A QUASI-CONSERVATIVE DYNAMICAL LOW-RANK ALGORITHM FOR THE VLASOV EQUATION
LUKAS EINKEMMER*‡ AND CHRISTIAN LUBICH*

Abstract. Numerical methods that approximate the solution of the Vlasov–Poisson equation by a low-rank representation have been considered recently. These methods can be extremely effective from a computational point of view, but contrary to most Eulerian Vlasov solvers, they do not conserve mass and momentum, neither globally nor in respecting the corresponding local conservation laws. This can be a significant limitation for intermediate and long time integration. In this paper we propose a numerical algorithm that overcomes some of these difficulties and demonstrate its utility by presenting numerical simulations.

Key words. low-rank approximation, conservative methods, projector splitting, Vlasov–Poisson equation

1. Introduction. Many plasma systems that are of interest in applications (such as in magnetic confined fusion or astrophysics) cannot be adequately described by fluid models. Instead kinetic models have to be employed. Since these models are posed in a $2d$-dimensional ($d = 1, 2, 3$) phase space, numerically solving kinetic equations on a grid is extremely expensive from a computational point of view. Thus, traditionally, particle methods have been employed extensively to approximate these types of problems (see, for example, [43]). However, particle methods suffer from excessive noise that makes it, for example, difficult to resolve regions with low phase space density. Due to the increase in computer performance, methods that directly discretize phase space, the so-called Eulerian approach, have recently seen increased interest [42, 17, 11, 38, 39, 6, 16, 11, 13, 12]. However, performing these simulations in higher dimensions is still extremely expensive. As a consequence, much effort has been devoted to efficiently implement these methods on high performance computing systems [10, 14, 24, 31, 10, 1, 13].

More recently, methods that use a low-rank approximation have emerged. In [9, 23] the Vlasov equation is first discretized in time and/or space, and then low-rank algorithms are applied to the discretized system. A different approach is taken in [15], where a low-rank projector-splitting is on top of the procedure. That is, the low-rank algorithm is applied before any time or space discretization is chosen. This results in small systems of $d$-dimensional advection equations (in either the space or the velocity variables, in an alternating fashion) that are then solved by spectral or semi-Lagrangian methods. The advantage of this approach is that the evolution equations that need to be solved numerically are directly posed in terms of the degrees of freedom of the low-rank representation. Thus, no intermediate tensors have to be constructed and no tensor truncation algorithms have to be employed. This also leads to increased flexibility in the choice of the time and space discretization methods.

Computing numerical solutions of high-dimensional evolutionary partial differential equations by dynamical low-rank approximation has only recently been considered for kinetic problems [15, 14]. However, such algorithms have been investigated extensively in quantum mechanics; see, in particular, [34, 33] for the MCTDH approach to molecular quantum dynamics in the chemical physics literature and [25, 26] for a computational mathematics point of view of this approach. Some uses of dynamical low-rank approximation in areas outside quantum mechanics are described in [36, 26, 32, 35]. In a general mathematical setting, dynamical low-rank approximation has been studied in [21, 22, 29]. A major algorithmic advance for the time integration was achieved with the projector-splitting methods first proposed in [28] for matrix differential equations and then developed further for various tensor formats in [20, 27, 28, 30]. In contrast to standard time-stepping methods, the projector-splitting methods have been shown to be robust to the typical presence of small singular values in the low-rank approximation [20]. The approach in [15, 14] and in the present paper is based on an adaptation of the projector-splitting method of [28] to kinetic equations.

While low-rank approximations can be very effective from a computational point of view, they destroy much of the physical structure of the problem under consideration. Important physical invariants, such as mass and momentum, are no longer conserved. Perhaps even more problematic is that the low-rank approximation does not take the corresponding local conservation laws into account. This can be a significant limitation for intermediate and long time integration. In this paper we propose a numerical algorithm that overcomes some of these difficulties and demonstrate its utility by presenting numerical simulations.

*Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D–72076 Tübingen, Germany. Email: {einkemmer,lubich}@na.uni-tuebingen.de
‡Department of Mathematics, University of Innsbruck, Austria
issue if these algorithms are to be used for long or even intermediate time integration.

This situation is in stark contrast with the state of the art for Eulerian Vlasov solvers, where significant
research has been conducted to conserve certain physical properties of the exact solution \[17, 11, 57, 5, 2, 3, 12\]. In particular, methods that conserve mass and momentum are commonly employed. However, to the
best of our knowledge, no low-rank algorithms are available that are able to conserve even linear invariants. Furthermore, it has recently been proposed to use low-rank numerical methods to solve fluid problems \[14\]. Also in this setting conservation of mass and momentum, a hallmark of traditional fluid solvers, is, of course, of great interest.

In this paper we will consider the Vlasov–Poisson equation

\[
\begin{align*}
\partial_t f(t, x, v) + v \cdot \nabla_x f(t, x, v) - E(f)(x) \cdot \nabla_v f(t, x, v) &= 0 \\
\nabla \cdot E(f)(x) &= -\int f(t, x, v) \, dv + 1, \\
\n\nabla \times E(f)(x) &= 0,
\end{align*}
\]  

(1.1)

which models the time evolution of a collisionless plasma in the electrostatic regime. This equation has
an infinite number of invariants (Casimir invariants). Here we will consider the linear invariants of mass
and momentum and the corresponding local conservation laws. In section 2, we will introduce the necessary
notation and describe the dynamical low-rank splitting algorithm for the Vlasov equation that was proposed
in \[15\]. We then derive a modification of that numerical method such that a projected version of the
continuity and momentum balance equation is satisfied (section 3). Subsequently we will discuss the global
conservation of mass and momentum in section 4. We will then consider the efficient implementation of these
methods (section 5). Finally, in section 6 we present numerical results for the Vlasov–Poisson equation. In
particular, we will demonstrate the efficiency of the proposed algorithms for a two-stream instability.

