COMMUN. MATH. SCI.
Vol. 6, No. 3, pp. 569–593

SECOND-ORDER SLOPE LIMITERS FOR THE SIMULTANEOUS LINEAR ADVECTION OF (NOT SO) INDEPENDENT VARIABLES*

QUANG HUY TRAN†

Abstract. We propose a strategy to perform second-order enhancement using slope-limiters for the simultaneous linear advection of several scalar variables. Our strategy ensures a discrete min-max principle not only for each variable but also for any number of non-trivial combinations of them, which represent *control variables*. This problem arises in fluid mechanics codes using the Arbitrary Lagrange-Euler formalism, where the additional monotonicity property on control variables is required by physical considerations within the *remap* step.

Key words. Linear advection, min-max principle, slope reconstruction, Arbitrary Lagrange-Euler (ALE) formalism

AMS subject classifications. 76M12, 65M06, 35L65

1. Introduction

We are concerned with the design of a second-order accurate scheme for the numerical solution of the system of linear advections

$$\partial_t \mathbf{\Psi} + u\partial_x \mathbf{\Psi} = \mathbf{0}, \tag{1.1}$$

where $u \in \mathbb{R}$ is a given velocity field and $\mathbf{\Psi} = (\psi^1, \ldots, \psi^P) \in \mathbb{R}_+^P$, with $P \in \mathbb{N}$, is the vector of positive unknowns. Although the components of $\mathbf{\Psi}$, called *main variables*, are independent from each other at the continuous level, our objective is to ensure a min-max principle at the discrete level not only for the main variables but also for a set of physically meaningful *control variables*

$$\mathbf{G}(\mathbf{\Psi}) = (G^1(\mathbf{\Psi}), \ldots, G^Q(\mathbf{\Psi})) \in \mathbb{R}^Q, \qquad Q \in \mathbb{N}. \tag{1.2}$$

The motivation for such a requirement usually comes from the context of the industrial application at hand and from the observation that, as a result of multiplication of (1.1) by $\nabla_{\mathbf{\Psi}}\mathbf{G}$, the control quantities are also transported by the advection system

$$\partial_t \mathbf{G}(\mathbf{\Psi}) + u\partial_x \mathbf{G}(\mathbf{\Psi}) = \mathbf{0}. \tag{1.3}$$

A simple but prominent example corresponds to $P=2$, $Q=1$ and $G(\Psi_1, \Psi_2) = \Psi_1 + \Psi_2$, which is called the *sum-problem* and will be addressed in Section 2.

In order to state the problem more accurately, let us recall some background material for the scalar linear advection

$$\partial_t \psi + u\partial_x \psi = 0, \tag{1.4}$$

in which the velocity u is assumed to be uniform and positive. Over a uniform grid with mesh-size Δx, the cells of which are indexed by i, and for a time-step Δt, we consider the explicit second-order accurate update formula [9]

$$\widehat{\psi}_i = \psi_i - \lambda \left\{ \left[\psi_i + \tfrac{1-\lambda}{2} D_i^{\psi} \right] - \left[\psi_{i-1} + \tfrac{1-\lambda}{2} D_{i-1}^{\psi} \right] \right\}, \tag{1.5}$$

*Received: January 23, 2008; accepted (in revised version): May 19, 2008. Communicated by Shi Jin.

†Département Mathématiques Appliquées, Institut Français du Pétrole, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France (Q-Huy.Tran@ifp.fr).

where D_i^ψ is the approximate slope (multiplied by Δx) of ψ in cell i, and

$$\lambda = \frac{u\Delta t}{\Delta x} \tag{1.6}$$

is the CFL ratio. It is well-known [4, 7] that if the slopes are suitably chosen via an appropriate limiter function

$$D_i^\psi = \widetilde{D}_i^\psi \equiv \Lambda(\psi_i - \psi_{i-1}, \psi_{i+1} - \psi_i), \tag{1.7}$$

then the discrete min-max principle

$$\widehat{\psi}_i \in \lfloor \psi_{i-1}, \psi_i \rceil \tag{1.8}$$

holds under the CFL condition $\lambda < 1$. We systematically use the notation $\lfloor \mathfrak{a}, \mathfrak{b} \rceil$ for the convex interval spanned by the real numbers $\mathfrak{a}$ and $\mathfrak{b}$. We recall that the conditions on the limiter function $(d_L, d_R) \in \mathbb{R}^2 \mapsto \Lambda(d_L, d_R) \in \mathbb{R}$ for the min-max principle (1.8) to hold are

$$\frac{2}{\lambda} d_L^- \leq \Lambda(d_L, d_R) \leq \frac{2}{\lambda} d_L^+ \tag{1.9a}$$

$$\frac{2}{1-\lambda} d_R^- \leq \Lambda(d_L, d_R) \leq \frac{2}{1-\lambda} d_R^+, \tag{1.9b}$$

for all pairs of "candidate" slopes (d_L, d_R), where we use the notations $r^- = \min(r, 0)$ and $r^+ = \max(r, 0)$ for the negative and positive part of any real number r. This automatically implies

$$\Lambda(d_L, d_R) = 0, \qquad \text{if } d_L d_R \leq 0. \tag{1.10}$$

Furthermore it is customary, albeit not compulsory, to impose additional constraints such as

$$\Lambda(d, d) = d, \qquad \text{for all } d \in \mathbb{R}, \tag{1.11}$$

to ensure consistency with exact second-order reconstruction and

$$\Lambda(d_L, d_R) = \Lambda(d_R, d_L), \qquad \text{for all } (d_L, d_R) \in \mathbb{R}^2, \tag{1.12}$$

to require symmetry. Examples will be provided in the upcoming section.

Of course, we are going to apply scheme (1.5) to numerically solve system (1.1) component-wise. If the slopes are limited as in (1.7), i.e.,

$$D_i^{\psi^p} = \widetilde{D}_i^{\psi^p} \equiv \Lambda(\psi_i^p - \psi_{i-1}^p, \psi_{i+1}^p - \psi_i^p) \tag{1.13}$$

for $1 \leq p \leq P$, then there holds discrete min-max principle component-wise

$$\widehat{\psi}_i^p \in \lfloor \psi_{i-1}^p, \psi_i^p \rceil. \tag{1.14}$$

As for the control variables, which are necessarily computed as

$$G_i^q = G^q(\mathbf{\Psi}_i), \qquad \widehat{G}_i^q = G^q(\widehat{\mathbf{\Psi}}_i) \tag{1.15}$$

for $1 \leq q \leq Q$, there is no reason that we should have the desired min-max principle

$$\widehat{G}_i^q \in \lfloor G_{i-1}^q, G_i^q \rceil, \tag{1.16}$$

insofar as the components of $\mathbf{\Psi}$ do not "see each other".

The min-max principle on control variables is a major challenge in many fluid mechanics codes using an ALE (Arbitrary Lagrange-Euler) method [2, 3], the remap phase of which consists in simultaneously advecting several supposedly independent variables. Such a requirement is essential for robustness. However, except for a partially successful attempt by VanderHeyden and Kashiwa [10] for a restricted setting of the fraction problem, we do not have knowledge of any thoroughly satisfactory solution. The present contribution demonstrates that the component-wise limitation (1.13) can be actually replaced by a more general framework

$$D_i^{\psi^p} = \Lambda^p(\mathbf{\Psi}_{i-1}, \mathbf{\Psi}_i, \mathbf{\Psi}_{i+1}) \tag{1.17}$$

which does guarantee (1.14) and (1.16) under the same CFL condition $\lambda < 1$. This newly proposed procedure can be extended to the case of a space-dependent velocity field $u = u(x)$, the sign of which is not necessarily constant. From a practical point of view, the new slopes (1.17) will be obtained from the old ones, computed by (1.13), through a projection mechanism which creates the opportunity for the various (main and control) variables to see each other. This projection mechanism is optimally designed in order for the new slopes to be as "close" as possible to the old slopes in some sense, so that sharp profiles can still be captured.

In order to convey the geometric insights that are at the root of the seemingly complex algebraic formalism of this work, we focus on the two simplest but most important examples encountered in the context of Euler-like fluid models: the sum problem Section 2 for a flame model [1] and the fraction problem Section 3 for a two-phase flow model [2]. Sec. 4 is devoted to the general problem, along with the numerical results for an example selected from real-life applications.

2. The sum problem

We consider the densities of two species, say, CH_4 and CO_2, as well as their sum, which represents the carbon tracer. Let us put

$$\mathbf{\Psi} = (\alpha, \beta) \in \mathbb{R}_+^2, \qquad G(\mathbf{\Psi}) = \alpha + \beta \in \mathbb{R}_+. \tag{2.1}$$

2.1. Uniform velocity. Assume $u(x) = u > 0$. From the current time level to the next one, the update formulae for (α, β) are

$$\widehat{\alpha}_i = \alpha_i - \lambda\{[\alpha_i + \tfrac{1-\lambda}{2} D_i^\alpha] - [\alpha_{i-1} + \tfrac{1-\lambda}{2} D_{i-1}^\alpha]\} \tag{2.2a}$$

$$\widehat{\beta}_i = \beta_i - \lambda\{[\beta_i + \tfrac{1-\lambda}{2} D_i^\beta] - [\beta_{i-1} + \tfrac{1-\lambda}{2} D_{i-1}^\beta]\}, \tag{2.2b}$$

where λ is defined in (1.6). Consider the *initial slopes*

$$\widetilde{D}_i^\alpha = \Lambda(\alpha_i - \alpha_{i-1}, \alpha_{i+1} - \alpha_i), \qquad \widetilde{D}_i^\beta = \Lambda(\beta_i - \beta_{i-1}, \beta_{i+1} - \beta_i), \tag{2.3}$$

inspired from the scalar case (1.7) and computed component-wise via a standard limiter function Λ, such as (see [4, 7] for more details)

- the *minmod* limiter

$$\operatorname{minmod}(d_L, d_R) = \max\{d_L^-; d_R^-\} + \min\{d_L^+; d_R^+\}; \tag{2.4}$$

- the *van Leer* limiter

$$\operatorname{VL}(d_L, d_R) = \frac{2 d_L^- d_R^-}{d_L^- + d_R^-} + \frac{2 d_L^+ d_R^+}{d_L^+ + d_R^+} \tag{2.5}$$

(being understood that at most one summand is not zero);

- the *superbee* limiter

$$\mathrm{SB}(d_L,d_R)=\mathrm{maxmod}(\mathrm{minmod}(2d_L,d_R),\mathrm{minmod}(d_L,2d_R)), \tag{2.6}$$

where minmod was defined in (2.4), and

$$\mathrm{maxmod}(e_L,e_R)=\min\{e_L^-,e_R^-\}+\max\{e_L^+,e_R^+\} \tag{2.7}$$

- the *hyperbee* limiter

$$\mathrm{HB}(d_L,d_R)=\mathrm{maxmod}(\mathrm{minmod}(\Phi d_L,d_R),\mathrm{minmod}(d_L,\Phi d_R)), \tag{2.8}$$

where $\Phi=\min\{\frac{2}{\lambda},\frac{2}{1-\lambda}\}$;

- the *ultrabee* limiter

$$\mathrm{UB}(d_L,d_R)=\mathrm{minmod}(\tfrac{2}{\lambda}d_L,\tfrac{2}{1-\lambda}d_R), \tag{2.9}$$

which is the "strongest" limiter that meets conditions (1.9); unlike the previous limiters, it does not satisfy the additional Propositions (1.11) and (1.12).

The following Lemma recalls a useful property, which expresses the basic fact given a pair of slopes that ensures the min-max principle component-wise, any pair with smaller amplitudes is also suitable for the min-max principle.

LEMMA 2.1. *If the slopes* (D_j^α,D_j^β) *in (2.2) satisfy*

$$[\widetilde{D}_j^\alpha]^-\le D_j^\alpha\le[\widetilde{D}_j^\alpha]^+,\qquad [\widetilde{D}_j^\beta]^-\le D_j^\beta\le[\widetilde{D}_j^\beta]^+ \tag{2.10}$$

for $j=i-1$ *and* $j=i$, *then* $\widehat{\alpha}_i\in\lfloor\alpha_{i-1},\alpha_i\rceil$ *and* $\widehat{\beta}_i\in\lfloor\beta_{i-1},\beta_i\rceil$.

