

A class of high-order discontinuous-Galerkin methods satisfying infinitely many entropy conditions with provable error estimates and strong convergence for general nonlinear conservation laws

September 3, 2026

Yuanzhe Wei and Chi-Wang Shu

Abstract. We propose a novel framework for deriving semi-discrete discontinuous-Galerkin (DG) methods using operator semigroups for scalar conservation laws, then apply it to construct a class of high-order OFDG-type schemes [13] satisfying infinitely many local entropy inequalities with general E-fluxes on non-uniform meshes. Such schemes are further generalized to systems of conservation laws in any number of space dimensions by using entropy stable numerical fluxes in the sense of [1]. Finally, we prove optimal error estimates for smooth solutions to nonlinear scalar conservation laws, and prove strong convergence for discontinuous solutions to strictly convex conservation laws via compensated compactness.

1. Introduction

Hyperbolic conservation laws arise throughout continuum mechanics and describe the propagation of nonlinear waves that may develop discontinuities even from smooth initial data. For such problems, a numerical method must reconcile several competing objectives: high-order accuracy in smooth regions, nonlinear stability near shocks, consistency with the entropy condition that selects the physically relevant weak solution, and convergence as the mesh is refined. Discontinuous Galerkin (DG) methods are particularly attractive in this setting because they combine high-order polynomial approximation, local conservation and geometric flexibility. Since the development of Runge–Kutta discontinuous Galerkin (RKDG) methods for nonlinear conservation laws [3], substantial effort has therefore been devoted to understanding their stability, entropy properties, and error estimates; see, for example, [1, 9, 19, 21] and the references therein.

Hyperbolic conservation laws generally admit infinitely many weak solutions. For scalar conservation laws, the only physically relevant solution, called the *entropy solution*, is a

Keywords: discontinuous Galerkin methods, entropy inequalities, nonlinear conservation laws, error estimates, compensated compactness.

special weak solution uniquely characterized by a family of conditions known as *entropy inequalities* [4]. Any numerical method that preserves sufficiently many entropy inequalities on the discrete level is particularly favorable since it can often be shown to converge strongly to the entropy solution following DiPerna's argument¹, see [2, 5]. First order monotone schemes inherit entropy inequalities for a broad class of convex entropies [8], whereas the situation for high-order approximations is more delicate. A classical semi-discrete DG method with a monotone numerical flux satisfies a local entropy inequality for the square entropy [9], but quadratic entropy alone does not encode the full family of entropy conditions associated with the underlying conservation law. On the other hand, the entropy-stable DG method developed in [1] admits an arbitrarily prescribed entropy other than the square entropy, but still cannot include more than one entropy. These developments motivate the following question: can one construct a high-order DG method that retains the familiar structure of the classical method while enforcing substantially more entropy inequalities, without sacrificing optimal accuracy for smooth solutions?

The starting point of this paper is an operator interpretation of the classical DG method. Let S_t denote the entropy-solution semigroup of the conservation law and let π denote the cellwise L^2 -projection onto the DG space. We show that the classical semi-discrete DG method with the Godunov flux can be recovered from the infinitesimal action of the composite operator $\pi S_{\Delta t}$ as $\Delta t \rightarrow 0$. This viewpoint separates two mechanisms that are usually intertwined in the standard weak-form derivation: the exact evolution $S_{\Delta t}$ carries the complete entropy structure of the conservation law, while the projection back to the finite-dimensional polynomial space determines which part of that structure survives in the discretization. In particular, the L^2 -projection is compatible with the square entropy, so that the classical DG scheme admits the square entropy inequality, which was first proved in [9]. This observation suggests modifying the projection in order to enforce additional entropy inequalities.

Following this idea, we introduce a nonlinear projection² that contracts both the square entropy and an arbitrarily prescribed strictly convex entropy. Its infinitesimal operator leads to a DG method consisting of the classical DG spatial operator plus a cellwise damping term acting only on the nonconstant modes. We then derive computable upper bounds for the damping coefficient. The resulting schemes preserve cell averages and conservation and, with general E-fluxes, satisfy local entropy inequalities for the prescribed entropy. More importantly, by choosing bounds that are uniform over suitable classes of convex entropies, the same numerical solution satisfies infinitely many local entropy inequalities. We also introduce simplified damping coefficients whose structure is closely related to the oscillation-free discontinuous Galerkin (OFDG) methodology of [11, 13], while retaining the entropy estimates needed in our analysis.

¹If strong convergence cannot be proved, a weaker result is still valid by the Lax–Wendroff theorem [10], which states that if the numerical solution converges strongly with bounded total variation, then the limit is the unique entropy solution.

²This nonlinear projection resembles the limiter in [12].

The operator construction extends beyond one-dimensional scalar equations. For one-dimensional systems, we formulate the method for a general strictly convex entropy pair and entropy-stable numerical fluxes in the sense of [1]. We further extend the framework to systems in arbitrary space dimensions on general convex, shape-regular meshes. In this setting, the additional damping again acts on the deviation from the element average, and entropy stability at element interfaces is supplied by numerical fluxes satisfying the entropy-stability condition used in [1]. Thus the same basic mechanism—entropy-compatible evolution followed by a suitably controlled projection—provides a unified construction for scalar equations and systems.

A central issue is whether the additional dissipation required by the entropy conditions degrades the high-order accuracy of DG methods. We show that it does not. For smooth solutions of general nonlinear scalar conservation laws on quasi-uniform meshes, the proposed semi-discrete schemes achieve the optimal L^2 error estimate $O(h^{k+1})$ when an upwind flux is used, and an $O(h^{k+1/2})$ estimate for a general monotone flux. The proof builds on the framework developed for nonlinear conservation laws in [21], together with estimates showing that the new damping terms are sufficiently high order in smooth regions.

We also study convergence when the exact solution contains discontinuities. For strictly convex scalar fluxes, assuming a uniform L^∞ bound on the numerical solutions and any E-flux, we prove strong convergence of the proposed approximations to the entropy solution. The argument uses the compensated-compactness framework of Tartar [18] and the convergence theory for approximate solutions of conservation laws developed by DiPerna [5]. In particular, the entropy dissipation built into the schemes (which is absent from the classical DG or OFDG) provides the necessary estimates for proving the compactness of entropy productions in H_{loc}^{-1} .

The main contributions of this paper can therefore be summarized as follows: we give an operator-semigroup derivation of the classical semi-discrete DG method; construct high-order DG schemes satisfying infinitely many local entropy inequalities; extend the construction to systems and multiple space dimensions; establish optimal-order error estimates for smooth scalar solutions; and prove strong convergence for discontinuous solutions of strictly convex scalar conservation laws. Numerical experiments are included to examine the accuracy and robustness of the proposed damping mechanisms and to illustrate the effect of the parameters in practice.

The class of DG schemes developed in this paper can be summarized as follows:

1D scalar: In Section 3 we introduce three schemes for solving 1D scalar conservation laws. They are named (Scheme B), (Scheme C) and (Scheme D) after their damping coefficients B_j , C_j and D_j . If the damping coefficients are chosen according to Corollary 3.13, (Scheme B) and (Scheme C) would satisfy infinitely many entropy inequalities. Moreover, (Scheme D) also satisfies infinitely many entropy inequalities if the numerical solution remains bounded. Because (Scheme D) is the easiest to implement, it is adopted in the numerical tests in Section 8.

1D system: In Section 4 we introduce three schemes for solving 1D systems of conservation laws. Exactly parallel to the 1D scalar case, they are named (**Scheme B'**), (**Scheme C'**) and (**Scheme D'**) after their damping coefficients B'_j , C'_j and D'_j . We use (**Scheme D'**) for the numerical tests in Section 8.

Multidimensional system: In Section 5 we introduce two schemes for solving systems of conservation laws in higher space dimension. Same as before, they are named (**Scheme B''**) and (**Scheme D''**) after their damping coefficients B''_j and D''_j . We use (**Scheme D''**) for the numerical tests in Section 8.

We shall use the general term *Operator DG* to denote any of these schemes mentioned above, since they are derived from our operator framework. They are closely related to the OFDG methods in the sense that they all contain a cellwise damping term acting on the nonconstant modes; the only difference is that, in our schemes, the damping coefficients have an additional term for controlling the oscillations on the interior of each cell, which provides the entropy estimates needed in our analysis. The remainder of the paper is organized as follows. Section 2 develops the operator framework and reviews the classical DG method. Section 3 constructs the Operator DG schemes for one-dimensional scalar conservation laws. Sections 4 and 5 extend the framework to one-dimensional and multi-dimensional systems. Section 6 proves the error estimates for smooth scalar solutions, and Section 7 establishes strong convergence via compensated compactness. The numerical results are presented thereafter, followed by concluding remarks in Section 9, and supplementary materials in the appendix concerning the HLL-based operator construction and sharper estimates for constants appearing in the entropy bounds.

2. An operator framework for the classical DG method

Consider the Cauchy problem for a nonlinear scalar conservation law in one space dimension (with I being any interval):

$$\begin{cases} \mathcal{U}_t(x, t) + f(\mathcal{U}(x, t))_x = 0, & x \in I, \quad t \geq 0. \\ \mathcal{U}(x, 0) = \mathcal{U}_0(x). \end{cases} \quad (2.1)$$

To solve (2.1) numerically, the interval I is divided into N disjoint sub-intervals $I_1, I_2, \dots, I_N$, where $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, $|I_j| = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}} = h_j$, $I = \cup_{j=1}^N I_j$. For any function $v(x)$ that is continuous on I_j , define $v_{j-\frac{1}{2}}^+ := v(x_{j-\frac{1}{2}}^+)$ and $v_{j+\frac{1}{2}}^- := v(x_{j+\frac{1}{2}}^-)$ to be the left and right end-point limit of v on I_j respectively. Define for $k \geq 1$ the space

$$\mathbb{V}^k := \{v \in L^2(I) : v|_{I_j} \in \mathbb{P}^k(I_j), 1 \leq j \leq N\}$$

of all piecewise (possibly discontinuous) polynomials on I that coincides with a k -th order polynomial on each cell I_j . From now on, we reserve the letter k to denote the polynomial order. The classical semi-discrete DG method with 1D $\mathbb{P}^k$ elements for solving the nonlinear

scalar conservation law (2.1) can be stated as follows: Seek $u(\cdot, t) \in \mathbb{V}^k$ such that

$$\int_{I_j} u_t(x, t) v(x) dx - \int_{I_j} f(u(x, t)) v_x(x) dx + \hat{f}_{j+\frac{1}{2}}(t) v_{j+\frac{1}{2}}^- - \hat{f}_{j-\frac{1}{2}}(t) v_{j-\frac{1}{2}}^+ = 0$$

(Classical DG)

holds for all $v \in \mathbb{V}^k$ and all $j = 1, 2, \dots, N$, where the numerical flux $\hat{f}_{j+\frac{1}{2}}(t) = \hat{f}(u(x_{j+\frac{1}{2}}^-, t), u(x_{j+\frac{1}{2}}^+, t))$ is Lipschitz continuous and consistent ($\hat{f}(u, u) = f(u)$). Clearly, (Classical DG) can be rewritten into the following form:

$$\frac{\partial u}{\partial t}(x, t) = F[u(\cdot, t)](x), \quad \forall x \in I, \quad (2.2)$$

where $F[v] \in \mathbb{V}^k$ for any $v \in \mathbb{V}^k$ is a piecewise polynomial whose coefficients depend on v . Moreover, (2.2) is an ODE system since $u(\cdot, t)$ lies in a finite dimensional space $\mathbb{V}^k$. The initial condition of (2.2) is taken as any projection of $\mathcal{U}_0(x)$ onto $\mathbb{V}^k$; this can be the L^2 -projection, the Gauss-Radau projection (see Section 6), or interpolation on point values. Let $\Phi_t : \mathbb{V}^k \mapsto \mathbb{V}^k$ denote the flow map of this ODE system, which maps any initial condition to the solution at time t . Since the right-hand side of the ODE is locally Lipschitz continuous, the solution is uniquely defined at least for t sufficiently small; moreover, one can show that the solution is always bounded (see the discussion below), so that Φ_t is well-defined for all $t \geq 0$. Then, (2.2) can be equivalently expressed as

$$\frac{\partial}{\partial t} \Phi_t u = F[\Phi_t u], \quad u \in \mathbb{V}^k.$$

(Classical DG) is most commonly derived by multiplying the test function v on both sides of (2.1), then integrate by parts on I_j , and finally replace the flux in the boundary terms by the numerical flux (which makes the scheme conservative) [3]. If the monotone flux is used, [9] showed that the numerical solution of (Classical DG) is L^2 -stable:

$$\|\Phi_{t_2} u\|_{L^2(I)} \leq \|\Phi_{t_1} u\|_{L^2(I)}, \quad \forall 0 \leq t_1 \leq t_2, \quad \forall u \in \mathbb{V}^k,$$

which is a property shared by the unique entropy solution of (2.1).

We now introduce a very different approach to deriving (Classical DG), which proves to be more natural and provides deeper insight into its fundamental structure. First, let us introduce the exact solution operator S_t , $t \geq 0$, associated with (2.1), which maps a given initial condition to the unique entropy solution at time t . In other words,

$$S_t \mathcal{U}_0(x) = \mathcal{U}(x, t), \quad \forall t \geq 0,$$

where $\mathcal{U}_0(x)$ and $\mathcal{U}(x, t)$ are the same as in (2.1). For 1D scalar conservation laws, S_t is well-defined for all $t \geq 0$ as a map from $L^\infty(\mathbb{R}) \cup L^1(\mathbb{R})$ to $L^\infty(\mathbb{R}) \cup L^1(\mathbb{R})$ [4]. Furthermore, let $\pi : L^2(I) \mapsto \mathbb{V}^k$ denote the L^2 -projection onto the space of piecewise polynomials, namely, for $w \in L^2(I)$, $\pi w \in \mathbb{V}^k$ is characterized by

$$\int_I (\pi w - w) v dx = 0, \quad \forall v \in \mathbb{V}^k.$$

Then we have the following theorem, which provides the alternative derivation of **(Classical DG)** when $\hat{f}$ is the Godunov flux:

Theorem 2.1. *For any $u \in \mathbb{V}^k$, we have*

$$\lim_{\Delta t \rightarrow 0} \frac{\pi S_{\Delta t} u - u}{\Delta t} = F[u], \quad (2.3)$$

uniformly on I , where $F[\cdot] \in \mathbb{V}^k$ is the right-hand side of (2.2) and $\hat{f}$ is the Godunov flux:

$$\hat{f}(u^-, u^+) = f(u^*), \quad (2.4)$$

where u^* is the value at $\frac{x}{t} = 0$ of the entropy solution to the Riemann problem with left and right states u^- and u^+ at $x = 0$. If u^- and u^+ are connected by a stationary shock, u^* may be taken as either u^- or u^+ , since the Rankine–Hugoniot condition yields $f(u^-) = f(u^+)$.

Proof. Since $u \in \mathbb{V}^k$, it is smooth on the interior of each I_j , but can have discontinuities at the cell interfaces. Let $\Delta t > 0$ be sufficiently small such that the evolutions of discontinuities at different cell interfaces do not interact, and no new discontinuity (including the discontinuity in derivatives of all orders) is formed. For the Riemann problem with left and right states $u_{j-\frac{1}{2}}^-$ and $u_{j-\frac{1}{2}}^+$, let $\lambda_{j-\frac{1}{2}}^-$ and $\lambda_{j-\frac{1}{2}}^+$ be lower and upper bounds, respectively, for all wave speeds in the Riemann solution. To be more conservative, we assume

$$\lambda_{j-\frac{1}{2}}^- \leq \min_{0 \leq \theta \leq 1} f'(\theta u_{j-\frac{1}{2}}^- + (1-\theta)u_{j-\frac{1}{2}}^+), \quad \lambda_{j-\frac{1}{2}}^+ \geq \max_{0 \leq \theta \leq 1} f'(\theta u_{j-\frac{1}{2}}^- + (1-\theta)u_{j-\frac{1}{2}}^+). \quad (2.5)$$

In each I_j , denote the polluted regions on the left and right by

$$I_j^+(\Delta t) := \left[x_{j-\frac{1}{2}}, x_{j-\frac{1}{2}} + \max\{0, \lambda_{j-\frac{1}{2}}^+\} \Delta t \right], \quad I_j^-(\Delta t) := \left[x_{j+\frac{1}{2}} - \min\{0, \lambda_{j+\frac{1}{2}}^-\} \Delta t, x_{j+\frac{1}{2}} \right]. \quad (2.6)$$

These are the parts of I_j that might be influenced by the discontinuity of u or its derivatives at $x_{j-\frac{1}{2}}$ and $x_{j+\frac{1}{2}}$, respectively. Thus, $S_{\Delta t} u$ remains smooth on $I_j \setminus (I_j^+(\Delta t) \cup I_j^-(\Delta t))$. Around each $x_{j-\frac{1}{2}}$, the discontinuity of u at $x_{j-\frac{1}{2}}$ can only influence the region $I_{j-1}^-(\Delta t) \cup I_j^+(\Delta t)$ by the time Δt .

Then, for any $v \in \mathbb{V}^k$, we have

$$\begin{aligned} & \int_{I_j} \frac{\pi S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \\ &= \int_{I_j} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \quad (\text{since } v|_{I_j} \text{ is a polynomial}) \\ &= \int_{I_j \setminus (I_j^+(\Delta t) \cup I_j^-(\Delta t))} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \\ & \quad + \int_{I_j^+(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx + \int_{I_j^-(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \end{aligned}$$

$$\begin{aligned}
&= - \int_{I_j} f(u)_x \cdot v(x) \, dx + O(\Delta t) \\
&\quad + \int_{I_j^+(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx + \int_{I_j^-(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx,
\end{aligned}$$

where the last equality follows from the fact that both u and $S_{\Delta t} u$ are smooth on $I_j \setminus (I_j^+(\Delta t) \cup I_j^-(\Delta t))$. The notation $O(\Delta t)$ stands for any term that can be upper bounded by $C\Delta t$, where $C > 0$ is a constant depending on the norms of derivatives of u on the interior of I_j . Since eventually we would let $\Delta t \rightarrow 0$ for a fixed u and on a fixed mesh, the dependency of the constant C on u or h_j 's need not concern us.

By taking $\Delta t \rightarrow 0$ in the last expression, we obtain

$$\begin{aligned}
\lim_{\Delta t \rightarrow 0} \int_{I_j} \frac{\pi S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx &= - \int_{I_j} f(u)_x \cdot v(x) \, dx + v_{j-\frac{1}{2}}^+ \lim_{\Delta t \rightarrow 0} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \, dx \\
&\quad + v_{j+\frac{1}{2}}^- \lim_{\Delta t \rightarrow 0} \int_{I_j^-(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \, dx. \tag{2.7}
\end{aligned}$$

To compute

$$\lim_{\Delta t \rightarrow 0} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \, dx,$$

we first observe that the limit would not change if we replace u in a neighborhood of $x_{j-\frac{1}{2}}$ by

$$\tilde{u}(x) = \begin{cases} u_{j-\frac{1}{2}}^-, & x < x_{j-\frac{1}{2}} \\ u_{j-\frac{1}{2}}^+, & x > x_{j-\frac{1}{2}} \end{cases} \tag{2.8}$$

since $S_{\Delta t} u(x) = S_{\Delta t} \tilde{u}(x) + O(\Delta t)$ for $x \in I_j^+(\Delta t)$. But $S_{\Delta t} \tilde{u}(x)$ is simply the Riemann solution with left and right states $u_{j-\frac{1}{2}}^-$ and $u_{j-\frac{1}{2}}^+$ at $x_{j-\frac{1}{2}}$. Because the Riemann solution is self-similar, set

$$R(\xi) := S_t \tilde{u}(x), \quad \xi = \frac{x - x_{j-\frac{1}{2}}}{t}. \tag{2.9}$$

By definition, all right-going waves of $S_{\Delta t} \tilde{u}(x)$ are contained in $I_j^+(\Delta t) = [x_{j-\frac{1}{2}}, x_{j-\frac{1}{2}} + a\Delta t]$, where $a = \max\{0, \lambda_{j-\frac{1}{2}}^+\}$. Then,

$$\begin{aligned}
\lim_{\Delta t \rightarrow 0} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \, dx &= \lim_{\Delta t \rightarrow 0} \int_{x_{j-\frac{1}{2}}}^{x_{j-\frac{1}{2}} + a\Delta t} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \, dx \\
&= \lim_{\Delta t \rightarrow 0} \int_{x_{j-\frac{1}{2}}}^{x_{j-\frac{1}{2}} + a\Delta t} \frac{S_{\Delta t} \tilde{u}(x) - \tilde{u}(x)}{\Delta t} \, dx \\
&= \int_0^a R(\xi) - u_{j-\frac{1}{2}}^+ \, d\xi. \tag{2.10}
\end{aligned}$$

Using the conservation law for the self-similar Riemann solution, we have

$$-\xi R'(\xi) + \frac{d}{d\xi} f(R(\xi)) = 0$$

in the sense of distributions. Hence

$$\frac{d}{d\xi} [\xi R(\xi) - f(R(\xi))] = R(\xi) \quad (2.11)$$

also holds in the sense of distributions, and therefore,

$$\int_0^a R(\xi) d\xi = aR(a) - f(R(a)) + f(R(0^+)).$$

Because a lies to the right of all Riemann waves,

$$R(a) = u_{j-\frac{1}{2}}^+.$$

Furthermore, by the definition of the Godunov flux,

$$f(R(0^+)) = \hat{f}(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+) = \hat{f}_{j-\frac{1}{2}}.$$

Consequently,

$$\int_0^a R(\xi) d\xi = au_{j-\frac{1}{2}}^+ - f(u_{j-\frac{1}{2}}^+) + \hat{f}_{j-\frac{1}{2}},$$

which, together with (2.10), implies

$$\lim_{\Delta t \rightarrow 0} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} dx = \int_0^a R(\xi) - u_{j-\frac{1}{2}}^+ d\xi = \hat{f}_{j-\frac{1}{2}} - f(u)_{j-\frac{1}{2}}^+. \quad (2.12)$$

In the same way, we can also prove

$$\lim_{\Delta t \rightarrow 0} \int_{I_j^-(\Delta t)} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} dx = -\hat{f}_{j+\frac{1}{2}} + f(u)_{j+\frac{1}{2}}^-. \quad (2.13)$$

Using (2.12) and (2.13) in (2.7) and integrate by parts on $\int_{I_j} f(u)_x \cdot v(x) dx$, we obtain (Classical DG). Since

$$\lim_{\Delta t \rightarrow 0} \frac{\pi S_{\Delta t} u - u}{\Delta t} \in \mathbb{V}^k$$

and v is an arbitrary element in $\mathbb{V}^k$, we have proved (2.3) holds point-wise, thus also uniformly because both sides of (2.3) are piecewise polynomials defined on the same grid. ■

A simple corollary from the above theorem is that the limit $\lim_{\Delta t \rightarrow 0} \frac{\pi S_{\Delta t} u - u}{\Delta t}$ exists for all $u \in \mathbb{V}^k$. In particular, $\lim_{\Delta t \rightarrow 0} \pi S_{\Delta t} u = u$ uniformly on I . Since $S_{\Delta t}$ and π both decrease the L^2 -norm, we have $\|\pi S_{\Delta t} u\|_{L^2(I)} \leq \|u\|_{L^2(I)}$, and the following corollary is immediate after passing to the limit as $\Delta t \rightarrow 0$:

Corollary 2.2. *Under periodic or compactly supported boundary conditions, (Classical DG) with Godunov flux is L^2 -stable in the sense that*

$$\|\Phi_{t_2} u\|_{L^2(I)} \leq \|\Phi_{t_1} u\|_{L^2(I)}, \quad \forall 0 \leq t_1 \leq t_2, \quad \forall u \in \mathbb{V}^k.$$

Proof. It suffices to estimate the time derivative of $\|\Phi_t u\|_{L^2(I)}^2$:

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\Phi_t u\|_{L^2(I)}^2 &= \int_I \Phi_t u(x) \cdot \frac{\partial}{\partial t} \Phi_t u(x) dx = \int_I \Phi_t u(x) \cdot F[\Phi_t u](x) dx \\ &= \lim_{\Delta t \rightarrow 0} \int_I \Phi_t u \cdot \frac{\pi S_{\Delta t} \Phi_t u - \Phi_t u}{\Delta t} dx \quad (\text{Theorem 2.1}) \\ &= \lim_{\Delta t \rightarrow 0} \int_I \frac{\pi S_{\Delta t} \Phi_t u + \Phi_t u}{2} \cdot \frac{\pi S_{\Delta t} \Phi_t u - \Phi_t u}{\Delta t} dx \\ &= \frac{1}{2} \lim_{\Delta t \rightarrow 0} \int_I \frac{(\pi S_{\Delta t} \Phi_t u)^2 - (\Phi_t u)^2}{\Delta t} dx \\ &= \frac{1}{2} \lim_{\Delta t \rightarrow 0} \frac{\|\pi S_{\Delta t} \Phi_t u\|_{L^2(I)}^2 - \|\Phi_t u\|_{L^2(I)}^2}{\Delta t} \leq 0, \end{aligned}$$

where the last inequality follows from the fact that $\|\pi S_{\Delta t} \Phi_t u\|_{L^2(I)}^2 \leq \|\Phi_t u\|_{L^2(I)}^2$. ■

In fact, since $S_{\Delta t} u$ is the entropy solution, it satisfies all entropy inequalities; and since π decreases the L^2 -norm on each cell, the scheme actually admits a local square-entropy inequality with boundary flux determined by the flux of the true solution $S_{\Delta t} u$, which was originally proved in [9]. The next corollary is not needed in the subsequent discussion, but nevertheless might be conceptually helpful.

Corollary 2.3. *Fix an arbitrary time $T > 0$ and $u \in \mathbb{V}^k$. Let n be a positive integer. Then,*

$$\lim_{n \rightarrow \infty} (\pi S_{\frac{T}{n}})^n u = \Phi_T u$$

uniformly on I .

Proof. The proof of Theorem 2.1 implies that the discrete map $u \mapsto \pi S_{\frac{T}{n}} u$ is a first-order time discretization of the ODE (2.2) with uniform time step $\frac{T}{n}$. Since the ODE itself is Lipschitz continuous, and for a fixed $u \in \mathbb{V}^k$, the first-order time derivative of $\Phi_t u(x)$ is uniformly bounded for all $x \in I$ and $0 \leq t \leq T$. The proof follows from a routine estimate through truncation errors. ■

In light of Corollary 2.3, the result of Corollary 2.2 is strikingly clear since

$$\|(\pi S_{\Delta t})^n u\|_{L^2(I)} \leq \|(\pi S_{\Delta t})^m u\|_{L^2(I)}, \quad \forall n \geq m, \quad \forall u \in \mathbb{V}^k.$$

It is worth noting that, since $S_{\Delta t}$ is the exact solution operator, $\pi S_{\Delta t}$ in the limit can only generate (Classical DG) with Godunov flux; however, a slightly modified operator can generate (Classical DG) with a general HLL flux. The discussion of how to modify $S_{\Delta t}$ and the computation of the limit is contained in Appendix A. This might be of interest if one wishes to generalize this operator framework to other types of PDEs.

3. The Operator DG method for 1D scalar conservation laws

Let $(U, \mathcal{F})$ be any fixed entropy pair. Assume $U \in C^3$ and strictly convex ($U'' > 0$ on $\mathbb{R}$). The entropy flux $\mathcal{F}$ is defined by

$$\mathcal{F}'(u) = U'(u)f'(u). \quad (3.1)$$

In addition, let

$$\psi(u) := U'(u)f(u) - \mathcal{F}(u), \quad (3.2)$$

so that $\psi'(u) = U''(u)f(u)$. Our next goal is to construct a scheme that satisfies two entropy inequalities: one for the square-entropy, and the other for $U(\cdot)$. In Section 2, we have shown that (**Classical DG**) with the Godunov flux can be generated by the operator $\pi S_{\Delta t}$ as $\Delta t \rightarrow 0$. The reason that it can only admit the square-entropy inequality is because π is the L^2 -projection, which can only decrease the integral of the square function on each cell but not for other convex functions. But $S_{\Delta t}$ admits all entropy pairs, whose potential is lost after applying π . Therefore, to realize the full potential of $S_{\Delta t}$, we should replace π by a different projection operator $\tilde{\pi}$ such that it can decrease both the integral of the square function and $U(\cdot)$. With this $\tilde{\pi}$, we shall derive a new scheme generated by $\tilde{\pi} S_{\Delta t}$, which will look very similar to the classical DG scheme with Godunov flux and satisfy entropy inequalities for both the square-entropy and $U(\cdot)$. Then, we will further modify it such that: (1) the Godunov flux can be replaced by any E-flux; (2) the scheme satisfies infinitely many entropy conditions; (3) it still retains its optimal order of accuracy.

In order to derive the scheme generated by the operator $\tilde{\pi} S_{\Delta t}$, we need to compute:

$$\begin{aligned} & \int_{I_j} \frac{\tilde{\pi} S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \\ &= \int_{I_j} \frac{\pi S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx + \int_{I_j} \frac{\tilde{\pi} S_{\Delta t} u(x) - \pi S_{\Delta t} u(x)}{\Delta t} \cdot v(x) \, dx \\ &= \int_{I_j} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx + \int_{I_j} \frac{\tilde{\pi} S_{\Delta t} u(x) - \pi S_{\Delta t} u(x)}{\Delta t} \cdot v(x) \, dx. \end{aligned} \quad (3.3)$$

For any function $v \in L^2(I)$, let $\overline{v}_j$ and $\overline{(\pi v)}_j$ denote the average of v and its L^2 -projection on I_j , respectively. Now, we define the new (nonlinear) projection $\tilde{\pi}$ as follows:³

$$\tilde{\pi} v(x) := \theta_j[v] \pi v(x) + (1 - \theta_j[v]) \overline{(\pi v)}_j, \quad \forall x \in I_j,$$

³This is the same limiter used in [12], where this limiter is simply applied after each time-stepping of the fully-discrete classical RKDG scheme. In their paper, the error estimate is done heuristically, and the accuracy of their scheme is observed to be sub-optimal in certain numerical tests. In our paper, however, we incorporate this limiter into the DG framework in a very different way so that optimal error estimates can be proved rigorously and also verified in numerical tests (see Section 6).

where

$$\theta_j[v] := \min \left\{ \frac{\int_{I_j} U(v(x)) \, dx - U(\bar{v}_j) h_j}{\int_{I_j} U(\pi v(x)) \, dx - U(\bar{v}_j) h_j}, 1 \right\}. \quad (3.4)$$

If the denominator in (3.4) is zero, we simply set $\theta_j[v] := 1$. Firstly, from the above definition we can see that $\tilde{\pi}v = v$ for all $v \in \mathbb{V}^k$, since in which case $\pi v = v$ so that $\theta_j[v] = 1$. Therefore, our new operator $\tilde{\pi}$ is indeed a projection. Note that $\tilde{\pi}$ still decreases the L^2 -norm on each cell, since it preserves the average and decreases the high frequency modes. The next lemma shows that $\tilde{\pi}$ also decreases the integral of $U(\cdot)$:

Lemma 3.1. *For all $v \in L^2(I)$, we have $0 \leq \theta_j[v] \leq 1$ and*

$$\int_{I_j} U(\tilde{\pi}v(x)) \, dx = \int_{I_j} U\left(\theta_j[v]\pi v(x) + (1 - \theta_j[v])\overline{(\pi v)}_j\right) \, dx \leq \int_{I_j} U(v(x)) \, dx. \quad (3.5)$$

Proof. When the denominator in (3.4) is zero, $\pi v(x) \equiv \bar{v}_j$ for all $x \in I_j$, so that (3.5) holds trivially; so we may assume otherwise. In order to apply Jensen's inequality to the left hand side of (3.5), we need $0 \leq \theta_j[v] \leq 1$. $\theta_j[v] \leq 1$ is already clear from the definition (3.4), so we only need to show

$$\frac{\int_{I_j} U(v(x)) \, dx - U(\bar{v}_j) h_j}{\int_{I_j} U(\pi v(x)) \, dx - U(\bar{v}_j) h_j} \geq 0.$$

This follows easily from applying Jensen's inequality to both the numerator and denominator and noting that $\bar{v}_j = (\pi v)_j$:

$$\begin{aligned} U(\bar{v}_j) &\leq \frac{1}{h_j} \int_{I_j} U(v(x)) \, dx, \\ U(\bar{v}_j) &= U\left(\overline{(\pi v)}_j\right) \leq \frac{1}{h_j} \int_{I_j} U(\pi v(x)) \, dx. \end{aligned}$$

Thus, $0 \leq \theta_j[v] \leq 1$ is proved. Now, by Jensen's inequality again,

$$\begin{aligned} \int_{I_j} U\left(\theta_j[v]\pi v(x) + (1 - \theta_j[v])\overline{(\pi v)}_j\right) \, dx &\leq \theta_j[v] \int_{I_j} U(\pi v(x)) \, dx + (1 - \theta_j[v])U\left(\overline{(\pi v)}_j\right) h_j \\ &= \theta_j[v] \int_{I_j} U(\pi v(x)) \, dx + (1 - \theta_j[v])U(\bar{v}_j) h_j. \end{aligned}$$

The right hand side is less than or equal to $\int_{I_j} U(v(x)) \, dx$ if and only if

$$\theta_j \leq \frac{\int_{I_j} U(v(x)) \, dx - U(\bar{v}_j) h_j}{\int_{I_j} U(\pi v(x)) \, dx - U(\bar{v}_j) h_j},$$

which proves the claim. ■

Applying this special projection $\tilde{\pi}$ to (3.3), we then have

$$\begin{aligned}
& \int_{I_j} \frac{\tilde{\pi} S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \\
&= \int_{I_j} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx - \frac{1 - \theta_j[S_{\Delta t} u(x)]}{\Delta t} \int_{I_j} (\pi S_{\Delta t} u(x) - \overline{(\pi S_{\Delta t} u)_j}) \cdot v(x) \, dx \\
&= \int_{I_j} \frac{S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx - \frac{1 - \theta_j[S_{\Delta t} u(x)]}{\Delta t} \int_{I_j} (S_{\Delta t} u(x) - \overline{(S_{\Delta t} u)_j}) \cdot v(x) \, dx. \quad (3.6)
\end{aligned}$$

Taking the limit of (3.6) as $\Delta t \rightarrow 0$ in the same way as in (2.7), we obtain the following semi-discrete scheme: seek $u \in \mathbb{V}^k$ such that for all $v \in \mathbb{V}^k$ we have⁴

$$\begin{aligned}
\int_{I_j} u_t v \, dx &= \int_{I_j} \lim_{\Delta t \rightarrow 0} \frac{\tilde{\pi} S_{\Delta t} u(x) - u(x)}{\Delta t} \cdot v(x) \, dx \\
&= - \int_{I_j} f(u)_x v \, dx + H_{j-\frac{1}{2}}^+ \cdot v_{j-\frac{1}{2}}^+ + H_{j+\frac{1}{2}}^- \cdot v_{j+\frac{1}{2}}^- - A_j[u; U] \int_{I_j} (u - \bar{u}_j) v \, dx,
\end{aligned} \quad (\text{Scheme A})$$

where $\bar{u}_j$ denotes the average of u on I_j ,

$$A_j[u; U] = \max \left\{ \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{\int_{I_j} U(\pi S_{\Delta t} u) \, dx - \int_{I_j} U(S_{\Delta t} u) \, dx}{\int_{I_j} U(\pi S_{\Delta t} u) \, dx - h_j U\left(\frac{1}{h_j} \int_{I_j} S_{\Delta t} u \, dx\right)}, 0 \right\}, \quad (3.7)$$

and

$$H_{j-\frac{1}{2}}^+ = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{I_j^+(\Delta t)} S_{\Delta t} u(x) - u(x) \, dx = \hat{f}_{j-\frac{1}{2}} - f(u)_{j-\frac{1}{2}}^+, \quad (3.8)$$

$$H_{j+\frac{1}{2}}^- = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{I_j^-(\Delta t)} S_{\Delta t} u(x) - u(x) \, dx = -\hat{f}_{j+\frac{1}{2}} + f(u)_{j+\frac{1}{2}}^-. \quad (3.9)$$

The expressions (3.8) and (3.9) are already proved in (2.12) and (2.13).