2. A low-rank projector-splitting integrator. We will start by summarizing the low-rank projector
splitting integrator for the Vlasov–Poisson equation introduced in \[15\]. It should be duly noted that this
algorithm neither respects the local conservation laws associated with mass or momentum, nor conserves
mass or momentum globally (this is also true for low-rank algorithms in \[9, 23\]).

We seek an approximation to the Vlasov–Poisson equation (1.1) in the following form:

\[
f(t, x, v) = \sum_{i,j=1}^{r} X_i(t, x) S_{ij}(t) V_j(t, v),
\]

with real coefficients \(S_{ij}(t) \in \mathbb{R}\) and with functions \(X_i(t, x)\) and \(V_j(t, v)\) that are orthonormal:

\[
\langle X_i, X_k \rangle_x = \delta_{ik} \quad \text{and} \quad \langle V_j, V_l \rangle_v = \delta_{jl},
\]

where \(\langle \cdot, \cdot \rangle_x\) and \(\langle \cdot, \cdot \rangle_v\) are the inner products on \(L^2(\Omega_x)\) and \(L^2(\Omega_v)\), respectively. The dependence of \(f\) on
the phase space variables \((x, v) \in \Omega = \Omega_x \times \Omega_v \subset \mathbb{R}^{2d}\) is approximated by the functions \(\{X_i: i = 1, \ldots, r\}\)
and \(\{V_j: j = 1, \ldots, r\}\), which depend only on the separated variables \(x \in \Omega_x\) and \(v \in \Omega_v \subset \mathbb{R}^d\), respectively.

Such an approach is efficient if the rank \(r\) can be chosen much smaller compared to the number of grid points
used to discretize \(X_i\) and \(V_j\) in space.

The dynamics of the Vlasov–Poisson equation is constrained to the corresponding low-rank manifold by
replacing (1.1) with an evolution equation

\[
\partial_t f = -P(f) (v \cdot \nabla_x f - E(f) \cdot \nabla_v f),
\]

where \(P(f)\) is the orthogonal projector onto the manifold. The projector can be written as

\[
P(f) g = P_{\mathcal{X}} g - P_{\mathcal{V}} P_{\mathcal{X}} g + P_{\mathcal{V}} g, \quad (2.1)
\]

where \(P_{\mathcal{X}}\) is the orthogonal projector onto the vector space \(\mathcal{X} = \text{span} \{X_i: i = 1, \ldots, r\}\) and \(P_{\mathcal{V}}\) is the orthogonal projector onto the vector space \(\mathcal{V} = \text{span} \{V_j: j = 1, \ldots, r\}\). Then, as first suggested in \[28\],
the dynamics is split into the three terms of equation (2.1). In the simplest case, the first-order Lie-Trotter splitting, we solve the equations

\[
\begin{align*}
\partial_t f &= -P_{\tau} (v \cdot \nabla_x f - E(f) \cdot \nabla_v f) \\
\partial_t f &= +P_{\tau} P_{\Omega} (v \cdot \nabla_x f - E(f) \cdot \nabla_v f) \\
\partial_t f &= -P_{\Omega} (v \cdot \nabla_x f - E(f) \cdot \nabla_v f)
\end{align*}
\]

(2.2) \hspace{1cm} (2.3) \hspace{1cm} (2.4)

one after the other. Now, let us define

\[
K_j(t, x) = \sum_i X_i(t, x) S_{ij}(t), \quad L_i(t, v) = \sum_j S_{ij}(t) V_j(t, v).
\]

The advantage of the splitting scheme then becomes that equation (2.2) only updates \(K_j\) (the \(V_j\) stay constant during that step), equation (2.3) only updates \(S_{ij}\) (the \(X_i\) and \(V_j\) stay constant during that step), and equation (2.4) only updates \(L_i\) (the \(X_i\) stay constant during that step). The corresponding evolution equations are derived in [15] and are of the following form:

\[
\begin{align*}
\partial_t K_j(t, x) &= -\sum_i c_{ij}^1 \cdot \nabla_x K_i(t, x) + \sum_l c_{ij}^2 \cdot E(K)(t, x) K_l(t, x) \\
\partial_t S_{ij}(t) &= \sum_{k,l} (c_{ij}^1 \cdot d_{ik}^2 - c_{ij}^2 \cdot d_{ik}^1) [E(S(t))] S_{kl}(t) \\
\partial_t L_i(t, v) &= \sum_k d_{ik}^1 [E(L(t, \cdot))] \cdot \nabla_v L_k(t, v) - \sum_k (d_{ik}^2 \cdot v) L_k(t, v)
\end{align*}
\]

(2.5) \hspace{1cm} (2.6) \hspace{1cm} (2.7)

The coefficients \(c_{ij}^1, c_{ij}^2\) and \(d_{ik}^1, d_{ik}^2\) are vector-valued but constant in space and, with the exception of \(d_{ik}^1\), also constant in time. They are given by integrals over \(\Omega_v\) and \(\Omega_x\), respectively; see [15] Section 2 for the details.

Assuming that the initial value is represented as \(f^0(x, v) = \sum_{i,j} X_i^0(x) S_{ij}^0 V_j^0(v)\), the algorithm with time step size \(\tau\) then proceeds in the following three steps.