Proof. This is a consequence of Sweby's analysis [7], by which an appropriate choice of the limiter function Λ allows one to express $\widehat{\alpha}_i$ as a convex combination of α_{i-1} and α_i (likewise for $\widehat{\beta}_i$). Conditions (2.10) say that the new slopes must be of the same sign as the old ones, while having smaller absolute values. ∎

The sum variable G is computed by $G_i=\alpha_i+\beta_i$ and $\widehat{G}_i=\widehat{\alpha}_i+\widehat{\beta}_i$. Because the scheme is nonlinear due to the use of limiter, it will be a mistake to take it for granted that the min-max principles on α and β always imply that on G: this is true only when the min (resp. max) of the sum is equal to the sum of the min values (resp. max values), which means that α and β both increase or both decrease from $i-1$ to i, as highlighted by the following Proposition.

PROPOSITION 2.2. *If* $(\alpha_i-\alpha_{i-1})(\beta_i-\beta_{i-1})\ge0$ *and if the slopes* (D_j^α,D_j^β) *satisfy (2.10) for* $j=i-1$ *and* $j=i$, *then* $\widehat{G}_i\in\lfloor G_{i-1},G_i\rceil$.

Proof. In the quarter-plane $(\alpha,\beta)\in\mathbb{R}_+\times\mathbb{R}_+$, let us depict the points

$$M_{i-1}=(\alpha_{i-1},\beta_{i-1}),\quad M_i=(\alpha_i,\beta_i),\quad \widehat{M}_i=(\widehat{\alpha}_i,\widehat{\beta}_i). \tag{2.11}$$

as in Figure 2.1. The min-max principles $\widehat{\alpha}_i\in\lfloor\alpha_{i-1},\alpha_i\rceil$ and $\widehat{\beta}_i\in\lfloor\beta_{i-1},\beta_i\rceil$, which follow from Lemma 2.1, amount to saying that $\widehat{M}_i$ belongs to the rectangle $\mathscr{R}_i$ whose opposite vertices are M_{i-1} and M_i and whose sides are parallel to the horizontal and vertical axes. Draw the lines $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$ defined by $\alpha+\beta=G_{i-1}$ and $\alpha+\beta=G_i$.

If $(\alpha_i-\alpha_{i-1})(\beta_i-\beta_{i-1})\ge0$, then the rectangle $\mathscr{R}_i$ is entirely included in the strip defined by the parallel lines $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$. Therefore, the isoline of $\alpha+\beta$

passing through $\widehat{M}_i$ lies between $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$, which is algebraically equivalent to $\widehat{G}_i \in \lfloor G_{i-1}, G_i \rceil$.

If $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) < 0$, the lines $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$ cut the rectangle $\mathscr{R}_i$ into three pieces, and it may happen that $\widehat{M}_i$ lies outside the strip, which violates the desired min-max principle. ∎

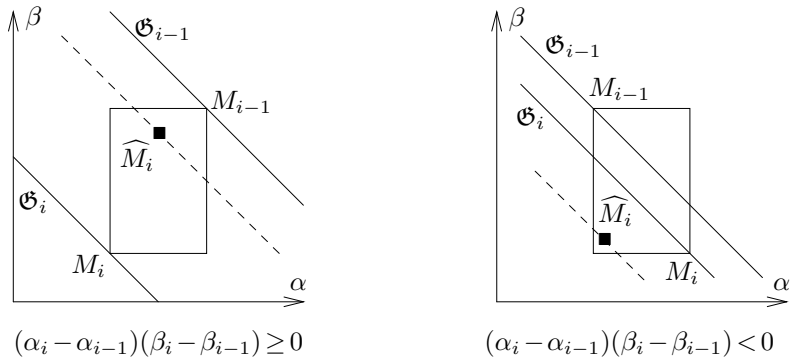

FIG. 2.1. *Geometric analysis of the min-max principle for the sum problem.*

To know what should be done for the case $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) < 0$, we introduce

$$G_i^m = \min\{G_{i-1}, G_i\}, \qquad G_i^M = \max\{G_{i-1}, G_i\}, \tag{2.12}$$

and seek sufficient conditions in terms of (D^α, D^β) so that $G_i^m \leq \widehat{G}_i \leq G_i^M$ at a given cell i. Unless otherwise indicated, it is assumed that $\lambda < 1$.

LEMMA 2.3. *For a given cell i, if*

$$\tfrac{2}{\lambda}(G_i - G_i^M) \leq D_i^\alpha + D_i^\beta \leq \tfrac{2}{\lambda}(G_i - G_i^m), \tag{2.13a}$$

$$-\tfrac{2}{1-\lambda}(G_{i-1} - G_i^m) \leq D_{i-1}^\alpha + D_{i-1}^\beta \leq -\tfrac{2}{1-\lambda}(G_{i-1} - G_i^M), \tag{2.13b}$$

then $G_i^m \leq \widehat{G}_i \leq G_i^M$.

Proof. Subtracting the convex decomposition $G_i^m = (1-\lambda)G_i^m + \lambda G_i^m$ from the sum of (2.2a) and (2.2b), we obtain

$$\widehat{G}_i - G_i^m = (1-\lambda)[G_i - G_i^m - \tfrac{\lambda}{2}(D_i^\alpha + D_i^\beta)] + \lambda[G_{i-1} - G_i^m + \tfrac{1-\lambda}{2}(D_{i-1}^\alpha + D_{i-1}^\beta)]. \tag{2.14}$$

To ensure $\widehat{G}_i - G_i^m \geq 0$, we split the right hand side into two parts and impose positivity to each summand. This leads to the right part of (2.13a) and the left part of (2.13b). We proceed similarly to impose negativity to $\widehat{G}_i - G_i^M$. ∎

The benefit of this splitting approach lies in the fact that the resulting conditions (2.13) are local: they do not couple the slopes at cell i with those at cell $i-1$, thus giving rise to a tractable procedure. It could be legitimately feared that imposing positivity separately in (2.14) yields excessively strong conditions, which might deteriorate accuracy. The miracle is that accuracy is preserved at a very good level, as shown by numerical results.

We are now in a position to formulate the new limitation procedure, which ensures the min-max principle for all cells.

THEOREM 2.4. *Given an initial choice* $(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta)$ *in accordance with (2.3), let* $\mathscr{G}_i\subset\mathbb{R}^2$ *be the set of all pairs* (D_i^α,D_i^β) *subject to the 6 linear inequality constraints*

$$[\widetilde{D}_i^\alpha]^-\le D_i^\alpha\le[\widetilde{D}_i^\alpha]^+,\quad [\widetilde{D}_i^\beta]^-\le D_i^\beta\le[\widetilde{D}_i^\beta]^+,\quad \mathfrak{m}_i\le D_i^\alpha+D_i^\beta\le\mathfrak{M}_i, \tag{2.15}$$

with

$$\mathfrak{m}_i=\max\{\tfrac{2}{\lambda}[G_i-G_{i-1}]^-;\ \tfrac{2}{1-\lambda}[G_{i+1}-G_i]^-\} \tag{2.16a}$$

$$\mathfrak{M}_i=\min\{\tfrac{2}{\lambda}[G_i-G_{i-1}]^+;\ \tfrac{2}{1-\lambda}[G_{i+1}-G_i]^+\}. \tag{2.16b}$$

For all $i\in\mathbb{Z}$, *define*

$$(D_i^\alpha,D_i^\beta)=\begin{cases}(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta) & \text{if } \widetilde{D}_i^\alpha\widetilde{D}_i^\beta>0\\ \Pi_{\mathscr{G}_i}(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta) & \text{otherwise}\end{cases} \tag{2.17}$$

where $\Pi_{\mathscr{G}_i}(.)$ *denotes the projection onto the convex set* $\mathscr{G}_i\subset\mathbb{R}^2$. *Then, we have the min-max principles*

$$\widehat{\alpha}_i\in\lfloor\alpha_{i-1},\alpha_i\rceil,\qquad \widehat{\beta}_i\in\lfloor\beta_{i-1},\beta_i\rceil,\qquad \widehat{G}_i\in\lfloor G_{i-1},G_i\rceil, \tag{2.18}$$

at every cell i *when updating* (α,β) *with scheme (2.2).*

Proof. The set $\mathscr{G}_i$ is obviously convex and nonempty, because it contains $(0,0)$. Therefore, definition (2.17) makes sense. Note that its shape depends on $(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta)$.

At a fixed cell i, if $\widetilde{D}_i^\alpha\widetilde{D}_i^\beta>0$, we necessarily have $(\alpha_i-\alpha_{i-1})(\beta_i-\beta_{i-1})>0$ on the grounds of the properties of standard limiter functions. According to Proposition 2.2, the default values $(D_i^\alpha,D_i^\beta)=(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta)$ are suitable. If $\widetilde{D}_i^\alpha\widetilde{D}_i^\beta\le0$, we are going to check conditions (2.13). From (2.15) and (2.16), we infer that

$$\tfrac{2}{\lambda}[G_i-G_{i-1}]^-\le\mathfrak{m}_i\le D_i^\alpha+D_i^\beta\le\mathfrak{M}_i\le\tfrac{2}{\lambda}[G_i-G_{i-1}]^+. \tag{2.19}$$

We claim that $[G_i-G_{i-1}]^-=G_i-G_i^M$ and $[G_i-G_{i-1}]^+=G_i-G_i^m$. To see this, we simply have to distinguish the two cases $G_{i-1}\le G_i$ and $G_{i-1}>G_i$. This establishes (2.13a).

If $\widetilde{D}_{i-1}^\alpha\widetilde{D}_{i-1}^\beta>0$, we also necessarily have $(\alpha_i-\alpha_{i-1})(\beta_i-\beta_{i-1})>0$ thanks to the properties of standard limiter functions. By virtue of Proposition 2.2, we conclude that there is no need to check (2.13b). If $\widetilde{D}_{i-1}^\alpha\widetilde{D}_{i-1}^\beta\le0$, we write (2.15) and (2.16) for cell $i-1$ and observe that

$$\tfrac{2}{1-\lambda}[G_i-G_{i-1}]^-\le\mathfrak{m}_{i-1}\le D_{i-1}^\alpha+D_{i-1}^\beta\le\mathfrak{M}_{i-1}\le\tfrac{2}{1-\lambda}[G_i-G_{i-1}]^+, \tag{2.20}$$

and once again argue that $[G_i-G_{i-1}]^-=G_i-G_i^M$ and $[G_i-G_{i-1}]^+=G_i-G_i^m$ to derive (2.13b). ∎

The coding of the projection operator $\Pi_{\mathscr{G}_i}$ in this problem can be made efficient through the explicit formulae provided by Proposition 2.5. Figure 2.2 illustrates a few situations for a locally increasing or decreasing behavior of G. Note that if a local extremum occurs, i.e., $(G_{i-1}-G_i)(G_{i+1}-G_i)>0$, by (2.16) we have $\mathfrak{m}_i=\mathfrak{M}_i=0$, hence $D_i^G=D_i^\alpha+D_i^\beta=0$. This testifies to a clipping mechanism on G in the proposed procedure.

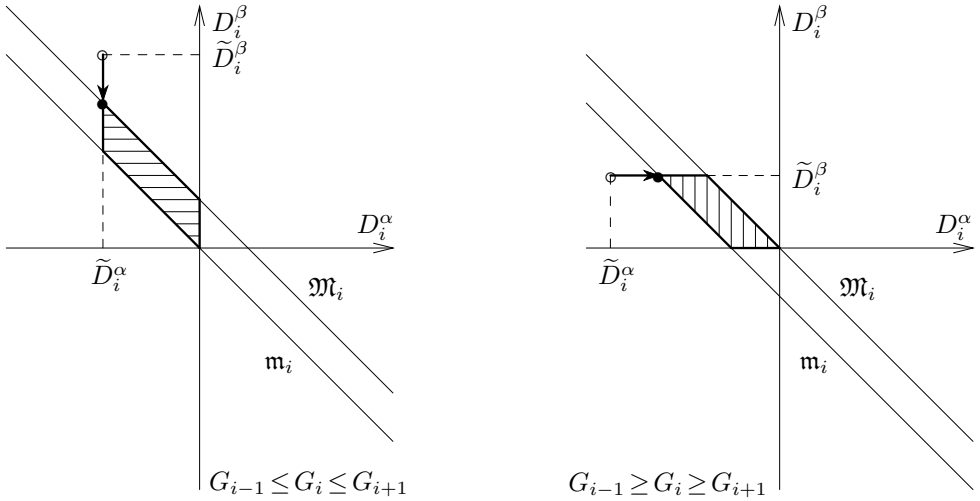

FIG. 2.2. *Projection of* $(\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta)$ *onto the convex set* $\mathscr{G}_i$ *for the sum problem.*

PROPOSITION 2.5. *The projection operator* $\Pi_{\mathscr{G}_i}$ *introduced in (2.17) can be carried out by the closed-form formulae*

$$D_i^\alpha = \widetilde{D}_i^\alpha - \{[\widetilde{D}_i^\alpha]^+ + [\widetilde{D}_i^\beta]^- - \mathfrak{M}_i\}^+ - \{[\widetilde{D}_i^\alpha]^- + [\widetilde{D}_i^\beta]^+ - \mathfrak{m}_i\}^- \tag{2.21a}$$

$$D_i^\beta = \widetilde{D}_i^\beta - \{[\widetilde{D}_i^\alpha]^- + [\widetilde{D}_i^\beta]^+ - \mathfrak{M}_i\}^+ - \{[\widetilde{D}_i^\alpha]^+ + [\widetilde{D}_i^\beta]^- - \mathfrak{m}_i\}^-. \tag{2.21b}$$

Proof. equations (2.21) are the algebraic translation of the geometric ideas presented in Figure 2.2. ∎

At this point, we wish to provide an intuitive argument on why the new slope-limiting procedure does not deteriorate too much accuracy. From formulae (2.21), we see that

- either $(D_i^\alpha, D_i^\beta) = (\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta)$, that is, we are optimal with respect to the main variables (α, β);
- or $D_i^G = D_i^\alpha + D_i^\beta$ is saturated at $\mathfrak{m}_i$ or $\mathfrak{M}_i$; by virtue of their values (2.17), this amounts to saying that $D_i^G = \mathrm{UB}(G_i - G_{i-1}, G_{i+1} - G_i)$ where UB denotes the UltraBee limiter function defined in (2.9). As a consequence, we are optimal with respect to the control variable G.

