Stabilizing Discontinuous Galerkin Methods Using Dafermos' Entropy Rate Criterion: I—One-Dimensional Conservation Laws

Simon-Christian Klein[1] ^(ORCID)

Abstract
A novel approach for the stabilization of the discontinuous Galerkin method based on the Dafermos entropy rate crition is presented. The approach is centered around the efficient solution of linear or nonlinear optimization problems in every timestep as a correction to the basic discontinuous Galerkin scheme. The thereby enforced Dafermos criterion results in improved stability compared to the basic method while retaining a high order of accuracy in numerical experiments for scalar conservation laws. Further modification of the optimization problem allows also to enforce classical entropy inequalities for the scheme. The proposed stabilization is therefore an alternative to flux-differencing to enforce entropy inequalities. As the shock-capturing abilities of the scheme are also enhanced is the method also an alternative to finite-volume subcells, artificial viscosity, modal filtering, and other shock capturing procedures in one space dimension. Tests are carried out for Burgers' equation.

Keywords Discontinuous Galerkin methods · Entropy stability · High-order methods · Finite-volume methods · Entropy rate criterion

Mathematics Subject Classification 35L03 · 35L45 · 35L65 · 35L67 · 65M08 · 65M12 · 65M20 · 76L05

1 Introduction

In this work, a novel shock-capturing approach for discontinuous Galerkin (DG) schemes is proposed. The first subsection of the introduction gives a short reminder of the basic theory of hyperbolic conservation laws and the Dafermos entropy rate criterion, which is the basis for the shock-capturing technique. A second subsection covers the basic discontinuous Galerkin

---

✉ Simon-Christian Klein
    simon-christian.klein@tu-bs.de

[1] Institute for Partial Differential Equations, TU Braunschweig, Braunschweig, Germany

Table 1 Used notation

A cell of the subdivision $\mathcal{T}$ of the domain Ω	T
The left and right boundaries of cell T	T_l, T_r
The space of ansatzfunctions for an cell T	V^T
L^2 projection of u onto V	$\mathbb{P}_V u$
Interpolation of u on V w.r.t. the collocation points $(\xi_k)_{k=1}^N$	$\mathbb{I}_V u$
Vector of nodal values in cell T at time t	$u^T(t)$
Ansatz function in cell T at position x and time t	$u^T(x, t)$
The total entropy in cell T	$E_{u,T}(t)$
The discrete total entropy in cell T	$E^T(t)$
Entropy variables in cell T	$\frac{\partial U}{\partial u}(u^T(x, t))$
Vector of nodal values of the entropy variables in cell T at t	$\frac{\partial U}{\partial u}^T(t)$
Interpolation of the entropy variables in cell T on V	$\frac{\partial U}{\partial u}^T(x, t)$
Numerical quadrature of $f(x)$ on the intervall $[a, b]$	$\int_a^b f(x)\,\mathrm{d}x = \sum_k \omega_k f(x_k)$
The canonical inner product between $a, b \in \mathbb{R}^n$	$a \cdot b$ or $(a)\cdot(b)$
The inner product between u and v on cell T	$\langle u, v\rangle_T$
The p norm of u in cell T	$\|u\|_{T,p}$
Exact solution to the initial condition $u(x, t_0)$ after $t - t_0$	$H(u(\cdot, t_0), t - t_0)$
Mean value of subcell k of N, u^T as initial condition	$u_k^{T,N}$

As a general rule, quantities with only t as an argument are a vector of nodal values at a certain time. Values with x and t in their argument list are functions that were evaluated at these values. Objects with T added as exponent are approximations of the quantity in the cell T

framework that is used as a base scheme for our construction. Section 2 outlines the general idea used for the correction, while Sects. 3 and 4 give a preliminary analysis of the properties of the modification. We close our presentation with numerical tests for Burgers' equation, a scalar conservation law, in Sect. 5 and an conclusion.

The used notation is summarized in Table 1.

1.1 Hyperbolic Conservation Laws

In this work we will consider one-dimensional systems of conservation laws for m conserved quantities [12, 30],

$$\frac{\partial u(x, t)}{\partial t} + \frac{\partial f \circ u(x, t)}{\partial x} = 0 \quad \text{for} \quad u(x, t) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^m, \quad f : \mathbb{R}^m \to \mathbb{R}^m, \quad (1)$$

where $f \circ u(x, t) = f(u(x, t))$ denotes the composition of f and u. As classical solutions to (1) can break down in finite time [30] one considers weak solutions u, which satisfy

$$\int_0^\infty \int_\mathbb{R} \left(u(x, t)\right)\cdot\left(\frac{\partial\varphi(x, t)}{\partial t}\right) + \left(f \circ u(x, t)\right)\cdot\left(\frac{\partial\varphi(x, t)}{\partial x}\right) \, \mathrm{d}x\,\mathrm{d}t$$
$$+ \int_\mathbb{R} (u(x, 0))\cdot(\varphi(x, 0)) \, \mathrm{d}x = 0. \tag{2}$$

This can be shown for classical solutions by multiplying with a suitable test function φ : $\mathbb{R} \times \mathbb{R} \to \mathbb{R}^m$, integrating over the domain and using integration by parts [29]. Sadly, weak solutions are not unique [28] and one therefore hopes that additional constraints in form of entropy inequalities,

$$\frac{\partial U \circ u}{\partial t} + \frac{\partial F \circ u}{\partial x} \leq 0, \tag{3}$$

single out the relevant solution [30]. In this inequality $U : \mathbb{R}^m \to \mathbb{R}$ is the convex entropy function and $F : \mathbb{R}^m \to \mathbb{R}$ the entropy flux satisfying

$$\frac{\mathrm{d}U}{\mathrm{d}u} \frac{\mathrm{d}f}{\mathrm{d}u} = \frac{\mathrm{d}F}{\mathrm{d}u}.$$

One can show that a pair of functions (U, F), satisfying this relation, induces an additional conservation law for smooth classical solutions. The entropy inequality for weak vanishing viscosity solution motivates the entropy inequality, as given above [29]. Entropy inequalities are not the only criterion one can use to shrink the set of admissible solutions. Dafermos proposed an entropy rate criterion in his seminal work [10]. After defining the total entropy in the domain,

$$E_u(t) = \int_{\mathbb{R}} U \circ u(x, t) \,\mathrm{d}x,$$

he conjectures that the physically relevant solution u is the one satisfying

$$\forall t > 0 : \quad \frac{\mathrm{d}E_u(t)}{\mathrm{d}t} \leq \frac{\mathrm{d}E_{\tilde{u}}(t)}{\mathrm{d}t}$$

compared to all other weak solutions $\tilde{u}$. See [9, 11, 15] for theoretical examples where the Dafermos criterion is able to reduce the amount of admissible solutions.

1.2 Discontinuous Galerkin Methods

Discontinuous Galerkin (DG) methods, first considered for time dependent hyperbolic problems in [2, 7], can be derived as a generalization of classical Finite Volume (FV) methods [34]. See [4, 5, 24] for a general introduction to DG methods. After the problem domain has been subdivided into a set $\mathcal{T}$ of disjoint cells one can select a basis of ansatzfunctions

$$B^T = \{\varphi_1^T(x), \varphi_2^T(x), \ldots, \varphi_M^T(x)\}$$

associated to every cell $T \in \mathcal{T}$. One further conjects that the numerical solution $u(x, t)$ is a linear combination of these basis functions on every cell and only piecewise continuous in space. Discontinuities are only allowed to occur at cell interfaces. We denote the space of ansatzfunctions on cell T by $V^T = \operatorname{span} B^T$. Often, the index T will be omitted. One therefore writes

$$u(x, t) = \sum_{T \in \mathcal{T}} \sum_{j \in T} u_j^T(t) \varphi_j^T(x) \chi^T(x)$$

as an ansatz for a time dependent function $u(\cdot, t)|_T \in V^T$ with the time dependent coefficients $u_j^T(t)$ and the characteristic functions $\chi^T(x)$ of every cell. A time evolution equation for our ansatz can be derived in every cell $T \in \mathcal{T}$ by multiplying the conservation law by another

function $v \in V^T$, using partial integration

$$\int_T \left(\frac{\partial v}{\partial x}\right) \cdot \left(f(u^T(x,t))\right) \, \mathrm{d}x - [v \cdot f(u^T(x,t))]_{T_l}^{T_r} - \int_T \left(v\right) \cdot \left(\frac{\partial u^T(x,t)}{\partial t}\right) \, \mathrm{d}x = 0,$$

(4)

and requiring that this equation holds for all $v \in V^T$ in the same way weak solutions are defined in (2). One can rewrite (4) as

$$\forall T \in \mathcal{T}, v \in V : \left\langle \frac{\partial v}{\partial x}, f \right\rangle_T - [v, f]_T - \left\langle v, \frac{\partial u^T}{\partial t} \right\rangle_T = 0$$

using the inner products

$$\langle u, v \rangle_T = \int_T (u) \cdot (v) \, \mathrm{d}x, \quad [v, f]_T = \int_{\partial T} (v) \cdot (f) \, \mathrm{d}O.$$

Evaluating these inner products for $u, v \in V^T$ is done using the basis representations $(u_k)_{k=1}^M$, $(v_k)_{k=1}^M$ of u and v and the Grammian matrices M^T and S^T associated with cell T,

$$M_{k,l}^T = \langle \varphi_k, \varphi_l \rangle_T, \quad S_{k,l}^T = \left\langle \frac{\partial \varphi_k}{\partial x}, \varphi_l \right\rangle_T.$$

Clearly, this is possible because of their linearity

$$\langle u(\cdot, t), v(\cdot, t) \rangle_T = \left\langle \sum_k u_k \varphi_k, \sum_l v_l \varphi_l \right\rangle_T = \sum_{k,l} u_k \langle \varphi_k, \varphi_l \rangle_T v_l = (u) \cdot (Mv).$$

A 2-norm is also induced by the inner product $\langle \cdot, \cdot \rangle$ and will later play an import role

$$\left\| u^T(t) \right\|_{T,2} = \sqrt{\langle u^T(t), u^T(t) \rangle_T} = \sqrt{(u^T(t)) \cdot (Mu^T(t))},$$

together with the integral functional on V^T

$$V^T \ni u^T \mapsto \int_T u^T(x,t) \, \mathrm{d}x = \left\langle 1, u^T \right\rangle_T = (1) \cdot (Mu^T).$$

Please note that the exact integration of u^T depends on the inclusion of constants into V [21], which will be assumed as given from now on. For nonlinear fluxes the chained function $f \circ u$ is not necessarily in V. We therefore must find a different method to evaluate the inner products involving f at least approximately. Approximating f in the space V via a projection $\mathbb{I}_V : \mathbb{R}^{\mathbb{R}} \to V$ with $\mathbb{I}_V = \mathbb{I}_V^2$ is a common method in this case, leading to the modified scheme

$$\forall v \in V : \left\langle \frac{\partial v}{\partial x}, \mathbb{I}_V f \right\rangle_T - [v, f]_T - \left\langle v, \frac{\partial u}{\partial t} \right\rangle_T = 0.$$

(5)

A suitable projection is the interpolation of f using suitable collocation points $\xi_k \in T$, i.e.

$$\mathbb{I}_V f(x) = \sum_k \varphi_k(x) f_k, \text{ with } (f_k)_{k=1}^M \in \mathbb{R}^M \text{ satisfying } \forall \xi_k : f(\xi_k) = \mathbb{I}_V f(\xi_k).$$

While one could also think of least squares projections $\mathbb{P}_V$, these are as hard to calculate as the sought after inner product [24]. In what follows, we use the Gauß–Lobatto points as collocation points [36] and assume that the matrices M^T, S^T are given with respect to the

Lagrange polynomials of ξ_k, i.e. our basis B are the Lagrange polynomials. The practical implementation should reside also to the usage of Legendre polynomials [24]. The values of u at the element boundaries are not uniquely determined and neither are their fluxes. One therefore uses monotone FV two-point fluxes as approximation of the flux over the boundary $f^* \approx f\left(\lim_{h\uparrow 0} u^T\left(x_{k+\frac{1}{2}} + h, t\right), \lim_{h\downarrow 0} u^T\left(x_{k+\frac{1}{2}} + h, t\right)\right)$, where the limits of $u^T(x, t)$ are entered into the two arguments. As we introduced constants into our approximation space,

$$\left\langle \mathbb{1}, \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\rangle_T = \underbrace{\left\langle \frac{\partial \mathbb{1}}{\partial x}, \mathbb{I}_V\, f \right\rangle_T}_{=0} - [\mathbb{1}, f^*]_T = f_l^* - f_r^*,$$

holds. The DG method can therefore be seen as a FV method that uses classical FV fluxes to determine the flux of conserved quantities between cells but also advances an ansatz inside the cells using a finite element ansatz. An especially popular choice for the inter cell flux is the (local) Lax–Friedrichs flux [6, 29]. The complete scheme can be rewritten in vector matrix notation as

$$M^T \frac{\mathrm{d}u^T}{\mathrm{d}t} = S^T f(u^T(t)) - \begin{pmatrix} \varphi_1^T(x_r) f_r^* - \varphi_1^T(x_l) f_l^* \\ \vdots \\ \varphi_N^T(x_r) f_r^* - \varphi_N^T(x_l) f_r^* \end{pmatrix}.$$

The resulting scheme can be integrated in time, for instance, by strong stability preserving Runge–Kutta methods [23, 44] or any other solver for ordinary differential equations. We will not consider the problem that if the semidiscrete scheme satisfies some bound or property will the numerical solution to the ODE system in general *not* satisfy this bound. Refer to [40] and [37] for an outline of the problems and a set of possible solutions. Classical DG type methods use total variation limiters [6] and troubled cell indicators that switch a cell into a first order FV mode [6] or reconstruct the ansatz from neighboring cells to enforce stability and improve their abilities to calculate discontinuous solutions. More recent methods using the summation by parts (SBP) property [3, 17, 41] in conjunction with flux differencing [16] allow to obtain stability results without the usage of limiters. Another option are the techniques of Abgrall, who introduced a general entropy correction term [1]. While the aforementioned results provide entropy stability, the resulting schemes often need additional stabilization to calculate solutions containing shocks. We will design two similar entropy correction techniques in this work for the usage in the DG framework based on Dafermos' entropy rate criterion that also serves as a shock-capturing technique. Prior usage of the Dafermos entropy rate criterion was based on solving the variational problem given by the entropy rate criterion in a convex subset of the fluxes [26]. In this work, the entropy rate criterion will be applied to the approximate solutions inside every cell of the DG scheme while the fluxes between cells will be left untouched. Therefore existing reasonable numerical fluxes can still be used. One should further note that the presented correction operations can be generalized to multiblock SBP schemes and the author will consider these methods in the future. The presented techniques will be tested here only without SBP-SAT and flux-differencing techniques to underline their abilities.

2 Dafermos Modification of DG Methods

In the last chapter the DG schemes were introduced. Let us denote an assumed exact solution operator as $H(u(\cdot, t_0), \tau)$ that maps an initial condition $u(\cdot, t_0)$ onto the solution at time $t_0 + \tau$. The finite dimensional vector space V extorts several approximations, like the projection $\mathbb{I}_V$ of nonlinear fluxes and that the time derivative of the coefficients $\frac{du^T}{dt} \in V$ satisfies the weak formulation only for $v \in V$. The time discretization is therefore in general different from the projection of $\frac{\partial H(u(\cdot,t),\tau)}{\partial \tau}$ onto V and also this projection differs from $\frac{\partial H(u(\cdot,t),\tau)}{\partial \tau}$. We therefore select a trajectory from the initial condition leading away from the exact solution u or its projection onto V. One further knows that weak solutions, which are standard for hyperbolic conservation laws and built into the DG method, allow more than one path for an exact weak solution. We will therefore modify the outlined vanilla DG scheme to rule out unwanted deviations from the assumed trajectory of the assumed exact entropy solution. Unwanted deviations are deviations that lead to instabilities, non-admissible solutions, oscillations or high approximation errors. It is therefore conjectured that changing the trajectory $\frac{du^T(t)}{dt}$ of our approximate solution to satisfy a modified Dafermos entropy rate criterion performs the needed deviation to rule out these problems. We begin the construction of our modification by the definition of a per cell total entropy as

$$E_{u,T}(t) = \int_T U(u^T(x,t)) \, dx = \left\langle \mathbb{1}, U(u^T(x,t)) \right\rangle_T \approx \left\langle \mathbb{1}, U^T(t) \right\rangle_T = E^T(t),$$

where $U^T(t) = U(u^T(t))$ is the vector composed of the values of $U(u)$ evaluated with the vector of nodal values $u^T(t)$. We will see that our numerical entropy functional for DG schemes is (strictly) convex and localy Lipschitz continuous under certain circumstances. One of the needed ingredients is a quadrature formula with perfect stability [19, 20].

