On local conservation of numerical methods for conservation laws*

Cengke Shi†, Chi-Wang Shu‡

June 3, 2017

Abstract. In this paper we introduce a definition of the local conservation property for numerical methods solving time dependent conservation laws, which generalizes the classical local conservation definition. The motivation of our definition is the Lax-Wendroff theorem, and thus we prove it for locally conservative numerical schemes per our definition in one and two space dimensions. Several numerical methods, including continuous Galerkin methods and compact schemes, which do not fit the classical local conservation definition, are given as examples of locally conservative methods under our generalized definition.

1 Introduction

Local conservation is a desired property for numerical methods solving conservation laws. The most important theoretical reason is the well-known Lax-Wendroff theorem (Lax and Wendroff [11]). It states that the finite volume method (in 1D) taking the conservation form would converge to a weak solution of the underlying conservation law, if the numerical solutions actually converge. It is easy to give counterexamples, e.g. in LeVeque [12], that finite volume methods not in conservation form could converge to non-weak-solutions with wrong shock speed. Kröner and Rokyta [9] and Kröner et al. [10] have extended the Lax-Wendroff theorem to finite volume methods on unstructured meshes for 2D conservation laws. Abgrall et al. [2] have further generalized the Lax-Wendroff theorem to residual schemes.

The finite volume methods and discontinuous Galerkin (DG) methods, see, e.g., [4], are by design locally conservative numerical schemes. We could easily extend the Lax-Wendroff theorem to DG methods by setting the test function to value 1 in one cell and 0 in all other cells. Continuous Galerkin (CG) methods are considered only globally conservative until Hughes et al. [8] showed that it is actually locally conservative. But their definition of flux is not consistent with ours in this paper, because in their definition at least one of the two neighboring subdomains has to be global (that is, its size is comparable to that of the whole domain and hence does not go to zero with mesh refinements) to get the uniqueness of the flux across the boundary between these two subdomains. Perot [13] showed the local conservation of the CG methods with a different flux definition that is consistent with our definition (section 3.4.3). See also Abgrall [1] for the discussion of local conservation of CG methods.

The notion of local conservation is widely known as the rate of change of a quantity (in the classical definition this is the total mass in the cell) being equal to the sum of locally defined

*Research supported by ARO grant W911NF-15-1-0226 and NSF grant DMS-1418750.

†Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. E-mail: cengke_shi@brown.edu.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. E-mail: shu@dam.brown.edu.

fluxes (exchange with neighbors), which is the idea underlying the physical conservation laws. However, the absence of a rigorous definition makes it arguable as to whether a numerical method is locally conservative or not. In this paper, we give a formal definition of the local conservation property for numerical methods in one and two space dimensions. The motivation for our definition is the requirements in the classical Lax-Wendroff theorem, and naturally we prove it for locally conservative numerical methods per our definition.

The rest of the paper is organized as follows. In section 2, we introduce the definition of local conservation in one space dimension, and prove the corresponding Lax-Wendroff type theorem. Some examples of locally conservative numerical methods in 1D according to our definition are given at the end of the section. Section 3 is about local conservation in two space dimensions. As in section 2, we give the formal definition, prove a Lax-Wendroff type theorem, and present some examples of locally conservative numerical methods in 2D.

2 One dimensional conservation laws

We first consider one dimensional scalar conservation laws:

$$\begin{aligned} u_t + f(u)_x = 0 \text{ in } \mathbb{R}^+ \times \mathbb{R} \\ u(\cdot, 0) = u_0 \text{ in } \mathbb{R}, \end{aligned} \tag{1}$$

where the flux function f is at least Lipschitz continuous. We would like to formally define the local conservation property for numerical schemes designed for the above conservation laws, and prove a Lax-Wendroff type theorem for locally conservative schemes.

2.1 Numerical schemes

We consider in our presentation only schemes with Euler forward time stepping since the spatial discretization is our primary concern, and thus the schemes read:

$$\frac{u_h(\cdot, t_{n+1}) - u_h(\cdot, t_n)}{\Delta t_n} = L\left(u_h(\cdot, t_n)\right), \tag{2}$$

where the time domain is discretized as $0 = t_0 < t_1 < t_2 \ldots$ with $t_n \to \infty$ as $n \to \infty$. Time steps are defined as $\Delta t_n = t_{n+1} - t_n,\ \Delta t = \max_{n \geq 0} \Delta t_n$. The numerical solution u_h is a function over the time-space domain $\mathbb{R}^+ \times \mathbb{R}$, which is consistent with the initial condition in the sense of (10). Since the scheme only defines the numerical solution when $t = t_n,\ \forall n \geq 0$, we expand u_h to be a constant over each time interval, i.e. $u_h(x,t) \equiv u_h(x, t_n),\ \forall t \in [t_n, t_{n+1})$. Therefore, we can denote functions over the space domain $u_h^n = u_h(\cdot, t),\ \forall t \in [t_n, t_{n+1})$. For example, u_h^n is a piecewise constant function for finite difference and finite volume methods, and a piecewise polynomial for Galerkin methods.

To define the local conservation property, we need (i) a partition of the spatial domain into intervals $\mathbb{R} = \cup_{j \in \mathbb{Z}} I_j$, (ii) a locally conserved quantity on each cell, and (iii) a flux on each interval endpoint. The intervals $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, and they satisfy the regularity condition: $|I_j| > c_0\, h$, where $|I_j|$ is the length of the interval, mesh size $h = \max_{j \in \mathbb{Z}} |I_j|$, and $c_0 > 0$ is independent of the mesh size. For simplicity, we now denote $\sum_{n=0}^{+\infty}$ as $\sum_n$, and $\sum_{j \in \mathbb{Z}}$ as $\sum_j$.