Numerical experiments presented in Section 2.3 will corroborate the above heuristic explanation.

2.2. Variable velocity field. The velocities $u_{i+1/2} = u(x_{i+1/2})$ are given at the edges. We choose to discretize the advection equation $\partial_t \psi + u \partial_x \psi = 0$ by the explicit scheme

$$\begin{aligned} \widehat{\psi}_i = \psi_i &- \lambda_{i-1/2}^- \tfrac{1-|\lambda|_i}{2} D_i^\psi - \lambda_{i-1/2}^+ \left\{ \psi_i - \left(\psi_{i-1} + \tfrac{1-|\lambda|_{i-1}}{2} D_{i-1}^\psi \right) \right\} \\ &- \lambda_{i+1/2}^+ \tfrac{1-|\lambda|_i}{2} D_i^\psi + \lambda_{i+1/2}^- \left\{ \psi_i - \left(\psi_{i+1} - \tfrac{1-|\lambda|_{i+1}}{2} D_{i+1}^\psi \right) \right\}, \end{aligned} \tag{2.22}$$

where

$$\lambda_{i\pm1/2}=\frac{u_{i\pm1/2}\Delta t}{\Delta x},\qquad |\lambda|_i=\lambda^+_{i-1/2}-\lambda^-_{i+1/2}. \tag{2.23}$$

Scheme (2.22) appears to be one of the many possible second-order extensions [5] of the first-order explicit scheme

$$\widehat{\psi}_i=\psi_i-\lambda^+_{i-1/2}(\psi_i-\psi_{i-1})-\lambda^-_{i+1/2}(\psi_{i+1}-\psi_i). \tag{2.24}$$

Before discussing about the min-max principle, we feel it essential to draw the reader's attention on what is exactly meant by "second-order" for a variable velocity field. A straightforward calculation shows that the modified equation equivalent to the first-order scheme (2.24) formally reads

$$\partial_t\psi+u\partial_x\psi=\frac{\Delta x}{2}(\partial_x u)(\partial_x\psi)+u\frac{\Delta x}{2}\partial_x\{[1-|\lambda(u)|]\partial_x\psi\}. \tag{2.25}$$

By taking into account the slope corrections, we find that the modified equation for (2.22) actually reads

$$\partial_t\psi+u\partial_x\psi=\frac{\Delta x}{2}|\lambda(u)|(\partial_x u)(\partial_x\psi). \tag{2.26}$$

Therefore, unless $\partial_x u=0$ in which case second-order accuracy rigorously holds true, there is no hope to get rid of all the first-order terms in the right-hand side of (2.25), and the scheme is only quasi second-order in space. As pointed out by LeVeque [5, §9.3.1], this phenomenon is commonplace and occurs for all allegedly "second-order" accurate schemes with a variable velocity. There does exist a genuinely second-order accurate scheme for the variable-velocity case that can be derived from the Lax-Wendroff approach [5, Exercise 9.2], the results of which are typically similar [5, §9.3.1] but the stencil of which does not fit into the context of ALE codes.

For clarity of language, let us give names to the various situations that may occur depending on the sign configuration $S(i)$ in the neighborhood of cell i.

I. $u_{i-1/2}>0$ and $u_{i+1/2}\geq 0$ (left to right propagation);
II. $u_{i-1/2}\leq 0$ and $u_{i+1/2}<0$ (right to left propagation);
III. $u_{i-1/2}\leq 0$ and $u_{i+1/2}\geq 0$ (source);
IV. $u_{i-1/2}>0$ and $u_{i+1/2}<0$ (sink).

The min-max principle on ψ reads

$$\widehat{\psi}_i\in\begin{cases}\lfloor\psi_{i-1},\psi_i\rceil & \text{if } S(i)=\text{I},\\ \lfloor\psi_{i+1},\psi_i\rceil & \text{if } S(i)=\text{II},\\ \lfloor\psi_{i-1},\psi_i,\psi_{i+1}\rceil & \text{if } S(i)=\text{III}\cup\text{IV}.\end{cases} \tag{2.27}$$

The aim of the game is the same as before: we transport $\psi=\alpha$ and $\psi=\beta$ by (2.22) but require the min-max principle (2.27) on $\psi=\alpha,\beta$ and G. Of course, we rely on the same kind of analysis as in the uniform velocity case, although the discussion becomes much more involved.

For simplicity, we make additional assumptions in order to have statements similar to the uniform case.

1. The CFL condition is about half the previous one. For all cell i, we have

$$|\lambda_{i-1/2}|+|\lambda_{i+1/2}|<1. \tag{2.28}$$

2. The standard limiter function Λ used to compute the initial slopes (2.3) at cell i is of *strength* lesser than $2/(1-|\lambda|_i)$, i.e.,

$$|\Lambda(d_{i-1/2},d_{i+1/2})| \le \frac{2}{1-|\lambda|_i}\min\{|d_{i-1/2}|,|d_{i+1/2}|\}, \tag{2.29}$$

where $d_{i-1/2}=\psi_i-\psi_{i-1}$ and $d_{i+1/2}=\psi_{i+1}-\psi_i$. This rules out the ultrabee limiter (2.9), but authorizes minmod (2.4), van Leer (2.5), superbee (2.6) and even hyperbee (2.8).

3. There is no sequence of *source-sink* (C–D) or *sink-source* (D–C) configuration over two consecutive cells. Put another way,

$$\nexists i\in\mathbb{Z} \mid \lambda_{i-1/2}\lambda_{i+1/2}<0 \quad \text{and} \quad \lambda_{i+1/2}\lambda_{i+3/2}<0. \tag{2.30}$$

Such a "saw-tooth" sequence can be avoided by refining the mesh sufficiently, provided that the velocity field $u(x)$ depends continuously on x.

In preparation for Theorem 2.6, we set

$$\Phi_i = \min\{\tfrac{2}{|\lambda|_i}, \tfrac{2}{1-|\lambda|_i}\} \tag{2.31}$$

and introduce the local bounds

$$\begin{aligned} G_i^m &= \mathbb{1}_{\{S(i)=\mathrm{I}\}}\ \min\{G_{i-1},G_i\} \\ &\quad + \mathbb{1}_{\{S(i)=\mathrm{II}\}}\min\{G_{i+1},G_i\} + \mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}}\min\{G_{i-1},G_i,G_{i+1}\} \qquad (2.32a)\\ G_i^M &= \mathbb{1}_{\{S(i)=\mathrm{I}\}}\ \max\{G_{i-1},G_i\} \\ &\quad + \mathbb{1}_{\{S(i)=\mathrm{II}\}}\max\{G_{i+1},G_i\} + \mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}}\max\{G_{i-1},G_i,G_{i+1}\}, \qquad (2.32b)\end{aligned}$$

where $\mathbb{1}_{\{.\}}$ is the characteristic function. Once all the G^m and G^M have been computed over the domain, we consider the set $\mathscr{G}_i$ of all pairs (D_i^α, D_i^β) that satisfy:

- For case I (left-to-right propagation)

$$[\widetilde{D}_i^\alpha]^- \le D_i^\alpha \le [\widetilde{D}_i^\alpha]^+, \quad [\widetilde{D}_i^\beta]^- \le D_i^\beta \le [\widetilde{D}_i^\beta]^+, \quad \mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i, \tag{2.33}$$

with

$$\begin{aligned} \mathfrak{m}_i &= \Phi_i\max\{G_i^m - G_i \ \ ; G_i \ - G_{i+1}^M\} \qquad (2.34a)\\ \mathfrak{M}_i &= \Phi_i\min\{G_i \ - G_{i+1}^m\, ; G_i^M - G_i\}. \qquad (2.34b)\end{aligned}$$

- For case II (right-to-left propagation)

$$[\widetilde{D}_i^\alpha]^- \le D_i^\alpha \le [\widetilde{D}_i^\alpha]^+, \quad [\widetilde{D}_i^\beta]^- \le D_i^\beta \le [\widetilde{D}_i^\beta]^+, \quad \mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i, \tag{2.35}$$

with

$$\begin{aligned} \mathfrak{m}_i &= \Phi_i\max\{G_i^m - G_i \ \ ; G_i \ - G_{i-1}^M\} \qquad (2.36a)\\ \mathfrak{M}_i &= \Phi_i\min\{G_i \ - G_{i-1}^m\, ; G_i^M - G_i\}. \qquad (2.36b)\end{aligned}$$

- For case III (source)

$$[\widetilde{D}_i^\alpha]^- \le D_i^\alpha \le [\widetilde{D}_i^\alpha]^+, \quad [\widetilde{D}_i^\beta]^- \le D_i^\beta \le [\widetilde{D}_i^\beta]^+, \quad \mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i, \tag{2.37}$$

with

$$\begin{aligned} \mathfrak{m}_i &= \Phi_i\max\{G_i - G_{i-1}^M\, ; G_i - G_i^M\, ; G_i^m - G_i\, ; G_{i+1}^m - G_i\} \qquad (2.38a)\\ \mathfrak{M}_i &= \Phi_i\min\{G_i - G_{i-1}^m\, ; G_i - G_i^m\, ; G_i^M - G_i\, ; G_{i+1}^M - G_i\}. \qquad (2.38b)\end{aligned}$$

- For case IV (sink), $\mathscr{G}_i = \mathbb{R}^2$.

THEOREM 2.6. *Given an initial choice $(\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta)$ in accordance with (2.3), (2.29), let $\mathscr{G}_i \subset \mathbb{R}^2$ be the convex set introduced above. For all $i \in \mathbb{Z}$, define*

$$(D_i^\alpha, D_i^\beta) = \begin{cases} (\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta) & \text{if } \widetilde{D}_i^\alpha \widetilde{D}_i^\beta > 0 \\ \Pi_{\mathscr{G}_i}(\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta) & \text{otherwise} \end{cases} \tag{2.39}$$

where $\Pi_{\mathscr{G}_i}(.)$ denotes the projection onto the convex set $\mathscr{G}_i \subset \mathbb{R}^2$. Then, under assumptions (2.28) and (2.30), we have the min-max principle (2.27) for $\psi = \alpha, \beta$ and G at every cell i when updating (α, β) with scheme (2.22).

