CONVERGENCE OF PARAOPT FOR GENERAL RUNGE-KUTTA TIME DISCRETIZATIONS*

FELIX KWOK[†], JULIEN SALOMON[‡], AND DJAHOU N. TOGNON [‡]

Abstract. ParaOpt is a time parallel method based on Parareal for solving optimality systems arising in optimal control problems. The method was presented in [M.J. Gander, F. Kwok and J. Salomon, *SIAM J. Sci. Comput.,* 42 (2020), A2773–A2802] together with a convergence analysis in the case where implicit Euler is used to discretize the differential equations governing the system dynamics. However, its convergence behaviour for higher order time discretizations has not been considered. In this paper, we use an operator norm analysis to prove that the convergence rate of ParaOpt applied to a linear-quadratic optimal control problem has the same order as the Runge-Kutta time integration method used, provided that a few auxiliary order conditions are satisfied. We illustrate our theoretical results with numerical examples, before showing an additional test case not covered by our analysis, namely, a nonlinear optimal control problem involving a Schrödinger type system.

Key words. ParaOpt, Optimal control, Preconditioning, Runge-Kutta methods

MSC codes. 65Y05, 49J15, 65F10

1. Introduction.

Time parallelization algorithms for solving ordinary differential equations (ODEs) or partial differential equations (PDEs) have reached a certain maturity since their emergence in the early 2000s [11, 29]. Current research on this class of algorithms mainly focuses on their use in conjunction with other types of procedures, such as data assimilation [9, 30, 7], multi-grid [18, 10, 35, 32, 15, 4], or control [25, 1, 34, 28, 21, 2, 5, 23, 27, 26].

In the recent paper [14], the authors introduced ParaOpt, a way of parallelizing the solving of the discrete Euler-Lagrange equations arising from an optimal control problem. The strategy consists of dividing the time interval into several sub-intervals, on which subdomain control problems can be defined. The original "global problem" is then equivalent to the continuity of the state and adjoint variables across time interval boundaries, and these continuity conditions can be expressed in terms of the subdomain solution operators. The algorithm is essentially a multiple shooting method and has the dual advantage of reducing the number of primary degrees of freedom to those of the state and adjoint at a few time points, while allowing the sub-interval problems to be solved in parallel. The main drawback is that even when linearized, the resulting Jacobian matrix is expensive to calculate explicitly and to solve by factorization. The idea of ParaOpt, inspired by Parareal [24, 12] for initial value problems, is to replace the expensive solve by a cheaper one that uses a coarser discretization in time. For an implicit Euler discretization, [14] contains a theoretical convergence analysis of the algorithm for linear dissipative problems: the convergence rate is found to be essentially an affine function of Δt, the time-step size for the coarse approximation. This first analysis of convergence of ParaOpt is based on a study of an eigenvalue problem. However, additional numerical results have shown the efficiency of ParaOpt for problems that are not of the dissipative type.

The goal of this paper is to investigate what happens to the convergence rate of ParaOpt when higher order discretizations is used for the optimal control problem. Unlike previous works which are mainly based on eigenvalues problems, our approach is based on operator analysis. Our framework is that of Runge-Kutta methods, where the approximation orders are central parameters. We will see that under appropriate assumptions, ParaOpt converges with a contraction factor proportional to Δt^k, where Δt and k are the time step and order of the coarser of the two time integration methods involved in the ParaOpt. In other words, ParaOpt performs better as an iterative method when higher order discretizations are used.

Our approach is strongly inspired by works dealing with the time discretization of optimality systems. In this field, an early work by Hager [16] on the convergence rate of discrete controls towards the continuous one showed the importance of using the same integrator for the objective function and the ODE under consideration. Otherwise, difficulties can arise if the discrete objective function is "incompatible" with the ODE method, as reported in [8], where a second-order approach is detailed. A more general analysis is presented in [17], where higher order Runge-Kutta methods are studied. In these works, a higher order discretization is required for the objective function

[*] Submitted to the editors DATE.

Funding: This work was funded by ANR/RGC ALLOWAPP project (ANR-19-CE46-0013/A-HKBU203/19) and ANR project DEEPNUM (ANR-21-CE23-0017).

[†] Département de mathématiques et de statistique, Université Laval, Québec G1V 0A6, Canada,(felix.kwok@mat.ulaval.ca)

[‡] Sorbonne Université, Université Paris Cité, CNRS, INRIA, Laboratoire Jacques-Louis Lions, LJLL, EPC ANGE, F-75005 Paris, France (julien.salomon@inria.fr, djahou-norbert.tognon@inria.fr),

to obtain solutions as accurate as can be expected when using Runge-Kutta. This introduces additional degrees of freedom to the control, which makes the discrete problem even larger and computationally more expensive to solve. Parallelization strategies, as the one of this paper, are therefore a must. Beyond accuracy, in some cases it is necessary to use specific Runge–Kutta methods to ensure that optimization leads to a physically correct solution. An example is given by the control of Schrödinger-type equations (such as discussed in Subsection 5.3) or transport-type equations (see [22]), for which a norm is preserved at the continuous level. If this norm is not preserved at the discrete level, the optimization method can take advantage of this property to artificially decrease or increase the norm in order to optimize the functional under consideration. This leads to non-physical solutions or numerical explosion and prevents the problem from being solved. One can overcome this issue by choosing a numerical scheme that preserves the norm, such as Crank-Nicolson. Having a sufficiently wide choice of numerical schemes is therefore a crucial property for the numerical coupling of solvers.

Our paper is organized as follows. In Section 2, we introduce our optimal control problem together with a corresponding general Runge-Kutta time discretization. Section 3 provides a detailed exposition of the ParaOpt procedure in this discrete framework. In Section 4, our main results are stated and proved. We conclude in Section 5 with some numerical experiments which include a nonlinear example.

2. Optimal control problem and its discretization.

We consider the linear-quadratic optimal control problem associated with the following cost functional

$$\mathfrak{J}(\nu) = \frac{1}{2}\left\|y(T) - y_{tg}\right\|^2 + \frac{\alpha}{2}\int_0^T \|\nu(t)\|^2 dt, \tag{2.1}$$

where α is a regularization parameter, y_{tg} is a target state and $\nu : [0,T] \longrightarrow \mathbb{R}^m$ (m is a positive integer) the control, which is assumed to be in $L^2([0,T],\mathbb{R}^m)$. The evolution of the state function $y : [0,T] \longrightarrow \mathbb{R}^r$ (r is a positive integer) is described by the system of ordinary differential equations (ODEs):

$$\dot{y}(t) = \mathcal{L}y(t) + \mathcal{B}\nu(t) \tag{2.2}$$

with initial condition $y(0) = y_{in}$ where $\mathcal{L} \in \mathbb{R}^{r\times r}$ and $\mathcal{B} \in \mathbb{R}^{r\times m}$. The ODEs system (2.2) may arise from a semi-discretization in space of a time-dependent partial differential equation (PDE) and $\|\cdot\|$ is derived from the standard inner product $\langle\cdot,\cdot\rangle$ of $\mathbb{R}^r$. It is well known (cf. [33, Ch. 4]) that for any initial state y_{in}, the problem of minimizing $\mathfrak{J}(\nu)$ subject to (2.2) and $y(0) = y_{in}$ has a unique solution $\nu^* \in L^2([0,T],\mathbb{R}^m)$, with the corresponding optimal trajectory $y^* \in H^1([0,T],\mathbb{R}^r)$ and therefore absolutely continuous. Moreover, the unique optimal control ν^* is characterized by first order optimality conditions that can be obtained formally from the Euler-Lagrange equations. To write them, we introduce the Langrangian

$$\mathfrak{L}(y,\lambda,\nu) = \mathfrak{J}(\nu) - \int_0^T \langle\lambda(t), \dot{y}(t) - \mathcal{L}y(t) - \mathcal{B}\nu(t)\rangle\, dt,$$

where λ is the adjoint state function. Differentiating with respect to ν and setting the derivative to zero leads to the relation

$$\nu^*(t) = -\frac{1}{\alpha}\mathcal{B}^T\lambda(t), \tag{2.3}$$

where λ^* is a solution of the adjoint equation

$$\begin{aligned} \dot{\lambda}(t) &= -\mathcal{L}^T\lambda(t), \qquad t \in (0,T) \\ \lambda(T) &= y^*(T) - y_{tg}. \end{aligned}$$

The adjoint λ^* is therefore a C^∞ function in t, and so is ν^* thanks to the relation (2.3). This in turn implies $y^* \in C^\infty([0,T],\mathbb{R}^r)$. These smoothness properties motivate the use of higher order Runge-Kutta discretizations, which we will present in the next section.

Substituting the optimal control (2.3) into (2.2) leads to the following reduced optimality system

$$\begin{gathered} \dot{y}(t) = \mathcal{L}y(t) - \frac{1}{\alpha}\mathcal{B}\mathcal{B}^T\lambda(t), \quad \dot{\lambda}(t) = -\mathcal{L}^T\lambda(t), \quad t \in (0,T) \\ y(0) = y_{in}, \quad \lambda(T) = y(T) - y_{tg}. \end{gathered} \tag{2.4}$$

2.1. Time discretization. Given $M_0 \in \mathbb{N}$, $M_0 > 0$, the time interval $[0, T]$ is discretized by $M_0 + 1$ grid points $t_0 = 0, \dots, t_{M_0} = T$ with $\delta t = T/M_0$ and $t_n = n\delta t$. We follow [17] and augment the ODE system $\dot{y} = \mathcal{L}y + \mathcal{B}\nu$ to include the integral term in the objective function $\mathfrak{J}$ as follows: we define $\mathsf{Y} : [0, T] \longrightarrow \mathbb{R}^{r+1}$, $\mathsf{Y} = (\mathsf{y}^0, \mathsf{y}^1, \dots, \mathsf{y}^r)^T$, such that

$$\text{(2.5)} \qquad \begin{cases} \dot{\mathsf{y}}^0(t) = \dfrac{\alpha}{2}\|\nu(t)\|^2, & \mathsf{y}^0(0) = 0, \\ \begin{pmatrix} \dot{\mathsf{y}}^1(t) \\ \vdots \\ \dot{\mathsf{y}}^r(t) \end{pmatrix} = \mathcal{L}\begin{pmatrix} \mathsf{y}^1 \\ \vdots \\ \mathsf{y}^r \end{pmatrix} + \mathcal{B}\nu(t), & \begin{pmatrix} \mathsf{y}^1(0) \\ \vdots \\ \mathsf{y}^r(0) \end{pmatrix} = y_{in}. \end{cases}$$

Then $y^i(t) = \mathsf{y}^i(t)$ for $i \geq 1$, and minimizing $\mathfrak{J}$ is equivalent to minimizing $C(\mathsf{Y}(T))$, where

$$C(\mathsf{Y}(t)) = \frac{1}{2}\|\mathsf{y}^{1:r}(t) - y_{tg}\|^2 + \mathsf{y}^0(t).$$

From now on, we identify $\mathsf{y}^{1:r}(t)$ with $y(t)$ to lighten the notation. We now apply an s-stage Runge-Kutta method of order $p \geq 1$ to the augmented system (2.5). Suppose its Butcher table is given by

$$\text{(RK)} \qquad \begin{array}{c|c} c & A \\ \hline & b^T \end{array}\ ;$$

with coefficients $A = [a_{i,j}]_{i,j=1}^s$, $b = (b_j)_{j=1}^s$ and $c = (c_j)_{j=1}^s$. Concretely, by applying this Runge-Kutta method to (2.5), we obtain the following discrete optimal control problem:

$$\text{(2.6a)} \qquad \text{minimize} \quad \frac{1}{2}\|y_{M_0} - y_{tg}\|^2 + \mathsf{y}^0_{M_0}$$

$$\text{(2.6b)} \qquad \text{subject to} \quad k_{n,i} = \mathcal{L}\left(y_n + \delta t \sum_{j=1}^s a_{ij}k_{n,j}\right) + \mathcal{B}\nu_{n,i}, \qquad i = 1, \dots, s,$$

$$\text{(2.6c)} \qquad \mathsf{y}^0_{n+1} = \mathsf{y}^0_n + \frac{\alpha\delta t}{2}\sum_{i=1}^s b_i\|\nu_{n,i}\|^2, \qquad n = 0, \dots, M_0 - 1,$$

$$\text{(2.6d)} \qquad y_{n+1} = y_n + \delta t\sum_{i=1}^s b_i k_{n,i}, \qquad n = 0, \dots, M_0 - 1.$$

Note that since the stage equations for the zeroth component y^0_n are particularly simple, we have written the recurrence relation for this component separately in (2.6c). From (2.6a) and (2.6c), we see that the discrete objective function to be minimized can be written as

$$\text{(2.7)} \qquad \mathfrak{J}_{\delta t}(\nu) = \frac{1}{2}\|y_{M_0} - y_{tg}\|^2 + \frac{\alpha\delta t}{2}\sum_{n=0}^{M_0-1}\sum_{j=1}^s b_j\|\nu_{n,j}\|^2.$$

We have thus discretized the integral in (2.1) in a way that is consistent with the Runge-Kutta method. Moreover, one can express the $k_{n,i}$ in (2.6b) as the solution of a linear system. Substituting this solution into (2.6d), we obtain

$$y_{n+1} = y_n + \delta t\begin{pmatrix} b_1 I & b_2 I & \cdots & b_s I \end{pmatrix}(\mathcal{I} - \delta t A \otimes \mathcal{L})^{-1}\begin{pmatrix} \mathcal{L}y_n + \mathcal{B}\nu_{n,1} \\ \vdots \\ \mathcal{L}y_n + \mathcal{B}\nu_{n,s} \end{pmatrix},$$

and where I and $\mathcal{I}$ are identity matrices of $\mathbb{R}^{r\times r}$ and $\mathbb{R}^{r\cdot s\times r\cdot s}$ respectively. Then, denoting by $\mathcal{W}$ and $\mathcal{W}_j$ the matrices

$$\text{(2.8)} \qquad \mathcal{W}_j := \sum_{i=1}^s b_i\mathcal{Z}_{i,j}, \quad \mathcal{W} := \sum_{j=1}^s \mathcal{W}_j,$$

where the matrices $\mathcal{Z}_{i,j}$ are the blocks of the inverse of $\mathcal{I} - \delta t A \otimes \mathcal{L}$, we get

$$\text{(2.9)} \qquad y_{n+1} = (I + \delta t\mathcal{W}\mathcal{L})y_n + \delta t\sum_{j=1}^s \mathcal{W}_j\mathcal{B}\nu_{n,j}.$$

Table 1

Order conditions for Runge-Kutta methods for optimal control. We assume $c_i = \sum_{j=1}^s a_{ij}$, $d_j = \sum_{i=1}^s b_i a_{ij}$.