**Step 1:** Solve equation (2.5) with initial value \(K_j(0, x) = K_j^0 = \sum_i X_i^0(x) S_{ij}^0\). Then perform a QR decomposition of \(K^1 = [K_1(\tau, \cdot), \ldots, K_r(\tau, \cdot)]\) to obtain \(X_1\) and \(S_{ij}^1\).

**Step 2:** Solve equation (2.6) with initial value \(S_{ij}(0) = S_{ij}^0\) to obtain \(S_{ij}^0 = S_{ij}^1\).

**Step 3:** Solve equation (2.7) with initial value \(L_i(0, v) = L_i^0 = \sum_j S_{ij}^0 V_j^0\). Then perform a QR decomposition of \(L^1 = [L_1(\tau, \cdot), \ldots, L_r(\tau, \cdot)]\) to obtain \(V_1\) and \(S_{ij}^1\).

The output of the algorithm is then the low-rank representation

\[
f(\tau, x, v) \approx f^1(x, v) = \sum_{i,j} X_i^1 S_{ij}^1 V_j^1.
\]

(2.8)

For a detailed derivation of this algorithm the reader is referred to [15]. We note that the extension to second order Strang splitting is immediate.

**3. Local conservation.** The Vlasov–Poisson equation (1.1) satisfies the continuity equation

\[
\partial_t \rho(t, x) + \nabla \cdot (\rho(t, x) u(t, x)) = 0
\]

(3.1)

and the momentum balance equation

\[
\partial_t (\rho(t, x) u(t, x)) + \nabla \cdot (\rho(t, x) u(t, x) \otimes u(t, x)) = -E(t, x) \rho(t, x),
\]

(3.2)

where

\[
\rho(t, x) = \int f(t, x, v) \, dv, \quad \rho(t, x) u(t, x) = \int v f(t, x, v) \, dv.
\]

(3.3)
From these equations, global conservation of mass and momentum is easily obtained by integrating in \(x\). Without the projection operators, equations (2.2)–(2.4) would satisfy the continuity equation (3.1) and the momentum balance equation (3.2). Overall this would ensure that the splitting scheme (without projection operators) respects the local conservation laws for mass and momentum. However, it can easily be seen that the projection operators destroy this property. In addition, as has already been pointed out in [15], global conservation of mass and momentum is lost as well.

A crucial observation that enables the following numerical method is the observation that the conserved quantities only depend on \(x\). While we cannot modify the algorithm such that the conservation laws are satisfied exactly (while keeping \(V_j\) constant in Step 1 and \(X_i\) constant in Step 3, and both \(X_i\) and \(V_j\) constant in Step 2), our goal is to derive a numerical method that satisfies the projected conservation laws for mass and momentum

\[
P_{X}(\partial_t \rho + \nabla \cdot (\rho u)) = 0, \quad P_{X}(\partial_t \rho + \nabla \cdot (\rho u \otimes u) + E\rho) = 0. \tag{3.3}
\]

The idea is to add to (2.2)–(2.4) corrections of the form

\[
\sum_{i,j} \lambda_{ij} X_i V_j, \tag{3.4}
\]

where the coefficients \(\lambda_{ij}\) are determined such that the projected continuity equation and the projected momentum balance hold true. This results in an overdetermined system for the \(\lambda_{ij}\) for which we seek the smallest solution in the Euclidean norm.

One might object at this point and argue that such a correction is unnecessarily restrictive. Certainly, one could envisage that for equation (2.5) and (2.7) an arbitrary function of \(x\) and \(v\), respectively, could be used as the correction. Unfortunately, as we will describe in more detail in Remark 2, this would introduce, for example, non-zero values in the density function at high velocities. This, clearly unphysical, artefact then pollutes the numerical solution. Thus, the benefit of the ansatz given in equation (3.4) is that the \(X_i\) and \(V_j\), which are already used to represent the numerical solution, are also used for the correction. Since the algorithm adapts the functions \(X_i\) and \(V_j\) in accordance with the solution, the artefact described above is avoided. This behavior is confirmed by numerical simulation.

In the following, the correction given in (3.4) will be made precise for the three steps of the splitting algorithm.

**Step 1:** We replace the evolution equation (2.5) by

\[
\partial_t K_j = \left\langle V_j^0, F(f) + \sum_{k,l} \lambda_{kl} X_k^0 V_l^0 \right\rangle_v
= \langle V_j^0, F(f) \rangle_v + \sum_k \lambda_{kj} X_k^0 \tag{3.5}
\]

with \(F(f) = -v \cdot \nabla_x f + E(f) \cdot \nabla_v f\) for \(f(t,x,v) = \int K_i(t,x) V_i^0(v)\), and \(\lambda_{kl}\) is yet to be determined. Note that the \(V_i\) are constant during that time step, and hence \(F(f)\) only depends on the \(K_i\). Now, we impose

\[
0 = P_{X} (\partial_t \rho + \nabla \cdot (\rho u)) = \sum_i X_i^0 \left[ \sum_j \lambda_{ij} \alpha_j + \sum_j \langle X_i^0 V_j^0, F(f) \rangle_{xv} \alpha_j + \langle X_i^0, \nabla \cdot (\rho u) \rangle_x \right], \tag{3.6}
\]

where \(\alpha_j = \int V_j^0 dv\), and

\[
0 = P_{X} (\partial_t (\rho u) + \nabla \cdot (\rho u \otimes u) + E\rho)
= \sum_i X_i^0 \left[ \sum_j \lambda_{ij} \beta_j + \sum_j \langle X_i^0 V_j^0, F(f) \rangle_{xv} \beta_j + \langle X_i^0, \nabla \cdot (\rho u \otimes u) \rangle_x + \langle X_i^0, E(f) \rangle_x \right], \tag{3.7}
\]

where \(\beta_j = \int v V_j^0 dv \in \mathbb{R}^d\). Together, equations (3.6) and (3.7) yield \((1 + d)r\) linear equations for the \(r^2\) unknowns \(\lambda_{ij}\) (We suppose \(r \geq 1 + d\) in the following). Since the equations for different \(i\) decouple, this
allows us to put this into matrix form as follows, with the row vector \( \alpha = (\alpha_1, \ldots, \alpha_r) \) and with the \( d \times r \) matrix \( \beta = (\beta_1, \ldots, \beta_r) \):

\[
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}
\lambda_{i(\cdot)} =
\begin{bmatrix}
b_i \\
d_i
\end{bmatrix}
\]