Although the explicit expression of $A_j[u; U]$ defined in (3.7) can generally be obtained, the resulting formula is too complicated to be implemented in practice. To make (Scheme A) computationally efficient, the following proposition is useful:

Proposition 3.2 (Monotonicity of Remainder). *If there exists $B_j[u; U]$ such that $A_j[u; U] \leq B_j[u; U]$ for all $u \in \mathbb{V}^k$ and $j = 1, \dots, N$, then replacing $A_j[u; U]$ by $B_j[u; U]$ in (Scheme A) can only decrease its entropy production rate $\frac{d}{dt} \int_{I_j} U(u)(x, t) \, dx$ on each cell I_j . We call this property “monotonicity of remainder.”*

⁴Here we follow the convention in the DG literature: when the context is clear, we suppress the variable t in the notation $u(\cdot, t) \in \mathbb{V}^k$, while keeping in mind that u is an element in $\mathbb{V}^k$ parameterized by $t \geq 0$.

Proof. Let $u \in \mathbb{V}^k$ and $z := \pi[U'(u)] \in \mathbb{V}^k$. Taking $v = z$ in (Scheme A), we obtain the entropy production rate for A_j as

$$\begin{aligned} \left(\frac{d}{dt} \int_{I_j} U(u) dx \right)_A &= \int_{I_j} u_t U'(u) dx = \int_{I_j} u_t \pi[U'(u)] dx = \int_{I_j} u_t z dx \\ &= - \int_{I_j} f(u)_x \cdot z dx + H_{j-\frac{1}{2}}^+ \cdot z_{j-\frac{1}{2}}^+ + H_{j+\frac{1}{2}}^- \cdot z_{j+\frac{1}{2}}^- - A_j[u; U] \int_{I_j} (u - \bar{u}_j) \cdot z dx. \end{aligned}$$

Similarly, the entropy production rate for B_j is

$$\left(\frac{d}{dt} \int_{I_j} U(u) dx \right)_B = - \int_{I_j} f(u)_x \cdot z dx + H_{j-\frac{1}{2}}^+ \cdot z_{j-\frac{1}{2}}^+ + H_{j+\frac{1}{2}}^- \cdot z_{j+\frac{1}{2}}^- - B_j[u; U] \int_{I_j} (u - \bar{u}_j) \cdot z dx.$$

Therefore,

$$\begin{aligned} &\left(\frac{d}{dt} \int_{I_j} U(u) dx \right)_B - \left(\frac{d}{dt} \int_{I_j} U(u) dx \right)_A = (A_j[u; U] - B_j[u; U]) \int_{I_j} (u - \bar{u}_j) \cdot \pi[U'(u)] dx \\ &= (A_j[u; U] - B_j[u; U]) \int_{I_j} (u - \bar{u}_j) \cdot U'(u) dx \quad (\text{since } u - \bar{u}_j \in \mathbb{P}^k(I_j)) \\ &= (A_j[u; U] - B_j[u; U]) \int_{I_j} (u - \bar{u}_j) \cdot (U'(u) - U'(\bar{u}_j)) dx \leq 0, \end{aligned}$$

where in the last equality, we used the fact that $u - \bar{u}_j$ is orthogonal to the constant $U'(\bar{u}_j)$. The last inequality follows from $A_j[u; U] \leq B_j[u; U]$ and $(u(x) - \bar{u}_j) \cdot (U'(u(x)) - U'(\bar{u}_j)) \geq 0$ since $U(\cdot)$ is convex. Thus, the proposition is proved. $\blacksquare$

It follows from Proposition 3.2 that we can instead derive an upper bound $B_j[u; U]$ for $A_j[u; U]$, and the scheme with B_j satisfies the entropy condition if A_j does. Moreover, as we will see shortly, if B_j is a bound uniform in an infinite class of entropies, the modified scheme will satisfy infinitely many entropy inequalities instead of just two (the prescribed $U(\cdot)$ and the original square-entropy). In the rest of this section, we will find a bound on $A_j[u; U]$ as tight as possible, so that the resulting scheme is still of optimal order of accuracy. To this end, we let $P_{i,j}(x)$ denote the i th Legendre polynomial normalized on I_j satisfying $\|P_{i,j}(x)\|_{L^2(I_j)} = 1$. Then,

$$\begin{aligned} &\pi S_{\Delta t} u(x) - u(x) \\ &= \sum_{i=0}^k P_{i,j}(x) \int_{I_j} P_{i,j}(x) S_{\Delta t} u(x) dx - u(x) \\ &= \sum_{i=0}^k P_{i,j}(x) \int_{I_j} P_{i,j}(x) (S_{\Delta t} u(x) - u(x)) dx \\ &= \Delta t \sum_{i=0}^k P_{i,j}(x) \left\{ - \int_{I_j} f(u)_x P_{i,j}(x) dx + P_{i,j}(x_{j+\frac{1}{2}}^-) H_{j+\frac{1}{2}}^- + P_{i,j}(x_{j-\frac{1}{2}}^+) H_{j-\frac{1}{2}}^+ \right\} + O(\Delta t^2) \end{aligned}$$

$$= -\Delta t \pi \circ [f(u)_x](x) + \Delta t \sum_{i=0}^k P_{i,j}(x) P_{i,j}(x_{j+\frac{1}{2}}^-) H_{j+\frac{1}{2}}^- + \Delta t \sum_{i=0}^k P_{i,j}(x) P_{i,j}(x_{j-\frac{1}{2}}^+) H_{j-\frac{1}{2}}^+ + O(\Delta t^2),$$

Therefore,

$$\begin{aligned} & \int_{I_j} U(\pi S_{\Delta t} u(x)) dx - \int_{I_j} U(u(x)) dx \\ &= \int_{I_j} U'(u(x)) \cdot (\pi S_{\Delta t} u(x) - u(x)) dx + O(\Delta t^2) \\ &= \int_{I_j} U'(u(x)) \cdot \left\{ -\Delta t \pi \circ [f(u)_x](x) + \Delta t \sum_{i=0}^k P_{i,j}(x) P_{i,j}(x_{j+\frac{1}{2}}^-) H_{j+\frac{1}{2}}^- \right. \\ & \quad \left. + \Delta t \sum_{i=0}^k P_{i,j}(x) P_{i,j}(x_{j-\frac{1}{2}}^+) H_{j-\frac{1}{2}}^+ \right\} dx + O(\Delta t^2) \\ &= -\Delta t \int_{I_j} U'(u(x)) \cdot \pi \circ [f(u)_x](x) dx + \Delta t \pi \circ [U'(u)](x_{j+\frac{1}{2}}^-) \cdot H_{j+\frac{1}{2}}^- \\ & \quad + \Delta t \pi \circ [U'(u)](x_{j-\frac{1}{2}}^+) \cdot H_{j-\frac{1}{2}}^+ + O(\Delta t^2). \end{aligned} \quad (3.10)$$

On the other hand,

$$\begin{aligned} & \int_{I_j} U(S_{\Delta t} u(x)) dx - \int_{I_j} U(u(x)) dx \\ &= \int_{I_j \setminus (I_j^+(\Delta t) \cup I_j^-(\Delta t))} U(S_{\Delta t} u(x)) - U(u(x)) dx + \int_{I_j^+(\Delta t)} U(S_{\Delta t} u(x)) - U(u(x)) dx \\ & \quad + \int_{I_j^-(\Delta t)} U(S_{\Delta t} u(x)) - U(u(x)) dx \\ &= -\Delta t \int_{I_j} U'(u(x)) \cdot f(u)_x dx + \Delta t G_{j-\frac{1}{2}}^+ + \Delta t G_{j+\frac{1}{2}}^- + O(\Delta t^2), \end{aligned} \quad (3.11)$$

where

$$G_{j-\frac{1}{2}}^+ := \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{I_j^+(\Delta t)} U(S_{\Delta t} u(x)) - U(u(x)) dx, \quad (3.12)$$

$$G_{j+\frac{1}{2}}^- := \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{I_j^-(\Delta t)} U(S_{\Delta t} u(x)) - U(u(x)) dx. \quad (3.13)$$

The exact expressions of $G_{j-\frac{1}{2}}^+$ and $G_{j+\frac{1}{2}}^-$ are not important, since they will be eliminated in the subsequent computations.

It follows from (3.10) and (3.11) that

$$\lim_{\Delta t \rightarrow 0} \frac{\int_{I_j} U(\pi S_{\Delta t} u(x)) dx - \int_{I_j} U(S_{\Delta t} u(x)) dx}{\Delta t}$$

$$\begin{aligned}
 &= \int_{I_j} U'(u) \cdot (f(u)_x - \pi \circ [f(u)_x]) \, dx + \pi \circ [U'(u)](x_{j+\frac{1}{2}}^-) \cdot H_{j+\frac{1}{2}}^- \\
 &\quad + \pi \circ [U'(u)](x_{j-\frac{1}{2}}^+) \cdot H_{j-\frac{1}{2}}^+ - G_{j+\frac{1}{2}}^- - G_{j-\frac{1}{2}}^+
 \end{aligned}$$

$$= \int_{I_j} U'(u) \cdot (f(u)_x - \pi \circ [f(u)_x]) \, dx \tag{3.14}$$

$$+ \pi \circ [U'(u) - U'(u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+) \cdot H_{j-\frac{1}{2}}^+ \tag{3.15}$$

$$+ \pi \circ [U'(u) - U'(u_{j+\frac{1}{2}}^-)](x_{j+\frac{1}{2}}^-) \cdot H_{j+\frac{1}{2}}^- \tag{3.16}$$

$$+ (U'(u_{j-\frac{1}{2}}^+) \cdot H_{j-\frac{1}{2}}^+ - G_{j-\frac{1}{2}}^+) \tag{3.17}$$

$$+ (U'(u_{j+\frac{1}{2}}^-) \cdot H_{j+\frac{1}{2}}^- - G_{j+\frac{1}{2}}^-). \tag{3.18}$$

This is the explicit expression for the numerator in (3.7). We first prove that both (3.17) and (3.18) in the last expression are non-positive.

Lemma 3.3.

$$U'(u_{j-\frac{1}{2}}^+) \cdot H_{j-\frac{1}{2}}^+ - G_{j-\frac{1}{2}}^+ \leq 0, \quad U'(u_{j+\frac{1}{2}}^-) \cdot H_{j+\frac{1}{2}}^- - G_{j+\frac{1}{2}}^- \leq 0.$$

Proof. It follows from (3.8) and (3.9) that

$$U'(u_{j-\frac{1}{2}}^+) \cdot H_{j-\frac{1}{2}}^+ = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{I_j^+(\Delta t)} U'(u(x)) \cdot (S_{\Delta t} u(x) - u(x)) \, dx,$$

$$U'(u_{j+\frac{1}{2}}^-) \cdot H_{j+\frac{1}{2}}^- = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{I_j^-(\Delta t)} U'(u(x)) \cdot (S_{\Delta t} u(x) - u(x)) \, dx.$$

From the definitions of $G_{j-\frac{1}{2}}^+$ and $G_{j+\frac{1}{2}}^-$ in (3.12) and (3.13), we see that it suffices to prove

$$\begin{aligned}
 &\int_{I_j^+(\Delta t)} U'(u(x)) \cdot (S_{\Delta t} u(x) - u(x)) \, dx - \int_{I_j^+(\Delta t)} U(S_{\Delta t} u(x)) - U(u(x)) \, dx \leq 0, \\
 &\int_{I_j^-(\Delta t)} U'(u(x)) \cdot (S_{\Delta t} u(x) - u(x)) \, dx - \int_{I_j^-(\Delta t)} U(S_{\Delta t} u(x)) - U(u(x)) \, dx \leq 0,
 \end{aligned}$$

since taking the limit does not change sign. By the convexity of $U(\cdot)$, the above two inequalities are trivial since

$$U'(u(x)) \cdot (S_{\Delta t} u(x) - u(x)) - (U(S_{\Delta t} u(x)) - U(u(x))) \leq 0, \quad \forall x \in I.$$

■

Next, we estimate (3.15) and (3.16).

Lemma 3.4. *For any $u \in \mathbb{V}^k$, we have*

$$|\pi \circ [U'(u) - U'(u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+)| \leq \frac{\beta_k}{2h_j} \max_{x \in I_j} |U'''(u(x))| \int_{I_j} |u(x) - \bar{u}_j|^2 \, dx,$$

$$|\pi \circ [U'(u) - U'(u_{j+\frac{1}{2}}^-)](x_{j+\frac{1}{2}}^-)| \leq \frac{\beta_k}{2h_j} \max_{x \in I_j} |U'''(u(x))| \int_{I_j} |u(x) - \bar{u}_j|^2 dx,$$

where $\beta_k \leq (k+1)^4$.

Proof. Since $U''(u_{j-\frac{1}{2}}^+)(u(x) - u_{j-\frac{1}{2}}^+)$ is a k -th order polynomial on I_j and vanishes at $x = x_{j-\frac{1}{2}}^+$,

$$\begin{aligned} & |\pi \circ [U'(u) - U'(u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+)| \\ &= |\pi \circ [U'(u) - U'(u_{j-\frac{1}{2}}^+) - U''(u_{j-\frac{1}{2}}^+)(u - u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+)| \\ &= \left| \sum_{i=0}^k P_{i,j}(x_{j-\frac{1}{2}}^+) \int_{I_j} P_{i,j}(x) [U'(u(x)) - U'(u_{j-\frac{1}{2}}^+) - U''(u_{j-\frac{1}{2}}^+)(u(x) - u_{j-\frac{1}{2}}^+)] dx \right| \\ &\leq \sum_{i=0}^k \|P_{i,j}(\cdot)\|_{L^\infty(I_j)}^2 \int_{I_j} |U'(u(x)) - U'(u_{j-\frac{1}{2}}^+) - U''(u_{j-\frac{1}{2}}^+)(u(x) - u_{j-\frac{1}{2}}^+)| dx \\ &\leq \sum_{i=0}^k \frac{2i+1}{h_j} \frac{\max_{x \in I_j} |U'''(u(x))|}{2} \int_{I_j} |u(x) - u_{j-\frac{1}{2}}^+|^2 dx \\ &= \frac{(k+1)^2}{2h_j} \max_{x \in I_j} |U'''(u(x))| \int_{I_j} |u(x) - u_{j-\frac{1}{2}}^+|^2 dx \\ &\leq \frac{(k+1)^4}{2h_j} \max_{x \in I_j} |U'''(u(x))| \int_{I_j} |u(x) - \bar{u}_j|^2 dx, \end{aligned}$$

where in the last step, the following norm equivalence on the space of all k -th order polynomials with zero average on I_j has been used:

$$\|p - p_{j-\frac{1}{2}}^+\|_{L^2(I_j)} \leq (k+1) \|p\|_{L^2(I_j)}, \quad \forall p \in \mathbb{P}^k(I_j) \text{ with } \bar{p}_j = 0.$$

The estimate for $|\pi \circ [U'(u) - U'(u_{j+\frac{1}{2}}^-)](x_{j+\frac{1}{2}}^-)|$ is proved in the same way. $\blacksquare$

Remark 3.5. A substantially improved estimate on β_k for small k can be found in Appendix B. In Proposition B.1 we show that $\beta_k \leq 2\mathcal{A}_k$, where the first few values of $\mathcal{A}_k$ are tabulated in Table B.1 therein.

Finally, we estimate (3.14).

Lemma 3.6. For any $\mathbf{u} \in \mathbb{V}^k$, we have

$$\begin{aligned} & \left| \int_{I_j} U'(u) \cdot (f(u)_x - \pi \circ [f(u)_x]) dx \right| \\ & \leq \frac{C_k}{2\sqrt{h_j}} \max_{x \in I_j} |U'''(u(x))| \|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} \int_{I_j} |u(x) - \bar{u}_j|^2 dx \end{aligned}$$

where $C_k > 0$ is the optimal constant such that

$$\|p\|_{L^4(I_j)} \leq \frac{\sqrt{C_k}}{h_j^{1/4}} \|p\|_{L^2(I_j)}, \quad \forall p \in \mathbb{P}^k(I_j). \quad (3.19)$$

Proof. Note that $U'(\bar{u}_j) + U''(\bar{u}_j)(u(x) - \bar{u}_j)$ is a k -th order polynomial on I_j , which is orthogonal to $f(u)_x - \pi \circ [f(u)_x]$. Therefore,

$$\begin{aligned} & \left| \int_{I_j} U'(u) \cdot (f(u)_x - \pi \circ [f(u)_x]) \, dx \right| \\ &= \left| \int_{I_j} (U'(u) - U'(\bar{u}_j) - U''(\bar{u}_j)(u - \bar{u}_j)) \cdot (f(u)_x - \pi \circ [f(u)_x]) \, dx \right| \\ &\leq \frac{\max_{x \in I_j} |U'''(u(x))|}{2} \int_{I_j} |u - \bar{u}_j|^2 |f(u)_x - \pi \circ [f(u)_x]| \, dx \\ &\leq \frac{\max_{x \in I_j} |U'''(u(x))|}{2} \|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} \sqrt{\int_{I_j} |u - \bar{u}_j|^4 \, dx} \\ (3.19) \quad &\leq \frac{C_k}{2\sqrt{h_j}} \max_{x \in I_j} |U'''(u(x))| \|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} \int_{I_j} |u(x) - \bar{u}_j|^2 \, dx. \end{aligned}$$

■

Remark 3.7. The first few values of C_k in (3.19) are

k	0	1	2	3	4	5	6	7
C_k	1	1.5	2.156	2.834	3.520	4.209	4.901	5.594

Table 1. Optimal constants C_k .

For a general $k \geq 0$ we have the convenient but non-sharp bound

$$C_k \leq k + 1,$$

since

$$\|p\|_{L^4(I_j)}^2 \leq \|p\|_{L^\infty(I_j)} \|p\|_{L^2(I_j)} \quad \text{and} \quad \|p\|_{L^\infty(I_j)} \leq \frac{k+1}{\sqrt{h_j}} \|p\|_{L^2(I_j)}.$$

Combining Lemma 3.3, 3.4 and 3.6, we obtain

$$\begin{aligned} & \lim_{\Delta t \rightarrow 0} \frac{\int_{I_j} U(\pi S_{\Delta t} u(x)) \, dx - \int_{I_j} U(S_{\Delta t} u(x)) \, dx}{\Delta t} \\ &\leq \frac{C_k}{2\sqrt{h_j}} \max_{x \in I_j} |U'''(u(x))| \|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} \int_{I_j} |u(x) - \bar{u}_j|^2 \, dx \end{aligned}$$

$$+ \frac{\beta_k}{2h_j} \max_{x \in I_j} |U'''(u(x))| \left(|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-| \right) \int_{I_j} |u(x) - \bar{u}_j|^2 dx. \quad (3.20)$$

By Hölder's Defect Formula [17] (also known as *the Jensen Gap*), we also have

$$\begin{aligned} & \lim_{\Delta t \rightarrow 0} \int_{I_j} U(\pi S_{\Delta t} u) dx - h_j U \left(\frac{1}{h_j} \int_{I_j} S_{\Delta t} u dx \right) \\ &= \int_{I_j} U(u) dx - h_j U(\bar{u}_j) \geq \frac{1}{2} \min_{x \in I_j} |U'''(u(x))| \int_{I_j} |u(x) - \bar{u}_j|^2, \end{aligned} \quad (3.21)$$

It follows from (3.20) and (3.21) that

$$\begin{aligned} A_j[u; U] &= \max \left\{ \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{\int_{I_j} U(\pi S_{\Delta t} u) dx - \int_{I_j} U(S_{\Delta t} u) dx}{\int_{I_j} U(\pi S_{\Delta t} u) dx - h_j U \left(\frac{1}{h_j} \int_{I_j} S_{\Delta t} u dx \right)}, 0 \right\} \\ &\leq \frac{\max_{x \in I_j} |U'''(u(x))|}{\min_{x \in I_j} |U'''(u(x))|} \left\{ \frac{C_k}{\sqrt{h_j}} \|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} + \frac{\beta_k}{h_j} \left(|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-| \right) \right\}. \end{aligned} \quad (3.22)$$

In the above, we implicitly used the following inequality:

$$\max\{a, 0\} \leq \max\{b, 0\}, \quad \forall a < b.$$

As the next theorem shows, we only need half of the bound in (3.22) to satisfy the entropy condition for $U(\cdot)$. Therefore, we define

$$B_j[u; U] := \frac{\max_{x \in I_j} |U'''(u(x))|}{\min_{x \in I_j} |U'''(u(x))|} \left\{ \frac{C_k}{2\sqrt{h_j}} \|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} + \frac{\beta_k}{2h_j} \left(|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-| \right) \right\}. \quad (3.23)$$

Now, replacing $A_j[u; U]$ in (Scheme A) with the $B_j[u; U]$ defined above yields the following new scheme: seek $u \in \mathbb{V}^k$ such that for all $v \in \mathbb{V}^k$ we have

$$\int_{I_j} u_t v dx = - \int_{I_j} f(u)_x v dx + H_{j-\frac{1}{2}}^+ \cdot v_{j-\frac{1}{2}}^+ + H_{j+\frac{1}{2}}^- \cdot v_{j+\frac{1}{2}}^- - B_j[u; U] \int_{I_j} (u - \bar{u}_j) v dx, \quad (\text{Scheme B})$$

where $H_{j-\frac{1}{2}}^+$ and $H_{j+\frac{1}{2}}^-$ are defined as

$$H_{j-\frac{1}{2}}^+ := \hat{f}_{j-\frac{1}{2}} - f(u)_{j-\frac{1}{2}}^+, \quad H_{j+\frac{1}{2}}^- := -\hat{f}_{j+\frac{1}{2}} + f(u)_{j+\frac{1}{2}}^-, \quad (3.24)$$

according to (3.8) and (3.9). After performing integration by parts on $\int_{I_j} f(u)_x \cdot v dx$, we get the equivalent weak formulation of (Scheme B) as

$$\int_{I_j} u_t v dx - \int_{I_j} f(u) v_x dx + \hat{f}_{j+\frac{1}{2}} \cdot v_{j+\frac{1}{2}}^- - \hat{f}_{j-\frac{1}{2}} \cdot v_{j-\frac{1}{2}}^+ = -B_j[u; U] \int_{I_j} (u - \bar{u}_j) v dx, \quad (3.25)$$

which is really (**Classical DG**) plus a remainder term on the right-hand side. Note that the numerical flux $\hat{f}$ computed using the expressions (3.8)–(3.9) is the Godunov flux. This follows from the fact that $S_{\Delta t}$ is the exact solution operator. Now, however, (**Scheme B**) has nothing to do with $S_{\Delta t}$ if we just define $H_{j-\frac{1}{2}}^+$ and $H_{j+\frac{1}{2}}^-$ by (3.24) instead of using (3.8)–(3.9); thus, (**Scheme B**) is still well-defined if any other numerical flux is used in the definitions (3.24).

Definition 3.8. $\hat{f}$ is an E-flux [15] if it is consistent ($\hat{f}(u, u) = f(u)$), locally Lipschitz continuous, and satisfies

$$(\hat{f}(u^-, u^+) - f(u))(u^+ - u^-) \leq 0$$

for all u between u^- and u^+ .

All monotone fluxes, such as the Godunov flux, local Lax-Friedrichs flux, or the more general class of HLL fluxes with suitable lower and upper wave-speed bounds (see (A.1)), are E-fluxes. The next theorem shows that (**Scheme B**) satisfies local entropy inequality for $U(\cdot)$, provided that $\hat{f}$ is any E-flux.

Theorem 3.9 (Local Entropy Inequality for (**Scheme B**)). *Suppose that $\hat{f}$ in (3.24) is a consistent E-flux. Then (**Scheme B**) satisfies the local semi-discrete entropy inequality*

$$\frac{d}{dt} \int_{I_j} U(u) dx + \hat{\mathcal{F}}_{j+\frac{1}{2}} - \hat{\mathcal{F}}_{j-\frac{1}{2}} \leq 0, \quad \forall u \in \mathbb{V}^k, \quad \forall j = 1, \dots, N, \quad (3.26)$$

where $\hat{\mathcal{F}}_{j+\frac{1}{2}} = \hat{\mathcal{F}}(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+)$ and

$$\hat{\mathcal{F}}(u^-, u^+) := U'(u^-) \hat{f}(u^-, u^+) - \psi(u^-), \quad (3.27)$$

with ψ defined by (3.2).

Proof. Set

$$z := \pi[U'(u)] \in \mathbb{V}^k.$$

Since $u_t \in \mathbb{V}^k$ and π is the L^2 -projection onto $\mathbb{V}^k$, we have

$$\int_{I_j} u_t z dx = \int_{I_j} u_t U'(u) dx = \frac{d}{dt} \int_{I_j} U(u) dx. \quad (3.28)$$

Taking $v = z$ in (**Scheme B**) therefore gives

$$\begin{aligned} \frac{d}{dt} \int_{I_j} U(u) dx &= - \int_{I_j} f(u)_x z dx + H_{j-\frac{1}{2}}^+ z_{j-\frac{1}{2}}^+ + H_{j+\frac{1}{2}}^- z_{j+\frac{1}{2}}^- \\ &\quad - B_j[u; U] \int_{I_j} (u - \bar{u}_j) z dx. \end{aligned} \quad (3.29)$$

We first rewrite the volume term. By the self-adjointness of the L^2 -projection,

$$\int_{I_j} f(u)_x z \, dx = \int_{I_j} \pi[f(u)_x] U'(u) \, dx.$$

Hence, using $\mathcal{F}'(u) = U'(u)f'(u)$,

$$\begin{aligned} - \int_{I_j} f(u)_x z \, dx &= - \int_{I_j} U'(u) f(u)_x \, dx + \int_{I_j} U'(u) (f(u)_x - \pi[f(u)_x]) \, dx \\ &= \mathcal{F}(u_{j-\frac{1}{2}}^+) - \mathcal{F}(u_{j+\frac{1}{2}}^-) + R_{j,1}, \end{aligned} \quad (3.30)$$

where

$$R_{j,1} := \int_{I_j} U'(u) (f(u)_x - \pi[f(u)_x]) \, dx. \quad (3.31)$$

Also, since $u - \bar{u}_j \in \mathbb{V}^k$ and has zero average,

$$\begin{aligned} \int_{I_j} (u - \bar{u}_j) z \, dx &= \int_{I_j} (u - \bar{u}_j) U'(u) \, dx \\ &= \int_{I_j} (u - \bar{u}_j) (U'(u) - U'(\bar{u}_j)) \, dx =: D_j. \end{aligned} \quad (3.32)$$

By the mean value theorem and strict convexity of U ,

$$D_j \geq m_j \int_{I_j} |u - \bar{u}_j|^2 \, dx \geq 0, \quad (3.33)$$

where

$$m_j := \min_{x \in I_j} |U''(u(x))|.$$

We next treat the interface terms. At $x_{j+\frac{1}{2}}$, it follows from (3.2) and (3.27) that

$$\begin{aligned} \hat{\mathcal{F}}_{j+\frac{1}{2}} &= U'(u_{j+\frac{1}{2}}^-) \hat{f}_{j+\frac{1}{2}} - \psi(u_{j+\frac{1}{2}}^-) \\ &= \mathcal{F}(u_{j+\frac{1}{2}}^-) + U'(u_{j+\frac{1}{2}}^-) (\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^-)) \\ &= \mathcal{F}(u_{j+\frac{1}{2}}^-) - U'(u_{j+\frac{1}{2}}^-) H_{j+\frac{1}{2}}^-. \end{aligned}$$

Consequently,

$$\begin{aligned} &- \mathcal{F}(u_{j+\frac{1}{2}}^-) + H_{j+\frac{1}{2}}^- z_{j+\frac{1}{2}}^- + \hat{\mathcal{F}}_{j+\frac{1}{2}} \\ &= H_{j+\frac{1}{2}}^- \pi[U'(u) - U'(u_{j+\frac{1}{2}}^-)](x_{j+\frac{1}{2}}^-) =: R_{j,3}. \end{aligned} \quad (3.34)$$

At $x_{j-\frac{1}{2}}$, put

$$u^- := u_{j-\frac{1}{2}}^-, \quad u^+ := u_{j-\frac{1}{2}}^+, \quad h := \hat{f}(u^-, u^+).$$

For convenience define the same numerical entropy-flux expression evaluated with the right state by

$$\hat{\mathcal{F}}^R(u^-, u^+) := U'(u^+)h - \psi(u^+).$$

Since $H_{j-\frac{1}{2}}^+ = h - f(u^+)$,

$$\begin{aligned} & \mathcal{F}(u_{j-\frac{1}{2}}^+) + H_{j-\frac{1}{2}}^+ z_{j-\frac{1}{2}}^+ - \hat{\mathcal{F}}^R(u^-, u^+) \\ &= H_{j-\frac{1}{2}}^+ \pi[U'(u) - U'(u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+) =: R_{j,2}. \end{aligned} \quad (3.35)$$

On the other hand,

$$\begin{aligned} \hat{\mathcal{F}}(u^-, u^+) - \hat{\mathcal{F}}^R(u^-, u^+) &= h(U'(u^-) - U'(u^+)) - (\psi(u^-) - \psi(u^+)) \\ &= \int_{u^-}^{u^+} U''(s)(f(s) - h) \, ds =: \mathcal{D}(u^-, u^+). \end{aligned} \quad (3.36)$$

Due to the property of E-flux and $U'' \geq 0$, we have

$$\mathcal{D}(u^-, u^+) \geq 0. \quad (3.37)$$

Thus, by (3.35),

$$\mathcal{F}(u_{j-\frac{1}{2}}^+) + H_{j-\frac{1}{2}}^+ z_{j-\frac{1}{2}}^+ - \hat{\mathcal{F}}_{j-\frac{1}{2}} = R_{j,2} - \mathcal{D}_{j-\frac{1}{2}}, \quad (3.38)$$

where

$$\mathcal{D}_{j-\frac{1}{2}} := \mathcal{D}(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+) \geq 0.$$

Adding $\hat{\mathcal{F}}_{j+\frac{1}{2}} - \hat{\mathcal{F}}_{j-\frac{1}{2}}$ to (3.29) and using (3.30), (3.34), (3.38), and (3.32), we obtain the exact identity

$$\begin{aligned} & \frac{d}{dt} \int_{I_j} U(u) \, dx + \hat{\mathcal{F}}_{j+\frac{1}{2}} - \hat{\mathcal{F}}_{j-\frac{1}{2}} \\ &= R_{j,1} + R_{j,2} + R_{j,3} - \mathcal{D}_{j-\frac{1}{2}} - B_j[u; U]D_j. \end{aligned} \quad (3.39)$$

It remains to estimate the three projection errors. Lemma 3.6 gives

$$|R_{j,1}| \leq \frac{C_k}{2\sqrt{h_j}} \max_{x \in I_j} |U'''(u(x))| \|f(u)_x - \pi[f(u)_x]\|_{L^2(I_j)} \int_{I_j} |u - \bar{u}_j|^2 \, dx, \quad (3.40)$$

and Lemma 3.4 gives

$$|R_{j,2}| + |R_{j,3}| \leq \frac{\beta_k}{2h_j} \max_{x \in I_j} |U'''(u(x))| (|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-|) \int_{I_j} |u - \bar{u}_j|^2 \, dx. \quad (3.41)$$

By the definition of $B_j[u; U]$, (3.40)–(3.41) imply

$$R_{j,1} + R_{j,2} + R_{j,3} \leq |R_{j,1}| + |R_{j,2}| + |R_{j,3}|$$

$$\begin{aligned}
&\leq B_j[u; U]m_j \int_{I_j} |u - \bar{u}_j|^2 \, dx \\
&\leq B_j[u; U]D_j,
\end{aligned} \tag{3.42}$$

where the last inequality follows from (3.33). Substituting (3.42) into (3.39) yields

$$\frac{d}{dt} \int_{I_j} U(u) \, dx + \hat{\mathcal{F}}_{j+\frac{1}{2}} - \hat{\mathcal{F}}_{j-\frac{1}{2}} \leq -\mathcal{D}_{j-\frac{1}{2}} \leq 0,$$

which proves (3.26). $\blacksquare$

Remark 3.10. The above proof is very much parallel to the proof for the square entropy $U(u) = \frac{u^2}{2}$ given in [9], the only difference being that the test function v is simply taken as $u = U'(u)$ therein. When $U(\cdot)$ is not the square entropy, however, we cannot take $v = U'(u)$ since $U'(u)$ is not a piecewise polynomial. Thus, the only way is by setting $v = \pi[U'(u)]$ and then estimating the projection error, which can be done easily thanks to Lemmas 3.4 and 3.6.

