SIAM J. NUMER. ANAL.
Vol. 54, No. 6, pp. 3523–3549

A POSTERIORI ANALYSIS OF FULLY DISCRETE METHOD OF LINES DISCONTINUOUS GALERKIN SCHEMES FOR SYSTEMS OF CONSERVATION LAWS*

ANDREAS DEDNER† AND JAN GIESSELMANN‡

Abstract. We present reliable a posteriori estimators for some fully discrete schemes applied to nonlinear systems of hyperbolic conservation laws in one space dimension with strictly convex entropy. The schemes are based on a method of lines approach combining discontinuous Galerkin spatial discretization with single- or multi-step methods in time. The construction of the estimators requires a reconstruction in time for which we present a very general framework first for ODEs and then applying the approach to conservation laws. The reconstruction does not depend on the actual method used for evolving the solution in time. Most importantly, in addition to implicit methods it also covers the wide range of explicit methods typically used to solve conservation laws. For the spatial discretization, we allow for standard choices of numerical fluxes. We use *reconstructions* of the discrete solution together with the *relative entropy* stability framework, which leads to error control in the case of smooth solutions. We study under which conditions on the numerical flux the estimate is of optimal order preshock. While the estimator we derive is computable and valid postshock for fixed mesh size, it will blow up as the mesh size tends to zero. This is due to a breakdown of the relative entropy framework when discontinuities develop. We conclude with some numerical benchmarking to test the robustness of the derived estimator.

Key words. conservation laws, a posteriori error estimates, method of lines, discontinuous Galerkin, relative entropy

AMS subject classifications. 65M60, 65M15, 35F50, 35L45, 35L65

DOI. 10.1137/15M1046265

1. Introduction.

Systems of hyperbolic balance laws are widely used in continuum mechanical modeling of processes in which higher order effects such as diffusion and dispersion can be neglected, with the Euler and the shallow water equations being prominent examples. A particular feature of these equations is the breakdown of smooth solutions to initial value problems for generic (smooth) initial data after finite time. After this "shock formation" discontinuous weak solutions are considered and attention is restricted to those satisfying a so-called entropy inequality. Due to the interest in discontinuous solutions, finite volume and discontinuous Galerkin (DG) spatial discretizations in space are state of the art [15, 23, 29, 18]. In the case of nonlinear systems, the number of theoretical results backing up these schemes is quite limited. A posteriori results for systems were derived in [26, 27] for front tracking and Glimm's schemes; see also [22]. A priori estimates for fully discrete Runge–Kutta DG schemes were obtained in [37] and an a posteriori error estimator for semidiscrete schemes was introduced in [14]. Other a priori and a posteriori results using the relative entropy method include [2, 20, 21]. In [17] the authors derive a poste-

*Received by the editors November 2, 2015; accepted for publication (in revised form) August 29, 2016; published electronically December 6, 2016.
http://www.siam.org/journals/sinum/54-6/M104626.html
Funding: The work of the first author was supported by the Royal Society. The work of the second author was partially supported by the German Research Foundation (DFG) via SFB TRR 75 "Tropfendynamische Prozesse unter extremen Umgebungsbedingungen."
†Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (a.s.dedner@warwick.ac.uk).
‡Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, D-70563 Stuttgart, Germany (jan.giesselmann@mathematik.uni-stuttgart.de).

riori estimates for space-time DG schemes in a goal oriented framework, provided certain dual problems are well-posed. All these results deal mainly with the preshock case. Arguably, this is due to the following reason: the well-posedness of generic initial value problems for (multidimensional) systems of hyperbolic conservation laws is quite open. While uniqueness of entropy solutions was expected by many researchers for a long time, it has been shown in recent years [4, 7] that entropy solutions of nonlinear multidimensional systems are not unique, in general. This severely restricts the range of cases in which a priori error estimates or convergent error estimators can be expected. This is in contrast to the situation for scalar hyperbolic problems, for which a priori convergence rates [12] and convergent a posteriori error estimators [16, 24, 8] are available postshock, based on a more discriminatory notion of the entropy solution [25]. Still, the simulation of multidimensional hyperbolic balance laws is an important field, and the numerically obtained results coincide well with experimental data. In the simulations, error indicators based on either entropy dissipation of the numerical solution [32] or nodal super convergence [1] are used. Our goal is to complement these results with a rigorous error estimator which, for certain types of numerical fluxes, is of optimal order, i.e., of the order of the true error, as long as the solution is smooth. In the case where there is no Lipschitz continuous solution, but (possibly several) discontinuous entropy solutions, our estimator is still an upper bound for the error of the method, but it does not converge to zero under mesh refinement.

Our work is based on the results for semi- (spatially) -discrete methods in [14] which we extend in two directions. Firstly, the results obtained here account for fully discrete, e.g., Runge–Kutta DG type, schemes, and secondly, we treat a larger class of numerical fluxes than was treated in [14]. Our results are optimal for a large class of central fluxes (e.g., of Richtmyer or Lax–Wendroff type) augmented with stabilization approaches such as artificial viscosity or flux limiting [28, 36]. Furthermore, we also prove that the error estimator is optimal for Roe-type numerical fluxes for systems of conservation laws. It is well known that only certain (stable) temporal discretizations are useful for the discretization of hyperbolic conservation laws and that for explicit methods the time step needs to satisfy a CFL condition. Our a posteriori analysis provides upper bounds for the errors for unstable as well as for stable methods. Indeed, for unstable methods the error will be large, which will be reflected by the error estimator.

Our work is based on reconstructions in space and time (see [30] for a general exposition on the idea of reconstruction-based error estimators.) and the relative entropy stability theory, going back to [5, 10]. While the idea of reconstruction-based a posteriori error estimates was extensively employed for implicit methods (see, e.g., [31]), it has not been used for explicit methods before. Thus, we will describe our temporal reconstruction approach for general systems of ODEs first. Our approach for reconstruction in the context of ODEs (which might be semidiscretizations of PDEs) differs from other error estimation approaches (e.g., [19, 31]) by being based on Hermite interpolation and by using information from old time steps. An advantage of this approach is that the reconstruction does not depend on the time stepping method used (covering, for example, both general IMEX, type Runge–Kutta or multistep methods). The spatial reconstruction follows the approach first described in [14]. As mentioned above, we extend the class of methods for which optimal convergence of the estimate can be shown. Furthermore, we prove that within the class of reconstructions considered here, our restriction on the flux is not only sufficient for optimal convergence but also necessary. We also show for a Roe-type flux that whether the error estimator is suboptimal or optimal depends on the choice of reconstruction.

The layout of the rest of this work is as follows. In section 2 we introduce a reconstruction approach for general single- or multi-step discretizations of ODEs and show that it leads to residuals of optimal order, i.e., of the same order as the error. This approach is applied to fully discrete DG schemes for systems of hyperbolic conservation laws in one space dimension endowed with a strictly convex entropy in section 3. We present numerical experiments in section 4.

2. Reconstructions and error estimates for ODEs. In this section we introduce reconstructions for general single- or multi-step methods approximating initial value problems for first order systems of ODEs. The general set-up will be an initial value problem

$$\mathrm{d}_t \mathbf{u} = \mathbf{f}(t, \mathbf{u}) \ \text{ on } (0,T), \ \mathbf{u}(0) = \mathbf{u}_0 \in \mathbb{R}^m \tag{2.1}$$

for some finite time $T > 0$ and a given map $\mathbf{f} : [0,T] \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$. We will usually assume that $\mathbf{f}$ is at least Lipschitz but we will specify our regularity assumptions on $\mathbf{f}$ later.

We assume that we have a numerical method which provides us with values $\{\mathbf{u}^n\}_{n=0}^N$ approximating the exact solution at times $\{t_n\}_{n=0}^N$ with $0 = t_0 < t_1 < \cdots < t_N = T$. Our aim is to construct a continuous reconstruction of the solution to the ODE based on the values $\{\mathbf{u}^n\}_{n=0}^N$ but not on the actual method used to obtain them. Our reconstruction approach is based on Hermite (polynomial) interpolation and the order of the employed polynomials depends on the convergence order of the method.

Also note that, while $\{\mathbf{u}^n\}_{n=0}^N$ are good approximations of the exact solution at the corresponding points in time for convergent methods, this is usually not true for intermediate values $\{\mathbf{u}^{n,i}\}_{n=0}^{N-1}$ which are frequently computed during, e.g., Runge–Kutta time steps. It is therefore in general a nontrivial task to include these in the reconstruction.

2.1. Reconstruction. To define our reconstruction we need to introduce some notation. By $\mathbb{P}_q(I,V)$ we denote the set of polynomials of degree at most q on some interval I with values in some vector space V, and $\mathbb{V}_q$ denotes a space of (possibly discontinuous) functions which are piecewise polynomials of degree less than or equal to q, i.e.,

$$\mathbb{V}_q := \{\mathbf{w} : [0,T] \to \mathbb{R}^m \,:\, \mathbf{w}|_{(t_{n-1},t_n)} \in \mathbb{P}_q((t_{n-1},t_n),\mathbb{R}^m) \text{ for } n = 1,\dots,N\}. \tag{2.2}$$

For any n and $\mathbf{w} \in \mathbb{V}_q$, we define traces by

$$\mathbf{w}(t_n^{\pm}) := \lim_{s \searrow 0} \mathbf{w}(t_n \pm s) \tag{2.3}$$

and element wise derivatives $\mathrm{d}_t^e \mathbf{w} \in \mathbb{V}_{q-1}$ by

$$(\mathrm{d}_t^e \mathbf{w})|_{(t_n,t_{n+1})} = \mathrm{d}_t(\mathbf{w}|_{(t_n,t_{n+1})}). \tag{2.4}$$

Let $\{\mathbf{u}^n\}_{n=0}^N$ denote the approximations of the solution of (2.1) at times $\{t_n\}_{n=0}^N$ computed using some single- or multi-step method. We will define the reconstruction $\hat{\mathbf{u}}$ as a C^0 or even C^1-function which is piece wise polynomial and whose polynomial degree matches the convergence order of the method. To define $\hat{\mathbf{u}}|_{[t_n,t_{n+1}]}$ the information $\mathbf{u}^n, \mathbf{u}^{n+1}, \mathbf{f}(\mathbf{u}^n), \mathbf{f}(\mathbf{u}^{n+1})$ is readily available, but for polynomial degrees above 3 we need additional conditions. Let $\hat{\mathbf{u}}^n$ denote the polynomial which coincides with

$\hat{\mathbf{u}}$ on $[t_n, t_{n+1}]$. Note that $\hat{\mathbf{u}}^n$ differs from $\hat{\mathbf{u}}|_{[t_n,t_{n+1}]}$ by being defined on all of $\mathbb{R}$. Then, $\hat{\mathbf{u}}|_{[t_n,t_{n+1}]}$ can be obtained by prescribing values of $\hat{\mathbf{u}}^n$, $(\hat{\mathbf{u}}^n)'$ at additional points, by prescribing additional derivatives of $\hat{\mathbf{u}}^n$ at t_n and t_{n+1}, or by a combination of both approaches.

For $(p, d, r) \in \mathbb{N}_0^3$, where $\mathbb{N}_0$ denotes the set containing all natural numbers and zero, we denote by $H(p, d, r)$ the reconstruction which fixes the value and the first $d+1$ derivatives of $\hat{\mathbf{u}}^n$ at $t_{n-p}, \dots, t_n$ and the value and the first $r+1$ derivatives of $\hat{\mathbf{u}}^n$ at t_{n+1}.

Before we make this more precise, let us note that we can express higher order derivatives of the solution $\mathbf{u}$ to (2.1) by evaluating $\mathbf{f}$ and its derivatives, e.g.,

$$\mathrm{d}_t^2\, \mathbf{u}(t_n) = \partial_t \mathbf{f}(t_n, \mathbf{u}(t_n)) + \mathrm{D}\, \mathbf{f}(t_n, \mathbf{u}(t_n)) \mathbf{f}(t_n, \mathbf{u}(t_n)), \tag{2.5}$$

where $\mathrm{D}\, \mathbf{f}$ is the Jacobian of $\mathbf{f}$ with respect to $\mathbf{u}$, provided $\mathbf{f}$ is sufficiently regular. We denote the corresponding expression for $\mathrm{d}_t^k\, \mathbf{u}(t_n)$ by $\mathbf{f}_k^n(\mathbf{u}(t_n))$. Note that instead of the exact solution $\mathbf{u}(t_n)$ we may also insert the approximation $\mathbf{u}^n$ into the expression (2.5).

Remark 1 (Regularity of $\mathbf{f}$). Subsequently assume $\mathbf{f} \in C^{\max\{d,r\}}((0,T)\times\mathbb{R}^m, \mathbb{R}^m)$. This is the amount of regularity required to define our reconstruction. It is not sufficient for the Runge–Kutta method to have (provably) an error of order $\mathcal{O}(\tau^q)$, where τ is the maximal time step size. When investigating the optimality of the residual in section 2.2 we will indeed require more regularity of $\mathbf{f}$.

Now we are in a position to define our reconstruction.

Definition 2 (Reconstruction for ODEs). *The $H(p, d, r)$ reconstruction $\hat{\mathbf{u}} \in \mathbb{V}_q$ with $q = (d+2)(p+1) + r + 1$ is determined by*

$$\begin{aligned} \hat{\mathbf{u}}|_{[t_n,t_{n+1}]} &= \hat{\mathbf{u}}^n|_{[t_n,t_{n+1}]} \text{ for } n = 0, \dots, N-1, \\ (\mathrm{d}_t)^k \hat{\mathbf{u}}^n(t_j) &= \mathbf{f}_k^j(\mathbf{u}^j), \text{ for } k = 0, \dots, d+1 \text{ and } j = n-p, \dots, n, \\ (\mathrm{d}_t)^k \hat{\mathbf{u}}^n(t_{n+1}) &= \mathbf{f}_k^{n+1}(\mathbf{u}^{n+1}) \text{ for } k = 0, \dots, r+1. \end{aligned} \tag{2.6}$$

Remark 3 (Start-up). Note that (strictly speaking) the $H(p, d, r)$ reconstruction is not defined on $[t_0, t_p]$. However, computing the numerical solution for the first $p+1$ time steps, we may use conditions at $\{t_0, \dots, t_p\}$ to define $\hat{\mathbf{u}}|_{[t_0,t_p]}$ in an analogous way.

By standard results on osculating polynomials and Hermite interpolation (see, e.g., [35]), we know the following.

