Research Article

Emil Kieri and Bart Vandereycken*

Projection methods for dynamical low-rank approximation of high-dimensional problems

DOI ..., Received ...; accepted ...

Abstract: We consider dynamical low-rank approximation on the manifold of fixed-rank matrices and tensor trains (also called matrix product states), and analyse projection methods for the time integration of such problems. First, under suitable approximability assumptions, we prove error estimates for the explicit Euler method equipped with quasi-optimal projections to the manifold. Then, we discuss the possibilities and difficulties with higher order explicit methods. In particular, we discuss ways for limiting rank growth in the increments, and robustness with respect to small singular values.

Keywords: tensor train, low rank approximation, tensor differential equations, projection methods

MSC 2010: 58J35, 65L05, 65L06, 65L70

1 Introduction

In this work we consider high-dimensional time-dependent problems. The problems could either be ordinary differential equations (ODEs), such as the chemical master equation [12], or partial differential equations (PDEs), such as the time-dependent Schrödinger equation [23] or parabolic problems. When PDEs are considered we apply a method of lines-approach, that is, we first discretise in space such that the problem is approximated by a system of ODEs. We assume, mostly for notational simplicity, that the problem considered is autonomous. Denoting $\dot{A}(t) = \mathrm{d}A/\mathrm{d}t$, the general form of our ODE is then

$$\dot{A}(t) = F(A(t)), \qquad A(0) = A_0, \tag{1}$$

Emil Kieri, Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, Bonn, Germany.

***Corresponding author: Bart Vandereycken,** Department of Mathematics, University of Geneva, Geneva, Switzerland. E-mail: bart.vandereycken@unige.ch.

on the Euclidean space $\mathcal{V} = \mathbb{R}^{N_1 \times \cdots \times N_d}$ of dth order $N_1 \times \cdots \times N_d$ tensors. We denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the standard Euclidean inner product and norm on $\mathcal{V}$. The generalisation of the methods and theory in this work to complex-valued tensors is straightforward. By adding t as an extra constant variable, any ODE can be brought to autonomous form. This can be done here as well although it requires extending each dimension N_i of A by one.

The characterising difficulty of high-dimensional problems is the exponential growth with the dimension of the amount of data and computational work; the space $\mathcal{V}$ has $\prod_{i=1}^{d} N_i$ degrees of freedom. This makes also seemingly simple problems computationally intractable when the dimension exceeds, say, three. Low rank approximation is one of the more promising approaches to tractable computation of high-dimensional problems. In computational chemistry, low rank methods such as Hartree–Fock and Multiconfigurational Time-Dependent Hartree (MCTDH) have become standard tools; see, e.g., [22].

1.1 Approximation by tensor trains

Tensors can be represented in different data-sparse formats, with different definitions of rank. In this work we consider the tensor train (TT) format [27, 26]. Before being rediscovered by the mathematical community it was known to physicists as matrix product states (MPS); see [29] for an overview. The rank of a TT $X \in \mathcal{V}$ is a $(d+1)$-tuple

$$\mathbf{r} = (r_0, \ldots, r_d) \quad \text{with} \quad r_j = \operatorname{rank} X^{\langle j \rangle}, \tag{2}$$

where $X^{\langle j \rangle} \in \mathbb{R}^{(N_1 \cdots N_j) \times (N_{j+1} \cdots N_d)}$ is the jth unfolding of X. The matrix $X^{\langle j \rangle}$ has the same elements as X, with the j first coordinate directions organised as rows and the others as columns. For $j = 0$ and $j = d$ we obtain a row or column vector, and consequently $r_0 = r_d = 1$. If $r_j \leq r$ and $N_j \leq N$ for all $j = 1, \ldots, d$, X can be represented by at most dr^2N numbers, breaking the exponential scaling with the dimension. More concretely, we can define order 3 tensors $C_j \in \mathbb{R}^{r_{j-1} \times N_j \times r_j}$ such that

$$X(i_1, \ldots, i_d) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_{d-1}=1}^{r_{d-1}} C_1(1, i_1, k_1) \cdots C_j(k_{j-1}, i_j, k_j) \cdots C_d(k_{d-1}, i_d, 1). \tag{3}$$

The modelling assumption in this work is that the solution $A(t)$ to (1) can be approximated by a TT of low rank. In other words, we expect the singular values of the unfoldings $A^{\langle j \rangle}$ to decay such that we can neglect most of them without losing too much information. We therefore look for an approximation

$Y(t)$ to the exact solution $A(t)$ which stays on the manifold

$$\mathcal{M}_{\mathbf{r}} = \{X \in \mathcal{V} \ : \ \text{TT-rank}(X) = \mathbf{r}\}$$

of tensors with fixed TT-rank $\mathbf{r}$. We do this using the Dirac–Frenkel time-dependent variational principle [4, 14, 17]:

$$\dot{Y}(t) = P(Y(t))F(Y(t)), \qquad Y(0) = Y_0 \in \mathcal{M}_{\mathbf{r}}, \tag{4}$$

where $P(Y)$ is the ℓ_2-orthogonal projection onto the tangent space $T_Y\mathcal{M}_{\mathbf{r}}$ of $\mathcal{M}_{\mathbf{r}}$ at $Y \in \mathcal{M}_{\mathbf{r}}$. This is the locally best approximation to the differential equation—we make the smallest possible perturbation of $\dot{Y}(t)$ such that $Y(t)$ stays on the manifold. If this perturbation is small we also get a bound on the ℓ_2 norm of the global error $\|Y(t) - A(t)\|$, see Lemma 1 below. $Y(t)$ will however in general not be the globally best approximation on $\mathcal{M}_{\mathbf{r}}$ to $A(t)$. The Dirac–Frenkel principle can equivalently be written using a Galerkin condition, as *find, for each* $t \in [0, T]$, $Y(t) \in \mathcal{M}_{\mathbf{r}}$ *such that*

$$\langle \dot{Y}(t), Z\rangle = \langle F(Y(t)), Z\rangle \quad \text{for all } Z \in T_{Y(t)}\mathcal{M}_{\mathbf{r}},$$

and $Y(0) = Y_0 \in \mathcal{M}_{\mathbf{r}}$.

1.2 Numerical challenges

Low rank approximation offers a reduction of the problem size which can enable the computational solution of problems which would otherwise be inaccessible. It does however not come without new challenges. Since the manifold $\mathcal{M}_{\mathbf{r}}$ is not linear, (4) is a non-linear problem even if F is linear. The projection may introduce additional difficulties also when the original problem already is non-linear. One main difficulty concerns the curvature of the manifold, which is unbounded. The local curvature grows without bound as we approach the boundary of the manifold, which consists of tensors of smaller rank, and the closure

$$\overline{\mathcal{M}_{\mathbf{r}}} = \{X \in \mathcal{V} \ : \ \text{TT-rank}(X) \leq \mathbf{r}\}$$

is no longer a smooth manifold. The local curvature can be quantified in terms of singular values of the matrix unfoldings of the tensor [20]. If the r_jth singular value $\sigma_{r_j}(X^{\langle j\rangle}) \geq \rho > 0$ for all j, and $Y \in \mathcal{M}_{\mathbf{r}}$ is close enough to X, then

$$\|(P(Y) - P(X))Z\| \leq \frac{c}{\rho}\|Y - X\|\|Z\| \tag{5}$$

for all $Z \in \mathcal{V}$ and some constant c which depends on the dimension, but not on Y or X. In Appendix A we show that this bound is essentially sharp in the sense that for any X, there exist a Y and a Z that attain the bound.

This strong local curvature in the presence of small singular values leads to a range of theoretical and practical difficulties. The difficulty which perhaps is the most relevant for this work is the stiffness induced to time-dependent problems on the manifold. The parametrisation (3) of the manifold is not unique, but if it is fixed at time $t = 0$ one can introduce gauge conditions such that the time-evolution of (4) has a unique parametrisation. One then gets ODEs which determine the evolution of the parametrisation [20, 31]. These ODEs are however stiff in the presence of small singular values since the Lipschitz constant of their right-hand side grows at least as $1/\rho$ due to (5). Therefore, a time step restriction $h \sim \rho$ will be necessary for explicit methods.

As an illustration we consider the two-dimensional case. Then, tensor trains coincide with matrices of the form $Y = USV^T \in \mathbb{R}^{N\times N}$, where $U, V \in \mathbb{R}^{N\times r}$ have orthonormal columns, and $S \in \mathbb{R}^{r\times r}$. The most common gauge conditions are $U^T\dot{U} = V^T\dot{V} = 0$, which lead to the system of ODEs

$$\begin{aligned} \dot{U} &= (I - UU^T)F(Y)VS^{-1}, \\ \dot{S} &= U^T F(Y)V, \\ \dot{V} &= (I - VV^T)F(Y)^T US^{-T}. \end{aligned} \tag{6}$$

Clearly, this system breaks down when S is singular. If S is nearly singular, the ODEs are very stiff leading to a severe step size restriction. This is, for example, illustrated numerically in [13, Fig. 1]. A popular way around this is regularisation. In the MCTDH method[1], S is commonly regularised as [2, 21]

$$S_{\text{reg}} = S + \epsilon_0 \exp(-S/\epsilon_0) \tag{7}$$

before it is inverted. The parameter ϵ_0 is small, often of the order 10^{-8}. This prevents the system from breaking down, but also modifies the problem and its solution. Furthermore, it still leads to systems with a large Lipschitz constant. In this paper, we will propose to change the integrator to address this problem more fundamentally.

