SIAM J. SCI. COMPUT.
Vol. 42, No. 2, pp. A666–A697

SPLITTING METHODS FOR ROTATIONS: APPLICATION TO VLASOV EQUATIONS*

JOACKIM BERNIER†, FERNANDO CASAS‡, AND NICOLAS CROUSEILLES§

Abstract. In this work, a splitting strategy is introduced to approximate two-dimensional rotation motions. Unlike standard approaches based on directional splitting which usually lead to a wrong angular velocity and then to large error, the splitting studied here turns out to be *exact* in time. Combined with spectral methods, the so-obtained numerical method is able to capture the solution to the associated partial differential equation with a very high accuracy. A complete numerical analysis of this method is given in this work. Then, the method is used to design highly accurate time integrators for Vlasov type equations: the Vlasov–Maxwell and the Vlasov-HMF systems. Finally, several numerical illustrations and comparisons with methods from the literature are discussed.

Key words. splitting, rotation, Vlasov equations, high-order time integrators

AMS subject classifications. 65M12, 65P10, 65Z05

DOI. 10.1137/19M1273918

1. Introduction.

The main goal of this work is to introduce a splitting strategy to deal with rotation motions and to apply it to construct efficient high-order time integrators for Vlasov type equations. The splitting is based on the fact that a rotation of angle θ can be decomposed into a product of three shear transformations,

$$\begin{pmatrix} 1 & -\tan\theta/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \sin\theta & 1 \end{pmatrix} \begin{pmatrix} 1 & -\tan\theta/2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = e^{\theta J}, \tag{1.1}$$

for $\theta \neq k\pi, k \in \mathbb{Z}^\star$ and where J is the fundamental symplectic matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{1.2}$$

Note that this decomposition into shear matrices can be derived using formal computations and already has been introduced in the image processing community (see [25, 28, 1, 30, 11]), in which several approaches have been developed to rotate an image on a computer screen. Moreover, this approach has also been used to design numerical methods for Gross–Pitaevskii equations (see [13] and [3, Lemma II.2]) in which this underlying splitting is used to solve exactly the harmonic oscillator.

*Submitted to the journal's Methods and Algorithms for Scientific Computing section July 10, 2019; accepted for publication (in revised form) December 17, 2019; published electronically March 17, 2020.

https://doi.org/10.1137/19M1273918

Funding: The work of the second author was supported by the Ministerio de Economia y Competitividad (Spain) through project MTM2016-77660-P (AEI/FEDER, UE), by Mobility Grant PRX18/00145, and by the Centre Henri Lebesgue. The work of the third author was supported by Enabling Research EUROFusion project MAGYK 2019-2020. The authors would like to thank P. Navaro for his help on the numerical part. This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom Research and Training Programme 2019–2020 under grant agreement 633053.

†Univ Rennes, INRIA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France (joackim.bernier@ens-rennes.fr).

‡IMAC and Departament de Matemàtiques, Universitat Jaume I, 12071 Castellón, Spain (fernando.casas@uji.es).

§Univ Rennes, INRIA-Rennes Bretagne Atlantique, MINGuS Project, IRMAR and ENS Rennes, Rennes, France (nicolas.crouseilles@inria.fr).

To make the link between (1.1) and the underlying partial differential equation, we aim to integrate the following two-dimensional transport equation:

$$\partial_t u = Jx \cdot \nabla u, \quad x \in \mathbb{R}^2, \tag{1.3}$$

with the initial condition $u(t = 0, x) = u^{in}(x)$. The exact solution of (1.3) at time t is $u(t, x) = u^{in}(e^{tJ}x)$, which is nothing but the rotation of angle t of the initial condition u^{in}. When the initial condition is not known analytically or when (1.3) is a part of a more complicated model, then one only has access to discrete information of the initial condition and a numerical method is required to approximate (1.3). Our goal in this work is to introduce a directional splitting inspired by (1.1) which is exact with respect to the time variable.

Obviously, standard finite differences or finite volumes based methods can be used to approximate the spatial direction x and coupled to Runge–Kutta strategies in time. However, this leads to methods which usually suffer from a strong CFL condition on the time step. Then, semi-Lagrangian methods are preferred, since they are free from the stability condition but still keep Eulerian accuracy (see [26, 18, 32, 14]). For (1.3), the feet of the characteristics can be computed exactly and a two-dimensional interpolation has to be performed to update the numerical unknown. However, high-dimensional interpolation is known to be nonconservative and it is obviously more demanding in terms of complexity and time. Then, splitting methods are very competitive since they reduce the problem into very simple one-dimensional linear transport equations which can be solved efficiently with semi-Lagrangian methods (using high-order or even spectral interpolation). Moreover, in a splitting procedure, the variable that does not appear in the derivative is just a parameter so that a very simple parallelization can be performed by distributing the computation on the processors according to the values of this parameter.

For rotation dynamics, however, the standard splitting strategy (like Strang or Lie splitting, for example) can induce some error since it involves a wrong rotational velocity (see [8]). Here, we propose a new splitting which enables us to solve (1.3) exactly in time (like in [13, 3]). Moreover, when this splitting is coupled with spectral methods (and under some assumptions detailed in what follows), the so-obtained method is able to capture to a very high accuracy the exact solution (spectral accuracy in practice). A complete proof of convergence of the fully discretized numerical method is performed. We will see that this strategy and some simple extensions turn out to be very efficient compared to standard methods when applied to the following problems. First, it enables us to design high-order (in time) methods for the Vlasov–Maxwell system. Second, when applied to the Vlasov-HMF model in the close-to-equilibrium regime (see [21]), this splitting turns out to be more efficient than the Strang one.

Concerning the Vlasov–Maxwell solvers, our goal was to improve the method introduced in [16] in which a splitting into three parts has been proposed. The exact treatment of the rotation enables us to solve exactly and efficiently the magnetic part which is then very helpful when designing high-order splitting methods for the full Vlasov–Maxwell system. The resulting schemes are fourth-order accurate in time and preserve the Gauss condition exactly. We also use the new splitting to approximate the solution of the Vlasov-HMF system, for which the close-to-equilibrium dynamics is driven by the linearized Hamiltonian part (see [21]). For such Hamiltonian, the new splitting has a good behavior (see [4]) and we compare its efficiency with the standard Strang splitting by studying perturbations of a nonhomogeneous equilibrium state.

The rest of the paper is organized as follows. First, the method is presented in the context of the numerical approximation of transport equation of the form (1.3) and a complete proof of convergence is performed with some numerical illustrations.

Then, the Vlasov–Maxwell system is presented and we explain how the new method is used to design high-order Vlasov–Maxwell solvers. Finally, some numerical results are given to show the benefit of the new method in the Vlasov context. Note that several technical details concerning the proofs or analyses can be found as supplementary material accompanying this paper.

2. Presentation of the method and its numerical analysis. In this section, we focus on the two-dimensional equation

$$\partial_t u = Jx \cdot \nabla u, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{2.1}$$

supplemented with an initial condition $u(t = 0, x) = u^{in}(x)$.

We intend to analyze the convergence of a splitting in time based numerical scheme coupled with a spectral method in space (i.e., in the x-direction). More precisely, we want to solve (2.1) on $[t_n, t_{n+1}]$; then we want to compute $u^{n+1}(x)$, an approximation of $u(t_{n+1}, x_1, x_2)$, the solution at time $t_{n+1} = t_n + \delta_t$ ($\delta_t > 0$ being the time step and $n \in \mathbb{N}$) of (2.1) with initial condition $u^{in}(x_1, x_2) = u(t_n, x_1, x_2)$ at time $t_n = n\delta_t, n \in \mathbb{N}$. To do so, we propose a new splitting in which each step is a shear transformation.

Let us introduce some notation. For a given 2×2 matrix A, we denote by $\exp(\delta_t Ax \cdot \nabla)u^n$ the solution at time t_{n+1} of

$$\left\{ \begin{array}{lcl} \partial_t u(t,x) & = & Ax \cdot \nabla u(t,x), \quad x \in \mathbb{R}^2, \\ u^{in}(x) & = & u^n(x). \end{array} \right. \tag{2.2}$$

Then, from (1.1), we search for $a, b \in \mathbb{R}$ so that the relation

$$e^{-\frac{a}{2}x_2\partial_{x_1}} e^{bx_1\partial_{x_2}} e^{-\frac{a}{2}x_2\partial_{x_1}} u^n = e^{\delta_t Jx\cdot\nabla} u^n \tag{2.3}$$

holds true, which can be written equivalently as

$$e^{A_1 x\cdot\nabla} e^{A_2 x\cdot\nabla} e^{A_1 x\cdot\nabla} u^n = e^{\delta_t Jx\cdot\nabla} u^n \tag{2.4}$$

with

$$A_1 = \left(\begin{array}{cc} 0 & -a/2 \\ 0 & 0 \end{array} \right), \qquad A_2 = \left(\begin{array}{cc} 0 & 0 \\ b & 0 \end{array} \right). \tag{2.5}$$

Using the method of characteristics, we have for (2.2)

$$e^{\delta_t Ax\cdot\nabla} u^n = u^n \circ e^{\delta_t A}, \quad \delta_t \geq 0,$$

so that (2.4) is nothing but $u^n(e^{A_1} e^{A_2} e^{A_1} x) = u^n(e^{\delta_t J} x)$. Since A_1 and A_2 are nilpotent matrices, their exponential follows readily,

$$e^{A_1} = \left(\begin{array}{cc} 1 & -a/2 \\ 0 & 1 \end{array} \right), \qquad e^{A_2} = \left(\begin{array}{cc} 1 & 0 \\ b & 1 \end{array} \right),$$

and it is clear from (1.1) that the choice $a = 2\tan(\delta_t/2)$ and $b = \sin(\delta_t)$ leads to an exact splitting in time (and this choice is unique), so that the scheme then is written as $u^{n+1}(x) = u^n(e^{A_1} e^{A_2} e^{A_1} x)$, with A_1 and A_2 given by (2.5). Let us remark that the usual Strang splitting corresponds to $a = b = \delta_t$.

In consequence, now we have to solve shear transformations, which are nothing but one-dimensional linear advections. We consider here using a pseudospectral method. To do so, we discretize a square, of size R, centered in 0 (i.e., $\left[-\frac{R}{2}, \frac{R}{2}\right]^2$) with a regular

grid with $N \in \mathbb{N}^*$ points per direction. Its stepsize is $h = R/N$. We denote this grid by $\mathbb{G}^2$, with

$$\mathbb{G} = h \left[\!\!\left[-\left\lfloor \frac{N-1}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor \right]\!\!\right]. \tag{2.6}$$

Then, we define the discrete partial Fourier transforms

$$\mathcal{F}_1 : \begin{cases} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\widehat{\mathbb{G}}\times\mathbb{G}}, \\ \boldsymbol{u} & \mapsto & h \displaystyle\sum_{g_1\in\mathbb{G}} \boldsymbol{u}_{g_1,g_2}\, e^{-ig_1\xi_1} \end{cases} \quad \text{and } \mathcal{F}_2 : \begin{cases} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}\times\widehat{\mathbb{G}}}, \\ \boldsymbol{u} & \mapsto & h \displaystyle\sum_{g_2\in\mathbb{G}} \boldsymbol{u}_{g_1,g_2}\, e^{-ig_2\xi_2} \end{cases},$$

where $\widehat{\mathbb{G}} = \eta \left[\!\!\left[-\left\lfloor \frac{N-1}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor \right]\!\!\right]$ stands for the set of discrete frequencies with $\eta = 2\pi/R$.

Now, we want to solve the continuous shear transformations ($\alpha \in \mathbb{R}$),

$$\partial_t u = \alpha x_2 \partial_{x_1} u, \quad \partial_t u = \alpha x_1 \partial_{x_2} u, \tag{2.7}$$

which are the basic building blocks of the splitting presented above. These shear transformations are particularly simple to solve and we shall use a pseudospectral method. Then, for any parameter $\alpha \in \mathbb{R}$, we introduce two pseudospectral shear transformations,

$$\mathcal{S}_1^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}^2}, \\ \boldsymbol{u} & \mapsto & \mathcal{F}_1^{-1}\left[e^{i\alpha\xi_1 g_2}\mathcal{F}_1\,\boldsymbol{u}\right] \end{array} \right. \tag{2.8}$$

and

$$\mathcal{S}_2^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}^2}, \\ \boldsymbol{u} & \mapsto & \mathcal{F}_2^{-1}\left[e^{i\alpha\xi_2 g_1}\mathcal{F}_2\,\boldsymbol{u}\right]. \end{array} \right. \tag{2.9}$$

Remark 2.1. If N is even, we have to pay attention to the mode $\frac{N}{2}$ associated to the frequency $\frac{\eta N}{2}$. Indeed, we can easily verify that $\mathcal{S}_i^\alpha \mathbb{R}^{\mathbb{G}^2} \subset \mathbb{R}^{\mathbb{G}^2}$ (for $i = 1, 2$) if and only if N is odd or $\alpha \in \mathbb{Z}$.

Finally, the numerical solution $(\boldsymbol{u}^n)_{n\in\mathbb{N}}$ of the numerical schemes we consider are defined by (for $\delta_t \neq k\pi, k \in \mathbb{Z}^\star$)

$$\begin{aligned} \boldsymbol{u}^n &= (\mathcal{L}_{\delta_t})^n\, u^{in}_{|\mathbb{G}^2} := \left(\mathcal{S}_2^{\delta_t}\mathcal{S}_1^{-\delta_t}\right)^n\, u^{in}_{|\mathbb{G}^2} && \text{(Lie)}, \\ \boldsymbol{u}^n &= (\mathcal{T}_{\delta_t})^n\, u^{in}_{|\mathbb{G}^2} := \left(\mathcal{S}_1^{-\delta_t/2}\mathcal{S}_2^{\delta_t}\mathcal{S}_1^{-\delta_t/2}\right)^n\, u^{in}_{|\mathbb{G}^2} && \text{(Strang)}, \\ \boldsymbol{u}^n &= (\mathcal{M}_{\delta_t})^n\, u^{in}_{|\mathbb{G}^2} := \left(\mathcal{S}_1^{-\tan(\delta_t/2)}\mathcal{S}_2^{\sin(\delta_t)}\mathcal{S}_1^{-\tan(\delta_t/2)}\right)^n\, u^{in}_{|\mathbb{G}^2} && \text{(New)}, \end{aligned} \tag{2.10}$$

where $u^{in}_{|\mathbb{G}^2}$ is the evaluation of the initial condition u^{in} on the grid $\mathbb{G}^2$. The rest of this section performs a numerical analysis of the splittings defined in (2.10).

2.1. Numerical analysis. We define some associated discrete Lebesgue norms. They are defined for $\boldsymbol{u} \in \mathbb{C}^{\mathbb{G}^2}$ by

$$\|\,\boldsymbol{u}\,\|^2_{L^2(\mathbb{G}^2)} = h^2 \sum_{g\in\mathbb{G}^2} |\,\boldsymbol{u}_g\,|^2 \qquad \text{and} \qquad \|\,\boldsymbol{u}\,\|_{L^\infty(\mathbb{G}^2)} = \max_{g\in\mathbb{G}^2} |\,\boldsymbol{u}_g\,|.$$

We also use the Schwartz space $\mathscr{S}(\mathbb{R}^2)$ and introduce a scale of spaces, denoted $(X^s)_{s\geq 0}$, defined by

$$X^s = \left\{ u \in L^2(\mathbb{R}^2),\ \|u\|^2_{X^s} := \int \langle x\rangle^{2s} |u(x)|^2 \mathrm{d}x + \int \langle \xi\rangle^{2s} |\mathscr{F}u(\xi)|^2 \mathrm{d}\xi < \infty \right\},$$

where $\langle x\rangle := \sqrt{1+|x|^2}$ and $\mathscr{F}u$ denotes the Fourier transform of u. This scale of spaces is well designed to estimate the consistency error of the pseudospectral method since it controls both the localization and the smoothness of the functions.

2.1.1. Consistency. First, we prove that the pseudospectral shear transformations (2.8) and (2.9) are consistent with the continuous ones (2.7). Let us remark that in addition to the analysis of the spectral consistency, we will also pay attention to the truncation R. The consistency error of the pseudospectral shear transformations is stated in the following proposition for $\mathcal{S}_1^\alpha$ but the result is also valid for $\mathcal{S}_2^\alpha$.

