A Posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws


It is advisable to refer to the publisher’s version if you intend to cite from the work.
Published version at: http://dx.doi.org/10.1137/140970999
To link to this article DOI: http://dx.doi.org/10.1137/140970999

Publisher: Society for Industrial and Applied Mathematics

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur
CentAUR

Central Archive at the University of Reading

Reading's research outputs online
A POSTERIORI ANALYSIS OF DISCONTINUOUS GALERKIN SCHEMES FOR SYSTEMS OF HYPERBOLIC CONSERVATION LAWS

JAN GIESELMANN†, CHARALAMBOS MAKRIDAKIS‡, AND TRISTAN PRYER§

Abstract. In this work we construct reliable a posteriori estimates for some semi- (spatially) discrete discontinuous Galerkin schemes applied to nonlinear systems of hyperbolic conservation laws. We make use of appropriate reconstructions of the discrete solution together with the relative entropy stability framework, which leads to error control in the case of smooth solutions. The methodology we use is quite general and allows for a posteriori control of discontinuous Galerkin schemes with standard flux choices which appear in the approximation of conservation laws. In addition to the analysis, we conduct some numerical benchmarking to test the robustness of the resultant estimator.

Key words. discontinuous Galerkin, a posteriori estimates, systems of hyperbolic conservation laws, relative entropy

AMS subject classifications. 65M60, 65M15, 35L65

DOI. 10.1137/140970999

1. Introduction. Hyperbolic conservation laws play an important role in many physical and engineering applications. One example is the description of nonviscous compressible flows by the Euler equations. Hyperbolic conservation laws in general only have smooth solutions up to some finite time even for smooth initial data. This makes their analysis and the construction of reliable numerical schemes challenging. The development of discontinuities poses significant challenges to their numerical simulation. Several successful schemes have been developed so far and are mainly based on finite differences, finite volume, and discontinuous Galerkin (dG) finite element schemes. For an overview on these schemes we refer to [GR96, Krö97, LeV02, Coc03, HW08] and their references. In this work we are interested in a posteriori error control of hyperbolic systems while solutions are still smooth. Our main tools are appropriate reconstructions of the dG schemes considered and relative entropy estimates.

The first systematic a posteriori analysis for numerical approximations of scalar conservation laws accompanied with corresponding adaptive algorithms can be traced back to [KO00, GM00]; see also [Coc03, DMO07] and their references. These estimates were derived by employing Kruzkov’s estimates. A posteriori results for systems were derived in [La08, La04] for front tracking and Glimm’s schemes; see also [KLY10].
For recent a posteriori analysis for well-balanced schemes for a damped semilinear wave equation we refer to [AG13].

We aim at providing a rigorous a posteriori error estimate for semidiscrete dG schemes applied to systems of hyperbolic conservation laws which are of optimal order. The extension of these results to fully discrete schemes is an important point which requires new ideas and is the subject of ongoing work. Our analysis is based on an extension of the reconstruction technique, developed mainly for discretizations of parabolic problems (see [Mak07] and references therein) to space discretizations in the hyperbolic setting. The main idea of the reconstruction technique is to introduce an intermediate function, which we will denote \( \tilde{u} \), which solves a perturbed partial differential equation (PDE). This perturbed PDE is constructed in such a way that this \( \tilde{u} \) is sufficiently close to both the approximate solution, denoted \( u_h \), and the exact solution to the conservation law, denoted \( u \). Then, typically

\[
\| u - u_h \| \leq \| u - \tilde{u} \| + \| \tilde{u} - u_h \| ,
\]

where \( \| \tilde{u} - u_h \| \) can be controlled explicitly and \( \| u - \tilde{u} \| \) is estimated using perturbation stability techniques. For systems of hyperbolic conservation laws admitting a convex entropy the relative entropy technique, introduced in [Daf79, DiP79], provides a natural stability framework in the case where one of the two functions involved in the analysis is a Lipschitz solution of the conservation law. This technique is based on the fact that usually systems of hyperbolic conservation laws are endowed with an entropy/entropy flux pair. For conservation laws describing physical systems this notion of entropy follows from the physical one. The entropy/entropy flux pair gives rise to an admissibility condition for weak solutions (cf. Definition 2.1), which leads to the notion of entropy solutions. It can also be used to define the notion of relative entropy between two solutions. In the case of a convex entropy the relative entropy can be used to control the \( L_2 \) distance. It can be used to obtain a stability result (Theorem 2.7), which implies uniqueness of Lipschitz solutions in the class of entropy solutions. One drawback of this stability framework is that a Gronwall type argument has to be employed such that the error estimate depends exponentially on time. There are two features of the relative entropy framework which need to be taken into account when constructing the reconstruction \( \tilde{u} \). If the relative entropy is to be used to compare \( u, \tilde{u} \), one of the two needs to be Lipschitz. As \( u \) may be discontinuous, \( \tilde{u} \) needs to be Lipschitz. Second, the relative entropy is an \( L_2 \) framework; thus, the residuals in the perturbed equation satisfied by \( \tilde{u} \) need to be in \( L_2 \).

Relative entropy techniques for the a priori error analysis of approximations of systems of conservation laws were first used in [AMT04]. For other works concerning analysis of schemes for systems of conservation laws see, e.g., [JR05, JR06]. For discontinuous Galerkin/Runge–Kutta (dGRK) schemes a priori estimates can be found in [ZS04, ZS06, ZS10]. In [HH02] the authors use a goal-oriented framework providing error indicators for a space-time dG scheme. These indicators are computable, provided that certain dual problems are well posed. Asymptotic nodal superconvergence is investigated in a series of papers; see [BA11] and references therein. In [DMO07] the authors provide an a posteriori estimate for the \( L_1 \) error of dGRK schemes approximating a scalar conservation law; see also [Ohl09] for an overview of a posteriori error analysis for hyperbolic conservation laws.

The novelty of this work is that it provides a posteriori estimates for some dG schemes for nonlinear systems of conservation laws. Notice we do not assume anything on the exact solution apart from the fact that it takes values on a compact set known
a priori. That said, the final estimate is 
conditional, i.e., holds under assumptions on
the approximation and its reconstruction (see [MN06, Mak07]), which can be verified
a posteriori. It must be noted, however, that our estimates are only robust before
the formation of shocks, in the case where the entropy solution is discontinuous, our
error estimator does not converge to zero if the meshwidth goes to zero. This is
explained in detail in Remark 5.7 and is an expected direct consequence of the fact
that in the relative entropy framework the Lipschitz constant of one of the solutions,
which are compared to each other, enters the error estimate. The extension of our
approach to the case of nonsmooth solutions is a very challenging problem which
is currently under investigation. The need of introducing reconstruction operators
imposes some restrictions on the permitted discrete fluxes used in the dG method.
The schemes falling under our framework include (but are not limited to) Godunov
schemes; see Remark 3.1 for further details. We present our analysis in the one-
dimensional case with periodic boundary conditions. An extension of our results to
several space dimensions would require a generalized reconstruction technique while
the other arguments would be analogous.

The remainder of this paper is organized as follows. In section 2 we give some
background on hyperbolic conservation laws and their stability via the relative entropy
method. In section 3 we describe the numerical schemes under consideration. In
section 4 we provide some background on reconstruction methods and we discuss the
reconstruction procedure which we employ here and study its properties. In section
5 we combine the reconstruction and the relative entropy methodology to derive an a
posteriori error estimate. In section 7 we show some numerical experiments employing
the estimates derived in section 5, studying their asymptotic properties. Finally, in
section 8 we conclude.

2. Preliminaries, conservation laws, and relative entropy. In this section we formalize our notation, introduce the model problem, and detail the relative entropy stability framework.

Given the standard Lebesgue space notation [Cia02, Eva98] we begin by introducing the Sobolev spaces. Let \( \Omega \subset \mathbb{R} \); then

\[
W^k_p(\Omega) := \{ \phi \in L^p(\Omega) : D^\alpha \phi \in L^p(\Omega) \text{ for } |\alpha| \leq k \},
\]

which are equipped with norms and seminorms

\[
\|u\|_{W^k_p(\Omega)} := \begin{cases} \left( \sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p} & \text{if } p \in [1, \infty), \\ \sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^\infty(\Omega)} & \text{if } p = \infty, \end{cases}
\]

\[
|u|_{W^k_p(\Omega)} := \left\| D^k u \right\|_{L^p(\Omega)},
\]

respectively, where derivatives \( D^\alpha \) are understood in a weak sense.

