MATHEMATICS OF COMPUTATION
Volume 87, Number 313, September 2018, Pages 2295–2320
http://dx.doi.org/10.1090/mcom/3296
Article electronically published on February 20, 2018

# STRONG-STABILITY-PRESERVING ADDITIVE LINEAR MULTISTEP METHODS

YIANNIS HADJIMICHAEL AND DAVID I. KETCHESON

Abstract. The analysis of strong-stability-preserving (SSP) linear multistep methods is extended to semi-discretized problems for which different terms on the right-hand side satisfy different forward Euler (or circle) conditions. Optimal perturbed and additive monotonicity-preserving linear multistep methods are studied in the context of such problems. Optimal perturbed methods attain larger monotonicity-preserving step sizes when the different forward Euler conditions are taken into account. On the other hand, we show that optimal SSP additive methods achieve a monotonicity-preserving step-size restriction no better than that of the corresponding nonadditive SSP linear multistep methods.

## 1. Introduction

We are interested in numerical solutions of initial value ODEs,

$$\begin{aligned} \boldsymbol{u}'(t) &= \boldsymbol{F}(\boldsymbol{u}(t)), \quad t \geq t_0, \\ \boldsymbol{u}(t_0) &= \boldsymbol{u}_0, \end{aligned} \tag{1.1}$$

where $\boldsymbol{F} : \mathbb{R}^m \to \mathbb{R}^m$ is a continuous function and $\boldsymbol{u} : [t_0, \infty) \to \mathbb{R}^m$ satisfies a monotonicity property

$$\|\boldsymbol{u}(t + \Delta t)\| \leq \|\boldsymbol{u}(t)\| \quad \forall \Delta t \geq 0, \tag{1.2}$$

with respect to some norm, semi-norm or convex functional $\|\cdot\| : \mathbb{R}^m \to \mathbb{R}$. In general, $\boldsymbol{F}(\boldsymbol{u}(t))$ may arise from the spatial discretization of partial differential equations, for example, hyperbolic conservation laws. A sufficient condition for monotonicity is that there exists some $\Delta t_{\mathrm{FE}} > 0$ such that the forward Euler condition

$$\|\boldsymbol{u} + \Delta t \boldsymbol{F}(\boldsymbol{u})\| \leq \|\boldsymbol{u}\|, \quad 0 \leq \Delta t \leq \Delta t_{\mathrm{FE}}, \tag{1.3}$$

holds for all $\boldsymbol{u} \in \mathbb{R}^m$. In this paper we focus on linear multistep methods (LMMs) for the numerical integration of (1.1). We denote by $\boldsymbol{u}_n$ the numerical approximation to $\boldsymbol{u}(t_n)$, evaluated sequentially at times $t_n = t_0 + n\Delta t$, $n \geq 1$. At step $n$, a $k$-step linear multistep method applied to (1.1) takes the form

$$\boldsymbol{u}_n = \sum_{j=0}^{k-1} \alpha_j \boldsymbol{u}_{n-k+j} + \Delta t \sum_{j=0}^{k} \beta_j \boldsymbol{F}(\boldsymbol{u}_{n-k+j}), \tag{1.4}$$

and if $\beta_k = 0$, then the method is explicit.

Received by the editor April 5, 2016, and, in revised form, December 6, 2016, and April 18, 2017.

2010 *Mathematics Subject Classification.* Primary, 65L06; Secondary, 65L05, 65M20.

This work was supported by the King Abdullah University of Science and Technology (KAUST), 4700 Thuwal, 23955-6900, Saudi Arabia.

We would like to establish a discrete analogue of (1.2) for the numerical solution $\boldsymbol{u}_n$ in (1.4). Consider the explicit case, and assume $\boldsymbol{F}$ satisfies the forward Euler condition (1.3) and all $\alpha_j, \beta_j$ are nonnegative. Then, convexity of $\|\cdot\|$ and the consistency requirement $\sum_{j=0}^{k-1} \alpha_j = 1$ imply that $\|\boldsymbol{u}_n\| \leq \max_j \|\boldsymbol{u}_{n-k+j}\|$ whenever $\Delta t \beta_j/\alpha_j \leq \Delta t_{\text{FE}}$ for each $j \in \{0, \ldots, k-1\}$. Hence, the monotonicity condition

$$\|\boldsymbol{u}_n\| \leq \max\{\|\boldsymbol{u}_{n-1}\|, \ldots, \|\boldsymbol{u}_{n-k}\|\} \tag{1.5}$$

is satisfied under a step-size restriction,

$$\Delta t \leq \mathcal{C}_{\text{LMM}} \Delta t_{\text{FE}}, \tag{1.6}$$

where $\mathcal{C}_{\text{LMM}} = \min_j \alpha_j/\beta_j$. The ratio $\alpha_j/\beta_j$ is taken to be infinity if $\beta_j = 0$. The above also holds for implicit methods under a technical rescaling of the coefficients[1]. See [3, Chapter 8] and references therein for a review of strong-stability-preserving linear multistep methods (SSP LMMs).

Most LMMs have one or more negative coefficients, so the foregoing analysis leads to $\mathcal{C}_{\text{LMM}} = 0$ and thus monotonicity condition (1.5) cannot be guaranteed by positive step sizes. Typical numerical methods for hyperbolic conservation laws $\boldsymbol{U}_t + \nabla \cdot \boldsymbol{f}(\boldsymbol{U}) = 0$ involve upwind-biased semi-discretizations of the spatial derivatives. In order to preserve monotonicity when using methods with negative coefficients for such semi-discretizations, downwind-biased spatial approximations may be used. Let $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ be, respectively, upwind- and downwind-biased approximations of $-\nabla \cdot \boldsymbol{f}(\boldsymbol{U})$. It is natural to assume that $\widetilde{\boldsymbol{F}}$ satisfies

$$\|\boldsymbol{u} - \Delta t \widetilde{\boldsymbol{F}}(\boldsymbol{u})\| \leq \|\boldsymbol{u}\|, \quad 0 \leq \Delta t \leq \Delta t_{\text{FE}} \tag{1.7}$$

for all $\boldsymbol{u} \in \mathbb{R}^m$. A linear multistep method that uses both $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ can be then written as

$$\boldsymbol{u}_n = \sum_{j=0}^{k-1} \alpha_j \boldsymbol{u}_{n-k+j} + \Delta t \sum_{j=0}^{k} \left( \beta_j \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) - \tilde{\beta}_j \widetilde{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \right). \tag{1.8}$$

If all coefficients $\alpha_j, \beta_j, \tilde{\beta}_j$ are nonnegative, then the SSP coefficient of method (1.8) is [10, Section 3]

$$\widetilde{\mathcal{C}}_{\text{LMM}} = \sup \left\{ r \mid \alpha_j - r(\beta_j + \tilde{\beta}_j) \geq 0 \right\} = \min_j \frac{\alpha_j}{\beta_j + \tilde{\beta}_j}.$$

If $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ satisfy (1.3) and (1.7), then the solution given by (1.8) satisfies (1.5) whenever the time step satisfies

$$\Delta t \leq \widetilde{\mathcal{C}}_{\text{LMM}} \Delta t_{\text{FE}}.$$

Downwind LMMs were originally introduced in [20, 21], with the idea that $\boldsymbol{F}$ be replaced by $\widetilde{\boldsymbol{F}}$ whenever $\beta_j < 0$. Optimal explicit linear multistep schemes of order up to six, coupled with efficient upwind and downwind WENO discretizations, were studied in [4]. Coefficients of optimal upwind- and downwind-biased methods together with a reformulation of the nonlinear optimization problem involved as a series of linear programming feasibility problems can be found in [10]. Bounds on

[1]For implicit methods convexity can be used after rescaling the coefficients of (1.4) by multiplying both sides by $1 - \alpha_k$, where $\alpha_k = \mathcal{C}_{\text{LMM}} \beta_k/(1 + \mathcal{C}_{\text{LMM}} \beta_k)$. Under this rescaling, the coefficients of (1.4) become $\bar{\alpha}_j = (1 - \alpha_k)\alpha_j$, $j \in \{0, \ldots, k-1\}$, $\bar{\alpha}_k = \alpha_k$, $\bar{\beta}_j = (1 - \alpha_k)\beta_j$, $j \in \{0, \ldots, k\}$, and hence consistency implies $\sum_{j=0}^{k} \bar{\alpha}_j = 1$.

the maximum SSP step size for downwind-biased methods have been analyzed in [11].

Method (1.8) can also be written in the perturbed form

(1.9)
$$\boldsymbol{u}_n = \sum_{j=0}^{k-1} \alpha_j \boldsymbol{u}_{n-k+j} + \Delta t \sum_{j=0}^{k} \left( \hat{\beta}_j \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) + \tilde{\beta}_j \left( \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) - \widetilde{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \right) \right),$$

where $\hat{\beta}_j = \beta_j - \tilde{\beta}_j$. We say method (1.9) is a perturbation of the LMM (1.4) with coefficients $\hat{\beta}_j$, and the latter is referred to as the *underlying method* for (1.9). By replacing $\widetilde{\boldsymbol{F}}$ with $\boldsymbol{F}$ in (1.9) one recovers the underlying method. The notion of a perturbed method can be useful beyond the realm of downwinding for hyperbolic PDE semi-discretizations. If $\boldsymbol{F}$ satisfies the forward Euler condition (1.3) for both positive and negative step sizes, then we can simply take $\widetilde{\boldsymbol{F}} = \boldsymbol{F}$. In such cases, the perturbed and underlying methods are the same, but analysis of a perturbed form of the method can yield a larger step size for monotonicity, giving more accurate insight into the behavior of the method. See [7] for a discussion of this in the context of Runge–Kutta methods, and see Example 2.17 herein for an example using multistep methods. As we will see in Section 2, the most useful perturbed LMMs (1.9) take a form in which either $\beta_j$ or $\tilde{\beta}_j$ is equal to zero for each value of $j \in \{0, \dots, k\}$. Thus $\widetilde{\mathcal{C}}_{\text{LMM}} = \min_j\{\alpha_j/\beta_j, \alpha_j/\tilde{\beta}_j\}$, and the class of perturbed LMMs (1.9) coincides with the class of downwind LMMs in [20, 21].

In this work, we adopt form (1.8) for perturbed LMMs and consider their application to the more general class of problems (1.1) for which $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ satisfy forward Euler conditions under different step-size restrictions:

$$\|\boldsymbol{u} + \Delta t \boldsymbol{F}(\boldsymbol{u})\| \le \|\boldsymbol{u}\| \quad \forall \boldsymbol{u} \in \mathbb{R}^m,\ 0 \le \Delta t \le \Delta t_{\text{FE}}, \tag{1.10a}$$

$$\|\boldsymbol{u} - \Delta t \widetilde{\boldsymbol{F}}(\boldsymbol{u})\| \le \|\boldsymbol{u}\| \quad \forall \boldsymbol{u} \in \mathbb{R}^m,\ 0 \le \Delta t \le \widetilde{\Delta t}_{\text{FE}}. \tag{1.10b}$$

For a fixed order of accuracy and number of steps, an optimal SSP method is defined to be any method that attains the largest possible SSP coefficient. The choice of optimal monotonicity-preserving method for a given problem will depend on the ratio $\Delta t_{\text{FE}}/\widetilde{\Delta t}_{\text{FE}}$. We analyze and construct such optimal methods. We illustrate by examples that perturbed LMMs with larger step sizes for monotonicity can be obtained when the different step sizes in (1.10) are accounted for.

The perturbed methods (1.8) are reminiscent of additive methods, and the latter can be analyzed in a similar way. Consider the problem

$$\boldsymbol{u}'(t) = \boldsymbol{F}(\boldsymbol{u}(t)) + \widehat{\boldsymbol{F}}(\boldsymbol{u}(t)),$$

where $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$ may represent different physical processes, such as convection and diffusion, or convection and reaction. Additive methods are expressed as

$$\boldsymbol{u}_n = \sum_{j=0}^{k-1} \alpha_j \boldsymbol{u}_{n-k+j} + \Delta t \sum_{j=0}^{k} \left( \beta_j \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) + \hat{\beta}_j \widehat{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \right),$$

where $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$ may satisfy the forward Euler condition (1.3) under possibly different step-size restrictions. We prove that optimal SSP explicit or implicit additive methods have coefficients $\beta_j = \hat{\beta}_j$ for each $j \in \{0, \dots, k\}$, hence they lie within the class of ordinary (not additive) LMMs.

The rest of the paper is organized as follows. In Section 2 we analyze the monotonicity properties of perturbed LMMs for which the upwind and downwind operators satisfy different forward Euler conditions. Optimal methods are derived, and their properties are discussed. Their effectiveness is illustrated by some examples. Additive linear multistep methods are presented in Section 3 where we prove that optimal SSP additive LMMs are equivalent to the corresponding nonadditive SSP LMMs. Monotonicity of IMEX linear multistep methods is discussed, and finally in Section 4 we summarize the main results.

## 2. Monotonicity-preserving perturbed linear multistep methods

The following example shows that using upwind- and downwind-biased operators allows the construction of methods that have positive SSP coefficients, even though the underlying methods are not SSP.

