Using \( K \)-Branch Entropy Solutions for Multivalued Geometric Optics Computations

Laurent Gosse

Istituto per le Applicazioni del Calcolo (Sezione di Bari), via G. Amendola, 122/F-70126 Bari, Italy
E-mail: l.gosse@area.ba.cnr.it

Received May 5, 2001; revised April 16, 2002

This paper is devoted to a numerical simulation of the classical WKB system arising in geometric optics expansions. It contains the nonlinear eikonal equation and a linear conservation law whose coefficient can be discontinuous. We address the problem of treating it in such a way that superimposed signals can be reproduced by means of the kinetic formulation of “multibranch solutions,” originally due to Brenier and Corrias. Some existence and uniqueness results are given, together with computational test cases of increasing difficulty displaying up to five multivaluations.

1. INTRODUCTION

We aim at computing efficiently highly oscillating solutions for linear second-order dispersive PDEs; we focus in the sequel, for instance, on the one-dimensional Schrödinger equation. More precisely, we seek plane wave solutions of the form \( \psi(t, x) = A(t, x) \exp(i \phi(t, x)/\hbar), \) satisfying for \( x \in \mathbb{R} \) (see \([17a, 24a, 33]\))

\[
\begin{align*}
-\hbar \partial_t \psi + \frac{\hbar^2}{2} \partial_{xx} \psi + V(x) \psi &= 0, \\
\end{align*}
\]

where \( V(x) \) is some given smooth potential. For this model, one can consider the semi-classical limit which consists of tuning the signals’ wavelengths according to the Planck constant \( \hbar \) being sent to zero. Plugging this ansatz inside (1) leads to the following relation between the wave phase \( \phi \) and its amplitude \( A \):

\[
\begin{align*}
-A \left( \partial_t \phi + \frac{(\partial_x \phi)^2}{2} - V(x) \right) + \frac{i \hbar}{2} (2 \partial_t A + A \partial_{xx} \phi + 2 \partial_x A \partial_x \phi) + \frac{\hbar^2}{2} \partial_{xx} A &= 0.
\end{align*}
\]

1 Partially supported by the EEC TMR projects “Viscosity solutions and applications” #ERBFMRXCT980234 and “Asymptotic methods in kinetic theories” #ERBFMRXCT970157.
Nullifying the expressions related to the first two powers of $h$ gives the well-known WKB (after Wentzel, Kramers, and Brillouin) system (see, e.g., [34, 46, 51, 53] for more detailed presentations)

$$\partial_t \varphi + \frac{(\partial_x \varphi)^2}{2} = V(x), \quad \partial_t (A^2) + \partial_x (A^2 \partial_x \varphi) = 0. \tag{2}$$

The phase $\varphi$ evolves according to the so-called eikonal equation, which is nonlinear. In sharp contrast, the intensity of the plane wave $|\psi|^2 = A^2$ is ruled by a linear conservation law whose velocity field is given by the $x$-derivative of $\varphi$. This expresses intensity preservation between two integral curves of $\partial_x \varphi$ in the states space. The system (2) is weakly coupled, as the first equation can be solved independently.

A fundamental question raised by (2) concerns the sense in which $\varphi$ and $A^2$ have to satisfy it. Recently, a precise mathematical theory of viscosity solutions [19, 37] has been settled in order to state precise uniqueness results for general Hamilton–Jacobi equations. Unfortunately, this class of weak solutions does not provide a convenient framework in which to treat geometric optics problems. Indeed, it has been shown in [31] that interpreting the eikonal equation within such a theory allows concentrations to develop in the intensity $A^2$ on the shock lines of the phase. This is clearly in contradiction of the a priori estimates for (1).

In Section 2, we address the problem of proposing a convenient definition for the solutions of (2). In particular, we want to preserve some kind of superposition principle which should mimic the one holding for the original linear equation (1). It is precisely this property which is lacking with viscosity solutions, as it is cancelled by entropy conditions stating (roughly speaking) that only the quickest signal has to be reproduced beyond the time caustics that have appeared in the system (2). Thus we recall in Section 2.1 the geometric solutions of the one-dimensional inviscid Burgers equation

$$\partial_t u + u \partial_x u = 0, \tag{3}$$

which is deduced from (2) by $x$-differentiation and assuming that $V' \equiv 0$. Following [1], we rely on the method of characteristics, which are smooth curves $t \mapsto X(t)$ along which $U(t) = \partial_x \varphi(t, X(t))$ satisfies the following differential system:

$$\dot{U} = V', \quad \dot{X} = U.$$

Its Hamiltonian reads $\mathcal{H}(x, u) = \frac{1}{2} u^2 - V$. One can thus define a “particles density” $f(t, x, u)$ as follows: differentiation along the trajectories induced by $\mathcal{H}$ leads to Liouville’s equation,

$$\partial_t f + \partial_u \mathcal{H} \cdot \partial_x f - \partial_x \mathcal{H} \cdot \partial_u f = 0. \tag{4}$$

Denoting by $\mu = A^2$ the intensity, we may therefore consider as realistic solutions for (2) the ones described by the kinetic problem

$$\partial_t f + u \partial_x f + V'(x) \partial_u f = 0. \tag{5}$$

Unfortunately, this equation’s unknown involves one supplementary variable.
It is a classical strategy to approximate a kinetic equation by a moment system taking advantage of a privileged (and simple!) dependence of \( f \) upon its velocity variable \( u \). It is at this level that, in Section 2.2, we introduce the recent formalism proposed in [10]. As was observed in [8], the geometric solution to

\[
\partial_t u + u \partial_x u = 0, \quad u(0, x) = u^0 \geq 0
\]

is given exactly by the one to the free transport equation,

\[
\partial_t f + \xi \partial_x f = 0, \quad f(0, x, \xi) = H(u^0(x) - \xi)H(\xi),
\]

where \( H \) stands for the Heaviside function. But as soon as \( u^0 \) is not everywhere increasing, folds develop in finite time and a correct expression for \( f \) becomes

\[
f(t, x, \xi) = \sum_{k=1}^{K(t)} (-1)^{k-1} H(u_k(t, x) - \xi), \quad u_k > u_{k+1},
\]

where \( K(t) \) is the number of branches \( u_k \) present in the solution at time \( t \). One remarkable observation in [10] is that in the case of \( K(t) \leq K \), the exact solution of (7) can be recovered at any time \( t > 0 \) from a moment system of \( K \) equations involving only the \( t, x \) variables. Relying on an entropy minimization principle, it is possible to close such a system at any time \( t \) by finding an expression for the \( K + 1 \)th moment knowing the \( K \) preceding ones under the constraint of satisfying some entropy conditions. Solutions to such a hyperbolic system give back the \( u_k(t, x) \), which are called \( K \)-branch entropy solutions to (6). This is equivalent to the adjunction of a singular source term in the right-hand side of (7):

\[
\partial_t f + \xi \partial_x f = (-1)^{K-1} \partial_x^K \tilde{m}, \quad f(t, x, \xi) = \sum_{k=1}^{K} (-1)^{k-1} H(u_k(t, x) - \xi).
\]

The nonnegative measure \( \tilde{m} \) expresses the fact that a large family of specific convex entropies is dissipated; it is zero as long as the geometric solution to (3) has not been folding more than \( K \) times. In the case of \( K = 1 \), it coincides with the one considered in the kinetic formulation of [39]: therefore a “1-branch solution” is but the Kružkov entropy solution to (3) or, equivalently, the viscosity solution to the eikonal equation. From this perspective, this article generalizes the previous work [31]. A possible interpretation we propose for (2) in this one-dimensional framework is thus given by the system

\[
\partial_t \tilde{m} + \partial_x F_K(\tilde{m}) = 0, \quad \partial_t \mu_k + \partial_x (u_k \mu_k) = 0, \quad k = 1, \ldots, K,
\]

where \((\tilde{m}, \tilde{u}) \in \mathbb{R}^K \times \mathbb{R}^K\) are, respectively, the \( K \) first moments from (8) and their associated intensities.