(3.8)

with

\[
b_i = -\sum_j \langle X_i^0 V_j^0, F(f) \rangle_{xv} \alpha_j - \langle X_i^0, \nabla \cdot (\rho u) \rangle_x
\]

\[
d_i = -\sum_j \langle X_i^0 V_j^0, F(f) \rangle_{xv} \beta_j - \langle X_i^0, \nabla \cdot (\rho u) \rangle_x - \langle X_i^0, E(f) \rangle_x.
\]

These systems of equations have (multiple) solutions if the rows of the matrix \([\alpha; \beta]\) are linearly independent. In order to minimize the magnitude of the correction that is applied, we seek the solution with the smallest Euclidean norm. This can be done easily and at negligible cost as the matrix is only of size \((1 + d) \times r\).

It is still necessary to compute the right hand side. We have

\[
\nabla \cdot (\rho u) = \sum_j \nabla K_j \cdot \beta_j,
\]

\[
\nabla \cdot (\rho u \otimes u) = \sum_j \nabla K_j \cdot \gamma_j,
\]

where \( \gamma_j = \int (v \otimes v) V_j^0 \, dv \). Since \( E \) and \( \rho \) have to be computed in any case and \( \alpha, \beta, \gamma \), on modern computer architectures, can be computed alongside the coefficients \( c^1 \) and \( c^2 \) at (almost) no extra cost, only the projections in \( x \) are of any concern from a computational point of view. These require the computation of \( r \) integrals and consequently \( O \left( r n^d \right) \) arithmetic operations when \( n \) quadrature points are used in each coordinate direction.

**Step 2:** We replace the evolution equation (2.6) by

\[
\partial_t S_{ij} = - \langle X_i^1 V_j^0, F(f) + \sum_{k,l} \lambda_{kl} X_k^1 V_l^0 \rangle_{xv}
\]

\[
= -\langle X_i^1 V_j^0, F(f) \rangle_{xv} - \lambda_{ij}
\]

(3.9)

for \( f(t, x, v) = \sum_{k,l} X_k^0(x) S_{kl}(t) V_l^0(v) \), so that \( F(f) \) depends only on the \( S_{kl} \), and where \( \lambda_{ij} \) is yet to be determined. Then we impose the constraints

\[
0 = P_{X^1} (\partial_t \rho - \nabla \cdot (\rho u)) = -\sum_i X_i^1 \left[ \sum_j \lambda_{ij} \alpha_j + \sum_j \langle X_i^1 V_j^0, F(f) \rangle_{xv} \alpha_j + \langle X_i^1, \nabla \cdot (\rho u) \rangle_x \right]
\]

(3.10)

with \( \alpha_j = \int V_j^0 \, dv \) and

\[
0 = P_{X^1} (\partial_t (\rho u) - \nabla \cdot (\rho u \otimes u) - E(\rho))
\]

\[
= -\sum_i X_i^1 \left[ \sum_j \lambda_{ij} \beta_j + \sum_j \langle X_i^1 V_j^0, F(f) \rangle_{xv} \beta_j + \langle X_i^1, \nabla \cdot (\rho u \otimes u) \rangle_x + \langle X_i^1, E(\rho) \rangle_x \right]
\]

(3.11)

with \( \beta_j = \int v V_j^0 \, dv \). Equations (3.10) and (3.11) yield \( (1 + d)r \) linear equations for the \( r^2 \) unknowns \( \lambda_{ij} \). Since the equations for different \( i \) decouple, we once again can put this into the form given by equation (3.8).

The only difference lies in the right hand side which is computed as follows:

\[
b_i = -\sum_j \langle X_i^1 V_j^0, F(f) \rangle_{xv} \alpha_j - \langle X_i^1, \nabla \cdot (\rho u) \rangle_x
\]

\[
d_i = -\sum_j \langle X_i^1 V_j^0, F(f) \rangle_{xv} \beta_j - \langle X_i^1, \nabla \cdot (\rho u) \rangle_x - \langle X_i^1, E(\rho) \rangle_x.
\]
As before, we seek the solution that minimizes the Euclidean norm of the $\lambda_{ij}$. This can be done efficiently as we only have to solve $r$ systems of size $(1 + d) \times r$. Computing the right-hand side requires $\langle X_1^1 V_0^0, F(f) \rangle_{xv}$, which has to be computed to conduct this splitting step in any case. Thus, only the projections in $\lambda$ with $i$ in each coordinate direction.