Lemma 4 (Properties of reconstruction). *Any $H(p, d, r)$ reconstruction $\hat{\mathbf{u}} \in \mathbb{V}_q$ with $q = (d+2)(p+1)+r+1$ as given in Definition 2 is well defined, computable and W_∞^1 in time. For higher values of d, r we even have that $\hat{\mathbf{u}}$ is $\min\{d+1, r+1\}$-times continuously differentiable.*

Remark 5 (Particular methods). In the work at hand we will restrict our attention to two classes of methods: either methods of type $H(p, 0, 0)$ or methods of types $H(0, d, d)$ and $H(0, d, d-1)$. Both methods need the same amount of extra storage to arrive at the same order. Methods of type $H(p, 0, 0)$ have the advantage that no derivatives of $\mathbf{f}$ need to be evaluated or approximated. In the case where, derivatives of $\mathbf{f}$ are explicitly known and can be cheaply evaluated, $H(0, d, d)$ has the advantage that no information from old time steps has to be accessed. We choose the numbers of conditions imposed at t_n and t_{n+1} to be equal, so that all the information created at t_{n+1} can be reused when $\hat{\mathbf{u}}|_{[t_{n+1},t_{n+2}]}$ is computed.

Remark 6 (Derivatives of $\mathbf{f}$). In many cases of importance, e.g., $\mathbf{f}$ being a spatial semidiscretization of a system of hyperbolic conservation laws, the explicit computation of derivatives of $\mathbf{f}$ might be infeasible or numerically expensive. Thus, for $H(0,d,d)$ with $d \geq 1$ it seems interesting to replace, e.g., $\mathrm{D}\,\mathbf{f}(\cdot,\mathbf{u})\mathbf{f}(\cdot,\mathbf{u})$ by an approximation thereof. We will elaborate upon this in Remark 12 and section 2.3.

Using any of the methods $H(p,d,r)$, the reconstruction $\hat{\mathbf{u}}$ is explicitly (and locally) computable, continuous, and piecewise polynomial, thus

$$\mathbf{R} := \mathrm{d}_t\,\hat{\mathbf{u}} - \mathbf{f}(\cdot,\hat{\mathbf{u}}) \in L_\infty(0,T) \tag{2.7}$$

is computable. Therefore, standard stability theory for ODEs implies our first a posteriori results.

LEMMA 7 (A posteriori estimates for ODEs). *Let* (2.1) *have an exact solution* $\mathbf{u}$ *and let* $\mathbf{f}$ *be Lipschitz with respect to* $\mathbf{u}$*, with Lipschitz constant* L*, on a neighborhood of the set of values taken by* $\mathbf{u}$ *and* $\hat{\mathbf{u}}$*. Then,*

$$\|\mathbf{u}-\hat{\mathbf{u}}\|_{L_\infty(0,T)} \leq (|\mathbf{u}_0-\hat{\mathbf{u}}(0)| + \|\mathbf{R}\|_{L_1(0,T)})e^{LT} \tag{2.8}$$

and

$$\|\mathbf{u}-\hat{\mathbf{u}}\|^2_{L_2(0,T)} \leq (|\mathbf{u}_0-\hat{\mathbf{u}}(0)|^2 + \|\mathbf{R}\|^2_{L_2(0,T)})e^{(2L+1)T} \tag{2.9}$$

with $\mathbf{R}$ *being defined in* (2.7).

Proof. Note that (2.1) and (2.7) imply

$$(\mathbf{u}-\hat{\mathbf{u}})_t = \mathbf{f}(\mathbf{u}) - \mathbf{f}(\hat{\mathbf{u}}) - \mathbf{R}. \tag{2.10}$$

Let $\operatorname{sign}(\mathbf{u}-\hat{\mathbf{u}})$ denote a vector whose ith component is the sign of the ith component of $\mathbf{u}-\hat{\mathbf{u}}$. We test (2.10) by $\operatorname{sign}(\mathbf{u}-\hat{\mathbf{u}})$ and obtain

$$\mathrm{d}_t\,|\mathbf{u}-\hat{\mathbf{u}}| = \big(\mathbf{f}(\mathbf{u})-\mathbf{f}(\hat{\mathbf{u}})\big)\operatorname{sign}(\mathbf{u}-\hat{\mathbf{u}}) - \mathbf{R}\operatorname{sign}(\mathbf{u}-\hat{\mathbf{u}}) \leq L|\mathbf{u}-\hat{\mathbf{u}}| + |\mathbf{R}|. \tag{2.11}$$

We obtain the first assertion of the lemma by applying Gronwall's lemma to (2.11). Testing (2.11) with $(\mathbf{u}-\hat{\mathbf{u}})$ implies

$$\frac{1}{2}\,\mathrm{d}_t\,|\mathbf{u}-\hat{\mathbf{u}}|^2 = \big(\mathbf{f}(\mathbf{u})-\mathbf{f}(\hat{\mathbf{u}})\big)(\mathbf{u}-\hat{\mathbf{u}}) - \mathbf{R}(\mathbf{u}-\hat{\mathbf{u}}) \leq \Big(L+\frac{1}{2}\Big)|\mathbf{u}-\hat{\mathbf{u}}|^2 + \frac{1}{2}|\mathbf{R}|^2. \tag{2.12}$$

By applying Gronwall's lemma to (2.12), we obtain

$$|\mathbf{u}(t)-\hat{\mathbf{u}}(t)|^2 \leq (|\mathbf{u}_0-\hat{\mathbf{u}}(0)|^2 + \|\mathbf{R}\|^2_{L_2(0,T)})e^{(2L+1)t} \tag{2.13}$$

for all $0 \leq t \leq T$. The second assertion of the lemma follows by integrating (2.13). ∎

Remark 8 (Influence of the Lipschitz constant of $\mathbf{f}$). The appearance of the Lipschitz constant of $\mathbf{f}$ in the error estimators in Lemma 7 cannot be avoided for general right-hand sides $\mathbf{f}$. However, for many particular right-hand sides it can be avoided, because better stability results for the underlying ODE are available. A particular example are right-hand sides stemming from the spatial discretization of systems of hyperbolic conservation laws, which we will investigate in section 3.

2.2. Optimality of the residual. In this section we investigate how the order of $\mathbf{R}$ is related to the order of the error of the numerical approximations $\mathbf{u}^n$. To this end we restrict ourselves to an equidistant time step $\tau > 0$. We also assume that there exists a constant $L > 1$ such that for all $k, l \in \mathbb{N}_0$,

$$\left|\partial_t^l \mathrm{D}^k \mathbf{f}(t,\mathbf{y}) - \partial_t^l \mathrm{D}^k \mathbf{f}(t,\bar{\mathbf{y}})\right| \leq L^{k+1} |\mathbf{y} - \bar{\mathbf{y}}| \quad \forall\, t \in [0,T],\ \mathbf{y}, \bar{\mathbf{y}} \in \mathbb{R}^m, \tag{2.14}$$

where on the left-hand side $|\cdot|$ denotes an appropriate norm on $\mathbb{R}^{m^{k+1}}$.

Remark 9 (Parameter dependence). Note that we explicitly keep track of the dependence of the subsequent estimates on L from (2.14), while we suppress all other constants using the Landau $\mathcal{O}$ notation. This is due to the fact that we will apply the results obtained here to spatial semidiscretizations of hyperbolic conservation laws in section 3. In that case the Lipschitz constant L depends on the spatial mesh width, which for practical computations is of the same order of magnitude as the time step size.

For a time integration method of order q we consider a reconstruction of type $H(0,d,r)$ such that $q = d + r + 3$ and $r \in \{d, d-1\}$. In this way the polynomial degree of the reconstruction coincides with the order of the method. The analysis for residuals stemming from reconstructions of type $H(p,0,0)$ is analogous.

For estimating the residual we make use of two auxiliary functions. Firstly, for $n = 0, \ldots, N-1$, we denote by $\tilde{\mathbf{u}}^n$ the exact solution to the initial value problem

$$\mathrm{d}_t\, \tilde{\mathbf{u}}^n(t) = \mathbf{f}(t, \tilde{\mathbf{u}}^n(t)) \quad \text{on } (t_n, t_{n+1}), \quad \tilde{\mathbf{u}}^n(t_n) = \mathbf{u}^n. \tag{2.15}$$

We assume from now on that $\mathbf{f}$ is indeed regular enough for the method at hand to be convergent of order q. In particular, we assume consistency errors to be of order $q+1$, i.e., there exists a constant $C = C(q) > 0$ such that

$$\left|\tilde{\mathbf{u}}^n(t_{n+1}) - \mathbf{u}^{n+1}\right| \leq C \left\| d_t^{q+1} \tilde{\mathbf{u}}^n \right\|_\infty \tau^{q+1}. \tag{2.16}$$

Secondly, for $n = 0, \ldots, N-1$, let $\mathbf{h}^n \in \mathbb{P}_q((t_n, t_{n+1}), \mathbb{R}^m)$ be the Hermite interpolation of $\tilde{\mathbf{u}}^n$, i.e.,

$$\begin{aligned} (\mathrm{d}_t^e)^k \mathbf{h}^n(t_n^+) &= \mathbf{f}_k^n(\tilde{\mathbf{u}}^n(t_n)) \text{ for } k = 0, \ldots, d+1, \\ (\mathrm{d}_t^e)^k \mathbf{h}^n(t_{n+1}^-) &= \mathbf{f}_k^{n+1}(\tilde{\mathbf{u}}^n(t_{n+1})) \text{ for } k = 0, \ldots, r+1. \end{aligned} \tag{2.17}$$

By standard results on Hermite interpolation [35] there exists a constant $C = C(q) > 0$ such that

$$\|\tilde{\mathbf{u}}^n - \mathbf{h}^n\|_{L_\infty(t_n,t_{n+1})} + \tau \|\tilde{\mathbf{u}}^n - \mathbf{h}^n\|_{W^1_\infty(t_n,t_{n+1})} = C \left\| d_t^{q+1} \tilde{\mathbf{u}}^n \right\|_\infty \tau^{q+1} \tag{2.18}$$

for $n = 0, \ldots, N-1$.

Let us write $\mathbf{R}^n$ instead of $\mathbf{R}|_{(t_n,t_{n+1})}$ for brevity. Because of (2.15) we may rewrite (2.7) as

$$\begin{aligned} \mathbf{R}^n &= \mathrm{d}_t(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n) + \mathrm{d}_t(\mathbf{h}^n - \tilde{\mathbf{u}}^n) \\ &\quad - (\mathbf{f}(\cdot, \hat{\mathbf{u}}|_{(t_n,t_{n+1})}) - \mathbf{f}(\cdot, \mathbf{h}^n)) - (\mathbf{f}(\cdot, \mathbf{h}^n) - \mathbf{f}(\cdot, \tilde{\mathbf{u}}^n)) =: \mathbf{R}_1^n + \mathbf{R}_2^n + \mathbf{R}_3^n + \mathbf{R}_4^n \end{aligned} \tag{2.19}$$

and (2.18) immediately implies that

$$\|\mathbf{R}_2^n\|_{L_\infty(0,T)} + \|\mathbf{R}_4^n\|_{L_\infty(0,T)} \leq C \left\| d_t^{q+1} \tilde{\mathbf{u}}^n \right\|_\infty \left(\tau^q + L\tau^{q+1}\right). \tag{2.20}$$

Regarding the estimates of $\mathbf{R}_1^n, \mathbf{R}_3^n$, we make use of the fact that $\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n \in \mathbb{P}_q((t_n,t_{n+1}),\mathbb{R}^m)$ satisfies

(2.21)
$$\begin{aligned}(\mathrm{d}_t^e)^k(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n)(t_n^+) &= 0 \text{ for } k = 0,\dots,d+1,\\ (\mathrm{d}_t^e)^k(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n)(t_{n+1}^-) &= \mathbf{f}_k^{n+1}(\mathbf{u}^{n+1}) - \mathbf{f}_k^{n+1}(\tilde{\mathbf{u}}^n(t_{n+1})) \text{ for } k = 0,\dots,r+1,\end{aligned}$$

and due to (2.14) and consistency of the underlying method (2.16),

$$\left|\mathbf{f}_k^{n+1}(\mathbf{u}^{n+1}) - \mathbf{f}_k^{n+1}(\tilde{\mathbf{u}}^n(t_{n+1}))\right| \le L^k C \left\| d_t^{q+1}\tilde{\mathbf{u}}^n \right\|_\infty \tau^{q+1} \text{ for } k = 0,\dots,r+1. \tag{2.22}$$

Lemma 10 (Stability of Hermite interpolation). *Let $n \in \{0,\dots,N-1\}$ and $\mathbf{v} \in \mathbb{P}_q((t_n,t_{n+1}),\mathbb{R}^m)$ satisfy*

$$\begin{aligned}(\mathrm{d}_t^e)^k\mathbf{v}(t_n^+) &= a_k \text{ for } k = 0,\dots,d+1,\\ (\mathrm{d}_t^e)^k\mathbf{v}(t_{n+1}^-) &= b_k \text{ for } k = 0,\dots,r+1\end{aligned} \tag{2.23}$$

with sequences $\{a_k\}_{k=0}^{d+1}, \{b_k\}_{k=0}^{r+1}$ satisfying

$$a_k, b_k \le cL^k\tau^{q+1}$$

for some $c, L, \tau > 0$. Then, it holds that

$$\|\mathbf{v}\|_{L_\infty(t_n,t_{n+1})} \le cC\sum_{k=0}^{q} L^k\tau^{q+1+k} \tag{2.24}$$

and

$$\|\mathrm{d}_t\,\mathbf{v}\|_{L_\infty(t_n,t_{n+1})} \le cC\sum_{k=0}^{q} L^k\tau^{q+k}, \tag{2.25}$$

where C is a constant only depending on q.

Proof. The functions $\{\psi_i\}_{i=0}^q$ with

$$\begin{aligned}\psi_i(t) &= (t-t_n)^i \quad \text{for } i = 0,\dots,d+1,\\ \psi_i(t) &= (t-t_n)^{d+1}(t-t_{n+1})^{i-d-1} \quad \text{for } i = d+2,\dots,q\end{aligned}$$

form a basis of $\mathbb{P}_q((t_n,t_{n+1}),\mathbb{R})$ and obviously

$$\|\psi_i\|_{L_\infty(t_n,t_{n+1})} = \mathcal{O}(\tau^i); \quad \|\mathrm{d}_t\,\psi_i\|_{L_\infty(t_n,t_{n+1})} = \mathcal{O}(\tau^{i-1}). \tag{2.26}$$

In this basis the coefficients of $\mathbf{v}$ are given by divided differences, i.e.,

$$\mathbf{v} = \sum_{i=0}^{d+1} \bar{\mathbf{v}}[\underbrace{t_n,\dots,t_n}_{(i+1)\text{-times}}]\psi_i + \sum_{i=d+2}^{q} \bar{\mathbf{v}}[\underbrace{t_n,\dots,t_n}_{(d+2)\text{-times}},\underbrace{t_{n+1},\dots,t_{n+1}}_{(i-d-1)\text{-times}}]\psi_i \tag{2.27}$$

with

$$\begin{aligned}\bar{\mathbf{v}}[\underbrace{t_n,\dots,t_n}_{i\text{-times}}] &= a_{i-1} \quad \text{for } i = 1,\dots,d+2,\\ \bar{\mathbf{v}}[\underbrace{t_{n+1},\dots,t_{n+1}}_{j\text{-times}}] &= b_{j-1} \quad \text{for } j = 1,\dots,r+2\end{aligned} \tag{2.28}$$

and

$$\bar{\mathbf{v}}[\overbrace{t_n,\dots,t_n}^{i\text{-times}},\overbrace{t_{n+1},\dots,t_{n+1}}^{j\text{-times}}] = \frac{\bar{\mathbf{v}}[\overbrace{t_n,\dots,t_n}^{(i-1)\text{-times}},\overbrace{t_{n+1},\dots,t_{n+1}}^{j\text{-times}}] - \bar{\mathbf{v}}[\overbrace{t_n,\dots,t_n}^{i\text{-times}},\overbrace{t_{n+1},\dots,t_{n+1}}^{(j-1)\text{-times}}]}{\tau} \tag{2.29}$$

for $i = 1,\dots,d+2,\ j = 1,\dots,r+2$.