In Section 2.3, we state a few theoretical results for the weakly coupled system (9) in case the branches \( u_k \) are endowed with some one-sided Lipschitz continuity. In particular, the \( L^p \) solutions constructed in [10, 56] do not permit application of the theory of [7] to tackle the \( K \) linear conservation equations. It turns out that the Jacobian \( F'_K \) can always be diagonalized using the Riemann invariants \( u_k \); the system is genuinely nonlinear, rich in the sense of [50], and decouples (for smooth solutions) into a set of \( K \) Burgers equations. For weak solutions
obtained as limits of Glimm-type approximations, some OSLC estimates are stated in [15] (see also [41]): any \( K \)-branch, entropy solution \( u_k \) is therefore asked to satisfy (3) in the smooth case, with Rankine–Hugoniot being supplemented with Olešćnik-type estimates when shocks are present. These constraints are actually the main tools both in proving uniqueness and stability results and in hinting at satisfying initializations for (9) as \( K \geq 2 \).

Section 3 is devoted to numerical implications for this kinetic approach. An essential difficulty in using the results of [10] in practical computations was initializing correctly \( f \) in (8). On the one hand, at time \( t = 0 \), the physical conditions give only a value corresponding to the first branch of phase, as only this one appears; on the other hand, system (9) can be sensitive to the choice of the two or three other ones it can handle [47]. Observing that \( K \)-branch solutions can be split into several solutions of Burgers equations relying on the finite superposition principle written in [48] sheds some light on this issue; of course, the initial monovalued intensity has to be also treated accordingly. This idea appears also in [4] for deriving a different numerical strategy. Other troubles come from the fact that for \( K > 2 \), the expression of \( F_K \) becomes so intricate it is only implicitly defined by means of the roots of some polynomial equations, so one has to give up using common Riemann-based numerical schemes: one alternative can be to use high-order central schemes [43].

The linear conservation laws are handled by means of the numerical algorithms introduced in [29]. Four test cases are presented in Sections 3 and 4: a smooth caustic (fold), two cusps of increasing difficulty, and one involving singularities’ interactions [17, 32]. Finally, we present in Section 4 the full computation of a convex lens taken from [21, 23, 24, 47, 48] and a short but classical nonhomogeneous test case is carried out in the Appendix.

We close this introduction by mentioning that other numerical approaches have been developed to treat efficiently the geometric optics system (2), mostly relying on the theory of viscosity solutions: we refer the reader to, e.g., [3

A recent survey of kinetic formulations is given in [54].

2. GEOMETRIC SOLUTIONS AND KINETIC FORMULATION

2.1. Blowup Loci for the Burgers Equation

In this section, following [1, 17], we are concerned with classical solutions to (6) in the neighborhood of some given point \( \bar{x} \in \mathbb{R} \). We shall always assume that \( \partial_x u^0(\bar{x}) < 0 \).

Defining the straight characteristic lines \( X = x_0 + t \cdot u^0(x_0) \) parametrized by \( (t, x_0) \), one sees that \( u \) remains constant along them. Moreover, \( \partial_t u \) taken along \( t \mapsto X(t, x_0) \) evolves according to the Riccati equation \( \dot{q} + q^2 = 0 \) and blows up for \( t = -(\partial_x u^0(x_0))^{-1} \). The blowup locus is

\[
\gamma = \{ (t, x) \in \mathbb{R}^+ \times \mathbb{R}; \; x = x_0 + t \cdot u^0(x_0) \text{ and } t = -(\partial_x u^0(x_0))^{-1} \}.
\]

(10)

Two generic cases occur in one space dimension.

(i) \( \partial_{x,t} u^0(x_0) \neq 0; \) \( \gamma \) is a smooth curve and the characteristics are tangent to it. By definition (10), the variation of \( (t, x) \in \gamma \) is \( (x', \dot{x}) = \left( \frac{\partial x}{\partial x_0}, \frac{\partial u}{\partial x_0} \right) (u^0(x_0), 1) \).

(ii) \( \partial_{x,t} u^0(x_0) = 0, \) \( \partial_{x,x} u^0(x_0) > 0; \) \( \gamma \) has a singularity pointing downward and there holds, \( (x', \dot{x}) = \left( \frac{\partial x}{\partial x_0}, \frac{\partial u}{\partial x_0} \right) (u^0(x_0), 1) \).
Therefore, \( \gamma \) is a local characteristics envelope: blowup occurs where two initially close characteristics touch each other. Case(i) is commonly referred to as a fold caustic and case (ii) a cusp caustic.

Let us consider the map (see also Section 2 in [17] and Section 5 in [32]):

\[
(t, x_0) \mapsto \Phi(t, x_0) = (t, \phi(t, x_0)), \quad \phi(t, x_0) = x_0 + t \cdot u^0(x_0), \quad v(t, x_0) = u^0(x_0).
\]  \( (11) \)

The characteristics method reduces to

\[
u(\Phi(t, x_0)) = v(t, x_0) = u^0(x_0), \quad \gamma = \Phi((t, x_0); \partial_{x_0} \phi(t, x_0) = 0)).
\]

**THEOREM 2.1 (Blowup Solutions [1]).** Suppose there exists a connected open set \( D \subset \mathbb{R}^+ \times \mathbb{R}^+ \) and \( \Psi \in C^0(\overline{D}) \) such that

\[
\Phi(\Psi(t, x)) = (t, x), \quad det(\Phi'(\Psi(t, x))) \neq 0 \text{ in } D.
\]

Then \( u(t, x) = v(\Psi(t, x)) \) is a \( C^2 \) solution of (6) in \( D \) called a “blowup solution.”

In the “fold case,” there are two possible choices, say \( \Psi^\pm \), and both are tangent to \( \gamma \). Inside the cusp, there are three possibilities for \( \Psi_k = 1, 2, 3 \). We shall speak accordingly about fold solutions or cusp solutions when dealing with the concrete examples of Sections 3 and 4. There exists also a very simple situation which could be called the point solution: suppose \( u^0(x) = ax + b, (a, b) \in \mathbb{R}^- \times \mathbb{R} \). Then the blowup locus \( \gamma \) reduces to the unique point \( (x = -b/a, t = -1/a) \), which is a prototype for a “hot focus point” in the sense of [51]. We do not develop further this case.

### 2.2. \( K \)-Branch Entropy Solutions to the Burgers Equation

In this section, we recall the construction of nonnegative “multibranch solutions” to (6) through a kinetic formulation: for detailed proofs of all the stated results, we refer the reader to [10] (see also [11, 39, 40]).

Let \( K \in \mathbb{N} \). We denote

\[
\Theta_K := \{0 \in C^0(\mathbb{R}) \cap \mathbb{R}; \partial^K \theta(\xi) \geq 0 \text{ (D')}\}
\]

and

\[
\mathcal{F}_{L>0} := \{f \in L^{\infty}, 0 \leq f \leq 1 \text{ a.e. with } Supp(\xi) \subset [0, L]\}.
\]

Each \( f \in \mathcal{F}_L \) induces a moments vector \( \tilde{m}(f) \in \mathbb{R}^K \) whose components read

\[
m_k(f) = \int_{\mathbb{R}^+} \xi^{k-1} f(\xi) \cdot d\xi, \quad k = 1, \ldots, K.
\]  \( (12) \)

It is therefore possible to define the set of “realizable moments"

\[
\mathcal{M}^L_K = \{\tilde{m} \in \mathbb{R}^K; \exists f \in \mathcal{F}_L \text{ such that } \tilde{m} = \tilde{m}(f)\}.
\]
onto which one treats the following minimization problem:

\[
J_k^*(\tilde{m}) = \inf_{f \in \mathcal{F}_L} \left\{ \int_{\mathbb{R}^+} \theta(\xi) f(\xi) \cdot d\xi \middle| \tilde{m}(f) = \tilde{m} \in \mathcal{M}_K^L \right. \text{ and } \theta \in \Theta_K \right\}.
\]

That is to say, one optimizes simultaneously the values of all possible \(\theta\)-moments among densities \(f \in \mathcal{F}_L\) once having fixed their \(K\) first moments (12) inside \(\mathcal{M}_K^L\). It is shown in [10] that for any \(\tilde{m} \in \mathcal{M}_K^L\), there exists a unique solution to this problem which is called the \(K\)-branch Maxwellian (see [56] for a precise definition). Moreover, it is independent of the choice of \(\theta \in \Theta_K\) and reads

\[
\mathcal{M}_{K,\tilde{m}}(u_1, \ldots, u_K, \xi) = \sum_{k=1}^K (-1)^{k-1} H(u_k - \xi), \quad u_k > u_{k+1} \geq 0,
\]

with \(H\) standing for the Heaviside function. Thus we can write the expression of the realizable moments: they are solutions of a nonlinear Vandermonde algebraic system. We can consider the map \(\tilde{m}, [0, L]^K \rightarrow \mathcal{M}_K^L\),

\[
m_j(u_1, \ldots, u_K) := \frac{1}{j} \sum_{k=1}^K (-1)^{k-1}(u_k)^j, \quad i = 1, \ldots, K,
\]

and \(\tilde{m}\) realizes a one-to-one \(C^\infty\) mapping of the \(u_k\)’s as long as \(u_k > u_{k+1}\) for all \(j\) under consideration. This leads to the following definition.