PROPOSITION 2.2. *For all $s>1$ and for all $M>0$, there exists $c>0$ such that for all $u\in\mathscr{S}(\mathbb{R}^2)$, $\alpha\in(-M,M)$, $R>0$, and $N\in\mathbb{N}^*$ we have*

$$\|\mathcal{S}_1^\alpha\,\boldsymbol{u}-\boldsymbol{v}\,\|_{L^2(\mathbb{G}^2)}\le c\;|\alpha|\frac{R^{-s}+h^s}{\sqrt{h}}\|u\|_{X^{s+6}},$$

where $\boldsymbol{u}=u_{|\mathbb{G}^2}$ and $\boldsymbol{v}=v_{|\mathbb{G}^2}$ with $v(x)=u(x_1+\alpha x_2,x_2)$.

Here the norm $\|\cdot\|_{X^{s+6}}$ is not optimal; it would be possible to have the same result with a norm $\|\cdot\|_{X^{s+\nu}}$ where $0<\nu<6$. The relevant thing in this is that to have a consistency error of order s with respect to the stepsize of the grid and the length of the box (up to the $h^{-1/2}$ factor) it is enough to consider functions a bit more well localized and smooth than functions in X^s.

Proof of Proposition 2.2. Applying the discrete Fourier–Plancherel isometry, we get

$$\|\mathcal{S}_1^\alpha\,\boldsymbol{u}-\boldsymbol{v}\,\|_{L^2(\mathbb{G}^2)}^2=\frac{h\eta}{2\pi}\sum_{(\xi_1,g_2)\in\widehat{\mathbb{G}}\times\mathbb{G}}|\,(\mathcal{F}_1\mathcal{S}_1^\alpha\,\boldsymbol{u})_{\xi_1,g_2}-(\mathcal{F}_1\,\boldsymbol{v})_{\xi_1,g_2}\,|^2. \tag{2.11}$$

Thus, we are going to expand $\mathcal{F}_1\,\boldsymbol{v}$ and $\mathcal{F}_1\,\boldsymbol{u}$ with respect to u. More precisely, we apply the Poisson formula to get

$$\begin{aligned}\mathcal{F}_1\,\boldsymbol{u}&=h\sum_{g_1\in h\mathbb{Z}}u(g_1,g_2)e^{-i\xi_1g_1}-h\sum_{g_1\in\mathbb{G}^c}u(g_1,g_2)e^{-i\xi_1g_1}\\&=\mathscr{F}_1u(\xi_1,g_2)+\sum_{k\in\mathbb{Z}^*}\mathscr{F}_1u\left(\xi_1+\frac{2k\pi}{h},g_2\right)-h\sum_{g_1\in\mathbb{G}^c}u(g_1,g_2)e^{-i\xi_1g_1},\end{aligned}$$

where $\mathscr{F}_1u(\xi_1,x_2)=\int u(x)e^{-i\xi_1x_1}\mathrm{d}x_1$ is the continuous Fourier transform of u along the first direction and $\mathbb{G}^c=h\mathbb{Z}\setminus\mathbb{G}$. Consequently, since $\mathscr{F}_1v(\xi_1,x_2)=e^{i\alpha\xi_1x_2}\mathscr{F}_1u$, we decompose the consistency error into three terms,

$$\begin{aligned}(\mathcal{F}_1\mathcal{S}_1^\alpha\,\boldsymbol{u})_{\xi_1,g_2}-(\mathcal{F}_1\,\boldsymbol{v})_{\xi_1,g_2}&=\sum_{k\in\mathbb{Z}^*}\left(1-e^{i\alpha\frac{2k\pi}{h}g_2}\right)e^{i\alpha\xi_1g_2}\mathscr{F}_1u\left(\xi_1+\frac{2k\pi}{h},g_2\right)&&(\boldsymbol{\varepsilon}^1_{\xi_1,g_2})\\&\quad+h\sum_{g_1\in\mathbb{G}^c}\left(1-e^{i\alpha\xi_1g_2}\right)u(g_1,g_2)e^{-i\xi_1g_1}&&(\boldsymbol{\varepsilon}^2_{\xi_1,g_2})\\&\quad+h\sum_{g_1\in\mathbb{G}^c}\left[u(g_1+\alpha g_2,g_2)-u(g_1,g_2)\right]e^{-i\xi_1g_1}&&(\boldsymbol{\varepsilon}^3_{\xi_1,g_2}).\end{aligned}$$

Now we bound each one of these three consistency errors.

Estimation of $\boldsymbol{\varepsilon}^1$: First, we have

$$
\begin{aligned}
|\,\boldsymbol{\varepsilon}^1_{\xi_1,g_2}\,| &\le \sum_{k\in\mathbb{Z}^*}\left|\alpha\frac{2k\pi}{h}g_2\right|\left|\mathscr{F}_1 u\left(\xi_1+\frac{2k\pi}{h},g_2\right)\right| \\
&= \sum_{k\in\mathbb{Z}^*}\left|\alpha\frac{2k\pi}{h}g_2\right|\left|\xi_1+\frac{2k\pi}{h}\right|^{-s-1}\left|\mathscr{F}_1(|\partial_{x_1}|^{s+1}u)\left(\xi_1+\frac{2k\pi}{h},g_2\right)\right| \\
&\le \sum_{k\in\mathbb{Z}^*}\left|\alpha\frac{2k\pi}{h}g_2\right|\left(\frac{(2|k|-1)\pi}{h}\right)^{-s-1}\left|\mathscr{F}_1(|\partial_{x_1}|^{s+1}u)\left(\xi_1+\frac{2k\pi}{h},g_2\right)\right| \\
&\le \frac{2|\alpha|}{\sqrt{1+g_2^2}}\left(\frac{\pi}{h}\right)^{-s}\sum_{k\in\mathbb{Z}^*}|k|\,(2|k|-1)^{-s-1}\left\|\mathscr{F}_1\left((1+x_2^2)\langle\partial_{x_1}\rangle^{s+1}u\right)\right\|_{L^\infty(\mathbb{R}^2)} \\
&\le C^2\frac{4|\alpha|\zeta(s)}{\sqrt{1+g_2^2}}\left(\frac{\pi}{h}\right)^{-s}\left\|\langle i\partial_{x_2}\rangle\langle x_1\rangle\langle x_2\rangle^2\langle i\partial_{x_1}\rangle^{s+1}u\right\|_{L^2(\mathbb{R}^2)},
\end{aligned}
$$

where ζ denotes the Riemann function, $C>0$ is a universal constant associated with the Sobolev embedding $L^\infty(\mathbb{R})\to H^1(\mathbb{R})$, and $\langle i\partial_{x_2}\rangle$,$\langle i\partial_{x_1}\rangle^{s+1}$ are naturally defined as Fourier multipliers. This estimate involves a norm of u that is neither usual nor isotropic. Furthermore, the estimates of $\boldsymbol{\varepsilon}^2$ and $\boldsymbol{\varepsilon}^3$ will lead to some other norms of this kind. Consequently, in order to get an estimate as readable as possible, we control these norms by the X^{s+6} norm. Such a control can be realized with classical techniques of pseudodifferential calculus. However, it would require introducing many notions and notation, like the Weyl quantization, some classical classes of symbols, and parametrix. Thus, since these estimates are not really crucial here, we omit details. Nevertheless, these bounds could be obtained by directly applying Theorems 1.2.16, 1.3.6, and 1.4.1 of [24].

Now, we observe that by monotonicity we have

$$
h\sum_{g_2\in\mathbb{G}\setminus\{0\}}\frac{1}{1+g_2^2}\le\int_{\mathbb{R}}\frac{1}{1+y^2}\mathrm{d}y\le\pi. \tag{2.12}
$$

Thus, since $\boldsymbol{\varepsilon}^1_{\xi_1,0}=0$, there exists a constant $c>0$, depending only on s, such that

$$
\frac{h\eta}{2\pi}\sum_{(\xi_1,g_2)\in\widehat{\mathbb{G}}\times\mathbb{G}}|\,\boldsymbol{\varepsilon}^1_{\xi_1,g_2}\,|^2\le c|\alpha|^2R^{-1}h^{2s}(\#\widehat{\mathbb{G}})\|u\|_{X^{s+6}}\le c|\alpha|^2h^{2s-1}\|u\|^2_{X^{s+6}}.
$$

Estimation of $\boldsymbol{\varepsilon}^2$: First, naturally, we control $\boldsymbol{\varepsilon}^2$ by

$$
|\,\boldsymbol{\varepsilon}^2_{\xi_1,g_2}\,|\le\alpha|\xi_1||g_2|\left|h\sum_{g_1\in\mathbb{G}^c}u(g_1,g_2)e^{-i\xi_1g_1}\right|. \tag{2.13}
$$

In order to absorb the factor ξ_1 on the left, we perform a discrete integration by parts. So, we assume that $\xi_1\neq0$, and we denote $\xi_1=k_1\eta$ and $g_1=g_2=n_1h$, where $k_1\in[\![-\lfloor\frac{N-1}{2}\rfloor,\lfloor\frac{N}{2}\rfloor]\!]$ and $n_1\in\mathbb{Z}\setminus[\![-\lfloor\frac{N-1}{2}\rfloor,\lfloor\frac{N}{2}\rfloor]\!]$.

Then we introduce $N^+=1+\lfloor N/2\rfloor$ and $N^-=-1-\lfloor(N-1)/2\rfloor$ so that we have

$$h\sum_{g_1\in\mathbb{G}^c} u(g_1,g_2)e^{-i\xi_1 g_1} = h\sum_{n_1\geq N^+} u(g_1,g_2)e^{-\frac{2i\pi n_1 k_1}{N}} + h\sum_{n_1\leq N^-} u(g_1,g_2)e^{-\frac{2i\pi n_1 k_1}{N}}$$

$$= h\sum_{n_1\geq N^+} \frac{u(g_1,g_2)-u(g_1+h,g_2)}{h} h\sum_{n=N^+}^{n_1} e^{-\frac{2i\pi n k_1}{N}} \qquad (E_+)$$

$$+ h\sum_{n_1\leq N^-} \frac{u(g_1,g_2)-u(g_1-h,g_2)}{h} h\sum_{n=n_1}^{N^-} e^{-\frac{2i\pi n k_1}{N}} \qquad (E_-).$$

To control (E_+), first we observe that since $0\leq |k_1|\leq N/2$, we have

$$\left| h\sum_{n=N^+}^{n_1} e^{-\frac{2i\pi n k_1}{N}} \right| \leq \frac{2h}{|1-e^{-\frac{2i\pi k_1}{N}}|} \leq c\frac{hN}{2\pi|k_1|} = \frac{c}{|\xi_1|},$$

where c is a universal constant. Then, by application of the mean value theorem, we get

$$|E_+| \leq \frac{c}{|\xi_1|(1+g_2^2)} \left(\sup_{x\in\mathbb{R}^2} |(1+x_2^2)|x_1|^{s+1}\partial_{x_1}u(x)| \right) h\sum_{n_1\geq N^+} g_1^{s+1}$$

$$\leq \frac{c_s}{|\xi_1|(1+g_2^2)} \|u\|_{X^{s+6}} R^{-s} \frac{1}{N}\sum_{n_1\geq 0} \left(\frac{N^+ + n_1}{N}\right)^{s+1} \qquad (2.14)$$

$$\leq \frac{c_s}{|\xi_1|(1+g_2^2)} \|u\|_{X^{s+6}} R^{-s} \frac{1}{N}\sum_{n_1\geq 0} \left(\frac{1}{2}+\frac{n_1}{N}\right)^{s+1},$$

where $c_s>0$ is a constant depending only on s. We recognize a Riemann sum, so we have

$$\frac{1}{N}\sum_{n_1\geq 0}\left(\frac{1}{2}+\frac{n_1}{N}\right)^{s+1} \xrightarrow[N\to\infty]{} 2^{-s-2}\int_1^\infty y^{-s-1}\mathrm{d}y = \frac{2^{-s-2}}{s}.$$

In particular, since this sequence converges, it is bounded by a constant depending only on s. Thus, we obtain the following bound for $|E_+|$:

$$|E_+| \leq \frac{c_s}{|\xi_1|(1+g_2^2)} \|u\|_{X^{s+6}} R^{-s},$$

where c_s is another constant depending only on s. Note that, by symmetry, the same control holds for E^-.

Finally, coming back to (2.13) and using (2.12), we have another constant, denoted c_s, depending only on s such that

$$\frac{h\eta}{2\pi} \sum_{(\xi_1,g_2)\in\widehat{\mathbb{G}}\times\mathbb{G}} |\,\varepsilon^2_{\xi_1,g_2}\,|^2 \leq c_s|\alpha|^2 R^{-1}R^{-2s}(\#\widehat{\mathbb{G}})\|u\|^2_{X^{s+6}} \leq c_s|\alpha|^2 h^{-1}R^{-2s}\|u\|^2_{X^{s+6}}. \qquad (2.15)$$

Estimation of $\boldsymbol{\varepsilon^3}$: Applying the mean value theorem, for any $g_1,g_2,\alpha\in\mathbb{R}$, there exists $m_{g_1,g_2,\alpha}$ in $[g_1;g_1+\alpha g_2]$ such that

$$u(g_1+\alpha g_2,g_2)-u(g_1,g_2) = \alpha g_2 \partial_{x_1}u(m_{g_1,g_2,\alpha},g_2).$$

$$|\,\varepsilon^3_{\xi_1,g_2}\,| \le h \sum_{g_1\in\mathbb{G}^c} |\alpha||g_2||\partial_{x_1} u(m_{g_1,g_2,\alpha}, g_2)| \tag{2.16}$$
$$\le h \sum_{g_1\in\mathbb{G}^c} \frac{|\alpha|}{\sqrt{1+g_2^2}} \left|\binom{m_{g_1,g_2,\alpha}}{g_2}\right|^{-s-1} \sup_{x\in\mathbb{R}^2} |(1+x_2^2)|x|^{s+1}\partial_{x_1} u(x)|.$$

To control the norm of the two-dimensional vector in the previous estimate, we use the following technical lemma, whose proof is postponed to the end of this proof.

LEMMA 2.3. *If $y_1, y_2, y_3, \lambda \in \mathbb{R}$ are such that $y_3 \in [y_1; y_1+\lambda y_2]$, then we have*

$$\left|\binom{y_3}{y_2}\right| \ge \frac{|y_1|}{\sqrt{1+\lambda^2}}.$$

Consequently, applying Lemma 2.3 to (2.16), we get

$$|\,\varepsilon^3_{\xi_1,g_2}\,| \le c_s \frac{|\alpha|}{\sqrt{1+g_2^2}} \left|\frac{R}{2\sqrt{1+M^2}}\right|^{-s-1} \|u\|_{X^{s+6}} \left(h \sum_{g_1\in\mathbb{G}^c} \left|\frac{2g_1}{R}\right|^{-s-1}\right),$$

where c_s is a constant depending only on s.

Then, carrying out the same procedure as in (2.14), we get another constant c_s depending only on s such that

$$\left(h \sum_{g_1\in\mathbb{G}^c} \left|\frac{2g_1}{R}\right|^{-s-1}\right) \le c_s R.$$

Thus we have the estimate

$$|\,\varepsilon^3_{\xi_1,g_2}\,| \le c_{s,M} \frac{|\alpha|}{\sqrt{1+g_2^2}} R^{-s} \|u\|_{X^{s+6}},$$

where $c_{s,M}$ is a constant depending only on s and M. Consequently, we can repeat the same argument as (2.15) to get

$$\frac{h\eta}{2\pi} \sum_{(\xi_1,g_2)\in\widehat{\mathbb{G}}\times\mathbb{G}} |\,\varepsilon^3_{\xi_1,g_2}\,|^2 \le c_{s,M}|\alpha|^2 h^{-1} R^{-2s} \|u\|^2_{X^{s+6}},$$

where $c_{s,M}$ is another constant depending only on s and M.

We conclude by summing the different contributions of ε^1, ε^2, and ε^3. ∎

Let us prove Lemma 2.3.

Proof of Lemma 2.3. If $0 \in [y_1; y_1+\lambda y_2]$, then we have $|y_1| \le \lambda |y_2|$ and so we get

$$\left|\binom{y_3}{y_2}\right| \ge |y_2| \ge \frac{|y_1|}{|\lambda|} \ge \frac{|y_1|}{\sqrt{1+\lambda^2}}.$$

Else we have $|y_3| = |y_1|$ or $|y_3| = |y_1+\lambda y_2|$. If $|y_3| = |y_1|$, then we have

$$\left|\binom{y_3}{y_2}\right| \ge |y_3| = |y_1| \ge \frac{|y_1|}{\sqrt{1+\lambda^2}}.$$

Else if $|y_3| = |y_1 + \lambda y_2|$, we have

$$\left|\begin{pmatrix} y_3 \\ y_2 \end{pmatrix}\right|^2 = y_2^2 + (y_1 + \lambda y_2)^2.$$

This last quantity is a second-order polynomial with respect to y_2. Thus its infimum can be determined explicitly. More precisely, we have

$$y_2^2 + (y_1 + \lambda y_2)^2 \geq \frac{|y_1|^2}{1+\lambda^2}. \qquad \square$$

2.1.2. Backward error analysis. We aim at describing the long time behavior of the splitting methods. So, we perform a general backward error analysis[1] for a large class of methods including Lie and Strang splittings but also the new splitting. Note that since we deal with a linear problem the expansions are convergent. This is the goal of the next proposition.