Proof. Since the proof is lengthy and relatively tedious, we are going to sketch out its beginning in order for the readers to grasp the key ideas. Applying the update formula (2.22) to $\psi = \alpha$ and β, then adding the equations together and substracting by G_i^m and G_i^M yields

$$\widehat{G}_i - G_i^m = A_i^m + B_{i-1}^m + C_{i+1}^m, \qquad \widehat{G}_i - G_i^M = A_i^M + B_{i-1}^M + C_{i+1}^M, \tag{2.40}$$

with

$$A_i^m = (1 - \lambda_{i-1/2}^+ + \lambda_{i+1/2}^-)(G_i - G_i^m) - (\lambda_{i+1/2}^+ + \lambda_{i-1/2}^-)\tfrac{1-|\lambda|_i}{2} D_i^G \tag{2.41a}$$

$$B_{i-1}^m = \lambda_{i-1/2}^+ (G_{i-1} - G_i^m) + \lambda_{i-1/2}^+ \tfrac{1-|\lambda|_{i-1}}{2} D_{i-1}^G \tag{2.41b}$$

$$C_{i+1}^m = \lambda_{i+1/2}^- (G_{i+1} - G_i^m) + \lambda_{i+1/2}^- \tfrac{1-|\lambda|_{i+1}}{2} D_{i+1}^G, \tag{2.41c}$$

and

$$A_i^M = (1 - \lambda_{i-1/2}^+ + \lambda_{i+1/2}^-)(G_i - G_i^M) - (\lambda_{i+1/2}^+ + \lambda_{i-1/2}^-)\tfrac{1-|\lambda|_i}{2} D_i^G \tag{2.42a}$$

$$B_{i-1}^M = \lambda_{i-1/2}^+ (G_{i-1} - G_i^M) + \lambda_{i-1/2}^+ \tfrac{1-|\lambda|_{i-1}}{2} D_{i-1}^G \tag{2.42b}$$

$$C_{i+1}^M = \lambda_{i+1/2}^- (G_{i+1} - G_i^M) + \lambda_{i+1/2}^- \tfrac{1-|\lambda|_{i+1}}{2} D_{i+1}^G, \tag{2.42c}$$

using the shorthand notation $D^G = D^\alpha + D^\beta$. In conformity with the splitting philosophy already explained for the uniform velocity case, we separately impose

$$A_i^m \geq 0, \qquad B_{i-1}^m \geq 0, \qquad C_{i+1}^m \geq 0 \tag{2.43a}$$

$$A_i^M \leq 0, \qquad B_{i-1}^M \leq 0, \qquad C_{i+1}^M \leq 0. \tag{2.43b}$$

We then express (2.43) in terms of D^G according to the sign configuration. In case I (resp. II), we drop out the identically vanishing and useless inequalties on $C_{i+1}^{m,M}$ (resp. $B_{i-1}^{m,M}$) and we shift the index for the inequalties on $B_{i-1}^{m,M}$ (resp. $C_{i+1}^{m,M}$) in order to ensure the min-max principle at the "receiving" neighbor $i+1$ (resp. $i-1$). In case III, we have to keep all the conditions and shift the index for them, because a source does have an influence on two receiving neighbors. In case IV, there is no need to change D_i^G because a sink does not have any influence on its neighbors and the min-max principle at a sink is actually ensured by conditions imposed to the two neighbors. ∎

Despite its apparent complexity, this procedure lends itself very well to numerical implementation. Instead of finding the image of the projection by hand, we can resort

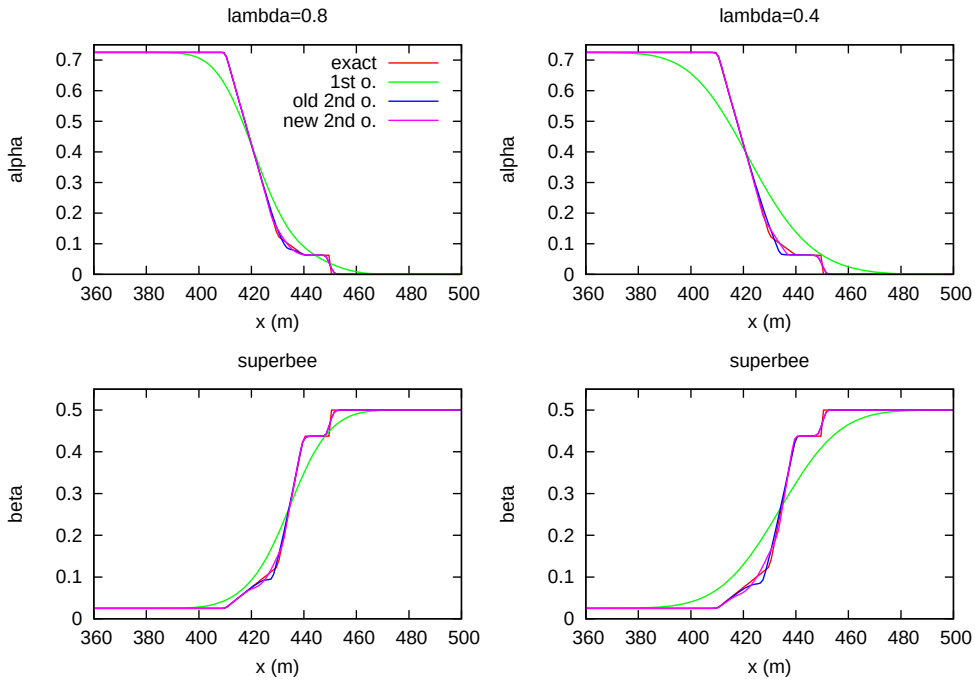

FIG. 2.3. *Main variables α (upper panels) and β (lower panels) for the sum problem.*

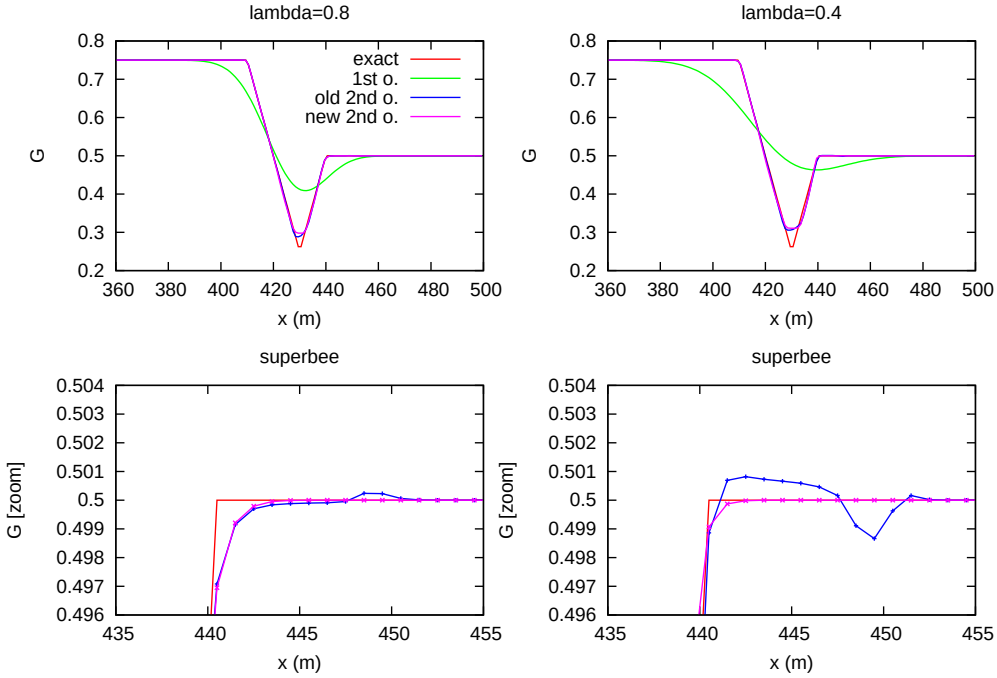

FIG. 2.4. *Control variable $G=\alpha+\beta$ for the sum problem.*

to a subroutine for quadratic minimization under linear inequality constraints. This will be addressed in Section 4.

REMARK 2.7. *It can be shown that in the* (D_i^α, D_i^β)*-plane, the set of points defined by the inequalities* $\mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i$ *in cases* I, II *and* III *always contain the strip*

$$\max\{[G_i - G_{i-1}]^-, [G_{i+1} - G_i]^-\} \le \frac{D_i^\alpha + D_i^\beta}{\Phi_i} \le \min\{[G_i - G_{i-1}]^+, [G_{i+1} - G_i]^+\}.$$

Therefore, if we consent to project onto a smaller convex set, it is possible to find the new slopes by explicit formulae similar to (2.21). The price to be paid is a slightly larger amount of dissipation.

2.3. Numerical results. In Figures 2.3 and 2.4, we compare the results of three different schemes and the exact solution for an experiment over the space-time domain $(x,t) \in [0,500\text{m}] \times [0,200\text{s}]$ with the positive velocity field $u = 2\text{m/s}$. The piecewise linear initial data

$$\begin{aligned}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}(x,t=0) = &\begin{bmatrix}0.725\\ 0.025\end{bmatrix}\mathbb{1}_{\{x<10\}} + \begin{bmatrix}1.025-0.03x\\ -0.025+0.005x\end{bmatrix}\mathbb{1}_{\{10<x<30\}}\\ &+ \begin{bmatrix}0.3125-0.00625x\\ -0.8125+0.03125x\end{bmatrix}\mathbb{1}_{\{30<x<40\}}\\ &+ \begin{bmatrix}0.0625\\ 0.4375\end{bmatrix}\mathbb{1}_{\{40<x<50\}} + \begin{bmatrix}0\\ 0.5\end{bmatrix}\mathbb{1}_{\{50<x\}}\end{aligned} \tag{2.44}$$

have been tailored in such a way that:

- α is decreasing and β is increasing, therefore we have $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) \le 0$ for all cell i in the domain, which activates the slope-projection mechanism all the time;
- both α and β are discontinuous at $x = 50\text{m}$, while being continuous with respect to x everywhere else;
- $G = \alpha + \beta$ exhibits a local minimum at $x = 30\text{m}$; moreover, G remains continuous across $x = 50\text{m}$.

The spatial domain is discretized by a uniform grid of size $\Delta x = 1\text{m}$. At the inlet boundary $x = 0\text{m}$, we maintain (α,β) at their left-most initial values $(0.725, 0.025)$. No special treatment is necessary at the outlet boundary $x = 500\text{m}$. Simulations are run with two values for the CFL ratio: $\lambda = 0.8$ and $\lambda = 0.4$.

The curves for the first-order accurate scheme are very much smeared out and are very sensitive to λ. Those for the two second-order accurate schemes are in very good agreement with the exact solution. What we mean by "old second-order" is the scheme with the initial slopes, computed component-wise. Of course, the "new second-order" is endowed with our coupling device for the slopes.

A close inspection of the curves reveals that in the vicinity of $x = 450\text{m}$ (which is equal to $50\text{m} + 2\text{m/s} \times 200\text{s}$), the old second-order accurate scheme exhibits spurious oscillations on the control variable G, as evidenced by the close-up in the lower panels of Figure 2.4. The smaller λ is, the stronger are the oscillations. As far as the new second-order is concerned, there is no violation of the min-max principle.

In order to assess how much the accuracy is sacrificed, we proceed to a study of convergence. Figure 2.5 displays the L^1 total relative error (that is, the sum of the L^1 relative errors on the two components α and β) of the computed solution versus the mesh size Δx on a log-log scale. The mesh size takes the decreasing values 2, 1, 0.5,

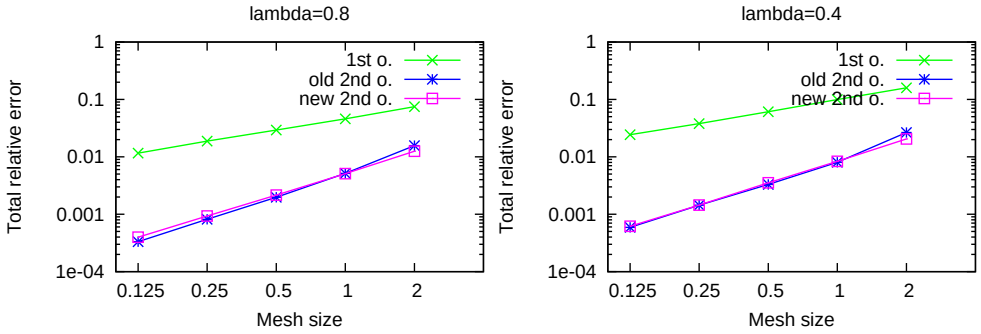

FIG. 2.5. *Accuracy measurements for the sum problem.*

0.25 and 0.125. We observe that the curves corresponding to the "old" scheme and the "new" one are very close to each other, and lie very far from that of the first-order scheme. This favorable feature is corroborated by Table 2.1, where we compute the order of convergence of the schemes by least-square fit.