**DEFINITION 2.1 (K-Multivalued Solutions [10]).** We call a K-multivalued solution any measurable function \([0, 1] \ni f(t, x, \xi)\) on \(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+\) satisfying in the sense of distributions the kinetic equation

\[
\partial_t f + \xi \partial_x f = (-1)^{K-1} \delta^K \tilde{m}, \quad f(t, x, \xi) = \mathcal{M}_{K,\tilde{m}(f)},
\]

where \(\tilde{m}\) is some nonnegative Radon measure on \(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+\).

Existence results for these \(K\)-multivalued solutions are provided in [10] by means of BGK approximations and also in [56], which relies on a more singular transport-collapse procedure [8]. Uniqueness for these very weak solutions has been recently proved only for special \(K\)’s [55, 57] without using any BV theory.

**THEOREM 2.2 (Existence of K-Multivalued Solutions [10]).** For all \(f_0(x, \xi) = \mathcal{M}_{K,\tilde{m}(f_0)}, f_0 \in \mathcal{F}_L \cap L^1(\mathbb{R} \times [0, L])\), there exist \((f^\epsilon)_{\epsilon \rightarrow 0} \in \mathcal{F}_L\) and nonnegative bounded Radon measures \((\tilde{m}^\epsilon)_{\epsilon \rightarrow 0}\) on \(\mathbb{R}^+ \times \mathbb{R} \times [0, L]\) such that

\[
f^\epsilon(t = 0, \cdot, \cdot) = f_0, \quad \partial_t f^\epsilon + \xi \partial_x f^\epsilon = \frac{1}{\epsilon} \left( \mathcal{M}_{K,\tilde{m}(f^\epsilon)} - f^\epsilon \right) \overset{\text{def}}{=} (-1)^{K-1} \delta^K \tilde{m}^\epsilon;
\]

along a subsequence if necessary, \(f^\epsilon \rightharpoonup f \in L^\infty(\mathbb{R}^+; L^1 \cap L^\infty(\mathbb{R} \times [0, L]))\) in \(L^\infty\) weak-*, and \(\tilde{m}^\epsilon \rightharpoonup \tilde{m}\) weakly. Moreover \(f\) is Maxwellian; that is, \(f = \mathcal{M}_{K,\tilde{m}(f)}\).

In the special cases of \(K = 1\) and 2, these statements coincide with the ones written in [8, 39, 40]. Indeed, such a “2-branch kinetic formulation” is equivalent to the weak entropy solutions of the isentropic Euler system for gas dynamics with the adiabatic exponent \(\gamma = 3\); see [56, 57] for further analysis of this model.
The same way as in [39, 40], we get equivalence between this kinetic formulation and the hyperbolic system ruling the time evolution of the realizable moments.

**THEOREM 2.3** (Brenier and Corrias [10]). A measurable function \( f(t, x, \xi) \in \mathcal{F}_{t>0} \) solution of (15) is a \( K \)-multivalued solution if and only if the following entropy inequalities hold in \( D' \):

\[
\forall \theta \in \Theta_K, \quad \partial_t \int_{\mathbb{R}^+} \theta(\xi) f(t, x, \xi) \cdot d\xi + \int_{\mathbb{R}^+} \xi \theta(\xi) f(t, x, \xi) \cdot d\xi \leq 0. \tag{16}
\]

Moreover, \( \tilde{m}(f)(t, x) \) is a weak entropy solution of the \( K \times K \) hyperbolic system:

\[
\partial_t m_k + \partial_x m_{k+1} = 0, \quad m_{K+1}(t, x) = J^0_K(\tilde{m}) = \int_{\mathbb{R}^+} \xi^K \mathcal{M}_{K,\tilde{m}}(\xi) \cdot d\xi;
\]

\[
\partial_t J^0_K(\tilde{m}) + \partial_x Z^0_K(\tilde{m}) \leq 0 \text{ with } \tilde{m} \mapsto J^0_K(\tilde{m}) \text{ convex.} \tag{17}
\]

The entropy flux is given by \( Z^0_K(\tilde{m}) = \int_{\mathbb{R}^+} \xi \theta(\xi) \mathcal{M}_{K,\tilde{m}}(\xi) \cdot d\xi. \)

For any function \( \theta \in \Theta_K, J^0_K(\tilde{m}), Z^0_K(\tilde{m}) \) provide a Lax entropy/entropy flux pair. The fact that \( \tilde{m} \mapsto J^0_K(\tilde{m}) \) is convex can be shown in the following way: let us take some \( \alpha \in [0, 1] \). We consider two moment vectors \( \tilde{m} \neq \tilde{m}' \) in \( \mathbf{M}_K^\prime \) and form

\[
f(\xi) = \alpha \mathcal{M}_{K,\tilde{m}} + (1 - \alpha) \mathcal{M}_{K,\tilde{m}'} = \int_{\mathbb{R}^+} \xi^{k-1} f(\xi) \cdot d\xi = (\alpha \tilde{m} + (1 - \alpha) \tilde{m}').
\]

By the very definition of the minimizer \( J^0_K \), one gets for any \( \theta \in \Theta_K \)

\[
J^0_K(\alpha \tilde{m} + (1 - \alpha) \tilde{m}') \leq \int_{\mathbb{R}^+} \theta(\xi) f(\xi) \cdot d\xi \leq \alpha J^0_K(\tilde{m}) + (1 - \alpha) J^0_K(\tilde{m}').
\]

System (17) is genuinely nonlinear and rich in the sense of [50] since it diagonalizes in Riemann coordinates. For smooth solutions, the \( u_k \)'s appearing in (13) are strong Riemann invariants for (17) and each one satisfies (3). System (17) is strictly hyperbolic if and only if they are all distinct; in this last case, \( u_k > u_{k+1} \) and the map \( \tilde{m} \) (14) realize a diffeomorphism; see also Theorem 12.1.1 in [50].

**DEFINITION 2.2** (\( K \)-Branch Entropy Solutions). A \( K \)-branch entropy solution to (6) is any set \( \tilde{u} \) of \( K \) nonnegative measurable functions \( u_k(t, x) \in L^\infty(\mathbb{R}^+ \times \mathbb{R}) \) such that \( u_k > u_{k+1} \) and for which (16) holds.

The relevance of this notion partly comes from the next result.

**THEOREM 2.4** (Finite Superposition Principle [48]). Suppose \( u_k(t, x), k = 1, \ldots, K \) is a set of weak entropy solutions to

\[
\partial_t u_k + \partial_x \left( \frac{u_k^2}{2} \right) = 0, \quad u_k(0, x) = u_k^0 > u_{k+1}^0 \geq 0,
\]

with each one being continuous and of locally bounded variation in \( x \in \mathbb{R} \). Then \( \tilde{u} := (u_k)_{k=1}^{K} \) is a \( K \)-branch entropy solution to (6).
For $K > 1$, $K$-branch entropy solutions generalize Kružkov’s notion of bounded entropy solution to (6); namely Definition 2.2 asks $\tilde{m}$ to dissipate the whole family of convex entropies in the form $J_\theta^K(\tilde{m})$ with $\theta \in \Theta_K$. The link with the geometric framework recalled in the preceding section lies in the fact that if $K \in \mathbb{N}$ is big enough, $u_k(t, x)$ and a blowup solution $v(\Psi_k(t, x))$ should coincide for a given $k \leq K$ on some common domain of definition where both satisfy the free transport equation (7); see also Theorem 3.5 in [10] and Section 4 in [51].