Note that in (Scheme B), U only appears in the quotient

$$q_j[u; U] := \frac{\max_{x \in I_j} |U'''(u(x))|}{\min_{x \in I_j} |U''(u(x))|} \tag{3.43}$$

in the expression of $B_j[u; U]$, and nowhere else! By monotonicity of remainder (Proposition 3.2), (Scheme B) still preserves the entropy inequality with a larger quotient.

Corollary 3.11. (Scheme B) also satisfies the local entropy inequality for the square-entropy $V(u) = \frac{u^2}{2}$, that is, inequality (3.26) with $(U, \mathcal{F})$ replaced by V and the corresponding square-entropy flux. This is nothing other than the well-known result for (Classical DG) proved in [9]. Thus, (Scheme B) with any $U(u) \neq \frac{u^2}{2}$ satisfies at least two local entropy inequalities.

Proof. This follows from $B_j[u; U] \geq B_j[u; V] \equiv 0$ and monotonicity of remainder. $\blacksquare$

Moreover, if U is uniformly convex, i.e. $\min_{w \in \mathbb{R}} U''(w) > 0$, then we can replace (3.43) in $B_j[u; U]$ by the larger quotient

$$\tilde{q}_j[u; U] := \frac{\max_{x \in I_j} |U'''(u(x))|}{\min_{w \in \mathbb{R}} U''(w)} \geq q_j[u; U]. \tag{3.44}$$

Denote the new $B_j[u; U]$ with $\tilde{q}_j[u; U]$ by $\tilde{B}_j[u; U]$. Clearly, $\tilde{B}_j[u; U] \geq B_j[u; U]$, so Theorem 3.9 also holds for $\tilde{B}_j[u; U]$ by monotonicity of remainder.

Corollary 3.12. Suppose $\min_{w \in \mathbb{R}} U''(w) = a > 0$ and we use $\tilde{B}_j[u; U]$ in (Scheme B). Then, for all uniformly convex function $V(\cdot)$ satisfying

$$a \leq V''(w), \quad |V'''(w)| \leq |U'''(w)|, \quad \forall w \in \mathbb{R}, \tag{3.45}$$

(Scheme B) also satisfies the local entropy inequality for V .

Proof. This follows from $\tilde{B}_j[u; U] \geq \tilde{B}_j[u; V]$ and monotonicity of remainder. ■

The next corollary states that for a special class of U , (Scheme B) with $\tilde{B}_j[u; U]$ can satisfy infinitely many local entropy inequalities.

Corollary 3.13. *Suppose we use $\tilde{B}_j[u; U]$ in (Scheme B) and take $U(u) = \sum_{i=1}^{\infty} a_i u^{2i}$ to be analytic on $\mathbb{R}$, such that $a_1 > 0$, $a_i \geq 0$ for all $i > 1$ and not all zero. Then, (Scheme B) also satisfies local entropy inequalities for all V defined by*

$$V(u) = \sum_{i=1}^{\infty} b_i u^{2i}, \quad \text{with } b_1 \geq a_1 \text{ and } 0 \leq b_i \leq a_i \text{ for all } i > 1.$$

Thus, with this U , (Scheme B) satisfies infinitely many entropy inequalities.

Proof. V satisfies (3.45), so the statement follows from Corollary 3.12. ■

The term $\|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)}$ in $B_j[u; U]$ (or $\tilde{B}_j[u; U]$) is easy to compute if $f(\cdot)$ is a polynomial. In particular, it vanishes when $f(\cdot)$ is linear. When $f(\cdot)$ is a general nonlinear function, however, this term might be computationally expensive. Since $f(u(x))$ is smooth on each compact interval I_j , by Bramble–Hilbert lemma and best approximation property of π ,

$$\|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)} \leq C(k, r) h_j^{r-1} \|\partial_x^r f(u)\|_{L^2(I_j)}, \quad k+1 \leq r \leq k+2, \quad (3.46)$$

where $C(k, r)$ is a constant only depending on k and r . The right-hand side is still zero if $f(\cdot)$ is linear, and when $f(\cdot)$ is nonlinear it can be computed with relative ease if the derivatives of $f(\cdot)$ have analytical formulae. In light of Proposition 3.2, we can replace $\|f(u)_x - \pi \circ [f(u)_x]\|_{L^2(I_j)}$ in $B_j[u; U]$ by this upper bound and get the new constant

$$C_j[u; U, r] := q_j[u; U] \left\{ \frac{C_k C(k, r) h_j^{r-\frac{3}{2}}}{2} \|\partial_x^r f(u)\|_{L^2(I_j)} + \frac{\beta_k}{2h_j} \left(|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-| \right) \right\}. \quad (3.47)$$

Replacing $B_j[u; U]$ by $C_j[u; U, r] \geq B_j[u; U]$ in the weak formulation of (Scheme B), we get yet another scheme that satisfies the local entropy inequality of U : seek $u \in \mathbb{V}^k$ such that for all $v \in \mathbb{V}^k$ we have

$$\int_{I_j} u_t v \, dx - \int_{I_j} f(u) v_x \, dx + \hat{f}_{j+\frac{1}{2}} \cdot v_{j+\frac{1}{2}}^- - \hat{f}_{j-\frac{1}{2}} \cdot v_{j-\frac{1}{2}}^+ = -C_j[u; U, r] \int_{I_j} (u - \bar{u}_j) v \, dx. \quad (\text{Scheme C})$$

Clearly, by monotonicity of remainder, Theorem 3.9 and Corollaries 3.11–3.13 also hold for (Scheme C). In particular, (Scheme C) admits infinitely many entropy inequalities if $q_j[u; U]$ is replaced by $\tilde{q}_j[u; U]$ with U defined as in Corollary 3.13. The following table lists the optimal $C(k, r)$ in (3.46) with $1 \leq k \leq 4$ and $k+1 \leq r \leq k+2$.

(k, r)	$C(k, r)$
(1, 2)	$\frac{1}{2\pi} \approx 0.159154943$
(1, 3)	0.0446961624
(2, 3)	0.0162145975
(2, 4)	0.00403144180
(3, 4)	0.00119363570
(3, 5)	0.000267584017
(4, 5)	0.0000684049992
(4, 6)	0.0000140442003

Table 2. Optimal constants $C(k, r)$.

For practical reasons, if we do not care about the choice of U , we may let

$$U(u) = U_\epsilon(u) = \frac{u^2}{2} + \epsilon V(u),$$

where $\epsilon > 0$ is any small constant and $V(u)$ is any convex function. Then, $q_j[u; U_\epsilon] \leq \epsilon \max|V'''(u)|$. In numerical implementations we can really replace $\max|V'''(u)|$ with a uniform constant whenever the numerical solution is observed to be uniformly bounded; therefore, $q_j[u; U_\epsilon]$ can be made arbitrarily small depending on how small ϵ is. As for the term $|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-|$, they can clearly be bounded by $|\llbracket u \rrbracket_{j-\frac{1}{2}}| + |\llbracket u \rrbracket_{j+\frac{1}{2}}|$ multiplied by the Local Lipschitz constant of $\hat{f}$ depending on $\|u\|_{L^\infty(I)}$. Therefore, we may replace $C_j[u; U, r]$ by

$$D_j[u; r, \epsilon_1, \epsilon_2] := \epsilon_1 h_j^{r-\frac{3}{2}} \|\partial_x^r f(u)\|_{L^2(I_j)} + \epsilon_2 \frac{|\llbracket u \rrbracket_{j-\frac{1}{2}}| + |\llbracket u \rrbracket_{j+\frac{1}{2}}|}{h_j}, \quad (3.48)$$

where $\epsilon_1, \epsilon_2 > 0$ are arbitrarily chosen. Then, we can define our last scheme in this section: seek $u \in \mathbb{V}^k$ such that for all $v \in \mathbb{V}^k$ we have

$$\int_{I_j} u_t v \, dx - \int_{I_j} f(u) v_x \, dx + \hat{f}_{j+\frac{1}{2}} \cdot v_{j+\frac{1}{2}}^- - \hat{f}_{j-\frac{1}{2}} \cdot v_{j-\frac{1}{2}}^+ = -D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (u - \bar{u}_j) v \, dx. \quad (\text{Scheme D})$$

This scheme is computationally efficient. Except for the first term in the definition of $D_j[u; r, \epsilon_1, \epsilon_2]$ involving $\|\partial_x^r f(u)\|_{L^2(I_j)}$, which is needed to bound the variation of u on the interior of each cell, (Scheme D) is in striking similarity with the OFDG scheme proposed in [13]. From Proposition 3.2 and Corollary 3.13, we deduce that (Scheme D) still admits infinitely many entropy inequalities (although they might be very close to the square-entropy) if the numerical solution remains bounded. In addition, more entropies are included as we increase ϵ_1 and ϵ_2 .

The optimal error estimates of (Scheme B), (Scheme C) and (Scheme D) for smooth solutions will be proved in Section 6, and the proof of strong convergence for general

discontinuous data will be given in Section 7. In the next two sections, we will provide generalizations of these schemes to one and higher dimensional systems of conservation laws.

4. The Operator DG method for 1D systems of conservation laws

Consider an $m \times m$ system of conservation laws in one space dimension:

$$\mathbf{u}_t(x, t) + \mathbf{f}(\mathbf{u}(x, t))_x = \mathbf{0}, \quad x \in I, \quad t \geq 0. \quad (4.1)$$

where $\mathbf{f} : \Omega' \mapsto \mathbb{R}^m$ is a smooth function defined on some open set $\Omega' \subset \mathbb{R}^m$. In what follows, $\mathbf{u}$ and $\mathbf{f}$ are viewed as column vectors. Let $(U, \mathcal{F})$ be any fixed entropy pair for (4.1). Assume $U \in C^3$ and strictly convex ($\nabla^2 U > 0$ on Ω'). The entropy flux $\mathcal{F}$ is defined by

$$\nabla \mathcal{F}(\mathbf{u})^T = \nabla U(\mathbf{u})^T \nabla \mathbf{f}(\mathbf{u}), \quad (4.2)$$

where the gradients $\nabla \mathcal{F}(\mathbf{u})$ and $\nabla U(\mathbf{u})$ are also viewed as column vectors, and $\nabla \mathbf{f}$ denotes the Jacobian matrix of $\mathbf{f}$. In addition, let

$$\psi(\mathbf{u}) := \nabla U(\mathbf{u})^T \mathbf{f}(\mathbf{u}) - \mathcal{F}(\mathbf{u}), \quad (4.3)$$

so that $\nabla \psi(\mathbf{u})^T = \mathbf{f}(\mathbf{u})^T \nabla^2 U(\mathbf{u})$. For clarity, we also adopt the notation $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b}$ for the dot product of two column vectors $\mathbf{a}$ and $\mathbf{b}$. Let $|\mathbf{a}|$ denote the Euclidean norm of the vector $\mathbf{a}$. Define $(\mathbb{V}^k)^m$ as the space of all vector-valued functions $\mathbf{v} : I \mapsto \mathbb{R}^m$ with each component belonging to $\mathbb{V}^k$. As in Section 3, we have the following estimates analogous to Lemmas 3.4 and 3.6.

Lemma 4.1. *For any $\mathbf{u} \in (\mathbb{V}^k)^m$, we have*

$$\begin{aligned} |\pi \circ [\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+)| &\leq \frac{m^{3/2}\beta_k}{2h_j} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 dx, \\ |\pi \circ [\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j+\frac{1}{2}}^-)](x_{j+\frac{1}{2}}^-)| &\leq \frac{m^{3/2}\beta_k}{2h_j} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 dx, \end{aligned}$$

where β_k may be chosen to be the same constant as in Lemma 3.4; in particular,

$$\beta_k = 2\mathcal{A}_k$$

is admissible, with $\mathcal{A}_k$ defined in (B.3). Moreover,

$$\max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| = \max_{\substack{1 \leq i_1, i_2, i_3 \leq m \\ \mathbf{w} \in \text{co}(\mathbf{u}(I_j))}} \left| \frac{\partial^3 U(\mathbf{w})}{\partial u_{i_1} \partial u_{i_2} \partial u_{i_3}} \right|$$

is the maximum over all possible third order partial derivatives of U in the convex hull of the states $\mathbf{u}(I_j) = \{\mathbf{u}(x) : x \in I_j\}$.

Proof. We prove the estimate at $x_{j-\frac{1}{2}}^+$; the estimate at the other endpoint follows by symmetry. Introduce the reference coordinate $\xi \in [-1, 1]$ by

$$x = x_j + \frac{h_j}{2}\xi,$$

and define the vector-valued polynomial

$$\mathbf{p}(\xi) := \mathbf{u}\left(x_j + \frac{h_j}{2}\xi\right) - \bar{\mathbf{u}}_j.$$

Then each component of $\mathbf{p}$ belongs to $\mathbb{P}^k([-1, 1])$ and has zero average. In addition,

$$\mathbf{u}(x) - \mathbf{u}_{j-\frac{1}{2}}^+ = \mathbf{p}(\xi) - \mathbf{p}(-1).$$

Set

$$M_j := \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})|.$$

For each component $r = 1, \dots, m$, Taylor's theorem with integral remainder gives

$$\begin{aligned} & [\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j-\frac{1}{2}}^+) - \nabla^2 U(\mathbf{u}_{j-\frac{1}{2}}^+)(\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+)]_r \\ &= \int_0^1 (1-s) \sum_{a,b=1}^m \frac{\partial^3 U}{\partial u_r \partial u_a \partial u_b} \left(\mathbf{u}_{j-\frac{1}{2}}^+ + s(\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+) \right) (\mathbf{u}_a - \mathbf{u}_{a,j-\frac{1}{2}}^+)(\mathbf{u}_b - \mathbf{u}_{b,j-\frac{1}{2}}^+) \, ds. \end{aligned}$$

Since the line segment joining $\mathbf{u}_{j-\frac{1}{2}}^+$ and $\mathbf{u}(x)$ is contained in $\text{co}(\mathbf{u}(I_j))$, it follows that

$$\begin{aligned} & \left| \nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j-\frac{1}{2}}^+) - \nabla^2 U(\mathbf{u}_{j-\frac{1}{2}}^+)(\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+) \right| \\ & \leq \frac{m^{3/2} M_j}{2} |\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+|^2. \end{aligned}$$

Indeed, for each component the absolute value is bounded by

$$\frac{M_j}{2} \left(\sum_{a=1}^m |\mathbf{u}_a - \mathbf{u}_{a,j-\frac{1}{2}}^+|^2 \right) \leq \frac{m M_j}{2} |\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+|^2,$$

and taking the Euclidean norm over the m components yields the factor $m^{3/2}$.

Each component of

$$\nabla^2 U(\mathbf{u}_{j-\frac{1}{2}}^+)(\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+)$$

belongs to $\mathbb{P}^k(I_j)$ and vanishes at $x_{j-\frac{1}{2}}^+$. Since π acts componentwise, its projected value at that endpoint is therefore zero. Using the endpoint reproducing kernel K_k^- from (B.1), we obtain

$$\left| \pi \circ [\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+) \right|$$

$$\begin{aligned}
 &\leq \int_{-1}^1 |K_k^-(\xi)| \left| \nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j-\frac{1}{2}}^+) - \nabla^2 U(\mathbf{u}_{j-\frac{1}{2}}^+)(\mathbf{u} - \mathbf{u}_{j-\frac{1}{2}}^+) \right| d\xi \\
 &\leq \frac{m^{3/2} M_j}{2} \int_{-1}^1 |K_k^-(\xi)| |\mathbf{p}(\xi) - \mathbf{p}(-1)|^2 d\xi.
 \end{aligned}$$

The scalar inequality defining $\mathcal{A}_k$ extends to vector-valued polynomials component-wise. This can be seen by writing $\mathbf{p} = (p_1, \dots, p_m)^T$ and using (B.3),

$$\begin{aligned}
 \int_{-1}^1 |K_k^-(\xi)| |\mathbf{p}(\xi) - \mathbf{p}(-1)|^2 d\xi &= \sum_{r=1}^m \int_{-1}^1 |K_k^-(\xi)| |p_r(\xi) - p_r(-1)|^2 d\xi \\
 &\leq \mathcal{A}_k \sum_{r=1}^m \int_{-1}^1 |p_r(\xi)|^2 d\xi \\
 &= \mathcal{A}_k \int_{-1}^1 |\mathbf{p}(\xi)|^2 d\xi.
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 &\left| \pi \circ [\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+) \right| \\
 &\leq \frac{m^{3/2} M_j \mathcal{A}_k}{2} \int_{-1}^1 |\mathbf{p}(\xi)|^2 d\xi \\
 &= \frac{m^{3/2} \mathcal{A}_k}{h_j} M_j \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 dx \\
 &= \frac{m^{3/2} \beta_k}{2h_j} M_j \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 dx,
 \end{aligned}$$

where in the last line we take $\beta_k = 2\mathcal{A}_k$.

For the right endpoint, use $K_k^+(\xi) = K_k^-(-\xi)$ and the transformation $\mathbf{p}(\xi) \mapsto \mathbf{p}(-\xi)$. The zero-average condition and the vector L^2 norm are preserved, so exactly the same constant $\mathcal{A}_k$, and hence the same β_k , applies. $\blacksquare$

Lemma 4.2. For any $\mathbf{u} \in (\mathbb{V}^k)^m$, we have

$$\begin{aligned}
 &\left| \int_{I_j} \nabla U(\mathbf{u}) \cdot (\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]) dx \right| \\
 &\leq \frac{m^{3/2} C_k}{2\sqrt{h_j}} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| \|\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]\|_{L^2(I_j)} \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 dx
 \end{aligned}$$

where $C_k > 0$ is the optimal constant from (3.19), and $\max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})|$ is the same maximum from Lemma 4.1.

Proof. Note that $\nabla U(\bar{\mathbf{u}}_j) + \nabla^2 U(\bar{\mathbf{u}}_j)(\mathbf{u}(x) - \bar{\mathbf{u}}_j)$ is a k -th order polynomial on I_j , which is orthogonal to $\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]$. Therefore,

$$\left| \int_{I_j} \nabla U(\mathbf{u}) \cdot (\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]) dx \right|$$

$$\begin{aligned}
&= \left| \int_{I_j} (\nabla U(\mathbf{u}) - \nabla U(\bar{\mathbf{u}}_j) - \nabla^2 U(\bar{\mathbf{u}}_j)(\mathbf{u} - \bar{\mathbf{u}}_j)) \cdot (\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]) \, dx \right| \\
&\leq \frac{m^{3/2}}{2} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| \int_{I_j} |\mathbf{u} - \bar{\mathbf{u}}_j|^2 |\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]| \, dx \\
&\leq \frac{m^{3/2}}{2} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| \|\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]\|_{L^2(I_j)} \sqrt{\int_{I_j} |\mathbf{u} - \bar{\mathbf{u}}_j|^4 \, dx} \\
&\stackrel{(3.19)}{\leq} \frac{m^{3/2} C_k}{2\sqrt{h_j}} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})| \|\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]\|_{L^2(I_j)} \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 \, dx.
\end{aligned}$$

■

In addition, Hölder's Defect Formula also holds in $\mathbb{R}^m$:

$$\int_{I_j} U(\mathbf{u}) \, dx - h_j U(\bar{\mathbf{u}}_j) \geq \frac{1}{2} \min_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\sigma_{\min}(\nabla^2 U(\mathbf{w}))| \int_{I_j} |\mathbf{u}(x) - \bar{\mathbf{u}}_j|^2 \, dx, \quad (4.4)$$

where $\sigma_{\min}(\nabla^2 U)$ denotes the (positive) minimum eigenvalue of the Hessian matrix $\nabla^2 U$, and $\min_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))}$ means taking the minimum over the convex hull of the states $\mathbf{u}(I_j) = \{\mathbf{u}(x) : x \in I_j\}$. Therefore, define

$$\begin{aligned}
B'_j[\mathbf{u}; U] &:= \frac{m^{3/2}}{2} \frac{\max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})|}{\min_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\sigma_{\min}(\nabla^2 U(\mathbf{w}))|} \times \\
&\quad \left\{ \frac{C_k}{\sqrt{h_j}} \|\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]\|_{L^2(I_j)} + \frac{\beta_k}{h_j} \left(|\mathbf{H}_{j-\frac{1}{2}}^+| + |\mathbf{H}_{j+\frac{1}{2}}^-| \right) \right\},
\end{aligned}$$

where

$$\mathbf{H}_{j-\frac{1}{2}}^+ := \hat{\mathbf{f}}_{j-\frac{1}{2}} - \mathbf{f}(\mathbf{u})_{j-\frac{1}{2}}^+, \quad \mathbf{H}_{j+\frac{1}{2}}^- := -\hat{\mathbf{f}}_{j+\frac{1}{2}} + \mathbf{f}(\mathbf{u})_{j+\frac{1}{2}}^-, \quad (4.5)$$

and $\hat{\mathbf{f}}$ is the numerical flux. Then, (Scheme B) can be directly generalized from 1D scalar equation to 1D system as follows: seek $\mathbf{u} \in (\mathbb{V}^k)^m$ such that for all $\mathbf{v} \in (\mathbb{V}^k)^m$ we have

$$\int_{I_j} \mathbf{u}_t \cdot \mathbf{v} \, dx = - \int_{I_j} \mathbf{f}(\mathbf{u})_x \cdot \mathbf{v} \, dx + \mathbf{H}_{j-\frac{1}{2}}^+ \cdot \mathbf{v}_{j-\frac{1}{2}}^+ + \mathbf{H}_{j+\frac{1}{2}}^- \cdot \mathbf{v}_{j+\frac{1}{2}}^- - B'_j[\mathbf{u}; U] \int_{I_j} (\mathbf{u} - \bar{\mathbf{u}}_j) \cdot \mathbf{v} \, dx.$$

(Scheme B')

After performing integration by parts on $\int_{I_j} \mathbf{f}(\mathbf{u})_x \cdot \mathbf{v} \, dx$, we also get the equivalent weak formulation of (Scheme B') as

$$\int_{I_j} \mathbf{u}_t \cdot \mathbf{v} \, dx - \int_{I_j} \mathbf{f}(\mathbf{u}) \cdot \mathbf{v}_x \, dx + \hat{\mathbf{f}}_{j+\frac{1}{2}} \cdot \mathbf{v}_{j+\frac{1}{2}}^- - \hat{\mathbf{f}}_{j-\frac{1}{2}} \cdot \mathbf{v}_{j-\frac{1}{2}}^+ = -B'_j[\mathbf{u}; U] \int_{I_j} (\mathbf{u} - \bar{\mathbf{u}}_j) \cdot \mathbf{v} \, dx. \quad (4.6)$$

Next, we borrow the definition of entropy stable numerical flux from [1]:

Definition 4.3. A consistent two-point numerical flux $\hat{\mathbf{f}}(\mathbf{u}^-, \mathbf{u}^+)$ is entropy stable for a given entropy function U if

$$\hat{\mathbf{f}}(\mathbf{u}^-, \mathbf{u}^+) \cdot (\nabla U(\mathbf{u}^+) - \nabla U(\mathbf{u}^-)) - (\psi(\mathbf{u}^+) - \psi(\mathbf{u}^-)) \leq 0. \quad (4.7)$$

It is shown in [1] that the class of HLL flux is entropy stable, which includes the local Lax–Friedrichs flux. The next theorem shows that (Scheme B') satisfies local entropy inequality for $U(\cdot)$, provided that $\hat{\mathbf{f}}$ is any entropy stable flux.

Theorem 4.4 (Local Entropy Inequality for (Scheme B')). *Suppose that $\hat{\mathbf{f}}$ in (4.5) is an entropy stable flux. Then (Scheme B') satisfies the local semi-discrete entropy inequality*

$$\frac{d}{dt} \int_{I_j} U(\mathbf{u}) dx + \hat{\mathcal{F}}_{j+\frac{1}{2}} - \hat{\mathcal{F}}_{j-\frac{1}{2}} \leq 0, \quad \forall \mathbf{u} \in (\mathbb{V}^k)^m, \quad \forall j = 1, \dots, N, \quad (4.8)$$

where $\hat{\mathcal{F}}_{j+\frac{1}{2}} = \hat{\mathcal{F}}(\mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+)$ and

$$\hat{\mathcal{F}}(\mathbf{u}^-, \mathbf{u}^+) := \nabla U(\mathbf{u}^-) \cdot \hat{\mathbf{f}}(\mathbf{u}^-, \mathbf{u}^+) - \psi(\mathbf{u}^-), \quad (4.9)$$

with ψ defined by (4.3).

Proof. This is essentially the same as the proof of Theorem 3.9. The only differences are that now we use the definition (4.7) of the entropy stable flux to show the positivity of (3.36), and use Lemmas 4.1–4.2 to bound $|R_{j,1}|$, $|R_{j,2}|$ and $|R_{j,3}|$. ■

Of course, similar to the 1D scalar case, we may use Bramble–Hilbert lemma (3.46) on the tricky term $\|\mathbf{f}(\mathbf{u})_x - \pi \circ [\mathbf{f}(\mathbf{u})_x]\|_{L^2(I_j)}$ in the definition of $B'_j[\mathbf{u}; U]$ and derive the following simplified schemes: seek $\mathbf{u} \in (\mathbb{V}^k)^m$ such that for all $\mathbf{v} \in (\mathbb{V}^k)^m$ we have

$$\begin{aligned} \int_{I_j} \mathbf{u}_t \cdot \mathbf{v} dx - \int_{I_j} \mathbf{f}(\mathbf{u}) \cdot \mathbf{v}_x dx + \hat{\mathbf{f}}_{j+\frac{1}{2}} \cdot \mathbf{v}_{j+\frac{1}{2}}^- - \hat{\mathbf{f}}_{j-\frac{1}{2}} \cdot \mathbf{v}_{j-\frac{1}{2}}^+ \\ = -C'_j[\mathbf{u}; U, r] \int_{I_j} (\mathbf{u} - \bar{\mathbf{u}}_j) \cdot \mathbf{v} dx \quad (\text{Scheme C'}) \end{aligned}$$

or

$$\begin{aligned} \int_{I_j} \mathbf{u}_t \cdot \mathbf{v} dx - \int_{I_j} \mathbf{f}(\mathbf{u}) \cdot \mathbf{v}_x dx + \hat{\mathbf{f}}_{j+\frac{1}{2}} \cdot \mathbf{v}_{j+\frac{1}{2}}^- - \hat{\mathbf{f}}_{j-\frac{1}{2}} \cdot \mathbf{v}_{j-\frac{1}{2}}^+ \\ = -D'_j[\mathbf{u}; r, \epsilon_1, \epsilon_2] \int_{I_j} (\mathbf{u} - \bar{\mathbf{u}}_j) \cdot \mathbf{v} dx, \quad (\text{Scheme D'}) \end{aligned}$$

where the damping coefficients are

$$C'_j[\mathbf{u}; U, r] := \frac{m^{3/2}}{2} \frac{\max_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\partial^3 U(\mathbf{w})|}{\min_{\mathbf{w} \in \text{co}(\mathbf{u}(I_j))} |\sigma_{\min}(\nabla^2 U(\mathbf{w}))|} \times$$

$$\left\{ C_k C(k, r) h_j^{r-\frac{3}{2}} \|\partial_x^r \mathbf{f}(\mathbf{u})\|_{L^2(I_j)} + \frac{\beta_k}{h_j} \left(|\mathbf{H}_{j-\frac{1}{2}}^+| + |\mathbf{H}_{j+\frac{1}{2}}^-| \right) \right\}$$

and

$$D'_j[\mathbf{u}; r, \epsilon_1, \epsilon_2] := \epsilon_1 h_j^{r-\frac{3}{2}} \|\partial_x^r \mathbf{f}(\mathbf{u})\|_{L^2(I_j)} + \epsilon_2 \frac{|\llbracket \mathbf{u} \rrbracket_{j-\frac{1}{2}}| + |\llbracket \mathbf{u} \rrbracket_{j+\frac{1}{2}}|}{h_j}$$

respectively with $k+1 \leq r \leq k+2$.

Because monotonicity of remainder (Proposition 3.2) still holds for systems of conservation laws, Theorem 4.4 also holds true for (Scheme C'), and also for (Scheme D') with sufficiently large ϵ_1, ϵ_2 if the numerical solution is uniformly bounded.

5. The Operator DG method for higher dimensional systems of conservation laws

Consider an $m \times m$ system of conservation laws in d space dimensions:

$$\frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial t} + \sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u}(\mathbf{x}, t))}{\partial x^{(i)}} = \mathbf{0}, \quad \mathbf{x} = (x^{(1)}, \dots, x^{(d)}) \in \Omega \subset \mathbb{R}^d, \quad t \geq 0. \quad (5.1)$$

where $\mathbf{f}^{(i)} : \Omega' \mapsto \mathbb{R}^m$ for $i = 1, \dots, d$ are smooth functions defined on some open set $\Omega' \subset \mathbb{R}^m$. In what follows, $\mathbf{u}$ and $\mathbf{f}^{(i)}$'s are viewed as column vectors. Let U be any fixed entropy for (5.1) with entropy fluxes $\{\mathcal{F}^{(i)}\}_{i=1}^d$ defined by

$$\nabla \mathcal{F}^{(i)}(\mathbf{u})^T = \nabla U(\mathbf{u})^T \nabla \mathbf{f}^{(i)}(\mathbf{u}), \quad i = 1 \dots, d. \quad (5.2)$$

Assume $U \in C^3$ and strictly convex. In addition, let

$$\psi^{(i)}(\mathbf{u}) := \nabla U(\mathbf{u})^T \mathbf{f}^{(i)}(\mathbf{u}) - \mathcal{F}^{(i)}(\mathbf{u}), \quad i = 1 \dots, d, \quad (5.3)$$

so that $\nabla \psi^{(i)}(\mathbf{u})^T = \mathbf{f}^{(i)}(\mathbf{u})^T \nabla^2 U(\mathbf{u})$.