We use the convention that when derivatives act on a vector valued multivariate
function, \( \mathbf{u}=(u_1, \ldots, u_d)^T \), it is meant componentwise, that is, \( \partial_x \mathbf{u}=(\partial_x u_1, \ldots, \partial_x u_d)^T \)
denotes a column vector. The derivative of a field, \( q \), say, with respect to the dependent
variable is denoted \( \mathbf{D}q = (\partial_{u_1} q(\mathbf{u}), \ldots, \partial_{u_d} q(\mathbf{u})) \) which is a row vector. The matrix of
second derivatives of \( q \) is

\[
\mathbf{D}^2 q(\mathbf{u}) := \begin{bmatrix}
\partial_{u_1,u_1} q(\mathbf{u}), & \ldots, & \partial_{u_1,u_d} q(\mathbf{u}) \\
\vdots & \ddots & \vdots \\
\partial_{u_d,u_1} q(\mathbf{u}), & \ldots, & \partial_{u_d,u_d} q(\mathbf{u})
\end{bmatrix}.
\]
For a vector field $f$, we denote its Jacobian by $Df$ which is also a $d \times d$ matrix and its Hessian as $D^2f$ which is given as a 3-tensor. We also make use of the following notation for time dependent Sobolev (Bochner) spaces:

$$L_\infty(0,T; W^k_p(\Omega)) := \left\{ u : [0,T] \rightarrow W^k_p(\Omega) : \sup_{t \in [0,T]} \| u(t) \|_{W^k_p(\Omega)} < \infty \right\}.$$

Let $U \subset \mathbb{R}^d$ convex be the state space. We consider the following first order (system of) conservation laws:

$$\partial_t u(x,t) + \partial_x f(u(x,t)) = 0 \text{ for } (x,t) \in (0,1) \times (0,\infty).$$

We complement (2.6) with the initial and boundary conditions

$$u(0,t) = u(1,t) \text{ for } t \in (0,\infty) \text{ and } u(x,0) = u_0(x) \text{ for } x \in (0,1)$$

for some function $u_0 \in L_\infty((0,1),U)$. The solution, which in general is only in $L_\infty((0,1) \times (0,\infty),U)$, takes values in the state space and we assume the flux function $f : U \rightarrow \mathbb{R}^d$ is at least $C^2(U)$. Throughout this paper we will assume that there is an entropy/entropy flux pair $(\eta,q)$ with $\eta \in C^2(U,\mathbb{R})$ strictly convex and $q \in C^1(U,\mathbb{R})$ associated to (2.6) in such a way that

$$Dq = D\eta Df.$$

The existence of an entropy flux implies that

$$(Df)^T D^2\eta = D^2\eta Df.$$  

It is readily verifiable that strong solutions of (2.6) satisfy the additional conservation law

$$\partial_t \eta(u) + \partial_x q(u) = 0.$$

For general background on hyperbolic conservation laws the reader is referred to [Daf10, LeF02]. Note that not every system of hyperbolic conservation laws admits a convex entropy/entropy flux pair (see [Daf10, sect. 5.4]), even if it is physically meaningful. The derivation of a posteriori error estimates for systems of hyperbolic conservation laws admitting only poly or quasi-convex entropies is beyond the scope of this work. It is common that solutions of (2.6) develop discontinuities after finite time. This motivates developing a notion of weak solution. As weak solutions, which satisfy the equation in the distributional sense, are not unique, attention is restricted to so-called entropy solutions $u \in L_\infty((0,1) \times (0,\infty),U)$. The concept of entropy solution guarantees uniqueness of solutions for scalar problems and can be interpreted as enforcing that solutions are compatible with the second law of thermodynamics. However, it is important to note that entropy solutions need not be unique for systems of conservation laws in multiple space dimensions even if these are endowed with a convex entropy [DLS10]. In this context it should be noted that the relative entropy technique (see Lemma 2.7) guarantees uniqueness for entropy solutions if and only if they are Lipschitz.
Definition 2.1 (entropy solution). A function $u \in L_\infty((0,1) \times [0,\infty), U)$ is said to be an entropy solution of the initial boundary value problem (2.6)–(2.7), with associated entropy/entropy flux pair $(\eta, q)$, if

$$
\int_0^\infty \int_0^1 u \cdot \partial_t \phi + f(u) \cdot \partial_x \phi \, dx \, dt + \int_0^1 u_0 \cdot \phi(\cdot,0) \, dx = 0 \quad \forall \phi \in C_c^\infty(S^1 \times [0,\infty), \mathbb{R}^d)
$$

and

$$
\int_0^1 \int_0^\infty \eta(u) \partial_t \phi + q(u) \partial_x \phi \, dx \, dt + \int_0^1 \eta(u_0) \phi(\cdot,0) \, dx \geq 0 \quad \forall \phi \in C_c^\infty(S^1 \times [0,\infty), [0,\infty)).
$$

Here $S^1$ (the 1-sphere) refers to the unit interval $[0,1]$ with matching endpoints.

Remark 2.2 (scalar case). In the scalar case entropy solutions are required to satisfy (2.12) for every convex entropy/entropy flux pair.

For $u \in L_\infty((0,1) \times (0,\infty), U)$ the distribution $\partial_t \eta(u) + \partial_x q(u)$ has a sign and therefore is a measure, i.e., we may replace the smooth test functions in Definition 2.1 by Lipschitz continuous ones. Stability of solutions and in particular uniqueness of Lipschitz solutions within the class of entropy solutions is obtained via relative entropy arguments; see [Daf10, Chap. 5] and references therein.

Definition 2.3 (relative entropy and entropy flux). We define the relative entropy, $\eta(u \mid v)$, and relative entropy flux, $q(u \mid v)$, of two generic vector valued functions $v$ and $w$ with values in $U$ to be

$$
\eta(v \mid w) := \eta(v) - \eta(w) - D\eta(w)(v - w),
$$

$$
q(v \mid w) := q(v) - q(w) - D\eta(w)(f(v) - f(w)).
$$

Note that $\eta(v \mid w)$ and $q(v \mid w)$ are not symmetric in $v, w$.

Assumption 2.4 (values in a compact set). We will assume throughout the paper that the exact solution $u$ of (2.6) takes values in $\mathcal{D}$, i.e.,

$$u(x,t) \in \mathcal{D} \quad \forall \ (x,t) \in (0,1) \times (0,\infty),$$

where $\mathcal{D}$ be a compact and convex subset of $U$.

Remark 2.5 (bounds on flux and entropy). Due to the regularity of $f$ and $\eta$ and the compactness of $\mathcal{D}$ there are constants $0 < C_T < \infty$ and $0 < C_\eta < C_\eta < \infty$ such that

$$
|v^T D f(u)| \leq C_T |v|^2, \quad C_\eta |v|^2 \leq v^T D^2 \eta(u)v \leq C_\eta |v|^2 \quad \forall v \in \mathbb{R}^d, u \in \mathcal{D},
$$

where $|\cdot|$ is the Euclidean norm for vectors. Note that $C_T, C_\eta$, and $C_\eta$, can be explicitly computed from $\mathcal{D}, f$, and $\eta$.

Lemma 2.6 (Gronwall inequality). Given $T > 0$, let $\phi(t) \in C^0([0,T])$ and $a(t), b(t) \in L_1([0,T])$ all be nonnegative functions with $b$ nondecreasing and satisfying

$$
\phi(t) \leq \int_0^t a(s) \phi(s) \, ds + b(t).
$$

Then

$$
\phi(t) \leq b(t) \exp \left( \int_0^t a(s) \, ds \right) \quad \forall t \in [0,T].
$$
As we will make use of a similar argument to derive our resultant a posteriori error estimate we give the proof of the following stability result, which can be found in [Daf10].

**Lemma 2.7** (L₂ stability [Daf10]). Let u be an entropy solution of (2.6)–(2.7) corresponding to initial data \( u₀ \) and \( v \) a Lipschitz solution of (2.6)–(2.7) corresponding to initial data \( v₀ \). Let \( u \) and \( v \) take values in \( \mathcal{D} \). Then there exist constants \( C₁, C₂ > 0 \) such that

\[
\|u(\cdot, t) - v(\cdot, t)\|_{L₂(t)} \leq C₁ \exp(C₂t) \|u₀ - v₀\|_{L₂(t)}.
\]

*Note that \( C₂ \) depends on the Lipschitz constant of \( v \).*

**Proof.** Note that \( v \) satisfies (2.12) as an equality. Thus, for any Lipschitz continuous, nonnegative test function \( \phi \) we have

\[
0 \leq \int_{0}^{∞} \int_{0}^{1} \partial_t \phi (\theta(u) - \theta(v)) + \partial_x \phi (\eta(u) - \eta(v)) \, dt \, dx \quad \text{for any } \phi \in L^1(0,1) \quad (2.18)
\]

Using the definition of relative entropy and relative entropy flux, we may reformulate this as

\[
0 \leq \int_{0}^{∞} \int_{0}^{1} \partial_t \phi (\theta(u) | v + D\eta(v)(u - v)) + \partial_x \phi (\eta(u) + D\eta(v)(f(u) - f(v))) \, dt \, dx
\]

\[
+ \int_{0}^{1} \phi(\cdot, 0) (\eta(u₀) - \eta(v₀)) \, dx.
\]

Upon using the Lipschitz continuous test function \( \phi = \phi D\eta(v) \) in (2.11) for \( u \) and \( v \), we obtain

\[
0 = \int_{0}^{∞} \int_{0}^{1} \partial_t (\phi D\eta(v))(u - v) + \partial_x (\phi D\eta(v))(f(u) - f(v)) \, dt \, dx
\]

\[
+ \int_{0}^{1} \phi(\cdot, 0) D\eta(v)(u₀ - v₀) \, dx.
\]

We use the product rule in (2.20) and combine it with (2.19) to obtain

\[
0 \leq \int_{0}^{∞} \int_{0}^{1} \partial_t \phi \theta(u) + \partial_x \phi \eta(u) \, dx \, dt
\]

\[
- \int_{0}^{∞} \int_{0}^{1} \phi(\partial_t v D²\eta(v)(u - v) + \partial_x v D²\eta(v)(f(u) - f(v))) \, dx \, dt + \int_{0}^{1} \phi(\cdot, 0) \eta(u₀ | v₀) \, dx.
\]