In particular, (2.27) shows that the coefficient of ψ_i is a divided difference with $i+1$ arguments. Our assumptions on a_k, b_k imply that divided differences containing only one argument j-times are bounded by terms of the form $cL^{j-1}\tau^{q+1}$, in particular

$$\bar{\mathbf{v}}[\underbrace{t_n,\dots,t_n}_{(j+1)\text{-times}}] \le cL^j\tau^{q+1} \tag{2.30}$$

for $j = 0,\dots,d+1$. Moreover, divided differences containing j_1-times t_n and j_2-times t_{n+1} are bounded by terms of the form

$$\sum_{l=1}^{\max\{j_1,j_2\}} cL^{l-1}\tau^{q+1-j_1-j_2+l}$$

due to our assumptions on a_k, b_k; (2.28) and (2.29). Thus, shifting the summation index,

$$\bar{\mathbf{v}}[\underbrace{t_n,\dots,t_n}_{(d+2)\text{-times}},\underbrace{t_{n+1},\dots,t_{n+1}}_{(i-d-1)\text{-times}}] \le \sum_{l=0}^{d+2} cL^l\tau^{q-i+l+1}, \tag{2.31}$$

we obtain the assertion of the lemma by combining (2.26), (2.30), and (2.31). ∎

Combining (2.21), (2.22), and Lemma 10 we obtain the following.

COROLLARY 11 (Bounds on residuals). *Let $\mathbf{R}_1^n, \mathbf{R}_3^n$ be defined as in* (2.19)*, then*

$$\|\mathbf{R}_1^n\|_{L_\infty(t_n,t_{n+1})} \le C\left\|d_t^{q+1}\tilde{\mathbf{u}}^n\right\|_\infty \sum_{k=0}^{q} L^k\tau^{q+k} \tag{2.32}$$

and

$$\|\mathbf{R}_3^n\|_{L_\infty(t_n,t_{n+1})} \le C\left\|d_t^{q+1}\tilde{\mathbf{u}}^n\right\|_\infty \sum_{k=0}^{q} L^{k+1}\tau^{q+1+k} \tag{2.33}$$

for $n = 0,\dots,N-1$.

Remark 12 (Approximation order). Note that in view of Lemma 10 we would still obtain that $\|R_1^n\|_{L_\infty}$ and $\|R_3^n\|_{L_\infty}$ are of optimal order if L in (2.34) were proportional to τ^{-1} instead of a constant.

Furthermore, a crucial point in the proof of optimality of $\|R_1^n\|_{L_\infty}$ and $\|R_3^n\|_{L_\infty}$ is that according to (2.21) and (2.22),

$$\begin{aligned} \left|(\mathrm{d}_t^e)^k(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n)(t_n^+)\right| &\le L^k\mathcal{O}(\tau^{q+1}) \text{ for } k = 0,\dots,d+1, \\ \left|(\mathrm{d}_t^e)^k(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n)(t_{n+1}^-)\right| &\le L^k\mathcal{O}(\tau^{q+1}) \text{ for } k = 0,\dots,r+1. \end{aligned} \tag{2.34}$$

This means we may change the prescribed values for $(\mathrm{d}_t^e)^k\hat{\mathbf{u}}(t_n^+)$ and $(\mathrm{d}_t^e)^k\hat{\mathbf{u}}(t_{n+1}^-)$ without compromising the quality of the reconstruction as long as

$$\begin{aligned} \left|(\mathrm{d}_t^e)^k(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n)(t_n^+)\right| &\le \mathcal{O}(\tau^{q+1-k}) \text{ for } k = 0,\dots,d+1, \\ \left|(\mathrm{d}_t^e)^k(\hat{\mathbf{u}}|_{(t_n,t_{n+1})} - \mathbf{h}^n)(t_{n+1}^-)\right| &\le \mathcal{O}(\tau^{q+1-k}) \text{ for } k = 0,\dots,r+1 \end{aligned} \tag{2.35}$$

remains valid. In other words, we can replace $\mathbf{f}_k^n(\mathbf{u}^n), \mathbf{f}_k^{n+1}(\mathbf{u}^{n+1})$ in (2.6) by (sufficiently accurate) approximations $\tilde{\mathbf{f}}_k^n[\mathbf{u}^n], \tilde{\mathbf{f}}_k^{n+1}[\mathbf{u}^{n+1}]$ thereof. In particular, we need the residual to be of order τ^q so that

$$\left|\mathbf{f}_k^m(\mathbf{u}^m) - \tilde{\mathbf{f}}_k^m[\mathbf{u}^m]\right| = \mathcal{O}(\tau^{q+1-k}).$$

We write $\tilde{\mathbf{f}}_k^m[\mathbf{u}^m]$ instead of $\tilde{\mathbf{f}}_k^m(\mathbf{u}^m)$ to indicate that this quantity might not only depend on $\mathbf{u}^m$, but also on $\mathbf{u}^{m-1}, \mathbf{u}^{m-2}$, etc. We will exploit this observation in section 2.3.

Now we are in a position to state the main result of this subsection.

THEOREM 13 (Optimality). *Let $\hat{\mathbf{u}} \in \mathbb{V}_q$ be the $H(p,d,r)$ reconstruction with $(p+1)(d+2)+r+1 = q$ of the solution to a qth order convergent single- or multistep method, approximating* (2.1)*, defined in* (2.6)*. Then, there exixts a constant $C = C(q) > 0$ such that the residual $\mathbf{R}$ defined in* (2.7) *satisfies*

$$\|\mathbf{R}\|_{L_\infty(0,T)} \le C \sup_n \left\| d_t^{q+1}\tilde{\mathbf{u}}^n \right\|_\infty \sum_{k=0}^{q+1} L^k \tau^{q+k}. \tag{2.36}$$

Proof. The assertion of the theorem follows upon combining (2.19), (2.20), and Corollary 11. ∎

This means (in the case of an ODE) that the residual is of the same order as the true error, since L and bounds for derivatives of solutions $\tilde{\mathbf{u}}^n$ of the auxiliary problems (2.15) can be expressed in terms of bounds on $\mathbf{f}$ and its derivatives which are independent of τ. We will see in the section 3 that the situation is more involved when we study fully discrete schemes for hyperbolic conservation laws, since the right-hand side of the (spatially) semidiscrete problem depends on h which is usually coupled to τ.

2.3. Approximation of f derivatives. The expressions $\mathbf{f}_k^n$ used in (2.6) might not be explicitly computable in many practical cases, e.g., in the case where the right-hand side stems from the spatial discretization of a system of hyperbolic conservation laws. We have already observed in Remark 12 that we may replace $\mathbf{f}_k^n$ in (2.6) by some approximation. Obviously, for any n there is only an interest in replacing $\mathbf{f}_k^n$ for $k \ge 2$, as $\mathbf{f}_0^n(\mathbf{u}^n) = \mathbf{u}^n$ and $\mathbf{f}_1^n(t_n,\mathbf{u}^n) = \mathbf{f}(\mathbf{u}^n)$ are readily computable and are probably computed anyway by the method used for time integration. Note that for $H(p,0,0)$ reconstructions, only $\mathbf{f}_0^n, \mathbf{f}_1^n$ appear in (2.6).

As the required order of accuracy of the approximations and also the different values of k for which $\mathbf{f}_k^n$ appears in (2.6) depend on q, we will present different approximation approaches for different values of $q > 3$.

2.3.1. Directional derivatives. In time integration methods of order $q = 4, 5$ reconstructions of types $H(0,1,0)$ and $H(0,1,1)$ require $\mathbf{f}_2^m(\mathbf{u}^m)$ for $m = n, n+1$, where

$$\mathbf{f}_2^m(\mathbf{u}^m) = \partial_t\mathbf{f}(t_m,\mathbf{u}^m) + \mathrm{D}\,\mathbf{f}(t_m,\mathbf{u}^m)\mathbf{f}(t_m,\mathbf{u}^m);$$

see (2.5). As we have seen in Remark 12 we may replace $\mathbf{f}_2^m$ by $\tilde{\mathbf{f}}_2^m$ as long as the error is of order $\mathcal{O}(\tau^{q-1})$ such that $\mathcal{O}(\tau^4)$ is admissible for $q = 4, 5$. Such an error is achieved by

$$\tilde{\mathbf{f}}_2^m[\mathbf{u}^m] := \frac{\mathbf{f}(t_m + \tau^2, \mathbf{u}^m) - \mathbf{f}(t_m - \tau^2, \mathbf{u}^m)}{2\tau^2} + \frac{\mathbf{f}(t_m, \mathbf{u}^m + \tau^2 \mathbf{f}(t_m, \mathbf{u}^m)) - \mathbf{f}(t_m, \mathbf{u}^m - \tau^2 \mathbf{f}(t_m, \mathbf{u}^m))}{2\tau^2} \tag{2.37}$$

for $m = n, n+1$. Note that for computing $\tilde{\mathbf{f}}_2^{n+1}[\mathbf{u}^{n+1}]$, four additional $\mathbf{f}$ evaluations (two if $\mathbf{f}$ does not depend on t) are required as the time integration scheme computes $\mathbf{f}(t_{n+1}, \mathbf{u}^{n+1})$. The value $\tilde{\mathbf{f}}_2^{n+1}[\mathbf{u}^{n+1}]$ computed for use at the right boundary in the time step from t_n to t_{n+1} can be reused in the next time step. Thus, this reconstruction approach requires four (two) additional $\mathbf{f}$ evaluations per time step.

2.3.2. Finite differences. Another way to obtain approximations of the derivatives $\mathbf{f}_k^{n+1}(\mathbf{u}^{n+1})$ is to use finite difference approximations of $d_t^{k-1}\mathbf{f}(t, \tilde{\mathbf{u}}(t))|_{t_{n+1}}$ in which $\tilde{\mathbf{u}}$ is some function interpolating $(t_{n-m}, \mathbf{u}^{n-m}), \dots, (t_n, \mathbf{u}^n), (t_{n+1}, \mathbf{u}^{n+1})$ for a sufficient number of points. It is preferable to choose those interpolated points as points which do not lie in the future of the point t_{n+1}, as this avoids the need to compute $\mathbf{f}$ evaluations before they are needed by the time integration scheme.

To be precise, let us elaborate the approach for $q = 4, 5$. We use the following backward finite difference stencil for first order derivatives:

$$\tilde{\mathbf{f}}_2^{n+1}[\mathbf{u}^{n+1}] := \frac{\frac{25}{12}\mathbf{f}_{n+1} - 4\mathbf{f}_n + 3\mathbf{f}_{n-1} - \frac{4}{3}\mathbf{f}_{n-2} + \frac{1}{4}\mathbf{f}_{n-3}}{\tau} \quad \text{for } n \geq 3, \tag{2.38}$$

where we used the abbreviation

$$\mathbf{f}_m := \mathbf{f}(t_m, \mathbf{u}^m). \tag{2.39}$$

This approach does not require any additional $\mathbf{f}$ evaluations but creates a need to store three previous $\mathbf{f}$ evaluations. While this creates (almost) no additional overhead in multi-step schemes, it is a certain overhead in one-step schemes like Runge–Kutta methods. The formula (2.38) does not allow for the computation of $\tilde{\mathbf{f}}_2^m[\mathbf{u}^m]$ for $m = 0, \dots, 3$, so a computation of our reconstruction on the first four time steps is not possible. However, after performing the first six time steps we may compute $\tilde{\mathbf{f}}_2^0[\mathbf{u}^0], \dots, \tilde{\mathbf{f}}_2^3[\mathbf{u}^3]$ using forward and central finite difference stencils.

Of course, higher order finite difference schemes can be used for the first and higher order derivatives. For example, using the sixth order one sided difference schemes for the first derivative and the fifth order scheme for the second derivative described in [13] allows for a reconstruction with $q = 7$. Using one sided differences makes it possible to reuse one approximation from one time step to the next. These methods require storing $\mathbf{f}_{n-6}, \dots, \mathbf{f}_{n-1}$. Note that the procedures discussed in [13] also allow us to construct higher order finite difference stencils, possibly including time steps of different lengths. However, for each combination of lengths of intervals a new stencil needs to be derived.

Remark 14 (Comparison of storage demands). In the case where $H(0, d, d)$ methods use the finite difference stencils described above, they do not need any additional $\mathbf{f}$ evaluations, but require the storage of $\mathbf{f}$ evaluations from previous time steps. In this sense they become comparable to $H(p, 0, 0)$ schemes, and we may compare the storage demands of both schemes. In the case of $q = 5$ we may use $H(0, 1, 1)$ requiring

the storage of three previous $\mathbf{f}$ evaluations, or $H(1,0,0)$ requiring the storage of one previous $\mathbf{f}$ evaluation. In the case of $q=7$ we may use $H(0,2,2)$ requiring the storage of six previous $\mathbf{f}$ evaluations, or $H(2,0,0)$ requiring the storage of two previous $\mathbf{f}$ evaluations.

From this perspective it seems that $H(p,0,0)$ schemes are more efficient than $H(0,d,d)$ schemes.

Remark 15 (Hermite–Birkhoff interpolation). Finally, note that Hermite–Birkhoff interpolation could also be used. As an example, the reconstruction suggested in [3] is similar to the approach presented in the work at hand. That reconstruction corresponds to fixing $\hat{\mathbf{u}} \in \mathbb{V}_3$ by prescribing three interpolation conditions for $\hat{\mathbf{u}}|_{(t_n,t_{n+1})}$:

$$\hat{\mathbf{u}}(t_n),\ \mathrm{d}_t^2\,\hat{\mathbf{u}}(t_{n+\frac{1}{2}});\hat{\mathbf{u}}(t_{n+1})\ . \tag{2.40}$$

While the values at the end points are readily available, some approximation is needed for the value of the second derivative at the interval midpoint. For this condition the following approximation is used:

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{f}(t_{n+\frac{1}{2}},\mathbf{u}(t_{n+\frac{1}{2}})) \approx \frac{1}{\tau}\big(\mathbf{f}(t_{n+1},\mathbf{u}^{n+1}) - \mathbf{f}(t_n,\mathbf{u}^n)\big),$$

resulting in an optimal scheme for $q=2$. A similar approximation order could be obtained by considering our reconstruction framework for the choice $H(0,0,-1)$, where the -1 should indicate that we prescribe the value but not any derivative of $\hat{\mathbf{u}}$ at t_{n+1}. Further Hermite–Birkhoff reconstructions for special Runge–Kutta methods are derived in [11, 19].