2.2 Definition of local conservation

We define a numerical scheme to be locally conservative if its solution satisfies the following conservation form (cf. [11]):

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t_n} + \frac{1}{|I_j|}\left(g_{j+\frac{1}{2}}(u_h^n) - g_{j-\frac{1}{2}}(u_h^n)\right) = 0. \tag{3}$$

Here $\bar{u}_j^n$ (generalized locally conserved quantity) and $g_{j+\frac{1}{2}}$ (generalized flux) both depend on the numerical solution locally. More precisely, they depend on $u_h^n(B_j)$. $B_j = \{\, x \in \mathbb{R} : |x - w_j| < c\,h\}$, where w_j is the midpoint of the interval I_j, and $c \geq 1$ is independent of the mesh size. Note that $\bar{u}_j^n$ here is not necessarily the cell average ($\frac{1}{|I_j|}\int_{I_j} u_h^n$) even though we use the traditional notation for cell averages. The conserved quantity and flux also need to be consistent and bounded in the following sense:

- *consistency*: if $u_h^n(x) \equiv u$, a constant, $\forall x \in B_j$, we have:

$$\bar{u}_j^n = u, \tag{4}$$
$$g_{j+\frac{1}{2}}(u_h^n) = f(u), \tag{5}$$

- *boundedness*: they are both bounded with respect to the L^∞-norm of the solution:

$$\left|\bar{u}_j^n - \bar{v}_j^n\right| \leq C\,\|u_h^n - v_h^n\|_{L^\infty(B_j)}, \tag{6}$$
$$\left|g_{j+\frac{1}{2}}(u_h^n) - g_{j+\frac{1}{2}}(v_h^n)\right| \leq C\,\|u_h^n - v_h^n\|_{L^\infty(B_j)} \tag{7}$$

  for two functions (u_h and v_h) in the numerical solution space.

To get global conservation from our definition, we also require the summation of the local conservation quantities being exactly the global conservation quantity:

$$\sum_j |I_j|\,\bar{u}_j^n = \int_{\mathbb{R}} u_h(x, t_n)\,dx,\ \forall n \geq 0. \tag{8}$$

It is easy to see that (8) together with (3) implies the following global conservation:

$$\int_{\mathbb{R}} u_h(x,t)\,dx = \int_{\mathbb{R}} u_h(x,0)\,dx,\ \ \forall t > 0. \tag{9}$$

We now have gathered all parts, and give the formal definition of local conservation.

Definition 2.1. *A numerical scheme of the form (2) is locally conservative if there are conserved quantities and fluxes, both of which locally depend on the numerical solution and satisfy (3)–(8).*

2.3 Lax-Wendroff type theorem

In this section, we prove a Lax-Wendroff type theorem for locally conservative numerical schemes in the sense of Definition 2.1. We first assume that the initial condition is weakly enforced in the numerical solution as follows:

$$\int_{\mathbb{R}} \left(u_0 - u_h^0\right) \phi \to 0, \text{ as } h \to 0, \ \forall \phi \in C_0^\infty(\mathbb{R}). \tag{10}$$

To get the convergence result, we also need an assumption similar to the bounded total variation of the numerical solution:

$$h \sum_j \max_{x \in B_j} |u_h^n(x) - u_h^n(w_j)| \to 0, \text{ as } h \to 0, \ \ \forall n \geq 0 \tag{11}$$

The assumption (11) on the numerical solution may seem odd at first glance, but it is actually implied by boundedness of total variation as Proposition 2.2 shows. Recall the definition of total variation:

$$\mathrm{TV}(u_h^n) = \sup_{\mathcal{P}} \sum_{i=-\infty}^{+\infty} |u_h^n(x_{i+1}) - u_h^n(x_i)|, \tag{12}$$

where $\mathcal{P}$ denotes the set of all partitions of the real line $\mathbb{R}$.

Proposition 2.2. *If the numerical solution u_h^n has uniformly bounded total variation, i.e.*

$$TV(u_h^n) < C, \ \ \forall n \geq 0, \tag{13}$$

u_h^n satisfies the assumption (11).

Proof. We can deduce that there are a finite number of, say no more than p (independent of mesh size), intervals in each neighborhood B_j from the regularity of the mesh and the definition of B_j. We define the point that attains the maximal value of $|u_h^n(x) - u_h^n(w_j)|$ in B_j as m_j, which is in some interval I_k with $|j-k| < p$.

We can therefore define a series of intervals $(I_{j_1}, I_{j_2} \dots I_{j_p})$ such that $m_j \in I_{j_p}$, where $j_1 = j$, and $|j_s - j_{s+1}| \leq 1$ $(\forall 1 \leq s < p)$. The summation in (11) can be estimated as follows:

$$\begin{aligned}
&h \sum_j \max_{x \in B_j} |u_h^n(x) - u_h^n(w_j)| = h \sum_j |u_h^n(m_j) - u_h^n(w_j)| \\
\leq &h \sum_j \left(\left|u_h^n(m_j) - u_h^n(w_{j_p})\right| + \sum_{s=1}^{p-1} \left|u_h^n(w_{j_s}) - u_h^n(w_{j_{s+1}})\right| \right) \\
\leq &h \sum_j \sup_{x \in I_{j_p}} \left|u_h^n(x) - u_h^n(w_{j_p})\right| + 2p\, h \sum_j |u_h^n(w_{j+1}) - u_h^n(w_j)| \\
\leq &4p\, h\, TV(u_h^n) \ \to 0, \text{ as } h \to 0.
\end{aligned}$$

In the above calculation, we used two obvious inequalities:

$$\sum_j \sup_{x \in I_j} |u_h^n(x) - u_h^n(w_j)| \leq TV(u_h^n) \tag{14}$$

$$\sum_j |u_h^n(w_{j+1}) - u_h^n(w_j)| \leq TV(u_h^n) \tag{15}$$

□

We now state and prove a Lax-Wendroff type theorem for our definition of local conservation. The proof is quite similar to the simplified proof of the Lax-Wendroff theorem in LeVeque [12].

