Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Total variation diminishing schemes in optimal control of scalar conservation laws

Soheil Hajian[1], Michael Hintermüller[1,2], Stefan Ulbrich[3]

submitted: March 30, 2017

[1] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
E-Mail: soheil.hajian@hu-berlin.de
hint@math.hu-berlin.de

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: michael.hintermueller@wias-berlin.de

[3] Technische Universität Darmstadt
Fachbereich Mathematik
Dolivostr. 15
64295 Darmstadt
Germany
E-Mail: ulbrich@mathematik.tu-darmstadt.de

No. 2383
Berlin 2017

2010 *Mathematics Subject Classification.* 49J20, 65M12, 65K10.

Key words and phrases. Optimal control of PDEs, adjoint equation, scalar conservation laws, TVD Runge-Kutta methods.

This research was supported by the German Research Foundation DFG through the SFB-TRR 154 and by the Research Center MATHEON through project C-SE5 and D-OT1 funded by the Einstein Center for Mathematics Berlin.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Total variation diminishing schemes in optimal control of scalar conservation laws

Soheil Hajian, Michael Hintermüller, Stefan Ulbrich

Abstract

In this paper, optimal control problems subject to a nonlinear scalar conservation law are studied. Such optimal control problems are challenging both at the continuous and at the discrete level since the control-to-state operator poses difficulties as it is, e.g., not differentiable. Therefore discretization of the underlying optimal control problem should be designed with care. Here the discretize-then-optimize approach is employed where first the full discretization of the objective function as well as the underlying PDE is considered. Then, the derivative of the reduced objective is obtained by using an adjoint calculus. In this paper total variation diminishing Runge-Kutta (TVD-RK) methods for the time discretization of such problems are studied. TVD-RK methods, also called strong stability preserving (SSP), are originally designed to preserve total variation of the discrete solution. It is proven in this paper that providing an SSP state scheme, is enough to ensure stability of the discrete adjoint. However requiring SSP for both discrete state and adjoint is too strong. Also approximation properties that the discrete adjoint inherits from the discretization of the state equation are studied. Moreover order conditions are derived. In addition, optimal choices with respect to CFL constant are discussed and numerical experiments are presented.

1 Introduction

We are considering an optimal control problem of type

$$\min_{u\in U_{\mathsf{ad}}} \int_{\mathbb{R}} G(y(x,T))\,\mathrm{d}x,$$

subject to (s.t.) a scalar conservation law, i.e.,

$$\begin{aligned} y_t + f(y)_x &= 0 && \text{in } \mathbb{R}\times\mathbb{R}_+,\\ y(x,0) &= u(x) && \text{in } \mathbb{R}. \end{aligned}$$

Here U_{ad} is called the admissible set and it is assumed to be non-empty, convex and closed. the state, $y(x,t)$, is considered to be the entropic (weak) solution of the scalar conservation law and $u(x)$, the control, is the initial data of the partial differential equation (PDE).

Although the definition of the optimal control problem seems simple, the PDE constraint (conservation law) poses severe difficulties for the analysis of such problems, both at the continuous and at the discrete level. The major problem is the possible formation of a shock in the state $y(x,t)$ at finite time even for very smooth initial data $u(x)$, when the flux function $f(\cdot)$ is non-linear. Moreover it is easy to show through examples that the control-to-state map is not Gâteaux differentiable when shocks are present. This poses a significant problem for obtaining the derivative of the cost functional of a (control) reduced version of the underlying optimal control problem. Luckily a generalized definition of the derivative, called “shift derivative”, for the control-to-state map has been derived by S. Ulbrich

DOI 10.20347/WIAS.PREPRINT.2383 Berlin 2017

[Ulb01, Ulb02] which implies Fréchet differentiability of the cost functional. Such differentiability results enable us to compute the derivative of the cost functional using an adjoint approach. In Section 2, we state the underlying optimal control in a rigorous context by recalling weak and entropic solutions of scalar conservation laws and their properties as well as the concept of shift differentiability.

The difficulties that arise from the nature of the PDE is also reflected at the discrete level, i.e., one should discretize the problem with care. Monotone schemes, that we recall in Section 3, are among successful discretizations for conservation laws and their theory is well-understood. We use monotone discretizations in space to obtain a semi-discrete formulation and then we discretize in time using a total variation diminishing (TVD) Runge-Kutta (RK) scheme. TVD-RK methods are a class of RK methods that guarantee, under quite mild assumption, that the discrete solution is total variation stable (also called "strongly stability preserving (SSP)" methods). We then obtain the fully discrete optimal control problem by discretizing the objective functional.

Similar to the continuous level, one can obtain the derivative of the cost functional using the adjoint calculus. The properties of the discrete adjoint are intimately related to the discretization of the discrete state. The TVD-RK method for the discrete state can be characterized by two sets of coefficients $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$. We show that the corresponding discrete adjoint is also obtained by a TVD-RK method where the coefficients are "conjugates" of $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$. Therefore we will study in Section 4 stability and approximation of the discrete adjoint. In particular we discover the following properties:

- Proposition 4.3: Imposing SSP on both, discrete state and discrete adjoint, is too strong and it results in a first-order time-discretization.
- Theorem 4.4: Imposing SSP on the discrete state is enough to give stability of the discrete adjoint.
- Theorem 4.5 and Theorem 4.6: Any two-stage second-order TVD-RK method for the discrete state results in a second-order adjoint approximation. Any three-stage third-order TVD-RK method for the discrete state results in a second-order adjoint approximation. Hager in [Hag00] showed that for certain third-order RK methods, the resulting discrete adjoint is only second-order. Theorem 4.6 shows that this is the case for the class of TVD-RK methods.

The study of the differentiability of the control-to-state map when shocks are present was started in [BG97]. In that paper it was shown that this map is in general not differentiable. A similar problem formulated in terms of a minimization task was studied in [JS99] where the derivative of the objective functional was obtained when the state is smooth and does not contain shocks. Later S. Ulbrich in [Ulb01, Ulb02] analyzed the shift-differentiability of the so-called control-to-state operator and showed Fréchet differentiability of the reduced objective functional in the presence of shocks. Moreover, an adjoint procedure to compute the mentioned derivative was introduced and analyzed.

For the nonlinear conservation law (1), it has been observed that not all RK methods can ensure TVD properties of the approximation, i.e., oscillations occur near discontinuities; see the example in [GS98, Section 2]. Shu and Osher in [SO88] constructed a *class* of RK methods that ensures the approximation to be TVD; the so-called TVD-RK methods. The main idea is to use convex combinations of the forward Euler method to construct a high-order approximation. If the Euler step is stable in some (semi-)norm, then under some mild conditions the convex combination of the Euler steps is stable, too. Order conditions also derived in [SO88] for second and third orders with two and three stages, respectively. We should also remark that the derivation of such conditions remains *formal* as often the solution of the hyperbolic problem does not possess the required regularity. It is proven in [GS98] that a fourth-order method with four stages does not exist. However a fourth order five stage method was

discovered in [SR02]. Ruuth and Spiteri showed in [RS02] that there does not exist methods beyond fourth order of any number of stages.

For the numerical treatment of such optimization problems and for a particular objective functional, M. Giles showed in [Gil03] that the classical Lax-Friedrichs scheme leads to a discrete adjoint which converges to a wrong solution. Later M. Giles and S. Ulbrich showed in [GU10a, GU10b, Ulb01] under a restrictive time-step that the classical Lax-Friedrichs scheme yields a convergent discrete adjoint. Higher-order discretizations based on relaxation of the conservation law was also introduced and studied by M. Herty and M. Banda in [BH12]. We finish this brief review of the literature by recalling an alternating descent method introduced in [CPZ08] as a solution technique for such optimization problems.

2 Optimal control of scalar conservation laws

An abstract optimal control problem can be formulated as

$$\min J(y) \quad \text{subject to} \quad S(u) = y, \quad u \in U_{\text{ad}}, \tag{P}$$

where y is called the *state* variable and u is the *control* variable, with the latter belonging to an admissible set U_{ad}. The objective functional is denoted by $J(\cdot)$ and it is assumed to be differentiable. The control and state variables are related through a control-to-state map $S(\cdot)$ which can be regarded as a solution operator of the underlying PDE. Obviously, the control-to-state $S(\cdot)$ influences existence and uniqueness of the optimal control problem as well as optimality conditions which characterize the optimal solution.

In this work, we consider $S(\cdot)$ to be the solution operator of the following one-dimensional scalar conservation law (Cauchy problem):

$$\begin{aligned} y_t + f(y)_x &= 0 && \text{in } \mathbb{R} \times \mathbb{R}_+, \\ y(x,0) &= u(x) && \text{in } \mathbb{R}, \end{aligned} \tag{1}$$

where $u \in \mathrm{L}^\infty(\mathbb{R})$ is the initial data with compact support in a bounded interval $K \subset \mathbb{R}$ and $y(x,t)$ is the *conserved* variable. This motivates to define the admissible set U_{ad} by

$$U_{\text{ad}} := \big\{ u \in \mathrm{L}^\infty(\mathbb{R}) : \text{supp}(u) \subset K, \|u\|_{\mathrm{BV}(\mathbb{R})} \leq C \big\}, \tag{2}$$

where $C > 0$ is a constant, $\text{supp}(\cdot)$ denotes the support of a function, i.e., $\text{supp}(u) = \{x \in \mathbb{R} : u(x) \neq 0\}$, and $\mathrm{L}^\infty(\mathbb{R})$ is the Lebesgue space of essentially bounded functions on $\mathbb{R}$ with norm $\|\cdot\|_{\mathrm{L}^\infty(\mathbb{R})}$. In this paper we assume that the so-called *flux* function $f(\cdot)$ is in $\mathrm{C}^2(\mathbb{R})$ and satisfies $f'' > 0$. A particular example is $f(y) = \frac{1}{2}y^2$ which gives rise to the inviscid Burgers' equation. Concerning the objective functional we study the so-called tracking type functional, i.e.,

$$J(y) := \int_{\mathbb{R}} G(y(x,T))\,\mathrm{d}x, \tag{3}$$

where $T > 0$ is a final time. A common example is $G(y(x,T)) := |y(x,T) - y_{\text{obs}}(x)|^2$ with $y_{\text{obs}} \in \mathrm{L}^2(\mathbb{R})$ given.

Classical solutions of (1) can be constructed by the method of characteristics. However, due to the possible non-linearity of the flux function f, classical solutions break down in *finite time* even for very

smooth initial data. Therefore we consider generalized (weak) solutions of (1) in integral form: The function $y \in \mathrm{L}^\infty(\mathbb{R} \times \mathbb{R}_+)$ is called a weak solution of (1) if it satisfies

$$\int_{\mathbb{R}\times\mathbb{R}_+} y\,\varphi_t + f(y)\,\varphi_x \,\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} u(x)\varphi(x,0)\,\mathrm{d}x = 0 \quad \forall \varphi \in \mathrm{C}_c^\infty(\mathbb{R}\times\mathbb{R}_+), \tag{4}$$

where $\mathrm{C}_c^\infty(\mathbb{R}\times\mathbb{R}_+)$ represents the space of infinitely differentiable functions with compact supports in $\mathbb{R}\times\mathbb{R}_+$, with $\mathbb{R}_+ = [0,+\infty)$. There might be more than one weak solution for given initial data. For instance, consider the following example: Let $f(y) = \frac{1}{2}y^2$, and $u(x) = 0$ for $x < 0$ and $u(x) = 1$ for $x \geq 0$ which yields a discontinuous initial data. Then, the following two weak solutions are known:

$$y_1(x,t) = \begin{cases} 0 & x < \frac{1}{2}t, \\ 1 & x \geq \frac{1}{2}t, \end{cases} \quad \text{and} \quad y_2(x,t) = \begin{cases} 0 & x < 0, \\ x/t & 0 \leq x < t, \\ 1 & t \leq x. \end{cases}$$

The lack of uniqueness can be overcome by picking the *physically relevant* or *entropy solution*. Therefore we impose an extra constraint on the weak solution: A function $y \in \mathrm{L}^\infty(\mathbb{R}\times\mathbb{R}_+)$ is called an entropy solution (in the sense of Kružkov [Kru70]) if it satisfies

$$\int_{\mathbb{R}\times\mathbb{R}_+} \eta(y)\,\varphi_t + q(y)\,\varphi_x \,\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} \eta(u(x))\varphi(x,0)\,\mathrm{d}x \geq 0 \quad \forall \varphi \in \mathrm{C}_c^\infty(\mathbb{R}\times\mathbb{R}_+),\ \varphi \geq 0, \tag{5}$$

where $\eta(y) := |y-k|$ and $q(y) = \operatorname{sign}(y-k)(f(y)-f(k))$ for all $k \in \mathbb{R}$. One can show that if y satisfies (5) then it is also a weak solution in the sense of (4); see for instance [CG09, Section 5.5]. Kružkov then shows that such an entropy solution is indeed unique. For the following result we introduce $\mathrm{BV}(\mathbb{R})$, the space of the functions of bounded variations on $\mathbb{R}$, i.e., $u \in \mathrm{BV}(\mathbb{R})$ iff $u \in \mathrm{L}^1(\mathbb{R})$ and $\mathrm{TV}(u) := \sup\{\int_{\mathbb{R}} u\,p'\,\mathrm{d}x : p \in \mathrm{C}_c^1(\mathbb{R}), |p| \leq 1 \text{ a.e. in } \mathbb{R}\}$. Here "a.e." stands for "almost everywhere". Endowed with $\|u\|_{\mathrm{BV}(\mathbb{R})} = \|u\|_{\mathrm{L}^1(\mathbb{R})} + \mathrm{TV}(u)$, it is a Banach space.

The following theorem is due to Kružkov [Kru70]. We refer the reader to [Eva10, Section 11.4.3] for a proof.

Theorem 2.1. *Let $u \in \mathrm{L}^\infty(\mathbb{R})$ in (5), then there exists a unique entropy solution y of (5) that satisfies $\|y(\cdot,t)\|_{\mathrm{L}^\infty(\mathbb{R})} \leq \|u\|_{\mathrm{L}^\infty(\mathbb{R})}$ for all $t > 0$. Moreover let, y_1 and y_2 be two entropy solutions corresponding to initial data $u_1, u_2 \in \mathrm{L}^\infty(\mathbb{R}) \cap \mathrm{L}^1(\mathbb{R})$, respectively. Then we have*

$$\|y_1(\cdot,t) - y_2(\cdot,t)\|_{\mathrm{L}^1(\mathbb{R})} \leq \|u_1 - u_2\|_{\mathrm{L}^1(\mathbb{R})} \quad \forall t > 0.$$

Finally if $u \in \mathrm{BV}(\mathbb{R})$ then $\mathrm{TV}(y(\cdot,t)) \leq \mathrm{TV}(u)$ for all $t > 0$.

The notion of entropy solutions enables us to state our optimal control problem in a meaningful way. Therefore, our optimal control problem (P) is completed by setting the solution operator $S(\cdot)$ to be the entropy solution of the conservation law at time T. We will see in Section 2.1 that $S(\cdot)$ is *not* Gâteaux differentiable when discontinuity (shocks) appear in the state y. Therefore a different approach in defining a derivative of the control-to-state map, called *shift derivative*, is derived which leads to Fréchet differentiability of the objective functional.

