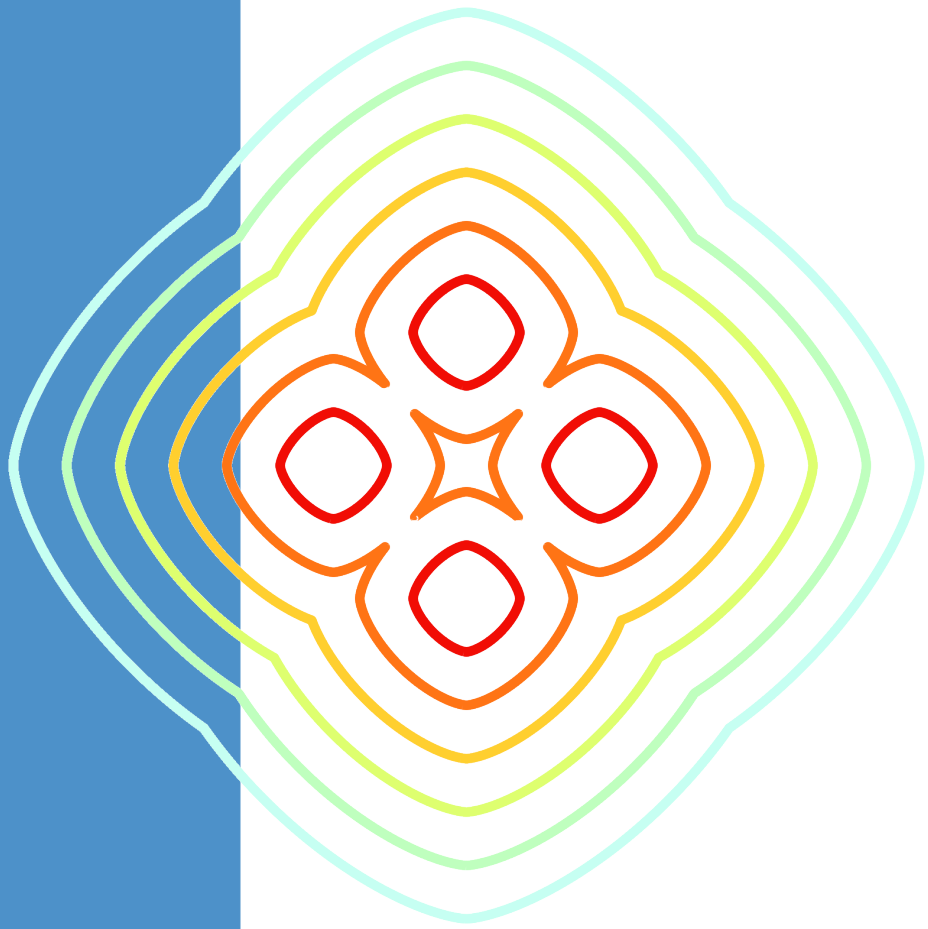
Numerical methods for Hamilton-Jacobi equations

Averil Prost
INSA Rouen Normandie

Master's thesis, 03/01/2022 - 08/31/2022
directed by
Olivier Bokanowski
Lab. Jacques-Louis Lions



INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

JJL

As the reader may have noticed, the title of this report reads *Numerical methods for Hamilton-Jacobi equations*. It is therefore no surprise if we begin with some words on Hamilton-Jacobi equations, and then present a few numerical methods. The motivation behind this study is to solve control problems, and more precisely, optimal path problems. Practical applications immediately come to mind, as driving cargo ships towards the moon with as few fuel as possible, or finding the shortest route towards the coffee machine.

In the first chapter, we precise the definitions and class of control problem that we adress, we draw the connection with Hamilton-Jacobi equations, and we detail a representation theorem in a particular case. In the second chapter, we investigate first a class of schemes for 1D advection equations, that are, in the constant speed case, both Hamilton-Jacobi equations and scalar conservation laws. The last section turns to a Lagrangian scheme for state-constrained problems, where the focus is put on the space of controls via internal approximations.

I wish to thank all the people that helps me to reduce my ignorance, and in particular the main interlocutors of this internship. First, let me mention Nathalie Bergame, who was terribly efficient on the administrative part. Warm thanks to Hasnaa Zidani, her clarity of mind and her useful advice. Special thanks to Nicolas Forcadel, who manages to spare time in a highly constrained environment and provides constant support. Finally, my gratitude goes to my tutor Olivier Bokanowski, a humble yet wise researcher who shared his insights, knowledge and pythonic wild trips with great generosity. I hope, and already know, that I go on with you three.

Contents

Chapter 1

Theoretical framework

1.1 Control problems

1.1.1 Definitions and assumptions

Given an horizon time $T > 0$, we will consider a set of trajectories guided by an ODE, of the form

$$\dot{y}_x^a(t) = f(y_x^a(t), a(t)), \quad y_x^a(0) = x \in \mathbb{R}^n. \tag{1.1}$$

Let us precise the meaning of each term. The function $a(\cdot)$ is taken in the following space.

(A1) **Regularity of a** Let $A \subset \mathbb{R}^p$ be a compact subset. It is assumed that $a \in \mathbb{A}_{[0,T]}$, where $\mathbb{A}_{[t_1,t_2]} \coloneqq L^\infty([t_1, t_2], A)$.

The dynamic $f(\cdot, \cdot)$ is taken as follows.

(A2) **Regularity of f** The function $f : \mathbb{R}^n \times \mathbb{R}^p \mapsto \mathbb{R}^n$ is Lipschitz-continuous with respect to both variables, i.e. there exists constants $[f]_x \geqslant 0$ and $[f]_a \geqslant 0$ such that

$$|f(x, a) - f(y, b)| \leqslant [f]_x |x - y| + [f]_a |a - b| \quad \forall (x, y, a, b) \in (\mathbb{R}^n)^2 \times (\mathbb{R}^p)^2$$

Moreover, it is assumed that $f(x, A) \coloneqq \{f(x, a) \mid a \in A\}$ is convex for any $x \in \mathbb{R}^n$.

Notice that by continuity of $f(x, \cdot)$, the set $f(x, A)$ is also compact.

With this definition, the function $(y, t) \mapsto f(y, a(t))$ is measurable in t and Lipschitz-continuous in y, so is a Caratheodory function. The theorem of Caratheodory ensures the existence of a local absolutely continuous solution over $[0, \theta]$ with $\theta > 0$. Moreover, for all $t \in [0, \theta]$,

$$|y_x^a(t) - x| \leqslant \int_0^t |f(y_x^a(s), a(s))| ds \leqslant \int_0^t |f(x, a(s))| + [f]_x |y_x^a(s) - x| ds \leqslant t|f(x, A)|_\infty + \int_0^t [f]_x |y_x^a(s) - x| ds$$

where $|f(x, A)|_\infty \coloneqq \max_{a \in A} |f(x, a)|$. By a Grönwall lemma, we have the bound (uniform in θ)

$$|y_x^a(t)| \leqslant |x| + t|f(x, A)|_\infty e^{t[f]_x} \leqslant |x| + T|f(x, A)|_\infty e^{T[f]_x} \tag{1.2}$$

and the boundedness ensures global existence of absolutely continuous trajectories $(y_x^a(t))_{t \in [0,T]} \in W^{1,1}([0, T], \mathbb{R}^n)$. We may now turn to some control problems of interest.

Mayer problem To each trajectory, we associate a cost

$$J_M : \mathbb{R}^n \times \mathbb{A}_{[0,T]} \mapsto J_M(x, a) \coloneqq \int_0^T l(y_x^a(t), a(t)) dt + \varphi(y_x^a(T)),$$

where $l \in L^1(\mathbb{R}^n \times A, \mathbb{R})$ is the running cost, and $\varphi : \mathbb{R}^n \to \mathbb{R}$ is the terminal cost. Our aim is the following:

$$\text{Find } a^* \in \mathbb{A}_{[0,T]} \text{ such that } J_M(x, a^*) = \inf_{a \in \mathbb{A}_{[0,T]}} J_M(x, a).$$

Bolza problem Mayer problems (with running cost) can be viewed as a Bolza problem (with $l \equiv 0$) by augmenting the dimension. Indeed, denote by

$$\xi^a_{x,z}(0) \coloneqq z, \quad \dot{\xi}^a_{x,z}(t) = l(t, y^a_x(t), a(t)).$$

If we assume that l satisfies the same regularity as f, then ξ is well-defined, and the problem is equivalent to the minimization (in $z = 0$) of

$$\widetilde{J}(x, z, a) \coloneqq \xi^a_{x,z}(T) - z + \varphi(y^a_x(T)).$$

Since we minimize J_M with respect to a, keeping (x, z) fixed, we may replace $\widetilde{J}$ by $J_B(x, z, a) \coloneqq \xi^a_{x,z}(T) + \varphi(y^a_x(T))$. The couple $(\xi^a_{x,z}(T), y^a_x(T))$ is the *augmented state* in a problem with only a terminal cost.

For this reason, in the sequel, we focus on Bolza problems with the following notations.

Definition 1 – Bolza problem For each $x \in \mathbb{R}^n$, find $a^* \in \mathbb{A}_{[0,T]}$ such that

$$J(x, a^*) \coloneqq \varphi(y^{a^*}_x(T)) = \inf_{a \in \mathbb{A}_{[0,T]}} \varphi(y^a_x(T)). \tag{1.3}$$

In the sequel, we consider the following terminal costs.

(A3) Regularity of φ The function $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ is Lipschitz-continuous with constant $[\varphi]$.

Remark 1 (Minimizer) *This regularity is enough to obtain the existence of a minimizer. Indeed, the bound* (1.2) *implies that* $|\dot{y}^a_x(\theta)| \leqslant |f(x, A)|_\infty| + [f]_x T|f(x, A)|_\infty e^{T[f]_x}$*, and the trajectories issued from a given* x *are equilipschitz and equibounded. Let*$(y^a_{x,k})_k$ *be a minimizing sequence for problem* (1.3)*. By Arzelà-Ascoli, up to a subsequence, it converges uniformly towards a Lipschitz trajectory* $(y_x(\theta))_{\theta \in [t,T]}$*. Since* φ *is continuous, the uniform convergence is enough to ensure that* y_x *minimises the cost. It stays to prove that a control* $a^* \in \mathbb{A}_{[t,T]}$ *can be found such that* $y_x = y^{a^*}_x$*. By contradiction,* $\dot{y}_x(\theta) \in f(y_x(\theta), A)$ *for all* $\theta \in [t, T]$*. By the convexity assumption in (A2), we may pick* $a_\theta \in A$ *such that* $\dot{y}_x(\theta) = f(y_x(\theta), a_\theta)$*. Let* $a^*(\cdot)$ *be a mesurable function such that* $\overset{\circ}{\forall}\theta \in [t, T]$*, it holds* $a^*(\theta) = a_\theta$*. For instance, the coordwise upper semicontinuous envelope of* $\theta \to a_\theta$*, which is measurable as a pointwise limit of continuous functions (owing to Baire's theorem on semicontinuous functions). Then* $a^* \in \mathbb{A}_{[t,T]}$*, and* y_x *coïncides with* $y^{a^*}_x$*.*

Finally, we may introduce our main object of study.

Definition 2 – Value function Given a Bolza problem associated with φ, we define $v : \mathbb{R}^n \times \mathbb{R}$ as

$$v(x, t) \coloneqq \min_{a \in \mathbb{A}_{[t,T]}} \{\varphi(y^a_x(T))\}.$$

The intelligence of the Hamilton-Jacobi-Bellman approach is to find an equation satisfied by v. The classical derivation of this PDE is introduced in the next section, in the slightly more general case of state constraints.

1.1.2 State constraints via obstacle problems

The following framework comes from [ABZ13]. We briefly present the general case, and then reduce to front propagation, which will be studied numerically in section (2.2.2).

The general idea The initial model is the Bolza problem with state constraints, whose value function v is defined by

$$\begin{cases} v(x,t) = \inf_{a(\cdot)\in\mathbb{A}_{[t,T]}} \{\varphi(y_x^a(T)) \mid y_x^a(\theta) \in K \ \forall \theta \in [t,T]\}, \\ \dot{y}_x^a(\theta) = f(y_x^a(\theta), a(\theta)), \quad y_x^a(t) = x. \end{cases}$$

The closed set $K \subset \mathbb{R}^n$ models the admissible states. Here, we consider K independant of t to simplify the presentation. Under this formulation, there is no guarantee that an admissible control $a(\cdot)$ exists. (Indeed, if $A = \{1\}$, $f(x,a) = a$ and $K = \mathbb{R}^-$, any trajectory y_x^a for $x > 0$ will stay outside of K.) In this case, we use the classical convention $\inf \emptyset = \infty$ to set $v(x,t) = \infty$. This hints that state-constraints problems do not yield very regular value functions. To circumvent this, an auxiliary problem is introduced using a level-set approach.

First, the set K is described as the level set $\{x \in \mathbb{R}^n \mid g(x) \leqslant 0\}$, where g is a Lipschitz function with constant $[g]$. The condition $y(t) \coloneqq y_x^a(t) \in K$ is equivalent to $g(y(t)) \leqslant 0$, and the admissibility of the trajectory $(y(t))_{t\in[0,T]}$ can be tested with $\max_{t\in[0,T]} g(y(t)) \leqslant 0$.

Secondly, the value function $v(x,t)$ is characterized by its epigraph $\{(\alpha,x,t) \in \mathbb{R}\times\mathbb{R}^n\times[0,T] \mid \alpha \geqslant v(x,t)\}$. The unknown of the auxiliary problem will be a level set function of this epigraph, namely

$$u(\alpha,x,t) \leqslant 0 \quad \iff \quad \alpha \geqslant v(x,t)$$

Let us try to find an expression for u by reformulating the condition $\alpha \geqslant v(x,t)$. For any x such that $v(x,t) \leqslant \alpha < \infty$, the set of admissible trajectories is nonempty. Thus, arguing as in remark (1) since g is Lipschitz-continuous, the infimum defining v is attained. We have

$$\begin{aligned} \{(\alpha,x,t) \mid v(x,t) \leqslant \alpha\} &= \left\{(\alpha,x,t) \mid \min_{a(\cdot)\in\mathbb{A}_{[t,T]}} \left\{\varphi(y_x^a(T)) - \alpha \mid \max_{\theta\in[t,T]} g(y_x^a(\theta)) \leqslant 0\right\} \leqslant 0\right\} \\ &= \left\{(\alpha,x,t) \mid \min_{a(\cdot)\in\mathbb{A}_{[t,T]}} \left\{(\varphi(y_x^a(T)) - \alpha) \bigvee \max_{\theta\in[t,T]} g(y_x^a(\theta))\right\} \leqslant 0\right\}. \end{aligned}$$

Then, a possible choice for u is

$$u(\alpha,x,t) \coloneqq \min_{a(\cdot)\in\mathbb{A}_{[t,T]}} \left\{(\varphi(y_x^a(T)) - \alpha) \bigvee \max_{\theta\in[t,T]} g(y_x^a(\theta))\right\}.$$

Clearly, βu is also an admissible choice for any $\beta > 0$.

Remark 2 (Link between u and v) *The value $v(x,t)$ can be recovered by*

$$v(x,t) = \inf_{\alpha\in\mathbb{R}} \{\alpha \mid u(\alpha,x,t) \leqslant 0\}. \tag{1.4}$$

Indeed, if $u(\alpha,x,t) > 0$ for all $\alpha \in \mathbb{R}$, then $\max_{\theta\in[t,T]} g(y_x^a(\theta)) > 0$ for all $a \in \mathbb{A}_{[t,T]}$ (since the other maximand is linear in α, thus unbounded). There exist no admissible trajectory issued from x, and under the classical convention for the minimum of empty set, $v(x,t) = \infty$. On the other hand, if $u(\alpha,x,t) \leqslant 0$ for a certain $\alpha \in \mathbb{R}$, it immediate to see that $u(\widetilde{\alpha},x,t) \geqslant 0$ for all $\widetilde{\alpha} \geqslant \alpha$. The minimum will be reached when $\alpha = \varphi(y_x^{a^*}(T))$ for an optimal a^*, that satisfies the constraint since $\max_{\theta\in[t,T]} g(y_x^{a^*}(\theta)) \leqslant 0$.

The case of moving fronts The motion of a closed curve, or equivalently, of the region that it delimitates, can be adressed by Hamilton-Jacobi equations. It may be interpretated as a reachability problem, where a target domain $D \coloneqq \{\varphi \leqslant 0\}$ is given, and the aim is to compute all the points from which D can be reached under a certain dynamic (see [BFZ10] for ample details). This backward reachable set problem writes

$$\begin{cases} -\partial_t u(x,t) + \max_{a\in A} \nabla u(x,t) f(x,a) = 0 & (x,t) \in \mathbb{R}^n \times]0,T[\\ u(x,T) = \varphi(x) & x \in \mathbb{R}^n \end{cases}$$

where for each $t \in [0,T]$, the reachable set is given by $\{u(x,t) \leqslant 0\}$. The introduction of a set K of admissible dynamics is interpreted as an "obstacle" $\mathbb{R}^n \setminus K$, that the dynamics cannot cross. State constraints may be treated with the above formulation, but the underlying focus on the 0-level set induces a simplification. Indeed, only the value $\alpha = 0$ is needed. We shall directly consider a level set function u defined as

$$u(x,t) \coloneqq \min_{a(\cdot)\in\mathbb{A}_{[t,T]}} \left\{ \varphi(y_x^a(T)) \bigvee \max_{\theta\in[t,T]} g(y_x^a(\theta)) \right\}.$$

The value $v(x,t)$ can also be recovered by expression (1.4), but since the 0-level sets of v and u coïncide, this is not even necessary to compute reachable sets.

Properties of u An essential advantage of the auxiliary problem is the following.

Lemma 1 – Regularity Suppose that φ, g and $f(\cdot,\cdot)$ are Lipschitz. Then $u(\cdot,\cdot,t)$ is Lipschitz-continuous uniformly over t.

Demonstration

Let a^* be an optimal control, such that $u(\alpha,x,t) = \left(\varphi(y_x^{a^*}(T)) - \alpha\right) \bigvee \max_{\theta\in[t,T]} g(y_x^{a^*}(\theta))$. Let $(\alpha_i, x_i) \in \mathbb{R}\times\mathbb{R}^n$ for $i \in \{1,2\}$. Using $\max(a,b) - \max(c,d) \leqslant \max(a-b, c-d)$, we get

$$\begin{aligned} u(\alpha_1,x_1,t) - u(\alpha_2,x_2,t) &\leqslant \left(\varphi(y_{x_1}^{a^*}(T)) - \varphi(y_{x_2}^{a^*}(T)) + \alpha_1 - \alpha_2\right) \bigvee \left(\max_{\theta\in[t,T]} [g(y_{x_1}^{a^*}(\theta)) - g(y_{x_2}^{a^*}(\theta))] \right) \\ &\leqslant ([\varphi] \vee [g]) \max_{\theta\in[t,T]} \left| y_{x_1}^{a^*}(\theta) - y_{x_2}^{a^*}(\theta) \right| + |\alpha_1 - \alpha_2| \end{aligned}$$

By a Grönwall lemma, it holds that $\left| y_{x_1}^{a^*}(\theta) - y_{x_2}^{a^*}(\theta) \right| \leqslant |x_1 - x_2| \exp(\theta [f]_x)$. Since $\theta \leqslant T$, $u(\alpha_1,x_1,t) - u(\alpha_2,x_2,t) \leqslant ([\varphi] \vee [g]) e^{[f]_x T} |x_1 - x_2| + |\alpha_1 - \alpha_2|$. Switching (α_1,x_1) and (α_2,x_2) gives the result. □

The first step towards an HJ equation is the following tool.