Let $\mathcal{T}_h$ be a partition of Ω . Assume every $K \in \mathcal{T}_h$ is convex and let ρ_K denote the diameter of the largest sphere inscribed in K . Let $|K|$ be the volume of K and

$$h = \max_{K \in \mathcal{T}_h} h_K, \quad h_K = \text{diam } K, \quad \gamma_K = \frac{h_K}{\rho_K}.$$

Define for $k \geq 1$ the space

$$(\mathbb{V}_d^k)^m := \{ \mathbf{v} \in (L^2(\Omega))^m : \mathbf{v}|_K \in (\mathbb{P}^k(K))^m, \forall K \in \mathcal{T}_h \}$$

of all piecewise (possibly discontinuous) polynomials on Ω that coincides with a k -th order polynomial on each element K . For any function $\mathbf{v}$ continuous on $\overline{K}$ and $\mathbf{x} \in \partial K$, denote the limit of $\mathbf{v}$ when approaching $\mathbf{x}$ from the interior of K by $\mathbf{v}(\mathbf{x}^{\text{int } K})$, and the limit from outside of K by $\mathbf{v}(\mathbf{x}^{\text{ext } K})$. Let π_h be the L^2 -projection onto $(\mathbb{V}_d^k)^m$, and denote the average of $\mathbf{u}$ on K by $\bar{\mathbf{u}}_K$. With similar proofs, one can show the following generalizations of Lemmas 4.1 and 4.2 to any element $K \in \mathcal{T}_h$:

Lemma 5.1. For any $K \in \mathcal{T}_h$ and $\mathbf{u} \in (\mathbb{V}_d^k)^m$, we have

$$|\pi_h \circ [\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}(\mathbf{x}^{\text{int } K}))](\mathbf{x}^{\text{int } K})| \leq \frac{C(k, m, d, \gamma_K)}{|K|} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\partial^3 U(\mathbf{w})| \int_K |\mathbf{u}(x) - \bar{\mathbf{u}}_K|^2 d\mathbf{x}$$

for all $\mathbf{x} \in \partial K$, and

$$\begin{aligned} & \left| \sum_{i=1}^d \int_K \nabla U(\mathbf{u}) \cdot \left(\frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} - \pi_h \circ \left[\frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \right] \right) d\mathbf{x} \right| \\ & \leq \frac{C(k, m, d, \gamma_K)}{\sqrt{|K|}} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\partial^3 U(\mathbf{w})| \left\| \sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} - \pi_h \circ \left[\sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \right] \right\|_{L^2(K)} \int_K |\mathbf{u}(x) - \bar{\mathbf{u}}_K|^2 d\mathbf{x}, \end{aligned}$$

where $C(k, m, d, \gamma_K) > 0$ is constant depending only on k, m, d, γ_K , and

$$\max_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\partial^3 U(\mathbf{w})| = \max_{\substack{1 \leq i_1, i_2, i_3 \leq m \\ \mathbf{w} \in \text{co}(\mathbf{u}(K))}} \left| \frac{\partial^3 U(\mathbf{w})}{\partial u_{i_1} \partial u_{i_2} \partial u_{i_3}} \right|$$

is the maximum over all possible third order partial derivatives of U over the convex hull of the states $\mathbf{u}(K) = \{\mathbf{u}(\mathbf{x}) : \mathbf{x} \in K\}$.

For $\mathbf{u} \in (\mathbb{V}_d^k)^m$ and using the constant $C(k, m, d, \gamma_K)$ from Lemma 5.1, define

$$\begin{aligned} B_K''[\mathbf{u}; U] &:= C(k, m, d, \gamma_K) \frac{\max_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\partial^3 U(\mathbf{w})|}{\min_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\sigma_{\min}(\nabla^2 U(\mathbf{w}))|} \times \\ & \quad \left\{ \frac{1}{\sqrt{|K|}} \left\| \sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} - \pi_h \circ \left[\sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \right] \right\|_{L^2(K)} + \frac{1}{|K|} \int_{\partial K} |\mathbf{H}_K(\mathbf{x}, \mathbf{n})| ds(\mathbf{x}) \right\}, \end{aligned}$$

where $\mathbf{n} = (n_1, \dots, n_d)^T$ is the unit outward normal with respect to ∂K ,

$$\mathbf{H}_K(\mathbf{x}, \mathbf{n}) := -\hat{\mathbf{f}}(\mathbf{u}(\mathbf{x}^{\text{int } K}), \mathbf{u}(\mathbf{x}^{\text{ext } K}), \mathbf{n}) + \sum_{i=1}^d n_i \mathbf{f}^{(i)}(\mathbf{u}(\mathbf{x}^{\text{int } K})), \quad (5.4)$$

and $\hat{\mathbf{f}}$ is the directional numerical flux. Then, (Scheme B') can be directly generalized from 1D system to d -dimension as follows: seek $\mathbf{u} \in (\mathbb{V}_d^k)^m$ such that for all $\mathbf{v} \in (\mathbb{V}_d^k)^m$ we have

$$\begin{aligned} \int_K \mathbf{u}_t \cdot \mathbf{v} d\mathbf{x} &= - \sum_{i=1}^d \int_K \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \cdot \mathbf{v} d\mathbf{x} + \int_{\partial K} \mathbf{H}_K(\mathbf{x}, \mathbf{n}) \cdot \mathbf{v}(\mathbf{x}^{\text{int } K}) ds(\mathbf{x}) \\ &\quad - B_K''[\mathbf{u}; U] \int_K (\mathbf{u} - \bar{\mathbf{u}}_K) \cdot \mathbf{v} d\mathbf{x}, \end{aligned} \quad (\text{Scheme B''})$$

After using divergence theorem on $\sum_{i=1}^d \int_K \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \cdot \mathbf{v} d\mathbf{x}$, we also get the equivalent weak formulation of (Scheme B'') as

$$\int_K \mathbf{u}_t \cdot \mathbf{v} d\mathbf{x} - \sum_{i=1}^d \int_K \mathbf{f}^{(i)}(\mathbf{u}) \cdot \frac{\partial \mathbf{v}}{\partial x^{(i)}} d\mathbf{x} + \int_{\partial K} \hat{\mathbf{f}}(\mathbf{u}(\mathbf{x}^{\text{int } K}), \mathbf{u}(\mathbf{x}^{\text{ext } K}), \mathbf{n}) \cdot \mathbf{v}(\mathbf{x}^{\text{int } K}) ds(\mathbf{x})$$

$$= -B_K''[\mathbf{u}; U] \int_K (\mathbf{u} - \bar{\mathbf{u}}_K) \cdot \mathbf{v} \, d\mathbf{x}. \quad (5.5)$$

Now, the definition (4.7) of entropy stable numerical flux is generalized to d -dimension following [1].

Definition 5.2. Given a normal vector $\mathbf{n} \in \mathbb{R}^d$, a directional numerical flux $\hat{\mathbf{f}}(\mathbf{u}^{\text{int } K}, \mathbf{u}^{\text{ext } K}, \mathbf{n})$ is consistent if

$$\hat{\mathbf{f}}(\mathbf{u}, \mathbf{u}, \mathbf{n}) = \sum_{i=1}^d n_i \mathbf{f}^{(i)}(\mathbf{u}). \quad (5.6)$$

It is called conservative if

$$\hat{\mathbf{f}}(\mathbf{u}^{\text{ext } K}, \mathbf{u}^{\text{int } K}, -\mathbf{n}) = -\hat{\mathbf{f}}(\mathbf{u}^{\text{int } K}, \mathbf{u}^{\text{ext } K}, \mathbf{n}). \quad (5.7)$$

A directional numerical flux is *entropy stable* for a given entropy $U(\cdot)$ if it is consistent and conservative and satisfies

$$\hat{\mathbf{f}}(\mathbf{u}^{\text{int } K}, \mathbf{u}^{\text{ext } K}, \mathbf{n}) \cdot \left(\nabla U(\mathbf{u}^{\text{ext } K}) - \nabla U(\mathbf{u}^{\text{int } K}) \right) - \sum_{i=1}^d n_i (\psi^{(i)}(\mathbf{u}^{\text{ext } K}) - \psi^{(i)}(\mathbf{u}^{\text{int } K})) \leq 0. \quad (5.8)$$

The class of HLL fluxes, including the local Lax-Friedrichs flux, are entropy stable. The next theorem shows that (Scheme B'') satisfies local entropy inequality for $U(\cdot)$, provided that $\hat{\mathbf{f}}$ is any entropy stable directional flux.

Theorem 5.3 (Local Entropy Inequality for (Scheme B'')). *Suppose that $\hat{\mathbf{f}}$ in (5.4) is an entropy stable directional flux. Then (Scheme B'') satisfies the local semi-discrete entropy inequality*

$$\frac{d}{dt} \int_K U(\mathbf{u}) \, d\mathbf{x} + \int_{\partial K} \hat{\mathcal{F}}(\mathbf{u}(\mathbf{x}^{\text{int } K}), \mathbf{u}(\mathbf{x}^{\text{ext } K}), \mathbf{n}) \, ds(\mathbf{x}) \leq 0, \quad \forall \mathbf{u} \in (\mathbb{V}_d^k)^m, \quad \forall K \in \mathcal{T}_h, \quad (5.9)$$

where

$$\begin{aligned} \hat{\mathcal{F}}(\mathbf{u}^{\text{int } K}, \mathbf{u}^{\text{ext } K}, \mathbf{n}) &:= \frac{\nabla U(\mathbf{u}^{\text{int } K}) + \nabla U(\mathbf{u}^{\text{ext } K})}{2} \cdot \hat{\mathbf{f}}(\mathbf{u}^{\text{int } K}, \mathbf{u}^{\text{ext } K}, \mathbf{n}) \\ &\quad - \sum_{i=1}^d n_i \frac{\psi^{(i)}(\mathbf{u}^{\text{int } K}) + \psi^{(i)}(\mathbf{u}^{\text{ext } K})}{2}, \end{aligned} \quad (5.10)$$

with $\psi^{(i)}$ defined by (5.3). Clearly, $\hat{\mathcal{F}}$ is consistent and conservative.

Proof. Fix $K \in \mathcal{T}_h$ and set

$$\mathbf{z} := \pi_h[\nabla U(\mathbf{u})] \in (\mathbb{V}_d^k)^m.$$

Since $\mathbf{u}_t|_K \in (\mathbb{P}^k(K))^m$ and π_h is the L^2 -projection,

$$\int_K \mathbf{u}_t \cdot \mathbf{z} \, d\mathbf{x} = \int_K \mathbf{u}_t \cdot \nabla U(\mathbf{u}) \, d\mathbf{x} = \frac{d}{dt} \int_K U(\mathbf{u}) \, d\mathbf{x}. \quad (5.11)$$

Taking $\mathbf{v} = \mathbf{z}$ in (Scheme B'') gives

$$\begin{aligned} \frac{d}{dt} \int_K U(\mathbf{u}) \, d\mathbf{x} &= - \sum_{i=1}^d \int_K \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \cdot \mathbf{z} \, d\mathbf{x} + \int_{\partial K} \mathbf{H}_K \cdot \mathbf{z}(\mathbf{x}^{\text{int } K}) \, ds \\ &\quad - B_K''[\mathbf{u}; U] \int_K (\mathbf{u} - \bar{\mathbf{u}}_K) \cdot \mathbf{z} \, d\mathbf{x}. \end{aligned} \quad (5.12)$$

By self-adjointness of the L^2 -projection,

$$\begin{aligned} &- \sum_{i=1}^d \int_K \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \cdot \mathbf{z} \, d\mathbf{x} \\ &= - \sum_{i=1}^d \int_K \nabla U(\mathbf{u}) \cdot \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \, d\mathbf{x} + R_{K,1} \\ &= - \int_{\partial K} \sum_{i=1}^d n_i \mathcal{F}^{(i)}(\mathbf{u}(\mathbf{x}^{\text{int } K})) \, ds + R_{K,1}, \end{aligned} \quad (5.13)$$

where

$$R_{K,1} := \int_K \nabla U(\mathbf{u}) \cdot \left\{ \sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} - \pi_h \left[\sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \right] \right\} \, d\mathbf{x}. \quad (5.14)$$

The last equality follows from (5.2) and the divergence theorem.

Moreover, since $\mathbf{u} - \bar{\mathbf{u}}_K \in (\mathbb{P}^k(K))^m$ and has zero average,

$$\int_K (\mathbf{u} - \bar{\mathbf{u}}_K) \cdot \mathbf{z} \, d\mathbf{x} = \int_K (\mathbf{u} - \bar{\mathbf{u}}_K) \cdot (\nabla U(\mathbf{u}) - \nabla U(\bar{\mathbf{u}}_K)) \, d\mathbf{x} =: D_K. \quad (5.15)$$

Put

$$m_K := \min_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} \sigma_{\min}(\nabla^2 U(\mathbf{w})).$$

Since $\bar{\mathbf{u}}_K \in \text{co}(\mathbf{u}(K))$, the segment joining $\bar{\mathbf{u}}_K$ and $\mathbf{u}(\mathbf{x})$ lies in $\text{co}(\mathbf{u}(K))$. Hence

$$\begin{aligned} &(\mathbf{u} - \bar{\mathbf{u}}_K) \cdot (\nabla U(\mathbf{u}) - \nabla U(\bar{\mathbf{u}}_K)) \\ &= \int_0^1 (\mathbf{u} - \bar{\mathbf{u}}_K)^T \nabla^2 U(\bar{\mathbf{u}}_K + \theta(\mathbf{u} - \bar{\mathbf{u}}_K)) (\mathbf{u} - \bar{\mathbf{u}}_K) \, d\theta \geq m_K |\mathbf{u} - \bar{\mathbf{u}}_K|^2, \end{aligned}$$

and therefore

$$D_K \geq m_K \int_K |\mathbf{u} - \bar{\mathbf{u}}_K|^2 \, d\mathbf{x} \geq 0. \quad (5.16)$$

For $\mathbf{x} \in \partial K$, write

$$\mathbf{u}^- := \mathbf{u}(\mathbf{x}^{\text{int } K}), \quad \mathbf{u}^+ := \mathbf{u}(\mathbf{x}^{\text{ext } K}), \quad \mathbf{h} := \hat{\mathbf{f}}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}),$$

and set

$$\mathcal{F}_{\mathbf{n}}(\mathbf{u}) := \sum_{i=1}^d n_i \mathcal{F}^{(i)}(\mathbf{u}), \quad \psi_{\mathbf{n}}(\mathbf{u}) := \sum_{i=1}^d n_i \psi^{(i)}(\mathbf{u}).$$

By (5.4), (5.3), and (5.10),

$$\begin{aligned} & -\mathcal{F}_{\mathbf{n}}(\mathbf{u}^-) + \mathbf{H}_K \cdot \nabla U(\mathbf{u}^-) + \hat{\mathcal{F}}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}) \\ &= \frac{1}{2} \left\{ \mathbf{h} \cdot (\nabla U(\mathbf{u}^+) - \nabla U(\mathbf{u}^-)) - (\psi_{\mathbf{n}}(\mathbf{u}^+) - \psi_{\mathbf{n}}(\mathbf{u}^-)) \right\} =: \frac{1}{2} \mathcal{E}_K(\mathbf{x}). \end{aligned} \quad (5.17)$$

The entropy stability condition (5.8) implies

$$\mathcal{E}_K(\mathbf{x}) \leq 0 \quad \text{on } \partial K. \quad (5.18)$$

Since constant vectors are preserved by π_h ,

$$\mathbf{z}(\mathbf{x}^{\text{int } K}) - \nabla U(\mathbf{u}^-) = \pi_h[\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}^-)](\mathbf{x}^{\text{int } K}).$$

Define

$$R_{K,2} := \int_{\partial K} \mathbf{H}_K(\mathbf{x}, \mathbf{n}) \cdot \pi_h[\nabla U(\mathbf{u}) - \nabla U(\mathbf{u}^-)](\mathbf{x}^{\text{int } K}) \, ds. \quad (5.19)$$

Integrating (5.17) over ∂K , then adding (5.19) to both sides yields

$$\begin{aligned} & - \int_{\partial K} \mathcal{F}_{\mathbf{n}}(\mathbf{u}^-) \, ds + \int_{\partial K} \mathbf{H}_K \cdot \mathbf{z}(\mathbf{x}^{\text{int } K}) \, ds + \int_{\partial K} \hat{\mathcal{F}}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}) \, ds \\ &= R_{K,2} + \frac{1}{2} \int_{\partial K} \mathcal{E}_K(\mathbf{x}) \, ds. \end{aligned} \quad (5.20)$$

Combining (5.12), (5.13), (5.15), and (5.20), we obtain

$$\begin{aligned} & \frac{d}{dt} \int_K U(\mathbf{u}) \, d\mathbf{x} + \int_{\partial K} \hat{\mathcal{F}}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}) \, ds \\ &= R_{K,1} + R_{K,2} + \frac{1}{2} \int_{\partial K} \mathcal{E}_K(\mathbf{x}) \, ds - B_K''[\mathbf{u}; U] D_K. \end{aligned} \quad (5.21)$$

By Lemma 5.1,

$$\begin{aligned} |R_{K,1}| &\leq \frac{C(k, m, d, \gamma_K)}{\sqrt{|K|}} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\partial^3 U(\mathbf{w})| \left\| \sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} - \pi_h \left[\sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \right] \right\|_{L^2(K)} \\ &\quad \times \int_K |\mathbf{u} - \bar{\mathbf{u}}_K|^2 \, d\mathbf{x}, \end{aligned} \quad (5.22)$$

and

$$|R_{K,2}| \leq \frac{C(k, m, d, \gamma_K)}{|K|} \max_{\mathbf{w} \in \text{co}(\mathbf{u}(K))} |\partial^3 U(\mathbf{w})| \left(\int_{\partial K} |\mathbf{H}_K(\mathbf{x}, \mathbf{n})| \, ds \right) \int_K |\mathbf{u} - \bar{\mathbf{u}}_K|^2 \, d\mathbf{x}. \quad (5.23)$$

Thus, the definition of $B_K''[\mathbf{u}; U]$ and (5.16) give

$$R_{K,1} + R_{K,2} \leq |R_{K,1}| + |R_{K,2}| \leq B_K''[\mathbf{u}; U] m_K \int_K |\mathbf{u} - \bar{\mathbf{u}}_K|^2 \, d\mathbf{x} \leq B_K''[\mathbf{u}; U] D_K. \quad (5.24)$$

Substituting this into (5.21) and using (5.18), we conclude that

$$\frac{d}{dt} \int_K U(\mathbf{u}) \, d\mathbf{x} + \int_{\partial K} \hat{\mathcal{F}}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}) \, ds \leq \frac{1}{2} \int_{\partial K} \mathcal{E}_K(\mathbf{x}) \, ds \leq 0,$$

which is precisely (5.9). $\blacksquare$

Similar as in the 1D system case, we may use Bramble-Hilbert lemma (3.46) on the term

$$\left\| \sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} - \pi_h \circ \left[\sum_{i=1}^d \frac{\partial \mathbf{f}^{(i)}(\mathbf{u})}{\partial x^{(i)}} \right] \right\|_{L^2(K)}$$

and derive the following simplified scheme: seek $\mathbf{u} \in (\mathbb{V}_d^k)^m$ such that for all $\mathbf{v} \in (\mathbb{V}_d^k)^m$ we have

$$\begin{aligned} \int_K \mathbf{u}_t \cdot \mathbf{v} \, d\mathbf{x} - \sum_{i=1}^d \int_K \mathbf{f}^{(i)}(\mathbf{u}) \cdot \frac{\partial \mathbf{v}}{\partial x^{(i)}} \, d\mathbf{x} + \int_{\partial K} \hat{\mathbf{f}}(\mathbf{u}(\mathbf{x}^{\text{int } K}), \mathbf{u}(\mathbf{x}^{\text{ext } K}), \mathbf{n}) \cdot \mathbf{v}(\mathbf{x}^{\text{int } K}) \, ds(\mathbf{x}) \\ = -D_K''[\mathbf{u}; r, \epsilon_1, \epsilon_2] \int_K (\mathbf{u} - \bar{\mathbf{u}}_K) \cdot \mathbf{v} \, d\mathbf{x}, \end{aligned}$$

(Scheme D'')

where

$$D_K''[\mathbf{u}; r, \epsilon_1, \epsilon_2] := \epsilon_1 \frac{h_K^{r-1}}{\sqrt{|K|}} \left(\sum_{i=1}^d \|\mathbf{f}^{(i)}(\mathbf{u})\|_{H^r(K)}^2 \right)^{1/2} + \frac{\epsilon_2}{|K|} \int_{\partial K} |\mathbf{H}_K(\mathbf{x}, \mathbf{n})| \, ds(\mathbf{x}), \quad k+1 \leq r \leq k+2.$$

Note that here we still integrate $|\mathbf{H}_K(\mathbf{x}, \mathbf{n})|$ along the boundary of K instead of $\|\mathbf{u}\|$. As before, (Scheme D'') also satisfies Theorem 5.3 for sufficiently large ϵ_1, ϵ_2 if the numerical solution is uniformly bounded.

6. Optimal error estimates to smooth solutions of general nonlinear scalar conservation laws

In this section, we show the optimal error estimates to smooth solutions of (Scheme B), (Scheme C) and (Scheme D) for general nonlinear scalar conservation laws. For the partition of the interval $I = \cup_{j=1}^N I_j$ with $|I_j| = h_j$, define

$$h := \max_{1 \leq j \leq N} h_j.$$

Assume the mesh is quasi-uniform, namely there exists a constant $C > 0$ such that

$$\max_{1 \leq j \leq N} h_j \leq C \min_{1 \leq j \leq N} h_j.$$

Theorem 6.1. Let $\tilde{u}(x, t)$ denote the exact solution of the 1D nonlinear scalar conservation law (2.1) with initial condition $\tilde{u}_0(x)$ and periodic or compactly supported boundary conditions. Suppose $f(\cdot)$ is sufficiently smooth, and $\exists T > 0$ such that $\tilde{u}$ is smooth for all $t \in [0, T]$. For any given integer $k \geq 1$, let $u(\cdot, t) \in \mathbb{V}^k$ denote the numerical solution of (Scheme B), (Scheme C) or (Scheme D) with a quasi-uniform mesh and initial condition $\pi \tilde{u}_0 \in \mathbb{V}^k$. If an upwind numerical flux is used, then for small enough h there holds the following optimal error estimate:

$$\max_{0 \leq t \leq T} \|\tilde{u}(\cdot, t) - u(\cdot, t)\|_{L^2(I)} \leq Ch^{k+1}. \quad (6.1)$$

Moreover, if a general monotone flux is used, then for small enough h there holds

$$\max_{0 \leq t \leq T} \|\tilde{u}(\cdot, t) - u(\cdot, t)\|_{L^2(I)} \leq Ch^{k+\frac{1}{2}}. \quad (6.2)$$

Here, $C > 0$ depends on the uniform bound on $\tilde{u}$ and its derivatives in $I \times [0, T]$, but not on u and h .

Proof. As it will be clear at the end, it suffices to show (6.1) and (6.2) for (Scheme D) with any $k+1 \leq r \leq k+2$ and any $\epsilon_1, \epsilon_2 > 0$. Recall that (Scheme D) is defined as follows: seek $u(\cdot, t) \in \mathbb{V}^k$ such that for all $v \in \mathbb{V}^k$ we have

$$\begin{aligned} \int_{I_j} u_t(x, t) v(x) dx - \int_{I_j} f(u(x, t)) v_x(x) dx + \hat{f}_{j+\frac{1}{2}}(t) \cdot v_{j+\frac{1}{2}}^- - \hat{f}_{j-\frac{1}{2}}(t) \cdot v_{j-\frac{1}{2}}^+ \\ = -D_j[u(\cdot, t); r, \epsilon_1, \epsilon_2] \int_{I_j} (u(x, t) - \bar{u}_j(t)) v(x) dx, \end{aligned} \quad (6.3)$$

where $\bar{u}_j(t)$ is the average of $u(x, t)$ on the cell I_j , $\hat{f}_{j+\frac{1}{2}}(t) = \hat{f}(u(x_{j+\frac{1}{2}}^-, t), u(x_{j+\frac{1}{2}}^+, t))$, and

$$D_j[u(\cdot, t); r, \epsilon_1, \epsilon_2] = \epsilon_1 h_j^{r-\frac{3}{2}} \|\partial_x^r f(u)(\cdot, t)\|_{L^2(I_j)} + \epsilon_2 \frac{|\llbracket u(\cdot, t) \rrbracket_{j-\frac{1}{2}}| + |\llbracket u(\cdot, t) \rrbracket_{j+\frac{1}{2}}|}{h_j} \quad (6.4)$$

is defined in (3.48). Set $e(x, t) := \tilde{u}(x, t) - u(x, t)$ and denote

$$e = \xi - \eta, \quad \xi := \mathbb{Q} \tilde{u} - u, \quad \eta := \mathbb{Q} \tilde{u} - \tilde{u}, \quad (6.5)$$

where the projection $\mathbb{Q}$ is

- (1) the L^2 -projection π if a general monotone flux is used;
- (2) if an upwind flux is used, $\mathbb{Q}$ is a modified Gauss-Radau projection that depends on the initial condition $\tilde{u}_0$: for any continuous function w on I , define $\mathbb{Q}w \in \mathbb{V}^k$ such that

$$\begin{cases} \int_{I_j} (\mathbb{Q}w - w) v dx = 0, & \forall v \in \mathbb{P}^{k-1}(I_j), \\ \mathbb{Q}w(x_{j+\frac{1}{2}}^-) = w(x_{j+\frac{1}{2}}), & \text{if } f'(\tilde{u}_0) > 0 \text{ on } I_j, \\ \mathbb{Q}w(x_{j-\frac{1}{2}}^+) = w(x_{j-\frac{1}{2}}), & \text{if } f'(\tilde{u}_0) < 0 \text{ on } I_j, \\ \mathbb{Q}w = \pi w, & \text{if } f'(\tilde{u}_0) \text{ has at least one zero on } I_j. \end{cases}$$

Because the points where $f'(\tilde{u}) = 0$ do not move with time as long as $\tilde{u}$ remains smooth, the definition of $\mathbb{Q}$ would not change even if we replace $f'(\tilde{u}_0)$ by the time-dependent $f'(\tilde{u})$ in the above three conditions, which is the original definition adopted in [21]. Since for all $v \in \mathbb{V}^k$, the exact solution $\tilde{u}$ satisfies

$$\int_{I_j} \tilde{u}_t(x, t) v(x) dx - \int_{I_j} f(\tilde{u}(x, t)) v_x(x) dx + f(\tilde{u}(x_{j+\frac{1}{2}}, t)) \cdot v_{j+\frac{1}{2}}^- - f(\tilde{u}(x_{j-\frac{1}{2}}, t)) \cdot v_{j-\frac{1}{2}}^+ = 0, \quad (6.6)$$

subtracting (6.3) from (6.6) yields the following equation for e :

$$\begin{aligned} & \int_{I_j} e_t(x, t) v(x) dx - \int_{I_j} (f(\tilde{u}(x, t)) - f(u(x, t))) v_x(x) dx + (f(\tilde{u}(x_{j+\frac{1}{2}}, t)) - \hat{f}_{j+\frac{1}{2}}(t)) \cdot v_{j+\frac{1}{2}}^- \\ & - (f(\tilde{u}(x_{j-\frac{1}{2}}, t)) - \hat{f}_{j-\frac{1}{2}}(t)) \cdot v_{j-\frac{1}{2}}^+ = D_j[u(\cdot, t); r, \epsilon_1, \epsilon_2] \int_{I_j} (u(x, t) - \bar{u}_j(t)) v(x) dx, \end{aligned} \quad (6.7)$$

which holds for all $v \in \mathbb{V}^k$. At any time $t \geq 0$, by taking $v(x) = \xi(x, t)$ in (6.7), we obtain

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \|\xi(\cdot, t)\|_{L^2(I_j)}^2 = \int_{I_j} \xi_t(x, t) \xi(x, t) dx \\ & = Z_j[\tilde{u}(\cdot, t), u(\cdot, t)] + D_j[u(\cdot, t); r, \epsilon_1, \epsilon_2] \int_{I_j} (u(x, t) - \bar{u}_j(t)) \xi(x, t) dx, \end{aligned} \quad (6.8)$$

where

$$Z_j[\tilde{u}, u] := \int_{I_j} \eta_t \xi dx + \int_{I_j} (f(\tilde{u}) - f(u)) \xi_x dx - (f(\tilde{u})_{j+\frac{1}{2}} - \hat{f}_{j+\frac{1}{2}}) \cdot \xi_{j+\frac{1}{2}}^- + (f(\tilde{u})_{j-\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}) \cdot \xi_{j-\frac{1}{2}}^+$$

is a functional acting on the functions $u, \tilde{u}$ at time t . It has been shown in [21] (see Lemmas 5.4, 5.5 and 5.7 therein) that

(1) if $\hat{f}$ is a general monotone flux,

$$\sum_j Z_j[\tilde{u}, u] \leq C \left(1 + \|\xi\|_{L^2(I)} + \frac{1}{h} \|e\|_{L^2(I)}^2 \right) \|\xi\|_{L^2(I)}^2 + C \left(1 + \frac{1}{h} \|e\|_{L^2(I)}^2 \right) h^{2k+1}; \quad (6.9)$$

(2) if $\hat{f}$ is an upwind flux,

$$\sum_j Z_j[\tilde{u}, u] \leq C \left(1 + \frac{1}{h} \|e\|_{L^2(I)} + \frac{1}{h^2} \|e\|_{L^2(I)}^2 \right) \|\xi\|_{L^2(I)}^2 + C \left(1 + \frac{1}{h^2} \|e\|_{L^2(I)}^2 \right) h^{2k+2}. \quad (6.10)$$

In (6.9) and (6.10), C depends on the uniform bound on $\tilde{u}$ and its derivatives in $I \times [0, T]$, but not on t, u and h .

In order to estimate the last term in (6.8), we need the following lemma:

Lemma 6.2. *Under the assumptions of Theorem 6.1, we have*

$$\sum_j D_j[u(\cdot, t); r, \epsilon_1, \epsilon_2] \int_{I_j} (u(x, t) - \bar{u}_j(t)) \xi(x, t) dx \leq C(\|u(\cdot, t)\|_{L^\infty(I)}) \left(\|\xi\|_{L^2(I)}^2 + h^{2k+2} \right), \quad (6.11)$$

where $C(\|u(\cdot, t)\|_{L^\infty(I)}) < \infty$ is a constant which depends on t only through the L^∞ -norm of u at time t ; it also depends on the uniform bound on $\tilde{u}$ and its derivatives in $I \times [0, T]$, but not on h .