3. Estimates for fully discrete schemes for conservation laws. Let us consider a spatially one dimensional, hyperbolic system of $m \in \mathbb{N}$ conservation laws on the flat one dimensional torus $\mathbb{T}$ complemented with initial data $\mathbf{u}_0 : \mathbb{T} \to \mathcal{U} \subset \mathbb{R}^m$ where the state space $\mathcal{U}$ is an open set:

$$\partial_t \mathbf{u} + \partial_x \mathbf{g}(\mathbf{u}) = 0 \text{ on } (0,T)\times\mathbb{T}, \quad \mathbf{u}(0,\cdot) = \mathbf{u}_0 \ \text{ on } \mathbb{T}. \tag{3.1}$$

We assume the flux function $\mathbf{g}$ to be in $C^2(\mathcal{U},\mathbb{R}^m)$.

We restrict ourselves to the case where (3.1) is endowed with a strictly convex entropy-entropy flux pair, i.e., there exists a strictly convex $\eta \in C^1(\mathcal{U},\mathbb{R})$ and $q \in C^1(\mathcal{U},\mathbb{R})$ satisfying

$$(\mathrm{D}\,\eta)\,\mathrm{D}\,\mathbf{g} = \mathrm{D}\,q. \tag{3.2}$$

It is straightforward to check that any classical solution $\mathbf{u}$ to (3.1) satisfies the companion conservation law

$$\partial_t \eta(\mathbf{u}) + \partial_x q(\mathbf{u}) = 0. \tag{3.3}$$

Classically, the study of weak solutions to (3.1) is restricted to so-called entropy solutions; see, e.g., [6].

Definition 16 (Entropy solution). *A weak solution* $\mathbf{u} \in L_\infty((0,T)\times\mathbb{T},\mathcal{U})$ *to* (3.1) *is called an entropy solution with respect to* (η,q) *if it weakly satisfies*

$$\partial_t \eta(\mathbf{u}) + \partial_x q(\mathbf{u}) \le 0. \tag{3.4}$$

It was believed for a long time that entropy solutions are unique. While this is true for entropy solutions to scalar problems (satisfying a more discriminatory entropy condition), at least in multiple space dimensions entropy solutions to, e.g., the Euler equations are not unique [7, 4].

While it is not entirely clear whether entropy solutions to (general) systems in one space dimension with generic initial data are unique, it is well known that the entropy inequality (3.4) gives rise to some stability results, which in particular imply weak-strong uniqueness; see [10, 5].

3.1. Reconstruction for fully discrete DG schemes. In the following we study fully discrete schemes approximating (3.1) and employing a method of lines approach. We assume that the spatial discretization is done using a DG method with qth order polynomials and that the temporal discretization is based on some single- or multi-step method of order r.

We use decompositions $-1 = x_0 < x_1 < \cdots < x_{M-1} < x_M = 1$ of the spatial domain and $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ of the temporal domain. In order to account for the periodic boundary conditions we identify x_0 and x_M. We define time steps $\tau_n := t_{n+1} - t_n$, a maximal time step $\tau := \max_n \tau_n$, spatial mesh sizes $h_{k+\frac{1}{2}} := x_{k+1} - x_k$, $h_k := (h_{k+\frac{1}{2}} + h_{k-\frac{1}{2}})/2$, and a maximal and minimal spatial step

$$h := \max_k h_{k+\frac{1}{2}}, \quad h_{\min} := \min_k h_{k+\frac{1}{2}},$$

where we assume that $\frac{h}{h_{\min}}$ is uniformly bounded for $h \to 0$. We will write $\int_{\mathcal{T}}$ instead of $\sum_{i=1}^{M} \int_{x_{i-1}}^{x_i}$.

Let us introduce the piecewise polynomial DG ansatz and test space:

$$\mathbb{V}_q^s := \{\mathbf{w} : [x_0, x_M] \to \mathbb{R}^m \,:\, \mathbf{w}|_{(x_{i-1},x_i)} \in \mathbb{P}_q((x_{i-1}, x_i), \mathbb{R}^m) \text{ for } 1 \le i \le M\}. \tag{3.5}$$

Then, the fully discrete scheme results as a single- or multi-step discretization of the semidiscrete scheme

$$\partial_t \mathbf{u}_h = -\boldsymbol{f}(\mathbf{u}_h), \tag{3.6}$$

where the (nonlinear) map $\boldsymbol{f} : \mathbb{V}_q^s \to \mathbb{V}_q^s$ is defined by requiring that for all $\mathbf{v}_h, \boldsymbol{\psi} \in \mathbb{V}_q^s$ it holds that

$$\int_{\mathbb{T}} \boldsymbol{f}(\mathbf{v}_h)\boldsymbol{\psi} \,\mathrm{d}\, x = -\int_{\mathcal{T}} \mathbf{g}(\mathbf{v}_h)\partial_x \boldsymbol{\psi} \,\mathrm{d}\, x + \sum_{i=0}^{M-1} \mathbf{G}(\mathbf{v}_h(x_i^-), \mathbf{v}_h(x_i^+)) \left[\!\left[\boldsymbol{\psi}\right]\!\right]_i, \tag{3.7}$$

where $\mathbf{G} : \mathcal{U} \times \mathcal{U} \to \mathbb{R}^m$ is a numerical flux function, $\left[\!\left[\boldsymbol{\psi}\right]\!\right]_i := \boldsymbol{\psi}(x_i^-) - \boldsymbol{\psi}(x_i^+)$ are jumps, and the notation for spatial traces is analogous to that for temporal traces; see (2.3). We will specify our assumptions on the numerical flux in Assumption 20.

Suppose the fully discrete numerical scheme allows us to compute a sequence of approximate solutions at points $\{t_n\}_{n=0}^N$ in time: $\mathbf{u}_h^0, \mathbf{u}_h^1, \mathbf{u}_h^2, \ldots, \mathbf{u}_h^N \in \mathbb{V}_q^s$. In order to make sense of its reconstruction, we define for any vector space V a space of piecewise polynomials in time by

$$\mathbb{V}_r^t(0, T; V) := \{\mathbf{w} : [0, T] \to V \,:\, \mathbf{w}|_{(t_n,t_{n+1})} \in \mathbb{P}_r((t_n, t_{n+1}), V)\}. \tag{3.8}$$

Using the methodology from section 2.1, we obtain a computable (temporal) reconstruction $\hat{\mathbf{u}}^t \in \mathbb{V}_r^t(0, T; \mathbb{V}_q^s)$.

Assumption 17 (Bounded reconstruction). For the remainder of this section we will suppose that there is some compact and convex $\mathfrak{K} \subset \mathcal{U}$ such that

$$\hat{\mathbf{u}}^t(t,x) \in \mathfrak{K} \quad \forall\ (t,x) \in [0,T] \times \mathbb{T}.$$

Remark 18 (Bounded reconstruction). Note that Assumption 17 is verifiable in an a posteriori fashion, since $\hat{\mathbf{u}}^t$ is explicitly computable. It is, however, not sufficient to verify $\mathbf{u}_h^n(x) \in \mathfrak{K}$ for all $n = 0, \dots, N$ and all $x \in \mathbb{T}$.

Remark 19 (Bounds on flux and entropy). Due to the regularity of $\mathbf{g}$ and η and the compactness of $\mathfrak{K}$, there exist constants $0 < C_{\overline{\boldsymbol{g}}} < \infty$ and $0 < C_{\underline{\eta}} < C_{\overline{\eta}} < \infty$ which can be explicitly computed from $\mathfrak{K}$, $\mathbf{g}$, and η such that

$$\left|\mathbf{v}^T \,\mathrm{H}\,\mathbf{g}(\mathbf{u})\mathbf{v}\right| \le C_{\overline{\boldsymbol{g}}} \left|\mathbf{v}\right|^2, \quad C_{\underline{\eta}} \left|\mathbf{v}\right|^2 \le \mathbf{v}^T \,\mathrm{H}\,\eta(\mathbf{u})\mathbf{v} \le C_{\overline{\eta}} \left|\mathbf{v}\right|^2 \ \forall\ \mathbf{v} \in \mathbb{R}^m, \mathbf{u} \in \mathfrak{K}, \tag{3.9}$$

where $|\cdot|$ is the Euclidean norm for vectors and H denotes Hessian matrices.

We can now define, as in the previous section, the temporal residual

$$\mathbf{R}^t := \partial_t \hat{\mathbf{u}}^t + \boldsymbol{f}(\hat{\mathbf{u}}^t) \tag{3.10}$$

with $\mathbf{R}^t \in L_2((0,T); \mathbb{V}_q^s)$. As $\hat{\mathbf{u}}^t$ is explicitly computable, so is $\mathbf{R}^t$. Note that for $r \ge 3$ we even have $\mathbf{R}^t \in C^0((0,T); \mathbb{V}_q^s)$.

Our spatial reconstruction of $\hat{\mathbf{u}}^t$ is based on [14]. To this end we restrict ourselves to two types of numerical fluxes $\mathbf{G}$.

Assumption 20 (Condition on the numerical flux). We assume that there exists a locally Lipschitz continuous function $\mathbf{w} : \mathcal{U} \times \mathcal{U} \to \mathcal{U}$ such that for any compact $K \subset \mathcal{U}$ there exists a constant $C_w(K) > 0$ with

$$|\mathbf{w}(\mathbf{a},\mathbf{b}) - \mathbf{a}| + |\mathbf{w}(\mathbf{a},\mathbf{b}) - \mathbf{b}| \le C_w(K)|\mathbf{a} - \mathbf{b}| \quad \forall\ \mathbf{a},\mathbf{b} \in K. \tag{3.11}$$

With this function the numerical flux $\mathbf{G}$ is of one of the two following types:

(i) $\mathbf{G}(\mathbf{a},\mathbf{b}) = \mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})) \quad \forall\ \mathbf{a},\mathbf{b} \in \mathcal{U}$;

(ii) $\mathbf{G}(\mathbf{a},\mathbf{b}) = \mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})) - \mu(\mathbf{a},\mathbf{b};h)h^\nu(\mathbf{b}-\mathbf{a}) \quad \forall\ \mathbf{a},\mathbf{b} \in \mathcal{U}$ for some $\nu \in \mathbb{N}_0$ and some matrix-valued function μ which has the property that for any compact $K \subset \mathcal{U}$ there exists a constant $\mu_K > 0$ so that $|\mu(\mathbf{a},\mathbf{b};h)| \le \mu_K(1 + \frac{|\mathbf{b}-\mathbf{a}|}{h})$ for h small enough.

Remark 21 (Restrictions on the numerical flux).

1. The condition imposed in Assumption 20 is stronger than the classical Lipschitz and consistency conditions.
2. The conditions do not, by any means, guarantee stability of the numerical scheme. In practical computations, interest focuses on numerical fluxes which satisfy one of the assumptions and lead to a stable numerical scheme, for obvious reasons.
3. The Lax–Wendroff and Richtmyer numerical fluxes, e.g.,

$$\mathbf{G}(\mathbf{a},\mathbf{b}) = \mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})), \quad \mathbf{w}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}+\mathbf{b}}{2} - \frac{\lambda}{2}(\mathbf{g}(\mathbf{b}) - \mathbf{g}(\mathbf{a})), \tag{3.12}$$

   satisfy Assumption 20 (i).
4. The Lax–Friedrichs flux

$$\mathbf{G}(\mathbf{a},\mathbf{b}) = \frac{1}{2}\Big(\mathbf{g}(\mathbf{a}) + \mathbf{g}(\mathbf{b})\Big) - \lambda(\mathbf{b}-\mathbf{a}) \tag{3.13}$$

satisfies Assumption 20 (ii) with $\mathbf{w}(\mathbf{a},\mathbf{b}) = \frac{1}{2}(\mathbf{a}+\mathbf{b})$, $\nu = 0$, and

$$\mu(\mathbf{a},\mathbf{b};h) = \lambda\mathbb{I} - \frac{\mathbf{g}(\mathbf{a}) - 2\mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})) + \mathbf{g}(\mathbf{b})}{2\|\mathbf{b}-\mathbf{a}\|^2} \otimes (\mathbf{b}-\mathbf{a}),$$

where $\mathbb{I}$ denotes the $m \times m$ identity matrix.

5. A Lax–Wendroff-type flux with artificial viscosity as described, for example, in [29, Chap. 16] satisfies Assumption 20 (ii) with $\nu = 1$. In fact, the analysis presented in the following applies to a large class of general artificial viscosity methods typically used with Lax–Wendroff-type fluxes, e.g., a discrete form of $h^2\partial_x\big(|\partial_x u|\partial_x u\big)$; see, e.g., [28, 34]. In the numerical examples we will use a discrete version of this viscosity given by $\nu = 1, \mu(\mathbf{a},\mathbf{b};h) = \mu_0|\mathbf{b}-\mathbf{a}|/h$, with $\mu_0 > 0$ constant.
6. The assumptions given here cover to a certain extent methods combining central with lower order upwind-type fluxes, e.g.,

$$\mathbf{G}(\mathbf{a},\mathbf{b}) = (1-\phi(h))\mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})) + \phi(h)\Big(\frac{1}{2}\big(\mathbf{g}(\mathbf{a})+\mathbf{g}(\mathbf{b})\big) - \lambda(\mathbf{b}-\mathbf{a})\Big)$$

as long as $\phi(h) = O(h)$ for smooth solutions, e.g., [36, 33]. To see this, take $\nu = 1$ and

$$\mu(\mathbf{a},\mathbf{b};h) = \lambda\mathbb{I} - \frac{\phi(h)}{2h}\frac{\mathbf{g}(\mathbf{a}) - 2\mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})) + \mathbf{g}(\mathbf{b})}{\|\mathbf{b}-\mathbf{a}\|^2} \otimes (\mathbf{b}-\mathbf{a}).$$

We will define a spatial reconstruction method which is applied to the temporal reconstruction $\hat{\mathbf{u}}^t(t,\cdot)$ for each $t \in (0,T)$ using the function $\mathbf{w}$ (which depends on the numerical flux under consideration) to obtain a continuous space-time reconstruction $\hat{\mathbf{u}}^{st}$.