Proposition 1 – Obstacle problem DPP Let $(\alpha,x,t) \in \mathbb{R}\times\mathbb{R}^n\times[0,T[$ and $h \in]0, T-t[$. Then

$$u(\alpha,x,t) = \min_{a(\cdot)\in\mathbb{A}_{[t,t+h]}} \max_{\theta\in[t,t+h]} g(y_x^a(\theta)) \bigvee u(\alpha, y_x^a(t+h), t+h). \tag{1.5}$$

Demonstration

Notice that for any control $a(\cdot) \in \mathbb{A}_{[t,T]}$, the restriction $a_{|[t_1,t_2]}(\cdot)$ on an interval $[t_1,t_2] \subset [t,T]$ belongs to $\mathbb{A}_{[t_1,t_2]}$. On the other hand, the concatenation of measurable functions is measurable, so $a(\cdot) \coloneqq a_1 \mathbb{1}_{[t,s[} + a_2 \mathbb{1}_{[s,T]} \in \mathbb{A}_{[t,T]}$ if $a_1 \in \mathbb{A}_{[t,s]}$ and $a_2 \in \mathbb{A}_{[s,T]}$. This allows us to cut and glue controls freely.
Moreover, for any $s \in]t,T[$, the definition of the trajectories (1.1) implies that $y_x^a(\theta) = y_{y_x^a(s)}^{a(\cdot-s)}(\theta - s)\ \forall \theta \in [s,T]$.
Using the decomposition $a = a_1 \mathbb{1}_{[t,s[} + a_2 \mathbb{1}_{[s,T]}$, the function u rewrites

$$u(\alpha,x,t) = \min_{a_1(\cdot)\in\mathbb{A}_{[t,s]}} \min_{a_2(\cdot)\in\mathbb{A}_{[s,T]}} \left\{ \left(\varphi(y_{y_x^{a_1}(s)}^{a_2}(T)) - \alpha\right) \bigvee \max_{\theta\in[t,s]} g(y_x^{a_1}(\theta)) \bigvee \max_{\theta\in[s,T]} g(y_{y_x^{a_1}(s)}^{a_2}(\theta)) \right\}.$$

Using that $\min_a \max(\alpha(a), \beta) = \max(\min_a \alpha(a), \beta)$ yields

$$\begin{aligned} u(\alpha,x,t) &= \min_{a_1(\cdot)\in\mathbb{A}_{[t,s]}} \max_{\theta\in[t,s]} g(y_x^{a_1}(\theta)) \bigvee \left[\min_{a_2(\cdot)\in\mathbb{A}_{[s,T]}} \left\{ \left(\varphi(y_{y_x^{a_1}(s)}^{a_2}(T)) - \alpha\right) \bigvee \max_{\theta\in[s,T]} g(y_{y_x^{a_1}(s)}^{a_2}(\theta)) \right\} \right] \\ &= \min_{a_1(\cdot)\in\mathbb{A}_{[t,s]}} \max_{\theta\in[t,s]} g(y_x^{a_1}(\theta)) \bigvee u(\alpha, y_x^{a_1}(s), s) \end{aligned}$$

the desired result. □

With this tool, we may derive an equation satisfied by u in the viscosity sense.

Proposition 2 – Obstacle problem HJ equation Suppose (A1) to (A3). Let $\Omega \coloneqq \mathbb{R} \times \mathbb{R}^n \times]0,T[$. The function $u = u(\alpha, x, t)$ is a viscosity solution of

$$\begin{cases} \min(-\partial_t u + \max_{a\in A} \left[-\nabla u \cdot f(x,a)\right], u - g(x)) = 0 & (\alpha, x, t) \in \Omega \\ u(\alpha, x, T) = (\varphi(x) - \alpha) \vee g(x) & (\alpha, x) \in \mathbb{R} \times \mathbb{R}^n \end{cases} \tag{1.6}$$

where $\nabla u = \partial_x u$ denotes the space gradient.

Since the Hamiltonian $H(x,p) \coloneqq \max_{a\in A} -p \cdot f(x,a)$ is convex and Lipschitz-continuous (as a supremum of linear equilipschitz functions), the viscosity solution will even be unique. We refer the reader to [ABZ13], Annex A, for the proof of the corresponding maximum principle in the case of obstacle problems. Here, we restrict to the verification of proposition (2).

Demonstration

The terminal condition is trivially satisfied. Let us verify the two points of the definition.

Subsolution Let $\varphi \in \mathcal{C}^1(\Omega)$ such that $u - \varphi$ reaches a global maximum in $(\alpha_0, x_0, t_0) \in \Omega$ with equality. Let $a(\cdot) \equiv a \in A$ be a constant control, and $h \in]0, T - t_0[$. Then

$$\begin{aligned} &\max_{\theta\in[t_0,t_0+h]} g(y^a_{x_0}(\theta)) \bigvee \varphi(\alpha_0, y^a_{x_0}(t_0+h), t_0+h) - \varphi(\alpha_0, x_0, t_0) \\ \geqslant &\max_{\theta\in[t_0,t_0+h]} g(y^a_{x_0}(\theta)) \bigvee u(\alpha_0, y^a_{x_0}(t_0+h), t_0+h) - u(\alpha_0, x_0, t_0) \geqslant 0. \end{aligned}$$

Using $\max(a,b) - c \geqslant 0 \iff \max(a-c, \frac{b-c}{h}) \geqslant 0$ for all $h > 0$, this yields

$$\left[\max_{\theta\in[t_0,t_0+h]} g(y^a_{x_0}(\theta)) - \varphi(\alpha_0, x_0, t_0)\right] \bigvee \left[\frac{\varphi(\alpha_0, y^a_{x_0}(t_0+h), t_0+h) - \varphi(\alpha_0, x_0, t_0)}{h}\right] \geqslant 0. \tag{1.7}$$

On one hand, the bound (1.2) gives (uniformly over $a(\cdot) \in \mathbb{A}_{[0,T]}$)

$$|y^a_{x_0}(\theta) - x_0| \leqslant \theta |f(x_0, A)|_\infty \exp([f]_x \theta) =: hC \quad \implies \quad \left|\max_{\theta\in[t_0,t_0+h]} g(y^a_{x_0}(\theta)) - g(x_0)\right| \leqslant [g]hC. \tag{1.8}$$

On the other hand, by the regularity of φ, the second maximand goes to $(\nabla\varphi \cdot f(x_0,a) + \partial_t\varphi)(\alpha_0, x_0, t_0)$ (the control a being constant). Then, taking the limit in $h \searrow 0$ in (1.7), we obtain

$$\begin{aligned} [g(x_0) - \varphi(\alpha_0, x_0, t_0)] \vee [(\nabla\varphi \cdot f(x_0,a) + \partial_t\varphi)(\alpha_0, x_0, t_0)] \geqslant 0 \\ \min\left(\varphi(\alpha_0, x_0, t_0) - g(x_0), (\max_{a\in A} -\nabla\varphi \cdot f(x_0,a) - \partial_t\varphi)(\alpha_0, x_0, t_0)\right) \leqslant 0 \end{aligned}$$

and u is a subsolution of (1.6).

Supersolution Let $\varphi \in \mathcal{C}^1(\Omega)$ such that $u - \varphi$ reaches a global minimum in $(\alpha_0, x_0, t_0) \in \Omega$ with equality. Let $\varepsilon > 0$ and $h > 0$, and consider $a^\varepsilon_h \in \mathbb{A}_{[0,T]}$ $\varepsilon h-$optimal for $u(\alpha_0, x_0, t_0)$, i.e. such that

$$\mathcal{J}_{\alpha_0,x_0,t_0}(h, a^\varepsilon_h, u) \coloneqq \max_{\theta\in[t_0,t_0+h]} g(y^{a^\varepsilon_h}_{x_0}(\theta)) \bigvee u(\alpha_0, y^{a^\varepsilon_h}_{x_0}(t_0+h), t_0+h) \leqslant u(\alpha_0, x_0, t_0) + \varepsilon h.$$

Then

$$\mathcal{J}_{\alpha_0,x_0,t_0}(h, a^\varepsilon_h, \varphi) - \varphi(\alpha_0, x_0, t_0) \leqslant \mathcal{J}_{\alpha_0,x_0,t_0}(h, a^\varepsilon_h, u) - u(\alpha_0, x_0, t_0) \leqslant \varepsilon h.$$

Using $\max(a,b) \leqslant \varepsilon h \implies a \leqslant \varepsilon h$ and $\frac{b}{h} \leqslant \varepsilon$, and the estimate (1.8), we deduce that

$$g(x_0) - \varphi(\alpha_0, x_0, t_0) - [g]hC \leqslant \varepsilon h \quad \text{and} \quad \frac{\varphi(\alpha_0, y_{x_0}^{a_h^\varepsilon}(t_0+h), t_0+h) - \varphi(\alpha_0, x_0, t_0)}{h} \leqslant \varepsilon. \tag{1.9}$$

Notice that

$$\begin{aligned} y_{x_0}^{a_h^\varepsilon}(t_0+h) - x_0 &= \int_{t_0}^{t_0+h} f(y_{x_0}^{a_h^\varepsilon}(\theta), a_h^\varepsilon(\theta))d\theta \\ &= \int_{t_0}^{t_0+h} [f]_x |y_{x_0}^{a_h^\varepsilon}(\theta) - x_0| \kappa(\theta) d\theta + \frac{h}{h}\int_{t_0}^{t_0+h} f(x_0, a_h^\varepsilon(\theta))d\theta \quad \text{for a certain } \kappa(\cdot) \in [-1,1], \\ &= [f]_x h^2 C\widetilde{\kappa} + hf(x_0, \overline{a}_h^\varepsilon) \quad \text{for a certain } \widetilde{\kappa} \in [-1,1], \text{ and } \overline{a}_h^\varepsilon \in A \text{ (since } f(x_0, A) \text{ is convex).} \end{aligned}$$

Since A is compact, the sequence $(\overline{a}_h^\varepsilon)_h$ admits a cluster point $\overline{a}^\varepsilon \in A$ when $h \searrow 0$. Up to a subsequence, we have $y_{x_0}^{a_h^\varepsilon}(t_0+h) = x_0 + hf(x_0, \overline{a}^\varepsilon) + o(h)$, and

$$\frac{\varphi(\alpha_0, y_{x_0}^{a_h^\varepsilon}(t_0+h), t_0+h) - \varphi(\alpha_0, x_0, t_0)}{h} = \nabla\varphi(\alpha_0, x_0, t_0) \cdot f(x_0, \overline{a}^\varepsilon) + \partial_t\varphi(\alpha_0, x_0, t_0) + O(h)$$

Plugging this in (1.9), and letting $h \searrow 0$, we get (for $\varphi = \varphi(\alpha_0, x_0, t_0)$)

$$\begin{cases} g(x_0) - \varphi \leqslant 0 \\ \partial_x\varphi \cdot f(x_0, \overline{a}^\varepsilon) + \partial_t\varphi \leqslant \varepsilon \end{cases} \iff \min\left(\varphi - g(x_0), -\partial_x\varphi \cdot f(x_0, \overline{a}^\varepsilon) - \partial_t\varphi + \varepsilon\right) \geqslant 0.$$

This implies in particular that $\min\left(\varphi - g(x_0), -\partial_t\varphi + \max_{a\in A}\left[-\partial_x\varphi \cdot f(x_0,a)\right] + \varepsilon\right) \geqslant 0$ for all $\varepsilon > 0$. Hence u is a supersolution of (1.6). □

The derivative of u with respect to α is not directly involved in the equation. Actually, the DPP shows that each α gives birth to an independant problem. This comes from the fact that the running cost l is identically zero, whereas a nontrivial l introduces a coupling between different values of α.

Example To conclude this section, let us consider a particular case in dimension $n = 1$. Suppose we want to minimize $|y_x^a(T)|$, with $\dot{y}_x^a \in [-1,1]$ and a constraint $|y_x^a| \geqslant \gamma$. We take

$$A = [-1,1], \quad f(x,a) = a, \quad \varphi(x) = |x|, \quad g(x) = \gamma - |x| \text{ for } \gamma > 0$$

The original Bolza problem is to find $v = v(x,t)$ such that

$$v(x,t) = \inf_{a\in\mathbb{A}_{[t,T]}} \varphi(y_x^a(T)) = \inf_{a\in\mathbb{A}_{[t,T]}} |y_x^a(T)|, \quad |y_x^a(\theta)| \geqslant \gamma \ \forall\theta \in [t,T].$$

Notice that whenever $|x| < \gamma$, there exist no admissible trajectory, and $v(x,t) = \infty$. Otherwise, it is clear that the trajectories will converge towards the boundary $\{|x| = \gamma\}$ at maximum speed, i.e.

$$a^*(t) = -\frac{x}{|x|}\mathbb{1}_{\{t\leqslant|x|-\gamma\}}, \quad y_x^{a^*}(t) = x - \min(t, |x|-\gamma)\frac{x}{|x|}, \quad \text{and } v(x,t) = \varphi(y_x^{a^*}(T)) = \max(|x| + t - T, \gamma).$$

Figure (1.1) illustrates the case $\gamma = T = 1$. Now, we consider the formulation of [ABZ13]. The auxiliary problem writes

$$u(\alpha, x, t) = \min_{a\in\mathbb{A}_{[t,T]}} \max_{\theta\in[t,T]} g(y_x^a(\theta)) \bigvee \left[\varphi(y_x^a(T)) - \alpha\right].$$

The associated HJB equation is

$$\min(u - g(x), -\partial_t u + |\nabla u|) = 0, \quad u(\alpha, x, T) \coloneqq (|x| - \alpha) \vee g(x). \tag{1.10}$$

We first consider the obstacle-free solution of $-\partial_t u_{of} + |\nabla u_{of}| = 0$, with terminal condition $u_{of}(\alpha, x, T) := (|x| - \alpha) \vee g(x)$. Using Lax-Oleinik formula (that will be introduced in section (1.2.2)), we get that

$$u_{of}(\alpha, x, t) = \max\left((-\alpha) \vee \frac{\gamma - \alpha}{2}, (|x| - (T-t))_+ - \alpha, \gamma - |x| - (T-t)\right).$$

(The value $(-\alpha) \vee \frac{\gamma-\alpha}{2}$ is the minimum of the terminal condition.) Using the DPP, we have that

$$u(\alpha, x, t) = \min_{a \in \mathbb{A}_{[t,T]}} \max_{\theta \in [t,T]} g(y_x^a(\theta)) \bigvee [\varphi(y_x^a(T)) - \alpha] \geqslant \min_{a \in \mathbb{A}_{[t,T]}} g(y_x^a(T)) \bigvee [\varphi(y_x^a(T)) - \alpha] = u_{of}(x, t)$$

and rewriting (1.10) as $u = \max(g(x), u + \partial_t u - |\nabla u|)$, we also have $u(\alpha, x, t) \geqslant g(x)$ for all (α, x, t). In our case, we actually have the equality

$$u(\alpha, x, t) = g(x) \vee u_{of}(\alpha, x, t) = \max\left((-\alpha) \vee \frac{\gamma - \alpha}{2}, (|x| - (T-t))_+ - \alpha, \gamma - |x|\right).$$

Indeed, for any (α, x), we can find a control $a \in \mathbb{A}_{[t,T]}$ such that

$$\max_{\theta \in [t,T]} g(y_x^a(\theta)) \bigvee [\varphi(y_x^a(T)) - \alpha] = g(x) \vee u_{of}(\alpha, x, t), \tag{1.11}$$

and this gives the inequality $u \leqslant g \vee u_{of}$. The curious reader can check that $a(\theta) := -\frac{x}{|x|} \mathbb{1}_{\{\theta \leqslant |x| - (\frac{\gamma+\alpha}{2})_+\}}$ satisfies (1.11). The representation of u in figure (1.1) may help.

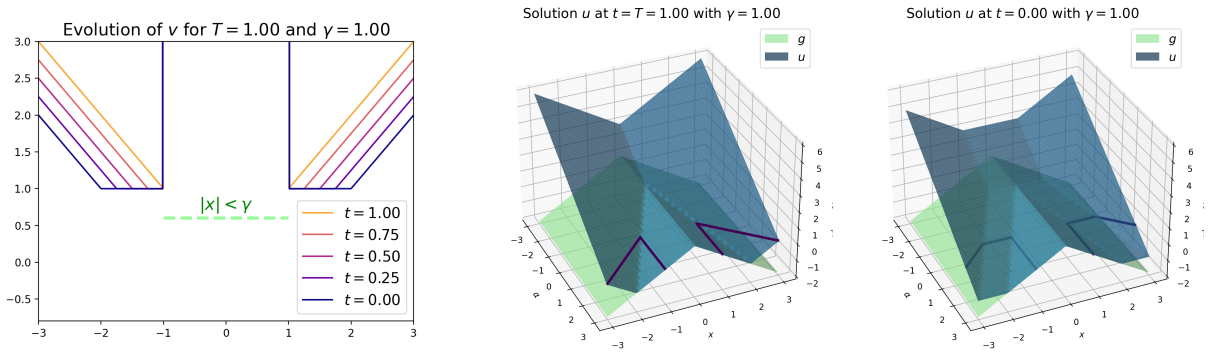


Figure 1.1: Different formulations of state-constrained problems

Left: graph of the original value function v. Right: terminal condition and initial value of the auxiliary value function u. The boundary of the level set $\{u \leqslant 0\}$ coïncides with the graph of v.

1.2 Elements of HJ theory

1.2.1 Fenchel transform

Let $H : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ be proper, lower semi-continuous and convex. The latter property yields that H majorizes all its tangent hyperplanes : for any $p \in \mathbb{R}^n$, the convex subdifferential $\partial H(p)$ is nonempty, and for $q \in \partial H(p)$, $H(r) \geqslant H(p) + \langle q, r - p \rangle$ for all $r \in \mathbb{R}^n$. On the other hand, we have

Lemma 2 – Tangent plane set Let $q \in \mathbb{R}^n$, and denote by T_q the set of points x such that $q \in \partial H(x)$. Then T_q is convex (and in particular, connected).

Demonstration

Let $x, y \in T_q$, and parametrize the segment $[x, y]$ by $z = \lambda x + (1-\lambda)y$, $\lambda \in [0,1]$. Then, by definition of the convex subdifferential,

$$\begin{aligned} H(z) &\geqslant H(x) + \langle q, z - x\rangle = H(x) + \langle q, (1-\lambda)(y-x)\rangle \\ H(z) &\geqslant H(y) + \langle q, z - y\rangle = H(y) + \langle q, \lambda(x-y)\rangle \end{aligned}$$

By convexity, we also have

$$H(z) \leqslant \lambda H(x) + (1-\lambda)H(y) \leqslant \lambda\left(H(z) - \langle q, (1-\lambda)(y-x)\rangle\right) + (1-\lambda)\left(H(z) - \langle q, \lambda(x-y)\rangle\right) = H(z)$$

and then $H(\lambda x + (1-\lambda)y) = \lambda H(x) + (1-\lambda)H(y)$. Let now z denote any point in $\mathbb{R}^n$: we have

$$\begin{aligned} H(z) - H(\lambda x + (1-\lambda)y) &= H(z) - \lambda H(x) - (1-\lambda)H(y) \\ &= \lambda\left[H(z) - H(x)\right] + (1-\lambda)\left[H(z) - H(y)\right] \\ &\geqslant \lambda\langle q, z - x\rangle + (1-\lambda)\langle q, z - y\rangle \\ &\geqslant \langle q, z - (\lambda x + (1-\lambda)y)\rangle \end{aligned}$$

and thus $q \in \partial H(\lambda x + (1-\lambda)y)$, and T_q is convex. □

The previous lemma geometrically says that at each hyperplane direction q, there correspond at most a convex set of points where H coïncides with $\langle q, x\rangle + \alpha$ for some α. The Fenchel transform can be interpreted as the limit object of this remark : it gives H as a pointwise maximum of hyperplanes, with the help of a function H^* giving the value of α associated to each q.