A related difficulty with small singular values appears in low-rank optimisation. An optimisation problem $\min_{X\in\Omega} J(X)$, with a closed convex set $\Omega \subset \mathcal{V}$ and a strictly convex functional $J : \mathcal{V} \to \mathbb{R}$, has a unique global minimum. If we search for a low-rank approximation to the minimum by restricting the feasible set to $\Omega \cap \overline{\mathcal{M}_{\mathbf{r}}}$, the problem is no longer convex, and we may have introduced new local minima. Except for certain very simple or restricted cases, convergence

1 The MCTDH method is formulated for tensors in the Tucker format. As for tensor trains, in two dimensions this format reduces to bounded rank matrices.

theory for optimisation algorithms is therefore only local; see, e.g., the local convergence analysis for alternating optimisation in [28]. Most importantly, we can only guarantee convergence towards the global minimum if the initial guess is within an $\mathcal{O}(\rho)$ distance from it.

Another largely open question in low-rank approximation is the one about approximability: Given a problem, how can we know if a reasonably accurate low-rank approximation exists? This question is resolved only for a small number of problems. For the Poisson problem, for example, approximability can be confirmed using exponential sums [3, 5]. See also [6, 7] for an overview of other examples. In this work, we will just assume approximability. For applications and numerical validation of dynamical low-rank to PDEs with blow-up or stochastic PDEs, we refer to [25, 24].

1.3 Contributions

As mentioned above, instead of solving (4) by regularising the ODE (6), we change the integrator in a more fundamental way. This is related to the so-called splitting projector integrators in [18, 19] that were shown in [13] to have no time step restriction due to small singular values. In this work we obtain integrators with similar properties but that are based on projected integrators. The simplicity is the main advantage of such projection methods. The projected Euler method is the first method one would try when confronted with the problem (4). We prove that projected Euler, as well as some higher order projected Runge–Kutta methods, are accurate and robust. We also believe our proof techniques are simpler than for the splitting projector integrators in [13].

The splitting integrators of [18, 19] are also accurate and robust, and albeit conceptually a little more involved, they are still fairly easy to implement. However, while they allow for arbitrary order, the robustness is only proven up to first order. In our case, we will be able to show higher order for certain methods. An important advantage of the splitting methods is that they retain some geometric properties of the continuous problem. If the ℓ_2 norm of the solution is conserved in the continuous problem, and norm-conserving methods are used to solve the substeps of the splitting scheme, then the splitting integrator will conserve the norm. This is not the case for our projection methods but as we will see in the numerical experiments, they can be considerably cheaper when comparing the ℓ_2 norm of the error.

2 Assumptions and approximability

In this section we state the assumptions we make on the problem, and discuss their implications on the solvability of the problem and on its low-rank approximability.

2.1 Assumptions

Our assumptions are the same as in [13]. First, we assume that F is Lipschitz continuous,

$$\|F(X) - F(Y)\| \leq L\|X - Y\| \quad \text{for all } X, Y \in \mathcal{V}. \tag{8}$$

This gives via the Picard–Lindelöf theorem (see, e.g., [11]) existence and uniqueness of a solution to (1), at least on some finite time-interval. We also assume that F satisfies the one-sided Lipschitz bound

$$\langle X - Y, F(X) - F(Y)\rangle \leq \ell \, \|X - Y\|^2 \quad \text{for all } X, Y \in \mathcal{V}, \tag{9}$$

or equivalently if F is C^1, that $\partial F/\partial Y$ has logarithmic norm bounded by ℓ. This bound follows directly from (8) with $\ell = L$, but for many problems, in particular for spatial semi-discretisations of partial differential equations, ℓ is much smaller than L. Note that ℓ can be negative. When $\ell \ll L$, much sharper error estimates can be proven if (9) is taken into account. In the complex-valued case, (9) is modified by taking the real part of the left-hand side.

To make higher order methods sensible, we also assume that the solution is sufficiently smooth. More precisely, when considering method of order p, we assume that

$$\frac{\mathrm{d}^{p+1}}{\mathrm{d}t^{p+1}} \Phi_F^t(Y)$$

can be uniformly bounded by a constant for all Y in a neighbourhood of the exact solution. Here, Φ_F^t denotes the flow of F, that is, the mapping $A(t) = \Phi_F^t(A_0)$ where $A(t)$ is the solution of (1) with initial value $A(0) = A_0$.

To assure low-rank approximability, we assume that F almost maps onto the tangent bundle of $\mathcal{M}_{\mathbf{r}}$:

$$\begin{gathered} \|F(Y) - P(Y)F(Y)\| \leq \varepsilon \quad \text{for all} \\ Y \in \mathcal{M}_{\mathbf{r}} \cap \{\text{suitable neighbourhood of the exact solution}\}. \end{gathered} \tag{10}$$

This assumption implies that the solution $Y(t)$ of (4) is an $\mathcal{O}(\varepsilon)$ perturbation of $A(t)$, the exact solution of (1); see Lemma 1 below. We call this difference the *modelling error*. Except for in pathological special cases, there is no way to avoid

it. The only way to make the modelling error smaller is to improve the model, which in our case means increasing the approximation rank.

We also assume that the solution $Y(t)$ of (4) stays on the manifold $\mathcal{M}_{\mathbf{r}}$ on the considered time interval $t \in [0, T]$. That is, we assume that the rank of $Y(t)$ does not drop, but is always $\mathbf{r}$. This is necessary for (4), and the tangent space projection $P(Y)$ in particular, to be well-defined. We still allow the smallest non-zero singular values to be arbitrarily small. We need a similar full approximation rank condition for the numerical approximation Y_i at all time steps. This will however be satisfied in practice due to numerical round-off.

2.2 Discussion on the approximability

In the following lemma, we use (10) to bound this modelling error. The result and its proof are standard, (see, e.g., [11, Thm. I.10.6]) but we include it here since we will encounter this kind of bounds a few times more throughout the paper.

Lemma 1. *Given the assumptions in Section 2.1, and with the error in the initial value bounded by* $\|A_0 - Y_0\| \leq \delta$*, the dynamical low-rank approximation* (4) *yields an error bounded by*

$$\|Y(t) - A(t)\| \leq \mathrm{e}^{\ell t}\delta + \varepsilon \int_0^t \mathrm{e}^{\ell s}\,\mathrm{d}s.$$

Proof. Denote $P^{\perp}(Y)Z = Z - P(Y)Z$. From the bound

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Y - A\|^2 &= \langle Y - A, P(Y)F(Y) - F(A)\rangle \\ &= \langle Y - A, F(Y) - F(A)\rangle - \langle Y - A, P^{\perp}(Y)F(Y)\rangle \\ &\leq \ell\|Y - A\|^2 + \|Y - A\|\|P^{\perp}(Y)F(Y)\|,\end{aligned}$$

we obtain the differential inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Y - A\| \leq \ell\|Y - A\| + \varepsilon.$$

Its solution satisfies (see, e.g., [11, Ch. I.10])

$$\|Y(t) - A(t)\| \leq \mathrm{e}^{\ell t}\|Y(0) - A(0)\| + \int_0^t \mathrm{e}^{\ell(t-s)}\varepsilon\,\mathrm{d}s. \qquad \square$$

The most direct approximability assumption would be to directly demand that the modelling error $\|Y(t) - A(t)\|$ is small. We use the condition (10) instead, mainly because it is easier to work with this assumption—it is a local assumption which matches well the local nature of the Dirac–Frenkel principle. Both conditions are difficult to verify a priori. If one can decompose F as

$$F = F_T + F_\varepsilon,$$

where $F_T : \mathcal{M}_{\mathbf{r}} \to T\mathcal{M}_{\mathbf{r}}$ maps onto the tangent bundle and $\|F_\varepsilon\| \leq \varepsilon$, then (10) obviously holds. A common example of terms mapping onto the tangent bundle are operators acting in a single coordinate direction. Since the tangent space is a linear space, linear combinations of such terms also map onto the tangent bundle. As an example, in the matrix case $A(t) \in \mathbb{R}^{N_1 \times N_2}$,

$$F_T(A) = M_1 A + A M_2 \in T_A \mathcal{M}_r \text{ for all } A \in \mathcal{M}_r,$$

and for any choice of $M_1 \in \mathbb{R}^{N_1 \times N_1}$ and $M_2 \in \mathbb{R}^{N_2 \times N_2}$. Most discretisations of the Laplace operator, in any dimension, are of this form. However, if the diffusion is anisotropic, or if a curvilinear coordinate transformation is employed, this structure will be lost and the diffusion operator no longer map onto the tangent bundle.