Definition 1 (Perfectly Stable Cubature [19, 20]) A quadrature or cubature formula ω_k is termed perfectly stable if

$$\forall k : \omega_k \geq 0$$

holds and the cubature formula is exact for constants, i.e. $\sum_k \omega_k = \mu(\Omega)$, where $\mu(\Omega) = \int_\Omega 1 \, dx$ shall be the measure of the integrated region Ω.

We will now state some useful properties of our discrete entropy functional condensed into the following lemma.

Lemma 1 *Let ω_k be a perfectly stable quadrature formula on the cell T. Then the nonlinear functional*

$$E^T(v) = \sum_k \omega_k U(v(x_k))$$

is local Lipschitz continuous. If U is a strictly convex entropy, then E^T is strictly convex. Further, the entropy functional is exact for constant functions and an upper bound for the entropy of the mean value in the cell

$$U(\overline{u}) = U\left(\frac{\int u_T \, dx}{\mu(T)}\right) \leq \frac{E^T(u)}{\mu(T)},$$

as used as per cell entropy in FV methods.

Proof We begin by showing that the entropy functional is convex, if U is convex. Let $\lambda \in [0, 1]$ and $u, v \in V$ be arbitrary. The perfect stability of the quadrature implies

$$E^T(\lambda u + (1-\lambda)v) = \sum_k \omega_k U(\lambda u(x_k) + (1-\lambda)v(x_k))$$

$$\leq \lambda \sum_k \omega_k U(u(x_k)) + (1-\lambda) \sum_k \omega_k U(v(x_k))$$

$$= \lambda E^T(u) + (1-\lambda)E^T(v).$$

If U is even strictly convex so is also E^T because in this case ≫strictly less≪ can be put in place of ≫less≪ in the above derivation if $u \neq v$. One further knows that E^T is local Lipschitz continuous in the $\|\cdot\|_\infty$ norm with constant $L = L_U$ if U is local Lipschitz continuous with constant $\mu(T)L_U$

$$\left| E^T(u) - E^T(v) \right| \leq \sum_k \omega_k \left| U(u(x_k)) - U(v(x_k)) \right|$$

$$\leq \sum_k \omega_k L_U \|u - v\|_\infty = \mu(T)L_U \|u - v\|_\infty.$$

To show that the functional is exact for constants we just note that our quadrature is exact for constants and therefore

$$E^T(\bar{v}) = \sum_k \omega_k U(\bar{v}(x_k)) = \sum_k \omega_k U(\bar{v}) = U(\bar{v})$$

holds. The last property follows because the quadrature is exact for constants and all $v \in V$ and the calculation of the mean value using this quadrature results in a convex combination of point evaluations

$$U(\bar{v}) = U\left(\frac{\sum_k \omega_k v(x_k)}{\sum_k \omega_k} \right) \leq \frac{\sum_k \omega_k U(v(x_k))}{\sum_k \omega_k} = \frac{E^T(v)}{\mu(T)}.$$

□

Clearly, one now aims to deviate into the reduction of this nonlinear but convex functional, or semidiscrecetely stated to minimize the derivative of this functional with respect to time

$$\frac{dE_{u,T}(t)}{dt} = \int_T \frac{dU}{du}\frac{\partial u}{\partial t} = \left\langle \frac{dU}{du}, \frac{\partial u}{\partial t} \right\rangle_T \approx \left\langle \frac{dU^T}{du}, \frac{du^T}{dt} \right\rangle_T \approx \frac{dE^T}{dt},$$

towards the smallest values by varying $\frac{du^T}{dt}$. It should be stressed, that our numerical approximation of the time derivative of $E_{u,T}$ is *not equivalent* to the time derivative of our numerical approximation of $E_{u,T}$, as the mass matrix M is in general not diagonal

$$\left\langle \frac{dU^T}{du}, \frac{du^T}{dt} \right\rangle_T = \left\langle \frac{dU^T}{du}, M\frac{du^T}{dt} \right\rangle \neq \left\langle \mathbb{1}, M\left(\frac{dU^T}{du}\frac{du^T}{dt} \right) \right\rangle = \left\langle \mathbb{1}, \frac{dU^T}{du}\frac{du^T}{dt} \right\rangle_T.$$

But as we are interested in discretizing the steepest entropy descent, and not the steepest discrete entropy descent, the first form will be used. Our modification will result in the following scheme in the discrete case in every step.

1. Given a state $u^T(t^n)$ calculate the next step $\tilde{u}^T(t^{n+1})$ using a vanilla RK-DG method as outlined in the last chapter.

2. Calculate an error estimate δ^T for the solution $\tilde{u}^T(t^{n+1})$, i.e. an reasonably small $\delta^T \geq 0$ with the property

$$\left\| H(u^T(t^n), \Delta t) - \tilde{u}^T(t^{n+1}) \right\|_T \leq \delta^T,$$

   without complete knowledge of the exact solution operator $H(\cdot, \Delta t)$.
3. Solve the optimization problem

$$u^T(t^{n+1}) = \mathrm{argmin}_{u \in Z} \, E^T(u), \text{ with}$$

$$Z = \left\{ u \in \mathbb{R}^n \,\middle|\, \langle 1, u \rangle_T = \langle 1, \tilde{u}^T(t^{n+1}) \rangle_T \wedge \left\| u - \tilde{u}^T(t^{n+1}) \right\|_T \leq \varepsilon(\delta^T) \right\}, \quad (6)$$

   where ε is a given function of the error δ and $u^T(t^n)$.

This is a direct statement of the modified Dafermos theorem pioneered by the author in [26]. The original Dafermos entropy rate criterion is thereby augmented with the three additional constraints:

- The introduced additional approximation error by the dissipation should be small, i.e. the dissipation should relate to the assumed error towards the exact solution. We would like to enforce $\lim_{\delta \to 0} \varepsilon = 0$.
- The resulting discretisation should still be conservative as defined in [43].
- The entropy rate criterion should not only hold global, but also local, i.e. in every cell. This mirrors the reduction of a global problem in the calculus of variations to the Euler-Lagrange equations, that state a local condition.

These three properties are engraved into the statement that the solution of the optimization problem should still have the same mean value in the cell, i.e. the basic FV method is unchanged, and that the error introduced by the dissipation is smaller than the bound $\varepsilon(\delta)$. We will therefore call this scheme Dafermos Runge–Kutta Discontinuous Galerkin (DRKDG). While the exact solution of this problem in every time step is not feasible it should be noted that, if a strictly convex entropy U is used, the resulting optimization problem has an unique solution, c.f. Sect. 4.

Remark 1 The exact effect of step 3 depends on the selected entropy and the selected norm used for the restriction, and in general the effect is not the same as scaling to the mean value. This can be seen from the example shown in Fig. 1 below. There, the situation is sketched in the hyperplane $\langle 1, \tilde{u}^T \rangle_T = \langle 1, u \rangle_T$. As this restriction is just a hyperplane are convex functions once more convex, especially the entropy and the norm. The next value of a vanilla DG method $\tilde{u}^T(t^{n+1})$ was added with the radius of a norm ball given by the error estimator. This can be thought of the set Z in this picture, as the entire situation is drawn in the hyperplane of constant mean value of u, i.e. the second condition for the set Z is automatically satisfied for all points in the plane. A minimizer of the entropy in the entire plane is given by the constant function $\bar{u}$ with the same mean value, as can be seen from Lemma 3. Scaling towards the mean value results in movement on a straight line between these two points. A solution $u^T(t^{n+1})$ of the optimization problem is on the other hand given by an intersection of the lowest contour line that touches the norm ball, and the norm ball around $\tilde{u}^T$. Obviously, this point is in general not part of the straight line. It should be noted that this could be the case if the contour lines of the entropy are circles them-self, but this is not even the case for scalar conservation laws, as the quadrature rule will in general have varying weights.
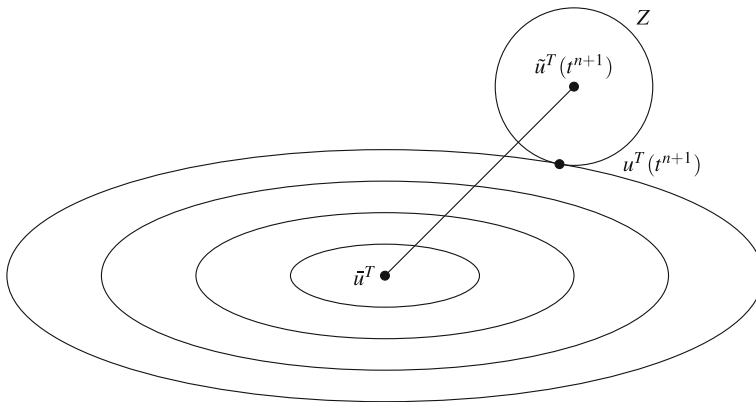
Fig. 1 Sketch of the solution of the optimization problem from step 3 of the DRKDG scheme

A related algorithm also can be given for a semi-discrete scheme in the following form.

1. Calculate the time derivative $\frac{\mathrm{d}\tilde{u}^T}{\mathrm{d}t}$ of the semidiscretisation using a vanilla DG scheme.
2. Estimate the error δ^T in cell T, i.e. find $\delta^T \geq 0$ with

$$\left\| \frac{\mathrm{d}\tilde{u}^T}{\mathrm{d}t} + \frac{\partial f(u^T(x,t))}{\partial x} \right\|_T \leq \delta^T$$

   in a suitable norm. In this case $\frac{\partial f(u^T(x,t))}{\partial x}$ must not be understood as a classical derivative, as it will exist only in a distributional sense.
3. Solve the optimization problem

$$\frac{\mathrm{d}u^T}{\mathrm{d}t} = \mathrm{argmin}_{u_t \in Z} \frac{\mathrm{d}E^T}{\mathrm{d}t}\left(u^T, u_t\right) \text{ with}$$
$$Z = \left\{ u_t \in \mathbb{R} \,\middle|\, \langle 1, u_t \rangle_T = 0 \wedge \left\| u_t - \frac{\mathrm{d}\tilde{u}_k}{\mathrm{d}t} \right\|_T \leq \varepsilon(\delta^T) \right\}, \tag{7}$$

   where $\varepsilon(\delta^T)$ is a given function of the error δ^T and $u^T(t)$.

We will call this second version Dafermos Discontinouous Galerkin (DDG). The optimization problem in this modified algorithm can be solved exactly, as we will see in the next chapter. Clearly, finding the correct error estimate and solving the optimization problems are intricate steps in the algorithm and we will often reside to using the semi-discrete form of the algorithm because the optimization problem, and the error estimate, are significantly simpler. Another motivation for this method can be made by the following observation. Assume, that the time derivative of the exact solution $\frac{\partial u}{\partial t} = \frac{\partial H(u^T(\cdot,t),\tau)}{\partial \tau}$ and the entropy variable $\frac{\mathrm{d}U(u)}{\mathrm{d}u}(u^T(\cdot,t))$ lie in the approximation space V. This implies $\frac{\partial u}{\partial t}$ exists almost everywhere. Then follows

$$\int_T \frac{\partial U(u^T(x,t))}{\partial t}\,\mathrm{d}x = \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\partial u}{\partial t} \right\rangle_T \leq F_l^T - F_r^T$$

in the sense of distributions in time for an entropy stable numerical FV flux f with numerical entropy flux F, from an integration of the entropy inequality (3) and the chain rule at all points where $\frac{\partial u}{\partial t}$ exists. Assume now that $\frac{\mathrm{d}u^T}{\mathrm{d}t}$ is the time derivative of a (numerical) approximation

of $u(x, t)$ inside T that satisfies

$$\left\| \frac{\partial u}{\partial t} - \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\|_{T,2} \leq \delta.$$

The Cauchy–Schwarz inequality [31] allows us to bound the entropy production in this case

$$\left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\mathrm{d}u^T}{\mathrm{d}t} - \frac{\partial u}{\partial t} \right\rangle_T \leq \left\| \frac{\mathrm{d}U}{\mathrm{d}u} \right\|_{T,2} \delta$$
$$\implies \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\rangle_T \leq \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\partial u}{\partial t} \right\rangle_T + \delta \left\| \frac{\mathrm{d}U}{\mathrm{d}u} \right\|_{T,2}$$

against the entropy production of the exact solution. If we now apply an entropy correction to $\frac{\mathrm{d}u^T}{\mathrm{d}t}$ by adding a deviation of length δ into the direction of the steepest entropy descent follows

$$\left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\mathrm{d}u^T}{\mathrm{d}t} - \delta \frac{\frac{\mathrm{d}U}{\mathrm{d}u}}{\left\| \frac{\mathrm{d}U}{\mathrm{d}u} \right\|} \right\rangle_T = \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\rangle_T - \delta \left\| \frac{\mathrm{d}U}{\mathrm{d}u} \right\|_{T,2}$$
$$\leq \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\partial u}{\partial t} \right\rangle_T + \delta \left\| \frac{\mathrm{d}U}{\mathrm{d}u} \right\|_{T,2} - \delta \left\| \frac{\mathrm{d}U}{\mathrm{d}u} \right\|_{T,2}$$
$$\leq F_l^T - F_r^T,$$

that the entropy production of our numerical solution is bounded by the entropy production of the exact solution. This would encourage us to use $\varepsilon(\delta) = \delta$ in the previous algorithm. Sadly, in general, we will not have an exact integration of the inner products as the exact solution and entropy will not lie in our space V. We are further restricted to entropy descent directions that do not change the mean value of the cell. The existence and regularity of u for multidimensional systems is part of ongoing research and a generalization of this observation to counteract these problems will be part of the next section.

Remark 2 Before we start the construction of a error indicator tailored to our needs in the next section it is worthwhile to mention previous literature on error estimates. A posterior error estimates can for example be constructed using Kruzkhovs uniqueness theory [27, 38]. As the basic Kruzkhov theory leans on the L^1 theory are these estimates also L^1 estimates, and therefore lead to more complicated variational descriptions of the correction presented above. For DDG, the L^2 together with the Cauchy-Schwarz in-equality leads to relatively simple solutions in the semi-discrete case, as we will see in Lemma 2. Further, one would be restricted to scalar conservation laws. A different approach, yielding L^2 error estimates for systems, is based on the relative entropy framework by Dafermos [8, 14]. These error estimates seem like a natural fit for our application, but numerical experiments show that they tend to infinity for discontinuous solutions [13, 18] under grid refinements, as the relative entropy framework assumes Lipschitz continuity of the solution. We would like to stress that readers should consider those methods for general error estimates with a solid mathematical foundation, as the error indicator in this work was designed specifically to prove an entropy inequality and follow Dafermos' entropy rate criterion. The base assumption of our estimate in the next section will be, that the Lax-Friedrichs scheme or a similar first order E-scheme converges for the given equation and initial data, which is an ad-hoc argument compared to the previously mentioned error estimates.