	$\lambda=0.8$	$\lambda=0.4$
1st order	0.66739	0.68137
old 2nd order	1.37746	1.34669
new 2nd order	1.24037	1.26302

TABLE 2.1. *Orders of convergence for the sum problem.*

3. The fraction problem

We now turn to the transport of a total density and a partial density, the quotient of the latter by the former being a mass fraction. More specifically, we put

$$\mathbf{\Psi}=(\rho,\kappa)\in\mathbb{R}_+^2,\qquad Y(\mathbf{\Psi})=\frac{\kappa}{\rho}\in[0,1]. \tag{3.1}$$

3.1. Uniform velocity. From a time level to the next one, the update formulae for $(\rho,\ \kappa)$ are

$$\widehat{\rho}_i=\rho_i-\lambda\{[\rho_i+\tfrac{1-\lambda}{2}D_i^\rho]-[\rho_{i-1}+\tfrac{1-\lambda}{2}D_{i-1}^\rho]\} \tag{3.2a}$$

$$\widehat{\kappa}_i=\kappa_i-\lambda\{[\kappa_i+\tfrac{1-\lambda}{2}D_i^\kappa]-[\kappa_{i-1}+\tfrac{1-\lambda}{2}D_{i-1}^\kappa]\}, \tag{3.2b}$$

where λ is defined in (1.6). Consider the *initial slopes*

$$\widetilde{D}_i^\rho=\Lambda(\rho_i-\rho_{i-1},\rho_{i+1}-\rho_i),\qquad \widetilde{D}_i^\kappa=\Lambda(\kappa_i-\kappa_{i-1},\kappa_{i+1}-\kappa_i). \tag{3.3}$$

For the same reasons as in Lemma 2.1, we have the following result.

LEMMA 3.1. *If the slopes (D_j^ρ,D_j^κ) in (3.2) satisfy*

$$[\widetilde{D}_j^\rho]^-\le D_j^\rho\le[\widetilde{D}_j^\rho]^+,\qquad [\widetilde{D}_j^\kappa]^-\le D_j^\rho\le[\widetilde{D}_j^\kappa]^+ \tag{3.4}$$

for $j=i-1$ and $j=i$, then $\widehat{\rho}_i\in\lfloor\rho_{i-1},\rho_i\rceil$ and $\widehat{\kappa}_i\in\lfloor\kappa_{i-1},\kappa_i\rceil$.

The control fraction Y is computed by $Y_i=\kappa_i/\rho_i$ and $\widehat{Y}_i=\widehat{\kappa}_i/\widehat{\rho}_i$. Contrary to the intuition, it will be erroneous to think that carrying out the slope reconstruction on ρ and Y solves the problem. Indeed, the min (resp. max) of a product is not the product of the min values (resp. max values).

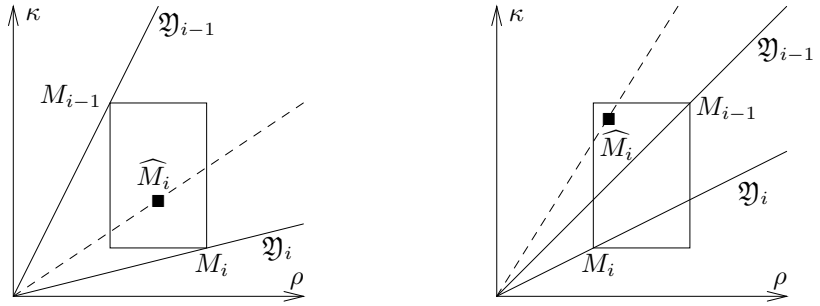

FIG. 3.1. *Geometric analysis of the min-max principle for the fraction problem.*

PROPOSITION 3.2. *If* $(\rho_i-\rho_{i-1})(\kappa_i-\kappa_{i-1})\le 0$ *and if the slopes* (D_j^ρ,D_j^κ) *satisfy (3.4) for* $j=i-1$ *and* $j=i$, *then* $\widehat{Y}_i\in\lfloor Y_{i-1},Y_i\rceil$.

Proof. In the quarter-plane $(\rho,\kappa)\in\mathbb{R}_+\times\mathbb{R}_+$, let us depict the points

$$M_{i-1}=(\rho_{i-1},\kappa_{i-1}),\quad M_i=(\rho_i,\kappa_i),\quad \widehat{M}_i=(\widehat{\rho}_i,\widehat{\kappa}_i). \tag{3.5}$$

as in Figure 3.1. The min-max principles $\widehat{\rho}_i\in\lfloor\rho_{i-1},\rho_i\rceil$ and $\widehat{\kappa}_i\in\lfloor\kappa_{i-1},\kappa_i\rceil$, which follow from Lemma 3.1, amount to saying that $\widehat{M}_i$ belongs to the rectangle $\mathscr{R}_i$ whose opposite vertices are M_{i-1} and M_i and whose sides are parallel to the horizontal and vertical axes. Draw the lines $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$ defined by $\kappa/\rho=Y_{i-1}$ and $\kappa/\rho=Y_i$.

If $(\rho_i-\rho_{i-1})(\kappa_i-\kappa_{i-1})\le 0$, then the rectangle $\mathscr{R}_i$ is entirely included in the cone of lines defined by the rays $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$. Therefore, the isoline of κ/ρ passing through $\widehat{M}_i$ lies between $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$, which is algebraically equivalent to $\widehat{Y}_i\in\lfloor Y_{i-1},Y_i\rceil$.

If $(\rho_i-\rho_{i-1})(\kappa_i-\kappa_{i-1})>0$, the rays $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$ cut the rectangle $\mathscr{R}_i$ into three pieces, and it may happen that $\widehat{M}_i$ lies outside the cone, which violates the desired min-max principle. ∎

To know what should be done for the case $(\rho_i-\rho_{i-1})(\kappa_i-\kappa_{i-1})>0$, we introduce

$$Y_i^m=\min\{Y_{i-1},Y_i\},\qquad Y_i^M=\max\{Y_{i-1},Y_i\}, \tag{3.6}$$

and seek sufficient conditions at a given cell i under the assumption $\lambda<1$.

LEMMA 3.3. *For a given cell* i, *if*

$$Y_i^M D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^M\rho_i) \le D_i^\kappa \le Y_i^m D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^m\rho_i), \tag{3.7a}$$

$$Y_i^m D_{i-1}^\rho - \tfrac{2}{1-\lambda}(\kappa_{i-1}-Y_i^m\rho_{i-1}) \le D_{i-1}^\kappa \le Y_i^M D_{i-1}^\rho - \tfrac{2}{1-\lambda}(\kappa_{i-1}-Y_i^M\rho_{i-1}), \tag{3.7b}$$

then $Y_i^m\le\widehat{Y}_i\le Y_i^M$.

Proof. A straightforward calculation shows that

$$\widehat{\kappa}_i-Y_i^m\widehat{\rho}_i=(1-\lambda)A_i^m+\lambda B_{i-1}^m,\qquad \widehat{\kappa}_i-Y_i^M\widehat{\rho}_i=(1-\lambda)A_i^M+\lambda B_{i-1}^M, \tag{3.8}$$

with

$$A_i^m = (\kappa_i \quad - Y_i^m \rho_i) \quad - \quad \tfrac{\lambda}{2}(D_i^\kappa \quad - Y_i^m D_i^\rho) \tag{3.9a}$$
$$A_i^M = (\kappa_i \quad - Y_i^M \rho_i) \quad - \quad \tfrac{\lambda}{2}(D_i^\kappa \quad - Y_i^M D_i^\rho) \tag{3.9b}$$
$$B_{i-1}^m = (\kappa_{i-1} - Y_i^m \rho_{i-1}) + \tfrac{1-\lambda}{2}(D_{i-1}^\kappa - Y_i^m D_{i-1}^\rho) \tag{3.9c}$$
$$B_{i-1}^M = (\kappa_{i-1} - Y_i^M \rho_{i-1}) + \tfrac{1-\lambda}{2}(D_{i-1}^\kappa - Y_i^M D_{i-1}^\rho). \tag{3.9d}$$

In order to ensure $\widehat{\kappa}_i - Y_i^m \widehat{\rho}_i \geq 0$ and $\widehat{\kappa}_i - Y_i^M \widehat{\rho}_i \leq 0$, our strategy consists in splitting the summands involved in (3.8). By forcibly imposing

$$A_i^m \geq 0, \quad A_i^M \leq 0, \quad B_{i-1}^m \geq 0, \quad B_{i-1}^M \leq 0, \tag{3.10}$$

we end up with the set of inequalities (3.7). ∎

THEOREM 3.4. *Given an initial choice $(\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa)$ in accordance with (3.3), let $\mathscr{Y}_i \subset \mathbb{R}^2$ be the set of all pairs (D_i^ρ, D_i^ρ) subject to the 8 linear inequality constraints*

$$[\widetilde{D}_i^\rho]^- \leq D_i^\rho \leq [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \leq D_i^\kappa \leq [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \leq D_i^\kappa \leq \mathfrak{M}_i(D_i^\rho), \tag{3.11}$$

with

$$\mathfrak{m}_i(D_i^\rho) = \max\{\, Y_i^M D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^M \rho_i)\,;\, Y_{i+1}^m D_i^\rho - \tfrac{2}{1-\lambda}(\kappa_i - Y_{i+1}^m \rho_i)\,\} \tag{3.12a}$$
$$\mathfrak{M}_i(D_i^\rho) = \min\{\, Y_i^m D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^m \rho_i)\,;\, Y_{i+1}^M D_i^\rho - \tfrac{2}{1-\lambda}(\kappa_i - Y_{i+1}^M \rho_i)\,\}. \tag{3.12b}$$

For all $i \in \mathbb{Z}$, define

$$(D_i^\rho, D_i^\kappa) = \begin{cases} (\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa) & \text{if } \widetilde{D}_i^\rho \widetilde{D}_i^\kappa < 0 \\ \Pi_{\mathscr{Y}_i}(\widetilde{D}_i^\rho, \widetilde{D}_i^\rho) & \text{otherwise} \end{cases} \tag{3.13}$$

where $\Pi_{\mathscr{Y}_i}(.)$ denotes the projection onto the convex set $\mathscr{Y}_i \subset \mathbb{R}^2$. Then, we have the min-max principles

$$\widehat{\rho}_i \in \lfloor \rho_{i-1}, \rho_i \rceil, \qquad \widehat{\kappa}_i \in \lfloor \kappa_{i-1}, \kappa_i \rceil, \qquad \widehat{Y}_i \in \lfloor Y_{i-1}, Y_i \rceil, \tag{3.14}$$

at every cell i when updating (ρ, κ) with scheme (3.2).

Proof. The proof is similar to that of Theorem 2.4. ∎

The practical implementation of the projection onto $\mathscr{Y}_i$ in this problem can be done via the explicit formulae given in Proposition 3.5. Figure 3.2 displays a few situations for a locally increasing or decreasing behavior of Y. It can be readily proven that the constraints $\mathfrak{m}_i(D_i^\rho) \leq D_i^\kappa \leq \mathfrak{M}_i(D_i^\rho)$ correspond in reality to a triangle in the (D_i^ρ, D_i^κ)-plane. The slopes of its sides are Y_{i-1}, Y_i and Y_{i+1}. The side with slope Y_i passes through the origin and connects the points $-\frac{2}{1-\lambda}(\rho_i, \kappa_i)$ and $\frac{2}{1-\lambda}(\rho_i, \kappa_i)$.

Should a local extremum occur, i.e., $(Y_{i-1} - Y_i)(Y_{i+1} - Y_i) > 0$, this triangle degenerates into the segment joining these two points. Hence, $D_i^\kappa = Y_i D_i^\rho$ whenever the projection operator $\Pi_{\mathscr{Y}_i}$ is activated, and we formally have $D_i^Y = (D_i^\kappa - Y_i D_i^\rho)/\rho_i = 0$. This testifies to a clipping mechanism on Y.