**Example 2.1.** Let $u'(t) = F(u(t))$ be a semi-discretization of $u_t + f(u)_x = 0$, where $F \approx -f(u)_x$. Consider the two-step, second-order explicit linear multistep method

$$u_n = \frac{1}{2}u_{n-2} - \frac{1}{4}\Delta t F(u_{n-2}) + \frac{1}{2}u_{n-1} + \frac{7}{4}\Delta t F(u_{n-1}). \tag{2.1}$$

The method has SSP coefficient equal to zero. Let us introduce a downwind-biased operator $\widetilde{F} \approx -f(u)_x$ such that (1.7) is satisfied. Then, a perturbed representation of (2.1) is

$$\begin{aligned} u_n = \frac{1}{2}u_{n-2} + \frac{1}{4}\Delta t F(u_{n-2}) - \frac{1}{2}\Delta t \widetilde{F}(u_{n-2}) \\ + \frac{1}{2}u_{n-1} + 2\Delta t F(u_{n-1}) - \frac{1}{4}\Delta t \widetilde{F}(u_{n-1}), \end{aligned} \tag{2.2}$$

in the sense that the underlying method (2.1) is retrieved from (2.2) by replacing $\widetilde{F}$ with $F$. The perturbed method has SSP coefficient $\widetilde{\mathcal{C}}_{\text{LMM}} = 2/9$. There are infinitely many perturbed representations of (2.1), but an optimal one is obtained by simply replacing $F$ with $\widetilde{F}$ at terms with negative coefficients in (2.1), yielding

$$u_n = \frac{1}{2}u_{n-2} - \frac{1}{4}\Delta t \widetilde{F}(u_{n-2}) + \frac{1}{2}u_{n-1} + \frac{7}{4}\Delta t F(u_{n-1}), \tag{2.3}$$

with SSP coefficient $\widetilde{\mathcal{C}}_{\text{LMM}} = 2/7$.

*Remark* 2.2. A LMM (1.4) has SSP coefficient $\mathcal{C}_{\text{LMM}} = 0$ if any of the following three conditions hold:

(1) $\alpha_j < 0$ for some $j$;
(2) $\beta_j < 0$ for some $j$;
(3) $\alpha_j = 0$ for some $j$ for which $\beta_j \neq 0$.

By introducing a downwind operator we can remedy the second condition, but not the first or the third. Most common methods, including the Adams–Bashforth, Adams–Moulton, and BDF methods, satisfy condition (1) or (3), so they cannot be made SSP via downwinding.

We consider a generalization of the perturbed LMMs described previously, by assuming different forward Euler conditions for the operators $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ (see (1.10)).

**Definition 2.3.** A perturbed LMM of the form (1.8) is said to be *strong-stability-preserving* (SSP) with SSP coefficients $(\mathcal{C}, \widetilde{\mathcal{C}})$ if conditions

$$\begin{aligned} \beta_j, \tilde{\beta}_j \geq 0, \quad j \in \{0, \ldots, k\}, \\ \alpha_j - r\beta_j - \tilde{r}\tilde{\beta}_j \geq 0, \quad j \in \{0, \ldots, k-1\}, \end{aligned} \tag{2.4}$$

hold for all $0 \leq r \leq \mathcal{C}$ and $0 \leq \tilde{r} \leq \widetilde{\mathcal{C}}$.

By plugging the exact solution into (1.8), setting $\widetilde{\boldsymbol{F}}(\boldsymbol{u}(t_n)) = \boldsymbol{F}(\boldsymbol{u}(t_n))$, and taking Taylor expansions around $t_{n-k}$, it can be shown that a perturbed LMM is order $p$ accurate if

$$\begin{aligned} &\sum_{j=0}^{k-1} \alpha_j = 1, \qquad \sum_{j=0}^{k-1} j\alpha_j + \sum_{j=0}^{k} (\beta_j - \tilde{\beta}_j) = k, \\ &\sum_{j=0}^{k-1} \alpha_j j^i + \sum_{j=0}^{k} (\beta_j - \tilde{\beta}_j) i j^{i-1} = k^i, \quad i \in \{2, \ldots, p\}. \end{aligned} \tag{2.5}$$

The step-size restriction for monotonicity of an SSP perturbed LMM is given by the following theorem.

**Theorem 2.4.** *Let $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ be given such that the forward Euler conditions* (1.10) *are satisfied for some $\Delta t_{\mathrm{FE}} > 0$ and $\widetilde{\Delta t}_{\mathrm{FE}} > 0$. Let a consistent perturbed LMM* (1.8) *be given with SSP coefficients $(\mathcal{C}, \widetilde{\mathcal{C}})$. Then the numerical solution satisfies the monotonicity condition* (1.5) *if*

$$0 \leq \Delta t \leq \min\{\mathcal{C}\, \Delta t_{\mathrm{FE}}, \widetilde{\mathcal{C}}\, \widetilde{\Delta t}_{\mathrm{FE}}\}. \tag{2.6}$$

*Proof.* Since the method is SSP with coefficients $(\mathcal{C}, \widetilde{\mathcal{C}})$, then conditions (2.4) hold with $r = \mathcal{C}$ and $\tilde{r} = \widetilde{\mathcal{C}}$. Let $\hat{\alpha}_j = \mathcal{C}\beta_j$ and $\alpha_j = \hat{\alpha}_j + \tilde{\alpha}_j$ for $j \in \{0, \ldots, k-1\}$. Then (2.4) yields $\tilde{\alpha}_j \geq \widetilde{\mathcal{C}}\tilde{\beta}_j$ and $\beta_j \geq 0$, $\tilde{\beta}_j \geq 0$. Define

$$\alpha_k \coloneqq \frac{\mathcal{C}\beta_k + \widetilde{\mathcal{C}}\tilde{\beta}_k}{1 + \mathcal{C}\beta_k + \widetilde{\mathcal{C}}\tilde{\beta}_k},$$

and multiply both sides of (1.8) by $1 - \alpha_k$. Rearranging terms, the perturbed LMM (1.8) can be expressed as

$$\boldsymbol{u}_n = \sum_{j=0}^{k} \left( \alpha_j^* \boldsymbol{u}_{n-k+j} + \Delta t \beta_j^* \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) - \Delta t \tilde{\beta}_j^* \widetilde{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \right), \tag{2.7}$$

where

$$\alpha_j^* = \begin{cases} (1-\alpha_k)\alpha_j, & \text{if } j \in \{0, \ldots, k-1\}, \\ \alpha_k, & \text{if } j = k, \end{cases} \qquad \begin{aligned} \beta_j^* &= (1-\alpha_k)\beta_j, \\ \tilde{\beta}_j^* &= (1-\alpha_k)\tilde{\beta}_j, \end{aligned} \quad j \in \{0, \ldots, k\}.$$

Note that $0 \leq \alpha_k < 1$, hence the nonnegativity of $\alpha_j, \beta_j, \tilde{\beta}_j$ implies that all $\alpha_j^*, \beta_j^*, \tilde{\beta}_j^*$ are also nonnegative. We can split $\alpha_j^*$, $j \in \{0, \ldots, k-1\}$ into two parts, such that $\alpha_j^* = \hat{\alpha}_j^* + \tilde{\alpha}_j^*$, where $\hat{\alpha}_j^* = (1-\alpha_k)\hat{\alpha}_j = \mathcal{C}\beta_j^*$, and $\tilde{\alpha}_j^* = (1-\alpha_k)\tilde{\alpha}_j \geq \widetilde{\mathcal{C}}\tilde{\beta}_j^*$. Consistency requires $\sum_{j=0}^{k-1} \alpha_j = 1$, hence $\sum_{j=0}^{k} \alpha_j^* = 1$. Also, let $\alpha_k^* = \hat{\alpha}_k^* + \tilde{\alpha}_k^*$,

where $\hat{\alpha}_k^* = \mathcal{C}\beta_k/(1+\mathcal{C}\beta_k+\widetilde{\mathcal{C}}\tilde{\beta}_k)$ and $\tilde{\alpha}_k^* = \widetilde{\mathcal{C}}\tilde{\beta}_k/(1+\mathcal{C}\beta_k+\widetilde{\mathcal{C}}\tilde{\beta}_k)$. Thus, the right-hand side of (2.7) can be expressed as a convex combination of forward Euler steps, yielding

$$\boldsymbol{u}_n = \sum_{j=0}^{k} \hat{\alpha}_j^* \Big( \boldsymbol{u}_{n-k+j} + \Delta t \frac{\beta_j^*}{\hat{\alpha}_j^*} \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) \Big) + \sum_{j=0}^{k} \tilde{\alpha}_j^* \Big( \boldsymbol{u}_{n-k+j} - \Delta t \frac{\tilde{\beta}_j^*}{\tilde{\alpha}_j^*} \widetilde{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \Big).$$

Taking norms and using convexity we have

$$\begin{aligned} \|\boldsymbol{u}_n\| \le & \sum_{j=0}^{k} \hat{\alpha}_j^* \Big\| \boldsymbol{u}_{n-k+j} + \Delta t \frac{\beta_j^*}{\hat{\alpha}_j^*} \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) \Big\| \\ & + \sum_{j=0}^{k} \tilde{\alpha}_j^* \Big\| \boldsymbol{u}_{n-k+j} - \Delta t \frac{\tilde{\beta}_j^*}{\tilde{\alpha}_j^*} \widetilde{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \Big\|. \end{aligned}$$

Under the step-size restriction $\Delta t \le \min\{\mathcal{C}\,\Delta t_{\mathrm{FE}}, \widetilde{\mathcal{C}}\,\widetilde{\Delta t}_{\mathrm{FE}}\}$ we get

$$\Delta t \frac{\beta_j^*}{\hat{\alpha}_j^*} \le \Delta t_{\mathrm{FE}} \quad \text{and} \quad \Delta t \frac{\tilde{\beta}_j^*}{\tilde{\alpha}_j^*} \le \widetilde{\Delta t}_{\mathrm{FE}},$$

for each $j \in \{0, \dots, k\}$. Since $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ satisfy (1.10a) and (1.10b), respectively, we have

$$(1-\alpha_k)\|\boldsymbol{u}_n\| \le \sum_{j=0}^{k-1} \hat{\alpha}_j^* \|\boldsymbol{u}_{n-k+j}\| + \sum_{j=0}^{k-1} \tilde{\alpha}_j^* \|\boldsymbol{u}_{n-k+j}\|,$$

and hence

$$\|\boldsymbol{u}_n\| \le \sum_{j=0}^{k-1} \alpha_j \|\boldsymbol{u}_{n-k+j}\| \le \max_{0\le j\le k-1} \|\boldsymbol{u}_{n-k+j}\| \sum_{j=0}^{k-1} \alpha_j.$$

Recall that $\sum_{j=0}^{k-1} \alpha_j = 1$ and therefore the monotonicity condition (1.5) follows. □

2.1. **Optimal SSP perturbed linear multistep methods.** We now turn to the problem of finding, among methods with a given number of steps $k$ and order of accuracy $p$, the largest SSP coefficients. Since $\mathcal{C}, \widetilde{\mathcal{C}}$ are continuous functions of the method's coefficients, we expect that the maximal step size (2.6) is achieved when $\mathcal{C} = \widetilde{\mathcal{C}}\,\widetilde{\Delta t}_{\mathrm{FE}}/\Delta t_{\mathrm{FE}}$. It is thus convenient to define $\xi \coloneqq \Delta t_{\mathrm{FE}}/\widetilde{\Delta t}_{\mathrm{FE}}$.

**Definition 2.5.** For a fixed $\xi \in [0, \infty)$ we say that a perturbed LMM (1.8) has *SSP coefficient*

$$\mathcal{C}(\xi) \coloneqq \sup\{r > 0 \mid \text{monotonicity conditions (2.4) hold with } \tilde{r} = \xi r\}$$

and its corresponding downwind SSP coefficient is $\widetilde{\mathcal{C}}(\xi) = \xi\,\mathcal{C}(\xi)$.

Whenever the set in Definition 2.5 is empty, then the method is non-SSP; in such cases we say the method has SSP coefficient equal to zero. In the next definition, we refer to a perturbed method (1.8) by its coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})$, and we write $\mathcal{P}_{k,p}$ to denote the set of all methods having at most $k$ steps and satisfying the order conditions up to (at least) order $p$.

**Definition 2.6.** An explicit (implicit) method in $\mathcal{P}_{k,p}$ is called *optimal* if no explicit (implicit) method in $\mathcal{P}_{k,p}$ has larger SSP coefficient. Given $\xi \in [0,\infty)$, we denote the largest SSP coefficient for explicit (implicit) $k$-step methods (1.8) of order $p$ by

$$\mathcal{C}_{k,p}(\xi) \coloneqq \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})\in\mathcal{P}_{k,p}} \Big\{ \mathcal{C}(\xi) > 0 \;\Big|\; \mathcal{C}(\xi) \text{ is the SSP coefficient of an explicit (implicit) method (1.8) with coefficients } (\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) \Big\}.$$

In the trivial case that the set over which the supremum is taken above is empty, we write $\mathcal{C}_{k,p}(\xi) = 0$.

Note that, because the inequalities involved are nonstrict (see (2.4)), the supremum in the last definition is always attained by some method.

Next we prove that for a given SSP perturbed LMM with SSP coefficient $\mathcal{C}(\xi)$, we can construct another SSP method (1.8) with the property that for each $j \in \{0,\dots,k\}$, either $\beta_j$ or $\tilde{\beta}_j$ is zero. Example 2.1 is an application of this result.

**Lemma 2.7.** *Consider a $k$-step perturbed LMM* (1.8) *of order $p$ with SSP coefficient $\mathcal{C}(\xi)$ for a given $\xi \in [0,\infty)$. Then, we can construct a $k$-step SSP method* (1.8) *of order $p$ with SSP coefficient at least $\mathcal{C}(\xi)$ that satisfies $\beta_j\tilde{\beta}_j = 0$ for each $j \in \{0,\dots,k\}$. Moreover, both perturbed methods correspond to the same underlying method.*

*Proof.* Suppose there exists a $k$-step SSP method (1.8) of order $p$ with SSP coefficient $\mathcal{C}(\xi)$ for some $\xi \in [0,\infty)$, such that $\beta_j \geq \tilde{\beta}_j > 0$ for $j \in J_1 \subseteq \{0,1,\dots,k\}$ and $\tilde{\beta}_j > \beta_j > 0$ for $j \in J_2 \subseteq \{0,1,\dots,k\}$. Clearly $J_1 \cap J_2 = \emptyset$. Define

$$\beta_j^* \coloneqq \begin{cases} \beta_j - \tilde{\beta}_j, & \text{if } j \in J_1, \\ 0, & \text{if } j \notin J_1, \end{cases} \qquad \tilde{\beta}_j^* \coloneqq \begin{cases} 0, & \text{if } j \notin J_2, \\ \tilde{\beta}_j - \beta_j, & \text{if } j \in J_2. \end{cases}$$

Observe that conditions (2.4) with $r = \mathcal{C}(\xi)$, $\tilde{r} = \widetilde{\mathcal{C}}(\xi)$ and the order conditions (2.5) are satisfied when $\beta_j, \tilde{\beta}_j$ are replaced by $\beta_j^*, \tilde{\beta}_j^*$. Therefore, the method with coefficients $(\boldsymbol{\alpha},\boldsymbol{\beta}^*,\tilde{\boldsymbol{\beta}}^*)$ has SSP coefficient at least $\mathcal{C}(\xi)$ and satisfies $\beta_j^*\tilde{\beta}_j^* = 0$ for each $j \in \{0,\dots,k\}$. Finally, the definition of $\beta_j^*$ and $\tilde{\beta}_j^*$ leaves $\beta_j - \tilde{\beta}_j$ invariant, thus substituting $\widetilde{\boldsymbol{F}} = \boldsymbol{F}$ in method (1.8) with coefficients $(\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})$ or $(\boldsymbol{\alpha},\boldsymbol{\beta}^*,\tilde{\boldsymbol{\beta}}^*)$ yields the same underlying method. □

The next Corollary is an immediate consequence of Lemma 2.7.