3 Construction of the Semidiscrete Scheme

We begin our search for error estimators with the design of an error estimator for the semi-discrete scheme, i.e., the error of the derivative $\frac{\mathrm{d}u^T}{\mathrm{d}t}$ in relation to $\frac{\partial u}{\partial t}$. We will afterwards generalize this semi-discrete case to design an error estimator for the discrete scheme. Estimating the error in FV methods was already done using linearized conservation laws in the sense of Friedrichs systems [45–47, 52] and numerical integration of the residual. We will use a similar procedure as error estimate for our semi-discrete problem, but will directly apply a quadrature to the error between our weak solution and the projection of a second suitably fine solution onto the space V of ansatz functions. This seems problematic first, and two questions arise. The first one is why to project a function if we could just use this function as the solution of our scheme, and the second concerns the cost of calculating this more accurate solution. Our reference solution will be the result of a subcell scheme [48]. We will see that the projection of this subcell scheme depends on the size h of the subcells in such a way that one easily goes over to the limit $h \to 0$. This limit will be a lot cheaper than the subcell scheme it replaces. We will first define our subcell scheme in the following way

Definition 2 [Finite Volume subcells] Let $u^T(x)$ be an ansatz on a cell $T = [x_l, x_r] \subset \mathbb{R}$ and let

$$x_l = x_{\frac{1}{2}} < x_1 < x_{\frac{3}{2}} < \cdots < x_N < x_{N+\frac{1}{2}} = x_r$$

be a subdivision of this cell into N subcells around x_k and with left boundary $x_{k-\frac{1}{2}}$ and right boundary $x_{k+\frac{1}{2}}$. Let us denote the space of piecewise constant functions on T with discontinuities at $x_{k+\frac{1}{2}}$ by

$$P^0_{(x_k)_k} = \{f : \mathbb{R} \to \mathbb{R} | \forall k \in \{1, \ldots, N\}, \forall x \in [x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}] : f(x) = f_k\}.$$

As a projection of our ansatz into this space shall the calculation of mean values

$$u^{T,N}_k = \frac{1}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}} \int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} u^T(x, t) \, \mathrm{d}x$$

be used. We define by

$$G : V \to P^0_{(x_k)_k}, u^T \mapsto \sum_{k=1}^{N} \chi_{[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}]} u^{T,N}_k$$

the projection of our ansatz function onto the space of piecewise constant functions.

If an arbitrary u^T is projected in this way onto a function in $P^0_{(x_k)_k}$ it can be used as an initial condition to a low order FV scheme on the subcells. The used numerical flux can be the same as the one used as intercell flux of the DG scheme. The solution of this scheme will, at least for small times, be a function in $\mathrm{L}^p(\Omega)$ at any fixed instant of time. We can moreover interpret it as a differentiable mapping

$$u^{T,N}(\cdot, t) : \mathbb{R} \to \mathrm{L}^p, t \mapsto \sum_{k=1}^{N} \chi_{[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}]}(x) u^T_k(t)$$

by the following definition. The result at any fixed time lends itself to be projected back onto the space V.

Definition 3 (Semi-discrete schemes as L^p functions) Let $f(u_l, u_r)$ be an entropy stable two point flux. We will interpret to a given initial state u_k^T the solution $u_k^{T,N}(t)$ of the semi-discrete scheme

$$\frac{\mathrm{d}u_k^{T,N}}{\mathrm{d}t} = \frac{f\left(u_{k-1}^{T,N}, u_k^{T,N}\right) - f\left(u_k^{T,N}, u_{k+1}^{T,N}\right)}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}}$$

as a function $u^{T,N}(\cdot, t) \in L^p(\Omega)$ by defining

$$u^{T,N}(x, t) = \sum_k \chi_{\left[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}\right]}(x) u_k^{T,N}(t).$$

We therefore also can give an interpretation of the time derivative of the scheme as an L^p function

$$\frac{\partial u^{T,N}(x, t)}{\partial t} = \sum_k \chi_{\left[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}\right]} \frac{f\left(u_{k-1}^{T,N}, u_k^{T,N}\right) - f\left(u_k^{T,N}, u_{k+1}^{T,N}\right)}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}}$$

$$= \sum_k \chi_{\left[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}\right]} \frac{\mathrm{d}u_k^{T,N}(t)}{\mathrm{d}t}.$$

The last definition allows us to devise approximate solutions to a given ansatz u^T as initial state by choosing an appropriate numerical flux $f(u_l, u_r)$, and we will often do so by choosing the same flux as in the DG scheme under consideration, projecting, solving, and projecting back. The limiting process $N \to \infty$ is delicate if the initial condition is discontinuous between cells, even at $t = 0$. The L^1 norm of the derivative for the cell stays bounded while all norms with higher exponents blow up. This can be seen for the L^2 norm for example, assuming a discontinuity at the left edge of the cell, as

$$\left\| \frac{\partial u^{T,N}(t)}{\partial t} \right\|_{T,2} = \sqrt{\sum_{k=1}^{N} \left(\frac{f\left(u_{k-1}^{T,N}, u_k^{T,N}\right) - f\left(u_k^{T,N}, u_{k+1}^{T,N}\right)}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}} \right)^2 (x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}})}$$

$$\geq \frac{\left| f\left(u_0^{T,N}, u_1^{T,N}\right) - f\left(u_1^{T,N}, u_2^{T,N}\right) \right|}{\sqrt{x_{1+\frac{1}{2}} - x_{\frac{1}{2}}}} \xrightarrow{N \to \infty} \infty$$

holds as a consequence of $x_{k+\frac{1}{2}} \to x_{k-\frac{1}{2}}$ in this case, because the fluxes are discontinuous at the edges. For the inner of the domain and also the edges, if $u^T(x, t)$ is differentiable at the cell interfaces, follows on the contrary by the consistency of the flux

$$\frac{f\left(u_{k-1}^{T,N}, u_k^{T,N}\right) - f\left(u_k^{T,N}, u_{k+1}^{T,N}\right)}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}} \xrightarrow{N \to \infty} -\frac{\partial f(u^T)}{\partial x}\left(\lim_{N \to \infty} x_k\right).$$

We will split the derivative of our approximate solution from now on into two parts which we will call the singular and the regular part. The singular part

$$(1 - R)\frac{\partial u^{T,N}}{\partial t} = \frac{f\left(u_0^{T,N}, u_1^{T,N}\right) - f\left(u_1^{T,N}, u_2^{T,N}\right)}{x_{\frac{3}{2}} - x_{\frac{1}{2}}} \chi_{\left[x_{\frac{1}{2}}, x_{\frac{3}{2}}\right]}(x)$$

$$+ \frac{f\left(u_{N-1}^{T,N}, u_N^{T,N}\right) - f\left(u_N^{T,N}, u_{N+1}^{T,N}\right)}{x_{N+\frac{1}{2}} - x_{N-\frac{1}{2}}} \chi_{\left[x_{N-\frac{1}{2}}, x_{N+\frac{1}{2}}\right]}(x)$$

is the derivative of the subcells next to the cell boundary of the outer big cell. The regular part

$$R\frac{\partial u^{T,N}}{\partial t} = \sum_{k=2}^{N-1} \chi_{\left[x_{k-\frac{1}{2}}, x_{k+\frac{1}{2}}\right]} \frac{f\left(u_{k-1}^{T,N}, u_k^{T,N}\right) - f\left(u_k^{T,N}, u_{k+1}^{T,N}\right)}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}}$$

shall be the derivative of the subcells using only the extrapolated inner polynomial of the cell. Although the 2-norm of our approximate solution derivative blows up under grid refinement, we can still project this approximate solution onto the function space of our ansatzes V for fixed N. One can even find a closed expression for the limit of this projection for $N \to \infty$, and this limit exists as the sequence of projections is bounded in the L^2 norm by the norm of the sequence of solutions prior to the projection in the L^1 norm

$$\left\|\mathbb{P}_V \frac{\partial u^{T,N}(x,t)}{\partial t}\right\|_{T,2} \leq \sum_{k=1}^{M} \left\|\frac{\varphi_k}{\langle \varphi_k, \varphi_k \rangle_{T,2}} \left\langle \varphi_k, \frac{\partial u^{T,N}}{\partial t}\right\rangle\right\|_{T,2} \leq \sum_{k=1}^{M} \frac{\|\varphi_k\|_{T,\infty} \left\|\frac{\partial u^{T,N}}{\partial t}\right\|_{T,1}}{\|\varphi_k\|_{T,2}}.$$

Because of our split into the regular and singular part we can estimate the norm of the time derivative as

$$\left\|\frac{\partial u^{T,N}}{\partial t}\right\|_{T,1} \leq \left\|R\frac{\partial u^{T,N}}{\partial t}\right\|_{T,1} + \left\|(1-R)\frac{\partial u^{T,N}}{\partial t}\right\|_{T,1}.$$

The norm of the regular part has to converge to $\frac{\partial f}{\partial x}$, as stated before, and therefore be bounded. The singular part is clearly bounded by

$$\left\|(1-R)\frac{\partial u^{T,N}}{\partial t}\right\|_{T,1}$$
$$\leq \left| f\left(u_0^{T,N}, u_1^{T,N}\right) - f\left(u_1^{T,N}, u_2^{T,N}\right)\right| + \left| f\left(u_{N-1}^{T,N}, u_N^{T,N}\right) - f\left(u_N^{T,N}, u_{N+1}^{T,N}\right)\right|.$$

We will now calculate a closed expression for the limit $N \to \infty$ for the sequence of these projections. For the regular part

$$R\frac{\partial u^{T,N}}{\partial t} \xrightarrow{N \to \infty} -\frac{\partial f}{\partial x} \implies \mathbb{P}_V R\frac{\partial u^{T,N}}{\partial t} \xrightarrow{N \to \infty} -\mathbb{P}_V \frac{\partial f}{\partial x}$$

follows because of the convergence in the L^2 norm of the regular part towards the strong spacial flux derivative. Calculating the L^2 projection of the singular part boils down to the

calculation

$$
\int_T \varphi_k(x) \chi_{\left[x_{\frac{1}{2}}, x_{\frac{3}{2}}\right]} \frac{f\left(u_0^{T,N}, u_1^{T,N}\right) - f\left(u_1^{T,N}, u_2^{T,N}\right)}{x_{\frac{3}{2}} - x_{\frac{1}{2}}} \, dx
$$

$$
= \int_{x_{\frac{1}{2}}}^{x_{\frac{3}{2}}} \varphi_k(x) \frac{f\left(u_0^{T,N}, u_1^{T,N}\right) - f\left(u_1^{T,N}, u_2^{T,N}\right)}{x_{\frac{3}{2}} - x_{\frac{1}{2}}} \, dx
$$

$$
= \varphi_k(\xi) \left(f\left(u_0^{T,N}, u_1^{T,N}\right) - f\left(u_1^{T,N}, u_2^{T,N}\right) \right), \quad \xi \in \left[x_{\frac{1}{2}}, x_{\frac{3}{2}}\right]
$$

$$
\xrightarrow{N\to\infty} \varphi_k(0) \left(f_l^* - f\left(u^T(x_l)\right) \right).
$$

This shows

$$
\mathbb{P}_V (1 - R) \frac{\partial u^{T,N}}{\partial t} \xrightarrow{N\to\infty} \sum_k \frac{\varphi_k(x_l)}{\langle \varphi_k, \varphi_k \rangle} \left(f_l^* - f\left(u^T(x_l)\right) \right)
$$

$$
+ \frac{\varphi_k(x_r)}{\langle \varphi_k, \varphi_k \rangle} \left(f\left(u^T(x_r)\right) - f_r^* \right).
$$

Here we used the mean value theorem of integration. Interestingly, the sequence of the projected approximate solution derivatives does not blow up in the norm but stays bounded also in all norms as our ansatz functions are bounded in all norms. We will use this knowledge later to calculate the distance between this projected approximate solution and the solution yielded by the DG method as an error estimate. This is possible as we can calculate the spatial derivative of our ansatz function $u^T(x, t)$, that is smooth in the inner of every cell, exactly as

$$
\frac{\partial u^T(x, t)}{\partial x} = \frac{\partial}{\partial x} \sum_{l=1} u_l^T(t) \varphi_l(x) = \sum_{l=1} u_l^T(t) \frac{\partial \varphi_l(x)}{\partial x}.
$$

One has therefore for $x \in \overset{\circ}{T}$

$$
\frac{\partial u}{\partial t} = -\frac{\partial f(u^T(x, t))}{\partial x} = -\frac{\partial f}{\partial u} \frac{\partial u^T}{\partial x}
$$

by usage of the chain rule and has therefore, in theory, access to the projection of the limit of the low order scheme.

Sadly, the exact evaluation of the error is still a nontrivial task as we would need to project the regular part of our assumed exact solutions derivative, calculated using the strong form, onto our vectorspace V. Because this is in general a function $u \in C^k$ is this a hard problem. Incidentally, the singular part can be projected easily, as we saw earlier. We will therefore estimate the norm of the distance between the limit of the projections and the DG scheme by

$$
\left\| \frac{du^T}{dt} - \mathbb{P}_V R \frac{\partial u^{T,N}}{\partial t} - \mathbb{P}_V (1 - R) \frac{\partial u^{T,N}}{\partial t} \right\|_p
$$

$$
= \left\| \mathbb{P}_V \left(\frac{du^T}{dt} - R \frac{\partial u^{T,N}}{\partial t} - \mathbb{P}_V (1 - R) \frac{\partial u^{T,N}}{\partial t} \right) \right\|_p
$$

$$
\leq \left\| \frac{du^T}{dt} - R \frac{\partial u^{T,N}}{\partial t} - \mathbb{P}_V (1 - R) \frac{\partial u^{T,N}}{\partial t} \right\|_p,
$$

where the estimate follows from the fact that $\frac{\mathrm{d}u^T}{\mathrm{d}t}$ and $\mathbb{P}_V(1-R)\frac{\partial u^{T,N}}{\partial t}$ lie in V, and are therefore fixed points of the projection, and the projection operator has a norm smaller than one. The author would like to stress that the function in the norm on the right hand side of the equation is as smooth as the (numerical) flux function and can be evaluated exactly at any point. It is therefore only logical to evaluate this norm in case of the 2 norm using numerical quadrature, which was done in the numerical tests. Quadrature rules of the Gauß–Legendre family were chosen in the numerical tests. We will from now on refer to the limit of the regular part of the approximate solution by the low-order scheme together with the limit of the irregular part projected onto V as the reference solution

$$\frac{\partial u^{\mathrm{ref}}}{\partial t} = R\frac{\partial u^{T,\infty}}{\partial t} + \mathbb{P}_V(1-R)\frac{\partial u^{T,\infty}}{\partial t}.$$

As explained before, our solver needs to solve an optimization problem after every time step for the discrete, or for every evaluation of the time derivative in the semi discrete algorithm. We will first consider the optimization problem of the semi discrete scheme as this will in fact also be a building block for our approximate solution to the optimization problem in the discrete case. The semi discrete problem is, luckily, a linear one and can be solved exactly, as we will see in the next lemma.

Lemma 2 *[Restricted gradien descent] A solution s of the semidiscrete optimization problem (7) is given by*

$$g = -\frac{\mathrm{d}U^T}{\mathrm{d}u}(u^T), \quad h = g - \frac{\langle 1, g\rangle_T}{\langle 1, 1\rangle_T}1, \quad s = \begin{cases} \frac{\varepsilon h}{\sqrt{\langle h,h\rangle_T}} & h \neq 0 \\ 0 & h = 0 \end{cases}$$

and this solution is unique if $\frac{\mathrm{d}U^T}{\mathrm{d}u} \neq c\,1$ holds, i.e. when the solution in the cell is not constant.