Proof. Within the proof of this lemma, we adopt the notation $C_\star$ to denote any such constant $C(\|u(\cdot, t)\|_{L^\infty(I)})$ described in the statement above, and suppress the variable t as we used to, since we may assume t is fixed. We still use C to denote those constants which are independent of $\|u(\cdot, t)\|_{L^\infty(I)}$ and h , but still depends on the uniform bound on $\tilde{u}$ and its derivatives in $I \times [0, T]$. Substituting $u = \mathbb{Q}\tilde{u} - \xi$ and $\bar{u}_j = \overline{(\mathbb{Q}\tilde{u})}_j - \bar{\xi}_j$ on the left-hand side of (6.11), we get

$$\begin{aligned} & \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (u - \bar{u}_j) \xi dx \\ &= \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (\mathbb{Q}\tilde{u} - \overline{(\mathbb{Q}\tilde{u})}_j) \xi dx - \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (\xi - \bar{\xi}_j) \xi dx \\ &= \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (\mathbb{Q}\tilde{u} - \overline{(\mathbb{Q}\tilde{u})}_j) \xi dx - \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (\xi - \bar{\xi}_j)^2 dx \\ &\leq \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \int_{I_j} (\mathbb{Q}\tilde{u} - \overline{(\mathbb{Q}\tilde{u})}_j) \xi dx \\ &\leq \sum_j D_j[u; r, \epsilon_1, \epsilon_2] \|\mathbb{Q}\tilde{u} - \overline{(\mathbb{Q}\tilde{u})}_j\|_{L^2(I_j)} \|\xi\|_{L^2(I_j)} \\ &\leq C \sum_j h_j^{3/2} D_j[u; r, \epsilon_1, \epsilon_2] \|\xi\|_{L^2(I_j)}, \end{aligned} \quad (6.12)$$

where the last inequality follows from

$$\|\mathbb{Q}\tilde{u} - \overline{(\mathbb{Q}\tilde{u})}_j\|_{L^2(I_j)} \leq \|\mathbb{Q}\tilde{u} - \tilde{u}\|_{L^2(I_j)} + \|\tilde{u} - \overline{(\tilde{u})}_j\|_{L^2(I_j)} + \|\overline{(\tilde{u} - \mathbb{Q}\tilde{u})}_j\|_{L^2(I_j)} \leq Ch_j^{3/2}.$$

Now we proceed to estimate $D_j[u; r, \epsilon_1, \epsilon_2]$. By triangle inequality,

$$\begin{aligned} \|\partial_x^r f(u)\|_{L^2(I_j)} &\leq \|\partial_x^r f(u) - \partial_x^r f(\mathbb{Q}\tilde{u})\|_{L^2(I_j)} \\ &\quad + \|\partial_x^r f(\mathbb{Q}\tilde{u}) - \partial_x^r f(\tilde{u})\|_{L^2(I_j)} \\ &\quad + \|\partial_x^r f(\tilde{u})\|_{L^2(I_j)}. \end{aligned} \quad (6.13)$$

Set $a_i = \partial_x^i u / i!$, $b_i = \partial_x^i (\mathbb{Q}\tilde{u}) / i!$ and $c_i = \partial_x^i \tilde{u} / i!$. By Faà di Bruno's formula [7],

$$\partial_x^r f(u) = \sum_{\substack{m_1, \dots, m_r \geq 0 \\ m_1 + 2m_2 + \dots + rm_r = r}} \frac{r!}{m_1! \cdots m_r!} f^{(m_1 + \dots + m_r)}(u) \prod_{i=1}^r (a_i)^{m_i}, \quad (6.14)$$

so that

$$\begin{aligned}
 & \partial_x^r f(u) - \partial_x^r f(\mathbb{Q}\tilde{u}) \\
 = & \sum_{\substack{m_1, \dots, m_r \geq 0 \\ m_1 + 2m_2 + \dots + rm_r = r}} \frac{r!}{m_1! \dots m_r!} \left(f^{(m_1 + \dots + m_r)}(u) - f^{(m_1 + \dots + m_r)}(\mathbb{Q}\tilde{u}) \right) \prod_{i=1}^r (a_i)^{m_i} \\
 & + \sum_{\substack{m_1, \dots, m_r \geq 0 \\ m_1 + 2m_2 + \dots + rm_r = r}} \frac{r!}{m_1! \dots m_r!} f^{(m_1 + \dots + m_r)}(\mathbb{Q}\tilde{u}) \left(\prod_{i=1}^r (a_i)^{m_i} - \prod_{i=1}^r (b_i)^{m_i} \right). \quad (6.15)
 \end{aligned}$$

Note that

$$\begin{aligned}
 \prod_{i=1}^r (a_i)^{m_i} - \prod_{i=1}^r (b_i)^{m_i} &= \sum_{i=1}^r ((a_i)^{m_i} - (b_i)^{m_i}) \prod_{l=1}^{i-1} (b_l)^{m_l} \prod_{l=i+1}^r (a_l)^{m_l} \\
 &= \sum_{i=1}^r (a_i - b_i) \left(\sum_{l=0}^{m_i-1} a_i^l b_i^{m_i-1-l} \right) \prod_{l=1}^{i-1} (b_l)^{m_l} \prod_{l=i+1}^r (a_l)^{m_l}, \quad (6.16)
 \end{aligned}$$

and also

$$\begin{aligned}
 \|a_i - b_i\|_{L^2(I_j)} &\leq \frac{C}{h_j^i} \|\xi\|_{L^2(I_j)}, \\
 \|a_i\|_{L^\infty(I_j)} &\leq \frac{C}{h_j^i} \|u\|_{L^\infty(I_j)} \leq \frac{C_\star}{h_j^i}, \\
 \|b_i\|_{L^\infty(I_j)} &\leq \frac{C}{h_j^i} \|\mathbb{Q}\tilde{u}\|_{L^\infty(I_j)} \leq \frac{C}{h_j^i}.
 \end{aligned}$$

Therefore, the exponent of $\frac{1}{h_j}$ in (6.16) is

$$\begin{aligned}
 & i + i(m_i - 1) + (m_1 + 2m_2 + \dots + (i-1)m_{i-1}) + ((i+1)m_{i+1} + \dots + rm_r) \\
 &= m_1 + 2m_2 + \dots + rm_r = r,
 \end{aligned}$$

which implies

$$\left\| \prod_{i=1}^r (a_i)^{m_i} - \prod_{i=1}^r (b_i)^{m_i} \right\|_{L^2(I_j)} \leq \frac{C_\star}{h_j^r} \|\xi\|_{L^2(I_j)}.$$

Also, by mean value theorem and the L^∞ -bound on u and $\mathbb{Q}\tilde{u}$,

$$\|f^{(m_1 + \dots + m_r)}(u) - f^{(m_1 + \dots + m_r)}(\mathbb{Q}\tilde{u})\|_{L^2(I_j)} \leq C_\star \|\xi\|_{L^2(I_j)}.$$

Hence, by (6.15) and the two inequalities above,

$$\|\partial_x^r f(u) - \partial_x^r f(\mathbb{Q}\tilde{u})\|_{L^2(I_j)} \leq \frac{C_\star}{h_j^r} \|\xi\|_{L^2(I_j)}. \quad (6.17)$$

To estimate $\|\partial_x^r f(\mathbb{Q}\tilde{u}) - \partial_x^r f(\tilde{u})\|_{L^2(I_j)}$, we proceed similarly. First, we compute

$$\begin{aligned} & \partial_x^r f(\mathbb{Q}\tilde{u}) - \partial_x^r f(\tilde{u}) \\ = & \sum_{\substack{m_1, \dots, m_r \geq 0 \\ m_1 + 2m_2 + \dots + rm_r = r}} \frac{r!}{m_1! \dots m_r!} \left(f^{(m_1 + \dots + m_r)}(\mathbb{Q}\tilde{u}) - f^{(m_1 + \dots + m_r)}(\tilde{u}) \right) \prod_{i=1}^r (c_i)^{m_i} \\ & + \sum_{\substack{m_1, \dots, m_r \geq 0 \\ m_1 + 2m_2 + \dots + rm_r = r}} \frac{r!}{m_1! \dots m_r!} f^{(m_1 + \dots + m_r)}(\mathbb{Q}\tilde{u}) \left(\prod_{i=1}^r (b_i)^{m_i} - \prod_{i=1}^r (c_i)^{m_i} \right). \end{aligned} \quad (6.18)$$

Next, we have

$$\prod_{i=1}^r (b_i)^{m_i} - \prod_{i=1}^r (c_i)^{m_i} = \sum_{i=1}^r (b_i - c_i) \left(\sum_{l=0}^{m_i-1} b_i^l c_i^{m_i-1-l} \right) \prod_{l=1}^{i-1} (c_l)^{m_l} \prod_{l=i+1}^r (b_l)^{m_l}, \quad (6.19)$$

It follows from Bramble–Hilbert lemma that

$$\begin{aligned} \|b_i - c_i\|_{L^2(I_j)} &\leq Ch_j^{k+1-i} |\tilde{u}|_{H^{k+1}(I_j)} \leq Ch_j^{k+\frac{3}{2}-i}, \\ \|b_i - c_i\|_{L^\infty(I_j)} &\leq Ch_j^{k+1-i} \|\partial_x^{k+1} \tilde{u}\|_{L^\infty(I_j)} \leq Ch_j^{k+1-i}, \\ \|c_i\|_{L^\infty(I_j)} &\leq C, \\ \|b_i\|_{L^\infty(I_j)} &\leq \|b_i - c_i\|_{L^\infty(I_j)} + \|c_i\|_{L^\infty(I_j)} \leq Ch_j^{\min\{0, k+1-i\}}. \end{aligned}$$

Recall that $i \leq r \leq k+2$ and $m_r \leq 1$. Therefore,

$$\left\| \prod_{i=1}^r (b_i)^{m_i} - \prod_{i=1}^r (c_i)^{m_i} \right\|_{L^2(I_j)} \leq Ch_j^{k+\frac{3}{2}-r}.$$

By mean value theorem and the L^∞ -bound on u and $\mathbb{Q}\tilde{u}$,

$$\|f^{(m_1 + \dots + m_r)}(\mathbb{Q}\tilde{u}) - f^{(m_1 + \dots + m_r)}(\tilde{u})\|_{L^2(I_j)} \leq C \|\eta\|_{L^2(I_j)}.$$

Hence, by (6.18) and the two inequalities above,

$$\|\partial_x^r f(\mathbb{Q}\tilde{u}) - \partial_x^r f(\tilde{u})\|_{L^2(I_j)} \leq C \|\eta\|_{L^2(I_j)} + Ch_j^{k+\frac{3}{2}-r}. \quad (6.20)$$

Finally, since $\tilde{u}$ is smooth, we have the trivial bound

$$\|\partial_x^r f(\tilde{u})\|_{L^2(I_j)} \leq Ch_j^{\frac{1}{2}}. \quad (6.21)$$

Combining (6.13), (6.17), (6.20) and (6.21), we get

$$\begin{aligned} h_j^r \|\partial_x^r f(u)\|_{L^2(I_j)} &\leq C_\star \|\xi\|_{L^2(I_j)} + h_j^r C \|\eta\|_{L^2(I_j)} + Ch_j^{\min\{k+\frac{3}{2}, r+\frac{1}{2}\}} \\ &\leq C_\star \|\xi\|_{L^2(I_j)} + h_j^r C \|\eta\|_{L^2(I_j)} + Ch_j^{k+\frac{3}{2}}, \end{aligned}$$

where the last inequality follows from $k+1 \leq r \leq k+2$. This takes care of the first term in (6.4). The second term in (3.48) can be easily bounded by

$$|\llbracket u \rrbracket|_{j-\frac{1}{2}} + |\llbracket u \rrbracket|_{j+\frac{1}{2}} \leq |\llbracket \xi \rrbracket|_{j-\frac{1}{2}} + |\llbracket \eta \rrbracket|_{j-\frac{1}{2}} + |\llbracket \xi \rrbracket|_{j+\frac{1}{2}} + |\llbracket \eta \rrbracket|_{j+\frac{1}{2}}$$

since the jumps of $\tilde{u}$ at cell interfaces are all zero. Therefore, taking into account $\|\eta\|_{L^2(I)} \leq Ch^{k+1}$, (6.12) can be estimated by

$$\begin{aligned} & C \sum_j h_j^{3/2} D_j[u; r, \epsilon_1, \epsilon_2] \|\xi\|_{L^2(I_j)} \\ & \leq C \sum_j h_j^r \|\partial_x^r f(u)\|_{L^2(I_j)} \|\xi\|_{L^2(I_j)} + C\sqrt{h} \sum_j \left(|\llbracket u \rrbracket|_{j-\frac{1}{2}} + |\llbracket u \rrbracket|_{j+\frac{1}{2}} \right) \|\xi\|_{L^2(I_j)} \\ & \leq C_\star \|\xi\|_{L^2(I)}^2 + Ch^r \|\eta\|_{L^2(I)}^2 + Ch^{2k+2} + C\sqrt{h} \sqrt{\sum_j \|\xi\|_{j+\frac{1}{2}}^2 + \sum_j \|\eta\|_{j+\frac{1}{2}}^2} \|\xi\|_{L^2(I)} \\ & \leq C_\star \|\xi\|_{L^2(I)}^2 + Ch^{2k+2} + C (\|\xi\|_{L^2(I)} + \|\eta\|_{L^2(I)}) \|\xi\|_{L^2(I)} \\ & \leq C_\star \|\xi\|_{L^2(I)}^2 + Ch^{2k+2} \end{aligned}$$

and the lemma is proved. $\blacksquare$

Now, we begin to prove (6.1) and (6.2) using a bootstrap argument. We first suppose

$$\|e(\cdot, t)\|_{L^2(I)} \leq h \quad \text{for all } t \in [0, T]. \quad (\text{Bootstrap Hypothesis})$$

Under this hypothesis, $\|u(\cdot, t)\|_{L^\infty(I)} \leq C$ for all $t \in [0, T]$. By Lemma 6.2,

$$\sum_j D_j[u(\cdot, t); r, \epsilon_1, \epsilon_2] \int_{I_j} (u(x, t) - \bar{u}_j(t)) \xi(x, t) dx \leq C \|\xi\|_{L^2(I)}^2 + Ch^{2k+2}, \quad (6.22)$$

where the constant C no longer depends on t . Moreover, it follows from (6.9) and (6.10) that

(1) if $\hat{f}$ is a general monotone flux,

$$\sum_j Z_j[\tilde{u}, u] \leq C \|\xi\|_{L^2(I)}^2 + Ch^{2k+1}.$$

Therefore, by (6.8) and (6.22),

$$\frac{1}{2} \frac{d}{dt} \|\xi(\cdot, t)\|_{L^2(I)}^2 \leq C \|\xi\|_{L^2(I)}^2 + Ch^{2k+1}.$$

Applying Gronwall's inequality, we obtain

$$\|\xi(\cdot, t)\|_{L^2(I)}^2 \leq \|\xi(\cdot, 0)\|_{L^2(I)}^2 + Ch^{2k+1} \leq Ch^{2k+1}, \quad \forall t \in [0, T],$$

so that

$$\|e(\cdot, t)\|_{L^2(I)} \leq \|\xi(\cdot, t)\|_{L^2(I)} + \|\eta(\cdot, t)\|_{L^2(I)} \leq Ch^{k+\frac{1}{2}}, \quad \forall t \in [0, T].$$

Now, we can take $h > 0$ sufficiently small so that $Ch^{k+\frac{1}{2}} \leq h$, which completes the bootstrap step and the proof.

(2) If $\hat{f}$ is an upwind flux,

$$\sum_j Z_j[\tilde{u}, u] \leq C \|\xi\|_{L^2(I)}^2 + Ch^{2k+2}.$$

We proceed as the monotone case and obtain

$$\|\xi(\cdot, t)\|_{L^2(I)}^2 \leq \|\xi(\cdot, 0)\|_{L^2(I)}^2 + Ch^{2k+2} \leq Ch^{2k+2}, \quad \forall t \in [0, T],$$

and

$$\|e(\cdot, t)\|_{L^2(I)} \leq \|\xi(\cdot, t)\|_{L^2(I)} + \|\eta(\cdot, t)\|_{L^2(I)} \leq Ch^{k+1}, \quad \forall t \in [0, T].$$

Now, we can take $h > 0$ sufficiently small so that $Ch^{k+1} \leq h$, which completes the bootstrap step and the proof.

Now, we have proved Theorem 6.1 for (Scheme D). Since u is uniformly bounded for all $t \in [0, T]$ under (Bootstrap Hypothesis), we can find ϵ_1, ϵ_2 such that $D_j[u; r, \epsilon_1, \epsilon_2] \geq B_j[u; U]$ and $D_j[u; r, \epsilon_1, \epsilon_2] \geq C_j[u; U, r]$ (recall this is how we constructed D_j so that we could use monotonicity of remainder). Thus, our estimate for $D_j[u; r, \epsilon_1, \epsilon_2]$ in Lemma 6.2 can be directly applied to $B_j[u; U]$ and $C_j[u; U, r]$, and in the same way we can prove Theorem 6.1 for (Scheme B) and (Scheme C). ■

6.1. Numerical tests for optimal order of convergence

Example 6.3. We first test the optimal order of convergence proved in Theorem 6.1 for (Scheme D) with polynomial degrees $k = 1, 2, 3$ and $k + 1 \leq r \leq k + 2$. Consider the non-linear Burgers equation

$$u_t + \left(\frac{u^2}{2} \right)_x = 0$$

with the initial condition $u_0(x) = \sin(x) + 0.5$. The computational domain is $I = [0, 2\pi]$ with periodic boundary conditions. This is the same example from [13]. The final time is taken to be $T = 0.6$, so that the exact solution remains smooth. We use the classic fourth order Runge–Kutta method for time discretizations. The CFL condition is $\Delta t = O(h)$. In all tests, we adopt the local Lax–Friedrichs numerical flux and set $\epsilon_1 = \epsilon_2 = 1$.

We test the accuracy on a series of uniform meshes, with the total number of cells equal to $N = 16, 32, \dots, 512$. The L^1 , L^2 and L^∞ errors and their orders of convergence for the cases $r = k + 1$ and $r = k + 2$ are shown in Tables 3 and 4, respectively. We can see that the numerical results matches with the statement of Theorem 6.1.

	N	L^1 error	Order	L^2 error	Order	L^∞ error	Order
$\mathbb{P}^1$	16	5.858936×10^{-2}	—	3.913706×10^{-2}	—	9.937107×10^{-2}	—
	32	1.375636×10^{-2}	2.0905	9.450200×10^{-3}	2.0501	2.300587×10^{-2}	2.1108
	64	3.271946×10^{-3}	2.0719	2.399096×10^{-3}	1.9779	7.577038×10^{-3}	1.6023
	128	7.845710×10^{-4}	2.0602	5.862501×10^{-4}	2.0329	2.055943×10^{-3}	1.8818
	256	1.921616×10^{-4}	2.0296	1.471028×10^{-4}	1.9947	5.553945×10^{-4}	1.8882
	512	4.746975×10^{-5}	2.0172	3.662547×10^{-5}	2.0059	1.346929×10^{-4}	2.0438
$\mathbb{P}^2$	16	5.579540×10^{-2}	—	6.537560×10^{-2}	—	2.125512×10^{-1}	—
	32	6.356686×10^{-3}	3.1338	9.017231×10^{-3}	2.8580	3.935147×10^{-2}	2.4333
	64	4.968015×10^{-4}	3.6775	8.178954×10^{-4}	3.4627	6.172782×10^{-3}	2.6724
	128	3.620624×10^{-5}	3.7784	5.919958×10^{-5}	3.7883	4.116191×10^{-4}	3.9065
	256	2.977955×10^{-6}	3.6038	5.265680×10^{-6}	3.4909	6.562058×10^{-5}	2.6491
	512	2.648582×10^{-7}	3.4910	4.668005×10^{-7}	3.4957	6.501662×10^{-6}	3.3353
$\mathbb{P}^3$	16	3.098316×10^{-1}	—	2.214555×10^{-1}	—	3.821968×10^{-1}	—
	32	6.271633×10^{-2}	2.3046	8.178122×10^{-2}	1.4372	1.780029×10^{-1}	1.1024
	64	1.298156×10^{-3}	5.5943	3.586751×10^{-3}	4.5110	1.591189×10^{-2}	3.4837
	128	1.287967×10^{-5}	6.6552	2.873509×10^{-5}	6.9637	1.597338×10^{-4}	6.6383
	256	2.857467×10^{-7}	5.4942	4.912621×10^{-7}	5.8702	2.823307×10^{-6}	5.8221
	512	9.106452×10^{-9}	4.9717	1.517097×10^{-8}	5.0171	9.072239×10^{-8}	4.9598

Table 3. Errors and convergence orders for (Scheme D) with polynomial degrees $k = 1, 2, 3$ and $r = k + 1$.

	N	L^1 error	Order	L^2 error	Order	L^∞ error	Order
$\mathbb{P}^1$	16	4.294674×10^{-2}	—	2.754407×10^{-2}	—	5.637620×10^{-2}	—
	32	1.099892×10^{-2}	1.9652	7.523647×10^{-3}	1.8722	2.174760×10^{-2}	1.3742
	64	2.785689×10^{-3}	1.9813	2.031890×10^{-3}	1.8886	6.226255×10^{-3}	1.8044
	128	7.101160×10^{-4}	1.9719	5.367464×10^{-4}	1.9205	1.692266×10^{-3}	1.8794
	256	1.802998×10^{-4}	1.9777	1.393244×10^{-4}	1.9458	4.389260×10^{-4}	1.9469
	512	4.559686×10^{-5}	1.9834	3.566824×10^{-5}	1.9657	1.115395×10^{-4}	1.9764
$\mathbb{P}^2$	16	1.854099×10^{-2}	—	1.849425×10^{-2}	—	6.238941×10^{-2}	—
	32	1.645980×10^{-3}	3.4937	1.991278×10^{-3}	3.2153	6.517240×10^{-3}	3.2590
	64	1.080701×10^{-4}	3.9289	1.303109×10^{-4}	3.9337	5.678523×10^{-4}	3.5207
	128	9.718192×10^{-6}	3.4751	1.100091×10^{-5}	3.5663	6.984597×10^{-5}	3.0233
	256	1.091877×10^{-6}	3.1539	1.259977×10^{-6}	3.1262	9.076374×10^{-6}	2.9440
	512	1.322708×10^{-7}	3.0452	1.559972×10^{-7}	3.0138	1.170951×10^{-6}	2.9544
$\mathbb{P}^3$	16	2.167622×10^{-1}	—	2.113418×10^{-1}	—	4.010803×10^{-1}	—
	32	5.608432×10^{-2}	1.9504	7.943122×10^{-2}	1.4118	1.981113×10^{-1}	1.0176
	64	6.803013×10^{-5}	9.6872	9.882538×10^{-5}	9.6506	3.113658×10^{-4}	9.3135
	128	1.058955×10^{-6}	6.0055	1.542502×10^{-6}	6.0015	8.865795×10^{-6}	5.1342
	256	2.144179×10^{-8}	5.6261	3.149298×10^{-8}	5.6141	2.284965×10^{-7}	5.2780
	512	6.680798×10^{-10}	5.0043	1.000724×10^{-9}	4.9759	8.485364×10^{-9}	4.7511

Table 4. Errors and convergence orders for (Scheme D) with polynomial degrees $k = 1, 2, 3$ and $r = k + 2$.

Example 6.4. Although Theorem 6.1 only covers 1D scalar conservation laws, we nevertheless proceed to the case of 2D scalar conservation laws. Consider the nonlinear 2D Burgers equation

$$u_t + \left(u^2\right)_x + \left(u^2\right)_y = 0$$

with the initial condition $u_0(x, y) = 0.5 \sin(2\pi(x + y))$. We test the optimal order of convergence for (Scheme D'') with polynomial degrees $k = 1, 2, 3$ and $k + 1 \leq r \leq k + 2$. The computational domain is $I = [0, 1] \times [0, 1]$ with periodic boundary conditions. This is the same example from [1]. The final time is taken to be $T = 0.05$, so that the exact solution remains smooth. We use the classic fourth order Runge–Kutta method for time discretizations. The CFL condition is $\Delta t = O(h)$. In all tests, we adopt the local Lax–Friedrichs numerical flux and set $\epsilon_1 = \epsilon_2 = 1$.

We test the accuracy on a series of uniform rectangular meshes with $N_x = N_y = 16, 32, \dots, 512$. The L^1 , L^2 and L^∞ errors and their orders of convergence for the cases $r = k + 1$ and $r = k + 2$ are shown in Tables 5 and 6, respectively. The convergence orders are indeed optimal. Note that the L^1 and L^2 errors are evaluated using a tensor-product 10-point Gauss–Legendre rule on every rectangular cell, while the L^∞ error uses a separate uniform point set that includes cell boundaries; super-convergence in L^1 and L^2 errors can be observed for $\mathbb{P}^3$ elements.

	N	L^1 error	Order	L^2 error	Order	L^∞ error	Order
$\mathbb{P}^1$	16	8.873109×10^{-3}	–	1.516498×10^{-2}	–	1.535248×10^{-1}	–
	32	2.219736×10^{-3}	1.999	4.294241×10^{-3}	1.820	5.490943×10^{-2}	1.483
	64	5.276571×10^{-4}	2.073	1.079269×10^{-3}	1.992	1.371005×10^{-2}	2.002
	128	1.250893×10^{-4}	2.077	2.612186×10^{-4}	2.047	2.932631×10^{-3}	2.225
	256	3.006254×10^{-5}	2.057	6.402969×10^{-5}	2.028	8.263882×10^{-4}	1.827
	512	7.332548×10^{-6}	2.036	1.593880×10^{-5}	2.006	2.199474×10^{-4}	1.910
$\mathbb{P}^2$	16	3.851715×10^{-3}	–	8.276021×10^{-3}	–	1.020042×10^{-1}	–
	32	5.841907×10^{-4}	2.721	1.750315×10^{-3}	2.241	2.323799×10^{-2}	2.134
	64	6.805218×10^{-5}	3.102	2.180787×10^{-4}	3.005	3.479090×10^{-3}	2.740
	128	6.682944×10^{-6}	3.348	2.203113×10^{-5}	3.307	4.685234×10^{-4}	2.893
	256	6.231348×10^{-7}	3.423	2.062058×10^{-6}	3.417	6.001436×10^{-5}	2.965
	512	6.120003×10^{-8}	3.348	2.041695×10^{-7}	3.336	7.359222×10^{-6}	3.028
$\mathbb{P}^3$	16	4.058691×10^{-2}	–	6.319777×10^{-2}	–	3.127916×10^{-1}	–
	32	1.042123×10^{-2}	1.961	2.798347×10^{-2}	1.175	1.871199×10^{-1}	0.741
	64	2.347751×10^{-3}	2.150	1.071306×10^{-2}	1.385	9.616082×10^{-2}	0.960
	128	7.096482×10^{-5}	5.048	8.266929×10^{-4}	3.696	1.557553×10^{-2}	2.626
	256	3.059249×10^{-7}	7.858	2.532318×10^{-6}	8.351	6.318595×10^{-5}	7.945
	512	1.004042×10^{-8}	4.929	6.985113×10^{-8}	5.180	2.634123×10^{-6}	4.584

Table 5. Errors and convergence orders for (Scheme D'') with polynomial degrees $k = 1, 2, 3$ and $r = k + 1$.

	N	L^1 error	Order	L^2 error	Order	L^∞ error	Order
$\mathbb{P}^1$	16	6.745922×10^{-3}	—	1.206151×10^{-2}	—	1.380574×10^{-1}	—
	32	1.762432×10^{-3}	1.936	3.512284×10^{-3}	1.780	4.127613×10^{-2}	1.742
	64	4.419675×10^{-4}	1.996	9.224084×10^{-4}	1.929	1.129637×10^{-2}	1.869
	128	1.108997×10^{-4}	1.995	2.388377×10^{-4}	1.949	3.236580×10^{-3}	1.803
	256	2.788790×10^{-5}	1.992	6.146162×10^{-5}	1.958	8.705389×10^{-4}	1.894
	512	7.020283×10^{-6}	1.990	1.567670×10^{-5}	1.971	2.256230×10^{-4}	1.948
$\mathbb{P}^2$	16	3.458565×10^{-3}	—	8.039311×10^{-3}	—	9.947845×10^{-2}	—
	32	3.873141×10^{-4}	3.159	1.131736×10^{-3}	2.829	1.898286×10^{-2}	2.390
	64	3.181137×10^{-5}	3.606	9.140592×10^{-5}	3.630	2.517905×10^{-3}	2.914
	128	3.184988×10^{-6}	3.320	9.412806×10^{-6}	3.280	3.038137×10^{-4}	3.051
	256	3.869309×10^{-7}	3.041	1.192756×10^{-6}	2.980	3.953387×10^{-5}	2.942
	512	4.895042×10^{-8}	2.983	1.557787×10^{-7}	2.937	5.673613×10^{-6}	2.801
$\mathbb{P}^3$	16	4.589640×10^{-2}	—	6.740899×10^{-2}	—	3.427352×10^{-1}	—
	32	9.132882×10^{-3}	2.329	2.828390×10^{-2}	1.253	2.050638×10^{-1}	0.741
	64	3.319369×10^{-5}	8.104	1.220094×10^{-4}	7.857	1.763706×10^{-3}	6.861
	128	8.950958×10^{-7}	5.213	3.976124×10^{-6}	4.939	6.332885×10^{-5}	4.800
	256	2.294565×10^{-8}	5.286	1.062076×10^{-7}	5.226	1.641234×10^{-6}	5.270
	512	7.024232×10^{-10}	5.030	2.994068×10^{-9}	5.149	1.017521×10^{-7}	4.012

Table 6. Errors and convergence orders for (Scheme D'') with polynomial degrees $k = 1, 2, 3$ and $r = k + 2$.

7. Strong convergence through compensated compactness for general strictly convex conservation laws

In this section we restrict attention to the scalar equation (2.1) with a strictly convex flux ($f'' > 0$ on $\mathbb{R}$). Let $h := \max_j h_j$ as in Section 6, but denote the numerical solution by u_h (instead of u) to emphasize its role as a sequence parameterized by h . We show in Theorem 7.4 that, for any initial data $\mathcal{U}_0 \in L^\infty(I)$ with $u_h(\cdot, 0) \rightarrow \mathcal{U}_0$ strongly in $L^2(I)$, the sequence of numerical solutions $\{u_h\}$ generated by (Scheme D) with $\epsilon_1 > 0$ converges strongly to the unique entropy solution as $h \rightarrow 0$ if $\{u_h\}$ is uniformly bounded in L^∞ . We prove it using the compensated compactness argument [5, 14, 18].

We make the following assumptions. The meshes are quasi-uniform, $k \geq 1$ is fixed, and u_h denotes a solution of (Scheme D) subject to periodic or compactly supported boundary conditions. For convenience, we only consider the case

$$r = k + 1, \quad \epsilon_1 > 0. \quad (7.1)$$

The same argument applies to $r = k + 2$, but $\epsilon_1 > 0$ must be satisfied. We assume that, for every $T > 0$,

$$\|u_h\|_{L^\infty(I \times (0, T))} \leq M_T \quad (7.2)$$

uniformly in h . We also assume that $f \in C^{k+2}$ and

$$f''(s) \geq c_f > 0, \quad |s| \leq M_T. \quad (7.3)$$

Note that c_f might depend on M_T if $f(\cdot)$ is not uniformly convex. Finally, $\hat{f}$ is assumed to be a consistent locally Lipschitz E-flux whose square-entropy dissipation is quantitatively coercive: if

$$\eta(s) := \frac{s^2}{2}, \quad q'(s) := s f'(s), \quad (7.4)$$

denotes the square-entropy pair⁵, then its interface dissipation $\mathcal{D}_\eta$ satisfies

$$\mathcal{D}_\eta(a, b) := \int_a^b (f(s) - \hat{f}(a, b)) \, ds. \quad (7.5)$$

This is (3.36) specialized to the square-entropy U , for which $U'' \equiv 1$. We assume that

$$\mathcal{D}_\eta(a, b) \geq c_{\hat{f}} |a - b|^3, \quad |a|, |b| \leq M_T. \quad (7.6)$$

This is crucial for deriving the weak-BV estimate (7.15) below. The following lemma shows that every E-flux (see Definition 3.8) satisfies (7.6).

Lemma 7.1. *Assume that (7.3) holds. Then every E-flux satisfies (7.6); in particular, one may take*

$$c_{\hat{f}} = \frac{c_f}{24}.$$

Proof. Let $a, b \in [-M_T, M_T]$. We prove the estimate with

$$c_{\hat{f}} := \frac{c_f}{24}, \quad (7.7)$$

which is independent of a, b , and of the particular E-flux. The case $a = b$ is immediate, so it remains to consider the two possible orientations of the interval.

Suppose first that $a < b$. Definition 3.8 gives

$$\hat{f}(a, b) \leq f(s) \quad \text{for every } s \in [a, b].$$

Consequently,

$$\hat{f}(a, b) \leq \min_{s \in [a, b]} f(s). \quad (7.8)$$

Let $m \in [a, b]$ be a point at which this minimum is attained. We claim that

$$f(s) - f(m) \geq \frac{c_f}{2} |s - m|^2, \quad s \in [a, b]. \quad (7.9)$$

Indeed, uniform convexity gives

$$f(s) \geq f(m) + f'(m)(s - m) + \frac{c_f}{2} |s - m|^2.$$

⁵Here we adopt the notation (η, q) instead of $(U, \mathcal{F})$ as the entropy-entropy flux pair to avoid confusion with the prescribed entropy U in the definition of (Scheme B) and (Scheme C) from the previous sections.

If $m \in (a, b)$, then $f'(m) = 0$. If $m = a$, then $f'(m) \geq 0$ and $s - m \geq 0$; if $m = b$, then $f'(m) \leq 0$ and $s - m \leq 0$. Thus in every case $f'(m)(s - m) \geq 0$, which proves (7.9). Using (7.5), (7.8), and (7.9), we obtain

$$\begin{aligned} \mathcal{D}_\eta(a, b) &= \int_a^b (f(s) - \hat{f}(a, b)) \, ds \\ &\geq \int_a^b (f(s) - f(m)) \, ds \\ &\geq \frac{c_f}{2} \int_a^b |s - m|^2 \, ds \\ &= \frac{c_f}{6} \left((m - a)^3 + (b - m)^3 \right). \end{aligned}$$

For $x, y \geq 0$, convexity of $z \mapsto z^3$ gives

$$x^3 + y^3 \geq 2 \left(\frac{x + y}{2} \right)^3 = \frac{(x + y)^3}{4}.$$

Taking $x = m - a$ and $y = b - m$ therefore yields

$$\mathcal{D}_\eta(a, b) \geq \frac{c_f}{24} (b - a)^3. \quad (7.10)$$

Suppose next that $a > b$. In this case Definition 3.8 gives

$$\hat{f}(a, b) \geq f(s) \quad \text{for every } s \in [b, a],$$

and hence

$$\hat{f}(a, b) \geq \max\{f(a), f(b)\}. \quad (7.11)$$

Let ℓ be the affine function joining the endpoint values of f on $[b, a]$:

$$\ell(s) := \frac{a - s}{a - b} f(b) + \frac{s - b}{a - b} f(a).$$

Since $\ell(s)$ is a convex combination of $f(a)$ and $f(b)$,

$$\ell(s) \leq \max\{f(a), f(b)\}. \quad (7.12)$$

Furthermore, $f'' \geq c_f$ implies

$$\ell(s) - f(s) \geq \frac{c_f}{2} (s - b)(a - s), \quad s \in [b, a]. \quad (7.13)$$

For completeness, this follows by applying the ordinary convexity inequality to the convex function $g(s) := f(s) - c_f s^2/2$: writing $s = (1 - \theta)b + \theta a$, where $\theta = (s - b)/(a - b)$, gives

$$g(s) \leq (1 - \theta)g(b) + \theta g(a),$$

and rearrangement is exactly (7.13).