Theorem 2.3. *Assume the solution to a locally conservative numerical scheme, with initial condition imposed as (10), satisfies the assumption (11). If u_h converges boundedly almost everywhere to some function u as $\Delta t, h \to 0$, then u is a weak solution to the conservation law (1), i.e.*

$$\int_{\mathbb{R}} u_0\,\phi \;+\int_{\mathbb{R}^+\times\mathbb{R}} u\,\phi_t \;+\int_{\mathbb{R}^+\times\mathbb{R}} f(u)\cdot\phi_x = 0,\ \forall\phi\in C_0^\infty\left(\mathbb{R}^+\times\mathbb{R}\right). \tag{16}$$

Proof. Multiplying $\Delta t_n\,|I_j|\,\phi(w_j,t_n)$ on both sides of (3) and summing it over all n, j, we have the following equality:

$$\sum_n\sum_j |I_j|\left(\bar{u}_j^{n+1}-\bar{u}_j^n\right)\phi_j^n+\sum_n\sum_j \Delta t_n\left(g_{j+\frac{1}{2}}(u_h^n)-g_{j-\frac{1}{2}}(u_h^n)\right)\phi_j^n=0 \tag{17}$$

where we denote $\phi(w_j,t_n)$ as ϕ_j^n, and $\phi(x,t_n)$ as $\phi^n(x)$. We can apply summation by parts on the first term and get:

$$\sum_n\sum_j |I_j|\left(\bar{u}_j^{n+1}-\bar{u}_j^n\right)\phi_j^n=-\sum_j |I_j|\,\bar{u}_j^0\,\phi_j^0\;-\sum_n\sum_j |I_j|\left(\phi_j^{n+1}-\phi_j^n\right)\bar{u}_j^{n+1}=-\mathrm{I}-\mathrm{II} \tag{18}$$

By using assumptions (4), (6), (10), (11), the fact that ϕ is compactly supported, and the bounded convergence theorem, we get the convergence of term I to the first integration in (16) as follows:

$$\begin{aligned}\sum_j |I_j|\,\bar{u}_j^0\,\phi_j^0 &= \int_{\mathbb{R}} u_h^0(x)\,\phi^0(x)\,dx+\sum_j\int_{I_j}\left(\bar{u}_j^0-u_h^0(x)\right)\phi^0(x)\,dx+O(h)\\ &=\int_{\mathbb{R}} u_0(x)\phi^0(x)\,dx+o(1)\end{aligned}$$

where $o(1)$ denotes a quantity which goes to zero when the mesh sizes go to zero. We similarly deduce the convergence of term II to the second integration in (16):

$$\begin{aligned}&\sum_n\sum_j |I_j|\left(\phi_j^{n+1}-\phi_j^n\right)\bar{u}_j^{n+1}\\ =&\sum_{n=1}^{+\infty}\sum_j\int_{t_n}^{t_{n+1}}\int_{I_j}\phi_t(x,t)\,u_h(x,t)\,dx\,dt+\sum_{n=1}^{+\infty}\sum_j\int_{t_n}^{t_{n+1}}\int_{I_j}\phi_t(x,t)\,\left(\bar{u}_j^n-u_h^n(x)\right)\,dx\,dt+O(h)\\ =&\int_{\mathbb{R}^+\times\mathbb{R}}\phi_t(x,t)\,u(x,t)\,dx\,dt+o(1)\end{aligned}$$

Notice that the second term in the middle line above is $o(1)$ because we can uniformly bound

$\phi_t(x,t)$ which has a compact support, and then get

$$\begin{aligned}
&\left|\sum_{n=1}^{+\infty}\sum_j \int_{t_n}^{t_{n+1}} \int_{I_j} \phi_t(x,t)\ \left(\bar{u}_j^n - u_h^n(x)\right)\, dx\, dt\right| \\
\leq & C_1 \sum_{n\in N_\phi}\sum_{j\in J_\phi} \int_{t_n}^{t_{n+1}} \int_{I_j} \left|\bar{u}_j^n - u_h^n(w_j)\right| + \left|u_h^n(w_j) - u_h^n(x)\right|\, dx\, dt \\
\leq & C_2 \max_{n\in N_\phi} \sum_{j\in J_\phi} |I_j| \max_{x\in B_j} |u_h^n(x) - u_h^n(w_j)| \\
= & o(1)
\end{aligned}$$

where N_ϕ is the index set corresponding to the support of $\phi_t(x,t)$, the first term on the second line is bounded by the term on the third line due to the consistency and boundedness of $\bar{u}_j^n$ (see (4) and (6)), and the last equality comes directly from the assumption (11). Lastly, we prove the convergence of the second summation in (17) to the third integration in (16):

$$\begin{aligned}
&\sum_n \sum_j \Delta t_n \left(g_{j+\frac{1}{2}}(u_h^n)\, \phi_j^n - g_{j-\frac{1}{2}}(u_h^n)\, \phi_j^n\right) \\
= & \sum_n \sum_j \Delta t_n \left(f(u_h^n(w_j))\ \left(\phi_j^n - \phi_{j+\frac{1}{2}}^n\right) - f(u_h^n(w_j))\ \left(\phi_j^n - \phi_{j-\frac{1}{2}}^n\right)\right) + \\
& \sum_n \sum_j \Delta t_n \left(\left(g_{j+\frac{1}{2}}(u_h^n) - f(u_h^n(w_j))\right)\ \left(\phi_j^n - \phi_{j+\frac{1}{2}}^n\right) - \left(g_{j-\frac{1}{2}}(u_h^n) - f(u_h^n(w_j))\right)\ \left(\phi_j^n - \phi_{j-\frac{1}{2}}^n\right)\right) \\
= & \mathrm{III} + \mathrm{IV},
\end{aligned}$$

where $\phi^n(x_{j+\frac{1}{2}})$ is denoted as $\phi_{j+\frac{1}{2}}^n$.