**Corollary 2.8.** *Let $k$, $p$ and $\xi$ be given such that $\mathcal{C}_{k,p}(\xi) > 0$. Then there exists an optimal SSP perturbed LMM* (1.8) *with SSP coefficient $\mathcal{C}_{k,p}(\xi)$ that satisfies $\beta_j\tilde{\beta}_j = 0$ for each $j \in \{0,\dots,k\}$.*

Optimal explicit $k$-step SSP perturbed LMMs (1.8) of first order are simply the explicit Euler method. To see that, first define

$$\gamma_j \coloneqq \alpha_j - r\beta_j - \tilde{r}\tilde{\beta}_j, \quad \text{for} \quad j \in \{0,\dots,k-1\}. \tag{2.8}$$

Then, the conditions of order one in (2.5) become

$$\sum_{j=0}^{k-1} \gamma_j + r\beta_j + \tilde{r}\tilde{\beta}_j = 1, \tag{2.9}$$

$$\sum_{j=0}^{k-1} j(\gamma_j + r\beta_j + \tilde{r}\tilde{\beta}_j) + \sum_{j=0}^{k-1} (\beta_j - \tilde{\beta}_j) = k. \tag{2.10}$$

Multiplying (2.9) by $k$, subtracting (2.10) and rearranging terms yields

$$\sum_{j=0}^{k-1} (k-j)\gamma_j + (r(k-j)-1)\,\beta_j + (\tilde{r}(k-j)+1)\,\tilde{\beta}_j = 0. \tag{2.11}$$

Monotonicity conditions (2.4) require that all $\gamma_j, \beta_j, \tilde{\beta}_j$ are nonnegative; thus, the first and last term in (2.11) are also nonnegative. In order (2.11) to hold, the coefficients of $\beta_j$ must be nonpositive, and therefore $r \le 1/(k-j)$. The maximum value $r = 1$ is attained when $j = k-1$, and in such case the only nonzero coefficient in (2.11) is $\beta_{k-1}$. From (2.8) and (2.9) we get that the optimal explicit $k$-step, first-order method (1.8) has nonzero coefficients $\alpha_{k-1} = \beta_{k-1} = 1$ and SSP coefficient $\mathcal{C}_{k,1}(\xi) = 1$ for all $\xi \in [0, \infty)$ and $k \ge 1$.

Arbitrary large SSP coefficient $\mathcal{C}(\xi)$ can be obtained for implicit first-order SSP methods (1.8). This was shown in [16, 19, 22] for SSP LMMs without downwinding. An optimal implicit $k$-step perturbed LMM (1.8) of first-order with $\mathcal{C}_{k,1}(\xi) = \infty$ for all $\xi \in [0, \infty)$ and $k \ge 1$ has coefficients

$$\alpha_j \ge 0,\ \beta_j = \tilde{\beta}_j = 0,\ j \in \{0, \dots, k-1\}, \quad \beta_k \ge 0,\ \tilde{\beta}_k \ge 0$$
$$\text{with } \sum_{j=0}^{k-1} \alpha_j = 1, \text{ and } \beta_k - \tilde{\beta}_k = k - \sum_{j=0}^{k-1} j\alpha_j.$$

If we consider optimal methods that satisfy $\beta_j\tilde{\beta}_j = 0$ for each $j \in \{0, \dots, k\}$, then $\tilde{\beta}_k$ must be set to zero. Otherwise, if $\beta_k = 0$, then $\tilde{\beta}_k = \sum_{j=0}^{k-1} j\alpha_j - k \le -1$ which violates the nonnegativity of the method's coefficients.

Based on Corollary 2.8 we have the following upper bound for the SSP coefficient of any perturbed LMM (1.8) of order greater than one. This extends [11, Theorem 2.2].

**Theorem 2.9.** *Given $\xi \in [0, \infty)$, any zero-stable perturbed LMM* (1.8) *of order greater than one satisfies $\mathcal{C}(\xi) \le 2$.*

*Proof.* Consider a second-order optimal SSP perturbed LMM with SSP coefficient $\mathcal{C} = \mathcal{C}(\xi)$ and $\widetilde{\mathcal{C}} = \xi\,\mathcal{C}(\xi)$ for some $\xi \in [0, \infty)$. Then, from Corollary 2.8 there exists an optimal method with at least SSP coefficient $\mathcal{C}$ and coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})$ such that $\beta_j\tilde{\beta}_j = 0$ for each $j \in \{0, \dots, k\}$.

Suppose $\xi > 0$ and define $\delta_j \coloneqq \beta_j + \xi\tilde{\beta}_j$ and

$$\sigma_j \coloneqq \begin{cases} 1, & \text{if } \tilde{\beta}_j = 0, \\ -1/\xi, & \text{if } \beta_j = 0, \end{cases}$$

where $j \in \{0, \dots, k\}$. Note that the nonnegativity of $\beta_j, \tilde{\beta}_j$ implies that there is at least one index $j$ such that $\delta_j > 0$ (otherwise, $\beta_j = \tilde{\beta}_j = 0$, for each $j \in \{0, \dots, k\}$,

and the method is not zero-stable[2]). Also, since either $\beta_j$ or $\tilde{\beta}_j$ is zero, then $\beta_j - \tilde{\beta}_j = \sigma_j \delta_j$ for each $j$. Let $\gamma_j = \alpha_j - \mathcal{C}\delta_j$ for $j \in \{0, \ldots, k-1\}$. Taking $p = 2$, $r = \mathcal{C}$, and $\tilde{r} = \widetilde{\mathcal{C}}$ in (2.5), the second-order conditions can be written as

$$\sum_{j=0}^{k-1} \gamma_j + \mathcal{C}\delta_j = 1, \tag{2.12}$$

$$\sum_{j=0}^{k-1} j\gamma_j + (j\mathcal{C} + \sigma_j)\delta_j = k - \sigma_k \delta_k, \tag{2.13}$$

$$\sum_{j=0}^{k-1} j^2\gamma_j + (j^2\mathcal{C} + 2j\sigma_j)\delta_j = k(k - 2\sigma_k \delta_k). \tag{2.14}$$

Multiplying (2.12), (2.13), and (2.14) by $-k^2$, $2k$ and $-1$, respectively, and adding all three expressions gives

$$\sum_{j=0}^{k-1} -(k-j)^2\gamma_j + \left(-\mathcal{C}(k-j)^2 + 2\sigma_j(k-j)\right)\delta_j = 0. \tag{2.15}$$

Since the method satisfies conditions (2.4) for $r = \mathcal{C}$ and $\tilde{r} = \widetilde{\mathcal{C}}$, then all coefficients $\gamma_j$ and $\delta_j$ are nonnegative. Therefore, there must be at least one index $j_0$ such that the coefficient of $\delta_{j_0}$ in (2.15) is nonnegative. Note that if $\beta_{j_0} = 0$, then $\sigma_{j_0} < 0$; hence it can only be that $\tilde{\beta}_{j_0} = 0$ and $\beta_{j_0} > 0$, yielding $\delta_{j_0} > 0$. Thus,

$$-\mathcal{C}(k - j_0)^2 + 2(k - j_0) \geq 0,$$

which implies

$$\mathcal{C} \leq \frac{2}{k - j_0} \leq 2, \tag{2.16}$$

since $k - j_0 \geq 1$.

If now $\xi = 0$, define $\delta_j \coloneqq \beta_j + \tilde{\beta}_j$ and $\sigma_j \coloneqq \operatorname{sign}(\beta_j - \tilde{\beta}_j)$ for each $j \in \{0, \ldots, k\}$. Using $\gamma_j = \alpha_j - \mathcal{C}\beta_j$ and performing the same algebraic manipulations as before we get

$$\sum_{j=0}^{k-1} -(k-j)^2\gamma_j - \mathcal{C}(k-j)^2\beta_j + 2(k-j)\sigma_j\delta_j = 0. \tag{2.17}$$

In the sum of (2.17), consider the two sets of indexes $J^0 = \{j \mid \beta_j = 0\}$ and $J^+ = \{j \mid \beta_j > 0\}$. If $\beta_j = 0$, then we have $\sigma_j\delta_j = -\tilde{\beta}_j \leq 0$ and therefore $-\mathcal{C}(k-j)^2\beta_j + 2\sigma_j(k-j)\delta_j \leq 0$; these terms are nonpositive for all indexes $j \in J^0$ and are considered together with terms $-(k-j)^2\gamma_j \leq 0$. As the sum is equal to zero, then the set $J^+$ must be nonempty. Hence, there is an index $j_0 \in J^+$ such that $\beta_{j_0} > 0$ and

$$-\mathcal{C}(k - j_0)^2\beta_{j_0} + 2(k - j_0)\sigma_{j_0}\delta_{j_0} \geq 0. \tag{2.18}$$

Since $\beta_{j_0} > 0$, then $\tilde{\beta}_{j_0} = 0$ and $\sigma_{j_0}\delta_{j_0} = \beta_{j_0} > 0$. We divide by $\beta_{j_0}$ in (2.18), reorganize terms and inequality (2.16) is obtained. □

[2]Let the first characteristic polynomial of (1.8) be $\rho(\zeta) = \zeta^k - \sum_{j=0}^{k-1} \alpha_j \zeta^j$. If $\beta_j = \tilde{\beta}_j = 0$, for each $j \in \{0, \ldots, k\}$, then a method of at least order one should satisfy $\rho(1) = \rho'(1) = 0$. Hence, in such case $\zeta = 1$ is a double root and the root condition for zero-stability is violated.

*Remark* 2.10. For given values $k, p, \xi$, it may be that there exists no method with positive SSP coefficients. However, from (2.4) and Theorem 2.9 if a method exists with bounded SSP coefficient, then the existence of an optimal method follows since the feasible region is compact.

The upper bound of the SSP coefficient can be only achieved in the case of implicit second-order SSP perturbed LMM (1.8). In view of the proof of Theorem 2.9, by taking $j_0 = k-1$ in (2.16) we have $\mathcal{C} = \mathcal{C}(\xi) = 2$. In that case $\beta_{k-1} > 0$ and we can choose all other terms in (2.15) (or (2.17) if $\xi = 0$) to be equal to zero. Then, from (2.12) and (2.13) we obtain that the nonzero coefficients are

$$\alpha_{k-1} = 1, \quad \beta_{k-1} = \beta_k = \frac{1}{2}.$$

Hence, an optimal $k$-step, second-order SSP perturbed LMM (1.8) with SSP coefficient $\mathcal{C}_{k,2}(\xi) = 2$ for all $\xi \in [0, \infty)$ and $k \geq 1$ is simply the trapezoidal rule. As we will see later, we can prove that the trapezoidal rule is in fact the *unique* optimal second-order SSP perturbed LMM . Therefore, downwinding does not result in any improvement to the SSP coefficient for the class of implicit second-order LMMs.

By combining conditions (2.4) and (2.5), and using (2.8), the problem of finding optimal SSP perturbed LMMs (1.8) can be formulated as a linear programming feasibility problem.

**LP 1.** For fixed $k \geq 1$, $p \geq 1$ and a given $\xi \in [0, \infty)$, determine whether there exist nonnegative coefficients $\gamma_j$, $j \in \{0, \ldots, k-1\}$ and $\beta_j, \tilde{\beta}_j$, $j \in \{0, \ldots, k\}$ such that

$$\begin{aligned} &\sum_{j=0}^{k-1} \gamma_j + r\beta_j + \tilde{r}\tilde{\beta}_j = 1, \qquad \sum_{j=0}^{k-1} j(\gamma_j + r\beta_j + \tilde{r}\tilde{\beta}_j) + \sum_{j=0}^{k} (\beta_j - \tilde{\beta}_j) = k, \\ &\sum_{j=0}^{k-1} (\gamma_j + r\beta_j + \tilde{r}\tilde{\beta}_j) j^i + \sum_{j=0}^{k} (\beta_j - \tilde{\beta}_j) i j^{i-1} = k^i, \quad i \in \{2, \ldots, p\}, \end{aligned} \tag{2.19}$$

for some value $r \geq 0$ and $\tilde{r} = \xi r$.

Expressing (2.19) in a compact form facilitates the analysis of the feasible problem LP 1. Let the vector

$$\boldsymbol{a}_j \coloneqq (1, j, j^2, \ldots, j^p)^\intercal \in \mathbb{R}^{p+1}, \tag{2.20}$$

and denote by $\boldsymbol{a}_j'$ the derivative of $\boldsymbol{a}_j$ with respect to $j$, namely

$$\boldsymbol{a}_j' = (0, 1, 2j, \ldots, pj^{p-1})^\intercal.$$

Define

$$\boldsymbol{b}_j^{\pm}(x) \coloneqq \begin{cases} \pm x \boldsymbol{a}_k', & \text{if } j = k, \\ \boldsymbol{a}_j \pm x \boldsymbol{a}_j', & \text{otherwise.} \end{cases} \tag{2.21}$$

The conditions (2.19) can be expressed in terms of vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_j^{\pm}(\cdot)$:

$$\sum_{j=0}^{k-1} \gamma_j \boldsymbol{a}_j + r \sum_{j=0}^{k} \beta_j \boldsymbol{b}_j^{+}(r^{-1}) + \tilde{r} \sum_{j=0}^{k} \tilde{\beta}_j \boldsymbol{b}_j^{-}(\tilde{r}^{-1}) = \boldsymbol{a}_k. \tag{2.22}$$

The number of nonzero coefficients of an optimal SSP perturbed LMM is given by Theorem 2.13. The proof of Theorem 2.13 relies on the following two lemmata.