2.3. Some Stability Results “à la Oleinik”

We now give a correct mathematical sense to the “coupled system,” describing time evolution for both the $K$ moments $\tilde{m}$ and the associated intensity vector $\tilde{\mu} := (\mu_k)_{k=1,...,K}$ (see (9)),

$$\partial_t \tilde{m} + \partial_x F_K(\tilde{m}) = 0, \quad \partial_t \tilde{\mu} + \partial_x (\tilde{u} \cdot \tilde{\mu}) = 0,$$

for any $K \in \mathbb{N}$. The intensities $\mu_k$ are solutions to linear conservation laws with possibly discontinuous coefficients: in order to apply the results from [7] (see also [45]), we need both a uniform $L^\infty$ bound and some one-sided Lipschitz estimate of each $u_k(t, \cdot)$. Unfortunately, this kinetic formulation does not provide so much space regularity for $K$-branch entropy solutions, [10, 56].

An alternative for constructing stronger solutions $\tilde{m}$ to (17) is the Glimm scheme [25]. However, we face a rather unusual obstacle in this as it is not clear that the flux functions $\tilde{m} \mapsto F_K(\tilde{m})$ are smooth enough for $K > 2$. Let us illustrate this point. According to (13), we get for $K = 2$

$$m_1 = u_1 = u_2, \quad m_2 = \frac{1}{2}((u_1)^2 - (u_2)^2).$$

To close the “2-moment system,” we must express the third moment $m_3$ by means of $m_1$, $m_2$. Following [10, 40], we notice that as soon as $m_1 > 0$,

$$u_1 = \frac{m_2}{m_1} + \frac{m_1}{2}, \quad u_2 = \frac{m_2}{m_1} - \frac{m_1}{2},$$

which implies that

$$m_3 = \frac{1}{3}((u_1)^3 - (u_2)^3) = \frac{(m_2)^2}{m_1} + \frac{(m_1)^3}{12}.$$

What saves the day is noticing that in this case, $u_1, u_2$ are solutions of an invertible linear algebraic system since

$$\frac{m_2}{m_1} = \frac{1}{2}(u_1 + u_2), \quad m_1 = u_1 - u_2.$$

Clearly, such a feature is already lost for $K = 3$. Indeed in [48], Runborg conjectures that deducing $\tilde{u}$ from $\tilde{m}$ amounts to solving in general two polynomal equations of degree $\frac{K}{2}$ (for $K$ even) or $\frac{K+1}{2}$ (for $K$ odd). The dependence of these roots (when they exist) upon the coefficients which are rational functions of the $m_k$’s has no reason to be smooth in large
domains of $\mathbb{R}^K$. Another observation is that the first $K - 1$ components of $F_k(\vec{m})$ are the $K - 1$ last ones of $\vec{m}$, but

$$
\nabla \vec{m} J_K^{(\xi)}(\vec{m}) = \partial_{\vec{m}} \left( \sum_{k=1}^{K} (-1)^{k-1} \int_{\mathbb{R}^+} \xi^K H(u_k - \xi) \cdot d\xi \right)
= K((u_1)^{K-1}, \ldots, (u_K)^{K-1}) \partial_{\vec{m}} \vec{u}.
$$

Where $\partial_{\vec{m}} \vec{u}$ is the inverse of the Vandermonde matrix (14), which is known to be badly conditioned if $K$ is big. So, following [15, 25], we make a hypothesis:

$(\bigstar)F_K$ is defined and $C^2$ is inside an open ball $\Omega \subset M_K^L$ for some $L > 0$. Inside this ball, the map $\vec{m}$ (14) is a diffeomorphism, and system (17) is strictly hyperbolic, genuinely nonlinear, and rich in all its characteristic fields.

In other words, $(\bigstar)$ asks the eventual weak solutions to (17) to remain inside a set of realizable moments $M_K^L$ in which $u_k > u_{k+1}$ for $k < K$ and $K > 1$. Indeed, if $K = 2$, keeping $u_1 > u_2 \geq 0$ is equivalent to $m_2 > m_1^2$, $m_1 > 0$, and this can be violated in a Riemann problem with a large enough initial jump since it corresponds, for instance, to the vacuum appearance in the isentropic Euler equations.

**Remark.** This problem is identified in [38] as the “loss of realizability of predicted moments.” It appears also in [6] as both the assumption (2.9) and the convexity assumption (CH2) which asks $\vec{m}$ to be a diffeomorphism. See also the assumptions of the main results in [20].

Under hypothesis $(\bigstar)$, (17) generates a standard Riemann semigroup $S$ on a closed domain $D_d \subset L^1 \cap BV(\mathbb{R}; \mathbb{R}^K)$ enjoying the following properties [13]:

(i) For all $\vec{m}_0, \vec{m}_0'$ in $D_d$ and $t > s > 0$

$$
\|S_t \vec{m} - S_s \vec{m}'\|_{L^1(\mathbb{R})} \leq R|t-s| + \|\vec{m}_0 - \vec{m}_0'\|_{L^1(\mathbb{R})}, \quad R \in \mathbb{R}^+.
$$

(ii) Every trajectory $t \mapsto \vec{m}(t, \cdot) = S_t \vec{m}_0$ is a weak entropy solution in the sense of Lax to (17), called standard Riemann solution (SRS).

(iii) Any weak solution obtained as a limit of Glimm or wavefront tracking approximations is a standard Riemann solution.

The positively invariant domain has the usual form

$$
D_d = cl_{L^1}[\vec{m} \in L^1(\mathbb{R}; \mathbb{R}^K) \text{ piecewise constant; } V(\vec{m}) + C_0 \|\vec{u}_0 - \vec{\bar{u}}\|_{L^\infty} Q(\vec{m}) \leq d];
$$

for some constants $C_0, d$ strictly positive. The functionals $V$ and $Q$ are the classical linear and quadratic ones used to control the variations of the Glimm approximations. $\vec{u}_0$ stands for the initial values from which $\vec{m}_0$ is deduced and $\vec{\bar{u}} \in (\mathbb{R}^+)^K$ is a trivial constant $K$-branch entropy solution such that $\vec{m}$ (with obvious notation) belongs to the domain of definition assumed in $(\bigstar)$.

We notice that when they exist, standard Riemann solutions admit the kinetic formulation (15) since the Lax shock conditions imply the entropy inequalities in (16). Still, we can speak also in this context about $K$-branch entropy solutions.

Taking into account the richness in all the characteristic fields of (17), we can rewrite the decay estimates for the positive waves from [15, 16, 41] in a simpler form, as the left
eigenvectors $\ell_k$ of $F_K$ are the gradients of the (strong) Riemann invariants $u_k$. Therefore, the signed measures read simply

$$v_k = \ell_k^T \cdot \partial \bar{m} = \partial_t u_k, \quad k = 1, \ldots, K.$$  

And we get that in the sense of measures on $\mathbb{R}$ and for some $\kappa > 0$,

$$\partial_t u_k(t, \cdot) \leq \frac{1}{\kappa t} + O(1)Q(\bar{m}_0) \quad \mathcal{M}(\mathbb{R}),$$  

for any standard Riemann solution obtained as a limit of Glimm or wavefront tracking approximations [14, 41]. This is enough to state a uniqueness result.

**Theorem 2.5.** Let $K \in \mathbb{N}$ and assume there exists $\tilde{u} \in \mathbb{R}^K$ a constant $K$-branch entropy solution to (6) in the neighborhood of which the hypotheses (♠) are satisfied. Suppose moreover that the initial data are chosen such that it holds that

$$\partial_t u_k^0 \leq C < +\infty \quad \mathcal{M}(\mathbb{R}), \quad k = 1, \ldots, K.$$  

Then, for $d > 0$ small enough and any $(\bar{m}_0, \bar{\mu}_0) \in D_d \times \mathcal{M}(\mathbb{R}; \mathbb{R}^K)$, there exists a unique pair $(\tilde{m}, \bar{\mu})(t, \cdot) \in D_d \times \mathcal{M}(\mathbb{R}; \mathbb{R}^K)$ of SRS/duality solutions to (18), $t > 0$.

**Proof.** Let us first assume that the Glimm Scheme for (17) remains inside an open domain around $\bar{m}$ where (♠) holds true. Then the decay of the Glimm functional implies the existence of a uniform constant $C \in \mathbb{R}^+$ such that for any $t > 0$,

$$TV_x(\bar{m}(t, \cdot)) \leq C \cdot TV_x(\bar{m}_0), \quad \|\bar{m}(t, \cdot) - \bar{m}\|_{L^\infty} \leq C \cdot \|\bar{m}_0 - \bar{m}\|_{L^\infty},$$  

which are bounded as $\bar{m}_0 \in D_d \subset L^\infty(\mathbb{R}; \mathbb{R}^K)$. One can therefore always adjust $d > 0$ to keep $\bar{m}$ inside a small ball where (♠) holds and ensures the existence of a unique standard Riemann semigroup for which the estimates (19) are valid. Thus we get existence and uniqueness of the duality solutions $\bar{\mu}$ by applying Theorem 4.2.5 in [7], thanks to the smoothness and the invertibility of the map $\bar{m}$ (14).