PROPOSITION 3.5. *The projection operator $\Pi_{\mathscr{Y}_i}$ introduced in (3.13) can be carried out via the following two-step procedure:*

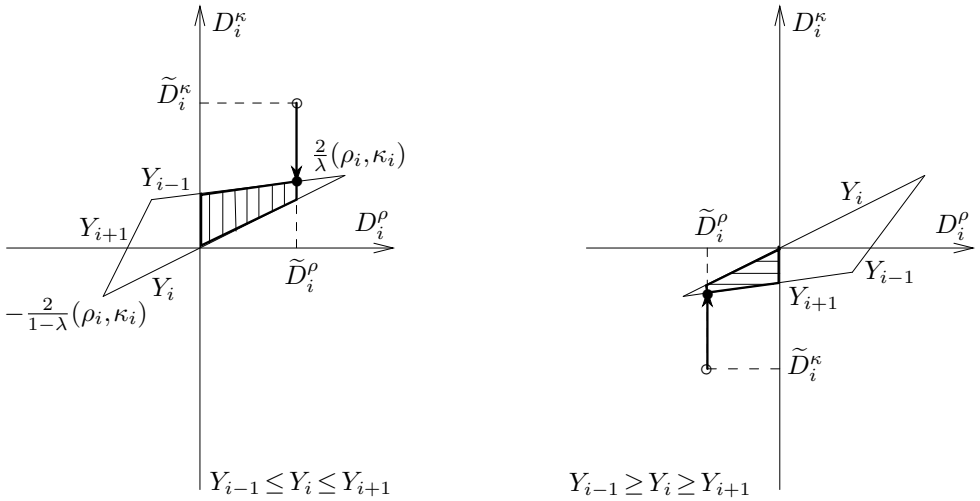

FIG. 3.2. *Projection of* $(\widetilde{D}_i^\rho,\widetilde{D}_i^\kappa)$ *onto the convex set* $\mathscr{Y}_i$ *for the fraction problem.*

1. Compute

$$\breve{D}_i^\kappa = \min\{[\widetilde{D}_i^\kappa]^+;\tfrac{2}{\lambda}\kappa_i\} + \max\{[\widetilde{D}_i^\kappa]^-;-\tfrac{2}{1-\lambda}\kappa_i\} \tag{3.15a}$$

$$\breve{D}_i^\rho = \min\{[\widetilde{D}_i^\rho]^+;\tfrac{2}{\lambda}\rho_i\} + \max\{[\widetilde{D}_i^\rho]^-;-\tfrac{2}{1-\lambda}\rho_i\}. \tag{3.15b}$$

2. Truncate

$$D_i^\kappa = \breve{D}_i^\kappa - \min\{[K_i^\uparrow]^+;[K_i^\uparrow - L_i^\uparrow]^+\} - \max\{[K_i^\downarrow]^-;[K_i^\downarrow - L_i^\downarrow]^-\} \tag{3.16a}$$

$$D_i^\rho = \breve{D}_i^\rho - \min\{[R_i^\uparrow]^+;[R_i^\uparrow - S_i^\uparrow]^+\} - \max\{[R_i^\downarrow]^-;[R_i^\downarrow - S_i^\downarrow]^-\}, \tag{3.16b}$$

with

$$K_i^\uparrow = [\breve{D}_i^\kappa]^+ - Y_i[\breve{D}_i^\rho]^+, \quad K_i^\downarrow = [\breve{D}_i^\kappa]^- - Y_i[\breve{D}_i^\rho]^- \tag{3.17a}$$

$$R_i^\uparrow = [\breve{D}_i^\rho]^+ - \frac{1}{Y_i}[\breve{D}_i^\kappa]^+, \quad R_i^\downarrow = [\breve{D}_i^\rho]^- - \frac{1}{Y_i}[\breve{D}_i^\kappa]^-, \tag{3.17b}$$

and

$$L_i^\uparrow = \operatorname{minmod}(\tfrac{2}{\lambda}(\kappa_i - Y_{i-1}\rho_i) + (Y_{i-1}-Y_i)[\breve{D}_i^\rho]^+, \tag{3.18a}$$

$$-\tfrac{2}{1-\lambda}(\kappa_i - Y_{i+1}\rho_i) + (Y_{i+1}-Y_i)[\breve{D}_i^\rho]^+) \tag{3.18b}$$

$$L_i^\downarrow = \operatorname{minmod}(\tfrac{2}{\lambda}(\kappa_i - Y_{i-1}\rho_i) + (Y_{i-1}-Y_i)[\breve{D}_i^\rho]^-, \tag{3.18c}$$

$$-\tfrac{2}{1-\lambda}(\kappa_i - Y_{i+1}\rho_i) + (Y_{i+1}-Y_i)[\breve{D}_i^\rho]^-) \tag{3.18d}$$

$$S_i^\uparrow = \operatorname{minmod}(\tfrac{2}{\lambda}(\rho_i - \tfrac{1}{Y_{i-1}}\kappa_i) + (\tfrac{1}{Y_{i-1}} - \tfrac{1}{Y_i})[\breve{D}_i^\kappa]^+, \tag{3.18e}$$

$$-\tfrac{2}{1-\lambda}(\rho_i - \tfrac{1}{Y_{i+1}}\kappa_i) + (\tfrac{1}{Y_{i+1}} - \tfrac{1}{Y_i})[\breve{D}_i^\kappa]^+) \tag{3.18f}$$

$$S_i^\downarrow = \operatorname{minmod}(\tfrac{2}{\lambda}(\rho_i - \tfrac{1}{Y_{i-1}}\kappa_i) + (\tfrac{1}{Y_{i-1}} - \tfrac{1}{Y_i})[\breve{D}_i^\kappa]^-, \tag{3.18g}$$

$$-\tfrac{2}{1-\lambda}(\rho_i - \tfrac{1}{Y_{i+1}}\kappa_i) + (\tfrac{1}{Y_{i+1}} - \tfrac{1}{Y_i})[\breve{D}_i^\kappa]^-). \tag{3.18h}$$

Proof. The above formulae are the algebraic translation of the geometric ideas sketched out in Figure 3.2. ∎

3.2. Variable velocity field. The setting is identical to that of the sum problem. Introduce the local bounds

$$\begin{aligned} Y_i^m =& \mathbb{1}_{\{S(i)=\mathrm{I}\}} \quad \min\{Y_{i-1}, Y_i\} \\ &+ \mathbb{1}_{\{S(i)=\mathrm{II}\}} \min\{Y_{i+1}, Y_i\} + \mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}} \min\{Y_{i-1}, Y_i, Y_{i+1}\} \qquad (3.19a) \\ Y_i^M =& \mathbb{1}_{\{S(i)=\mathrm{I}\}} \quad \max\{Y_{i-1}, Y_i\} \\ &+ \mathbb{1}_{\{S(i)=\mathrm{II}\}} \max\{Y_{i+1}, Y_i\} + \mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}} \max\{Y_{i-1}, Y_i, Y_{i+1}\}. \qquad (3.19b) \end{aligned}$$

Once all the Y^m and Y^M have been computed over the domain, we consider the set $\mathscr{Y}_i$ of all pairs (D_i^ρ, D_i^κ) that satisfy:

- For case I (left-to-right propagation)

$$[\widetilde{D}_i^\rho]^- \le D_i^\rho \le [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \le D_i^\kappa \le [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho), \qquad (3.20)$$

  with

$$\begin{aligned} \mathfrak{m}_i(D_i^\rho) =& \max\{ Y_i^M D_i^\rho + \Phi_i(\kappa_i - Y_i^M \rho_i) \, ; \, Y_{i+1}^m D_i^\rho - \Phi_i(\kappa_i - Y_{i+1}^m \rho_i) \} \qquad (3.21a) \\ \mathfrak{M}_i(D_i^\rho) =& \min\{ Y_i^m D_i^\rho + \Phi_i(\kappa_i - Y_i^m \rho_i) \, ; \, Y_{i+1}^M D_i^\rho - \Phi_i(\kappa_i - Y_{i+1}^M \rho_i) \}. \qquad (3.21b) \end{aligned}$$

- For case II (right-to-left propagation)

$$[\widetilde{D}_i^\rho]^- \le D_i^\rho \le [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \le D_i^\kappa \le [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho), \qquad (3.22)$$

  with

$$\begin{aligned} \mathfrak{m}_i(D_i^\rho) =& \max\{ Y_i^M D_i^\rho + \Phi_i(\kappa_i - Y_i^M \rho_i) \, ; Y_{i-1}^m D_i^\rho - \Phi_i(\kappa_i - Y_{i-1}^m \rho_i) \} \qquad (3.23a) \\ \mathfrak{M}_i(D_i^\rho) =& \min\{ Y_i^m D_i^\rho + \Phi_i(\kappa_i - Y_i^m \rho_i) \, ; Y_{i-1}^M D_i^\rho - \Phi_i(\kappa_i - Y_{i-1}^M \rho_i) \}. \qquad (3.23b) \end{aligned}$$

- For case III (source)

$$[\widetilde{D}_i^\rho]^- \le D_i^\rho \le [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \le D_i^\kappa \le [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho), \qquad (3.24)$$

  with

$$\begin{aligned} \mathfrak{m}_i(D_i^\rho) =& \max\{ Y_{i-1}^M D_i^\rho + \Phi_i(\kappa_i - Y_{i-1}^M \rho_i) ; Y_i^m D_i^\rho - \Phi_i(\kappa_i - Y_i^m \rho_i); \\ & \qquad Y_i^M D_i^\rho + \Phi_i(\kappa_i - Y_i^M \rho_i) \, ; Y_{i+1}^m D_i^\rho - \Phi_i(\kappa_i - Y_{i+1}^m \rho_i) \} \qquad (3.25a) \\ \mathfrak{M}_i(D_i^\rho) =& \min\{ Y_{i-1}^m D_i^\rho + \Phi_i(\kappa_i - Y_{i-1}^m \rho_i) ; Y_i^m D_i^\rho + \Phi_i(\kappa_i - Y_i^m \rho_i); \\ & \qquad Y_i^M D_i^\rho - \Phi_i(\kappa_i - Y_i^M \rho_i) \, ; Y_{i+1}^M D_i^\rho - \Phi_i(\kappa_i - Y_{i+1}^M \rho_i) \}. \qquad (3.25b) \end{aligned}$$

- For case IV (sink), $\mathscr{Y}_i = \mathbb{R}^2$.

THEOREM 3.6. *Given an initial choice $(\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa)$ in accordance with (3.3), (2.29), let $\mathscr{Y}_i \subset \mathbb{R}^2$ be the convex set introduced above. For all $i \in \mathbb{Z}$, define*

$$(D_i^\rho, D_i^\kappa) = \begin{cases} (\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa) & \text{if } \widetilde{D}_i^\rho \widetilde{D}_i^\kappa < 0 \\ \Pi_{\mathscr{Y}_i}(\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa) & \text{otherwise} \end{cases} \qquad (3.26)$$

where $\Pi_{\mathscr{Y}_i}(.)$ denotes the projection onto the convex set $\mathscr{Y}_i \subset \mathbb{R}^2$. Then, under assumptions (2.28) and (2.30), we have the min-max principle (2.27) for $\psi = \rho, \kappa$ and Y at every cell i when updating (ρ, κ) with scheme (2.22).

Proof. The proof is similar to that of Theorem 2.6. ∎

Again, we recommend a minimization subroutine to perform the projection.

REMARK 3.7. *In the* (D_i^ρ, D_i^κ)*-plane, let* A *and* B *be the points located at*

$$\mathtt{A} = -\frac{1}{\Phi_i}(\rho_i, \kappa_i), \qquad \mathtt{B} = \frac{1}{\Phi_i}(\rho_i, \kappa_i).$$

It can be shown that, the set of points defined by the inequalities $\mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho)$ *in cases* I, II *and* III *is the segment* AB *if* $(Y_i - Y_{i-1})(Y_i - Y_{i+1}) \ge 0$. *For* $(Y_i - Y_{i-1})(Y_i - Y_{i+1}) < 0$, *this domain always contains the triangle* ABC, *in which the slope of* (AC) *is* Y_{i+1} *and the slope of* (CB) *is* Y_{i-1}. *Therefore, if we consent to project onto a smaller convex set, it is possible to find the new slopes by explicit formulae similar to (3.15)–(3.18).*

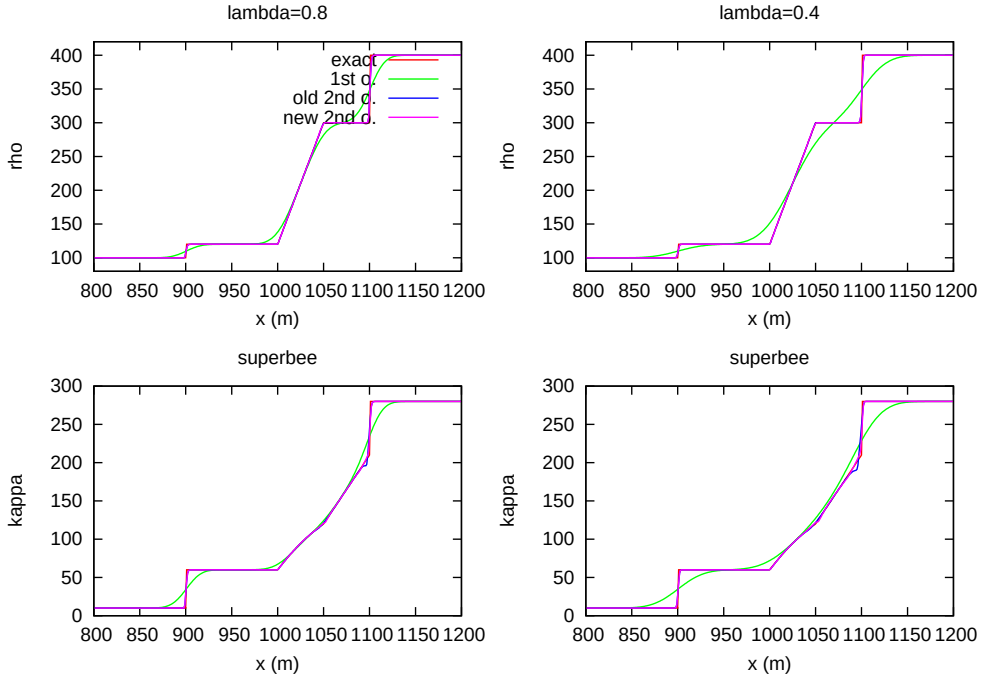

FIG. 3.3. *Main variables* ρ *(upper panels) and* κ *(lower panels) for the fraction problem.*