Using \( \partial_t v = -Df(v) \partial_x v \) and (2.9) we find

\[
0 \leq \int_{0}^{∞} \int_{0}^{1} \partial_t \phi \theta(u) + \partial_x \phi \eta(u) \, dx \, dt
\]

\[
- \int_{0}^{∞} \int_{0}^{1} \phi(\partial_x v D²\eta(v)(f(u) - f(v) - Df(v)(u - v)) + \int_{0}^{1} \phi(\cdot, 0) \eta(u₀ | v₀) \, dx.
\]
Now we fix $t > 0$. Then for every $0 < s < t$ and $\varepsilon > 0$ we consider the test function

\begin{equation}
\phi(x, \sigma) = \begin{cases} 
1 & : \sigma < s, \\
1 - \frac{\sigma - s}{\varepsilon} & : s < \sigma < s + \varepsilon, \\
0 & : \sigma > s + \varepsilon.
\end{cases}
\end{equation}

In this case we infer from (2.22)

\begin{equation}
0 \leq -\frac{1}{\varepsilon} \int_s^{s+\varepsilon} \int_0^1 \eta(u | v) \, dx \, dt \\
- \int_0^1 \int_0^1 \phi(\partial_x vD^2\eta(v)(f(u) - f(v) - Df(v)(u - v))) \, dx \, dt + \int_0^1 \eta(u_0 | v_0) \, dx.
\end{equation}

When sending $\varepsilon \to 0$ we find for all points $s$ of $L_\infty$-weak-*-continuity of $\eta(u(\cdot, \sigma))$ in $(0, t)$ that

\begin{equation}
0 \leq -\int_0^1 \eta(u(x, s) | v(x, s)) \, dx \\
- \int_0^s \int_0^1 \partial_x vD^2\eta(v)(f(u) - f(v) - Df(v)(u - v)) \, dx \, dt + \int_0^1 \eta(u_0 | v_0) \, dx.
\end{equation}

Upon using (2.14) we infer that for almost all $s \in (0, t)$

\begin{equation}
C_2^2 \|u(\cdot, s) - v(\cdot, s)\|_{L_2(I)}^2 \leq C_\eta \|u_0 - v_0\|_{L_2(I)}^2 \\
+ C_\eta C_\gamma \int_0^1 \|v(\cdot, \sigma)\|_{W^{1, \infty}(I)} \|u(\cdot, \sigma) - v(\cdot, \sigma)\|_{L_2(I)}^2 \, d\sigma.
\end{equation}

This equation, in fact, holds for all $s \in (0, t)$ as $u$ is weakly lower semicontinuous. Since $v$ is Lipschitz continuous, applying Gronwall’s lemma completes the proof.

3. The semidiscrete scheme. In this section we introduce the class of semi-discrete problem which we consider in this contribution.

We will discretize (2.6) in space using consistent dG finite element methods. Let $I := [0, 1]$ be the unit interval and choose $0 = x_0 < x_1 < \cdots < x_N = 1$. We denote $I_n = [x_n, x_{n+1}]$ to be the $n$th subinterval and let $h_n := x_{n+1} - x_n$ be its size. Let $\mathbb{P}^p(I)$ be the space of polynomials of degree less than or equal to $p$ on $I$; then we denote

\begin{equation}
\mathbb{V}_p := \{g : I \to \mathbb{R}^d : g_i|_{I_n} \in \mathbb{P}^p(I_n) \text{ for } i = 1, \ldots, d, \ n = 0, \ldots, N - 1\},
\end{equation}

where $g = (g_1, \ldots, g_d)^T$, to be the usual space of piecewise $p$th degree polynomials for vector valued functions over $I$. In addition we define jump and average operators such that

\begin{equation}
[g]_n := g(x_n^-) - g(x_n^+) := \lim_{s \searrow 0} g(x_n - s) - \lim_{s \nearrow 0} g(x_n + s),
\end{equation}

\begin{equation}
\{\!\{g\}!\}_n := \frac{1}{2} (g(x_n^-) + g(x_n^+)) := \frac{1}{2} \left( \lim_{s \searrow 0} g(x_n - s) + \lim_{s \nearrow 0} g(x_n + s) \right).
\end{equation}
A POSTERIORI ANALYSIS FOR CONSERVATION LAWS

We will examine the following class of semidiscrete numerical schemes [GR96, Krö97, HW08], where \( u_h \in C^1([0,T), V_p) \) is determined such that

\[
0 = \sum_{n=0}^{N-1} \int_{I_n} \left( \partial_t u_h \cdot \phi + \partial_x f(u_h) \cdot \phi \right) \, dx \\
+ \sum_{n=0}^{N-1} \left( F(u_h(x_n^-), u_h(x_n^+)) \cdot \left[ \phi \right]_n - \left[ f(u_h) \cdot \phi \right]_n \right) \quad \forall \phi \in V_p.
\]

(3.3)

In what follows we will assume that (3.3) has a solution and in particular that \( u_h \) takes values in \( U \). We also set

\[
\llbracket u_h \rrbracket_0 := u_h(x_N^-) - u_h(x_N^+) ; \quad \llbracket u_h \rrbracket_0 := \frac{u_h(x_N^+) + u_h(x_N^-)}{2}
\]

to account for the periodic boundary conditions. Here \( F : U^2 \subset \mathbb{R}^{2d} \to \mathbb{R}^d \) is a numerical flux function. We restrict our attention to a certain class of numerical flux functions. We impose that there exists a function

\[
w : U \times U \to U \text{ such that } F(u, v) = f(w(u, v))
\]

and that there exists a constant \( L > 0 \) such that \( w \) satisfies

\[
|w(u, v) - u| \leq L |u - v| , \quad |w(u, v) - v| \leq L |u - v| \quad \forall u, v \in U.
\]

(3.6)

Remark 3.1 (restriction of fluxes). The reason for the restriction on the choice of fluxes will be made apparent later. Our assumptions are met by Godunov schemes employing exact Riemann solvers. For approximate Riemann solvers there are two classes [LeV02, sect. 12.3]. Our assumption is generally satisfied for the class in which the numerical flux is computed by evaluating the exact flux on some intermediate state extracted from an approximate Riemann solution. For the second class, which encompasses, e.g., the Roe scheme, the situation is more involved.

Let us look at some numerical fluxes in special cases. In the case of the inviscid Burgers’ equation, i.e., \( f(u) = \frac{u^2}{2} \), our condition is not satisfied for the local and global Lax–Friedrichs scheme. For the local Lax–Friedrichs scheme the numerical flux reads

\[
F(a, b) = \frac{1}{2} (a^2 + b^2) + \max(|a|, |b|)(a - b),
\]

(3.7)

which is negative for \( a = 0 \) and \( b > 0 \). Therefore there can be no \( w \in U \) satisfying \( f(w) = F(0, b) \). The argument for the global Lax–Friedrichs scheme is analogous.

For the inviscid Burgers’ equation both the Roe and the Engquist–Osher flux satisfy our condition, with

\[
w_{EO}(a, b) = \left( \frac{1}{2} a^2 (1 + \text{sgn}(a)) + \frac{1}{2} b^2 (1 - \text{sgn}(b)) \right)^{1/2}
\]

(3.8)

and

\[
w_{Roe}(a, b) = \left( \frac{1}{2} a^2 (1 + \text{sgn}(a + b)) + \frac{1}{2} b^2 (1 - \text{sgn}(a + b)) \right)^{1/2}.
\]

(3.9)
The situation is far more complicated for nonlinear systems. In fact, for the $p$-system which is given by
\[
\begin{align*}
\partial_t u - \partial_x v &= 0, \\
\partial_t v - \partial_x p(u) &= 0
\end{align*}
\]
for some function $p$ with $p' > 0$, the question whether the Roe scheme fits into our framework hinges on whether $p$ is surjective.

4. **Reconstruction and projection operators.** To analyze the scheme (3.3) we introduce reconstructions which we denote by $\hat{u}$ and $\hat{f}$. For brevity we will omit the time dependency of all quantities in this section.

**Definition 4.1 (reconstruction of $u_h$).** The reconstruction $\hat{u}$ is the unique element of $\mathbb{V}_{p+1}$ such that
\[
\begin{align*}
\sum_{n=0}^{N-1} \int_{I_n} \hat{u} \cdot \phi \, dx &= \sum_{n=0}^{N-1} \int_{I_n} u_h \cdot \phi \, dx \quad \forall \phi \in \mathbb{V}_{p-1} \\
\text{and} \\
\hat{u}(x^+_n) &= w(u_h(x^-_n), u_h(x^+_n)) \quad \text{and} \\
\hat{u}(x^-_{n+1}) &= w(u_h(x^-_{n+1}), u_h(x^+_n)) \quad \forall n \in [0, N-1],
\end{align*}
\]
recalling that $u_h(x^-_0) := u_h(x^-_N)$ and $u_h(x^+_0) := u_h(x^+_N)$.

