$ and $V(T,x) = 1$ for all $T \geq 0,\ x \in \mathcal{K}^c$, satisfying (i)-(iii). It is straightforward that $V(T,x) = \widetilde{\vartheta}(T,x)$ for all $T \geq 0, x \in \mathcal{K}^c$. We prove in two steps that $V(T,x) = \widetilde{\vartheta}(T,x)$ for $T > 0$ and $x \in \mathcal{K}$.

Step 1. We first prove that $V(T,x) \geq \widetilde{\vartheta}(T,x)$.

If $V(T,x) = 1$ then the inequality is obvious. We suppose that $V(T,x) < 1$, then as $V(T,x) = 1$ for all $x \in \mathcal{K}^c$, we obtain by (i) and (ii):

$$\forall (T,x) \in (]0,+\infty[\times\mathbb{R}^n),\ \forall (p_t,p_x) \in \partial_- V(T,x), \qquad p_t + \max_{\alpha\in\mathcal{A}}\{-f(x,\alpha)\cdot p_x\} \geq 0.$$

We can view V as the value function of a free state constraint problem and by the results of [8, Lemma 4.3 and Theorem 3.2] we deduce that there exists a trajectory y_x starting at x such that $V(T,x) \geq V(T-t, y_x(t)),\ \forall t \in [0,T]$. Then y_x satisfies the state constraint as $V(T-t,y_x(t)) < 1\ \forall t \in [0,T]$ and $\{z \in \mathbb{R}^n,\ V(T-t,z) < 1\} \subset \mathcal{K}$ by hypothesis. We obtain

$$V(T,x) \geq V(0,y_x(T)) = \varphi(y_x(T)) \geq \widetilde{\vartheta}(T,x).$$

Step 2. We prove that $V(T,x) \leq \widetilde{\vartheta}(T,x)$ for $x \in \mathcal{K}$. Then it suffices to show that $V(T,x) \leq \varphi(y_x(T))$ for every admissible trajectory y_x starting at x.

1) We suppose that $\{y_x(T) \in \partial\mathcal{K}$ and satisfies (OP) or $y_x(T) \in \overset{\circ}{\mathcal{K}}\}$, and that $\{x \in \partial\mathcal{K}$ and satisfies (OP) or $x \in \overset{\circ}{\mathcal{K}}\}$. As $\varphi(y_x(T)) = \liminf_{\substack{\tau\to 0^+ \\ \xi\to y_x(T),\ \xi\in\overset{\circ}{\mathcal{K}}}} V(\tau,\xi)$ then let $\tau_i \to 0^+$ and $\xi_i \to y_x(T)$, $\xi_i \in \overset{\circ}{\mathcal{K}}$ be such that

$$\varphi(y_x(T)) = \lim_{i\to\infty} V(\tau_i,\xi_i).$$

By [11, Lemma 4.1 (ii)] there exists a sequence y_i of admissible trajectories such that $y_i(T-\tau_i) = \xi_i$, $y_i(t) \in \overset{\circ}{\mathcal{K}},\ \forall t \in [0, T-\tau_i]$ and $||y_i - y_x||_{L^\infty([0,T-\tau_i])} \to_{i\to\infty} 0$.

By the Filippov theorem [2], we can extend the y_i to $[0,T]$ and choose $\sigma_i \in]T-\tau_i, T[,\ \varepsilon_i > 0$ such that $y_i(t) + \varepsilon_i\overline{\mathcal{B}}$ [1] $\in \overset{\circ}{\mathcal{K}},\ \forall t \in [0,\sigma_i]$. Then by [11, Lemma 4.2] we obtain

$$V(T,y_i(0)) \leq V(\tau_i,\xi_i)$$

[1] $\overline{\mathcal{B}}$ is the closed unit ball.

and hence we get

$$V(T,x) = V(T, y_x(0)) \leq \lim_i V(T, y_i(0)) \leq \lim_i V(\tau_i, \xi_i) = \varphi(y_x(T)).$$

2) Now we suppose that $y_x(T) \in \partial\mathcal{K}$ is such that (SIP) is satisfied. Then y_x is obviously stationnary by Lemma 5.4. Hence

$$\varphi(y_x(T)) = \varphi(x) = 1 \geq V(T,x).$$

3) The remaining case is when $x \in \partial\mathcal{K}$ and satisfies (SIP) and $\{y_x(T) \in \overset{\circ}{\mathcal{K}}$ or $y_x(T) \in \partial\mathcal{K}$ and satisfies (OP)$\}$. Let $\tau_i \to 0^+$ and $\xi_i := y_x(\tau_i)$ then by case 1) we have that $y_x(t),\ t \in [\tau_i, T]$ satisfies

$$V(T - \tau_i, \xi_i) \leq V(0, y_x(T)) = \varphi(y_x(T)).$$

As V is l.s.c. then we deduce that

$$V(T,x) \leq \liminf_i V(T - \tau_i, \xi_i) \leq \varphi(y_x(T)).$$

Combining steps 1 and 2 we obtain $V \equiv \widetilde{\vartheta}$. □

Let us comment the (SIP) constraint that we add. Let $x \in \partial\mathcal{K}$ be a boundary point such that the (SIP) constraint is satisfied. Then the vector field is repulsive around x and admissible trajectories will never reach x (the only trajectory touching the boundary at x is the one starting at x). As a consequence no trajectory arrives in x. If $0 \in f(x, \mathcal{A})$ then the corresponding stationnary trajectory does not play any role in the problem.

Remark 5.5. *The (SIP) constraint means that for a given boundary point $x \in \partial\mathcal{K}$, if no outward pointing trajectory exists then all trajectories are strictly inward pointing or stationnary. In particular there are no boundary admissible trajectories which reach the target.*

Notice that this theorem holds in particular for some transport equations as in the following example

Example 5.6. *We consider the equation*

$$v_t(t,x) + (|x_1|, 0) \cdot v_x(t,x) = 0,\ \forall t \geq 0, x \in \mathcal{K}$$

and the domain $\mathcal{K} := \overline{\mathcal{B}}((0,0), 3)$ the closed ball centered at the origin, and $\mathcal{C} = \overline{\mathcal{B}}((0,0), 1)$. The final cost φ is given by:

$$\varphi(x) := \begin{cases} 0 \text{ if } x \in \mathcal{C}, \\ 1 \text{ otherwise.} \end{cases}$$

Notice that the vector field is strictly outward pointing on $\{x \in \partial\mathcal{K}, x_1 > 0\}$ and strictly inward pointing on $\{x \in \partial\mathcal{K}, x_1 < 0\}$. On the two remaining points of $\partial\mathcal{K}$ it is null.