Demanding that the remainder F_ε is small is arguably a strong assumption. If the manifold has tiny, high-frequency wiggles, the best approximation on it could stay close to $A(t)$ but at the same time (10) is violated. On the other hand, (10) does appear to hold in neighbourhoods of the exact solutions of many interesting problems. In Figure 1 we plot $\|P^\perp(Y(t))F(Y(t))\|$ against t for the numerical solution of the hyperbolic problem

$$\begin{aligned} u_t &= -\mathbf{b} \cdot \nabla u + 0.8u^2, \qquad \mathbf{x} \in (-\pi, \pi)^d, \quad t > 0, \\ u(0, \mathbf{x}) &= \prod_{j=1}^{d} \mathrm{e}^{-x_j^2}, \end{aligned} \tag{11}$$

with periodic boundary conditions and $\mathbf{b} = (1, 1, \ldots, 1)$. We consider the problem in four dimensions, and discretise the gradient with second order upwind finite differences. We use 64 spatial grid points per dimension and the approximation rank $\mathbf{r} = (1, 5, 5, 5, 1)$, and time-step on the low-rank manifold using projected Euler (which will be introduced in Section 4) with time step $h = 1/200$. It is not obvious a priori that the right-hand side almost maps onto the tangent bundle, as the non-linear term in the right-hand side, which in the spatially discrete case is a Hadamard product, squares the rank. Still, we see in Figure 1 that the normal component stays moderate. The norm is scaled by the spatial step size

to mimic the continuous L^2 norm:

$$\|\mathbf{u}\|^2 = \Delta x^d \sum_{k_1=1}^{N_1} \cdots \sum_{k_d=1}^{N_d} u_{k_1,\ldots,k_d}^2. \tag{12}$$

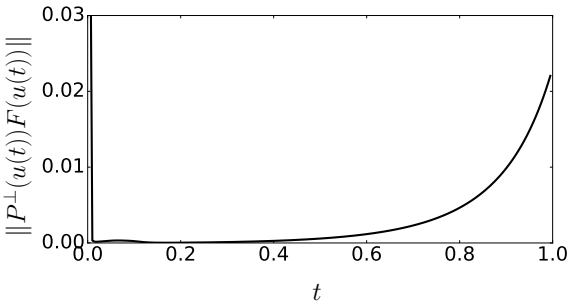

Fig. 1: Component in the normal space of the right-hand side of (11), plotted against time. $\|F(u(t))\|$ ranges between 3.2 and 9.0 over the same time interval, and the norm of the non-linear term ranges between 0.62 and 4.2. The spike near $t = 0$ is due to the initial data being rank-deficient.

3 An idealised projection method

The central idea in this work is to solve problem (4) using projection methods [10, Ch. IV.4]. A standard projection method typically behaves as follows:

1. Take one time step with a one-step method, most likely (unless in very special cases) leaving the manifold.
2. Retract or project the new solution back to the manifold.

The discussion on projection methods in [10] concerns problems which are quite different from ours. They are of low or moderate dimension, and the curvature of the manifold stays nicely bounded. Furthermore, the solution to the original problem stays on the manifold, that is, there is no modelling error. Evolution on the manifold, which, for example, might be the set of constant energy or angular momentum, is a property of the exact solution. These differences make our problems more difficult. As our problems are of very high dimension, we

must take care so that the amount of data does not grow unreasonably in any intermediate step. We must also treat tangent space projections with care, so that the strong curvature of the manifold does not give rise to time step restrictions.

To implement a projection method we need a mapping $\mathcal{R} : \mathcal{V} \to \mathcal{M}_{\mathbf{r}}$ that satisfies for all $A \in \mathcal{V}$:

$$\|\mathcal{R}(A) - A\| \leq C_{\mathcal{R}} \|\mathcal{P}_{\mathcal{M}_{\mathbf{r}}}(A) - A\|, \tag{13}$$

where $C_{\mathcal{R}}$ does not depend on A and such that

$$\mathcal{P}_{\mathcal{M}_{\mathbf{r}}}(A) \ni \operatorname*{arg\,min}_{X \in \overline{\mathcal{M}_{\mathbf{r}}}} \|X - A\|$$

is the best approximation of A on $\overline{\mathcal{M}_{\mathbf{r}}}$. Such mappings are called quasi-optimal projections. Since they also extend the domain of definition of retractions in manifold algorithms [1] form the tangent space $T_A\mathcal{M}_{\mathbf{r}}$ to the full space $\mathcal{V}$, we call $\mathcal{R}$ an extended retraction.

The best approximation for TT is known to exist [7, Thm. 11.56], which means it is available in our theoretical argumentation, but in general it is highly impractical to compute it. The best approximation need not be unique since $\overline{\mathcal{M}_{\mathbf{r}}}$ is not a convex set. In that case, $\mathcal{P}_{\mathcal{M}_{\mathbf{r}}}(A)$ means any best approximation. However, $\mathcal{P}_{\mathcal{M}_{\mathbf{r}}}(A)$ is unique if A is sufficiently close to $\mathcal{M}_{\mathbf{r}}$. A practical, extended retraction is given by successively truncating the singular value decompositions (SVDs) of the unfoldings $A^{\langle j \rangle}$, $j = 1, \dots, d$. This method is known as TT-SVD [26]. It has quasi-optimality constant $C_{\mathcal{R}} = \sqrt{d-1}$, and can be computed efficiently if A has larger but still moderate rank, and is represented in the TT format. See also [7].

Using the exact flow of F, we shall now construct the following *idealised projection method*: for time step h, compute

$$Y_{i+1} = \mathcal{R}(\Phi_F^h(Y_i)). \tag{14}$$

This is not a practical method, first, because we usually have no means of computing the exact flow, and second, because the exact solution will in general have full rank before retracting, making the method prohibitively expensive—recall that only working with low-rank tensors is our way of making computations tractable. The method, however, brings the idea about. It satisfies the following error estimate without a restriction due to small singular values.

Theorem 2. *Under the assumptions of Section 2.1 and assuming* (13)*, and with the bound* $\|Y_0 - A_0\| \leq \delta$ *of the error in the initial data, the idealised projection method* (14) *satisfies the error estimate*

$$\|Y_n - A(nh)\| \leq C(\delta + \varepsilon)$$

on the finite time-interval $0 \le nh \le T$*, for all* $0 < h \le h_0$*. The constant* C *depends only on* ℓ*,* T*,* h_0 *and* $C_{\mathcal{R}}$*, but not on* h*.*

Proof. Let Φ^h_{PF} be the flow of (4). From Lemma 1 with $\delta = 0$, we get

$$\|\Phi^h_{PF}(Y_i) - \Phi^h_F(Y_i)\| \le \varepsilon h \max\{1, \mathrm{e}^{\ell h}\}.$$

Since $\Phi^h_{PF}(Y_i) \in \mathcal{M}_{\mathbf{r}}$, the best approximation satisfies the same bound:

$$\|\mathcal{P}_{\mathcal{M}_{\mathbf{r}}}(\Phi^h_F(Y_i)) - \Phi^h_F(Y_i)\| \le \varepsilon h \max\{1, \mathrm{e}^{\ell h}\}. \tag{15}$$

By the quasi-optimality (13) of $\mathcal{R}$, we therefore obtain the bound

$$\|e_{i+1}\| = \|\mathcal{R}(\Phi^h_F(Y_i)) - \Phi^h_F(Y_i)\| \le Ch\varepsilon, \quad C = C_{\mathcal{R}} \max\{1, \mathrm{e}^{\ell h}\} \tag{16}$$

for the local error $e_{i+1} = Y_{i+1} - \Phi^h_F(Y_i)$.

To bound the global error $E = Y_n - \Phi^{nh}_F(Y_0)$, we use a standard Lady Windermere's fan argument with error transport along the exact solution curves, as described in [11, Ch. II.3]. First, we expand into the telescoping sum

$$\begin{aligned} E &= Y_n - \Phi^{nh}_F(A_0) \\ &= \sum_{i=1}^{n} \left(\Phi^{(n-i)h}_F(Y_i) - \Phi^{(n-i+1)h}_F(Y_{i-1}) \right) + \Phi^{nh}_F(Y_0) - \Phi^{nh}_F(A_0). \end{aligned}$$

Each of the terms

$$\begin{aligned} E_i &= \Phi^{(n-i)h}_F(Y_i) - \Phi^{(n-i)h}_F(\Phi^h_F(Y_{i-1})), \ i = 1, \dots, n, \\ E_0 &= \Phi^{nh}_F(Y_0) - \Phi^{nh}_F(A_0), \end{aligned}$$

can now be bounded as in the proof of Lemma 1 using the logarithmic norm. In particular,

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi^t_F(X) - \Phi^t_F(Y)\|^2 &= \langle \Phi^t_F(X) - \Phi^t_F(Y), F(\Phi^t_F(X)) - F(\Phi^t_F(Y)) \rangle \\ &\le \ell \|\Phi^t_F(X) - \Phi^t_F(Y)\|^2, \end{aligned}$$

which leads to

$$\|\Phi^t_F(X) - \Phi^t_F(Y)\| \le \mathrm{e}^{\ell t}\|X - Y\|.$$

Together with (16), this results in the bounds

$$\begin{aligned} &\|E_i\| \le \mathrm{e}^{\ell(n-i)h}\|e_i\| \le h\varepsilon C \mathrm{e}^{\ell(n-i)h}, \ i = 1, \dots, n, \\ &\|E_0\| \le \mathrm{e}^{\ell nh}\delta. \end{aligned} \tag{17}$$

The global error can be then bounded as

$$\|Y_n - \Phi_F^{nh}(Y_0)\| \leq \sum_{i=0}^{n} \|E_i\| \leq \mathrm{e}^{\ell nh}\delta + C\varepsilon \sum_{i=1}^{n} h\mathrm{e}^{\ell(n-i)h}.$$

Bounding the Riemann sum as (see also [11, Ch. II.3])

$$\sum_{i=1}^{n} h\mathrm{e}^{\ell(n-i)h} \leq \begin{cases} \int_0^{nh} \mathrm{e}^{\ell(nh-t)}\,\mathrm{d}t = (\mathrm{e}^{\ell nh} - 1)/\ell & \text{if } \ell > 0, \\ nh & \text{if } \ell = 0, \\ \int_0^{nh} \mathrm{e}^{\ell(nh-n-t)}\,\mathrm{d}t = \mathrm{e}^{-\ell h}(\mathrm{e}^{\ell nh} - 1)/\ell & \text{if } \ell < 0, \end{cases}$$

we end up with a bound independent of h for all $h \leq h_0$. This gives the desired result (but with a different constant C). □

As mentioned above, this method is mostly of theoretical value since the exact flow is not at our disposal. To make it practical, one can use a one-step method to approximate it. In this work, we shall consider explicit Runge–Kutta methods with s stages that applied to (1) can be written as

$$\begin{aligned} k_j &= F(A_i + h\sum_{l=1}^{j-1} a_{jl}k_l), \quad j = 1, \dots, s, \\ A_{i+1} &= A_i + h\sum_{j=1}^{s} b_j k_j, \end{aligned} \tag{18}$$

and that are of order p, that is, for all $0 < h \leq h_0$:

$$\|A_{i+1} - \Phi_F^h(A_i)\| \leq C_L h^{p+1}, \tag{19}$$

with C_L independent of h.