**Lemma 2.11.** *Consider a matrix*

$$A(r) = \left[\begin{array}{c|c|c} & & \\ \boldsymbol{\psi}_1(r) & \dots & \boldsymbol{\psi}_m(r) \\ & & \end{array}\right] \in \mathbb{R}^{n\times m},$$

*where the columns $\boldsymbol{\psi}_j(r) \in \mathbb{R}^n$, $j \in \{1,\dots,m\}$ are functions of a variable $r \in \mathbb{R}$ and $m \geq n$. Let $\boldsymbol{c} \in \mathbb{R}^n$ and $r^* \in \mathbb{R}$ be given. Let each $\boldsymbol{\psi}_j(r)$ be a continuous function of $r$ for values in some neighborhood of $r^*$, and let the system $A(r^*)\boldsymbol{x} = \boldsymbol{c}$ have a nonnegative solution with at least $n$ strictly positive elements*

$$x_{i_1}, x_{i_2}, \dots, x_{i_s}, \quad n \leq s \leq m.$$

*If the set $\{\boldsymbol{\psi}_{i_1}(r^*), \dots, \boldsymbol{\psi}_{i_s}(r^*)\}$ spans $\mathbb{R}^n$, then there exists $\epsilon > 0$ such that the system $A(r^* + \epsilon)\boldsymbol{x} = \boldsymbol{c}$ also has a nonnegative solution.*

*Proof.* Let $\boldsymbol{c} \in \mathbb{R}^n$ and $r^* \in \mathbb{R}$ be given, and assume that $A(r^*)\boldsymbol{x} = \boldsymbol{c}$ has a solution $\boldsymbol{x} \geq 0$ with at least $n$ strictly positive elements

$$x_{i_1}, x_{i_2}, \dots, x_{i_s}, \quad n \leq s \leq m.$$

Let $S(r) = \{\boldsymbol{\psi}_{i_1}(r), \dots, \boldsymbol{\psi}_{i_s}(r)\}$ be a subset of the columns of $A(r) \in \mathbb{R}^{n\times m}$, and assume $S(r^*)$ spans $\mathbb{R}^n$. Then, there exists a subset of the columns in $S(r^*)$ that forms a basis for $\mathbb{R}^n$, so we can permute the columns of $A(r)$ in such a way that the first $n$ columns are in $S(r)$ and for $r = r^*$ they are linearly independent. This yields $A_p(r) = [B(r) \mid N(r)]$, where $B(r^*) \in \mathbb{R}^{n\times n}$ has full rank and $N(r) \in \mathbb{R}^{n\times(m-n)}$. We can permute the entries of $\boldsymbol{x}$ in the same way, yielding $\boldsymbol{x}_p = (\boldsymbol{x}_B, \boldsymbol{x}_N)$, where $\boldsymbol{x}_B \in \mathbb{R}^n$ is a strictly positive vector, $\boldsymbol{x}_N \in \mathbb{R}^{m-n}$ is nonnegative, and $A_p(r^*)\boldsymbol{x}_p = \boldsymbol{c}$. A simple calculation gives $\boldsymbol{x}_B = B^{-1}(r^*)\big(\boldsymbol{c} - N(r^*)\boldsymbol{x}_N\big)$. With this motivation we define $\hat{\boldsymbol{x}}_B(r) \coloneqq B^{-1}(r)\big(\boldsymbol{c} - N(r)\boldsymbol{x}_N\big)$ and $\hat{\boldsymbol{x}}_p(r) \coloneqq (\hat{\boldsymbol{x}}_B(r), \boldsymbol{x}_N)$. Direct calculation shows that $A_p(r)\hat{\boldsymbol{x}}_p(r) = \boldsymbol{c}$ as long as $\hat{\boldsymbol{x}}_p(r)$ is well-defined. By continuity, $B^{-1}(r^*+\epsilon)$ exists for small enough $\epsilon > 0$, so $\hat{\boldsymbol{x}}_B(r)$ is well-defined in some neighborhood of $r^*$. Moreover, $\hat{\boldsymbol{x}}_B(r)$ is (in some neighborhood of $r^*$) a continuous function of $r$, so continuity implies that $\hat{\boldsymbol{x}}_B(r^* + \epsilon) > 0$, for $\epsilon$ sufficiently small. Therefore, $\hat{\boldsymbol{x}}_p(r^* + \epsilon)$ is nonnegative and

$$\begin{aligned} A_p(r^* + \epsilon)\hat{\boldsymbol{x}}_p(r^* + \epsilon) &= \left(B(r^* + \epsilon) \;\middle|\; N(r^* + \epsilon)\right)\begin{pmatrix} \hat{\boldsymbol{x}}_B(r^* + \epsilon) \\ \boldsymbol{x}_N \end{pmatrix} \\ &= B(r^* + \epsilon)\big(B^{-1}(r^* + \epsilon)\big(\boldsymbol{c} - N(r^* + \epsilon)\boldsymbol{x}_N\big)\big) + N(r^* + \epsilon)\boldsymbol{x}_N \\ &= \boldsymbol{c}. \end{aligned}$$

□

The next lemma is a consequence of Carathéodory's theorem [17, Theorem 17.1], which states that a vector $\boldsymbol{x}$ belongs to the convex hull of a set $S \subseteq \mathbb{R}^n$, if and only if $\boldsymbol{x}$ can be expressed as a convex combination of $n + 1$ vectors in $S$. The proof of Lemma 2.12 appears in Appendix A.

**Lemma 2.12.** *Consider a set $S = \{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_m\}$ of distinct vectors $\boldsymbol{\psi}_j \in \mathbb{R}^n$, $j \in \{1,\dots,m\}$. Let $C = conv(S)$ be the convex hull of $S$. Then the following statements hold:*

(a) *Any nonzero vector in $C$ can be expressed as a nonnegative linear combination of linearly independent vectors in $S$.*
(b) *Suppose the vectors in $S$ lie in the hyperplane $\{(1, \boldsymbol{v}) \mid \boldsymbol{v} \in \mathbb{R}^{n-1}\}$ of $\mathbb{R}^n$. Then any nonzero vector in $C$ can be expressed as a convex combination of linearly independent vectors in $S$.*

We can now characterize the coefficients of an optimal perturbed LMM as follows.

**Theorem 2.13.** *Let $k, p$ be positive integers such that $0 < \mathcal{C}_{k,p}(\xi) < \infty$ for a given $\xi \in [0, \infty)$. Then there exists an optimal perturbed LMM* (1.8) *with SSP coefficient $\mathcal{C}_{k,p}(\xi)$ that has at most $p$ nonzero coefficients in the set*

$$\{\gamma_0, \dots, \gamma_{k-1}, \beta_0, \dots, \beta_k, \tilde{\beta}_1, \dots, \tilde{\beta}_k\}.$$

*Proof.* Let $0 < \mathcal{C}_{k,p}(\xi) < \infty$ for given $k$, $p$ and $\xi$. Since the inequalities involved in (2.4) are not strict, then the supremum in Definition 2.6 is always attained by some method. Consider an optimal LMM (1.8) with coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})$ and SSP coefficient $0 < \mathcal{C}_{k,p}(\xi) < \infty$, for a given $\xi \in [0, \infty)$. From Corollary 2.8 an optimal method can be chosen such that $\beta_j \tilde{\beta}_j = 0$ for each $j \in \{0, \dots, k\}$. Using (2.8) we can perform a change of variables and consider the vector of coefficients $\boldsymbol{x} = \big(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}\big) \in \mathbb{R}^{3k+2}$, where $\boldsymbol{x} \geq 0$.

We will show that $\boldsymbol{x}$ has at most $p$ nonzero coefficients. Suppose to the contrary that $\boldsymbol{x}$ has at least $p+1$ nonzero coefficients

$$\gamma_{i_1}, \dots, \gamma_{i_m}, \beta_{j_1}, \dots, \beta_{j_n}, \tilde{\beta}_{l_1}, \dots, \tilde{\beta}_{l_s},$$

where $0 \leq i_1 < \cdots < i_m \leq k-1$, $0 \leq j_1 < \cdots < j_n \leq k$ and $0 \leq l_1 < \cdots < l_s \leq k$, such that $m + n + s \geq p + 1$. Let $\tilde{r} = \xi r$; then the system (2.22) can be written as

$$A(r)\boldsymbol{x} = \boldsymbol{a}_k, \tag{2.23}$$

where

$$A(r) = \left[\begin{array}{c|c|c|c|c|c|c|c|c} \boldsymbol{a}_0 & \dots & \boldsymbol{a}_{k-1} & r\boldsymbol{b}_0^+\left(\frac{1}{r}\right) & \dots & r\boldsymbol{b}_k^+\left(\frac{1}{r}\right) & \xi r\boldsymbol{b}_0^-\left(\frac{1}{\xi r}\right) & \dots & \xi r\boldsymbol{b}_k^-\left(\frac{1}{\xi r}\right) \end{array}\right],$$

and the solution $\boldsymbol{x} = \boldsymbol{x}(r)$ depends on $r$. In particular, the coefficients of an optimal method solve (2.23) with $r = \mathcal{C}_{k,p}(\xi)$, and $\boldsymbol{x}\big(\mathcal{C}_{k,p}(\xi)\big) \geq 0$. Define the set

$$S(r) = \left\{\boldsymbol{a}_{i_1}, \dots, \boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^+\left(\frac{1}{r}\right), \dots, \boldsymbol{b}_{j_n}^+\left(\frac{1}{r}\right), \boldsymbol{b}_{l_1}^-\left(\frac{1}{\xi r}\right), \dots, \boldsymbol{b}_{l_s}^-\left(\frac{1}{\xi r}\right)\right\}.$$

*Case* 1. Assume $S(r)$ spans $\mathbb{R}^{p+1}$ for $r = \mathcal{C}_{k,p}(\xi)$. By our assumption, $\boldsymbol{x}$ has at least $p+1$ nonzero elements; thus by using Lemma 2.11 there exists $\epsilon > 0$ such that the system (2.23) has a nonnegative solution $\boldsymbol{x}^*$ for $r = \mathcal{C}_{k,p}(\xi) + \epsilon$. This contradicts the optimality of the method, since we can construct a $k$-step SSP perturbed LMM of order $p$ with coefficients given by $\boldsymbol{x}^*$ and SSP coefficient $\mathcal{C}_{k,p}(\xi) + \epsilon$.

*Case* 2. Now, assume that the set $S(r)$ does not span $\mathbb{R}^{p+1}$ for $r = \mathcal{C}_{k,p}(\xi)$; therefore, the largest size of a linearly independent subset of $S\big(\mathcal{C}_{k,p}(\xi)\big)$ is $p$.

*Case* 2a. If the method is explicit, then $\beta_k = \tilde{\beta}_k = 0$, and from (2.20) and (2.21) the vectors in set $S\big(\mathcal{C}_{k,p}(\xi)\big)$ lie in the hyperplane $\{(1, v) \mid v \in \mathbb{R}^p\} \subset \mathbb{R}^{p+1}$, since none of these vectors have index equal to $k$. Moreover, from the first condition of (2.19) and equation (2.22) the vector $\boldsymbol{a}_k$ lies in the convex hull of $S(r)$ for $r = \mathcal{C}_{k,p}(\xi)$. Therefore, from part (b) of Lemma 2.12, vector $\boldsymbol{a}_k$ can be expressed as a convex combination of linearly independent vectors in $S\big(\mathcal{C}_{k,p}(\xi)\big)$. Such a set can have no more than $p$ elements, so it must be expressible as a convex combination of at most $p$ vectors in $S\big(\mathcal{C}_{k,p}(\xi)\big)$.

*Case* 2b. If the method is implicit, assume without loss of generality that $\beta_k > 0$ and $\tilde{\beta}_k = 0$. Again, by using the first condition of (2.19) we have that the sum of the coefficients in (2.22) divided by $\boldsymbol{a}_k(1 + r\beta_k)$ sum to unity. Therefore,

the vector $\boldsymbol{a}_k/(1+r\beta_k)$ belongs to the convex hull of $S(r)$, for $r = \mathcal{C}_{k,p}(\xi)$, and thus from part (a) of Lemma 2.12 it can be written as a nonnegative linear combination linearly independent vectors in $S\big(\mathcal{C}_{k,p}(\xi)\big)$. Again, such a set can have no more than $p$ elements. □

Furthermore, uniqueness of optimal perturbed LMMs can be established under certain conditions on the vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_j^{\pm}$. The following lemma is a generalization of [12, Lemma 3.5].