Proof We beginn our proof by showing that $s \in Z$ holds. Clearly,

$$\langle 1, s\rangle_T = \frac{\varepsilon}{\sqrt{\langle h,h\rangle_T}}\left(\langle 1, g\rangle_T - \langle 1, g\rangle_T\frac{\langle 1, 1\rangle_T}{\langle 1, 1\rangle_T}\right) = 0$$

shows that the solution lies in the linear subspace $V \supset W = \{v \in V \mid \langle 1, v\rangle_T = 0\}$. The rescaling from h to s also implies

$$\|s\|_{T,2} = \sqrt{\langle s, s\rangle_T} = \varepsilon\frac{\sqrt{\langle h,h\rangle_T}}{\sqrt{\langle h,h\rangle_T}} = \varepsilon.$$

The set $Z = W \cap \overline{B}_\varepsilon$ is clearly the intersection of vectors of length less than or equal ε and the subspace W and hence s lies inside this vectorspace. After the admissibility of s was established we can take care of the optimality. We will first show that for the restriction $g \in \overline{B}_\varepsilon$ is a scaled version of $g = -\frac{\mathrm{d}U^T}{\mathrm{d}u}$ indeed the optimal descent direction. We will afterwards show that the optimal descent direction is $s = \mathbb{P}_w\,g$ if only directions in the subspace W are considered. Clearly,

$$\left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, \varepsilon\frac{g}{\|g\|_{T,2}}\right\rangle_T = \left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, -\varepsilon\frac{\frac{\mathrm{d}U^T}{\mathrm{d}u}}{\left\|\frac{\mathrm{d}U^T}{\mathrm{d}u}\right\|_{T,2}}\right\rangle_T = -\varepsilon\left\|\frac{\mathrm{d}U^T}{\mathrm{d}u}\right\|_{T,2}$$

holds, because $\frac{\mathrm{d}U^T}{\mathrm{d}u}$ and s are colinear. Let now $v \in V$ be arbitrary with $\|v\|_{T,2} = \varepsilon$. The Cauchy-Schwarz inequality implies

$$\left|\left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, v\right\rangle_T\right| \leq \left\|\frac{\mathrm{d}U^T}{\mathrm{d}u}\right\|_T\varepsilon.$$

It therefore follows

$$\left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, s \right\rangle_T \leq \left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, v \right\rangle_T,$$

and this inequality is strict for $\frac{\mathrm{d}U^T}{\mathrm{d}u} \neq c\,\mathbb{1}$ and $v \neq s$, as in this case $s \neq 0$ and $\frac{\mathrm{d}U^T}{\mathrm{d}u} \neq 0$ follows and the Cauchy-Schwarz inequality is equal only if $\frac{\mathrm{d}U^T}{\mathrm{d}u}$ and s are colinear in this case. We note in passing that $\langle \cdot, \cdot \rangle_T$, as a inner product on V, also is an inner product on W and hence the Cauchy-Schwartz inequality applies for all elements in W. We can further decompose V into $W^\perp$ and W and if at least one vector $v \in W$ of two vectors $v, w \in V$ is from W follows

$$\langle v, w \rangle = \langle \mathbb{P}_W\, v + \mathbb{P}_{W^\perp}\, v, w \rangle = \langle \mathbb{P}_W\, v, w \rangle.$$

we can therefore conclude that the aforementioned proof can be reread with $\mathbb{P}_W \frac{\mathrm{d}U^T}{\mathrm{d}u}$ and still applies, if only vectors $v \in W$ are allowed, or equivalently $\mathbb{P}_W \frac{\mathrm{d}U^T}{\mathrm{d}u}$ is entered instead of $\frac{\mathrm{d}U^T}{\mathrm{d}u}$. This in turn equals s. □

Remark 3 Abgrall and collaborators in [1] showed that the entropy correction terms derived for the residual distribution schemes to enforce entropy conservation can be interpreted as solutions to optimization problems. These correction terms or the respective solutions to the optimization problems lead to the same descent directions, albeit their optimization problems are different. A closer inspection reveals that the optimization problems are in fact more or less dual to ours. Still, we are not interested in entropy conservation but high dissipation restricted by error bounds.

Remark 4 The expression s in lemma 2 above is discontinuous at $h = 0$ as it jumps from a vector of length ε to 0. Similar problems appear when WENO weights [32] are computed and we therefore use and advise to use

$$s = \frac{\varepsilon h}{\sqrt{\langle h, h \rangle_T + c}}$$

as a stand in term. The constant $c \approx 10^{-30}$ averts a division by zero and removes a conditional expression in the same way WENO weights are computed in a stable manner while avoiding a division by zero. If ε is in the region of the relative machine precision the correction can be omitted.

Knowledge of the solution also allows us to give a lower bound for δ that makes our scheme even entropy dissipative in the classical sense.

Theorem 1 *[Classical entropy inequality for DDG] Assume that a monotone entropy stable FV flux $f(u_l, u_r)$ and a strictly convex and twice continuously differentiable entropy U is used. Let ε be determined by the following expression, depending on u*

$$\varepsilon \geq \lim_{N \to \infty} \frac{\|h\|_{T,2}}{\left\langle h, \frac{\mathrm{d}U^T}{\mathrm{d}u} \right\rangle_T} \left(\delta(N) \left\| \widetilde{\frac{\mathrm{d}U^T}{\mathrm{d}u}} \right\|_{T,2} + \delta_U \left\| \frac{\partial u^{T,N}}{\partial t} \right\|_{T,1} \right),$$

$$\delta(N) = \left\| \frac{\mathrm{d}u^T}{\mathrm{d}t} - \frac{\partial u^{\mathrm{ref},N}}{\partial t} \right\|_{T,2}, \quad \delta_U = \left\| \frac{\mathrm{d}U}{\mathrm{d}u}\left(u^T\right) - \frac{\mathrm{d}U^T}{\mathrm{d}u} \right\|_{T,\infty}. \tag{8}$$

Then hold the two entropy inequalities

$$\frac{\mathrm{d}E_{u,T}}{\mathrm{d}t} \leq F_l^T - F_r^T, \quad \frac{\mathrm{d}E_{u,T}}{\mathrm{d}t}(t_0) \leq \lim_{N\to\infty} \frac{\mathrm{d}}{\mathrm{d}t} \int_T U\left(u^{T,N}\right) \mathrm{d}x,$$

for our modified scheme

$$\frac{\mathrm{d}u^{T,D}}{\mathrm{d}t} = \frac{\mathrm{d}u^T}{\mathrm{d}t} - s.$$

The first inequality states that our solution satisfies a discrete entropy inequality, while the second one states that the entropy decreases faster than the entropy of the limit solution of the subcell scheme.

Proof Let us remark before we start with our proof, that our corrections to the base scheme have zero mean value, i.e. the mean value of the time derivative of our scheme and of the base scheme are the same. We will exploit this behavior in what follows and use

$$\frac{\mathrm{d}u^T}{\mathrm{d}t} = \overline{\frac{\mathrm{d}u^T}{\mathrm{d}t}} + \widetilde{\frac{\mathrm{d}u^T}{\mathrm{d}t}}$$

as a notation for the splitting of a function g into a constant function $\overline{g}$ with the mean values of g and a function $\tilde{g}$ representing the variation of g around it's mean value. If $u^T = \overline{u^T}$ holds we can look at the special case of our base FV scheme as in this case $\frac{\mathrm{d}U}{\mathrm{d}u}(u(x_k)) = \frac{\mathrm{d}U}{\mathrm{d}u}(\bar{u})$

$$\frac{\mathrm{d}E_{u,T}(u(t))}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}, \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\rangle_T = \frac{\mathrm{d}U}{\mathrm{d}u}(\bar{u}) \left\langle \mathbb{1}, \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\rangle_T = \frac{\mathrm{d}U}{\mathrm{d}u}(\bar{u})(f_l^T - f_r^T) \leq (F_l^T - F_r^T)$$

is satisfied. Here we used that the entropy flux $F(u_l, u_r)$ of the entropy stable flux $f(u_l, u_r)$ satisfies

$$\frac{\mathrm{d}U}{\mathrm{d}u}(u_m)(f(u_l, u_m) - f(u_m, u_r)) = \frac{\mathrm{d}U}{\mathrm{d}u}\frac{\mathrm{d}u_m}{\mathrm{d}t} = \frac{\mathrm{d}U}{\mathrm{d}t} \leq F(u_l, u_m) - F(u_m, u_r).$$

See for example [49, 50]. This entropy inequality also allows us to proof

$$\lim_{N\to\infty} \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}(u^{T,N}), \frac{\mathrm{d}u^{T,N}}{\mathrm{d}t} \right\rangle = \lim_{N\to\infty} \frac{\mathrm{d}}{\mathrm{d}t} \int_T U(u^{T,N}(x,t)) \, \mathrm{d}x \leq F_l^T - F_r^T \qquad (9)$$

for arbitrary $u^T(x,t)$, as this holds for the summed contributions of the FV subcell scheme in the cell for fixed N, and also in the limit. Clearly holds also

$$\lim_{N\to\infty} \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}(u^T), \frac{\partial u^{T,N}}{\partial t} \right\rangle = \lim_{N\to\infty} \left\langle \frac{\mathrm{d}U}{\mathrm{d}u}\left(u^{T,N}\right), \frac{\partial u^{T,N}}{\partial t} \right\rangle$$

as $\left\| \frac{\partial u^{T,N}}{\partial t} \right\|_{T,1}$ stays bounded while $\left\| \frac{\mathrm{d}U}{\mathrm{d}u}\left(u^T\right) - \frac{\mathrm{d}U}{\mathrm{d}u}\left(u^{T,N}\right) \right\|_{T,\infty} \xrightarrow{N\to\infty} 0$ holds as the ansatz is continuous in the cell. We will assume from now on $u^T \neq \bar{u}$ and therefore also that

$$h = g - \frac{\langle \mathbb{1}, g \rangle}{\langle \mathbb{1}, \mathbb{1} \rangle} \mathbb{1} \neq 0$$

because $\frac{\mathrm{d}U}{\mathrm{d}u}(u^T) = 0$ is only possible for a single $u \in \mathbb{R}^n$ and $u \neq \bar{u}$ implies $\frac{\partial U}{\partial u} \neq \overline{\frac{\partial U}{\partial u}}$ as the entropy is strictly convex. We will now generalize our argument from the last section concerning the entropy dissipativity of our approximate time derivative $\frac{\mathrm{d}u^T}{\mathrm{d}t}$. It is imperative to first concentrate on the case where N is finite as done before for the entropy inequality for

the subcell scheme and derive appropriate bounds. We will afterwards go over to the limit to prove the theorem. Let us denote by $\frac{dU}{du}(u^T(x,t))$ the exact value of the entropy variables associated with the numerical solution $u^T(t)$ while

$$\frac{dU^T}{du}(u^T(x,t)) = \sum_{j\in T} \varphi_j(x)\frac{dU}{du}(u^T(x_j,t)), \text{ where } \varphi_j(x) \text{ are Lagrange polynomials,}$$

shall be the interpolation of $\frac{dU}{du}(u^T(x,t))$ in the space V. We can therefore state the error made in the prediction of the entropy dissipation by interchanging the exact entropy functional with the one living in our approximation space

$$\left\langle \frac{dU}{du}\left(u^T(\cdot,t)\right), \frac{\partial u^{T,N}}{\partial t} \right\rangle_T$$
$$= \left\langle \frac{dU^T}{du}, \frac{\partial u^{T,N}}{\partial t} \right\rangle_T + \underbrace{\left\langle \frac{dU}{du}\left(u^T(\cdot,t)\right) - \frac{dU^T}{du}(\cdot,t), \frac{\partial u^{T,N}}{\partial t} \right\rangle_T}_{\leq \delta_U \left\| \frac{\partial u^{T,N}}{\partial t} \right\|_{T,1}}.$$

Because $\frac{dU^T}{du}$ is from V we can exchange $\frac{\partial u^{T,N}}{\partial t}$ for $\frac{\partial u^{\mathrm{ref},N}}{\partial t}$

$$\left\langle \frac{dU^T}{du}, \frac{\partial u^{T,N}}{\partial t} \right\rangle = \left\langle \frac{dU^T}{du}, \frac{\partial u^{\mathrm{ref},N}}{\partial t} \right\rangle$$

without any penalty as $\frac{\partial u^{T,N}}{\partial t} - \frac{\partial u^{\mathrm{ref},N}}{\partial t} \in V^\perp$ holds. If we also swap the reference time derivative for our scheme derivative $\frac{du^{T,N}}{dt}$, we find

$$\left\langle \frac{dU^T}{dt}, \frac{\partial u^{\mathrm{ref},N}}{\partial t} \right\rangle_T$$
$$= \left\langle \frac{dU^T}{du} + \widetilde{\frac{dU^T}{du}}, \frac{\partial u^{\mathrm{ref},N}}{\partial t} + \widetilde{\frac{\partial u^{\mathrm{ref},N}}{\partial t}} \right\rangle_T$$
$$= \left\langle \frac{dU^T}{du} + \widetilde{\frac{dU^T}{du}}, \frac{\partial u^{\mathrm{ref},N}}{\partial t} \right\rangle_T + \left\langle \frac{dU^T}{du}, \widetilde{\frac{\partial u^{\mathrm{ref},N}}{\partial t}} \right\rangle_T + \left\langle \widetilde{\frac{dU^T}{du}}, \widetilde{\frac{\partial u^{\mathrm{ref},N}}{\partial t}} \right\rangle_T$$
$$\underbrace{\hspace{4cm}}_{\text{same mean value}} \quad \underbrace{\hspace{3cm}}_{\text{orthogonality}}$$
$$= \underbrace{\left\langle \frac{dU^T}{du} + \widetilde{\frac{dU^T}{du}}, \frac{du^T}{dt} \right\rangle_T + \left\langle \frac{dU^T}{du}, \widetilde{\frac{du^T}{dt}} \right\rangle_T + \left\langle \widetilde{\frac{dU^T}{du}}, \widetilde{\frac{du^T}{dt}} \right\rangle_T}_{\left\langle \frac{dU^T}{du}, \frac{du^T}{dt} \right\rangle_T}$$
$$+ \underbrace{\left\langle \widetilde{\frac{dU^T}{du}}, \widetilde{\frac{\partial u^{\mathrm{ref},N}}{\partial t}} - \widetilde{\frac{du^T}{dt}} \right\rangle_T}_{\leq \left\| \widetilde{\frac{dU^T}{du}} \right\|_{T,2} \delta(N)}.$$

Please note that we used several facts of our reference solution and our approximate solution to sharpen this bound to only depend on the variation of the entropy variables and the variation of the solution around their respective mean values. This was possible after the inner product

was split into the respective inner products of the mean values and variations around the mean values with each other. As our reference solution $\frac{\partial u^{\text{ref}}}{\partial t}$ and $\frac{du^T}{dt}$ share the same mean values we can easily swap one for the other in the first inner product. The second inner product in this decomposition can also be swapped, as it is by definition zero. This follows from the fact that the mean values, calculated with respect to the inner product of T, are orthogonal to the variations. The only penalty that has to be bounded consists of the inner product of the variations. Combining the previous steps leads us to

$$
\left| \left\langle \frac{dU}{du}\left(u^T(t)\right), \frac{\partial u^{T,N}}{\partial t} \right\rangle_T - \left\langle \frac{dU^T}{du}(t), \frac{du^T}{dt} \right\rangle_T \right|
$$
$$
\leq \left| \left\langle \widetilde{\frac{dU^T}{du}}(t), \widetilde{\frac{\partial u^{\text{ref},N}}{\partial t}} - \widetilde{\frac{du^T}{dt}} \right\rangle_T \right| + \left| \left\langle \frac{dU}{du} - \frac{dU^T}{du}, \frac{\partial u^{T,N}}{\partial t} \right\rangle_T \right|
$$
$$
\leq \left\| \widetilde{\frac{dU^T}{du}} \right\|_{T,2} \delta(N) + \delta_U \left\| \frac{\partial u^{T,N}}{\partial t} \right\|_{T,1}
$$

as an upper bound for the difference of the exact entropy dissipation and the entropy dissipation of our approximate solution. If ε is set according to the value given above, one finds that the following inequality holds as the additional entropy production that can be bounded using the error δ and δ_U can be fully counteracted by the entropy dissipation of the steepest descent direction

$$
\left\langle \frac{dU^T}{du}, \frac{du^T}{dt} - s_N \right\rangle_T = \left\langle \frac{dU^T}{du}, \frac{du^T}{dt} - \varepsilon_N \frac{h}{\|h\|} \right\rangle_T = \left\langle \frac{dU^T}{du}, \frac{du^T}{dt} \right\rangle_T - \varepsilon \frac{\left\langle h, \frac{dU^T}{du} \right\rangle_T}{\|h\|_{T,2}}
$$
$$
= \left\langle \frac{dU^T}{du}, \frac{du^T}{dt} \right\rangle_T - \left(\delta \left\| \widetilde{\frac{dU^T}{du}} \right\|_{T,2} + \delta_U \left\| \frac{\partial u^{T,N}}{\partial t} \right\|_{T,1} \right) \leq \left\langle \frac{dU}{du}\left(u^T(t)\right), \frac{\partial u^{T,N}}{\partial t} \right\rangle .
$$