Order	IVP conditions	Optimal control conditions
1	$\sum_{i=1}^s b_i = 1$	
2	$\sum_{i=1}^s d_i = \frac{1}{2}$	
3	$\sum_{i=1}^s c_i d_i = \frac{1}{6}$, $\sum_{i=1}^s b_i c_i^2 = \frac{1}{3}$	$\sum_{i=1}^s d_i^2/b_i = \frac{1}{3}$
4	$\sum_{i=1}^s b_i c_i^3 = \frac{1}{4}$, $\sum_{i,j=1}^s b_i c_i a_{ij} c_j = \frac{1}{8}$,	$\sum_{i=1}^s d_i^3/b_i^2 = \frac{1}{4}$, $\sum_{i,j=1}^s d_i a_{ij} d_j/b_j = \frac{1}{8}$,
	$\sum_{i=1}^s d_i c_i^2 = \frac{1}{12}$, $\sum_{i,j=1}^s d_i a_{ij} c_j = \frac{1}{24}$	$\sum_{i=1}^s c_i d_i^2/b_i = \frac{1}{12}$, $\sum_{i,j=1}^s b_i c_i a_{ij} d_j/b_j = \frac{5}{24}$

To obtain the first-order optimality conditions, let us introduce the discrete Lagrangian

$$\mathfrak{L}_{\delta t}(y,\lambda,\nu) = \mathfrak{J}_{\delta t}(\nu) - \sum_{n=0}^{M_0-1} \langle \lambda_{n+1}, y_{n+1} - (I + \delta t \mathcal{W}\mathcal{L})y_n - \delta t \sum_{j=1}^{s} \mathcal{W}_j \mathcal{B} \nu_{n,j} \rangle.$$

The discrete Euler-Lagrange equations then become

$$\begin{aligned} y_{n+1} &= (I + \delta t \mathcal{W}\mathcal{L})y_n + \delta t \sum_{j=1}^{s} \mathcal{W}_j \mathcal{B} \nu_{n,j}, \\ \lambda_n &= (I + \delta t \mathcal{W}\mathcal{L})^T \lambda_{n+1}, \\ \lambda_{M_0} &= y_{M_0} - y_{tg}, \\ \alpha b_j \nu_{n,j} &= -\mathcal{B}^T \mathcal{W}_j^T \lambda_{n+1}. \end{aligned} \tag{2.10}$$

We now use the last equation to eliminate the control and to obtain the following discrete version of the reduced optimality system (2.4):

$$y_{n+1} = (I + \delta t \mathcal{W}\mathcal{L})y_n - \frac{\delta t}{\alpha} \left(\sum_{j=1}^{s} \frac{1}{b_j} \mathcal{W}_j \mathcal{B}\mathcal{B}^T \mathcal{W}_j^T \right) \lambda_{n+1} \tag{2.11}$$

$$\lambda_n = (I + \delta t \mathcal{W}\mathcal{L})^T \lambda_{n+1}, \tag{2.12}$$

with the initial and final conditions $y_0 = y_{in}$ and $\lambda_{M_0} = y_{M_0} - y_{tg}$ respectively.

For any given (not necessarily optimal) control function $\nu \in C^\infty([0,T],\mathbb{R}^r)$, if we let $\nu_{n,j} = \nu(t_n + c_j \delta t)$ in (2.6b), then the implicit function theorem implies that the system (2.6b) has a unique solution $(k_{n,i})_{1\le i\le s}$ for δt sufficiently small, and the initial-value problem (IVP) order conditions (see Table 1) imply that Y_n converges to $Y(t_n)$ with order p. However, if one is free to choose the values $\nu_{n,j}$ to minimize (2.7) in the discrete sense, it does not automatically follow that y_n and $\nu_{n,j}$ converge to $y^*(t_n)$ and $\nu^*(t_n + c_j\delta t)$ with order p. Indeed, Hager showed in [17] that additional order conditions are required for convergence of the the discrete state and adjoint at the correct order. We therefore assume that our Runge-Kutta method satisfies all the order conditions in Table 1 up to order p, both in the IVP and in the optimal control sense. In section 4, we will adapt Hager's proof of consistency in [17] to our linear-quadratic problem in order to bound the norms of certain matrices, which in turn will allow us to estimate the convergence rate of ParaOpt for these higher order discretizations.

3. Time Parallelization using ParaOpt.

The solution of the coupled forward-backward system (2.11)–(2.12) can be parallelized using the ParaOpt method [14]. A simplified version for the discrete linear-quadratic control problem is given below. Let us consider the subdivision of $[0,T]$ into L sub-intervals $[T_\ell, T_{\ell+1}]$ with uniform length ΔT that satisfies $T_\ell = \ell \Delta T, \ell = 0,\dots,L$, and $\Delta T = M\delta t$, These quantities are illustrated (among others that will be introduced later) on Figure 1.

We start by eliminating interior unknowns in $[T_\ell, T_{\ell+1}]$, i.e., the unknowns that are not located at $T_0, T_1, \dots, T_L$. For $0 \le n_1 \le n_2 \le M$, (2.12) implies that

$$\lambda_{n_2 - n} = [(I + \delta t \mathcal{W}\mathcal{L})^T]^n \lambda_{n_2}, \tag{3.1}$$

and combining (2.11) and (2.12), we obtain

$$y_{n_2} = (I + \delta t \mathcal{W}\mathcal{L})^{n_2 - n_1} y_{n_1} - \frac{\delta t}{\alpha} \sum_{n=0}^{n_2 - n_1 - 1} (I + \delta t \mathcal{W}\mathcal{L})^n \left(\sum_{j=1}^{s} \frac{1}{b_j} \mathcal{W}_j \mathcal{B}\mathcal{B}^T \mathcal{W}_j^T \right) \lambda_{n_2 - n}. \tag{3.2}$$

Substituting (3.1) into (3.2) then leads to

$$\begin{aligned} y_{n_2} =& (I+\delta t\mathcal{W}\mathcal{L})^{n_2-n_1} y_{n_1} \\ &- \frac{\delta t}{\alpha}\left(\sum_{n=0}^{n_2-n_1-1}(I+\delta t\mathcal{W}\mathcal{L})^n\left(\sum_{j=1}^{s}\frac{1}{b_j}\mathcal{W}_j\mathcal{B}\mathcal{B}^T\mathcal{W}_j^T\right)[(I+\delta t\mathcal{W}\mathcal{L})^T]^n\right)\lambda_{n_2}. \end{aligned} \tag{3.3}$$

Furthermore, setting $n_1 = (\ell-1)M$, $n_2 = \ell M$ and introducing the notation $Y_\ell := y_{\ell M}$ and $\Lambda_\ell := \lambda_{\ell M}$ into (3.1) and (3.3) yields

$$Y_0 = y_0$$

$$-\mathcal{P}_{\delta t}Y_{\ell-1} + Y_\ell + \frac{1}{\alpha}\mathcal{R}_{\delta t}\Lambda_\ell = 0, \qquad 1 \le \ell \le L, \tag{3.4}$$

$$\Lambda_{\ell-1} - \mathcal{P}_{\delta t}^T\Lambda_\ell = 0, \qquad 2 \le \ell \le L, \tag{3.5}$$

$$-Y_L + \Lambda_L = -y_{tg}, \tag{3.6}$$

where

$$\mathcal{P}_{\delta t} := (I+\delta t\mathcal{W}\mathcal{L})^M, \tag{3.7}$$

$$\mathcal{R}_{\delta t} := \delta t\sum_{n=0}^{M-1}(I+\delta t\mathcal{W}\mathcal{L})^n\left(\sum_{j=1}^{s}\frac{1}{b_j}\mathcal{W}_j\mathcal{B}\mathcal{B}^T\mathcal{W}_j^T\right)[(I+\delta t\mathcal{W}\mathcal{L})^T]^n. \tag{3.8}$$

In matrix form, this system reads

$$\mathcal{M}_{\delta t}X = \mathbf{f}, \tag{3.9}$$

with

$$\mathcal{M}_{\delta t} := \left(\begin{array}{cccc|cccc} I & & & & 0 & & & \\ -\mathcal{P}_{\delta t} & \ddots & & & \mathcal{R}_{\delta t}/\alpha & & \ddots & \\ & \ddots & \ddots & & & & \ddots & 0 \\ & & -\mathcal{P}_{\delta t} & I & & & & \mathcal{R}_{\delta t}/\alpha \\ \hline & & & & I & -\mathcal{P}_{\delta t}^T & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \ddots & -\mathcal{P}_{\delta t}^T \\ & & & -I & & & & I \end{array}\right), X := \left(\begin{array}{c} Y_0 \\ \vdots \\ \vdots \\ Y_L \\ \hline \Lambda_1 \\ \vdots \\ \vdots \\ \Lambda_L \end{array}\right), \ \mathbf{f} := \begin{pmatrix} y_{in} \\ 0 \\ \\ \vdots \\ \\ 0 \\ -y_{tg} \end{pmatrix}.$$

The matrix blocks $\mathcal{P}_{\delta t}$ and $\mathcal{R}_{\delta t}$ can be interpreted as follows.

- $\mathcal{P}_{\delta t}$ represents the forward propagator that maps Y_i to the discrete Runge-Kutta approximation of $y(T_{i+1})$ in the following initial value problem:

  $$\dot{y} = \mathcal{L}y \quad \text{on } [T_i, T_{i+1}], \qquad y(T_i) = Y_i.$$

  Note that this continuous problem can be solved explicitly to give

  $$y(t) = e^{(t-T_i)\mathcal{L}}Y_i \implies y(T_{i+1}) = e^{\Delta T\mathcal{L}}Y_i =: \mathcal{P}_0Y_i, \tag{3.10}$$

  where $\Delta T = T_{i+1} - T_i$. Here, the notation $e^{t\mathcal{L}}$ denotes the matrix exponential operator of $t\mathcal{L}$, so that $e^{t\mathcal{L}}Y_i$ is defined by

  $$e^{t\mathcal{L}}\ Y_i = \sum_{n=0}^{\infty}\frac{1}{n!}(t\mathcal{L})^nY_i.$$

- $\mathcal{P}_{\delta t}^T$, the transpose of $\mathcal{P}_{\delta t}$, can be interpreted as a discretization of the backward propagator that maps the adjoint Λ_{i+1} to the discrete Runge-Kutta approximation of $\lambda(T_i)$ in the adjoint problem

  $$\dot{\lambda} = -\mathcal{L}^T\lambda \quad \text{on } [T_i, T_{i+1}], \qquad \lambda(T_{i+1}) = \Lambda_{i+1},$$

  provided that the optimal control conditions in Table 1 are satisfied, see [17]. This continuous problem has an exact solution given by

  $$\lambda(t) = e^{(T_{i+1}-t)\mathcal{L}^T}\Lambda_{i+1} \implies \lambda(T_i) = e^{\Delta T\mathcal{L}^T}\Lambda_{i+1} = \mathcal{P}_0^T\Lambda_{i+1}. \tag{3.11}$$

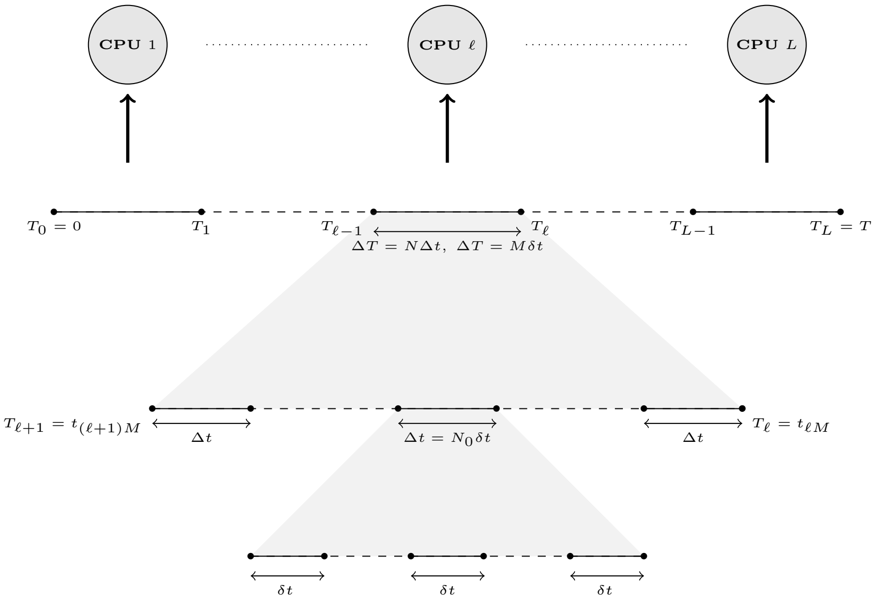

FIG. 1. *Hierarchy of the time parallel discretization: sub-intervals $[T_\ell, T_{\ell+1}]$ of length ΔT, coarse grid with time step Δt and finally fine grid with time step δt.*

- $\mathcal{R}_{\delta t}$ represents the backward-forward propagator that maps Λ_{i+1}, the adjoint at T_{i+1}, to the discrete Runge-Kutta approximation of $-y(T_{i+1})$ in the following coupled forward-backward problem:

$$\begin{aligned} &\dot{y} = \mathcal{L}y - \mathcal{B}\mathcal{B}^T\lambda \quad &&\text{on } [T_i, T_{i+1}], \qquad &&y(T_i) = 0,\\ &\dot{\lambda} = -\mathcal{L}^T\lambda &&\text{on } [T_i, T_{i+1}], &&\lambda(T_{i+1}) = \Lambda_{i+1}. \end{aligned}$$

This problem is of the same form as (2.4), but constrained to the interval $[T_i, T_{i+1}]$ and with $\alpha = 1$; since the problem is linear in Λ_{i+1}, an explicit division by α as in (3.4) brings us back to the correct operator for the original problem. By substituting the exact solution for $\lambda(t)$ in (3.11), we can integrate the ODE in y to obtain

$$y(t) = -\int_{T_i}^{t} e^{(t-\tau)\mathcal{L}}\mathcal{B}\mathcal{B}^T\lambda(\tau)\,d\tau = -\int_0^{t-T_i} e^{(t-T_i-\tau)\mathcal{L}}\mathcal{B}\mathcal{B}^T\lambda(\tau+T_i)\,d\tau$$

which implies

$$-y(T_{i+1}) = \left(\int_0^{\Delta T} e^{(\Delta T-\tau)\mathcal{L}}\mathcal{B}\mathcal{B}^T e^{(\Delta T-\tau)\mathcal{L}^T}\,d\tau\right)\Lambda_{i+1} =: \mathcal{R}_0\Lambda_{i+1}. \tag{3.12}$$

Note that $\mathcal{R}_0$ is symmetric positive semi-definite (and possibly positive definite), just like $\mathcal{R}_{\delta t}$.