DEFINITION 22 (Space-time reconstruction). *Let $\hat{\mathbf{u}}^t$ be the temporal reconstruction, as in Definition 2, of a sequence $\{\mathbf{u}_h^n\}_{n=0}^N \subset \mathbb{V}_q^s$ computed from a method of lines scheme consisting of a DG discretization in space using a numerical flux satisfying Assumption 20 and some single- or multi-step method in time. Then, the space-time reconstruction $\hat{\mathbf{u}}^{st}(t,\cdot) \in \mathbb{V}_{q+1}^s$ is defined by requiring*

$$\begin{aligned} &\int_{\mathbb{T}} (\hat{\mathbf{u}}^{st}(t,\cdot) - \hat{\mathbf{u}}^t(t,\cdot)) \cdot \boldsymbol{\psi} = 0 \quad \forall \boldsymbol{\psi} \in \mathbb{V}_{q-1}^s, \\ &\qquad \hat{\mathbf{u}}^{st}(t,x_k^{\pm}) = \mathbf{w}(\hat{\mathbf{u}}^t(t,x_k^-),\hat{\mathbf{u}}^t(t,x_k^+)) \quad \forall k. \end{aligned} \tag{3.14}$$

Remark 23 (Choice of reconstruction).

1. We believe this reconstruction to be the most meaningful choice for general systems and fluxes as in Assumption 20 (i) and (ii). We will see in Lemma 34 that this choice may lead to a suboptimal residual in case $\nu = 0$. We cannot rule out the possibility that another reconstruction choice exists which leads to an estimator of optimal order for this type of flux also in the case $\nu = 0$.
2. We will see in section 3.3 that for strictly hyperbolic systems it is indeed possible to obtain an estimator of optimal order with a Roe-type flux. But how to extend this construction to more general fluxes is not obvious. We will for example, demonstrate the difficulties in the case of a Lax–Friedrichs-type flux.

LEMMA 24 (Properties of space-time reconstruction). *Let $\hat{\mathbf{u}}^{st}$ be the space-time reconstruction defined by* (3.14). *Then, for each $t \in [0,T]$, the function $\hat{\mathbf{u}}^{st}(t,\cdot)$ is well defined and locally computable. Moreover,*

$$\hat{\mathbf{u}}^{st} \in W^1_\infty(0,T;\mathbb{V}^s_{q+1} \cap C^0(\mathbb{T})).$$

As $\hat{\mathbf{u}}^{st}$ is a piecewise polynomial and continuous in space, it is also Lipschitz continuous in space.

Proof. The facts that for each $t \in [0,T]$ the function $\hat{\mathbf{u}}^{st}(t,\cdot)$ is well defined, locally computable, and continuous follow from [14, Lem. 4.3]. Assumptions 20 and 17 imply that $\mathbf{w}$ is Lipschitz on the set of values taken by $\hat{\mathbf{u}}^t$. Thus, the Lipschitz continuity in time of $\hat{\mathbf{u}}^t$ translates into Lipschitz continuity in time of $\hat{\mathbf{u}}^{st}$. ∎

Since $\hat{\mathbf{u}}^{st}$ is computable and Lipschitz continuous in space and time, we may define a computable residual

$$\mathbf{R}^{st} := \partial_t \hat{\mathbf{u}}^{st} + \partial_x \mathbf{g}(\hat{\mathbf{u}}^{st}) \tag{3.15}$$

with $\mathbf{R}^{st} \in L_2((0,T)\times\mathbb{T},\mathbb{R}^m)$.

We can now formulate our main a posteriori estimate.

THEOREM 25 (A posteriori error bound). *Let $\mathbf{g} \in C^2(\mathcal{U},\mathbb{R}^m)$ and let $\mathbf{u}$ be an entropy solution of* (3.1) *with periodic boundary conditions. Let $\hat{\mathbf{u}}^{st}$ be the space-time reconstruction of the solution $\{\mathbf{u}_h^n\}_{n=0}^N$ of a fully discrete DG scheme, defined according to Definition* 22. *Provided $\mathbf{u}$ takes only values in $\mathfrak{K}$, then for $n = 0,\dots,N$ the error between the numerical solution $\mathbf{u}_h^n$ and $\mathbf{u}(t_n,\cdot)$ satisfies*

$$\begin{aligned}\|\mathbf{u}(t_n,\cdot) - \mathbf{u}_h^n\|^2_{L_2(\mathbb{T})} &\le 2\left\|\hat{\mathbf{u}}^{st}(t_n,\cdot) - \mathbf{u}_h^n\right\|^2_{L_2(\mathbb{T})} \\ &\quad + 2C_{\underline{\eta}}^{-1}\left(\left\|\mathbf{R}^{st}\right\|^2_{L_2((0,t_n)\times\mathbb{T})} + C_{\overline{\eta}}\left\|\mathbf{u}_0 - \hat{\mathbf{u}}^{st}(0,\cdot)\right\|^2_{L_2(\mathbb{T})}\right) \\ &\quad \times \exp\left(\int_0^{t_n} \frac{C_{\overline{\eta}}C_{\overline{\boldsymbol{g}}}\left\|\partial_x\hat{\mathbf{u}}^{st}(s,\cdot)\right\|_{L_\infty(\mathbb{T})} + C^2_{\overline{\eta}}}{C_{\underline{\eta}}}\,\mathrm{d}\,s\right).\end{aligned} \tag{3.16}$$

Proof. The theorem follows from (3.15) in exactly the same way as [14, Thm. 5.5] follows from [14, eq. (5.2)]. ∎

Remark 26 (Computation of the estimator). Note that if $\mathbf{w}$ is not any smoother than Lipschitz continuous, $\hat{\mathbf{u}}^{st}|_{(t_n,t_{n+1})}$ is also only Lipschitz continuous in time, although $\hat{\mathbf{u}}^t|_{(t_n,t_{n+1})}$ is a polynomial. In this case the evaluation of $\|\mathbf{R}^{st}\|^2_{L_2((0,t_n)\times\mathbb{T})}$ with high precision is numerically extremely expensive. Thus, we aim to find smooth choices for $\mathbf{w}$ in our test.

Remark 27 (Discontinuous entropy solutions).

1. The reader may note that the estimate in Theorem 25 does not require the entropy solution $\mathbf{u}$ to be continuous. However, in the case where $\mathbf{u}$ is discontinuous, $\|\partial_x\hat{\mathbf{u}}^{st}(s,\cdot)\|_{L_\infty(\mathbb{T})}$ is expected to scale like h^{-1}. Therefore, the estimator in (3.16) will (at best) scale like $h^{q+1}\exp(h^{-1})$, which diverges for $h \to 0$. So, in particular, the estimator diverges for $h \to 0$ if the entropy solution is discontinuous.
2. The fact that the estimator does not converge to zero for $h,\tau \to 0$ in the case of a discontinuous entropy solution results from using the relative entropy

framework, which does not guarantee uniqueness of an entropy solution. Indeed, it was shown in [7] that, in general, entropy solutions may be nonunique for the Euler equations in several space dimensions.

3. It is well known that for nonlinear problems the DG method is not stable in the case of discontinuities in the solution. Thus we cannot expect the error to converge to zero and therefore the estimator cannot be expected to converge either. Stabilizing limiters can be included in the framework but that is outside the scope of this paper and will be investigated in a subsequent study.

3.2. Optimality of the error estimator. In this section we address optimality of the error estimator, i.e., whether the estimator is of the same order (in h and τ) as the true error of the numerical scheme in the case where the scheme converges. Throughout this section we will assume that the entropy solution is smooth, since otherwise no a priori error analysis is available and the relative entropy framework breaks down.

Assumption 28 (Stability of the semidiscrete scheme). We will also assume that the solutions $\tilde{\mathbf{u}}^n$ to the auxiliary initial value problems for the (spatially) semidiscrete schemes

$$\partial_t \tilde{\mathbf{u}}^n = -\boldsymbol{f}(\tilde{\mathbf{u}}^n) \quad \text{in } (t_n, t_{n+1}), \quad \tilde{\mathbf{u}}^n(t_n) = \mathbf{u}^n, \tag{3.17}$$

with $\mathbf{f}$ from (3.7) satisfy

$$\sup_n \left\| d_t^{r+1} \tilde{\mathbf{u}}^n \right\|_{L^\infty(t_n, t_{n+1})} \leq C \tag{3.18}$$

for some $C > 0$ independent of τ, h.

Remark 29 (Stability of the semidiscrete scheme). In view of the fact that the $\tilde{\mathbf{u}}^n$ are solutions to the semidiscrete scheme approximating the entropy solution $\mathbf{u}$ of a hyperbolic conservation law which is independent of τ, h, Assumption 28 is a condition on the stability of the semidiscrete scheme when applied to initial values $\mathbf{u}^n$. Due to the fact that the initial values $\mathbf{u}^n$ enter into the assumption, it is also related to the stability of the fully discrete scheme.

Note that Assumption 28 is not related to the reliability of the estimator, i.e., to the fact that it provides an upper bound for the error. It is only related to whether or not the estimator decreases with the same order as the true error. This question is only relevant for convergent and, hence, stable schemes.

We will now consider the residual in more detail. Note that, adding zero, we may combine (3.10) and (3.15) in order to obtain

$$\mathbf{R}^{st} := \partial_t(\hat{\mathbf{u}}^{st} - \hat{\mathbf{u}}^t) + \partial_x \mathbf{g}(\hat{\mathbf{u}}^{st}) - \mathbf{f}(\hat{\mathbf{u}}^t) + \mathbf{R}^t =: \mathbf{R}^s + \mathbf{R}^t. \tag{3.19}$$

In this way we may decompose the residual into a "spatial" and a "temporal" part. The analysis of [37] shows that the spatial part of the error of Runge–Kutta DG discretizations of systems of hyperbolic conservation laws is $\mathcal{O}(h^{q+\gamma})$, with q the polynomial degree of the DG scheme and γ depending on the numerical flux. For general monotone fluxes $\gamma = \frac{1}{2}$ and for upwind-type fluxes γ is improved to 1. When stating our optimality result below, we will assume that the true error of our scheme is $\mathcal{O}(h^{q+\gamma} + \tau^r)$. We first consider the case of fluxes satisfying Assumption 20(i).

THEOREM 30 (Optimality of residuals). *Let a numerical scheme which is of order r in time and uses qth order DG in space with a numerical flux satisfying Assumption* 20 (i) *be given. Let the temporal and spatial mesh size comply with a CFL-type restriction $\tau = \mathcal{O}(h_{min})$. Let the residual $\mathbf{R}^{st}$ be defined by* (3.15) *and* (3.14). *Then, it is of optimal order $\mathcal{O}(h^{q+\gamma} + \tau^r)$ with $\gamma \in \{\frac{1}{2}, 1\}$, provided the exact error is of this order, and the Lipschitz constant L of $\mathbf{f}$ defined in* (3.7) *in the sense of* (2.14) *behaves like $h_{\min}^{-1}$.*

Proof. Due to (3.19) it is sufficient that $\mathbf{R}^s$ and $\mathbf{R}^t$ are of optimal order. The temporal residual $\mathbf{R}^t$ is of the type of residuals investigated in section 2. Invoking Theorem 13 we obtain

$$\|\mathbf{R}^t\|_{L_\infty(0,T;L_2(\mathbb{T}))} \leq C \sup_n \left\| d_t^{r+1} \tilde{\mathbf{u}}^n \right\|_{L_\infty(t_n,t_{n+1})} \sum_k L^k (\tau^r + h^{q+\gamma}) \tau^k \tag{3.20}$$

with $\tilde{\mathbf{u}}^n$ from (3.17). The condition on L and the CFL condition ensure that $L^k \tau^k = \mathcal{O}(1)$. Combining this with Assumption 28, we obtain that $\mathbf{R}^t$ is of optimal order. The spatial residual $\mathbf{R}^s$ is of the form of residuals investigated in [14] and can be estimated as in [14, Lem. 6.2] with $\mathbf{u}_h$ being replaced by $\hat{\mathbf{u}}^t$. Arguments similar to [14, Rem. 6.6] show that $\mathbf{R}^s$ is of optimal order. ∎

Remark 31 (Spatial residual). Any computable reconstruction $\tilde{\mathbf{u}} \in W_\infty^1((0,T) \times \mathbb{T}, \mathcal{U})$ of the numerical solution gives rise to an error estimate of the form (3.16). Our particular reconstruction and (at the same time) the condition on the numerical flux in Assumption 20 (i) are driven by our desire for the spatial part $\mathbf{R}^s$ of the residual to be of optimal order. This part of the residual, and its optimality, was extensively investigated in [14]. We will investigate the effects of added artificial viscosity in Theorem 32 and Lemma 34, showing that low order viscosity leads to a suboptimal convergence of the estimator.

Let us turn our attention to numerical fluxes satisfying Assumption 20 (ii). Our goal is to ascertain the effect of artificial viscosity on the order of the residual. Note that we define the reconstruction by (3.14) as before, accounting for $\mathbf{w}$ but not for the artificial viscosity. In section 3.3 we will show for a Roe-type flux that for one flux there might be different choices of $\mathbf{w}$ leading to different values of ν. It is not obvious whether this approach can be generalized so that optimal reconstructions can be obtained for more general numerical fluxes.

To simplify the presentation we will assume in the following that $\tau = \mathcal{O}(h_{min})$ and that the order of the time stepping method is compatible with the order of the space discretization, i.e., $r = q + 1$. We will first show how an upper bound for the residual depends on the order ν of the artificial viscosity.

THEOREM 32 (Conditional optimality of residuals). *For $q \geq 1$ let a numerical scheme which is order $q+1$ in time and uses qth order DG in space with a numerical flux satisfying Assumption* 20 (ii) *be given. Let the residual $\mathbf{R}^{st}$ be defined by* (3.15) *and* (3.14). *Let the exact error be of order $\mathcal{O}(h^{q+\gamma})$. Then, $\mathbf{R}^{st}$ is of order $\mathcal{O}(h^{q+\gamma} + h^{q+\gamma+\nu-1})$ with $\gamma \in \{\frac{1}{2}, 1\}$, provided the Lipschitz constant L of $\mathbf{f}$ defined in* (3.7) *in the sense of* (2.14) *behaves like $h_{\min}^{-1}$.*

Remark 33 (Conditional optimality). Note that the rate proven here is optimal for $\nu \geq 1$ but suboptimal otherwise. We will show in Lemma 34 that the $\mathcal{O}(h^{q+\gamma+\nu-1})$ part of the residual is actually present.

Proof. Let $\hat{\mathbf{u}}^t$ denote the temporal reconstruction of the numerical solution. Note that we will suppress time dependence in the notation in the remainder of this proof.