The Strong Inward Pointing (SIP) constraint In the continuous context, the (SIP) constraint induces continuity of the solution [13, 6]. Clearly, such a result may not be obtained when the final cost φ is only l.s.c.: $\widetilde{\vartheta}$ in this case can not be better than l.s.c.
We suppose that $\mathcal{K}$ is invariant by the dynamics f:

$$\mathcal{A}(x) = \mathcal{A}, \quad \forall x \in \partial\mathcal{K}, \qquad (SV)$$

with

$$\mathcal{A}(x) = \{\alpha \in \mathcal{A},\ \exists\, \varepsilon > 0,\ y_x(t) \in \mathcal{K},\ \dot{y}_x(t) = f(y_x(t), \alpha)\ \forall t \in [0, \varepsilon],\ y_x(0) = x\}. \qquad (5.2.2)$$

Hypothesis (SV) means that all trajectories are viable in $\mathcal{K}$ (they stay in $\mathcal{K}$ all the time). This is in particular verified if (SIP) is satisfied for all $x \in \partial\mathcal{K}$.

Under (SV), $\widetilde{\vartheta}$ coincides on $\mathcal{K}$ with the value function

$$V(T, x) := \inf\{\varphi(y_x(T)),\ \dot{y}_x(t) = f(y_x(t), \alpha(t)), a.a.\ t \in [0, T],\ y_x(0) = x\},$$

i.e.

$$\widetilde{\vartheta}(T, x) = V(T, x),\ \forall T \geq 0,\ x \in \mathcal{K}.$$

As well known from the works of Barron and Jensen [7] and of Frankowska [8], V is the unique viscosity solution of (5.2.3):

$$v_t(t, x) + \max_{\alpha \in \mathcal{A}}\{-f(x, \alpha) \cdot v_x(t, x)\} = 0,\ t > 0,\ x \in \mathbb{R}^n, \qquad (5.2.3a)$$

$$v(0, x) = \varphi(x), \qquad x \in \mathbb{R}^n. \qquad (5.2.3b)$$

Hence $\widetilde{\vartheta}$ may also be characterized as the restriction of V to $\mathcal{K}$.

5.3 A reformulation of the problem

We are interested in this section in proving that $\widetilde{\vartheta}$ is a solution of a HJB equation independantly from the set $\mathcal{K}$. This question is in fact of practical interest when dealing with the numerical approximation of $\widetilde{\vartheta}$.

We transform in the sequel the dynamics of $\widetilde{\mathcal{P}}_{T,x}$ such that we get an equivalent problem $\mathcal{P}_{T,x}$ free from explicit state constraints:

$$\mathcal{P}_{T,x} \qquad \begin{cases} \text{Minimize } \varphi(y_x(T)),\ \alpha \in A, \\ \dot{y}_x(t) \in \mathcal{F}(y_x(t)),\ a.a.\ t \in [0, T], \\ y_x(0) = x, \end{cases}$$

with $\mathcal{F}$ defined on $\mathbb{R}^n$ by:

$$\mathcal{F}(y) := \{\lambda f(y, \alpha),\ \alpha \in \mathcal{A},\ \lambda \in \Lambda(y)\}, \qquad (5.3.1)$$

and the set valued map Λ defined on $\mathbb{R}^n$ by:

$$\Lambda(y) := \begin{cases} \{1\} & \text{if } y \in \overset{\circ}{\mathcal{K}}, \\ [0,1] & \text{if } y \in \partial\mathcal{K}, \\ \{0\} & \text{if } y \in \mathcal{K}^c. \end{cases} \tag{5.3.2}$$

Then we rather study this new problem on $\mathbb{R}^n$. Notice that the two problems remain strongly linked: we can see that $\widetilde{\mathcal{P}}_{T,x}$ and $\mathcal{P}_{T,x}$ have the same value function. Our idea is that as $\mathcal{P}_{T,x}$ is free from state constraints, then its value function should verify an HJB equation on $\mathbb{R}^n$.

Let us introduce the following modified trajectories:

$$\begin{array}{l} \dot{y}_x(t) = \lambda(t) \cdot f(y_x(t), \alpha(t)), \ (\alpha(t), \lambda(t)) \in \mathcal{A} \times \Lambda(y_x(t)) \text{ a.a. } t \in [0,T], \\ y_x(0) = x, \end{array} \tag{5.3.3}$$

We set for all $\alpha \in A$, $x \in \mathbb{R}^n$ and $T \geq 0$,

$$S^{\alpha}_{[0,T]}(x) := \{y_x, \ \exists \lambda \in L^{\infty}(\mathbb{R}^+, [0,1]), \text{ s.t. } (y_x, \lambda, \alpha) \text{ satisfies (5.3.3)}\}. \tag{5.3.4}$$

Notice that this set is never empty for any given starting point $x \in \mathbb{R}^n$ and any control $\alpha \in A$.

Remark 5.7. *The controls α and λ are qualitatively different here:*

$$\alpha \in A := L^{\infty}(\mathbb{R}^+, \mathcal{A}),$$

whereas λ depends on the state:

$$\lambda(t) \in \Lambda(y(t)) \ a.a. \ t \in [0,T].$$

It is easy to check that under hypotheses (H2), the mapping $\mathcal{F}$ is upper-semi-continuous[2] (u.s.c.) and for all $y \in \mathbb{R}^n$, $\mathcal{F}(y) \neq \emptyset$ is a compact convex set. Hence, by using [1, chap. 2.1, theorem 3], there exists an absolutely continuous solution of the differential inclusion:

$$\dot{y}_x(t) \in \mathcal{F}(y_x(t)) \text{ a.a. } t \geq 0, \qquad y_x(0) = x. \tag{5.3.5}$$

Moreover, thanks to [1, chap. 1.14, corollary 1] we can extract a control $\alpha \in A$, $\lambda \in L^{\infty}(\mathbb{R}^+, [0,1])$ such that (y_x, λ, α) satisfy (5.3.3).
Conversely, if (y_x, λ, α) satisfies (5.3.3) , then clearly y_x is a solution of (5.3.5).