By using the definition (4)-(7) and the assumption (11), we have that term IV vanishes as the mesh size goes to zero. The convergence of term III to the last integration of (16) can be deduced by using the local conservation definition and the assumption (11) as follows:

$$\begin{aligned}
\mathrm{III} &= -\sum_n \sum_j \Delta t_n f(u_h^n(w_j))\ \left(\phi_{j+\frac{1}{2}}^n - \phi_{j-\frac{1}{2}}^n\right) \\
&= -\sum_n \sum_j \int_{t_n}^{t_{n+1}} \int_{I_j} f(u_h^n(x))\, \partial_x \phi(x,t_n)\, dx\, dt + o(1) \\
&= -\int_{\mathbb{R}^+\times\mathbb{R}} f(u)\phi_x + o(1)
\end{aligned}$$

The proof is concluded by combining the convergence of terms I–IV. □

2.4 Examples of locally conservative numerical methods

We first point out that the DG methods (see, e.g., [14]), as well as standard finite difference and finite volume methods taking conservation form obviously have the local conservation property. In the next few subsections, we show some other examples of locally conservative numerical methods.

2.4.1 Compact (finite difference) schemes

Compact schemes are non-standard finite difference methods, for which the numerical solution usually is defined on the grid points of a uniform mesh. We can extend the definition of the numerical solution to be a piecewise constant function defined over the whole real line. A compact scheme for the conservation law (1) can be written as (Cockburn and Shu [5]):

$$\frac{u_h^{n+1}(w_j) - u_j^n(w_j)}{\Delta t_n} + \frac{1}{\Delta x}\left(A^{-1}\, B\, f(u_h^n)\right)_j = 0 \tag{19}$$

where A and B are local linear operators. We can define the conserved quantity $\bar{u}_j^n = (A\, u_h^n)_j$ and the flux $g_{j+\frac{1}{2}}(u_h^n)$ satisfying

$$(B\, f(u_h^n))_j = g_{j+\frac{1}{2}}(u_h^n) - g_{j-\frac{1}{2}}(u_h^n). \tag{20}$$

It is not hard to check that the compact scheme is locally conservative with the conserved quantity and flux we just defined.

2.4.2 Non-standard finite volume methods

Zhang et al. [15] have proposed a non-standard finite volume scheme, in order to design high order maximum-principle-satisfying schemes for solving convection-diffusion equations. The degrees of freedom for this method are the so called "double cell averages" as opposed to the usual cell averages for standard finite volume methods. Suppose the real line is uniformly decomposed into intervals $\mathbb{R} = \cup_{j\in\mathbb{Z}} I_j$. The cells $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ whose midpoints are denoted as x_j. The scheme applied on conservation law (1) can be written as (see (3.19) in [15]):

$$\frac{\bar{\bar{u}}_j^{n+1} - \bar{\bar{u}}_j^n}{\Delta t} + \frac{1}{\Delta x}\sum_{\alpha=1}^{3} w_\alpha \left(\hat{f}^\alpha_{j+\frac{1}{2}} - \hat{f}^\alpha_{j-\frac{1}{2}}\right) = 0 \tag{21}$$

where w_α are quadrature weights and $\hat{f}^\alpha_{j+\frac{1}{2}}$ are numerical fluxes (clearly satisfying (5) and (7)) on the quadrature points. The double cell averages are defined as:

$$\bar{\bar{u}}_j^n = \frac{1}{\Delta x^2}\int_{I_j}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} u_h^n(\xi)\, d\xi\, dx, \tag{22}$$

where u_h^n is the underlying numerical solution. By using the following equality

$$\bar{\bar{u}}_j^n = \frac{1}{\Delta x^2}\int_{x_{j-1}}^{x_j} u_h^n(x)(x - x_{j-1})\, dx + \frac{1}{\Delta x^2}\int_{x_j}^{x_{j+1}} u_h^n(x)(x_{j+1} - x)\, dx, \tag{23}$$

we can easily show that the double cell averages satisfy conditions (4), (6), and (8). Therefore the scheme (21) is locally conservative.

2.4.3 Continuous Galerkin method

Let $U_h^k = \{v \in H^1(\mathbb{R}) : v \in P_k(e_j), \forall j \in \mathbb{Z}\}$, where the set of intervals$\{e_j = [y_j, y_{j+1}] : j \in \mathbb{Z}\}$ is a partition of the real line and $P_k(e_j)$ denotes the set of polynomials over the cell e_j. The CG method reads: find $u_h(\cdot, t_n) \in U_h^k$, such that for all test function $v_h \in U_h^k$

$$\int_{\mathbb{R}} \frac{u_h^{n+1} - u_h^n}{\Delta t_n} v_h - \int_{\mathbb{R}} f(u_h^n) v_h' = 0. \tag{24}$$

To define local conservation, we need a different partition of the domain $\mathbb{R} = \cup_{j\in\mathbb{Z}} I_j = \cup_{j\in\mathbb{Z}} [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, where $x_{j+\frac{1}{2}} = \frac{1}{2}(y_j + y_{j+1})$. We take the test function v_h as the shape function v_j, which is a piecewise linear function satisfying $v_j(y_j) = \frac{1}{|I_j|},\ v_j(y_l) = 0,\ \forall l \neq j$. The generalized conserved quantity is defined as: $\bar{u}_j^n = \int_{\mathbb{R}} u_h^n v_j$, and the generalized flux is defined as $g_{j+\frac{1}{2}}(u_h^n) = \frac{1}{|e_j|}\int_{e_j} f(u_h^n)$. It is easy to check that the above definitions satisfy the conditions (3)–(8), and thus the CG method is locally conservative. The result in this section can be generalized to CG methods with local basis satisfying the partition of unity property by a similar argument to that in section 3.4.3. The B-spline finite element methods (see, e.g., [6]) is one example.