**Corollary 2.1.** Let $K = 2$; there exists $d > 0$ such that the same conclusion holds for any pair $(\bar{m}_0, \bar{\mu}_0) \in D_d \times \mathcal{M}(\mathbb{R}; \mathbb{R}^2)$ satisfying (20).

**Proof.** In this special case, $J(\ell_k^T)\bar{m}$ is known explicitly and $C^\infty$ is known for $m_2 > \frac{(m_1)^2}{2}, m_1 > 0$. Thus $\tilde{u}$ can be any constant 2-branch entropy solution in the sense of Definition 2.2 and it is enough to ask for $TV_x(\tilde{\mu}_0)\|\tilde{u}_0 - \tilde{u}\|_{L^\infty}$ small thanks to the richness of system (17) and the regularity of the map $\bar{m}$ (14).

**Remark.** Theorem 2 in [15] implies that the estimates (19) are sufficient for any weak solution to (17) to coincide with a trajectory of the semigroup $S$ and therefore to define the intensities $\bar{\mu}$ in a unique way in the sense of duality [7, 31]. Remark 4.2.6 in [7] ensures that $\bar{\mu}_0 \geq 0$ implies $\bar{\mu}(t, \cdot) \geq 0$ for all $t > 0$.

The uniqueness result of [15] in this last case is stronger than the one of [57], as it contains Lipschitz estimates, but its range of application is narrower since it is concerned with the “small solutions” of $D_d$. As noted in [57], the system (17) does not belong to the Temple class. Anyway, OSLC regularity is necessary for [7, 45] and this implies $BV_{loc}$ smoothness in space for the $K$-branch entropy solutions.
In the case of $K = 1$, we recover part of the theory developed in [31] doing a by-product of the Olešnik estimates [44] (or semiconcavity stability [19, 37]) with the results of [7]. Since no wave interactions occur and $F_1$ is a $C^\infty(\mathbb{R})$ strictly convex map, we have uniqueness and stability for very general initial data and also a broad class of numerical approximations.

Numerical experience suggests that for reasonable $\bar{u}$’s, convenient smoothness domains for $F_3$ do exist. Spurious complex roots appear generally together with wild oscillations before the general breakdown of the scheme when too-big CFL numbers are prescribed.

3. NUMERICAL STUDY OF HOMOGENEOUS PROBLEMS

3.1. General Numerical Discretizations

Starting from here, we introduce a uniform Cartesian grid defined by the positive parameters $\Delta x$ and $\Delta t$ standing, respectively, for the mesh size and the time step. We denote, for $(j, n) \in \mathbb{Z} \times \mathbb{N}$, $x_j = j \Delta x$, $x_{j+1/2} = (j + 1/2) \Delta x$, $t^n = n \Delta t$, and a generic computational cell

$$T^n_j = [t^n, t^{n+1}] \times [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}].$$

As usual, the parameter $\lambda$ refers to $\Delta t/\Delta x$. In the present paper, we do not try to give rigorous convergence results for such sophisticated models describing multiphase phenomena. We rather suggest some possible numerical approaches in order to simulate efficiently the $K$-branch entropy solutions studied in the preceding section. For an easy introduction to the numerics of one-phase problems, we refer the reader to [30]. Concerning the precise form of the flux functions $F_K$, $2 \leq K \leq 4$, the reader can consult the appendices in [47, 48].

First, we give some notation: for a given $K > 1$, the grid functions $\bar{m}^n_j$, $\bar{u}^n_j$ stand for some numerical approximations of $\bar{m}(t^n, x_j)$ and $\bar{u}(t^n, x_j)$ on each $T^n_j$. Then, following [31], we introduce the basic Lax–Friedrichs (LxF) scheme which reads

$$\bar{m}^{n+1}_j = \frac{1}{2} (\bar{m}^n_{j+1} + \bar{m}^n_{j-1}) - \frac{\lambda}{2} (F_K(\bar{m}^n_{j+1}) - F_K(\bar{m}^n_{j-1})), \quad K \in \mathbb{N}. \tag{21}$$

However, in supersonic regimes, we shall also consider the Godunov scheme, which, for this special case, reduces to the simple upwind scheme. Assuming the $u_k$’s are all positive, as in the preceding section, it reads

$$\bar{m}^{n+1}_j = \bar{m}^n_j - \lambda (F_K(\bar{m}^n_j) - F_K(\bar{m}^n_{j-1})) \tag{22}$$

and contains already far less numerical viscosity than the LxF one (21). More-accurate approximations can come out of a second-order scheme like the Nessyahu–Tadmor (NT) one [43]. It can be written in a predictor–corrector form as follows:

$$\bar{m}^{n+\frac{1}{2}}_j = \bar{m}^n_j - \frac{\lambda}{2} (F'_K)^n_j, \tag{23}$$

$$\bar{m}^{n+1}_j = \frac{1}{2} (\bar{m}^{n+\frac{1}{2}}_{j+1} + \bar{m}^{n+\frac{1}{2}}_{j-1}) + \frac{1}{4} ((\bar{m}'_j)^n_{j+1} + (\bar{m}'_j)^n_{j-1}) - \frac{\lambda}{2} (F_K(\bar{m}^{n+\frac{1}{2}}_{j+1}) - F_K(\bar{m}^{n+\frac{1}{2}}_{j-1})).$$

The numerical derivatives are evaluated componentwise as in [47] using a slightly modified
MinMod limiter,
\[(\bar{m})^*_j = \text{MM} \left( 2(\bar{m}^*_j+1 - \bar{m}^*_j), \frac{1}{2}(\bar{m}^*_j+1 - \bar{m}^*_j-1), 2(\bar{m}^*_j - \bar{m}^*_j-1) \right),\]
with
\[
\text{MM}(x, y, z) = \begin{cases} 
\min(x, y, z) & \text{if } x, y, z > 0, \\
\max(x, y, z) & \text{if } x, y, z < 0, \\
0 & \text{otherwise.}
\end{cases}
\]
Whenever possible, it is of interest to compute the multivalued phase \( \varphi \) in order to complete the reconstruction of the WKB ansatz. This can be done by numerical integration if the physical problem generates the same number of multivaluations as the chosen \( K \)-moment system can handle. In this case, one has to follow each \( K \)-branch entropy solution \( u_k(t, \cdot) \) until it touches \( u_{k+1}(t, \cdot) \) for all \( 1 \leq k \leq K - 1 \). Reconstructing the multivalued intensity will be done accordingly, switching from branch \( \mu_k(t, \cdot) \) to the next one, \( \mu_{k+1}(t, \cdot) \), at the same points. Another criterion lies in the fact that some branches of the intensity vector \( \hat{\mu} \) should blow up when approaching a caustic.

To update the intensity vectors, we relied on [29, 31] and selected a positivity-preserving upwind scheme able to compute singular solutions free from spurious spikes and oscillations. In its three-point stencil version, it reads
\[
\hat{\mu}^{n+1}_j = \hat{\mu}^n_j - \lambda \left( \langle A^n_{j+\frac{1}{2}}, \mu^n_{j+\frac{1}{2}} \rangle_{\mathbb{R}^2} - \langle A^n_{j-\frac{1}{2}}, \mu^n_{j-\frac{1}{2}} \rangle_{\mathbb{R}^2} \right),
\]
\[
A^n_{j+\frac{1}{2}} = (\bar{a}^n_{j+\frac{1}{2},0}, \bar{a}^n_{j+\frac{1}{2},1}) \in (\mathbb{R}^K)^2, \quad \mu^n_{j+\frac{1}{2}} = (\hat{\mu}^n_j, \hat{\mu}^n_{j+1}) \in (\mathbb{R}^K)^2,
\]
and a convenient choice for \( A^n_{j+\frac{1}{2}} \) has to be drawn in order to get a good numerical approximation of (18). As in [31], and to share the same CFL as (22), we select
\[
\bar{a}^n_{j+\frac{1}{2},0} = \frac{1}{2} \max(0, \bar{a}^n_{j+1} + \bar{a}^n_j), \quad \bar{a}^n_{j+\frac{1}{2},1} = \frac{1}{2} \min(0, \bar{a}^n_{j+1} + \bar{a}^n_j),
\]
but any convex combination of the adjacent Riemann invariants could be admitted [29]. The reason for this lies in the fact that, even if not rigorously established, discrete one-sided Lipschitz conditions in the sense of [12] should be preserved.