Step 3: We replace the evolution equation (2.5) by

$$
\partial_t L_i = \left( X_1^1, F(f) + \sum_{kl} \lambda_{kl} X_k^1 V_l^0 \right) x
= \langle X_1^1, F(f) \rangle_x + \sum_l \lambda_{il} V_l^0
$$

(3.12)

for $f(t,x,v) = \sum_k X_k^1(x) L_k(t,v)$, so that $F(f)$ depends only on the functions $L_k$, and where $\lambda_{ij}$ is yet to be determined. Then we impose the constraints

$$
0 = P_{X^1} (\partial_t \rho + \nabla \cdot (\rho u)) = \sum_i X_i^1 \left[ \sum_l \lambda_{il} \alpha_l + \langle X_1^1, F(f) \rangle_{xv} + \langle X_1^1, \nabla \cdot (\rho u) \rangle_x \right]
$$

(3.13)

with $\alpha_l = \int V_l^0 \, dv$ and

$$
0 = P_{X^1} (\partial_t (\rho u) + \nabla \cdot (\rho u \otimes u) + E \rho)
= \sum_i X_i^1 \left[ \sum_l \lambda_{il} \beta_l + \langle X_1^1, v F(f) \rangle_{xv} + \langle X_1^1, \nabla \cdot (\rho u) \rangle_x - \langle X_1^1, E \rho \rangle_x \right]
$$

(3.14)

with $\beta_l = \int v V_l^0 \, dv$. As before, equations (3.13) and (3.14) yield $(1 + d)r$ linear equations for the $r^2$ unknowns $\lambda_{ij}$. We can once again put this into the form given by equation (3.8) with right-hand side

$$
b_i = -\langle X_1^1, F(f) \rangle_{xv} - \langle X_1^1, \nabla \cdot (\rho u) \rangle_x
$$

$$
d_i = -\langle X_1^1, v F(f) \rangle_{xv} - \langle X_1^1, \nabla \cdot (\rho u) \rangle_x + \langle X_1^1, E \rho \rangle_x.
$$

As before, this can be done efficiently as the matrix involved is small and the right-hand side can be efficiently computed alongside the coefficients that are needed for the low-rank splitting algorithm.

Note that in the third step we have

$$
b_i = \sum_k \langle X_1^1, \nabla X_k^1 \rangle_{xv} \cdot \int v L_k \, dv - \sum_k \langle X_1^1, E X_k^1 \rangle_{xv} \int \nabla v L_k \, dv - \sum_k \langle X_1^1, \nabla X_k^1 \rangle_{xv} \cdot \int v L_k \, dv
= 0,
$$

where we have assumed that the $L_k$ go to zero as $|v| \to \infty$. Thus, step 3 already satisfies the continuity equation.

Remark 1. At first sight it looks more natural to use $K_1$ and $L_i$ instead of $X_0^0$ and $V_i^0$ in step 1 and 3. These are the quantities that are updated in that step of the algorithm. The correction would then also reflect the corresponding changes that occur as the subflows are advanced in time. However, note that in actual numerical simulations $S$ can be very ill-conditioned. Now, since $K_j = \sum_i X_i S_{ij}$, the smallest singular value of $K = (K_1, \ldots, K_r)$ is equal to that of $S$. Specifically, this is a problem for momentum conservation as many problems start with zero or very small momentum. This then changes over time as the algorithm selects appropriate basis functions which carry a non-zero momentum. However, since initially the contribution of these functions to $K$ (contrary to $X$) is very small, the coefficients in the correction have to become large. This implies that the correction overall becomes quite large. Choosing $X_0^0$ instead of $K_j$, as we have done here, solves this issue. The situation is analogous for $V_i^0$ and $L_i$.

Remark 2. Let us now consider a correction $R_i(v)$ for equation (2.5)

$$
\partial_t L_i = \langle X_1^1, F(f) \rangle_x + R_i(v).
$$
This correction is more general than the ansatz we made in equation (3.4). As before, our goal is to determine the smallest $R_i(v)$ such that the local conservation laws are satisfied. Since $R_i(v)$ has more degrees of freedom, after the space discretization has been performed, in principle, a smaller correction could be obtained. Thus, this seems like a promising approach. In this remark we will restrict ourselves, for simplicity, only to the continuity equation. To obey the continuity equation the correction has to satisfy

$$\int R_i(v) \, dv = - \langle X_i^1, F(f) \rangle_{xv} - \langle X_i^1, \nabla \cdot (\rho u) \rangle_x.$$ 

Minimizing the correction in the $L^2$ norm immediately yields

$$R_i = \frac{-1}{|\Omega_v|} \left[ \langle X_i^1, F(f) \rangle_{xv} + \langle X_i^1, \nabla \cdot (\rho u) \rangle_x \right],$$

where $|\Omega_v|$ is the volume of the domain in the $v$-direction. Note, in particular, that $R_i$ is independent of $v$. Thus, the correction equally distributes the defect in velocity space. In the case of the Vlasov equation, however, the density function $f$ is expected to decay to zero for large velocities. On the other hand, the described correction would introduce non-zero densities for large velocities, which is clearly an unphysical artefact. The correction considered in this paper, i.e. equation (3.4), only allows linear combination of $V_i^0$. This avoids the problems stated above as the $V_i^0$ are already used to represent the numerical solution and thus decay to zero. In fact, any property of the $V_i^0$ that is invariant under taking linear combinations, is preserved by our approach.

4. Global conservation. The algorithm developed above satisfies a projected version of the local conservation law. For mass conservation this is stated as

$$P_X (\partial_t \rho + \nabla \cdot (\rho u)) = 0.$$ 

However, contrary to the continuous formulation, conservation of mass cannot be deduced from this expression by simply integrating in $x$. In fact, conservation of mass, in general, is violated for the scheme described in the previous section. The situation for momentum is similar.

Since we have an underdetermined system of equations it is, in principle, possible to add an equation that enforces global conservation of mass and momentum. This has to be done for each step in the splitting algorithm.