Because the integral in (7.5) is now oriented from a to b , equations (7.11)–(7.13) give

$$\begin{aligned}\mathcal{D}_\eta(a, b) &= \int_b^a (\hat{f}(a, b) - f(s)) \, ds \\ &\geq \int_b^a (\ell(s) - f(s)) \, ds \\ &\geq \frac{c_f}{2} \int_b^a (s - b)(a - s) \, ds \\ &= \frac{c_f}{12} (a - b)^3 \\ &\geq \frac{c_f}{24} |a - b|^3.\end{aligned}$$

Together with (7.10), this proves

$$\mathcal{D}_\eta(a, b) \geq c_{\hat{f}} |a - b|^3, \quad |a|, |b| \leq M_T,$$

with $c_{\hat{f}} = c_f/24$, and hence proves (7.6) for every E-flux. $\blacksquare$

In what follows, we denote by C any constant that may depend on k, T but independent of h . We first record the estimates used in the proof; they cannot be obtained without the first term in the definition of D_j (3.48), which is why we need to assume $\epsilon_1 > 0$ in (7.1).

Lemma 7.2 (Entropy dissipation and consistency). *Under the preceding assumptions, the numerical solutions satisfy*

$$\int_0^T \sum_j D_j[u_h; k+1, \epsilon_1, \epsilon_2] \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2 \, dt \leq C, \quad (7.14)$$

$$\int_0^T \sum_j |[[u_h]]_{j+\frac{1}{2}}|^3 \, dt \leq C, \quad (7.15)$$

$$\int_0^T \sum_j D_j[u_h; k+1, \epsilon_1, \epsilon_2] (h_j)^3 \, dt \leq Ch. \quad (7.16)$$

Moreover, with the piecewise smooth function ρ defined on I by

$$\rho(x) := f(u_h)_x(x) - \pi[f(u_h)_x](x) \quad \forall x \in I_j, \quad (7.17)$$

one has

$$\|\rho\|_{L^2(I_j)} \leq C(h_j)^{1/2} D_j. \quad (7.18)$$

Here and below $D_j = D_j[u_h; k+1, \epsilon_1, \epsilon_2]$.

Proof. Assume $I = [x_L, x_R]$. Taking $v = u_h$ in (Scheme D), proceeding as in [9] or in the proof of Theorem 3.9 and using

$$\int_{I_j} (u_h - \bar{u}_{h,j}) u_h \, dx = \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2,$$

we obtain the following local square-entropy inequality:

$$\frac{d}{dt} \int_{I_j} \eta(u_h) dx + \hat{q}_{j+\frac{1}{2}}(t) - \hat{q}_{j-\frac{1}{2}}(t) + \frac{1}{2} \left(\mathcal{D}_{\eta,j-\frac{1}{2}} + \mathcal{D}_{\eta,j+\frac{1}{2}} \right) + D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2 \leq 0,$$

where $\mathcal{D}_{\eta,j+\frac{1}{2}} = \mathcal{D}_{\eta}(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+)$ and

$$\hat{q}_{j+\frac{1}{2}}(t) = \frac{1}{2} (u_{j+\frac{1}{2}}^- + u_{j+\frac{1}{2}}^+) \hat{f}_{j+\frac{1}{2}} - \frac{1}{2} \left(\int_{j+\frac{1}{2}}^{u_{j+\frac{1}{2}}^-} f(\xi) d\xi + \int_{j+\frac{1}{2}}^{u_{j+\frac{1}{2}}^+} f(\xi) d\xi \right).$$

Summing it over j yields the global square-entropy inequality

$$\frac{d}{dt} \int_I \eta(u_h) dx + \hat{q}_R(t) - \hat{q}_L(t) + \sum_j \frac{1}{2} \left(\mathcal{D}_{\eta,j-\frac{1}{2}} + \mathcal{D}_{\eta,j+\frac{1}{2}} \right) + \sum_j D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2 \leq 0.$$

Integrate in time and apply the assumption $\|u_h\|_{L^\infty(I \times (0,T))} \leq M_T$,

$$\begin{aligned} \int_0^T \sum_j \frac{1}{2} \left(\mathcal{D}_{\eta,j-\frac{1}{2}} + \mathcal{D}_{\eta,j+\frac{1}{2}} \right) dt + \int_0^T \sum_j D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2 dt \\ \leq \int_I \eta(u_h(x, 0)) dx + \int_0^T (\hat{q}_L(t) - \hat{q}_R(t)) dt \leq C. \end{aligned} \quad (7.19)$$

By (7.6),

$$\int_0^T \sum_j \frac{1}{2} \left(\mathcal{D}_{\eta,j-\frac{1}{2}} + \mathcal{D}_{\eta,j+\frac{1}{2}} \right) dt \geq \frac{c_{\hat{f}}}{2} \int_0^T \sum_j \|\llbracket u_h \rrbracket_{j+\frac{1}{2}}\|^3 dt,$$

so that

$$\frac{c_{\hat{f}}}{2} \int_0^T \sum_j \|\llbracket u_h \rrbracket_{j+\frac{1}{2}}\|^3 dt + \int_0^T \sum_j D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2 dt \leq C.$$

This proves (7.14) and (7.15).

For the first term in the definition of D_j , map I_j affinely to the reference interval. The set of polynomials of degree at most k that are bounded by M_T is a bounded subset of a finite-dimensional space. Hence by Faà di Bruno's formula (6.14), the boundedness of the derivatives of f on $[-M_T, M_T]$, and scaling back to I_j give

$$\|\partial_x^{k+1} f(u_h)\|_{L^2(I_j)} \leq C(h_j)^{-(k+1)} \sum_{i=1}^{k+1} \|f^{(i)}(u_h)\|_{L^2(I_j)} \leq C(h_j)^{-k-\frac{1}{2}}.$$

Since the jumps are uniformly bounded by (7.2), the definition (3.48) consequently implies

$$0 \leq D_j \leq C(h_j)^{-1}. \quad (7.20)$$

Quasi-uniformity now yields

$$\int_0^T \sum_j D_j (h_j)^3 dt \leq C \sum_j h_j^2 \leq Ch,$$

which is (7.16). Finally, (7.18) follows from (3.46) with $r = k + 1$ and

$$D_j \geq \epsilon_1(h_j)^{k-\frac{1}{2}} \|\partial_x^{k+1} f(u_h)\|_{L^2(I_j)}.$$

■

Lemma 7.3 (Compact entropy productions). *The two sequences*

$$\partial_t u_h + \partial_x f(u_h), \quad \partial_t \eta(u_h) + \partial_x q(u_h) \quad (7.21)$$

are relatively compact in $H_{\text{loc}}^{-1}(I \times (0, T))$.

Proof. Fix a bounded Lipschitz open set $O \Subset I \times (0, T)$. All estimates below are restricted to the finitely many space-time cells meeting O . We use the distributional convention

$$\langle \partial_t u_h + \partial_x f(u_h), \varphi \rangle := - \iint_O (u_h \varphi_t + f(u_h) \varphi_x) \, dx \, dt, \quad (7.22)$$

$$\langle \partial_t \eta(u_h) + \partial_x q(u_h), \varphi \rangle := - \iint_O (\eta(u_h) \varphi_t + q(u_h) \varphi_x) \, dx \, dt \quad (7.23)$$

for every $\varphi \in C_c^\infty(O)$. In what follows, $\mathcal{M}(O)$ denotes the space of finite signed Radon measures on the open set O . Equivalently, $\mathcal{M}(O) = C_0(O)^*$. For any $\mu \in \mathcal{M}(O)$, define its norm as $\|\mu\|_{\mathcal{M}(O)} := |\mu|(O)$.

Step 1: the decomposition for the residual of (7.22) on each cell. Let π_j denote the $L^2(I_j)$ -projection onto $\mathbb{P}^k(I_j)$, and define the two polynomials $\ell_j^-, \ell_j^+ \in \mathbb{P}^k(I_j)$ by

$$\int_{I_j} \ell_j^- v \, dx = v(x_{j-\frac{1}{2}}^+), \quad \int_{I_j} \ell_j^+ v \, dx = v(x_{j+\frac{1}{2}}^-) \quad \forall v \in \mathbb{P}^k(I_j). \quad (7.24)$$

One can actually obtain the explicit expressions for ℓ_j^-, ℓ_j^+ as

$$\ell_j^-(x) := \sum_{i=0}^k \sqrt{\frac{2i+1}{h_j}} (-1)^i P_{i,j}(x), \quad \ell_j^+(x) := \sum_{i=0}^k \sqrt{\frac{2i+1}{h_j}} P_{i,j}(x).$$

where $P_{i,j}(x)$ is the i th Legendre polynomial normalized on I_j with $\|P_{i,j}\|_{L^2(I_j)} = 1$. Therefore, $\int_{I_j} \ell_j^- \varphi \, dx = \pi_j \varphi(x_{j-\frac{1}{2}}^+)$ and $\int_{I_j} \ell_j^+ \varphi \, dx = \pi_j \varphi(x_{j+\frac{1}{2}}^-)$. Recall

$$H_{j-\frac{1}{2}}^+ := \hat{f}_{j-\frac{1}{2}} - f(u_h)_{j-\frac{1}{2}}^+, \quad H_{j+\frac{1}{2}}^- := f(u_h)_{j+\frac{1}{2}}^- - \hat{f}_{j+\frac{1}{2}}. \quad (7.25)$$

Using the strong form of (Scheme D) and the defining property (7.24), we obtain the following identity in $\mathbb{P}^k(I_j)$:

$$(u_h)_t|_{I_j} + \pi_j[f(u_h)_x|_{I_j}] = H_{j-\frac{1}{2}}^+ \ell_j^- + H_{j+\frac{1}{2}}^- \ell_j^+ - D_j(u_h - \bar{u}_{h,j}). \quad (7.26)$$

Therefore, with

$$\rho_j := \rho|_{I_j} = f(u_h)_x|_{I_j} - \pi_j[f(u_h)_x|_{I_j}] \in C^{k+1}(I_j),$$

the classical residual in the interior of I_j is

$$((u_h)_t + f(u_h)_x)|_{I_j} = \rho_j + H_{j-\frac{1}{2}}^+ \ell_j^- + H_{j+\frac{1}{2}}^- \ell_j^+ - D_j(u_h - \bar{u}_{h,j}). \quad (7.27)$$

Because $f(u_h)$ is discontinuous at the interfaces, its distributional derivative contains additionally the term

$$\sum_j \llbracket f(u_h) \rrbracket_{j+\frac{1}{2}} \delta_{x_{j+\frac{1}{2}}}.$$

Combining this fact with (7.27) gives the exact decomposition on I :

$$\partial_t u_h + \partial_x f(u_h) = R_h^{\text{vol}} + R_h^{\text{damp}} + R_h^{\text{int}} \quad (7.28)$$

in $\mathcal{D}'(O)$, where

$$R_h^{\text{vol}} := \sum_j \rho_j \mathbf{1}_{I_j}, \quad (7.29)$$

$$R_h^{\text{damp}} := - \sum_j D_j(u_h - \bar{u}_{h,j}) \mathbf{1}_{I_j}, \quad (7.30)$$

$$R_h^{\text{int}} := \sum_j \left(H_{j-\frac{1}{2}}^+ \ell_j^- + H_{j+\frac{1}{2}}^- \ell_j^+ \right) \mathbf{1}_{I_j} + \sum_j \llbracket f(u_h) \rrbracket_{j+\frac{1}{2}} \delta_{x_{j+\frac{1}{2}}}. \quad (7.31)$$

Step 2: R_h^{vol} tends to zero in $H^{-1}(O)$. In addition to (7.18), we need a bound containing the within-cell oscillation. Set

$$v_j := u_h - \bar{u}_{h,j}.$$

Taylor expansion at $\bar{u}_{h,j}$ gives

$$f(u_h) = f(\bar{u}_{h,j}) + f'(\bar{u}_{h,j})v_j + \mathcal{R}_j, \quad (7.32)$$

where the integral form of the remainder is

$$\mathcal{R}_j = v_j^2 \int_0^1 (1-\theta) f''(\bar{u}_{h,j} + \theta v_j) d\theta. \quad (7.33)$$

On the bounded range (7.2), put

$$C_f := \max_{|s| \leq M_T} f''(s).$$

Together with (7.3), (7.33) implies the pointwise comparison

$$\frac{c_f}{2} |v_j|^2 \leq \mathcal{R}_j \leq \frac{C_f}{2} |v_j|^2. \quad (7.34)$$

The first two terms on the right-hand side of (7.32) belong to $\mathbb{P}^k(I_j)$. Their derivatives therefore belong to $\mathbb{P}^{k-1}(I_j) \subset \mathbb{P}^k(I_j)$ and are reproduced exactly by π_j . Consequently,

$$\rho_j = (\mathcal{R}_j)_x - \pi_j[(\mathcal{R}_j)_x]. \quad (7.35)$$

Since π_j is the L^2 -orthogonal projection,

$$\|\rho_j\|_{L^2(I_j)} \leq \|(\mathcal{R}_j)_x\|_{L^2(I_j)}. \quad (7.36)$$

We next establish an inverse-type estimate for $\|(\mathcal{R}_j)_x\|_{L^2(I_j)}$. Differentiating (7.32) in x gives

$$(\mathcal{R}_j)_x = (f'(u_h) - f'(\bar{u}_{h,j}))(v_j)_x.$$

By the fundamental theorem of calculus,

$$f'(u_h) - f'(\bar{u}_{h,j}) = v_j \int_0^1 f''(\bar{u}_{h,j} + \theta v_j) \, d\theta,$$

and hence

$$|(\mathcal{R}_j)_x| \leq C_f |v_j(v_j)_x| = \frac{C_f}{2} |(v_j^2)_x|. \quad (7.37)$$

Although $\mathcal{R}_j$ is generally not a polynomial, v_j^2 belongs to $\mathbb{P}^{2k}(I_j)$. The polynomial inverse inequality can therefore be applied to v_j^2 , giving

$$\|(v_j^2)_x\|_{L^2(I_j)} \leq C h_j^{-1} \|v_j^2\|_{L^2(I_j)}.$$

Moreover, the lower bound in (7.34) yields

$$\|v_j^2\|_{L^2(I_j)} \leq \frac{2}{c_f} \|\mathcal{R}_j\|_{L^2(I_j)}.$$

Combining these two inequalities with (7.36) and (7.37), we obtain

$$\|\rho_j\|_{L^2(I_j)} \leq C h_j^{-1} \|\mathcal{R}_j\|_{L^2(I_j)}, \quad (7.38)$$

where C depends on k and C_f/c_f , but is independent of h_j .

Finally, the upper bound in (7.34) and the polynomial inverse estimate

$$\|u_h - \bar{u}_{h,j}\|_{L^\infty(I_j)} \leq C h_j^{-1/2} \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}$$

give

$$\begin{aligned} \|\rho_j\|_{L^2(I_j)} &\leq C h_j^{-1} \|\mathcal{R}_j\|_{L^2(I_j)} \\ &\leq C h_j^{-1} \|u_h - \bar{u}_{h,j}\|_{L^\infty(I_j)} \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)} \\ &\leq C (h_j)^{-\frac{3}{2}} \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2. \end{aligned} \quad (7.39)$$

Multiplying (7.18) and (7.39) yields

$$h_j^2 \|\rho_j\|_{L^2(I_j)}^2 \leq C h_j D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2. \quad (7.40)$$

Moreover, $\rho_j \perp \mathbb{P}^0(I_j)$. Thus, if $\varphi \in C_c^\infty(O)$ and $\bar{\varphi}_j(t)$ is its average over I_j , then

$$\begin{aligned} |\langle R_h^{\text{vol}}, \varphi \rangle| &= \left| \int_0^T \sum_j \int_{I_j} \rho_j(\varphi - \bar{\varphi}_j) \, dx \, dt \right| \\ &\leq C \left(\int_0^T \sum_j h_j^2 \|\rho_j\|_{L^2(I_j)}^2 \, dt \right)^{1/2} \|\varphi_x\|_{L^2(O)} \\ &\leq Ch^{1/2} \|\varphi\|_{H^1(O)} \end{aligned}$$

by the cellwise Poincaré inequality, (7.14), (7.40), and quasi-uniformity. Consequently,

$$R_h^{\text{vol}} \longrightarrow 0 \quad \text{strongly in } H^{-1}(O). \quad (7.41)$$

Step 3: R_h^{damp} tends to zero in $H^{-1}(O)$. Using again the zero mean of $u_h - \bar{u}_{h,j}$, the cellwise Poincaré inequality, (7.14), and (7.20), we obtain

$$\begin{aligned} |\langle R_h^{\text{damp}}, \varphi \rangle| &\leq \int_0^T \sum_j D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)} \|\varphi - \bar{\varphi}_j\|_{L^2(I_j)} \, dt \\ &\leq C \left(\int_0^T \sum_j D_j \|u_h - \bar{u}_{h,j}\|_{L^2(I_j)}^2 \, dt \right)^{1/2} \\ &\quad \times \left(\int_0^T \sum_j D_j h_j^2 \|\varphi_x\|_{L^2(I_j)}^2 \, dt \right)^{1/2} \\ &\leq Ch^{1/2} \|\varphi\|_{H^1(O)}. \end{aligned}$$

Hence

$$R_h^{\text{damp}} \longrightarrow 0 \quad \text{strongly in } H^{-1}(O). \quad (7.42)$$

Step 4: the conservation-law interface remainder R_h^{int} . For clarity of notation, let $u_{j+\frac{1}{2}}^\pm = u_h(x_{j+\frac{1}{2}}^\pm, t)$ and $\varphi_{j+\frac{1}{2}} = \varphi(x_{j+\frac{1}{2}}, t)$. From (7.24), the complete contribution at $x_{j+\frac{1}{2}}$ is

$$\begin{aligned} \langle R_{h,j+\frac{1}{2}}^{\text{int}}, \varphi \rangle &= \int_0^T \left\{ (f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}})(\pi_j \varphi)(x_{j+\frac{1}{2}}^-) + (\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+))(\pi_{j+1} \varphi)(x_{j+\frac{1}{2}}^+) \right. \\ &\quad \left. + (f(u_{j+\frac{1}{2}}^+) - f(u_{j+\frac{1}{2}}^-))\varphi_{j+\frac{1}{2}} \right\} dt \\ &= \int_0^T \left\{ (f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}})[(\pi_j \varphi)(x_{j+\frac{1}{2}}^-) - \varphi_{j+\frac{1}{2}}] \right. \\ &\quad \left. + (\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+))[(\pi_{j+1} \varphi)(x_{j+\frac{1}{2}}^+) - \varphi_{j+\frac{1}{2}}] \right\} dt. \end{aligned} \quad (7.43)$$

Indeed, the terms multiplying $\varphi_{j+\frac{1}{2}}$ cancel because

$$f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}} + \hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+) + f(u_{j+\frac{1}{2}}^+) - f(u_{j+\frac{1}{2}}^-) = 0.$$

Consistency and local Lipschitz continuity imply

$$|f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}}| + |\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+)| \leq C |u_{j+\frac{1}{2}}^+ - u_{j+\frac{1}{2}}^-|. \quad (7.44)$$

For $p' > 2$, we claim that the following inequality holds

$$|(\pi_j \varphi)(x_{j+\frac{1}{2}}^\pm) - \varphi_{j+\frac{1}{2}}| \leq C h_j^{1-\frac{1}{p'}} \|\varphi_x(\cdot, t)\|_{L^{p'}(I_j)}, \quad (7.45)$$

where C is independent of φ . Here is the detailed scaling argument for proving (7.45). Let

$$F_j(\xi) := x_j + \frac{h_j}{2} \xi, \quad -1 \leq \xi \leq 1,$$

and set

$$\hat{\varphi}(\xi, t) := \varphi(F_j(\xi), t).$$

If $\hat{\pi}$ denotes the $L^2(-1, 1)$ -projection onto $\mathbb{P}^k(-1, 1)$, then the affine invariance of the L^2 -projection gives

$$(\pi_j \varphi)(F_j(\xi), t) = (\hat{\pi} \hat{\varphi})(\xi, t). \quad (7.46)$$

Since $x_{j+\frac{1}{2}} = F_j(1)$ and $\hat{\pi}$ reproduces constants,

$$\begin{aligned} |(\pi_j \varphi)(x_{j+\frac{1}{2}}^-, t) - \varphi_{j+\frac{1}{2}}| &= |\hat{\pi}[\hat{\varphi}(\cdot, t) - \hat{\varphi}(1, t)](1)| \\ &\leq C \|\hat{\varphi}(\cdot, t) - \hat{\varphi}(1, t)\|_{L^\infty(-1, 1)}. \end{aligned} \quad (7.47)$$

The last inequality follows because $g \mapsto (\hat{\pi} g)(1)$ is a bounded linear functional on $L^\infty(-1, 1)$. For every $\xi \in [-1, 1]$, the fundamental theorem of calculus and Hölder's inequality give

$$|\hat{\varphi}(\xi, t) - \hat{\varphi}(1, t)| \leq \int_{\xi}^1 |\partial_{\zeta} \hat{\varphi}(\zeta, t)| \, d\zeta \leq 2^{1-\frac{1}{p'}} \|\partial_{\xi} \hat{\varphi}(\cdot, t)\|_{L^{p'}(-1, 1)}.$$

Since

$$\partial_{\xi} \hat{\varphi} = \frac{h_j}{2} \varphi_x \circ F_j$$

and $dx = (h_j/2)d\xi$, scaling yields

$$\|\partial_{\xi} \hat{\varphi}(\cdot, t)\|_{L^{p'}(-1, 1)} = \left(\frac{h_j}{2}\right)^{1-\frac{1}{p'}} \|\varphi_x(\cdot, t)\|_{L^{p'}(I_j)}. \quad (7.48)$$

Substituting (7.48) into (7.47) proves (7.45). The proof at the left endpoint is identical. Notice that this one-dimensional estimate only requires $p' > 1$; we choose $p' > 2$ because its conjugate exponent $p = p'/(p' - 1)$ then satisfies $1 < p < 2$, which is the range used below.

Put

$$b_{j+\frac{1}{2}}(t) := |\llbracket u_h \rrbracket_{j+\frac{1}{2}}|, \quad a_j(t) := \|\varphi_x(\cdot, t)\|_{L^{p'}(I_j)}.$$

At the interface $x_{j+\frac{1}{2}}$, equations (7.43), (7.44), and (7.45) imply

$$|\langle R_{h,j+\frac{1}{2}}^{\text{int}}, \varphi \rangle| \leq C \int_0^T b_{j+\frac{1}{2}}(t) \left[h_j^{1-\frac{1}{p'}} a_j(t) + h_{j+1}^{1-\frac{1}{p'}} a_{j+1}(t) \right] dt. \quad (7.49)$$

Define

$$A_{j+\frac{1}{2}}(t) := a_j(t) + a_{j+1}(t).$$

Summing (7.49) over all interfaces contained in $\mathcal{O}$ gives

$$|\langle R_h^{\text{int}}, \varphi \rangle| \leq Ch^{1-\frac{1}{p'}} \int_0^T \sum_{(x_{j+\frac{1}{2}}, t) \in \mathcal{O}} b_{j+\frac{1}{2}}(t) A_{j+\frac{1}{2}}(t) dt. \quad (7.50)$$

In what follows we shall simply write $\sum_j$ instead of $\sum_{(x_{j+\frac{1}{2}}, t) \in \mathcal{O}}$. We now apply Hölder's inequality with exponents 3 and $3/2$ on the product of Lebesgue measure in time and counting measure on the interface set:

$$\int_0^T \sum_j b_{j+\frac{1}{2}} A_{j+\frac{1}{2}} dt \leq \left(\int_0^T \sum_j b_{j+\frac{1}{2}}^3 dt \right)^{1/3} \left(\int_0^T \sum_j A_{j+\frac{1}{2}}^{3/2} dt \right)^{2/3}. \quad (7.51)$$

By (7.15), the first factor is bounded uniformly in h . Moreover,

$$A_{j+\frac{1}{2}}^{3/2} \leq 2^{1/2} (a_j^{3/2} + a_{j+1}^{3/2})$$

implies

$$\sum_j A_{j+\frac{1}{2}}(t)^{3/2} \leq C \sum_j a_j(t)^{3/2}. \quad (7.52)$$

For completeness, let N_h be the number of cells meeting the spatial projection of $\mathcal{O}$. Quasi-uniformity gives $N_h \leq Ch^{-1}$. Because $p' > 2 > 3/2$, discrete Hölder's inequality with conjugate exponents $2p'/3$ and $2p'/(2p' - 3)$ yields

$$\sum_j a_j(t)^{3/2} \leq \left(\sum_j a_j(t)^{p'} \right)^{\frac{3}{2p'}} N_h^{1-\frac{3}{2p'}} \leq Ch^{-1+\frac{3}{2p'}} \left(\sum_j a_j(t)^{p'} \right)^{\frac{3}{2p'}}. \quad (7.53)$$

Set

$$B(t) := \sum_j a_j(t)^{p'} = \sum_j \int_{I_j} |\varphi_x(x, t)|^{p'} dx.$$

Since $\gamma := 3/(2p') < 1$, Hölder's inequality in time gives

$$\int_0^T B(t)^\gamma dt \leq T^{1-\gamma} \left(\int_0^T B(t) dt \right)^\gamma = T^{1-\frac{3}{2p'}} \|\varphi_x\|_{L^{p'}(\mathcal{O})}^{3/2}. \quad (7.54)$$

Combining (7.52)–(7.54) and then raising to the power $2/3$, we obtain

$$\left(\int_0^T \sum_j A_{j+\frac{1}{2}}(t)^{3/2} dt \right)^{2/3} \leq Ch^{-\frac{2}{3}+\frac{1}{p'}} T^{\frac{2}{3}-\frac{1}{p'}} \|\varphi_x\|_{L^{p'}(\mathcal{O})}. \quad (7.55)$$

Substituting (7.51) and (7.55) into (7.50), and using (7.15), yields

$$|\langle R_h^{\text{int}}, \varphi \rangle| \leq Ch^{1-\frac{1}{p'}} h^{-\frac{2}{3}+\frac{1}{p'}} \|\varphi_x\|_{L^{p'}(O)} = Ch^{1/3} \|\varphi_x\|_{L^{p'}(O)} \leq Ch^{1/3} \|\varphi\|_{W^{1,p'}(O)}. \quad (7.56)$$

Thus, with $1/p + 1/p' = 1$,

$$R_h^{\text{int}} \longrightarrow 0 \quad \text{strongly in } W^{-1,p}(O), \quad 1 < p < 2. \quad (7.57)$$

Combining (7.28), (7.41), (7.42), and (7.57), we conclude that the first production in (7.21) is relatively compact in $W^{-1,p}(O)$ for every $1 < p < 2$. Its upgrade to $H^{-1}(O)$ is carried out in Step 6.

Step 5: the decomposition of the square-entropy production (7.23). Since $\eta'(u_h) = u_h$, the chain rule is valid in each open cell and

$$(\partial_t \eta(u_h) + \partial_x q(u_h))|_{I_j} = (u_h((u_h)_t + f(u_h)_x))|_{I_j}.$$

Multiplying (7.27) by u_h and adding the distributional interface term $\llbracket q(u_h) \rrbracket_{j+\frac{1}{2}} \delta_{x_{j+\frac{1}{2}}}$ gives

$$\partial_t \eta(u_h) + \partial_x q(u_h) = -\mu_h^{\text{cell}} - \mu_h^{\text{int}} + S_h^{\text{vol}} + S_h^{\text{damp}} + S_h^{\text{int}}, \quad (7.58)$$

where the two nonnegative measures are

$$\mu_h^{\text{cell}} := \sum_j D_j |u_h - \bar{u}_{h,j}|^2 \mathbf{1}_{I_j} \, dx \, dt, \quad (7.59)$$

$$\mu_h^{\text{int}} := \sum_j \mathcal{D}_\eta(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) \delta_{x_{j+\frac{1}{2}}} \, dt. \quad (7.60)$$

and the space-time distributions are

$$S_h^{\text{vol}} := \sum_j u_h \rho_j \mathbf{1}_{I_j}, \quad (7.61)$$

$$S_h^{\text{damp}} := - \sum_j D_j \bar{u}_{h,j} (u_h - \bar{u}_{h,j}) \mathbf{1}_{I_j}, \quad (7.62)$$

$$S_h^{\text{int}} := \sum_j u_h \left(H_{j-\frac{1}{2}}^+ \ell_j^- + H_{j+\frac{1}{2}}^- \ell_j^+ \right) \mathbf{1}_{I_j} + \sum_j \left(\llbracket q(u_h) \rrbracket_{j+\frac{1}{2}} + \mathcal{D}_\eta(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) \right) \delta_{x_{j+\frac{1}{2}}}. \quad (7.63)$$

Note that the total contribution at the interface $x_{j+\frac{1}{2}}$ is

$$\begin{aligned} & (f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}}) [\pi_j(u_h \varphi)(x_{j+\frac{1}{2}}^-) - u_{j+\frac{1}{2}}^- \varphi_{j+\frac{1}{2}}] \\ & + (\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+)) [\pi_{j+1}(u_h \varphi)(x_{j+\frac{1}{2}}^+) - u_{j+\frac{1}{2}}^+ \varphi_{j+\frac{1}{2}}] \\ & + \left\{ u_{j+\frac{1}{2}}^- [f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}}] + u_{j+\frac{1}{2}}^+ [\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+)] + q(u_{j+\frac{1}{2}}^+) - q(u_{j+\frac{1}{2}}^-) \right\} \varphi_{j+\frac{1}{2}}. \end{aligned} \quad (7.64)$$

By $q'(s) = sf'(s)$, the quantity in braces equals

$$(u_{j+\frac{1}{2}}^+ - u_{j+\frac{1}{2}}^-) \hat{f}_{j+\frac{1}{2}} - \int_{u_{j+\frac{1}{2}}^-}^{u_{j+\frac{1}{2}}^+} f(s) \, ds = -\mathcal{D}_\eta(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+).$$

This is precisely the measure $-\mu_h^{\text{int}}$. The remaining two terms define S_h^{int} . In addition, μ_h^{cell} and S_h^{damp} arise from $-D_j u_h(u_h - \bar{u}_{h,j}) = -D_j |u_h - \bar{u}_{h,j}|^2 - D_j \bar{u}_{h,j}(u_h - \bar{u}_{h,j}) = -\mu_h^{\text{cell}} + S_h^{\text{damp}}$.

Inequalities (7.19) and (7.14), which are the direct consequences of the global square-entropy inequality, imply

$$\|\mu_h^{\text{cell}}\|_{\mathcal{M}(O)} + \|\mu_h^{\text{int}}\|_{\mathcal{M}(O)} \leq C. \quad (7.65)$$

We now identify and estimate the three remaining terms in (7.58). First, because $|u_h| \leq M_T$,

$$S_h^{\text{vol}} := \sum_j u_h \rho_j \mathbf{1}_{I_j}$$

can be estimated in the same way as the proof of (7.41):

$$\begin{aligned} |\langle S_h^{\text{vol}}, \varphi \rangle| &= \left| \int_0^T \sum_j \int_{I_j} u_h \rho_j (\varphi - \bar{\varphi}_j) \, dx \, dt \right| \\ &\leq C \left(\int_0^T \sum_j h_j^2 \|\rho_j\|_{L^2(I_j)}^2 \, dt \right)^{1/2} \|\varphi_x\|_{L^2(O)} \\ &\leq Ch^{1/2} \|\varphi\|_{H^1(O)}. \end{aligned} \quad (7.66)$$

Thus $S_h^{\text{vol}} \rightarrow 0$ strongly in $H^{-1}(O)$. Also, $S_h^{\text{damp}} := -\sum_j D_j \bar{u}_{h,j}(u_h - \bar{u}_{h,j}) \mathbf{1}_{I_j}$ tends to zero strongly in $H^{-1}(O)$ by Step 3 and $|\bar{u}_{h,j}| \leq M_T$.