PROPOSITION 2.4. *If $a, b \in \mathbb{R}$ satisfy $ab < 2$, then*

$$e^{bx_1\partial_{x_2}} e^{-ax_2\partial_{x_1}} = e^{JL_{a,b}x\cdot\nabla} \tag{2.17}$$

and

$$e^{-\frac{a}{2}x_2\partial_{x_1}} e^{bx_1\partial_{x_2}} e^{-\frac{a}{2}x_2\partial_{x_1}} = e^{JS_{a,b}x\cdot\nabla}, \tag{2.18}$$

where

$$L_{a,b} = \mu_{a,b}\begin{pmatrix} b & \frac{ab}{2} \\ \frac{ab}{2} & a \end{pmatrix} \qquad \text{and} \qquad S_{a,b} = \mu_{a,b}\begin{pmatrix} b & 0 \\ 0 & a(1-\frac{ab}{4}) \end{pmatrix} \tag{2.19}$$

with $\mu_{a,b} = F(ab(1-ab/4))$, where F is the continuous function on $(-\infty, 1]$ defined by

$$F(x) = \begin{cases} \frac{\arcsin(\sqrt{x})}{\sqrt{x}} & \text{if} \quad 0 < x \leq 1, \\ \frac{\operatorname{asinh}(\sqrt{-x})}{\sqrt{-x}} & \text{if} \quad x < 0, \\ 1 & \text{if} \quad x = 0. \end{cases}$$

Proof. Considering the transport equation (2.2) which can be solved with the method of characteristics, we have

$$e^{tAx\cdot\nabla} u_0 = u^{in} \circ e^{tA}.$$

Thus, noting that this formula maps an action on the left to an action on the right, (2.17) is equivalent to

$$\exp\begin{pmatrix} 0 & -a \\ 0 & 0 \end{pmatrix} \exp\begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix} = e^{JL_{a,b}},$$

with J given by (1.2). These exponentials of matrices can be written as shear transforms. So (2.17) is equivalent to

$$P_{a,b} := \begin{pmatrix} 1 & -a \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} = \begin{pmatrix} 1-ab & -a \\ b & 1 \end{pmatrix} = e^{JL_{a,b}}. \tag{2.20}$$

[1] The reader can refer to [20] for an overview on backward error analysis.

Similarly, (2.18) is equivalent to

$$\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = e^{JS_{a,b}}. \tag{2.21}$$

First, we prove that if (2.20) holds with $L_{a,b}$ given by (2.19), then (2.20) also holds with $S_{a,b}$ given by (2.19). Indeed, observing that a Lie splitting is always conjugate to the Strang splitting we have

$$\begin{aligned} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} &= \begin{pmatrix} 1 & a/2 \\ 0 & 1 \end{pmatrix} P_{a,b} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & a/2 \\ 0 & 1 \end{pmatrix} e^{JL_{a,b}} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

But $\left(\begin{smallmatrix} 1 & -a/2 \\ 0 & 1 \end{smallmatrix}\right)$ is symplectic, i.e.,

$$^{\mathrm{t}}\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} J \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = J.$$

Thus, since symplectic transforms map Hamiltonian systems on Hamiltonian systems (see, e.g., section VI of [20]), we have

$$\begin{pmatrix} 1 & a/2 \\ 0 & 1 \end{pmatrix} e^{JL_{a,b}} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = \exp\left(J \, ^{\mathrm{t}}\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} L_{a,b} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \right) = e^{JS_{a,b}},$$

where $S_{a,b}$ is given by (2.19).

So, now we aim at proving (2.20). First, the existence of such an $L_{a,b}$ is ensured by the following lemma (an elementary proof is given at the end of this proof).

LEMMA 2.5. *If a, b are small enough, there exists a symmetric matrix $L_{a,b}$ such that $L_{a,b}$ goes to 0 as (a, b) goes to 0 and $P_{a,b} = \exp(JL_{a,b})$.*

Then, we have to determine a formula for $L_{a,b}$. Since $L_{a,b}$ is a symmetric matrix, $e^{JL_{a,b}}$ is a Hamiltonian flow at time 1. A fortiori, $L_{a,b}$ is a constant of motion. So we have

$$^{\mathrm{t}}(e^{JL_{a,b}}) L_{a,b} e^{JL_{a,b}} = L_{a,b}.$$

But, by construction, $e^{JL_{a,b}} = P_{a,b}$, so $L_{a,b}$ is an eigenvector associated with the eigenvalue 1 of the linear application

$$R_{a,b} : \left\{ \begin{array}{lcl} S_2(\mathbb{R}) & \to & S_2(\mathbb{R}), \\ Q & \mapsto & {}^{\mathrm{t}}P_{a,b} Q P_{a,b}. \end{array} \right.$$

By a straightforward calculation, we observe that

$$Q_{a,b} = \begin{pmatrix} b & \frac{ab}{2} \\ \frac{ab}{2} & a \end{pmatrix} \text{ satisfies } R_{a,b}(Q_{a,b}) = Q_{a,b}. \tag{2.22}$$

However, the following lemma holds (it is proven at the end of this proof).

LEMMA 2.6. *If $0 < ab < 4$ the eigenspace of $R_{a,b}$ associated with the eigenvalue 1 is of dimension 1.*

Consequently, we deduce that if $0 < ab < 4$, then there exists $\mu_{a,b} \in \mathbb{R}$ such that

$$L_{a,b} = \mu_{a,b} Q_{a,b}. \tag{2.23}$$

Now, we just have to determine $\mu_{a,b}$. Since $L_{a,b}$ is symmetric, it is diagonalizable in an orthonormal basis, i.e.,

$$\exists \lambda \in \mathbb{R}^2, \exists \Omega \in O_2(\mathbb{R}),\ L_{a,b} = \Omega^{-1} \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \Omega = \Omega^{-1} D \Omega.$$

So, since J and Ω commute, we have

$$P_{a,b} = \Omega^{-1} e^{JD} \Omega.$$

Since we assume that $0 < ab < 4$, we deduce from (2.23) that $L_{a,b}$ is either positive or negative. In particular we have $\lambda_1 \lambda_2 > 0$. Thus we can define the symplectic matrix

$$K = \begin{pmatrix} \sqrt[4]{\lambda_1/\lambda_2} & \\ & \sqrt[4]{\lambda_2/\lambda_1} \end{pmatrix}.$$

This matrix satisfies $\sqrt{\lambda_1\lambda_2}\,{}^{\mathrm{t}}KK = D$ and $J\,{}^{\mathrm{t}}K = K^{-1}J$. Thus, we have

(2.24)
$$P_{a,b} = (K\Omega)^{-1} e^{\sqrt{\lambda_1\lambda_2} J} (K\Omega) = (K\Omega)^{-1} \begin{pmatrix} \cos\left(\sqrt{\lambda_1\lambda_2}\right) & -\sin\left(\sqrt{\lambda_1\lambda_2}\right) \\ \sin\left(\sqrt{\lambda_1\lambda_2}\right) & \cos\left(\sqrt{\lambda_1\lambda_2}\right) \end{pmatrix} (K\Omega).$$

In particular, we have

$$\operatorname{Tr} P_{a,b} = 2\cos\left(\sqrt{\lambda_1\lambda_2}\right) = 2\cos\left(\sqrt{\det L_{a,b}}\right) = 2\cos\left(\mu_{a,b}\sqrt{\det Q_{a,b}}\right).$$

As a consequence, since $\sqrt{\det L_{a,b}}$ goes to zero when (a,b) goes to 0, we deduce from a straightforward calculation that if ab is small enough, then

$$\mu_{a,b} = \pm F(ab(1 - ab/4)).$$

Finally, we have to determine the sign of $\mu_{a,b}$. First, observe that by continuity, we have either $\mu_{a,b} > 0$ for all a, b small enough satisfying $ab > 0$ or $\mu_{a,b} < 0$ for all a, b small enough satisfying $ab > 0$. This second case is excluded because when a goes to zero we have

$$e^{-F(a^2(1-a^2/4))JQ_{a,a}} = e^{-aJ+\mathcal{O}(a^2)} = P_{-a,-a} + \mathcal{O}(a^2) \neq P_{a,a} + \mathcal{O}(a^2).$$

To conclude, we have proved that if $ab > 0$ and (a,b) is small enough, then

$$P_{a,b} = e^{F(ab(1-ab/4))JQ_{a,b}}.$$

Furthermore, this relation is analytic with respect to a and b, so it can be extended to all $a, b \in \mathbb{R}$ such that $ab < 2$. Indeed, under this assumption we have $ab(1 - ab/4) \in (-\infty, 1)$, which is the domain of analyticity of F. ∎

Let us prove the two lemmas used within the proof.

Proof of Lemma 2.5. Notice that if a and b are small enough, $P_{a,b}$ defined in (2.20) is close to the identity. Consequently, it admits a logarithm $M_{a,b} \in M_2(\mathbb{R})$, defined by

$$M_{a,b} = \sum_{n\in\mathbb{N}} \frac{(-1)^n}{n+1} (P_{a,b} - I_2)^n$$

and satisfying

$$e^{M_{a,b}} = P_{a,b}.$$

A fortiori, we have $\exp(\operatorname{Tr} M_{a,b}) = \det P_{a,b} = 1$. Hence we have $\operatorname{Tr} M_{a,b} = 0$. Furthermore, the following application defines an isomorphism of vector spaces (it is an injection between two spaces of dimension 3):

$$\begin{cases} S_2(\mathbb{R}) & \to & \mathfrak{sl}_2(\mathbb{R}), \\ L & \mapsto & JL, \end{cases}$$

where $\mathfrak{sl}_2(\mathbb{R}) = \{M \in M_2(\mathbb{R}) \mid \operatorname{Tr} M = 0\}$. As a consequence, there exists a symmetric matrix $L_{a,b} \in S_2(\mathbb{R})$ such that

$$M_{a,b} = JL_{a,b}. \qquad \square$$

Proof of Lemma 2.6. Since $0 < ab < 4$, $Q_{a,b}$ is either positive or negative, and, as a consequence, the following Euclidean norm is well defined on $S_2(\mathbb{R})$:

$$\forall K \in S_2(\mathbb{R}),\ \|K\|_{a,b}^2 := \int_{\mathbb{R}^2} ({}^{\mathrm{t}}xKx)^2 e^{-|\,{}^{\mathrm{t}}xQ_{a,b}x|}\mathrm{d}x.$$

Since $\det P_{a,b} = 1$, computing $\|R_{a,b}^{-1}K\|_{a,b}$, we deduce from a change of variables and from (2.22) that

$$\forall K \in S_2(\mathbb{R}),\ \|R_{a,b}K\|_{a,b} = \|K\|_{a,b}.$$

This relation means that $R_{a,b}$ is an isometry for the Euclidean norm $\|\cdot\|_{a,b}$. A fortiori, we have $\det R_{a,b} = \pm 1$. But, since $R_{0,0} = I_2$ and $(a,b) \mapsto \det R_{a,b}$ is a continuous map, we deduce that $\det R_{a,b} = 1$. Consequently, $R_{a,b}$ is a rotation in a space of dimension 3. So, there are only two possibilities: either $R_{a,b}$ is the identity or the eigenspace of $R_{a,b}$ associated with the eigenvalue 1 is of dimension 1.

To conclude, we just have to verify that $R_{a,b}$ is not the identity. First, we observe that $P_{a,b}$ is not a scalar matrix, so there exists $x \in \mathbb{R}^2$ such that x is not an eigenvector of $P_{a,b}$. Then, we consider a vector $y \in \mathbb{R}^2 \setminus \{0\}$ such that x and y are orthogonal. By construction, we have

$${}^{\mathrm{t}}yP_{a,b}x \neq 0.$$

Consequently, if $K = y\,{}^{\mathrm{t}}y \in S_2(\mathbb{R})$, we have

$${}^{\mathrm{t}}xR_{a,b}(K)x = ({}^{\mathrm{t}}yP_{a,b}x)^2 \neq 0 = ({}^{\mathrm{t}}yx)^2 = {}^{\mathrm{t}}xKx.$$

Thus, we have $R_{a,b}(K) \neq K$. $\square$

The classical splitting formulas of Lie and Strang correspond to the choice $a = b = \delta_t$ in (2.17) and (2.18). However, as mentioned in the introduction, these choices are not necessarily the best. For the Strang like splittings, a straigthforward calculation proves that there exists an optimal choice for which the splitting is exact. This choice can be obtained by direct formal calculations by assuming a decomposition of the rotation matrix and we then can write the following lemma.

LEMMA 2.7. *If $\delta_t \in (-\pi, \pi)$ then we have*

$$S_{2\tan(\delta_t/2),\sin(\delta_t)} = \delta_t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that due to the nondiagonal terms of $L_{a,b}$, it is impossible to design an exact splitting based on the Lie splitting.

2.1.3. Convergence. We now consider the convergence of the pseudospectral splittings (2.10) to approximate our problem (2.1). Then, for a discrete initial condition $\boldsymbol{u} = u^{in}_{|\mathbb{G}^2}$, the numerical solution at time $t_n = n\delta_t$ ($n \in \mathbb{N}$) is given by n compositions of the operators defined in (2.10). For instance, for the standard Strang splitting, the numerical solution at time t_n is $(\mathcal{T}_{\delta_t})^n \boldsymbol{u}$. In the following theorem, we show that, up to a spectral spatial error, the dynamics generated by the Strang (resp., Lie) pseudospectral method $\mathcal{T}_{\delta_t}$ (resp., $\mathcal{L}_{\delta_t}$) corresponds to the exact solution of some modified equations over very long times.

THEOREM 2.8. *For all $s > 0$ there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$, and all $\delta_t \in [-1, 1]$ with $t_n = n\delta_t$, we have*

$$\left\| (\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} \right\|_{L^2(\mathbb{G}^2)} \leq c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}}$$

and

$$\left\| (\mathcal{T}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{T}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} \right\|_{L^2(\mathbb{G}^2)} \leq c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}},$$

where $\boldsymbol{u} = u_{|\mathbb{G}^2}$, $S^{\mathcal{L}}_{\delta_t} := L_{\delta_t,\delta_t}/\delta_t = I_2 + \mathcal{O}(\delta_t)$, and $S^{\mathcal{T}}_{\delta_t} := S_{\delta_t,\delta_t}/\delta_t = I_2 + \mathcal{O}(\delta_t^2)$. The definitions of $S_{a,b}$ and $L_{a,b}$ are given by (2.19) *in Proposition* 2.4*, whereas $\mathcal{L}_{\delta_t}$ and $\mathcal{T}_{\delta_t}$ are given by* (2.10).

Proof. We focus only on proving the convergence estimate for the Lie splitting. The same proof could be applied to prove the estimate for the Strang splitting.

Let $\boldsymbol{\varepsilon}_n \in L^2(\mathbb{G}^2)$ be the consistency error at time t_n. It is defined by

$$\boldsymbol{\varepsilon}_n = \mathcal{L}_{\delta_t} \left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} - \left(e^{t_{n+1} J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2}.$$

As usual, for linear schemes, the convergence error is given by

$$(\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} = \sum_{k=0}^{n-1} \mathcal{L}_{\delta_t}^{n-1-k} \boldsymbol{\varepsilon}_k .$$

Here, $\mathcal{L}_{\delta_t}$ is an isometry of $L^2(\mathbb{G}^2)$, so we have

$$\left\| (\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} \right\|_{L^2(\mathbb{G}^2)} \leq \sum_{k=0}^{n-1} \| \boldsymbol{\varepsilon}_k \|_{L^2(\mathbb{G}^2)} \leq n \sup_{k\in\mathbb{N}} \| \boldsymbol{\varepsilon}_k \|_{L^2(\mathbb{G}^2)}.$$

Thus, we just have to bound $\boldsymbol{\varepsilon}_k$. Using formulas of Proposition 2.4, we decompose $\boldsymbol{\varepsilon}_k$ into two consistency errors for the pseudospectral shear transformations:

$$\begin{aligned}\boldsymbol{\varepsilon}_k = \mathcal{S}_2^{\delta_t} \Big[\mathcal{S}_1^{-\delta_t} \left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} - \left(e^{-\delta_t x_2 \partial_{x_1}} e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} \Big] \\ + \mathcal{S}_2^{\delta_t} \left(e^{-\delta_t x_2 \partial_{x_1}} e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2} - \left(e^{\delta_t x_1 \partial_{x_2}} e^{-\delta_t x_2 \partial_{x_1}} e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \right)_{|\mathbb{G}^2}.\end{aligned}$$

Then applying Proposition 2.2 and using that $\mathcal{S}_2^{\delta_t}$ is an isometry on $L^2(\mathbb{G}^2)$, we get a constant $c > 0$, depending only on $s > 0$ such that

(2.25)

$$\| \boldsymbol{\varepsilon}_k \|_{L^2(\mathbb{G}^2)} \leq c\, |\delta_t| \frac{R^{-s} + h^s}{\sqrt{h}} \left(\| e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \|_{X^{s+6}} + \| e^{-\delta_t x_2 \partial_{x_1}} e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u \|_{X^{s+6}} \right).$$

Now, we introduce a natural lemma to control these norms, whose proof is given at the end of the proof.