Step 1: We impose

$$0 = \partial_t \int \rho \, dx = \sum_{ij} \kappa_i \lambda_{ij} \alpha_j + \sum_j \langle V_j^0, F(f) \rangle_{xv},$$

where $\alpha_j = \int V_j^0 \, dv$ and $\kappa_i = \int X_i^0 \, dx$, and

$$0 = \partial_t \int \rho u \, dx = \sum_{ij} \kappa_i \lambda_{ij} \beta_j + \sum_j \langle V_j^0, F(f) \rangle_{xv} \beta_j,$$

where $\beta_j = \int v V_j^0 \, dv$. This adds $1 + d$ linear equations to the $2(1 + d)r$ linear equations (3.8) required for the local conservation laws. Note that in contrast to these equations all the $\lambda_{ij}$ are coupled to each other. Thus, we have to solve a single system of size $(1 + d)(2r + 1) \times r^2$. We will discuss the computational ramifications later in this section.

Step 2: We impose

$$0 = -\partial_t \int \rho \, dx = \sum_{ij} \kappa_i \lambda_{ij} \alpha_j + \sum_{ij} \kappa_i \langle X_i^1 V_j^0, F(f) \rangle_{xv} \alpha_j,$$

where $\alpha_j = \int V_j^0 \, dv$ and $\kappa_i = \int X_i^1 \, dx$, and

$$0 = -\partial_t \int \rho u \, dx = \sum_{ij} \gamma_i \lambda_{ij} \beta_j + \sum_{ij} \kappa_i \langle X_i^1 V_j^0, F(f) \rangle_{xv} \beta_j.$$
where $\beta_j = \int vV_j^0 \, dv$.

**Step 3:** We impose

$$0 = \partial_t \int \rho \, dx = \sum_{ij} \kappa_i \lambda_{ij} \alpha_j + \sum_i \kappa_i \langle X_1^i, F(f) \rangle_{xv},$$

where $\alpha_j = \int V_j^0 \, dv$ with $\kappa_i = \int X_1^i \, dx$, and

$$0 = \partial_t \int \rho u \, dx = \sum_{ij} \kappa_i \lambda_{ij} \beta_j + \sum_i \kappa_i \langle vX_1^i, F(f) \rangle_{xv},$$

where $\beta_j = \int vV_j^0 \, dv$.

The problem with this approach is that there is no guarantee that the resulting linear system even has a solution. This is most easily demonstrated by considering step 3 in our algorithm. In this case $b_i = 0$ (see section 3). Now, let us consider the rank 2 function on the domain $[0, 2\pi] \times \mathbb{R}$ given by

$$X_1^1(x) = \frac{2}{\sqrt{3\pi}} \cos^2 x, \quad X_2^1(x) = \frac{1}{\sqrt{\pi}} \sin(2x), \quad V_1^0(v) = e^{-v^2} (\pi/2)^{1/4}, \quad V_2^0(v) = 2v e^{-v^2} (\pi/2)^{1/4}.$$  

This gives $\alpha = (\sqrt{2\pi}, 0)$ and $\kappa = (2\sqrt{\pi}/3, 0)$. Thus,

$$\lambda_{11} = 0.$$  

Since $\beta_1 = 0$, we have

$$\langle X_1^1, F(f) \rangle_{xv} = \langle X_1^1, \nabla X_2^1 \rangle \cdot \beta_2$$  

$$\propto \int \cos^2(x) \cos(2x) \, dx$$

$$\neq 0.$$  

This is in contradiction to the condition of global mass conservation. Thus, it is not possible to both satisfy the continuity equation and obtain global conservation of mass. We have only considered conservation of mass here, but the same behavior is observed for the momentum as well. We have the following options:

**Local:** We enforce only the local conservation laws, while minimizing the Euclidean norm of the correction.

**Global:** We enforce only the global conservation laws, while minimizing the Euclidean norm of the correction.

**Combined:** We try to find the best approximation to both the local conservation laws and the global conservation of mass and momentum. This results in a linear least square problem for the correction. The different equations can be weighted to either focus on the local conservation laws or the global conservation of mass and momentum.

All of these configurations will be considered in section 6. However, before proceeding let us discuss the computational cost of the combined approach. We have to compute an underdetermined (but incompatible) linear least square problem with $r^2$ unknowns and $(1 + d)(2r + 1)$ data. This problem can be solved by computing the Moore–Penrose pseudo-inverse which requires a QR decomposition of $A^T$. Thus, it requires at most $O(r^4)$ arithmetic operations which is typically small compared to the cost of the low-rank algorithm itself.

5. **Efficient implementation.** In the proposed algorithm correction terms are added to the three evolution equations. This implies that our correction is a continuous function of time for the respective subflows. However, in order to increase performance it is often of interest to use a specifically tailored numerical method for solving these subflows. For example, methods based on fast Fourier techniques (FFT) and semi-Lagrangian schemes have been proposed in [15]. To employ these algorithms while still maintaining the conservation laws for mass and momentum is not necessarily straightforward. Thus, we will now introduce
a procedure that allows us to apply our correction independent of the specific time integration strategy that is chosen for solving the evolution equations (3.5), (3.9), (3.12). The approach outlined here is similar to the projection schemes described in [8].