3.3. Numerical results. In Figures 3.3 and 3.4, we compare the results of three different schemes and the exact solution for an experiment over the space-time domain $(x,t) \in [0, 1200\text{m}] \times [0, 400\text{s}]$ with the positive velocity field $u = 2\text{m/s}$. The initial data

$$\begin{aligned}\begin{bmatrix}\rho\\ Y\end{bmatrix}(x, t=0) = &\begin{bmatrix}100\\ 0.1\end{bmatrix}\mathbb{1}_{\{x<100\}} + \begin{bmatrix}120\\ 0.5\end{bmatrix}\mathbb{1}_{\{100<x<200\}}\\ &+ \begin{bmatrix}-600+3.6x\\ 0.9-0.002x\end{bmatrix}\mathbb{1}_{\{200<x<250\}}\\ &+ \begin{bmatrix}300\\ -1.1+0.006x\end{bmatrix}\mathbb{1}_{\{250<x<300\}} + \begin{bmatrix}400\\ 0.7\end{bmatrix}\mathbb{1}_{\{300<x\}}\end{aligned} \tag{3.27}$$

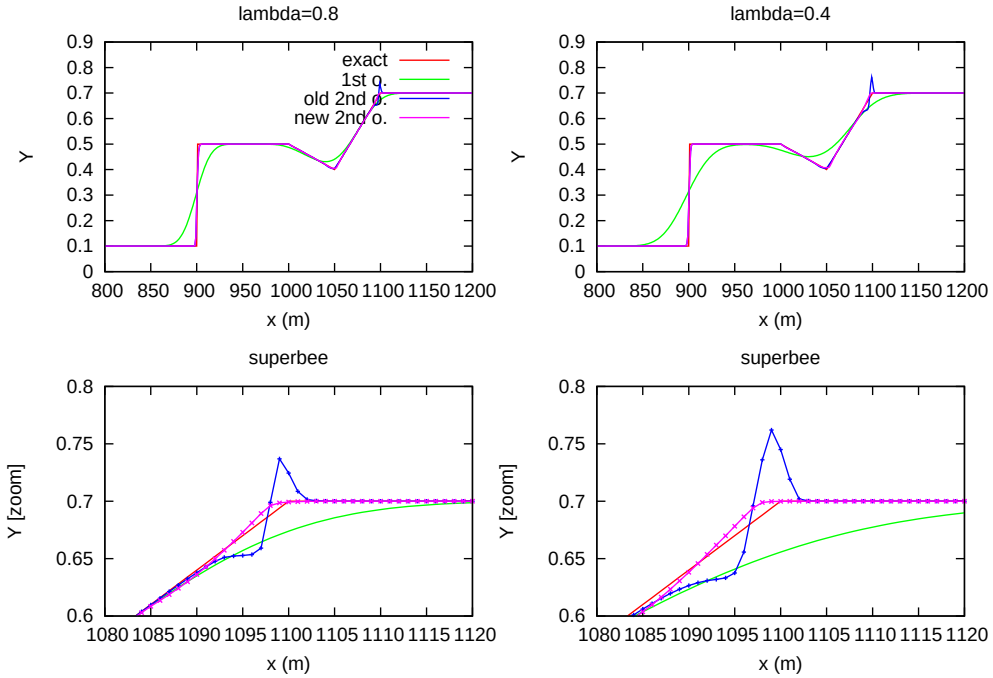

FIG. 3.4. *Control variable $Y=\kappa/\rho$ for the fraction problem.*

have been tailored in such a way that:

- ρ and $\kappa=Y\rho$ are both increasing, therefore we have $(\rho_i-\rho_{i-1})(\kappa_i-\kappa_{i-1})\geq 0$ for all cell i in the domain, which activates the slope-projection mechanism all the time;
- both ρ and κ are discontinuous at $x=100$m and at $x=300$m, while being continuous with respect to x everywhere else;
- Y exhibits a local minimum at $x=250$m; moreover, Y is discontinuous at $x=100$m but remains continuous across $x=300$m.

The space domain is discretized by a uniform grid of size $\Delta x=1$m. At the inlet boundary $x=0$m, we maintain (ρ,κ) at their left-most initial values $(100,10)$. No special treatment is necessary at the outlet boundary $x=1200$m. Simulations are run with two values for the CFL ratio: $\lambda=0.8$ and $\lambda=0.4$.

The curves for the first-order scheme are very much smeared out and turn out to be very sensitive to λ. Those for the two second-order accurate schemes are in very good agreement with the exact solution. The labels "old second-order" and "new second-order" have the same meaning as in section 2.3. We see that in the vicinity of $x=1100$m (which is equal to $300\text{m}+2\text{m/s}\times 400\text{m}$), the old second-order accurate scheme does not comply with the min-max principle on the control variable Y. The violation increases as the CFL ratio λ decreases. As for the new second-order accurate scheme, it does not exhibit any oscillation on Y, as testified by the lower panel of Figure 3.4.

The behavior of the accuracy is shown in the study of convergence of Figure 3.5, where the total L^1-relative error (that is, the sum of the L^1-relative errors on the two

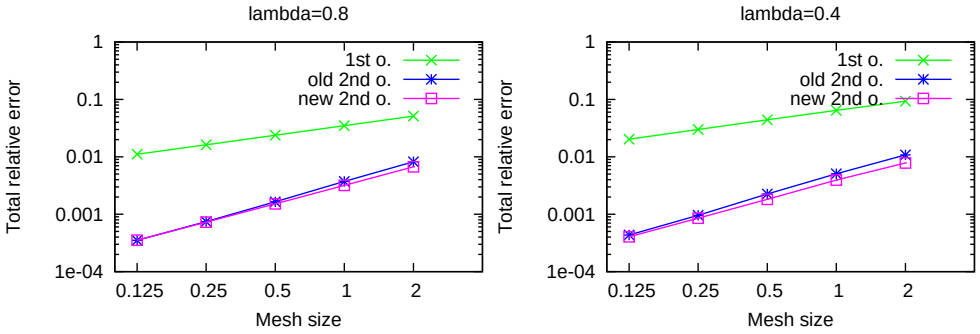

FIG. 3.5. *Accuracy measurements for the fraction problem.*

components ρ and κ) of the computed solution is displayed versus the mesh size Δx on a log-log scale. The mesh size takes the sequence of decreasing values 2, 1, 0.5, 0.25 and 0.125. Again, there is no significant discrepancy between the curves corresponding to the "old" scheme and the "new" one. According to Table 3.1, where we compute the order of convergence of the schemes by least-square fit, the convergence of the new second-order accurate scheme is somewhat slower than that of the old second-order accurate scheme.

	$\lambda = 0.8$	$\lambda = 0.4$
1st order	0.55155	0.55049
old 2nd order	1.14250	1.16673
new 2nd order	1.06267	1.07509

TABLE 3.1. *Orders of convergence for the fraction problem.*

4. The general problem

We go back to the problem stated in the Introduction. The ideas presented for the sum problem and the fraction problem can be carried over to the case of several control variables $\mathbf{G}$, each of them being a first-order rational fraction with respect to $\mathbf{\Psi} \in \mathbb{R}_+^P$, that is,

$$G^q(\mathbf{\Psi}) = \frac{a_0^q + a_1^q \psi^1 + \ldots + a_P^q \psi^P}{b_0^q + b_1^q \psi^1 + \ldots + b_P^q \psi^P}, \tag{4.1}$$

with $b_p^q \geq 0$ for all $1 \leq p \leq P$, $1 \leq q \leq Q$. This class of homographic functions is wide enough to represent a vast majority of control variables in real-life applications.

4.1. Uniform velocity.

Assuming $u > 0$, we define

$$(\psi_i^q)^m = \min\{\psi_{i-1}^q, \psi_i^q\}, \qquad (\psi_i^q)^M = \max\{\psi_{i-1}^q, \psi_i^q\}, \tag{4.2}$$

and

$$(G_i^q)^m = \min\{G_{i-1}^q, G_i^q\}, \qquad (G_i^q)^M = \max\{G_{i-1}^q, G_i^q\}. \tag{4.3}$$

Our objective is to find the slopes $\mathbf{D}_i = (D_i^1, \ldots, D_i^P)$, which should be as close as possible to the initial slopes $\widetilde{\mathbf{D}}_i = (\widetilde{D}_i^1, \ldots, \widetilde{D}_i^P)$ computed component-wise by a standard limiter function, so that by updating $\mathbf{\Psi}$ with (1.5), we have not only

$$(\psi_i^q)^m \leq \widehat{\psi}_i^q \leq (\psi_i^q)^M, \tag{4.4}$$

but also

$$(G_i^q)^m \leq \widehat{G}_i^q = G^q(\widehat{\mathbf{\Psi}}_i) \leq (G_i^q)^M. \tag{4.5}$$

Getting rid of the denominator in G^q, the above condition can be cast into two linear inequalities involving $(\mathbf{\Psi}_{i-1}, \mathbf{\Psi}_i)$ and $(\mathbf{D}_{i-1}, \mathbf{D}_i)$. The splitting strategy enables us to break these inequalities into local conditions which do not couple $\mathbf{D}_{i-1}$ and $\mathbf{D}_i$. These conditions, once gathered, express that we must project the initial guess $\widetilde{\mathbf{D}}_i$ onto a convex set $\mathscr{G}_i \subset \mathbb{R}^P$ defined by $2P$ bound constraints (to ensure monotonicity on $\mathbf{\Psi}$) and $4Q$ non-trivial linear inequalities (to ensure monotonicity on $\mathbf{G}$).

To carry out this projection, we reformulate the projection operator as a quadratic minimization problem

$$\min_{\mathbf{D}_i \in \mathscr{G}_i} \tfrac{1}{2} \|\mathbf{D}_i - \widetilde{\mathbf{D}}_i\|^2 \tag{4.6}$$

subject to linear inequality constraints. We recall that by virtue of Hilbert's theorem about projection onto a convex non-empty set, there is a unique solution to problem (4.6). In the context of the applications we have in mind, the Euclidean norm does make sense, insofar as the components of Ψ are homogeneous to a density. We advocate the use of an existing subroutine, e.g., the QL algorithm by Schittkowski [6], the advantage of which lies in its fast convergence. Moreover, it can be initialized with $\mathbf{D}_i = \widetilde{\mathbf{D}}_i$, which is not necessarily a feasible point.

Before launching the optimization procedure, however, we have to carefully determine the regions in the $\mathbf{D}_i$-space for which the min-max principle on $\mathbf{G}$ is automatically guaranteed and for which there is no need to perform projection (for the sum problem, this region is $D_i^\alpha D_i^\beta > 0$ and for the fraction problem, this is $D_i^\rho D_i^\kappa < 0$). This crucial preliminary step is meant to maintain sharp profiles. It can only be done on a case-by-case basis.

4.2. Variable velocity field. The ideas remain the same as in the uniform velocity case, but the calculations are trickier. On one hand, the definitions of the local bounds depend on the sign configuration at the edges of each cell. On the other hand, the inequalities to be split now involve $\mathbf{D}_{i-1}, \mathbf{D}_i$ and $\mathbf{D}_{i+1}$. As a consequence, after imposing positivity or negativity to the summands separately, we end up with more than $4Q$ non-trivial combinations for case III (source). Nevertheless, this is not a difficulty because the hard part of the job is done by the optimization subroutine.