All the arguments we give are uniform in time. Our argument is based on an observation in [31] which states that

$$\sum_i h_i |\left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i|^2 = \mathcal{O}(h^{2q+2\gamma}), \tag{3.21}$$

i.e., this sum is of the order of the square of the (spatial) L_2-norm of the true error. Let us note that (3.21) implies

$$\max_i \left| \frac{\left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i}{h_i} \right|^2 \lesssim \frac{1}{h^3} \max_i \left(h_i \left| \left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i \right|^2 \right) \lesssim \frac{1}{h^3} \sum_i h_i |\left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i|^2 \lesssim \mathcal{O}(h^{2q+2\gamma-3}). \tag{3.22}$$

Since $q \geq 1$ and $\gamma \geq \frac{1}{2}$, (3.22) implies

$$\max_i \left| \frac{\left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i}{h_i} \right| \lesssim 1. \tag{3.23}$$

We will not consider the full spatial residual $\mathbf{R}^s$, since, following the arguments in the proof of [14, Lem. 6.2], the only part which changes as Assumption 20 (i) is replaced by 20 (ii) is the estimate of the L_2-norm of

$$\mathbf{R}_g := \mathcal{P}_q[\partial_x \mathbf{g}(\hat{\mathbf{u}}^{st})] - \mathbf{f}(\hat{\mathbf{u}}^t) \in \mathbb{V}_q^s, \tag{3.24}$$

where $\mathcal{P}_q$ denotes L_2-orthogonal projection into $\mathbb{V}_q^s$ and we note that $\mathbf{f}(\hat{\mathbf{u}}^t)$, defined by (3.7), corresponds to $\partial_x \hat{\mathbf{g}}$ in [14]. We will only consider this part (3.24) of the spatial residual in the sequel. Using integration by parts we obtain

$$\begin{aligned}
\int_{\mathcal{T}} |\mathbf{R}_g|^2 \,\mathrm{d}\, x &= \int_{\mathcal{T}} (\mathcal{P}_q[\partial_x \mathbf{g}(\hat{\mathbf{u}}^{st})] - \mathbf{f}(\hat{\mathbf{u}}^t)) \cdot \mathbf{R}_g \,\mathrm{d}\, x \\
&= \int_{\mathcal{T}} (\partial_x \mathbf{g}(\hat{\mathbf{u}}^{st}) - \mathbf{f}(\hat{\mathbf{u}}^t)) \cdot \mathbf{R}_g \,\mathrm{d}\, x \\
&= \int_{\mathcal{T}} (\mathbf{g}(\hat{\mathbf{u}}^t) - \mathbf{g}(\hat{\mathbf{u}}^{st})) \cdot \partial_x^e \mathbf{R}_g \,\mathrm{d}\, x + \sum \left(\mathbf{G}((\hat{\mathbf{u}}^t)^{\pm}) - \mathbf{g}(\mathbf{w}((\hat{\mathbf{u}}^t)^{\pm})) \right) \cdot \left[\!\left[\mathbf{R}_g\right]\!\right] \\
&= \int_{\mathcal{T}} (\mathbf{g}(\hat{\mathbf{u}}^t) - \mathbf{g}(\hat{\mathbf{u}}^{st})) \cdot \partial_x^e \mathbf{R}_g \,\mathrm{d}\, x + h^\nu \sum \mu((\hat{\mathbf{u}}^t)^-, (\hat{\mathbf{u}}^t)^+; h_i) \left[\!\left[\hat{\mathbf{u}}^t\right]\!\right] \cdot \left[\!\left[\mathbf{R}_g\right]\!\right] \\
&=: E_1 + E_2,
\end{aligned} \tag{3.25}$$

where $(\hat{\mathbf{u}}^t)^{\pm}$ is an abbreviation for $((\hat{\mathbf{u}}^t)^-, (\hat{\mathbf{u}}^t)^+) := (\hat{\mathbf{u}}^t(x_i^-), \hat{\mathbf{u}}^t(x_i^+))$. Using the same trick as in the proof of [14, Lem. 6.2] we obtain

$$\begin{aligned}
|E_1| \lesssim \left|\hat{\mathbf{u}}^t\right|_{W_\infty^1(\mathcal{T})} \left\|\hat{\mathbf{u}}^{st} - \hat{\mathbf{u}}^t\right\|_{L_2(\mathbb{T})} \|\mathbf{R}_g\|_{L_2(\mathbb{T})} \\
+ \left(\sum_{i=1}^M \frac{1}{h_{i-\frac{1}{2}}^2} \int_{x_{i-1}}^{x_i} \left|\hat{\mathbf{u}}^{st} - \hat{\mathbf{u}}^t\right|^4 \mathrm{d}\, x \right)^{\frac{1}{2}} \|\mathbf{R}_g\|_{L_2(\mathbb{T})}.
\end{aligned} \tag{3.26}$$

In order to bound E_2 we employ the Cauchy–Schwarz inequality, trace inequalities [9, Lem. 1.46], and (3.23) so that we get

$$\begin{aligned}
|E_2| &\lesssim \mu_{\mathfrak{K}} h^\nu \left(\sum_{i=0}^{M-1} \left(1 + \frac{|\left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]|}{h_i} \right)^2 \frac{1}{h_i} \left| \left[\!\left[\hat{\mathbf{u}}^t\right]\!\right] \right|^2 \right)^{\frac{1}{2}} \left(\sum_{i=0}^{M-1} h_i |\left[\!\left[\mathbf{R}_g\right]\!\right]|^2 \right)^{\frac{1}{2}} \\
&\lesssim \mu_{\mathfrak{K}} h^\nu \left(\sum_{i=0}^{M-1} \frac{1}{h_i} \left| \left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i \right|^2 \right)^{\frac{1}{2}} \|\mathbf{R}_g\|_{L_2(\mathbb{T})}.
\end{aligned} \tag{3.27}$$

Inserting (3.26) and (3.27) into (3.25) and dividing both sides by $\|\mathbf{R}_g\|_{L_2(\mathbb{T})}$, we obtain

$$\|\mathbf{R}_g\|_{L_2(\mathbb{T})} \le \left|\hat{\mathbf{u}}^t\right|_{W^1_\infty} \left\|\hat{\mathbf{u}}^{st} - \hat{\mathbf{u}}^t\right\|_{L_2(\mathbb{T})} + \left(\sum_{i=1}^{M} \frac{1}{h^2_{i-\frac{1}{2}}} \int_{x_{i-1}}^{x_i} \left|\hat{\mathbf{u}}^{st} - \hat{\mathbf{u}}^t\right|^4 \mathrm{d}\, x\right)^{\frac{1}{2}} + \mu_{\mathfrak{K}} h^\nu \left(\sum_{i=0}^{M-1} \frac{1}{h_i} \left|\left[\!\left[\hat{\mathbf{u}}^t\right]\!\right]_i\right|^2\right)^{\frac{1}{2}}. \tag{3.28}$$

Following the arguments of [14, Rem. 6.6] we see that the first two terms on the right-hand side of (3.28) are of order $\mathcal{O}(h^{q+\gamma})$, while the last term, which has no counterpart in the analysis presented in [14], is of order $\mathcal{O}(h^{\nu-1+q+\gamma})$. ∎

3.3. Optimal reconstruction for a Roe-type flux. The goal of this section is twofold. Firstly, we will study a Roe-type flux with a reconstruction based on a simple average showing that this can lead to suboptimal convergence of the residual. The flux with this choice of $\mathbf{w}$ fits into the framework of Assumption 20 (ii) with $\nu = 0$, thus we show that our estimate in Theorem 32 is sharp. Secondly, we show that for the same flux it is possible to find a more involved version of $\mathbf{w}$, leading to a reconstruction with optimal order residual. We hope also to convince the reader of the difficulty of finding such a reconstruction for general numerical flux functions.

In the following we consider a strictly hyperbolic system of the form (3.1) using a numerical flux of Roe type. We fix two elements $\mathbf{a}, \mathbf{b}$ in a convex, compact subset $K \subset \mathcal{U}$ and denote their average with $\mathbf{c} := \frac{1}{2}(\mathbf{a} + \mathbf{b})$. Let $A := D\mathbf{g}(\mathbf{c})$ be the flux Jacobian at this point. Due to the hyperbolicity of the system, A is diagonalizable: $LAR = D$ with a diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$, consisting of the eigenvalues and matrices R containing the right eigenvectors $\mathbf{r}_i$ as columns and L with the left eigenvectors $\mathbf{l}_i$ as rows. We choose the right and the left eigenvectors to be dual to each other, i.e., $\mathbf{l}_i \cdot \mathbf{r}_j = \delta_{ij}$. Consider now the numerical flux function of Roe type given by

$$\mathbf{G}(\mathbf{a}, \mathbf{b}) = \mathbf{g}(\mathbf{c}) + \frac{1}{2}\left|A\right|(\mathbf{a} - \mathbf{b}), \tag{3.29}$$

where $|A| = R\,|D|\,L$ and $|D| := \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_m|)$. Note that, as we restrict ourselves to the strictly hyperbolic case, appropriately normalized right and left eigenvalues are C^1 vector fields on $\mathcal{U}$ and, thus, there is a bound $C^* > 0$ on $|L| \cdot |R|$ depending on K, but not on the specific choice of $\mathbf{a}, \mathbf{b} \in K$.

First we will demonstrate that, in general, taking $\mathbf{w}(\mathbf{a}, \mathbf{b}) = \mathbf{c}$ does not lead to an optimal error estimate. It is easy to see that this choice leads to a μ which satisfies Assumption 20 (ii) with $\nu = 0$. Taking, for example, the scalar case and noting $|A| = |g'(c)|$, we find

$$|\mu(a, b; h)| = \frac{|g(c) - G(a, b)|}{|b - a|} = \frac{1}{2}|g'(c)|, \tag{3.30}$$

which is clearly bounded on any compact subset of the state space. The following lemma shows that the suboptimal rate stated in Theorem 32 is sharp and that therefore optimal order for the residual is in general only guaranteed if the artificial viscosity term is chosen with $\nu \ge 1$. Taking this result together with the observation made above, it is clear that $\mathbf{w}(\mathbf{a}, \mathbf{b}) = \mathbf{c}$ will in general not lead to an optimal rate of convergence of the residual.

LEMMA 34 (Suboptimality). *Consider the scalar linear problem*

$$u_t + u_x = 0$$

and a numerical scheme which is of order 2 *in time and uses first order DG in space on an equidistant mesh of size* h *with a numerical flux satisfying Assumption* 20 (ii) *with* $\nu = 0$ *and* $\mu(a,b;h) = \mu_0 > 0$. *Let the temporal and spatial mesh size comply with a CFL-type restriction* $\tau = \mathcal{O}(h)$. *Then, the norm of the residual* $\mathbf{R}^{st}$, *defined by* (3.15) *and* (3.14), *is bounded from below by terms of order* h^γ *even if the error of the method is* $\mathcal{O}(h^{1+\gamma})$.

Proof. We use the same notation as in the proof of Theorem 32. As argued there, it is sufficient to show that $\|\mathbf{R}_g\|_{L_2(\mathbb{T})}$ is bounded from below by terms of order h^γ. All the other terms are of higher order and can, therefore, not cancel with $\mathbf{R}_g$. The assumptions of the lemma at hand are a special case of those of Theorem 32 so that we have (3.25). Using $g(u) = u$ and $(3.14)_1$, we obtain, analogous to (3.25),

$$\int_{\mathbb{T}} R_g \phi \,\mathrm{d}\, x = \int_{\mathcal{T}} (\mathcal{P}_1[\partial_x g(\hat{u}^{st})] - f(\hat{u}^t))\phi \,\mathrm{d}\, x = \sum \mu_0 \left[\!\left[\hat{u}^t\right]\!\right] [\![\phi]\!] \tag{3.31}$$

for all $\phi \in \mathbb{V}_1^s$. Since an orthonormal basis of $\mathbb{V}_1^s$ is given by $\{\phi_j, \psi_j\}_{j=0}^{M-1}$ with

$$\begin{aligned} \phi_j(x) &:= \begin{cases} \sqrt{\frac{1}{h}} & : \quad x \in [x_j, x_{j+1}], \\ 0 & : \quad \text{else} \end{cases} \\ \psi_j(x) &:= \begin{cases} \sqrt{\frac{3}{h^3}}(x - \frac{x_j + x_{j+1}}{2}) & : \quad x \in [x_j, x_{j+1}], \\ 0 & : \quad \text{else} \end{cases} \end{aligned} \tag{3.32}$$

we get, for any fixed $t \in [0,T]$,

$$\begin{aligned} &\left\|\mathcal{P}_1[\partial_x g(\hat{u}^{st})] - f(\hat{u}^t)\right\|_{L_2}^2 \\ &\quad = \left(\int_{\mathcal{T}} \Big(\mathcal{P}_1[\partial_x g(\hat{u}^{st})] - f(\hat{u}^t)\Big)\phi_j\right)^2 + \left(\int_{\mathcal{T}} \Big(\mathcal{P}_1[\partial_x g(\hat{u}^{st})] - f(\hat{u}^t)\Big)\psi_j\right)^2 \\ &\quad = \mu_0 \sum_j \left(\frac{1}{h}\big(-\left[\!\left[\hat{u}^t\right]\!\right]_j + \left[\!\left[\hat{u}^t\right]\!\right]_{j+1}\big)^2 + \frac{3}{h}\big(\left[\!\left[\hat{u}^t\right]\!\right]_j + \left[\!\left[\hat{u}^t\right]\!\right]_{j+1}\big)^2\right) \\ &\quad = \frac{\mu_0}{2h} \sum_j \left(4\big(\left[\!\left[\hat{u}^t\right]\!\right]_{j+1}\big)^2 + 4\left[\!\left[\hat{u}^t\right]\!\right]_{j+1}\left[\!\left[\hat{u}^t\right]\!\right]_j + 4\big(\left[\!\left[\hat{u}^t\right]\!\right]_j\big)^2\right) \\ &\quad \geq \frac{\mu_0}{h} \sum_j \left(2\big(\left[\!\left[\hat{u}^t\right]\!\right]_{j+1}\big)^2 + 2\big(\left[\!\left[\hat{u}^t\right]\!\right]_j\big)^2\right) = \frac{2\mu_0}{h} \sum_j \big(\left[\!\left[\hat{u}^t\right]\!\right]_j\big)^2. \end{aligned} \tag{3.33}$$

According to the arguments given in [14, Rem. 6.6], the lower bound derived in (3.33) is of order h^γ even if the error of the method is $\mathcal{O}(h^{1+\gamma})$. ∎

We conclude this section by showing that we can choose a function $\mathbf{w}(\mathbf{a},\mathbf{b})$ such that (3.29) satisfies Assumption 20 (ii) with $\nu = 1$. Using Theorem 32, we thus obtain a reconstruction of optimal order. We restrict ourselves to a DG scheme with polynomials of degree $q \geq 1$ such that, following the discussion in the proof of Theorem 32, we can restrict our attention to the case of arguments $\mathbf{a}, \mathbf{b} \in \mathcal{U}$ of the numerical flux satisfying

$$\left|\mathbf{a} - \mathbf{b}\right| \lesssim h. \tag{3.34}$$

In order to define the reconstruction, we define the characteristic decomposition of $\mathbf{a}, \mathbf{b}$ given by $\boldsymbol{\alpha} := L\mathbf{a}, \boldsymbol{\beta} := L\mathbf{b}$ and will study a $\mathbf{w}$ of the form $\mathbf{w}(\mathbf{a},\mathbf{b}) := R\boldsymbol{\omega}(\boldsymbol{\alpha},\boldsymbol{\beta})$.