Definition 3 – Fenchel transform (or convex dual) Suppose $H : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is lower semi-continuous (lsc) and proper. Its Legendre-Fenchel transform $H^* : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is defined as

$$H^*(q) \coloneqq \sup_{p \in \mathbb{R}^n} \left(\langle q, p\rangle - H(p)\right).$$

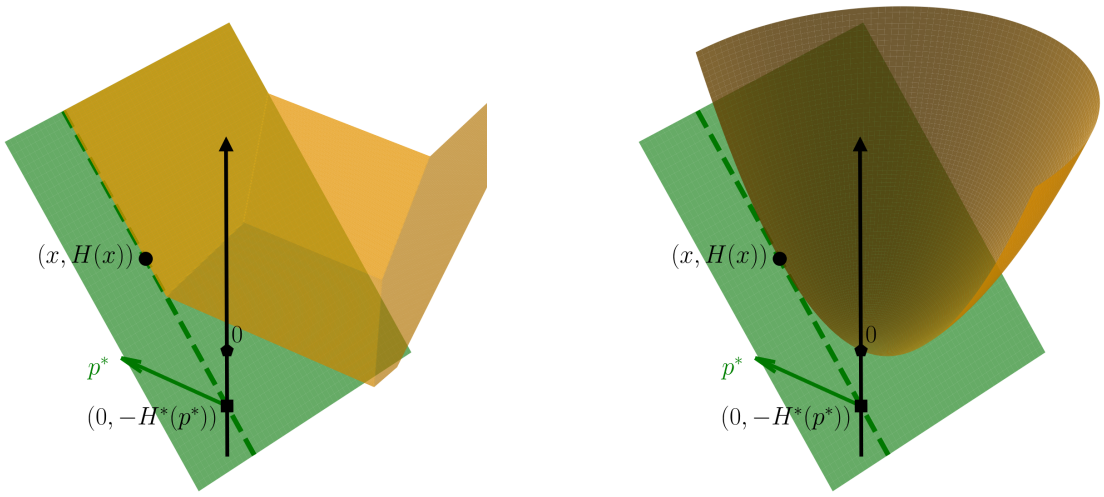


Figure 1.2: Illustration of the Fenchel transform.

Left: a function H (in yellow) and the plan defining its Fenchel transform (in green). The value $H^*(p^*)$ is computed so that $H(x)$ coïncides with $\langle p^*, x\rangle - H^*(p^*)$. In this case, H is locally equal to a plan $q(x)$, and H^* is equal to $-q(0)$. Right: the limit case when H becomes smooth.

Before going further, let us give some canonical examples (all computations are left to the reader). If $H(p) = p \cdot b$, then $H^*(q) = \delta_b(q)$ the convex indicator of the singleton $\{q\}$, that is, $H^*(q) = 0$ is $q = b$, and $H^*(q) = \infty$ otherwise. In particular, if $H \equiv 0$, its Legendre transform is equal to ∞ everywhere but in 0. If $H(p) = \|p\|$, then $H^*(q) = \delta_{\overline{\mathscr{B}(0,1)}}(q)$, the indicator of the unit ball. If $H(p)$ goes to $-\infty$ superlinearly when $\|p\|$ grows, than $H^* \equiv \infty$. Interestingly, $H(p) \coloneqq \frac{1}{2}\|p\|^2$ yields $H^* = H$. In general, if ∇H is invertible and H^* is attained, the Euler condition gives $H^*(x) = \langle x, \nabla H^{-1}(x)\rangle - H(x)$.

Proposition 3 – Properties of the Legendre transform Suppose that H is convex, lsc and proper. Then H^* is convex, lower semi-continuous and proper, and

$$H(p) = \sup_{q\in\mathbb{R}^n} \left(\langle q, p\rangle - H^*(q)\right).$$

Demonstration

The author wish to thank Peter R. Wolenski for his help on the "properness" of H^*.

H^* is proper Let $(x, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ such that $H(x) \leqslant \alpha$: since H is proper, such a couple can be found (the epigraph is nonempty). By convexity of H, epi(H) is convex: by lower semi-continuity, it is closed. Let $z < H(x)$, and denote by $(\overline{x}, \overline{z}) \in \mathbb{R}^d \times \mathbb{R}$ the (unique) orthogonal projection of the point (x, z) on epi(H). By definition, for any $(y, \beta) \in$epi(H),

$$\left\langle \begin{pmatrix} x \\ z \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{z} \end{pmatrix}, \begin{pmatrix} y \\ \beta \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{z} \end{pmatrix} \right\rangle \leqslant 0 \tag{1.12}$$

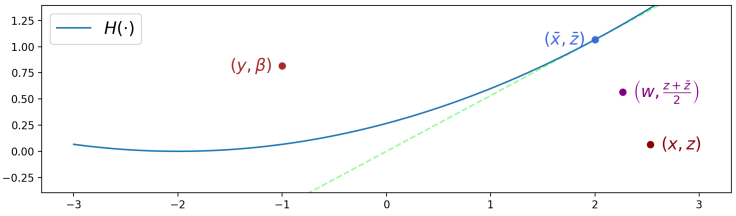
In particular, $(x, H(x))$ belongs to epi(H), and we deduce that $\|x - \overline{x}\|^2 \leqslant \langle \overline{z} - z, H(z) - \overline{z}\rangle$. Since (x, z) does not belong to epi(H), $x - \overline{x} \neq 0$, and we get $\overline{z} - z \neq 0$. Suppose $z > \overline{z}$. Define $w = x + (x - \overline{x})$, so that $x = \frac{w+\overline{x}}{2}$, and $z < H(x) \leqslant \frac{H(w)+H(\overline{x})}{2} \leqslant \frac{H(w)+\overline{z}}{2}$. Then $0 < z - \overline{z} \leqslant \frac{1}{2}\left(H(w) - \overline{z}\right)$, and we have

$$\left\langle \begin{pmatrix} x \\ z \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{z} \end{pmatrix}, \begin{pmatrix} x + (x - \overline{x}) \\ H(w) \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{z} \end{pmatrix} \right\rangle \leqslant 0 \quad \Longrightarrow \quad \langle z - \overline{z}, H(w) - \overline{z}\rangle \leqslant -2\|x - \overline{x}\|^2 < 0$$

so $z - \overline{z} < 0$, which is absurd. Then $z < \overline{z}$. Coming back to (1.12), for any $(ya) \in$ epi (H), we have

$$\langle x - \overline{x}, y\rangle + \langle z - \overline{z}, H(y)\rangle \leqslant \langle x - \overline{x}, y\rangle + \langle z - \overline{z}, \beta\rangle \leqslant \langle x - \overline{x}, \overline{x}\rangle + \langle z - \overline{z}, \overline{z}\rangle < \infty.$$

We may divide by $\overline{z} - z > 0$, and pass to the supremum on $y \in \mathbb{R}^n$ to obtain $H^*\left(\frac{x-\overline{x}}{\overline{z}-z}\right) < \infty$. Then H^* is proper.



Other properties Let $(y^\nu, \alpha^\nu)_\nu \subset$epi$(H^*)$ be a sequence converging towards (y, α). Then for all $x \in \mathbb{R}^n$,

$$\langle x, y^\nu\rangle - H(x) \leqslant H^*(y^\nu) \leqslant \alpha^\nu$$

Since the scalar product is continuous, we may pass to the limit $\nu \to \infty$ and obtain $\langle x, y\rangle - H(x) \leqslant \alpha$ for all $x \in \mathbb{R}^n$, and thus $H^*(y) \leqslant \alpha$. Then $(y, \alpha) \in$epi(H^*), which is closed, and H^* is lower semi-continuous.
The convexity of H^* is direct from that of the supremum. Finally, we have

$$(H^*)^*(p) = \sup_{q\in\mathbb{R}^n} \left(\langle q, p\rangle - H^*(q)\right) = \sup_{q\in\mathbb{R}^n} \left(\langle q, p\rangle - \sup_{r\in\mathbb{R}^n} \left(\langle q, r\rangle - H(r)\right)\right)$$

$$= \sup_{q\in\mathbb{R}^n} \inf_{r\in\mathbb{R}^n} \left(\langle q, p-r\rangle + H(r)\right)$$
$$\leqslant \inf_{r\in\mathbb{R}^n} \sup_{q\in\mathbb{R}^n} \left(\langle q, p-r\rangle + H(r)\right) \leqslant H(p) \quad (\text{for } r=p)$$

On the other hand, let $\overline{q} \in \partial H(p)$ the convex sub-differential. Then, by convexity,

$$H(r) \geqslant H(p) + \langle r-p, \overline{q}\rangle \quad \Longrightarrow \quad \langle \overline{q}, p\rangle - (\langle \overline{q}, r\rangle - H(r)) = \langle \overline{q}, p-r\rangle + H(r) \geqslant H(p) \quad \forall r \in \mathbb{R}^n$$

Let $\varepsilon > 0$, and r^ε be such that $\langle \overline{q}, r^\varepsilon\rangle - H(r^\varepsilon) \geqslant H^*(\overline{q}) - \varepsilon$. Then

$$(H^*)^*(p) + \varepsilon \geqslant \langle \overline{q}, p\rangle - H^*(\overline{q}) + \varepsilon \geqslant \langle \overline{q}, p\rangle - (\langle \overline{q}, r^\varepsilon\rangle - H(r^\varepsilon)) \geqslant H(p)$$

Letting $\varepsilon \searrow 0$, we conclude by double inequality that $H(p) = (H^*)^*(p)$. □

1.2.2 Lax-Oleinik representation

This section relies on the notes of Pierre-Louis Lions [Lio82]. Let $\Omega \subset \mathbb{R}^n$ be open. We focus on

$$\begin{cases} H(\nabla u(x)) = \nu(x) & \text{in } \Omega \\ u(x) = u_b(x) & \text{on } \partial\Omega \end{cases} \tag{1.13}$$

The Lax-Oleinik representation is akin to spectral analysis of linear equations. The analogy is detailed and justified in max-plus algebra, where linearity is replaced by sup-linearity - *provided the Hamiltonian is convex*. We do not dive into max-plus, but the interested reader may enjoy [McE06]. We first precise the assumptions on the data, and build a formal representation of the solution, and then justify it with viscosity arguments.

Framework

In the sequel, we consider the following properties.

(A4) **Hamiltonian structure** $H(\cdot)$ is supposed convex, continuous and proper.

In particular, H is lsc, and H^* enjoys all the properties listed in proposition (3).

(A5) **Source term structure** ν is supposed proper and lipschitz with constant $[\nu]$. If Ω is unbounded, then ν is supposed bounded with constant $\|\nu\|_\infty$. Moreover, $\nu \geqslant \inf_{p\in\mathbb{R}^n} H(p)$.

Notice that we may always shift the source term ν and the Hamiltonian to obtain $\nu \leqslant 0$. This has a "physical" meaning if ν is interpreted as a weight on the space, with $\nu \equiv \nu_0 > 0$ corresponding to the euclidian geometry.

Consider now a solution u of (1.13). Let $x \in \Omega$ and $y \in \overline{\Omega}$, and suppose that we can formally write

$$u(x) - u(y) = \int_0^T \left\langle \nabla u(\xi(t)), \dot{\xi}(t)\right\rangle dt, \quad \text{for } T \in \mathbb{R}^+ \text{ and a path } \xi \in W^{1,1}([0,T], \overline{\Omega}) \text{ with } \xi(0) = y,\ \xi(T) = x.$$

The genius of Lax-Oleinik analysis is to introduce the convex dual by writing

$$\begin{aligned} u(x) - u(y) &= \int_0^T \left\langle \nabla u(\xi(t)), \dot{\xi}(t)\right\rangle - H^*\left(\dot{\xi}(t)\right) + H^*\left(\dot{\xi}(t)\right) dt \\ &\leqslant \int_0^T H\left(\nabla u(\xi(t))\right) + H^*\left(\dot{\xi}(t)\right) dt = \int_0^T \nu\left(\xi(t)\right) + H^*\left(\dot{\xi}(t)\right) dt \end{aligned}$$

This development is only formal for several reasons : first, we did not assume any regularity over u or ξ, and secondly, the convex dual may be unbounded, thus the right hand side infinite. Anyway, we use the following definition.

Definition 4 – Candidate solution For better readability, we define the set of trajectories $\mathcal{C}_T$ and the optical length $\mathcal{L}$ as

$$\mathcal{C}_T(y,x) = \left\{\xi \in \mathcal{C}([0,T],\overline{\Omega}) \mid \xi(0) = y,\ \xi(T) = x\right\}, \quad J(x,p) \coloneqq \nu(x) + H^*(p)$$
$$\mathcal{L}(y,x) = \inf_{T,\xi}\left\{\int_0^T J\left(\xi(t),\dot{\xi}(t)\right)dt \mid T \in \mathbb{R}^+,\ \xi \in \mathcal{C}_T(y,x)\right\}$$

We define a candidate function $u : \overline{\Omega} \mapsto \mathbb{R}$ by

$$u(x) \coloneqq \inf_{y\in\partial\Omega}\{u_b(y) + \mathcal{L}(y,x)\}$$

Let us give a few examples to fix the ideas. If $H(p) = \|p\|$, we have $H^*(q) = \delta_{\overline{\mathscr{B}(0,1)}}(q)$, and the functional $\mathcal{L}$ reduces to

$$\mathcal{L}(y,x) \coloneqq \inf_{T,\xi}\left\{\int_0^T \nu(\xi(t))dt \mid |\dot{\xi}(t)| \leqslant 1, \xi(0) = y,\ \xi(T) = x\right\}.$$

In the particular case $\nu \equiv \nu_0 > 0$, this becomes a minimal time problem over $1-$Lipschitz trajectories linking y to x, and the minimum is attained for the straight line, with $T = \|x-y\|$. Then the candidate solution is

$$u(x) = \inf_{y\in\partial\Omega}\{u_b(y) + \nu_0\|x-y\|\}.$$

In the case $H(p) = p\cdot b$, with $b \in \mathbb{R}^b$ constant, we have $H^*(q) = \delta_b(q)$, and the only allowed $\dot{\xi}$ is b. Then, the candidate solution is a minimum over the half-line $x + b\mathbb{R}^-$.

So far, there is no reason to find a boundary point on this half-line. More generally, to ensure that u is finite-valued, we need to have $\partial\Omega$ "well-placed" with respect to the characteristic directions. Since this is a geometric property that has no chance to be deduced from the equation, we assume it.

(A6) Boundary placement The domain Ω is such that for all x, there exists $y \in \partial\Omega$ with $\mathcal{L}(y,x) < \infty$.

With this assumption, u is finite, and in the case $\nu \equiv \nu_0 > 0$, the candidate solution is given by

$$u(x) = u_b(y) + \nu_0\|x-y\|, \quad \|x-y\| = \min\left\{\|x-z\| \mid x - z \in b\mathbb{R}^+,\ z \in \partial\Omega\right\}.$$

Our next concern is the satisfaction of the boundary condition: do we have $u(x) = u_b(x)$ on $\partial\Omega$? This may not be trivial: in dimension 2, consider

$$\nu \equiv 1, \quad H(x,p) = \|p\|, \quad \Omega = \mathscr{B}(0,1), \quad u_b(x) = \alpha x_1$$

Then $\mathcal{L}(y,x) = \|y-x\|$, and

$$u(x) = \inf_{y\in\partial\mathscr{B}(0,1)}\{u_b(y) + \mathcal{L}(y,x)\} = \inf_{y\in\partial\mathscr{B}(0,1)}\{\alpha y_1 + \|y-x\|\}$$

and in $x = (0,1)^t$, one can verify that $u_b(x) = 0$, but (choosing $y = (1,0)^t$) $u(x) \leqslant \sqrt{2} + \alpha < 0$ for sufficiently small α. The boundary value oscillates fast enough to create discontinuities in u. To prevent this, we enforce the

(A7) Boundary value structure The boundary condition u_b must satisfy $u_b(x) \leqslant \mathcal{L}(x,y) + u_b(y)$ for all $x, y \in \partial\Omega$.

In particular, if $\mathcal{L}(y,x) = \|y-x\|$, (A7) means that u_b is 1-Lipschitz. If $\eta \equiv 0$ and $\min H^* = 0$, the optical length $\mathcal{L}(y,x)$ vanishes, and the boundary condition is chosen constant.

(Another) dynamic programming approach

In order to prove that u is indeed a viscosity solution, we make use of the following dynamical programming principle (DPP).

Proposition 4 – DPP Let $x \in \Omega$, and $\omega \subset \Omega$ be an open neighbourhood of x. Then

$$u(x) = \inf_{z \in \partial\omega} \{\mathcal{L}(z,x) + u(z)\}$$

The proof highlights the role of $\mathcal{L}(y,x)$ as a semi-distance on the set $\overline{\Omega}$ (eventually weighted by ν). Indeed, we have the following lemma.

Lemma 3 – Triangular inequality Let $x, y, z \in \overline{\Omega}$. Then $\mathcal{L}(y,x) \leqslant \mathcal{L}(y,z) + \mathcal{L}(z,x)$.