3 Two-dimensional conservation laws

We study numerical methods for two-dimensional scalar conservation laws of the form:

$$\begin{aligned} u_t + \nabla \cdot \boldsymbol{f}(u) = 0 \text{ in } \mathbb{R}^+ \times \mathbb{R}^2 \\ u(\cdot, 0) = u_0 \text{ in } \mathbb{R}^2, \end{aligned} \tag{25}$$

where the flux function $\boldsymbol{f} = (f_1,\ f_2)$, and both of its components are at least Lipschitz continuous. We similarly define the local conservation property and prove a Lax-Wendroff type theorem here.

3.1 Numerical schemes

We again consider in our presentation only schemes with Euler forward time stepping since the spatial discretization is our primary concern, and thus the schemes read:

$$\frac{u_h(\cdot, t_{n+1}) - u_h(\cdot, t_n)}{\Delta t_n} = L\left(u_h(\cdot, t_n)\right), \tag{26}$$

where the time domain is discretized as in 1D case. The numerical solution is a constant over each time interval, i.e. $u_h(\cdot, t) = u_h(\cdot, t_n),\ \forall t \in [t_n, t_{n+1})$, and thus we can denote $u_h^n = u_h(\cdot, t_n)$.

To define the local conservation property, we need (i) a partition of the spatial domain into cells $\mathbb{R}^2 = \cup_{j\in J} T_j$, (ii) a locally conserved quantity on each cell, and (iii) a flux on each cell interface. The cells are polygons with at most p edges, and satisfy the regularity condition: $|T_j| > c_0\, h^2$, where $|T_j|$ is the area of the cell, and the mesh size $h = \max_{j\in J} \mathrm{diam}\,(T_j)$. For simplicity, we again denote $\sum_{n=0}^{\infty}$ as $\sum_n$, and $\sum_{j\in J}$ as $\sum_j$.

3.2 Definition of local conservation

We say a numerical scheme is locally conservative if its solution satisfies the following conservation form equality (cf. [9, 10, 2]):

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t_n} + \frac{1}{|T_j|} \sum_{l \in N_j} \sum_{s=1}^{c_{jl}} g_{jl}^{(s)}(u_h^n) = 0. \tag{27}$$

Here, N_j denotes the set of cell indices of all cells adjacent to T_j, i.e. $N_j = \{l \in J : T_j \cap T_l \neq \varnothing\}$. The cell interface between T_j and T_l has c_{jl} line segments. If $\{T_j : j \in J\}$ is a triangulation, c_{jl} is always 1. If $\{T_j : j \in J\}$ is the dual mesh (see, e.g., Bush and Ginting [3]) of a triangulation, c_{jl} equals to two for all the boundaries. $\bar{u}_j^n$ (generalized conserved quantity) and g_{jl} (generalized flux) both depend on the numerical solution locally. More precisely, they depend on $u_h^n(B_j)$. $B_j = \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x} - w_j| < c\,h\}$, where the center of a cell $w_j = \frac{1}{|T_j|}\int_{T_j} \boldsymbol{x}\, d\boldsymbol{x}$, and $c\,(>1)$ is independent of the mesh size. The conserved quantity and flux also need to be consistent and bounded in the following sense:

- *consistency*: if $u_h^n(\boldsymbol{x}) \equiv u$, a constant, $\forall \boldsymbol{x} \in B_j$, we have:

$$\bar{u}_j^n = u, \tag{28}$$

$$g_{jl}^{(s)}(u_h^n) = \boldsymbol{\nu}_{jl}^{(s)} \cdot \boldsymbol{f}(u), \tag{29}$$

  where $\boldsymbol{\nu}_{jl}^{(s)} = \left|S_{jl}^{(s)}\right| \boldsymbol{n}_{jl}^{(s)}$, and $S_{jl}^{(s)}$ is the s-th line segment on the boundary between elements T_j and T_l. $\boldsymbol{n}_{jl}^{(s)}$ is the normal vector on $S_{jl}^{(s)}$ pointing toward T_l.

- *boundedness*: they are both bounded in terms of the L^∞-norm of the numerical solution:

$$\left|\bar{u}_j^n - \bar{v}_j^n\right| \leq C\, \|u_h^n - v_h^n\|_{L^\infty(B_j)}, \tag{30}$$

$$|g_{jl}(u_h^n) - g_{jl}(v_h^n)| \leq C\, h\, \|u_h^n - v_h^n\|_{L^\infty(B_j)}. \tag{31}$$

As a flux, g_{jl} must be unique on cell interfaces, i.e.

$$g_{jl}(u_h^n) + g_{lj}(u_h^n) = 0, \ \ \forall j \in J,\, l \in N_j. \tag{32}$$

To get global conservation from our definition, we also require the summation of the local conservation quantities being exactly the global conservation quantity:

$$\sum_j |T_j|\, \bar{u}_j^n = \int_{\mathbb{R}^2} u_h^n, \ \forall n \geq 0. \tag{33}$$

It is easy to see that (33) together with (27) and (32) implies the following global conservation:

$$\int_{\mathbb{R}^2} u_h(\cdot, t) = \int_{\mathbb{R}^2} u_h(\cdot, 0), \ \ \forall t > 0. \tag{34}$$

We now have all the parts and give the formal definition of local conservation.