LEMMA 2.9. *For all $\kappa > 0$ and all $s > 0$, there exists a constant $c > 0$ such that if $\tau \in GL_2(\mathbb{R})$ satisfies*

$$\forall x \in \mathbb{R}^2,\ \kappa^{-1}|x| \le |\tau(x)| \le \kappa|x|, \tag{2.26}$$

then for all $u \in \mathscr{S}(\mathbb{R}^2)$ we have

$$\|u \circ \tau\|_{X^s} \le c\ \|u\|_{X^s}.$$

Since if $A \in M_2(\mathbb{R})$, then $e^{(Ax\cdot\nabla)}u = u^{in} \circ e^A$, to get an estimate of the two norms of (2.25) by $\|u\|_{X^{s+6}}$ using Lemma 2.9 and to conclude this proof, we just have to get estimates of the form (2.26) for $\tau = \left(\begin{smallmatrix} 1 & -\delta_t \\ 0 & 1 \end{smallmatrix}\right)$ and $\tau = \exp(tJS^{\mathcal{L}}_{\delta_t})$, uniformly with respect to $t \in \mathbb{R}$ and δ_t satisfying $|\delta_t| \le 1$.

On the one hand, since $\left(\begin{smallmatrix} 1 & -\delta_t \\ 0 & 1 \end{smallmatrix}\right)^{-1} = \left(\begin{smallmatrix} 1 & \delta_t \\ 0 & 1 \end{smallmatrix}\right)$ and $\delta_t \in [-1,1]$ which is compact, by continuity, we get $\kappa > 0$ such that

$$\forall \delta_t \in [-1,1], \forall x \in \mathbb{R}^2,\ \kappa^{-1}|x| \le \left|\begin{pmatrix} 1 & \delta_t \\ 0 & 1 \end{pmatrix} x\right| \le \kappa|x|.$$

On the other hand, we observe that the quadratic form associated with $S^{\mathcal{L}}_{\delta_t}$ is a constant of the motion of $\exp(tJS^{\mathcal{L}}_{\delta_t})$: for all $t \in \mathbb{R}$ and all $\delta_t \in [-1,1]$ we have

$$\forall x \in \mathbb{R}^2,\ {}^{\mathrm{t}}\left(e^{tJS^{\mathcal{L}}_{\delta_t}}x\right) S^{\mathcal{L}}_{\delta_t} e^{tJS^{\mathcal{L}}_{\delta_t}}x = {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x. \tag{2.27}$$

Furthermore, for all $\delta_t \in [-1,1]$, we have

$$\det S^{\mathcal{L}}_{\delta_t} = \delta_t^{-2} \arcsin^2\left(\sqrt{\delta_t^2(1-\delta_t^2/4)}\right) > 0.$$

So, $S^{\mathcal{L}}_{\delta_t}$ is either positive or negative. Thus, since $(S^{\mathcal{L}}_{\delta_t})_{1,1} > 0$, it is positive. Then, since $\delta_t \mapsto S^{\mathcal{L}}_{\delta_t}$ is a continuous map, $S^{\mathcal{L}}_{\delta_t}$ and $(S^{\mathcal{L}}_{\delta_t})^{-1}$ are bounded uniformly with respect to $\delta_t \in [-1,1]$. Consequently, there exists $\kappa > 0$ such that for all $\delta_t \in [-1,1]$ and all $x \in \mathbb{R}^2$ we have

$$\kappa^{-1}\ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x \le \kappa^{-1}|S^{\mathcal{L}}_{\delta_t}||x|^2 \le |x|^2 \le \kappa|(S^{\mathcal{L}}_{\delta_t})^{-1}|^{-1}|x|^2 \le \kappa\ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x.$$

Thus we deduce of (2.27) that for all $t \in \mathbb{R}$, all $\delta_t \in [-1,1]$ and all $x \in \mathbb{R}^2$ we have

$$\left|e^{tJS^{\mathcal{L}}_{\delta_t}}x\right|^2 \le \kappa\ {}^{\mathrm{t}}\left(e^{tJS^{\mathcal{L}}_{\delta_t}}x\right) S^{\mathcal{L}}_{\delta_t} e^{tJS^{\mathcal{L}}_{\delta_t}}x = \kappa\ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x \le \kappa^2|x|^2$$

and

$$\left|e^{tJS^{\mathcal{L}}_{\delta_t}}x\right|^2 \ge \kappa^{-1}\ {}^{\mathrm{t}}\left(e^{tJS^{\mathcal{L}}_{\delta_t}}x\right) S^{\mathcal{L}}_{\delta_t} e^{tJS^{\mathcal{L}}_{\delta_t}}x = \kappa^{-1}\ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x \ge \kappa^{-2}|x|^2. \qquad \square$$

Proof of Lemma 2.9. We have to bound $\|\langle x\rangle^s(u\circ\tau)\|_{L^2(\mathbb{R}^2)}$ and $\|\langle\xi\rangle^s\mathscr{F}(u\circ\tau)\|_{L^2(\mathbb{R}^2)}$. However, a straightforward calculation shows that

$$\mathscr{F}(u \circ \tau) = |\det\tau|^{-1}\,(\mathscr{F}u) \circ {}^t\tau^{-1},$$

and (2.26) is clearly equivalent to

$$|\tau| \le \kappa \text{ and } |\tau^{-1}| \le \kappa.$$

Thus, since $|\tau| = |{}^t\tau|$, if we get a bound on $\|\langle x\rangle^s(u\circ\tau)\|_{L^2(\mathbb{R}^2)}$, uniform with respect to τ, we also get a bound on $\|\langle\xi\rangle^s\mathscr{F}(u\circ\tau)\|_{L^2(\mathbb{R}^2)}$ uniform with respect to τ.

Finally, to bound $\|\langle x\rangle^s(u\circ\tau)\|_{L^2(\mathbb{R}^2)}$, we just have to apply a change of coordinates:

$$\|\langle x\rangle^s(u\circ\tau)\|_{L^2(\mathbb{R}^2)}=\sqrt{|\det\tau|^{-1}}\|\langle\tau(x)\rangle^s u\|_{L^2(\mathbb{R}^2)}\leq\sqrt{|\det\tau^{-1}|}\langle\tau\rangle^s\|u\|_{X^s}\leq\kappa^{s+1}\|u\|_{X^s}.$$

∎

As a corollary of Theorem 2.8, we deduce the convergence error of these methods.

COROLLARY 2.10. *For all $s>0$ and all $h_0>0$, there exists $c>0$ such that for all $N\in\mathbb{N}^*$, all $R>0$, all $u\in\mathscr{S}(\mathbb{R}^2)$, all $n\in\mathbb{N}$, and all $\delta_t\in[-1,1]$ and $h=R/N\leq h_0$, denoting $t_n=n\delta_t$, we have*

$$\left\|(\mathcal{L}_{\delta_t})^n\,\boldsymbol{u}-\left(e^{t_nJx\cdot\nabla}u\right)_{|\mathbb{G}^2}\right\|_{L^2(\mathbb{G}^2)}\leq c\;t_n\frac{R^{-s}+h^s}{\sqrt{h}}\|u\|_{X^{s+6}}+c\;|e^{t_nJ}-e^{t_nJS^{\mathcal{L}}_{\delta_t}}|\|u\|_{X^4}$$

and

$$\left\|(\mathcal{T}_{\delta_t})^n\,\boldsymbol{u}-\left(e^{t_nJx\cdot\nabla}u\right)_{|\mathbb{G}^2}\right\|_{L^2(\mathbb{G}^2)}\leq c\;t_n\frac{R^{-s}+h^s}{\sqrt{h}}\|u\|_{X^{s+6}}+c\;|e^{t_nJ}-e^{t_nJS^{\mathcal{T}}_{\delta_t}}|\|u\|_{X^4},$$

where $\boldsymbol{u}=u_{|\mathbb{G}^2}$, $\mathcal{L}_{\delta_t}$ and $\mathcal{T}_{\delta_t}$ are given by (2.10)

Before proving this result, we introduce the following technical lemma.

LEMMA 2.11. *There exists a universal constant $c>0$ such that for all $v\in H^2(\mathbb{R}^2)$, all $R>0$, and all $N\in\mathbb{N}^*$ we have*

$$\|v_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)}\leq\|u\|_{L^2(\mathbb{R}^2)}+c\;h^2\|\Delta u\|_{L^2(\mathbb{R}^2)}.$$

Proof. First, we apply the Poisson formula and the discrete Fourier–Plancherel isometry to get

$$\|v_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)}\leq\|v_{|h\mathbb{Z}^2}\|_{L^2(h\mathbb{Z}^2)}=\frac{1}{2\pi}\left\|\sum_{k\in\mathbb{Z}^2}\mathscr{F}v\left(\cdot+\frac{2k\pi}{h}\right)\right\|_{L^2((-\frac{\pi}{h},\frac{\pi}{h})^2)}.$$

Then we observe that if $k\in\mathbb{Z}^2\setminus\{0\}$ and $\xi\in(-\frac{\pi}{h},\frac{\pi}{h})^2$, then

$$\left|\xi+\frac{2k\pi}{h}\right|\geq\frac{\pi}{h}(2|k|-\sqrt{2}).$$

Thus, we control $\|v_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)}$ by

$$\frac{1}{2\pi}\|\mathscr{F}v\|_{L^2((-\frac{\pi}{h},\frac{\pi}{h})^2)}+\frac{h^2}{2\pi^3}\sum_{k\in\mathbb{Z}^2\setminus\{0\}}\frac{1}{(2|k|-\sqrt{2})^2}\|(|\xi|^2\mathscr{F}v)\left(\cdot\;+\frac{2k\pi}{h}\right)\|_{L^2((-\frac{\pi}{h},\frac{\pi}{h})^2)}.$$

Finally, applying the Cauchy–Schwarz inequality and the Chasles relation, we control the second term by

$$\frac{h^2}{2\pi^3}\||\xi|^2\mathscr{F}v\|_{L^2(\mathbb{R}^2)}\sqrt{\sum_{k\in\mathbb{Z}^2\setminus\{0\}}\frac{1}{(2|k|-\sqrt{2})^4}}.$$

∎

Proof of Corollary 2.10. We only focus on proving the convergence estimate for the Lie splitting, the case of the Strang splitting being similar. Applying Theorem 2.8 and the triangle inequality, we have

$$\|(\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leq c\, t_n \frac{R^{-s}+h^s}{\sqrt{h}}\|u\|_{X^{s+6}} + E_{u,\delta_t,n,\mathbb{G}}$$

with

$$E_{u,\delta_t,n,\mathbb{G}} = \left\| \left(e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla}u - e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\right\|_{L^2(\mathbb{G}^2)}.$$

Consequently, to prove the corollary, we just have to bound $E_{u,\delta_t,n,\mathbb{G}}$ by $|e^{t_n J} - e^{t_n JS^{\mathcal{L}}_{\delta_t}}|\|u\|_{X^4}$. Since $h \leq h_0$, applying Lemma 2.11, we get a constant $c>0$, depending only on $h_0>0$, such that

$$E_{u,\delta_t,n,\mathbb{G}} \leq c\, \left\|(1-\Delta)\left(e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla}u - e^{t_n Jx\cdot\nabla}u\right)\right\|_{L^2(\mathbb{R}^2)}.$$

Then applying the Fourier–Plancherel isometry, we get

$$E_{u,\delta_t,n,\mathbb{G}} \leq \frac{c}{2\pi}\left\|(1+|\xi|^2)\left(\mathscr{F}u\circ {}^{\mathrm{t}}\left(e^{-t_n JS^{\mathcal{L}}_{\delta_t}}\right) - \mathscr{F}u\circ {}^{\mathrm{t}}\left(e^{-t_n J}\right)\right)\right\|_{L^2(\mathbb{R}^2)}.$$

Then introducing a Taylor remainder under its integral form, we have

$$\begin{aligned} E_{u,\delta_t,n,\mathbb{G}} &\leq \left\|(1+|\xi|^2)\int_0^1 \nabla_\xi \mathscr{F}u(y_{\alpha,\xi,n,\delta_t})\cdot {}^{\mathrm{t}}\left(e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}\right)\xi \,\mathrm{d}\alpha\right\|_{L^2(\mathbb{R}^2)} \\ &\leq |e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}| \max_{\alpha\in(0,1)} \|(1+|\xi|^2)^{3/2}\nabla_\xi \mathscr{F}u(y_{\alpha,\xi,n,\delta_t})\|_{L^2(\mathbb{R}^2)}, \end{aligned} \tag{2.28}$$

where $y_{\alpha,\xi,n,\delta_t} = {}^{\mathrm{t}}M_{\alpha,n,\delta_t}\xi$ and $M_{\alpha,n,\delta_t} = I_2 - \alpha(e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J})$.

Now, we distinguish two cases. If $|e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}| \leq 1/2$, then we deduce $|M_{\alpha,n,\delta_t} - I_2| \leq \frac{1}{2}$. Consequently, we have

$$|\det M_{\alpha,n,\delta_t}| \geq \kappa \text{ and } |M^{-1}_{\alpha,n,\delta_t}| \leq 2,$$

where κ is a universal constant.

Thus, by performing a natural change of coordinates, we get

$$\begin{aligned} E_{u,\delta_t,n,\mathbb{G}} &\leq \frac{|e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|}{\kappa}\left\|(1+|\,{}^{\mathrm{t}}M_{\alpha,n,\delta_t}{}^{-1}\xi|^2)^{3/2}\nabla_\xi\mathscr{F}u\right\|_{L^2(\mathbb{R}^2)} \\ &\leq 8\frac{|e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|}{\kappa}\left\|(1+|\xi|^2)^{3/2}\nabla_\xi\mathscr{F}u\|_{L^2(\mathbb{R}^2)} \leq c|e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|\right\| u\|_{X^4}, \end{aligned}$$

where $c>0$ is a universal constant.

Finally, we have to consider the case where $|e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}| \geq 1/2$. Applying Lemma 2.11 and the Fourier–Plancherel isometry we get two constant $c,\kappa>0$ depending only on h_0 such that

$$\begin{aligned} E_{u,\delta_t,n,\mathbb{G}} &\leq \|\,\boldsymbol{u}\,\|_{L^2(\mathbb{G}^2)} + \left\|\left(e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\right\|_{L^2(\mathbb{G}^2)} \leq c\|(1-\Delta)u\|_{L^2(\mathbb{R})} \\ &\leq \kappa|e^{-t_n JS^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|\|u\|_{X^4}. \end{aligned}$$

∎

Next, we focus on the new splitting $\mathcal{M}_{\delta_t}$. We provide a theorem showing that its dynamics corresponds, up to a spectral spatial error, to the rotation with the exact speed, for very long times.

THEOREM 2.12. *For all $s, \nu > 0$ there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$, and all $\delta_t \in \mathbb{R}$ satisfying $|\delta_t| < \pi - \nu$, denoting $t_n = n\delta_t$, we have*

$$\left\| (\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J x\cdot\nabla} u\right)_{|\mathbb{G}^2} \right\|_{L^2(\mathbb{G}^2)} \leq c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}}, \tag{2.29}$$

where $\boldsymbol{u} = u_{|\mathbb{G}^2}$, and $\mathcal{M}_{\delta_t}$ is given by (2.10).