The given lower bound for ε behaves significantly better than one would think. A primary reason for this is that h is co-linear to $\widetilde{\frac{dU^T}{du}}$ and relates to $\frac{dU^T}{du}$ via an orthogonal projection. Therefore is the expression

$$
\frac{\|h\|_{T,2} \left\| \widetilde{\frac{dU^T}{du}} \right\|_{T,2}}{\left\langle h, \frac{dU^T}{du} \right\rangle_T} = \frac{\|h\|_{T,2} \left\| \widetilde{\frac{dU^T}{du}} \right\|_{T,2}}{\left\langle h, \widetilde{\frac{dU^T}{du}} \right\rangle_T + \left\langle h, \overline{\frac{dU^T}{du}} \right\rangle_T}
$$

bounded. We showed earlier, that the limit of the reference solution exists. We can therefore go over to the limit and conclude that

$$
\left\langle \frac{dU^T}{du}, \frac{du^{T,D}}{dt} \right\rangle_T = \left\langle \frac{dU^T}{du}, \frac{du^T}{dt} - s \right\rangle_T \leq \lim_{N \to \infty} \frac{d}{dt} \int_T U\left(u^{T,N}\right) dx
$$

holds in the limit. The last step in our proof is a comparison principle. We combine the last equation with (9) to find

$$
\left\langle \frac{dU^T}{du}, \frac{du^{T,D}}{dt} \right\rangle_T = \left\langle \frac{dU^T}{du}, \frac{du}{dt} - s \right\rangle_T \leq F_l^T - F_r^T .
$$

□

4 Construction of the Discrete Scheme

Designing an error estimator for the discrete scheme is a more delicate issue. The observation

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \left\| u^{\text{ref}}(\cdot, t) - u^T(\cdot, t) \right\|_{T,2} &= \frac{\mathrm{d}}{\mathrm{d}t} \sqrt{\left\langle u^{\text{ref}}(\cdot, t) - u^T(\cdot, t), u^{\text{ref}}(\cdot, t) - u^T(\cdot, t) \right\rangle_T} \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}t} \left\langle u^{\text{ref}}(\cdot, t) - u^T(\cdot, t), u^{\text{ref}}(\cdot, t) - u^T(\cdot, t) \right\rangle_T}{2 \sqrt{\left\langle u^{\text{ref}}(\cdot, t) - u^T(\cdot, t), u^{\text{ref}}(\cdot, t) - u^T(\cdot, t) \right\rangle_T}} \\
&= \frac{\left\langle u^{\text{ref}} - u^T, \frac{\partial u^{\text{ref}}}{\partial t} - \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\rangle_T}{\left\| u^{\text{ref}} - u^T \right\|_{T,2}} \leq \left\| \frac{\partial u^{\text{ref}}}{\partial t} - \frac{\mathrm{d}u^T}{\mathrm{d}t} \right\|_{T,2} = \delta_t^T
\end{aligned}
$$

allows us to estimate the total error made in cell T if we know the error between the derivative of the reference solution and the exact solution,

$$
\left\| u^{\text{ref}}(x, t) - u^T(x, t) \right\|_{T,2} \leq \int_0^t \left\| \frac{\partial u^{\text{ref}}}{\partial t}(\tau) - \frac{\mathrm{d}u^T}{\mathrm{d}t}(\tau) \right\|_{T,2} \mathrm{d}\tau = \int_0^t \delta_\tau^T \, \mathrm{d}\tau.
$$

But during the integration of such an error estimate in time the two solutions will in general start to drift apart from each other. Calculating a reference derivative would therefore need the knowledge of the exact solution. We will assume therefore that the total error between reference solution and numerical solution stays small enough to warrant us using the reference derivative at time τ as calculated from the solution $u^T(\cdot, \tau)$ and not with respect to $u^{\text{ref}}(\cdot, \tau)$. As in the other cases, we will reside to numerical quadrature for the calculation of this quantity. The integrand was already used as an error estimate in the semidiscrete case. For the calculation of the outer integral in the time direction it is worthwhile to consider the connection between Runge–Kutta time integration and numerical quadrature. If for example the SSPRK33 solver [44]

$$
\begin{aligned}
u^{(1)} &= u^{(0)} + \Delta t L(u^{(0)}) \\
u^{(2)} &= u^{(0)} + \frac{1}{4} \Delta t L(u^{(0)}) + \frac{1}{4} \Delta t L(u^{(1)}) \\
u^{(3)} &= u^{(0)} + \frac{1}{6} \Delta t L(u^{(0)}) + \frac{1}{6} \Delta t L(u^{(1)}) + \frac{2}{3} \Delta t L(u^{(2)})
\end{aligned}
$$

is used one sees clearly, that this in fact a numerical quadrature of $\frac{\mathrm{d}u}{\mathrm{d}t} = L(u)$ where first using the left sided Newton-Cotes formula an approximation $u^{(1)}$ for the solution at Δt is calculated. Next an approximation $u^{(2)}$ for the solution at time $\Delta t/2$ is calculated using the trapezoidal rule

$$
\int_0^{\Delta t} \frac{\mathrm{d}u}{\mathrm{d}t} \, \mathrm{d}t \approx \frac{L(u(0)) + 4L(u(\Delta t/2)) + L(u(\Delta t))}{6},
$$

and as a last step a better approximation of the rightmost value is calculated using the Simpson rule and the two precalculated approximations. If our semi-discrete error estimate is local Lipschitz continuous, which it clearly is, it is therefore a logical decision to choose the quadrature of the time integrator as error estimate quadrature. One can therefore also reuse the interim results of the time integrator. Our discrete error estimate therefore reads as

$$
\delta^T = \Delta t \frac{\delta_t^T \left(u^{(0)} \right) + 4\delta_t^T \left(u^{(2)} \right) + \delta_t^T \left(u^{(1)} \right)}{6}, \tag{10}
$$

and is third order accurate, as is the timesteping algorithm.

Lemma 3 *[Existence of at most one solution] The optimization problem stated for the DRKDG method (6) possesses a unique solution if a strictly convex entropy functional is used.*

Proof The set Z is closed and bounded, hence compact, and the functional $E^T(u)$ continuous. Therefore there exists a u in Z with $\forall v \in Z : E^T(u) \leq E^T(v)$. Concerning the uniqueness, the strict convexity of $E^T(\cdot)$ implies that if $u, v \in V$, with $u \neq v$, would exist with $E^T(u) = E^T(v)$ and $\forall w \in W : E^T(u) \leq E^T(w)$ a contradiction for $\lambda = \frac{1}{2}$ and $w = \lambda u + (1 - \lambda)v$ would arise as

$$E^T(w) = E^T(\lambda u + (1 - \lambda)v) < \lambda E^T(u) + (1 - \lambda)E^T(v) = E^T(u)$$

would follow from the strict convexity in this case. □

Solving the optimization problem for the discrete case is more intricate than in the semidiscrete case. While a solution for the simple entropy $U(u) = \frac{u^2}{2}$ could be computed by hand, the computation for complicated entropies is not feasible. We will therefore concentrate on the numerical approximation of a solution. A simple, yet effective, procedure seems to be the gradient descent

$$u^{T,n+1} = u^{T,n} - \lambda h$$

with appropriately chosen step size λ [35]. The descent direction will be, as in the semidiscrete case, the solution to the steepest descent problem from Lemma 2. As we would like to solve this problem in every complete RK step should the cost of the nonlinear solver be of the same magnitude as the repeated evaluations of the semidiscretisation for the Runge–Kutta time integration method. This limits the number of allowed steps and we generally try to use the same amount of gradient steps r as there are stages in the RK method. We therefore propose

$$\lambda = \frac{\varepsilon}{r \, \|h\|_{T,2}}$$

as this limits the maximal distance between $u^{T,0}$ and the final result $u^{T,s}$ to

$$\left\| u^{T,0} - u^{T,r} \right\|_{T,2} \leq \sum_{n=1}^{r} \left\| u^{T,n} - u^{T,n-1} \right\|_{T,2} \leq \sum_{n=1}^{r} \frac{\varepsilon \, \|h\|_{T,2}}{r \, \|h\|_{T,2}} = \varepsilon.$$

One restriction has to be made, as this selection could lead to a diverging sequence of steps if λ is large as ε is large. We therefore restrict the step size to counteract this problem. Let L be a Lipschitz bound on the entropy variables in the norm on T. Then follows from [35, equation 1.2.5]

$$
\begin{aligned}
E^T\left(u^{T,n+1}\right) &\leq E^T\left(u^{T,n}\right) + \left\langle \frac{dU^T}{du}, \lambda h \right\rangle_T + \frac{L}{2}\lambda^2 \\
&= E^T\left(u^{T,n}\right) - \lambda \underbrace{\left\langle \frac{dU^T}{du}, -h \right\rangle_T}_{\geq 0} + \frac{L}{2}\lambda^2.
\end{aligned}
$$

This implies that a descent happens whenever the step size is chosen as

$$\lambda < 2 \frac{\left\langle \frac{dU^T}{du}, -h \right\rangle_T}{L}.$$

Table 2 Overview of the schemes that were tested against each other in the numerical tests section

Option	DDG	DRK-DG	Godunov
Time integrator	SSPRK33	SSPRK33	Forward Euler
Timestep	$0.5 = \frac{\Delta t (p^2+1)}{\Delta x} c_{max}$	$0.5 = \frac{\Delta t (p^2+1)}{\Delta x} c_{max}$	$\frac{\Delta t}{\Delta x} c_{max} = 0.25$
Correction length $\varepsilon(\delta)$	From (8)	From (10), $\varepsilon(\delta) = \delta$	None
Collocation points	$p+1$ pt. Gauß-Lob	$p+1$ pt. Gauß-Lob	None
Error estimate	$p+2$ pt. Gauß-Leg	$p+2$ pt. Gauß-Leg	None
Mass matrix	Exact	Exact	Not applicable

Their respective parameters concerning time integration are given together with the used quadratures and collocation points. The Godunov method is used to calculate reference solutions on a fine mesh. The reference solutions for the numerical convergence analysis were calculated using the method of characteristics to achieve the desired accuracy

In fact, in the implementation

$$\lambda \leq \frac{3 \left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, -h \right\rangle_T}{2L}$$

is used as a compromise as

$$\lambda \leq \frac{\left\langle \frac{\mathrm{d}U^T}{\mathrm{d}u}, -h \right\rangle_T}{L}$$

minimizes this bound.

5 Numerical Tests

Our numerical tests were carried out to shed some light on the following topics and possible problems:

- Does the added error estimate controlled dissipation stabilize the schemes enough to use them for shock capturing calculations?
- Is the numerical Dafermos entropy rate criterion satisfied, even though only approximate error indicators are used?
- Are the added corrections small enough when a smooth solution is calculated to not destroy the high order approximation?
- Does the added dissipation reduce or enlarge the possible time step?

Before detailing the results, let us state that all of these questions can be answered with promising results, although the question of small approximation errors needs further research, especially for long integration times. The tested methods, i.e. their used quadrature rules, collocation points, time integration and similar design decisions are given in the Table 2.

Remark 5 The procedure is not for free, but not to expensive either. Additional costs compared to the standard DG scheme are the evaluation of the error estimator, and the evaluation of the entropy variables at every node. The needed operations for both of these actions scale linear with the number of nodes per cell, apart from the evaluation of the ansatz function at the nodes for the quadrature of the error estimate, that needs a quadratic amount of operations, as

a matrix vector multiplication is needed. While the evaluation of the entropy variables can be costly, one gains a semi-discrete entropy inequality. Flux-differencing, for example, is based on the evaluation of two-point fluxes with entered nodal values and coefficients forming an anti-symmetric matrix. Therefore, $n^2/2$ evaluations of in general non-trivial two-point fluxes are needed, which can outweigh the cost of the presented approach. Abgrall [1] uses not an error estimate but a direct calculation of the entropy dissipation needed to produce an entropy conservative scheme. As stated in Remark 3, this results in a correction with a similar direction, but a different overall effect and size. A recent comparison showed [33], that the correction of Abgrall only results in a 50 percent runtime increse of the unmodified DG scheme, while flux-differencing resulted in a 180 percent runtime increase, and a similar outcome could be expected for the method presented here. As this work focuses not on the efficient implementation of the new technique but is a first proof of concept a detailed cost analysis is out of scope for the present publication.

5.1 Calculation of Discontinuous Solutions

The DRKDG and DDG schemes were first tested for Burgers' equation

$$\Omega = [0, 2), \quad f(u) = \frac{u^2}{2}, \quad U(u) = u^2$$

on a periodic domain with the square entropy and the local Lax-Friedrichs flux. The tested initial conditions were

$$u_1(x, 0) = \sin(\pi x) + \frac{1}{2}, \quad u_2(x, 0) = \begin{cases} -x & x \in [0, 1) \\ 2 - x & x \in [1, 2) \end{cases}$$

and it is well known that u_1 results in a discontinuous solution in finite time and crashes a RK-DG method without further stabilization quite easily. The second initial condition results in a rarefaction of the initial discontinuity. Because the sonic point of the flux is also part of this rarefaction we will be able to analyze the behavior of the scheme in this sometimes troublesome situation [51].

First of all, both test cases were successfully run by both schemes until $t = 100$, where the simulation was stopped, and nothing indicates that the simulation could not be run further. Moreover, as can be seen from the results in Figs. 2 and 3, their solutions seem to be essentially free of oscillations, as only the polynomials in up to three shocked cells oscillate slightly. This evidence of a robust scheme is only hampered slightly by the tests concerning the sonic point glitch. The cells around the sonic point of the flux clearly show a problematic feature of the base Lax-Friedrichs scheme that can't be corrected using the devised method, consult also Figs. 4 and 5. Future improvements could therefore be based around modifications of the intercell flux. Note that while only tests for polynomials $p = 6$ are shown, the results look comparable for all orders satisfying $2 \leq p \leq 10$, and higher orders were not tested. The case $p = 1$ showed a limitation of the method, as in this case the entropy dissipation in the cell results in a slope limiting, as this is the only degree of freedom, that also happens in smooth areas.