Let us make some comments about the dynamics transformation above. Notice that if x is in $\mathcal{K}^c := \mathbb{R}^n \backslash \mathcal{K}$, then there is no admissible trajectory for $(\widetilde{\mathcal{P}}_{T,x})$ starting at x and staying in $\mathcal{K}$ until reaching $\mathcal{C}$, so $\widetilde{\vartheta}(T,x) = 1$ and we can stop all trajectories starting at x without modifying $\widetilde{\vartheta}$. Now if $x \in \mathcal{K}$ and all trajectories leave $\mathcal{K}$ before reaching $\mathcal{C}$ or keep moving inside $\mathcal{K}$ until T without reaching $\mathcal{C}$, then $\widetilde{\vartheta}(T,x) = 1$. Moreover, when a trajectory leaves $\mathcal{K}$ before reaching $\mathcal{C}$, it is no more interesting to follow its evolution as we know that either $\widetilde{\vartheta}(T,x) = 1$ or there exists another trajectory that reaches $\mathcal{C}$ before T (and we would rather follow this latter trajectory). Once again we can stop the trajectories starting at x as soon as they leave $\mathcal{K}$ and this does not modify $\widetilde{\vartheta}$.

[2] $\mathcal{F}$ is u.s.c. if for all y and for all open set N such that $\mathcal{F}(y) \subset N$, $\exists M$ a neighborhood of y such that $\mathcal{F}(M) \subset N$.

Remark 5.8. *We can emphasize the fact that all trajectories of $\mathcal{P}_{T,x}$ are stationary outside $\mathcal{K}$. These trajectories do not have any influence on the control problem.*

The link between the original and the modified trajectories is clarified in the following lemma.

Lemma 5.9. *Let $x \in \mathbb{R}^n$, $T \in [0, +\infty[$, $\alpha \in A$ and $\widetilde{y}_x \in S^{\alpha}_{[0,T]}(x)$ Then there exists a solution z of*

$$\begin{cases} \dot{z}(t) = f(z(t), \alpha(t)), \ a.a.\ t \geq 0, \\ z(0) = x. \end{cases} \tag{5.3.6}$$

and $S \in [0, T]$ such that

$$\{z(t),\ t \in [0, S]\} = \{\widetilde{y}_x(t),\ t \in [0, T]\}.$$

Proof. We explicit here the existence of the trajectory $z(.)$. We set

$$\gamma(t) := \int_0^t \lambda(\tau) d\tau, \quad \forall t \in [0, T],$$

and consider the function built as follows:

$$a : t \to \inf\{\tau,\ 0 \leq \tau \leq t,\ \gamma(\tau) = \gamma(t)\}.$$

We show in the sequel the existence of a generalized inverse q defined almost everywhere of the function γ. Let

$$S := \int_0^T \lambda(\tau) d\tau,$$

and q be defined by: let $\eta = \gamma(t)$, we set $q(\eta) := a(t),\ \forall \eta \in [0, S]$. Then q is well defined, in fact if $\gamma(t) = \gamma(t')$ with $t < t'$ then $\beta(\tau) = 0$ *a.a.* $\tau \in [t, t']$ this implies $\gamma(\tau) = \gamma(t)\ \forall \tau \in [t, t']$. and we get

$$\begin{array}{rcl} a(t') & = & \inf\{\tau,\ 0 \leq \tau \leq t',\ \gamma(\tau) = \gamma(t')\}, \\ & = & \inf\{\tau,\ 0 \leq \tau \leq t,\ \gamma(\tau) = \gamma(t)\}, \\ & = & a(t). \end{array}$$

Hence $q : [0, S] \to [0, T]$ satisfies $q(\gamma(t)) = a(t)\ \forall t \in [0, T]$ and $\forall \eta \in [0, S]$,

$$\begin{array}{rcl} \gamma(q(\eta)) & = & \gamma(q(\gamma(t))), \\ & = & \gamma(a(t)), \\ & = & \gamma(t), \\ & = & \eta. \end{array}$$

We check now that

$$z(\eta) := \widetilde{y}(q(\eta)),\ \eta \in [0, S],$$

is the solution of (5.3.6) associated to the control

$$\alpha(q(\eta)),\ \forall \eta \in [0, S].$$

We will use for this proof

Lemma 5.10. *q is an increasing differentiable a.e.function and it satisfies*

$$\dot{q}(\eta).\lambda(q(\eta)) = 1 \ a.a \ \eta \in [0, S].$$

Proof. Notice that for a given t in $]0, T]$, either $\exists \varepsilon > 0$ such that $\gamma_{[t-\varepsilon,t]} = constant$ or $\exists \varepsilon > 0$ such that $\gamma_{[t-\varepsilon,t]}$ is strictly increasing. In the second case $\beta(\tau) > 0$ *a.a.* $\tau \in [t-\varepsilon, t]$ leads to $a(\tau) = \tau$. Hence $q(\gamma(\tau)) = \tau$. Furthermore as $\gamma(q(\eta)) = \eta$ and γ is strictly monotonous, hen q is inversible and differentiable a.e.on $\gamma([t-\varepsilon, t])$. q is moreover increasing. Let

$$\Omega = \{t \in [0, T], \ a(t) < t\},$$

then $\gamma(\Omega)$ is the set of points where q is eventually non differentiable. As q is differentiable on $\gamma([0, T] \setminus \Omega)$, it remains to show that $\gamma(\Omega)$ is of null measure. The open set $\overset{\circ}{\Omega}$ can be written as a countable union of disjoint open intervals,

$$\overset{\circ}{\Omega} = \cup_n I_n, \ I_n :=]t_n, r_n[.$$

γ is constant on each interval $\gamma(I_n) = \gamma(t_n)$. Finally

$$\begin{array}{rcl} \gamma(\Omega) & \subset & \gamma(\partial\Omega) \cup (\overset{\circ}{\Omega}), \\ & \subset & \gamma(\partial\Omega) \cup (\cup_n \gamma(I_n)), \\ & \subset & \gamma(\partial\Omega) \cup (\cup_n \gamma(t_n)). \end{array}$$

The sets I_n being countable, the border $\partial\Omega = \cup_n \partial I_n$ is also countable. We conclude that $\gamma(\Omega)$ is of measure zero. We have hence proved that q is differentiable a.e. on $[0, S]$. □

As shown in lemme 5.10, z is differentiable a.e. and $\dot{z}(\eta) = \dot{\widetilde{y}}_x(q(\eta)).\dot{q}(\eta) = f(z(\eta), \alpha(\eta))$. It remains to show that z is absolutely continuous. We can easily prove that $|\widetilde{y}_x(t)| \leq L$ with $L := (|x| + k_1 T)e^{k_1 T}$. Let $\tau, \ r$ be in $[0, S]$, $r \leq \tau$, then

$$\begin{array}{rcl} |z(\tau) - z(r)| & = & |\widetilde{y}_x(q(\tau)) - \widetilde{y}_x(q(r))|, \\ & \leq & \int_{q(r)}^{q(\tau)} |\lambda(t).f(\widetilde{y}_x(t), \alpha(t))| dt, \\ & \leq & \int_{q(r)}^{q(\tau)} \lambda(t).k_1(1 + |\widetilde{y}_x(t)|) dt, \\ & \leq & k_1(1 + L)(\gamma(q(\tau)) - \gamma(q(r))), \\ & \leq & k_1(1 + L)|\tau - r|. \end{array}$$

Hence z is Lipschitz and consequently absolutely continuous. □

Notice also the following result that will be very useful in the next section.