Demonstration

For all $z \in \overline{\Omega}$, and $s \in [0,T]$, the path $\xi(t) = \xi_1(t)\mathbb{I}_{[0,s]} + \xi_2(s+t)\mathbb{I}_{[s,T]}$ belongs to $\mathcal{C}_T(y,x)$ if $\xi_1 \in \mathcal{C}^s(y,z)$ and $\xi_2 \in \mathcal{C}^{T-s}(z,x)$. Consider ξ_1 $\varepsilon-$optimal for $\mathcal{L}(y,z)$ and ξ_2 $\varepsilon-$optimal for $\mathcal{L}(z,x)$: then

$$\begin{aligned}
\mathcal{L}(y,x) &= \inf_{T,\xi} \left\{ \int_0^T J(\xi(t), \dot{\xi}(t)) dt \mid T \in \mathbb{R}^+,\ \xi \in \mathcal{C}_T(y,x) \right\} \\
&\leqslant \int_0^s J(\xi_1(t), \dot{\xi}_1(t)) dt + \int_0^{T-s} J(\xi_2(t), \dot{\xi}_2(t)) dt \\
&\leqslant \mathcal{L}(y,z) + \varepsilon + \mathcal{L}(z,x) + \varepsilon
\end{aligned}$$

and taking $\varepsilon \searrow 0$, we obtain the desired result. □

Demonstration of the DPP

The first inequality comes from the triangular inequality : let $z \in \partial\omega$. Then

$$u(x) = \inf_{y \in \partial\Omega} \{\mathcal{L}(y,x) + u_b(y)\} \leqslant \inf_{y \in \partial\Omega} \{\mathcal{L}(y,z) + \mathcal{L}(z,x) + u_b(y)\}.$$

Taking the infimum on $z \in \partial\omega$, we find

$$u(x) \leqslant \inf_{z \in \partial\omega} \left\{ \mathcal{L}(z,x) + \inf_{y \in \partial\Omega} \{\mathcal{L}(y,z) + u_b(y)\} \right\} = \inf_{z \in \partial\omega} \{\mathcal{L}(z,x) + u(z)\}.$$

On the other hand, consider $y^\varepsilon \in \partial\Omega$ an $\varepsilon-$optimal point for u, and $(T^{\varepsilon,\eta}, \xi^{\varepsilon,\eta})$ an $\eta-$optimal couple for $\mathcal{L}(y^\varepsilon, x)$:

$$u(x) \geqslant \int_0^{T^{\varepsilon,\eta}} J\left(\xi^{\varepsilon,\eta}(t), \dot{\xi}^{\varepsilon,\eta}(t)\right) dt - \varepsilon - \eta, \quad \xi^{\varepsilon,\eta} \in \mathcal{C}^{T^{\varepsilon,\eta}}(y^\varepsilon, x).$$

By continuity of $\xi^{\varepsilon,\eta}$, there exists $s \in [0, T^{\varepsilon,\eta}]$ such that $\xi^{\varepsilon,\eta}(s) \in \partial\omega$. Then

$$\begin{aligned}
u(x) &\geqslant u_b(y^\varepsilon) + \int_0^s J\left(\xi^{\varepsilon,\eta}(t), \dot{\xi}^{\varepsilon,\eta}(t)\right) dt + \int_0^{T^{\varepsilon,\eta}-s} J\left(\xi^{\varepsilon,\eta}(s+t), \dot{\xi}^{\varepsilon,\eta}(s+t)\right) dt - \varepsilon - \eta \\
&\geqslant u(\xi^{\varepsilon,\eta}(s)) + \mathcal{L}(\xi^{\varepsilon,\eta}(s), x) - \varepsilon - \eta \\
&\geqslant \inf_{z \in \partial\omega} u(z) + \mathcal{L}(z,x) - \varepsilon - \eta
\end{aligned}$$

and taking $\eta \searrow 0$, then $\varepsilon \searrow 0$, we find the desired result. □

We now turn to the main result of this section.

Proposition 5 The candidate solution $u(x) \coloneqq \inf_{y\in\partial\Omega}\{u_b(y) + \mathcal{L}(y,x)\}$ is a viscosity solution of (1.13).

Demonstration

We proceed in three steps. First, the DPP yields that u is a sub-solution. Then, we show that it is a pointwise maximum of all Lipschitz subsolutions, using essentially the same arguments as in the formal development. We conclude by contradiction.

u is a sub-solution Let $\varphi \in \mathcal{C}^1(\overline{\Omega})$ such that $u - \varphi$ attains a maximum at $x \in \Omega$, with $u(x) = \varphi(x)$. Let $b \in \text{ dom } H^*$, and choose $T = \inf\{t > 0 \mid x - bt \in \partial\Omega\}$. Define $\xi(t) = (x - bT)\frac{T-t}{T} + \frac{t}{T}x = x + (t-T)b$ the straight trajectory of constant derivative b. Then for all $h > 0$,

$$\varphi(x-hb) - \varphi(x) \geqslant u(x-hb) - u(x) \geqslant -\mathcal{L}(x-hb, x) \geqslant -\int_{T-h}^{T} \nu(\xi(t)) + H^*(b)dt$$

Thus

$$\frac{\varphi(x-hb)-\varphi(x)}{h} + H^*(b) \geqslant -\frac{1}{h}\int_{T-h}^{T} \nu(x+(t-T)b)dt$$

By the regularity of φ, the left hand side converges to $\langle\nabla\varphi(x), -b\rangle + H^*(b)$ as $h \searrow 0$. Since ν is lipschitz, the right hand side converges towards $-\nu(x)$. Multiplicating by -1, we find

$$\langle\nabla\varphi(x), b\rangle - H^*(b) \leqslant \nu(x) \quad \forall b \in \text{ dom } H^* \quad \Longrightarrow \quad \sup_{b\in \text{ dom } H^*} \langle\nabla\varphi(x), b\rangle - H^*(b) = H(\nabla\varphi(x)) \leqslant \nu(x)$$

and u is a sub-solution.

u is the pointwise maximum of Lipschitz sub-solutions Let v be a Lipschitz-continuous sub-solution of (1.13). Then, by Rademacher theorem, it is almost everywhere differentiable, and is particular, it is absolutely continuous. Let $x \in \Omega$, and a triple $(y, T, \xi) \in \partial\Omega \times \mathbb{R}_+^* \times \mathcal{C}_T(y,x)$ admissible, i.e such that $\int_0^T H^*(\dot{\xi}(t))dt < \infty$. Such a triple exists since H^* is proper, and $\partial\Omega$ "well-placed" by (A6). We have

$$\begin{aligned} v(x) - v(y) &= \int_0^T p(\xi(t))\cdot\dot{\xi}(t)dt && \text{for } p(z) \in D^+v(z) \\ &\leqslant \int_0^T p(\xi(t))\cdot\dot{\xi}(t) - H^*(\dot{\xi}(t)) + H^*(\dot{\xi}(t))dt \\ &\leqslant \int_0^T H(p(\xi(t))) + H^*(\dot{\xi}(t))dt \leqslant \int_0^T \nu(\xi(t)) + H^*(\dot{\xi}(t))dt \end{aligned}$$

since v is a sub-solution. Notice that the inequality is trivially satisfied if (y, T, ξ) sends the right hand side to ∞. Taking the infimum over such (T, ξ), then the supremum over y, yields $v(x) \leqslant \inf_{y\in\partial\Omega} u_b(y) + \mathcal{L}(y,x) = u(x)$.

u is super-solution We argue by contradiction. Suppose that u is not a super-solution of (1.13). Then there exists $x \in \Omega$ and $p \in D^-u(x)$ such that $H(p) < \nu(x)$, with $D^-u(x)$ the subgradient of u. Let $Q(y) \coloneqq u(x) + \langle p, y-x\rangle + \delta - \gamma|y-x|^2$. By continuity of H, we have

$$\nabla Q(y) = p - \gamma(y-x), \quad H\left(\nabla Q(y)\right) - n(y) = H\left(p - \gamma(y-x)\right) - \nu(y) < 0$$

for $y \in V_r$ a neighbourhood of x, with γ and r small enough. By definition, Q is a sub-solution in V_r. Define $Z(y) = \begin{cases} \max(u(y), Q(y)) & y \in V_r \\ u(y) & \text{otherwise} \end{cases}$. We first justify that for δ and r small enough, the maximum does not induce any discontinuity : indeed,

$$u(y) \geqslant u(x) + \langle p, y-x\rangle + o\left(|y-x|^2\right)$$

$$
\begin{aligned}
&\geqslant Q(y) - \delta + \gamma\,|x-y|^2 + o\left(|y-x|^2\right) \\
&\geqslant Q(y) - \delta + \gamma\frac{r^2}{4} + o\left(|y-x|^2\right) && \text{for } y \in V_r \setminus V_{r/2} \\
&\geqslant Q(y) - \delta + \gamma\frac{r^2}{8} && \text{for } r \text{ small enough} \\
&\geqslant Q(y) + \gamma\frac{r^2}{16} && \text{for } \delta \text{ small enough}
\end{aligned}
$$

and $u(y) - Q(y) > 0$ on $V_r \setminus V_{r/2}$, yielding $Z = u$ on this set. Z is then a subsolution of (1.13), and it is Lipschitz by construction. Since $\delta > 0$, there exists a neighbourhood of x where

$$
Z(y) - u(y) \geqslant u(x) + \langle p, y-x\rangle + \delta + o(|y-x|^2) - u(x) - [u]|y-x| \geqslant \delta - C|y-x| > 0
$$

which contradicts the maximality of u. Then u is a viscosity solution of (1.13). □

Far more general results are presented in [Lio82], concerning Hamiltonians that may depend on x and u. Since then, the theory of second-order equations has been rigorously developped, and the present section is by no mean a correct introduction to the modern state of the viscosity solutions. However, it may provide insights on an historical class of equations, and a very nice tool to build analytical solutions to the examples coming in section (2.2.2).

Chapter 2

Numerical methods

This chapter is splitted into two independant parts. The first presents a class of numerical schemes designed for the dimension 1, completely mesh-based, and best applied in smooth cases. The second targets high dimension, removes every connection with a mesh, and tackles nonsmooth problems. Both adresses the world of Hamilton-Jacobi equations, or the (strongly connected) transport equations.

2.1 Flux-limited schemes

Let $b : \mathbb{R}^n \mapsto \mathbb{R}$ be a Lipschitz-continuous function. We look for $u : \mathbb{R}^n \times \mathbb{R}^+$ solution of the advection equation

$$\begin{cases} \partial_t u(x,t) + \partial_x u(x,t) \cdot b(x) = 0 & (x,t) \in \mathbb{R}^n \times [0,T[\\ u(x,0) = u_0(x) & x \in \mathbb{R}^n \end{cases} \tag{2.1}$$

For $x \in \mathbb{R}^n$, we denote by $y_x(t)$ the solution of $\partial_t y(t) = b(y(t))$ with initial condition $y(0) = x$. The solution of (2.1) is then given by $u(x,t) = u_0(y_x(-t))$.

Let us simplify this problem by considering the scalar case $n = 1$, and taking a constant speed $b(x) \equiv b$. Notice that since the equation is linear, one can always consider $b \geqslant 0$, and define $v(x,t) \coloneqq u(x,-t)$ to treat negative speed.

The numerical approximation of (2.1) begins with finite difference schemes. Again, we restrict the problem to explicit schemes, where the time derivative is discretized with an explicit Euler scheme. It is well-known that finite difference discretization of the space derivative must "look for the information from where it comes" to be stable, that is, discretize with decentered schemes going *upwind* the advection. For instance, an upwind scheme for (2.1) could be

$$\frac{V_j^{n+1} - V_j^n}{\Delta t} + \frac{V_j^n - V_{j-1}^n}{\Delta x} = 0, \quad V_j^0 \coloneqq u_0(x_j)$$

where $V_j^n \simeq v(x_j, t^n)$ is the approximation of the solution on a grid $(x_j, t^n)_{j,n}$. Unfortunately, Taylor development shows that this scheme is consistant at order 2 with an advection equation where a diffusion coefficient is introduced. The immediate consequence is that the solution slowly but surely converges towards a constant. This drawback is known as "numerical diffusion", and the aim of this section is to find stable schemes that does not suffer from it.

2.1.1 Godunov-type schemes

Let $(x_j)_{j\in\mathbb{Z}} = \Delta x\, \mathbb{Z}$ be a regular space mesh of step $\Delta x > 0$, and $(t^n)_{n\in\mathbb{N}} = \Delta t\, \mathbb{N}$ be a time mesh of step $\Delta t > 0$. Uppercase letters V^n will denote vectors of unknowns at time t^n, whereas lowercase letters v, u will denote functions, u being the solution to (2.1). We denote by $\mathcal{F}$ a generic class of functions containing u.

Definition 5 – Iterative scheme We design a class of schemes characterized by

$$V^n = \underline{\mathcal{A}\mathcal{T}_{\Delta t}\mathcal{R}} \dots \underline{\mathcal{A}\mathcal{T}_{\Delta t}\mathcal{R}}\mathcal{A}v_0 = (\mathcal{A}\mathcal{T}_{\Delta t}\mathcal{R})^n \mathcal{A}v_0 \tag{2.2}$$

where $\mathcal{A} : \mathcal{F} \to \mathbb{R}^{\mathbb{Z}}$ is an averaging operator, $\mathcal{R} : \mathbb{R}^{\mathbb{Z}} \to \mathcal{F}$ is a reconstruction operator, and $\mathcal{T}_{\Delta t} : \mathcal{F} \to \mathcal{F}$ is a transport operator.

Godunov-type schemes are very general, and the sequence of operators can (theoritically) be chosen freely. The operators $\mathcal{A}$ and $\mathcal{R}$ converts functions into vectors and vice versa, for example by interpolation or best fit. The transport operator should be close to the semigroup operator associated with (2.1). In the case of scalar advection with positive constant speed, we give precise definitions of $\mathcal{A}$, $\mathcal{R}$ and $\mathcal{T}_{\Delta t}$, taken from the nice presentation of [Roe87]. The general case follows the same idea, with the use of approximated characteristic lines, or more generally, a Riemann solver for the operator $\mathcal{T}_{\Delta t}$.

Averaging Let c_j denote the cell of width Δx centered in x_j, i.e. $c_j = [x_j - \frac{\Delta x}{2}, x_j + \frac{\Delta x}{2}[$, and $I_{\Delta x}(x) = \mathbf{1}_{c_j}(x)/\Delta x$ be the normalized indicator of c_0. The operator $\mathcal{A}$ turns a function into a vector by the following cell-averaging :

$$\mathcal{A} : u \in L^1_{loc}(\mathbb{R}) \mapsto \mathcal{A}u \in \mathbb{R}^{\mathbb{Z}}, \quad (\mathcal{A}u)_j = (u * I_{\Delta x})(x_j) = \frac{1}{\Delta x}\int_{x_j - \frac{\Delta x}{2}}^{x_j + \frac{\Delta x}{2}} u(y)dy$$

Reconstruction The second operator wants to behave as a left inverse of $\mathcal{A}$, which would lead to an exact scheme. Numerically, we define

$$\mathcal{R} : V \in \mathbb{R}^{\mathbb{Z}} \mapsto \mathcal{R}V \in L^1_{loc}(\mathbb{R}), \quad \mathcal{R}V(x) = \sum_j \left(V_j + \varphi_j(V)\frac{x - x_j}{2\Delta x} \right) \mathbf{1}_{c_j}(x)$$

with $\varphi_j(V)$ a slope, which will be precised later. Note that the values of $\mathcal{R}V$ on each cell are centered in V_j : this choice implies that $(\mathcal{A}\mathcal{R}V)_j = V_j$, so that $\mathcal{R}$ is a right inverse of $\mathcal{A}$ in $\mathcal{R}^{\mathbb{Z}}$. This property would hold with any choice of centered reconstruction, such as the piecewise approximation $\mathcal{R}V(x) = V_j^n + \mathrm{sign}(x - x_j)d_j$ for $x \in c_j$.

Transport Finally, we define (for an arbitrary $h > 0$)

$$\mathcal{T}_h : v \in L^1_{loc}(\mathbb{R}) \mapsto \mathcal{T}_h v L^1_{loc}(\mathbb{R}), \quad \mathcal{T}_h v(x) = v(x - bh).$$

The operator $\mathcal{T}_{\Delta t}$, abbreviated $\mathcal{T}$ by now, gives the exact solution of the transport problem (2.1) with constant speed on a single time step Δt.

We now have a composed operator $\mathcal{A}\mathcal{T}\mathcal{R}$ that goes from the discrete space $\mathbb{R}^{\mathbb{Z}}$ to a discrete space $\mathbb{R}^{\mathbb{Z}}$, that is, a numerical scheme. One might think that this definition is not convenient to manipulate, but fortunately, the implementation is eased by the following expression.

Lemma 4 – Numerical expression Suppose $\nu \coloneqq b\Delta t/\Delta x \leqslant 1$. The scheme (2.2) is equivalent to

$$\frac{V_j^{n+1} - V_j^n}{\Delta t} + b\frac{V_{j+1/2}^n - V_{j-1/2}^n}{\Delta x} = 0 \quad \text{where } V_{j+1/2}^n \coloneqq V_j^n + \frac{1-\nu}{2}\varphi_j, \tag{2.3}$$

$$\text{and } V_j^0 = \frac{1}{\Delta x}\int_{x_j - \frac{\Delta x}{2}}^{x_j + \frac{\Delta x}{2}} v_0(y)dy. \tag{2.4}$$

Demonstration

The initial condition V^0 is exactly $\mathcal{A}v_0$. Given a sequence $\left(V_j^n\right)_j$, we apply $\mathcal{ATR}$ to compute the expression of V^{n+1}. We have

$$\begin{aligned} V_j^{n+1} = \left(\mathcal{ATR}V^n\right)_j &= \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\mathcal{TR}V^n\right)(x)dx = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\mathcal{R}V^n\right)(x-\nu\Delta x)dx \\ &= \frac{1}{\Delta x}\int_{x_{j-1/2}-\nu\Delta x}^{x_{j+1/2}-\nu\Delta x} \left(\mathcal{R}V^n\right)(x)dx \end{aligned}$$

Assuming $\nu \leqslant 1$, we have at most two different cells intersecting $[x_{j-1/2}-\nu\Delta x, x_{j+1/2}-\nu\Delta x[$. Splitting the integral in two, we find that

$$\begin{aligned} V_j^{n+1} &= \frac{1}{\Delta x}\int_{x_{j-1/2}-\nu\Delta x}^{x_{j-1/2}} \left(V_{j-1}^n + \varphi_{j-1}\frac{x-x_{j-1}}{2\Delta x}\right)dx + \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}-\nu\Delta x} \left(V_j^n + \varphi_j\frac{x-x_j}{2\Delta x}\right)dx \\ &= \nu V_{j-1}^n + \frac{\varphi_{j-1}}{2\Delta x}\left(\frac{\left(x_{j-1/2}-x_{j-1}\right)^2}{\Delta x} - \frac{\left(x_{j-1/2}-\nu\Delta x-x_{j-1}\right)^2}{\Delta x}\right) + (1-\nu)V_j^n \\ &+ \frac{\varphi_j}{2\Delta x}\left(\frac{\left(x_{j+1/2}-\nu\Delta x-x_j\right)^2}{\Delta x} - \frac{\left(x_{j-1/2}-x_j\right)^2}{\Delta x}\right) \\ &= \nu V_{j-1}^n + (1-\nu)V_j^n + \frac{\varphi_{j-1}}{2}\left(\frac{1}{2} - \left(\frac{1}{4}-\nu\right)^2\right) + \frac{\varphi_j}{2}\left(\left(\frac{1}{4}-\nu\right)^2 - \frac{1}{2}\right) \\ &= \nu V_{j-1}^n + (1-\nu)V_j^n - \frac{\nu(1-\nu)}{2}\left(\varphi_j - \varphi_{j-1}\right) \end{aligned}$$

This can be written as $V_j^{n+1} - V_j^n + \nu\left([V_j^n + \frac{1-\nu}{2}\varphi_j] - [V_{j-1}^n + \frac{1-\nu}{2}\varphi_{j-1}]\right) = 0$, and the given expression comes from $\nu = b\frac{\Delta t}{\Delta x}$. □

We use the notation $\mathcal{T}u(x,t)$ in place of $\mathcal{T}(u(\cdot,t))(x)$. The exact solution u satisfies the equivalent equation $u(\cdot,t+h) - \mathcal{T}_h u(\cdot,t) = 0$. Using this expression, we can give a consistency error estimate, as soon as the reconstruction step is efficient enough.