5.2 Numerical Test of the Dafermos Entropy Rate Criterion and Semidiscrete Entropy Inequalities

In [26] the author tested several schemes for their compatibility with Dafermos' entropy rate criterion. A followup paper [25] also tested if a similar family of solvers respects the
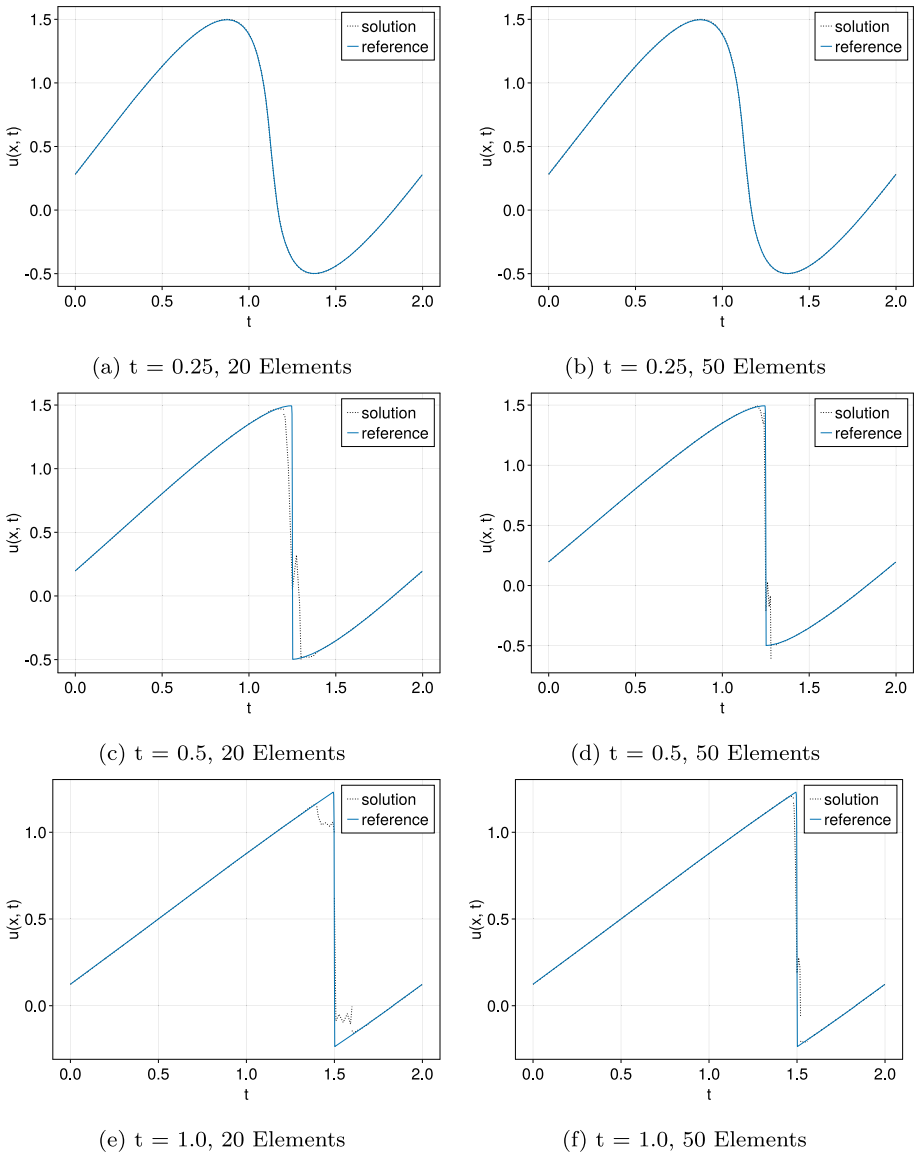
Fig. 2 Solution to the first initial condition using the semi-discrete scheme DDG and polynomial degree $p = 6$

classical entropy inequality, i.e. if the schemes are entropy dissipative. We would like to test the new DG scheme presented in this publication for the same two entropy criteria, i.e. Dafermos' entropy rate criterion and classical entropy inequalities. Our first test case with initial condition $u_1(x, 0)$ will be used, which coincides with the test case used in the first publication.

The semi-discrete entropy inequality was calculated for the first test case at every time-step and the decadic logarithm of the positive and negative deviation from the entropy equality was

(a) t = 0.25, 20 Elements

(b) t = 0.25, 50 Elements

(c) t = 0.5, 20 Elements

(d) t = 0.5, 50 Elements

(e) t = 1.0, 20 Elements

(f) t = 1.0, 50 Elements
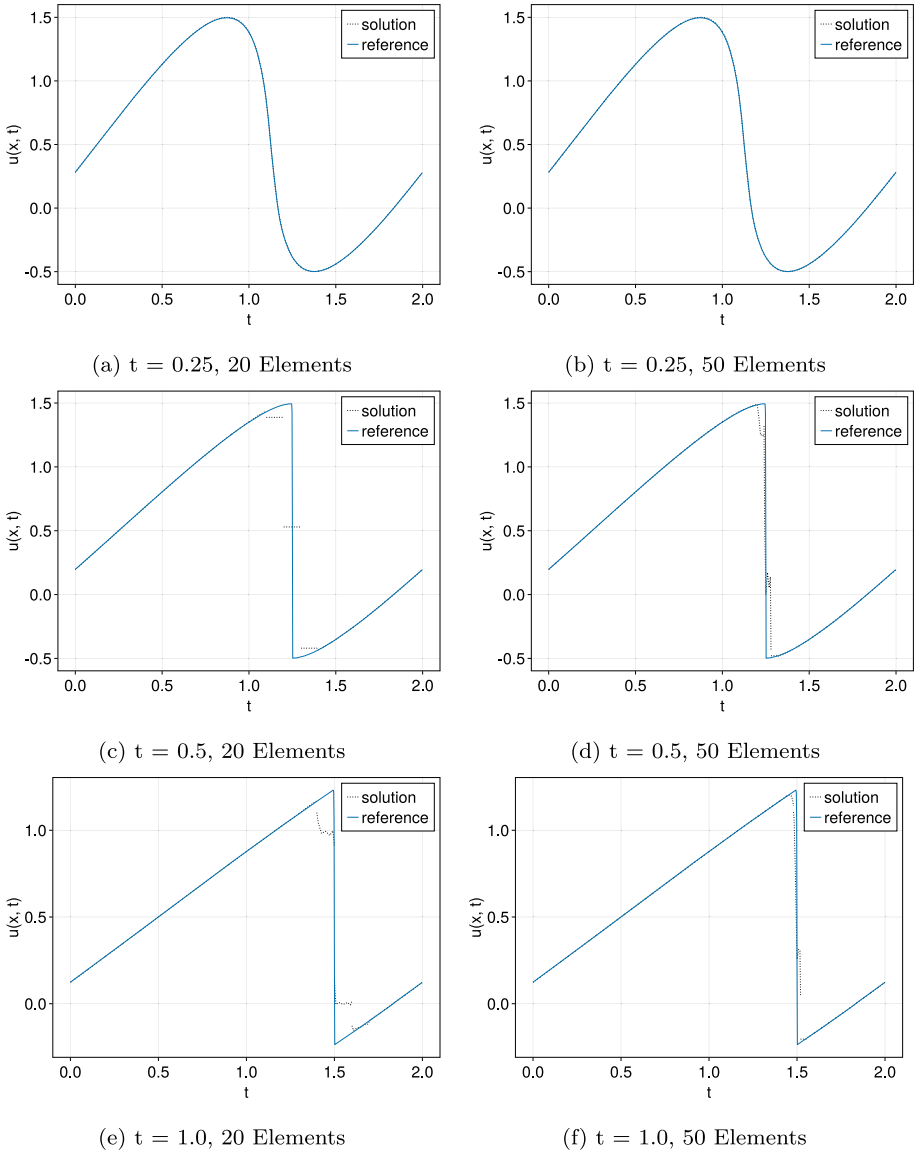
Fig. 3 Solution to the first initial condition by the fully discrete scheme DRKDG and polynomial degree $p = 6$

plotted as a heat map, see Fig. 6. This was deemed necessary as error bounds stemming from numerical quadrature are used in the implementation, hence the proved entropy inequality does not apply in a strict sense but under the assumption of vanishing quadrature errors. Still, the hope is that the deviations from this entropy equality are small. And interestingly this seems to be true. A positive violation is only appearing with a magnitude of 10^{-16}, around the machine precision used for the calculation and only in extremely smooth areas of the solution - indicating that these smooth areas indeed lead to an entropy equality, as dictated by theory. Moreover, we would like to find evidence of a numerical solution aligned

(a) t = 0.3, 20 Elements   (b) t = 0.3, 50 Elements
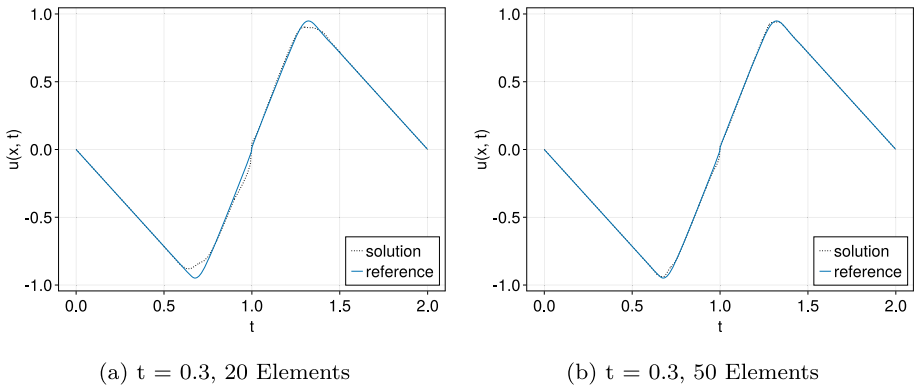
Fig. 4 Solution to the second initial condition calculated by the DDG scheme using polynomial degree $p = 6$

(a) t = 0.3, 20 Elements   (b) t = 0.3, 50 Elements
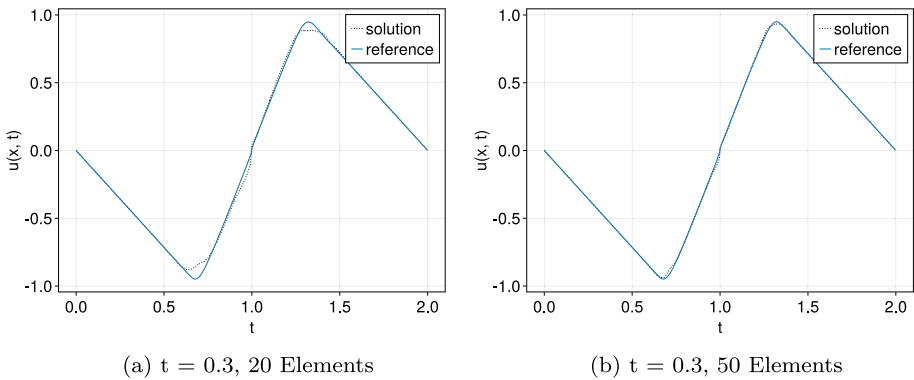
Fig. 5 Solution to the second initial condition calculated by the fully discrete scheme DRKDG using polynomials of degree $p = 6$

with Dafermos' entropy rate criterion. Therefore the same method as in [26] was used. There, the total entropy was calculated at fixed instants in time for the reference as well the numerical solution. Then, plots of these total entropies over time were used to compare several solvers, and it was found that some entropy dissipative solvers were able calculate approximate solutions with a higher total entropy that also reduced slower than the total entropy of the reference solution. One can therefore conclude that schemes whose entropy lies above the entropy of the reference solution seem to violate Dafermos' entropy criterion under two additional assumptions. First, that the entropy dissipation in the limit will be comparable, and second, that the reference solution in fact satisfies Dafermos' entropy rate criterion. The last assumption is provably true for the entropy solutions to one-dimensional scalar conservation laws with convex flux, as follows from the existence proof [39]. The total entropies of both solvers from this publication can be seen in Fig. 7. The solution of an extremely fine grained Godunov scheme with $N = 10000$ cells was added as a reference solution, to which the entropy of the numerical solutions by the modified DG schemes can be compared to. Both curves align nearly perfectly, with one exception. The entropy of the DG method starts to drop at nearly the same rate as the one of the Godunov method, but at a slightly earlier time. While the rates, apart from small oscillations in the DG method,

Logarithm to the basis 10 of the negative violation of the entropy equality for $p = 3$

(a)

$$\log_{10}\left(-\min\left(0, \frac{\mathrm{d}E_{u,T}}{\mathrm{d}t} + F_r^T - F_l^T\right)\right)$$

Logarithm to the basis 10 of the positive violation of the entropy equality for $p = 3$

(b)

$$\log_{10}\left(\max\left(0, \frac{\mathrm{d}E_{u,T}}{\mathrm{d}t} + F_r^T - F_l^T\right)\right)$$

Logarithm to the basis 10 of the negative violation of the entropy equality for $p = 6$

(c)

$$\log_{10}\left(-\min\left(0, \frac{\mathrm{d}E_{u,T}}{\mathrm{d}t} + F_r^T - F_l^T\right)\right)$$

Logarithm to the basis 10 of the positive violation of the entropy equality $p = 6$

(d)

$$\log_{10}\left(\max\left(0, \frac{\mathrm{d}E_{u,T}}{\mathrm{d}t} + F_r^T - F_l^T\right)\right)$$

Fig. 6 Negative and positive violation of the entropy equality for test case u_1 and polynomial degree $p = 3$ and $p = 6$. The positive violation is of the same magnitude as the machine precision

are comparable, the total entropy of the DG method is always lower than the one of the Godunov method. Against the first instinct, this is not an unpleasant result. The reason being that the DG method is based around a piece-wise polynomial approximation space. Therefore the method starts every timestep from an approximate solution and this solution seems to be unable to resolve the sharp discontinuity and instead smears it over up to 3 cells. This, together with conservation, seems to force a lower total entropy for the approximate solution. One can therefore conclude that our plan to enforce Dafermos' entropy rate criterion in DG schemes was successful.

5.3 Numerical Convergence Analysis

Our deviations from the basic DG scheme are only limited by an assumed error estimate. Future work could therefore be concentrated on deriving convergence rate estimates for smooth solutions. We will collect some numerical evidence that such convergence rate esti-
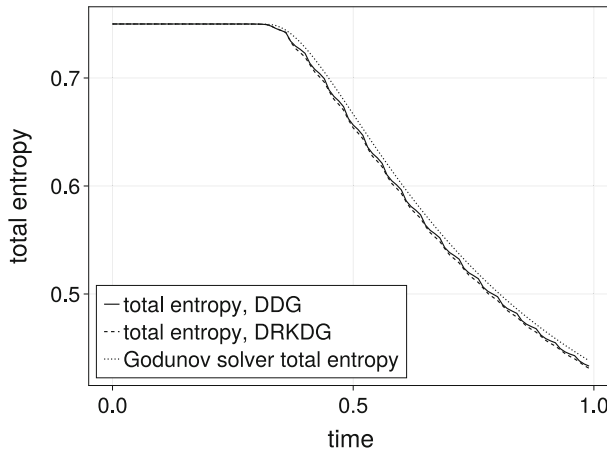
(a) p = 3
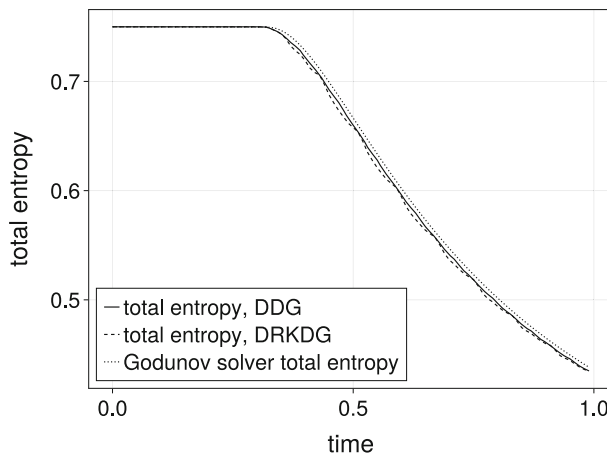
(b) p = 6