Using this setting, existence of a minimizer can be shown for the underlying problem; see Appendix. We also refer to [CPZ08, Ulb99] for a more general case. Uniqueness is, however, not guaranteed

since we can construct control examples that lead to the same state which minimizes the objective functional: Indeed, consider the following controls

$$u_1 = \begin{cases} 1 & -2 \le x < 0, \\ -1 & 0 \le x \le 2, \\ 0 & \text{otherwise}, \end{cases} \qquad u_2 = \begin{cases} 1 & -2 \le x < -1, \\ -x & -1 \le x \le 1, \\ -1 & 1 \le x \le 2, \\ 0 & \text{otherwise}, \end{cases}$$

for Burgers' equation, i.e., $f(y) = \frac{1}{2}y^2$. We fix the final time $T = 2$. One can construct the corresponding entropy solution for each initial data. A direct calculation by the method of characteristics shows that at time $t \ge T$ both initial data give

$$y(x,t) = \begin{cases} \frac{1}{t}(x+2) & -2 \le x \le 0, \\ \frac{1}{t}(-x+2) & 0 < x \le 2, \\ 0 & \text{otherwise}. \end{cases}$$

Setting $y_{\text{obs}} := y(x,T)$, we then have $J(S(u_1)) = J(S(u_2)) = 0$, i.e., two optimal controls for (P).

Remark 2.2 (flux identification problem)**.** A different, yet similar optimal control problem of conservation laws can be formulated in which the control variable is the flux function and the initial data is fixed. More precisely we look for a flux function $f \in \mathcal{F}_{\text{ad}}$, where $\mathcal{F}_{\text{ad}}$ is the admissible set, minimizing a given objective functional. For instance f may have a closed analytical form, perhaps transforming the optimal control problem into $\mathbb{R}^n$. In this case the solution operator is, i.e., defined as $y = S(f)$. The existence of the minimum can be proven by a continuity result due to Lucier [Luc86]

$$\|S(f)(\cdot,t) - S(g)(\cdot,t)\|_{\mathrm{L}^1(\mathbb{R})} \le t\|f-g\|_{\text{Lip}}\,\|u\|_{\mathrm{BV}},$$

and an assumption on the compactness of $\mathcal{F}_{\text{ad}}$; see [JS99, Section 2.2] for details. Uniqueness is again not guaranteed in general; see the discussion in [JS99].

2.1 Shift differentiability and adjoint calculus

As we have seen, the control-to-state operator is a delicate object and needs special care for the *forward problem* (1) to be well-posed. We have also seen that although the optimal control problem (P) admits a minimizer, uniqueness cannot be expected in general. it is not necessarily unique. We now turn our attention on characterizing such a minimum. For the numerical treatment, first-order characterizations of such minimizers are of importance. This is our next subject of study.

A first step in deriving optimal conditions for the problem is investigating differentiability of the objective functional. Note that the objective functional defined by (3) is well behaved, in the sense that G is sufficiently smooth and allows existence of a solution and only the solution operator $S(\cdot)$ may cause problems. Thus, using the chain rule, we must investigate differentiability of $S(\cdot)$. It turns out that in the presence of shocks, $S(\cdot)$ is not Gâteaux differentiable as we illustrate in the following example borrowed from [BG97].

Example 2.1. Suppose the initial data is $u_\epsilon = (1+\epsilon)x \cdot \mathbf{1}_{[0,1]}(x)$ where $\epsilon \in \mathbb{R}$ and $\mathbf{1}_{[0,1]}(x)$ denotes the indicator function of the interval $[0,1]$. Consider the Burgers' equation. We can construct the entropy solution by the method of characteristics:

$$y_\epsilon(x,t) = \frac{(1+\epsilon)x}{1+(1+\epsilon)t} \cdot \mathbf{1}_{[0,\sqrt{1+(1+\epsilon)t}]}(x).$$

Note that the shock position now depends on the perturbation, that is, $x_s(t) = \sqrt{1+(1+\epsilon)t}$, and at time $t = 0$ the derivative exists

$$v(x) = \lim_{\epsilon \to 0} \epsilon^{-1}(u_\epsilon - u_0) = x \cdot \mathbf{1}_{[0,1]}(x).$$

However at time $t > 0$ there is no such derivative $v(x) \in \mathrm{L}^1(\mathbb{R})$. In fact $\epsilon^{-1}(y_\epsilon(\cdot,t) - y_0(\cdot,t))$ converges as $\epsilon \to 0$ in the sense of distributions to

$$\frac{x}{(1+t)^2} \cdot \mathbf{1}_{[0,\sqrt{1+t}]} + \frac{t}{2(t+1)} \delta_{\sqrt{1+t}}, \tag{6}$$

where δ_x is the Dirac delta function located at x.

Note that in the previous example, the distributional derivative has a continuous part and a singular part due to a shift in the shock location. Moreover the perturbed shock location, $x_s(t) = \sqrt{1+(1+\epsilon)t}$, is differentiable with respect to the perturbation and the solution, $y_\epsilon(x,t)$, vary differentiably in the left and right of the shock. This motivates a first order approximation of (6) introduced in [Ulb01, Ulb02].

Let $u \in U \subset \mathcal{U}$ where U is an open set and $\mathcal{U}$ a Banach space. Moreover u is such that $y_u := S(u)$ has bounded variation and its support is in $K \subset \mathbb{R}$. For the sake of presentation, suppose for instance that u has compact support, contains a shock at $x_s(0)$ and is a C^1 function on either side of the shock. Suppose the shock remains in the solution up to time T and its position is given by $x_s(t)$. We perturb u by $\delta u \in U$ such that $y_{u+\delta u} := S(u+\delta u) \in \mathrm{L}^\infty(\mathbb{R}) \cap \mathrm{BV}(\mathbb{R})$. Then we can define a first order approximation of $y_{u+\delta u} - y_u$ by the *shift variation*

$$\delta S(\delta y, s) := \delta y + \operatorname{sign}(s) [\![y(x_s(T),T)]\!] \mathbf{1}_{[x_s(T), x_s(T)+s]} \in \mathrm{L}^1(K),$$

where $[\![y(x_s(T),T)]\!] = y(x_s^-(T),T) - y(x_s^+(T),T)$ and $(\delta y, s) \in \mathrm{L}^1(K) \times \mathbb{R}$ depends linearly on δu. Here δy corresponds to the variation of the solution in the continuous part and s is a linear approximation of the shock shift (e.g., in Example 2.1, $x_s(t)$ is differentiable with respect to ϵ). More precisely we suppose that there exists a bounded linear operator

$$T(u) \in \mathcal{L}(\mathcal{U}, \mathrm{L}^r(K) \times \mathbb{R}) \quad r \in (1, \infty],$$

such that $(\delta y, s) := T(u) \cdot \delta u$. We say y_u is *shift differentiable* at u if

$$\lim_{\delta u \to 0} \|\delta u\|_{\mathcal{U}}^{-1} \cdot \|y_{u+\delta u} - y_u - \delta S(T(u) \cdot \delta u)\|_{\mathrm{L}^1(K)} = 0.$$

It is proven that shift differentiability implies Fréchet differentiability of the objective functional, see for instance [Ulb01, Section 3.2.2]. Moreover it is shown in [Ulb01, Theorem 3.3.2] that, under some technical assumptions, entropy solutions are continuously shift-differentiable and therefore the objective functional is Fréchet differentiable.

Although the shift differentiability result is a useful tool in proving differentiability of the objective functional using sensitivities, it is often more convenient to obtain the objective functional's derivative using an adjoint calculus. For the conservation laws with smooth solution, F. James and M. Sepúlveda derived such an adjoint based derivative; see [JS99, Section 2.4] and the reference therein. Later S. Ulbrich generalized the adjoint calculus to the case where the solution contains *shocks*. Here we only state the result and refer the interested reader to [Ulb01, Ulb02] for details. For a formal derivation see [GU10a].

Suppose the perturbation δu has the same structure as u, i.e., it contains a shock discontinuity at $x_s(0)$ and is piecewise C^1 on either side of the shock. Then the derivative of $\mathcal{J}(u) := J(S(u))$ in the direction of the δu (perturbation in the initial data) is given by

$$\mathcal{J}'(u)\delta u = \int_{\mathbb{R}} p(x,0)\,\delta u(x)\,\mathrm{d}x, \tag{7}$$

where $p(x,t)$ is the adjoint variable that satisfies the following backward equation with final-time condition

$$\begin{array}{rcll} p_t + f'(y)p_x & = & 0 & \text{in } \mathbb{R}\times(0,T), \\ p(x,T) & = & G'(y(x,T)) & \text{in } \mathbb{R}. \end{array} \tag{8}$$

In case $y(x,t)$ contains a shock which travels along $x_s(t)$, we need to impose an *interior* boundary condition for the adjoint along $x_s(t)$:

$$p(x_s(t),t) = \frac{[\![G(y)]\!]}{[\![y]\!]}\Big|_{(x_s(T),T)} \quad \forall t\in[0,T]. \tag{9}$$

The adjoint equation (8) is backward in time and is a non-conservative hyperbolic PDE with final datum and has been studied by many authors, see for instance [JS99] and [Ulb03]. Let us first consider the case when $p(x,T)\in \mathrm{Lip}_{loc}(\mathbb{R})$, e.g., shocks at the final time are smoothed in the objective functional. A one-sided Lipschitz-continuity condition (OSLC) on $f'(y)$, i.e.,

$$\operatorname*{ess\,sup}_{x\neq z}\Big(\frac{f'(y(x,t))-f'(y(z,t))}{x-z}\Big)^+ \leq m(t),$$

where $m(t)\in \mathrm{L}^1(0,T)$, guarantees that the generalized backward characteristics starting from time T do not intersect. Then existence of at least one Lipschitz continuous solution can be guaranteed. However, uniqueness is not ensured; see [Con67]. Uniqueness can be proved for the so-called *reversible solutions* which we briefly recall. Let us define the set $\mathcal{E}$ as the set of exceptional solutions, i.e., Lipschitz continuous solutions of (8) where $p(x,T)=0$. Then the support of the exceptional solutions is defined as

$$\mathcal{V}_e := \big\{(x,t)\in\mathbb{R}\times(0,T) : \exists p\in\mathcal{E}, p(x,t)\neq 0\big\}.$$

A *reversible solution* is then a Lipschitz continuous solution of (8) which is locally constant in $\mathcal{V}_e$; see [BJ98, Definition 4.1.4].

We should mention that reversible solutions can be also defined for discontinuous Borel functions as end data $p(x,T)$ that are pointwise everywhere the limit of a bounded sequence $(p_n^T)\in \mathrm{C}^{0,1}(\mathbb{R})$, i.e., bounded in $\mathrm{C}(\mathbb{R})\cap W^{1,1}_{loc}(\mathbb{R})$; see [Ulb03]. In this case, reversible solutions can be defined as broad solutions along the generalized backward characteristics, which automatically ensures the internal boundary condition (9), if the end date (9) are used at $(x_s(T),T)$.

The following theorem states properties of the reversible solution and is proved in [Ulb03, Theorem 14] for the general case of hyperbolic balance laws. For simplicity of the presentation, we adapt the result in [Ulb03, Theorem 14] to our setting (see also [BJ98]). For the case of a discontinuous end data, we refer the reader to [Ulb03, Corollary 15].

Theorem 2.3. *Let $f'(y(x,t))\in \mathrm{L}^\infty(\mathbb{R}\times(0,T))$ and satisfies OSLC. Then the following holds: For all $p(x,T)\in \mathrm{C}^{0,1}(\mathbb{R})$ there exists a unique reversible solution p of (8). Moreover, $p\in \mathrm{C}^{0,1}(\mathbb{R}\times[0,T])$*

and solves (8) a.e. on $\mathbb{R} \times (0,T)$*. Finally, for all* $t \in [0,T]$*,* $z_1 < z_2$ *and* $0 \le s < \hat{s} \le T$ *with*

$$\begin{aligned} I &:= [z_1, z_2], \quad I_s^{\hat{s}} := [s, \hat{s}] \times I, \\ J &:= \big[z_1 - (T-t)\|f'(y)\|_{L^\infty(\mathbb{R}\times(0,T))}, z_2 + (T-t)\|f'(y)\|_{L^\infty(\mathbb{R}\times(0,T))}\big], \\ J_t^T &:= [t,T] \times J, \end{aligned}$$

the following estimates hold:

$$\begin{aligned} \|p(\cdot,t)\|_{B(I)} &\le \|p(\cdot,T)\|_{B(J)}, \\ \|p_x(\cdot,t)\|_{L^1(I)} &\le \|p_x(\cdot,T)\|_{L^1(J)}, \\ \|p_t\|_{L^1(I_s^{\hat{s}})} &\le (\hat{s}-s)\|f'(y)\|_{L^\infty(I_s^{\hat{s}})}\|p_x\|_{L^\infty(s,\hat{s};L^1(I))}, \end{aligned}$$

where $B(I)$ *is the Banach space of the bounded functions equipped with the* $\sup$*-norm.*

We now demonstrate the use of the adjoint calculus by the following example.

Example 2.2. Consider Burgers' equation, i.e., $f(y) := \frac{1}{2}y^2$, on the domain $\Omega = (-1,1)$, and the initial data is set to be

$$u(x) := \begin{cases} 1 & -1 \le x < 0, \\ -1 & 0 < x \le 1, \end{cases}$$

with a shock discontinuity at $x = 0$. For the boundary conditions we set $y(-1,t) = 1$ and $y(1,t) = -1$ for $t > 0$. It is easy to see that the entropic solution equals the initial data, i.e., $y(x,t) = u(x)$ for all $t > 0$. Hence the shock position remains at $x = 0$, i.e., $x_s(t) = 0$ for $t > 0$. We now compute the adjoint state using (8) and (9). For this purpose, let $G(y) := \frac{1}{2}|y|^2$. Then we have $[\![G(y)]\!]|_{t=T} = 0$ and $G'(y) = y$. Then for the final-time condition of the adjoint equation, we obtain

$$p(x,T) := \begin{cases} 1 & -1 \le x < 0, \\ -1 & 0 < x \le 1, \end{cases} \quad \text{and} \quad p(0,T) = 0.$$

and for the "interior" boundary condition we have $p(0,t) = 0$ for $t \in [0,T]$. We fix $T = \frac{1}{2}$.

By the method of characteristics we can construct $p(x,t)$ for $t < T$. More precisely, along all straight lines $x(t)$ with the derivative $\dot{x}(t) = y(x(t),t)$, the adjoint state is constant. This implies that, at $t = 0$, we have

$$p(x,0) = \begin{cases} 1 & -1 < x < -\frac{1}{2}, \\ 0 & -\frac{1}{2} \le x \le \frac{1}{2}, \\ -1 & -\frac{1}{2} < x < 1. \end{cases}$$

See also Figure 1.

3 Numerical methods

In this section we describe how we discretize the optimal control problem (P). In this paper we follow the *discretize-then-differentiate* approach, that is, we first fully discretize the optimization problem and then derive the optimality conditions for the resulting *finite dimensional* optimization problem. In particular we should consider in detail the discretization of the conservation law (1). It is important that one makes sure that the discretization of the forward problem converges to the unique entropy solution

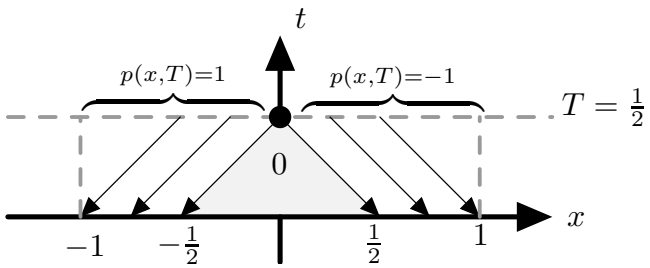

Figure 1: Construction of the adjoint using the method of characteristics in space-time domain. Note that the adjoint has a constant value in the gray area due to the discontinuity of the state at time T.

as well as the inherited adjoint discretization to the continuous adjoint state. We refer the reader to [LeV90] for an introduction to numerical methods for such PDEs.