Proposition 6 – Consistency of the scheme Suppose that the operator $\mathcal{R}$ is of order 2 in the following sense : for any $v \in \mathcal{C}^2(\mathbb{R})$, if $V \coloneqq (v(x_j))_j$ is the interpolation of v, then

$$|\mathcal{R}V(x) - v(x)| \leqslant C\Delta x^2$$

Then if $u \in \mathcal{C}^2(\Omega)$ is a solution to (2.1), we have for any $n \in \mathbb{N}$

$$\left|\mathcal{TRA}u(x,t^n) - u(x,t^{n+1})\right| \leqslant \widetilde{C}\Delta x^2$$

Demonstration

The function u satisfies $u(\cdot,t^{n+1}) = \mathcal{T}u(\cdot,t^n)$, so that

$$\left|\mathcal{TRA}u\left(y,t^n\right) - u(y,t^{n+1})\right| = \left|\mathcal{RA}u\left(y-\nu\Delta x\right) - uyx - \nu\Delta x)\right|.$$

Let $x \coloneqq y - \nu\Delta x$. We have

$$|\mathcal{RA}u(x) - u(x)| \leqslant \left|\mathcal{RA}u\left(x,t^n\right) - \left(u(\cdot,t^n) * I_{\Delta x}\right)(x)\right| + \left|\left(u(\cdot,t^n) * I_{\Delta x}\right)(x) - u(x,t^n)\right|.$$

The sequence $\mathcal{A}u(\cdot,t^n)$ is exactly the interpolation on $(x_j)_j$ of the function $u(\cdot,t^n) * I_{\Delta x}$. The hypothese on $\mathcal{R}$

then bounds the first term by $C\Delta x^2$. On the other hand,

$$|(u(\cdot,t^n)*I_{\Delta x})(x)-u(x,t^n)| = \left|\frac{1}{\Delta x}\int_{x-\Delta x/2}^{x+\Delta x/2} u(y)dy - u(x,t^n)\right| \quad \text{and with a Taylor expansion,}$$
$$= \left|\frac{1}{\Delta x}\int_{x-\Delta x/2}^{x+\Delta x/2} u(x,t^n)+(y-x)\partial_x u(x,t^n)+O((y-x)^2)dy - u(x,t^n)\right|$$
$$= \left|u(x,t^n)+0+O\left(\Delta x^2\right)dy - u(x,t^n)\right| = O\left(\Delta x^2\right)$$

and owing to the smoothness of u, for sufficiently small Δx, $O(\Delta x^2)\leqslant \Delta x^2\widetilde{C}$, leading to the result. □

Corollary 1 – Choice of the slope Let $v\in\mathcal{C}^2(\Omega)$, and $V=(v(x_j))_j$. Suppose each $(\varphi_j)_j$ is chosen such that

$$\frac{\varphi_j(V)}{2\Delta x} = v'(x_j)+O(\Delta x)$$

Then the scheme defined by $(\varphi_j)_j$ is consistant of order 2.

The result follows from the expression

$$\mathcal{R}V(x) = \sum_j \left(V_j+\varphi_j(V)\frac{x-x_j}{2\Delta x}\right)\mathbf{1}_{c_j}(x) = \sum_j \left(v(x_j)+v'(x_j)\left(x-x_j\right)+(x-x_j)O(\Delta x)\right)\mathbf{1}_{c_j}(x)$$

which is an approximation of order 2 of v. In particular, any choice of

$$\frac{\varphi_j(V)}{2\Delta x} = (1-\lambda_j)\frac{V_j-V_{j-1}}{\Delta x}+\lambda_j\frac{V_{j+1}-V_j}{\Delta x}, \quad \lambda_j\in[0,1]$$

leads to such an order. Note that there is no need to compute the $(\varphi_j)_j$ with the same expression for all j : this opens the door for non-linear schemes, where each φ_j is independently evaluated based on local information.

2.1.2 Stability

Following the previous section, we consider a general scheme defined by

$$\frac{V_j^{n+1}-V_j^n}{\Delta t}+b_j\frac{V_{j+\frac{1}{2}}^n-V_{j-\frac{1}{2}}^n}{\Delta x}=0, \quad V_{j+\frac{1}{2}}^n \coloneqq V_j^n+\frac{1-\nu_j}{2}\varphi_j\left(V_{j+1}^n-V_j^n\right), \quad V^0=\mathcal{A}v_0 \tag{2.5}$$

Here, we allow b to vary with x_j, since the results are equivalent. We make use of the following definition :

Definition 6 – Local stability Suppose $b_j\geqslant 0\ \forall j$. A scheme is locally stable if

$$V_j^{n+1}\in[V_j^n,V_{j-1}^n] \quad \forall j\in\mathbb{Z},\ \forall n\in\mathbb{N}^*$$

Along the steps of [Swe84], [Lag00] or [Bok05], we wish to choose the maximal φ that defines a locally stable scheme. In particular, no regularity assumptions are made on V. The answer is known as the UltraBee limiter : let $r_j\coloneqq(V_j-V_{j-1})/(V_{j+1}-V_j)$ be the local ratio. The values $\varphi_j\coloneqq\varphi^{UB}(r_j,\nu_j)$ are chosen, with

$$\varphi^{UB}(r,\nu)=\max\left(0,\min\left(\frac{2}{1-\nu},r\frac{\nu}{2}\right)\right)=\frac{2}{1-\nu}\mathbb{P}_{[0,1]}\left(\frac{1-\nu}{\nu}r\right)$$

We combine the proof with a slightly stronger result.

Proposition 7 – Stability by comparison Any φ satisfying $0 \leqslant \varphi(r,\nu) \leqslant \varphi^{UB}(r,\nu)$ for all $r \in \overline{\mathbb{R}}$ and $\nu \in [0,1]$ defines a locally stable scheme by $\varphi_j = \varphi(r_j, \nu_j)$.

Demonstration

For clarity, we consider $\widetilde{\varphi}_j \coloneqq \frac{1-\nu_j}{2}\varphi_j$. The UltraBee is defined by $\widetilde{\varphi_j} = \mathbb{P}_{[0,1]}\left(\frac{1-\nu_j}{\nu_j} r_j\right)$. Definition (2.5) yields

$$V_j^{n+1} = V_j^n - \nu_j \left(V_j^n - V_{j-1}^n + \widetilde{\varphi}_j \left(V_{j+1}^n - V_j^n\right) - \widetilde{\varphi}_{j-1}\left(V_j^n - V_{j-1}^n\right)\right)$$

We discuss the case $V_j^n - V_{j-1}^n \geqslant 0$ first. By positivity of $\widetilde{\varphi}_{j-1}$ and ν_j,

$$V_j^{n+1} \geqslant V_j^n - \nu_j \left(V_j^n - V_{j-1}^n\right) - \nu\widetilde{\varphi}_j \left(V_{j+1}^n - V_j^n\right) \geqslant V_j^n - \nu_j \left(V_j^n - V_{j-1}^n\right) - \nu_j \widetilde{\varphi}_j^{UB}\left(V_{j+1}^n - V_j^n\right)$$

Note that $\widetilde{\varphi}^{UB}$ is always inferior to $\frac{1-\nu_j}{\nu_j} r_j$, and thus

$$V_j^{n+1} \geqslant V_j^n - \nu_j \left(V_j^n - V_{j-1}^n\right) - (1-\nu_j)\left(V_j^n - V_{j-1}^n\right) = V_{j-1}^n$$

On the other hand, if $r_j = \left(V_j^n - V_{j-1}^n\right) / \left(V_{j+1}^n - V_j^n\right)$ is negative, $\widetilde{\varphi}^{UB} = 0$ implies $\widetilde{\varphi} = 0$. We then have

$$V_j^{n+1} \leqslant V_j^n - \nu_j \left(1 - \widetilde{\varphi}_{j-1}\right)\left(V_j^n - V_{j-1}^n\right) \leqslant V_j^n + \nu_j \left(\widetilde{\varphi}_{j-1}^{UB} - 1\right)\left(V_j^n - V_{j-1}^n\right) \leqslant V_j^n$$

owing to $\widetilde{\varphi}^{UB} \leqslant 1$. The other case is symmetric. □

This nice result is completed by the following one.

Proposition 8 – TVD schemes A locally stable scheme is Total Variation Diminishing, i.e.

$$\sum_j \left|V_j^{n+1} - V_{j-1}^{n+1}\right| \leqslant \sum_j \left|V_j^n - V_{j-1}^n\right|$$

Demonstration

Local stability is equivalent to the existence of $(\lambda_j)_j \subset [0,1]$ such that $V_j^{n+1} = \lambda_j V_j^n + (1-\lambda_j)V_{j-1}^n$. Then

$$\begin{aligned} TV(V^{n+1}) = \sum_j \left|V_j^{n+1} - V_{j-1}^{n+1}\right| &= \sum_j \left|\lambda_j V_j^n + (1-\lambda_j)V_{j-1}^n - \lambda_{j-1}V_{j-1}^n - (1-\lambda_{j-1})V_{j-2}^n\right| \\ &= \sum_j \left|\lambda_j\left(V_j^n - V_{j-1}^n\right) + (1-\lambda_{j-1})\left(V_{j-1}^n - V_{j-2}^n\right)\right| \\ &\leqslant \sum_j \left(\lambda_j \left|V_j^n - V_{j-1}^n\right| + (1-\lambda_{j-1})\left|V_{j-1}^n - V_{j-2}^n\right|\right) \end{aligned}$$

and with a summation by parts, we find that $TV(V^{n+1}) \leqslant \sum_j \left|V_j^n - V_{j-1}^n\right| = TV(V^n)$. □

We now know how to choose the slopes φ_j such that the scheme is locally stable, and in the regions of regularity, we also have a family of second-order consistent schemes. Sweby [Swe84] introduced a convenient representation to compare φ functions as variables of r, where second-order schemes appears as a subregion of the stable schemes. Various examples may be found on this link.

2.1.3 Exact advection property

In addition to being a limit case for stability, the UltraBee scheme enjoys a curious property.

We need to define a class of "simple", or "unitary" waves : we make use of the following terminology.

Definition 7 – Shock wave Let $a \in \mathbb{R}$. The associate shock wave is $f : x \in \mathbb{R} \mapsto f(x) = \mathbf{1}_{\{x \leqslant a\}}(x)$.

Let $u_0(x) = \mathbf{1}_{\{x \leqslant 0\}}$ be a shock wave, and define $w := u_0 * I_{\Delta x}$. Then, by definition, the first iteration of any scheme of the form (2.5) will be to interpolate w on the mesh, i.e. $V^0 = (w(x_j))_j$.

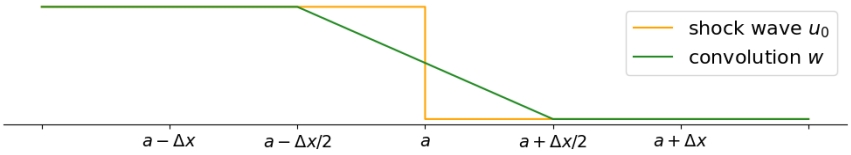


Figure 2.1: Initial function and its convolution by $I_{\Delta x}$

However, the UB limiter behaves particularly well.

Proposition 9 – UltraBee transport property At any iteration n, V^n is the interpolation of $\mathcal{T}^n w$, the convolution of the exactly transported shock wave.

In particular, the set of transition points $w^{-1}(]0,1[)$ has a width of Δx : this implies that at any n, the UltraBee scheme will propagate the shock wave with no more than one transition point. We proceed by building a scheme with the exact advection property, and we prove that it coïncides with the UB on shock waves.

Demonstration

The proof can be carried out with the reconstruction $\mathcal{R}$ involving piecewise linear functions. We prefer to use another interpretation, slightly more intuitive - and way less pedestrian. We define

$$\mathcal{R} : V\mathbb{R}^{\mathbb{Z}} \mapsto \mathcal{R}V \in L^1_{loc}, \quad \mathcal{R}V(x) = \sum_{c_j} \mathbf{1}_{\left\{x \leqslant x_{j-1/2} + \Delta x V_j\right\}}(x) \mathbf{1}_{c_j}(x)$$

The operator $\mathcal{R}$ "fills every cell" from left to right with the total disponible mass. This is well-adapted to our case, since $\mathcal{R}w$ will exactly fall back on v_0 : $\mathcal{R}$ is a left inverse of $\mathcal{A}$ on the class of shock waves. Thus, a scheme (S) of the form (2.2) will exactly propagate this class. The expression of V^{n+1} can be deduced similarly to the piecewise linear case : assuming $\nu \leqslant 1$, we compute

$$\begin{aligned} V_j^{n+1} = (\mathcal{A}\mathcal{T}\mathcal{R}V^n)_j &= \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathcal{R}V^n(x - \nu\Delta x)dx \\ &= \frac{1}{\Delta x} \int_{x_{j-1/2}-\nu\Delta x}^{x_{j-1/2}} \mathbf{1}_{\{x \leqslant x_{j-3/2} + \Delta x V_{j-1}^n\}}(x)dx + \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}-\nu\Delta x} \mathbf{1}_{\{x \leqslant x_{j-1/2} + \Delta x V_j^n\}}(x)dx \\ &= \max\left(0, V_{j-1}^n - (1-\nu)\right) + \min\left(1-\nu, V_j^n\right) \end{aligned}$$

by discussion on the integration intervals.
Suppose now that $(V^n)_j$ is an interpolation of w on a certain regular mesh (x_j), $x_j = j\Delta x + h$, $h \in [0, \Delta x[$ being free. We know that there exists an unique j such that

$$V_k^n = 1 \ \ \forall k < j, \quad V_j^n \in [0,1[, \quad V_k^n = 0 \ \ \forall k > j$$

Consequently, the reconstructed function is $\mathcal{R}w(x) = \mathbf{1}_{\{x \leqslant x_{j-1/2} + \Delta x V_j^n\}}(x)$, and the scheme (S) reduces to

$$V_k^{n+1} = V_k^n \ \ \forall k \notin \{j, j+1\}, \quad V_j^{n+1} = \nu + \min\left(1-\nu, V_j^n\right), \quad V_{j+1}^{n+1} = \max\left(0, V_j^n - (1-\nu)\right)$$

We know discuss the values V^{n+1} computed by the UltraBee scheme. By local stability, we have $V_k^{n+1} \in [V_k^n, V_{k-1}^n] = V_k^n$ for $k \notin \{j, j+1\}$, since the approximation is locally constant. On both sides $j \pm 1$, we need that

either $V^n_{j\pm1} = V^n_{j\pm2}$, or $r_{j\pm1} \leqslant 0$. In both cases, V^{n+1}_j and V^{n+1}_{j+1} will only depend on V^n_j and φ_j. Denoting again $\widetilde{\varphi} = \frac{1-\nu}{2}\varphi$, we have

$$V^{n+1}_j = V^n_j - \nu\left(V^n_j + \widetilde{\varphi}_j\Big(V^n_{j+1} - V^n_j\Big) - V^n_{j-1} - \overbrace{\widetilde{\varphi}_{j-1}(V^n_j - V^n_{j-1})}^{=0}\right) = V^n_j + \nu - \nu V^n_j\,(1-\widetilde{\varphi}_j)$$
$$V^{n+1}_{j+1} = V^n_{j+1} - \nu\Big(V^n_{j+1} - \underbrace{\widetilde{\varphi}_{j+1}\left(V^n_{j+2} - V^n_{j+1}\right)}_{=0} - V^n_j - \widetilde{\varphi}_j\left(V^n_{j+1} - V^n_j\right)\Big) = \nu V^n_j\,(1-\widetilde{\varphi}_j)$$

so that the only parameter is $\widetilde{\varphi}_j = \mathbb{P}_{[0,1]}\left(\frac{1-\nu}{\nu} r_j\right)$. The disposition $V^n_{j+1} \leqslant V^n_j \leqslant V^n_{j-1}$ implies $r_j \geqslant 0$. The remaining cases are delimited by V^n_j such that

$$\frac{1-\nu}{\nu}\frac{V^n_j - V^n_{j-1}}{V^n_{j+1} - V^n_j} = 1 \quad \Longleftrightarrow \quad (V^n_j - 1)(1-\nu) = -\nu V^n_j \quad \Longleftrightarrow \quad V^n_j = 1-\nu$$

In case $V^n_j \leqslant 1-\nu$, $\widetilde{\varphi}_j = 1$ and we find by both ways $V^{n+1}_j = V^n_j + \nu$, $V^{n+1}_{j+1} = 0$. On the other hand, if $V^n_j \geqslant 1-\nu$, we find

$$\nu V^n_j\,(1-\widetilde{\varphi}_j) = \nu V^n_j\left(1 - \frac{1-\nu}{\nu}\frac{V^n_j - 1}{-V^n_j}\right) = \nu V^n_j + (1-\nu)(V^n_j - 1) = V^n_j - (1-\nu)$$

and both schemes lead to $V^{n+1}_j = 1$ and $V^{n+1}_{j+1} = V^n_j - (1-\nu)$. Since the initialisation step is identical, we conclude by induction that the UltraBee coïncides with the constructed scheme (S) on shock waves. □

Remark 3 (Generalisation) *This property holds for any function that can be locally seen as a shock wave. More precisely, the proof holds if $V^n_{j-2} = V^n_{j-1}$ and $V^n_{j+1} = V^n_{j+2}$: we infer that any sequence V^n piecewise constant on sets of at least two consecutive points, with at most one transition point between each value, will stay on the graph of a convolution with a piecewise constant function.*

Remark 4 (Limit case) *The argument carries over if we only have $r_{j\pm1} \leqslant 0$, instead of $V_{j\pm1} = V_{j\pm2}$. An example is given by the limit case $u_0 = \mathbf{1}_{[-\Delta x, \Delta x[}$: the values V^n will stay on the graph of $(\mathcal{T}^n u_0) * I_{\Delta x}$. Since the width of the level set $\{w = 1\}$ is exactly of Δx, it is just possible for two consecutive points to take the same value.*

Integral schemes

Let us take a step back, and consider a more general Hamilton-Jacobi equation. Let $H : \mathbb{R} \mapsto \mathbb{R}$ be the Hamiltonian. We look for u satisfying

$$\begin{cases} \partial_t u + H(\partial_x u) = 0 \\ u(\cdot, 0) = u_0 \end{cases} \tag{2.6}$$

Xu and Shu [SX05] have proposed a class of "Hamiltonian corrected" limited schemes for this equation. Let φ be a limiter, and define

$$\frac{V^{n+1}_j - V^n_j}{\Delta t} + \hat{H}_j = 0, \quad \hat{H}_j = H\left(\frac{V^n_j - V^n_{j-1}}{\Delta x}\right)(1-\widetilde{\varphi}_j) + \widetilde{\varphi}_j H\left(\frac{V^n_{j+1} - V^n_j}{\Delta x}\right) \tag{2.7}$$
$$\widetilde{\varphi}_j = \frac{1-\nu}{2}\varphi(r_j, \nu_j), \quad r_j = \frac{\Delta_j V^n - \Delta_{j-1} V^n}{\Delta_{j+1} V^n - \Delta_j V^n}, \quad \Delta_j V^n = V^n_j - V^n_{j-1} \tag{2.8}$$

In the linear case $H(p) = bp$, (2.6) coïncides with (2.1). Correcting the Hamiltonian instead of the fluxes leads to the following property.