**Lemma 2.14.** *Consider an optimal perturbed LMM* (1.8) *with SSP coefficient* $\mathcal{C} = \mathcal{C}_{k,p}(\xi) > 0$ *and* $\widetilde{\mathcal{C}} = \xi\,\mathcal{C}_{k,p}(\xi)$ *for a given* $\xi \in [0,\infty)$. *Let the indexes*

$$0 \le i_1 < \cdots < i_m \le k-1, \quad 0 \le j_1 < \cdots < j_n \le k, \quad 0 \le l_1 < \cdots < l_s \le k,$$

*where* $m+n+s \le p$ *be such that* $\gamma_{i_1},\dots,\gamma_{i_m},\beta_{j_1}\dots,\beta_{j_n},\tilde{\beta}_{l_1},\dots,\tilde{\beta}_{l_s}$ *are the positive coefficients in* (1.8). *Let us also denote the sets* $I = \{0,\dots,k\}$, $J_1 = \{i_1,\dots,i_m\}$, $J_2 = \{j_1,\dots,j_n\}$, $J_3 = \{l_1,\dots,l_s\}$. *Assume that the function*

$$F(\boldsymbol{v}) = \operatorname{DET}\left(\boldsymbol{v}, \boldsymbol{a}_{i_1},\dots,\boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^{+}\Big(\frac{1}{\mathcal{C}}\Big),\dots,\boldsymbol{b}_{j_n}^{+}\Big(\frac{1}{\mathcal{C}}\Big), \boldsymbol{b}_{l_1}^{-}\Big(\frac{1}{\widetilde{\mathcal{C}}}\Big),\dots,\boldsymbol{b}_{l_s}^{-}\Big(\frac{1}{\widetilde{\mathcal{C}}}\Big)\right)$$

*is either strictly positive or strictly negative, simultaneously for all* $\boldsymbol{v} = \boldsymbol{a}_i$, $i \in I \setminus (J_1 \cup \{k\})$, $\boldsymbol{v} = \boldsymbol{b}_j^{+}(1/\mathcal{C})$, $j \in I \setminus J_2$ *and* $\boldsymbol{v} = \boldsymbol{b}_l^{-}\big(1/\widetilde{\mathcal{C}}\big)$, $l \in I \setminus J_3$. *Then* (1.8) *is the unique optimal* $k$*-step SSP perturbed LMM of order* $p$.

*Proof.* Assume there exists another optimal $k$-step method (1.8) of order least $p$ with coefficients $(\boldsymbol{\alpha^*},\boldsymbol{\beta^*},\tilde{\boldsymbol{\beta}}^*)$. Define $\gamma_i^* \coloneqq \alpha_i^* - \mathcal{C}\beta_i^* - \widetilde{\mathcal{C}}\tilde{\beta}_i^*$, $i \in \{0,\dots,k-1\}$, then by the monotonicity conditions (2.4) and order conditions (2.5) we have

$$\begin{aligned} \gamma_i^* &\ge 0, && i \in \{0,\dots,k-1\},\\ \beta_j^* \ge 0, \tilde{\beta}_l^* &\ge 0, && j \in \{0,\dots,k\},\ l \in \{0,\dots,k\},\\ \sum_{i=0}^{k-1}\gamma_i^*\boldsymbol{a}_i + \mathcal{C}\sum_{j=0}^{k}\beta_j^*\boldsymbol{b}_j^{+}\Big(\frac{1}{\mathcal{C}}\Big) &+ \widetilde{\mathcal{C}}\sum_{l=0}^{k}\tilde{\beta}_l^*\boldsymbol{b}_l^{-}\Big(\frac{1}{\widetilde{\mathcal{C}}}\Big) = \boldsymbol{a}_k. \end{aligned}$$

Since the method (1.8) with coefficients $(\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})$ is optimal, then $\boldsymbol{a}_k$ can be also written as a nonnegative linear combination of vectors in the set

$$\left\{\boldsymbol{a}_{i_1},\dots,\boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^{+}\Big(\frac{1}{\mathcal{C}}\Big)\dots,\boldsymbol{b}_{j_n}^{+}\Big(\frac{1}{\mathcal{C}}\Big), \boldsymbol{b}_{l_1}^{-}\Big(\frac{1}{\widetilde{\mathcal{C}}}\Big),\dots,\boldsymbol{b}_{l_s}^{-}\Big(\frac{1}{\widetilde{\mathcal{C}}}\Big)\right\}. \tag{2.24}$$

Consider the first condition in (2.19) and (2.20). In the case the optimal method is explicit, then the vector $\boldsymbol{a}_k$ belongs in the convex hull of the set (2.24). If the optimal method is implicit, assume without loss of generality that $\beta_k > 0$ and $\tilde{\beta}_k = 0$. Then, the vector $\boldsymbol{a}_k/(1+\mathcal{C}\beta_k)$ belongs in the convex hull of the set (2.24).

Therefore, from Lemma 2.12 the vectors in (2.24) are linearly independent. Hence,

$$\begin{aligned}
0 =& \operatorname{DET}\left(\boldsymbol{a}_k, \boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^+\left(\tfrac{1}{\mathcal{C}}\right), \ldots, \boldsymbol{b}_{j_n}^+\left(\tfrac{1}{\mathcal{C}}\right), \boldsymbol{b}_{l_1}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right), \ldots, \boldsymbol{b}_{l_s}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right)\right) \\
=& \sum_{i=0}^{k-1} \gamma_i^* \operatorname{DET}\left(\boldsymbol{a}_i, \boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^+\left(\tfrac{1}{\mathcal{C}}\right), \ldots, \boldsymbol{b}_{j_n}^+\left(\tfrac{1}{\mathcal{C}}\right), \boldsymbol{b}_{l_1}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right), \ldots, \boldsymbol{b}_{l_s}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right)\right) \\
&+ \mathcal{C} \sum_{j=0}^{k} \beta_j^* \operatorname{DET}\left(\boldsymbol{b}_j^+\left(\tfrac{1}{\mathcal{C}}\right), \boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^+\left(\tfrac{1}{\mathcal{C}}\right), \ldots, \boldsymbol{b}_{j_n}^+\left(\tfrac{1}{\mathcal{C}}\right), \boldsymbol{b}_{l_1}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right), \ldots, \boldsymbol{b}_{l_s}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right)\right) \\
&+ \widetilde{\mathcal{C}} \sum_{l=0}^{k} \tilde{\beta}_l^* \operatorname{DET}\left(\boldsymbol{b}_l^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right), \boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}, \boldsymbol{b}_{j_1}^+\left(\tfrac{1}{\mathcal{C}}\right), \ldots, \boldsymbol{b}_{j_n}^+\left(\tfrac{1}{\mathcal{C}}\right), \boldsymbol{b}_{l_1}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right), \ldots, \boldsymbol{b}_{l_s}^-\left(\tfrac{1}{\widetilde{\mathcal{C}}}\right)\right).
\end{aligned}$$

By positivity of coefficients $\gamma_i^*, \beta_j^*, \tilde{\beta}_j^*$ and the assumptions of the lemma, we have $\gamma_i^* = 0$, $i \notin J_1$, $\beta_j^* = 0$, $j \notin J_2$ and $\tilde{\beta}_l^* = 0$, $l \notin J_3$. Linear independence of the vectors in (2.24) implies that $\gamma_i^* = \gamma_i$, $i \in J_1$ and $\beta_j^* = \beta_j$, $j \in J_2$ and $\tilde{\beta}_l^* = \tilde{\beta}_l$, $l \in J_3$ and the statement of the lemma is proved. □

Lemma 2.14 can be applied to prove that the trapezoidal rule is the unique optimal perturbed LMM (1.8) of second-order with $\mathcal{C}_{k,2} = 2$ for $k \geq 1$. The trapezodium rule has coefficients $\beta_{k-1} = \beta_k = 1/2$, whereas all other $\gamma_j, \beta_j, \tilde{\beta}_j$ are equal to zero. Therefore, according to Lemma 2.14 it is sufficient to check the sign of

$$F(\boldsymbol{v}) = \operatorname{DET}\left(\boldsymbol{v}, \boldsymbol{b}_{k-1}^+\left(\tfrac{1}{2}\right), \boldsymbol{b}_k^+\left(\tfrac{1}{2}\right)\right) \tag{2.25}$$

for all $\boldsymbol{v} = \boldsymbol{a}_i$, $i \in \{0, \ldots, k-1\}$, $\boldsymbol{v} = \boldsymbol{b}_j^+(1/2)$, $j \in \{0, \ldots, k-2\}$ and $\boldsymbol{v} = \boldsymbol{b}_l^-(1/(2\xi))$, $l \in \{0, \ldots, k\}$ for all $\xi \in [0, \infty)$. After some simple calculations we can show that all determinants (2.25) are strictly positive.

Fixing the number of steps $k$, and the order of accuracy $p$, the feasibility problem LP 1 has been numerically solved for different values of $\xi$ by using `linprog` from `MATLAB`'s optimization toolbox. Optimal explicit and implicit perturbed LMMs (not shown here) were found for $k \in \{1, \ldots, 50\}$ and $p \in \{1, \ldots, 15\}$. For $\xi = 1$, the SSP coefficients of optimal methods coincide with those in [10, Section 3]. All code to generate the coefficients of SSP perturbed LMMs and SSP IMEX methods discussed in Section 3 is available at `https://github.com/numerical-mathematics/ssp-almm_RR`.

*Remark* 2.15. In all cases we have investigated, the SSP coefficient $\mathcal{C}(\xi)$ (see Definition 2.5) is a strictly decreasing function of $\xi$. Similarly, the corresponding SSP coefficient $\widetilde{\mathcal{C}}(\xi)$ is strictly increasing. Assume that $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ satisfy (1.10), and fix the number of stages and order of accuracy. The monotone behavior of $\mathcal{C}(\xi)$ and $\widetilde{\mathcal{C}}(\xi)$ suggests that the optimal perturbed LMM obtained by considering the different step sizes in (1.10) allows *larger* step sizes for monotonicity than what is allowed by the optimal downwind SSP method, obtained just by taking the minimum of the two forward Euler step sizes. If we use an optimal downwind LMM, then a sufficient condition for monotonicity is $\Delta t \leq \widetilde{\mathcal{C}}_{\text{LMM}} \Delta t_{\text{FE}}$ if $\xi < 1$, or $\Delta t \leq \widetilde{\mathcal{C}}_{\text{LMM}} \widetilde{\Delta t}_{\text{FE}}$ if $\xi > 1$. On the other hand, using a perturbed LMM a sufficient step-size restriction is given by $\Delta t \leq \mathcal{C}(\xi) \Delta t_{\text{FE}} = \widetilde{\mathcal{C}}(\xi) \widetilde{\Delta t}_{\text{FE}}$. Since $\widetilde{\mathcal{C}}_{\text{LMM}} = \mathcal{C}(1) < \mathcal{C}(\xi)$ if $\xi < 1$ and $\widetilde{\mathcal{C}}_{\text{LMM}} < \widetilde{\mathcal{C}}(\xi)$ if $\xi > 1$, then the perturbed LMM allows larger step sizes. This

behavior is shown in Figure 1 for the class of explicit two-step, second-order SSP perturbed LMMs.

*Remark* 2.16. The dependence of the SSP coefficient $\mathcal{C}(\xi)$ with respect to $\xi$ can be explained in view of inequalities (2.4) and forward Euler conditions (1.10). As the step-size restriction in (1.10a) becomes more severe, then $\xi = \Delta t_{\mathrm{FE}}/\widetilde{\Delta t}_{\mathrm{FE}}$ approaches zero. However, since $\tilde{r} = \xi r$, inequalities (2.4) depend less on coefficients $\tilde{\beta}_j$ enabling larger SSP coefficients to be obtained. On the other hand, as the step-size restriction of forward Euler condition (1.10b) is stricter, then $\xi$ tends to infinity and coefficients $\tilde{\beta}_j$ must approach zero in order (2.4) to hold. In other words, the best possible SSP method in this case would be a method without downwind and thus the SSP coefficient $\mathcal{C}(\xi)$ approaches the corresponding SSP coefficient of traditional LMMs (1.4). Finally, for a fixed order of accuracy the SSP coefficient $\mathcal{C}(\xi)$ tends to $\widetilde{\mathcal{C}}_{\mathrm{LMM}}$ as we increase the number of steps.

2.2. **Examples.** Here we illustrate the effectiveness of perturbed LMMs by presenting two examples. We consider the following assumptions:

(A1) Condition (1.3) holds only for operator $\boldsymbol{F}$.
(A2) Conditions (1.10) hold for $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ under a step-size restriction $\Delta t \le \min\{\Delta t_{\mathrm{FE}}, \widetilde{\Delta t}_{\mathrm{FE}}\}$.
(A3) Conditions (1.10) hold for $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ under different step-size restrictions.

In the literature, traditional SSP LMMs applied to problems satisfying assumption (A1) have been extensively studied; for example, see [9, 12, 13]. Downwind SSP LMMs [10, 11, 18, 20, 21] were introduced for problems that comply with assumption (A2), whereas methods for problems satisfying assumption (A3) are the topic of this work.