Lemma 5.11. *For all $x \in \mathbb{R}^n$, there exists $\varepsilon > 0$ such that for all $\alpha \in A$ and all $y_x \in S^{\alpha}_{[0,\varepsilon]}(x)$ we have*

$$\Lambda(y_x(t)) \subset \Lambda(x), \ \forall t \in [0, \varepsilon[.$$

Proof. If $x \in \partial\mathcal{K}$ then for all $\alpha \in A$, $\varepsilon > 0$ and $y_x \in S^\alpha_{[0,\varepsilon]}(x)$ we have $\Lambda(y_x(\theta)) \subset \Lambda(x) = [0,1],\ \forall\theta \in [0,\varepsilon]$.
If $x \in \overset{\circ}{\mathcal{K}}$ then there exists ε' such that $\mathcal{B}(x,\varepsilon') \subset \overset{\circ}{\mathcal{K}} \setminus \mathcal{C}$ and we can take

$$\varepsilon := \inf\{t \geq 0,\ y_x \in S^\alpha_{[0,t]}(x),\ y_x(t) \in \partial\mathcal{B}(x,\varepsilon'),\ \alpha \in A\}.$$

We get for all $y_x \in S^\alpha_{[0,\varepsilon]}(x)$ and $\theta \in [0,\varepsilon[$,

$$\Lambda(y_x(\theta)) = \{1\} = \Lambda(x).$$

If $x \in \mathbb{R}^n \backslash \mathcal{K}$ then for all $\varepsilon > 0$ and for all $y_x \in S^\alpha_{[0,\varepsilon]}(x)$, $y_x(\theta) = x$, $\forall\theta \in [0,\varepsilon]$. Then $\Lambda(y_x(\theta)) = \{0\} = \Lambda(x)$. □

Let ϑ denote the value function of $(\mathcal{P}_{T,x})$,

$$\vartheta(T,x) := \inf(\mathcal{P}_{T,x}).$$

It is well known in control theory [14] that under assumptions (H0)-(H2), $\mathcal{P}_{T,x}$ admits a minimizer.

Lemma 5.12. *Under assumptions (H0)-(H2), there exists an optimal control for the problem $(\mathcal{P}_{T,x})$ and the value function ϑ is l.s.c.*

5.4 The HJB equation

As a first step to derive the HJB equation, the classical procedure is to prove that the value function ϑ satisfies the Dynamic Programming Principle (DPP) and the Backward Dynamic Programming Principle (BDPP). Let us first introduce the set of backward controls

$$A^- := \{\alpha \in L^\infty(\mathbb{R}^-, \mathcal{A})\},$$

and the set of backward trajectories, associated to $\alpha \in A^-$ and arriving at x, as the solutions of:

$$\dot{y}_x(t) \in \mathcal{F}(y_x(t)) \ a.a.\ t \in \mathbb{R}^-, \quad y_x(0) = x.$$

Identically to forward trajectories, we will use the notation (5.3.4) for the backward trajectories associated to the control $\alpha \in A^-$. We can by now state

Proposition 5.13. *Assume (H1)-(H2), then*
• *The value function ϑ satisfies the DPP: for all $x \in \mathbb{R}^n$, $T \geq 0$ and all $\tau \in]0,T]$,*

$$\vartheta(T,x) = \inf_{\substack{\alpha\in A \\ y_x \in S^\alpha_{[0,\tau]}(x)}} \vartheta(T-\tau, y_x(\tau)).$$

• *The value function ϑ satisfies the BDPP: for all $T \geq 0$ and all $\tau \geq 0$,*

$$\vartheta(T,x) \geq \vartheta(T+\tau, y_x(-\tau)),\ \forall\alpha \in A^-,\ y_x \in S^\alpha_{[0,-\tau]}(x).$$

Let us define for a function $\phi \in C^1(\mathbb{R}\times\mathbb{R}^n)$ and $t>0,\ x\in\mathbb{R}^n$ the Hamiltonian,

$$\mathcal{H}(\phi)(t,x) := \max_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}} \{-\lambda f(x,\alpha).\phi_x(t,x)\}.$$

Following the standard technique of the proof given by Bardi and Capuzzo-Dolcetta [3] and using Lemma 5.11, it is not difficult to prove the following theorem which uses the notion of interior Hamiltonian $\mathcal{H}_{int}$ introduced by Ishii and Koike [12]. We define $\mathcal{H}_{int}$ whenever the set $\mathcal{A}(x)\neq\emptyset$ ($\mathcal{A}(x)$ is defined by (5.2.2)) by

$$\mathcal{H}_{int}(\phi)(t,x) := \max_{\substack{\alpha\in\mathcal{A}(x)\\ \lambda\in\Lambda(x)}} \{-\lambda f(x,\alpha).\phi_x(t,x)\}.$$

Theorem 5.14. *Assume (H0)-(H2), then the value function ϑ satisfies the initial condition*

$$\liminf_{\substack{\tau\to 0^+\\ \xi\in\overset{\circ}{\mathcal{K}},\ \xi\to x}} \vartheta(\tau,\xi) = \varphi(x),\ \forall x\in\mathcal{K}. \tag{5.4.1}$$

and the HJB equation:

$$\vartheta_t(t,x) + \max_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}} \{-\lambda f(x,\alpha).\vartheta_x(t,x)\} = 0,\ t\in]0,+\infty[,\ x\in\mathbb{R}^n, \tag{5.4.2}$$

in the following sense:
1) For all $\phi\in C^1(\mathbb{R}\times\mathbb{R}^n)$, if $(t,x)\in]0,+\infty[\times\mathbb{R}^n$ is a minimum of $\vartheta-\phi$ on $]0,+\infty[\times\mathbb{R}^n$, then

$$\phi_t(t,x)+\mathcal{H}(\phi)(t,x)\geq 0.$$

2) For all $\phi\in C^1(\mathbb{R}\times\mathbb{R}^n)$, and all minimum $(t,x)\in]0,+\infty[\times\mathbb{R}^n$ of $\vartheta-\phi$ on $]0,+\infty[\times\mathbb{R}^n$,
i) If $x\in\overset{\circ}{\mathcal{K}}\cup\mathcal{K}^c$, then

$$\phi_t(t,x)+\mathcal{H}(\phi)(t,x)\leq 0.$$

ii) If $x\in\partial\mathcal{K}$ and $\mathcal{A}(x)\neq\emptyset$, then

$$\phi_t(t,x)+\mathcal{H}_{int}(\phi)(t,x)\leq 0.$$

Remark 5.15. *When $\mathcal{K}=\mathbb{R}^n$, then by theorem 5.14 we get that ϑ is a l.s.c. viscosity solution of the HJB equation (5.4.2).*