In order to solve (3.9) numerically, we consider another time step Δt with $\Delta T = N\Delta t$ such that $\delta t \leq \Delta t \leq \Delta T$ as shown on Figure 1. Applying the ParaOpt (see [14]) to (3.9) gives the following iterative linear solver

$$\mathcal{M}_{\Delta t}\left(X^{k+1} - X^k\right) = \mathbf{f} - \mathcal{M}_{\delta t}X^k, \tag{3.13}$$

or equivalently

$$X^{k+1} = \mathcal{M}_{\Delta t}^{-1}\left(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\right)X^k + \mathcal{M}_{\Delta t}^{-1}\mathbf{f}. \tag{3.14}$$

The matrix $\mathcal{M}_{\Delta t}$ is defined the same way as $\mathcal{M}_{\delta t}$, except that its constitutive blocks $\mathcal{P}_{\Delta t}$ and $\mathcal{R}_{\Delta t}$ are obtained via a discretization with time step size Δt instead of δt. If the iterative method (3.14) converges to a vector X, then we have

$$X = \left(\mathcal{I} - \mathcal{M}_{\Delta t}^{-1}\mathcal{M}_{\delta t}\right)X + \mathcal{M}_{\Delta t}^{-1}\mathbf{f} \Longleftrightarrow \mathcal{M}_{\Delta t}^{-1}\mathcal{M}_{\delta t}X = \mathcal{M}_{\Delta t}^{-1}\mathbf{f},$$

meaning that the matrix $\mathcal{M}_{\Delta t}$ is a preconditioner for solving the linear system (3.9). Consequently, when Δt equals δt, we have $\mathcal{M}_{\Delta t} = \mathcal{M}_{\delta t}$, and the iteration (3.14) becomes a direct solver for the linear system (3.9). In the more general case where $\delta t < \Delta t$, a convergence analysis is presented in section 4.

The computational cost of solving (3.14) at each iteration is, of course, affected by the numerical method considered, in the expected sense, i.e., for a given δt, the numerical solution of the evolution equations involved in the problem will be, for example, twice as costly for an second order Runge–Kutta method as for an Euler method. A related question is whether one can exploit the structure of specific discretizations to speed up the computation of the matrix-vector product with $\mathcal{M}_{\Delta t}$ or the construction of the preconditioner, using for example diagonalization or spectral deferred correction techniques [13, 6]. However, this question goes beyond the scope of our article, since the answer will depend specifically on the discretization scheme in consideration; a detailed study will be the subject of future work.

4. Convergence factor analysis. In this section, we analyze of the convergence factor of (3.14) by estimating the operator norm of the iteration matrix $\mathcal{M}_{\Delta t}^{-1}(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})$. To do so, we must first define the vector space on which this matrix acts, as well as the vector norm on this space. Let us introduce the discrete vector spaces $\mathbb{E}_0$ and $\mathbb{E}_1$, defined by

$$\mathbb{E}_i := \{Y = (Y_\ell)_{\ell=i,\dots,L} : Y_\ell \in \mathbb{R}^r, \ \ \|Y\|_{\Delta T}^2 = \Delta T \sum_{\ell=i}^{L} \|Y_\ell\|^2\} \qquad \text{for } i = 0 \text{ or } 1.$$

We consider each element of $\mathbb{E}_i$ to be a block column vector and write $\mathbb{E} := \mathbb{E}_0 \times \mathbb{E}_1$. Given the structure of the matrices $\mathcal{M}_{\delta t}$ and $\mathcal{M}_{\Delta t}$, one readily sees that the first and the last block rows of the matrix $\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}$ vanish. In other words, if $X = \begin{pmatrix} Y \\ \Lambda \end{pmatrix} \in \mathbb{E}$ lies in the range of $\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}$, then its first and last block rows must vanish, meaning that $Y_0 = \Lambda_L = 0$. Thus, if we let Π be the projector on $\mathbb{E}$ defined by $\Pi \begin{pmatrix} Y \\ \Lambda \end{pmatrix} := (0, Y_1^T, \dots, Y_L^T, \Lambda_1^T, \dots, \Lambda_{L-1}^T, 0)^T$, we obtain

$$\mathcal{M}_{\Delta t}^{-1}(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}) = \mathcal{M}_{\Delta t}^{-1}\Pi(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}). \tag{4.1}$$

Let us endow the vector space $\mathbb{E}$ with the following weighted L^2-norm $\|.\|_*$

$$\|X\|_*^2 = \|Y\|_{\Delta T}^2 + \alpha^{-2} \|\Lambda\|_{\Delta T}^2 = \Delta T \left(\sum_{\ell=0}^{L} \|Y_\ell\|^2 + \alpha^{-2} \sum_{\ell=1}^{L} \|\Lambda_\ell\|^2 \right), \tag{4.2}$$

where the purpose of the α-dependent weighting will be apparent later. The corresponding matrix norm is then given by

$$\|\mathcal{M}_{\Delta t}\|_* = \inf \{\sigma \in \mathbb{R} : \|\mathcal{M}_{\Delta t} X\|_* \le \sigma \|X\|_* \ \forall X \in \mathbb{E}\}.$$

Our convergence analysis will require the following two ingredients, whose derivation will be the content of the next two subsections.

1. *Truncation error estimates:* We use the order conditions in Table 1 to obtain bounds on $\|\mathcal{P}_{\delta t} - \mathcal{P}_0\|$ and $\|\mathcal{R}_{\delta t} - \mathcal{R}_0\|$ as a function of δt. We can then use the triangle inequality to estimate $\|\Delta\mathcal{P}\|$ and $\|\Delta\mathcal{R}\|$, where $\Delta\mathcal{P} := \mathcal{P}_{\Delta t} - \mathcal{P}_{\delta t}$ and $\Delta\mathcal{R} := \mathcal{R}_{\Delta t} - \mathcal{R}_{\delta t}$. This in turn can be used to estimate $\|\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\|_*$, since the only blocks that appear in this difference matrix are scalar multiples of $\Delta\mathcal{P}$ and $\Delta\mathcal{R}$.
2. *Stability estimate:* we first estimate $\left\|\mathcal{M}_{\Delta t}^{-1}\Pi\right\|_*$. Then from (4.1), we deduce that

$$\|\mathcal{M}_{\Delta t}^{-1}(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})\|_* \le \left\|\mathcal{M}_{\Delta t}^{-1}\Pi\right\|_* \|\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\|_*,$$

which, when combined with step 1, immediately yields a convergence estimate for ParaOpt.

4.1. Truncation error estimates. Our first task is to estimate the local truncation errors $\mathcal{P}_{\delta t} - \mathcal{P}_0$ and $\mathcal{R}_{\delta t} - \mathcal{R}_0$ as a function of δt. Our argument is essentially based on [17], but with two major differences:

- In [17], Hager shows that the local truncation error is bounded by $c\delta t^p$ with a constant c that depends on the exact solution and its derivatives, without explicitly specifying these dependencies. Here, we show explicitly how the truncation error depends on the initial data Y_i or final data Λ_{i+1}, which enables us to bound the operator norms of $\mathcal{P}_{\delta t} - \mathcal{P}_0$ and $\mathcal{R}_{\delta t} - \mathcal{R}_0$.
- In [17], Hager considers a general problem of convex minimization under nonlinear ODE constraints, so additional assumptions are needed to ensure well-posedness and regularity of the solution. Here, we exploit the linear-quadratic structure of the problem, so these assumptions are automatically satisfied. Moreover, we have explicit formulas for the exact solution, which further simplify our proof.

THEOREM 4.1. *Let $T > 0$. For any $\Delta T \in (0,T]$, let $\mathcal{P}_{\delta t}$, $\mathcal{R}_{\delta t}$, $\mathcal{P}_0$ and $\mathcal{R}_0$ be defined as in (3.7), (3.8), (3.10) and (3.12) respectively. If the Runge-Kutta method (RK) satisfies both the IVP and optimal control order conditions in Table 1 up to order p, then there exist constants $c_{\mathcal{P}} > 0$ and $c_{\mathcal{R}} > 0$, which are independent of δt and ΔT (but can depend on T, $\mathcal{L}$, etc.), such that*

$$\|\mathcal{P}_{\delta t} - \mathcal{P}_0\| \le c_{\mathcal{P}}\delta t^p \quad \textit{and} \quad \|\mathcal{R}_{\delta t} - \mathcal{R}_0\| \le c_{\mathcal{R}}\delta t^p, \tag{4.3}$$

where $\|\cdot\|$ denotes the spectral norm, i.e., the operator norm associated with the Euclidean norm in $\mathbb{R}^r$.

Proof. Since the ODE system is time-invariant, we can assume without loss of generality that we are working with the time sub-interval $[0,\Delta T]$, where $\Delta T = M\delta t$. For given $Y_0, \Lambda_{\Delta T} \in \mathbb{R}^r$, we define

$$\begin{aligned} y^*(t) &= e^{t\mathcal{L}}Y_0 - \left(\int_0^t e^{(t-\tau)\mathcal{L}}\mathcal{B}\mathcal{B}^T e^{(\Delta T-\tau)\mathcal{L}^T}\,d\tau\right)\Lambda_{\Delta T},\\ \lambda^*(t) &= e^{(\Delta T - t)\mathcal{L}^T}\Lambda_{\Delta T}, \end{aligned}$$

so that

$$y^*(\Delta T) = \begin{cases} \mathcal{P}_0 Y_0 & \text{if } \Lambda_{\Delta T} = 0,\\ -\mathcal{R}_0\Lambda_{\Delta T} & \text{if } Y_0 = 0. \end{cases}$$

Then $z(t) := \begin{pmatrix} y^*(t)\\ \lambda^*(t)\end{pmatrix}$ is the solution of the initial value problem

$$\dot z(t) = \begin{bmatrix} \mathcal{L} & -\mathcal{B}\mathcal{B}^T\\ 0 & -\mathcal{L}^T \end{bmatrix} z(t), \qquad z(0) = \begin{pmatrix} Y_0\\ e^{\Delta T\mathcal{L}^T}\Lambda_{\Delta T}\end{pmatrix}. \tag{4.4}$$

Let $t_k = k\delta t$. By [17, §3], the Runge-Kutta method (2.11), (2.12) can be written in the form

$$z_{k+1} = z_k + \delta t\sum_{i=1}^s b_i g_i(t_k, z_k, \delta t),$$

where g_i is the ith stage of the Runge-Kutta method; note that each g_i is linear in z_k, since the problem (4.4) is linear. We now define

$$\zeta_k(\delta t) := z(t_k) + \delta t\sum_{i=1}^s b_i g_i(t_k, z(t_k), \delta t),$$

so that the local truncation error can be written as

$$\tau_k(\delta t) := \frac{1}{\delta t}(z(t_{k+1}) - \zeta_k(\delta t)).$$

A standard argument (cf. [19, §II.3]) then gives the error estimate

$$\begin{aligned} \|z(t_k) - z_k\| &\le e^{K\Delta T}\left(\|z(0) - z_0\| + \Delta T \max_{0\le \ell\le k-1}\|\tau_\ell(\delta t)\|\right)\\ &\le e^{KT}\left(\|z(0) - z_0\| + T\max_{0\le\ell\le k-1}\|\tau_\ell(\delta t)\|\right), \end{aligned} \tag{4.5}$$

where $K > 0$ is a Lipschitz constant for the g_i with respect to the second argument, which is independent of $z(0)$ in our case because the problem (4.4) is linear. It remains to show that

$$\|\tau_k(\delta t)\| \le \delta t^p(C_1\|Y_0\| + C_2\|\Lambda_{\Delta T}\|) \qquad \text{for all } 0 \le k \le \Delta T/\delta t = N_0, \tag{4.6}$$

from which we immediately deduce the bounds (4.3) by letting $\Lambda_{\Delta T} = 0$ and $Y_0 = 0$ respectively.