Proof. By carrying out the same proof as in Theorem 2.8, we could easily prove that for all $s, \nu > 0$ there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$, and all $\delta_t \in \mathbb{R}$ satisfying $|\delta_t| < \pi - \nu$, denoting $t_n = n\delta_t$, we have

$$\left\| (\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{M}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2} \right\|_{L^2(\mathbb{G}^2)} \leq c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}},$$

where $\boldsymbol{u} = u_{|\mathbb{G}^2}$ and $S^{\mathcal{M}}_{\delta_t} := S_{2\tan(\delta_t/2),\sin(\delta_t)}/\delta_t$, where $S_{a,b}$ is given by (2.19).

Thus, to conclude this proof, we just have to observe that by Lemma 2.7 we have $S_{2\tan(\delta_t/2),\sin(\delta_t)} = \delta_t I_2$. ∎

Remark 2.13. For all $\boldsymbol{u} \in L^2(\mathbb{G}^2)$ we have $\|\boldsymbol{u}\|_{L^\infty(\mathbb{G}^2)} \leq h^{-1} \|\boldsymbol{u}\|_{L^2(\mathbb{G}^2)}$; thus (2.29) gives a control of convergence error with the discrete L^∞ norm for very long times:

$$\left\| (\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J x\cdot\nabla} u\right)_{|\mathbb{G}^2} \right\|_{L^\infty(\mathbb{G}^2)} \leq c\, t_n \frac{R^{-s} + h^s}{h^{3/2}} \|u\|_{X^{s+6}}.$$

2.2. Numerical illustrations. In this subsection, we intend to illustrate the different results obtained previously, namely the spatial accuracy of the pseudospectral method and the time accuracy of the time splitting.

Spatial accuracy. First, we present some numerical results to illustrate the estimates obtained in Proposition 2.2. To do so, we consider the function

$$u(x) = \exp\left(-\frac{|x|^2}{2}\right), \quad x = (x_1, x_2) \in \mathbb{R}^2,$$

which is shifted by $\alpha = 0.01$. We denote $v_{|\mathbb{G}^2}$, where $v(x) = u(x_1 + \alpha x_2, x_2)$ the exact shifted solution, and we compute the (discrete) L^2 norm of the difference between $\mathcal{S}_1^\alpha u_{|\mathbb{G}^2}$ and $v_{|\mathbb{G}^2}$. The spatial grid $\mathbb{G}^2$ is defined by (2.6), where $h = R/N$, $R = 15$, and different values of N are considered to check the spatial accuracy. The results are displayed in Figure 1. One can observe that for large h (or small N), the term R^{-s} is negligible and the term h^s gives the exponential decreasing of the error which is the typical behavior of spectral methods. On the contrary, for very small values of h (or large values of N), the term $R^{-s}/h^{-1/2}$ becomes prominent even if the error is quite small (around 10^{-11}).

Time accuracy. In this part, we give some numerical illustrations of the efficiency of the new splitting. To do so, we consider the equation

$$\partial_t u = Jx \cdot \nabla_x u, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{2.30}$$

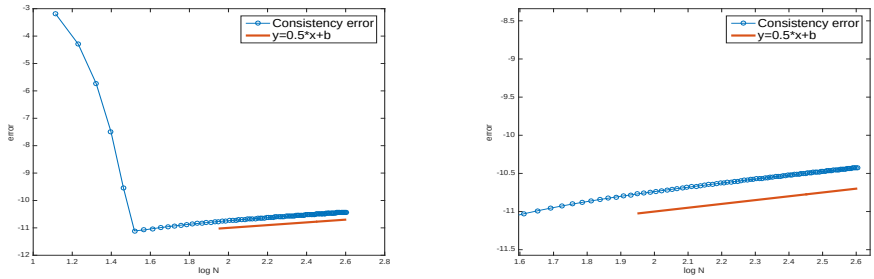

FIG. 1. *Spatial error (*log-log *scale) as a function of the number of points N between the exact shifted solution and the numerical approximation. The right figure is a zoom.*

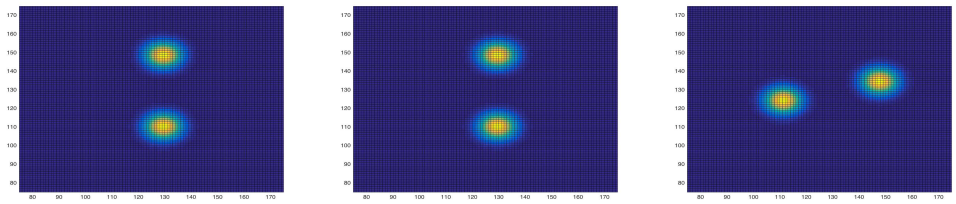

FIG. 2. *Solution $u(T,x)$ of* (2.30)*. Left: Exact solution $u(T,x)$. Middle: Numerical solution obtained by the new splitting. Right: Numerical solution obtained by the Strang splitting.*

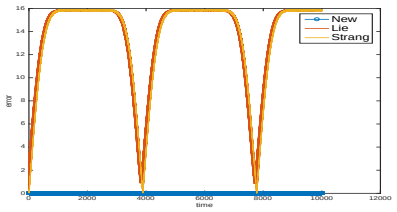

FIG. 3. *Time history of the relative errors between the solution of* (2.30) *and the numerical solution obtained by the new splitting, the Lie splitting, and the Strang splitting.*

with the initial condition

$$u^{in}(x) = \frac{1}{2\pi\beta}\left[\exp\left(-\frac{(x_1-0.3)^2}{\beta}\right)+\exp\left(-\frac{(x_1+0.3)^2}{\beta}\right)\right]\exp\left(-\frac{x_2^2}{\beta}\right),$$

with $\beta = 0.01$. The spatial truncated domain $[-2,2]^2$ is discretized with the grid $\mathbb{G}^2$ defined by (2.6) with $R = 4$ and a space step $h = R/N$, $N = 243 = 3^5$. The time step is $\delta_t \approx 0.139$ and the final time is $T = 10^5$ (the number of iterations is 71888). In Figures 2–5, some results are displayed where we compare the exact solution, the solution given by $(\mathcal{T}_{\delta_t})^n\, u^{in}_{|\mathbb{G}^2}$ (Strang splitting and spectral interpolation), the solution given by $(\mathcal{L}_{\delta_t})^n\, u^{in}_{|\mathbb{G}^2}$ (Lie splitting and spectral interpolation), and the solution given by the new method $(\mathcal{M}_{\delta_t})^n\, u^{in}_{|\mathbb{G}^2}$ (see (2.10)). First, in Figure 2, the three solutions are plotted at the final time. We can observe that the exact solution and the solution obtained by the new method are very close, whereas the solution obtained by the Strang splitting is not good due to the fact that the angular velocity of the Strang method is not exact. To make precise these observations, we plot in Figure 3 (Figure 4

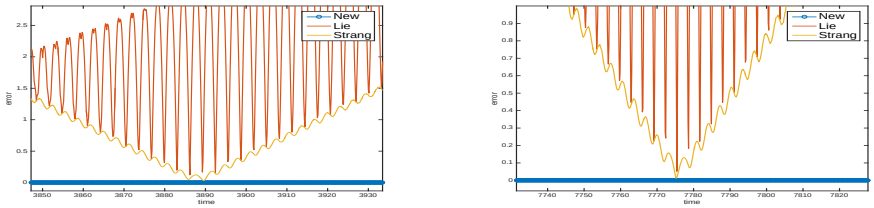

FIG. 4. *Time history of the relative errors (zoom of Figure* 3 *around* $\bar{T} \approx 3188$ *(left) and* $\bar{T} \approx 2 \times 3188$ *(right)).*

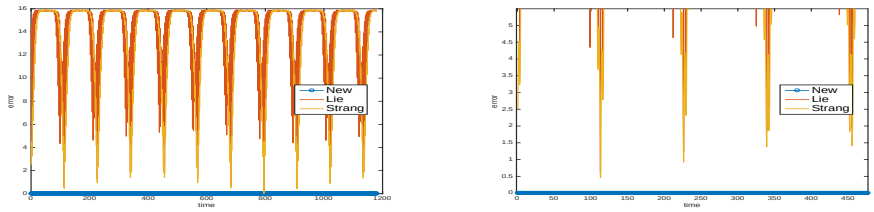

FIG. 5. *Time history of the relative errors between the solution of* (2.30) *and the numerical solution obtained by the new splitting, the Lie splitting, and the Strang splitting with* $\delta_t = \pi/4$. *The right figure is a zoom of the left one around* $k\bar{T}$ *with* $\bar{T} \approx 113, k = 1, 2, 3, 4$.

is a zoom) the relative L^∞ error of the new method and the Strang and the Lie methods. The error produced by the new method is close to 10^{-13}, which is the spectral error. On the contrary, the Strang and Lie methods periodically produce an error of order one. This is due to its wrong angular velocity: the solution moves away from the exact solution producing large error and at some times, the Strang method recovers the exact solution so that the error becomes very small. These times can be computed from the above analysis. Indeed, from the rotation speed of the Strang splitting $\omega_{\delta_t} = \frac{\arcsin(\sqrt{\delta_t^2(1-\delta_t^2/4)})}{\delta_t}$, we deduce that the exact solution (which rotates with a speed $\omega_{ex} = 1$) and the numerical solution obtained by the Strang method will coincide every times $\bar{T}$ such that $t^n + \omega_{ex}\bar{T} = t^n + \omega_{\delta_t}\bar{T}$ $[\pi]$ (the factor π (instead of a factor 2π) is due to our choice of a symmetric initial condition). Then, we have $\bar{T} = \pi/(\omega_{\delta_t} - 1)$, which gives with our choice of time step $\delta_t \approx 0.139$, $\bar{T} \approx 3888$. We observe a very good agreement in Figures 3, 4, and 5 (for which $\delta_t = \pi/4$ and then $\bar{T} \approx 113$).

Finally, we study the performance of the new method. Indeed, we compare the new splitting and a direct two-dimensional solution of (2.30). The direct resolution is done by a semi-Lagrangian type strategy: at each time step, we first compute exactly the feet of the characteristics equations and we then use a two-dimensional spectral interpolation by means of the nonuniform fast Fourier transform (the so-called nufft procedure introduced in [19]). We checked that this approach also leads to spectral accuracy, and we want here to compare the two spectral methods in terms of CPU time with respect to the total number of points N^2 (N being the number of points per direction). The results are displayed in Figure 6: the time execution (for 10 iterations) for both methods (new splitting and nufft) as a function of N^2 (for $N = 2^5, \ldots, 2^{11}$), in log–log scale. Let us mention that these runs have been conducted in a serial way (in particular using FFT libraries of Julia), on a machine whose characteristics are

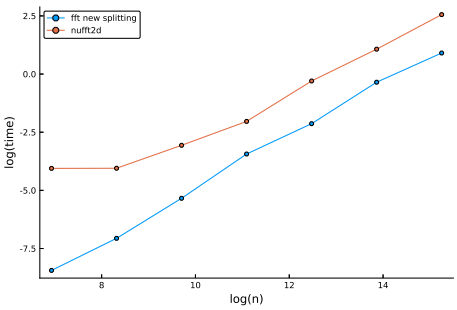

FIG. 6. *Time execution as a function of the total number of points (*log–log *scale). Blue: new method (new splitting and one-dimensional fast Fourier transform). Red: exact computation of the feet of the characteristics and two-dimensional nonuniform fast Fourier transform.*

the following: Intel Xeon CPU E5-2630 v3 @ 2.40GHz. Even if both methods have the same complexity $\mathcal{O}(N^2 \log(N))$, the new approach clearly has a smaller constant (around 10 times smaller). Moreover, in such a splitting procedure, a simple and efficient parallelization can be performed since the variable that does not appear in the derivative is just a parameter.

3. Application to the Vlasov–Maxwell equations. In this section, we intend to apply the above splitting to the context of the 1+1/2 Vlasov–Maxwell system. Indeed, the time discretization of this system is based on a time splitting, and one of the pieces (the so-called magnetic part) corresponds to a rotation in the velocity direction due to the presence of the self-consistent electromagnetic field. Then, instead of using a Strang splitting like in [16], we shall use the exact splitting presented in the previous section, so that this magnetic part will be solved exactly in time and with a spectral accuracy in the velocity directions. This is very helpful for designing high-order methods for the full Vlasov–Maxwell system. After introducing the 1+1/2 Vlasov–Maxwell system we intend to solve, the splitting method introduced in [16] is recalled and then high-order methods dedicated to systems split into three parts are introduced.

3.1. Reduced 1+1/2 Vlasov–Maxwell equations. We consider the phase space $(x_1, v_1, v_2) \in L \times \mathbb{R}^2$, where $L = \mathbb{R}/2\pi\mathbb{Z}$ is a one-dimensional torus, and the unknown functions $f(t, x_1, v_1, v_2)$, $B(t, x_1)$, and $E(t, x_1) = (E_1, E_2)(t, x_1)$, which are determined by solving the following system of evolution equations:

$$\begin{aligned}
&\partial_t f + v_1 \partial_{x_1} f + E \cdot \nabla_v f - BJv \cdot \nabla_v f = 0,\\
&\partial_t B = -\partial_{x_1} E_2,\\
&\partial_t E_2 = -\partial_{x_1} B - \int_{\mathbb{R}^2} v_2 f(t, x_1, v) \mathrm{d}v + \overline{\mathcal{J}}_2(t),\\
&\partial_t E_1 = -\int_{\mathbb{R}^2} v_1 f(t, x_1, v) \mathrm{d}v + \overline{\mathcal{J}}_1(t),
\end{aligned} \tag{3.1}$$

where $v = (v_1, v_2)$, $\overline{\mathcal{J}}_i(t) = 1/|L| \int_L \int_{\mathbb{R}^2} v_i f(t, x_1, v) \mathrm{d}x_1 \mathrm{d}v, i = 1, 2$ ($|L|$ denotes the measure of L), and J denotes the symplectic matrix (1.2). This reduced system, which has been considered in several former studies (see [9, 12, 16]), has to be supplemented with the Gauss condition

$$\partial_{x_1} E_1(t, x_1) = \int_{\mathbb{R}^2} f(t, x_1, v) \mathrm{d}v - 1 \ \forall t \geq 0, \tag{3.2}$$

and with initial conditions $f(t=0,x_1,v)=f^{in}(x_1,v), E_2(t=0,x_1)=E_2^{in}(x_1)$ and $B(t=0,x_1)=B^{in}(x_1)$. Notice that $E_1^{in}(x_1)$ is implied by the Gauss condition (3.2) at the initial time.

3.2. Splitting method. Here we propose to use the splitting method introduced in [16] by reformulating the Vlasov–Maxwell system into

$$\frac{dF}{dt}=\mathcal{H}_E(F)+\mathcal{H}_f(F)+\mathcal{H}_B(F), \qquad F(0)=F^{in},$$

where the fields $\mathcal{H}_E(F),\mathcal{H}_f(F)$, and $\mathcal{H}_B(F)$ will be as written below. We denote by $F(\delta_t)=(f,E_1,E_2,B)(\delta_t)$ the solution of the Vlasov–Maxwell system (3.1). This solution can be formally written as $F(\delta_t)=\varphi_{\delta_t}(F^{in}):=\exp((\mathcal{H}_E+\mathcal{H}_f+\mathcal{H}_B)\delta_t)F^{in}$, where $F^{in}=(f^{in},E_1^{in},E_2^{in},B^{in})$ denotes the initial condition.