Notice however that when $\mathcal{K}=\mathbb{R}^n$ then $\Lambda(x)=\{1\}$ for all $x\in\mathbb{R}^n$ and we recover the dynamics f for $\mathcal{P}_{T,x}$, in particular Λ is continuous in this case. But in fact this result remains valid in some cases where Λ is only u.s.c. For example we consider the finite horizon target problem

$$\begin{array}{lll} \mathcal{T}_{T,x}: & \textit{Minimize} \quad \varphi(y_x(t)), & t\leq T, \alpha\in A,\\ & \dot{y}_x(\tau)=f(y_x(\tau),\alpha(\tau)) & \textit{a.a. } \tau\in[0,t],\ y_x(0)=x,\\ & y_x(\tau)\in\mathcal{K} \quad \forall\tau\in[0,t]. & \end{array}$$

Then we may convert $\mathcal{T}_{T,x}$ into a RDV problem of $\mathcal{P}_{T,x}$ type by setting:

$$\begin{array}{ll} \textit{Minimize} & \varphi(y_x(T)),\ \alpha \in A, \\ & y_x \in S^{\alpha}_{[0,T]}(x) \ \textit{with}\ \Lambda\ \textit{given by}\ (5.4.3). \end{array}$$

$$\Lambda(y) := \begin{cases} \{0\} & \textit{if } y \in \mathcal{K}^c, \\ [0,1] & \textit{if } y \in \mathcal{C} \cup \partial\mathcal{K}, \\ \{1\} & \textit{if } y \in \overset{\circ}{\mathcal{K}} \setminus \mathcal{C}. \end{cases} \tag{5.4.3}$$

Hence when $\mathcal{K} = \mathbb{R}^n$, the value function $\vartheta_{\mathcal{C}}(T,x) := \inf(\mathcal{T}_{T,x})$ is a l.s.c. viscosity solution of (5.4.2) eventhough the set valued map Λ defined by (5.4.3) is only u.s.c.

More generally the control set $\Lambda(x)$ must be chosen such that the RDV problem admits non statinnary trajectories starting from x for all $x \in \mathcal{K}$. These trajectories will provide the DPP and consequently the HJB formulation in $\mathcal{K}$.

When $\mathcal{K} \neq \mathbb{R}^n$ We consider now a particular final cost function φ defined by

$$(H0') \qquad \varphi(x) := \begin{cases} 0 \text{ if } x \in \mathcal{C}, \\ 1 \text{ otherwise.} \end{cases}$$

Our interest for this specific final cost is motivated in particular by the minimum time problem handled in chapter 4.

Let $\chi_{\mathcal{K}}$ be defined by

$$\chi_{\mathcal{K}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{K}, \\ 1 & \text{otherwise.} \end{cases}$$

As shown in Theorem 5.14 2), the second inequality is not obtained for all controls on the boundary points. This is shown in the following example.

Example 5.16. *Let $n = 1$, $\mathcal{C} = [0,1]$, $\mathcal{K} = [-1,1]$ and $f(x,\alpha) = 1$. Then*

$$\vartheta(t,x) = \begin{cases} \chi_{[-t,1]}(x) & \textit{if } t \in [0,1], \\ \chi_{[-1,1]}(x) & \textit{if } t \geq 1, \end{cases}$$

Now we take $t \geq 1$, $x = -1$ and ϕ regular such that $\phi(s,z) = -1 - z$ locally around (t,x). We have $\phi(t,x) = \vartheta(t,x) = 0$, $\vartheta - \phi$ minimal at (t,x) and $\phi_t(t,x) + \mathcal{H}(\phi)(t,x) = 1 > 0$.

It is clear that the modified dynamics formulation fails to give a characterization of the value function ϑ on $\mathbb{R}^n$ independently from the set $\mathcal{K}$. Eventhough we have no more any explicit state constraint in $\mathcal{P}_{T,x}$, the constraint appears in the dynamics $\mathcal{F}$ which is only u.s.c. Hence the HJB equation that we obtain in Theorem 5.14 is still depending on $\mathcal{K}$.

We propose here to prove alternatively to the HJB equation (5.4.2) that ϑ satisfies the inequality given in

Theorem 5.17. *We assume (H0'), (H1) and (H2). Let $t > 0$, $x \in \mathbb{R}^n$ and let ϕ be a regular function such that $\vartheta - \phi$ admits a minimum at $(t,x) \in]0,+\infty[\times\mathbb{R}^n$ on $]0,+\infty[\times\mathbb{R}^n$. Then*

$$\min\{\phi_t(t,x) + \mathcal{H}(\phi)(t,x), \vartheta(t,x) - \chi_{\mathcal{K}}(x)\} = 0. \tag{5.4.4}$$

Proof. We can suppose that $\vartheta(t,x) = \phi(t,x)$.
•As already shown in Theorem 5.14 (1), we have $\phi_t(t,x) + \mathcal{H}(\phi)(t,x) \geq 0$. We also have clearly $\vartheta(t,x) - \chi_{\mathcal{K}}(x) \geq 0$.
•Now we have to prove that the equality is fulfilled. We distinguish different cases depending on the position of x:
i) if $x \in \overset{\circ}{\mathcal{K}} \cup \mathcal{K}^c$, then by Theorem 5.14 (2) i) we get directly $\phi_t(t,x) + \mathcal{H}(\phi)(t,x) \leq 0$.
ii) if $x \in \partial\mathcal{K}$, then if $\vartheta(t,x) = 0$ we get $\vartheta(t,x) - \chi_{\mathcal{K}}(x) = 0$. Otherwise if $\vartheta(t,x) = 1$, then as ϑ is l.s.c, there exists an open neighborhood $\mathcal{N}(t,x)$ of (t,x) such that $\vartheta \equiv 1$ on $\mathcal{N}(t,x)$. Hence as $1 - \phi(s,z)$ attains a minimum at (t,x) on $\mathcal{N}(t,x)$, then $\nabla\phi(t,x) = 0$ and consequently $\phi_t(t,x) + \mathcal{H}(\phi)(t,x) = 0$.