**Definition 4.2 (reconstruction of $f(u_h)$).** The reconstruction $\hat{f}$ is the unique element of $\mathbb{V}_p$ such that
\[
\begin{align*}
\sum_{n=0}^{N-1} \int_{I_n} \partial_x \hat{f} \cdot \phi \, dx &= \sum_{n=0}^{N-1} \int_{I_n} \partial_x f(u_h) \cdot \phi \, dx \\
&\quad + \sum_{n=0}^{N-1} (f(u_h(x^-_n), u_h(x^+_n))) \cdot \langle \phi \rangle - \langle f(u_h) \cdot \phi \rangle_n \\
(4.4) \\
\text{coupled with the skeletal “boundary” conditions that} \\
\hat{f}(x^+_n) &= f(u_h(x^-_n), u_h(x^+_n)) \quad \forall n \in [0, N-1].
\end{align*}
\]

**Lemma 4.3 (continuity and orthogonality).** The reconstructions $\hat{u}$ and $\hat{f}$ given in Definitions 4.1, and 4.2, respectively, are continuous and $\hat{f}$ satisfies the orthogonality property
\[
\sum_{n=0}^{N-1} \int_{I_n} (\hat{f} - f(u_h)) \cdot \phi \, dx = 0 \quad \forall \phi \in \mathbb{V}_{p-1}.
\]

**Proof.** The continuity of $\hat{u}$ follows from (4.2)–(4.3). To prove the continuity of $\hat{f}$ we choose $\phi$ as the $i$th unit vector on $I_n$ and zero elsewhere. Then, upon letting $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_d)^T$ and $f = (f_1, \ldots, f_d)^T$ we obtain from (4.4)
\[
\begin{align*}
\hat{f}_i(x^-_{n+1}) - \hat{f}_i(x^+_n) &= f_i(u_h(x^-_{n+1}), u_h(x^-_n), u_h(x^+_n)) - f_i(u_h(x^-_n), u_h(x^+_n)) \\
&\quad + f_i(u_h(x^-_{n+1}), u_h(x^+_n)) + f_i(u_h(x^-_n), u_h(x^+_n)).
\end{align*}
\]
This implies
\[ \hat{f}_i(x_{n+1}) = f_i(w(u_h(x_{n+1}^-), u_h(x_{n+1}^+))) \]
due to (4.5). This shows the continuity of \( \hat{f} \). Using integration by parts in (4.4) we have that the boundary terms cancel due to our choice of \( \hat{f}(x_n^+) \) and (4.8). Hence, we find
\[
\sum_{n=0}^{N-1} \int_{I_n} \hat{f} \cdot \partial_x \phi \, dx = \sum_{n=0}^{N-1} \int_{I_n} f(u_h) \cdot \partial_x \phi \, dx \quad \forall \phi \in \mathbb{V}_p,
\]
concluding the proof.

**Definition 4.4 (L_2 projection).** We define \( P_p : [L_2(I)]^d \to \mathbb{V}_p \) to be the \( L_2 \) orthogonal projection to \( \mathbb{V}_p \), that is,
\[ \int_I \psi \cdot \phi \, dx = \int_I P_p \psi \cdot \phi \, dx \quad \forall \phi \in \mathbb{V}_p. \]
If \( \psi \in W^{p+1}_\infty(I) \) the operator is well known [Cia02] to satisfy the following estimate in \( L_\infty \):
\[ \| \psi - P_p \psi \|_{L_\infty(I_n)} \leq C_p h^{p+1}_n \| \psi \|_{W^{p+1}_\infty} \forall n = 0, \ldots, N - 1. \]

**Remark 4.5 (restriction of fluxes revisited).** The assumption on the numerical flux functions (3.5) is posed such that we can choose our reconstructions \( \hat{u}, \hat{f} \) such that \( \hat{f}(x_n) = f(\hat{u}(x_n)) \) for all \( n \). This is needed for the proof of Lemma 6.2 and it will be elaborated upon in Remark 6.3.

**5. A posteriori control based on computation of local reconstructions.**
In this section we make use of the reconstruction operators from section 4 to construct a posteriori bounds for the generic numerical scheme (3.3). This allows us, using the relative entropy stability framework, to state a fully computable a posteriori bound.

Using these reconstructions \( \hat{u} \) (given in Definition 4.1) and \( \hat{f} \) (given in Definition 4.2) we can rewrite our scheme as
\[ 0 = \sum_{n=0}^{N-1} \int_{I_n} \partial_x u_h \cdot \phi \, dx + \sum_{n=0}^{N-1} \int_{I_n} \partial_x \hat{f} \cdot \phi \, dx \quad \forall \phi \in \mathbb{V}_p. \]
Since we have that \( \partial_x u_h \) and \( \partial_x \hat{f} \) are piecewise polynomials of degree \( p \) we may write (5.1) as a pointwise equation
\[ \partial_t \hat{u} + \partial_x \hat{f}(\hat{u}) = \partial_x f(\hat{u}) - \partial_x \hat{f} + \partial_x \hat{u} - \partial_t u_h =: R. \]
Using the relative entropy technique we obtain the following preliminary error estimate.

**Lemma 5.1 (error bound for the reconstruction).** Let \( u \) be the entropy solution of (2.6), (2.7); then the difference between \( u \) and the reconstruction \( \hat{u} \) satisfies
\[
C_2 \| u(\cdot, s) - \hat{u}(\cdot, s) \|_{L_2(I)}^2 \leq C_\gamma \| u_0 - \hat{u}_0 \|_{L_2(I)}^2 + \| R \|_{L_2(I \times (0,s))}^2 + (C_\gamma C_\gamma \| \hat{u} \|_{W^{1,\infty}} + C_\gamma^2) \int_0^s \| u(\cdot, \sigma) - \hat{u}(\cdot, \sigma) \|_{L_2(I)}^2 \, d\sigma
\]
for every \( s \in (0, \infty) \), provided \( \hat{u} \) takes values in \( \Omega \).
Proof. Since $\tilde{u}$ is Lipschitz continuous, we multiply (5.2) by $D\eta(\tilde{u})$ and find for any Lipschitz continuous, nonnegative test function $\phi$
\begin{align}
0 \leq \int_0^1 \int_0^1 \partial_t \phi(\eta(u) - \eta(\tilde{u})) + \partial_x \phi(q(u) - q(\tilde{u})) - \phi D\eta(\tilde{u}) R \, dx \, dt \\
+ \int_0^1 \phi(\cdot, 0)(\eta(u_0) - \eta(\tilde{u}_0)) \, dx.
\end{align}

Using the definition of relative entropy and relative entropy flux, we may reformulate this as
\begin{align}
0 \leq \int_0^1 \int_0^1 \partial_t \phi(\eta(u | \tilde{u}) + D\eta(\tilde{u})(u - \tilde{u})) + \partial_x \phi(q(u | \tilde{u}) + D\eta(\tilde{u})(f(u) - f(\tilde{u}))) \, dx \, dt \\
- \int_0^1 \int_0^1 \phi D\eta(\tilde{u}) R \, dx \, dt + \int_0^1 \phi(\cdot, 0)(\eta(u_0) - \eta(\tilde{u}_0)) \, dx.
\end{align}

Using the Lipschitz continuous test function $\phi = \phi D\eta(\tilde{u})$ in (2.11) and (5.2) we obtain
\begin{align}
0 = \int_0^1 \int_0^1 \partial_t(\phi D\eta(\tilde{u}))(u - \tilde{u}) + \partial_x(\phi D\eta(\tilde{u}))(f(u) - f(\tilde{u})) - \phi D\eta(\tilde{u}) R \, dx \, dt \\
+ \int_0^1 \phi(\cdot, 0)D\eta(\tilde{u}(\cdot, 0))(u_0 - \tilde{u}_0) \, dx.
\end{align}

We use the product rule in (5.6) and combine it with (5.5) to obtain
\begin{align}
0 \leq \int_0^1 \int_0^1 \partial_t \phi \eta(u | \tilde{u}) + \partial_x \phi q(u | \tilde{u}) \, dx \, dt \\
- \int_0^1 \int_0^1 \phi(\partial_x \tilde{u} D^2 \eta(\tilde{u})(u - \tilde{u}) + \partial_x \tilde{u} D^2 \eta(\tilde{u})(f(u) - f(\tilde{u}))) \, dx \, dt \\
+ \int_0^1 \phi(\cdot, 0)\eta(u_0 | \tilde{u}_0) \, dx.
\end{align}

Using the fact that $\partial_x \tilde{u} = -Df(\tilde{u})\partial_x \tilde{u} + R$ and (2.9) we find
\begin{align}
0 \leq \int_0^1 \int_0^1 \partial_t \phi \eta(u | \tilde{u}) + \partial_x \phi q(u | \tilde{u}) \, dx \, dt \\
- \int_0^1 \int_0^1 \phi(\partial_x \tilde{u} D^2 \eta(\tilde{u})(f(u) - f(\tilde{u}) - Df(\tilde{u})(u - \tilde{u}))) \, dx \, dt \\
- \int_0^1 \int_0^1 \phi(u - \tilde{u})^T D^2 \eta(\tilde{u}) R \, dx \, dt + \int_0^1 \phi(\cdot, 0)\eta(u_0 | \tilde{u}_0) \, dx.
\end{align}