Now, we want to use a splitting method to approximate the system (3.1). To do so, we shall use the splitting introduced in [16, 17] based on a decomposition into three parts corresponding respectively to the fields $\mathcal{H}_E(F),\mathcal{H}_f(F)$, and $\mathcal{H}_B(F)$. Then, a first order Lie method based on this decomposition is written $\chi_{\delta_t}(F^{in})=\varphi_{\delta_t}(F^{in})+\mathcal{O}({\delta_t}^2)$ with

$$\chi_{\delta_t}=\varphi_{\delta_t}^{[\mathcal{H}_E]}\circ\varphi_{\delta_t}^{[\mathcal{H}_f]}\circ\varphi_{\delta_t}^{[\mathcal{H}_B]}, \tag{3.3}$$

where $\varphi_{\delta_t}^{[\mathcal{H}_E]}$, $\varphi_{\delta_t}^{[\mathcal{H}_f]}$, $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ denote the exact solutions corresponding to the fields $\mathcal{H}_E$, $\mathcal{H}_f$, and $\mathcal{H}_B$. Using this notation, the adjoint [20] of the Lie method $\chi_t^\star$ is written

$$\chi_{\delta_t}^\star=\varphi_{\delta_t}^{[\mathcal{H}_B]}\circ\varphi_{\delta_t}^{[\mathcal{H}_f]}\circ\varphi_{\delta_t}^{[\mathcal{H}_E]}. \tag{3.4}$$

In the following we write the equations associated to the fields $\mathcal{H}_E,\mathcal{H}_f$, and $\mathcal{H}_B$:

$$\begin{aligned}
&\varphi_{\delta_t}^{[\mathcal{H}_E]}: \ \partial_t f+E\cdot\nabla_v f=0, \ \partial_t E=0, \ \partial_t B=-\partial_{x_1}E_2,\\
&\varphi_{\delta_t}^{[\mathcal{H}_f]}: \ \partial_t f+v_1\partial_{x_1}f=0, \ \partial_t E=-\int_{\mathbb{R}^2} vf\mathrm{d}v+\overline{\mathcal{J}}, \ \partial_t B=0,\\
&\varphi_{\delta_t}^{[\mathcal{H}_B]}: \ \partial_t f-BJv\cdot\nabla_v f=0, \ \partial_t E_1=0, \ \partial_t E_2=-\partial_{x_1}B, \ \partial_t B=0.
\end{aligned}$$

Then, as mentioned in [16], $\varphi_{\delta_t}^{[\mathcal{H}_E]}$ and $\varphi_{\delta_t}^{[\mathcal{H}_f]}$ can be computed exactly in time and efficiently in phase space using spectral methods. However, the computation of $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ was performed using a Strang splitting. Instead, we shall use the new splitting $\mathcal{M}_{\delta_t}$ introduced above to compute exactly in time $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ and efficiently in phase space using spectral methods. Let us remark that the application of the new splitting to the $\mathcal{H}_B$ part requires a slight modification. Indeed, to solve $\partial_t f-BJv\cdot\nabla_v f=0$ (with B constant in time during this part) on one time step δ_t from an initial condition f^{in} (defined on the velocity grid), we will use the new splitting with a modified time step $B\delta_t$ to capture the right rotation speed, i.e., $(\mathcal{M}_{B\delta_t})$ with $(\mathcal{M}_{\delta_t})$ defined by (2.10).

Based on the fact that each step can be computed exactly in time, we now look for efficient integration methods for systems separable into three parts which enable us to design efficient high-order methods in time. A simple and efficient way to achieve this goal is to consider compositions of a first-order method with its adjoint computed at fractional stepsizes. This is the main subject of the next part.

3.3. Composition methods for systems separable into three parts. To simplify the presentation, we restrict ourselves to ordinary differential equations. The so-obtained composition methods will then be used within the Vlasov–Maxwell framework.

Let us consider the ODE

$$\frac{dx}{dt}(t) = u(x(t)), \qquad x(0) = x^{in} \in \mathbb{R}^D, \tag{3.5}$$

with $D \in \mathbb{N}^\star$, whose exact solution at time $t = \delta_t$ will be denoted by $x(\delta_t) = \varphi_{\delta_t}(x^{in})$. We are interested in problems where u in (3.5) can be split into three parts,

$$u(x) = u_a(x) + u_b(x) + u_c(x),$$

in such a way that the exact flows $\varphi^{[a]}_{\delta_t}$, $\varphi^{[b]}_{\delta_t}$, $\varphi^{[c]}_{\delta_t}$, corresponding to u_a, u_b, u_c, can be computed exactly. One might consider then splitting methods of the form

$$\varphi^{[a]}_{a_s\delta_t} \circ \varphi^{[b]}_{b_s\delta_t} \circ \varphi^{[c]}_{c_s\delta_t} \circ \cdots \circ \varphi^{[a]}_{a_1\delta_t} \circ \varphi^{[b]}_{b_1\delta_t} \circ \varphi^{[c]}_{c_1\delta_t} \tag{3.6}$$

and fix the coefficients $a_i, b_i, c_i,\ i = 1, \ldots, s$, so that it provides an approximation of order, say, p. It turns out, however, that the number of order conditions to be satisfied by these parameters grows very rapidly with the order. Thus, time-symmetric schemes of order $p = 4$ (resp., $p = 6$) require solving 11 (resp., 56) conditions. A more convenient way consists in considering compositions of χ_{δ_t} and its adjoint $\chi^\star_{\delta_t}$, with

$$\chi_{\delta_t} = \varphi^{[a]}_{\delta_t} \circ \varphi^{[b]}_{\delta_t} \circ \varphi^{[c]}_{\delta_t} \qquad \text{and} \qquad \chi^\star_{\delta_t} = \varphi^{[c]}_{\delta_t} \circ \varphi^{[b]}_{\delta_t} \circ \varphi^{[a]}_{\delta_t}. \tag{3.7}$$

More specifically, we construct integrators within the family

$$\mathcal{G}_1 \equiv \left\{ \psi_{\delta_t} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \cdots \circ \chi_{\alpha_{2s-1}\delta_t} \circ \chi^\star_{\alpha_{2s}\delta_t} :\ s \geq 1,\ (\alpha_j)_{1\leq j\leq 2s} \in \mathbb{R}^{2s} \right\}, \tag{3.8}$$

where χ_{δ_t} and $\chi^\star_{\delta_t}$ are given by (3.7), so that

$$\chi_{\delta_t}(x^{in}) = \varphi_{\delta_t}(x^{in}) + \mathcal{O}(\delta_t^2), \tag{3.9}$$

and an analogous relation for $\chi^\star_{\delta_t}$. Composition integrators $\psi_{\delta_t} \in \mathcal{G}_1$ are time-symmetric (self-adjoint) whenever they have left-right palindromic sequences of coefficients α_i, i.e., if $\alpha_{2s+1-i} = \alpha_i,\ \ i = 1, \ldots, s$ [20].

Notice that one could achieve methods of order p within this family even if only first-order approximations to the flows $\varphi^{[a]}_{\delta_t}$, $\varphi^{[b]}_{\delta_t}$, and $\varphi^{[c]}_{\delta_t}$ are available, as long as one is able to construct the corresponding adjoint $\chi^\star_{\delta_t}$.

Remark 3.1. Another well-known class $\mathcal{G}_2$ of integrators is formed by compositions

$$\mathcal{G}_2 = \left\{ \psi_{\delta_t} = \phi_{\alpha_1\delta_t} \circ \cdots \circ \phi_{\alpha_s\delta_t} :\ s \geq 1,\ (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s \right\}, \tag{3.10}$$

where $\phi_{\delta_t} : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is any second-order self-adjoint integrator. Notice that if ϕ_{δ_t} is chosen as $\phi_{\delta_t} = \chi_{\delta_t/2} \circ \chi^\star_{\delta_t/2}$, then $\mathcal{G}_2$ is contained in $\mathcal{G}_1$. These integrators also enjoy the time-symmetric property if $\alpha_{s+1-i} = \alpha_i,\ i = 1, \ldots, s$.

3.4. Analysis of composition methods for systems separable into three parts. Next we formally analyze the new composition methods. To do that, it is convenient to introduce the graded Lie algebra associated with the vector field defining the ODE (3.5) and its corresponding exact flow φ_{δ_t}. As is well known, for each infinitely differentiable map $g : \mathbb{R}^D \longrightarrow \mathbb{R}$, $g(\varphi_{\delta_t}(x))$ admits an expansion of the form

$$g(\varphi_{\delta_t}(x)) = \mathrm{e}^{\delta_t F}[g](x) = g(x) + \sum_{k\geq 1} \frac{{\delta_t}^k}{k!} F^k[g](x), \qquad x \in \mathbb{R}^D,$$

where F is the vector field associated with u,

$$F = \sum_{i=1}^{D} u_i(x)\, \frac{\partial}{\partial x_i}.$$

Similarly, for the basic first-order method χ_{δ_t} defined by (3.9), one has $g(\chi_{\delta_t}(x)) = \mathrm{e}^{Y_{\delta_t}}[g](x)$ with $Y_{\delta_t} = \sum_{k\geq 1} {\delta_t}^k Y_k$, and for its adjoint $\chi^\star_{\delta_t} \equiv \chi^{-1}_{-\delta_t}$ one has $g(\chi^\star_{\delta_t}(x)) = e^{-Y_{-\delta_t}}[g](x)$. Then, one can formally compute the operator series associated to any integrator $\psi_{\delta_t} \in \mathcal{G}_1$ defined by (3.8)

$$\Psi_{\delta_t} = \exp(Y_{\delta_t\alpha_1}) \exp(-Y_{-\delta_t\alpha_2}) \cdots \exp(Y_{\delta_t\alpha_{2s-1}}) \exp(-Y_{-\delta_t\alpha_{2s}}).$$

By repeated application of the Baker–Campbell–Hausdorff formula we can express formally Ψ_{δ_t} as the exponential of an operator F_{δ_t},

$$\Psi_{\delta_t} = \mathrm{e}^{F_{\delta_t}} \qquad \text{with} \qquad F_{\delta_t} = \sum_{k\geq 1} {\delta_t}^k F_k,$$

${\delta_t}^k F_k \in \mathcal{L}_k$ for each $k \geq 1$ and $\mathcal{L} = \bigoplus_{k\geq 1} \mathcal{L}_k$ is the graded Lie algebra generated by the vector fields $\{\delta_t Y_1, \delta_t^2 Y_2, \delta_t^3 Y_3, \ldots\}$, where, by consistency, $Y_1 = F$. Notice that

$$\begin{aligned} Y_{\delta_t\alpha_i} &= \delta_t\alpha_i Y_1 + (\delta_t\alpha_i)^2 Y_2 + (\delta_t\alpha_i)^3 Y_3 + \cdots, \\ -Y_{-\delta_t\alpha_i} &= \delta_t\alpha_i Y_1 - (\delta_t\alpha_i)^2 Y_2 + (\delta_t\alpha_i)^3 Y_3 - \cdots \end{aligned}$$

so that

$$\begin{aligned} \Psi_{\delta_t} = \exp\big(&\delta_t w_1 Y_1 + {\delta_t}^2 w_2 Y_2 + {\delta_t}^3 (w_3 Y_3 + w_{12}[Y_1, Y_2]) \\ &+ {\delta_t}^4 (w_4 Y_4 + w_{13}[Y_1, Y_3] + w_{112}[Y_1, [Y_1, Y_2]]) + \mathcal{O}({\delta_t}^5)\big), \end{aligned}$$

where $w_1, w_2, \ldots$ are polynomials in the coefficients α_i. In particular, we have

$$\begin{aligned} w_1 &= \sum_{i=1}^{2s} \alpha_i, \qquad w_2 = \sum_{i=1}^{2s} (-1)^i \alpha_i^2, \\ w_3 &= \sum_{i=1}^{2s} \alpha_i^3, \qquad w_4 = \sum_{i=1}^{2s} (-1)^i \alpha_i^4, \\ w_{12} &= \frac{1}{2} \left(\sum_{i=1}^{2s-1} (-1)^{i+1} \alpha_i^2 \sum_{j=i+1}^{2s} \alpha_j + \sum_{i=1}^{2s-1} \alpha_i \sum_{j=i+1}^{2s} (-1)^j \alpha_j^2 \right). \end{aligned} \tag{3.11}$$

3.5. Methods of order 4. To construct symmetric time-integration schemes of order 4 within the family $\mathcal{G}_1$ we only have to solve equations $w_1 = 1$ (for consistency) and the third-order conditions $w_3 = w_{12} = 0$, where w_1, w_3, w_{12} are given by (3.11) and conditions at even order ($w_2 = w_4 = 0$) are automatically verified by symmetry.

It is then clear that we need at least $2s = 6$ maps (or *stages*). It turns out, however, that methods with the minimum $s = 3$ do not usually provide the best efficiency. In other words, considering additional stages (and thus some free parameters) leads to more efficient schemes, even when the computational cost per step is also higher. The difficulty then lies in the way the free parameters are fixed according with some previously chosen optimization criterion. In this respect, several objective functions have been considered in the literature. In particular we mention the following [7] (let us recall that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s}$):

$$\mathcal{E}_1(\boldsymbol{\alpha}) = \sum_{i=1}^{2s} |\alpha_i| \qquad \text{and} \qquad \mathcal{E}_2(\boldsymbol{\alpha}) = 2s \left| \sum_{i=1}^{2s} \alpha_i^5 \right|^{1/4}. \tag{3.12}$$

The quantity $\mathcal{E}_2$ is usually the dominant error term for a number of problems. The criterion we follow here will be to look for symmetric methods with small values of $\mathcal{E}_1$ which, in addition, have also small values of $\mathcal{E}_2$. This has been shown to lead to efficient methods when the processing technique is considered [7]. In what follows, we consider composition methods in the class $\mathcal{G}_1$ with $s = 3, 4, 5, 6$ (see (3.8)) which have been designed by optimizing both functions $\mathcal{E}_1$ and $\mathcal{E}_2$.

Case $s = 3$. The integrator reads

$$\psi_{\delta_t}^{[3]} = \chi_{\alpha_1\delta_t} \circ \chi_{\alpha_2\delta_t}^{\star} \circ \chi_{\alpha_3\delta_t} \circ \chi_{\alpha_3\delta_t}^{\star} \circ \chi_{\alpha_2\delta_t} \circ \chi_{\alpha_1\delta_t}^{\star} \tag{3.13}$$

and the unique (real) solution to the order conditions $w_1 = 1, w_3 = w_{12} = 0$, is given by

$$\alpha_1 = \alpha_2 = \frac{1}{2(2 - 2^{1/3})}, \qquad \alpha_3 = \frac{1}{2} - 2\alpha_1.$$

If $\chi_{\delta_t} = \varphi_{\delta_t}^{[a]} \circ \varphi_{\delta_t}^{[b]} \circ \varphi_{\delta_t}^{[c]}$, then it involves 13 maps (the minimum number). The values of the objective functions are $\mathcal{E}_1(\boldsymbol{\alpha}) \simeq 4.40483$ and $\mathcal{E}_2(\boldsymbol{\alpha}) \simeq 4.55004$.

Remark 3.2. Notice that this corresponds to the familiar scheme of Yoshida [31]

$$\psi_{\delta_t} = \phi_{\gamma\delta_t/2} \circ \phi_{\beta\delta_t} \circ \phi_{\gamma\delta_t/2}$$

in $\mathcal{G}_2$ with $\gamma = 1/(2 - 2^{1/3})$. Moreover, this method is also recovered in [23] when considering splitting methods of the form (3.6).

Case $s = 4$. The composition is

$$\psi_{\delta_t}^{[4]} = \chi_{\alpha_1\delta_t} \circ \chi_{\alpha_2\delta_t}^{\star} \circ \chi_{\alpha_3\delta_t} \circ \chi_{\alpha_4\delta_t}^{\star} \circ \chi_{\alpha_4\delta_t} \circ \chi_{\alpha_3\delta_t}^{\star} \circ \chi_{\alpha_2\delta_t} \circ \chi_{\alpha_1\delta_t}^{\star}, \tag{3.14}$$

involving 17 maps. Now we have a free parameter, which we take as α_1. The minima of both $\mathcal{E}_1$ and $\mathcal{E}_2$ are achieved at approximately $\alpha_1 = 0.358$, and so the coefficients are

$$\begin{aligned} &\alpha_1 = 0.358, && \alpha_2 = -0.47710242361717810834, \\ &\alpha_3 = 0.35230499471528197958, && \alpha_4 = 0.26679742890189612876 \end{aligned}$$

with $\mathcal{E}_1(\boldsymbol{\alpha}) \simeq 2.9084$ and $\mathcal{E}_2(\boldsymbol{\alpha}) \simeq 3.1527$. This scheme has not appeared previously in the literature.

Case $s = 5$. Now the composition

$$\psi_{\delta_t}^{[5]} = \chi_{\alpha_1\delta_t} \circ \chi_{\alpha_2\delta_t}^{\star} \circ \chi_{\alpha_3\delta_t} \circ \chi_{\alpha_4\delta_t}^{\star} \circ \chi_{\alpha_5\delta_t} \circ \chi_{\alpha_5\delta_t}^{\star} \circ \chi_{\alpha_4\delta_t} \circ \chi_{\alpha_3\delta_t}^{\star} \circ \chi_{\alpha_2\delta_t} \circ \chi_{\alpha_1\delta_t}^{\star} \tag{3.15}$$

involves 21 maps when applied to a system separable into three parts. By carrying out a similar analysis we conclude that the best solution according to the criterion adopted is achieved when

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2(4-4^{1/3})}, \qquad \alpha_5 = \frac{1}{2} - 4\alpha_1,$$

which give $\mathcal{E}_1(\boldsymbol{\alpha}) \simeq 2.3159$ and $\mathcal{E}_2(\boldsymbol{\alpha}) \simeq 2.6111$.