All our numerical results have been obtained by means of these aforementioned discretizations. Of course, despite the fact that the construction of \( K \)-branch entropy solutions in Section 2.2 was made only for nonnegative values, we shall compute in practice general solutions belonging to \( \mathbb{R}^K \) since the aforementioned kinetic formalism can be applied in this context up to minor changes [9] (see also [40, 57]). When displaying antiderivatives of \( K \)-branch entropy solutions, we shall always use the letter “p” to refer to the corresponding (multivalued) phases.

### 3.2. A Fold Caustic with a Shadow Zone

We simulate a fold singularity, for instance, with the initial datum for (3):
\[
u^0(x) = -x^2, \quad \mu^0(x) \equiv 1; \quad x \in [0, 2].
\]
Using the notations of Section 2.1, we get
\[ \Phi(t, x_0) = (t, x_0 - t(x_0)^2), \quad \gamma = \left\{ x = \frac{1}{4t} \right\}, \]
together with
\[ \Psi^\pm(t, x) = \left(0, \frac{1 \pm \sqrt{\Delta}}{2t} \right), \quad \Delta = 1 - 4tx. \]

Inside a domain where \( \Delta \) is nonnegative, one can therefore define the twofold solutions of (6) in the sense of Theorem 2.1,
\[ u^\pm(t, x) = u^0(\Psi^\pm(t, x)) = u^0 \left(\frac{1 \pm \sqrt{\Delta}}{2t}\right), \quad (27) \]
where the minus sign corresponds to the incident wave. The intensities are easily deduced since \( z^\pm(t, x) = \int x \cdot u^\pm(t, x) \cdot dx \) satisfy a transport equation even in the case where solutions are weak [7]. Thus one finds that
\[ \mu^\pm(t, x) = |\partial_x \Psi^\pm(t, x)|u^0(\Psi^\pm(t, x)) = \frac{1}{\sqrt{1 - 4tx}}u^0(\Psi^\pm(t, x)). \quad (28) \]

As expected, WKB signals blow up on the caustic curve \( \gamma \). In the sequel, we shall refer to (27), (28) as the ray-traced solutions to (2) with initial data (26); see the rays geometry on Fig. 1. We simulated this problem using the 2-moment system corresponding to the case \( K = 2 \); for this choice, (18) can handle the right number of phases. Another feature of the choice (26) is that (17) is always in supersonic regime and the results displayed in Fig. 2 were obtained using the Godunov scheme, as in (22), together with the upwind scheme (24), (25). This discretization turns out to be more robust than the NT scheme (23) close to the degeneracy points \( u_1 \simeq u_2 \) where (17) is not strictly hyperbolic anymore. In order to

\[ \text{FIG. 1. Ray geometry for (3) with data (26), } x \in [0, 2], \ t \in [0, 0.7]. \]
take full advantage of the decoupling property of the $u_k$'s in the smooth region and to get them to satisfy discrete OSLC conditions, we selected the following initialization for (18) with a small $\varepsilon > 0$:

$$
\begin{align*}
    u_1(t = 0, x) &= u^0(x), \\
    u_2(t = 0, x) &= u^0(2) - \varepsilon; \\
    \mu_1(t = 0, x) &= \mu^0(x).
\end{align*}
$$

(29)

In order to get correct results for the reflected intensity, we follow ideas from [24] and impose the boundary condition on the caustic $\gamma$; since $\Phi(0.25, 2) = \gamma(0.25)$,

$$
\begin{align*}
    \mu_2(t, \gamma(t) - 0) &= \mu_1(t, \gamma(t) - 0), \\
    t &\geq 0.25.
\end{align*}
$$

(30)

In [5], a fix relying on [36, 42] is proposed in order to handle such a shadow zone more correctly by means of a local modification of the classical geometric optics ansatz. At the practical level, one has to locate the caustic curve; one way out is to look for an index...
value \( j_0 \) such that \( |\tilde{\mu}_{j_0+1} - \tilde{\mu}_{j_0}| > C \cdot \Delta x \), where \( C \) is a (big) positive constant and \( x_{j_0} \) lies in a neighborhood of an extremum of \( \tilde{\mu}^n \). We used values for \( C \) ranging from 10 to 100 to get the results shown in Fig. 2. In order to reconstruct the phase, it is sufficient to take numerical antiderivatives of the 2-branch entropy solutions on the right domains. Namely, one has to follow \( u_1(t, \cdot) \) until it touches \( u_2(t, \cdot) \), and then to follow backward \( u_2(t, \cdot) \) until it reaches the moving endpoint \( x = 2 - 4t \) located ahead of the fold. The parameters used were \( \Delta x = 7.10^{-3}, \Delta t = 15.10^{-4}, \) except for the fine grid for which \( \Delta x = 2 \cdot 10^{-3} \).

We decided also to make a convergence analysis for the 2-branch entropy solutions and the corresponding intensities. Hence for the incident signal at time \( t^n = 0.5 \) we looked at the decay as \( \Delta x \to 0 \) of the quantities

\[
\sum_{x_j \in [0, \gamma(t^n)]} \Delta x |u^-(t^n, x_j) - u_1(t^n, x_j)|, \quad \sum_{x_j \in [0, \gamma(t^n)]} \Delta x |\mu^-(t^n, x_j) - \mu_1(t^n, x_j)|,
\]

and together with (when \( t \geq 0.25 \))

\[
\sum_{x_j \in [2-4t^n, \gamma(t^n)]} \Delta x |u^+(t^n, x_j) - u_2(t^n, x_j)|, \quad \sum_{x_j \in [2-4t^n, \gamma(t^n)]} \Delta x |\mu^+(t^n, x_j) - \mu_2(t^n, x_j)|
\]

for the reflected one. Despite the fact that theoretical results from [29] only ensure a weak convergence for numerical approximations of linear conservation laws, we found the decays presented in Fig. 3. The numerical rate of convergence is approximately 0.5 for the 2-branch entropy solutions and 0.05 for the corresponding intensities. We refer the reader to [48, 3] for related numerical experiments.

### 3.3. A Cusp Caustic Generated by a Focus

Figure 2 in [31] displays a blowup of the WKB ansatz for a focusing problem with the free Schrödinger equation when interpreted in the sense of viscosity/duality, [19, 37, 7], which corresponds here to the choice \( K = 1 \).
Therefore, we take it again as an example to study our present approach in a situation leading to a cusp caustic; we select a Gaussian pulse

\[ u^0(x) = -8 \tanh(x), \quad \mu^0(x) = \exp(-x^2); \quad x \in [-5, 5]. \]  

(31)

According to Section 2.1, we observe self-interference,

\[ \Phi(t, x_0) = \Phi(t, x_0 - 8t \cdot \tanh(x_0)), \quad \gamma(t) = \Phi(t, x_0 = \text{atanh} \left( \pm \sqrt{1 - \frac{1}{8t}} \right)). \]

and \((x = 0, t = 0.125)\) is a “cool focus” in the sense of [51]. In this case, we cannot give an analytical expression for \(\Psi_{k=1, 2, 3}\); we use instead Newton’s algorithm to find these values numerically. Concerning the intensities, we observe that

\[ \mu_k(t, x) = |\partial(x)\Psi_k(t, x)|\mu^0(\Psi_k(t, x)) = \left| \frac{1}{(\partial(x)\Phi) \circ \Psi_k(t, x)} \right| \mu^0(\Psi_k(t, x)). \]

Therefore \(\tilde{u}(t, x)\) and \(\tilde{\mu}(t, x)\) can be known exactly and we refer to these values as to the ray-traced solutions to (2), (31); see the ray geometry in Fig. 4. In Fig. 5, we present the numerical results obtained from (18) with \(K = 3\) by means of the NT scheme (23) and (24). We initialized the 3-branch entropy solutions as follows, with \(\varepsilon > 0\) a small number:

\[ u_1(t = 0, x) = u^0(x) + \varepsilon, \quad u_2(t = 0, x) = u^0(x), \quad u_3(t = 0, x) = u^0(x) - \varepsilon. \]  

(32)