Probably the most straight-forward way to obtain a projection method is

$$\begin{aligned} k_j &= F(Y_i + h\sum_{l=1}^{j-1} a_{jl}k_l), \quad j = 1, \dots, s, \\ Y_{i+1} &= \mathcal{R}\big(Y_i + h\sum_{j=1}^{s} b_j k_j\big). \end{aligned} \tag{20}$$

Using similar techniques as in the projected Euler case (see Section 4 below), this method can be shown to have a global error of $\mathcal{O}(h^p + \varepsilon)$. Except for the unavoidable modelling error ε, the bound reflects the right order. Unfortunately, the computational feasibility of this approach depends very much on the function F. For the vast majority of problems, however, F will be an operator that

increases the rank of its argument. For example, for a simple Laplacian, the ranks are doubled, whereas Hadamard products multiply the ranks. Representations of F using TT-matrices, or matrix product operators (see, e.g., [33]) also increase the rank of F linearly. This approach therefore soon becomes impractical for higher-order methods since the ranks of the internal stages k_j grow exponentially with the number of stages s.

In the next two sections, we will show how one can use tangent space projections and retractions to limit the rank growth of projected Runge–Kutta methods, and analyse the accuracy of such methods. This enables methods of higher order.

4 A practical projected Euler method

The most elementary projection method is projected Euler. Applied to (4), it reads

$$Y_{i+1} = \mathcal{R}(Y_i + hP(Y_i)F(Y_i)). \tag{21}$$

As long as F can be evaluated in an efficient way for Y_i of low rank, time stepping can be done efficiently. In addition, if $F(Y_i)$ is the sum of several terms $F_j(Y_i)$, we can compute $P(Y_i)$ by sequentially summing each projected term $P(Y_i)F_j(Y_i)$. This is usually cheaper and since the tangent space $T_Y\mathcal{M}_{\mathbf{r}}$ is a vector space, there is less risk of numerical cancellation this way, compared to retracting $F(Y_i)$ (see also [16, Sect. 3.2]). Finally, computing the retraction is also efficient since the elements of $T_Y\mathcal{M}_{\mathbf{r}}$ are tensors of at rank at most $2\mathbf{r}$, which also applies to the argument of $\mathcal{R}$ in (21) since

$$Y_i + hP(Y_i)F(Y_i) = P(Y_i)(Y_i + hF(Y_i)).$$

The analysis of (21) is fairly straightforward. It satisfies the following local and global error estimates.

Lemma 3. *Under the assumptions of Theorem 2, the local error of the projected Euler method* (21) *is bounded by*

$$\|Y_{i+1} - \Phi_F^h(Y_i)\| \leq Ch(\varepsilon + h),$$

where the constant C is independent of h for all $h \leq h_0$.

We postpone the proof of this lemma, and immediately state the global error estimate as Theorem 4 below. Compared to the idealised method in Theorem 2, we see that the price to pay to obtain a practical method is an additional term

$\mathcal{O}(h)$ in the global error. This is to be expected when using explicit Euler. In addition, there is again no step size restriction due to small singular values.

Theorem 4. *Under the assumptions of Theorem 2, the projected Euler method* (21) *satisfies the error estimate*

$$\|Y_n - A(nh)\| \leq C(\delta + \varepsilon + h)$$

on the finite time-interval $0 \leq nh \leq T$*, for all* $0 < h \leq h_0$*. The constant* C *depends only on* L*,* T*,* h_0*, and* $C_{\mathcal{R}}$*, but not on* h*.*

Proof. The proof is similar as for Theorem 2, but with the different local error from Lemma 3:

$$e_i = Y_i - \Phi_F^h(Y_{i-1}), \qquad \|e_i\| \leq Ch(\varepsilon + h).$$

Using the same notation as in the proof of Theorem 2, the global error is then estimated as

$$\|Y_n - \Phi_F^{nh}(Y_0)\| \leq \sum_{i=0}^{n} \|E_i\| \leq \mathrm{e}^{\ell nh}\delta + C(\varepsilon + h)\sum_{i=1}^{n} h\mathrm{e}^{\ell(n-i)h},$$

and the result follows again by bounding the Riemann sum. □

We now prove the local error.

Proof of Lemma 3. Let us write

$$Y_i + hP(Y_i)F(Y_i) = Y_i + hF(Y_i) - hP^{\perp}(Y_i)F(Y_i),$$

and note that $\|hP^{\perp}(Y_i)F(Y_i)\| \leq h\varepsilon$ by assumption (10). Since Euler's method has order one, we have for all $h \leq h_0$ that

$$\|\Phi_F^h(Y_i) - (Y_i + hF(Y_i))\| \leq C_L h^2,$$

with C_L independent of h (but it might depend on L). This means that (21) can be rewritten as

$$Y_{i+1} = \mathcal{R}(\Phi_F^h(Y_i) + h\Delta), \qquad \text{with } \|\Delta\| = \mathcal{O}(\varepsilon + h). \tag{22}$$

Hence, one step of the projected Euler is an idealised projection method (14) but for a perturbation of the flow.

To bound the local error

$$e_{i+1} = Y_{i+1} - \Phi_F^h(Y_i) = \mathcal{R}(\Phi_F^h(Y_i) + h\Delta) - \Phi_F^h(Y_i),$$

let us introduce the notation

$$Z = \Phi_F^h(Y_i) + h\Delta \qquad \text{and} \qquad \widetilde{Z} = \Phi_F^h(Y_i).$$

Using (13) and by definition of $\mathcal{P}_{\mathcal{M}_\mathbf{r}}$,

$$\begin{aligned}\|\mathcal{R}(Z) - Z\| &\le C_\mathcal{R} \min_{X \in \overline{\mathcal{M}_\mathbf{r}}} \|X - \widetilde{Z} + (\widetilde{Z} - Z)\| \\ &\le C_\mathcal{R}(\|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(\widetilde{Z}) - \widetilde{Z}\| + \|\widetilde{Z} - Z\|).\end{aligned}$$

From this we obtain the useful result

$$\|\mathcal{R}(Z) - \widetilde{Z}\| \le C_\mathcal{R}\|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(\widetilde{Z}) - \widetilde{Z}\| + (1 + C_\mathcal{R})\|Z - \widetilde{Z}\|. \tag{23}$$

Since $\widetilde{Z} = \Phi_F^h(Y_i)$, the first term can be bounded as (15) in the proof of Theorem 2. We therefore obtain as bound for the local error

$$\|e_{i+1}\| \le C_\mathcal{R} h\varepsilon \max\{1, \mathrm{e}^{\ell h}\} + (1 + C_\mathcal{R})h\|\Delta\| \le Ch(\varepsilon + \|\Delta\|), \tag{24}$$

with C independent of h if $h \le h_0$. Since $\|\Delta\| = \mathcal{O}(\varepsilon + h)$, this gives the desired result. □

The local error in Lemma 3 does not depend on the curvature of $\mathcal{M}_\mathbf{r}$, that is, on the singular values of Y_i. This may seem surprising given the essentially tight bound (5). However, the recurring assumption in this paper is that we assume that the solution can locally be well approximated by low rank. A similar robustness property also holds for the best rank r approximation of a matrix A (see, e.g., [8, Lemma 4.1]):

$$\|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(A + E) - \mathcal{P}_{\mathcal{M}_\mathbf{r}}(A)\| \le 2(\|A - \mathcal{P}_{\mathcal{M}_\mathbf{r}}(A)\| + \|E\|).$$

From (23), one can generalise this result to quasi-optimal retractions of tensors as

$$\|\mathcal{R}(A + E) - \mathcal{R}(A)\| \le (1 + C_\mathcal{R})(\|A - \mathcal{R}(A)\| + \|E\|).$$

However, while it clearly shows that $\mathcal{R}$ is well behaved for good approximations, we did not see directly a way to exploit it.

Remark 5. *Consider the equation $\mathcal{A}(X) - B$ with $\mathcal{A}$ a linear and symmetric positive definite operator on $\mathcal{V}$. Given a preconditioner $\mathcal{M}$ of the same type, one can try to find low-rank solutions to $\mathcal{A}(X) - B$ by integrating the gradient flow*

$$\dot{X} = -P(X)[\mathcal{M}^{-1}(\mathcal{A}(X) - B)].$$

When the projected Euler method (21) *is used, we get the "geometric" version of the preconditioned Richardson iteration, as introduced in [16, (3.6)]. It was shown in [16] that this version is typically much more efficient than without tangent space projection.*

5 Projected Runge–Kutta methods of higher order

Recall that the scheme (20) is formally a high-order projected Runge–Kutta scheme but due to the rank growth of the intermediate stages, it becomes quickly computationally prohibitive. In this section, we will construct more efficient methods that have more stages but with a limited rank growth. We do this by projecting onto the tangent space and retracting back to the manifold.