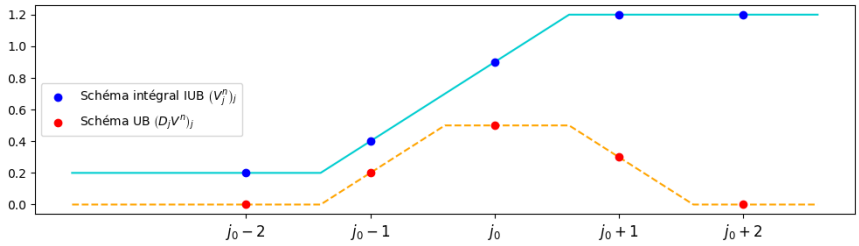


Figure 2.2: Values of the integral scheme and of the differences

Lemma 5 – Integral scheme Suppose b constant. The differences $(\Delta_j V^n)_{j,n}$ satisfy a limited scheme of the form (2.5), with φ as a limiter.

Demonstration

Definition (2.7) is equivalent to $V_j^{n+1} = V_j^n - \nu\left(\Delta_j V^n + \varphi_j\left(\Delta_{j+1}V^n - \Delta_j V^n\right)\right)$. Substracting V_{j-1}^{n+1}, we find

$$\Delta_j V^{n+1} = \Delta_j V^n - \nu\left(\left[\Delta_j V^n + \varphi_j\left(\Delta_{j+1}V^n - \Delta_j V^n\right)\right] - \left[\Delta_{j-1}V^n + \varphi_{j-1}\left(\Delta_j V^n - \Delta_{j-1}V^n\right)\right]\right)$$

The ratio r_j is evaluated in variables $\Delta_j V^n$, so that the definition exactly matches a limited scheme (2.5). □

Nice properties of integral schemes can be deduced from this relation. We use the following definition :

Definition 8 – Broken-line function Let $a, b, c, d \in \mathbb{R}^4$. A broken-line function is of the form $f(x) = \min\left(ax + b, cx + d\right)$.

Without loss of generality, we consider $a < c$, so that $f(x) = (ax+b)\mathbf{1}_{\{x\leqslant x_0\}}(x) + (cx+d)\mathbf{1}_{\{x>x_0\}}(x)$, with $x_0 = \frac{d-b}{a-c}$ the junction point.

Corollary 2 – Advection of broken-line functions The integral scheme associated to φ^{UB} (referred as (IUB)) is exact on broken-line functions.

Indeed, suppose that $(V_j^n)_j$ is an interpolation of a broken-line function f. The differences $(\Delta_j V^n)_j$ are then the interpolation of the shock wave $f'(x) = a\mathbf{1}_{\{x\leqslant x_0\}}(x) + c\mathbf{1}_{\{x>x_0\}}(x)$, eventually with a transition point $\Delta_j V^n$ if $x_{j-1} < x_0 < x_j$. The unicity of this point implies that for every iteration n and every point j, either the backward difference $\Delta_j V^n$ or the forward difference $\Delta_{j+1}V^n$ is in $\{a, c\}$, and every value (x_j, V_j^n) belongs to $(x, ax + b)$ or $(x, cx + d)$.

As before, this generalises for a class of piecewise linear functions of at least 3 consecutive aligned values (and thus two consecutive equal differences). The limit case is reached on functions equal to $u_0 * I_{2\Delta x}$, where u_0 is a shock wave, as represented on figure (2.2).

Remark 5 (Non-stability) *The exact advection of broken-line functions contradicts the local stability. Indeed, consider an interpolation of f that does not pass through the junction point. We may always choose ν such that the junction is exactly interpolated at $n+1$, falling outside the local intervals. Although (IUB) is still total variation bounded in dimension 1, owing to the local stability of $(\Delta_j V^n)_j$, this desirable property is lost on splitting schemes.*

We conclude this section with some insights gained from numerical simulations, yet (to our knowledge) unproved. Both the UltraBee and its integral equivalent are exact (in a certain sense) on a class of functions. We

observe that whenever u_0 is outside of this class, the numerical solution is quickly "projected" on this subspace, creating staircase-like structures. This has been observed, for instance, by Deprès-Lagoutière [Lag00].

We believe that all schemes defined by a limiter φ with the pointwise inequality $\varphi_{NB} \leqslant \varphi \leqslant \varphi_{UB}$ are exact on a certain class of function, where φ_{UB} is associated with the UltraBee, and φ_{NB} is the upper limit of the region of order 2, given by

$$\varphi^{NB}(r,\nu) \coloneqq \max\left(0, \min\left(\frac{1-\nu}{\nu}r, \frac{1-\nu}{2}\right), \min\left(\frac{1-\nu}{2}r, 1\right)\right).$$

Moreover, we believe that numerical solutions of such schemes will converge towards the said class of function, which is made by eigenvectors of the discrete operator $\mathcal{T}_{\Delta t}$. Numerical propagation could then be seen as an iterated power method on a nonlinear operator, and solutions will converge towards the eigenvectors of maximal eigenvalues. It is still not clear how to characterize these eigenvectors, and even it the case of the UltraBee, it is not trivial to deduce the shape of an eigenvector from the expression of the scheme. This may be a perspective for future work.

2.2 Neural networks

We now turn to a completely different topic. Neural networks are becoming a family of numerical methods with an active community, usual empirical practices, and growing theoritical results. Efforts are made to obtain density results in large functional spaces (see [Hor91] for multilayers perceptrons), to characterize the behavior of network spaces with respect to size (as in [HB]), or to obtain rates of convergence (see for instance Chap. 16 of [GKKW02], among other useful comments).

The historical definition of a neural network reflects its goal, which is to classify regions of the space. However, one may try to approximate functions with neural networks - provided one knows how to choose the activation function accordingly. In this point of view, neural networks are finite-dimensional spaces of approximation, which may be tried in place of finite elements or polynomial basis, for instance. We follow this general idea in the case of semi-lagrangian schemes.

2.2.1 Definitions and scheme

What is a neural network We begin by a general definition.

Definition 9 – Feedforward neural network Let $l \in \mathbb{N}^*$ be the *number of layers*, and $(n_0, \dots, n_l) \in (\mathbb{N}^*)^{l+1}$ be a sequence of dimensions. We denote by $\sigma_i : \mathbb{R}^{n_i} \mapsto \mathbb{R}^{n_i}$ the i^{th} *activation function*, and by $\mathcal{L}_i : \mathbb{R}^{n_{i-1}} \mapsto \mathbb{R}^{n_i}$ an affine transformation, with $i \in [\![1, l]\!]$. The associated neural network is a function from $\mathbb{R}^{n_0}$ to $\mathbb{R}^{n_l}$, given by

$$\mathcal{R}(x) = \sigma_l \circ \mathcal{L}_l \circ \cdots \circ \sigma_1 \circ \mathcal{L}_1(x) \quad \forall x \in \mathbb{R}^{n_0}$$

In the sequel, we will often restrict to $n_0 = n$, $n_l \in \{1, n\}$, and $n_i = N_n \in \mathbb{N}^*$ for all $i \in [\![1, l-1]\!]$. We call N_n the *number of neurons*.

Activation functions are often coordwise functions, in the sense that $\sigma(x)_i = \overline{\sigma}(x_i)$ for a certain $\overline{\sigma} : \mathbb{R} \mapsto \mathbb{R}$. Famous examples include the Rectified Linear unit (ReLu) $\overline{\sigma}(x) = \max(x, 0)$, the sigmoid function $\overline{\sigma}(x) = \frac{1}{1+\exp(-x)}$, or simply the identity map. Theorem 16.1 in [GKKW02] states that under mild assumption on $\overline{\sigma}$, the class of neural network is dense in continuous functions for a weighted L^2 norm. Such results are part of a corpus of "universal approximation theorems", with various degrees of precision (see for instance Table 1 of [TSB20] for a classification of some estimates).

Besides coordwise functions, one may need to deduce a multidimensional result from the inputs. For instance, the softmax function returns a vector of probability, that may be interpreted as degree of confidence for the classification of an item. Finally, the GroupSort function acts blockwise by sorting the inputs. This curious choice

leads to a desirable property : by bounding the coefficients of the affine transformations $\mathcal{L}_i$, one can ensure that the neural network will be Lipschitz-continuous. This stands on the 1-Lipschitz continuity of the GroupSort function. As the proof is quite fun, we present it here.

Proposition Let $\sigma : \mathbb{R}^k \to \mathbb{R}^k$ be the sorting function in decreasing order, given by

$$\sigma(x)_1 \coloneqq \max_{i\in[\![1,k]\!]} x_i =: x_{i_1}, \quad \sigma(x)_2 \coloneqq \max_{i\in[\![1,k]\!], i\neq i_1} x_i, \quad \dots \quad \sigma(x)_k \coloneqq x_{i_k},\ i_k \notin \{i_1,\dots,i_{k-1}\}.$$

Then σ is 1-lipschitz for $\|\cdot\|_\infty$, i.e $\max_{i\in[\![1,k]\!]} |\sigma(x)_i - \sigma(y)_i| \leqslant \max_{i\in[\![1,k]\!]} |x_i - y_i|$ for all $(x,y) \in (\mathbb{R}^k)^2$.

Demonstration

Let $x, y \in \mathbb{R}^k$. We call p (resp. q) a permutation of indices such that $\sigma(x)_i = x_{p(i)}$ for all i (resp $\sigma(y)_j = y_{q(j)}$ for all j). Let $i_0 \in [\![1,k]\!]$ such that $\|\sigma(x)-\sigma(y)\|_\infty = |\sigma(x)_{i_0} - \sigma(y)_{i_0}|$: w.l.o.g, we assume that $\sigma(x)_{i_0} - \sigma(y)_{i_0} \geqslant 0$. Consider the set of indexes (in green in the illustrative table below)

$$\mathcal{S} \coloneqq \{p(1),\dots,p(i_0),q(i_0),\dots,q(k)\}$$

Since $\mathcal{S}$ is made of $k+1$ elements in $[\![1,k]\!]$, at least one of them appears twice. Since $p(i) \neq p(j)$ for $i \neq j$, we can find $l \geqslant i_0$ and $m \leqslant i_0$ such that $p(l) = q(m)$. By the well-ordering of the vectors, we have

$$\|\sigma(x) - \sigma(y)\|_\infty = \sigma(x)_{i_0} - \sigma(y)_{i_0} \leqslant \sigma(x)_l - \sigma(y)_{i_0} \leqslant \sigma(x)_l - \sigma(y)_m = x_{p(l)} - y_{q(m)} \leqslant \|x-y\|_\infty$$

and σ is 1-lipschitz.

$$\begin{array}{ccccccccccccccccccc}
x_{p(1)} & & & & & & x_{p(l)} & & & & & & x_{p(i_0)} & & & & x_{p(m)} & & & & & & x_{p(k)} \\
\uparrow & & & & & & \uparrow & & & & & & \uparrow & & & & \uparrow & & & & & & \uparrow \\
\sigma(x)_1 & \geqslant & \dots & \geqslant & & & \sigma(x)_l & \geqslant & \dots & \geqslant & & & \sigma(x)_{i_0} & \geqslant & \dots & \geqslant & \sigma(x)_m & \geqslant & \dots & \geqslant & & & \sigma(x)_k \\
 & & & & & & & & & & & & \wedge\!\!\vee & & & & & & & & & & \\
\sigma(y)_1 & \geqslant & \dots & \geqslant & & & \sigma(y)_l & \geqslant & \dots & \geqslant & & & \sigma(y)_{i_0} & \geqslant & \dots & \geqslant & \sigma(y)_m & \geqslant & \dots & \geqslant & & & \sigma(y)_k \\
\downarrow & & & & & & \downarrow & & & & & & \downarrow & & & & \downarrow & & & & & & \downarrow \\
y_{q(1)} & & & & & & y_{q(l)} & & & & & & y_{q(i_0)} & & & & y_{q(m)} & & & & & & y_{q(k)}
\end{array}$$

□

Using the chain rule, it is now possible to control the Lipschitz constant of networks if the $L^\infty-$norm of the affine transformations is appropriately controlled. This can be useful for numerical schemes whose convergence properties depend on the smoothness of the approximations. The next section develops what we mean by *numerical schemes* using neural networks.

How do we use them This section briefly describes the intervention of neural networks in a numerical scheme. Rigorous definitions and developments will be found in [**?**].

We consider an obstacle problem in "physically high" dimension, i.e $n \in [\![2,8]\!]$. Our aim is to implement a classical Lagrangian scheme, with a time discretization of the dynamic programming principle (1.5). In the case of front propagation, the value function $(x,t) \to V(x,t)$ satisfies

$$\begin{cases} V(x,t) = \min_{a\in\mathbb{A}_{[t,t+h]}} \max_{\theta\in[t,t+h]} g(y_x^a(\theta)) \bigvee V(y_x^a(t+h), t+h) & (x,t) \in \mathbb{R}^n \times]0,T[,\ h \in]0, T-t[\\ V(x,T) \coloneqq g(x) \vee \phi(x), \quad \dot{y}_x^a(\theta) = f(y_x^a(\theta), a(\theta)), \quad y_x^a(t) \coloneqq x. \end{cases}$$

The discretization proceeds in 3 steps. First, a time mesh is introduced. Then, the space of control is approximated by a finite-dimensional space, using an average criterion in the minimization. Finally, the continuous criterion is approximated with a discrete counterpart.

Step 1 Let $\Delta t > 0$, and $t_k := k\Delta t$ for $k \in [\![0, N]\!]$. The first discretized cost function is

$$J_k : \mathbb{R}^n \times \mathbb{A}_{[t_k,t_{k+1}]} \mapsto J_k(x, a) := \max_{\theta\in[t_k,t_{k+1}]} g(y_x^a(\theta)) \bigvee V^{k+1}(y_x^a(t_{k+1})).$$

The time-discrete value functions $V^k \sim V(\cdot, t_k)$ are defined by

$$\begin{cases} V^k(x) = \min_{a\in\mathbb{A}_{[t_k,t_{k+1}]}} J_k(x, a) \quad x \in \mathbb{R}^n,\ k \in [\![0, N]\!] \\ V^N(x) := g(x) \vee \phi(x). \end{cases} \tag{2.9}$$

Step 2 The next idea is to approximate *open-loop controls* of $\mathbb{A}_{[t_k,t_{k+1}]}$ by a sequence of *closed-loop controls* (or *feedback controls*) in $\mathcal{A} := L^\infty(\mathbb{R}^n, A)$. This implies two changes.

– Given $a \in A$ and an initial time t, the characteristics will be approximated by

$$\partial_t \hat{y}_x^a(\theta) = f(\hat{y}_x^a(\theta), a), \quad \hat{y}_x^a(t) = x.$$

This corresponds to a restriction to constant controls $a(\cdot) \equiv a \in A$. Consequently, we define $\hat{J}_k$ by

$$\hat{J}_k : \mathbb{R}^n \times A \mapsto \hat{J}_k(x, a) := \max_{\theta\in[t_k,t_{k+1}]} g(\hat{y}_x^a(\theta)) \bigvee V^{k+1}(\hat{y}_x^a(t_{k+1})). \tag{2.10}$$

– We turn to a global minimization criterion instead of a pointwise one. To this end, let μ be a probability measure on $\mathbb{R}^n$ that is nowhere degenerate, i.e. $\mu(x) > x$ for all $x \in \mathbb{R}^n$.

This intermediate scheme needs two sequences $(V^k)_{k\in[\![0,N]\!]}$ and $(a^k)_{k\in[\![0,N]\!]} \subset \mathcal{A}$ such that

$$\begin{cases} a^k \in \operatorname*{argmin}_{a\in\mathcal{A}} \int_{\mathbb{R}^n} \hat{J}_k(x, a(x))\mu(dx), \quad V^k(x) := \hat{J}_k(x, a^k(x)),\ k \in [\![0, N]\!] \\ V^N(x) := g(x) \vee \phi(x). \end{cases}$$

This functional formulation opens the door to internal approximations of $\mathcal{A}$.

Step 3 The implementation of the scheme requires the following two approximations:

– Feedback controls will be approximated by a finite-dimensional space $\hat{\mathcal{A}} \subset \mathcal{A}$. The space could be chosen among finite elements, discontinuous galerkin, spectral approximations, max-plus pseudo-basis... We will illustrate the choice of neural networks.

– The measure μ is replaced by an empirical measure $\hat{\mu}(x) = \sum_{i=1}^M \delta_{X_i}(x)$, for a set of iid realisations (X_i) of a random variable $X \sim \mu$.

These modifications lead to the (final) scheme.

Definition 10 – Lagragian scheme Let $\hat{\mathcal{A}} \subset L^\infty(\Omega)$ be a finite-dimensional space for the controls, $M \in \mathbb{N}$ a number of samples, and define the sequence $(\hat{V}^n)_{n\in[\![0,N]\!]} \simeq (V^n)_{n\in[\![0,N]\!]}$ by

$$\begin{cases} \hat{V}^N(x) := g(x) \vee \phi(x), \\ \hat{a}^k \in \operatorname*{argmin}_{\hat{a}\in\hat{\mathcal{A}}} \frac{1}{M} \sum_{i=1}^M \hat{J}_k(X_i, \hat{a}(X_i)), \quad \hat{V}^k(x) := J_k(x, \hat{a}^k(x)). \end{cases} \tag{2.11}$$

where $\hat{J}_k$ is defined in (2.10).

Since the problem is in finite horizon $T > 0$, each approximation $\hat{V}^k$ can be written as a composition of g, φ and the controls $\hat{a}^k, \dots, \hat{a}^{N-1}$. This allows us to store only the feedback controls, and to use a "full lagrangian" representation of $\hat{V}$.

Main result and comments Suppose $N \in \mathbb{N}$ the number of time iterations is fixed. Then, we have the following:

Proposition 10 – Convergence for fixed $N \in \mathbb{N}$ Let $(V^n)_{n\in[\![0,N]\!]}$ be the exact solution of the discrete problem (2.9), and define $(\hat{V}^n)_{n\in[\![0,N]\!]}$ by the numerical scheme (2.11). For clarity, we denote by $\hat{\mathcal{A}}_P$ the finite-dimensional space for the controls, with P a parameter such that $\overline{\lim_{P\to\infty} \hat{\mathcal{A}}_P} = \mathcal{C}(\Omega)$ (for instance, the size of the networks, the number of basis functions...). Then

$$\lim_{P\to\infty} \max_{0\leqslant n\leqslant N} |V^n - \hat{V}^n|_{L^1(\Omega)} = 0.$$

Let us comment on this result. All details can be found in [**?**]. The construction of the scheme naturally lead to discrete $L^1(\hat{\mu})$ estimates, since we have

$$\int_{x\in\Omega} \left[\hat{V}^k(x) - v(x,t^k)\right] \hat{\mu}(dx) = \min_{\hat{a}\in\hat{\mathcal{A}}} \int_{x\in\Omega} \left[\hat{J}_k(x,a) - J_k(x,a^*)\right] \hat{\mu}(dx).$$

In fine, one wishes to bound the error in function of $\left|V^{k+1} - v(\cdot,t^{k+1})\right|_{L^1} + \varepsilon_k$, with a consistency error ε_k that can be controlled. An idea is to replace the empirical measure $\hat{\mu}$ by μ (who may be taken absolutely continuous with respect to the Lebesgue measure), up to an error of order $M^{-1/2}$ (with M the number of samples of μ). Developping the expression of $\hat{J}_k$ and J_k, one gets to study

$$\int_{x\in\Omega} \left[\max_{\theta\in\Theta}(g(\hat{y}_x^a(\theta)) - g(y_x^{a^*}(\theta)))\right] \bigvee \left[\hat{V}^{k+1}(\hat{y}_x^a(t^{k+1})) - v(y_x^{a^*}(t^{k+1}),t^{k+1})\right] \mu(dx).$$

The set Θ is a discretisation of $[t^k,t^{k+1}]$. Owing to the Lipschitz regularity of g, the first maximand is directly controlled by the error $|\hat{y}_x^a(\theta) - y_x^a(\theta)|$. The maximand in $\hat{V}^{k+1}$ may be splitted into

$$\left[\hat{V}^{k+1}(\hat{y}_x^a(t^{k+1})) - v(\hat{y}_x^a(t^{k+1}),t^{k+1})\right] + \left[v(\hat{y}_x^a(t^{k+1}),t^{k+1}) - v(y_x^{a^*}(t^{k+1}),t^{k+1})\right].$$

The first summand looks one change of variable away from the error at time t^{k+1}, and the second one is controlled by $\left|\hat{y}_x^a(t^{k+1}) - y_x^a(t^{k+1})\right|$ since v is Lipschitz.