We start with the evolution equation for $K_j$ which is given as follows

$$\partial_t K_j = \langle V_j^0, F(f) \rangle_v + \sum_k \lambda_{kj} X^0_k, \quad \text{for} \quad f = \sum_l K_l V^0_l.$$  

Now, we split this equation into

$$\partial_t K_j = \langle V_j^0, F(f) \rangle_v$$

and

$$\partial_t K_j = \sum_k \lambda_{kj} X^0_k.$$  

Equation (5.1) is identical to what has to be solved in case of the original low-rank algorithm described in section 3 (i.e. the algorithm without correction). Thus, starting from an appropriate initial value $K^0_j$ we compute an approximation at time $\tau$, where $\tau$ is the time step size. This value is henceforth denoted by $K^*_j$. Now, instead of solving equation (5.2) we consider the following approximation

$$\frac{K^*_j - K^0_j}{\tau} = \sum_k \lambda_{kj} X^0_k.$$  

It remains to derive the conditions on $\lambda_{kj}$ under which the (discretized) conservation laws are satisfied. We have

$$\rho^1 - \rho^0 + \tau \nabla \cdot (\rho^0 u^0) = \sum_j (K^*_j - K^0_j) \alpha_j + \tau \sum_j (\nabla K^0_j) \cdot \beta_j$$

$$= \sum_j (K^*_j - K^0_j) + \tau \sum_k \lambda_{kj} X^0_k \alpha_j + \tau \sum_j (\nabla K^0_j) \cdot \beta_j,$$

where $\alpha_j = \int V^0_j \, dv$ and $\beta_j = \int v V^0_j \, dv$. Now, we apply the projection onto $X^0$ to obtain

$$0 = P_{X^0} (\rho^1 - \rho^0 + \tau \nabla \cdot (\rho^0 u^0)) = \sum_k X^0_k \left[ \tau \sum_k \lambda_{kj} \alpha_j + \sum_j \langle X^0_k, K^*_j - K^0_j \rangle x \alpha_j + \tau \sum_j \langle X^0_k, \nabla K^0_j \rangle x \cdot \beta_j \right].$$

This is the analogue to equation (3.6).

For the momentum balance equation we have

$$\rho^1 u^1 - \rho^0 u^0 + \tau \nabla \cdot (\rho^0 u^0 \otimes u^0) + \tau E^0 \rho^0 = \sum_j (K^*_j - K^0_j) \beta_j + \tau \sum_j (\nabla K^0_j) \cdot \gamma_j + \tau E^0 \rho^0$$

$$= \sum_j (K^*_j - K^0_j) + \tau \sum_k \lambda_{kj} X^0_k \beta_j + \tau \sum_j (\nabla K^0_j) \cdot \gamma_j + \tau E^0 \rho^0,$$

where $\gamma_j = \int (v \otimes v) V^0_j \, dv$ and we have used $E^0$ to denote the electric field at the beginning of the time step. Applying the projection onto $X^0$ we obtain

$$0 = P_{X^0} (\rho^1 u^1 - \rho^0 u^0 + \tau \nabla \cdot (\rho^0 u^0 \otimes u^0) + \tau E^0 \rho^0)$$

$$= \sum_k X^0_k \left[ \tau \sum_j \lambda_{kj} \beta_j + \sum_j \langle X^0_k, K^*_j - K^0_j \rangle x \beta_j + \tau \sum_j \langle X^0_k, \nabla K^0_j \rangle x \cdot \gamma_j + \tau \langle X^0_k, E^0 \rho^0 \rangle x \right].$$
where is the analogue to equation [3.7]

In fact, these equations are precisely in the form of [3.8]. Only the right-hand side

\[
\begin{align*}
    b_i &= -\sum_j \left( X_k^0, \frac{K_j^0 - K_0^0}{\tau} \right)_x \alpha_j - \sum_j \langle X_k^0, \nabla K_j^0 \rangle_x : \beta_j \\
    d_i &= -\sum_j \left( X_k^0, \frac{K_j^0 - K_0^0}{\tau} \right)_x \beta_j - \sum_j \langle X_k^0, \nabla K_j^0 \rangle_x : \gamma_j - \sum_j \langle X_k^0, E^0 K_j^0 \rangle_x \alpha_j
\end{align*}
\]

is modified. Thus, there is no additional difficulty in implementing this approach.

A similar procedure can be applied to step 2 and 3 of the splitting algorithm. For step 2 we obtain equation [3.8] with

\[
\begin{align*}
    b_i &= \sum_j \frac{S_{ij}^* - S_{ij}^\dagger}{\tau} \alpha_j - \sum_j \langle X_j^1, \nabla X_j^1 \rangle_x S_{ij} \beta_j \\
    d_i &= \sum_j \frac{S_{ij}^* - S_{ij}^\dagger}{\tau} \beta_j - \sum_j \langle X_j^1, \nabla X_j^1 \rangle_x S_{ij} \gamma_j - \sum_j \langle E^0, X_j^1 \rangle_x S_{ij} \alpha_j
\end{align*}
\]

and for step 3 we obtain equation [3.8] with (denoting the integral \( \langle g \rangle_v = \int g \, dv \))

\[
\begin{align*}
    b_i &= -\left( \frac{L_j^* - L_j^0}{\tau} \right)_v \langle X_j^1, \nabla X_j^1 \rangle_x \cdot (v L_j^0) \\
    d_i &= -\left( \frac{L_j^* - L_j^0}{\tau} \right)_v \langle X_j^1, \nabla X_j^1 \rangle_x \cdot \langle (v \otimes v) L_j^0 \rangle_v - \sum_j \langle X_j^1, E X_j^1 \rangle_x \langle L_j^0 \rangle_v.
\end{align*}
\]

Thus, we are able to apply the procedure introduced in section 3 independent of the specific numerical discretization. This has the added benefit that the correction and the associated coefficients only need to be computed once for each step of the splitting algorithm. The only downside here is that we have traded the continuous version of the conservation laws for a discretized version.