To this end, we first write the standard Runge–Kutta scheme (18) applied to F in its equivalent form

$$\begin{aligned} \widetilde{Z}_j &= A_i + h\sum_{l=1}^{j-1} a_{jl} F(\widetilde{Z}_l), \quad j = 1, \dots, s, \\ A_{i+1} &= A_i + h\sum_{j=1}^{s} b_j F(\widetilde{Z}_j). \end{aligned} \tag{25}$$

The equivalence follows by identifying $k_j = F(\widetilde{Z}_j)$. To obtain a fully projected integrator, we apply (25) to the vector field $X \mapsto P(\mathcal{R}(X))F(\mathcal{R}(X))$. With Y_i as initial value, we obtain our *projected Runge–Kutta* method

$$\begin{aligned} Z_j &= Y_i + h\sum_{l=1}^{j-1} a_{jl} P(\mathcal{R}(Z_l))F(\mathcal{R}(Z_l)), \quad j = 1, \dots, s, \\ Y_{i+1} &= \mathcal{R}(Y_i + h\sum_{j=1}^{s} b_j P(\mathcal{R}(Z_j))F(\mathcal{R}(Z_j)). \end{aligned} \tag{26}$$

Such schemes are sometimes called internal projection methods [9] since they extend the domain of the vector field $Y \mapsto P(Y)F(Y)$ from the manifold $\mathcal{M}_{\mathbf{r}}$ to the whole space $\mathcal{V}$ by projecting all the intermediate stages. This is needed to have a well-defined tangent space projection throughout the scheme.

For efficient implementation, (26) can also be written in the more usual notation with stages:

$$\begin{aligned} \eta_1 &= Y_i, \\ \kappa_j &= P(\eta_j)F(\eta_j), \quad j = 1, \dots, s, \\ \eta_j &= \mathcal{R}(Y_i + h\sum_{l=1}^{j-1} a_{jl}\kappa_l), \quad j = 2, \dots, s, \\ Y_{i+1} &= \mathcal{R}(Y_i + h\sum_{j=1}^{s} b_j\kappa_j). \end{aligned} \tag{27}$$

Since $\eta_j \in \mathcal{M}_{\mathbf{r}}$ and $\kappa_j \in T_{\eta_j}\mathcal{M}_{\mathbf{r}}$, the rank of κ_j is at most $2\mathbf{r}$. This way, the retraction is applied to tensors of rank at most $2s\mathbf{r}$, which is considerably less than in (20).

5.1 Non-linear Schrödinger equation

Below, we illustrate the performance of projected Runge–Kutta methods with a numerical example. We consider a non-linear Schrödinger equation on a two-dimensional lattice [30], where $A\colon [0,T] \to \mathbb{C}^{n\times n}$ evolves according to

$$\mathrm{i}\dot{A} = -\frac{1}{2}(BA + AB) - \alpha|A|^2 A. \tag{28}$$

The cubic nonlinearity is taken element-wise. The matrix $B = \operatorname{tridiag}(1,0,1)$ models the coupling of the lattice sites. We use a lattice of size $n = 100$ and the initial data

$$A_{j,k}(0) = \exp\Big(-\frac{(j-\mu_1)^2}{\sigma^2} - \frac{(k-\nu_1)^2}{\sigma^2}\Big) + \exp\Big(-\frac{(j-\mu_2)^2}{\sigma^2} - \frac{(k-\nu_2)^2}{\sigma^2}\Big),$$

where $\sigma = 10$, $\mu_1 = 60$, $\mu_2 = 50$, $\nu_1 = 50$, and $\nu_2 = 40$.

We solve (28) for approximation rank $r = 8$ with a range of time steps h using projected Runge–Kutta methods of first, second, and third order (PRKp). Apart from the explicit Euler method, we build projected versions of the following Runge–Kutta methods:

$$\begin{aligned}
&\text{PRK1:} && b_1 = 1 \\
&\text{PRK2:} && a_{21} = 1,\ b_1 = b_2 = \tfrac{1}{2}, \\
&\text{PRK3:} && a_{21} = \tfrac{1}{3},\ a_{31} = 0,\ a_{32} = \tfrac{2}{3},\ b_1 = \tfrac{1}{4},\ b_2 = 0,\ b_3 = \tfrac{3}{4}.
\end{aligned} \tag{29}$$

Recall that PRK1 is our projected Euler method from Section 4 and PRK2 is constructed from Heun's method. Since the rank of $A(0)$ is exactly two, the resulting initial value problem does not satisfy the assumptions from Section 2.1 that $A(t)$ has to be of rank at least r. To obtain a full-rank initial value (with very small singular values), we propagate (28) from $t = 0$ to $t = 0.01$ and start our experiment from there. The relative error is computed at time $T = 5$ in Frobenius norm, compared to a full-rank reference solution computed with `ode45` in MATLAB with tolerance 10^{-12}.

The results are shown in Figure 2, in two separate setups with $\alpha = 0.1$ and $\alpha = 0.3$. When h is reduced, the error initially decays for all methods, until a point where the modelling error by low rank becomes dominant and the error stagnates. Before stagnation, we clearly see that for $\alpha = 0.1$, the error converges

with order p as hoped for when using PRKp. For $\alpha = 0.3$, the modelling error is larger and stagnation happens already at a larger time step. This is likely the reason why PRK3 does not show third order convergence. In addition, we see that for h sufficiently small, the projected methods compute a rank r approximation of $A(T)$ that is nearly optimal and that higher order methods are more accurate for fixed h.

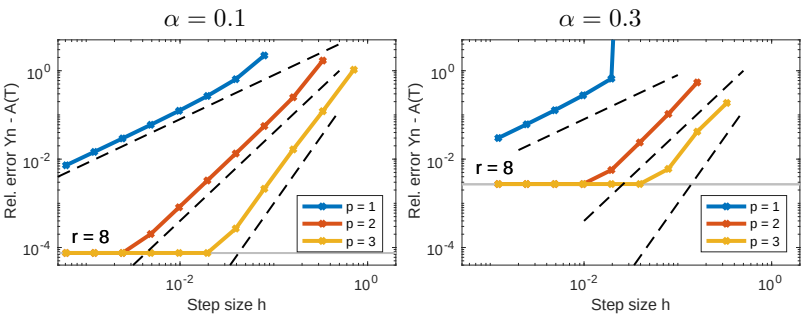

Fig. 2: Time-stepping for (28) with the projected Runge–Kutta methods (29) for rank $r = 8$. On each figure, the error is visible in full line from top to bottom for PRK1-2-3. The dashed lines indicate h^p for $p = \{1, 2, 3\}$. In grey line is the error of the best rank r approximation of $A(T)$.

5.2 Error analysis

We will now prove that the projection methods in the experiment above (see also Figure 2) retain their classical order of accuracy up to the modelling error $\mathcal{O}(\varepsilon)$. In addition to the usual order p of the Runge–Kutta scheme (25), our Theorem 6 below also uses the *stage orders* $q_1, \dots q_s$. They are defined from the local errors of the $\widetilde{Z}_j$, that is, for all $h \leq h_0$,

$$\|\widetilde{Z}_j - \Phi_F^{c_j h}(A_i)\| \leq C_L h^{q_j+1}, \qquad j = 1, \dots, s, \tag{30}$$

where $c_j = \sum_{l=1}^{j-1} a_{jl}$. The constant C_L is independent of h.

Given the coefficients of the Runge–Kutta method, q_j can be found by verifying the following quadrature relations (see, e.g., [11, Lemma II.7.5]):

$$\sum_{l=1}^{j-1} a_{jl} c_l^{\tau-1} = \frac{c_j^\tau}{\tau}, \qquad \tau = 1, \dots, q_j. \tag{31}$$

For the schemes from above, we have the following orders.

$$\begin{aligned} &\text{PRK1:} && q_1 = 0 \\ &\text{PRK2:} && q_1 = 0,\ q_2 = 1, \\ &\text{PRK3:} && q_1 = 0,\ q_2 = 1,\ q_3 = 2. \end{aligned} \tag{32}$$

For explicit Runge–Kutta methods of higher order p, these stage orders q_i are significantly smaller than p. Usually, one has $q_1 \leq q_2 \leq \cdots \leq q_s$ but this is not always true for methods with many stages.

Theorem 6. *Let* (27) *be a projected Runge–Kutta method with s stages, based on an explicit Runge–Kutta method of order p and stage orders $q_1 \leq q_2 \leq \cdots \leq q_s$. Denote*

$$q = \begin{cases} \min(p, q_2 + 1), & \text{if } b_2 \neq 0, \\ \min(p, q_3 + 1, q_2 + 2), & \text{if } b_2 = 0. \end{cases}$$

Then, under the assumptions of Theorem 2, the global error is bounded by

$$\|Y_n - A(nh)\| \leq C(\delta + \varepsilon + h^q),$$

on the finite time-interval $0 \leq nh \leq T$, for all $0 < h \leq h_0$. The constant C depends only on C_L, h_0, L, $C_{\mathcal{R}}$, s, $C_A = \max_{ij} |a_{ij}|$, and $C_B = \max_i |b_i|$, but not on h.

From (29) and (32), we see that PRK1, 2, and 3 indeed retain their classical order, that is, $q = p$. Unfortunately, we cannot generalise to higher order due to the $q_2 + 2$ term in Theorem 6. This can be understood by studying the quadrature rule defined by the nodes c_j and weights b_j. Due to the limited number of degrees of freedom in explicit methods, it is impossible to raise q_2 much higher. See also Remark 8 on how to obtain $q = 4$.

Like for the other projection methods, the proof for the global error in Theorem 6 follows immediately from the local error. To this end, we first need the following technical lemma regarding the perturbation of (26) compared to (25).