Definition 3.1. *A numerical scheme of the form (26) is locally conservative if there exist conserved quantities and fluxes, both of which locally depend on the numerical solution and satisfy (27)–(33).*

3.3 Lax-Wendroff type theorem

In this section, we prove a Lax-Wendroff type theorem for numerical schemes satisfying Definition 3.1. The technique of the proof is similar to that in Kröner and Rokyta [9] and Kröner et al. [10]. We first need to weakly enforce the initial condition as follows:

$$\int_{\mathbb{R}^2} \left(u_0 - u_h^0\right) \phi \to 0, \text{ as } h \to 0, \ \forall \phi \in C_0^\infty\left(\mathbb{R}^2\right). \tag{35}$$

To get the convergence result, we also need an assumption similar to the bounded total variation of the numerical solution:

$$h^2 \sum_j \max_{x \in B_j} |u_h^n(x) - u_h^n(w_j)| \to 0, \text{ as } h \to 0, \ \ \forall n \geq 0, \tag{36}$$

which is a 2D generalization of (11). This assumption clearly holds for piecewise smooth functions with smooth regions divided by curves.

Theorem 3.2. *Assume the solution to a locally conservative numerical scheme with initial condition imposed as (35) satisfies the assumption (36). If u_h converges boundedly almost everywhere to some function u as $\Delta t, h \to 0$, then u is a weak solution to the conservation law (25), i.e.*

$$\int_{\mathbb{R}^2} u_0\, \phi \ + \int_{\mathbb{R}^+ \times \mathbb{R}^2} u\, \phi_t \ + \int_{\mathbb{R}^+ \times \mathbb{R}^2} \boldsymbol{f}(u) \cdot \nabla \phi = 0, \ \forall \phi \in C_0^\infty\left(\mathbb{R}^+ \times \mathbb{R}^2\right). \tag{37}$$

Proof. Multiplying $\Delta t_n \left|T_j\right| \phi(w_j, t_n)$ on both sides of (27) and summing it over all n, j, we have the following equality:

$$\sum_n \sum_j \left|T_j\right| \left(\bar{u}_j^{n+1} - \bar{u}_j^n\right) \phi_j^n \ + \Delta t_n \sum_n \sum_j \sum_{l \in N_j} \sum_{s=1}^{c_{jl}} g_{jl}^{(s)}(u_h^n) \phi_j^n = 0, \tag{38}$$

where we denote $\phi(w_j, t_n)$ as ϕ_j^n, and $\phi(x, t_n)$ as $\phi^n(x)$. The convergence of the first summation in (38) to the first two integrations in (37) can be proved exactly the same as in Theorem 2.3.

We now prove the convergence of the second summation in (38) to the third integration in (37) by using definition (28)–(32) together with the assumption (36):

$$\begin{aligned}
\sum_n \sum_j \sum_{l \in N_j} \sum_{s=1}^{c_{jl}} \Delta t_n\, g_{jl}^{(s)}(u_h^n)\, \phi_j^n &= \sum_n \sum_j \sum_{l \in N_j} \sum_{s=1}^{c_{jl}} \Delta t_n\, g_{jl}^{(s)}(u_h^n) \ \left(\phi_j^n - \phi^n(z_{jl}^{(s)})\right) \\
&= -\sum_n \sum_j \sum_{l \in N_j} \sum_{s=1}^{c_{jl}} \Delta t_n\, \boldsymbol{\nu}_{jl}^{(s)} \cdot \boldsymbol{f}(u_h^n(w_j))\, \phi^n(z_{jl}^{(s)}) \\
&\quad + \sum_n \sum_j \sum_{l \in N_j} \sum_{s=1}^{c_{jl}} \Delta t_n \left(g_{jl}^{(s)}(u_h^n) - g_{jl}^{(s)}(u_h^n(w_j))\right) \left(\phi_j^n - \phi^n(z_{jl}^{(s)})\right) \\
&= -\sum_n \sum_j \Delta t_n \left|T_j\right| \boldsymbol{f}(u_h^n(w_j)) \cdot \nabla \phi^n(w_j) \ + o(1) \\
&= -\int_{\mathbb{R}^+ \times \mathbb{R}^2} \boldsymbol{f}(u) \cdot \nabla \phi \ + o(1)
\end{aligned}$$

where $z_{jl}^{(s)}$ is the midpoint of $S_{jl}^{(s)}$. In the above calculation, we used the fact that

$$\sum_{l\in N_j}\sum_{s=1}^{c_{jl}} \boldsymbol{\nu}_{jl}^{(s)} = 0, \tag{39}$$

$$\sum_{j}\sum_{l\in N_j}\sum_{s=1}^{c_{jl}} g_{jl}(u_h^n)\,\phi^n(z_{jl}^{(s)}) = 0, \tag{40}$$

and Lemma 3.3 from [10]. □

Lemma 3.3. *If $\phi \in C_0^\infty\left(\mathbb{R}^2\right)$, we have the approximate integration by parts formula:*

$$|T_j|\,\nabla\phi(w_j) = \sum_{l\in N_j}\sum_{s=1}^{c_{jl}} \phi(z_{jl}^{(s)})\boldsymbol{\nu}_{jl}^{(s)} + O(h^3). \tag{41}$$

3.4 Examples of locally conservative numerical methods

As in the one dimensional case, the DG methods (see, e.g., [4]), standard finite volume and finite element methods taking conservation form are obviously locally conservative per Definition 3.1. We show in this section that the two dimensional counterpart of numerical methods in section 2.4 are also locally conservative.