To prove (4.6), we compare the Taylor expansions

$$\begin{aligned} z(t_{k+1}) = z(t_k + \delta t) &= z(t_k) + \delta t z'(t_k) + \cdots + \frac{\delta t^p}{p!}z^{(p)}(t_k) + \cdots,\\ \zeta_k(\delta t) &= \zeta_k(0) + \delta t\zeta_k'(0) + \cdots + \frac{\delta t^p}{p!}\zeta_k^{(p)}(0) + \cdots. \end{aligned}$$

Since $\tau_k = O(\delta t^p)$ by the order conditions, we deduce that $z^{(q)}(t_k) = \zeta^{(q)}(0)$ for $q = 0, 1, \ldots, p$. By expanding $\delta t \tau_k(\delta t) = z(t_{k+1}) - \zeta_k(\delta t)$ up to order $p-1$ and writing the pth order remainder term in integral form, we deduce that

$$\begin{aligned} \|\tau_k(\delta t)\| &= \frac{\|z(t_k+\delta t) - \zeta_k(\delta t)\|}{\delta t} = \frac{1}{\delta t}\left\| \int_0^{\delta t} \frac{z^{(p)}(t_k+\eta) - \zeta_k^{(p)}(\eta)}{(p-1)!}(\delta t - \eta)^{p-1}\, d\eta \right\| \\ &\leq \frac{\delta t^{p-2}}{(p-1)!}\int_0^{\delta t} \|z^{(p)}(t_k+\eta) - \zeta^{(p)}(\eta)\|\, d\eta \\ &\leq \frac{\delta t^{p-2}}{(p-1)!}\left(\int_0^{\delta t} \|z^{(p)}(t_k+\eta) - z^{(p)}(t_k)\|\, d\eta + \int_0^{\delta t} \|\zeta^{(p)}(\eta) - \zeta^{(p)}(0)\|\, d\eta \right). \end{aligned} \tag{4.7}$$

Let $\mathcal{A}$ be the matrix that multiplies $z(t)$ in (4.4). Then for any integer $q \geq 0$, we can write

$$z^{(q)}(t_k+\eta) = \mathcal{A}^q z(t_k+\eta) = \mathcal{A}^q e^{(t_k+\eta)\mathcal{A}} z(0).$$

By letting $q = p+1$, we see that $z^{(p)}(t_k+\eta)$ is differentiable, and therefore Lipschitz with respect to η, so there exists $K_1 > 0$ such that $\|z^{(p)}(t_k+\eta) - z^{(p)}(t_k)\| \leq K_1\eta\|z(0)\|$ for all $0 \leq \eta \leq \delta t$. Similarly, when δt is small enough, $\zeta_k^{(p)}(\eta)$ is a rational function of η, since $\zeta_k(\delta t)$ is the solution of a linear system of the form (2.6b), with δt entering linearly into the coefficients of the linear system. Moreover, $\zeta_k^{(p)}(\eta)$ is defined everywhere on $[0, \delta t]$, and is thus Lipschitz there; it is also linear in $z(t_k)$, so there exists $K_2 > 0$ such that $\|\zeta^{(p)}(\eta) - \zeta^{(p)}(0)\| \leq K_2\eta\|z(t_k)\| = K_2\eta\|e^{t_k\mathcal{A}}z(0)\|$ for all $0 \leq \eta \leq \delta t$. Substituting these inequalities into (4.7) leads to the required inequality

$$\|\tau_k(\delta t)\| \leq C\delta t^p\|z(0)\| \leq \delta t^p(C_1\|Y_0\| + C_2\|\Lambda_{\Delta T}\|),$$

where C_1 and C_2 can be made independent of ΔT by taking the maximum over all $\Delta T \in (0, T]$. Finally, to bound $\|z(0) - z_0\|$, we note that since the first component of z_0 is exact, we have

$$\|z(0) - z_0\| = \|e^{\Delta T \mathcal{L}^T}\Lambda_{\Delta T} - \lambda_0\|,$$

where λ_0 is obtained by a pth order Runge-Kutta method applied to the backward linear ODE $\dot{\lambda} = -\mathcal{L}^T\lambda$, $\lambda(\Delta T) = \Lambda_{\Delta T}$. It follows that λ_0 is linear in $\Lambda_{\Delta T}$, and since we use exact final conditions to compute λ_0, a similar argument to the above shows that

$$\|e^{\Delta T\mathcal{L}^T}\Lambda_{\Delta T} - \lambda_0\| \leq C_3\delta t^p\|\Lambda_{\Delta T}\|$$

for some $C_3 > 0$, where we can again make C_3 independent of ΔT (but still dependent on T). Substituting this and (4.7) into (4.5) with $k = M$ yields

$$\begin{aligned} \left\| \begin{pmatrix} y^*(\Delta T) - y_M \\ \lambda^*(\Delta T) - \lambda_M \end{pmatrix} \right\| = \|z(t_M) - z_M\| &\leq \delta t^p\big(e^{K\Delta T}C_1\Delta T\|Y_0\| + e^{K\Delta T}(C_2\Delta T + C_3)\|\Lambda_{\Delta T}\|\big) \\ &\leq \delta t^p\big(\underbrace{e^{KT}C_1T}_{c_{\mathcal{P}}}\|Y_0\| + \underbrace{e^{KT}(C_2T + C_3)}_{c_{\mathcal{R}}}\|\Lambda_{\Delta T}\|\big), \end{aligned}$$

with $c_{\mathcal{P}}$ and $c_{\mathcal{R}}$ independent of ΔT. We finally obtain the bounds (4.3) by letting either $Y_0 = 0$ or $\Lambda_{\Delta T} = 0$. □

Next, we define the operators

$$\Delta\mathcal{P} := \mathcal{P}_{\Delta t} - \mathcal{P}_{\delta t}, \quad \text{and} \quad \Delta\mathcal{R} := \mathcal{R}_{\Delta t} - \mathcal{R}_{\delta t},$$

which appear as subblocks of the matrix $\mathcal{M}_{\delta t} - \mathcal{M}_{\Delta t}$. We recall that δt and Δt are the fine and coarse time steps and that $\Delta t = N_0\delta t$ for some integer $N_0 \geq 2$, so that the fine grid can be viewed as a refinement of the coarse grid. Then by the triangle inequality, we have

$$\|\Delta\mathcal{P}\| \leq \|\mathcal{P}_{\Delta t} - \mathcal{P}_0\| + \|\mathcal{P}_{\delta t} - \mathcal{P}_0\| \leq c_{\mathcal{P}}(\Delta t^p + \delta t^p) \leq 2c_{\mathcal{P}}\Delta t^p, \tag{4.8}$$

and similarly for $\|\Delta\mathcal{R}\|$. The following theorem allows us to bound the norm of $\mathcal{M}_{\delta t} - \mathcal{M}_{\Delta t}$.

Theorem 4.2. *Let ΔT be fixed. Let $\mathcal{M}_{\Delta t}$ and $\mathcal{M}_{\delta t}$ be the ParaOpt matrices* (3.9) *obtained from the system* (3.4)–(3.8) *with time steps Δt and δt respectively. Then there exists $c_{\mathcal{M}} > 0$ independent of Δt and N_0 such that*

$$\|\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\|_* \leq c_{\mathcal{M}}\Delta t^p.$$

Proof. Let $X, E \in \mathbb{E}$ with

$$X = \begin{pmatrix} Y \\ \Lambda \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} F \\ G \end{pmatrix},$$

for some $F \in \mathbb{R}^{r(L+1)}$ and $G \in \mathbb{R}^{rL}$ such that $(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})\, X = E$. We will bound $\|E\|_*$ in terms of $\|X\|_*$. Writing the block rows of $(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})X = E$ explicitly leads to the componentwise system

$$F_0 = 0,$$

$$-\Delta\mathcal{P} Y_{\ell-1} + \frac{1}{\alpha}\Delta\mathcal{R}\Lambda_\ell = F_\ell, \qquad \ell = 1, \dots, L, \tag{4.9}$$

$$-\Delta\mathcal{P}^T \Lambda_{\ell+1} = G_\ell, \qquad \ell = 1, \dots, L-1, \tag{4.10}$$

$$G_L = 0.$$

Taking norms in (4.10) and using the definition of matrix norms immediately yields

$$\|G_\ell\|^2 \leq \|\Delta\mathcal{P}\|^2 \|\Lambda_{\ell+1}\|^2, \qquad \ell = 1, \dots, L-1. \tag{4.11}$$

Doing the same for (4.9) and applying the triangle inequality, we get, for $\ell = 1, \dots, L$,

$$\|F_\ell\| \leq \|\Delta\mathcal{P}\| \, \|Y_{\ell-1}\| + \alpha^{-1} \|\Delta\mathcal{R}\| \, \|\Lambda_\ell\|.$$

Taking squares on both sides then yields

$$\|F_\ell\|^2 \leq \|\Delta\mathcal{P}\|^2 \|Y_{\ell-1}\|^2 + \alpha^{-2} \|\Delta\mathcal{R}\|^2 \|\Lambda_\ell\|^2 + 2\alpha^{-1} \|\Delta\mathcal{P}\| \, \|\Delta\mathcal{R}\| \, \|Y_{\ell-1}\| \, \|\Lambda_\ell\|.$$

We now bound the last term on the right hand side using the arithmetic-geometric mean inequality

$$2\alpha^{-1} \|\Delta\mathcal{P}\| \, \|\Delta\mathcal{R}\| \, \|Y_{\ell-1}\| \, \|\Lambda_\ell\| \leq \|\Delta\mathcal{R}\|^2 \|Y_{\ell-1}\|^2 + \alpha^{-2} \|\Delta\mathcal{P}\|^2 \|\Lambda_\ell\|^2$$

in order to obtain

$$\|F_\ell\|^2 \leq \left(\|\Delta\mathcal{P}\|^2 + \|\Delta\mathcal{R}\|^2\right)\left(\|Y_{\ell-1}\|^2 + \alpha^{-2} \|\Lambda_\ell\|^2\right). \tag{4.12}$$

We now have all the ingredients for bounding $\|E\|_*^2$, which by definition (4.2) is given by

$$\|E\|_*^2 = \|F\|_{\Delta T}^2 + \alpha^{-2}\|G\|_{\Delta T}^2 = \Delta T \left(\sum_{\ell=0}^{L} \|F_\ell\|^2 + \alpha^{-2} \sum_{\ell=1}^{L} \|G_\ell\|^2\right).$$

The terms in F can be bounded using (4.12):

$$\sum_{\ell=1}^{L} \|F_\ell\|^2 \leq \left(\|\Delta\mathcal{P}\|^2 + \|\Delta\mathcal{R}\|^2\right) \sum_{\ell=1}^{L} \left(\|Y_{\ell-1}\|^2 + \alpha^{-2} \|\Lambda_\ell\|^2\right),$$

which, together with $F_0 = 0$, implies that

$$\|F\|_{\Delta T}^2 \leq \Delta T \left(\|\Delta\mathcal{P}\|^2 + \|\Delta\mathcal{R}\|^2\right)\left(\sum_{\ell=0}^{L} \|Y_\ell\|^2 + \alpha^{-2} \sum_{\ell=1}^{L} \|\Lambda_\ell\|^2\right) = \left(\|\Delta\mathcal{P}\|^2 + \|\Delta\mathcal{R}\|^2\right) \|X\|_*^2. \tag{4.13}$$

Similarly, we can bound the terms in G using (4.11) and the fact that $G_L = 0$:

$$\|G\|_{\Delta T}^2 \leq \|\Delta\mathcal{P}\|^2 \, \Delta T \sum_{\ell=1}^{L-1} \|\Lambda_{\ell+1}\|^2 \leq \|\Delta\mathcal{P}\|^2 \, \Delta T \sum_{\ell=1}^{L} \|\Lambda_\ell\|^2 = \|\Delta\mathcal{P}\|^2 \, \|\Lambda\|_{\Delta T}^2. \tag{4.14}$$

Combining (4.13) and (4.14) therefore leads to

$$\begin{aligned} \|E\|_*^2 &= \|F\|_{\Delta T}^2 + \alpha^{-2} \|G\|_{\Delta T}^2 \\ &\leq \left(\|\Delta\mathcal{P}\|^2 + \|\Delta\mathcal{R}\|^2\right) \|X\|_*^2 + \alpha^{-2} \|\Delta\mathcal{P}\|^2 \, \|\Lambda\|_{\Delta T}^2 \\ &\leq \left(2\, \|\Delta\mathcal{P}\|^2 + \|\Delta\mathcal{R}\|^2\right) \|X\|_*^2. \end{aligned}$$

But thanks to Theorem 4.1 and the triangle inequality, we know that (cf. (4.8))

$$\|\Delta\mathcal{P}\| \leq 2c_{\mathcal{P}} \Delta t^p, \qquad \|\Delta\mathcal{R}\| \leq 2c_{\mathcal{R}} \Delta t^p.$$

Hence, by defining $c_{\mathcal{M}} = \sqrt{8c_{\mathcal{P}}^2 + 4c_{\mathcal{R}}^2}$, we can write

$$\|E\|_* = \|(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})X\|_* \leq c_{\mathcal{M}} \Delta t^p \|X\|_*$$ □

which finally leads us to conclude that $\|(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})\|_* \leq c_{\mathcal{M}} \Delta t^p$, as required.

4.2. Stability estimate. We keep the notation of the previous section and now turn our attention to bounding $\|\mathcal{M}_{\Delta t}^{-1}\Pi\|_*$. Let $X = \mathcal{M}_{\Delta t}^{-1}\Pi E$, where $X, E \in \mathbb{E}$ satisfy

$$X = \begin{pmatrix} Y \\ \Lambda \end{pmatrix} \quad \text{and} \quad \Pi E = \begin{pmatrix} F \\ G \end{pmatrix} \quad \text{with } F_0 = G_L = 0.$$

In other words, we have $\mathcal{M}_{\Delta t}X = \Pi E$, and we wish to bound $\|X\|_*$ in terms of $\|\Pi E\|_*$, which is always less than or equal to $\|E\|_*$. Using the definition of $\mathcal{M}_{\Delta t}$ in (3.9), we can write down the block rows of $\mathcal{M}_{\Delta t}X = \Pi E$ explicitly to obtain the componentwise system

$$Y_0 = 0, \tag{4.15}$$

$$-\mathcal{P}_{\Delta t}Y_{\ell-1} + Y_\ell + \alpha^{-1}\mathcal{R}_{\Delta t}\Lambda_\ell = F_\ell, \qquad \ell = 1, \dots, L, \tag{4.16}$$

$$\Lambda_{\ell-1} - \mathcal{P}_{\Delta t}^T\Lambda_\ell = G_{\ell-1}, \qquad \ell = 2, \dots, L, \tag{4.17}$$

$$\Lambda_L - Y_L = 0. \tag{4.18}$$

We first show that one can transform (4.15)–(4.18) into an equivalent system involving Y_L as the only unknown.

Lemma 4.3. *Let $(Y_\ell)_{\ell=0}^L$ and $(\Lambda_\ell)_{\ell=1}^L$ satisfy (4.15)–(4.18). Then for $\ell = 1, \dots, L$, we have*

$$\Lambda_\ell = \left(\mathcal{P}_{\Delta t}^T\right)^{L-\ell} Y_L + \sum_{j=\ell}^{L-1} \left(\mathcal{P}_{\Delta t}^T\right)^{j-\ell} G_j, \tag{4.19}$$

$$Y_\ell = \sum_{j=1}^{\ell} \mathcal{P}_{\Delta t}^{\ell-j} \left[F_j - \alpha^{-1}\mathcal{R}_{\Delta t} \left(\left(\mathcal{P}_{\Delta t}^T\right)^{L-j} Y_L + \sum_{k=j}^{L-1} \left(\mathcal{P}_{\Delta t}^T\right)^{k-j} G_k \right) \right], \tag{4.20}$$

where Y_L satisfies the equation $(\mathcal{I} + \alpha^{-1}\mathcal{U})Y_L = S$ with

$$\mathcal{U} = \sum_{\ell=1}^{L} \mathcal{P}_{\Delta t}^{L-\ell}\mathcal{R}_{\Delta t}(\mathcal{P}_{\Delta t}^T)^{L-\ell}, \qquad S = \sum_{\ell=1}^{L} \mathcal{P}_{\Delta t}^{L-\ell} \left(F_\ell - \alpha^{-1} \sum_{j=\ell}^{L-1} \mathcal{R}_{\Delta t} \left(\mathcal{P}_{\Delta t}^T\right)^{j-\ell} G_j \right). \tag{4.21}$$

Moreover, we have $\|Y_L\| \le \|S\|$.