Now we fix $t > 0$; then for every $0 < s < t$ and $\varepsilon > 0$ we consider the test function $\phi(x, \sigma)$ given in (2.23). In this case we infer from (2.22)
\begin{align}
0 \leq -\frac{1}{\varepsilon} \int_s^{s+\varepsilon} \int_0^1 \eta(u | \tilde{u}) \, dx \, dt - \int_0^1 \int_0^1 \phi(\partial_x \tilde{u} D^2 \eta(\tilde{u})(f(u) - f(\tilde{u}) - Df(\tilde{u})(u - \tilde{u}))) \, dx \, dt \\
- \int_0^1 \int_0^1 \phi(u - \tilde{u})^T D^2 \eta(\tilde{u}) R \, dx \, dt + \int_0^1 \eta(u_0 | \tilde{u}_0) \, dx.
\end{align}
When sending $\varepsilon \to 0$ we find for all points $s$ of $L^\infty$-weak-*-continuity of $\eta(u(\cdot, \sigma))$ in $(0, t)$ that

\begin{equation}
0 \leq -\int_0^1 \eta(u(x, s) \mid \bar{u}(x, s)) \, dx - \int_0^s \int_0^1 \partial_x \bar{u} D^2 \eta(\bar{u}) (f(u) - f(\bar{u}) - Df(\bar{u})(u - \bar{u})) \, dx \, dt
\end{equation}

\begin{equation}
- \int_0^s \int_0^1 (u - \bar{u}) D^2 \eta(\bar{u}) R \, dx \, dt + \int_0^1 \eta(u_0 \mid \bar{u}_0) \, dx \geq 0.
\end{equation}

Upon using (2.14) and the convexity of $\Theta$ we infer that for almost all $s \in (0, t)$

\begin{equation}
C_{\eta} \|u(\cdot, s) - \bar{u}(\cdot, s)\|_{L^2(I)}^2 \leq C_{\eta} \|u_0 - \bar{u}_0\|_{L^2(I)}^2 + \|R\|_{L^2(I\times(0, s))}^2
\end{equation}

\begin{equation}
+ (C_{\eta}^2 \|\bar{u}\|_{W^{1, \infty}} + C_{\eta}^2) \int_0^s \|u(\cdot, \sigma) - \bar{u}(\cdot, \sigma)\|_{L^2(I)}^2 \, d\sigma.
\end{equation}

This equation, in fact, holds for all $s \in (0, t)$ as $u$ is weakly lower semicontinuous. \hfill \square

Remark 5.2 (values of $\bar{u}$). Note that the condition that $\bar{u}$ takes values in $\mathcal{D}$ can be verified in an a posteriori fashion, as $\bar{u}$ can be explicitly computed.

Proposition 5.3 (Legendre polynomials [AW05]). Let $l_k$ denote the $k$th Legendre polynomial on $(-1, 1)$, and $l_k^n$ its transformation to the interval $I_n$, i.e.,

\begin{equation}
l_k^n(x) = l_k \left( \frac{x - x_n}{h_n} \right) - 1 \right).
\end{equation}

Let $\alpha_k := \partial_x l_k(1)$. Then $l_k^n$ has the following properties:

\begin{equation}
(-1)^k l_k^n(x_n) = l_k^n(x_{n+1}) = 1,
\end{equation}

\begin{equation}
(-1)^{k+1} h_n \partial_x l_k^n(x_n) = h_n \partial_x l_k^n(x_{n+1}) = 2 \alpha_k,
\end{equation}

\begin{equation}
\int_{I_n} l_k^n(x) l_k^n(x) \, dx = \frac{2h_n}{2k+1} \delta_{kj} \leq h_n,
\end{equation}

\begin{equation}
|l_k^n(x)| \leq 1 \quad \forall x \in I_n.
\end{equation}

Lemma 5.4 (Legendre representation of $\bar{u}$). The reconstruction $\bar{u}$ given by Definition 4.1 satisfies the following representation for all $x \in I_n$:

\begin{equation}
(\bar{u} - u_h)(x) = \frac{1}{2} \left( (-1)^p \left( w(u_h(x_n^-), u_h(x_n^+)) - u_h(x_n^+) \right)
\end{equation}

\begin{equation}
+ w(u_h(x_{n+1}^-), u_h(x_{n+1}^+)) - u_h(x_{n+1}^-) \right) l_k^n(x)
\end{equation}

\begin{equation}
+ \frac{1}{2} \left( (-1)^{p+1} \left( w(u_h(x_n^-), u_h(x_n^+)) - u_h(x_n^+) \right)
\end{equation}

\begin{equation}
+ w(u_h(x_{n+1}^-), u_h(x_{n+1}^+)) - u_h(x_{n+1}^-) \right) l_{p+1}^n(x),
\end{equation}

where $l_k^n$ and $l_{p+1}^n$ are the rescaled Legendre polynomials from Proposition 5.3. Therefore,

\begin{equation}
\|\bar{u} - u_h\|_{L^2(I_n)}^2 \leq L^2 h_n \left( \|u_h\|^2_n + \|u_h\|^2_{n+1} \right)
\end{equation}
for real numbers

and

\[
\| \partial^k u \|_{L^\infty(I_n)} \leq \| \partial^k u_h \|_{L^\infty(I_n)} + L \frac{1}{b_k} \left( \| u_h \|_n + \| u_h \|_{n+1} \right),
\]

where \( b_k := |l_p|_{k,\infty} + |l_{p+1}|_{k,\infty} \).

**Proof.** Letting \( \tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_d)^T \) and \( u_h = ((u_h)_1, \ldots, (u_h)_d)^T \) and writing \( \tilde{u}_i|_{I_n} \) and \( (u_h)_i|_{I_n} \) as linear combinations of Legendre polynomials we see that (4.1) implies

\[
(\tilde{u}_i - (u_h)_i)(x) = \alpha l_p^n(x) + \beta l_{p+1}^n(x) \quad \forall \ x \in I_n
\]

for real numbers \( \alpha, \beta \) depending on \( i \) and \( n \). Using (5.13) and the boundary conditions on \( \tilde{u} \) (4.2)–(4.3) we obtain

\[
\alpha(-1)^p - \beta(-1)^p = \tilde{u}_i(x_n^+) - (u_h)_i(x_n^+) = w_i((u_h)_1(x_n^-), (u_h)_d(x_n^-)) - (u_h)_i(x_n^-)
\]

and

\[
\alpha + \beta = \tilde{u}_i(x_n^-) - (u_h)_i(x_n^-) = w_i((u_h)_1(x_n^-), (u_h)_d(x_n^-)) - (u_h)_i(x_n^-).
\]

Since

\[
\begin{pmatrix}
(-1)^p & (-1)^{p+1} \\
1 & 1
\end{pmatrix}^{-1} = \frac{1}{2} \begin{pmatrix}
(-1)^p & 1 \\
(-1)^{p+1} & 1
\end{pmatrix}
\]

we obtain (5.17). Equations (5.18) and (5.19) are immediate consequences of (5.17) upon using (5.13)–(5.16). \( \Box \)

**Theorem 5.5** (a posteriori error bound). Let \( f \in C^2(U, \mathbb{R}^d) \) satisfy (2.10) and let \( u \) be an entropy solution of (2.6) with periodic boundary conditions. Let \( \tilde{u} \) take values in \( \mathcal{D} \). Then for \( 0 \leq t \leq T \) the error between the numerical solution \( u_h \) and \( u \) satisfies

\[
\| u(\cdot,t) - u_h(\cdot,t) \|_{L^2(I)}^2 \leq 2 \| \tilde{u}(\cdot,t) - u_h(\cdot,t) \|_{L^2(I)}^2
\]

\[
+ 2C_2^{-1} \left( \| R \|_{L^2(I \times (0,t))} + C_\pi \| u_0 - \tilde{u}_0 \|_{L^2(I)} \right)
\]

\[
\times \exp \left( \int_0^t \frac{C_\pi C_\pi \| \partial_x \tilde{u}(\cdot,s) \|_{L^\infty(I)} + C_2^2}{C_2} \, ds \right).
\]

**Proof.** Combining Lemmas 2.6 and 5.1 we obtain

\[
\| u(\cdot,t) - \tilde{u}(\cdot,t) \|_{L^2(I)}^2 \leq C_2^{-1} \left( \| R \|_{L^2(I \times (0,t))} + C_\pi \| u_0 - \tilde{u}_0 \|_{L^2(I)} \right)
\]

\[
\times \exp \left( \int_0^t \frac{C_\pi C_\pi \| \partial_x \tilde{u}(\cdot,s) \|_{L^\infty(I)} + C_2^2}{C_2} \, ds \right).
\]

The triangle inequality and (5.25) imply the assertion of the theorem. \( \Box \)

**Remark 5.6** (values of \( \tilde{u} \)). The \( L^\infty \) estimates based on (5.17) can be employed to verify a posteriori that \( \tilde{u} \) takes values in \( \mathcal{D} \).
Remark 5.7 (discontinuous entropy solutions). The estimate in Theorem 5.5 does not require the entropy solution $u$ to be continuous. However, in case $u$ is discontinuous $\|\partial_x \tilde{u}(\cdot, s)\|_{L^\infty(I)}$ is expected to behave like $O(h^{-1})$. Therefore, the estimator in (5.24) will (at best) be $O(h^{p+1} \exp(h^{-1}))$ which diverges for $h \to 0$. Thus, the estimator in (5.24) is expected not to converge for $h \to 0$ if the entropy solution is discontinuous. The same is true for the estimator derived in Theorem 6.5. This is a consequence of the use of the relative entropy framework and the fact that the entropy solution does not need to be unique if it is not Lipschitz.

Remark 5.8 (comparison to the scalar case).