3.4.1 Compact (finite difference) schemes

The compact schemes in the two space dimensions has their solutions defined on a 2D uniform structured grid with size $\Delta x,\ \Delta y$. We can extend the solution by making it a constant on a rectangular centered at each grid point. The size of each rectangle is $\Delta x \times \Delta y$. We denote the grid points as $\{w_{ij} \in \mathbb{R}^2 : i, j \in \mathbb{Z}\}$, and the schemes read:

$$\frac{u_h^{n+1}(w_{ij}) - u_h^n(w_{ij})}{\Delta t} + \frac{1}{\Delta x}(A_x^{-1}\,B_x\,f_1(u))_{ij} + \frac{1}{\Delta y}(A_y^{-1}B_y f_2(u))_{ij} = 0 \tag{42}$$

where the operator A_x is the operator for the 1D compact scheme applied in the x direction, and the same for other three operators. We can rewrite the scheme (42) into a conservation form as in Cockburn and Shu [5]:

$$\frac{\bar{u}_{ij}^{n+1} - \bar{u}_{ij}^n}{\Delta t_n} + \frac{1}{\Delta x}\left(\hat{f}^n_{i+\frac{1}{2},j} - \hat{f}^n_{i-\frac{1}{2},j}\right) + \frac{1}{\Delta y}\left(\hat{g}^n_{i,j+\frac{1}{2}} - \hat{g}^n_{i,j-\frac{1}{2}}\right) = 0, \tag{43}$$

where $\bar{u}_{ij}^n = A_y A_x u$, and the flux are similarly defined as in the 1D case. For example,

$$\hat{f}^n_{i+\frac{1}{2},j} = \frac{1}{2}A_y\left(f(u_h^n(w_{i+1,j})) + f(u_h^n(w_{ij}))\right),$$
$$\hat{g}^n_{i,j+\frac{1}{2}} = \frac{1}{2}A_x\left(g(u_h^n(w_{i,j+1})) + g(u_h^n(w_{ij}))\right),$$

if $(Bu)_i = \frac{1}{2}(u_{i+1} - u_{i-1})$. We could easily check that the compact scheme in 2D is locally conservative with the form (43).

3.4.2 Non-standard finite volume methods

A straightforward generalization of the non-standard finite volume methods (21) on 2D uniform rectangular meshes are given in [15] (see (4.3)):

$$\frac{\bar{\bar{u}}_{ij}^{n+1} - \bar{\bar{u}}_{ij}^{n}}{\Delta t} + \frac{1}{\Delta x}\left(\hat{f}_{i+\frac{1}{2},j} - \hat{f}_{i-\frac{1}{2},j}\right) + \frac{1}{\Delta y}\left(\hat{g}_{i,j+\frac{1}{2}} - \hat{g}_{i,j-\frac{1}{2}}\right) = 0 \tag{44}$$

where $\bar{\bar{u}}_{ij}^{n}$ is the 2D version of double cell average:

$$\bar{\bar{u}}_{ij}^{n} = \frac{1}{\Delta x^2}\frac{1}{\Delta y^2} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{y-\frac{\Delta y}{2}}^{y+\frac{\Delta y}{2}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} u(\xi,\eta)\, d\xi\, dx\, d\eta\, dy. \tag{45}$$

The 2D uniform rectangular mesh is denoted as: $\mathbb{R}^2 = \cup_{i,j\in\mathbb{Z}}[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$. We can similarly get the local conservation property for this scheme by using the following equality on double cell averages:

$$\begin{aligned}
\bar{\bar{u}}_{ij}^{n} =& \frac{1}{\Delta x^2}\frac{1}{\Delta y^2}\int_{x_{i-1}}^{x_i}\int_{y_{j-1}}^{y_j} u_h^n(x,y)(x-x_{i-1})(y-y_{j-1})\, dx\, dy + \\
& \frac{1}{\Delta x^2}\frac{1}{\Delta y^2}\int_{x_{i-1}}^{x_i}\int_{y_j}^{y_{j+1}} u_h^n(x,y)(x-x_{i-1})(y_{j+1}-y)\, dx\, dy + \\
& \frac{1}{\Delta x^2}\frac{1}{\Delta y^2}\int_{x_i}^{x_{i+1}}\int_{y_{j-1}}^{y_j} u_h^n(x,y)(x_{i+1}-x)(y-y_{j-1})\, dx\, dy + \\
& \frac{1}{\Delta x^2}\frac{1}{\Delta y^2}\int_{x_i}^{x_{i+1}}\int_{y_j}^{y_{j+1}} u_h^n(x,y)(x_{i+1}-x)(y_{j+1}-y)\, dx\, dy,
\end{aligned}$$

where x_i and y_j are midpoints of the the intervals in each dimension.

3.4.3 Continuous Galerkin method

Let $U_h^k = \{v \in H^1(\mathbb{R}^2) : v \in P_k(e_i), \forall i \in \mathbb{Z}\}$, where $\{e_i,\, i \in \mathbb{Z}\}$ is a triangulation of the whole space. The CG method reads: find $u_h(\cdot, t_n) \in U_h^k$, such that for all test function $v_h \in U_h^k$

$$\int_{\mathbb{R}^2} \frac{u_h^{n+1} - u_h^n}{\Delta t_n} v_h - \int_{\mathbb{R}^2} \boldsymbol{f}(u_h^n) \cdot \nabla v_h = 0. \tag{46}$$

The flux defined in Hughes et al. [8] does not fall in our definition since it needs at least one of the two regions to be global to get the uniqueness of flux between the two regions. The flux and conserved quantity in Perot [13] does satisfy the requirement of Definition 3.1, and we briefly reproduce his formulation here.