Observe that \(u_1 > u_2 > u_3\), and as long as the solution remains smooth, each \(u_k(t, \cdot)\) has to satisfy (3) up to the accuracy of the selected numerical scheme. Beyond the shock appearance in \(x = 0\), an \(N\)-wave surrounded by two shocks (also present in \(u_1\) and \(u_3\)) travelling at characteristic speed develops in \(u_2\).
Concerning the intensities, we thus deduce from (32) and (30)

\[ \mu_1(t = 0, x) = \mu_2(t = 0, x) = \mu_3(t = 0, x) = \mu^0(x), \]
\[ \mu_2(t, \gamma(t) - 0) = \mu_1(t, \gamma(t) - 0), \quad \gamma(t) \geq 0 \quad \text{and} \quad t \geq 0.125, \]
\[ \mu_2(t, \gamma(t) + 0) = \mu_3(t, \gamma(t) + 0), \quad \gamma(t) \leq 0 \quad \text{and} \quad t \geq 0.125, \]

with identical notation. We observe that $\mu_1, \mu_3$ are quite well rendered when compared to the ray-traced solution and also $\mu_2$, which is first shrunk (before the shock) and then expanded (beyond it); see also [34]. Finally, the reconstruction of the trivalued phase $\varphi$ is fairly good for the whole time interval. The viscosity solution is the part of the graph laying down the multivaluation; see again Fig. 2 in [31]. We used $\Delta x = 0.06$, $\Delta t = 0.0075$, and $\varepsilon = \Delta x/10$. Then we computed the intensities after the caustic on finer grid (see Fig. 6).
A history for $\varphi$ and $\vec{\mu}$ is displayed in Fig. 7. This can be seen as the correct behavior one would expect instead of the one presented in Fig. 2 in [31]. We emphasize finally that initializations of the type $\text{(32)}$, $\text{(33)}$ are also convenient for the computation of unsteady shocks (as is seen in Section 3.4).

We then tried to simulate the test case $\text{(31)}$ with the 2-moment system, as its structure is much simpler. In this context, since the system $\text{(17)}$ handles fewer values than the physical problem generates, there are many possible ways to initialize it, depending on how the user decides to represent a trivalued signal by means of a 2-branch entropy solution. Following [47], we selected the following choice to get the results shown in Fig. 8 ($1_A$ stands for the characteristic function of a set $A$):

$$u_1(t = 0, x) = u^0(x)1_{x < 0} + u^0(-x)1_{x > 0}, \quad u_2(t = 0, x) = u^0(-x)1_{x < 0} + u^0(x)1_{x > 0},$$

$$\mu_1(t = 0, x) = \mu^0(x)1_{x < 0}, \quad \mu_2(t = 0, x) = \mu^0(x)1_{x > 0}.$$ 

But the other forthcoming choices also lead to reasonable results, as they just correspond to another representation of a complex signal by means of a too-simple ansatz. For instance, this one is inspired by an alternative to $\text{(32)}$,

$$u_1(t = 0, x) = u^0(x)1_{x < 0} + u^0(0)1_{x > 0}, \quad u_2(t = 0, x) = u^0(0)1_{x < 0} + u^0(x)1_{x > 0}, \quad \text{(34)}$$

and now, this last one comes from the fold test case $\text{(29)}$ with a small $\epsilon > 0$,

$$u_1(t = 0, x) = u^0(x), \quad u_2(t = 0, x) = u^0(5) - \epsilon.$$ 

We do not claim that they are the only reasonable initial values for $\text{(18)}$, $\text{(31)}$ with the choice $K = 2$. Each representation gives, rather acceptable values for the intensities: of course, the two shocks appearing in $u_1$, $u_2$ induce concentrations in Dirac masses in $\mu_1$, $\mu_2$ as we refine the computational grid. This drawback is not a consequence of any numerical approximation but comes directly from the fact that the selected value of $K$ is too low and some superimposed signals collide instead of crossing each other (as already experienced in [31]). We finally refer to [47, 48] for other experiments on focusing problems.
3.4. A Double Cusp Singularity

This last test case is inspired by a practical situation studied in [21, 23, 24, 31, 47, 48], namely the focusing of light rays behind a thin convex lens. For the time being, we limit ourselves to consider the following initial datum:

\[ u^0(x) = -4(\tanh(x - 2) + \tanh(x + 2)), \quad \mu^0(x) = 1; \quad x \in [-2.5, 2.5]. \] (35)

The main difficulty lies in the fact that the signal develops up to five phases around \( x = 0 \).
for $t \geq 0.4$; see the ray geometry in Fig. 9. The geometrical solutions to (3), (35) can be derived by means of computations similar to the ones in Section 3.3. We did not try to simulate numerically this problem (18), (35) with the $K = 5$ moment closure, as it would be too expensive, cf. the Appendix in [47] for an expression of the resulting flux functions. Anyway, one can compute all the branches developed by (35) with a 3-branch moment

FIG. 8. Two-branch entropy solutions and intensity for a plane wave in the cusp caustic with the two-phases moment closure: $T = 0.07$ (left), $T = 0.7$ (right).

FIG. 9. Ray geometry for (3) with data (35); $x \in [-3, 3], t \in [0, 0.7]$. 
FIG. 10. Phase evolution for the double cusp caustic (35): $T = 0.3$ (left), $T = 0.7$ (right).

FIG. 11. Three-branch entropy solutions (left) and intensity (right) for the double cusp caustic (35) with the three-phases moment closure: $T = 0.1$ (up), $T = 0.3$ (middle), $T = 0.7$ (bottom).
moment closure \((K = 3)\) exploiting the fact that \(x = 0\) is a center of symmetry for \(\tilde{u}\). To this end, we restrict ourselves to the nonnegative branches and use the initialization

\[
\begin{align*}
    u_1(t = 0, \cdot) &= u^0 1_{x<0} + \varepsilon, & u_2(t = 0, \cdot) &= u^0 1_{x<0}, & u_3(t = 0, \cdot) &= u^0 1_{x<0} - \varepsilon, \\
    \mu_1(t = 0, \cdot) &= \mu_2(t = 0, \cdot) = \mu_3(t = 0, \cdot) = \mu^0 1_{x<0}; \\
    \mu_2(t, \gamma(t) - 0) &= \mu_1(t, \gamma(t) - 0), & \mu_2(t, \gamma(t) + 0) &= \mu_3(t, \gamma(t) + 0).
\end{align*}
\]

(36)

The remaining part of the signal is reconstructed by symmetry (see Figs. 10 and 11). It can be seen in Fig. 10 that the multivalued phase is quite well rendered even in the neighborhood of \(x = 0\), where it involves up to five values. Looking at the \(K\)-branch entropy solutions in Fig. 11, one notices again that the second branch, \(u_2\), is not accurately computed for \(T \simeq 0.3\). The computation of the intensities suffers from this fact, especially the middle branch \(\mu_2\) moving with \(u_2\). Concerning \(\mu_1\) and \(\mu_3\), correct blowups are observed when compared to a ray-traced solution. The numerical runs have been achieved relying on the Godunov scheme (22) and the upwind discretization (24) with the parameters \(\Delta x = 0.01, \varepsilon = \Delta t = \Delta x/10\).

4. SIMULATION OF A THIN CONVEX LENS

4.1. The Paraxial Problem

This approach may be extended to some cases of the steady 2D Helmholtz equation under a paraxial assumption, as in [4, 31, 47]. We recall briefly how such a problem can be carried out; more details are to be found, for instance, in [24, 46]. One starts from a 2D Helmholtz equation written in the following form:

\[
\Delta u + k^2 \eta^2 u = 0, \quad (x, y) \in \mathbb{R}^2.
\]

(37)

The quantity \(k\) refers to the frequency and is assumed to be big; \(\eta\) is a smooth positive function of \(x, y\) called the refraction index of the considered medium. When looking for oscillating plane wave solutions to (37), \(u(x, y) = A(x, y) \exp(ik\varphi(x, y))\), \(k \to +\infty\), a steady version of the WKB system (2) is derived:

\[
(\partial_x \varphi)^2 + (\partial_y \varphi)^2 = \eta^2, \quad \partial_t (A^2 \partial_x \varphi) + \partial_x (A^2 \partial_y \varphi) = 0.
\]

(38)

The first one is the well-known eikonal equation for the phase and the other is a linear conservation law expressing the fact that the intensity \(A^2\) is preserved inside “ray tubes.” In order to rewrite (38) in such a way that it matches our framework, we make the so-called paraxial assumption, which reads \(\partial_x \varphi > 0\). Hence, we can consider the following system, where \(\mu = A^2 \partial_x \varphi\):

\[
\begin{align*}
    \partial_y \varphi - \sqrt{\eta^2 - (\partial_x \varphi)^2} &= 0, & \partial_t \mu + \partial_x \left( \frac{\partial_x \varphi}{\sqrt{\eta^2 - (\partial_x \varphi)^2}} \right) &= 0.
\end{align*}
\]

(39)

In a homogeneous medium for which \(\eta \equiv 1\), one can introduce \(u = \partial_x \varphi\) and is naturally led to deal with a system of the form (where we substituted the variable \(t\) for \(y\) for ease of reading)

\[
\begin{align*}
    \partial_t u - \partial_x (\sqrt{1 - u^2}) &= 0, & \partial_t \mu + \partial_x \left( \frac{u \mu}{\sqrt{1 - u^2}} \right) &= 0,
\end{align*}
\]
and the kinetic formulation of [10] extends rather naturally up to more-involved calculations in deducing the expression of the resulting flux functions $F_K$.