Lemma 7. *Under the assumptions of Theorem 2 and assuming $q_1 \leq q_2 \leq \cdots \leq q_s$ are the stage orders of the order p Runge–Kutta scheme* (25)*, let $A_i = Y_i$ and denote $\widetilde{q}_j = \min(q_j, q_2 + 1)$. Then* (26) *satisfies for $0 < h \leq h_0$ and $j = 1, \ldots, s$*

the following bounds:

$$\|Z_j - \widetilde{Z}_j\| \leq \begin{cases} 0 & \text{if } j = 1, \\ Ch(\varepsilon + h^{q_2+1}) & \text{if } j \neq 1, \end{cases} \tag{33}$$

$$\|P(\mathcal{R}(Z_j))F(\mathcal{R}(Z_j)) - F(\widetilde{Z}_j)\| \leq \begin{cases} C\varepsilon & \text{if } j = 1, \\ C(\varepsilon + h^{\widetilde{q}_j+1}) & \text{if } j \neq 1, \end{cases} \tag{34}$$

where C depends only on C_L, L, $C_{\mathcal{R}}$, h_0, s, and $C_A = \max_{ij} |a_{ij}|$.

Proof. By definition of an explicit Runge–Kutta method, $Z_1 = Y_i = \widetilde{Z}_1$. Observing that $\mathcal{R}(Y_i) = Y_i$ due to quasi-optimality, we have $F(\mathcal{R}(Z_1)) = F(Y_i) = F(\widetilde{Z}_1)$. Hence, the approximability assumption (10) gives

$$\|P(\mathcal{R}(Z_1))F(\mathcal{R}(Z_1)) - F(\widetilde{Z}_1)\| \leq \varepsilon.$$

Hence, the statements of the lemma are true for $j = 1$.

We show the rest by induction using the short-hand notation

$$P_i = P(\mathcal{R}(Z_i)),\ F_i = F(\mathcal{R}(Z_i)),\ \widetilde{F}_i = F(\widetilde{Z}_i). \tag{35}$$

Let $j \geq 2$ and assume (33)–(34) are true up to and including $j - 1$. By definitions (25)–(26) of the schemes, it holds that

$$\|Z_j - \widetilde{Z}_j\| \leq h \sum_{l=1}^{j-1} |a_{jl}| \|P_l F_l - \widetilde{F}_l\|. \tag{36}$$

The induction hypothesis on (34) for $l = 1, \dots, j - 1$ then gives

$$\|Z_j - \widetilde{Z}_j\| \leq C_A C h(s\varepsilon + h^{\widetilde{q}_2+1} + \cdots + h^{\widetilde{q}_{j-1}+1}).$$

Since $\widetilde{q}_2 = q_2$ and $q_l \leq q_{l+1}$, it follows that $q_2 \leq \widetilde{q}_3 \leq \cdots \leq \widetilde{q}_{j-1}$. Absorbing higher powers of h in a constant C_Z for $0 < h \leq h_0$, we can write

$$\|Z_j - \widetilde{Z}_j\| \leq C_Z h(\varepsilon + h^{q_2+1}).$$

This establishes (33) for j.

To show (34) for j, it suffices to bound $\|\mathcal{R}(Z_j) - \widetilde{Z}_j\|$ since by Lipschitz continuity of F and the approximability assumption (10), we have

$$\begin{aligned} \|P_j F_j - \widetilde{F}_j\| &\leq \|P_j F_j - F_j\| + \|F_j - \widetilde{F}_j\| \\ &\leq \varepsilon + L\|\mathcal{R}(Z_j) - \widetilde{Z}_j\|. \end{aligned}$$

In turn, we can now directly use (23) from the proof of Lemma 3:

$$\|\mathcal{R}(Z_j) - \widetilde{Z}_j\| \le C_{\mathcal{R}} \|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(\widetilde{Z}_j) - \widetilde{Z}_j\| + (1 + C_{\mathcal{R}}) \|\widetilde{Z}_j - Z_j\|.$$

This second term above is (33). For the first, write $\widetilde{Z}_j = \Phi_F^{c_j h}(Y_i) + h\Delta_j$ with $\|\Delta_j\| = C_L h^{q_j}$ due to the stage order condition (30). Then (15) in the proof of Theorem 2 shows that

$$\begin{aligned} \|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(\widetilde{Z}_j) - \widetilde{Z}_j\| &\le \|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(\Phi_F^{c_j h}(Y_i)) - \Phi_F^{c_j h}(Y_i)\| + h\|\Delta_j\| \\ &\le \varepsilon h \max\{1, \mathrm{e}^{\ell c_j h}\} + C_L h^{q_j+1}. \end{aligned} \tag{37}$$

Collecting all the bounds, we arrive at

$$\|P(\mathcal{R}(Z_j))F(\mathcal{R}(Z_j) - F(\widetilde{Z}_j)\| \le C_F h((1+h)\varepsilon + h^{q_j} + h^{q_2+1}) + \varepsilon.$$

Absorbing again higher powers of h in the constant, we have shown (34) for j. Tracing the constants C_Z and C_F through the proof, it is clear that they can be taken as stated in the lemma. □

With this lemma at hand, we can more easily estimate the local error.

Proof of Theorem 6. Using the same notation as in (35), we shall bound the local error of (26) as follows:

$$\|Y_{i+1} - \Phi_F^h(Y_i)\| \le \|Y_{i+1} - \widetilde{Y}_{i+1}\| + \|\widetilde{Y}_{i+1} - \Phi_F^h(Y_i)\|, \tag{38}$$

where $\widetilde{Y}_{i+1} = Y_i + h\sum_{j=1}^s b_j \widetilde{F}_j$ is one step of (25) with $A_i = Y_i$. Since $Y_{i+1} = \mathcal{R}(Y_i + h\sum_{j=1}^s b_j P_j F_j)$, we can use (23) from the proof of Lemma 3 to bound the first term above as

$$\begin{aligned} \|Y_{i+1} - \widetilde{Y}_{i+1}\| &\le C_{\mathcal{R}} \|\mathcal{P}_{\mathcal{M}_\mathbf{r}}(\widetilde{Y}_{i+1}) - \widetilde{Y}_{i+1}\| + (1 + C_{\mathcal{R}}) \|h \sum_{j=1}^s (b_j P_j F_j - \widetilde{F}_j)\| \\ &\le (1 + C_{\mathcal{R}}) \big(\varepsilon h \max\{1, \mathrm{e}^{\ell h}\} + \widetilde{C}_L h^p + h C_B \sum_{j=1}^s \|P_j F_j - \widetilde{F}_j\| \big). \end{aligned}$$

Here, we formally used the same bound as in (37). Now applying Lemma 7, the local error becomes

$$\|Y_{i+1} - \widetilde{Y}_{i+1}\| \le Ch(\varepsilon + h^{\widetilde{q}}), \qquad \widetilde{q} = \begin{cases} q_2 + 1 & \text{if } b_2 \neq 0, \\ \min(q_3 + 1, q_2 + 2) & \text{if } b_2 = 0. \end{cases}$$

since $q_2 = \widetilde{q}_2 \le \widetilde{q}_3 \le \cdots \le \widetilde{q}_s$. The second term in (38) is simply the local error (19). The global error is now obtained in the same way as in the proof of Theorem 2. □

Remark 8. *It is possible to improve Theorem 6 if $a_{j2} = 0$ for all $j \geq 4$ and $b_2 = b_3 = 0$. The final order then satisfies $q = \min(p, q_4 + 1, q_3 + 2, q_2 + 3)$. For example, the (embedded) RK rule 6(5)9b from [32] gives $q = 4$. See the numerical experiments in the next section for details.*

6 Numerical experiments

In this section, we verify numerically the theorems proven above, in particular, the dependence of the convergence on the smallest singular value. We also compare to directly solving the factored formulation (6) by standard methods, and to the specialised projector splitting method from [18].

6.1 Synthetic example

We use an example very similar to that in [13] that allows us to clearly estimate the order. Let D be a diagonal matrix with diagonal entries $d_1, \dots, d_n$, and let Ω_1, Ω_2 be two random skew-symmetric matrices of size $n \times n$. We consider the explicit time dependent matrix

$$A(t) = e^{t\Omega_1} e^{tD} e^{t\Omega_2} \tag{39}$$

with singular values $\sigma_j(t) = e^t d_j$. We take $\log_{10} d_1, \dots, \log_{10} d_n$ to be equidistantly distributed with $d_1 = 1$ and $d_n = \varepsilon$ for a given $\varepsilon < 1$.

We now apply dynamical low-rank for a given rank r on the time interval $0 \leq t \leq 1$ with $F(A(t)) = \dot{A}(t)$. We solve a series of problems for $\varepsilon = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ to investigate the behaviour of the integrators when the singular values of $A(t)$ converge to zero. In Figure 3, the computational results are visible for $r = 5$ and $n = 50$. Clearly, there is step size restriction for explicit Runge–Kutta methods applied to the factored formulation[2]. For our projected Runge–Kutta schemes, the error can be bounded as $C(\varepsilon + h^2)$ and $C(\varepsilon + h^3)$, respectively, and with a constant C uniform in h and ε, as predicted by Theorem 6.

As next experiment, we test two higher order methods. The first, denoted as RK4, is the classical 4th order Runge–Kutta method given in Table 1. Verification

2 Regularisation of S made little difference. For the experiments, we replaced S^{-1} by its Moore–Penrose pseudo-inverse after having put the singular values of S smaller than $10^{-16}\|S\|_2$ to zero. Similar results are obtained using (7).