The extra time incurred by the latter depends on the size of P and Q. Numerical experiments reveal that for a typical multi-specie flow model ($P \approx 10$, $Q \approx 5$), such as in [1,8], the CPU ratio never exceeds 2. Besides, we have to be aware of the fact that the remap phase contributes little to the overall computational time of the nonlinear Euler code. From this global point of view, the reward brought by the min-max principle on the main and control variables is worth an increase by a factor 2 in the CPU time of the remap phase, which is not really significant!

4.3. A selected example. Consider the following "two-sum four-species" problem. Let $\mathbf{\Psi} = (\alpha, \beta, \gamma, \delta) \in \mathbb{R}_+^4$ and

$$\mathbf{G}(\mathbf{\Psi}) = (e, f) = (7\alpha + \gamma, 2\beta + 2\gamma + \delta). \tag{4.7}$$

The densities α, β, γ, δ respectively correspond to the four species C_7H_{16}, O_2, CO_2, H_2O, related to each other through the reversible chemical reaction

$$C_7H_{16} + 11O_2 \rightleftarrows 7CO_2 + 8H_2O \tag{4.8}$$

which takes place at the known rate

$$\mathscr{R}(\alpha, \beta, \gamma, \delta) = K_+ \alpha \beta^{11} - K_- \gamma^7 \delta^8. \tag{4.9}$$

This implies the evolution equations

$$\partial_t \alpha + u \partial_x \alpha = - \quad \mathscr{R}(\alpha, \beta, \gamma, \delta) \tag{4.10a}$$
$$\partial_t \beta + u \partial_x \beta = -11 \mathscr{R}(\alpha, \beta, \gamma, \delta) \tag{4.10b}$$
$$\partial_t \gamma + u \partial_x \gamma = \quad 7 \mathscr{R}(\alpha, \beta, \gamma, \delta) \tag{4.10c}$$
$$\partial_t \delta + u \, \partial_x \delta = \quad 8 \mathscr{R}(\alpha, \beta, \gamma, \delta). \tag{4.10d}$$

By linear combinations of (4.10), it can be inferred that

$$\partial_t e + u \partial_x e = \partial_t f + u \partial_x f = 0, \tag{4.11}$$

which highlights e and f as control variables. In the present case, e is the carbon tracer, and f the oxygen tracer. Since the third component γ is involved in the definition of both tracers e and f, this problem cannot be decomposed into two independent one-sum problems.

As in Section 2, we consider the space-time domain $(x,t) \in [0, 500\text{m}] \times [0, 200\text{s}]$, over which linear advection is performed at the velocity $u = 2\text{m/s}$. The initial data

$$\begin{aligned} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} (x, t=0) = & \begin{bmatrix} 0.35 \\ 0.2 \\ 0.025 \\ 1.55 \end{bmatrix} \mathbb{1}_{\{x<10\}} + \begin{bmatrix} 0.45 - 0.01x \\ 0.175 + 0.0025x \\ 0.0125 + 0.00125x \\ 1.85 - 0.03x \end{bmatrix} \mathbb{1}_{\{10<x<30\}} \\ & + \begin{bmatrix} 0.3 - 0.005x \\ -1.1 + 0.045x \\ -1.6 + 0.055x \\ 1.1 - 0.005x \end{bmatrix} \mathbb{1}_{\{30<x<40\}} \\ & + \begin{bmatrix} 0.1 \\ 0.7 \\ 0.6 \\ 0.9 \end{bmatrix} \mathbb{1}_{\{40<x<50\}} + \begin{bmatrix} 0.05 \\ 0.75 \\ 0.95 \\ 0.1 \end{bmatrix} \mathbb{1}_{\{50<x\}} \end{aligned} \tag{4.12}$$

have been designed so that the slope-projection mechanism is always activated. The main variables $\alpha, \beta, \gamma, \delta$ are monotone and discontinuous at $x = 50\text{m}$, while the control variables e, f are continous across this point and exhibit local minima. The space domain is discretized by a uniform grid of size $\Delta x = 1\text{m}$. Since we are primarily interested in the transportation of data, we do not prescribe any chemical reaction here. In other words, we set $\mathscr{R} = 0$. Nevertheless, in order to get closer to "real" operating conditions, we run the simulations with the CFL ratio $\lambda = 0.1$: indeed,

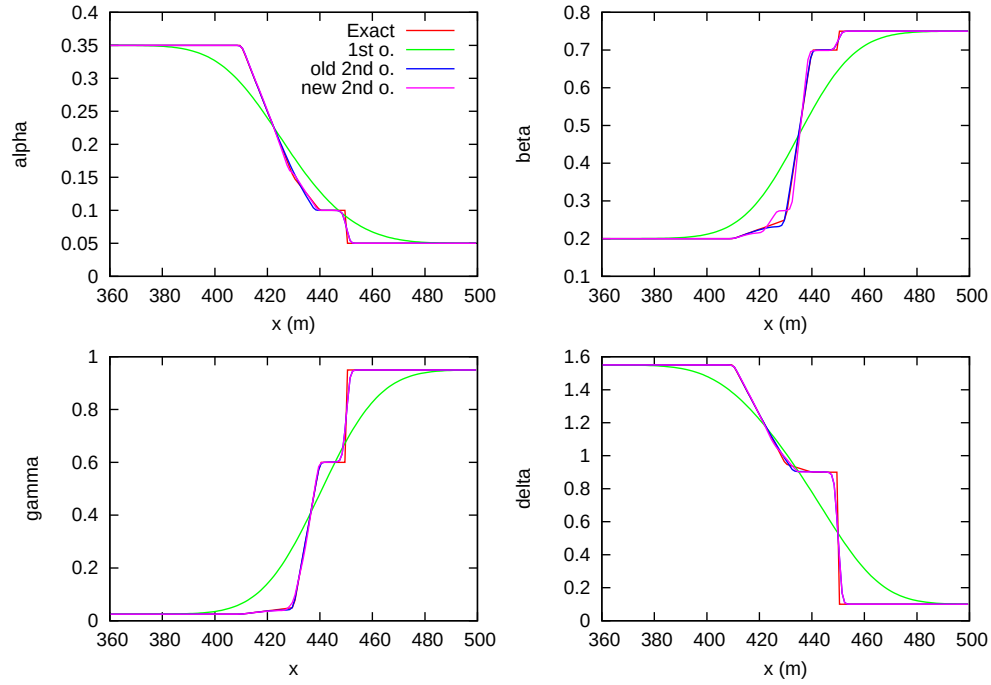

FIG. 4.1. *Main variables $\alpha,\beta,\gamma,\delta$ for the two-sum four-species problem.*

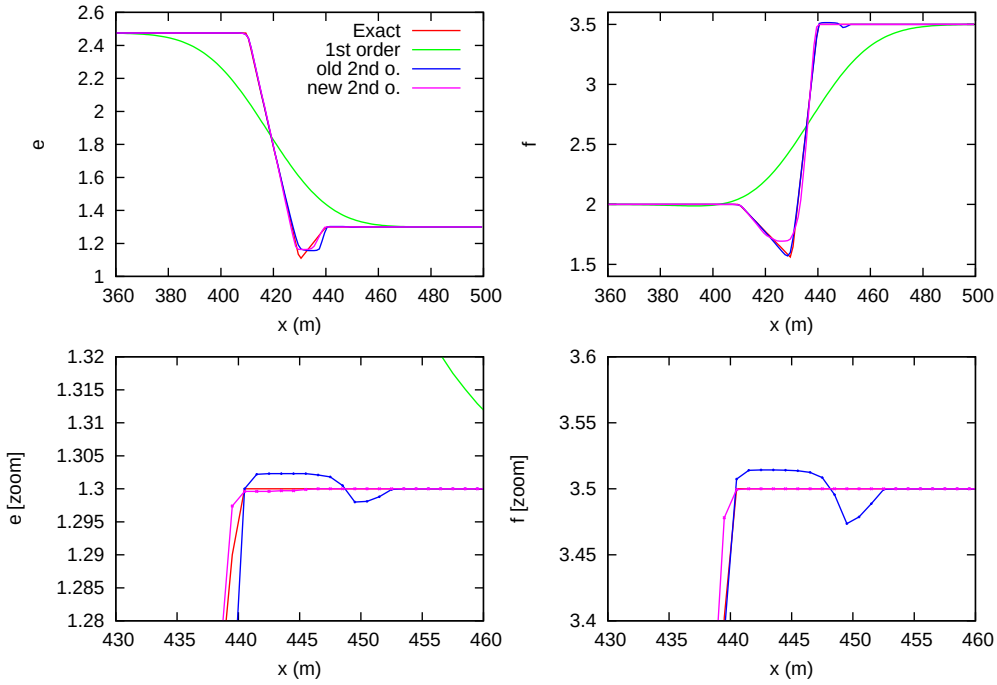

FIG. 4.2. *Control variables e (left) and f (right) for the two-sum four-species problem.*

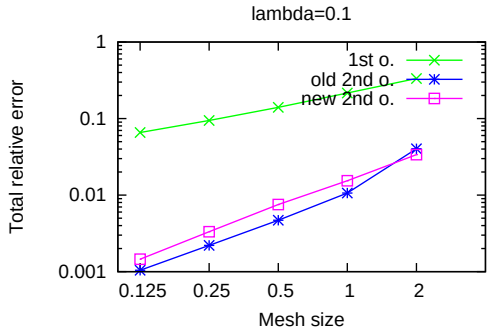

FIG. 4.3. *Accuracy measurements for the two-sum four-species problem.*

chemical reactions usually induce very small time-steps of this order of magnitude. At the inlet boundary $x = 0$m, we maintain $(\alpha, \beta, \gamma, \delta)$ at their left-most initial values $(0.35, 0.2, 0.025, 1.55)$. No special treatment is necessary at the outlet boundary $x =$ 500m.

As shown in Figs 4.1 and 4.2, the first-order accurate scheme is so dissipative that the local minima in e and f are not correctly captured. The two second-order accurate schemes are in very good agreement with the exact solution, but the "old" one using decoupled slopes has spurious oscillations in the region where the tracers e, f have to remain constant. The "new" one enables us to avoid this phenomenon, at the price of a less sharp capturing of the local minima.

The behavior of the total L^1-relative error (that is, the sum of the L^1-relative errors on the four main variables α, β, γ, δ) is shown in the study of convergence of Figure 4.3, equipped with a log-log scale. The difference between the "old" scheme and the "new" one is now stronger and more visible than in the two previous problems (sum and fraction). However, the accuracy remains at a very acceptable level, compared to the first-order scheme. The numerical orders of convergence, obtained by regression, are:

1st order	0.58790
old 2nd order	1.28209
new 2nd order	1.12846

5. Conclusion

We hope the slope-reconstruction methodology proposed in this paper, based on a mathematically sound analysis while being not too much expensive, will be helpful to the practitioners who have to daily face similar problems. Current works are in progress in order to extend this approach to multi-dimensional problems in a rigorous way.

Acknowledgement. The author is grateful to Frédéric Coquel and Bruno Scheurer for several stimulating discussions. I also wish to thank the two anonymous reviewers for their helpful comments.

REFERENCES

[1] O. Colin and A. Benkenida, *The 3-zones extended coherent flame model (ECFM3Z) for computing premixed/diffusion combustion*, Oil & Gas Sci. Tech., 59, 593–609, 2004.
[2] F. Coquel, Q.L. Nguyen, M. Postel and Q.H. Tran, *Entropy-satisfying relaxation method with large time-steps for Euler IBVPs*, Submitted, 2007.
[3] C.W. Hirt, A.A. Amsden and J.L. Cook, *An arbitrary Lagrangian-Eulerian computing method for all flow speeds*, J. Comput. Phys., 14, 227–25, 19743.
[4] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures in Mathematics, ETH Zürich, Birkhäuser Verlag, Berlin, 1992.
[5] R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 31, 2002.
[6] K. Schittkowski, *NLPQL: A Fortran subroutine solving constrained nonlinear programming problems*, Ann. Oper. Res., 5, 485–500, 1986.
[7] P.K. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal., 21, 995–1011, 1984.
[8] Q.H. Tran and B. Scheurer, *High-order monotonicity-preserving compact schemes for linear advection on 2-d irregular meshes*, J. Comput. Phys., 175, 454–486, 2002.
[9] B. van Leer, *Towards the ultimate conservative difference scheme V: A second-order sequel to Godunov's method*, J. Comput. Phys., 32, 101–136, 1979.
[10] W.B. VanderHeyden and B.A. Kashiwa, *Compatible fluxes for van Leer advection*, J. Comput. Phys., 146, 1–28, 1998.