Proof. The recurrence (4.17) can be unrolled to obtain

$$\begin{aligned}
\Lambda_{L-1} &= G_{L-1} + \mathcal{P}_{\Delta t}^T\Lambda_L \\
\Lambda_{L-2} &= G_{L-2} + \mathcal{P}_{\Delta t}^T(G_{L-1} + \mathcal{P}_{\Delta t}^T\Lambda_L) \\
&\ \vdots \\
\Lambda_\ell &= (\mathcal{P}_{\Delta t}^T)^{L-\ell}\Lambda_L + \sum_{j=\ell}^{L-1}(\mathcal{P}_{\Delta t}^T)^{j-\ell}G_j, \qquad \ell = 1, \dots, L-1.
\end{aligned}$$

Replacing Λ_L in the above by Y_L, which is equal to Λ_L by (4.18), yields the expression (4.19), which is also valid for $\ell = L$ because the sum would be empty in this case. Next, we solve the forward recurrence (4.16) starting from $Y_0 = 0$ to obtain

$$Y_\ell = \sum_{j=1}^{\ell} \mathcal{P}_{\Delta t}^{\ell-j}(F_j - \alpha^{-1}\mathcal{R}_{\Delta t}\Lambda_j), \qquad \ell = 1, \dots, L. \tag{4.22}$$

Substituting the expression of Λ_j from (4.19) into the above leads to (4.20). In particular, when $\ell = L$, we get

$$Y_L = \sum_{j=1}^{L} \mathcal{P}_{\Delta t}^{L-j} \left[F_j - \alpha^{-1}\mathcal{R}_{\Delta t} \left(\left(\mathcal{P}_{\Delta t}^T\right)^{L-j} Y_L + \sum_{k=j}^{L-1} \left(\mathcal{P}_{\Delta t}^T\right)^{k-j} G_k \right) \right].$$

Moving all terms containing Y_L to the left-hand side leads to the system $(I + \alpha^{-1}\mathcal{U})Y_L = S$, with $\mathcal{U}$ and S as defined in (4.21). Finally, since $\mathcal{R}_{\Delta t}$ is symmetric positive semi-definite (see the sentence immediately after (3.12)), so is $\mathcal{U}$; we thus have

$$\|Y_L\|^2 \le \langle Y_L, Y_L\rangle + \alpha^{-1}\langle Y_L, \mathcal{U}Y_L\rangle = \langle Y_L, S\rangle \le \|Y_L\|\,\|S\|.$$

Dividing both sides by $\|Y_L\|$ leads to $\|Y_L\| \le \|S\|$, as required. □

We will also need the following lemma.

Lemma 4.4. *Let $T>0$ be fixed. Then for all $0\le \Delta T\le T$ and $\Delta t>0$ small enough, there exist positive constants C_1, C_2 independent of ΔT (but which can depend on T) such that*

$$\|\mathcal{P}_{\Delta t}\|\le 1+C_1\Delta T \qquad \text{and} \qquad \|\mathcal{R}_{\Delta t}\|\le C_2\Delta T. \tag{4.23}$$

Proof. We start by proving that there exists $C_0>0$ such that $\|\mathcal{P}_0\|\le 1+C_0\Delta T$. Indeed, by Definition (3.10), we have $\mathcal{P}_0 y_0 = y(\Delta T)$, where $y(t)$ is the solution of $\dot{y}=\mathcal{L}y$, $y(0)=y_0$. We therefore have

$$\mathcal{P}_0 y_0 = y_0 + \int_0^{\Delta T}\dot{y}(t)\,dt = y_0+\int_0^{\Delta T}\mathcal{L}y(t)\,dt.$$

But $y(t)=e^{t\mathcal{L}}y_0$, so $\|\mathcal{L}y(t)\|$ can be bounded uniformly by $C_0\|y_0\|$ with $C_0=\max_{t\in[0,T]}\|\mathcal{L}e^{t\mathcal{L}}\|$. We therefore have

$$\|\mathcal{P}_0y_0\|\le\|y_0\|+C_0\Delta T\|y_0\|,$$

which implies $\|\mathcal{P}_0\|\le 1+C_0\Delta T$. In fact, by replacing ΔT with a general t in $\mathcal{P}_0=e^{\Delta T\mathcal{L}}$, we have actually shown that $\|e^{t\mathcal{L}}\|\le 1+C_0t$ for all $0\le t\le T$ (where C_0 depends on T), a fact that will be used to bound $\|\mathcal{R}_{\Delta t}\|$ later.

We now recall the result of Theorem 4.1, which asserts that $\|\mathcal{P}_{\Delta t}-\mathcal{P}_0\|\le c_{\mathcal{P}}\Delta t^p$ for some constant $c_{\mathcal{P}}$. By choosing Δt small enough such that $c_{\mathcal{P}}\Delta t^p\le C_0\Delta T$, we obtain

$$\|\mathcal{P}_{\Delta t}\|\le\|\mathcal{P}_0\|+\|\mathcal{P}_{\Delta t}-\mathcal{P}_0\|\le 1+2C_0\Delta T,$$

which is the first bound in (4.23) with $C_1=2C_0$. For the bound on $\|\mathcal{R}_{\Delta t}\|$, we again go through $\|\mathcal{R}_0\|$, which by Definition (3.12) is given by

$$\mathcal{R}_0=\int_0^{\Delta T}e^{(\Delta T-\tau)\mathcal{L}}\mathcal{B}\mathcal{B}^Te^{(\Delta T-\tau)\mathcal{L}^T}\,d\tau=\int_0^{\Delta T}e^{\tau\mathcal{L}}\mathcal{B}\mathcal{B}^Te^{\tau\mathcal{L}^T}\,d\tau.$$

Having shown earlier that $\|e^{t\mathcal{L}}\|\le 1+C_0t$ for all $0\le t\le T$, we deduce

$$\begin{aligned}\|\mathcal{R}_0\|&\le\|\mathcal{B}\mathcal{B}^T\|\int_0^{\Delta T}(1+C_0t)^2\,d\tau\\&=\|\mathcal{B}\mathcal{B}^T\|\cdot\frac{(1+C_0\Delta T)^3-1}{3C_0}=\Delta T\|\mathcal{B}\mathcal{B}^T\|\left(1+C_0\Delta T+\frac{C_0^2\Delta T^2}{3}\right).\end{aligned}$$

We therefore have $\|\mathcal{R}_0\|\le K\Delta T$ with $K=\|\mathcal{B}\mathcal{B}^T\|\left(1+C_0T+\frac{1}{3}C_0^2T^2\right)$. Finally, we choose Δt small enough so that $c_{\mathcal{R}}\Delta t^p\le K\Delta T$, where $c_{\mathcal{R}}$ is defined in Theorem 4.1. This theorem then allows us to conclude that

$$\|\mathcal{R}_{\Delta t}\|\le\|\mathcal{R}_0\|+\|\mathcal{R}_{\Delta t}-\mathcal{R}_0\|\le 2K\Delta T,$$

so the second bound in (4.23) holds with $C_2=2K$. □

We are now ready to state and prove the following theorem.

Theorem 4.5. *Let $\Delta T=T/L$ and $\alpha>0$ be given. Then, there exists $c_{\mathcal{M}^{-1}}>0$ independent of Δt such that*

$$\left\|\mathcal{M}_{\Delta t}^{-1}\Pi\right\|_*\le\frac{c_{\mathcal{M}^{-1}}}{\Delta T}(1+\alpha^{-1}).$$

Proof. The proof consists of three steps.

1. *Estimation of $\|Y_L\|$ in Lemma 4.3.* By Lemma 4.3, it suffices to estimate $\|S\|$ for the vector S defined in (4.21). We first observe that $S=S_1+\alpha^{-1}S_2$, where

$$S_1:=\sum_{\ell=1}^{L}\mathcal{P}_{\Delta t}^{L-\ell}F_\ell,\quad\text{and}\qquad S_2:=-\sum_{\ell=1}^{L}\sum_{j=\ell}^{L-1}\mathcal{P}_{\Delta t}^{L-\ell}\mathcal{R}_{\Delta t}\left(\mathcal{P}_{\Delta t}^T\right)^{j-\ell}G_j.$$

To bound $\|S_1\|$, we use the first inequality in (4.23) and then the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}\|S_1\|&\le\sum_{\ell=1}^{L}\|\mathcal{P}_{\Delta t}\|^{L-\ell}\|F_\ell\|\\&\le\left(\sum_{\ell=1}^{L}(1+C_1\Delta T)^{2(L-\ell)}\right)^{1/2}\left(\sum_{\ell=1}^{L}\|F_\ell\|^2\right)^{1/2}=\left(\frac{(1+C_1\Delta T)^{2L}-1}{(1+C_1\Delta T)^2-1}\right)^{1/2}\frac{\|F\|_{\Delta T}}{\Delta T^{1/2}}.\end{aligned}$$

Applying the inequality $1 + C_1\Delta T \leq e^{C_1 \Delta T}$ to the numerator, we deduce that

$$\|S_1\| \leq \left(\frac{e^{2C_1 L\Delta T} - 1}{2C_1\Delta T^2 + C_1^2 \Delta T^3}\right)^{1/2} \|F\|_{\Delta T} \leq \frac{1}{\Delta T}\left(\frac{e^{2C_1 T} - 1}{2C_1}\right)^{1/2} \|F\|_{\Delta T}.$$

We now bound S_2 again using (4.23) and Cauchy-Schwarz:

$$\begin{aligned}
\|S_2\| &\leq \|\mathcal{R}_{\Delta t}\| \sum_{\ell=1}^{L}\sum_{j=\ell}^{L-1} \|\mathcal{P}_{\Delta t}\|^{L-\ell} \left\|\mathcal{P}_{\Delta t}^T\right\|^{j-\ell} \|G_j\| \\
&\leq C_2\Delta T \left(\sum_{\ell=1}^{L}\sum_{j=\ell}^{L-1} (1 + C_1\Delta T)^{2(L+j-2\ell)}\right)^{1/2} \left(\sum_{\ell=1}^{L}\sum_{j=\ell}^{L-1} \|G_j\|^2\right)^{1/2} \\
&\leq C_2\Delta T \left(\sum_{\ell=1}^{L}\sum_{j=1}^{L} (1 + C_1\Delta T)^{4(L-\ell)}\right)^{1/2} \left(\sum_{\ell=1}^{L}\sum_{j=1}^{L} \|G_j\|^2\right)^{1/2} \\
&= C_2\Delta T L \left(\sum_{\ell=1}^{L} (1 + C_1\Delta T)^{4(L-\ell)}\right)^{1/2} \left(\sum_{j=1}^{L} \|G_j\|^2\right)^{1/2} \\
&= C_2 T \left(\frac{(1 + C_1\Delta T)^{4L} - 1}{(1 + C_1\Delta T)^4 - 1}\right)^{1/2} \frac{\|G\|_{\Delta T}}{\Delta T^{1/2}} \\
&\leq \frac{C_2 T}{\Delta T}\left(\frac{e^{4C_1 T} - 1}{4C_1}\right)^{1/2} \|G\|_{\Delta T}.
\end{aligned}$$

Finally, combining the bounds for $\|S_1\|$ and $\|S_2\|$ yields

$$\|Y_L\| \leq \|S\| \leq \frac{c_S}{\Delta T}\left(\|F\|_{\Delta T} + \alpha^{-1}\|G\|_{\Delta T}\right), \tag{4.24}$$

where $c_S = \max\left\{\left(\frac{e^{2C_1 T} - 1}{2C_1}\right)^{\frac{1}{2}}, C_2 T\left(\frac{e^{4C_1 T} - 1}{4C_1}\right)^{\frac{1}{2}}\right\}$.

2. *Estimation of* $\|\Lambda_\ell\|$ *and* $\|Y_\ell\|$ *in Lemma 4.3.* From (4.19), we get

$$\begin{aligned}
\|\Lambda_\ell\| &\leq \left\|\mathcal{P}_{\Delta t}^T\right\|^{L-\ell} \|Y_L\| + \sum_{j=\ell}^{L-1} \left\|\mathcal{P}_{\Delta t}^T\right\|^{j-\ell} \|G_j\| \\
&\leq (1 + C_1\Delta T)^{L-\ell}\|Y_L\| + \sum_{j=\ell}^{L-1} (1 + C_1\Delta T)^{j-\ell}\|G_j\| \\
&\leq (1 + C_1\Delta T)^{L-\ell}\|Y_L\| + \sum_{j=1}^{L} (1 + C_1\Delta T)^{L-\ell}\|G_j\| \\
&\overset{(*)}{\leq} (1 + C_1\Delta T)^{L-\ell}\left(\|Y_L\| + L^{1/2}\Big(\sum_{j=1}^{L}\|G_j\|^2\Big)^{1/2}\right) \\
&= (1 + C_1\Delta T)^{L-\ell}\left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right),
\end{aligned} \tag{4.25}$$

where the inequality $(*)$ is due to Cauchy-Schwarz. Likewise, from (4.22) we get

$$\begin{aligned}
\|Y_\ell\| &\leq \sum_{j=1}^{\ell} \left\|\mathcal{P}_{\Delta t}^T\right\|^{\ell-j} \|F_j\| + \alpha^{-1}\|\mathcal{R}_{\Delta t}\| \sum_{j=1}^{\ell} \left\|\mathcal{P}_{\Delta t}^T\right\|^{\ell-j} \|\Lambda_j\| \\
&\leq \sum_{j=1}^{\ell} (1 + C_1\Delta T)^{\ell-j}\|F_j\| + \alpha^{-1}C_2\Delta T \sum_{j=1}^{\ell} (1 + C_1\Delta T)^{L+\ell-2j}\left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right) \\
&\leq \sum_{j=1}^{L} (1 + C_1\Delta T)^{L-j}\|F_j\| + \alpha^{-1}C_2\Delta T\left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right)\sum_{j=1}^{L} (1 + C_1\Delta T)^{2(L-j)}.
\end{aligned}$$

Just as in the calculation for bounding $\|S_1\|$, we have

$$\sum_{j=1}^{L}(1+C_1\Delta T)^{2(L-j)} \le \frac{e^{2C_1T}-1}{2C_1\Delta T} \le \frac{c_{\mathcal{S}}}{\Delta T},$$

from which we deduce that

$$\|Y_\ell\| \le \frac{c_{\mathcal{S}}^{1/2}}{\Delta T}\|F\|_{\Delta T} + \alpha^{-1}C_2c_{\mathcal{S}}\left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right).$$