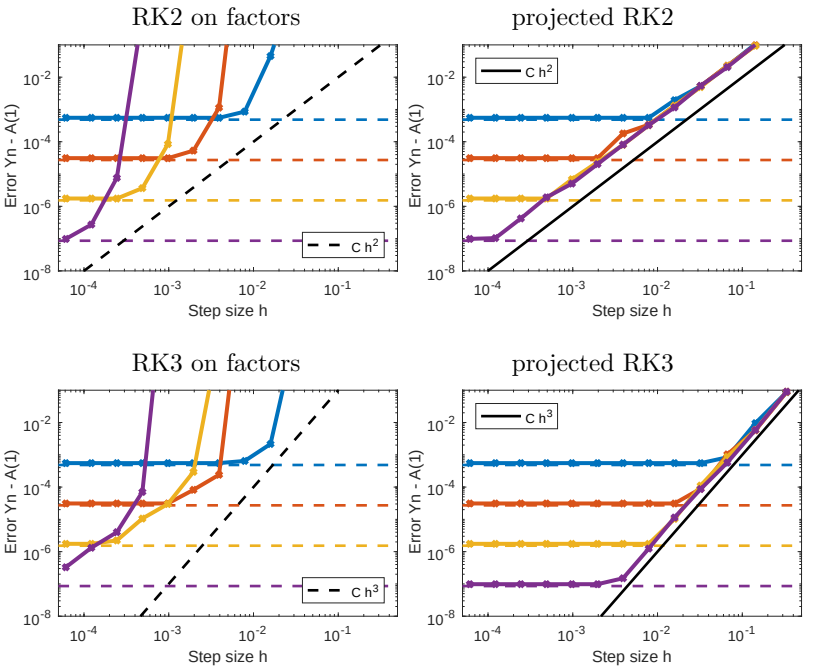

Fig. 3: Time-stepping for (39). Left: standard Runge–Kutta for (6). Right: projected Runge–Kutta methods (26). On each figure, the error is visible in full line from top to bottom for $\varepsilon = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. In dashed line is the error of the best approximation $A(1)$.

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Table 1: Butcher tableau of the classical RK4 method.

0								
$\frac{1}{8}$	$\frac{1}{8}$							
$\frac{1}{6}$	$\frac{1}{18}$	$\frac{1}{9}$						
$\frac{1}{4}$	$\frac{1}{16}$	0	$\frac{3}{16}$					
$\frac{1}{2}$	$\frac{1}{4}$	0	$-\frac{3}{4}$	1				
$\frac{3}{5}$	$\frac{134}{625}$	0	$-\frac{333}{625}$	$\frac{476}{625}$	$\frac{98}{625}$			
$\frac{4}{5}$	$-\frac{98}{1875}$	0	$\frac{12}{625}$	$\frac{10736}{13125}$	$-\frac{1936}{1875}$	$\frac{22}{21}$		
1	$\frac{9}{50}$	0	$\frac{21}{25}$	$-\frac{2924}{1925}$	$\frac{74}{25}$	$-\frac{15}{7}$	$\frac{15}{22}$	
	$\frac{11}{144}$	0	0	$\frac{256}{693}$	0	$\frac{125}{504}$	$\frac{125}{528}$	$\frac{5}{72}$

Table 2: Butcher tableau of rule 6(5)9b from [32].

of (31) gives for the internal stage order

$$q_1 = 0,\ q_2 = q_3 = 1,\ q_4 = 2.$$

Hence, Theorem 6 predicts $q = 2$ since $b_2 \neq 0$.

For the second method, denoted as RK6, we take the rule 6(5)9b from [32]. This is an embedded Runge–Kutta method so we restrict to the rule for $p = 6$. It that has the desired property explained in Remark 8 to obtain a higher order q. In particular, we have

$$q_1 = 0,\ q_2 = 1,\ q_3 = 2,\ q_4 = q_5 = q_6 = q_7 = q_8 = 3.$$

Hence, we obtain $q = 4$ by extending the proof of Theorem 6.

In Figure 4, we repeated the experiment from above but now with $\varepsilon = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$ to better see the faster convergence. Both plots show that initially the error decays as Ch^p but with C dependent on ε. To have a uniform behaviour in ε, the order is lowered to Ch^q. Furthermore, the order for q predicted by the theory seems to be the correct, that is, $q = 2$ for RK4 and $q = 4$ for RK6.

6.2 Non-linear Schrödinger equation

We consider again the non-linear Schrödinger equation from Section 5.1. For the same choice of parameters as before, we compare the projected Runge–Kutta methods of orders $p = 2, 4$, and 6. The result is visible in Figure 5 and we can clearly see that the convergence is robust when increasing the rank r, or

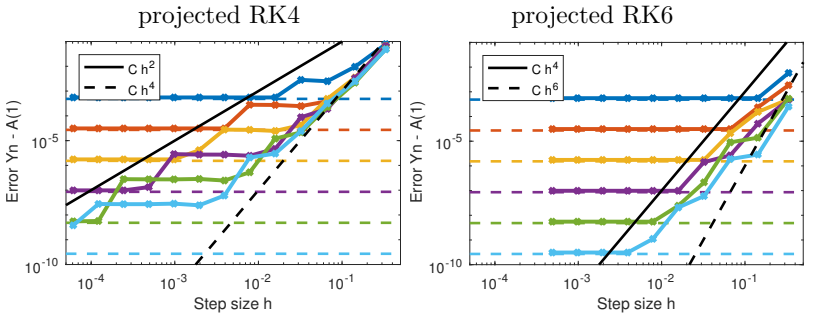

Fig. 4: Time-stepping for (39) with projected Runge–Kutta (26). Left: classical RK4. Right: rule 6(5)9b. On each figure, the error is visible in full line from top to bottom for $\varepsilon = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$. In dashed line is the error of the best approximation $A(1)$.

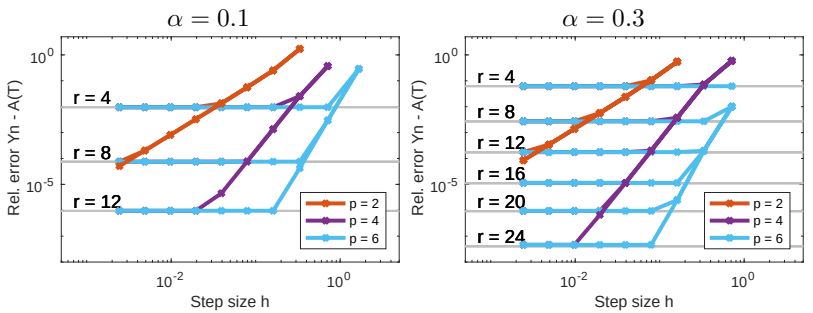

Fig. 5: Time-stepping for (28) with projected Runge–Kutta (26). On each figure, the error is visible in full line from top to bottom for ranks $r = \{4, 8, 12, 16, 20, 24\}$ and orders $p = \{2, 4, 6\}$. In grey line is the error of the best rank r approximation of $A(T)$.

equivalently, when decreasing the smallest singular value. Furthermore, contrary to the synthetic example from above, there seems to be no order reduction for $p = 4$ and 6. This is probably due to the larger modelling error. In addition, we again observe that the rank r approximations constructed by our methods are very close to the corresponding best rank r approximations of the exact solution when h is sufficiently small.

Next, we compare our projected methods with the projector splitting method from [18], denoted as KSL1 and KSL2 for their first and second order formulations. In [13] it has been shown that KSL1 and KSL2 are also robust to small singular values. The computationally most expensive steps in KSL are the solutions of initial value problems of the form $\dot{U} = F(UV_0^T)V_0$ with $F(Y) = \frac{\mathrm{i}}{2}(BY + YB + \alpha|Y|^2Y)$ for a constant $n \times r$ matrix V_0 and a $n \times r$ time-dependent matrix

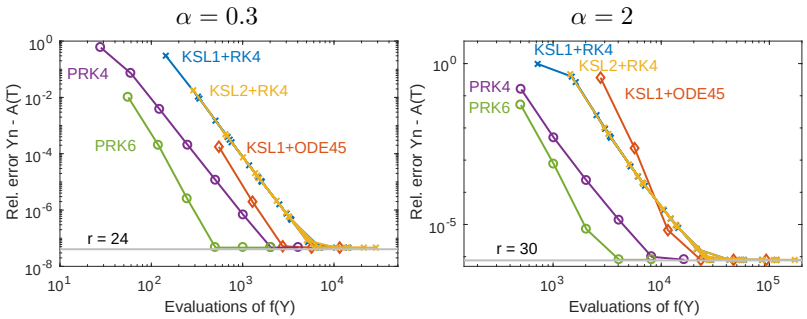

Fig. 6: Performance plot for (28). On each figure, the error is visible in full line of the projected Runge–Kutta (26) of order 4 and 6 (denoted as PRK4 and PRK6) and the projector splitting integrator from [18] (denoted as KSL1+RK4, KSL2+RK4, and KSL1+ODE45). In grey line is the error of the best rank r approximation of $A(T)$

$U(t)$; see [18] for details. To solve these subsystems, we used standard RK4 with fixed step $\tilde{h} = \ell^{-1}h$ for $\ell \in \{5, 10, 15, 20\}$. In addition, we also used `ode45` from MATLAB with relative tolerance 10^{-8}.