Since we are interested in studying how time discretization, using TVD-RK methods, influences quality of the overall scheme, we first discretize the conservation law in space and then in time by a TVD-RK method.

3.1 Spatial discretization

Let us first partition the domain, say $\mathbb{R}$ with non-overlapping intervals, $I_j := (x_{j-1/2}, x_{j+1/2}]$, where $x_{j-1/2} < x_{j+1/2}$ for all $j \in \mathbb{Z}$. The so-called mesh size is denoted by $h_j := x_{j+1/2} - x_{j-1/2}$. We denote the semi-discrete approximation at time t by

$$\boldsymbol{y}(t) := (\ldots, y_{j-1}(t), y_j(t), y_{j+1}(t), \ldots) \in \ell^\infty(\mathbb{Z}).$$

More precisely, $y_j(t) \in \mathbb{R}$ is an approximation to the cell-average of the true solution, i.e.,

$$y_j(t) \approx \frac{1}{h_j} \int_{x_{j-1/2}}^{x_{j+1/2}} y(x,t)\,\mathrm{d}x.$$

We then discretize in space using a *conservative scheme*, by choosing a *numerical flux function* $\hat{f}(y_j(t), y_{j+1}(t))$ that approximates $f(y(x_{j+1/2}, t))$. Then the semi-discrete numerical scheme is obtained by solving the following ordinary differential equation (ODE) in time:

$$\frac{\mathrm{d}}{\mathrm{d}t} y_j(t) + \frac{1}{h_j}\Big(\hat{f}(y_j(t), y_{j+1}(t)) - \hat{f}(y_{j-1}(t), y_j(t))\Big) = 0 \quad \forall j \in \mathbb{Z}. \tag{10}$$

The simplest time discretization for (10) is given by the forward Euler scheme. For this purpose, we partition the time direction into time slabs t_n for $n \in \mathbb{Z}_+$, where $t_n < t_{n+1}$. For simplicity we assume a uniform time-step, i.e., $t_{n+1} - t_n = k$ for all $n \in \mathbb{Z}_+$ and similarly a uniform mesh-size, i.e., $h_j = h$ for all $j \in \mathbb{Z}$. Then the fully discrete system using the forward Euler discretization reads

$$y_j^{n+1} = y_j^n - \frac{k}{h}\Big(\hat{f}(y_j^n, y_{j+1}^n) - \hat{f}(y_{j-1}^n, y_j^n)\Big) \quad \forall j \in \mathbb{Z}, n \in \mathbb{Z}_+, \tag{11}$$

where $\boldsymbol{y}^n := (\ldots, y_{j-1}^n, y_j^n, y_{j+1}^n, \ldots)$ is an approximation to $\boldsymbol{y}(t_n)$. Later, in Section 4, we discretize (10) by a high-order TVD-RK method.

The choice of the numerical flux $\hat{f}(\cdot,\cdot)$ is crucial since it effects convergence properties of the method. Note that not only are we interested in convergence to a weak solution but, also we require convergence to the entropy solution. Hence we require that the method satisfy a discrete version of (5) as

well as other properties like L^1-contraction, total variation diminishing etc. In fact, *monotone schemes* satisfy such requirements and converge to the entropy solution; see for instance [LeV90, Chapter 15]. We, however, need yet another condition: the numerical scheme should be differentiable. This enables us to derive optimality conditions at the discrete level.

Motivated by the differentiability issue addressed above, the following numerical fluxes are used in this paper:

$$\begin{aligned} \hat{f}_{\mathsf{LF}}(a,b) &:= \tfrac{1}{2}(f(a)+f(b)) - \tfrac{\gamma}{2}\tfrac{h}{k}(b-a) && \text{Lax-Friedrichs (LF)},\\ \hat{f}_{\mathsf{EO}}(a,b) &:= f(0) + \textstyle\int_0^a f'(s)^+ \,\mathrm{d}s + \int_0^b f'(s)^- \,\mathrm{d}s && \text{Engquist-Osher (EO)}, \end{aligned} \tag{12}$$

where $\gamma \in (0,1)$ and we define $f'(s)^+ := \max(0, f'(s))$ and $f'(s)^- := \min(0, f'(s))$. We note that the classical Lax-Friedrichs method uses $\gamma = 1$. However, due to the stability requirements for the discrete adjoint we impose $\gamma < 1$; see Section 3.3. The Lax-Friedrichs scheme is monotone provided that the time-step satisfies the CFL condition

$$\frac{k}{h} \sup_{|y|\le M} |f'(y)| \le \gamma, \tag{13}$$

where $M = \max_x |u(x)|$. We define the total variation semi-norm for $\boldsymbol{y}$ by

$$|\boldsymbol{y}|_{\mathsf{TV}} := \sum_{j=-\infty}^{\infty} |y_{j+1} - y_j|.$$

It is well-known that monotone schemes are also TVD; see [LeV90, Chapter 15.7]. Therefore we have $|\boldsymbol{y}^{n+1}|_{\mathsf{TV}} \le |\boldsymbol{y}^n|_{\mathsf{TV}}$. M. Giles and S. Ulbrich in [GU10a, GU10b] proved that for the Lax-Friedrichs method, provided $k \le \gamma \cdot h^{2-q}$ for $0 < q < 1$, both forward and adjoint approximations converge to the their respective continuous counterparts. This is due to adding more grid points near the shock position as the mesh is refined. This, however, results in a very restrictive time-step requirement.

Since the support of the initial data is assumed to be in a bounded set, recall (2), and it is known that the solution of the conservation law has a finite speed of propagation, we can consider the problem on a bounded domain, denoted by Ω only. Therefore, the discretization can be written using finite dimensional operators. Let us suppose that $\Omega = (a,b)$ is partitioned into N elements, I_j for $j = 1,\dots,N$, with $x_{1/2} = a$ and $x_{N+1/2} = b$. Then we can define a finite dimensional space $V_h := \mathbb{R}^N$ as approximation space, and we denote the approximation at time t_n by $\boldsymbol{y}_h^n = (y_1^n, y_2^n, \dots, y_N^n) \in V_h$. This allows us to express the underlying scheme using a *non-linear* discrete operator, $F_h : V_h \to V_h$, which is defined by

$$\big[F_h(\boldsymbol{w})\big]_j := \hat{f}(w_j, w_{j+1}) - \hat{f}(w_{j-1}, w_j) \quad \forall j = 1,\dots,N, \quad \boldsymbol{w} \in V_h. \tag{14}$$

The fully discrete version of (11) with forward Euler time integration can then be written as

$$\boldsymbol{y}_h^{n+1} = \boldsymbol{y}_h^n - \frac{k}{h} F_h(\boldsymbol{y}_h^n), \quad \boldsymbol{y}_h^0 = \boldsymbol{u}_h. \tag{15}$$

Differentiability properties of $F_h(\cdot)$ will be exploited later in Section 3.2 for deriving an adjoint discretization. Now we state Lax-Friedrichs and Engquist-Osher differentiability in the following proposition whose proof we defer to the appendix.

Proposition 3.1. *Let the flux function be $f(\cdot) \in \mathrm{C}^2$. Then the respective $F_h(\boldsymbol{w})$ for the Lax-Friedrichs and Engquist-Osher schemes at $\boldsymbol{w} \in V_h$ is Fréchet differentiable, i.e., there exists a linear bounded operator $F_h'(\boldsymbol{w}) : V_h \to V_h$ such that in direction $\boldsymbol{v} \in V_h$ we have $[F_h'(\boldsymbol{w})\boldsymbol{v}]_j = g_{j,j+1} - g_{j-1,j}$ with*

$$g^{\mathsf{LF}}_{j,j+1} := \tfrac{1}{2}\big[f'(w_{j+1})v_{j+1} + f'(w_j)v_j\big] - \tfrac{\gamma}{2}\tfrac{h}{k}(v_{j+1} - v_j) \qquad \text{(LF)},$$

$$g^{\mathsf{EO}}_{j,j+1} := \tfrac{1}{2}\big[f'(w_{j+1})v_{j+1} + f'(w_j)v_j\big] - \tfrac{1}{2}(|f'(w_{j+1})|v_{j+1} - |f'(w_j)|v_j) \quad \text{(EO)},$$

for $i = j, \dots, N$. Moreover, for their transposes we have

$$\big[F_h'(\boldsymbol{w})^\top \boldsymbol{v}\big]_j = \gamma\tfrac{h}{k}v_j - \tfrac{1}{2}\big(\gamma\tfrac{h}{k} + f'(w_j)\big)v_{j+1} - \tfrac{1}{2}\big(\gamma\tfrac{h}{k} - f'(w_j)\big)v_{j-1} \qquad \text{(LF)},$$

$$\big[F_h'(\boldsymbol{w})^\top \boldsymbol{v}\big]_j = |f'(w_j)|v_j - \tfrac{1}{2}\big(|f'(w_j)| + f'(w_j)\big)v_{j+1} - \tfrac{1}{2}\big(|f'(w_j)| - f'(w_j)\big)v_{j-1} \quad \text{(EO)}.$$

We will see in Section 3.2 how properties of $F_h(\cdot)$ influence the discrete adjoint variable. In particular we are interested in total variation diminishing properties of the discrete adjoint.

3.2 Discrete optimal control problem and adjoint calculus

In this section we state the discrete optimization problem, derive the discrete adjoint and study its properties when a forward Euler time integration is employed together with the spatial discretization discussed in Section 3.1.

As before, we denote the final time by T. For simplicity, we partition the time direction in a way that there exists n_T such that $T = n_T \cdot k$. In order to ease the notation, we concatenate approximations $\boldsymbol{y}_h^n$ at different times $n = 1, \dots, n_T$ to obtain $\boldsymbol{y}_h \in (V_h)^{n_T+1}$:

$$\boldsymbol{y}_h := (\boldsymbol{y}_h^0, \boldsymbol{y}_h^1, \boldsymbol{y}_h^2, \dots, \boldsymbol{y}_h^{n_T})^\top. \tag{16}$$

The discretized objective functional is then given by

$$J_h(\boldsymbol{y}_h) := \sum_{j=1}^{N} h\, G(y_j^{n_T}),$$

where $\boldsymbol{y}_h$ is obtained by (15). We then consider the following finite dimensional optimal control problem

$$\min J_h(\boldsymbol{y}_h) \quad \text{subject to} \quad S_h(\boldsymbol{u}_h) = \boldsymbol{y}_h^{n_T}, \quad \boldsymbol{u}_h \in V_h \cap U_{\mathsf{ad}}, \tag{DP}$$

where $S_h(\cdot)$ is the discrete control-to-state operator which is defined by successive (n_T times) application of (15) to $\boldsymbol{u}_h$.

For deriving the discrete adjoint, it is more convenient to consider (15) as constraint instead of $S_h(\cdot)$. For this purpose, we define the equality constraint at time t_n by $L_h^n : (V_h)^{n_T+1} \to V_h$ with

$$L_h^n(\boldsymbol{y}_h) := -\boldsymbol{y}_h^n + \boldsymbol{y}_h^{n-1} - \frac{k}{h}F_h(\boldsymbol{y}_h^{n-1}) \quad \text{for } n = 1, \dots, n_T,$$

and $L_h^0(\boldsymbol{y}_h, \boldsymbol{u}_h) = -\boldsymbol{y}_h^0 + \boldsymbol{u}_h$. We then collect all time-step contributions and state the discrete constraint as $L_h(\boldsymbol{y}_h, \boldsymbol{u}_h) = 0$ where $L_h : (V_h)^{n_T+1} \times V_h \to (V_h)^{n_T+1}$ with

$$L_h(\boldsymbol{y}_h, \boldsymbol{u}_h) = \Big(L_h^0(\boldsymbol{y}_h, \boldsymbol{u}_h), L_h^1(\boldsymbol{y}_h), L_h^2(\boldsymbol{y}_h), \cdots, L_h^{n_T}(\boldsymbol{y}_h)\Big)^\top.$$

Then we define the Lagrangian for the finite dimensional problem (DP) by

$$\mathcal{L}(\boldsymbol{y}_h, \boldsymbol{p}_h, \boldsymbol{u}_h) := J_h(\boldsymbol{y}_h) + h\, \boldsymbol{p}_h^\top \cdot L_h(\boldsymbol{y}_h, \boldsymbol{u}_h),$$

where $\boldsymbol{p}_h \in (V_h)^{n_T+1}$ is enumerated like $\boldsymbol{y}_h$ in (16).

Recall that from Proposition 3.1 we know that $F_h(\cdot)$ is differentiable when Lax-Friedrichs and Engquist-Osher schemes are used. Differentiability of $F_h(\cdot)$ implies differentiability of $L_h(\cdot,\cdot)$ with respect to the first argument. More precisely, a direct calculation shows that this derivative, denoted by $L'_{\boldsymbol{y}_h}(\cdot,\cdot)$, has a lower triangular structure:

$$L'_{\boldsymbol{y}_h}(\boldsymbol{y}_h, \boldsymbol{u}_h) = \begin{bmatrix} -I & & & & \\ I - \frac{k}{h}F'_h(\boldsymbol{y}_h^1) & -I & & & \\ & \ddots & \ddots & & \\ & & I - \frac{k}{h}F'_h(\boldsymbol{y}_h^{n_T-1}) & -I & \\ & & & I - \frac{k}{h}F'_h(\boldsymbol{y}_h^{n_T}) & -I \end{bmatrix},$$

and therefore $L'_{\boldsymbol{y}_h}(\cdot,\cdot)$ is invertible. As a consequence, the constraint $L_h^n(\boldsymbol{y}_h) = 0$ satisfies the linear independence constraint qualification (LICQ) at any feasible point of (DP). Hence, at a solution pair (y_h, u_h) there exists a Lagrange multiplier (adjoint state) p_h.

Now, we can compute the discrete derivative of the objective functional using a discrete adjoint calculus. The concatenated discrete adjoint equation is given

$$\nabla J_h(\boldsymbol{y}_h) + h\, L'_{\boldsymbol{y}_h}(\boldsymbol{y}_h, \boldsymbol{u}_h)^\top \boldsymbol{p}_h = 0. \tag{17}$$

Note that $L'_{\boldsymbol{y}_h}(\boldsymbol{y}_h, \boldsymbol{u}_h)^\top$ is upper triangular, and for a fixed $\boldsymbol{y}_h$ one can solve (17) successively from $\boldsymbol{p}_h^{n_T}$ to $\boldsymbol{p}_h^0$, i.e., backward in time similar to the continuous adjoint; see (8). From the definition of the discrete objective function we get

$$\nabla J_h(\boldsymbol{y}_h) = \big(\boldsymbol{0}, \boldsymbol{0}, \cdots, \boldsymbol{0}, h\, G'(\boldsymbol{y}_h^{n_T})\big)^\top,$$

and therefore $\boldsymbol{p}_h^{n_T} = G'(\boldsymbol{y}_h^{n_T})$ which is the discrete counterpart for the final-time condition in (8). For $n = n_T, \ldots, 0$ we have

$$\boldsymbol{p}_h^{n_T} = G'(\boldsymbol{y}_h^{n_T}), \quad \boldsymbol{p}_h^n = \Big[I - \frac{k}{h}F'_h(\boldsymbol{y}_h^{n+1})^\top\Big]\boldsymbol{p}_h^{n+1} \quad \text{for } n = (n_T - 1), \ldots, 0. \tag{18}$$

which is the discrete counterpart of the adjoint PDE (8). The gradient of the discrete reduced objective function in the direction of $\delta\boldsymbol{u}_h \in V_h$ is obtained by

$$\nabla \mathcal{J}_h(\boldsymbol{u}_h)^\top \delta\boldsymbol{u}_h = h \sum_{j=1}^{N} p_j^0\, \delta u_j. \tag{19}$$

Remark 3.2 (Convergence of the discrete adjoint)**.** Note, however, that at the discrete level, the interior boundary condition (9) does not appear. Hence a natural question is to ask whether the discrete adjoint converges to the continuous one or not. It is shown in [Gil03] through a numerical experiments that the adjoint obtained from Lax-Friedrichs scheme converges to a wrong solution as the mesh is refined. It is also observed that if numerical diffusion is chosen such that the shock in $\boldsymbol{y}_h^{n_T}$ spreads into more cells, then the discrete adjoint converges. This is proven in [GU10a, GU10b].