**Example 2.17.** Consider the ODE problem

$$\begin{aligned} u'(t) &= u(t)^2\big(u(t)-1\big), \quad t \ge 0, \\ u(t_0) &= u_0. \end{aligned} \tag{2.26}$$

The right-hand side is Lipschitz continuous in $u$ in a close interval containing $[0,1]$. Thus, there exists a unique solution and it is easy to see that existence holds for all $t$. Therefore, if $u(t_0) = 0$ or $u(t_0) = 1$, then $u(t) = 0$ or $u(t) = 1$, respectively, for all $t$. If $u_0 \in [0,1]$, uniqueness implies that $u(t) \in [0,1]$ for all $t$. It can also be shown that if $u \in [0,1]$, then

$$\begin{aligned} &0 \le u + \Delta t\, u^2(u-1) \le 1 \quad \text{for } 0 \le \Delta t \le 4, \\ &0 \le u - \Delta t\, u^2(u-1) \le 1 \quad \text{for } 0 \le \Delta t \le 1. \end{aligned}$$

Applying method (1.8) where $\boldsymbol{F} = u^2(u-1)$, it is natural to take $\widetilde{\boldsymbol{F}} = \boldsymbol{F}$, and then we have that (1.10) holds with $\Delta t_{\mathrm{FE}} = 4$ and $\widetilde{\Delta t}_{\mathrm{FE}} = 1$. For method (2.1), in practice we observe that $u_n \in [0,1]$ whenever $\Delta t \le 8/7$. The method has $\mathcal{C}_{\mathrm{LMM}} = 0$, so applying only assumption (A1) above we cannot expect a monotone solution under any step size. Using assumption (A2), and writing the method in the form (2.3) (notice that perturbations do not change the method at all in this case, since $\widetilde{\boldsymbol{F}} = \boldsymbol{F}$) we obtain a step-size restriction $\Delta t \le \widetilde{\mathcal{C}}_{\mathrm{LMM}} \min\{\Delta t_{\mathrm{FE}}, \widetilde{\Delta t}_{\mathrm{FE}}\} = 2/7$, as $\widetilde{\mathcal{C}}_{\mathrm{LMM}} = 2/7$. Finally, using assumption (A3) to take into account the different

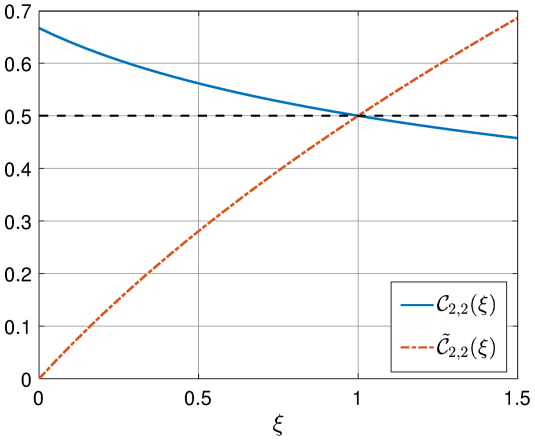



FIGURE 1. Functions $\mathcal{C}_{2,2}(\xi)$ and $\widetilde{\mathcal{C}}_{2,2}(\xi)$ for the class of explicit two-step, second-order perturbed LMMs. The dotted line shows $\widetilde{\mathcal{C}}_{\text{LMM}} = \mathcal{C}_{2,2}(1)$ for this particular class of methods.

forward Euler step sizes for $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$, we obtain the step-size restriction $\Delta t_{\max} = \widetilde{\mathcal{C}}_{\text{LMM}} \Delta t_{\text{FE}} = 8/7$, which matches the experimental observation.

An even larger step-size restriction can be achieved by finding the optimal perturbed LMM among the class of two-step, second-order perturbed LMMs. In this case $\xi = \Delta t_{\text{FE}}/\widetilde{\Delta} t_{\text{FE}} = 4$ and the optimal perturbed LMM has SSP coefficient $\mathcal{C}_{2,2}(4) = 0.3465$, thus the numerical solution is guaranteed to lie in the interval $[0, 1]$ if the step size is at most $\Delta t_{\max} = \mathcal{C}_{2,2}(4)\Delta t_{\text{FE}} = 1.386$. The coefficients of this method can be found in Appendix B.

For purely hyperbolic problems the spatial discretizations are usually chosen in such a way that $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ satisfy (1.10) under the same step-size restriction. However, in many other cases (e.g., advection-reaction problems) this is not the case, as shown in Example 2.19. First, we mention the following lemma which is an extension of [1, Proposition 5.4]; its proof can be found in Appendix A.

**Lemma 2.18.** *Consider the function*

$$f(u) = \sum_{i=1}^{n} f_i(u),$$

*and assume that there exist* $\epsilon_i > 0$ *such that* $||u + \tau f_i(u)|| \leq ||u||$ *for* $0 \leq \tau \leq \epsilon_i$, $i \in \{1, \ldots, n\}$, *where* $||\cdot||$ *is a convex functional. Then* $||u + \tau f(u)|| \leq ||u||$ *for* $0 \leq \tau \leq \epsilon$, *where*

$$\epsilon = \left( \sum_{i=1}^{n} \frac{1}{\epsilon_i} \right)^{-1}.$$

**Example 2.19.** Consider the LeVeque and Yee problem [1, 14],

$$U_t + f(U)_x = s(U), \quad U(x, 0) = U_0(x), \quad x \in \mathbb{R},\ t \geq 0,$$

where $s(U) = -\mu U(U-1)(U-\frac{1}{2})$ and $\mu > 0$. Let $u_i(t) \approx U(x_i, t)$; then first-order upwind semi-discretization yields

$$u'(t) = F(u(t)) = D(u(t)) + S(u(t)), \quad u(0) = u_0, \quad t > 0,$$

where

$$D_i(u) = -\frac{f(u_i) - f(u_{i-1})}{\Delta x}, \qquad S_i(u) = s(u_i).$$

Consider also the downwind discretizations

$$\widetilde{D}_i(u) = -\frac{f(u_{i+1}) - f(u_i)}{\Delta x}, \qquad \widetilde{S}_i(u) = s(u_i),$$

and let $\widetilde{F} = \widetilde{D} + \widetilde{S}$. If $u \in [0,1]$, it can be easily shown that

$$0 \le u + \Delta t\, S(u) \le 1 \quad \text{for } 0 \le \Delta t \le \Delta t_{\mathrm{FE}} = \frac{2}{\mu},$$
$$0 \le u - \Delta t\, \widetilde{S}(u) \le 1 \quad \text{for } 0 \le \Delta t \le \widetilde{\Delta t}_{\mathrm{FE}} = \frac{16}{\mu}.$$

Using Lemma 2.18 we then have that

$$0 \le u + \Delta t\, F(u) \le 1 \quad \text{for } 0 \le \Delta t \le \Delta t_{\mathrm{FE}} = \frac{2\tau}{2+\mu\tau},$$
$$0 \le u - \Delta t\, \widetilde{F}(u) \le 1 \quad \text{for } 0 \le \Delta t \le \widetilde{\Delta t}_{\mathrm{FE}} = \frac{16\tau}{16+\mu\tau},$$

where $\tau > 0$ is such that

$$0 \le u + \Delta t\, D(u) \le 1 \quad \text{for } 0 \le \Delta t \le \tau,$$
$$0 \le u - \Delta t\, \widetilde{D}(u) \le 1 \quad \text{for } 0 \le \Delta t \le \tau.$$

Note that $\Delta t_{\mathrm{FE}} < \widetilde{\Delta t}_{\mathrm{FE}}$ for all positive values of $\mu$ and $\tau$. Therefore, under assumptions (A1) and (A2) above, the forward Euler step size must be $\Delta t_{\mathrm{FE}} = 2\tau/(2+\mu\tau)$ so that the numerical solution is stable. Let

$$\xi = \Delta t_{\mathrm{FE}} / \widetilde{\Delta t}_{\mathrm{FE}} = \frac{16+\mu\tau}{8(2+\mu\tau)},$$

then for all $\xi < 1$ we have $\widetilde{\mathcal{C}}_{\mathrm{LMM}} = \mathcal{C}(1) < \mathcal{C}(\xi)$; hence not considering SSP perturbed LMMs will always result in a stricter step-size restriction. Suppose $\mu$ is relatively small so that the problem is not stiff and explicit methods could be used. For instance, among the class of explicit two-step, second-order LMMs, there is no classical SSP method and the optimal downwind method has SSP coefficient $\widetilde{\mathcal{C}}_{\mathrm{LMM}} = 1/2$. Let for example $\mu\tau = 2/3$, then the step-size bound for downwind SSP methods such that the solution remains in $[0,1]$ is $\Delta t \le \widetilde{\mathcal{C}}_{\mathrm{LMM}} \Delta t_{\mathrm{FE}} = 0.375\tau$. Using the optimal two-step, second-order SSP perturbed LMM (see Appendix B) larger step sizes are allowed, since for $\xi = 25/32$ we have $\Delta t \le \mathcal{C}(\frac{25}{32})\, \Delta t_{\mathrm{FE}} = 0.3928\tau$.

## 3. Monotonicity of additive linear multistep methods

Following the previous example, it is natural to study the monotonicity properties of additive methods applied to problems which consist of components that

describe different physical processes. A $k$-step additive LMM for the solution of the initial value problem

$$\begin{aligned} \boldsymbol{u}'(t) &= \boldsymbol{F}(\boldsymbol{u}(t)) + \widehat{\boldsymbol{F}}(\boldsymbol{u}(t)), \quad t \geq t_0, \\ \boldsymbol{u}(t_0) &= \boldsymbol{u}_0, \end{aligned} \tag{3.1}$$

takes the form

$$\boldsymbol{u}_n = \sum_{j=0}^{k-1} \alpha_j \boldsymbol{u}_{n-k+j} + \Delta t \sum_{j=0}^{k} \left( \beta_j \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) + \hat{\beta}_j \widehat{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \right). \tag{3.2}$$

The method is explicit if $\beta_k = \hat{\beta}_k = 0$ and implicit if none of $\beta_k$ or $\hat{\beta}_k$ is equal to zero. It can be shown that method (3.2) is order $p$ accurate if

$$\begin{gathered} \sum_{j=0}^{k-1} \alpha_j = 1, \qquad \sum_{j=0}^{k-1} j\alpha_j + \sum_{j=0}^{k} \beta_j = k, \qquad \sum_{j=0}^{k-1} j\alpha_j + \sum_{j=0}^{k} \hat{\beta}_j = k, \\ \sum_{j=0}^{k-1} \alpha_j j^i + \sum_{j=0}^{k} \beta_j i j^{i-1} = k^i, \quad \sum_{j=0}^{k-1} \alpha_j j^i + \sum_{j=0}^{k} \hat{\beta}_j i j^{i-1} = k^i, \quad i \in \{2, \dots, p\}. \end{gathered} \tag{3.3}$$

The operators $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$ generally approximate different derivatives and also have different stiffness properties. We extend the analysis of monotonicity conditions for LMMs by assuming that $\boldsymbol{F}$ and $\widehat{\boldsymbol{F}}$ satisfy

$$\|\boldsymbol{u} + \Delta t \boldsymbol{F}(\boldsymbol{u})\| \leq \|\boldsymbol{u}\| \quad \forall \boldsymbol{u} \in \mathbb{R}^m,\ 0 \leq \Delta t \leq \Delta t_{\mathrm{FE}}, \tag{3.4a}$$

$$\|\boldsymbol{u} + \Delta t \widehat{\boldsymbol{F}}(\boldsymbol{u})\| \leq \|\boldsymbol{u}\| \quad \forall \boldsymbol{u} \in \mathbb{R}^m,\ 0 \leq \Delta t \leq \widehat{\Delta t}_{\mathrm{FE}}, \tag{3.4b}$$

respectively.

**Definition 3.1.** An additive LMM (3.2) is said to be strong-stability-preserving (SSP) if the following monotonicity conditions

$$\begin{aligned} \beta_j, \hat{\beta}_j &\geq 0, \quad j \in \{0, \dots, k\}, \\ \alpha_j - r\beta_j - \hat{r}\hat{\beta}_j &\geq 0, \quad j \in \{0, \dots, k-1\}. \end{aligned} \tag{3.5}$$

hold for $r \geq 0$ and $\hat{r} \geq 0$. For a fixed $\xi = \hat{r}/r$ the method has SSP coefficients $(\mathcal{C}(\xi), \widehat{\mathcal{C}}(\xi))$, where

$$\mathcal{C}(\xi) = \sup \{ r > 0 \mid \text{monotonicity conditions (3.5) hold with } \hat{r} = \xi r \} \tag{3.6}$$

and $\widehat{\mathcal{C}}(\xi) = \xi\, \mathcal{C}(\xi)$.

As in Section 2, it is clear that whenever the set in (3.6) is empty, then the method is non-SSP; in such cases we say the method has SSP coefficient equal to zero.

Define the vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_j^+(\cdot) \in \mathbb{R}^{p+1}$ as in (2.20) and (2.21). Then, by using the substitution

$$\gamma_j \coloneqq \alpha_j - r\beta_j - \hat{r}\hat{\beta}_j \quad \text{for} \quad j \in \{0, \dots, k-1\}, \tag{3.7}$$

the order conditions (3.3) can be expressed in terms of vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_j^+$:

$$\sum_{j=0}^{k-1}(\gamma_j + \hat{r}\hat{\beta}_j)\boldsymbol{a}_j + \sum_{j=0}^{k} r\beta_j \boldsymbol{b}_j^+(r^{-1}) = \boldsymbol{a}_k, \tag{3.8a}$$

$$\sum_{j=0}^{k-1}(\gamma_j + r\beta_j)\boldsymbol{a}_j + \sum_{j=0}^{k} \hat{r}\hat{\beta}_j \boldsymbol{b}_j^+(\hat{r}^{-1}) = \boldsymbol{a}_k. \tag{3.8b}$$

The above equations suggest a change of variables. Instead of considering the method's coefficients in terms of the column vectors

$$\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_{k-1})^\intercal, \quad \boldsymbol{\beta} = (\beta_0, \dots, \beta_k)^\intercal, \quad \hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_k)^\intercal,$$

and the order conditions independent of $r$ and $\hat{r}$, one can consider the coefficients $\boldsymbol{\gamma}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$ under the substitution (3.7). Let $\hat{r} = \xi r$, then the order conditions can be written as functions of $r$. In particular, the system of $p+1$ equations (3.8a) can be written as $A(r)\boldsymbol{x}(r) = \boldsymbol{a}_k$, where

$$A(r) = \left[\begin{array}{c|c|c|c|c|c} \boldsymbol{a}_0 & \dots & \boldsymbol{a}_{k-1} & r\boldsymbol{b}_0^+(r^{-1}) & \dots & r\boldsymbol{b}_{k-1}^+(r^{-1}) \quad r\boldsymbol{b}_k^+(r^{-1}) \end{array}\right]$$

and $\boldsymbol{x}(r) = \big(\boldsymbol{\delta}(r), \boldsymbol{\beta}\big) \in \mathbb{R}^{2k+1}$ with $\delta_j(r) = \gamma_j + \xi r\hat{\beta}_j$, $j \in \{0, \dots, k-1\}$. Define the feasible set

$$P(r) = \{\boldsymbol{x} \in \mathbb{R}^{2k+1} \mid A(r)\boldsymbol{x}(r) = \boldsymbol{a}_k,\ \boldsymbol{x}(r) \geq 0\}. \tag{3.9}$$

For a given $\xi$, if there exists a $k$-step, $p$-order accurate SSP additive LMM (3.2) with SSP coefficient $\mathcal{C}(\xi)$, then $P\big(\mathcal{C}(\xi)\big)$ is nonempty.