To define $\boldsymbol{\omega} = \big(\omega_i(\alpha_i, \beta_i)\big)_{i=1}^m$, we consider a (not strictly) monotone smooth function $\chi : \mathbb{R} \to \mathbb{R}$ such that $\chi(z) = 0$ for $z < -1$ and $\chi(z) = 1$ for $z > 1$. We define $\chi_h(z) := \chi(z/h)$. Then

$$\omega_i(\alpha_i, \beta_i) = \chi_h(\lambda_i)\ \alpha_i + \Big(1 - \chi_h(\lambda_i)\Big)\ \beta_i. \tag{3.35}$$

Note that $\boldsymbol{\omega}$ provides upwinding of the characteristic variables smoothed out so that the reconstruction is smooth enough to allow for an efficient computation of the integrated residual.

Finally, we can now define the function μ:

$$\mu(\mathbf{a}, \mathbf{b}; h) = \frac{\mathbf{g}(\mathbf{w}(\mathbf{a}, \mathbf{b})) - \mathbf{G}(\mathbf{a}, \mathbf{b})}{h\|\mathbf{a} - \mathbf{b}\|^2} \otimes (\mathbf{b} - \mathbf{a}).$$

In order to show that μ can be bounded such that Assumption 20 (ii) is satisfied with $\nu = 1$, we need to prove

$$\big|\mathbf{g}(\mathbf{w}(\mathbf{a}, \mathbf{b})) - \mathbf{G}(\mathbf{a}, \mathbf{b})\big| \overset{!}{\lesssim} h|\mathbf{b} - \mathbf{a}| \quad \text{for } \mathbf{a}, \mathbf{b} \in K \text{ with } \big|\mathbf{a} - \mathbf{b}\big| \lesssim h. \tag{3.36}$$

Using (3.29) and the Taylor expansion, we obtain

$$\begin{aligned}\big|\mathbf{g}(\mathbf{w}(\mathbf{a}, \mathbf{b})) - \mathbf{G}(\mathbf{a}, \mathbf{b})\big| \le\ & \big|A(\mathbf{w}(\mathbf{a}, \mathbf{b}) - \mathbf{c}) - \frac{1}{2}|A|(\mathbf{a} - \mathbf{b})\big| \\ & + \big|(\mathbf{w}(\mathbf{a}, \mathbf{b}) - \mathbf{c})^T\, \mathrm{H}\,\mathbf{g}(\boldsymbol{\xi})(\mathbf{w}(\mathbf{a}, \mathbf{b}) - \mathbf{c})\big| =: S_1 + S_2,\end{aligned}$$

where $\mathrm{H}\,\mathbf{g}$ is a tensor of third order consisting of Hessians of the components of $\mathbf{g}$, and $\boldsymbol{\xi}$ is a convex combination of $\mathbf{w}(\mathbf{a}, \mathbf{b})$, and $\mathbf{c}$.

Defining $\boldsymbol{\gamma} := L\mathbf{c} = \frac{1}{2}(\boldsymbol{\alpha} + \boldsymbol{\beta})$, we can bound S_1 by

$$S_1 = \big|RD(\boldsymbol{\omega}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \boldsymbol{\gamma}) - \frac{1}{2}R|D|(\boldsymbol{\alpha} - \boldsymbol{\beta})\big| \le |R|\ |\boldsymbol{\delta}| \tag{3.37}$$

with $\boldsymbol{\delta} := D(\boldsymbol{\omega}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \boldsymbol{\gamma}) - \frac{1}{2}|D|(\boldsymbol{\alpha} - \boldsymbol{\beta})$.

Now we consider each component of the vector $\boldsymbol{\delta}$ separately, distinguishing two cases.

Case one: $|\boldsymbol{\lambda_i}| \ge \boldsymbol{h}$. We show the computation for $\lambda_i \ge h$, and $\lambda_i \le -h$ is analogous:

$$\lambda_i\big(\omega_i(\alpha_i, \beta_i) - \gamma_i\big) - \frac{1}{2}|\lambda_i|(\alpha_i - \beta_i) \ =\ \lambda_i(\alpha_i - \gamma_i) - \frac{1}{2}\lambda_i(\alpha_i - \beta_i) \ =\ 0. \tag{3.38}$$

Case two: $-\boldsymbol{h} \le \boldsymbol{\lambda_i} \le \boldsymbol{h}$. Now $\big|\lambda_i\big(\omega_i(\alpha_i, \beta_i) - \gamma_i\big) - \frac{1}{2}|\lambda_i|(\alpha_i - \beta_i)\big| \le h|\alpha_i - \beta_i|$. Combining both cases, we get that each component of $\boldsymbol{\delta}$ can be bounded by $h|\alpha_i - \beta_i|$ and thus

$$S_1 \le h|\boldsymbol{\alpha} - \boldsymbol{\beta}| \lesssim h|\mathbf{a} - \mathbf{b}|.$$

In order to bound S_2, we observe

$$\begin{aligned}S_2 \le \big|\mathrm{H}\,\mathbf{g}(\boldsymbol{\xi})\big|\big|\mathbf{w}(\mathbf{a}, \mathbf{b}) - \mathbf{c}\big|^2 &\le \big|\mathrm{H}\,\mathbf{g}(\boldsymbol{\xi})\big|\,|R|^2\,\big|\boldsymbol{\omega}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \boldsymbol{\gamma}\big|^2 \\ &\le \big|\mathrm{H}\,\mathbf{g}(\boldsymbol{\xi})\big|\,|R|^2\,\big|\boldsymbol{\alpha} - \boldsymbol{\beta}\big|^2 \lesssim \big|\mathrm{H}\,\mathbf{g}(\boldsymbol{\xi})\big|h\big|\mathbf{a} - \mathbf{b}\big|,\end{aligned} \tag{3.39}$$

so it remains to show that $|\mathrm{H}\,\mathbf{g}(\boldsymbol{\xi})|$ is bounded uniformly in h for h small enough. It is sufficient to show that $\boldsymbol{\xi}$ is in some compact subset of $\mathcal{U}$ for h small enough. Since K is compact and $\mathcal{U}$ is open, there exists $\varepsilon > 0$ such that $K_\varepsilon := \{x \in \mathbb{R}^m : \mathrm{dist}(x, K) \le \varepsilon\}$ is a compact subset of $\mathcal{U}$. By (3.39) and (3.34), we know $\mathbf{c} \in K$ and $|\boldsymbol{\xi} - \mathbf{c}| \le |\mathbf{w} - \mathbf{c}| \lesssim h$ so that $\boldsymbol{\xi} \in K_\varepsilon$ for h small enough, which completes the proof of (3.36).

Remark 35 (Difficulty for general fluxes). For the Roe-type flux studied here, an optimal reconstruction is possible since the artificial viscosity term $|A|(\mathbf{a}-\mathbf{b})$ matches the linear term in the Taylor expansion of $\mathbf{g}$ so that the vector $\boldsymbol{\delta}$ vanishes in the first case studied above. In the second case it is crucial that we have the same eigenvalue in both terms contributing to δ_i so that both are proportional to $h|\mathbf{a}-\mathbf{b}|$. Take, for example, a viscosity term typically used in local Lax–Friedrichs-type fluxes, i.e., of the form $\max_k|\lambda_k|(\mathbf{a}-\mathbf{b})$. Now consider a situation with $\lambda_1 = 0$, which implies $|\delta_1| = \max_k|\lambda_k|\ |\alpha_1-\beta_1|$. Now if $\alpha_1 \neq \beta_1$ and, for example, $\lambda_2 = 1$, then δ_1 can only be bounded by $|\mathbf{a}-\mathbf{b}|$ without the factor of h, which would be required for an optimal reconstruction.

4. Numerical experiments. In the following we will show some numerical tests verifying the convergence rates presented in the previous section. As pointed out above, smooth reconstructions are preferable since they allow for an efficient computation of the integrated residual. One numerical flux which gives good results as long as the entropy solution is quite regular and allows for optimal estimates with a smooth $\mathbf{w}$ is the Richtmyer flux (3.12), with or without an additional diffusion term. We have performed tests using the following numerical flux:

$$\mathbf{G}(\mathbf{a},\mathbf{b}) = \mathbf{g}(\mathbf{w}(\mathbf{a},\mathbf{b})) - \mu|\mathbf{b}-\mathbf{a}|(\mathbf{b}-\mathbf{a})\ , \quad \mathbf{w}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}+\mathbf{b}}{2} - \frac{\lambda}{2}(\mathbf{g}(\mathbf{b})-\mathbf{g}(\mathbf{a})), \tag{4.1}$$

where $\lambda = \frac{\tau}{h}$ and $\mu = \frac{1}{2}$ or $\mu = 0$. The artificial viscosity provides a simple approximation of a viscosity of the form $h^2\partial_x\big(|\partial_x u|\partial_x u\big)$. Thus we are either in the case of Assumption 20 (ii) ($\mu = \frac{1}{2}$) with $\nu = 1$ or in the case of Assumption 20 (i) ($\mu = 0$), so we can expect optimal convergence of the residuals. We studied both $\mu = 0$ and $\mu = \frac{1}{2}$ in (4.1) but found the difference to be negligible, so we will only show results for one of those choices in the following.

If not stated otherwise, we use an explicit Runge–Kutta method of order r matching the order of the spatial discretization. The temporal reconstruction is based on the Hermite polynomial $H(p,0,0)$ or $H(p,0,-1)$, with p chosen to match the rate of the spatial scheme. In the following we show results with polynomials in space of degrees $q = 1, 2, 3$, and 4. For the temporal reconstruction to match the order of the spatial scheme we thus take the following reconstructions: $H(0,0,-1), H(0,0,0), H(1,0,-1)$, and $H(1,0,0)$, respectively (see Definition 2). Assuming that $u^n, u^{n+1}, f(u^n), f(u^{n+1})$ will need to be stored in the implementation of the time stepping scheme, additional memory is only required for $q \geq 3$. For $q = 3, 4$, additionally u^{n-1} and $f(u^{n-1})$ are required for computing the reconstruction on $[t_n, t_{n+1}]$.

The higher order temporal reconstructions cannot be used during the first steps of the simulation, since they rely on data from previous time steps. In the simulations shown here we have decided to compute the residuals starting from $t = 0.01$ onwards so that we always have enough information to compute the reconstructions. Another approach would have been to compute a sufficient number of steps and use forward Hermite interpolation to start off the simulation. Our analysis would easily extend to cover this case as well.

In all the following we study both the norm of the residual and the error. In each figure the left image shows the values using a logarithmic scale and the corresponding experimental orders of convergence are shown on the right. In the case of systems, the l^2-norm of the error in all components is shown. Note that we do not show the full estimator from (3.16). We only show values of the term $\|\mathbf{R}^{st}\|_{L_2((0,t_n)\times\mathbb{T})}$ involving the residual. Note that we have ignored the exponential factor and error in the initial

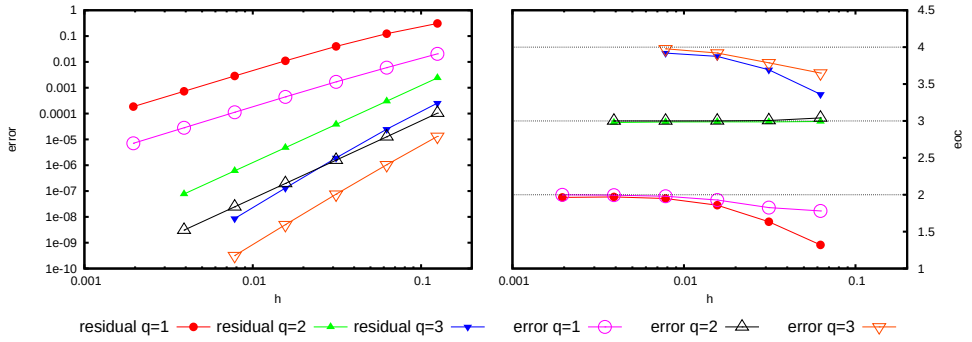

FIG. 1. *Error and residuals for the linear advection problem and polynomial degree* $q = 1, \dots, 3$ *using flux* (4.1) *with* $\mu = \frac{1}{2}$.

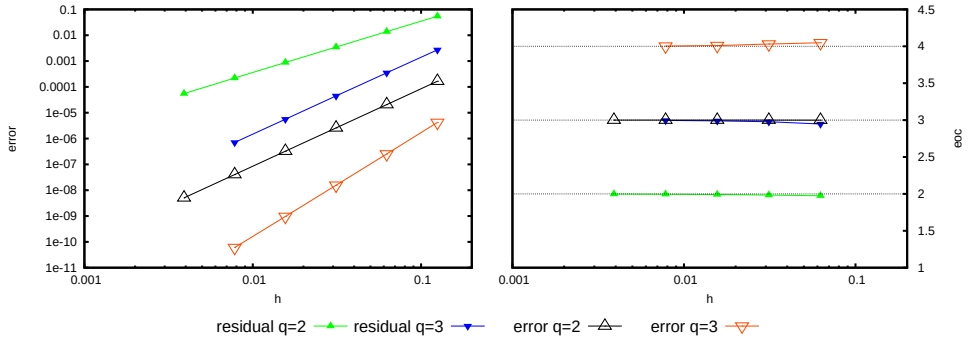

FIG. 2. *Error and residuals for the linear advection problem and polynomial degree* $q = 2, 3$ *using a local Lax–Friedrichs type flux.*

conditions. Also note that in all our tests the first term in the estimator, involving the error of the spatial reconstruction, was of the same order as the residual.

4.1. Linear problem. Consider the linear scalar conservation law

$$\partial_t u + 8\partial_x u = 0$$

in the domain $[0, 2]$ with periodic boundary conditions and initial condition $u(0, x) = 1 - \frac{1}{2}\cos(\pi x)$. The problem is solved up to time $T = 0.4$. We compute the experimental order of convergence starting our simulations with $h = 0.125$ and $\tau = 0.02$ and reducing both h and τ by a factor of 2 in each step. This leads to a CFL constant of about 0.13, which is sufficiently small for all polynomial degrees used in the following simulations.