We are not aware of a similar result for the Engquist-Osher scheme. However, it has been shown in [Ulb01] that for the case where the end data is Lipschitz continuous, e.g., by using a smoothed end state in the tracking functional, the discrete adjoint converges to the continuous adjoint. This motivates the use of two separate solvers for the discrete forward and adjoint problems, e.g., by using a TVDRK-MUSCL scheme, when the end data is smoothed. We should mention that this approach results in an inexact discrete adjoint.

3.3 Stability of the discrete adjoint

We would like to examine monotonicity of the discrete adjoint by checking whether or not it is TVD. The discrete adjoint, satisfying (18), can be written in a simplified form for analysis as

$$p_j^n = A_{j,0}\, p_j^{n+1} + A_{j,1}\, p_{j+1}^{n+1} + A_{j,-1}\, p_{j-1}^{n+1}, \tag{20}$$

where

$$A_{j,l}^{\mathsf{LF}} := \begin{cases} \frac{\gamma}{2} - \frac{k}{2h} f'(w_j) & \text{for } l = -1, \\ 1 - \gamma & \text{for } l = 0, \\ \frac{\gamma}{2} + \frac{k}{2h} f'(w_j) & \text{for } l = 1, \end{cases} \qquad A_{j,l}^{\mathsf{EO}} := \begin{cases} \frac{k}{2h}\big(|f'(w_j)| - f'(w_j)\big) & \text{for } l = -1, \\ 1 - \frac{k}{h}|f'(w_j)| & \text{for } l = 0, \\ \frac{k}{2h}\big(|f'(w_j)| + f'(w_j)\big) & \text{for } l = 1. \end{cases}$$

Note that, provided the CFL condition (13) is satisfied, we have $A_{j,l}^{\mathsf{LF}} \geq 0$ and $A_{j,l}^{\mathsf{EO}} \geq 0$. Moreover, observe that by construction we have

$$\sum_{l=-1}^{1} A_{j,l} = 1.$$

Taking absolute values on both sides in (20) and using the above properties, we obtain

$$|p_j^n| \leq A_{j,0}|p_j^{n+1}| + A_{j,1}|p_{j+1}^{n+1}| + A_{j,-1}|p_{j-1}^{n+1}| \leq \max_{l=-1,0,1} |p_{j+l}^{n+1}| \leq \|\boldsymbol{p}_h^{n+1}\|_\infty,$$

which implies L^∞ stability. We now show that the discrete adjoint scheme with a forward Euler time discretization is also TVD. We first rewrite (20) in the following form

$$p_j^n = p_j^{n+1} + B_{j,0}(p_j^{n+1} - p_{j-1}^{n+1}) + B_{j,1}(p_{j+1}^{n+1} - p_j^{n+1}), \tag{21}$$

with

$$B_{j,l}^{\mathsf{LF}} := \begin{cases} -\frac{\gamma}{2} + \frac{k}{2h} f'(w_j) & \text{for } l = 0, \\ \frac{\gamma}{2} + \frac{k}{2h} f'(w_j) & \text{for } l = 1, \end{cases} \qquad B_{j,l}^{\mathsf{EO}} := \begin{cases} -\frac{k}{2h}|f'(w_j)| + \frac{k}{2h} f'(w_j) & \text{for } l = 0, \\ \frac{k}{2h}|f'(w_j)| + \frac{k}{2h} f'(w_j) & \text{for } l = 1. \end{cases}$$

Then, Harten's Lemma [Har83a] guarantees TVD properties of the discrete adjoint scheme.

Lemma 3.3 (Harten's Lemma)**.** *Suppose a finite difference scheme can be written as*

$$w_j = v_j + B_{j,0} \cdot (v_j - v_{j-1}) + B_{j,1} \cdot (v_{j+1} - v_j),$$

where $B_{j,0}$ and $B_{j,1}$ are arbitrary nonlinear functions of v_j, v_{j+1}, v_{j-1} and satisfy

$$B_{j,0} \leq 0, \quad B_{j,1} \geq 0, \quad B_{j,1} - B_{j+1,0} \leq 1.$$

Then we have $|\boldsymbol{w}_h|_{\mathsf{TV}} \leq |\boldsymbol{v}_h|_{\mathsf{TV}}$.

Observe that in our case, the above conditions on $B_{j,0}$ and $B_{j,1}$ are satisfied provided a $(1-\gamma)$-CFL condition for Lax-Friedrichs and a $\frac{1}{2}$-CFL condition for Engquist-Osher are satisfied, respectively, i.e.,

$$\frac{k^{\mathsf{LF}}}{h} \sup_{|y|\leq M} |f'(y)| \leq (1-\gamma), \quad \frac{k^{\mathsf{EO}}}{h} \sup_{|y|\leq M} |f'(y)| \leq \frac{1}{2}. \tag{22}$$

This is the reason for the choice of $\gamma \in (0,1)$ in the Lax-Friedrichs scheme. Obviously the optimal CFL condition is

$$\mathsf{CFL}^* = \max_{\gamma\in(0,1)} \min\{\gamma, 1-\gamma\} = \frac{1}{2}$$

Lemma 3.3 now guarantees that the discrete adjoint obtained from forward Euler time discretization is TVD, i.e.,

$$\big|\boldsymbol{p}_h^n\big|_{\mathsf{TV}} \leq \big|\boldsymbol{p}_h^{n+1}\big|_{\mathsf{TV}}. \tag{23}$$

Note that this property is shared by the continuous adjoint in Theorem 2.3.

4 Strong stability preserving time discretizations

In this section, we examine the effect of using RK methods for discretizing the underlying problem instead of using the forward Euler method.

A TVD-RK method is defined by convex combinations of forward Euler steps which are parametrized by two sets of coefficients: $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$ for $i = 1, ..., s$ and $j = 0, ..., (s-1)$ where s is the number of RK stages. A TVD-RK time stepping is then defined as follows:

1 Set the initial stage: $\boldsymbol{w}^{(0)} = \boldsymbol{y}_h^n$,

2 Compute for each stage $i = 1, \dots, s$:

$$\boldsymbol{w}^{(i)} = \sum_{j=0}^{i-1} \alpha_{ij} \boldsymbol{w}^{(j)} - \frac{k}{h} \beta_{ij} F_h(\boldsymbol{w}^{(j)}), \tag{24}$$

3 Set the next time-step approximation: $\boldsymbol{y}_h^{n+1} = \boldsymbol{w}^{(s)}$.

Moreover, we impose the following constraints over $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$:

$$\alpha_{ij}, \beta_{ij} \geq 0, \quad \sum_{j=0}^{i-1} \alpha_{ij} = 1, \quad \big(\beta_{ij} \neq 0 \implies \alpha_{ij} \neq 0\big). \tag{25}$$

The following result is due to Shu and Osher [SO88] and shows that the TVD-RK method is stable if the forward Euler (15) is stable. We call such methods *strong stability preserving* (SSP) since we have stability with respect to stage variables, too.

Proposition 4.1. *Suppose the time-step k is chosen such that the forward Euler discretization is stable, i.e.,*

$$\|\boldsymbol{w} - \frac{k}{h} \frac{\beta_{ij}}{\alpha_{ij}} F_h(\boldsymbol{w})\| \leq \|\boldsymbol{w}\| \quad \forall \boldsymbol{w} \in V_h, \tag{26}$$

for all $i = 1, \dots, s$, $j = 0, \dots, (s-1)$ and $\alpha_{ij} \neq 0$. Here, $\|\cdot\|$ is a non-negative homogeneous convex function, e.g., a norm or a semi-norm. Moreover assume that the conditions in (25) are satisfied. Then for the TVD-RK method we have

$$\|\boldsymbol{w}^{(i)}\| \leq \max_{j=0,\dots,(i-1)} \|\boldsymbol{w}^{(j)}\| \quad \text{for } i = 1, \dots, s,$$

and consequently $\|\boldsymbol{y}^{n+1}\| \leq \|\boldsymbol{y}^n\|$.

In order to highlight the technical differences between the proof technique of [SO88], which relies on $\sum_{j=0}^{i-1} \alpha_{ij} = 1$, the property not available for the discrete adjoint in Section 4.1, we display the short proof.

Proof. First observe that if $\alpha_{ij} = 0$ then the contribution of $w^{(j)}$ is zero. So we can rewrite (24) by

$$\boldsymbol{w}^{(i)} = \sum_{\{j:\alpha_{ij}\neq 0\}} \alpha_{ij} \Big(\boldsymbol{w}^{(j)} - \frac{k}{h} \frac{\beta_{ij}}{\alpha_{ij}} F_h(\boldsymbol{w}^{(j)}) \Big) \quad \forall i = 1, \dots, s.$$

Table 1: Table of coefficients for TVD-RK methods of order two and three.

order	α_{ij}			β_{ij}			$\min \alpha_{ij}/\beta_{ij}$
2	1			1			1
	1/2	1/2		0	1/2		
3	1			1			1
	3/4	1/4		0	1/4		
	1/3	0	2/3	0	0	2/3	

We then take $\|\cdot\|$ from both sides and use convexity as well as our assumption on the stability of the Euler step: for all $i = 1, \dots, s$ we have

$$\|\boldsymbol{w}^{(i)}\| \le \sum_{\{j:\alpha_{ij}\neq 0\}} \alpha_{ij}\|\boldsymbol{w}^{(j)}\| \le \Big(\max_{j=0,\dots,(i-1)} \|\boldsymbol{w}^{(j)}\|\Big)\Big(\sum_{\{j:\alpha_{ij}\neq 0\}} \alpha_{ij}\Big) = \max_{j=0,\dots,(i-1)} \|\boldsymbol{w}^{(j)}\|,$$

where we also used positivity of α_{ij}, β_{ij} and $\sum_{j=0}^{i-1} \alpha_{ij} = 1$. The proof is completed by induction. □

Let us denote the forward Euler time-step by k_{FE}. Then the stable TVD-RK time-step is bounded by

$$k \le \Big(\min_{\alpha_{ij},\beta_{ij}\neq 0} \frac{\alpha_{ij}}{\beta_{ij}}\Big) k_{\mathsf{FE}}. \tag{27}$$

Therefore one can optimize the coefficients α_{ij} and β_{ij} to maximize the constant $\min_{\alpha_{ij},\beta_{ij}\neq 0} \frac{\alpha_{ij}}{\beta_{ij}}$. In Table 1, we show such optimal TVD-RK methods of two and three stages.

We now rewrite our discrete optimization problem (DP) using a TVD-RK time discretization. First we redefine $\boldsymbol{y}_h^n$ to be suitable for the TVD-RK method: the collection of all stage approximations at time-slab t_n is given by

$$\boldsymbol{y}_h^n := (\boldsymbol{y}_h^{n,0}, \boldsymbol{y}_h^{n,1}, \dots, \boldsymbol{y}_h^{n,s})^\top \in W_h := (V_h)^{s+1},$$

where $\boldsymbol{y}_h^{n,l}$ is the approximation at time t_n, at stage $l = 0, \dots, s$. Then we concatenate contributions from all time-steps to get

$$\boldsymbol{y}_h := (\boldsymbol{y}_h^0, \boldsymbol{y}_h^1, \dots, \boldsymbol{y}_h^{n_T})^\top \in (W_h)^{n_T+1}.$$

Let us denote the forward Euler operator by $H_{ij} : V_h \to V_h$ with

$$H_{ij}(\boldsymbol{v}_h) := \begin{cases} \boldsymbol{v}_h - \frac{k}{h}\frac{\beta_{ij}}{\alpha_{ij}} F_h(\boldsymbol{v}_h) & \text{if } \alpha_{ij} > 0, \\ 0 & \text{if } \alpha_{ij} = 0. \end{cases}$$

for all $\boldsymbol{v}_h \in V_h$ which is differentiable with derivative in direction of $\boldsymbol{u}_h \in V_h$ given by

$$H'_{ij}(\boldsymbol{v}_h)\boldsymbol{u}_h := \begin{cases} \big[I - \frac{k}{h}\frac{\beta_{ij}}{\alpha_{ij}} F'_h(\boldsymbol{v}_h)\big]\boldsymbol{u}_h & \text{if } \alpha_{ij} > 0, \\ 0 & \text{if } \alpha_{ij} = 0. \end{cases}$$

Then the equality constraint generated by TVD-RK scheme at time t_n and stage l is denoted by $L_h^{n,l} : (W_h)^{n_T+1} \to W_h$ with

$$L_h^{0,0}(\boldsymbol{y}_h, \boldsymbol{u}_h) := -\boldsymbol{y}_h^{0,0} + \boldsymbol{u}_h,$$

$$L_h^{n,0}(\boldsymbol{y}_h) := -\boldsymbol{y}_h^{n,0} + \boldsymbol{y}_h^{n-1,s},$$

$$L_h^{n,l}(\boldsymbol{y}_h) := -\boldsymbol{y}_h^{n,l} + \textstyle\sum_{j=0}^{l-1} \alpha_{lj} H_{lj}(\boldsymbol{y}_h^{n,j}) \quad \text{for } n = 1, \dots, n_T, \quad \text{and } l = 1, \dots, s.$$

The fully discrete scheme, i.e., spatial discretization with the TVD-RK method, can be expressed as $L_h(\boldsymbol{y}_h, \boldsymbol{u}_h) = 0$ where

$$L_h(\boldsymbol{y}_h, \boldsymbol{u}_h) = \Big(L_h^{0,0}(\boldsymbol{y}_h, \boldsymbol{u}_h), L_h^{0,1}(\boldsymbol{y}_h), \cdots, L_h^{0,s}(\boldsymbol{y}_h), \ldots, L_h^{n_T,s}(\boldsymbol{y}_h)\Big)^\top \in (W_h)^{n_T+1}.$$

As before we define the Lagrangian function by

$$\mathcal{L}(\boldsymbol{y}_h, \boldsymbol{p}_h, \boldsymbol{u}_h) := J_h(\boldsymbol{y}_h) + h\, \boldsymbol{p}_h^\top \cdot L_h(\boldsymbol{y}_h, \boldsymbol{u}_h),$$

where $\boldsymbol{p}_h \in W_h^{n_T+1}$ has the same structure as $\boldsymbol{y}_h$. Note that since $H_{ij}(\cdot)$ is differentiable we can conclude that $L_h(\cdot,\cdot)$ is also differentiable with respect to the first argument. Moreover, similar to the case with forward Euler step, it has again a lower triangular structure:

$$L'_{\boldsymbol{y}_h}(\boldsymbol{y}_h, \boldsymbol{u}_h) = \begin{bmatrix} G(\boldsymbol{y}_h^0) & & & & \\ K & G(\boldsymbol{y}_h^1) & & & \\ & \ddots & \ddots & & \\ & & K & G(\boldsymbol{y}_h^{n_T-1}) & \\ & & & K & G(\boldsymbol{y}_h^{n_T}) \end{bmatrix},$$

where $G(\cdot)$ is the linearized contribution of the TVD-RK scheme, i.e.,

$$G(\boldsymbol{y}_h^n) := \begin{bmatrix} -I & & & & \\ \alpha_{10} H'_{10}(\boldsymbol{y}_h^{n,0}) & -I & & & \\ \alpha_{20} H'_{20}(\boldsymbol{y}_h^{n,0}) & \alpha_{21} H'_{21}(\boldsymbol{y}_h^{n,1}) & -I & & \\ & \vdots & & \ddots & \\ \alpha_{s,0} H'_{s,0}(\boldsymbol{y}_h^{n,0}) & & & \alpha_{s,s-1} H'_{s,s-1}(\boldsymbol{y}_h^{n,s-1}) & -I \end{bmatrix}, \tag{28}$$

and K links the approximation from the previous time-step to the current one:

$$K := \begin{bmatrix} O & \cdots & O & I \\ O & \cdots & O & O \\ \vdots & & \vdots & \vdots \\ O & \cdots & O & O \end{bmatrix}. \tag{29}$$

As before, the discrete adjoint satisfies,

$$\nabla J_h(\boldsymbol{y}_h) + h L_{\boldsymbol{y}_h}(\boldsymbol{y}_h, \boldsymbol{u}_h)^\top \boldsymbol{p}_h = 0,$$

which is again backward in time. As before, the above constraint satisfies LICQ. Let us illustrate the above adjoint equation with an example.