The difficulty lies is the lack of regularity of the optimal control a^*. Indeed, in the feedback formulation, the exact characteristics satisfy $\dot{y}_x^a(t) = f(y_x^a(t), a(y_x^a(t)))$, and the composition $y \mapsto f(y,a(y)) =: f^a(y)$ is, in general, discontinuous. Using Fillippov's theory of differential inclusions (see [FA88]), one may still define absolutely continuous trajectories. However,

- available numerical schemes for $\hat{y}$ are at most of order $O(\Delta t)$. This means that at best, $|\hat{y}_x^a(\theta) - y_x^a(\theta)| \leqslant C\Delta t$. Then, the sum of the errors on all time steps will be of order $O(1)$, which is not enough.
- It is not possible to perform a classical change of variable $x \to y_x^a(\theta) =: y$ with bounded jacobian.

We tried to tackle both problems by regularization. Briefly speaking, the regularized controls $a^\varepsilon := a^* * \rho_\varepsilon$ (with ρ_ε a convolution kernel) are proven to be near-optimal controls in an integral norm. Then, the error may be analyzed using these approximations, and one gets enough regularity to proceed by induction. It is clear that the bounds are degenerating when $\varepsilon \searrow 0$, but one still gets a convergence result for a fixed number of time steps N, when the limit of regularization parameter ε is carefully compensated by a growing size of the approximation space $\hat{\mathcal{A}}$.

Another possible direction would be to play on the measure μ. In particular, one wants to use the sequence $(\hat{\mu}_k)_k$ of discrete invariant measures, satisfying

$$\int_{x\in\mathbb{R}^n} f(\hat{y}_x^{a^*}(\theta))\hat{\mu}_k(dx) = \int_{y\in\mathbb{R}^n} f(y)\hat{\mu}_{k+1}(dy)$$

The lack of regularity translates here in a degeneresccence of μ_k. More precisely, $\hat{\mu}_k$ may be concentrated on sets of Lebesgue measure 0. This would not be a problem if the L^∞ error $\hat{V}^k - v(\cdot,t^k)$ was controlled, but in general, there is no hope to get such bounds with the proposed scheme, and we did not investigate this direction.

Anyway, it is worth testing the scheme numerically. The next section gives a few results of an implementation with neural networks. The objective is to reach dimensions up to $n = 8$, in which mesh-based approximations are far too expensive. The output of the scheme may be used to localize the exact solution, and initialize finer methods.

2.2.2 Numerical exploration

We begin by studying a class of problems whose analytical solution is known. Then, in a particular case, we add an obstacle, and describe how we can obtain a reference solution. In this section, the obstacle problems with terminal condition are converted to initial value problems by the transformation $u(x,t) \coloneqq v(x, T-t)$.

Hölder eikonal equation with drift

Let $(k, k^*) \in [1, \infty]$ be conjugate exponents, i.e $\frac{1}{k} + \frac{1}{k^*} = 1$. We denote the $k-$norm as $\|p\|_k = \left(\sum_{i=1}^n p_i^k\right)^{1/k}$.

Definition 11 – Hölder eikonal equation with drift Given an initial condition u_0, a vector $b \in \mathbb{R}^n$ and a constant $c \geqslant 0$, we look for u such that

$$\begin{cases} \partial_t u + \langle \nabla u, b \rangle + c\|\nabla u\|_{k^*} = 0 & (x,t) \in \mathbb{R}^n \times]0,T[\\ u(x,0) = u_0 & x \in \mathbb{R}^n \end{cases} \tag{2.12}$$

Let $\mathscr{B}_k(0,1)$ be the open unit ball for $\|\cdot\|_k$, and remember that $\|p\|_{k^*} = \max_{q \in \overline{\mathscr{B}_k(0,1)}} \langle p, q \rangle$. Then, (2.12) can be formulated as a control problem by

$$\begin{cases} \partial_t u + \langle \nabla u, b \rangle + c \max_{a \in \overline{\mathscr{B}_k(0,1)}} \langle \nabla u, a \rangle = 0 & (x,t) \in \mathbb{R}^n \times]0,T[\\ u(x,0) = u_0 & x \in \mathbb{R}^n \end{cases}$$

Lemma 6 – Solution Let $u_0(x) = \|x - \xi\|_k + \alpha_{\min}$. Then the analytical solution u and the optimal control a^* are given by

$$u(x,t) = \left(\|x - bt - \xi\|_k - ct\right)_+ + \alpha_{\min}, \qquad a^* = \frac{x - bt - \xi}{\max(ct, \|x - bt - \xi\|_k)}$$

Demonstration

Since all Hölder norms are convex, proper and lower semi-continuous, their Fenchel transform are well-defined. Let $H(p) = \langle p, b \rangle + c\|p\|_{k^*}$: then

$$H^*(q) = \sup_{p \in \mathbb{R}^n} \langle p, q - b \rangle - H(p) \leqslant \sup_{p \in \mathbb{R}^n} \|p\|_{k^*}\|q - b\|_k - c\|p\|_{k^*} = \begin{cases} 0 & \text{if } \|q - b\|_k \leqslant c \\ \infty & \text{otherwise} \end{cases}$$

and the Fenchel transform of H is the indicator of $\overline{\mathscr{B}_k(b,c)}$. Using Lax-Oleinik formula, we find

$$u(x,t) = \min_{q \in \overline{\mathscr{B}_k(b,c)}} u_0(x - tq) = \min_{q \in \overline{\mathscr{B}_k(0,1)}} \|x - bt - \xi - ctq\|_k + \alpha_{\min}$$

If $\|x - bt - \xi\|_k \leqslant ct$, the minimum is equal to $\alpha_{\min}$ and attained in $ctq = x - bt - \xi$. Otherwise, by the second triangular inequality,

$$\|x - bt - \xi - ctq\|_k \geqslant \|x - bt - \xi\|_k - ct\|q\|_k \geqslant \|x - bt - \xi\|_k - ct$$

and this becomes an equality for $q \coloneqq \frac{x - bt - \xi}{\|x - bt - \xi\|_k}$. This yields the desired result. □

The representation proposition (5) allows us to build more elaborate problems, with $u_0 = \min_{i\in\mathbb{I}} u_{0,i}$. Indeed,

$$u(x,t) = \inf_{y\in\partial\Omega}\{u_0(y) + \mathcal{L}(y,x)\} = \inf_{y\in\partial\Omega}\left\{\min_{i\in\mathbb{I}} u_{0,i}(y) + \mathcal{L}(y,x)\right\} = \min_{i\in\mathbb{I}}\inf_{y\in\partial\Omega} u_{0,i}(y) + \mathcal{L}(y,x) = \min_{i\in\mathbb{I}} u_i(x,t).$$

where u_i is the solution of (2.12) initialized with $u_{0,i}$. This "min-linearity" is heavily exploited in min-plus algebra, and leads to analogues of the spectral methods. The curious reader may enjoy [McE06] on this topic.

Numerical results We begin by a 2D case, with the following data :

$$p = \frac{3}{2}, \quad u_0(x) = \min_{\theta\in\left\{0,\frac{\pi}{2},\pi,\frac{3\pi}{2}\right\}} \|x - 2e^{i\theta}\|_k - \frac{1}{2}, \quad b = 0, \quad c = 1$$

We use a 3-layers feedforward architecture, with activation function ReLu. Minimization is tackled with a stochastic gradient algorithm, with 1000 gradient steps per minimization. At each iteration, 2000 samples X_i are uniformly drawn over Ω. The time discretization is uniform, with 5 steps running from $t = 0$ to $T = 1.2$.

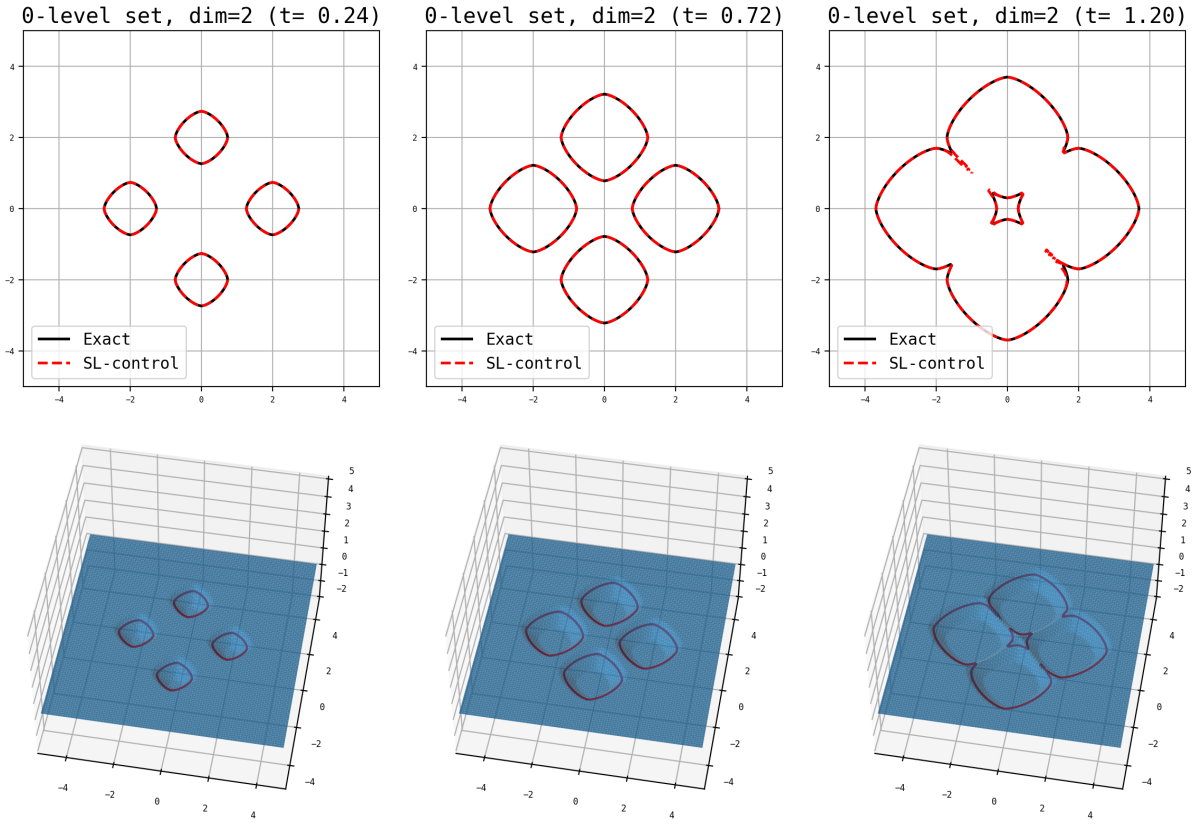


Figure 2.3: 0 level set of the solution, and truncated value function $\hat{V}$ for $p = 3/2$.

Results are collected in figure (2.3). Here, the value function is trucated on display for aesthetics.

The door problem

We now consider $k = 2$, and add an obstacle. More precisely, we use the following notations and parameters :

– $\xi \in \mathbb{R}^n$ will be the origin point.

– $d \in \mathbb{S}^{n-1}$ is the main direction of the example. $b = \|b\| d$ is the drift, and $c \geqslant 0$ the speed coefficient.

– For each $x \in \mathbb{R}^n$, we define $x_{\perp d}$ as $x_{\perp d} = x - \xi - \langle x - \xi, d \rangle d$.

– Given a real number $\alpha_{\min}$, the initial condition is $u_0(x) = \|x - \xi\| + \alpha_{\min}$.

– Given c_e, c_x, g_c, $g_{\max}$ and $g_{\min}$, the obstacle function is

$$g(x) := \min\left(g_{\max} - c_e \left|\langle x - \xi, d\rangle - g_c\right|, c_x \left|x_{\perp d}\right| + g_{\min}\right)$$

Definition 12 – Door problem We want to find u s. t.

$$\min\left(\partial_t u(x,t) + \max_{a \in \mathscr{B}(b,c)} a \cdot \nabla u(x,t), u(x,t) - g(x)\right) = 0 \qquad (x,t) \in \mathbb{R}^d \times]0,T[\qquad (2.13)$$

$$u(x,0) = u_0 \vee g(x) \qquad x \in \mathbb{R}^d \qquad (2.14)$$

The example name is justified in figure (2.4) : for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, the level set $\{g = \alpha\}$ is a wall pierced by a square door.

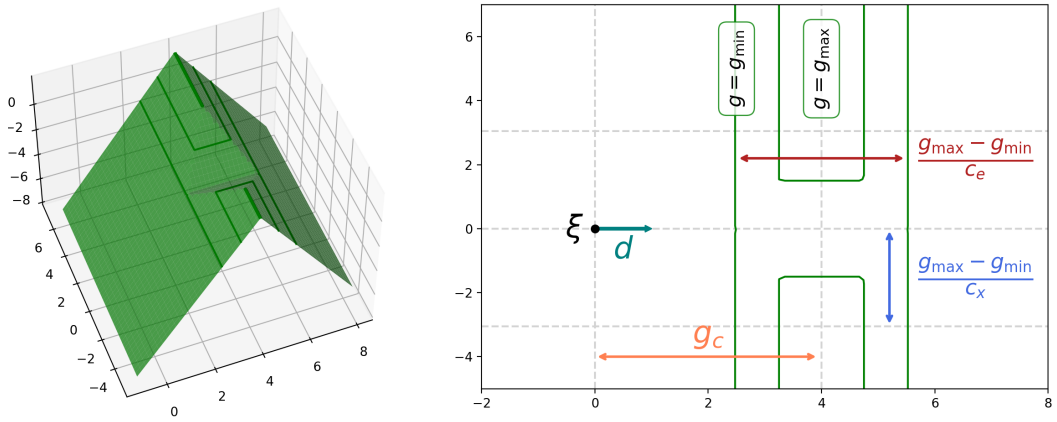


Figure 2.4: Obstacle function and illustration of its parameters.

(A8)

We suppose that

– $\alpha_{\min} \leqslant g_{\min}$. If $\alpha_{\min} > g_{\max}$, the obstacle function is always inferior to $u(x,t)$, and $u = u_{of}$.

– The obstacle is completely located at one side of the initial condition : for simplicity, we suppose that

$$\max_{\alpha \in [\alpha_{\min}, g_{\max}]} \max_{x \in \{u_0 = \alpha\}} \langle x - \xi, d\rangle \leqslant \min_{\alpha \in [\alpha_{\min}, g_{\max}]} \min_{x \in \{g = \alpha\}} \langle x - \xi, d\rangle$$

This is not very restrictive, since we can always take g_c large enough to move the obstacle away from the (bounded) initial conditions $\{u_0 = \alpha\}_{\alpha \leqslant g_{\max}}$.

From lemma (6) with $k = 2$, we know that the obstacle-free solution is $u_{of} = (|x - bt - \xi| - ct)_+ + \alpha_{\min}$.

Since the right hand side of (2.13) is vanishes identically, we know that obstacle-free characteristics will be straight lines. Now, consider an initial front $\{u_0 = l\}$, and the associated door $\{g = l\}$. Some points behind the obstacle are not reachable by a straight line : the geodesic will start from a point of the initial front, touch the

frame of the door, and then reach its destination (see figure (2.5)). This translates in a reaching time equation, that we now develop.

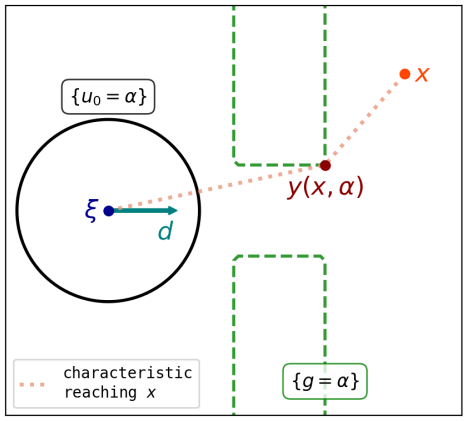


Figure 2.5: Characteristics in the presence of a door

Definition 13 – Reaching times Let $x \in \Omega$, and consider the front $\{u_0 = l\}$, where $u_0(y) := \|y-x\|+r$. The set of reaching times of a point y by the front is defined as

$$T_g^r(y, x, l) := \{t \in \mathbb{C} \mid u(y, t) = l\}, \quad u \text{ solution of (2.13) with initial condition } u_0$$

We denote by T^r the set of reaching times of the obstacle-free problem, corresponding to T_g^r for $g \equiv -\infty$.

The *physical* reaching time is the element of T_g^r of smallest positive real part, if such an element exist.

Let $\forall \alpha \in]g_{\max}, \alpha_{\min}[$. The value $\alpha = u(x, t)$ gives the level set of u_0 such that $T_g^{\alpha_{\min}}(x, \xi, u(x, t)) = t$. Denote by $y(x, \alpha)$ the point of $\{g = \alpha\}$ closest to x and still attainable in straight line by the front $\{u_0 = \alpha\}$: we have

$$T_g^{\alpha_{\min}}(x, \xi, \alpha) = T^{\alpha_{\min}}(y(x, \alpha), \xi, \alpha) + T^{\alpha}(x, y(x, \alpha), \alpha)$$

The good news is that obstacle-free reaching times are easily computable. We may then obtain α as the solution of the reaching time equation

$$T^{\alpha_{\min}}(y(x, \alpha), \xi, \alpha) + T^{\alpha}(x, y(x, \alpha), \alpha) = t \tag{2.15}$$

Numerical resolution of the reaching time equation This section is devoted to a numerical trick to avoid infinite reaching times. After some elementary geometry, we obtain the following expression for $y = y(x, \alpha)$:

$$y(x, \alpha) = \xi + \left(g_c + \frac{g_{\max} - \alpha}{c_e}\right) d + \frac{\alpha - g_{\min}}{c_x} \frac{x_{\perp d}}{|x_{\perp d}|}$$

Let us compute $T^{\alpha_{\min}}(y, x, \alpha)$. By definition, we look for t such that $\alpha = u(x, t) = (\|x - \xi - bt\| - ct)_+ + \alpha_{\min}$. Instead of this nonlinear problem, we consider the (non-equivalent) formulation

$$(\alpha - \alpha_{\min} + ct)^2 = \|x - \xi - bt\|^2 \tag{2.16}$$

Taking the square may enrich the set of solutions. Anyway, we can now develop and solve a quadratic equation in t. Table (2.1) gives an illustration of the possible cases.