□

Remark 5.18. *A comparable equation to (5.4.4) appears in [4] in the context of stopping time problems with discontinuous obstacle function ψ.*

5.5 Contingent epiderivatives

We introduce here the notion of contingent epiderivative used in [8, 10, 11] and verify that some results of Frankowska et al. extend (without any additional difficulty) to the case when the dynamics $\mathcal{F}$ defined in section 5.3 is u.s.c.
Let $\mathcal{U} : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be an extended function. The contingent epiderivative of $\mathcal{U}$ at $x \in Dom\,\mathcal{U}$ in the direction $u \in \mathbb{R}^n$ is

$$D_{\uparrow}\mathcal{U}(x)(u) = \liminf_{\substack{h\to 0+ \\ u'\to u}} \frac{\mathcal{U}(x+hu') - \mathcal{U}(x)}{h}.$$

Remark 5.19. *The contingent epiderivative defined above is also called lower generalized Dini derivative or again upper contingent derivative in control theory literature [3].*

5.5.1 Some properties of the value function

It is not difficult to prove using the arguments of Frankowska [8, Theorem 3.2 and Theorem 3.3] the following properties satisfied by the value function. In particular these properties investigate the relationship between epiderivatives and the DPP satisfied by ϑ. This principle induces clearly the monotonicity of ϑ along admissible trajectories: for all $t \geq 0$, $x \in \mathbb{R}^n$, $\alpha \in A$ and $y_x \in S^{\alpha}_{[0,t]}(x)$, $\tau \in]0,t]$,

$$\vartheta(t,x) \leq \vartheta(t-\tau, y_x(\tau)),$$

and Theorem 5.12 proves, under (H0)-(H2), the existence of an optimal trajectory such that ϑ is constant along this trajectory: there exists $\alpha \in A$ and $y_x \in S^{\alpha}_{[0,t]}(x)$ such that for all $\tau \in [0,t]$,

$$\vartheta(t-\tau, y_x(\tau)) = \vartheta(t,x).$$

Lemma 5.20. *The value function ϑ satisfies:*
i)

- *For all $x \in \mathbb{R}^n \backslash \partial\mathcal{K}$, for all $t \geq 0$,* $\sup_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\{0,1\}\cap\Lambda(x)}} D_{\uparrow}\vartheta(t,x)(1,-\lambda f(x,\alpha)) \leq 0.$
- *For all $x \in \partial\mathcal{K}$, for all $t \geq 0$,* $D_{\uparrow}\vartheta(t,x)(1,0) \leq 0$, *and*
  whenever $\mathcal{A}(x) \neq \emptyset$, $\sup_{\alpha\in\mathcal{A}(x)} D_{\uparrow}\vartheta(t,x)(1,-f(x,\alpha)) \leq 0.$

ii) For all $x \in \mathbb{R}^n$, for all $t \geq 0$, $\inf_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}} D_{\uparrow}\vartheta(t,x)(-1,\lambda f(x,\alpha)) \leq 0.$

Proof. For sake of clarity we prove i)
By the backward DPP, for all $\tau > 0$ and all $\alpha \in \mathcal{A}$, $y_x \in S^{\alpha}_{[0,-\tau]}(x)$ we have

$$\vartheta(t,x) \geq \vartheta(t+\tau, y_x(-\tau)),$$

where $y_x(-\tau) = x + \int_0^{-\tau} \lambda(\theta) f(y_x(\theta),\alpha)d\theta$ and λ is the associated control to y_x.

If $x \in \overset{\circ}{\mathcal{K}}$ then $\exists \tau > 0$ such that $y_x(-\theta) \in \overset{\circ}{\mathcal{K}} \backslash \mathcal{C}$, $\forall \theta \in [0,\tau]$. Hence $\Lambda(x) = \{1\} = \Lambda(y_x(-\theta))$. We obtain then that y_x is an admissible backward trajectory with $\lambda \equiv 1$. We get

$$\frac{\vartheta(t+\tau, y_x(-\tau)) - \vartheta(t,x)}{\tau} \leq 0.$$

By the continuity of f and y_x, $\frac{y_x(-\tau)-x}{\tau} = \frac{1}{\tau}\int_0^{-\tau} f(y_x(\theta),\alpha)d\theta \to_{\tau\to 0} -f(x,\alpha)$. Finally

$$\liminf_{\substack{\tau\to 0+\\ u'\to(1,-f(x,\alpha))}} \frac{\vartheta((t,x)+\tau u') - \vartheta(t,x)}{\tau} \leq 0,$$

which leads to $D_{\uparrow}\vartheta(t,x)(1,-\lambda f(x,\alpha)) \leq 0$, $\forall\alpha \in \mathcal{A}$, $\forall\lambda \in \Lambda(x)$. Now, if $x \in \mathcal{K}^c$ then $\Lambda(x) = \{0\}$. Applying the backward DPP for the trajectory $y_x(-\tau) \equiv x$, we get

$$\sup_{\alpha\in\mathcal{A},\ \lambda\in\Lambda(x)} D_{\uparrow}(t,x)(1,-\lambda f(x,\alpha)) \leq 0.$$

Finally, if $x \in \partial\mathcal{K}$ then $\Lambda(x) = [0,1]$. Let λ be in $]0,1]$ and $z(-\tau) := y_x(\frac{-\tau}{\lambda})$.
if y_x is admissible then the trajectory z is also an admissible trajectory associted to $\lambda \equiv 1$. As previously, we prove using the backward DPP that

$$D_{\uparrow}\vartheta(t,x)(1,-f(x,\alpha)) \leq 0.$$

Now if $\lambda \equiv 0$ then $y_x(-\tau) := x$ is admissible, and as previously this leads to

$$D_{\uparrow}\vartheta(t,x)(1,0) \leq 0.$$

□

5.5.2 Properties of contingent epiderivatives

We recall in Lemma 5.21 and Lemma 5.22 some properties satisfied by contingent epiderivatives and proved in [8, Theorem 3.2 and Theorem 3.3]. These properties extend to our case where $\mathcal{F}$ is only u.s.c. and prove how to recover the DPP from the inequalities on contingent derivatives.

Lemma 5.21. *Let $\mathcal{U} : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ be a l.s.c. function and assume that hypotheses (H1)-(H2) are satisfied, then the following statements are equivalent:*
i) For all $(t,x) \in Dom\,\mathcal{U}$, $\inf_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}} D_\uparrow\mathcal{U}(t,x)(-1,\lambda f(x,\alpha)) \leq 0$.
ii) For all $(t,x) \in \mathbb{R}^+ \times \mathbb{R}^n$, there exist $\alpha \in A$ and $y_x \in S^\alpha_{[0,t]}(x)$ such that

$$\mathcal{U}(t-\tau, y_x(\tau)) \leq \mathcal{U}(t,x), \quad \forall \tau \in [0,t].$$

Lemma 5.21 shows the equivalence between inequality i) and the existence of an optimal trajectory y_x.