Consider the class of additive LMMs (3.2) having at most $k$ steps and order of accuracy at least $p$, denoted by $\mathcal{A}_{k,p}$. Since we would like to obtain the method with the largest possible SSP coefficient, then for a fixed $k \geq 1$, $p \geq 1$ and a given $\xi \in [0, \infty)$, we define optimal additive LMMs as follows.

**Definition 3.2.** An explicit (implicit) additive LMM (3.2) in $\mathcal{A}_{k,p}$ is called *optimal* if no explicit (implicit) method in $\mathcal{A}_{k,p}$ has larger SSP coefficient. Given $\xi \in [0, \infty)$, we denote the largest SSP coefficient for $k$-step explicit (implicit) additive LMMs (3.2) of order $p$ by

$$\mathcal{C}_{k,p}(\xi) = \sup_{(\boldsymbol{\alpha},\boldsymbol{\beta},\hat{\boldsymbol{\beta}})\in\mathcal{A}_{k,p}} \Big\{\mathcal{C}(\xi) > 0 \mid \mathcal{C}(\xi) \text{ is the SSP coefficient of an explicit (implicit) method (3.2) with coefficients } (\boldsymbol{\alpha}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}})\Big\}.$$

In the trivial case that the set over which the supremum is taken above is empty, we write $\mathcal{C}_{k,p}(\xi) = 0$.

The following theorem and lemma characterize the coefficients of an optimal additive LMM.

**Theorem 3.3.** *Let $k \geq 1$, $p \geq 1$ be given such that $0 < \mathcal{C}_{k,p}(\xi) < \infty$ for a given $\xi \in [0, \infty)$. Then, there exists a $k$-step optimal SSP additive LMM (3.2) of order $p$ with at most $p$ nonzero coefficients $\delta_j$, $\beta_i$, where $\delta_j = \alpha_j - \mathcal{C}_{k,p}(\xi)\beta_j$, $j \in \{0, \dots, k-1\}$ and $i \in \{0, \dots, k\}$.*

*Proof.* Let $0 < \mathcal{C}_{k,p}(\xi) < \infty$, where $k \geq 1$, $p \geq 1$ and $\xi \in [0, \infty)$ are given. Since the inequalities involved in (3.5) are not strict, the supremum in Definition 3.2 is always attained by some method. Consider an optimal $k$-step SSP additive LMM (3.2) of order $p$ with SSP coefficient $\mathcal{C}_{k,p}(\xi)$. Define $\gamma_j = \alpha_j - \mathcal{C}_{k,p}(\xi)\beta_j - \hat{\mathcal{C}}_{k,p}(\xi)\hat{\beta}_j$ and $\delta_j = \gamma_j + \hat{\mathcal{C}}_{k,p}(\xi)\hat{\beta}_j$ for $j \in \{0, \ldots, k-1\}$. Then the vector $\boldsymbol{x} = (\boldsymbol{\delta}, \boldsymbol{\beta}) \in \mathbb{R}^{2k+1}$ belongs to the feasible set (3.9) when $r = \mathcal{C}_{k,p}(\xi)$.

If $\boldsymbol{x}$ has $p$ or fewer nonzero entries, we are done. Suppose $\boldsymbol{x}$ has more than $p$ nonzero entries and let $S(r)$ be the set of columns of the matrix $A(r)$ in (3.9) corresponding to the nonzero elements of $\boldsymbol{x}$.

We distinguish two cases. First, assume that the set $S\big(\mathcal{C}_{k,p}(\xi)\big)$ does not span $\mathbb{R}^{p+1}$. Then, similarly to Case 2 of the proof of Theorem 2.13, the vectors $\boldsymbol{a}_k$ (explicit case) and $\boldsymbol{a}_k/(1 + r\beta_k)$ (implicit case) belong to the convex hull of $S(r)$, for $r = \mathcal{C}_{k,p}(\xi)$. Then from Lemma 2.12, $\boldsymbol{a}_k$ and $\boldsymbol{a}_k/(1 + \mathcal{C}_{k,p}(\xi)\beta_k)$ can be written, respectively, as a convex and nonnegative combination of $p$ vectors in $S\big(\mathcal{C}_{k,p}(\xi)\big)$. Therefore, there exists a different method with the SSP coefficient $\mathcal{C}_{k,p}(\xi)$ and at most $p$ nonnegative coefficients $\delta_j$, $j \in \{0, \ldots, k-1\}$ and $\beta_i$, $i \in \{0, \ldots, k\}$. On the other hand, if $S\big(\mathcal{C}_{k,p}(\xi)\big)$ spans $\mathbb{R}^{p+1}$, then by using Lemma 2.11 there exists $\epsilon > 0$ and $\boldsymbol{x}^* = (\boldsymbol{\delta}^*, \boldsymbol{\beta}^*)$ with nonnegative entries such that $A(\mathcal{C}_{k,p}(\xi) + \epsilon)\boldsymbol{x}^* = \boldsymbol{a}_k$. For each index $j$ in $\boldsymbol{x}^*$ such that $\delta_j^* > 0$, we can choose $\gamma_j^*$ so that $\beta_j^* = \hat{\beta}_j^*$. Then, $\boldsymbol{x}^*$ satisfies (3.8b) as well. But this contradicts the optimality of the method since we have constructed a $k$-step SSP additive LMM of order $p$ with coefficients given by $\boldsymbol{x}^*$ and SSP coefficient $\mathcal{C}_{k,p}(\xi) + \epsilon$. ☐

**Lemma 3.4.** *For a given $k \geq 1$, $p \geq 1$ an optimal additive LMM* (3.2) *has $\beta_j = \hat{\beta}_j$ for all $j \in \{0, \ldots, k\}$.*

*Proof.* Consider an optimal method (3.2) of order $p$. From Theorem 3.3 at most $p$ coefficients $\delta_j, \beta_i$, $j \in \{0, \ldots, k-1\}$, $i \in \{0, \ldots, k\}$ are nonzero. Let $\boldsymbol{v} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$, then $\boldsymbol{v}$ has at most $p$ nonzero elements. Subtracting the order conditions (3.8) results in

$$\sum_{i \in I} v_i \bar{\boldsymbol{a}}_i = 0,$$

where $I$ is the set of distinct indexes for which $v_i$'s are nonzero. The vectors $\bar{\boldsymbol{a}}_i = (1, i, \ldots, i^{p-1})^\intercal$, $i \in I$ are linearly independent (see [5, Chapter 21]), therefore $\boldsymbol{v}$ must be identically equal to zero. Hence, $\beta_j = \hat{\beta}_j$ for each $j \in \{0, \ldots, k\}$. ☐

The main result of this section relies on Theorem 3.3 and Lemma 3.4 and is given by the following theorem.

**Theorem 3.5.** *For a given $k \geq 1$, $p \geq 1$ an optimal additive LMM with SSP coefficient $\mathcal{C}_{k,p}$ and corresponding SSP coefficient $\widehat{\mathcal{C}}_{k,p}$ is equivalent to the optimal $k$-step optimal SSP LMM* (1.4) *of order $p$ with SSP coefficient $\mathcal{C}_{k,p} + \widehat{\mathcal{C}}_{k,p}$.*

*Proof.* Consider an optimal method (3.2) of order $p$ with SSP coefficient $\mathcal{C}_{k,p}$ and $\widehat{\mathcal{C}}_{k,p} = \xi\,\mathcal{C}_{k,p}$ for some $\xi \in [0, \infty)$. From Lemma 3.4 we have $\beta_j = \hat{\beta}_j$ for each $j \in \{0, \ldots, k\}$, therefore monotonicity conditions (3.5) yield $\min_j \frac{\alpha_j}{\beta_j} = \mathcal{C}_{k,p} + \widehat{\mathcal{C}}_{k,p}$. Thus the additive LMM is equivalent to the optimal $k$-step SSP LMM method of order $p$ with SSP coefficient $\mathcal{C}_{k,p} + \widehat{\mathcal{C}}_{k,p}$. ☐

3.1. **Monotone IMEX linear multistep methods.** Based on Theorem 3.5, it is only interesting to consider Implicit-Explicit (IMEX) SSP LMMs. Such methods are particularly useful for initial value problems (3.1) where $\boldsymbol{F}$ represents a nonstiff or mild stiff part of the problem, and $\widehat{\boldsymbol{F}}$ a stiff term for which implicit integration is required. The following theorem provides sufficient conditions for monotonicity for the numerical solution of an IMEX method.

**Theorem 3.6.** *Consider the additive problem* (3.1) *for which* $\boldsymbol{F}$ *and* $\widehat{\boldsymbol{F}}$ *satisfy* (3.4)*, for some* $\Delta t_{\mathrm{FE}} > 0$ *and* $\widehat{\Delta t}_{\mathrm{FE}} > 0$*. Let an IMEX LMM* (3.2) *with coefficients* $\beta_k = 0$*,* $\hat{\beta}_k \neq 0$ *be strong-stability-preserving with SSP coefficients* $(\mathcal{C}(\xi), \widehat{\mathcal{C}}(\xi))$ *for* $\xi = \Delta t_{\mathrm{FE}}/\widehat{\Delta t}_{\mathrm{FE}}$*. Then, the numerical solution satisfies the monotonicity condition* (1.5) *under a step-size restriction*

$$\Delta t \leq \min\{\mathcal{C}\,\Delta t_{\mathrm{FE}}, \widehat{\mathcal{C}}\,\widehat{\Delta t}_{\mathrm{FE}}\}. \tag{3.10}$$

*Proof.* The proof is similar to that of Theorem 2.4. □

As in Section 2, the minimum step size in (3.10) occurs when $\mathcal{C}\,\Delta t_{\mathrm{FE}} = \widehat{\mathcal{C}}\,\widehat{\Delta t}_{\mathrm{FE}}$. For a given $k \geq 1$ and $p \geq 1$, we would like to find the largest possible value $\mathcal{C}_{k,p}(\xi)$ such that an optimal IMEX method is SSP with coefficients $(\mathcal{C}_{k,p}, \mathcal{C}_{k,p}\,\Delta t_{\mathrm{FE}}/\widehat{\Delta t}_{\mathrm{FE}})$. Setting $\xi \coloneqq \Delta t_{\mathrm{FE}}/\widehat{\Delta t}_{\mathrm{FE}}$, and combining the inequalities (3.5) and the order conditions (3.3), we can form the following optimization problem:

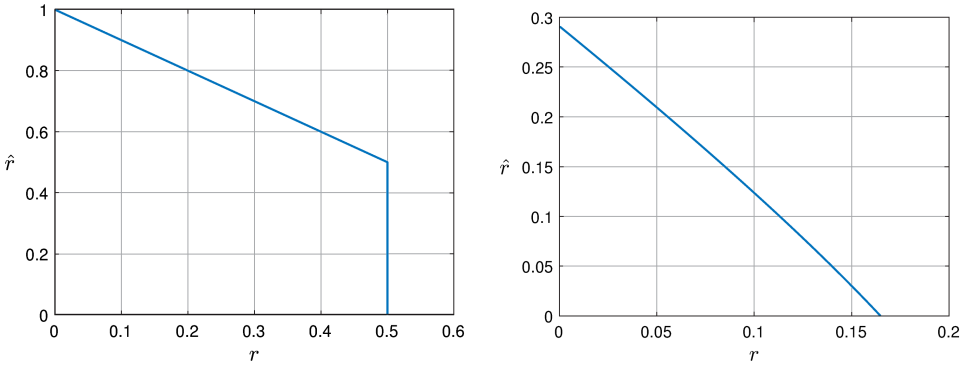
$$\max_{\{\boldsymbol{\gamma},\boldsymbol{\beta},\hat{\boldsymbol{\beta}},r\}} r, \qquad \text{subject to}$$

(3.11)
$$\begin{cases} \displaystyle\sum_{j=0}^{k-1} \gamma_j + r(\beta_j + \xi\hat{\beta}_j) = 1, \quad \sum_{j=0}^{k-1} \big(\gamma_j + r(\beta_j + \xi\hat{\beta}_j)\big) j + \beta_j = k, \\ \displaystyle\sum_{j=0}^{k-1} \big(\gamma_j + r(\beta_j + \xi\hat{\beta}_j)\big) j^i + \beta_j i j^{i-1} = k^i, & i \in \{2,\dots,p\}, \\ \displaystyle\sum_{j=0}^{k-1} (\beta_j - \hat{\beta}_j) - \hat{\beta}_k = 0, \quad \sum_{j=0}^{k-1} (\beta_j - \hat{\beta}_j) j^i - \hat{\beta}_k k^i = 0, & i \in \{1,\dots,p-1\}, \\ \gamma_j \geq 0, \beta_j \geq 0, & j \in \{0,\dots,k-1\}, \\ \hat{\beta}_j \geq 0, & j \in \{0,\dots,k\}, \\ r \geq 0. \end{cases}$$

By using bisection in $r$, the optimization problem (3.11) can be viewed as a sequence of linear feasible problems, as suggested in [10]. We solved the above problem using `linprog` in `MATLAB` with the active-set algorithm and found optimal IMEX SSP methods for $k \in \{1,\dots,50\}, p \in \{1,\dots,15\}$ and for different values of $\xi$. Similarly to additive Runge–Kutta methods [6], we can define the feasibility SSP region of IMEX SSP methods for a fixed $k \geq 1$ and $p \geq 1$ by

$$R_{k,p} = \big\{(r,\hat{r}) \mid \xi \in \mathbb{R}^+ \text{ and monotonicity conditions (3.5) hold for } r \geq 0, \hat{r} = \xi r\big\}.$$

For instance, the feasibility SSP regions for three-step, second-order and six-step, fourth-order IMEX methods are shown in Figure 2.




(a) Three-step, second-order IMEX SSP region (b) Six-step, fourth-order IMEX SSP region

FIGURE 2. SSP regions of IMEX LMMs. The intersection of the line $\hat{r} = \xi r$, $\xi > 0$ with the boundary of the SSP region corresponds to an optimal IMEX LMM with SSP coefficient $\mathcal{C}(\xi)$.