In Figure 1 we show the values and experimental orders of convergence of the error and the residual for different polynomial degrees. It is evident that for all polynomial degrees $q = 1, \dots, 3$ both the error and the residual converge with the expected order of $q + 1$. In Figure 2 we use a Lax–Friedrichs numerical flux function of the form (3.13). For both $q = 2$ and $q = 3$ the suboptimal convergence of the residual predicted by the theory presented in the previous section is evident. Finally we investigate the influence of the temporal reconstruction on the convergence of the

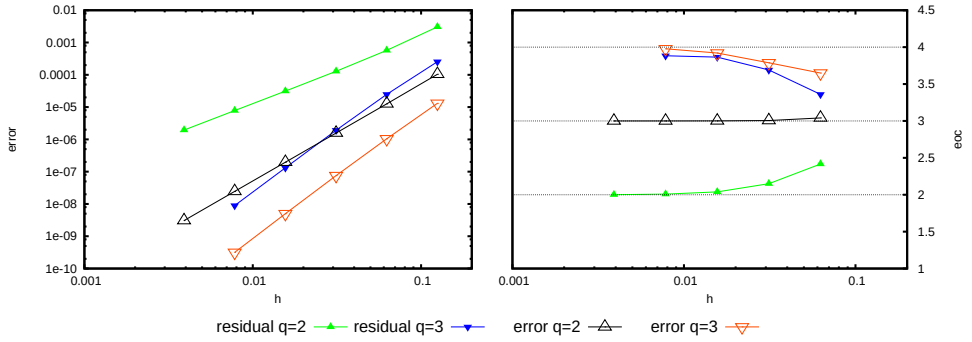

FIG. 3. *Error and residuals for the linear advection problem and polynomial degree* $q = 2, 3$ *using flux* (4.1) *with* $\mu = \frac{1}{2}$ *but a temporal reconstruction of lower order.*

residual in Figure 3. We use a Hermite interpolation one degree lower than our theory demands. For $q = 2$ the quadratic reconstruction is clearly not sufficient for optimal convergence of the residual. For $q = 3$ the results are not conclusive, since the residual is still converging with order 4 for this simple test case.

4.2. Euler equation. We conclude our numerical experiments with some tests using the compressible Euler equations of gas dynamics with an ideal pressure law with adiabatic constant $\gamma = 1.4$. We use the same domain and initial grid as in the previous test case. The time step on the coarsest grid is set to be $\tau = 0.008$. The initial conditions consist of a constant density and velocity of $\rho = 1, u = 0.81$ and a sinusoidal pressure wave $p(x) = 1.3 - \frac{1}{2}\cos(\pi x)$. We again use the flux (4.1) with $\mu = 0$. In Figure 4 a few snapshots in time of the density are displayed. Note how the nonlinearities in the Euler equations lead to the development of shocks between the shown times 1.2 and 1.4. Note that our estimator is not convergent (for $h \to 0$) for discontinuous solutions; cf. Remark 27. The residual calculated for the density clearly captures the onset of the shocks and grows by many orders of magnitude once the shock has fully developed; see Figure 4. Note that we use no limiters here, so the method is beginning to show signs of instabilities within the shock regions.

We do not have an exact solution in this case and therefore we cannot compute the convergence rate of the scheme directly. But since the solution clearly remains smooth up to $t = 1$, we can expect our residual to converge with optimal order up to this point in time. This is confirmed by the results shown in Figure 5.

5. Conclusions. In this work we extended previous results on a posteriori error estimation based on relative entropy to fully discrete schemes. The resulting estimator requires the computation of first a temporal and then a spatial reconstruction of the solution. The temporal reconstruction is based on a Hermite interpolation and is independent of the time stepping method used. The spatial reconstruction uses the idea presented in [14]. As noted there, the reconstruction and the flux have to be chosen carefully to guarantee that the estimator is optimal in the sense that it converges with the same order as the scheme (in case the scheme converges). We managed to generalize the assumptions on the numerical flux, greatly extending the class of schemes for which we can prove optimal convergence of the estimator. Lax–Wendroff-type fluxes with artificial diffusion are now included as well as the Roe flux if upwinding in the characteristic variables is used to define the reconstruction—simple

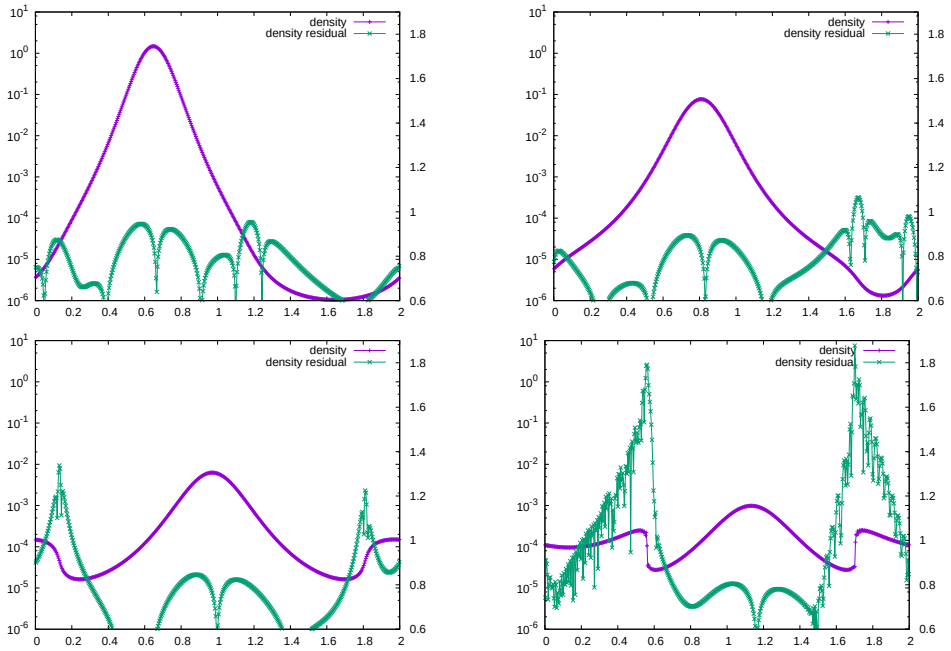

FIG. 4. *Time evolution of the density for the Euler test case for $q = 2$. From left to right and top to bottom: $t = 0.8, 1.0, 1.2, 1.4$. We show the density with scaling indicated on the right axis. In addition, we show the residual in the density using the left axis in log scaling.*

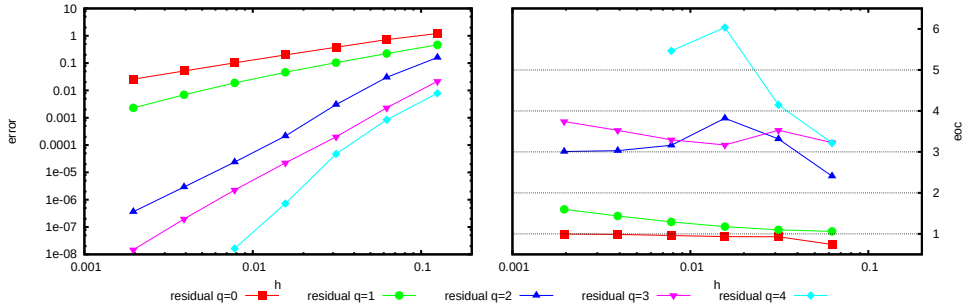

FIG. 5. *Error and residuals for the Euler problem and polynomial degree $q = 0, \ldots, 4$ using flux* (4.1) *with $\mu = 0$.*

averaging, on the other hand, leads to suboptimal convergence, as we have shown. So we have managed to prove that in general the requirements stated are not only sufficient but in fact necessary, and are demonstrated suboptimality numerically.

In future work, we plan to extend the results to higher space dimensions. Also, we will investigate how the residual can be used to drive grid adaptation and possibly to detect regions of discontinuities in the solution required in the design of stabilization methods. These could either be based on local artificial diffusion or on limiters typically used together with DG methods.

REFERENCES

[1] R. Abgrall, D. Aregba, C. Berthon, M. Castro, and C. Parés, *Preface [Special issue: Numerical approximations of hyperbolic systems with source terms and applications]*, J. Sci. Comput., 48 (2011), pp. 1–2, doi:10.1007/s10915-011-9497-z.

[2] C. Arvanitis, C. Makridakis, and A. E. Tzavaras, *Stability and convergence of a class of finite element schemes for hyperbolic systems of conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 1357–1393, doi:10.1137/S0036142902420436.

[3] E. Bänsch, F. Karakatsani, and C. Makridakis, *A posteriori error control for fully discrete Crank–Nicolson schemes*, SIAM J. Numer. Anal., 50 (2012), pp. 2845–2872, doi:10.1137/110839424.

[4] E. Chiodaroli, *A counterexample to well-posedness of entropy solutions to the compressible Euler system*, J. Hyperbolic Differ. Equ., 11 (2014), pp. 493–519, doi:10.1142/S0219891614500143.

[5] C. M. Dafermos, *The second law of thermodynamics and stability*, Arch. Ration. Mech. Anal., 70 (1979), pp. 167–179, doi:10.1007/BF00250353.

[6] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] 325, 3rd ed., Springer, Berlin, 2010, doi:10.1007/978-3-642-04048-1.

[7] C. De Lellis and L. Székelyhidi, Jr., *On admissibility criteria for weak solutions of the Euler equations*, Arch. Ration. Mech. Anal., 195 (2010), pp. 225–260, doi:10.1007/s00205-008-0201-x.

[8] A. Dedner, C. Makridakis, and M. Ohlberger, *Error control for a class of Runge–Kutta discontinuous Galerkin methods for nonlinear conservation laws*, SIAM J. Numer. Anal., 45 (2007), pp. 514–538, doi:10.1137/050624248.

[9] D. A. Di Pietro and A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*, Mathématiques et applications, Springer, Heidelberg, 2012.

[10] R. J. DiPerna, *Uniqueness of solutions to hyperbolic conservation laws*, Indiana Univ. Math. J., 28 (1979), pp. 137–188, doi:10.1512/iumj.1979.28.28011.

[11] W. H. Enright, K. R. Jackson, S. P. Nørsett, and P. G. Thomsen, *Interpolants for Runge–Kutta formulas*, ACM Trans. Math. Software, 12 (1986), pp. 193–218, doi:10.1145/7921.7923.

[12] R. Eymard, T. Gallouët, M. Ghilani, and R. Herbin, *Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes*, IMA J. Numer. Anal., 18 (1998), pp. 563–594, doi:10.1093/imanum/18.4.563.

[13] B. Fornberg, *Generation of finite difference formulas on arbitrarily spaced grids*, Math. Comp., 51 (1988), pp. 699–706, doi:10.2307/2008770.

[14] J. Giesselmann, C. Makridakis, and T. Pryer, *A posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws*, SIAM J. Numer. Anal., 53 (2015), pp. 1280–1303, doi:10.1137/140970999.

[15] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Applied Mathematical Sciences 118, Springer, New York, 1996.

[16] L. Gosse and C. Makridakis, *Two a posteriori error estimates for one-dimensional scalar conservation laws*, SIAM J. Numer. Anal., 38 (2000), pp. 964–988, doi:10.1137/S0036142999350383.

[17] R. Hartmann and P. Houston, *Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 979–1004, doi:10.1137/S1064827501389084.

[18] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*, Texts in Applied Mathematics 54, Springer, New York, 2008, doi:10.1007/978-0-387-72067-8.

[19] D. J. Higham, *Runge–Kutta defect control using Hermite–Birkhoff interpolation*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 991–999, doi:10.1137/0912053.

[20] V. Jovanović and C. Rohde, *Finite-volume schemes for Friedrichs systems in multiple space dimensions: A priori and a posteriori error estimates*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 104–131, doi:10.1002/num.20026.

[21] V. Jovanović and C. Rohde, *Error estimates for finite volume approximations of classical solutions for nonlinear systems of hyperbolic balance laws*, SIAM J. Numer. Anal., 43 (2006), pp. 2423–2449, doi:10.1137/S0036142903438136.

[22] H. Kim, M. Laforest, and D. Yoon, *An adaptive version of Glimm's scheme*, Acta Math. Sci. Ser. B Engl. Ed., 30 (2010), pp. 428–446, doi:10.1016/S0252-9602(10)60057-4.

[23] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley–Teubner Series Advances in Numerical Mathematics, John Wiley & Sons, Chichester, UK, 1997.

[24] D. Kröner and M. Ohlberger, *A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions*, Math. Comp., 69 (2000), pp. 25–39, doi:10.1090/S0025-5718-99-01158-8.

[25] S. N. Kružkov, *First order quasilinear equations with several independent variables*, Mat. Sb. (N.S.), 81 (1970), pp. 228–255.

[26] M. Laforest, *A posteriori error estimate for front-tracking: Systems of conservation laws*, SIAM J. Math. Anal., 35 (2004), pp. 1347–1370, doi:10.1137/S0036141002416870.

[27] M. Laforest, *An a posteriori error estimate for Glimm's scheme*, in Hyperbolic Problems: Theory, Numerics, Applications, Springer, Berlin, 2008, pp. 643–651.

[28] A. Lapidus, *A detached shock calculation by second-order finite differences*, J. Comput. Phys., 2 (1967), pp. 154–177.

[29] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2002, doi:10.1017/CBO9780511791253.

[30] C. Makridakis, *Space and time reconstructions in a posteriori analysis of evolution problems*, ESAIM Proc., 21 (2007), pp. 31–44, doi:10.1051/proc:072104.

[31] C. Makridakis and R. H. Nochetto, *A posteriori error analysis for higher order dissipative methods for evolution problems*, Numer. Math., 104 (2006), pp. 489–514, doi:10.1007/s00211-006-0013-6.

[32] G. Puppo and M. Semplice, *Numerical entropy and adaptivity for finite volume schemes*, Commun. Comput. Phys., 10 (2011), pp. 1132–1160, doi:10.4208/cicp.250909.210111a.

[33] J. Qiu, B. C. Khoo, and C.-W. Shu, *A numerical study for the performance of the Runge–Kutta discontinuous Galerkin method based on different numerical fluxes*, J. Comput. Phys., 212 (2006), pp. 540–565, doi:10.1016/j.jcp.2005.07.011.

[34] J. Reisner, J. Serencsa, and S. Shkoller, *A space-time smooth artificial viscosity method for nonlinear conservation laws*, J. Comput. Phys., 235 (2013), pp. 912–933.

[35] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Texts in Applied Mathematics 12, 3rd ed., Springer, New York, 2002.

[36] E. F. Toro and S. J. Billett, *Centred TVD schemes for hyperbolic conservation laws*, IMA J. Numer. Anal., 20 (2000), pp. 47–79, doi:10.1093/imanum/20.1.47.

[37] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws*, SIAM J. Numer. Anal., 44 (2006), pp. 1703–1720, doi:10.1137/040620382.