For the purpose of defining local conservation, we use a different partition $\mathbb{R} = \cup_{j\in J} T_j$, where $\{T_j : j \in J\}$ is the dual mesh of the original triangulation $\{e_i,\, i \in \mathbb{Z}\}$. The dual mesh, see, e.g., Bush and Ginting [3], is constructed by connecting the midpoint of edges and centroids of cells in the original triangulation. Since every cell in the dual mesh corresponds to a vertex in the original triangulation, we denote $\{n_j : j \in J\}$ as the set of vertices. The shape functions are defined as

$\{v_j \in U_h^1 : v_j(n_j) = 1, v_j(n_l) = 0, \forall l \neq j\}$. Because the partition of unity property, $\sum_{j\in J} v_j \equiv 1$ on $\mathbb{R}^2$, we have the following equality:

$$\nabla v_j = \sum_{l\in N_k} v_l \nabla v_j - v_j \nabla v_l \,, \text{ in } \mathbb{R}^2, \tag{47}$$

which leads to a rewriting of the scheme (46) if we take the test function as the shape function:

$$\frac{1}{\Delta t_n}\left(\frac{1}{|T_j|}\int_{\mathbb{R}^2} u_h^{n+1}\, v_j - \frac{1}{|T_j|}\int_{\mathbb{R}^2} u_h^n\, v_j\right) - \sum_{l\in N_j} \frac{1}{|T_j|}\int_{\mathbb{R}^2} v_l\, \boldsymbol{f}(u_h^n)\cdot\nabla v_j - v_j\, \boldsymbol{f}(u_h^n)\cdot\nabla v_l = 0. \tag{48}$$

We can check, with routine algebraic calculation, that the 2D CG method (46) is locally conservative in the sense of Definition 3.1 if we define the conserved quantity and flux as follows:

$$\bar{u}_j^n = \frac{1}{|T_j|}\int_{\mathbb{R}^2} u_h^n\, v_j, \tag{49}$$

$$g_{jl}^{(s)}(u_h^n) = \left(\frac{\left|S_{jl}^{(s)}\right|}{\left|S_{jl}^{(1)}\right| + \left|S_{jl}^{(2)}\right|}\right) \sum_{l\in N_j} \frac{1}{|T_j|}\int_{\mathbb{R}^2} v_j\, \boldsymbol{f}(u_h^n)\cdot\nabla v_l - v_l\, \boldsymbol{f}(u_h^n)\cdot\nabla v_j. \tag{50}$$

Note that c_{jl} is always 2 here because $\{T_j,\, j \in J\}$ is the dual mesh of a triangulation. According to the above argument, all CG methods with local basis satisfying partition of unity are locally conservative. The 2D B-spline finite element methods (see, e.g., [7]) with tensor product B-spline basis is one example.

4 Conclusions

In this paper, we have given a rigorous definition of the local conservation property for numerical methods solving conservation laws in both one and two space dimensions. Lax-Wendroff type theorems are stated and proved for locally conservative numerical schemes thus defined. Our definition serves as a general framework for the discussion of local conservation property. We have shown that the CG methods among others are considered locally conservative in our framework. Hence, their solutions would converge to a weak solution of the underlying conservation law if they do converge.

References

[1] R. Abgrall, On a class of high order schemes for hyperbolic problems. In S. Y. Jang, Y. R. Kim, D.-W. Lee, and I. Yie, editors, Proceedings of the International Congress of Mathematicians, KYUNG MOON SA, Seoul, 2014, 699–725.

[2] R. Abgrall, K. Mer-Nkonga, B. Nkonga, A Lax-Wendroff type theorem for residual schemes. In M. M. Hafez and J.-J. Chattot, editors, Innovative Methods for Numerical Solutions of Partial Differential Equations, World Scientific, Singapore, 2002, 243–266.

[3] L. Bush, V. Ginting, On the application of the continuous Galerkin finite element method for conservation problems. SIAM J. Sci. Comput. 35:6, A2953–A2975 (2013).

[4] B. Cockburn, C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case. Math. Comp. 54, 545–581 (1990).

[5] B. Cockburn, C.-W. Shu, Nonlinearly stable compact schemes for shock calculations. SIAM J. Numer. Anal. 31:3, 607–627 (1994).

[6] V. Ginting, R. Johnson, Locally conservative B-spline finite element methods for two-point boundary value problems. Procedia Comput. Sci. 80, 1279–1290 (2016).

[7] K. Höllig, Finite Element Methods with B-Splines. Society for Industrial and Applied Mathematics, 2003.

[8] T. J.R. Hughes, G. Engel, L. Mazzei, M. G. Larson, The continuous Galerkin method is locally conservative. J. Comput. Phys. 163, 467–488 (2000).

[9] D. Kröner, M. Rokyta, Convergence of upwind finite volume schemes for scalar conservation laws in two dimensions. SIAM J. Numer. Anal. 31:2, 324–343 (1994).

[10] D. Kröner, M. Rokyta, M. Wierse, A Lax-Wendroff type theorem for upwind finite volume schemes in 2-D. East-West J. Numer. Math. 4, 279–292 (1996).

[11] P. Lax, B. Wendroff, Systems of conservation laws. Comm. Pure Appl. Math. 13, 217–237 (1960).

[12] R. J. LeVeque, Numerical Methods for Conservation Laws. Springer Basel AG, 1992.

[13] J. B. Perot, Discrete conservation properties of unstructured mesh schemes. Annu. Rev. Fluid Mech. 43, 299–318 (2011).

[14] C.-W. Shu, Discontinuous Galerkin methods: general approach and stability. In S. Bertoluzza, S. Falletta, G. Russo, and C.-W. Shu, Numerical Solutions of Partial Differential Equations, Birkhäuser, Basel, 2009, 149–201.

[15] X. Zhang, Y. Liu, C.-W. Shu, Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations. SIAM J. Sci. Comput. 34(2), A627–A658 (2012).