As mentioned in [8, Section 2.2] the SSP coefficients of IMEX SSP methods in the case that the forward Euler ratio $\xi = \Delta t_{\text{FE}}/\widehat{\Delta t}_{\text{FE}}$ is equal to one are not large. The same seems to hold when considering SSP IMEX methods for additive problems (3.1) satisfying (3.4) for any values $\xi \geq 0$ (see Figure 2). Thus, instead of requiring both parts of an IMEX method to be SSP, one can impose SSP conditions only on the explicit part and optimize stability properties for the implicit method. Second-order methods among this class of methods have been studied in [2], whereas in [8] higher order IMEX methods with optimized stability features were constructed based on general monotonicity and boundedness properties of the explicit component.

## 4. Conclusion and future work

We have investigated a generalization of the linear multistep methods with upwind- and downwind-biased operators introduced in [20,21], by considering problems in which the downwind operator satisfies a forward Euler condition with different step-size restriction than that of the upwind operator. We expressed the perturbed LMMs in an additive form and analyzed their monotonicity properties. By optimizing in terms of the upwind and downwind Euler step sizes, methods with larger SSP step sizes are obtained for such problems. We studied additive problems in the same framework, and we have shown that when both parts of the method are explicit (or both parts are implicit), the optimal additive SSP methods lie within the class of traditional (nonadditive) SSP linear multistep methods. Finally, we have seen that IMEX SSP methods for additive problems allow relatively small monotonicity-preserving step sizes.

The concepts of additive splitting and downwind semi-discretization can be combined to yield downwind IMEX LMMs of the form (applying downwinding to the nonstiff term):

$$\begin{aligned}\boldsymbol{u}_n = &\sum_{j=0}^{k-1} \alpha_j \boldsymbol{u}_{n-k+j} + \Delta t \sum_{j=0}^{k-1} \left( \beta_j \boldsymbol{F}(\boldsymbol{u}_{n-k+j}) - \tilde{\beta}_j \widetilde{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}) \right) \\ &+ \Delta t \sum_{j=0}^{k} \hat{\beta}_j \widehat{\boldsymbol{F}}(\boldsymbol{u}_{n-k+j}),\end{aligned} \tag{4.1}$$

where $\boldsymbol{F}$ and $\widetilde{\boldsymbol{F}}$ satisfy the forward Euler conditions (1.10) and the explicit part is an SSP perturbed LMM. Preliminary results show that it is possible to obtain second-order IMEX linear multistep methods with two or three steps, where the implicit part is A-stable and the explicit part is an optimal SSP perturbed LMM. This generalization allows the construction of new IMEX methods with fewer steps for a given order of accuracy and with larger SSP coefficients (for the explicit component). Moreover, the best possible IMEX method can be chosen based on the ratio of forward Euler step sizes of the nonstiff term in (3.1). Also, it is worth investigating the possibility of obtaining A($\alpha$)-stable implicit parts whenever A-stability is not feasible. Work on optimizing the stability properties of the IMEX methods (4.1) is ongoing and will be presented in a future work. Analysis of SSP perturbed LMMs with variable step sizes and monotonicity properties of perturbed LMMs with special starting procedures can also be studied.

## Appendix A. Proofs of lemmata in section 2

In this section we present the proofs of some technical lemmata that were omitted in the previous sections.

*Proof of Lemma* 2.12. Consider a set of distinct vectors $S = \{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_m\}$ in $\mathbb{R}^n$. Let a nonzero vector $\boldsymbol{z} \in C$ be given, where $C = \operatorname{conv}(S)$ is the convex hull of $S$. Then there exist nonnegative coefficients $\lambda_j$ that sum to unity such that

$$\boldsymbol{z} = \sum_{j=1}^{m} \lambda_j \boldsymbol{\psi}_j.$$

If $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_m$ are linearly independent, it must be that $m \le n$ and both parts (a) and (b) of the lemma hold trivially. Therefore, assume the vectors in $S$ are linearly dependent. Then, we can find $\mu_j$ not all zero and at least one which is positive, such that

$$\sum_{j=1}^{m} \mu_j \boldsymbol{\psi}_j = 0.$$

Define

$$\nu \coloneqq \min_{1 \le j \le m} \left\{ \frac{\lambda_j}{\mu_j} \mid \mu_j > 0 \right\} = \frac{\lambda_{j_0}}{\mu_{j_0}},$$

then we have $\nu \mu_j \le \lambda_j$ for all $j \in \{1, \dots, m\}$, where equality holds for at least $j = j_0$. Let $\tilde{\lambda}_j = \lambda_j - \nu \mu_j$ for $j \in \{1, \dots, m\}$. By the choice of $\nu$, all coefficients $\tilde{\lambda}_j$

are nonnegative and at least one of them is equal to zero. Note that

$$\boldsymbol{z} = \sum_{j=1}^{m} \lambda_j \boldsymbol{\psi}_j - \nu \sum_{j=1}^{m} \mu_j \boldsymbol{\psi}_j = \sum_{j=1}^{m} \tilde{\lambda}_j \boldsymbol{\psi}_j,$$

hence $\boldsymbol{z}$ can be expressed as a nonnegative linear combination of at most $m-1$ vectors in $S$. The above argument can be repeated until $\boldsymbol{z}$ is written as a nonnegative linear combination of linearly independent vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_r$, $r \leq n$. This proves part (a).

For part (b), suppose $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$ are linearly dependent and belong in $\{(1, \boldsymbol{v}) \mid \boldsymbol{v} \in \mathbb{R}^{n-1}\}$. Then, any nonzero vector $\boldsymbol{z} \in C$ has the form $(1, \boldsymbol{v})^\intercal$, $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and from part (a) can be written as a nonnegative combination of at most $n$ linearly independent vectors in $S$ with coefficients $\tilde{\lambda}_j$. In addition $\sum_{j=1}^{m} \tilde{\lambda}_j = 1$, since the first component of vector $\boldsymbol{z}$ and all $\boldsymbol{\psi}_j$, $j \in \{1, \ldots, m\}$ is equal to unity. □

*Proof of Lemma* 2.18. Let $p_i(u; \epsilon_i) \coloneqq u + \epsilon_i f_i(u)$, then we have

$$f_i(u) = \frac{p_i(u; \epsilon_i) - u}{\epsilon_i} \quad \text{for } i \in \{1, \ldots, n\}.$$

Using $\sum_{i=1}^{n} \epsilon/\epsilon_i = 1$ and the assumption of the lemma, it can be shown that

$$\begin{aligned} \|u + \epsilon f(u)\| &= \left\| u + \sum_{i=1}^{n} \frac{\epsilon}{\epsilon_i} \left(p_i(u; \epsilon_i) - u\right) \right\| \\ &= \left\| \sum_{i=1}^{n} \frac{\epsilon}{\epsilon_i} p_i(u; \epsilon_i) \right\| \\ &\leq \sum_{i=1}^{n} \frac{\epsilon}{\epsilon_i} \|u\| = \|u\|. \end{aligned}$$

The rest of the proof relies on [15, Lemma II.5.1]. If $0 < \tau < \epsilon$, then there exist $0 < \rho < 1$ such that $\tau = (1-\rho)\epsilon$. Then $u + \tau f(u) = u + (1-\rho)\epsilon f(u) = \rho u + (1-\rho)\left(u + \epsilon f(u)\right)$ and hence

$$\begin{aligned} \|u + \tau f(u)\| - \|u\| &\leq \rho\|u\| + (1-\rho)\|u + \epsilon f(u)\| - \|u\| \\ &= (1-\rho)\left(\|u + \epsilon f(u)\| - \|u\|\right) \\ &\leq \|u + \epsilon f(u)\| - \|u\|. \end{aligned}$$

This implies that $\|u + \tau f(u)\| \leq \|u + \epsilon f(u)\|$. If $\tau = 0$ or $\tau = \epsilon$, then the inequality $\|u + \tau f(u)\| \leq \|u\|$ is trivial; hence the lemma's result holds for all $0 \leq \tau \leq \epsilon$. □

## Appendix B. Coefficients of methods used in Section 2.2

The coefficients of the optimal explicit two-step, second-order SSP perturbed LMM used in Example 2.17 are given by:

$$\begin{aligned} &\alpha_0 = 0.409332709113745, \quad \beta_0 = 0.0, && \tilde{\beta}_0 = 0.295333645443128, \\ &\alpha_1 = 0.590667290886257, \quad \beta_1 = 1.704666354556872, && \tilde{\beta}_1 = 0.0. \end{aligned}$$

The method is related to a ratio of forward Euler step sizes $\Delta t_{\text{FE}}/\widetilde{\Delta t}_{\text{FE}} = 4$, and has SSP coefficient $\mathcal{C}_{2,2}(4) = 0.3465$.

The optimal explicit two-step, second-order SSP perturbed LMM used in Example 2.19 has coefficients

$$\begin{array}{lll} \alpha_0 = 0.169849709137948, & \beta_0 = 0.0, & \tilde{\beta}_0 = 0.415075145431026, \\ \alpha_1 = 0.830150290862053, & \beta_1 = 1.584924854568973, & \tilde{\beta}_1 = 0.0, \end{array}$$

and is related to a ratio of forward Euler step sizes $\Delta t_{\mathrm{FE}}/\widetilde{\Delta t}_{\mathrm{FE}} = 25/32$ with SSP coefficient $\mathcal{C}_{2,2}(\frac{25}{32}) = 0.5238$.

## Acknowledgments

The authors would like to thank the anonymous referees for their suggestions that significantly improved the paper. Also, they would like to thank Lajos Lóczi and Inmaculada Higueras for carefully reading this manuscript and making valuable comments.

## References

[1] R. Donat, I. Higueras, and A. Martínez-Gavara, *On stability issues for IMEX schemes applied to 1D scalar hyperbolic equations with stiff reaction terms*, Math. Comp. **80** (2011), no. 276, 2097–2126, DOI 10.1090/S0025-5718-2011-02463-4. MR2813350

[2] T. Gjesdal, *Implicit-explicit methods based on strong stability preserving multistep time discretizations*, Appl. Numer. Math. **57** (2007), no. 8, 911–919, DOI 10.1016/j.apnum.2006.09.001. MR2331084

[3] S. Gottlieb, D. Ketcheson, and C.-W. Shu, *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011. MR2789749

[4] S. Gottlieb and S. J. Ruuth, *Optimal strong-stability-preserving time-stepping schemes with fast downwind spatial discretizations*, J. Sci. Comput. **27** (2006), no. 1-3, 289–303, DOI 10.1007/s10915-005-9054-8. MR2285782

[5] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. MR1927606

[6] I. Higueras, *Strong stability for additive Runge-Kutta methods*, SIAM J. Numer. Anal. **44** (2006), no. 4, 1735–1758, DOI 10.1137/040612968. MR2257125

[7] I. Higueras, *Positivity properties for the classical fourth order Runge-Kutta method*, Proceedings of the MCalv6(5) Conference, Monogr. Real Acad. Ci. Exact. Fís.-Quím. Nat. Zaragoza, vol. 33, 2010, pp. 125–139. MR3346421

[8] W. Hundsdorfer and S. J. Ruuth, *IMEX extensions of linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys. **225** (2007), no. 2, 2016–2042, DOI 10.1016/j.jcp.2007.03.003. MR2349693

[9] W. Hundsdorfer, S. J. Ruuth, and R. J. Spiteri, *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal. **41** (2003), no. 2, 605–623, DOI 10.1137/S0036142902406326. MR2004190

[10] D. I. Ketcheson, *Computation of optimal monotonicity preserving general linear methods*, Math. Comp. **78** (2009), no. 267, 1497–1513, DOI 10.1090/S0025-5718-09-02209-1. MR2501060

[11] D. I. Ketcheson, *Step sizes for strong stability preservation with downwind-biased operators*, SIAM J. Numer. Anal. **49** (2011), no. 4, 1649–1660, DOI 10.1137/100818674. MR2831065

[12] H. W. J. Lenferink, *Contractivity preserving explicit linear multistep methods*, Numer. Math. **55** (1989), no. 2, 213–223, DOI 10.1007/BF01406515. MR987386

[13] H. W. J. Lenferink, *Contractivity-preserving implicit linear multistep methods*, Math. Comp. **56** (1991), no. 193, 177–199, DOI 10.2307/2008536. MR1052098

[14] R. J. LeVeque and H. C. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys. **86** (1990), no. 1, 187–210, DOI 10.1016/0021-9991(90)90097-K. MR1033905

[15] R. H. Martin Jr., *Nonlinear Operators and Differential Equations in Banach Spaces*, Pure and Applied Mathematics, Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1976. MR0492671

[16] O. Nevanlinna and W. Liniger, *Contractive methods for stiff differential equations. I*, BIT **18** (1978), no. 4, 457–474, DOI 10.1007/BF01932025. MR520755
[17] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1996.
[18] S. J. Ruuth and W. Hundsdorfer, *High-order linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys. **209** (2005), no. 1, 226–248, DOI 10.1016/j.jcp.2005.02.029. MR2145787
[19] J. Sand, *Circle contractive linear multistep methods*, BIT **26** (1986), no. 1, 114–122, DOI 10.1007/BF01939367. MR833836
[20] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput. **9** (1988), no. 6, 1073–1084, DOI 10.1137/0909073. MR963855
[21] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys. **77** (1988), no. 2, 439–471, DOI 10.1016/0021-9991(88)90177-5. MR954915
[22] M. N. Spijker, *Contractivity in the numerical solution of initial value problems*, Numer. Math. **42** (1983), no. 3, 271–290, DOI 10.1007/BF01389573. MR723625

4700 King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

*Current address*: Eötvös Loránd University, MTA-ELTE Numerical Analysis and Large Networks Research Group, Pázmány Péter sétány 1/C, Budapest H-1117, Hungary; and Budapest University of Technology and Economics, Department of Differential Equations, Building H, Egry József utca 1, Budapest H-1111, Hungary

*Email address*: hadjimy@cs.elte.hu

4700 King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia.

*Email address*: david.ketcheson@kaust.edu.sa