1. From Theorem 5.5 it becomes clear that it is desirable for $C_{\eta}/C_{\eta}$ to be as small as possible. While this ratio is prescribed by $\eta$ and $\mathcal{D}$ in the systems case, there is freedom in the choice of $\eta$ in the scalar case. Choosing $\eta(u) = \frac{1}{2}u^2$ implies the optimal ratio $C_{\eta}/C_{\eta} = 1$.

2. The estimate in Theorem 5.5 blows up after a shock has formed. This is a direct consequence of the fact that (discontinuous) entropy solutions of general systems of hyperbolic conservation laws are not unique [Daf10, p. 282]. The a posteriori estimates in the scalar case [KO00, GM00, DMO07, Ohl09] allow for shocks to form. This is because in the scalar case the estimators rely on the $L_1$-contraction principle, which does not require Lipschitz continuity of any of the compared solutions and avoids an exponential dependence of the estimators on time. These estimates and the $L_1$-contraction principle as such are a consequence of the fact that there are infinitely many convex entropies in the scalar case.

3. The estimates in the scalar case are, in general, not optimal before shocks form. Due to the generality of the Kruzkov stability theory, the a posteriori estimates based on this can be seen as a worst-case scenario, allowing for arbitrarily many shocks and not accounting for any preshock regime [GM00, Rem. 2].

6. A posteriori control based on the discrete solution. In this section we use the bound given in Theorem 5.5 to construct a fully computable a posteriori quantity which depends only upon the discrete solution itself in that no reconstructions ever need to be explicitly computed.

We begin by noting that $\mathbf{R}$ can be explicitly computed locally in every cell using only information from that cell and traces from the adjacent cells. Still we would like to estimate $\|\mathbf{R}\|_{L^2}$ by quantities only involving $u_h$. There are two reasons for doing this: First we expect the new bound to be computationally cheaper. Second, we will use this new form to argue why we expect our estimator to be of optimal order.

**Lemma 6.1 (inverse inequality [Cia02, c.f.]).** For every $k \in \mathbb{N}$ there is a constant $C_{\text{inv}} > 0$ such that for any interval $J \subset \mathbb{R}$ and any $\phi \in \mathbb{P}^k(J)$ the following inequality is satisfied:

$$
\|\partial_x \phi\|_{L^2(J)} \leq \frac{C_{\text{inv}}}{|J|} \|\phi\|_{L^2(J)}.
$$

**Lemma 6.2 (a posteriori control on $\mathbf{R}$).** Let $f \in C^{p+2}(U, \mathbb{R}^d)$ and satisfy (2.14). It then holds that

$$
\|\mathbf{R}\|_{L^2(I)}^2 \leq 3(E_1 + E_2 + E_3)
$$
with

\begin{equation}
E_1 := \sum_{n=0}^{N-1} h_n L^2 \left( \| [\partial_t u_h]_n \|^2 + \| [\partial_t u_h]_{n+1} \|^2 \right),
\end{equation}

\begin{equation}
E_2 := \sum_{n=0}^{N-1} 4h_n L^2 \left( \| u_h \|^2 + \| u_h \|_{n+1}^2 \right) \left( L \frac{\| u_h \|_n + \| u_h \|_{n+1}}{h_n} + \| \partial_x u_h \|_{L_\infty(I_n)} \right) C \left( \frac{\| P \|_{L_\infty(I_n)}}{h_n} \right) + 2h_n \left( \frac{p+1}{k} \right) \left( \left( \frac{p^2}{k} \right) \frac{\| \partial_x^k u_h \|_{L_\infty(I_n)} + L \frac{p^2}{k} \| \partial_x u_h \|_{L_\infty(I_n)} \right) \right),
\end{equation}

\begin{equation}
E_3 := 2C^2 \sum_{n=0}^{N-1} \frac{1}{h_n} \left( \| u_h \|^4 + \| u_h \|^4 \right),
\end{equation}

where \( b := \| l_p \|_{W^{p+1}} + 2 \| l_{p+1} \|_{W^{p+1}}. \)

**Proof.** Recalling the definition of \( R \)

\begin{equation}
R := \partial_t \hat{u} + \partial_x \hat{f}(\hat{u}) = \partial_x \hat{f}(\hat{u}) - \partial_x \hat{f} + \partial_t \hat{u} - \partial_t u_h,
\end{equation}

we begin by splitting \( R \) into three quantities via the \( L_2 \) projection of \( \partial_x \hat{f}(\hat{u}) \), that is,

\begin{equation}
R = \partial_t (\hat{u} - u_h) + (\partial_x \hat{f}(\hat{u}) - \partial_t \hat{u}) + (\partial_x \hat{f}(\hat{u}) - \partial_t \hat{u}) =: R_1 + R_2 + R_3,
\end{equation}

and bounding each of these individually.

Forming the time derivative of (5.17) we immediately obtain

\begin{equation}
\| R_1 \|^2_{L_2(I_n)} = \| \partial_t (\hat{u} - u_h) \|^2_{L_2(I_n)} \leq L^2 h_n \left( \| [\partial_t u_h]_n \|^2 + \| [\partial_t u_h]_{n+1} \|^2 \right).
\end{equation}

For the term involving \( R_2 \) we further split the term and evaluate derivatives, giving

\begin{equation}
\| \partial_x \hat{f}(\hat{u}) - \partial_x \hat{f}(\hat{u}) \|_{L_2(I_n)} \leq \| \partial_x \hat{f}(\hat{u}) - \partial_x \hat{f}(\hat{u}) \|_{L_2(I_n)}
\end{equation}

\begin{equation}
+ \| \partial_x \hat{f}(\hat{u}) - \partial_x \hat{u} \|_{L_2(I_n)} \leq 2 \| \partial_x \hat{u} \|_{L_\infty(I_n)} C \| \hat{u} - u_h \|_{L_2(I_n)} + \| \partial_x \hat{u} \|_{L_\infty(I_n)} C \| \hat{u} - u_h \|_{L_2(I_n)}
\end{equation}

since the \( L_2 \) projection is stable and satisfies \( \| \partial_x g \|_{L_2(I_n)} \leq \| g \|_{L_2(I_n)} \) for any \( g \in L_2(\Omega) \).

In addition from (4.11) we have that

\begin{equation}
\| \partial_x \hat{f}(\hat{u}) - \partial_x u_h \|_{L_\infty(I_n)} \leq C \| \partial_x \hat{u} \|_{W^{p+1}(I_n)}.
\end{equation}
By the product rule we have inside \( I_n \)

\[
\partial_x^{p+1} (Df(u_h) \partial_x \tilde{u}) = \sum_{k=0}^{p+1} \binom{p+1}{k} (\partial_x^{k+1} \tilde{u}) (\partial_x^{p+1-k} Df(u_h))
\]

(6.9)

\[
= \sum_{k=0}^{p} \binom{p+1}{k} (\partial_x^{k+1} \tilde{u}) (\partial_x^{p+1-k} Df(u_h))
\]

as \( \tilde{u} \in \mathbb{V}_{p+1} \). Using the properties of the derivatives of the reconstruction (5.19) in

(6.9) we have that

\[
h_n^{p+1} \left\| \partial_x^{p+1} (Df(u_h) \partial_x \tilde{u}) \right\|_{L_\infty(I_n)} 
\]

\[
\leq h_n^{p+1} \sum_{k=0}^{p} \binom{p+1}{k} \left\| \partial_x^{k+1} \tilde{u} \right\|_{L_\infty(I_n)} \left\| \partial_x^{p+1-k} Df(u_h) \right\|_{L_\infty(I_n)} 
\]

(6.10)

\[
\leq \sum_{k=0}^{p} \binom{p+1}{k} \left( h_n^{p+1} \left\| \partial_x^{k+1} u_h \right\|_{L_\infty(I_n)} 
\]

\[
+ L h_n^{p-k} b_{k+1} \left( \left\| u_h \right\|_n + \left\| u_h \right\|_{n+1} \right) \left\| \partial_x^{p+1-k} Df(u_h) \right\|_{L_\infty(I_n)} \right).
\]

Inserting (6.10) into (6.8) gives

\[
\left\| \mathcal{P}_p (Df(u_h) \partial_x \tilde{u}) - Df(u_h) \partial_x \tilde{u} \right\|_{L_\infty(I_n)} 
\]

\[
\leq C_p \sum_{k=0}^{p} \binom{p+1}{k} \left( h_n^{p+1} \left\| \partial_x^{k+1} u_h \right\|_{L_\infty(I_n)} 
\]

\[
+ L h_n^{p-k} b_{k+1} \left( \left\| u_h \right\|_n + \left\| u_h \right\|_{n+1} \right) \left\| \partial_x^{p+1-k} Df(u_h) \right\|_{L_\infty(I_n)} \right).
\]

(6.11)

Therefore, we can infer from (6.7) that

\[
\left\| R_2 \right\|_{L_2(I_n)}^2 \leq 8 C_p L^2 h_n \left( \left\| u_h \right\|_n^2 + \left\| u_h \right\|_{n+1}^2 \right) \left\| \partial_x \tilde{u} \right\|_{L_\infty(I_n)} 
\]

\[
+ 2 C_p^2 h_n^{p+1} \sum_{k=0}^{p} \binom{p+1}{k} \left( h_n^{p+1} \left\| \partial_x^{k+1} u_h \right\|_{L_\infty(I_n)} 
\]

\[
+ L h_n^{p-k} b_{k+1} \left( \left\| u_h \right\|_n + \left\| u_h \right\|_{n+1} \right) \left\| \partial_x^{p+1-k} Df(u_h) \right\|_{L_\infty(I_n)} \right)^2.
\]