Applying the inequality $(a+b)^2 \le 2(a^2+b^2)$ then gives

$$\|Y_\ell\|^2 \le \frac{2c_{\mathcal{S}}}{\Delta T^2}\|F\|_{\Delta T}^2 + 2\alpha^{-2}C_2^2c_{\mathcal{S}}^2\left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right)^2. \tag{4.26}$$

3. Estimation of $\|X\|_* = \left\|\mathcal{M}_{\Delta t}^{-1}\Pi E\right\|_*$. By the definition of $\|\cdot\|_*$, we have

$$\|X\|_*^2 = \|Y\|_{\Delta T}^2 + \alpha^{-2}\|\Lambda\|_{\Delta T}^2 = \Delta T\left(\sum_{\ell=0}^{L}\|Y_\ell\|^2 + \alpha^{-2}\sum_{\ell=1}^{L}\|\Lambda_\ell\|^2\right).$$

Substituting (4.25) and (4.26) into the above leads to

$$\begin{aligned}\|X\|_*^2 &\le \frac{2c_{\mathcal{S}}L}{\Delta T}\|F\|_{\Delta T}^2 + \left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right)^2\left(2L\Delta T\alpha^{-2}C_2^2c_{\mathcal{S}}^2 + \alpha^{-2}\Delta T\sum_{\ell=1}^{L}(1+C_1\Delta T)^{2(L-\ell)}\right)\\ &\le \frac{2Tc_{\mathcal{S}}}{\Delta T^2}\|F\|_{\Delta T}^2 + \alpha^{-2}\left(\|Y_L\| + \frac{T^{1/2}}{\Delta T}\|G\|_{\Delta T}\right)^2(2TC_2^2c_{\mathcal{S}}^2 + c_{\mathcal{S}}).\end{aligned}$$

Inserting the bound (4.24) into the above then gives

$$\|X\|_*^2 \le \frac{2Tc_{\mathcal{S}}}{\Delta T^2}\|F\|_{\Delta T}^2 + (2TC_2^2c_{\mathcal{S}}^2 + c_{\mathcal{S}})\left(\frac{\alpha^{-1}c_{\mathcal{S}}}{\Delta T}\|F\|_{\Delta T} + \frac{\alpha^{-1}T^{1/2} + \alpha^{-2}c_{\mathcal{S}}}{\Delta T}\|G\|_{\Delta T}\right)^2$$

Applying once more the inequality $(a+b)^2 \le 2(a^2+b^2)$ and simplifying, we obtain

$$\begin{aligned}\|X\|_*^2 &\le \frac{\|F\|_{\Delta T}^2}{\Delta T^2}\left(2Tc_{\mathcal{S}} + 2\alpha^{-2}c_{\mathcal{S}}^3 + 4\alpha^{-2}TC_2^2c_{\mathcal{S}}^4\right)\\ &\quad + \frac{\alpha^{-2}\|G\|_{\Delta T}^2}{\Delta T^2}\left(4Tc_{\mathcal{S}} + 4\alpha^{-2}c_{\mathcal{S}}^3 + 8\alpha^{-2}TC_2^2c_{\mathcal{S}}^4 + 8T^2C_2^2c_{\mathcal{S}}^2\right)\\ &\le \frac{\|F\|_{\Delta T}^2 + \alpha^{-2}\|G\|_{\Delta T}^2}{\Delta T^2}\cdot 4c_{\mathcal{S}}(1+2TC_2^2c_{\mathcal{S}})\left(T + \alpha^{-2}c_{\mathcal{S}}^2\right)\\ &\le \frac{K^2(T^{1/2} + \alpha^{-1}c_{\mathcal{S}})^2\|E\|_*^2}{\Delta T^2},\end{aligned}$$

where $K^2 = 4c_{\mathcal{S}}(1+2TC_2^2c_{\mathcal{S}})$. Taking square roots finally yields

$$\|X\|_* = \left\|\mathcal{M}_{\Delta t}^{-1}\Pi E\right\|_* \le \frac{c_{\mathcal{M}^{-1}}}{\Delta T}(1+\alpha^{-1})\|E\|_*,$$

with $c_{\mathcal{M}^{-1}} = \max\{KT^{1/2}, Kc_{\mathcal{S}}\}$, as required. □

The main theorem of this paper is now simply a consequence of the results proven above.

THEOREM 4.6. *Let $\alpha > 0$ be fixed, and suppose that the interval $[0,T]$ is subdivided into L subintervals of length ΔT, which is then further discretized with step sizes δt and Δt that are sufficiently small. If a Runge-Kutta method (RK) satisfying both the IVP and optimal control conditions of order p is used to form the matrices $\mathcal{M}_{\Delta t}$ and $\mathcal{M}_{\delta t}$, then there exists a constant $c_\rho > 0$ independent of δt, Δt, ΔT and α such that the convergence factor ρ of the iteration matrix $\mathcal{M}_{\Delta t}^{-1}(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})$ satisfies*

$$\rho \le \frac{c_\rho(1+\alpha^{-1})}{\Delta T}\Delta t^p. \tag{4.27}$$

Proof. Since the spectral radius is smaller than any operator norm, we have

$$\rho \leq \|\mathcal{M}_{\Delta t}^{-1}(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})\|_* = \|\mathcal{M}_{\Delta t}^{-1}\Pi(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})\|_* \leq \|\mathcal{M}_{\Delta t}^{-1}\Pi\|_* \|\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\|_*.$$

The result then follows directly from Theorems 4.2 and 4.5 if we define $c_\rho := c_{\mathcal{M}} c_{\mathcal{M}^{-1}}$. □

Remark 4.7. Note that two different Runge-Kutta methods can be used for the coarse and fine solvers. As an example, let us consider two different Runge-Kutta methods RK_1 and RK_2 (see (RK)) that satisfy both IVP and the optimal control conditions up to orders p and q, respectively. Without loss of generality, we can use RK_1 and RK_2 to discretize the optimal control problem (2.5) on the fine and coarse grids, respectively. As a result, (4.3) holds for each grid and, for $\kappa = \min\{p, q\}$, we derive the estimate corresponding to (4.8) as follows

$$\|\Delta\mathcal{P}\| \leq \|\mathcal{P}_{\Delta t} - \mathcal{P}_0\| + \|\mathcal{P}_{\delta t} - \mathcal{P}_0\| \leq c_{\mathcal{P}} \left(\Delta t^q + \delta t^p\right) \leq 2c_{\mathcal{P}} \Delta t^\kappa,$$

similarly for $\|\Delta\mathcal{R}\|$. Consequently, the truncation error estimate of Theorem 4.2 becomes

$$\|\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\|_* \leq c_{\mathcal{M}} \Delta t^\kappa, \tag{4.28}$$

and the stability estimate of Theorem 4.5 continues to hold. Hence, under the assumptions of Theorem 4.6, the convergence factor ρ of the iteration matrix $\mathcal{M}_{\Delta t}^{-1}\left(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\right)$ satisfies

$$\rho \leq \frac{c_\rho(1+\alpha^{-1})}{\Delta T} \Delta t^\kappa. \tag{4.29}$$

Remark 4.8. In order to obtain good speedup in practical implementations, one should choose a coarse discretization in time that is much cheaper to integrate than the fine discretization, so as to reduce the cost of computing matrix-vector products with $\mathcal{P}_{\Delta t}$, $\mathcal{P}_{\Delta t}^T$ and $\mathcal{R}_{\Delta t}$. This can be done by either increasing the time step size Δt, or by reducing the order of the integrator. Moreoever, the preconditioning step requires solving a linear system with $\mathcal{M}_{\Delta t}$: this should not be done by assembling and factoring the matrix explicitly, but by using an inner preconditioned iteration instead, see for instance [3] for details.

Remark 4.9. The estimate (4.27) holds for an arbitrary choice of Δt and δt, as long as $\delta t \leq \Delta t$. In the specific but relatively common case of $\delta t = \Delta t / N_0$ with $N_0 \geq 2$, Theorem A.1 shows that $\|\Delta\mathcal{P}\| \leq c_{\mathcal{P}}(\Delta t - \delta t)\Delta t^{p-1}$ and $\|\Delta\mathcal{R}\| \leq c_{\mathcal{R}}(\Delta t - \delta t)\Delta t^{p-1}$. Therefore, the error estimate in Theorem 4.2 can be refined to give $\|\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\| \leq c_{\mathcal{M}}(\Delta t - \delta t)\Delta t^{p-1}$, which leads to the following tighter estimate of the convergence factor

$$\rho \leq \frac{c_\rho(1+\alpha^{-1})}{\Delta T}(\Delta t - \delta t)\Delta t^{p-1}. \tag{4.30}$$

One can derive a similar estimate for the case when different sets of coefficients in (RK) are used for the fine and coarse grids.

Remark 4.10. Using arguments similar to the proof of Theorem 4.2 and Theorem 4.5, we can show that the condition number of $\mathcal{M}_{\Delta t}$ with respect to the norm $\|\cdot\|_*$ is bounded by $\gamma(1+\alpha^{-1})^2/\Delta T$, where γ is a constant that depends on $\mathcal{L}$ and T, but not on Δt and ΔT. The details of the proof can be found in Theorem A.2.

5. Numerical results.

In this section, we investigate numerically the theoretical estimate (4.27), i.e., the order of the convergence factor ρ with respect to Δt. We perform these experiments first on a linear ODE, then on a linear PDE discretized in space, and finally on a nonlinear control problem associated with a Schrödinger-type equation. We use MATLAB `R2021b` as our numerical computing environment.

5.1. Linear ODE example.

We start by considering an academic example in which the dynamics are given by (2.2). The matrix $\mathcal{L}$ and $\mathcal{B}$ are generated randomly as follows: we let $\mathcal{L} = \mathtt{rd} \cdot \mathtt{rd}^T$, where `rd` is a 10×10 random matrix with entries chosen uniformly in the interval $[0, 1]$. Similarly, we define $\mathcal{B}$ to be a 10×5 matrix with randomly generated entries following the same distribution. We will discretize the ODE in time using the Runge-Kutta methods presented in Table 2. Three of them satisfy both the IVP and optimal control conditions up to their respective orders: RK2, then SDIRK (Singly Diagonal Implicit Runge-Kutta) with $\gamma = (3+\sqrt{3})/6$ (see [20]), and finally Gauss-Lobatto (see [19]). The fourth method, namely RK3, satisfies the IVP conditions up to order 3, but the optimal control ones are only satisfied up to order 2.

For the parameters $T = 10^{-2}$, $\alpha = 10^{-1}$, $L = 10$ and $\Delta T = T/L$, we use the `eig` function in Matlab to compute the spectral radius ρ of the iteration matrix $\mathcal{M}_{\Delta t}^{-1}(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t})$. We do this for various Δt in four different cases:

TABLE 2

Butcher tables for various Runge-Kutta methods used in the linear ODE and PDE examples, with their respective IVP orders p indicated.

RK2, $p = 2$

0	0	0	0
$1/2$	$1/2$	0	0
1	0	1	0
	$1/4$	$1/2$	$1/4$

RK3, $p = 3$

0	0	0	0
$1/2$	$1/2$	0	0
$3/4$	0	$3/4$	0
	$2/9$	$1/3$	$4/9$

SDIRK, $p = 3$

γ	γ	0
$1-\gamma$	$1-2\gamma$	γ
	$1/2$	$1/2$

Gauss-Lobatto (GL), $p = 4$

0	0	0	0
$1/2$	$1/4$	$1/4$	0
1	0	1	0
	$1/6$	$2/3$	$1/6$

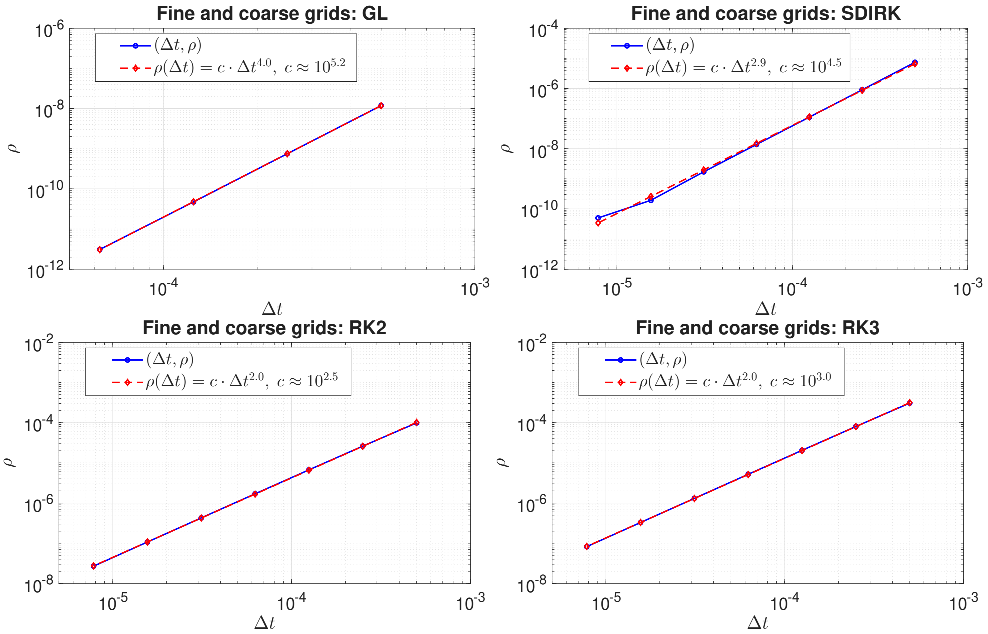

FIG. 2. *Convergence factor for various Runge-Kutta methods presented in Table 2 for a positive definite operator $\mathcal{L}$.*

- we use Gauss-Lobatto with fixed $\delta t = \Delta T/2^{10}$ and $\Delta t = \Delta T/2^k, k = 1, \ldots, 4$;
- we use SDIRK, RK2 and RK3 with fixed $\delta t = \Delta T/2^{16}$ and $\Delta t = \Delta T/2^k, 1, \ldots, 7$.