Remark 3.3. This method also belongs to $\mathcal{G}_2$ since it can be written as

$$\psi_{\delta_t} = \phi_{\gamma\delta_t} \circ \phi_{\gamma\delta_t} \circ \phi_{\beta\delta_t} \circ \phi_{\gamma\delta_t} \circ \phi_{\gamma\delta_t}$$

with coefficients $\gamma = 2\alpha_1, \beta = 2\alpha_5$, originally proposed in [27].

Case $s = 6$. Analogously we have considered a composition involving three free parameters and 25 maps:

$$\psi_{\delta_t}^{[6]} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^\star_{\alpha_4\delta_t} \circ \chi_{\alpha_5\delta_t} \circ \chi^\star_{\alpha_6\delta_t} \circ \chi_{\alpha_6\delta_t} \circ \chi^\star_{\alpha_5\delta_t} \circ \chi_{\alpha_4\delta_t} \circ \chi^\star_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^\star_{\alpha_1\delta_t}. \tag{3.16}$$

A solution leading to small values of $\mathcal{E}_1$ and $\mathcal{E}_2$ is

$$\begin{aligned} &\alpha_1 = \alpha_2 = \frac{3}{20}, && \alpha_3 = \frac{17}{100}, \\ &\alpha_4 = -0.2628463256938681137, && \alpha_5 = 0.16217658484020533783, \\ &\alpha_6 = 0.13066974085366277593 \end{aligned}$$

with $\mathcal{E}_1(\boldsymbol{\alpha}) \simeq 2.0513$ and $\mathcal{E}_2(\boldsymbol{\alpha}) \simeq 2.4078$. Notice how the values of $\mathcal{E}_1(\boldsymbol{\alpha})$ and $\mathcal{E}_2(\boldsymbol{\alpha})$ are reduced by considering additional stages. This particular scheme is also presented here for the first time.

Remark 3.4. Although the optimization criterion we have adopted here usually leads to good methods, one can find schemes in the literature with larger values of $\mathcal{E}_1$ and $\mathcal{E}_2$ which are very efficient in practice. Thus, in particular, we mention the fourth-order splitting method designed in [6] which, once written as a method in $\mathcal{G}_1$, also involves $s = 6$ stages.

4. Numerical results for the Vlasov type equations. In this section, we show some numerical results to illustrate the efficiency and performance of the methods previously derived. We focus on Vlasov applications by considering the Vlasov–Maxwell system as well as Vlasov–HMF system.

4.1. Vlasov–Maxwell system. The composition methods introduced in the previous sections can then be used to derive a global fourth-order method for the Vlasov–Maxwell equation. As an example, the Yoshida (or triple-jump) method ($s = 3$) in the Vlasov–Maxwell context is written

$$\psi_{\delta_t}^{[3]} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^\star_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^\star_{\alpha_1\delta_t},$$

with $\alpha_1 = \alpha_2 = \frac{1}{2(2-2^{1/3})}, \alpha_3 = \frac{1}{2} - 2\alpha_1$, and where χ_{δ_t} and $\chi^\star_{\delta_t}$ are given by (3.3) and (3.4), respectively. Then, if we denote by F^n an approximation at time $t^n = n\delta_t, n \in \mathbb{N}$, of the Vlasov–Maxwell solution $F(t^n)$, we have $F^n = (\psi_{\delta_t}^{[3]})^n(F^{in})$, and F^n is a fourth-order approximation of $F(t^n)$. The other fourth-order methods $\psi_{\delta_t}^{[s]}, s = 4, 5, 6,$

are defined by (3.14), (3.15), and (3.16) in subsection 3.5. We also define the standard Strang splitting $\psi_{\delta_t}^{[2]}$ which, with our notation, is written

$$\begin{aligned}\psi_{\delta_t}^{[2]} &= \chi_{\delta_t/2} \circ \chi_{\delta_t/2}^{\star} = \varphi_{\delta_t/2}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_E]} \\ &= \varphi_{\delta_t/2}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_E]}.\end{aligned}$$

The Strang splitting for a decomposition into three parts involves five maps since, as usual, the first and the last map can be concatenated.

We present some numerical results to illustrate the efficiency of the different methods. First of all, we used the methods $\psi_{\delta_t}^{[s]}, s = 2, 3, 4, 5, 6$. In this context, one goal is to compare the new exact splitting for the rotation applied to the field $\mathcal{H}_B$ to a standard Strang method. In the methods $\psi_{\delta_t}^{[s]}$, the flow $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ is then approximated by the Strang splitting $\mathcal{T}_{\delta_t}^{[\mathcal{H}_B]}$ given by (2.10). This means that in this method, χ_{δ_t} is now replaced by $\tilde{\chi}_{\delta_t}$ defined by $\tilde{\chi}_{\delta_t} = \varphi_{\delta_t}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t}^{[\mathcal{H}_f]} \circ \mathcal{T}_{\delta_t}^{[\mathcal{H}_B]}$. The global Strang splitting is then defined by $\tilde{\psi}_{\delta_t}^{[2]}$ with $\tilde{\psi}_{\delta_t}^{[2]} = \tilde{\chi}_{\delta_t/2} \circ \tilde{\chi}_{\delta_t/2}^{\star}$, and the definition of $\tilde{\psi}_{\delta_t}^{[s]}$ for $s = 3, 4, 5, 6$ follows directly. Let us remark that even if the magnetic part $\mathcal{H}_B$ is not solved exactly in time, the global method $\tilde{\psi}_{\delta_t}^{[s]}$ still has the same order as $\psi_{\delta_t}^{[s]}$ (i.e., of order 2 for $s = 2$ or of order 4 for $s = 3, 4, 5, 6$). We then want to investigate the impact of this approximation on the global error of the so-obtained splitting.

To do so, we consider the initial condition for (3.1),

$$f^{in}(x_1, v_1, v_2) = \frac{1}{\pi v_{\mathrm{th}}^2 \sqrt{T_r}} \mathrm{e}^{-(v_1^2 + v_2^2/T_r)/v_{\mathrm{th}}^2} (1 + \alpha \cos(k x_1)),$$

and $B^{in}(x_1) = 10 + 3\cos(kx_1), E_2^{in}(x_1) = 0$. We consider $\alpha = 10^{-4}, k = 0.4, v_{\mathrm{th}} = 0.02, k = 0.4$, and $T_r = 12$. The phase space domain is $(x_1, v_1, v_2) \in [0, 2\pi/k] \times [-1, 1]^2$ and the number of points is $N_x = 8$ in space and $N_v = 513$ per direction in velocity. The runs are performed up to a final time $T = 2$ and different values of the time step δ_t are considered between 10^{-3} to 0.4. The results are given in Figure 7, where we have plotted the L^∞ error on the total energy with respect to δ_t/M, where M is the number of maps. The total energy (which is conserved with time at the continuous level) is defined by

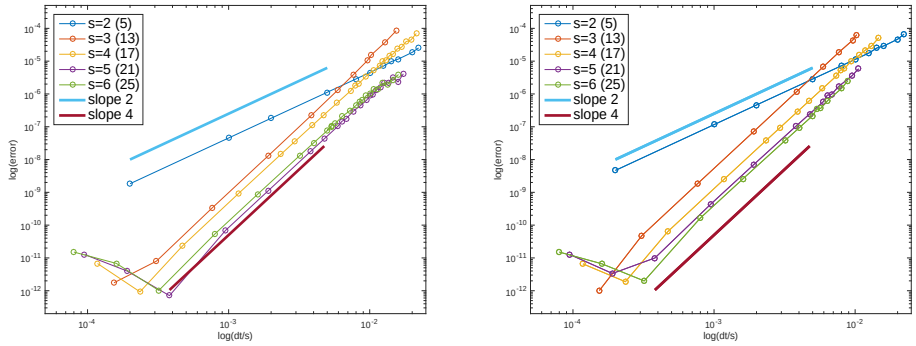

FIG. 7. *Efficiency diagrams obtained by different composition methods* $\psi_{\delta_t}^{[s]}, s = 2, 3, 4, 5, 6$ *(left) and* $\tilde{\psi}_{\delta_t}^{[s]}, s = 2, 3, 4, 5, 6$ *(right) for the Vlasov–Maxwell system. The number of maps for each method is indicated in parentheses.*

$$\mathcal{H}(t) = \frac{1}{2}\int_L |E|^2 \mathrm{d}x + \frac{1}{2}\int_L |B|^2 \mathrm{d}x + \frac{1}{2}\int_{L\times\mathbb{R}^2} |v|^2 f \mathrm{d}v\mathrm{d}x \tag{4.1}$$

with $L = [0, 2\pi/k]$, and the error we consider is

$$\text{err} := \max_{t\in[0,T]} \Big| \frac{\mathcal{H}(t) - \mathcal{H}(0)}{\mathcal{H}(0)} \Big|. \tag{4.2}$$

Let us remark that other conserved quantities (like the L^2 norm of f) may not be affected by the use of high-order time integrators (see [15]).

First, one can see that the order of convergence is well recovered for all the methods but some fourth-order methods present better efficiency. For instance, the two methods corresponding to $s = 5$ and $s = 6$ are clearly the best and are much more efficient than the triple jump method ($s = 3$) or the Strang one ($s = 2$) even if they involve a larger number of maps. Second, we can observe than the error produced by the methods $\psi_{\delta_t}^{[s]}$ (i.e., when the exact splitting is used for the part $\mathcal{H}_B$) is smaller than the error performed by the methods $\tilde{\psi}_{\delta_t}^{[s]}$ (i.e. when a Strang splitting is used for the part $\mathcal{H}_B$). Note that in Figure 7 the lines indicating the order are kept fixed. For the Strang method the ratio between the error produced by $\tilde{\psi}_{\delta_t}^{[2]}$ and $\psi_{\delta_t}^{[2]}$ is about 2.5, whereas the ratio between the error produced by $\tilde{\psi}_{\delta_t}^{[5]}$ and $\psi_{\delta_t}^{[5]}$ is about 6 (the same ratio is observed between $\tilde{\psi}_{\delta_t}^{[6]}$ and $\psi_{\delta_t}^{[6]}$). Let us remark that, for a given method, the cost of a $\tilde{\psi}_{\delta_t}^{[s]}$ method is the same as that of a $\psi_{\delta_t}^{[s]}$ method.

We end this subsection by considering other splitting methods from the literature, namely the splitting methods of the form (3.6) from [2] which assume that each subpart is solved exactly, which is our case when the exact splitting is used for the magnetic part. The results are displayed in Figure 8, where we have tested second-order methods (AK 3-2 and AK 5-2 involve 9 maps), a fourth-order method (AK 11-4 involves 21 maps), and even a sixth-order method (AY 15-6 involves 29 maps). We refer to [2] for more details on these methods. As previously we also added $\psi_{\delta_t}^{[2]}$ (second order) and $\psi_{\delta_t}^{[5]}$ (fourth order) for comparison, whereas slopes 2 and 4 are the same as in Figure 7. First, we observe that AK 3-2 is the best second-order method. The third-order PP method is not very attractive in this context compared to second-order methods. Second, among the two fourth-order methods (AK 11-4 and $\psi_{\delta_t}^{[5]}$), the

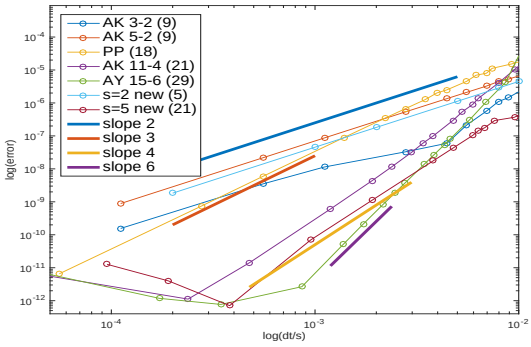

FIG. 8. *Efficiency diagrams obtained by different methods from* [2] *and* $\psi_{\delta_t}^{[2]}$, $\psi_{\delta_t}^{[5]}$ *for the Vlasov–Maxwell system. The order lines "slope* 2*" and "slope* 4*" are the same as in Figure* 7*. The number of maps for each method is indicated in parentheses.*

method $\psi_{\delta_t}^{[5]}$ offers better efficiency since the error is about 5 times smaller. Finally, the method AY 15-6 offers sixth-order accuracy but this extra accuracy is apparent only for very small time steps.

4.2. Vlasov–Maxwell system: Long time test. We now present a test to highlight the fact that the new methods are able to capture the long time dynamics of the Vlasov–Maxwell solution. Then, we consider the same initial condition as in the previous test,

$$f^{in}(x_1, v_1, v_2) = \frac{1}{2\pi\beta} \mathrm{e}^{-v_2^2/\beta} \left[\mathrm{e}^{-(v_1-0.1)^2/\beta} + \mathrm{e}^{-(v_1+0.3)^2/\beta} \right],$$

where $(x_1, v_1, v_2) \in [0, 2\pi] \times [-1, 1]^2$ and we have chosen $\beta = 0.002$. The electric fields E_1^{in} and E_2^{in} are set to zero, whereas the magnetic field is prescribed as $B^{in}(x_1) = 1 + 0.0001 \sin(x_1)$. The number of points in space is $N_x = 32$, whereas we took $N_v = 257$ points per velocity direction.

We compare the $\psi_{\delta_t}^{[5]}$ method (which is the best method according to the previous tests) with the two second-order splittings $\psi_{\delta_t}^{[2]}$ (referred to as $s = 2$ new) and $\tilde{\psi}_{\delta_t}^{[2]}$ (referred as Strang). Let us recall that these two second-order splittings differ only in the solving of the magnetic part. The time step is chosen as $\delta_t = 0.125$ or $\delta_t = 0.025$ so that we will compare $\psi_{\delta_t}^{[5]}$ and $\psi_{\delta_t}^{[2]}$ at a fixed computational cost and the final time is $t_f = 500$.

In Figure 9 (left), we plot the time evolution of the relative total energy given by (4.2) for the two second-order splittings with a small time step ($\delta_t = 0.025$). The cost of these two methods is the same, but we can see that the relative total energy is better preserved for the $\psi_{\delta_t}^{[2]}$ method (about 2×10^{-9}) compared to the standard $\tilde{\psi}_{\delta_t}^{[2]}$ method (about 2×10^{-6}). On the right part of Figure 9, we compare the methods for which the rotation is solved exactly, namely $\psi_{\delta_t}^{[2]}$ (with $\delta_t = 0.025$ and 0.125) and $\psi_{\delta_t}^{[5]}$ (with $\delta_t = 0.125$). Let us mention that $\psi_{\delta_t}^{[2]}$ for $\delta_t = 0.025$ and $\psi_{\delta_t}^{[5]}$ for $\delta_t = 0.125$ have the same number of stages and so as the same computational cost. We can observe that the high-order method $\psi_{\delta_t}^{[5]}$ preserves very well the total energy (about 6×10^{-12}), which confirms the results obtained in the previous subsection.

Finally, the time history of first mode of E_1 is plotted in Figure 10, for $\psi_{\delta_t}^{[2]}$ and $\psi_{\delta_t}^{[5]}$ methods, with $\delta_t = 0.125$. We can observe that after a linear phase during which the amplitude of the mode grows exponentially, a saturation phase is well captured by the two methods, even if the saturation level is not the same (see the zoomed inset figure for $t \in [400, 500]$). Refining the time step by considering $\delta_t = 0.025$ enables the $\psi_{\delta_t}^{[2]}$

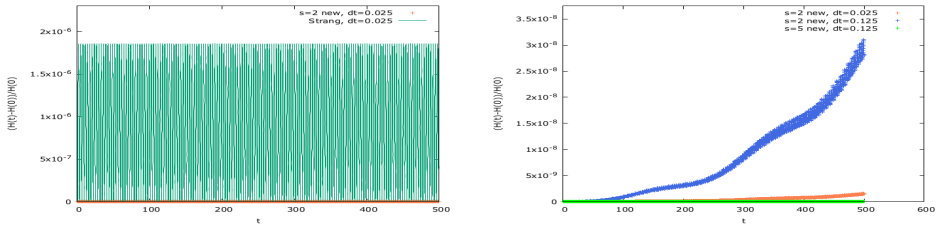

FIG. 9. *Time history of the relative total energy. Left:* $\psi_{\delta_t}^{[2]}$ *and* $\tilde{\psi}_{\delta_t}^{[2]}$ *with* $\delta_t = 0.025$. *Right:* $\psi_{\delta_t}^{[2]}$ *with* $\delta_t = 0.125$ *and* $\delta_t = 0.025$, $\psi_{\delta_t}^{[5]}$ *with* $\delta_t = 0.125$.