Finally, we estimate S_h^{int} . By (7.64), we have

$$\begin{aligned} \langle S_h^{\text{int}}, \varphi \rangle &= \int_0^T \sum_j \left\{ (f(u_{j+\frac{1}{2}}^-) - \hat{f}_{j+\frac{1}{2}}) [\pi_j(u_h \varphi)(x_{j+\frac{1}{2}}^-) - u_{j+\frac{1}{2}}^- \varphi_{j+\frac{1}{2}}] \right. \\ &\quad \left. + (\hat{f}_{j+\frac{1}{2}} - f(u_{j+\frac{1}{2}}^+)) [\pi_{j+1}(u_h \varphi)(x_{j+\frac{1}{2}}^+) - u_{j+\frac{1}{2}}^+ \varphi_{j+\frac{1}{2}}] \right\} dt \end{aligned}$$

Since π_j reproduces u_h , the L^∞ -stability of projection gives

$$\begin{aligned} |\pi_j(u_h \varphi)(x_{j+\frac{1}{2}}^\pm) - u_{j+\frac{1}{2}}^\pm \varphi_{j+\frac{1}{2}}| &\leq C \|u_h(\varphi - \varphi_{j+\frac{1}{2}})\|_{L^\infty(I_j)} \\ &\leq Ch_j^{1-\frac{1}{p'}} \|u_h\|_{L^\infty(I_j)} \|\varphi_x\|_{L^{p'}(I_j)} \end{aligned} \quad (7.67)$$

for $p' > 1$. Now, repeating the explicit Hölder estimates leading to (7.56), gives

$$S_h^{\text{int}} \longrightarrow 0 \quad \text{strongly in } W^{-1,p}(O), \quad 1 < p < 2. \quad (7.68)$$

Step 6: application of Murat's lemma. We use the following form of Murat's lemma [14].

Murat's lemma. Let $O \subset \mathbb{R}^2$ be bounded. Suppose that a sequence $\{L_n\}$ satisfies:

- (i) $\{L_n\}$ is bounded in $W^{-1,r}(O)$ for some $r > 2$;
- (ii) for some $1 < p < 2$, $\{L_n\}$ is relatively compact in $W^{-1,p}(O)$.

Then $\{L_n\}$ is relatively compact in $H^{-1}(O)$. A frequently used sufficient condition for (ii) is a decomposition

$$L_n = K_n + M_n,$$

where $\{K_n\}$ is relatively compact in $W^{-1,p}(O)$ and $\{M_n\}$ is bounded in $\mathcal{M}(O)$. Indeed, in two space-time dimensions,

$$\mathcal{M}(O) \hookrightarrow W^{-1,p}(O) \quad \text{compactly for every } 1 < p < 2. \quad (7.69)$$

After extracting a subsequence, both K_n and M_n therefore converge in $W^{-1,p}(O)$, which proves (ii). Finally, interpolation between $W^{-1,p}$ and $W^{-1,r}$ gives

$$\|L_n - L_m\|_{H^{-1}(O)} \leq C \|L_n - L_m\|_{W^{-1,p}(O)}^\theta \|L_n - L_m\|_{W^{-1,r}(O)}^{1-\theta},$$

where $0 < \theta < 1$ is determined by

$$\frac{1}{2} = \frac{\theta}{p} + \frac{1-\theta}{r}.$$

Thus every subsequence that is Cauchy in $W^{-1,p}$ is Cauchy in H^{-1} , which proves the lemma.

We now verify these hypotheses. Define

$$L_{1,h} := \partial_t u_h + \partial_x f(u_h), \quad L_{2,h} := \partial_t \eta(u_h) + \partial_x q(u_h).$$

For $\varphi \in C_c^\infty(O)$,

$$\begin{aligned} |\langle L_{1,h}, \varphi \rangle| &\leq \|u_h\|_{L^\infty(O)} \|\varphi_t\|_{L^1(O)} + \|f(u_h)\|_{L^\infty(O)} \|\varphi_x\|_{L^1(O)}, \\ |\langle L_{2,h}, \varphi \rangle| &\leq \|\eta(u_h)\|_{L^\infty(O)} \|\varphi_t\|_{L^1(O)} + \|q(u_h)\|_{L^\infty(O)} \|\varphi_x\|_{L^1(O)}. \end{aligned}$$

The uniform bound (7.2) and boundedness of O imply that, for every finite $r > 2$,

$$\|L_{1,h}\|_{W^{-1,r}(O)} + \|L_{2,h}\|_{W^{-1,r}(O)} \leq C_r. \quad (7.70)$$

This verifies condition (i) for both entropy productions.

For $L_{1,h}$, Steps 2 and 3 give terms converging strongly in H^{-1} , while Step 4 gives a term converging strongly in $W^{-1,p}$. Thus (7.28) shows directly that $\{L_{1,h}\}$ is relatively compact in $W^{-1,p}(O)$.

For $L_{2,h}$, define

$$K_h := S_h^{\text{vol}} + S_h^{\text{damp}} + S_h^{\text{int}}, \quad M_h := -\mu_h^{\text{cell}} - \mu_h^{\text{int}}.$$

Equations (7.66) and (7.68) show that $\{K_h\}$ is relatively compact in $W^{-1,p}(O)$, while (7.65) shows that $\{M_h\}$ is bounded in $\mathcal{M}(O)$. By (7.69), $\{M_h\}$ is relatively compact in $W^{-1,p}(O)$. Hence (7.58) implies that $\{L_{2,h}\}$ is relatively compact in $W^{-1,p}(O)$, proving condition (ii).

Murat's lemma and (7.70) now imply that both $\{L_{1,h}\}$ and $\{L_{2,h}\}$ are relatively compact in $H^{-1}(O)$. Since $O \Subset I \times (0, T)$ was arbitrary, the two sequences in (7.21) are relatively compact in $H_{\text{loc}}^{-1}(I \times (0, T))$. ■

We can now state the strong-convergence result.

Theorem 7.4 (Strong convergence for a uniformly convex flux). *Let $k \geq 1$ and let u_h be the solution of (Scheme D) with $\epsilon_1 > 0$ subject to periodic or compactly supported boundary conditions. Assume (7.1)–(7.6) hold, and that the initial projections converge strongly to $\mathcal{U}_0$ in $L^2(I)$, i.e.,*

$$u_h(\cdot, 0) \longrightarrow \mathcal{U}_0 \quad \text{strongly in } L^2(I).$$

Then every sequence $h \rightarrow 0$ contains a subsequence, still denoted by u_h , and a function $\mathcal{U} \in L^\infty(I \times (0, T))$ such that

$$u_h \longrightarrow \mathcal{U} \quad \text{strongly in } L_{\text{loc}}^p(I \times (0, T)) \quad \text{for every } 1 \leq p < \infty. \quad (7.71)$$

The limit is a distributional solution of (2.1) and satisfies the limiting square-entropy inequality. Since $f(\cdot)$ is strictly convex, the square-entropy inequality is enough to guarantee uniqueness of solution [6], so that the entire family u_h converges to the unique entropy solution of (2.1).

Proof. By (7.2), we can extract a subsequence of u_h and the associated family of Young measures $\nu_{x,t}$ supported in $[-M_T, M_T]$ such that $g(u_h) \rightharpoonup \langle \nu_{x,t}, g \rangle$ weakly-* in L^∞ , for any continuous function g [4, 5]. By Lemma 7.3, we can apply the div-curl lemma [18] to the two entropy pairs

$$\begin{aligned} \eta_1(s) &= s, & q_1(s) &= f(s), \\ \eta_2(s) &= \frac{s^2}{2}, & q_2(s) &= q(s), & q'(s) &= s f'(s), \end{aligned}$$

and obtain

$$\langle \nu, \eta_1 \rangle \langle \nu, q_2 \rangle - \langle \nu, \eta_2 \rangle \langle \nu, q_1 \rangle = \langle \nu, \eta_1 q_2 - \eta_2 q_1 \rangle.$$

Equivalently, this can be written in its symmetrized form as

$$\iint_{\mathbb{R}^2} C(a, b) \, d\nu_{x,t}(a) \, d\nu_{x,t}(b) = 0, \quad (7.72)$$

where

$$\begin{aligned} C(a, b) &= (a - b)(q(a) - q(b)) - \frac{a^2 - b^2}{2}(f(a) - f(b)) \\ &= \frac{1}{2} \int_b^a \int_b^a (s - r)(f'(s) - f'(r)) \, dr \, ds. \end{aligned} \quad (7.73)$$

By (7.3),

$$(s - r)(f'(s) - f'(r)) \geq c_f(s - r)^2.$$

Hence,

$$C(a, b) \geq \frac{cf}{12} |a - b|^4. \quad (7.74)$$

Equations (7.72) and (7.74) imply

$$\iint_{\mathbb{R}^2} |a - b|^4 \, d\nu_{x,t}(a) \, d\nu_{x,t}(b) = 0.$$

Thus $\nu_{x,t}$ is a Dirac mass for almost every (x, t) . Hence $\nu_{x,t} = \delta_{\mathcal{U}(x,t)}$ and $u_h \rightarrow \mathcal{U}$ locally in measure. Since $\{|u_h|^p\}$ is uniformly integrable on every compact subset, Vitali's theorem gives

$$u_h \rightarrow \mathcal{U} \quad \text{strongly in } L_{\text{loc}}^p, \quad 1 \leq p < \infty.$$

In Lemma 7.3 we showed that $(u_h)_t + f(u_h)_x \rightarrow 0$ in $H_{\text{loc}}^{-1}(I \times (0, T))$. Together with the strong convergence of u_h , this implies $\mathcal{U}_t + f(\mathcal{U})_x = 0$ in distributions, and the square-entropy inequality can be passed to the limit as well. It remains to identify the initial trace. Let $\varphi \in C_c^\infty(I \times [0, T))$. The estimates in Steps 2–4 of Lemma 7.3 remain valid for such a test function, since they only use spatial derivatives of φ and estimates integrated over $0 < t < T$. Hence, writing $R_h := (u_h)_t + f(u_h)_x$, we have

$$|\langle R_h, \varphi \rangle| \leq C_\varphi (h^{1/2} + h^{1/3}) \rightarrow 0.$$

Because u_h is differentiable in time and $\varphi(\cdot, T) = 0$, integration by parts in time gives

$$\langle R_h, \varphi \rangle = - \int_0^T \int_I (u_h \varphi_t + f(u_h) \varphi_x) \, dx \, dt - \int_I u_h(x, 0) \varphi(x, 0) \, dx.$$

Fix $\delta > 0$. On $I \times (\delta, T)$, the strong convergence already proved above and continuity of f on $[-M_T, M_T]$ allow passage to the limit in the space–time integral. On $I \times (0, \delta)$, the uniform L^∞ bound on u_h and $f(u_h)$ makes the absolute value of that integral at most $C_\varphi \delta$, uniformly in h . Thus, first letting $h \rightarrow 0$ and then $\delta \downarrow 0$, and using $u_h(\cdot, 0) \rightarrow \mathcal{U}_0$ strongly in $L^2(I)$, we obtain

$$\int_0^T \int_I (\mathcal{U} \varphi_t + f(\mathcal{U}) \varphi_x) \, dx \, dt + \int_I \mathcal{U}_0(x) \varphi(x, 0) \, dx = 0.$$

Therefore $\mathcal{U}$ has initial value $\mathcal{U}_0$ in the distributional trace sense, and the limit is the unique entropy solution to (2.1). $\blacksquare$

8. Numerical results

Recall the definitions (3.47) and (3.48) of the damping coefficients in (Scheme C) and (Scheme D):

$$C_j[u; U, r] := q_j[u; U] \left\{ \frac{C_k C(k, r) h_j^{r - \frac{3}{2}}}{2} \|\partial_x^r f(u)\|_{L^2(I_j)} + \frac{\beta_k}{2h_j} \left(|H_{j-\frac{1}{2}}^+| + |H_{j+\frac{1}{2}}^-| \right) \right\},$$

$$D_j[u; r, \epsilon_1, \epsilon_2] := \epsilon_1 h_j^{r-\frac{3}{2}} \|\partial_x^r f(u)\|_{L^2(I_j)} + \epsilon_2 \frac{|\llbracket u \rrbracket_{j-\frac{1}{2}}| + |\llbracket u \rrbracket_{j+\frac{1}{2}}|}{h_j},$$

where $k+1 \leq r \leq k+2$. The first term on the right-hand side of either C_j and D_j hinders the efficiency of the numerical scheme, but is nevertheless indispensable in the proof of local entropy inequalities and the strong convergence. In Section 7, we proved strong convergence only for strictly convex conservation laws; but for general non-convex conservation laws, the convergence of the scheme usually depends on the choice of ϵ_1, ϵ_2 . This section discusses the effects of ϵ_1 on numerical results and the question of whether it can be set to zero.

First of all, it can be seen from Tables 1, 2 and B.1 that the constant $C_k C(k, r)$ is much smaller in magnitude compared to $\beta_k = 2\mathcal{A}_k$ as k increases. Indeed, the ratio $C_k C(k, r)/\beta_k$ is recorded in the following table for the first few k :

(k, r)	$C_k C(k, r)$	$\beta_k = 2\mathcal{A}_k$	$C_k C(k, r)/\beta_k$
(1,2)	0.2387324145	4.3703703704	$5.46252135 \times 10^{-2}$
(1,3)	0.06704424360	4.3703703704	$1.53406320 \times 10^{-2}$
(2,3)	0.03495228188	12.1950289930	$2.86610896 \times 10^{-3}$
(2,4)	0.008690199690	12.1950289930	$7.12601806 \times 10^{-4}$
(3,4)	0.003382471073	28.6965574262	$1.17870274 \times 10^{-4}$
(3,5)	0.0007582675327	28.6965574262	$2.64236410 \times 10^{-5}$
(4,5)	0.0002407636609	52.7920626087	$4.56060341 \times 10^{-6}$
(4,6)	0.00004943108133	52.7920626087	$9.36335481 \times 10^{-7}$

Table 7. The ratio $C_k C(k, r)/\beta_k$ for $k+1 \leq r \leq k+2$.

Therefore, assuming that our constant β_k is close to its optimal value, and taking into account that $|H_{j+\frac{1}{2}}^\pm| \approx |f'(u)| \cdot |\llbracket u \rrbracket_{j+\frac{1}{2}}|$ we deduce that ϵ_1 and ϵ_2 should also satisfy

$$\frac{\epsilon_1}{\epsilon_2} \approx \frac{C_k C(k, r)}{\beta_k \max |f'(u)|}, \quad (8.1)$$

or at least both sides of (8.1) should be asymptotically in the same order of magnitude. This suggests that, for small values of ϵ_2 , we may take ϵ_1 to be equally small during computations if the numerical solution itself is observed to be uniformly bounded. This is observed in numerical tests for non-convex conservation laws shown below. Moreover, it has also been observed that if ϵ_2 is sufficiently large, the scheme would always converge for non-convex tests with $\epsilon_1 = 0$. The really interesting case is when we use a moderate value of ϵ_2 , say around 10, then the smallest value of ϵ_1 for the scheme to converge satisfies (8.1). Indeed, this is demonstrated by the following example.

Example 8.1. Consider the Buckley-Leverett equation with the non-convex flux

$$f(u) = \frac{4u^2}{4u^2 + (1-u)^2}.$$

The computational domain is $I = [-1, 1]$ and the initial condition is

$$u_0(x) = \begin{cases} 1, & -\frac{1}{2} \leq x \leq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We evolve the solution up to $T = 0.3$. The reference solution is shown in Figure 1, which was computed by the first-order monotone scheme with Godunov flux and a uniform mesh of size $N = 20000$.

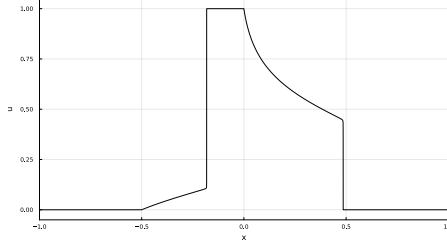


Figure 1

Next, we compute the solution using (Scheme D) with a uniform mesh of size $N = 275$. The numerical solutions with $\epsilon_1 = 0$, $\epsilon_2 = 10$ and $k = 1, 2, 3, 4$ are shown in Figure 2. The global Lax–Friedrichs flux with $\max|f'(u)| \approx 2.34$ and classical RK4 were used in all tests. Note that none of them converges to the correct solution.

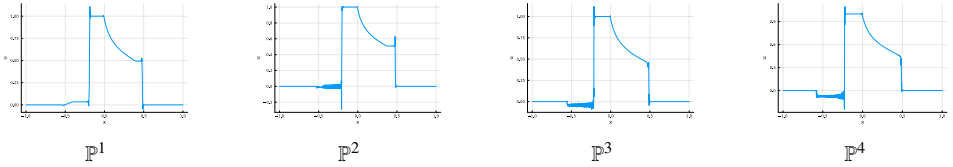


Figure 2. Solutions of (Scheme D) with $\epsilon_1 = 0$, $\epsilon_2 = 10$ and $k = 1, 2, 3, 4$. We use a uniform mesh with $N = 275$. None of the above solutions converges to the correct solution.

Now, keep $\epsilon_2 = 10$ and start slightly increasing ϵ_1 . We observe that when ϵ_1 reaches a critical value, the scheme would converge correctly. These numerical solutions for (Scheme D) with $r = k + 1$ are shown in Figure 3, and those with $r = k + 2$ are shown in Figure 4.

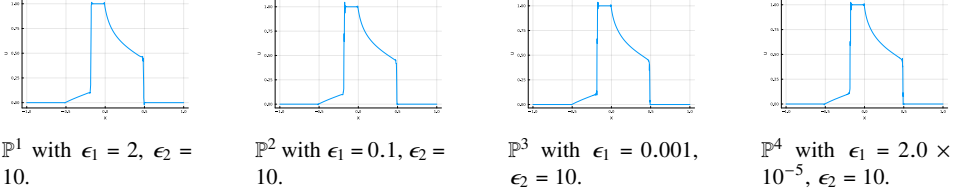


Figure 3. Solutions of (Scheme D) with $k = 1, 2, 3, 4$ and $r = k + 1$. We still use a uniform mesh with $N = 275$. Now, all of the above solutions agree with the correct solution.

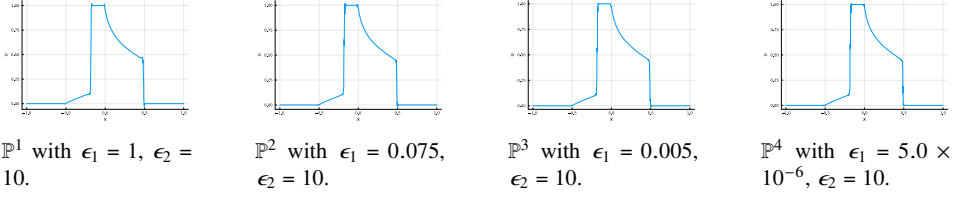


Figure 4. Solutions of (Scheme D) with $k = 1, 2, 3, 4$ and $r = k + 2$. We still use a uniform mesh with $N = 275$. All of the above solutions agree with the correct solution.

If we compute the ratio of the observed ϵ_1 and ϵ_2 from Figures 3 and 4, and then compare it with the value on the right-hand side of (8.1), we would find that they indeed share the same order of magnitude asymptotically, as is shown in Table 8. With this smallest ϵ_1 ,

(k, r)	$\frac{\epsilon_1}{\epsilon_2}$	$\frac{C_k C(k, r)}{\beta_k \max f'(u) }$
(1,2)	2.0×10^{-1}	2.3×10^{-2}
(1,3)	1.0×10^{-1}	6.6×10^{-3}
(2,3)	1.0×10^{-2}	1.2×10^{-3}
(2,4)	7.5×10^{-3}	3.0×10^{-4}
(3,4)	1.0×10^{-4}	5.0×10^{-5}
(3,5)	5.0×10^{-4}	1.1×10^{-5}
(4,5)	2.0×10^{-6}	1.9×10^{-6}
(4,6)	5.0×10^{-7}	4.0×10^{-7}

Table 8

we can reduce ϵ_2 significantly such that the scheme still converges, as shown in Figure 5 for (Scheme D) with $r = k + 2$. In particular, when $k = 4$ we may set $\epsilon_1 = 5.0 \times 10^{-6}$ and $\epsilon_2 = 1.0 \times 10^{-6}$.

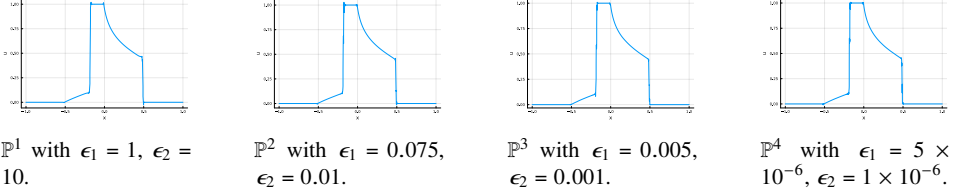


Figure 5. Solutions of (Scheme D) with $k = 1, 2, 3, 4$ and $r = k + 2$. We still use a uniform mesh with $N = 275$. All of the above solutions agree with the correct solution.

Example 8.2. Consider the non-convex flux

$$f(u) = \sin(u).$$

The computational domain is $I = [-5, 5]$ and the initial condition is

$$u_0(x) = \begin{cases} \frac{\pi}{64}, & x < 0, \\ \frac{255\pi}{64}, & x \geq 0. \end{cases}$$

We evolve the solution up to $T = 4$. The reference solution is shown in Figure 6, which was computed by the first-order monotone scheme with Godunov flux and a uniform mesh of size $N = 20000$. This example is borrowed from [16], where it is used to demonstrate the failure of convergence for the classical DG scheme.

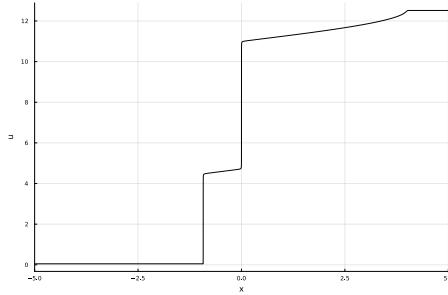


Figure 6

Next, we compute the solution using (Scheme D) with a uniform mesh of size $N = 275$. The numerical solutions with $\epsilon_1 = 0, \epsilon_2 = 1$ and $k = 1, 2, 3, 4$ are shown in Figure 7. The global Lax–Friedrichs flux with $\max|f'(u)| = 1$ and classical RK4 were used in all tests. None of them converges to the correct solution.

Then, while keeping $\epsilon_2 = 1$, we set $\epsilon_1 = 1$ for $k = 1, 2$, $\epsilon_1 = 0.1$ for $k = 3$, and $\epsilon_1 = 0.01$ for $k = 4$. The results for $r = k + 1$ and $r = k + 2$ are shown in Figures 8 and 9 respectively. The scheme now converges correctly to the entropy solution.

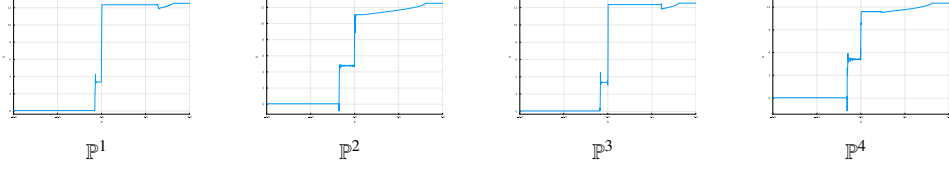


Figure 7. Solutions of (Scheme D) with $\epsilon_1 = 0$, $\epsilon_2 = 1$ and $k = 1, 2, 3, 4$. We use a uniform mesh with $N = 275$. None of the above solutions converges to the correct solution.

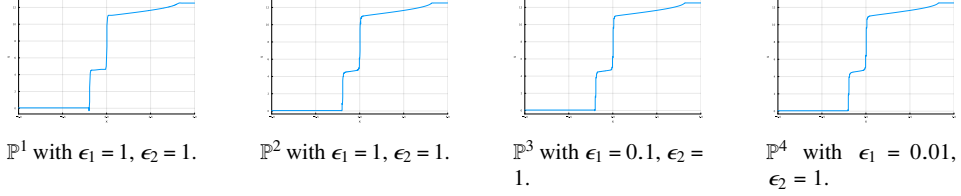


Figure 8. Solutions of (Scheme D) for $k = 1, 2, 3, 4$ and $r = k + 1$. We still use a uniform mesh with $N = 275$. Now we have convergence to the correct solution.

Example 8.3. Consider a Riemann problem for one-dimensional Euler equations known as the *Sod shock tube problem*. The heat capacity ratio is taken as $\gamma = 1.4$. The initial condition is

$$(\rho, u, p) = \begin{cases} (1, 0, 1), & x < 0.5, \\ (0.125, 0, 0.1), & x \geq 0.5. \end{cases}$$

The computational domain is $I = [0, 1]$ with transmissive boundary conditions. We use a uniform mesh of size $N = 100$. The solution is computed up to $T = 0.2$ using (Scheme D') with $k = 1, 2, 3$ and $r = k + 1$. The local Lax–Friedrichs flux and SSP–RK3 were used in all tests. We set $\epsilon_2 = 10$ for all tests. In Figure 10, the solutions on the left are obtained with $\epsilon_1 = 0$ for all $k = 1, 2, 3$. On the right, we set $\epsilon_1 = 10$ for $k = 1$, $\epsilon_1 = 1$ for $k = 2$, and $\epsilon_1 = 0.01$ for $k = 3$. In this example, no huge difference can be noted between $\epsilon_1 = 0$ and $\epsilon_1 > 0$. The

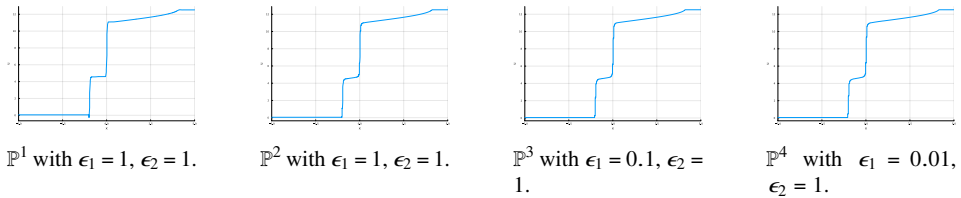


Figure 9. Solutions of (Scheme D) for $k = 1, 2, 3, 4$ and $r = k + 2$. We still use a uniform mesh with $N = 275$. Now we have convergence to the correct solution.

reference solution is computed by the fifth order finite difference WENO method with $N = 4096$.

Example 8.4. Consider the Shu-Osher problem for one-dimensional Euler equations defined on $I = [-5, 5]$. The initial condition is

$$(\rho, u, p) = \begin{cases} (3.857143, 2.629369, 10.33333), & x < -4, \\ (1 + 0.2 \sin(5x), 0, 1), & x \geq -4. \end{cases}$$

The fixed post-shock state is imposed at the left boundary, and outflow boundary conditions are used on the right. The heat capacity ratio is taken as $\gamma = 1.4$. We use a uniform mesh of size $N = 400$. The solution is computed up to $T = 1.8$ using (Scheme D') with $k = 1, 2, 3$ and $r = k + 1$. The local Lax–Friedrichs flux and SSP–RK3 were used in all tests. We set $\epsilon_2 = 10$ for all tests. In Figure 11, the solutions on the left are obtained with $\epsilon_1 = 0$ for all $k = 1, 2, 3$. On the right, we set $\epsilon_1 = 2$ for $k = 1$, $\epsilon_1 = 1$ for $k = 2$, and $\epsilon_1 = 0.01$ for $k = 3$. Again, no huge difference can be noted between $\epsilon_1 = 0$ and $\epsilon_1 > 0$. The reference solution is computed by the fifth order finite difference WENO method with $N = 4096$.

Example 8.5. In this test, we compute the double Mach reflection problem designed by [20]. It involves a Mach 10 shock impinging on a reflecting wall at a 60° angle. The computational domain is $\Omega = [0, 4] \times [0, 1]$, with the initial condition

$$(\rho_0, u_0, v_0, p_0) = \begin{cases} (8, \frac{33\sqrt{3}}{8}, -\frac{33}{8}, 116.5), & \text{if } y > \sqrt{3}(x - \frac{1}{6}) \quad (\text{post-shock}) \\ (1.4, 0, 0, 1), & \text{if } y < \sqrt{3}(x - \frac{1}{6}) \quad (\text{pre-shock}). \end{cases}$$

The inflow and outflow boundary conditions are prescribed at the left and right boundary, respectively. The top boundary values are set to equal the exact solution, which is time-dependent. On the bottom boundary, the (constant) post shock values are prescribed on the segment $\{0 \leq x < \frac{1}{6}, y = 0\}$, while reflective boundary conditions are imposed on the segment $\{\frac{1}{6} < x \leq 4, y = 0\}$. We use a rectangular mesh of size $h_x = h_y = \frac{1}{240}$. The numerical solution is computed up to $T = 0.2$ using (Scheme D'') with $\epsilon_1 = 0$, $\epsilon_2 = 0.4$. SSP–RK3 is used for time integration. No positivity-preserving limiter is needed. The computed values of ρ and p are shown in Figure 12. One can see that the scheme already has enough viscosity even without the flux derivative term in the damping coefficient D_K'' . In addition, the results are very similar to those of the OFDG scheme from [11], although the damping coefficients are not exactly the same.

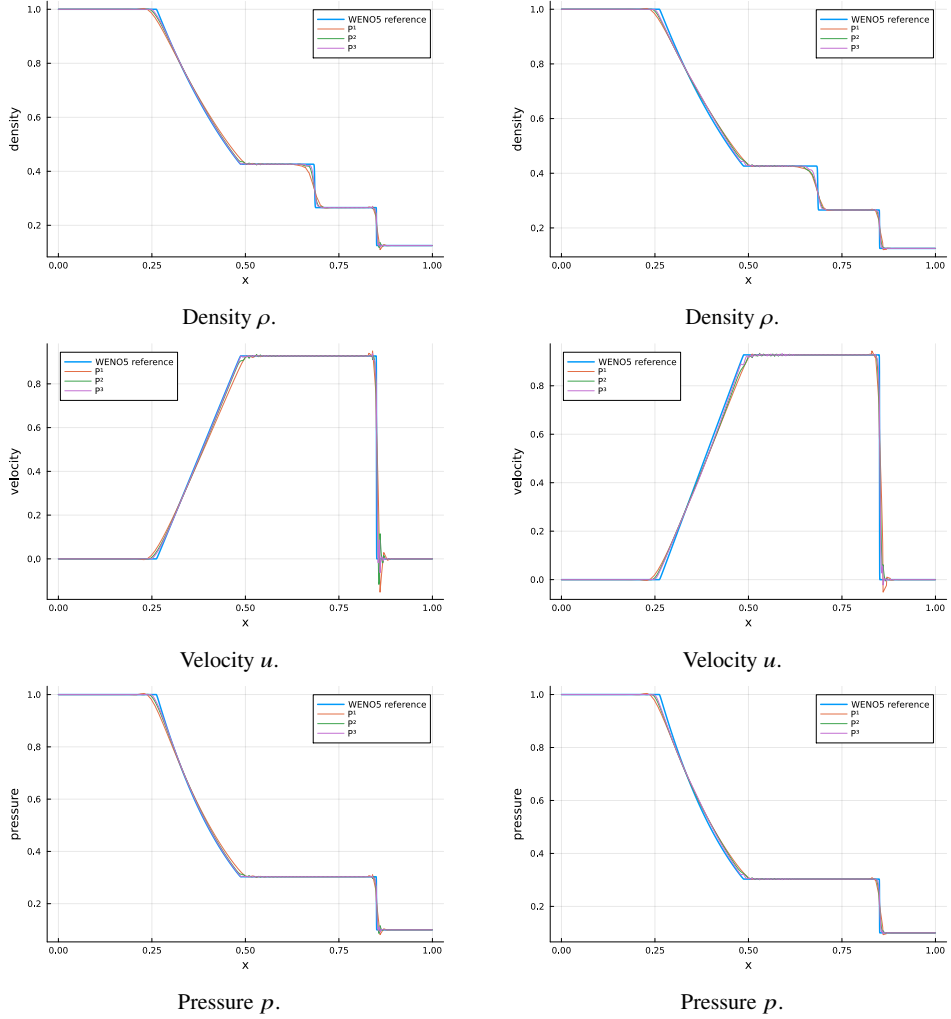


Figure 10. Solutions of (Scheme D') with $k = 1, 2, 3$ and $r = k + 1$. We use a uniform mesh with $N = 100$. We set $\epsilon_2 = 10$ for all tests. The solutions on the left are obtained with $\epsilon_1 = 0$ for all $k = 1, 2, 3$. On the right, we set $\epsilon_1 = 10$ for $k = 1$, $\epsilon_1 = 1$ for $k = 2$, and $\epsilon_1 = 0.01$ for $k = 3$.