Example 4.1 (case $s = 2$)**.** Consider the discrete adjoint for a TVD-RK discretization with two stages. Let $n < n_T$. Then the discrete adjoint satisfies the following backward in time linear system:

$$\begin{array}{cccccc} \boldsymbol{p}_h^{n,0} & -\alpha_{10} H'_{10}(\boldsymbol{y}_h^{n,0})^\top \boldsymbol{p}_h^{n,1} & -\alpha_{20} H'_{20}(\boldsymbol{y}_h^{n,0})^\top \boldsymbol{p}_h^{n,2} & = & \mathbf{0}, \\ & \boldsymbol{p}_h^{n,1} & -\alpha_{21} H'_{21}(\boldsymbol{y}_h^{n,1})^\top \boldsymbol{p}_h^{n,2} & = & \mathbf{0}, \\ & & \boldsymbol{p}_h^{n,2} & = & \boldsymbol{p}_h^{n+1,0}, \end{array}$$

which is a TVD-RK scheme with coefficients $\alpha^\star_{ij}$ and $\beta^\star_{ij}$. In fact, the TVD-RK coefficients for the discrete adjoint equation are related to the TVD-RK of the discrete state as follows

$$\begin{array}{llll} \alpha^*_{10} = \alpha_{21}, & & \beta^*_{10} = \beta_{21}, & \\ \alpha^*_{20} = \alpha_{20}, & \alpha^*_{21} = \alpha_{10}, & \beta^*_{20} = \beta_{20}, & \beta^*_{21} = \beta_{10}. \end{array}$$

Observe the way how the coefficients of the TVD-RK scheme are transformed by this conjugation process. We refer to the coefficients of the adjoint TVD-RK scheme as conjugate coefficients. This transformation in the table of adjoint TVD-RK coefficients might pose some restrictions on the choice of the TVD-RK method in the first place. We conclude this section by the following proposition.

Proposition 4.2. *Suppose we discretize-then-optimize the problem (P) and a TVD-RK time discretization is used for the discrete state variable with coefficients α_{ij} and β_{ij}. Then the discrete adjoint is also obtained by a TVD-RK method with coefficients α^*_{ij} and β^*_{ij} such that*

$$\alpha^*_{ij} = \alpha_{s-j,s-i}, \quad \beta^*_{ij} = \beta_{s-j,s-i} \text{ for } i = 1, \dots, s, \text{ and } j = 0, \dots, (s-1). \tag{30}$$

4.1 Stability of the discrete adjoint

Given the result of Proposition 4.2, our first idea is to impose SSP on both the discrete state and the adjoint variables, i.e., imposing (25) on $\{\alpha_{ij}, \beta_{ij}\}$ and $\{\alpha^*_{ij}, \beta^*_{ij}\}$. In other words, the following conditions on $\{\alpha_{ij}, \beta_{ij}\}$ should hold:

$$\alpha_{ij}, \beta_{ij} \geq 0, \quad \sum_{j=0}^{i-1} \alpha_{ij} = 1, \quad \sum_{i=j+1}^{s} \alpha_{ij} = 1, \quad \big(\beta_{ij} \neq 0 \implies \alpha_{ij} \neq 0\big). \tag{31}$$

This however turns out to be too strong as the following proposition clarifies.

Proposition 4.3. *Suppose we discretize-then-optimize the problem (P) with a TVD-RK method. If we require SSP for both discrete state and discrete adjoint, then the TVD-RK method is at most* first-order*. More precisely, the TVD-RK coefficients are*

$$\alpha_{i,j} = \begin{cases} 1 & \text{if } j = i-1, \\ 0 & \text{otherwise,} \end{cases} \qquad \beta_{i,j} = \begin{cases} \beta_{i,i-1} & \text{if } j = i-1, \\ 0 & \text{otherwise,} \end{cases}$$

which gives a concatenation of forward Euler steps.

Proof. We need to identify sets of $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$ that satisfy (31). Thus, all coefficients need to be non-negative. Let $i = 1$, then $\alpha_{10} = 1$. Now let $j = 0$ and observe that $\sum_{i=j+1}^{s} \alpha_{ij} = 1$. This implies $\alpha_{i,0} = 0$ for all $i = 2, \dots, s$ since all $\alpha_{ij} \geq 0$ and $\alpha_{10} = 1$. Now let $i = 2$ and observe that $\sum_{j=0}^{i-1} \alpha_{ij} = 1$. However we just showed that $\alpha_{20} = 0$ which implies $\alpha_{21} = 1$. We now repeat the same process by letting $j = 1$ and consider the constraint $\sum_{i=j+1}^{s} \alpha_{ij} = 1$. Continuing, we conclude that $\alpha_{i,i-1} = 1$ and the other coefficients are zero. Since $\alpha_{i,j} = 0$ for $j < i-1$ we conclude from the last requirement of (31) that $\beta_{ij} = 0$ for $j < i-1$. Therefore the only free parameters are $\beta_{i,i-1}$. However, such a TVD-RK scheme is equivalent to the concatenation of Euler steps, instead of a combination. Finally, the concatenation of Euler steps yields a first-order method. □

As we shall see next, imposing (31) is not necessary. Let us consider a two-stage TVD-RK that ensures SSP only for the state discretization. That is, TVD-RK satisfies only (25). Moreover suppose that the time-step k is chosen such that adjoint stability of the forward Euler step for discrete adjoint holds. For instance, for the discrete adjoint from Example 4.1 we have

$$\begin{aligned} \|\boldsymbol{p}^{n,1}\| &\leq \alpha_{21}\|\boldsymbol{p}^{n,2}\|, \\ \|\boldsymbol{p}^{n,0}\| &\leq \alpha_{20}\|\boldsymbol{p}^{n,2}\| + \alpha_{10}\|\boldsymbol{p}^{n,1}\|, \end{aligned}$$

where $\|\cdot\|$ is a non-negative homogeneous convex function, e.g., a semi-norm or a norm.

For SSP of the TVD-RK scheme we would take the maximum of (semi-)norms up to the $(i-1)^{\text{st}}$ stage and use the assumption that the sum of the coefficients in each stage equals *one*. However here we do not have that $\alpha_{20} + \alpha_{10} = 1$. Instead, we can substitute the first inequality into the second and obtain

$$\|\boldsymbol{p}^{n,0}\| \leq (\alpha_{20} + \alpha_{21}\,\alpha_{10})\|\boldsymbol{p}^{n,2}\| = \|\boldsymbol{p}^{n,2}\|,$$

which holds true since $\alpha_{10} = 1$ and $\alpha_{20} + \alpha_{21} = 1$. This shows that for two-stage SSP TVD-RK methods we have stability at each *time-step* which is weaker than stability at each *stage*; observe that in Proposition 4.1 stability is achieved at each stage and therefore at each time-step. We can generalize this observation to an arbitrary s-stage TVD-RK method.

Theorem 4.4. *Suppose the state equation is discretized with an SSP TVD-RK method. Moreover suppose that k is chosen such that it ensures forward Euler stability for both the discrete state and adjoint. Then the discrete adjoint is stable in each time-step for an arbitrary s-stage method., i.e.,*

$$\|\boldsymbol{p}^{n,0}\| \leq \|\boldsymbol{p}^{n,s}\|,$$

where $\|\cdot\|$ is a non-negative homogeneous convex function, e.g., a semi-norm or a norm.

Proof. Since we require Euler step stability, we have for each stage

$$\|\boldsymbol{p}^{n,\ell}\| \leq \sum_{i=\ell+1}^{s} \alpha_{i\ell}\,\|\boldsymbol{p}^{n,i}\| \quad \text{for } \ell = 0, \ldots, s-1.$$

Let $\ell = 0$ and recall that $\alpha_{10} = 1$. Then, using the above inequality we have

$$\|\boldsymbol{p}^{n,0}\| \leq \sum_{i=1}^{s} \alpha_{i0}\,\|\boldsymbol{p}^{n,i}\| = \|\boldsymbol{p}^{n,1}\| + \sum_{i=2}^{s} \alpha_{i0}\|\boldsymbol{p}^{n,i}\| \leq \sum_{i=2}^{s}(\alpha_{i1} + \alpha_{i0})\|\boldsymbol{p}^{n,i}\|.$$

Now isolating the term with $i = 2$ and noting that $\alpha_{21} + \alpha_{20} = 1$ we have

$$\|\boldsymbol{p}^{n,0}\| \leq \|\boldsymbol{p}^{n,2}\| + \sum_{i=3}^{s}(\alpha_{i1} + \alpha_{i0})\|\boldsymbol{p}^{n,i}\| \leq \sum_{i=3}^{s}(\alpha_{i2} + \alpha_{i1} + \alpha_{i0})\|\boldsymbol{p}^{n,i}\|.$$

We repeat this procedure to obtain $\|\boldsymbol{p}^{n,0}\| \leq \sum_{i=\ell'}^{s} \big(\sum_{j=0}^{\ell'-1} \alpha_{ij}\big)\|\boldsymbol{p}^{n,i}\|$ for all $\ell' = 1, \ldots, s$. We then choose $\ell' = s$ and obtain $\|\boldsymbol{p}^{n,0}\| \leq (\sum_{j=0}^{s-1} \alpha_{sj})\|\boldsymbol{p}^{n,s}\| = \|\boldsymbol{p}^{n,s}\|$ which completes the proof. □

Theorem 4.4 shows that any s-stage TVD-RK method for the discrete state yields a stable TVD-RK method for the discrete adjoint. However the discrete adjoint is proved to be stable at each time-step instead of each stage.

4.2 Order conditions for the discrete adjoint

In this section we study approximation properties of the scheme for the discrete adjoint. We focus on deriving order conditions for the discrete adjoint scheme. For this purpose, extra conditions on the TVD-RK method applied to the state discretization are needed to ensure high-accuracy of the adjoint scheme. We also check which methods provide optimal CFL constants. For simplicity of the notation we consider TVD-RK methods with a conjugate coefficient table as in Proposition 4.2 for the following linear problem

$$\dot{\boldsymbol{p}}(t) + R(t)\boldsymbol{p}(t) = 0, \quad \boldsymbol{p}(T) = \boldsymbol{p}_T \in V_h. \tag{32}$$

Here $R(t)$ is a linear operator and defined as

$$R(t) := -\frac{1}{h} F_h'(\boldsymbol{y}(t))^\top, \tag{33}$$

where $\boldsymbol{y}(t)$ is the solution of the semi-discrete problem (10), i.e., $\dot{\boldsymbol{y}}(t) = -\frac{1}{h} F_h(\boldsymbol{y}(t))$. For completeness we state derivatives of $R(\cdot)$:

$$\begin{aligned} \dot{R}(t) &= \tfrac{1}{h^2} F_h''(\boldsymbol{y}(t))^\top F_h(\boldsymbol{y}(t)), \\ \ddot{R}(t) &= -\tfrac{1}{h^3}\big[F_h''(\boldsymbol{y}(t))^\top F_h'(\boldsymbol{y}(t))^\top F_h(\boldsymbol{y}(t)) + F_h'''(\boldsymbol{y}(t))^\top F_h(\boldsymbol{y}(t))^2\big]. \end{aligned} \tag{34}$$

Observe that $R(t)$ at ℓth stage is approximated by $\frac{-1}{h} F_h'(\boldsymbol{y}_h^{n,\ell})$, see for instance Example 4.1 and the definition of $H_{i\ell}'(\boldsymbol{y}_h^{n,\ell})$. Since we consider the local error of the TVD-RK method in the time interval $[t_n, t_{n+1}]$, we let $\boldsymbol{y}_h^{n,0} = \boldsymbol{y}(t_n)$. For the analysis of the TVD-RK method we need approximation properties of $\frac{-1}{h} F_h'(\boldsymbol{y}_h^{n,\ell})$. A direct calculation gives

$$-\tfrac{1}{h} F_h'(\boldsymbol{y}_h^{n,0})^\top = R(t_{n+1}) - k\dot{R}(t_{n+1}) + \tfrac{1}{2}k^2\ddot{R}(t_{n+1}) + O(k^3), \tag{35}$$

$$\begin{aligned} -\tfrac{1}{h} F_h'(\boldsymbol{y}_h^{n,1})^\top = {} & R(t_{n+1}) - (1-\beta_{10})k\dot{R}(t_{n+1}) \\ & -\tfrac{1}{2}\tfrac{k^2}{h^3}(1-\beta_{10})^2 F_h'''(\boldsymbol{y}(t_{n+1}))^\top F_h^2(\boldsymbol{y}(t_{n+1})) \\ & -\tfrac{1}{2}\tfrac{k^2}{h^3}(1-2\beta_{10}) F_h''(\boldsymbol{y}(t_{n+1}))^\top F_h'(\boldsymbol{y}(t_{n+1}))^\top F_h(\boldsymbol{y}(t_{n+1})) + O(k^3), \end{aligned} \tag{36}$$

and

$$\begin{aligned} -\tfrac{1}{h} F_h'(\boldsymbol{y}_h^{n,2})^\top = {} & R(t_{n+1}) - (1-\psi)k\dot{R}(t_{n+1}) \\ & -\tfrac{1}{2}\tfrac{k^2}{h^3}(1-\psi)^2 F_h'''(\boldsymbol{y}(t_{n+1}))^\top F_h^2(\boldsymbol{y}(t_{n+1})) \\ & -\tfrac{1}{2}\tfrac{k^2}{h^3}(1-2\psi+2\beta_{21}\beta_{10}) F_h''(\boldsymbol{y}(t_{n+1}))^\top F_h'(\boldsymbol{y}(t_{n+1}))^\top F_h(\boldsymbol{y}(t_{n+1})) \\ & +O(k^3), \end{aligned} \tag{37}$$

where

$$\psi := \beta_{20} + \beta_{21} + \alpha_{21}\beta_{10}. \tag{38}$$

Note that since the inner stages of the RK are low order, we have first order approximation of $R(\cdot)$ in (36) and (37).