We also prove using the arguments of [8, Theorem 3.3] the following lemma:

Lemma 5.22. *Let $\mathcal{U} : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ be a l.s.c. function and assume that hypotheses (H1)-(H2) are satisfied, then i) and ii) are equivalent*
i) For all $t \geq 0$ and all $x \in \overset{\circ}{\mathcal{K}}$,

$$\sup_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\{0,1\}\cap\Lambda(x)}} D_\uparrow\mathcal{U}(t,x)(1,-\lambda f(x,\alpha)) \leq 0. \tag{5.5.1}$$

ii) For all $(t,x) \in \mathbb{R}^+ \times \overset{\circ}{\mathcal{K}}$ and $\alpha \in A$, $y_x \in S^\alpha_{[0,t]}(x)$ with $y_x(\tau) \in \overset{\circ}{\mathcal{K}}$ for all $\tau \in [0,t]$, $\mathcal{U}(t,x) \leq \mathcal{U}(t-\tau, y_x(\tau))$, $\forall \tau \in [0,t]$.

Remark 5.23. *Notice that Lemma 5.22 concerns only interior trajectories. In fact for a trajectory touching the boundary $\partial\mathcal{K}$ (especially trajectories which arrive on $\partial\mathcal{K}$ then leave it or stay on $\partial\mathcal{K}$ during some time), we need to approach it by trajectories from the interior. The (OP) qualification constraint*

$$\forall x \in \partial\mathcal{K},\ \exists \alpha \in \mathcal{A},\ \lambda \in \Lambda(x) \ \ s.t.\ \nabla h_j(x)\cdot\lambda f(x,\alpha) > 0,\ \forall j \in I(x), \qquad (OP)$$

insures that we could construct such an approaching admissible trajectory. This sequence allows in [11] to extend Lemma 5.22 and treat all admissible trajectories. Hence in [11, Theorem 2.1] Frankowska and Vinter show that under (OP), a function $\mathcal{U}$ satisfying (5.5.1) for all $x \in \mathcal{K}$ is monotonous along all admissible trajectories.
It appears then that the (OP) constraint is not only sufficient but also inevitable to recover the DPP characterizing the value function ϑ, from contingent epiderivative inequalities.

5.6 The $\mathcal{F}$-contingent characterization

We have seen in the previous section that using epiderivatives, we may recover the monotonicity of a given function $\mathcal{U}$ along interior trajectories (Lemma 5.22) and the existence of an optimal trajectory (Lemma 5.21). The missing information concerns the monotonicity of $\mathcal{U}$ along boundary admissible trajectories, and requires necessarily qualification constraints on $\partial\mathcal{K}$. We introduce here a new notion of derivative akin to contingent epiderivatives. This derivative will allow to complete the lacking monotonicity information and to get a uniqueness result (in terms of epiderivatives) without any controllability assumption.

Definition 5.24. *Let $\mathcal{U} : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be an extended function. Then the $\mathcal{F}$-contingent epiderivative of $\mathcal{U}$ at (t, x) is:*

$$D^{\mathcal{F}}_{\uparrow}\mathcal{U}(t,x) = \sup_{\substack{y_x \in S^{\alpha}_{[0,-\tau]}(x) \\ \tau>0,\ \alpha\in A}} \liminf_{h\to 0^+} \frac{\mathcal{U}(t+h, y_x(-h)) - \mathcal{U}(t,x)}{h}.$$

Lemma 5.25. *Let $\mathcal{U} : \mathbb{R}^+ \mapsto \{0, 1\}$ be a l.s.c. function satisfying for all $t \in \mathbb{R}^+$,*

$$D_+\mathcal{U}(t) := \liminf_{h\to 0^+} \frac{\mathcal{U}(t+h) - \mathcal{U}(t)}{h} \leq 0,$$

then $\mathcal{U}$ is a decreasing function.

Proof.

Suppose that there exists $t_1 < t_2$ such that $\mathcal{U}(t_1) < \mathcal{U}(t_2)$. , then as $\mathcal{U}$ takes only values 0 and 1, $\mathcal{U}(t_1) = 0$ and $\mathcal{U}(t_2) = 1$.
As $\mathcal{U}$ is l.s.c, there exists a neighborhood $\mathcal{N}(t_2)$ of t_2 such that $\mathcal{U}(t) = 1 \quad \forall t \in \mathcal{N}(t_2)$. Let

$$t := \sup\{t',\ t_1 \leq t' \leq t_2 \text{ such that } \mathcal{U}(t') = 0\},$$

then $\mathcal{U}(t) = 0$.
Let $h_n \to 0^+$ be such that

$$D_+\mathcal{U}(t) = \lim_{h_n\to 0^+} \frac{\mathcal{U}(t+h_n) - \mathcal{U}(t)}{h_n},$$

then $\mathcal{U}(t + h_n) = 1$ for $h_n \leq t_2 - t$ and consequently

$$\frac{\mathcal{U}(t+h_n) - \mathcal{U}(t)}{h_n} = \frac{1}{h_n} \to +\infty \text{ when } h_n \to 0^+,$$

which contradicts the hypothesis. □

We prove in the next Lemma that if the $\mathcal{F}$-contingent epiderivative is negative for a given l.s.c. function $\mathcal{U}$, then $\mathcal{U}$ is monotone along admissible trajectories.

Lemma 5.26. *Let $\mathcal{U}$ be a l.s.c. function taking values in $\mathbb{R}$ and satisfying*

$$\forall t \geq 0,\ \forall x \in \mathbb{R}^n,\ D_{\uparrow}^{\mathcal{F}}\mathcal{U}(t,x) \leq 0.$$

Then for all $(t_0, x_0) \in \mathbb{R}^+ \times \mathbb{R}^n,\ \alpha \in A,\ y_{x_0} \in S^{\alpha}_{[0,t_0]}(x_0)$,

$$t \mapsto \mathcal{U}(t, y_{x_0}(t_0 - t)),$$

is a decreasing function for $t \in [0, t_0]$.