6. Numerical results. In this section we will present numerical results for a two-stream instability. Specifically, we consider the domain \([0, 10\pi] \times [-9, 9]\) and impose the initial value

\[
    f_0(x, v) = \frac{1}{2\sqrt{2\pi}} \left( e^{-\frac{(v-v_0)^2}{2}} + e^{-\frac{(v+v_0)^2}{2}} \right) (1 + \alpha \cos(kx)),
\]

where \(\alpha = 10^{-3}\), \(k = \frac{1}{5}\), and \(v_0 = 2.4\). Periodic boundary conditions are used in both the \(x\)- and the \(v\)-direction. This setup models two beams propagating in opposite directions and is an unstable equilibrium. Small perturbations in the initial particle-density function eventually force the electric energy to increase exponentially. This is called the linear regime. At some later time saturation sets in (the nonlinear regime). This phase is characterized by nearly constant electric energy and significant filamentation of the phase space. This test problem has been considered in [9, 23] and [15] in the context of low-rank approximations.

It has been established there that low-rank approximations of relatively small rank are sufficient in order to resolve the linear regime. However, once saturation sets in, the reference solution (computed using a full grid simulation) shows only small oscillations in the electric field. For the low-rank approximation, however, oscillations with significant amplitude can be observed. Since filamentation makes it very difficult to efficiently resolve the small structures in this regime (the \(L^\infty\) error will be large for any numerical method), we consider it a good test example for the conservative method developed in this work.

In Figure 6.1 numerical simulation of the two-stream instability for rank \(r = 10\) are shown for the algorithm without correction (labeled low-rank), the correction that exactly satisfies the local projected continuity equations described in section 4 (labeled local), the algorithm of section 4 that combines both local and global corrections (labeled combined), and the algorithm that conserves mass and momentum exactly.
Figure 6.1. Numerical simulations of the two-dimensional two-stream instability with rank $r = 10$ are shown. The Strang splitting algorithm with a time step size $\tau = 0.025$ is employed. In both the $x$ and $v$-directions 128 grid points are used. As a comparison, a direct Eulerian simulation (based on a spectral method) is also shown.

but does not satisfy the local continuity equations (labeled global). In addition, the full grid simulation is shown (labeled full grid). We observe that all methods show excellent agreement in the linear regime. In the nonlinear regime the local correction shows the best performance (the least amount of oscillations). The performance of the combined approach is also significantly better compared to the uncorrected algorithm and the global correction. The uncorrected algorithm clearly performs worst.

Figure 6.1 also shows the error in mass, momentum, energy, and the $L^2$ norm. We see that although the local correction results in a significant improvement with respect to the qualitative behavior of the electric field, the errors in mass and momentum are still comparable to the uncorrected algorithm. As has been discussed in section 4, in general, satisfying both the local continuity equations and the global invariants is
Figure 6.2. Numerical simulations of the two-dimensional two-stream instability with rank $r = 15$ are shown. The Strang splitting algorithm with a time step size $\tau = 0.025$ is employed. In both the $x$ and $v$-directions 128 grid points are used. As a comparison, a direct Eulerian simulation (based on a spectral method) is also shown.

not possible. We clearly see this in the numerical simulation. Nevertheless, the combined approach results in a significant reduction in the error in mass and momentum (by approximately two orders of magnitude).

Now, we increase the rank to $r = 15$ and consider a longer time interval (up to $t = 300$). The numerical results are shown in Figure 6.2. It can be observed very clearly that the uncorrected algorithm as well as the global correction result in qualitatively wrong results (the electric energy decreases by more than two orders of magnitude). On the other hand, the local correction and the combined approach keep the electric energy stable until the final time of the simulation. With respect to the conservation of the invariants the same conclusion as above can be drawn.

As has been mentioned in section 4 the combined approach can be adjusted to either be closer to the
Figure 6.3. Numerical simulations of the two-dimensional two-stream instability with rank $r = 15$ are shown. For the combined approach we also show numerical results for $w = 10^{-2}$ and $w = 10^{-4}$. A weight of $w = 0$ corresponds to the global correction and a weight of $w = 1$ to the combined correction described in section 4. The Strang splitting algorithm with a time step size $\tau = 0.025$ is employed. In both the $x$ and $v$-directions 128 grid points are used. As a comparison, a direct Eulerian simulation (based on a spectral method) is also shown.

local correction or the global correction. The results in Figure 6.3 show how we can trade-off the error in mass and momentum and the error in the local conservation laws. We clearly see that the solution deteriorates as the error in the conservation laws increases.

REFERENCES

[1] J. Bigot, V. Grandgirard, G. Latu, C. Passeron, F. Rozar, and O. Thomine. Scaling GYSELA code beyond 32K-cores on Blue Gene/Q. In ESAIM: Proceedings, volume 43, pages 117–135, 2013.
[38] J.M. Qiu and C.W. Shu. Positivity preserving semi-Lagrangian discontinuous Galerkin formulation: theoretical analysis and application to the Vlasov–Poisson system. *J. Comput. Phys.*, 230(23):8386–8409, 2011.

[39] J.A. Rossmanith and D.C. Seal. A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov–Poisson equations. *J. Comput. Phys.*, 230(16):6203–6232, 2011.

[40] F. Rozar, G. Latu, and J. Roman. Achieving memory scalability in the GYSELA code to fit exascale constraints. In *Parallel Processing and Applied Mathematics*, pages 185–195, 2013.

[41] N. J. Sircombe and T. D. Arber. VALIS: A split-conservative scheme for the relativistic 2D Vlasov–Maxwell system. *J. Comput. Phys.*, 228(13):4773–4788, 2009.

[42] 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(2):201–220, 1999.

[43] J. P. Verboncoeur. Particle simulation of plasmas: review and advances. *Plasma Physics and Controlled Fusion*, 47(5A):A231, 2005.