The results are shown in Figure 2, where we plot ρ on a logarithmic scale as a function of Δt in blue, with its linear regression in red. We observe that for the three methods that satisfy both the IVP and control order conditions, namely Gauss-Lobatto, SDIRK and RK2, ρ behaves like Δt^p for their respective orders p, which is consistent with (4.27). The behaviour is different for RK3, which satisfies the IVP conditions up to order $p = 3$ but does not satisfy the optimal control condition $\sum d_i^2/b_i = 1/3$ (see Table 1). The order of ρ with respect to Δt is found by regression to be 2, meaning that the optimal control conditions are necessary to get the third order behaviour.

5.2. A linear PDE example. In this test, we tackle the optimal control problem considered in [14], where the dynamics (2.2) are governed by the heat equation

$$\dot{y} - \Delta y = \mathcal{B}\nu, \tag{5.1}$$

with $y = y(x, t)$ is defined on $\Omega = [0, 1] \times [0, T]$, periodic boundary conditions along $\partial\Omega$, and $T = 10^{-2}$. We also set $\alpha = 10^{-1}$. The initial and target states are

$$\begin{aligned} y_{in} &= \exp(-100(x - 1/2)^2), \\ y_{tg} &= \frac{1}{2}\exp(-100(x - 1/4)^2) + \frac{1}{2}\exp(-100(x - 3/4)^2). \end{aligned}$$

The operator $\mathcal{B}$ is the indicator function of the sub-interval $\Omega_c = [1/3, 2/3]$ of Ω. We consider a semi-discretization in space of (5.1) using second-order centered finite-difference with $r = 50$ grid points.

We again let $L = 10$ and $\Delta T = T/L$. We fix $\delta t = \Delta T/2^{16}$ and vary Δt while making sure to satisfy the appropriate CFL condition whenever an explicit Runge-Kutta method is used. For each of the four methods listed in Table 2, we again compute the convergence factor of ParaOpt,

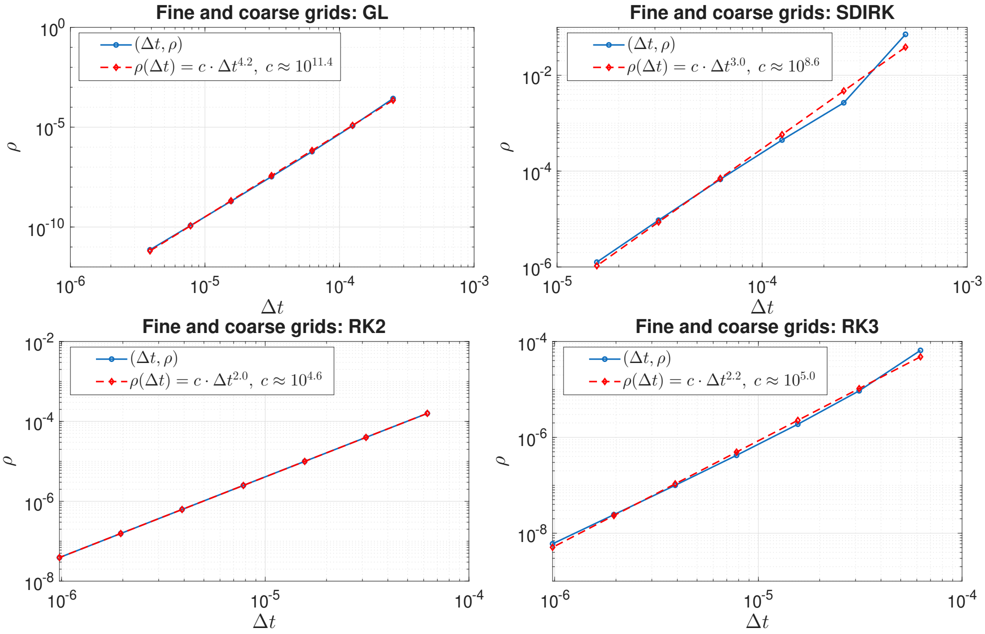

FIG. 3. *Order of the convergence factor Δt for various Runge-Kutta methods presented in Table 2 for (5.1).*

that is, the spectral radius ρ of the iteration matrix $\mathcal{M}_{\Delta t}^{-1}\left(\mathcal{M}_{\Delta t} - \mathcal{M}_{\delta t}\right)$ for the various Δt shown below:

- Gauss-Lobatto with $\Delta t = \Delta T/2^k$, $k = 2, \dots, 8$,
- SDIRK with $\Delta t = \Delta T/2^k$, $k = 1, \dots, 6$,
- RK2 with $\Delta t = \Delta T/2^k$, $k = 4, \dots, 10$,
- RK3 with $\Delta t = \Delta T/2^k$, $k = 4, \dots, 10$.

The results are shown in Figure 3. As in the previous section, we observe that since Gauss-Lobatto, SDIRK, and RK2 satisfy both the IVP conditions and the optimal control conditions up to order p, the observed order of convergence correspond to the predicted order. We also observe that this is not the case with RK3, since the optimal control conditions are not satisfied.

Next, we use different Runge-Kutta methods in the fine and coarse grids to simulate (4.29). For this instance, we consider three different test cases :

- we use SDIRK on the fine grid and implicit Euler (IE) on the coarse grid for $\Delta t = \Delta T/2^k, k = 3, \dots, 9$,
- we use Gauss-Lobatto on the fine grid and explicit Euler (EE) on the coarse grid for $\Delta t = \Delta T/2^k, k = 3, \dots, 10$,
- we use Gauss-Lobatto (GL) on the fine grid and RK2 on the coarse grid for $\Delta t = \Delta T/2^k, k = 2, \dots, 10$.

The results are presented in Figure 4, where we observe that the order of ρ with respect to Δt is determined by the lower order method. Our experiments are therefore consistent with (4.29).

We finally consider a case which is not covered by our analysis. In this test, the quadrature formula and the Runge-Kutta method (RK) used for discretizing the two ODE systems in (2.5) are independent in the sense that the coefficient b_j used in the Runge-Kutta method are different from the ones used in the quadrature part of (2.7), i.e., the formulas (2.6c) and (2.6d) do not use the same coefficients. We nonetheless assume that the quadrature nodes c_i are located at the same time points as the Runge-Kutta stages, since we would otherwise not be able to eliminate the control from the discrete optimality system. The results are shown in Figure 5, where we use RK2 for the quadrature formula and Gauss-Lobatto as (RK) to discretize the two ODE systems in (2.5). We see that the order of ρ is the minimum of the orders of both methods. The analysis for this case will be investigated in future work.

5.3. A nonlinear optimal control problem. In this section, we consider a nonlinear control problem involving the Schrödinger equation. More precisely, we minimize the following cost functional

$$\min_{\nu} \mathfrak{J}(\nu) := -\mathrm{Re}\left(\langle y(T), y_{tg}\rangle\right) + \frac{\alpha}{2}\int_0^T \nu(t)^2 dt, \tag{5.2}$$
$$\text{subject to} \qquad \dot{y}(t) = -i\mathcal{H}[\nu(t)]y(t) \quad \text{on } [0,T], \quad y(0) = y_{in}.$$

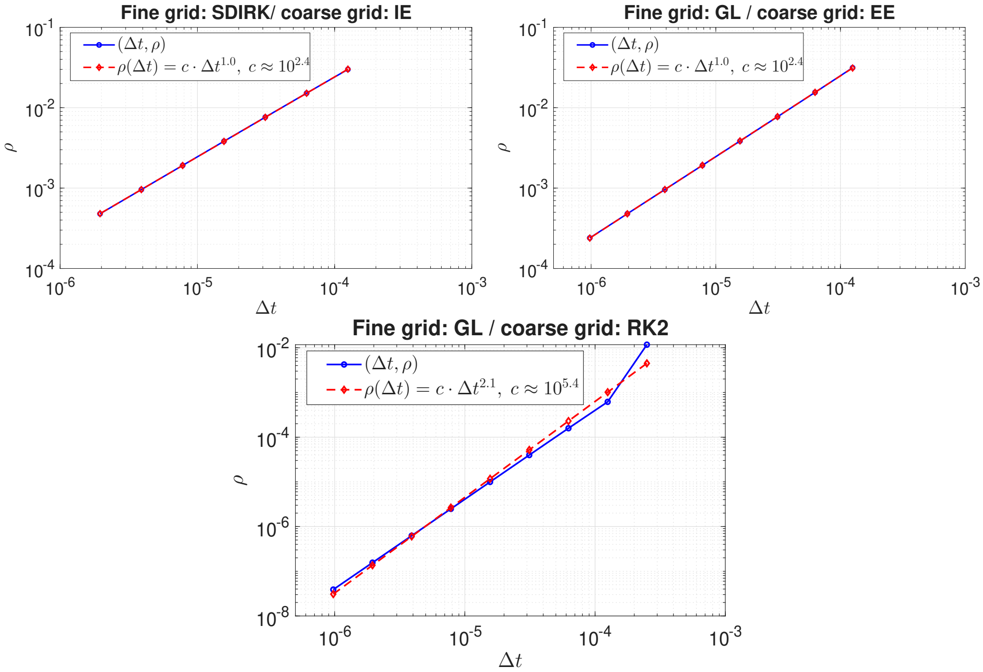

FIG. 4. *Order of the convergence factor when one uses two different Runge-Kutta methods on the fine and coarse grids for discretizing* (2.5)*, where EE represents explicit Euler and IE represents implicit Euler*

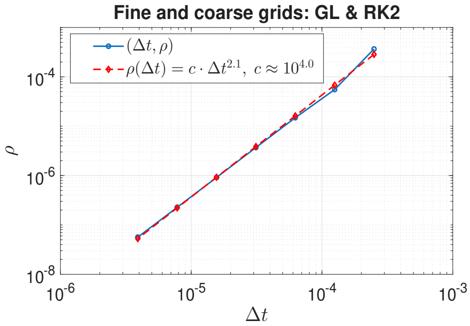

FIG. 5. *Order of the convergence factor when two different Runge-Kutta methods are used for discretizing* (2.1) *and* (2.2)*.*

The dynamics involved in (5.2) are that of a system of coupled spin $-1/2$ particles. The complete physical description can be found in [31]. The control ν consists of magnetic fields that act independently on one spin. We choose to focus on the case of five coupled spins whose interaction is encoded by the following Hamiltonian:

$$\mathcal{H}[\nu(t)] = \mathcal{L} + \sum_{k=1}^{5} \left[\nu_x^k(t) I_x^{(k)} + \nu_y^k(t) I_y^{(k)} \right]$$

where

$$\mathcal{L} = 2\pi J_p \left(I_z^{(1)} I_z^{(2)} + I_z^{(1)} I_z^{(3)} + I_z^{(2)} I_z^{(3)} + I_z^{(2)} I_z^{(5)} + I_z^{(3)} I_z^{(4)} \right),$$

the operators $I_x^{(k)}$ and $I_y^{(k)}$ are (up to a factor) Pauli matrices which only act on the kth spin:

$$I_x = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \quad I_y = \begin{pmatrix} 0 & -\frac{i}{2} \\ \frac{i}{2} & 0 \end{pmatrix}, \quad I_z = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix},$$

and $J_p = 140$ is the uniform coupling constant between the spin. The efficiency of ParaOpt in the case of non-linear dynamics has been studied numerically in [14] in the case of the Lotka-Volterra system. Here, we instead focus on the convergence behavior for a second order discretization of (5.2).

We use the discretization setting of $[0, T]$ presented in Section 2.1. We use the Crank-Nicolson

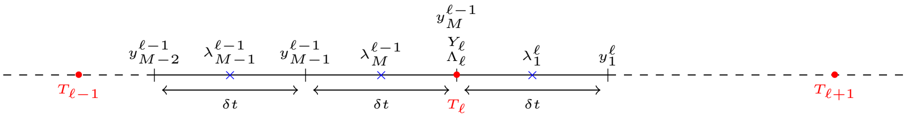

FIG. 6. *The stencil on the fine grids around T_ℓ.*

method to discretize both $\mathfrak{J}$ and the constraint in (5.2), that is, we have

$$\mathfrak{J}_{\delta t}(\nu) = -\mathrm{Re}\left(\langle y_{M_0}, y_{tg}\rangle\right) + \frac{\alpha}{2}\delta t \sum_{n=0}^{M_0-1}\sum_{k=1}^{5}\left([\nu_{x,n}^k]^2 + [\nu_{y,n}^k]^2\right),$$

and

$$\left(I + \mathcal{H}_0(\nu_n)\right) y_{n+1} = \left(I - \mathcal{H}_0(\nu_n)\right) y_n,$$

where $\mathcal{H}_0(\nu_n) = \frac{i}{2}\delta t \mathcal{H}(\nu_n)$ and $\nu_n = \nu(t_n + \delta t/2)$. Introducing the Lagrange function

$$\mathfrak{L}_{\delta t} = \mathfrak{J}_{\delta t}(\nu) - \mathrm{Re}\left(\sum_{n=0}^{M_0-1}\langle \lambda_{n+1}, \left(I + \mathcal{H}_0(\nu_n)\right) y_{n+1} - \left(I - \mathcal{H}_0(\nu_n)\right) y_n\rangle\right),$$

the Euler-Lagrange equations and elimination of the control give the following optimality system: $y(0) = y_{in}$ and

$$\left[I + \tilde{\mathcal{H}}_0(\lambda_{n+1}, y_{n+1}, y_n)\right] y_{n+1} = \left[I - \tilde{\mathcal{H}}_0(\lambda_{n+1}, y_{n+1}, y_n)\right] y_n,\ n = 0, \dots, M_0 - 1, \tag{5.3}$$

$$\left[I + \tilde{\mathcal{H}}_0^*(\lambda_n, y_n, y_{n-1})\right] \lambda_n = \left[I - \tilde{\mathcal{H}}_0^*(\lambda_{n+1}, y_{n+1}, y_n)\right] \lambda_{n+1},\ n = 1, \dots, M_0 - 1, \tag{5.4}$$

$$\left[I + \tilde{\mathcal{H}}_0^*(\lambda_{M_0}, y_{M_0}, y_{M_0-1})\right] \lambda_{M_0} = -\,\bar{y}_{tg}, \tag{5.5}$$

where $\bar{y}_{tg}$ is the complex conjugate of y_{tg} and $\tilde{\mathcal{H}}_0^*$ denote the adjoint matrix of $\tilde{\mathcal{H}}_0$ defined as follows:

$$\begin{aligned}\tilde{\mathcal{H}}_0(\lambda_{n+1}, y_{n+1}, y_n) :=& i\frac{\delta t}{2}\mathcal{L} + i\frac{\delta t}{4\alpha}\sum_{k=1}^{5}\mathrm{Im}\left(\left\langle \lambda_{n+1}, I_x^{(k)}(y_{n+1} + y_n)\right\rangle\right) I_x^{(k)} \\ &+ i\frac{\delta t}{4\alpha}\sum_{k=1}^{5}\mathrm{Im}\left(\left\langle \lambda_{n+1}, I_y^{(k)}(y_{n+1} + y_n)\right\rangle\right) I_y^{(k)}.\end{aligned}$$

With this discretization, λ is located in the middle of $[t_n, t_{n+1}]$, which means the adjoint is not defined at the interface $T_\ell = t_{M\ell}$ of the sub-intervals. We therefore need to extend the discrete adjoint to the interface in a way that is consistent with the continuous propagator and with the discrete equation (5.4). To do so, we add two equations, one for each sub-interval: for $[T_{\ell-1}, T_\ell]$, we add

$$\Lambda_\ell = \left[I + \tilde{\mathcal{H}}_0^*\left(\lambda_M^{\ell-1}, y_M^{\ell-1}, y_{M-1}^{\ell-1}\right)\right] \lambda_M^{\ell-1}, \tag{5.6}$$

which describes the propagation of the adjoint from the final condition Λ_ℓ over a distance of $\delta t/2$ to yield $\lambda_M^{\ell-1}$. On $[T_\ell, T_{\ell+1}]$, the adjoint at T_ℓ is denoted by $\mathcal{Q}(Y_\ell, \Lambda_{\ell+1})$ and is obtained by propagating λ_1^ℓ over a distance of $\delta t/2$. We therefore add the equation

$$\mathcal{Q}(Y_\ell, \Lambda_{\ell+1}) = \left[I - \tilde{\mathcal{H}}_0^*\left(\lambda_1^\ell, y_1^\ell, Y_\ell\right)\right] \lambda_1^\ell. \tag{5.7}$$

With these additions, we can now impose the matching condition

$$\Lambda_\ell - \mathcal{Q}(Y_\ell, \Lambda_{\ell+1}) = 0,$$

whose equivalence to (5.4) can be seen by substituting (5.6) and (5.7) into the above.