The performance of the methods applied to $\alpha = 0.3$ and 2 is visible in Figure 6. The results of all values of $\tilde{h}$ are depicted together for KSL1+RK4 in the same colour, and similarly for KSL2+RK4. As a measure of work we have used the number of evaluations of $F(Y) = \frac{\mathrm{i}}{2}(BY + YB + \alpha|Y|^2Y)$. This is fair since, in both KSL and the projected Runge–Kutta methods, the computationally dominating parts are calculating $|Y|^2Y$ and the projections $U^T(|Y|^2Y)$ and $(|Y|^2Y)V$ for $n \times r$ matrices U and V. Indeed, even when the cubic term $|Y|^2Y$ is directly constructed as a rank r^3 matrix (see, e.g., [7]), we need about $O(nr^3)$ and $O(nr^4)$ flops for these operations. The rest of the computational time is about $O(nr^2)$ in KSL and $O(ns^2r^2)$ in an s-stage projected Runge–Kutta. Typically $s \ll r$ ($s \leq 8$ and $r \geq 24$ in our numerics) and so the calculation associated to $|Y|^2Y$ easily dominates. This remains true if accelerated techniques for recompressing the Hadamard product are used, as in [15].

Figure 6 shows that PRK6 and PRK4 are both more efficient then any combination of KSL since, on average, they need about 10 times less function evaluations for the same accuracy. Observe that there is no noticeable difference between KSL1 and KSL2, which was already observed in [13, §4.1]. In the left plot, KSL with `ode45` was more efficient than fixed time stepping since the relative tolerance of 10^{-8} for the adaptive time stepper corresponds well with the best approximation error. On the right, 10^{-8} is too low and that method becomes less efficient. This shows that the integration of the substeps in KSL has to be done efficiently.

Acknowledgment: The second author thanks Daniel Kressner for initial discussions about projected integrators in the context of low-rank matrices.

Funding: B.V. was partly supported by SNF project 159856 entitled "Analyse numérique".

A Tightness of curvature bound

We here show that the bound (5) is sharp, in the sense that for any X, we can choose Y and Z such that the bound is attained. To keep the notation manageable, we restrict ourselves to the case of square matrices.

We consider the manifold $\mathcal{M}_r$ of real $N \times N$ matrices of rank r. Let $X \in \mathcal{M}_r$ be any matrix on the manifold, and construct its SVD $X = USV^T$ such that $U, V \in \mathbb{R}^{N\times r}$ have orthonormal columns u_i and v_i, respectively, and $S \in \mathbb{R}^{r\times r}$ is a diagonal matrix with elements $s_{ii} = \sigma_i > 0$ in decreasing order. Next, choose $Y \in \mathcal{M}_r$ such that $Y = US\widetilde{V}^T$, where also $\widetilde{V} \in \mathbb{R}^{N\times r}$ has orthonormal columns $\tilde{v}_i$, and $v_i = \tilde{v}_i$ for $i = 1, \dots, r-1$. Then,

$$Y - X = \sigma_r u_r (\tilde{v}_r - v_r)^T,$$

and thereby

$$\|Y - X\|^2 = \operatorname{tr}((Y - X)^T (Y - X)) = 2\sigma_r^2 (1 - \tilde{v}_r^T v_r). \tag{40}$$

Now take any $Z \in \mathbb{R}^{N\times N}$. Since $P(X)Z = UU^T Z + ZVV^T - UU^T ZVV^T$ (see, e.g., [18, eq. (2.5)]), we get

$$(P(Y) - P(X))Z = (I - UU^T)Z(\widetilde{V}\widetilde{V}^T - VV^T) = (I - UU^T)Z(\tilde{v}_r\tilde{v}_r^T - v_r v_r^T).$$

We choose $Z = \tilde{u}\tilde{v}_r^T$, where $\tilde{u}$ is normalised and orthogonal to the columns of U. Then, $\|Z\| = 1$ and

$$\|(P(Y) - P(X))Z\|^2 = \|\tilde{u}(\tilde{v}_r^T - \tilde{v}_r^T v_r v_r^T)\|^2 = 1 - (\tilde{v}_r^T v_r)^2.$$

Since v_r and $\tilde{v}_r$ are normalised, $|\tilde{v}_r^T v_r| \leq 1$. We choose $\tilde{v}_r$ such that $0 < \tilde{v}_r^T v_r < 1$. Then,

$$\|(P(Y) - P(X))Z\| = \sqrt{1 - (\tilde{v}_r^T v_r)^2} \geq \sqrt{1 - \tilde{v}_r^T v_r}.$$

and by (40) and $\|Z\| = 1$, we get

$$\|(P(Y) - P(X))Z\| \geq \frac{1}{\sqrt{2}\sigma_r}\|Y - X\|\|Z\|,$$

which shows the claim that (5) is essentially a tight estimate.

References

[1] P.-A. Absil and I. V. Oseledets. Low-rank retractions: a survey and new results. *Comput. Optim. Appl.*, 62:5–29, 2015.

[2] M. Beck, A. Jäckle, G. Worth, and H.-D. Meyer. The multiconfiguration time-dependent Hartree (MCTDH) method: A highly efficient algorithm for propagating wavepackets. *Phys. Rep.*, 324:1–105, 2000.

[3] D. Braess and W. Hackbusch. Approximation of $1/x$ by exponential sums in $[1, \infty)$. *IMA J. Numer. Anal.*, 25:685–697, 2005.

[4] P. A. M. Dirac. Note on exchange phenomena in the Thomas atom. *Math. Proc. Cambridge Philos. Soc.*, 26:376–385, 1930.

[5] L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72:247–265, 2004.

[6] L. Grasedyck, D. Kressner, and C. Tobler. A literature survey of low-rank tensor approximation technique. *GAMM-Mitt.*, 36:53–78, 2013.

[7] W. Hackbusch. *Tensor Spaces and Numerical Tensor Calculus*. Springer, 2012.

[8] W. Hackbusch. New estimates for the recursive low-rank truncation of block-structured matrices. *Num. Math.*, 132(2):303–328, 2016.

[9] E. Hairer. Solving differential equations on manifolds. Lecture notes, University of Geneva, 2011.

[10] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, Berlin, 2002.

[11] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer, Berlin, 2nd edition, 1993.

[12] T. Jahnke and W. Huisinga. A dynamical low-rank approach to the chemical master equation. *Bull. Math. Biol.*, 70:2283–2302, 2008.

[13] E. Kieri, C. Lubich, and H. Walach. Discretized dynamical low-rank approximation in the presence of small singular values. *SIAM J. Numer. Anal.*, 54:1020–1038, 2016.

[14] P. Kramer and M. Saraceno. *Geometry of the Time-Dependent Variational Principle in Quantum Mechanics*, volume 140 of *Lecture Notes in Physics*. Springer, Berlin, 1981.

[15] D. Kressner and L. Periša. Recompression of Hadamard products of tensors in Tucker format. *SIAM J. Sci. Comput.*, 39(5):A1879–A1902, 2017.

[16] D. Kressner, M. Steinlechner, and B. Vandereycken. Preconditioned low-rank

Riemannian optimization for linear systems with tensor product structure. *SIAM J. Sci. Comput.*, 38:A2018–A2044, 2016.

[17] C. Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. European Math. Soc., Zürich, 2008.

[18] C. Lubich and I. V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT*, 54:171–188, 2014.

[19] C. Lubich, I. V. Oseledets, and B. Vandereycken. Time integration of tensor trains. *SIAM J. Numer. Anal.*, 53:917–941, 2015.

[20] C. Lubich, T. Rohwedder, R. Schneider, and B. Vandereycken. Dynamical approximation by hierarchical Tucker and tensor-train tensors. *SIAM J. Matrix Anal. Appl.*, 34:470–494, 2013.

[21] U. Manthe, H. Meyer, and L. S. Cederbaum. Wave-packet dynamics within the multiconfiguration Hartree framework: General aspects and application to NOCl. *J. Chem. Phys.*, 97:3199–3213, 1992.

[22] H.-D. Meyer, F. Gatti, and G. A. Worth, editors. *Multidimensional Quantum Dynamics: MCTDH Theory and Applications*. Wiley, Weinheim, 2009.

[23] H.-D. Meyer, U. Manthe, and L. S. Cederbaum. The multi-configurational time-dependent Hartree approach. *Chem. Phys. Lett.*, 165:73–78, 1990.

[24] F. Musharbash, E. Nobile and T. Zhou. Error analysis of the dynamically orthogonal approximation of time dependent random PDEs. *SIAM J. Sci. Comput.*, 37(2):A776–A810, 2015.

[25] A. Nonnenmacher and C. Lubich. Dynamical low-rank approximation: applications and numerical experiments. *Math. Comput. Simulation*, 79(4):1346–1357, 2008.

[26] I. V. Oseledets. Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33:2295–2317, 2011.

[27] I. V. Oseledets and E. E. Tyrtyshnikov. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM J. Sci. Comput.*, 31:3744–3759, 2009.

[28] T. Rohwedder and A. Uschmajew. On local convergence of alternating schemes for optimization of convex problems in the tensor train format. *SIAM J. Numer. Anal.*, 51:1134–1162, 2013.

[29] U. Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Ann. Physics*, 326:96–192, 2011.

[30] A. Trombettoni and A. Smerzi. Discrete solitons and breathers with dilute Bose–Einstein condensates. *Phys. Rev. Lett.*, 86:2353–2356, 2001.

[31] A. Uschmajew and B. Vandereycken. The geometry of algorithms using hierarchical tensors. *Linear Algebra Appl.*, 439:133–166, 2013.

[32] J. H. Verner. Some Runge–Kutta formula pairs. *SIAM J. Numer. Anal.*, 28:496–511, 1991.

[33] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac. Matrix product density operators: Simulation of finite-temperature and dissipative systems. *Phys. Rev. Lett.*, 93:207204, 2004.