Fig. 7 Total entropy of the solution calculated by the DDG and DRKDG schemes in comparison with the total entropy of a solution calculated by a Godunov solver

mates are possible. This will be done by an experimental convergence analysis. As we will see, our schemes retain a high order of accuracy if high polynomial orders are used but this order can be one degree lower than the expected order of the base schemes [24, 53]. Once more Burgers' equation was solved with the initial conditions

$$u_s(x, 0) = 1 + \frac{\sin(\pi x)}{10}, \quad u_l(x, 0) = 1 + \frac{\sin(\pi x)}{50}$$

and a periodic domain $\Omega = [0, 2)$. The solutions of these problems are smooth on $t = [0, 1]$ for the initial condition u_s and $t = [0, 8]$ for the initial condition u_l. We look at the error at $t_{\text{end}} = 1$ or $t_{\text{end}} = 8$ of the solution calculated with $N \in \{10, 15, 20, 25, 30, 40, 50\}$ cells and internal polynomials of order $p \in \{3, 4, 5, 6, 7\}$. The relatively small maximum amount of cells was chosen as it was feared that otherwise the floating point accuracy could interfere
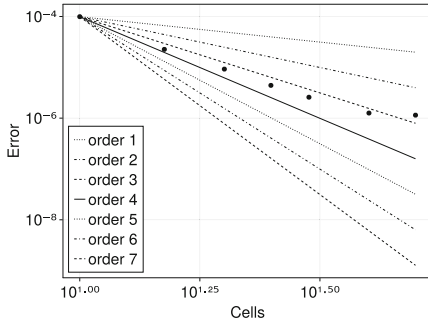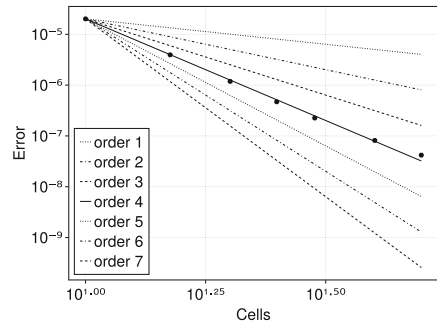
(a) Convergence analysis for polynomial order 3 in the 1-Norm.
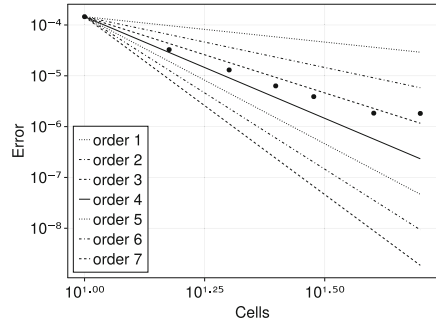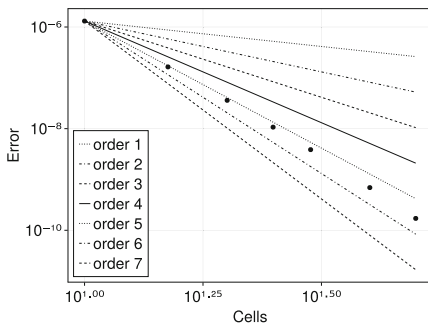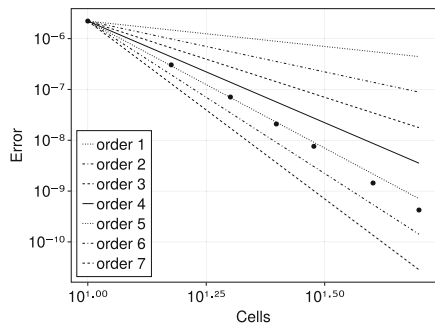
(b) Convergence analysis for polynomial order 3 in the 2-Norm.
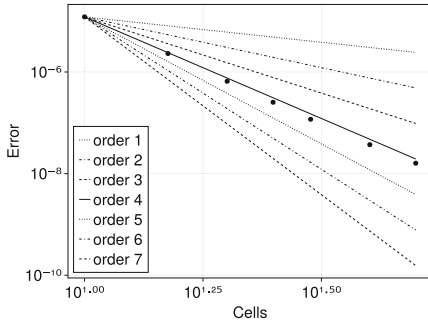
(c) Convergence analysis for polynomial order 4 in the 1-Norm.
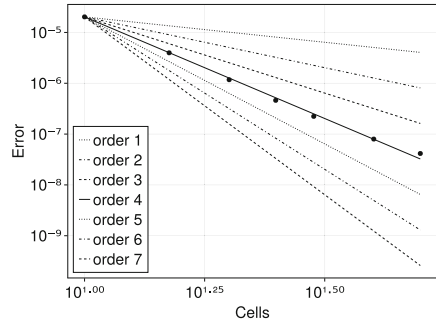
(d) Convergence analysis for polynomial order 4 in the 2-Norm.
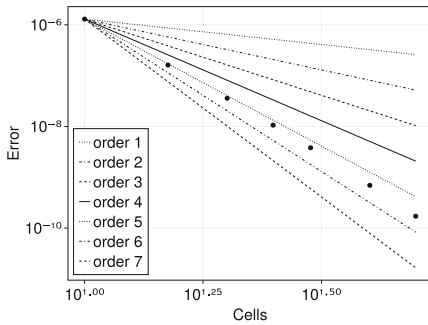
(e) Convergence analysis for polynomial order 5 in the 1-Norm.
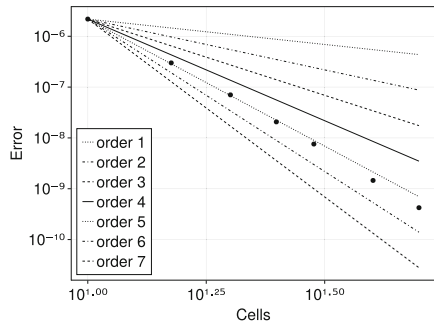
(f) Convergence analysis for polynomial order 5 in the 2-Norm.

Fig. 8 Convergence analysis for the semi-discrete scheme DDG, initial condition u_s solved up to $t = 1$

with the calculation of the error estimates, as the errors are already near the square root of the relative machine precision. Reference solutions were calculated with the method of characteristics. Time integration was carried out once more using the SSPRK33 method. The time step was not selected to keep $\lambda = \frac{\Delta t}{\Delta x}$ constant, but as $\lambda (\Delta x)^2 = \Delta t$ as this allow us to observe up to sixth order convergence rate without being limited by the accuracy of the time
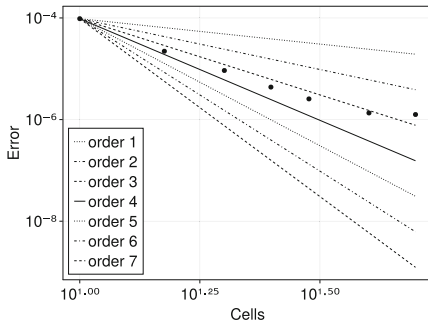
(a) Convergence analysis for polynomial order 3 in the 1-Norm for the discrete scheme.
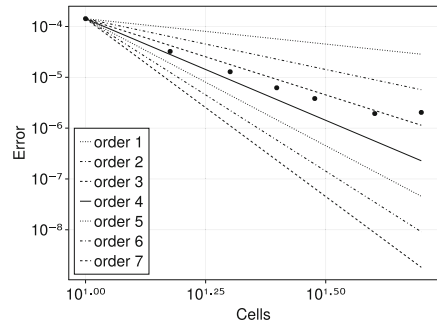
(b) Convergence analysis for polynomial order 3 in the 2-Norm for the discrete scheme.
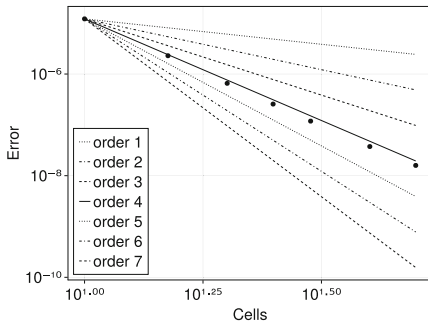
(c) Convergence analysis for polynomial order 4 in the 1-Norm for the discrete scheme.

(d) Convergence analysis for polynomial order 4 in the 2-Norm for the discrete scheme.

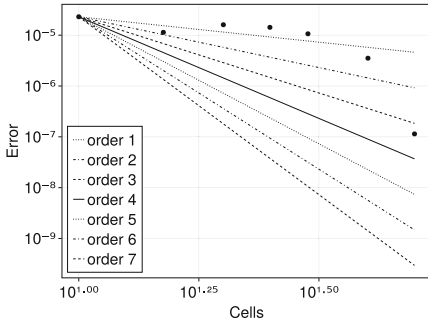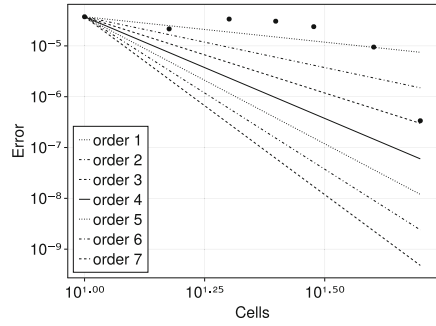(e) Convergence analysis for polynomial order 5 in the 1-Norm for the discrete scheme.

(f) Convergence analysis for polynomial order 5 in the 2-Norm for the discrete scheme.

Fig. 9 Convergence analysis for the fully discrete scheme DRKDG, initial condition u_s solved up to $t = 1$

integration method. The method displays p-th order of accuracy in our tests in Figs. 8 and 9 for the initial condition u_s. Note that orders above $p = 5$ are not shown, but also converge.

While this is a promising result some caution is advised, when using long integration times, as for example for u_l. The author witnessed that lower order versions of the scheme had problems to converge with high order. For example a scheme with $p = 5$ only converged with first order, consult Figs. 11 and 10. For order $p = 6$ and higher the problem vanished. These
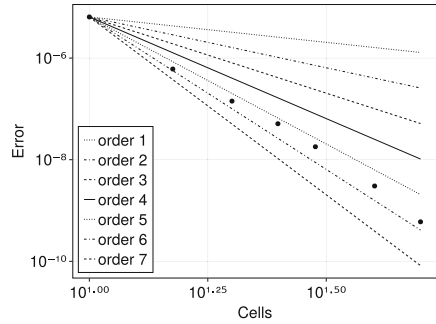
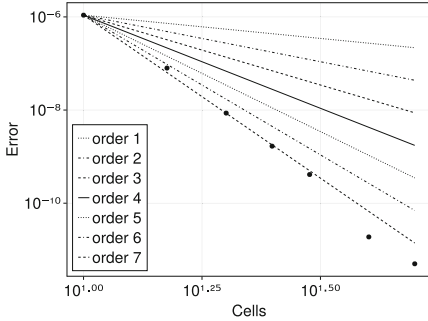(a) Convergence analysis for polynomial order 5 in the 1-Norm for the discrete scheme.

(b) Convergence analysis for polynomial order 5 in the 2-Norm for the discrete scheme.
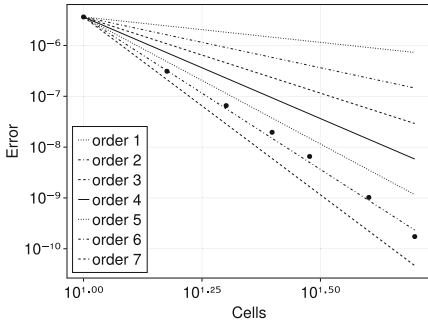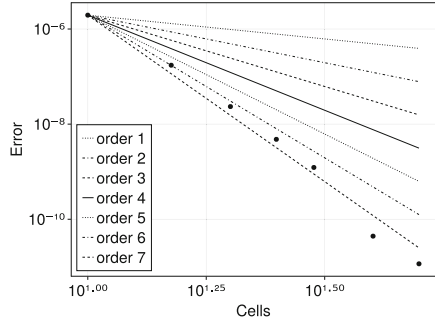
(c) Convergence analysis for polynomial order 6 in the 1-Norm for the discrete scheme.

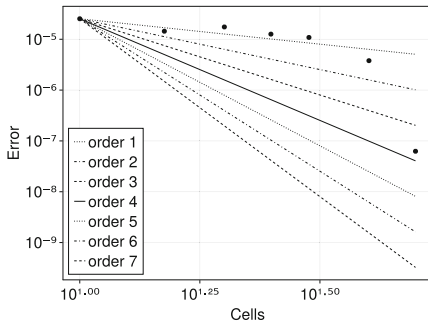(d) Convergence analysis for polynomial order 6 in the 2-Norm for the discrete scheme.

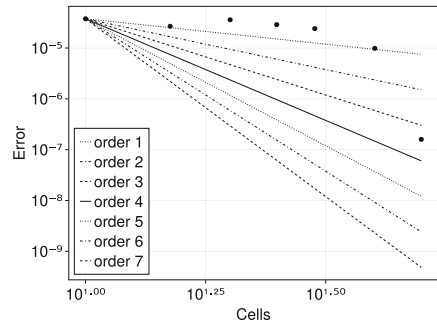(e) Convergence analysis for polynomial order 7 in the 1-Norm for the discrete scheme.

(f) Convergence analysis for polynomial order 7 in the 2-Norm for the discrete scheme.

Fig. 10 Convergence analysis for the fully discrete scheme DRKDG, initial condition u_l solved up to $t = 8$

problems could have several reasons. One of them being that the method was implemented without any modal filtering present. Higher polynomial orders could have lead to smaller aliasing errors and therefore mitigated the problem in the tests. Future tests will include modal filtering as an additional building block after the theoretical compatibility of modal filtering and the DG entropy descent was explored. See [22, 42] concerning modal filtering (Figs. 12 and 13).
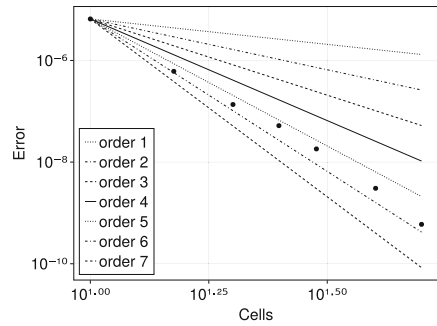
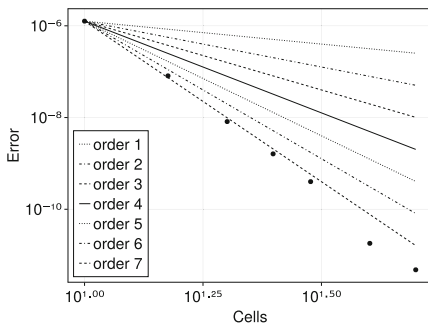(a) Convergence analysis for polynomial order 5 in the 1-Norm.

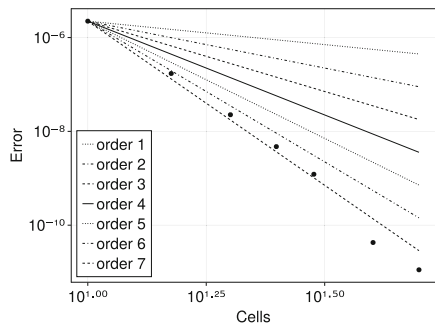(b) Convergence analysis for polynomial order 5 in the 2-Norm.
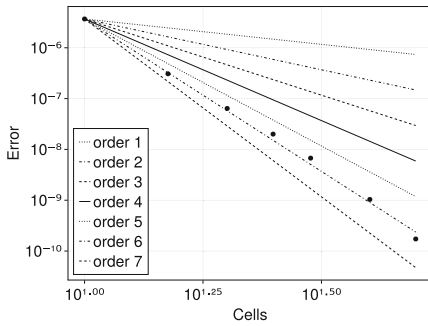
(c) Convergence analysis for polynomial order 6 in the 1-Norm.

(d) Convergence analysis for polynomial order 6 in the 2-Norm.

(e) Convergence analysis for polynomial order 7 in the 1-Norm.

(f) Convergence analysis for polynomial order 7 in the 2-Norm.