A Taylor expansion of the solution of (32) at time t_{n+1} with the time-step $(-k)$ yields

$$\begin{aligned} \boldsymbol{p}(t_n) = {} & \boldsymbol{p}(t_{n+1}) + kR(t_{n+1})\boldsymbol{p}(t_{n+1}) + \tfrac{k^2}{2}\big[R^2(t_{n+1}) - \dot{R}(t_{n+1})\big]\boldsymbol{p}(t_{n+1}) \\ & -\tfrac{k^3}{6}\big[2\dot{R}(t_{n+1})R(t_{n+1}) + R(t_{n+1})\dot{R}(t_{n+1}) - R^3(t_{n+1}) - \ddot{R}(t_{n+1})\big]\boldsymbol{p}(t_{n+1}) \\ & +O(k^4), \end{aligned} \tag{39}$$

We will compare the TVD-RK method with conjugate coefficients against the Taylor expansion in (39).

4.2.1 Two stage methods

We now study the approximation properties of a two stage scheme. In this vein, Shu and Osher derived in [SO88] order conditions for TVD-RK methods with two stages. As before, let $\{\alpha_{ij}\}, \{\beta_{ij}\}$ be the coefficients of a TVD-RK scheme for the state equation. Then the method is second-order if it satisfies (25) and the following order conditions

$$\begin{aligned} \alpha_{21}\beta_{10} + \alpha_{10}\beta_{21} + \beta_{20} &= 1, \\ \beta_{10}\beta_{21} &= \tfrac{1}{2}. \end{aligned} \tag{40}$$

We need to find extra conditions on $\{\alpha_{ij}\}$ and $\{\beta_{ij}\}$ to obtain a second-order approximation for the discrete adjoint too. For the TVD-RK method with a conjugate coefficient table applied to (32) we have (see Example 4.1):

$$\begin{aligned}\boldsymbol{p}_h^{n,0} &= \big[\alpha_{10}+k\beta_{10}R(t_n)\big]\boldsymbol{p}_h^{n,1}+\big[\alpha_{20}+k\beta_{20}R(t_n)\big]\boldsymbol{p}_h^{n,2},\\ \boldsymbol{p}_h^{n,1} &= \big[\alpha_{21}+k\beta_{21}R(t_{n+1})-k^2\beta_{21}(1-\beta_{10})\dot{R}(t_{n+1})\big]\boldsymbol{p}_h^{n,2}+O(k^3).\end{aligned}$$

Eliminating $\boldsymbol{p}_h^{n,1}$, we simplify the above equations and obtain

$$\begin{aligned}\boldsymbol{p}_h^{n,0} &= (\alpha_{20}+\alpha_{10}\alpha_{21})\boldsymbol{p}_h^{n,2}+(\alpha_{10}\beta_{21}+\alpha_{21}\beta_{10}+\beta_{20})kR(t_{n+1})\boldsymbol{p}_h^{n,2}\\ &\quad+\Big[\beta_{10}\beta_{21}R^2(t_{n+1})-(\beta_{20}+\alpha_{21}\beta_{10}+\alpha_{10}\beta_{21}(1-\beta_{10}))\dot{R}(t_{n+1})\Big]k^2\boldsymbol{p}_h^{n,2}+O(k^3).\end{aligned}\tag{41}$$

Recall that $\alpha_{10}=1$ and $\alpha_{20}+\alpha_{21}=1$. Then using these facts in the above expansion we get

$$\begin{aligned}\boldsymbol{p}_h^{n,0} &= \boldsymbol{p}_h^{n,2}+(\alpha_{10}\beta_{21}+\alpha_{21}\beta_{10}+\beta_{20})kR(t_{n+1})\boldsymbol{p}_h^{n,2}\\ &\quad+\Big[\beta_{10}\beta_{21}R^2(t_{n+1})-(\beta_{20}+\alpha_{21}\beta_{10}+\alpha_{10}\beta_{21}(1-\beta_{10}))\dot{R}(t_{n+1})\Big]k^2\boldsymbol{p}_h^{n,2}+O(k^3).\end{aligned}\tag{42}$$

We compare (42) to the Taylor expansion of the exact solution (39) and require the following conditions on the coefficients:

$$\begin{aligned}\alpha_{21}\beta_{10}+\alpha_{10}\beta_{21}+\beta_{20} &= 1,\\ \beta_{10}\beta_{21} &= \tfrac{1}{2},\\ \beta_{20}+\alpha_{21}\beta_{10}+\alpha_{10}\beta_{21}(1-\beta_{10}) &= \tfrac{1}{2}.\end{aligned}\tag{43}$$

Note that the first and second conditions are satisfied due to (40). Moreover, the first two conditions imply that the third condition is automatically satisfied. Therefore any second order TVD-RK method applied to the forward problem results in a second order approximation of the discrete adjoint. The above arguments prove the following theorem.

Theorem 4.5. *Suppose a second-order two-stage TVD-RK method is used to discretize the state equation. Then the corresponding TVD-RK method for the adjoint equation is consistent and is* second-order*. Moreover, The optimal CFL constant is* one*.*

4.2.2 Three stage methods

In this section we analyze approximation properties of the third-order three stages TVD-RK methods. The following conditions should be satisfied to ensure a third-order discrete state:

$$\alpha_{32}=1-\alpha_{31}-\alpha_{30},\tag{44}$$

$$\beta_{32}=\frac{3\beta_{10}-2}{6\psi(\beta_{10}-\psi)},\tag{45}$$

$$\beta_{21}=\frac{1}{6\beta_{10}\beta_{32}},\tag{46}$$

$$\beta_{31}=\frac{\frac{1}{2}-\alpha_{32}\beta_{10}\beta_{21}-\psi\beta_{32}}{\beta_{10}},\tag{47}$$

$$\beta_{30}=1-\alpha_{31}\beta_{10}-\alpha_{32}\psi-\beta_{31}-\beta_{32},\tag{48}$$

$$\beta_{20}=\psi-\alpha_{21}\beta_{10}-\beta_{21}.\tag{49}$$

The free parameters are $\alpha_{21}, \alpha_{30}, \alpha_{31}, \beta_{10}$ and ψ; see the discussion in [SO88] for details. The same analysis as in Section 4.2.1, and employing inner stages approximation properties (35)-(37) and comparing with coefficients of the Taylor expansion in (39), yields that the following conditions must be satisfied for the discrete adjoint:

$$\alpha_{31}\beta_{10} + \alpha_{21}\alpha_{32}\beta_{10} + \alpha_{32}(\beta_{20}+\beta_{21}) + \beta_{30} + \beta_{31} + \beta_{32} = 1 \quad \text{(first order)}, \tag{50}$$

$$\left.\begin{aligned} \alpha_{32}\beta_{10}\beta_{21} + \beta_{10}\beta_{31} + \alpha_{21}\alpha_{32}\beta_{10} + \beta_{32}(\beta_{20}+\beta_{21}) &= \tfrac{1}{2} \quad (\text{for } R^2) \\ \alpha_{31}\beta_{10} + \alpha_{21}\alpha_{32}\beta_{10} + \alpha_{32}\beta_{20} + \beta_{30} + \beta_{32}(1-\psi) & \\ +(\alpha_{32}\beta_{21} + \beta_{31})(1-\beta_{10}) &= \tfrac{1}{2} \quad (\text{for } \dot{R}) \end{aligned}\right\} \text{(second order)}, \tag{51}$$

$$\left.\begin{aligned} \beta_{10}\beta_{21}\beta_{32} &= \tfrac{1}{6} \quad (\text{for } R^3) \\ \beta_{20}\beta_{32} + \beta_{10}\beta_{31} + (1-\beta_{10})\beta_{21}\beta_{32} + \beta_{10}\beta_{21}\alpha_{32} + \beta_{10}\beta_{32}\alpha_{21} &= \tfrac{1}{3} \quad (\text{for } \dot{R}R) \\ (1-\psi)\big(\beta_{32}\beta_{20} + \beta_{10}\alpha_{21}\beta_{32} + \beta_{32}\beta_{21}\big) & \\ +(1-\beta_{10})\big(\beta_{21}\beta_{10}\alpha_{32} + \beta_{10}\beta_{31}\big) &= \tfrac{1}{6} \quad (\text{for } R\dot{R}) \\ \alpha_{31}\beta_{10} + \alpha_{21}\alpha_{32}\beta_{10} + \alpha_{32}\beta_{20} + (\alpha_{32}\beta_{21}+\beta_{31})(1-\beta_{10})^2 & \\ +\beta_{30} + \beta_{32}(1-\psi)^2 &= \tfrac{1}{3} \quad (\text{for } F'''F^2) \\ \alpha_{31}\beta_{10} + \alpha_{21}\alpha_{32}\beta_{10} + \alpha_{32}\beta_{20} + (\alpha_{32}\beta_{21}+\beta_{31})(1-2\beta_{10}) & \\ +\beta_{30} + \beta_{32}(1-2\psi+2\beta_{21}\beta_{10}) &= \tfrac{1}{3} \quad (\text{for } F''F'F) \end{aligned}\right\} \text{(third order)}. \tag{52}$$

One can show that if conditions (44)-(49) are satisfied then (50)-(52) are automatically satisfied except for the term associated to $R\dot{R}$: Let us define $A := (1-\psi)\big(\beta_{32}\beta_{20} + \beta_{10}\alpha_{21}\beta_{32} + \beta_{32}\beta_{21}\big) + (1-\beta_{10})\big(\beta_{21}\beta_{10}\alpha_{32} + \beta_{10}\beta_{31}\big)$. Then we have

$$A = (1-\psi)\psi\beta_{32} + (1-\beta_{10})\Big(\frac{1}{2} - \psi\beta_{32}\Big) = \frac{1}{2} + \psi(\psi-\beta_{10})\beta_{32} - \frac{1}{2}\beta_{10} = \frac{1}{2} + \frac{1}{2}\beta_{10} - \frac{1}{6} - \frac{1}{2}\beta_{10} = \frac{1}{3},$$

which cannot be equal to $\frac{1}{6}$ for any choice of parameters. Therefore the discrete adjoint TVD-RK method is at most second order. The above arguments prove the following theorem.

Theorem 4.6. *Suppose a third-order three-stage TVD-RK method is used to discretize the state equation. Then the corresponding TVD-RK method for the adjoint equation is consistent and is at most* second-order*. Moreover, the optimal CFL constant is* one*.*

4.2.3 Fifth stage (fourth-order) method

We have already mentioned that a fourth-order TVD-RK method with four stages does not exist, see [GS98, RS02]. Using a non-linear programming computer code, a fourth-order with five stages TVD-RK method has been found in [SR02, Appendix B]. It has been shown by Hager in [Hag00, Proposition 6.1] that when a general four stage fourth-order RK method is applied to the forward problem of an optimal control problem of ODEs, then the corresponding discrete adjoint is also automatically fourth-order. Inspired by this result we would like to check whether or not the fourth-order with five stage TVD-RK method of [RS02] generates a fourth-order discrete adjoint too.

In order to check the order conditions for this particular TVD-RK method, we use the Butcher's table of the mentioned TVD-RK method and check if the order conditions in [Hag00, Table 1] are satisfied. A direct calculation shows that the discrete adjoint is only second order. This is however not surprising since the TVD-RK method is obtained by a non-linear programming (with order conditions for the forward problem as constraints).

4.3 Convergence of the TVD-RK method

In this section we discuss convergence of the discrete adjoint obtained from a TVD-RK method. We follow the framework established in [Ulb01, Chapter 6.4]. We consider the case when the end data of the continuous adjoint is smooth, e.g., $p(x,T) \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R})$, and the space discretization is either Lax-Friedrichs or Engquist-Osher method which yields a monotone scheme when combined with forward Euler method in time; see for instance [LeV90, Chapter 15.7].

Consider the TVD-RK schemes with either Lax-Friedrichs or Engquist-Osher discretization and suppose the time-step k is chosen such that the forward Euler discretizations (26) are monotone and their corresponding discrete adjoint scheme is TVD, see (22), (27). The first ingredient for the proof is to show that the TVD-RK scheme is again a monotone scheme for the forward problem, i.e., given any initial data y_j^n, w_j^n, at time t_n we have

$$y_j^n \geq w_j^n \quad \forall j \implies y_j^{n+1} \geq w_j^{n+1} \quad \forall j.$$

Suppose that until ℓ-th stage we have $y_j^{n,i} \geq w_j^{n,i}$ for all $i = 1, \ldots, \ell$. Since $y_j^{n,\ell+1}$ is equal to a convex combination of forward Euler methods and each forward Euler method is a monotone scheme, we can conclude that $y_j^{n,\ell+1} \geq w_j^{n,\ell+1}$.

The second ingredient is to show that after eliminating the inner stages we have a conservative scheme. Let us demonstrate this for a two-stage method: observe that the application of an Euler step at the first stage can be written as

$$y_j^{n,1} = \mathcal{H}(y_{j-1}^{n,0}, y_j^{n,0}, y_{j+1}^{n,0}) := \alpha_{10} y_j^{n,0} - \beta_{10} \frac{k}{h} \big[\hat{f}(y_j^{n,0}, y_{j+1}^{n,0}) - \hat{f}(y_{j-1}^{n,0}, y_j^{n,0}) \big]. \tag{53}$$

Then the second stage can be written in the conservative form

$$y_j^{n+1} = y_j^n - \frac{k}{h} \Big[\tilde{f}(y_{j-1}^n, y_j^n, y_{j+1}^n, y_{j+2}^n) - \tilde{f}(y_{j-2}^n, y_{j-1}^n, y_j^n, y_{j+1}^n) \Big], \tag{54}$$

where the numerical flux is defined by

$$\begin{aligned} \tilde{f}(y_{j-1}, y_j, y_{j+1}, y_{j+2}) := {} & (\alpha_{21}\beta_{10} + \beta_{20}) \hat{f}(y_j, y_{j+1}) \\ & + \alpha_{10}\beta_{21} \hat{f}(\mathcal{H}(y_{j-1}, y_j, y_{j+1}), \mathcal{H}(y_j, y_{j+1}, y_{j+2})), \end{aligned}$$

after eliminating the inner stage. We state the convergence result in the following proposition.

Proposition 4.7. *Suppose the final end data of the continuous adjoint satisfies $p(x,T) \in \mathrm{C}^{0,1}(\mathbb{R})$. Moreover, suppose that the space discretization is either the Lax-Friedrichs or Engquist-Osher method and that the time is discretized using a TVD-RK method. Let the time-step $k = c\,h$ be chosen such that the forward Euler discretizations* (26) *are monotone and their corresponding discrete adjoint scheme is TVD, see* (22), (27)*. Then the discrete adjoint is convergent to the unique reversible solution as $k = c\,h \to 0$, i.e.,*

$$p_h \to p \quad \text{in } B([0,T]; \mathrm{L}^r(I)),$$

where $B([0,T]; L^r(I))$ is the space of bounded functions equipped with the $\sup$-norm and with values in $L^r(I)$ for $r \in [1,\infty)$ and $I := (-R,R)$ for all $R > 0$.