Proof. Let y_{x_0} be in $S^{\alpha}_{[0,t_0]}(x_0)$, $t \in [0, t_0]$ and $x = y_{x_0}(t_0 - t)$, then

$$D_+\mathcal{U}(t) = \liminf_{h\to 0^+} \frac{\mathcal{U}(t+h, y_{x_0}(t_0 - t - h)) - \mathcal{U}(t, y_{x_0}(t_0 - t))}{h},$$

can be written

$$D_+\mathcal{U}(t) = \liminf_{h\to 0^+} \frac{\mathcal{U}(t+h, z(-h)) - \mathcal{U}(t,x)}{h},$$

where $z(-h) = y(t_0 - t - h) \in S^{\alpha}_{[0,-h]}(x)$, for $h \geq 0$. Then by hypothesis, $D_+\mathcal{U}(t) \leq 0$ and by lemma 5.25 we get that $t \mapsto \mathcal{U}(t, y_{x_0}(t_0 - t))$ is decreasing. □

We can now state

Theorem 5.27. *Let $\mathcal{U} : \mathbb{R}^+ \times \mathbb{R}^n \to \{0, 1\}$ be a l.s.c. function satisfying $\mathcal{U}(t,x) = 1$ for all $t \geq 0$ and $x \in \mathcal{K}^c$, and*

- $\liminf_{\substack{\tau\to 0^+ \\ \xi\to x,\ \xi\in\mathring{\mathcal{K}}}} \mathcal{U}(\tau,\xi) = \mathcal{U}(0,x) = \varphi(x), \quad \forall x \in \mathcal{K},$
- $\inf_{\substack{\alpha\in\mathcal{A} \\ \lambda\in\Lambda(x)}} D_{\uparrow}\mathcal{U}(t,x)(-1, \lambda f(x,\alpha)) \leq 0, \qquad \forall t \geq 0,\ x \in \mathbb{R}^n,$
- $\sup_{\substack{\alpha\in\mathcal{A} \\ \lambda\in\{0,1\}\cap\Lambda(x)}} D_{\uparrow}\mathcal{U}(t,x)(1, -\lambda f(x,\alpha)) \leq 0, \quad \forall t \geq 0,\ x \in \mathring{\mathcal{K}},$
- $D_{\uparrow}^{\mathcal{F}}\mathcal{U}(t,x) \leq 0, \qquad \forall t \geq 0,\ x \in \partial\mathcal{K}.$

Then $\mathcal{U} \equiv \vartheta$.

Proof. Suppose that t, x is such that $\vartheta(t,x) = 1$. By lemma 5.21, there exists $\alpha \in A$ and $y_x \in S^{\alpha}_{[0,t]}(x)$ such that $\mathcal{U}(t-\tau, y_x(\tau)) \leq \mathcal{U}(t,x), \quad \forall \tau \in [0,t]$. For $\tau = t$ we get $\varphi(y_x(t)) = \mathcal{U}(0, y_x(t)) \leq \mathcal{U}(t,x)$. On the other hand, by the DPP $\vartheta(t,x) \leq \vartheta(0, y_x(t)) := \varphi(y_x(t))$, and consequently $\vartheta(t,x) \leq \mathcal{U}(t,x)$. Then $\mathcal{U}(t,x) = 1$.

Now if $\vartheta(t,x) = 0$ then there exists an admissible trajectory $y_x \in S^{\alpha}_{[0,t]}(x)$ such that $y_x(t) \in \mathcal{C}$. Using lemma 5.26 , we get that $a(\tau) := \mathcal{U}(\tau, y_x(t-\tau))$ is decreasing. Hence we have for $\tau = t$,

$$\mathcal{U}(t,x) \leq \mathcal{U}(0, y_x(t))$$

Finally as $\mathcal{U}(0, y_x(t)) = \varphi(y_x(t)) = 0$ then we get $\mathcal{U}(t,x) = 0$. □

Remark 5.28. *In [8, Theorem 3.3], the sup on contingent epiderivatives*

$$\sup_{\substack{\alpha\in\mathcal{A}\\ \lambda\in\Lambda(x)}} D_{\uparrow}\mathcal{U}(t,x)(1,-\lambda f(x,\alpha)) \leq 0,$$

was sufficient to deduce the monotonicity of $\mathcal{U}$ along admissible trajectories (in the case $\mathcal{K} = \mathbb{R}^n$). We have seen that this is not the case here without any (OP) qualification constraint. Alternatively, the last inequality in Theorem 5.27 allows to recover this information.

Bibliography

[1] J. P. Aubin and A. Cellina. *Differential inclusions*, volume 264.

[2] J. P. Aubin and H. Frankowska. *Set-valued analysis*. Birkhauser, 1990.

[3] M. I. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and viscosity solutions of Hamilton Jacobi Bellman equations*. Birkhäuser Boston, 1997.

[4] G. Barles and B. Perthame. Discontinuous solutions of deterministic stopping time problems. *M2AN Mathematical modelling and numerical analysis*, 21(4):557–579, 1987.

[5] G. Barles and B. Perthame. Exit time problems in optimal control and vanishing viscosity method. *SIAM J. Control and Optim.*, 26(5):1133–1148, 1988.

[6] G. Barles and B. Perthame. Comparison principle for Dirichlet type Hamilton Jacobi equations and singular perturbations of degenerated elliptic equations. *Appl. Math. Optim.*, 21:21–44, 1990.

[7] E. N. Barron and R. Jensen. Semicontinuous viscosity solutions for Hamilton Jacobi equations with convex hamiltonian. *Comm. Partial Differential equations*, 15:1713–1742, 1990.

[8] F. Frankowska. Lower semi-continuous solutions of hamilton-jacobi-equations. *SIAM J.Control Optim.*, 31:257–272, 1993.

[9] F. Frankowska and S. Plaskacz. Semicontinuous solutions of hamilton-jacobi equations with state constraints. *Differential inclusions and optimal control, Lecture notes in nonlinear analysis*, 2:145–161, 1998.

[10] F. Frankowska and S. Plaskacz. Semicontinuous solutions of Hamilton Jacobi equations with degenerate state constraints. *JMAA*, pages 818–838, 2000.

[11] F. Frankowska and R. B. Vinter. Existence of neighboring feasible trajectories: applications to dynamic programming for state constrained optimal control problems. *I.Optim.Theory Appl.*, 104:27–40, 2000.

[12] H. Ishii and S. Koike. A new formulation of state constraint problems for first order pdes. *SIAM J. Control and Optimization*, 34(2):554–571, 1996.

[13] H. M. Soner. Optimal control with state space constraint. *SIAM Journal of Control and Optimization*, 24(3):552–561, 1986.

[14] J. Warga. Relaxed variational problems. *J. Mathematical Analysis and Applications*, 4:111–128, 1962.