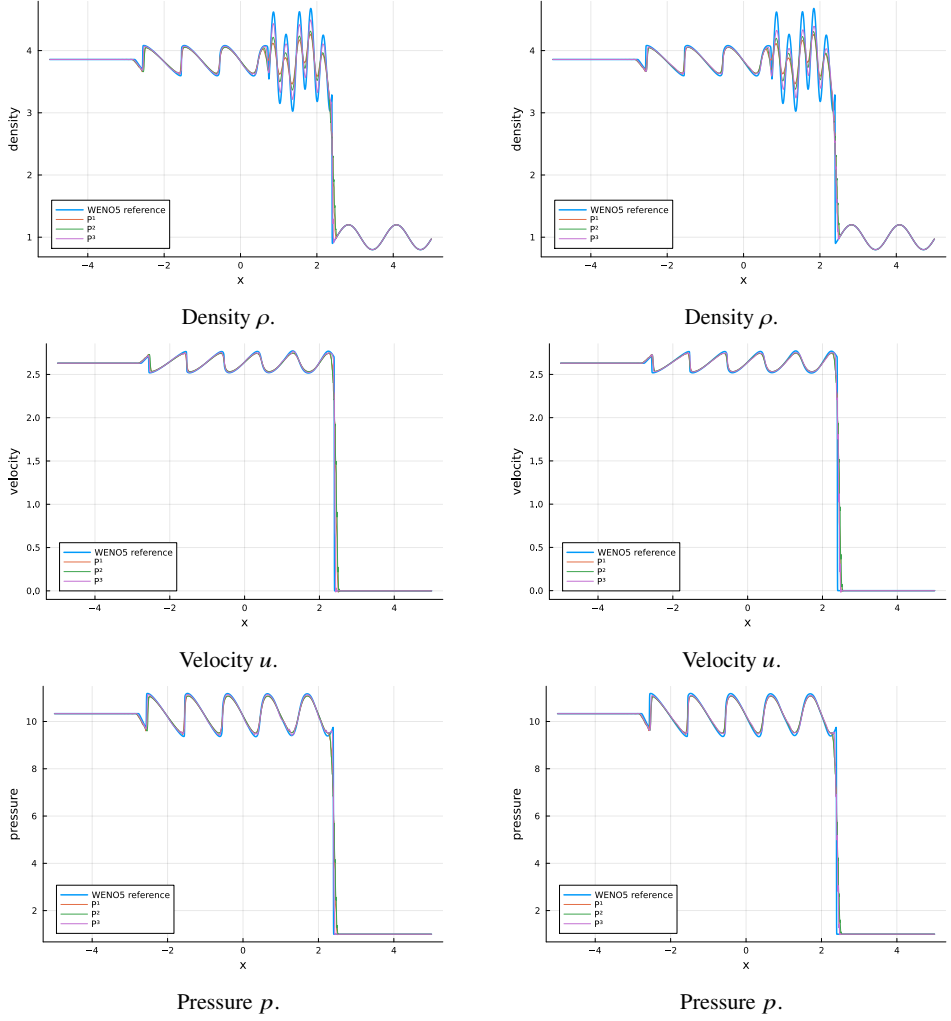


Figure 11. Solutions of (Scheme D') with $k = 1, 2, 3$ and $r = k + 1$. We use a uniform mesh with $N = 400$. We set $\epsilon_2 = 10$ for all tests. The solutions on the left are obtained with $\epsilon_1 = 0$ for all $k = 1, 2, 3$. On the right, we set $\epsilon_1 = 2$ for $k = 1$, $\epsilon_1 = 1$ for $k = 2$, and $\epsilon_1 = 0.01$ for $k = 3$.

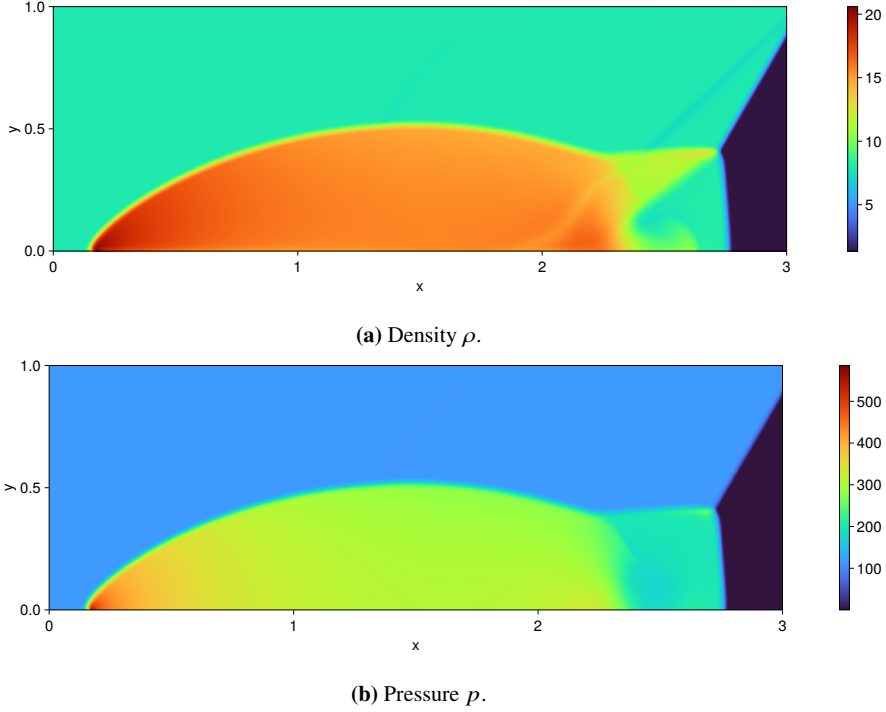


Figure 12. The numerical solution at $T = 0.2$ using (Scheme D'') with $\epsilon_1 = 0$, $\epsilon_2 = 0.4$.

Example 8.6. In this test, we compute a Mach 10 shock diffracting at a 120° degree [1, 22]. The computational domain and the unstructured triangular mesh with $h = \frac{1}{4}$ are presented in Figure 13. The shock is initially located at $x = 3.4$ and $6 \leq y \leq 11$, moving into undisturbed air with a density of 1.4 and a pressure of 1. Boundary conditions are inflow at the left/top boundary (in accordance with the exact shock motion), and outflow at the right/bottom boundary. The numerical solution is computed up to $T = 0.9$ using (Scheme D'') with $\epsilon_1 = 0$, $\epsilon_2 = 0.4$ and $h = \frac{1}{40}$. We use the third-order modified exponential Runge–Kutta method from [11]. No positivity-preserving limiter is needed. The computed values of ρ and p are shown in Figure 14. Again, the scheme already has enough viscosity without the flux derivative term in the damping coefficient D_K'' , and the performance is very close to the OFDG scheme from [11].

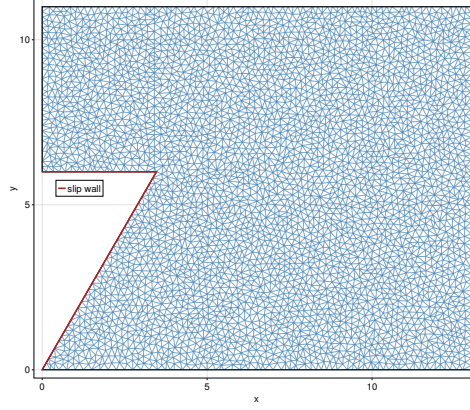


Figure 13. Illustration of the computational domain and the triangular mesh with $h = \frac{1}{4}$.

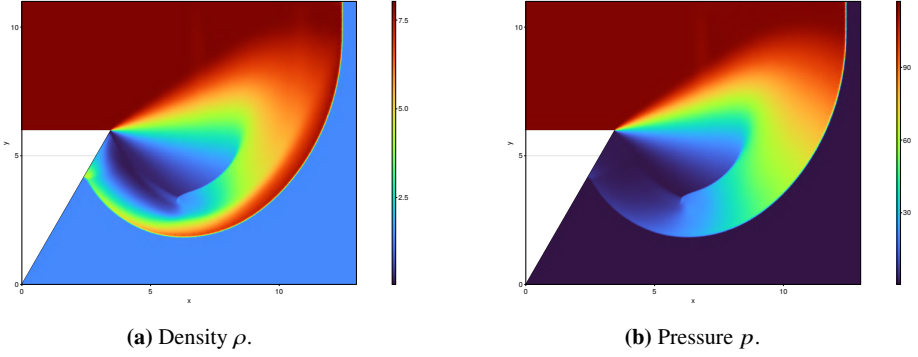


Figure 14. The numerical solution at $T = 0.9$ using (Scheme D'') with $\epsilon_1 = 0$, $\epsilon_2 = 0.4$ and $h = \frac{1}{40}$.

9. Concluding remarks

In this paper, we have introduced an operator framework for constructing high-order DG methods for nonlinear hyperbolic conservation laws. The framework first gives an alternative derivation of the classical semi-discrete DG method from the infinitesimal action of an evolution operator followed by the L^2 -projection. Replacing the usual linear projection by an entropy-contracting nonlinear projection then produces the Operator DG methods studied here. The resulting correction to the classical DG scheme has a particularly simple form: it damps only the nonconstant polynomial modes within each cell and therefore preserves cell averages and conservation, much like the OFDG schemes. This construction led to the schemes (Scheme B)–(Scheme D) for one-dimensional scalar equations, their counterparts (Scheme B')–(Scheme D') for one-dimensional systems, and the multidimensional schemes (Scheme B'') and (Scheme D'') on general convex, shape-regular meshes.

For 1D scalar conservation laws, the damping coefficients can be chosen so that a single numerical solution satisfies infinitely many local entropy inequalities when an E-flux is used. For systems, the same mechanism is compatible with a prescribed strictly convex entropy and an entropy-stable numerical flux. The analysis also shows that the additional entropy dissipation does not destroy high-order accuracy: for smooth 1D scalar solutions, we proved theoretically the optimal L^2 error bound of order $k + 1$ with an upwind flux and order $k + \frac{1}{2}$ with a general monotone flux. For discontinuous solutions of 1D scalar conservation laws with a strictly convex flux, the entropy dissipation supplies the compactness estimates needed in the compensated-compactness argument. Under the stated uniform boundedness assumption, the numerical solutions converge strongly to the unique entropy solution. Thus, the damping term is not merely a device for suppressing oscillations; it also provides the analytical mechanism that links the high-order approximation to the correct entropy solution.

The numerical experiments support these conclusions and clarify the roles of the two parts of the practical damping coefficient D_j . The smooth one- and two-dimensional tests exhibit the expected high-order accuracy. For the non-convex scalar problems, using only the interface-jump contribution can fail to approximate the correct entropy solution when the amount of damping is moderate, whereas restoring a suitable flux-derivative contribution recovers the correct solution. At the same time, the Sod, Shu–Osher, double Mach reflection, and shock-diffraction computations show that the simplified system schemes can robustly resolve strong shocks and fine-scale structures. In the multidimensional Euler tests, the interface contribution alone was sufficient for the chosen parameters, and no positivity-preserving limiter was required.

Several questions remain open. A fully discrete entropy and convergence theory that includes the time integrator would complement the present semi-discrete analysis. It would also be desirable to relax the uniform L^∞ assumption in the compensated-compactness argument, to extend strong-convergence and optimal-error results to systems and multiple space dimensions, and to understand entropy selection in order to ensure the convergence for general non-convex conservation laws. Nonlinear projections other than $\tilde{\pi}$ need to be explored for generating schemes with additional properties; the operator framework also needs to be extended to other types of PDEs. On the computational side, efficient evaluation or approximation of the flux-derivative term, and adaptive selection of ϵ_1 and ϵ_2 are natural directions for making the Operator DG schemes more robust and economical for large-scale applications.

Disclosure on the use of artificial intelligence

During the preparation of this work, ChatGPT was used for (1) improving the readability and language quality; (2) assisting with the proof of compensated compactness presented in Section 7, in particular the proof of Lemma 7.3; (3) programming the Julia codes for numerical tests; (4) computing the optimal coefficients C_k , $C(k, r)$ and $\mathcal{A}_k$ recorded in

Tables 1, 2 and B.1. All mathematical derivations, results, and proofs were independently verified and validated by the human authors.

Funding. This research was supported in part by NSF grant DMS-2309249.

References

- [1] Tianheng Chen and Chi-Wang Shu, *Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws*, Journal of Computational Physics **345** (2017), 427–461.
- [2] B. Cockburn, F. Coquel, and P. G. LeFloch, *Convergence of the finite volume method for multidimensional conservation laws*, SIAM Journal on Numerical Analysis **32** (1995), no. 3, 687–705.
- [3] Bernardo Cockburn, Suchung Hou, and Chi-Wang Shu, *The runge-kutta local projection discontinuous galerkin finite element method for conservation laws. iv: The multidimensional case*, Mathematics of Computation **54** (1990), no. 190, 545–581.
- [4] Constantine M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, 2 ed., pp. 87–120, 136–141, 511–519, Springer-Verlag, Berlin Heidelberg, 2005.
- [5] Ronald J. DiPerna, *Convergence of approximate solutions to conservation laws*, Archive for Rational Mechanics and Analysis **82** (1983), no. 1, 27–70.
- [6] Edwige Godlewski and Pierre-Arnaud Raviart, *Hyperbolic Systems of Conservation Laws*, pp. 1–48, Ellipses, 1991.
- [7] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner, *Solving Ordinary Differential Equations I Nonstiff Problems*, 2 ed., pp. 149–150, Springer-Verlag, Berlin Heidelberg, 1993.
- [8] Amiram Harten, James M. Hyman, Peter D. Lax, and Barbara Keyfitz, *On finite-difference approximations and entropy conditions for shocks*, Communications on Pure and Applied Mathematics **29** (1976), no. 3, 297–322.
- [9] Guangshan Jiang and Chi-Wang Shu, *On a cell entropy inequality for discontinuous Galerkin methods*, Mathematics of Computation **62** (1994), no. 206, 531–538.
- [10] Randall J. LeVeque, *Numerical methods for conservation laws*, pp. 129–133, Birkhäuser Basel, 1992.
- [11] Yong Liu, Jianfang Lu, and Chi-Wang Shu, *An essentially oscillation-free discontinuous galerkin method for hyperbolic systems*, SIAM Journal on Scientific Computing **44** (2022), no. 1, A230–A259.
- [12] Yuchang Liu, Wei Guo, Yan Jiang, and Zheng Sun, *A limiter-based approach to construct high-order fully-discrete entropy stable explicit dg schemes for hyperbolic conservation laws*, arXiv preprint arXiv:2603.27478 (2026).
- [13] Jianfang Lu, Yong Liu, and Chi-Wang Shu, *An oscillation-free discontinuous galerkin method for scalar hyperbolic conservation laws*, SIAM Journal on Numerical Analysis **59** (2021), no. 3, 1299–1324.
- [14] François Murat, *Compacité par compensation*, Annali della Scuola Normale Superiore di Pisa, Classe di Scienze **5** (1978), no. 3, 489–507.
- [15] Stanley Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM Journal on Numerical Analysis **21** (1984), no. 2, 217–235.

- [16] Jing-Mei Qiu and Chi-Wang Shu, *Convergence of high order finite volume weighted essentially nonoscillatory scheme and discontinuous galerkin method for nonconvex conservation laws*, SIAM Journal on Scientific Computing **31** (2008), no. 1, 584–607.
- [17] J. Michael Steele, *The Cauchy-Schwarz Master Class: An introduction to the art of mathematical inequalities*, pp. 94–96, Cambridge University Press, Cambridge, 2004.
- [18] Luc Tartar, *Compensated compactness and applications to partial differential equations*, Non-linear Analysis and Mechanics: Heriot-Watt Symposium, vol. 4, Research Notes in Mathematics, no. 39, Pitman, Boston, 1979, pp. 136–212.
- [19] Yuanzhe Wei and Chi-Wang Shu, *L^2 -stability of explicit second order Runge-Kutta discontinuous Galerkin method for nonlinear conservation laws*, Journal of Scientific Computing **105** (2025), 12.
- [20] Paul Woodward and Phillip Colella, *The numerical simulation of two-dimensional fluid flow with strong shocks*, Journal of Computational Physics **54** (1984), no. 1, 115–173.
- [21] Qiang Zhang and Chi-Wang Shu, *Error estimates to smooth solutions of runge-kutta discontinuous galerkin methods for scalar conservation laws*, SIAM Journal on Numerical Analysis **42** (2005), no. 2, 641–666.
- [22] Xiangxiong Zhang, Yinhua Xia, and Chi-Wang Shu, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, Journal of Scientific Computing **50** (2012), no. 1, 29–62.

Appendix A Operator for generating the classical DG method with HLL flux

Although Theorem 2.1 holds only for the Godunov flux, the statement can actually be generalized to the class of HLL fluxes, which includes the local Lax–Friedrichs flux. Previously, we assumed $S_{\Delta t}$ as the true solution operator, which would produce the exact Godunov flux in the expressions (2.12) and (2.13). Now, the idea is to construct a modified version of S_t , denoted by S_t^{HLL} , such that it satisfies the following three properties:

- (a.1) For sufficiently small $\Delta t > 0$, $S_{\Delta t}^{\text{HLL}}u = S_{\Delta t}u$ on the interior of each I_j , except for two small regions at the end-points with length $O(\Delta t)$.
- (a.2) (2.12) and (2.13) now gives the expressions $\hat{f}_{j-\frac{1}{2}}^{\text{HLL}} - f(u)_{j-\frac{1}{2}}^+$ and $-\hat{f}_{j+\frac{1}{2}}^{\text{HLL}} + f(u)_{j+\frac{1}{2}}^-$, where $\hat{f}^{\text{HLL}}$ is the HLL flux.
- (a.3) Under periodic or compactly supported boundary conditions, $S_{\Delta t}^{\text{HLL}}$ decreases all entropies: for any convex function $U(\cdot)$,

$$\int_I U(S_{\Delta t}^{\text{HLL}}u(x)) \, dx \leq \int_I U(u(x)) \, dx, \quad \forall \Delta t \geq 0, \quad u \in \mathbb{V}^k.$$

This is a property shared by $S_{\Delta t}$.

Clearly, if such an $S_{\Delta t}^{\text{HLL}}$ can be found, the proof of Theorem 2.1 would still go through with $S_{\Delta t}$ replaced by $S_{\Delta t}^{\text{HLL}}$ and the Godunov flux replaced by $\hat{f}^{\text{HLL}}$. Moreover, if $S_{\Delta t}^{\text{HLL}}$ is locally defined, we may expect the resulting scheme to satisfy certain local entropy inequalities on account of (a.3). We now proceed to construct such an $S_{\Delta t}^{\text{HLL}}$.

As in the proof of Theorem 2.1, suppose $u \in \mathbb{V}^k$. Let $S_{\Delta t}u$ denote the true solution of the conservation law with the initial condition u after evolving a time $\Delta t > 0$ sufficiently small, such that the evolutions of discontinuities at different cell interfaces don't interact, and no new discontinuity in u and its derivatives is formed. Let λ^- and λ^+ denote the lower and upper bounds for all wave speeds of the Riemann solution with left and right states u^- and u^+ . The HLL flux is defined by:

$$\hat{f}^{\text{HLL}}(u^-, u^+) := \begin{cases} f(u^-) = f(u^+), & a^- = a^+ = 0 \\ \frac{a^+ f(u^-) - a^- f(u^+) + a^- a^+ (u^+ - u^-)}{a^+ - a^-}, & \text{otherwise,} \end{cases} \quad (\text{A.1})$$

where $a^- = \min\{0, \lambda^-\} \leq 0$ and $a^+ = \max\{0, \lambda^+\} \geq 0$. The local Lax–Friedrichs flux is a special case by choosing $-\lambda^- = \lambda^+$. Let $I_j^\pm(\Delta t)$ be the same as in (2.6), and define

$$I_{j-\frac{1}{2}}(\Delta t) := I_{j-1}^-(\Delta t) \cup I_j^+(\Delta t). \quad (\text{A.2})$$

In each $I_{j-\frac{1}{2}}(\Delta t)$, we replace the value of $S_{\Delta t}u$ by its average over $I_{j-\frac{1}{2}}(\Delta t)$. We call this modified function $S_{\Delta t}^{\text{HLL}}u$. Namely,

$$S_{\Delta t}^{\text{HLL}}u(x) := \begin{cases} \frac{1}{|I_{j-\frac{1}{2}}(\Delta t)|} \int_{I_{j-\frac{1}{2}}(\Delta t)} S_{\Delta t}u(x) dx, & \text{if } |I_{j-\frac{1}{2}}(\Delta t)| > 0 \text{ and } x \in I_{j-\frac{1}{2}}(\Delta t) \\ S_{\Delta t}u(x), & \text{otherwise.} \end{cases} \quad (\text{A.3})$$

Since the intervals (A.2) are disjoint for sufficiently small Δt , this defines $S_{\Delta t}^{\text{HLL}}u$ unambiguously.

With $S_{\Delta t}^{\text{HLL}}$ thus defined, (a.1) is easily verified, since we only modified $S_{\Delta t}u$ in the region $I_{j-\frac{1}{2}}(\Delta t)$ with length $O(\Delta t)$ around each interface. $S_{\Delta t}^{\text{HLL}}$ also satisfies (a.3), since taking average can only decrease the total entropy by Jensen's inequality. Moreover, $S_{\Delta t}^{\text{HLL}}$ satisfies (a.2), as can be seen in the proof of the following analogue of Theorem 2.1:

Theorem A.1. *For any $u \in \mathbb{V}^k$, we have*

$$\lim_{\Delta t \rightarrow 0} \frac{\pi S_{\Delta t}^{\text{HLL}}u - u}{\Delta t} = F[u], \quad (\text{A.4})$$

where $F[\cdot] \in \mathbb{V}^k$ is from the right-hand side of (2.2) with $\hat{f}^{\text{HLL}}$.

Proof. Fix an arbitrary j . Following the proof of Theorem 2.1, we first obtain

$$\begin{aligned} \lim_{\Delta t \rightarrow 0} \int_{I_j} \frac{\pi S_{\Delta t}^{\text{HLL}}u(x) - u(x)}{\Delta t} \cdot v(x) dx &= - \int_{I_j} f(u)_x \cdot v(x) dx + v_{j-\frac{1}{2}}^+ \lim_{\Delta t \rightarrow 0} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t}^{\text{HLL}}u(x) - u(x)}{\Delta t} dx \\ &\quad + v_{j+\frac{1}{2}}^- \lim_{\Delta t \rightarrow 0} \int_{I_j^-(\Delta t)} \frac{S_{\Delta t}^{\text{HLL}}u(x) - u(x)}{\Delta t} dx \end{aligned} \quad (\text{A.5})$$

for any $v \in \mathbb{V}^k$. Up to $O(\Delta t)$, the value of $S_{\Delta t}u$ near $x_{j-\frac{1}{2}}$ can be approximated by the Riemann solution $S_{\Delta t}\tilde{u}$ defined in (2.8). Let $\lambda_{j-\frac{1}{2}}^-$ and $\lambda_{j-\frac{1}{2}}^+$ denote the lower and upper bounds for all wave speeds of this Riemann solution. Let $a^- = \min\{0, \lambda_{j-\frac{1}{2}}^-\} \leq 0$ and $a^+ = \max\{0, \lambda_{j-\frac{1}{2}}^+\} \geq 0$. If $(a^+ - a^-) \neq 0$, by definition (A.3) we have

$$\int_{I_j^+(\Delta t)} S_{\Delta t}^{\text{HLL}} u(x) dx = \frac{a^+}{a^+ - a^-} \int_{I_{j-\frac{1}{2}}^-(\Delta t)} S_{\Delta t} u(x) dx = \frac{a^+}{a^+ - a^-} \int_{I_{j-\frac{1}{2}}^-(\Delta t)} S_{\Delta t} \tilde{u}(x) dx + O(\Delta t^2),$$

where $\tilde{u}(x)$ is defined in (2.8). Hence,

$$\begin{aligned} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t}^{\text{HLL}} u(x) - u(x)}{\Delta t} dx &= \frac{a^+}{a^+ - a^-} \int_{I_{j-\frac{1}{2}}^-(\Delta t)} \frac{S_{\Delta t} \tilde{u}(x) - u_{j-\frac{1}{2}}^+}{\Delta t} dx + O(\Delta t) \\ &= \frac{a^+}{a^+ - a^-} \int_{a^-}^{a^+} R(\xi) - u_{j-\frac{1}{2}}^+ d\xi + O(\Delta t). \end{aligned} \quad (\text{A.6})$$

where $R(\xi)$ is from (2.9). Integrating the conservation law for the self-similar Riemann solution (2.11) yields

$$\int_{a^-}^{a^+} R(\xi) d\xi = a^+ u_{j-\frac{1}{2}}^+ - a^- u_{j-\frac{1}{2}}^- - f(u_{j-\frac{1}{2}}^+) + f(u_{j-\frac{1}{2}}^-).$$

Applying this relation in (A.6), we obtain

$$\lim_{\Delta t \rightarrow 0} \int_{I_j^+(\Delta t)} \frac{S_{\Delta t}^{\text{HLL}} u(x) - u(x)}{\Delta t} dx = \hat{f}_{j-\frac{1}{2}}^{\text{HLL}} - f(u_{j-\frac{1}{2}}^+), \quad (\text{A.7})$$

where $\hat{f}_{j-\frac{1}{2}}^{\text{HLL}} = \hat{f}_{j-\frac{1}{2}}^{\text{HLL}}(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+)$ is the HLL flux from (A.1). If $(a^+ - a^-) = 0$, namely $a^- = a^+ = 0$, then (A.7) holds trivially since both sides are zero. In the same way, we can also prove

$$\lim_{\Delta t \rightarrow 0} \int_{I_j^-(\Delta t)} \frac{S_{\Delta t}^{\text{HLL}} u(x) - u(x)}{\Delta t} dx = -\hat{f}_{j+\frac{1}{2}}^{\text{HLL}} + f(u_{j+\frac{1}{2}}^-). \quad (\text{A.8})$$

The theorem then follows from applying (A.7) and (A.8) in (A.5) and integrate by parts. As a remark, (A.7) and (A.8) imply that $S_{\Delta t}^{\text{HLL}}$ satisfies (a.2), which enables us to extend the result of Theorem 2.1 to the HLL flux. $\blacksquare$

By property (a.3), Corollary 2.2 still holds with Godunov flux replaced by $\hat{f}^{\text{HLL}}$, without any change to the proof. In addition, since on each cell I_j , the scheme only uses local information from I_j and its two neighboring cells, the local square-entropy inequality also holds for $\hat{f}^{\text{HLL}}$, see [9] or Theorem 3.9. Finally, Corollary 2.3 can also be shown to hold in this case.

Appendix B Improved estimate of β_k from Lemma 3.4

In this section, we improve the estimate on the constant β_k from Lemma 3.4. The bound $\beta_k \leq (k+1)^4$ obtained in the proof of Lemma 3.4 is the product of two separate sharp inequalities, but the corresponding optimizers do not coincide. A better estimate is obtained by using one single inequality.

Let $\xi \in [-1, 1]$ be the reference coordinate on I_j , namely

$$x = x_j + \frac{h_j}{2}\xi,$$

and let P_i denote the standard Legendre polynomial of degree i , normalized by $P_i(1) = 1$. The reproducing kernel for evaluation at the left endpoint of the L^2 -projection onto $\mathbb{P}^k([-1, 1])$ is

$$K_k^-(\xi) := \sum_{i=0}^k \frac{2i+1}{2} (-1)^i P_i(\xi). \quad (\text{B.1})$$

Indeed, for every $g \in L^2(-1, 1)$,

$$(\hat{\pi}_k g)(-1) = \int_{-1}^1 K_k^-(\xi) g(\xi) d\xi, \quad (\text{B.2})$$

where $\hat{\pi}_k$ is the L^2 -projection onto $\mathbb{P}^k([-1, 1])$.

We define

$$\mathcal{A}_k := \sup_{\substack{p \in \mathbb{P}^k([-1, 1]) \\ \int_{-1}^1 p(\xi) d\xi = 0, p \neq 0}} \frac{\int_{-1}^1 |K_k^-(\xi)| |p(\xi) - p(-1)|^2 d\xi}{\int_{-1}^1 |p(\xi)|^2 d\xi}. \quad (\text{B.3})$$

Since the admissible space in (B.3) is finite dimensional, the supremum is attained.

Proposition B.1. *For $k \geq 1$, the constant in Lemma 3.4 may be chosen so that*

$$\beta_k \leq 2\mathcal{A}_k. \quad (\text{B.4})$$

The same constant applies to the left and right endpoints.

Proof. We prove the estimate at $x_{j-\frac{1}{2}}^+$; the other endpoint follows by symmetry. Set

$$p(\xi) := u\left(x_j + \frac{h_j}{2}\xi\right) - \bar{u}_j.$$

Then $p \in \mathbb{P}^k([-1, 1])$ and

$$\int_{-1}^1 p(\xi) d\xi = 0.$$

Moreover,

$$u(x) - u_{j-\frac{1}{2}}^+ = p(\xi) - p(-1).$$

Let

$$M_j := \max_{x \in I_j} |U'''(u(x))|.$$

Taylor's theorem gives, pointwise on I_j ,

$$\left| U'(u) - U'(u_{j-\frac{1}{2}}^+) - U''(u_{j-\frac{1}{2}}^+)(u - u_{j-\frac{1}{2}}^+) \right| \leq \frac{M_j}{2} |u - u_{j-\frac{1}{2}}^+|^2.$$

Again, we use the fact that $U''(u_{j-\frac{1}{2}}^+)(u - u_{j-\frac{1}{2}}^+) \in \mathbb{P}^k(I_j)$ and vanishes at $x_{j-\frac{1}{2}}^+$, so its projected value at that endpoint is zero. After mapping to $[-1, 1]$ and using (B.2), we therefore obtain

$$\left| \pi \circ [U'(u) - U'(u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+) \right| \leq \frac{M_j}{2} \int_{-1}^1 |K_k^-(\xi)| |p(\xi) - p(-1)|^2 d\xi \leq \frac{M_j \mathcal{A}_k}{2} \int_{-1}^1 |p(\xi)|^2 d\xi.$$

The change of variables gives

$$\int_{-1}^1 |p(\xi)|^2 d\xi = \frac{2}{h_j} \int_{I_j} |u(x) - \bar{u}_j|^2 dx.$$

Consequently,

$$\left| \pi \circ [U'(u) - U'(u_{j-\frac{1}{2}}^+)](x_{j-\frac{1}{2}}^+) \right| \leq \frac{\mathcal{A}_k}{h_j} M_j \int_{I_j} |u(x) - \bar{u}_j|^2 dx.$$

Comparison with the normalization in Lemma 3.4 yields $\beta_k \leq 2\mathcal{A}_k$.

For the right endpoint, the corresponding kernel is

$$K_k^+(\xi) = \sum_{i=0}^k \frac{2i+1}{2} P_i(\xi) = K_k^-(-\xi).$$

The transformation $p(\xi) \mapsto p(-\xi)$ preserves both the zero-average condition and the L^2 norm, and maps the right-endpoint quotient onto the quotient in (B.3). Hence the same constant $\mathcal{A}_k$ applies. $\blacksquare$

The constant $\mathcal{A}_k$ can be computed from a $k \times k$ generalized eigenvalue problem. Since p has zero average, write

$$p(\xi) = \sum_{m=1}^k a_m P_m(\xi), \quad \mathbf{a} = (a_1, \dots, a_k)^T.$$

Then

$$p(\xi) - p(-1) = \sum_{m=1}^k a_m (P_m(\xi) - (-1)^m).$$

Define the symmetric matrices $M^{(k)}$ and $G^{(k)}$ by

$$M_{mn}^{(k)} := \int_{-1}^1 |K_k^-(\xi)| (P_m(\xi) - (-1)^m) (P_n(\xi) - (-1)^n) d\xi, \quad (\text{B.5})$$

$$G_{mn}^{(k)} := \frac{2}{2m+1} \delta_{mn}, \quad 1 \leq m, n \leq k. \quad (\text{B.6})$$

By Legendre orthogonality,

$$\int_{-1}^1 |p(\xi)|^2 d\xi = \mathbf{a}^T G^{(k)} \mathbf{a},$$

whereas the numerator in (B.3) is $\mathbf{a}^T M^{(k)} \mathbf{a}$. Therefore

$$\mathcal{A}_k = \max_{\mathbf{a} \neq 0} \frac{\mathbf{a}^T M^{(k)} \mathbf{a}}{\mathbf{a}^T G^{(k)} \mathbf{a}} = \lambda_{\max}(M^{(k)}, G^{(k)}). \quad (\text{B.7})$$

Thus $2\lambda_{\max}(M^{(k)}, G^{(k)})$ is an explicit computable upper bound for β_k . Solving the symmetric generalized eigenvalue problem (B.7) then gives the values in Table B.1.

k	$\mathcal{A}_k$	$2\mathcal{A}_k$	$(k+1)^4$
1	$\frac{59}{27} \approx 2.1851851852$	4.3703703704	16
2	6.0975144965	12.1950289930	81
3	14.3482787131	28.6965574262	256
4	26.3960313043	52.7920626087	625
5	44.2424749310	88.4849498620	1296
6	67.2523038234	134.5046076468	2401
7	97.2694197715	194.5388395430	4096

Table B.1. Improved computable bounds $\beta_k \leq 2\mathcal{A}_k$ from (B.4), compared with the bound $(k+1)^4$ used in Lemma 3.4.

Yuanzhe Wei

Division of Applied Mathematics, Brown University, Box F, 182 George Street, Providence, RI, 02912, USA; yuanzhe_wei@brown.edu

Chi-Wang Shu

Division of Applied Mathematics, Brown University, Box F, 182 George Street, Providence, RI, 02912, USA; chi-wang_shu@brown.edu

Author IDs: zbMATH [shu.chi-wang](https://zbmath.org/authors/?id=shu.chi-wang) MR 242268 ORCID 0000-0001-7720-9564