Proof. We give the main steps of the proof from [Ulb01, Chapter 6.4]. First, we eliminate the inner stages and write the TVD-RK method in a conservative form (see (54) for the two stage method):

$$y_j^{n+1} = y_j^n - \frac{k}{h}\big[\tilde{f}_{j+1/2}^n - \tilde{f}_{j-1/2}^n\big], \quad \tilde{f}_{j+1/2}^n := \tilde{f}(y_{j-K+1}^n, \dots, y_{j+K}^n).$$

Therefore it is of the form (6.9) in [Ulb01]. Then the discrete adjoint can be written as

$$p_j^n = p_j^{n+1} + \frac{k}{h} \sum_{i=1-K}^{K} a_{j-i+1/2,i}^n \, (p_{j-i+1}^{n+1} - p_{j-i}^{n+1}), \tag{55}$$

where $a_{j-i+1/2,i}^n = \partial_{y_i} \tilde{f}_{j-i+1/2}^n$. For the L^∞-stability, it is more convenient to write the discrete adjoint scheme in the following form:

$$p_j^n = \sum_{i=-K}^{K} B_{j,i}^n p_{j-i}^{n+1},$$

where $B_{j,-K}^n = \frac{k}{h} a_{j+K-\frac{1}{2},1-K}^n$, $B_{j,K}^n = -\frac{k}{h} a_{j-K+\frac{1}{2},K}^n$ and

$$B_{j,i}^n = \delta_{0,i} + \frac{k}{h}(a_{j-i-\frac{1}{2},i+1}^n - a_{j-i+\frac{1}{2},i}^n), \quad -K < i < K,$$

where $\delta_{0,i}$ is the Kronecker delta. For the stability of the total variation, however, it is more convenient to write the discrete adjoint scheme in the form

$$(p_{j+1}^n - p_j^n) = \sum_{i=-K}^{K} C_{j,i}^n (p_{j-i+1}^{n+1} - p_{j-i}^{n+1}), \tag{56}$$

where $C_{j,-K}^n = \frac{k}{h} a_{j+K+\frac{1}{2},1-K}^n$, $C_{j,K}^n = -\frac{k}{h} a_{j-K+\frac{1}{2},K}^n$ and

$$C_{j,i}^n = \delta_{0,i} + \frac{k}{h}(a_{j-i+\frac{1}{2},i+1}^n - a_{j-i+\frac{1}{2},i}^n), \quad -K < i < K.$$

Monotonicity of the TVD-RK method applied to the forward problem ensures that $B_{j,i}^n \geq 0$ (see [Ulb01, Lemma 6.4.2]). Moreover, the TVD-RK adjoint scheme is TVD by Theorem 4.4 and maps by (55) constant values $p_{j-K}^{n+1} = \ldots = p_{j+K}^{n+1} = c$ to the same value $p_j^n = c$. Hence, [Har83b, Theorem 2.1] yields that the TVD-RK adjoint scheme is monotonicity perserving and thus (56) implies $C_{j,i}^n \geq 0$. This shows that condition (1) in [Ulb01, 6.4.1] is satisfied.

It remain to show that assumptions (D1), (D2) and (D3) of [Ulb01, 6.4.1] hold. (D1) is consistency of the numerical flux which holds in our case; observe for instance that $\tilde{f}(y, y, y, y) = f(y)$ in (53) since $\alpha_{21}\beta_{10} + \alpha_{10}\beta_{21} + \beta_{20} = 1$. (D2) is the convergence of the discrete state to the entropy solution of the state which holds as our discretization is TVD. (D3) is an OSLC condition which holds for Lax-Friedrichs and Engquist-Osher discretization (see [Ulb01, Lemma 6.5.2 and Lemma 6.5.5] for details). Then we can apply Theorem 6.4.4 and Theorem 6.4.6 in [Ulb01] to show convergence of the discrete adjoint to the unique reversible solution. □

5 Numerical experiments

In this section we perform numerical experiments on TVD-RK methods for computing the discrete adjoint state. We show through numerical experiments that the adjoint scheme obtained from discretization of the forward problem using a TVD-RK method is stable.

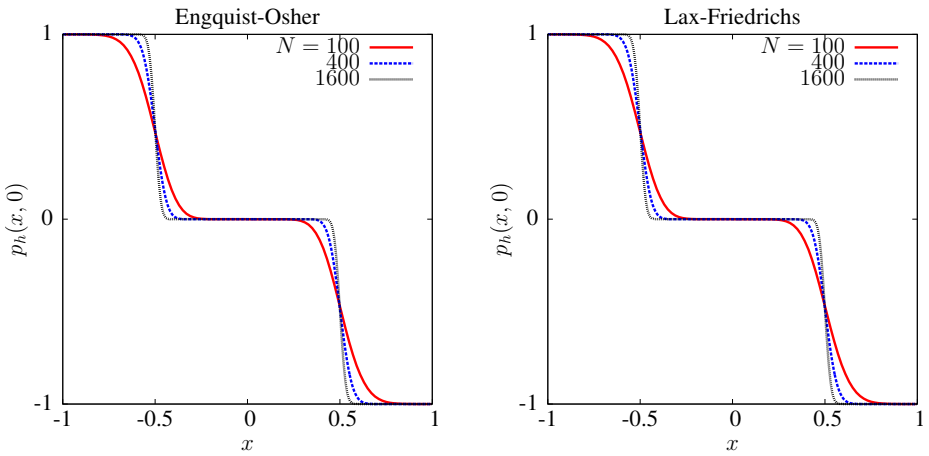

Figure 2: Discrete adjoint computed using Engquist-Osher and Lax-Friedrichs schemes with two-stage TVD-RK method.

Let us consider the configuration of Example 2.2. The domain is set to be $\Omega = (-1, 1)$, the flux function $f(y) := \frac{1}{2}y^2$ and $u(x) = 1$ for $x \in [-1, 0)$ and $u(x) = -1$ for $x \in [0, 1]$. We set the function $G(y(x, T)) := \frac{1}{2}|y(x, T)|^2$. For the boundary condition we choose $y(-1, t) = 1$ and $y(1, t) = -1$. We then use the Lax-Friedrichs and Engquist-Osher schemes to compute $\{\boldsymbol{y}_h^n\}_{n=1}^{n_T}$ with the time-step

$$k = \frac{1}{2}\Big(\min_{\alpha_{ij},\beta_{ij}\neq 0} \frac{\alpha_{ij}}{\beta_{ij}}\Big)\gamma\, h,$$

where $\gamma = \frac{1}{2}$ is the optimal CFL constant for both the forward and adjoint discretization as discussed in Section 3.3. Here, as before, h is the mesh parameter and is inversely proportional to the number of cells N. We use the TVD-RK methods of Table 1. As shown in Theorem 4.4 the discrete adjoint is also stable provided the TVD-RK method for the forward problem is SSP. Moreover we expect that the discrete adjoint approximates the continuous adjoint at $t = 0$,

$$p(x, 0) = \begin{cases} 1 & -1 < x < -\frac{1}{2}, \\ 0 & -\frac{1}{2} \le x \le \frac{1}{2}, \\ -1 & -\frac{1}{2} < x < 1, \end{cases}$$

even though the discrete adjoint scheme does *not* impose the "interior" boundary condition (9).

In Figure 2, we observe that for both the Engquist-Osher and Lax-Friedrichs methods with a two-stage TVD-RK method, we obtain a stable TVD discrete adjoint. Note that in the interval $x \in (-\frac{1}{2}, \frac{1}{2})$, the discrete adjoint has the correct value, i.e., $p(x, 0) = 0$, and the shock location is correct as well. Moreover, the discrete adjoint converges as we refine the mesh. Identical results are obtained using a three-stage TVD-RK method. The reason for this is that the Lax-Friedrichs and Engquist-Osher methods are low order methods while the time discretization is high order and the leading error is due to the spatial discretization.

In Figure 3, we observe that the total variation of the discrete adjoint at any time $t < T$ deviates from the final time discrete adjoint, i.e., $\boldsymbol{p}_h^{n_T}$, only up to machine precision. This implies that the TVD-RK method for the discrete adjoint is TV stable, even though it is not SSP.

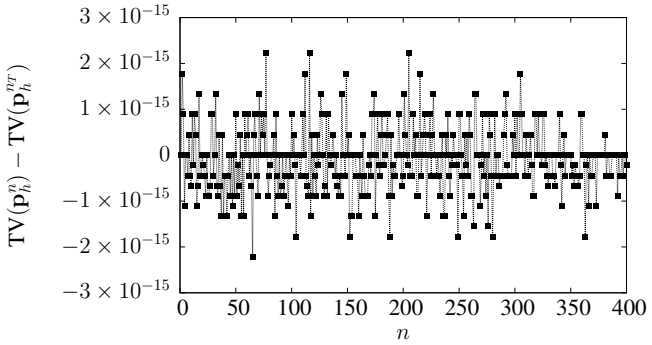

Figure 3: The difference between total variation of the discrete adjoint at time t_n and total variation of the final time discrete adjoint.

5.1 Giles' test case

In this section we perform numerical experiments on an optimal control problem which was first proposed by M. Giles in [Gil03]. Let us choose the setting of the problem of the previous section except for the objective functional which we now choose to be $G(y) := y^5 - y$. Note that since $y(x,T) = 1$ for $x \in [-1,0)$ and $y(x,T) = -1$ for $x \in (0,1]$ we have $G'(y(x,T)) = 4$ for $x \in [-1,1] \setminus \{0\}$. Moreover since there is a shock in the solution at time T at $x = 0$ we should impose an "interior" boundary condition for the adjoint state. Since $[\![G(y(x,T))]\!] = 0$, we should set the "interior" boundary condition to

$$p(0,t) = 0 \quad \forall t \in [0,T].$$

Therefore the adjoint state at time $t = 0$ reads

$$p(x,0) = \begin{cases} 4 & -1 < x < -\frac{1}{2}, \\ 0 & -\frac{1}{2} \le x \le \frac{1}{2}, \\ 4 & -\frac{1}{2} < x < 1. \end{cases}$$

If we choose the discretize-then-differentiate approach, we do not impose such an "interior" boundary condition for the discrete adjoint. It has been first observed by M. Giles in [Gil03] that the Lax-Friedrichs scheme provides a discrete adjoint that converges to a wrong adjoint; see Remark 3.2. We perform the numerical experiment for the Engquist-Osher scheme with a two-stage TVD-RK method and the number of cells $N = 800$. In Figure 4, we observe that discrete adjoint has the correct value for $x \in [-1,-\frac{1}{2}) \cup (\frac{1}{2},1]$ but it has a wrong value for $x \in [-\frac{1}{2},\frac{1}{2}]$, i.e., $p_h(x,0) = 0.25$ for $x \in [-\frac{1}{2},\frac{1}{2}]$. The wrong value in this region does not improve under refinement and the approximation converges to the value 0.25 in this region. Note that the final-time condition $p_h(0,T)$ has a Dirac delta shape at $x = 0$ due to the non-linearity of the objective functional with the value of 0.25. This is precisely the value that is transported backward in time.

It is shown in [GU10a, GU10b] that the Lax-Friedrichs scheme converges to the correct adjoint provided a restrictive time-step of type $k = O(h^{2-q})$, for $0 < q < 1$; see Figure 4 (right). Observe that in Figure 5 that Lax-Friedrichs method converges in the shock funnel with $k = O(h^{1.2})$.

Such a time-step increases diffusion in the scheme and allows more grid points to enter the shock and hence it leads to convergence of the discrete adjoint to the correct solution. This, however, is not possible with the classical definition of the Engquist-Osher scheme as such a diffusion is not present.

In Figure 6, we plot the deviation of the total variation of the discrete adjoint from the total variation of

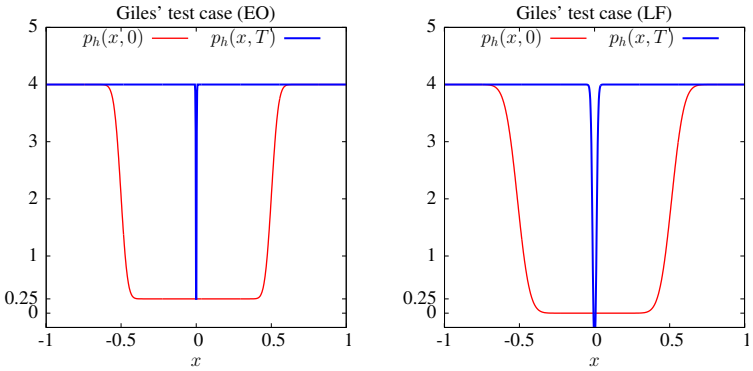

Figure 4: The discrete adjoint computed using Engquist-Osher (left) and Lax-Friedrichs (right) schemes and a two-stage TVD-RK method for Giles' test case.

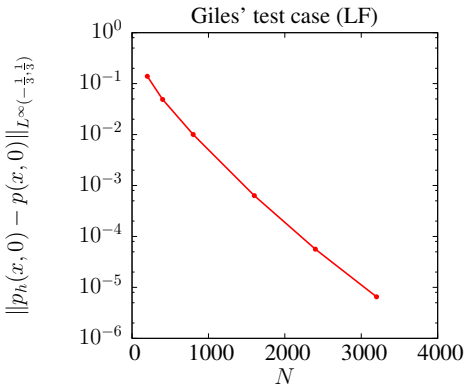

Figure 5: Convergence of the discrete adjoint computed using the Lax-Friedrich method.

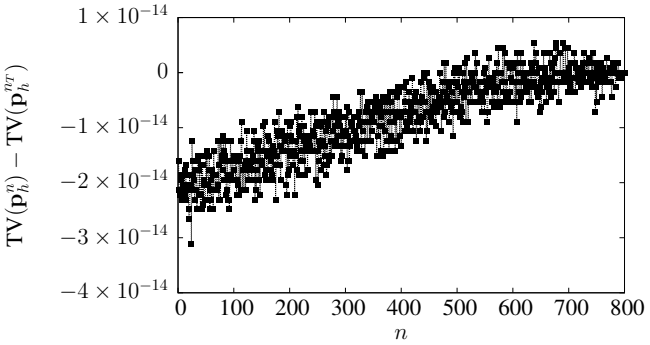

Figure 6: The difference between total variation of the discrete adjoint at time t_n and total variation of the final time discrete adjoint for Giles' test case.

$p_h(x,T)$. Observe that we have $p_h(x,t) \leq p_h(x,T)$ for $t \in [0,T]$ up to machine precision which agrees with the theoretical stability result for TVD-RK methods.

5.2 A numerical optimal control problem

In this section we solve an optimization task using the gradient information obtained from a discrete adjoint.

We set the objective functional to

$$J(y) := \frac{1}{2}\int_{-1}^{1} |y(x,T) - y_{\text{obs}}(x)|^2 \mathrm{d}x,$$

with the final time $T = \frac{1}{2}$,

$$y_{\text{obs}}(x) := \begin{cases} 2x - \frac{1}{2} & \frac{1}{4} \leq x \leq \frac{3}{4}, \\ 0 & \text{otherwise}, \end{cases}$$

and Burgers' flux function $f(y) = \frac{1}{2}y^2$. Given the current control u_j (at iteration j) we compute the discrete adjoint at time $T = 0$, i.e., $p_j(x,0)$ and choose $\delta u = -\eta_j p_j(x,0)$ where $\eta_j \in \mathbb{R}_+$ is a parameter to ensure $J(y_h(u_{j+1})) < J(y_h(u_j))$. Then the updated control at iteration $j+1$ reads

$$u_{j+1}(x) = u_j(x) - \eta_j\, p_j(x,0). \tag{57}$$

We choose η_j by checking Armijo's condition and a back-tracking procedure, i.e.,

$$J(y_h(u_{j+1})) \leq J(y_h(u_j)) - c_{\text{opt}}\eta_j \|p_j(x,0)\|^2_{\mathrm{L}^2(-1,1)}, \tag{58}$$

where $c_{\text{opt}} \in (0,1)$ is the Armijo's constant. If the above Armijo's condition is not satisfied we choose a smaller α by

$$\alpha_{\text{new}} := \rho\, \alpha_{\text{old}},$$

where $0 < \rho < 1$ and recheck (58). In this numerical experiment, we choose $\rho = 0.95$, $c_{\text{opt}} = 0.5$ and the initial $\alpha = 0.5$. As initial guess we fix

$$u_0(x) := \begin{cases} (\frac{3}{4} + x)(\frac{1}{2} - x) & -\frac{3}{4} \leq x \leq \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

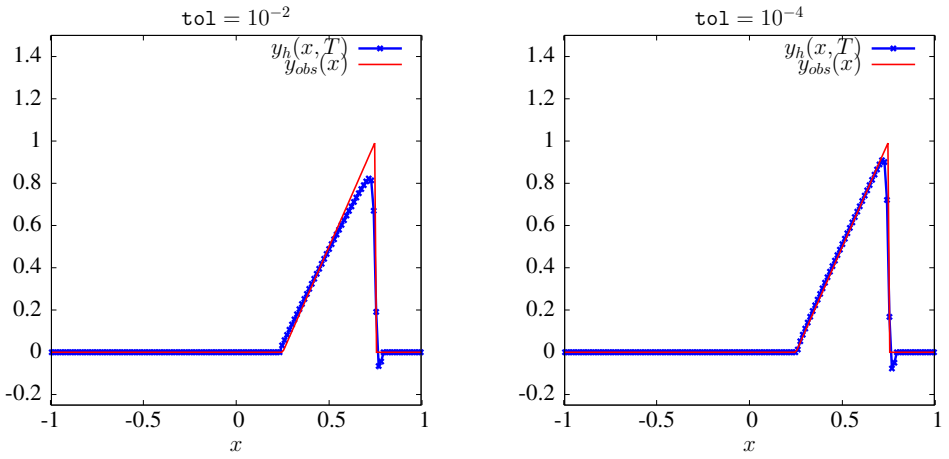

Figure 7: The state variable $y_h(x,T)$ obtained from optimization algorithm with $\texttt{tol} = 10^{-2}$ and $\texttt{tol} = 10^{-4}$.