Fig. 11 Convergence analysis for the semi-discrete scheme DDG, initial condition u_l solved up to $t = 8$

5.4 Experimental Analysis of the Timestep Restriction

Sadly, the vanilla DG method suffers from low time step restrictions, i.e. the CFL number, the upper bound on the grid constant $\lambda = \frac{\Delta t}{\Delta x}$ with respect to the highest signal velocity, drops drastically with the used order [2, 7, 24]. Several methods have been devised to counteract this problem and since the limiting base FV scheme of our discretization does not have such a
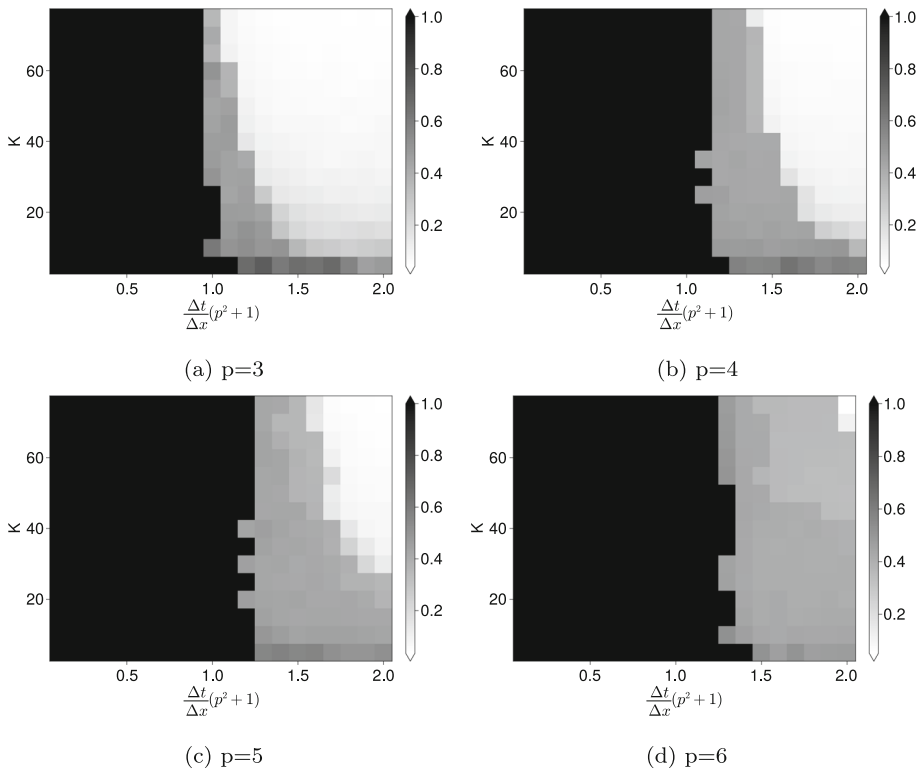
Fig. 12 Heatmap of the achieved maximum simulation time before the solution blows up for test case u_1 in relation to the used CFL number and the number of cells. The semidiscrete scheme DDG was used

low time step restriction we will test which effect our modifications have on the allowed time step. We use the first initial condition given, that was used as a test for discontinuous solutions. Note that the standard DG scheme blows up at $t \approx 0.3$ when a shock forms independently of the time step. We test the scheme with different CFL numbers and orders. We can therefore identify, if the maximum allowed time step scales linearly with the grid size and if, and how their ratio depends on the order of the used cells.

The results for our modified DG schemes are shown in Figs. 12 and 13. The shown heatmaps correspond to the simulation time at which a blowup occurred. The black areas correspond to the maximum simulation time of $t = 1$. There, no blowup occurred up to this time, where the simulation was stopped. Interestingly, the DRKDG scheme is able to use significantly bigger time steps than the DDG scheme, when the Grid is fine enough. Gains of a factor of 4 for polynomials of degrees 3 to 6 can be seen in Fig. 13 compared to Fig. 12. A problem in this regard concerns the accuracy of solutions calculated in this way. The simulations do not crash, but are pulled down to first order when used at to big time steps. This is especially dangerous if these schemes are used with the rule of thump to run them with time steps only barely stable.

6 Conclusion

In this publication, the author used error estimates between the exact entropy variables and the approximate entropy variables, error estimates between a fine sub cell scheme and the

(a) p=3

(b) p=4

(c) p=5

(d) p=6

Fig. 13 Heatmap of the achieved maximum simulation time before the solution blows up for test case u_1 in relation to the used CFL number and the number of cells for the discrete scheme DRKDG

approximate solution, and error estimates between a projection of the exact solution to the DG approximation space to control the entropy in DG methods. It was thereby possible to bound the derivative of the total entropy with the derivative of the total entropy of the limit of the sub cell scheme. This serves as a numerical treatment of the Dafermos entropy criterion in its original form with the sub cell solution as a reference weak solution to which the modified DG solution is compared concerning its total entropy dissipation. If one conjects that this solution already satisfies the Dafermos criterion, it follows further that the approximate solution of the DG method satisfies this criterion. Further, the author conjectured in a previous work that a numerical Dafermos entropy rate criterion should be centered around allowing only a minimization of the entropy rate if the additional residual produced by this action is small. This is especially important as the method would otherwise be able to dissipate entropy to fast by enlarging approximation errors. Because the scheme was designated as a correction to the classical DG method, where the correction is the biggest possible entropy dissipation in the allowed error margin, could the method be also interpreted as correctly implementing this criterion. As a side note a classical entropy inequality is also satisfied if one conjectures the correctness of the error estimators and scales the maximum descent direction using the formulae derived by the author. The same technique was also used to construct a fully discrete variant of the scheme. Apart from being able to calculate smooth solutions with high order of accuracy, as the standard DG method, shocks are also handled well without any additional stabilization. Especially oscillations in shocked cells are nearly invisible and do not produce

oscillations in other cells. Future work will focus on the application to multiple dimensions and systems of conservation laws. The author is sure that the derived error estimates can be also useful for other techniques in use with DG methods. These could be for example positivity preserving correction terms or artificial viscosity.

Data Availability The commented implementation of the schemes is available under https://github.com/simonius/dgdafermos.

Declarations

Competing interests The author has no relevant financial or non-financial interests to disclose.

References

1. Abgrall, R., Öffner, P., Ranocha, H.: Reinterpretation and extension of entropy correction terms for residual distribution and discontinuous Galerkin schemes: application to structure preserving discretization. J. Comput. Phys. **453**, 24 (2022). https://doi.org/10.1016/j.jcp.2022.110955
2. Chavent, G., Cockburn, B.: The local projection $P^0 - P^1$-discontinuous-Galerkin finite element method for scalar conservation laws. RAIRO Modélisation Math. Anal. Numér. **23**(4), 565–592 (1989). https://doi.org/10.1051/m2an/1989230405651
3. Chen, T., Shu, C.-W.: Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws. J. Comput. Phys. **345**, 427–461 (2017). https://doi.org/10.1016/j.jcp.2017.05.025
4. Chen, T., Shu, C.-W.: Review of entropy stable discontinuous galerkin methods for systems of conservation laws on unstructured simplex meshes. Trans. Appl. Math. (2020)
5. Cockburn, B., Johnson, C., Shu, C.-W., Tadmor, E.: Advanced numerical approximation of nonlinear hyperbolic equations. Lectures given at the 2nd session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Cetraro, Italy, June 23–28, 1997, volume 1697. Berlin: Springer (1998). https://doi.org/10.1007/BFb0096351
6. Cockburn, B.: An introduction to the discontinuous Galerkin method for convection-dominated problems. In: Advanced numerical approximation of nonlinear hyperbolic equations. Lectures given at the 2nd session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Cetraro, Italy, June 23–28, 1997, pp. 151–268. Berlin: Springer (1998)
7. Cockburn, B., Shu, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II: general framework. Math. Comput. **52**(186), 411–435 (1989). https://doi.org/10.2307/2008474
8. Dafermos, C.M.: The second law of thermodynamics and stability. Arch. Ration. Mech. Anal. **70**, 167–199 (1979). https://doi.org/10.1007/BF00250353
9. Dafermos, C.M.: Maximal dissipation in equations of evolution. J. Differ. Equ. **252**(1), 567–587 (2012). https://doi.org/10.1016/j.jde.2011.08.006

10. Dafermos, C.M.: The entropy rate admissibility criterion for solutions of hyperbolic conservation laws. J. Differ. Equ. **14**, 202–212 (1972)
11. Dafermos, C.M.: A variational approach to the Riemann problem for hyperbolic conservation laws. Discrete Contin. Dyn. Syst. **23**(1–2), 185–195 (2009). https://doi.org/10.3934/dcds.2009.23.185
12. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics, vol. 325. Springer, Berlin (2016). https://doi.org/10.1007/978-3-662-49451-6
13. Dedner, A., Giesselmann, J.: A posteriori analysis of fully discrete method of lines discontinuous Galerkin schemes for systems of conservation laws. SIAM J. Numer. Anal. **54**(6), 3523–3549 (2016). https://doi.org/10.1137/15M1046265
14. Diperna, R.J.: Uniqueness of solutions to hyperbolic conservation laws. Indiana Univ. Math. J. **28**, 137–188 (1979). https://doi.org/10.1512/iumj.1979.28.28011
15. Eduard, F.: Maximal dissipation and well-posedness for the compressible Euler system. J. Math. Fluid Mech. **16**(3), 447–461 (2014). https://doi.org/10.1007/s00021-014-0163-8
16. Fisher, T., Carpenter, M.: High-order entropy stable finite difference schemes for nonlinear conservation laws. Finite domains. NASA Technical Report, 02 2013. https://doi.org/10.1016/j.jcp.2013.06.014
17. Gassner, G.J.: A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods. SIAM J. Sci. Comput. **35**(3), a1233–a1253 (2013). https://doi.org/10.1137/120890144
18. Giesselmann, J., Makridakis, C., Pryer, T.: A posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws. SIAM J. Numer. Anal. **53**(3), 1280–1303 (2015). https://doi.org/10.1137/140970999
19. Glaubitz, J.: Stable high order quadrature rules for scattered data and general weight functions. SIAM J. Numer. Anal. **58**(4), 2144–2164 (2020). https://doi.org/10.1137/19M1257901
20. Glaubitz, J.: Stable high-order cubature formulas for experimental data. J. Comput. Phys. **447**, 22 (2021). https://doi.org/10.1016/j.jcp.2021.110693
21. Glaubitz, J., Gelb, A.: Stabilizing radial basis function methods for conservation laws using weakly enforced boundary conditions. J. Sci. Comput. **87**(2), 29 (2021). https://doi.org/10.1007/s10915-021-01453-8
22. Glaubitz, J., Öffner, P., Sonar, T.: Application of modal filtering to a spectral difference method. Math. Comput. **87**(309), 175–207 (2018). https://doi.org/10.1090/mcom/3257
23. Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**, 05 (2001). https://doi.org/10.1137/S003614450036757X
24. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods. Algorithms, Analysis, and Applications, volume 54 of Texts Appl. Math. New York, NY: Springer (2008)
25. Hillebrand, D., Klein, S.C., Öffner, P.: Comparison to control oscillations in high-order Finite Volume schemes via physical constraint limiters, neural networks and polynomial annihilation. arXiv:2203.00297 (2022)
26. Klein, S.-C.: Using the Dafermos entropy rate criterion in numerical schemes. BIT Numer. Math. 1–29 (2022)
27. Kruzhkov, S.N.: First order quasilinear equations with several independent variables. Mat. Sb., Nov. Ser., **81**, 228–255 (1970)
28. Lax, P.D.: Weak solutions of nonlinear hyperbolicequations and their numerical computation. Commun. Pure Appl. Math. **7**, 159–193 (1954)
29. Lax, P.D.: Shock Waves and Entropy. Contributions to Nonlinear Functional Analysis, pp. 603–634 (1971)
30. Lax, P.D.: Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves. SIAM (1973)
31. Lax, P.D.: Functional Analysis, Wiley Interscience (2002)
32. Liu, X.D., Osher, S., Chan, T.: Weighted essentially non-oscillatory schemes. J. Comput. Phys. **115**(1), 200–212 (1994). https://doi.org/10.1006/jcph.1994.1187
33. Markert, J., Gassner, G.: Comparison of different entropy stabilization techniques for discontinuous Galerkin spectral element methods. **12** (2022). https://doi.org/10.23967/eccomas.2022.066
34. Morton, K.W., Sonar, T.: Finite volume methods for hyperbolic conservation laws. Acta Numerica **16**, 155–238 (2007). https://doi.org/10.1017/S0962492906300013
35. Nesterov, Y.: Introductory lectures on convex optimization. A Basic Course, volume 87 of Appl. Optim. Boston: Kluwer Academic Publishers (2004)
36. Öffner, P., Ranocha, H.: Error boundedness of discontinuous Galerkin methods with variable coefficients. J. Sci. Comput. **79**(3), 1572–1607 (2019). https://doi.org/10.1007/s10915-018-00902-1
37. Öffner, P., Torlo, D.: Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes. Appl. Numer. Math. **153**, 15–34 (2020). https://doi.org/10.1016/j.apnum.2020.01.025

38. Ohlberger, M.: A review of a posteriori error control and adaptivity for approximations of non-linear conservation laws. Int. J. Numer. Methods Fluids **59**(3), 333–354 (2009). https://doi.org/10.1002/fld.1686

39. Oleĭnik, O.A.: Construction of a generalized solution of the Cauchy problem for a quasi-linear equation of first order by the introduction of 'vanishing viscosity'. Translated by George Birink. Transl., Ser. 2, Am. Math. Soc., **33**, 277–283 (1963). https://doi.org/10.1090/trans2/033/08

40. Ranocha, H.: On strong stability of explicit Runge–Kutta methods for nonlinear semibounded operators. IMA J. Numer. Anal. (2020). https://doi.org/10.1093/imanum/drz070

41. Ranocha, H., Öffner, P., Sonar, T.: Summation-by-parts operators for correction procedure via reconstruction. J. Comput. Phys. **311**, 299–328 (2016). https://doi.org/10.1016/j.jcp.2016.02.009

42. Ranocha, H., Glaubitz, J., Öffner, P., Sonar, T.: Stability of artificial dissipation and modal filtering for flux reconstruction schemes using summation-by-parts operators. Appl. Numer. Math. **128**, 1–23 (2018). https://doi.org/10.1016/j.apnum.2018.01.019

43. Shi, C., Shu, C.-W.: On local conservation of numerical methods for conservation laws. Comput. Fluids **169**, 3–9 (2018)

44. Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)

45. Sonar, T.: Strong and weak norm refinement indicators based on the finite element residual for compressible flow computation. I: the steady case. IMPACT Comput. Sci. Eng. **5**(2), 111–127 (1993). https://doi.org/10.1006/icse.1993.1005

46. Sonar, T., Süli, E.: A dual graph-norm refinement indicator for finite volume approximations of the Euler equations. Numer. Math. **78**(4), 619–658 (1998). https://doi.org/10.1007/s002110050328

47. Sonar, T., Warnecke, G.: On a posteriori error indication based on finite differences in triangular grids. ZAMM Z. Angew. Math. Mech., 78:s47–s48 (1998)

48. Sonntag, M., Munz, C.-D.: Shock capturing for discontinuous Galerkin methods using finite volume subcells. In: Fuhrmann, J., Ohlberger, M., Rohde, C., (eds.), Finite Volumes for Complex Applications VII-Elliptic, Parabolic and Hyperbolic Problems, pp. 945–953, Cham. Springer International Publishing (2014)

49. Tadmor, E.: The numerical viscosity of entropy stable schemes for systems of conservation laws. Math. Comput. **49**, 91–103 (1987)

50. Tadmor, E.: Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. Acta Numerica 451–512 (2003)

51. Tang, H.: On the sonic point glitch. J. Comput. Phys. **202**, 507–532 (2005). https://doi.org/10.1016/j.jcp.2004.07.013

52. Thomas, I., Sonar, T.: On a second order residual estimator for numerical schemes for nonlinear hyperbolic conservation laws. J. Comput. Phys. **171**(1), 227–242 (2001). https://doi.org/10.1006/jcph.2001.6784

53. Zhang, Q., Shu, C.-W.: Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. SIAM J. Numer. Anal. **42**(2), 641–666 (2004). https://doi.org/10.1137/S0036142902404182