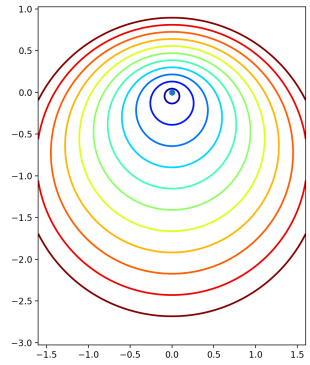
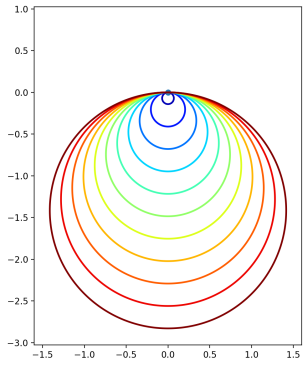
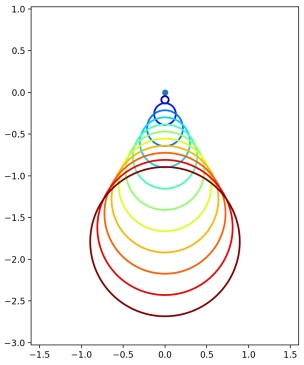
	controllable case	critical case	non controllable case
domain reaching time	$\|b\| < c$ always real	$\|b\| = c$ might be complex	$\|b\| > c$ might be complex
typical solution			

Table 2.1: Classification of behaviors

Remark 6 (Insights on complex roots) *Let us illustrate the solutions of* (2.16). *Suppose* $n = 2$, $\xi = (0,0)^t$, $c = 1$ *and* $b = (2,0)^t$. *We choose* $\alpha = 0$ *and* $\alpha_{\min} = -1$, *so that the initial front it the unit circle. Characteristics will radiate from* $\xi_0 = \xi - \frac{0-1}{1}b = -b$. *Computing the real and imaginary parts of* t, *we find the results of figure (2.6). Using the definition of* $\|a\|$ *as* $\left(\sum_{i=1}^n a_i^2\right)^{1/2}$, *one can indeed verify that both roots*

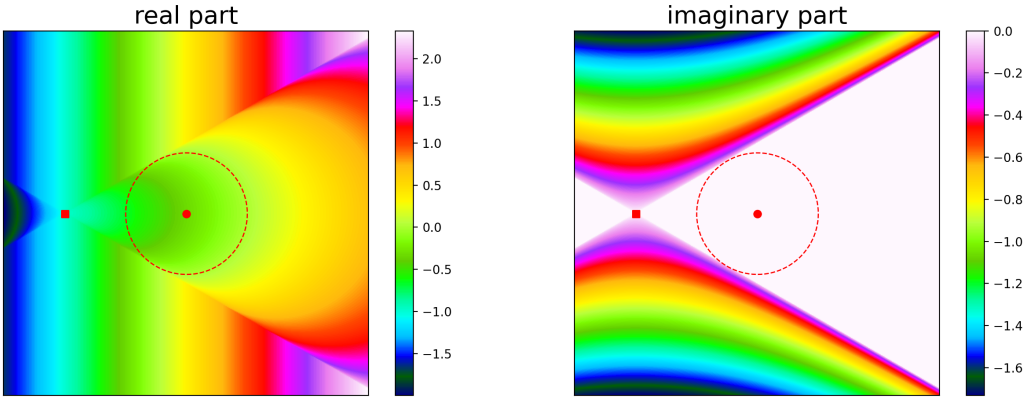


Figure 2.6: Representation of one root of the quadratic equation (2.16).

ξ and ξ_0 are respectively the dot and the square. The initial front is represented by a red circle.

satisfy the reaching time definition. This may not have a physical meaning, but it it of great use in numerical approximation.

Bare algorithm Notice that $T^r(y,x,l)$ is a function of $l-r$, so that $T^\alpha(x,y,\alpha) = T^0(x,y,0)$. For any given (x,t), we use Newton algorithm. The skeleton of the algorithm is the following.

Algorithm – Newton

`Determine a initial value` $\alpha^0 = \alpha^0(x,t)$.
`While (stopping criteria not satisfied)`
`Compute` $J(\alpha^n) := T^{\alpha_{\min}}(y(x,\alpha),\xi,\alpha) + T^0(x,y(x,\alpha),0) - t$

Compute

$$J'(\alpha^n) = \partial_a T^{\alpha_{\min}}\left(y(x,\alpha)^n, \xi, \alpha^n\right) + \left\langle \nabla_y T^{\alpha_{\min}}\left(y(x,\alpha^n), \xi, \alpha^n\right) + \nabla_x T^0\left(x, y(x,\alpha^n), 0\right), \partial_\alpha y(x,\alpha^n)\right\rangle$$

Set $\alpha^{n+1} = \alpha^n - J(\alpha^n)/J'(\alpha^n)$

The stopping criteria is triggered when $|J(\alpha^n)| \leqslant \varepsilon$ for a small fixed ε. In practice, we take $\varepsilon = 10^{-7}$.

This algorithm works well when $\|b\| \leqslant c$. In this case, reaching times are always real : any point can be reached by any moving front (the system is controllable). The uncontrollable case is theoritically similar, but turned out to be numerically more challenging, due to the existence of several solutions to the reaching time equation (2.15). We used *a-priori* bounds on α to enforce uniqueness of the solution, but the efficiency of Newton algorithm is greatly affected by the reparametrization. Since the implementation managed to meet our needs, we stopped the development at this point, but further work could consider using an interior point method to enforce box constraints on α with higher efficiency.

Numerical results The following examples are extracted from [?]. We first consider the controllable case, with parameters

$$\xi = (-3, 0, \dots, 0)^t, \quad d = (1, 0, \dots, 0)^t, \quad \|b\| = 0.5, \quad c = 1, \quad \alpha_{\min} = -1,$$
$$c_e = 1, \quad c_x = 1.5, \quad g_{\max} = 2, \quad g_{\min} = -2, \quad g_c = 4, \quad T = 7.$$

We use neural networks with 3 layers of 60 neurons with full ReLu activation function, and $N = 8$ time steps. The evolution of the 0-level set is given in figure (2.7). To measure the error between the reference solution and the approximation, we use L^∞ and L^1 errors on the 2d plan of representation. The "global" error refers to a computation on the whole plan, while the "local" error concentrates around the 0 level-set. More precisely,

$$e_d^1 \coloneqq \frac{\sum_{i=1}^K |u_{ex}(x_i,T) - U^N(x_i)| \mathbb{1}_{i\in\Omega_d}}{\sum_{i=1}^K \mathbb{1}_{i\in\Omega_d}} \sim \frac{|u_{ex}(\cdot,T) - U^N|_{L^1(\Omega_d)}}{|\Omega_d|},$$
$$e_d^\infty \coloneqq \max_{i\in[\![1,K]\!]} |u_{ex}(x_i,T) - U^N(x_i)| \sim |u_{ex}(\cdot,T) - U^N|_{L^\infty(\Omega_d)},$$

where $d \in \{\mathsf{loc}, \mathsf{glob}\}$, $\Omega_{\mathsf{glob}} = \Omega$ the computation domain intersected with the plan of representation, and $\Omega_{\mathsf{loc}} \coloneqq \{|u_{ex}| \leqslant 0.1\}$. Errors are given in table (2.2).

Parameters		Global errors		Local errors		Time
dimension	Stochastic gradient iterations	L_∞	L_1 rel.	L_∞	L_1 rel.	
2	50000	1.24e-01	2.94e-03	8.81e-02	5.21e-03	3h09
4	100000	2.49e-01	4.70e-03	8.67e-02	5.45e-03	6h51
6	400000	8.74e-01	3.70e-02	1.09e-01	1.01e-02	35h07

Table 2.2: Errors in the controllable case.

We turn to the non-controllable case, with new parameters

$$\xi = (-12, 0, \dots, 0)^t, \quad \|b\| = 1.0, \quad c = 0.5, \quad T = 16.$$

Errors are given in table (2.3). The propagation of the 0-level set is illustrated in figure (2.8).

Parameters		Global errors		Local errors		Time
dimension	Stochastic gradient iterations	L_∞	L_1 rel.	L_∞	L_1 rel.	
2	50000	2.66e-01	5.99e-03	1.19e-01	4.61e-02	3h02
4	100000	3.90e-01	6.77e-03	1.16e-01	2.69e-02	8h13
6	400000	9.69e-01	1.09e-02	1.78e-01	2.88e-02	35h20

Table 2.3: Errors in the non-controllable case.

More details, other examples, and a comparison with two other schemes are found in [**?**]. To conclude on this numerical section, we found that the lagrangian scheme is promising for physically high dimensions. In particular, it could be used to initialize gradient descent methods.

Two main difficulties are remaining: first, the minimization of the nonconvex nonlinear cost is a computational challenge. More efficient algorithms are needed. There is hope that the efforts deployed in understanding neural networks will provide such tools in a near future.

The second problem lies in the lack of regularity of the control. The discontinuities are penalizing both the approximation of feedback controls, and the recursive estimate of the error. To overcome it, one needs discontinuous spaces of approximation, and numerical schemes for ODEs with discontinuous left hand side of order stricly greater than 1. To our knowledge, such schemes are not currently available: we let this problem open for the sagacious reader.

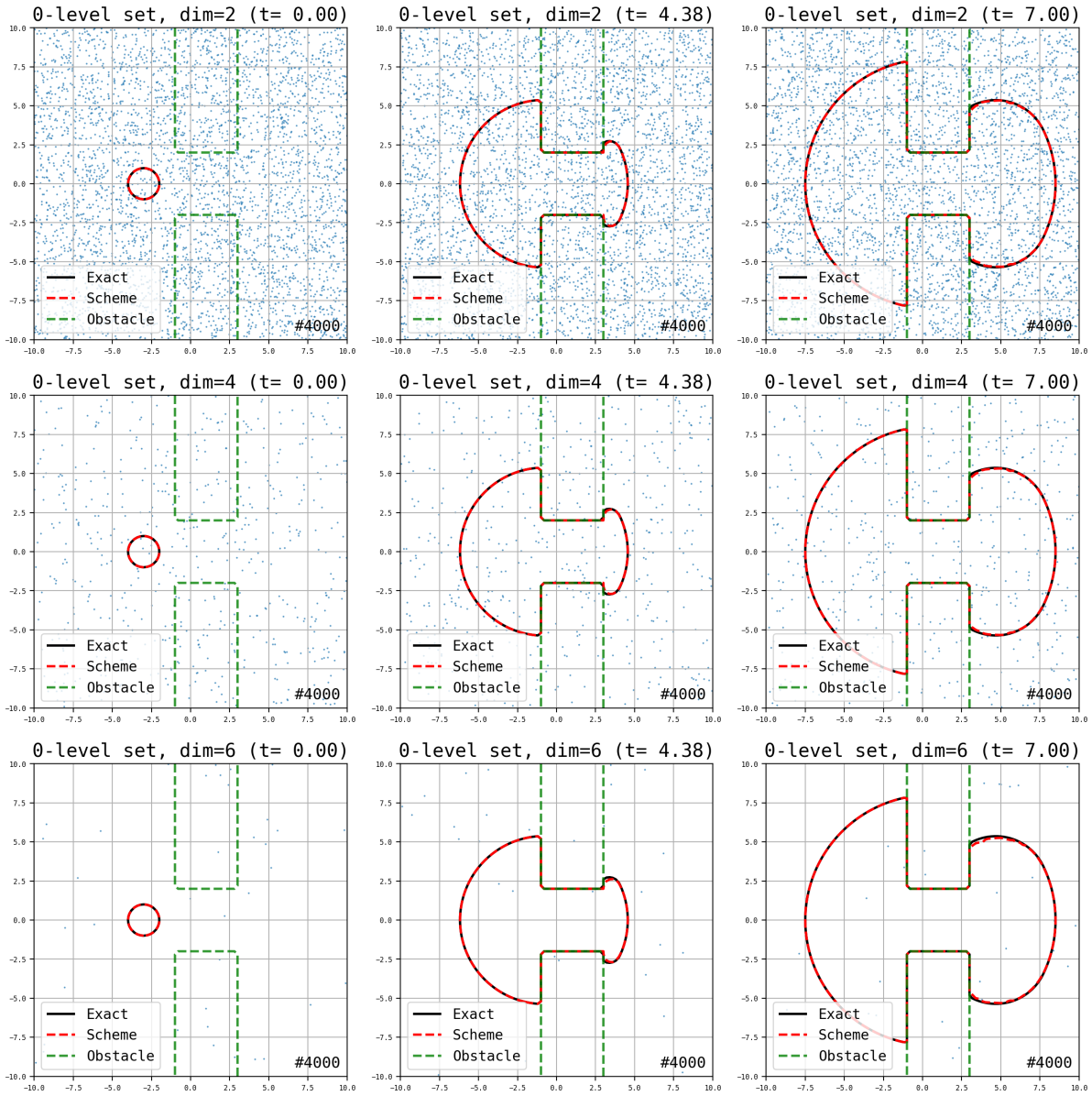


Figure 2.7: Results in the controllable case, in dimension 2, 4 and 6.

The scheme used is the full lagrangian scheme developped in section (2.2.1). The 0-level set of the exact solution is displayed in black, and the numerical approximation is plotted in red. Blue dots represents the location of the $M = 4000$ samples drawn to evaluate μ. In dimension higher than 2, only the points in a layer close to the 2D cut are plotted. As the dimension grows, the relative volume of this layer decreases, and less points are visible despite a constant cardinal.

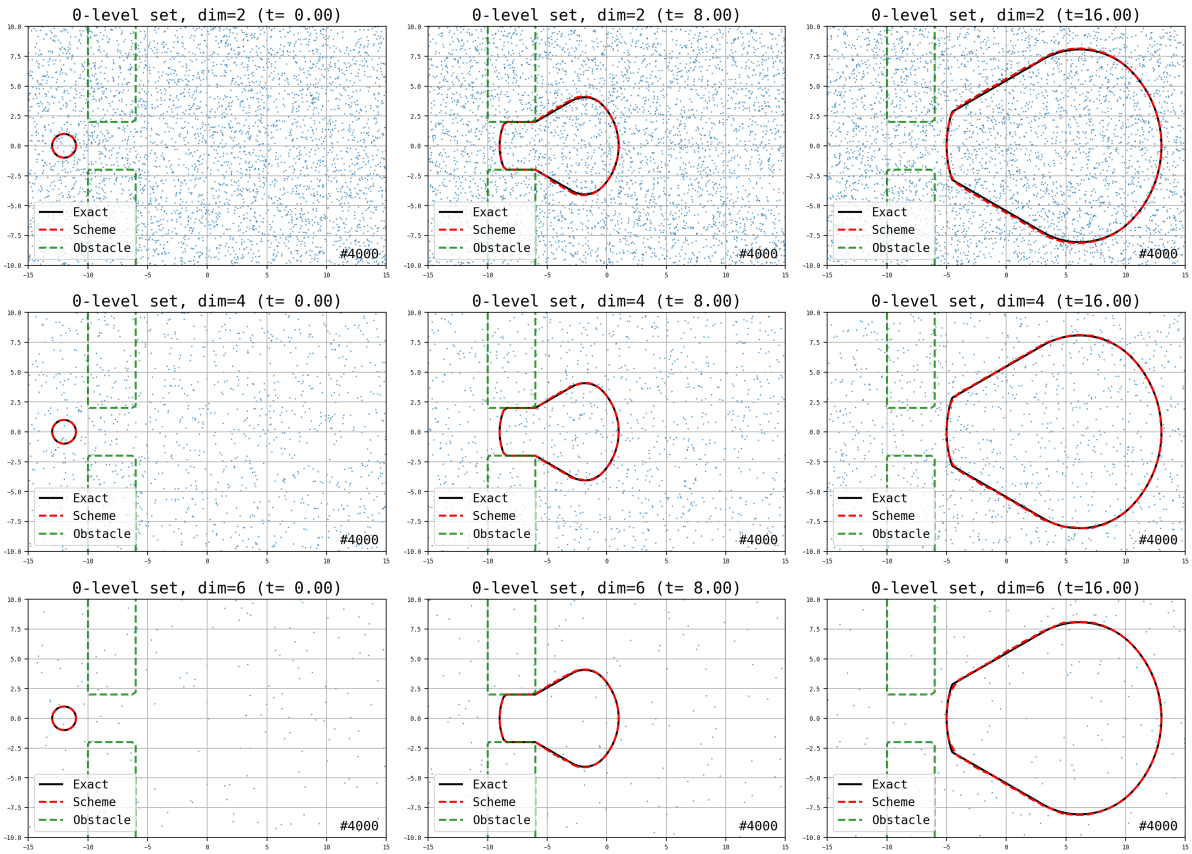


Figure 2.8: Results in the non-controllable case, in dimension 2, 4 and 6.

Here, the lack of controllability translates into nonreachable regions even outside the obstacle. For instance, a point near $(-5, \pm 5)$ *will not be reached by the moving front, contrary to the previous case.*

Conclusion

For the patient reader, this report may have been a reminder of some notions about Hamilton-Jacobi equations, a presentation of obstacle problems, and a introduction to one particular Lagrangian scheme using neural networks. For the author, it has been a joyful dive into this class of first-order equations, its lack of regularity, and nonlinear ideas to approximate it numerically. Fortunately, I now have more questions than at the beginning of this work, and there is hope to find beauty along the road towards the answers.

Bibliography

[ABZ13] Albert Altarovici, Olivier Bokanowski, and Hasnaa Zidani. A general Hamilton-Jacobi framework for non-linear state-constrained control problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(2):337–357, April 2013.

[BFZ10] Olivier Bokanowski, Nicolas Forcadel, and Hasnaa Zidani. Reachability and Minimal Times for State Constrained Nonlinear Problems without Any Controllability Assumption. *SIAM Journal on Control and Optimization*, 48(7):4292–4316, January 2010.

[Bok05] Olivier Bokanowski. Non monotone schemes for HJB equations. page 43, 2005.

[FA88] Aleksei Fedorovich Filippov and Felix Medland Arscott. Differential Equations with Discontinuous Righthand Sides, 1988.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York, New York, NY, 2002.

[HB] Christoph Hertrich and Amitabh Basu. Towards Lower Bounds on the Depth of ReLU Neural Networks. page 13.

[Hor91] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[Lag00] Frédéric Lagoutière. *Modélisation mathématique et résolution numérique de problèmes de fluides compressibles à plusieurs constituants*. PhD thesis, 2000.

[Lio82] P. L. Lions. *Generalized Solutions of Hamilton-Jacobi Equations*. Number 69 in Research Notes in Mathematics. Pitman, Boston, 1982.

[McE06] William M. McEneaney. *Max-plus Methods for Nonlinear Control and Estimation*. Systems and Control. Birkhäuser, Boston, 2006.

[Roe87] P. L. Roe. Upwind differencing schemes for hyperbolic conservation laws with source terms. In Claude Carasso, Denis Serre, and Pierre-Arnaud Raviart, editors, *Nonlinear Hyperbolic Problems*, volume 1270, pages 41–51. Springer Berlin Heidelberg, Berlin, Heidelberg, 1987.

[Swe84] P. K. Sweby. High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws. *SIAM Journal on Numerical Analysis*, 21(5):995–1011, October 1984.

[SX05] Chi-Wang Shu and Zhengfu Xu. Anti-diffusive High Order WENO Schemes for Hamilton-Jacobi Equations. *Methods and Applications of Analysis*, 12(2):169–190, 2005.

[TSB20] Ugo Tanielian, Maxime Sangnier, and Gerard Biau. Approximating Lipschitz continuous functions with GroupSort neural networks. 2020.