(6.12)

Using the fact that

\[
\left\| \partial_x \tilde{u} \right\|_{L_\infty(I_n)} \leq L \left( \left\| u_h \right\|_n + \left\| u_h \right\|_{n+1} \right) \left/ h_n \right.
\]

(6.13)

\[
\left( \left\| u_h \right\|_n + \left\| u_h \right\|_{n+1} \right) \left/ h_n \right.
\]

(6.12) implies the desired estimate for \( \left\| R_2 \right\|_{L_2(I)}^2 \).
To conclude we will estimate the term containing $R_3$. Note that $R_3 \in \mathbb{V}_p$. Using the definitions of $\hat{u}$ and $\hat{f}$ as well as integration by parts we find

$$
|R_3|^2_{L_2(I)} = \sum_{n=0}^{N-1} \int_{I_n} |R_3|^2 \, dx = \sum_{n=0}^{N-1} \int_{I_n} \left( \mathcal{P}_p (\partial_x f(\tilde{u})) - \partial_x \tilde{f} \right) \cdot R_3 \, dx \\
= \sum_{n=0}^{N-1} \int_{I_n} \left( \partial_x f(\tilde{u}) - \partial_x \tilde{f} \right) \cdot R_3 \, dx \\
= \sum_{n=0}^{N-1} \int_{I_n} \left( \partial_x f(\tilde{u}) - \partial_x f(u_h) \right) \cdot R_3 \, dx \\
- \sum_{n=0}^{N-1} \int_{I_n} \left( f(w(u_h(x_n^-), u_h(x_n^+))) \right) \, dx.
$$

(6.14)

Now upon integrating by parts, we see that

$$
|R_3|^2_{L_2(I)} = - \sum_{n=0}^{N-1} \int_{I_n} (f(\tilde{u}) - f(u_h)) \cdot \partial_x R_3 \, dx.
$$

(6.15)

Using the orthogonality property (4.1) taking $\phi = Df\mathcal{P}_0 u_h$ we have that

$$
|R_3|^2_{L_2(I)} \leq \sum_{n=0}^{N-1} \int_{I_n} \left( \frac{2}{\beta!} \int_0^1 (1-t)D^\beta f(u_h + t(\tilde{u} - u_h)) \, dt \right) \, dx
$$

$$
\leq C_{inv} C_T |u_h|_{W_1^0} \| \tilde{u} - u_h \|_{L_2(I)} \| R_3 \|_{L_2(I)}
$$

$$
+ C_{inv} C_T \left( \sum_{n=0}^{N-1} \frac{1}{h_n^2} \int_{I_n} |\tilde{u} - u_h|^4 \, dx \right)^{1/2} \| R_3 \|_{L_2(I)},
$$

by the inverse inequality (6.1), where $D^\beta f$ is the partial derivative of $f$ specified by the multiindex $\beta$. Note that $|u_h|_{W_1^0}$ in (6.16) is to be understood as $\max_{n=0, \ldots, N-1} |u_h|_{L_w(I_n)}$. Therefore,

$$
|R_3|_{L_2(I)} \leq C_{inv} C_T \left( |u_h|_{W_1^0} \| \tilde{u} - u_h \|_{L_2(I)} + \sqrt{\sum_{n=0}^{N-1} \frac{1}{h_n^2} \int_{I_n} |\tilde{u} - u_h|^4 \, dx} \right).
$$

(6.17)

In view of the boundedness of the Legendre polynomials and (5.17) this implies

$$
|R_3|_{L_2(I)} \leq C_{inv} C_T \left( |u_h|_{W_1^0} \sqrt{\sum_{n=0}^{N-1} h_n L^2 \left( |u_h|_{n}^2 + |u_h|_{n+1}^2 \right)}
$$

$$
+ \sqrt{\sum_{n=0}^{N-1} \frac{1}{h_n} L^4 \left( |u_h|_{n}^4 + |u_h|_{n+1}^4 \right) \, dx},
$$

(6.18)

concluding the proof.
Remark 6.3 (general numerical fluxes). The assumption on the numerical fluxes (3.5) was used in the above proof in order to estimate $R_3$. If we used more general numerical fluxes we would get additional contributions in the estimate (6.14) which would not be of optimal order in general. In particular, it is not sufficient for the numerical fluxes to be consistent and monotone.

Lemma 6.4 (stability of the reconstruction). Let $f \in C^{p+2}(U, \mathbb{R}^d)$ satisfy (2.10) and let $\mathbf{u}$ be an entropy solution of (2.6) with periodic boundary conditions. Then, provided $\hat{\mathbf{u}}$ takes values in $\mathfrak{D}$, for $0 \leq t \leq T$ the error between the reconstruction $\hat{\mathbf{u}}$ and $\mathbf{u}$ satisfies

\begin{equation}
\|\mathbf{u}(\cdot, t) - \hat{\mathbf{u}}(\cdot, t)\|_{L^2(I)}^2 \leq C_{2}^{-1} E(t) \exp \left( \int_0^t \frac{C_{2} L_{2}(I)}{C_{2}} \int_{\sigma} \|\partial_\sigma \hat{\mathbf{u}}(\cdot, \sigma)\|_{L^\infty(I)}^2 + \frac{C_{2}}{C_{2}} \|\partial_\tau \hat{\mathbf{u}}(\cdot, \sigma)\|_{L^\infty(I)}^2 \, d\sigma \right) \tag{6.19}
\end{equation}

with

\begin{equation}
E(t) := C_{\gamma} \|\mathbf{u}(\cdot, 0) - \hat{\mathbf{u}}(\cdot, 0)\| + \int_0^t \left( E_1 + E_2 + E_3 \right) \, dt \tag{6.20}
\end{equation}

with $E_i$ defined as in Lemma 6.2.

Proof. The proof follows by combining Lemmas 5.1 and 6.2.

Theorem 6.5 (a posteriori error estimate). Let $f \in C^{p+2}(U, \mathbb{R}^d)$ and $\mathbf{u}$ be the entropy solution of (2.6) with periodic boundary conditions. Let $\hat{\mathbf{u}}$ take values in $\mathfrak{D}$. Then, for $0 \leq t \leq T$ the error between the numerical solution $\mathbf{u}_h$ and $\mathbf{u}$ satisfies

\begin{equation}
\|\mathbf{u}(\cdot, t) - \mathbf{u}_h(\cdot, t)\|_{L^2(I)}^2 \leq C_{2}^{-1} E(t) \exp \left( \int_0^t \frac{C_{2} L_{2}(I)}{C_{2}} \int_{\sigma} \|\partial_\sigma \hat{\mathbf{u}}(\cdot, \sigma)\|_{L^\infty(I)}^2 + \frac{C_{2}}{C_{2}} \|\partial_\tau \hat{\mathbf{u}}(\cdot, \sigma)\|_{L^\infty(I)}^2 \, d\sigma \right) \tag{6.21}
\end{equation}

\[ + L_2^2 \sum_n h_n \left( \|\mathbf{u}_h(\cdot, t)\|_{n}^2 + \|\mathbf{u}_h(\cdot, t)\|_{n+1}^2 \right), \]

where $E$ is defined as in Lemma 6.4.

Proof. The proof follows from Lemmas 5.4 and 6.4.

Remark 6.6 (optimality of the estimator). Assume that the entropy solution $\mathbf{u}$ and its time derivative $\partial_t \mathbf{u}$ are $p + 1$ times continuously differentiable in space and

\begin{equation}
\|\mathbf{u} - \mathbf{u}_h\|_{L^\infty(0, T; L^2(I))} + \|\partial_t \mathbf{u} - \partial_t \mathbf{u}_h\|_{L^\infty(0, T; L^2(I))} \leq C h^{p+\gamma} \tag{6.22}
\end{equation}

for some $\gamma \in \left( \frac{1}{2}, 1 \right)$; compare to [ZS10, Thm. 5.1] for examples of schemes satisfying these rates. In that case one can verify that $\|\partial_\sigma \hat{\mathbf{u}}\|_{L^\infty(0, T; L^\infty(I))}$ remains bounded for $h$ small enough. Further, by employing arguments similar to [MN06, Rem. 3.6] adapted to our spatially discrete case we can show that

\[ \sum_n h_n \left( \|\partial_\sigma \mathbf{u}_h\|_{n}^2 + \|\partial_\tau \mathbf{u}_h\|_{n+1}^2 \right) \leq C h^{2p+2\gamma} \]

and

\[ \sum_n h_n \left( \|\mathbf{u}_h\|_{n}^2 + \|\mathbf{u}_h\|_{n+1}^2 \right) \leq C h^{2p+2\gamma}, \]

where $h = \max_n h_n$. As, in addition,

\[ \frac{1}{h_n} \left( \|\mathbf{u}_h\|_{n} + \|\mathbf{u}_h\|_{n+1} \right) \]

is expected to be bounded, we expect $E$ in (6.21) to be of order $h^{2p+2\gamma}$ and the exponential term in (6.21) to be bounded uniformly in $h$. Therefore, we claim that
our error estimator is of optimal order, for sufficiently smooth solutions. This is supported by numerical evidence in section 7.