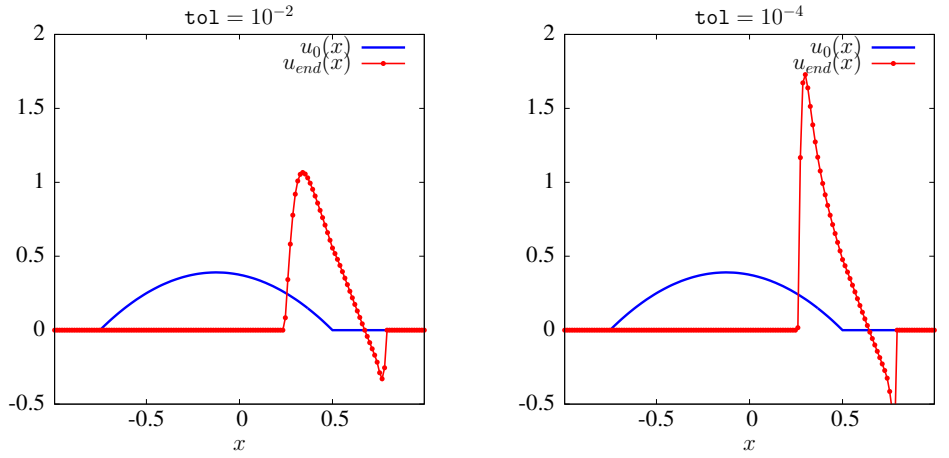

Figure 8: The control variable $u(x)$ obtained from optimization algorithm with $\texttt{tol} = 10^{-2}$ and $\texttt{tol} = 10^{-4}$.

We choose the Engquist-Osher method for spatial discretization and Heun's second order TVD-RK method. The time step is chosen as $k = 0.25h$ and the stopping criteria is taken to be $\|\nabla \mathcal{J}_h(u_h)\|_{\mathrm{L}^2(\Omega)} = \|p_j(x,0)\|_{\mathrm{L}^2(\Omega)} \leq \texttt{tol}$. In our experiments we set $\texttt{tol} = 10^{-2}$ and 10^{-4}.

In Figure 7, we observe that the numerical algorithm seems to converge to the true solution. That is, it captures correctly the shock location at $x = \frac{3}{4}$ and also the rare-faction. There are numerical artifacts at $x = \frac{3}{4}$ which vanish as we reduce the tolerance tol. The corresponding initial guess for the control variable and the final control variable are plotted in Figure 8.

6 Conclusion

In this paper we studied TVD-RK methods for the numerical treatment of the optimal control problems in a discretize-then-optimize approach. We have shown that a TVD-RK discretization of the state equation yields a TVD-RK method for the discrete adjoint with a conjugate coefficient table. We then showed that requiring SSP for both discrete state and adjoint is too strong and results in a first order method. Luckily, imposing SSP for the discrete state is sufficient to obtain stability of the discrete

adjoint. This result holds for an arbitrary s-stage TVD-RK method. Moreover, thanks to the linearity of the adjoint equation, the TVD-RK method for the discrete adjoint is consistent. We also studied the approximation properties of the discrete adjoint and showed that for a second order two-stage method applied to the forward problem, we obtain a second order discrete adjoint too. However for the third order three-stage method applied, we only obtain a second order discrete adjoint. Our theoretical results were finally illustrated by numerical experiments.

We would like to finish this conclusion by mentioning that the convergence of the discrete adjoint to the continuous adjoint is an interesting question of its own.

A Existence of a minimizer

Proposition A.1. *Let $G(y) = \frac{1}{2}|y(x,T) - y_{obs}(x)|^2$. Then the optimal control problem subject to the conservation law (1) has a solution in the admissible set U_{ad}.*

Proof. We begin with the continuity of the objective functional. Suppose $y(x,t)$ and $w(x,t)$ are the entropic solution of (1) with the initial data $u(x) \in U_{\mathsf{ad}}$ and $v(x) \in U_{\mathsf{ad}}$, respectively. Then we have

$$|J(y) - J(w)| = \frac{1}{2}\Big| \int_{\mathbb{R}} (y-w)(y+w-2y_{\mathsf{obs}})\Big| \leq \|y-w\|_{\mathrm{L}^1(\mathbb{R})}\|y+w-2y_{\mathsf{obs}}\|_{\mathrm{L}^\infty(\mathbb{R})}.$$

Then by L^1-contraction of the entropic solutions we have

$$\begin{aligned} |J(y) - J(w)| &\leq \|y(\cdot,T) - w(\cdot,T)\|_{\mathrm{L}^1(\mathbb{R})}(\|y(\cdot,T)\|_{\mathrm{L}^\infty(\mathbb{R})} + \|w(\cdot,T)\|_{\mathrm{L}^\infty(\mathbb{R})} + c) \\ &\leq \|u-v\|_{\mathrm{L}^1(\mathbb{R})}(\|u\|_{\mathrm{L}^\infty(\mathbb{R})} + \|v\|_{\mathrm{L}^\infty(\mathbb{R})} + c) \\ &\leq C\|u-v\|_{\mathrm{L}^1(\mathbb{R})}, \end{aligned}$$

where in the last step we used the L^∞ stability of the entropic solutions and the fact that u and v are in U_{ad} and therefore we have a uniform bound on the L^∞-norm.

Observe that $J(y) \geq 0$ and therefore a minimizing sequence denoted by $\{u_i\}$ exists. Since U_{ad} is a compact set in L^1, one can obtain a subsequence denoted by $\{u_{i_j}\}$ that converges strongly in $\mathrm{L}^1(\mathbb{R})$ to $u^\star \in U_{\mathsf{ad}}$ as $j \to \infty$. Then using continuity of the objective functional we have

$$\inf_{u \in U_{\mathsf{ad}}} J(y) = \lim_{j\to\infty} J(y(u_{i_j})) = J(y(u^\star)) \quad \text{for } u^\star \in U_{\mathsf{ad}},$$

which shows existence of the minimizer. □

We would like to mention also that for a more general admissible set, e.g.,

$$U_{\mathsf{ad}} := \big\{u \in \mathrm{L}^\infty(\mathbb{R}) : \mathrm{supp}(u) \in K, \|u\|_{\mathrm{L}^\infty(\mathbb{R})} \leq C\big\}, \tag{59}$$

and an assumption on the uniform convexity of the flux function $f(\cdot)$, one can also obtain an existence result. We refer the reader to [CPZ08] for the proof.

B Proof of Proposition 3.1

Proof. Since $f(\cdot)$ is C^2 we have $f(w+v) = f(w) + f'(w)v + \frac{1}{2}f''(z)v^2$ for $w, v \in \mathbb{R}$ and some $z \in (w, w+v)$. For the Lax-Friedrichs flux, a direct calculation shows

$$\hat{f}(w_j + v_j, w_{j+1} + v_{j+1}) = \hat{f}(w_j, w_{j+1}) + g^{\mathsf{LF}}_{i+1,i} + O(v_j^2) + O(v_{j+1}^2) + O(v_{j-1}^2).$$

For Engquist-Osher we have

$$\hat{f}(w_j + v_j, w_{j+1} + v_{j+1}) = \hat{f}(w_j, w_{j+1}) + \int_{w_j}^{w_j+v_j} f'(s)^+ \,\mathrm{d}s + \int_{w_{j+1}}^{w_{j+1}+v_{j+1}} f'(s)^- \,\mathrm{d}s.$$

Note that we can write $f'(x)^+ = \frac{1}{2}(f'(x) + |f'(x)|)$ and $f'(x)^- = \frac{1}{2}(f'(x) - |f'(x)|)$. Recall the definition of $g^{\mathsf{EO}}_{j,j+1}$ in the proposition and define the residual $r(\boldsymbol{w}, \boldsymbol{v}) : V_h \times V_h \to V_h$ by $[r(\boldsymbol{w}, \boldsymbol{v})]_j := [F_h(\boldsymbol{w} + \boldsymbol{v})]_j - [F_h(\boldsymbol{w})]_j - [F_h'(\boldsymbol{w})\boldsymbol{v}]_j$. Then we have $[r(\boldsymbol{w}, \boldsymbol{v})]_j = q_{j,j+1} - q_{j-1,j}$ where

$$q_{j,j+1} = \int_{w_j}^{w_j+v_j} f'(s)^+ - f'(w_j)^+ \,\mathrm{d}s + \int_{w_{j+1}}^{w_{j+1}+v_{j+1}} f'(s)^- - f'(w_{j+1})^- \,\mathrm{d}s.$$

For the first term on the right-hand side we have

$$\Big| \int_{w_j}^{w_j+v_j} f'(s)^+ - f'(w_j)^+ \,\mathrm{d}s \Big| \leq \max_{z \in (w_j, w_j+v_j)} |f'(z)^+ - f'(w_j)^+| \cdot |v_j| \leq C\, v_j^2,$$

since $f'(x)^+$ is Lipschitz continuous. The proof for the second term of $q_{j,j+1}$ is similar.

Hence we showed that for both Lax-Friedrichs and Engquist-Osher fluxes we have $[r(\boldsymbol{w}, \boldsymbol{v})]_j = O(v_j^2) + O(v_{j+1}^2) + O(v_{j-1}^2)$, respectively. Taking the ℓ^1-norm of $r(\boldsymbol{w}, \boldsymbol{v})$ one obtains

$$\|r(\boldsymbol{w}, \boldsymbol{v})\|_{\ell^1} \leq C\|\boldsymbol{v}\|_{\ell^2}^2 \leq C\|\boldsymbol{v}\|_{\ell^1}^2,$$

where C is independent of the mesh parameter. Multiplying both sides by h gives

$$\|r(w, v)\|_{\mathrm{L}^1(\Omega)} \leq C\|v\|_{\mathrm{L}^1(\Omega)}\, \|\boldsymbol{v}\|_{\ell^1},$$

which shows that $\|r(w, v)\|_{\mathrm{L}^1(\Omega)}/\|v\|_{\mathrm{L}^1(\Omega)} \to 0$ uniformly when $\boldsymbol{v}$ converges to zero. This completes the proof for the derivative. A direct calculation yields the formula for the transpose. □

References

[BG97] A. Bressan and G. Guerra. Shift-differentiability of the flow generated by a conservation law. *Discrete Contin. Dynam. Systems*, 3(1):35–58, 1997.

[BH12] M. K. Banda and M. Herty. Adjoint IMEX-based schemes for the numerical solution of optimal control problems governed by conservation laws. In *Hyperbolic problems—theory, numerics and applications. Volume 1*, volume 17 of *Ser. Contemp. Appl. Math. CAM*, pages 297–303. World Sci. Publishing, Singapore, 2012.

[BJ98] F. Bouchut and F. James. One-dimensional transport equations with discontinuous coefficients. *Nonlinear Anal.*, 32(7):891–933, 1998.

[CG09] G. A. Chechkin and A. Yu. Goritsky. S. N. Kruzhkov's lectures on first-order quasilinear PDEs. In *Analytical and numerical aspects of partial differential equations*, pages 1–67. Walter de Gruyter, Berlin, 2009. Translated from the Russian by Boris P. Andreianov.

[Con67] E. D. Conway. Generalized solutions of linear differential equations with discontinuous coefficients and the uniqueness question for multidimensional quasilinear conservation laws. *J. Math. Anal. Appl.*, 18:238–251, 1967.

[CPZ08] C. Castro, F. Palacios, and E. Zuazua. an alternating descent method for the optimal control of the inviscid Burgers equation in the presence of shocks. *Mathematical Models and Methods in Applied Sciences*, 18(03):369–416, 2008.

[Eva10] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[Gil03] M. Giles. Discrete adjoint approximations with shocks. In *Hyperbolic problems: theory, numerics, applications*, pages 185–194. Springer, Berlin, 2003.

[GS98] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67(221):73–85, 1998.

[GU10a] M. Giles and S. Ulbrich. Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals. *SIAM J. Numer. Anal.*, 48(3):882–904, 2010.

[GU10b] M. Giles and S. Ulbrich. Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint approximations and extensions. *SIAM J. Numer. Anal.*, 48(3):905–921, 2010.

[Hag00] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numer. Math.*, 87(2):247–282, 2000.

[Har83a] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49(3):357–393, 1983.

[Har83b] Ami Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49(3):357–393, 1983.

[JS99] F. James and M. Sepúlveda. Convergence results for the flux identification in a scalar conservation law. *SIAM J. Control Optim.*, 37(3):869–891 (electronic), 1999.

[Kru70] S. N. Kružkov. First order quasilinear equations with several independent variables. *Mat. Sb. (N.S.)*, 81 (123):228–255, 1970.

[LeV90] R. J. LeVeque. *Numerical methods for conservation laws*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1990.

[Luc86] B. J. Lucier. A moving mesh numerical method for hyperbolic conservation laws. *Math. Comp.*, 46(173):59–69, 1986.

[RS02] S. J. Ruuth and R. J. Spiteri. Two barriers on strong-stability-preserving time discretization methods. In *Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala)*, volume 17, pages 211–220. Springer, 2002.

[SO88] C.-W. Shu and S. Osher. Efficient implementation of essentially nonoscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.

[SR02] R. J. Spiteri and S. J. Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM J. Numer. Anal.*, 40(2):469–491 (electronic), 2002.

[Ulb99] S. Ulbrich. On the existence and approximation of solutions for the optimal control of nonlinear hyperbolic conservation laws. In *Optimal control of partial differential equations (Chemnitz, 1998)*, volume 133 of *Internat. Ser. Numer. Math.*, pages 287–299. Birkhäuser, Basel, 1999.

[Ulb01] S. Ulbrich. Optimal control of nonlinear hyperbolic conservation laws with source terms. *Technische Universität München*, 2001.

[Ulb02] S. Ulbrich. A sensitivity and adjoint calculus for discontinuous solutions of hyperbolic conservation laws with source terms. *SIAM J. Control Optim.*, 41(3):740–797 (electronic), 2002.

[Ulb03] S. Ulbrich. Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws. *Systems Control Lett.*, 48(3-4):313–328, 2003. Optimization and control of distributed systems.