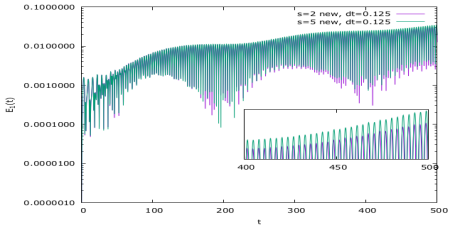

FIG. 10. *Time history of the first mode of* E_1 *(semi-log scale):* $\psi_{\delta_t}^{[2]}$ *and* $\psi_{\delta_t}^{[5]}$ *with* $\delta_t = 0.125$.

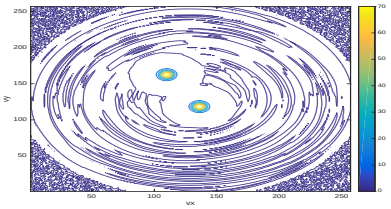

FIG. 11. *Velocity dependence of* $\int f(t = 500, x, v_x, v_y)dx$ *at the final time, using* $\psi_{\delta_t}^{[5]}$ *with* $\delta_t = 0.125$.

method to recover the results obtained by $\psi_{\delta_t}^{[5]}$. In Figure 11 the velocity dependence of $\int_0^{2\pi} f(t = 500, x, v_x, v_y)dx$ is plotted. The use of Fourier interpolation in the velocity directions may induce some noise that contaminates the numerical solution (see [22]) and other strategies can be used such as high-order piecewise polynomial interpolation (see [10]). As illustrated in Figure 11, this spurious effect is not too important in our case.

4.3. Vlasov-HMF system. Our goal is to solve numerically the Vlasov-HMF model satisfied by $f(t, x, v)$, $(x, v) \in L \times \mathbb{R}$, with $L = \mathbb{R}/2\pi\mathbb{Z}$ (see [21])

$$\partial_t f + \{f, H[f]\} = 0, \tag{4.3}$$

where $\{f, g\} = \partial_x f \partial_v g - \partial_v f \partial_x g$ and $H[f]$ is given by $H[f] = \frac{v^2}{2} - \Phi[f](x)$. Finally, the potential is defined by

$$\Phi[f](x) = \cos x \int_{L\times\mathbb{R}} \cos(y) f(y, u)\mathrm{dydu} + \sin x \int_{L\times\mathbb{R}} \sin(y) f(y, u)\mathrm{dydu}. \tag{4.4}$$

We consider the following stationary solution:

$$f^{eq}(x, v) = \gamma e^{-\beta\left(\frac{v^2}{2} - M_0 \cos x\right)} \text{ with } M_0 = \int_{L\times\mathbb{R}} \cos(y) f^{eq}(y, u)\mathrm{dydu}, \tag{4.5}$$

where $\gamma, \beta, M_0 \in \mathbb{R}$ will be explicitly given below. Following [21], the long time behavior of (4.3) is driven by the linearized Hamiltonian part, i.e., $\partial_t f + \{f, H[f^{eq}]\} = 0$, with $H[f^{eq}] = \frac{v^2}{2} - M_0 \cos(x)$. We recognize the pendulum Hamiltonian for which a slight modification of the new splitting is able to capture the rotation phenomena with high accuracy compare to standard Strang splitting (see [4]). In this HMF context, the material introduced before has to be slightly modified.

First, let us introduce the discretization of the phase space $L \times [-v_{\max}, v_{\max}]$, with $v_{\max} > 0$ a truncation of the velocity direction. We consider $\mathbb{G}_x := h_x [\![0, N_x - 1]\!]$

the space grid (with $h_x = L/N_x$ the stepsize and $N_x \in \mathbb{N}^*$ the number of points) and $\mathbb{G}_v := h_v [\![-\lfloor (N_v-1)/2 \rfloor, \lfloor N_v/2 \rfloor]\!]$ the speed grid (with $h_v = 2v_{\max}/N_v$ the stepsize and $N_v \in \mathbb{N}^*$ the number of points). We also introduce the set of discrete frequencies: $\widehat{\mathbb{G}}_x = \eta_x [\![-\lfloor (N_x-1)/2 \rfloor, \lfloor N_x/2 \rfloor]\!]$ and $\widehat{\mathbb{G}}_v = \eta_v [\![-\lfloor (N_v-1)/2 \rfloor, \lfloor N_v/2 \rfloor]\!]$ with $\eta_x = 2\pi/L$ and $\eta_v = \pi/v_{\max}$. Then, we define the discrete partial Fourier transforms

$$\mathcal{F}_1 : \begin{cases} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v}, \to & \mathbb{C}^{\widehat{\mathbb{G}}_x \times \mathbb{G}_v}, \\ \boldsymbol{u} & \mapsto h_x \displaystyle\sum_{g_1 \in \mathbb{G}_x} \boldsymbol{u}_{g_1,g_2}\, e^{-i g_1 \xi_1}, \end{cases} \text{ and } \mathcal{F}_2 : \begin{cases} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v} \to & \mathbb{C}^{\mathbb{G}_x \times \widehat{\mathbb{G}}_v}, \\ \boldsymbol{u} & \mapsto h_v \displaystyle\sum_{g_2 \in \mathbb{G}_v} \boldsymbol{u}_{g_1,g_2}\, e^{-i g_2 \xi_2}, \end{cases}$$

whereas the shears are now defined by

$$\mathcal{S}_1^\alpha : \begin{cases} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v}, \to & \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v}, \\ \boldsymbol{u} & \mapsto \mathcal{F}_1^{-1} \left[e^{i\alpha \xi_1 g_2} \mathcal{F}_1\, \boldsymbol{u} \right], \end{cases} \text{ and } \widetilde{\mathcal{S}}_2^\alpha : \begin{cases} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v}, \to & \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v}, \\ \boldsymbol{u} & \mapsto \mathcal{F}_2^{-1} \left[e^{i\alpha \xi_2 E[\boldsymbol{u}]_{g_1}} \mathcal{F}_2 \boldsymbol{u} \right], \end{cases}$$

where $E[\boldsymbol{u}]_{g_1}$ is deduced from the relation $E[\boldsymbol{u}](x) = -\partial_x \Phi[\boldsymbol{u}](x)$ and (4.4)

$$\begin{aligned} E[\boldsymbol{u}]_{g_1} = \sin(g_1 h_x) h_x h_v \sum_{(g_1,g_2) \in \mathbb{G}_x \times \mathbb{G}_v} \cos(g_1 h_x)\, \boldsymbol{u}_{g_1,g_2} \\ - \cos(g_1 h_x) h_x h_v \sum_{(g_1,g_2) \in \mathbb{G}_x \times \mathbb{G}_v} \sin(g_1 h_x)\, \boldsymbol{u}_{g_1,g_2}\,. \end{aligned} \tag{4.6}$$

Then, at time $t^n = n\delta_t$, we denote by f^n an approximation of the solution $f(t^n)$ on the phase space grid computed by the Strang splitting $\widetilde{\mathcal{T}}_{\delta_t}$ and the new splitting $\widetilde{\mathcal{M}}_{\delta_t}$ which are defined by

$$\begin{aligned} f^{n+1} &= \widetilde{\mathcal{T}}_{\delta_t}\, f^n := \mathcal{S}_1^{-\delta_t/2} \widetilde{\mathcal{S}}_2^{\delta_t} \mathcal{S}_1^{-\delta_t/2}\, f^n && \text{(Strang)}, \\ f^{n+1} &= \widetilde{\mathcal{M}}_{\delta_t}\, f^n \\ &:= \mathcal{S}_1^{-t_c \tan(\delta_t/(2t_c))} \widetilde{\mathcal{S}}_2^{t_c \sin(\delta_t/t_c)} \mathcal{S}_1^{-t_c \tan(\delta_t/(2t_c))}\, f^n && \text{(New)}, \end{aligned} \tag{4.7}$$

where $f^0 := f^{in}$, and $t_c = \frac{1}{\sqrt{M_0}}$ is the characteristic time of the Vlasov-HMF model which has been introduced to capture the correct angular velocity. Let us remark that the electric field $E[f]$ has to be solved using (4.6) before the shear $\widetilde{\mathcal{S}}_2^\alpha$ in the splittings defined previously.

To evaluate the performance of the new splitting compare to the Strang one, we consider an initial condition f^{in} as a perturbation of the equilibrium solution (4.5) (with $\beta = 10, M_0 = 0.9455421864232981$ and $\alpha = 0.0001194365987897421$)

$$f^{in}(x,v) = f^{eq}(x,v)(1 + \varepsilon \cos(x)), \qquad (x,v) \in [-\pi,\pi] \times \mathbb{R},$$

with $\varepsilon = 10^{-3}$. We consider a truncated velocity domain of $[-8,8]$, and the number of points in the spatial direction is $N_x = 128$, whereas we considered $N_v = 256$ points in the velocity direction, and the final time is $T = 25$. Note that the splitting can also be coupled to a semi-Lagrangian method; the shears $\mathcal{S}_1$ and $\widetilde{\mathcal{S}}_2$ have to be modified accordingly (see [5], for instance).

We look at the L^∞ error between a reference distribution function (obtained with the new splitting with a small time step $\delta_t = T/1000$) and the one obtained by Strang or new splitting given by (4.7) (with $t_c = 1.0283940255$) for different time steps $\delta_t \in \{T/50, T/100, T/150, T/200, T/250\}$. The results are displayed in Figure 12 in log-log scale. First we observe that, as expected, the two methods are second-order accurate in time. But, one can remark that the error produced by the new splitting is much smaller than the error produced by the Strang splitting, at the same cost (the number of maps is the same for the two methods).

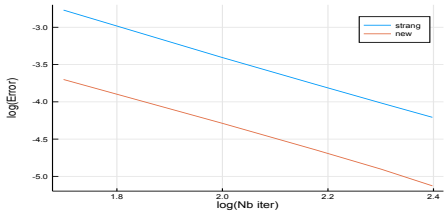

FIG. 12. *Error as a function of the number of iterations for the HMF-Poisson system. Comparison of the Strang splitting and the new splitting.*

5. Conclusion. In this work, we have studied a directional splitting which preserves exactly the rotations and we have applied it to the PDE context. A careful numerical analysis of this splitting coupled with spectral interpolation techniques has been performed. These results are illustrated by some numerical experiments. Then, this step serves as a building block of a splitting for the Vlasov–Maxwell system. Indeed, this system can be split into three parts which, thanks to this new splitting, can all be solved exactly. New high-order composition methods are then designed to accurately and efficiently solve the full Vlasov–Maxwell system. Numerical results show the good behavior of these methods.

Acknowledgments. FC would like to express his gratitude to the Université de Rennes 1 for its hospitality. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1] E. ANDRES, *The quasi-shear rotation*, in Discrete Geometry for Computer Imagery, S. Miguet, A. Montanvert, and S. Ubéda, eds., Lecture Notes in Comput. Sci. 1176, Springer, Berlin, 1996.

[2] W. AUZINGER, H. HOFSTÄTTER, D. KETCHESON, AND O. KOCH, *Practical splitting methods for the adaptive integration of nonlinear evolution equations. Part* I: *Construction of optimized schemes and pairs of schemes*, BIT, 57 (2017), pp. 55–74.

[3] P. BADER AND S. BLANES, *Fourier methods for the perturbed harmonic oscillator in linear and nonlinear Schrödinger equations*, Phys. Rev. E, 83 (2011), 046711.

[4] K. BEAUCHARD AND F. MARBACH, *personnal communication*, 2019.

[5] N. BESSE AND M. MEHRENBERGER, *Convergence of classes of high-order semi-Lagrangian schemes for the Vlasov-Poisson system*, Math. Comp., 77 (2008), pp. 93–123.

[6] S. BLANES AND P. C. MOAN, *Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, J. Comput. Appl. Math., 142 (2002), pp. 313–330.

[7] S. BLANES, F. CASAS, AND A. MURUA, *Composition methods for differential equations with processing*, SIAM J. Sci. Comput., 27 (2006), pp. 1817–1843.

[8] J. P. BORIS, *Relativistic plasma simulation-optimization of a hybrid code*, in Proceedings of the Fourth Conference on Numerical Simulations of Plasmas held at the Naval Research Laboratory, Washington, DC, 1970.

[9] F. CALIFANO, F. PEGORARO, S. V. BULANOV, AND A. MANGENEY, *Kinetic saturation of the Weibel instability in a collisionless plasma*, Phys. Rev. E, 57 (1998), 704.

[10] F. CHARLES, B. DESPRES, AND M. MEHRENBERGER, *Enhanced convergence estimates for semi-Lagrangian schemes. Application to the Vlasov-Poisson equation*, SIAM J. Numer. Anal., 51 (2013), pp. 840–863.

[11] B. CHEN AND A. KAUFMAN, 3*D volume rotation using shear transformation*, Graphical Models, 62 (2000), pp. 308–322.

[12] Y. CHENG, I. M. GAMBA, F. LI, AND P. J. MORRISON, *Discontinuous Galerkin methods for Vlasov-Maxwell equations*, SIAM J. Numer. Anal., 52 (2014), pp. 1017–1049.

[13] S. A. CHIN AND E. KROTSCHECK, *Fourth-order algorithms for solving the imaginary-time Gross-Pitaevskii equation in a rotating anisotropic trap*, Phys. Rev. E, 72 (2005), 036705.

[14] N. Crouseilles, M. Mehrenberger, and E. Sonnendrücker, *Conservative semi-Lagrangian schemes for Vlasov equations*, J. Comput. Phys., 229 (2010), pp. 1927–1953.
[15] F. Casas, N. Crouseilles, E. Faou, and M. Mehrenberger, *High-order Hamiltonian splitting for Vlasov-Poisson equations*, Numer. Math., 135 (2017), pp. 769–801.
[16] N. Crouseilles, L. Einkemmer, and E. Faou, *Hamiltonian splitting for the Vlasov-Maxwell equations*, J. Comput. Phys., 283 (2015), pp. 224–240.
[17] N. Crouseilles, L. Einkemmer, and E. Faou, *An asymptotic preserving scheme for the relativistic Vlasov-Maxwell equations in the classical limit*, Comput. Phys. Commun., 209 (2016), pp. 13–26.
[18] F. Filbet, E. Sonnendrücker, and P. Bertrand, *Conservative numerical schemes for the Vlasov equation*, J. Comput. Phys., 172 (2001), pp. 166–187.
[19] L. Greengard and J.-Y. Lee, *Accelerating the nonuniform fast Fourier transform*, SIAM Rev., 46 (2004), pp. 443–454.
[20] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Ser. Comput. Math. 31, 2006.
[21] R. Horsin, *Comportement en temps long d'équations de type Vlasov: études mathématiques et numériques*, thèse de l'Université de Rennes I, 2017.
[22] A. J. Klimas and W. M. Farrell, *A splitting algorithm for Vlasov simulation with filamentation filtration*, J. Comput. Phys., 110 (1994), pp. 150–163.
[23] P. V. Koseleff, *Exhaustive search of symplectic integrators using computer algebra*, in Integration Algorithms and Classical Mechanics. Fields Inst. Commun. 10, AMS, Providence, RI, 1996, pp. 103–120.
[24] F. Nicola and L. Rodino, *Global Pseudo-Differential Calculus on Euclidean Spaces*, Pseudo Diff. Oper. 4, Birkhäuser Verlag, Basel, 2010.
[25] A. W. Paeth, *A fast algorithm for general raster rotation*, in Proceedings of Graphics Interface, Vancouver, 1986, pp. 77–81.
[26] E. Sonnendrücker, J. Roche, P. Bertrand, and A. Ghizzo, *The semi-Lagrangian method for the numerical resolution of the Vlasov equation*, J. Comput. Phys., 149 (1999), pp. 201–220.
[27] M. Suzuki, *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*, Phys. Lett. A., 146 (1990), pp. 319–323.
[28] A. Tanaka, *A rotation method for raster image using skew transformation*, in Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 1986, pp. 272–277.
[29] E. Weibel, *Spontaneously growing transverse waves in a plasma due to an anisotropic velocity distribution*, Phys. Rev. Lett., 2 (1959), 83.
[30] J. Welling, W. Eddy, and T. Young, *Rotation of 3D volumes by Fourier-interpolated shears* Graphical Models, 68 (2006), pp. 356–370.
[31] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A, 150 (1990), pp. 262–268.
[32] M. Zerroukat, N. Wood, and A. Staniforth, *The parabolic spline method (PSM) for conservative transport problems*, Internat. J. Numer. Methods Fluids, 51 (2006), pp. 1297–1318.