4.2. Numerical Results

The problem we want to simulate consists in computing the geometrical solutions to (39) with the following data: $x \in [-1, 1]$, $t \in [0, 2]$, $D = (\frac{-0.5}{0.3})^2 + (\frac{x}{0.8})^2$, and

$$\varphi(t = 0, x) \equiv 0, \quad \mu(t = 0, x) \equiv 1, \quad \eta(t, x) = \begin{cases} 
\frac{4}{\pi - \cos(\pi D)} & \text{if } D < 1, \\
1 & \text{in the other cases.}
\end{cases}$$ (40)

Several difficulties arise from this set of data. First, the refraction index is not constant and the kinetic formulation of Section 2.2 does not apply straightforwardly. Second, the ray-traced solution develops five phases in the neighborhood of the focal point of the lens located around $t = x = 1$ before settling with three; see [47], p. 82.

There is however a point to notice in order to treat efficiently this situation; namely, $K$-multivalued solutions are not needed in the computational domain where $D < 1$, that is to say, where the refraction index is not constant. Before and inside the lens, the solution to (39) is still smooth: thus it can be computed easily by relying on the numerical schemes proposed in [31]. Moreover, the problem is still endowed with the symmetry properties which have been usefully exploited in Section 3.4: hence it can be treated by means of the 3-moment closure (the expression of the flux functions $F_K$ is to be found in Appendix A.2 in [48]). Thus it is sufficient to mimic the approach of Section 3.4 in order to produce the numerical results displayed in Fig. 12; the parameters used were $\Delta x = 0.004$, $\Delta t = 0.003$.

In order to reconstruct the phase, one has to be careful only at one level. More precisely, if $u = \partial_t \varphi$, the correct phase is given by

$$\varphi(t, x) = t + \int_{-1}^{x} u(t, s) \cdot ds, \quad x \in [-1, 1].$$

And the reason is that $\partial_t \varphi = 1 + \int_{-1}^{x} \partial_s \sqrt{1 - (\partial_s \varphi)^2} \cdot ds = 1 + \sqrt{1 - (\partial_s \varphi)^2} - \sqrt{1}$ because the computational domain is big enough to have $\partial_s \varphi = 0$ in $x = \pm 1$. Then we display in Fig. 13 the values obtained from schemes (23) and (24). A standard computation relying

---

**FIG. 12.** $K$-branch entropy solutions $\vec{u}$ outside the convex lens (40): $T = 1.35$ (left), $T = 1.85$ (right). Dotted lines refer to the positive part of the ray-traced solution.
FIG. 13. Multivalued phase \( \varphi(T, \cdot) - T \) (left) and amplitude \( A(T, \cdot) \) (right) for a plane wave outside the convex lens (40). (Top to bottom) \( T = 0.85, 1.05, 1.35, 1.85 \).
on an interpretation of (39) within the theory of viscosity/duality solutions can be checked in [31], Fig. 5. In the present case, we observe the formation of the five valuations in the phase and the intensity.

5. CONCLUSION

We presented in this paper a reasonable “shock-capturing” compromise to compute the multivalued signals one would expect out of geometric optics expansions, for instance in the case of the free 1D Schrödinger and a steady 2D Helmholtz equations. As an extension, fully nonhomogeneous problems can be considered within a similar framework: for instance, a potential with wells for the Schrödinger equation would lead from (4) to a $K$-moment system of the form

$$
\partial_t m_k + \partial_x m_{k+1} = (k - 1)V'(x)m_{k-1}, \quad k = 1, \ldots, K,
$$

(A.1)

to which the same closure formalism can be applied. This amounts in general to adding another source term in the kinetic equation (8). At the theoretical level, one should then use the SRS solutions built in [2] and Theorem 2.5 will hold relying on the modified one-sided estimates established in [26]. The numerical analysis can be delicate, as the use of accurate well-balanced [27] discretizations may be necessary [9]. In contrast, a truly multidimensional approach seems still to lie out of reach; some alternatives have been recently proposed in, e.g., [3, 5, 22, 34, 35, 51, 52].

APPENDIX: A SHORT NONHOMOGENEOUS EXAMPLE

We wish to present briefly a way to solve an interesting nonhomogeneous test case suggested by Brenier, [9]; namely, we consider the Liouville equation (4) in the special case $\mathcal{H}(x, u) = \frac{1}{2}(u^2 + x^2)$. This leads to a classical Vlasov equation,

$$
\partial_t f + \partial_u \mathcal{H} \cdot \partial_x f - x \partial_u f = 0,
$$

which describes a rigid rotation in the phase space at a unit angular speed. An exact solution is to be found in [18]. We reduce it by means of the 2-moment closure to the following system of balance laws (see (41)):

$$
\partial_t m_1 + \partial_x m_2 = 0, \quad \partial_t m_2 + \partial_x \left( \frac{(m_2)^2}{m_1} + \frac{(m_1)^3}{12} \right) = -x m_1.
$$

(A.1)

It turns out that such a system can hardly be numerically simulated by means of a conventional time-splitting Godunov scheme for which the convection step and the source term are treated independently. However, following [28], it is possible to modify the homogeneous discretization (22) in order to tackle properly this problem: indeed, keeping $u_1$ and $u_2$ positive, one can make use of

$$
\tilde{m}^{n+1}_j = \tilde{m}^n_j - \lambda \left( F_2(\tilde{m}^n_j) - F_2(\tilde{m}^n_{j-\frac{1}{2}}) \right),
$$

where the state $\tilde{m}^n_{j-\frac{1}{2}}$ is deduced from $\tilde{m}^n_{j-1}$ following, as usual, the steady-state equations of (A.1). It is at this level that the richness of (A.1) plays a role since these last equations
FIGURE 14. Two-branch entropy solutions for $V(x) = -\frac{1}{2}x^2$: $T = 0.07$ (left), $T = 0.7$ (right).

can be easily integrated in the Riemann coordinates. One finds within the notation of (14) the following values to be used in (A.2):

$$(u_{1,2})_{n}^{j} = \sqrt{(u_{1,2})_{n}^{j-1}} - 2x \cdot \Delta x - \Delta x^2,$$

$$\vec{m}^{n}_{j-\frac{1}{2}} = \vec{m}((u_{1})^{n}_{j-\frac{1}{2}}, (u_{2})^{n}_{j-\frac{1}{2}}).$$

The results displayed in Fig. 14 were obtained with $\Delta x = 0.01$ and fitting the time step $\Delta t$ in order to keep a constant CFL value of 1. The initial data are

$$u_{1}^{0} = 0.15 + \Delta x + \max(0, -5(x + 1)(x + 0.7)),$$

$$u_{2}^{0} = 0.15 + \min(0, 5(x + 1)(x + 0.7)).$$

ACKNOWLEDGMENTS

An anonymous referee and Professor Yann Brenier made constructive comments which led to a genuine improvement of the present article. The author also thanks Professors I. Capuzzo-Dolcetta, M. Falcone, G. T. Kossioris, and G. Toscani for having suggested this problem and providing the environment where this research has been conducted.

REFERENCES

9. Y. Brenier, personal communication.


34. S. Jin and X. Li, Multi-phase computations of the semiclassical limit of the Schrödinger equation and related problems: Whitham vs. Wigner, submitted for publication.