Remark 6.7 (localizable estimates). As can be seen in [Daf10] the relative entropy stability estimate in Lemma 2.7 can be localized in the sense that there is a computable $c > 0$ depending on $\mathcal{O}$ such that for every $[a, b] \subset I$ and $t > 0$

\[(6.23) \quad \|u(\cdot, t) - v(\cdot, t)\|_{L^2([a, b])} \leq C_1 \exp(C_2 t) \|u_0 - v_0\|_{L^2([a-ct, b+ct])}\]

with $C_2$ depending on $\|\partial_x v\|_{L^\infty((x, s): x \in [a-ct, b+ct])}$. This, in particular, shows that the arguments presented above allow for the construction of localized a posteriori error estimates.

7. Numerical experiments. In this section we study the numerical behavior of the error indicators and compare this behavior with the true error on two model problems. The coding was done in MATLAB under the framework provided by [HW08].

Remark 7.1 (computed a posteriori estimators). We study the behavior of two a posteriori quantities. The first estimator, given in Theorem 6.5, is dependent only upon the discrete solution

\[(7.1) \quad \delta_t^1 := \left( L^2 \sum_n h_n \left( \|u_n(\cdot, t)\|_n^2 + \|u_{n+1}(\cdot, t)\|_{n+1}^2 \right) + C^{-1}_2 E(t) \exp(\kappa_1(u_h)) \right)^{1/2}\]

with

\[(7.2) \quad \kappa_1(u_h) := \int_0^t \max_n \left( \|\partial_x u_h\|_{L^\infty(I_n)} + LK/h_n \left( \|u_h\|_n + \|u_{h+1}\|_{n+1} \right) \right) \frac{C_2^2}{C_2} \, ds\]

and $E(t)$ given in Lemma 6.4, $L$ is the Lipschitz constant of the numerical fluxes (3.6), and $K = \max \left( \|p\|_{L^\infty(-1,1)} : \|p+1\|_{L^\infty(-1,1)} \right)$.

The second estimator, given in Theorem 5.5, is determined by computing the discrete reconstruction operators and is

\[(7.3) \quad \delta_t^2 := \left( \|u_h - \tilde{u}\|_{L^2(I)}^2 + 2C^{-1}_2 \left( \|R\|_{L^2(I \times (0,t))}^2 + C_2^2 \|u_0 - \tilde{u}_0\|_{L^2(I)}^2 \right) \exp(\kappa_2(\tilde{u})) \right)^{1/2}\]

where

\[(7.4) \quad \kappa_2(\tilde{u}) := \int_0^t \frac{C_2^2 \|\partial_x \tilde{u}(\cdot, s)\|_{L^\infty(I)} + C_2^2}{C_2} \, ds\]

and $R$ is given in (5.2).

The constants for both estimators are readily computable as detailed in Remark 7.4 for each of the test cases; as such both quantities are estimators.

Definition 7.2 (estimated order of convergence). Given two sequences $a(i)$ and $h(i) \searrow 0$, we define estimated order of convergence to be the local slope of the log $a(i)$ versus log $h(i)$ curve, i.e.,

\[(7.5) \quad \text{EOC}(a, h; i) := \frac{\log(a(i+1)/a(i))}{\log(h(i+1)/h(i))}\]
A POSTERIORI ANALYSIS FOR CONSERVATION LAWS

Fig. 1. Numerical results for the dGRK scheme with Engquist–Osher fluxes approximating (7.8) the solution to Burgers’ equation. In each subfigure we plot both estimators, $E_1, E_2$, together with the error, $e$, on a logarithmic scale against time. We also show the estimated orders of convergence and the effectivity indices over time.

Definition 7.3 (effectivity index). The main tool deciding the quality of an estimator is the effectivity index, which is the ratio of the error and the estimator, i.e.,

$$EI^i(t_n) := \frac{\max_{t \in [0,t_n]} E^i_t}{\|u - u_h\|_{L_{\infty}(0,t_n;L_2(S^1))}}$$

for $i = 1, 2$. 

(a) Results for $P^0$ elements. Notice both estimators are robust; however, $E_2$ has a slightly lower effectivity index.

(b) Results for $P^1$ elements. Notice both estimators are robust; however, $E_2$ has a slightly lower effectivity index.
Fig. 2. Numerical results for the dGRK scheme with Engquist–Osher fluxes approximating the solution to Burgers’ equation with initial condition $u(x, 0) = -\sin(x)$. We study the behavior of the estimators as the solution approaches blowup at $t = 1$. Notice that before shock time the estimators behave robustly. As shock time approaches the estimators blow up at a rate which increases as the meshsize decreases.

Remark 7.4 (computation of constants). The constants appearing in the estimators $\delta_i^n$ are readily computable; $C_{\eta}$ and $C_{\tau}$ represent the absolute values of the minimum and maximum eigenvalues of $D^2\eta$ on $\mathcal{O}$. In addition $C_{\tau} := (\sum_i C_{\eta}^i)^{1/2}$, where $C_{\eta}^i$ is an upper bound for the absolute values of the eigenvalues of the $i$th component of $f$.

In both tests below we choose an explicit fourth order Runge–Kutta method for the temporal discretization. To test the asymptotic behavior of the estimators given in Theorems 5.5 and 6.5 we use a uniform timestep and uniform meshes that are fixed with respect to time. Hence for each test we have $V^n = V^0 = V$ and $\tau_n = \tau(h)$ for all $n \in [1 : N]$. We fix the polynomial degree $p$ and two parameters $k, c$ and then compute a sequence of solutions with $h = h(i) = 2^{-i}$, and $\tau = ch^k$ for a sequence of refinement levels $i = l, \ldots, L$.

7.1. Test 1: The scalar case—invviscid Burgers’ equation. We conduct a benchmarking experiment using the invviscid (scalar) Burgers’ equation

$$\partial_t u + \partial_x \left( \frac{u^2}{2} \right) = 0.$$  

Using an initial condition $u(x, 0) = -\sin(x)$ over an interval $I = [-\pi, \pi]$. It can be verified that, before shock formation, the exact solution can be represented by an infinite sum of Bessel functions, that is,

$$u(x, t) = -2 \sum_{k=1}^{\infty} \frac{J_k(kt)}{kt} \sin(kx),$$

where $J_k$ denotes the $k$th Bessel function. Note this is a decaying sequence, hence we may approximate the solution by taking a truncation of this series.

We discretize the problem (7.7) using the dG scheme (3.3) together with Engquist–Osher type fluxes. These fluxes satisfy the assumptions (3.5)–(3.6) as shown in Remark 3.1.

For this problem we may take $\eta(u) = \frac{1}{2}u^2$, and hence it is readily verified that $C_{\eta} = C_{\tau} = C_{\tau} = 1$. 

\[72x710\]j. GiesseLMann, C. Makridakis, and t. pryer
(a) Results for $P^0$ elements. Notice both estimators are robust; however, $\delta^2_t$ has a slightly lower effectivity index.

(b) Results for $P^1$ elements. Notice both estimators are robust; however, $\delta^2_t$ has a slightly lower effectivity index.

Fig. 3. Numerical results for the dGRK scheme with Roe fluxes approximating (7.8) the solution to the $p$-system. In each subfigure we plot both estimators, $\delta^1_t, \delta^2_t$, together with the error, $e$, on a logarithmic scale against time. We also show the estimated orders of convergence and the effectivity indices over time.

In Figures 1(a) and 1(b) we examine the asymptotic behavior of the estimators and error for the solution given by (7.8). In Figure 2 we study the behavior of the estimators when a shock forms.
7.2. Test 2: The system case—the $p$-system. In this case we conduct some benchmarking using the $p$-system, given by

\begin{align}
0 &= \partial_t u - \partial_x v, \\
0 &= \partial_t v - \partial_x (p(u)).
\end{align}

(7.9)

We choose an initial condition $u(x, 0) = -v(x, 0) = \exp \left( -10|x|^2 \right)$ over an interval $I = [-5, 5]$.

We discretize (7.9) using the dG scheme (3.3) with a Roe flux (as described in Remark 3.1). This class of fluxes satisfies the assumption on the fluxes (3.6) assuming $p$ is surjective. We take $p(u) = u^3 + u$.

For this problem we have $\eta(u, v) = W(u) + \frac{1}{2}v^2$, where $W$ is a primitive of $p$. Thus the eigenvalues are $3u^2 + 1$ and $1$. Suppose $u \in [-a, a]$; then we have that $C_\eta = 1$ and $C_f = 1 + 3a^2$. Similarly, $C_{\eta} = (6a)^{1/2}$.

To generate an exact solution to this problem we introduce a source term into the second equation in (7.9). We choose the source term in such a way that

\begin{align}
(7.10) \\
(7.11)
\end{align}

The results are summarized in Figures 3(a) and 3(b).

8. Conclusion. In this work we introduced a methodology for deriving a posteriori bounds for semidiscrete discontinuous Galerkin schemes approximating systems of hyperbolic conservation laws. The methodology is applicable whenever solutions to the system remain Lipschitz continuous and we have numerically demonstrated that the a posteriori estimator is robust in this case. When shocks develop the relative entropy stability theory breaks down and, although the estimator remains computable, it contains, as expected, a constant that blows up as the meshsize goes to zero. The extension of this approach to the postshock case remains a challenging problem.

REFERENCES


A POSTERIORI ANALYSIS FOR CONSERVATION LAWS