In this numerical test, we define y_{in} and y_{tg} as the first columns of $I_x^{(1)}$ and $I_x^{(5)}$ respectively. We set $T = 10/J_p, \alpha = 1, L = 10$, and fix the fine discretization step to $\delta t = 1/80J_p$. We show in Figure 7 the convergence rate of the ParaOpt for various values of the ratio $r = \delta t/\Delta t$. The L^∞ error is defined as the maximum of the difference between the state and adjoint values obtained

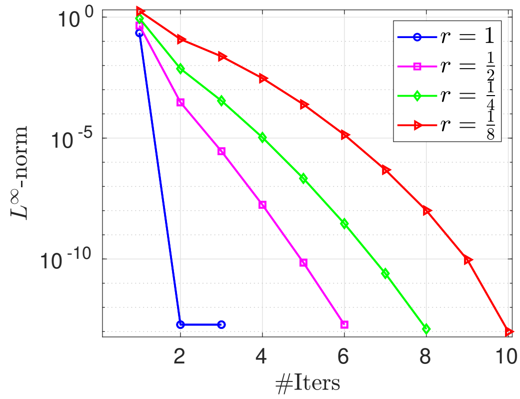

FIG. 7. *L^∞-norm error for various values of the ratio $r = \frac{\delta t}{\Delta t}$.*

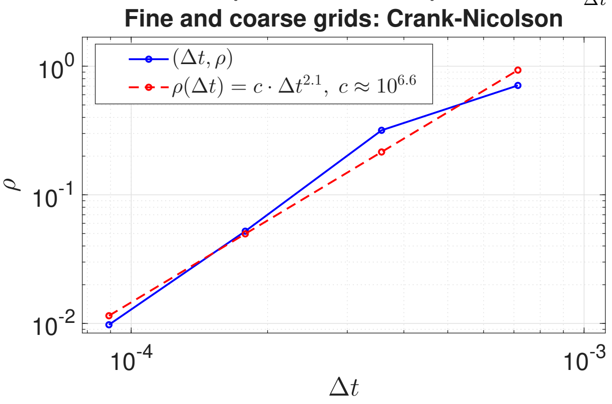

FIG. 8. *Order of the convergence factor of ParaOpt with respect to Δt.*

from a converged fine-grid solution, and the values obtained at each inexact Newton iteration using the ParaOpt .

Fast convergence in two iterations has been observed for $r = 1$, since the ParaOpt actually corresponds to the exact Newton method. When $r \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$, the approximation of the Jacobian becomes coarser, which explains the slower convergence. Denoting by a_r, k_r the slope and the number of iterations associated with the curve of ratio r on Figure 7 and by ρ the convergence factor of ParaOpt applied to (5.2), we observe that $|a_r| \approx \left(c\Delta t_r^2\right)^{k_r}$, where c is a positive real constant. It follows that for $\rho_r^{k_r} = |a_r|$, we obtain $\rho_r \approx c\Delta t_r^2$, that is, ρ is of order 2 (see Figure 8). This order is equal to the one of the Crank-Nicolson method used to discretize both $\mathfrak{J}$ and the constraint in (5.2). This is consistent with the result of Theorem 4.6 for a linear case problem, meaning that our results also hold in nonlinear cases.

L	# Iter	CPU time	Parallel computing time	Speedup	Efficiency
1	2	571.2983	571.2947	1.0	100%
2	3	117.9937	117.9925	4.84	242%
4	7	42.4054	42.4046	13.47	336,75%
8	9	12.8253	12.8235	44.54	556.75%
16	13	7.6024	7.6020	75.15	469.7%

TABLE 3

Performance of ParaOpt algorithm: total computing time CPU time, parallel computing time only in seconds, speedup (CPU time($L = 1$)/CPU time(L)) and efficiency (100 × speedup/L).

We also study how well the ParaOpt algorithm scales for solving the control problem (5.2). All computations were run in MATLAB R2021 on a Dell Precision 7780 laptop with 20 CPU cores and an NVIDIA RTX 2000 Ada GPU. The results are shown in Table 3, where we report the total computing time, the parallel computing time without communication, and the number of outer Newton iterations required to reach a tolerance of 10^{-10}. The total time corresponds to running the code implementing the ParaOpt algorithm. The parallel computing time is the time required to run the Newton method, where the local solves on the fine and coarse grids are performed in parallel using MATLAB's `parfor`. Our tests show very good scalability: when we double the number of processors, the total computing time is reduced by more than a factor of two. We obtain such efficiency because, as L increases, the local parallel solver become much faster. These results are similar to those presented in Table 2 of [14] for an optimal control problem involving Lotka–Volterra dynamics, solved using the implicit Euler method.

Appendix A. Additional results.

THEOREM A.1. *Let Δt and $N_0 \geq 2$ be fixed with $\delta t = \Delta t/N_0$. Then, there exist $c_{\mathcal{P}} > 0$ and $c_{\mathcal{R}} > 0$ independent of Δt and δt such that*

$$\|\Delta\mathcal{P}\| \leq c_{\mathcal{P}}(\Delta t - \delta t)\Delta t^{p-1} \quad \text{and} \quad \|\Delta\mathcal{R}\| \leq c_{\mathcal{R}}(\Delta t - \delta t)\Delta t^{p-1}.$$

Proof. We consider the mapping $\varphi : (0,1)^2 \longrightarrow (0,\infty)$ defined as follows

$$\varphi(\Delta t, \delta t) := \frac{\|\Delta\mathcal{P}\|}{\Delta t - \delta t}, \quad \text{for } \delta t < \Delta t. \tag{A.1}$$

The triangular inequality applied to the right-hand side of (A.1) leads to

$$\varphi(\delta t, \Delta t) \leq \frac{1}{\Delta t - \delta t}\left(\|\mathcal{P}_{\Delta t} - \mathcal{P}_0\| + \|\mathcal{P}_{\delta t} - \mathcal{P}_0\|\right).$$

Using the inequalities in Theorem 4.1 on $\mathcal{P}_{\delta t}$ and $\mathcal{P}_0$ by replacing δt by Δt , $\|\mathcal{P}_{\Delta t} - \mathcal{P}_0\| \leq c_{\Delta T}\Delta t^p$, so that,

$$\varphi(\delta t, \Delta t) \leq \frac{c_{\mathcal{P}}}{(\Delta t - \delta t)}\left(\Delta t^p + \delta t^p\right) \leq \frac{2c_{\mathcal{P}}}{(\Delta t - \delta t)}\Delta t^p.$$

Setting $\delta t = \Delta t/N_0$ for $N_0 \geq 2$ leads $\frac{1}{\Delta t - \delta t} = \frac{N_0}{\Delta t(N_0 - 1)} \leq \frac{2}{\Delta t}$, so that,

$$\varphi(\delta t, \Delta t) \leq 4c_{\mathcal{P}}\Delta t^{p-1}. \tag{A.2}$$

Substituting (A.1) into (A.2) gives rise to

$$\|\Delta\mathcal{P}\| \leq 4(\Delta t - \delta t)c_{\mathcal{P}}\Delta t^{p-1}.$$

Considering the inequalities in Theorem 4.1 on $\mathcal{R}_{\delta t}$ and $\mathcal{R}_0$ by replacing δt by Δt, we can proceed analogously to obtain

$$\|\Delta\mathcal{R}\| \leq 4\,(\Delta t - \delta t)\,c_{\mathcal{R}}\Delta t^{p-1}.$$

And the proof is complete. ∎

THEOREM A.2. *Let $0 < \Delta t \leq \Delta T \leq 1$. Then there exists a constant $\gamma > 0$ depending on the operator $\mathcal{L}$ and the time horizon T, but independent of Δt, ΔT and α, such that the condition number of $\mathcal{M}_{\Delta t}$ with respect to the norm $\|\cdot\|_*$ is bounded by*

$$\operatorname{cond}(\mathcal{M}_{\Delta t}) = \|\mathcal{M}_{\Delta t}\|_*\|\mathcal{M}_{\Delta t}^{-1}\|_* \leq \frac{\gamma(1+\alpha^{-1})^2}{\Delta T}.$$

Proof. To bound $\|\mathcal{M}_{\Delta t}\|_*$, we proceed like in Theorem 4.2: we consider $X = \begin{pmatrix} Y \\ \Lambda \end{pmatrix}$ and $E = \begin{pmatrix} F \\ G \end{pmatrix}$ such that $E = \mathcal{M}_{\Delta t}X$ and bound $\|E\|_*$ in terms of $\|X\|_*$. From the definition of the matrix $\mathcal{M}_{\Delta t}$, we see that $E = X + \Delta E$, where $\Delta E = \begin{pmatrix} \Delta F \\ \Delta G \end{pmatrix}$ satisfies

$$\begin{aligned}
&\Delta F_0 = 0, && \\
&\Delta F_\ell = -\mathcal{P}_{\Delta t}Y_{\ell-1} + \frac{1}{\alpha}\mathcal{R}_{\Delta t}\Lambda_\ell, && \ell = 1,\dots,L, \\
&\Delta G_\ell = -\mathcal{P}_\ell^T\Lambda_{\ell+1}, && \ell = 1,\dots,L-1, \\
&\Delta G_L = -Y_L. &&
\end{aligned}$$

These are almost the same equations as in Theorem 4.2, except we replaced $\Delta\mathcal{P}$ and $\Delta\mathcal{R}$ by $\mathcal{P}_{\Delta t}$ and $\mathcal{R}_{\Delta t}$, and that $\Delta G_L = -Y_L$ instead of 0. Therefore, the same calculation as in the theorem shows that

$$\begin{aligned}
\|\Delta F\|_{\Delta T}^2 &\leq (\|\mathcal{P}\|_{\Delta T}^2 + \|\mathcal{R}\|_{\Delta T}^2)\|X\|_*^2, \\
\|\Delta G\|_{\Delta T}^2 &\leq \|\mathcal{P}_{\Delta T}\|^2\|\Lambda\|_{\Delta T}^2 + \Delta T\|Y_L\|^2.
\end{aligned}$$

Combining the above inequalities gives

$$\begin{aligned}\|\Delta E\|_*^2 &\le \left(2\|\mathcal{P}_{\Delta t}\|^2 + \|\mathcal{R}_{\Delta t}^2\|\right)\|X\|_*^2 + \alpha^{-2}\Delta T\|Y_L\|^2\\ &\le \left(2\|\mathcal{P}_{\Delta t}\|^2 + \|\mathcal{R}_{\Delta t}\|^2 + \alpha^{-2}\right)\|X\|_*^2\\ &\le (c_1+\alpha^{-1})^2\|X\|_*^2,\end{aligned}$$

where $c_1^2 = 2\|\mathcal{P}_{\Delta t}\|^2 + \|\mathcal{R}_{\Delta t}\|^2 \le 3 + c_1'\Delta T \le 3 + c_1'T$ is obtained using the estimates (4.23) on $\mathcal{P}_{\Delta t}$ and $\mathcal{R}_{\Delta t}$. Therefore, $\|E\|_* \le (1+c_1+\alpha^{-1})\|X\|_*$, which implies $\|\mathcal{M}_{\Delta t}\|_* \le \gamma_1(1+\alpha^{-1})$ for some constant $\gamma_1 > 0$ depending on C_1, C_2 and T, but not on Δt and ΔT.

We now bound $\|\mathcal{M}_{\Delta t}^{-1}\|_*$. Since $\mathcal{M}_{\Delta t}^{-1} = \mathcal{M}_{\Delta t}^{-1}\Pi + \mathcal{M}_{\Delta t}^{-1}(\mathcal{I}-\Pi)$ and we already have an estimate of $\|\mathcal{M}_{\Delta t}^{-1}\Pi\|_*$ from Theorem 4.5, it suffices to estimate $\|\mathcal{M}_{\Delta t}^{-1}(\mathcal{I}-\Pi)\|_*$. Let $X = \mathcal{M}_{\Delta t}^{-1}(\mathcal{I}-\Pi)E$, i.e., $\mathcal{M}_{\Delta t}X = (\mathcal{I}-\Pi)E$. Then the blocks of E and X satisfy the recurrence

$$\begin{aligned} Y_0 &= F_0, &&\\ -\mathcal{P}_{\Delta t}Y_{\ell-1} + Y_\ell + \alpha^{-1}\mathcal{R}_{\Delta t}\Lambda_\ell &= 0, && \ell = 1,\dots,L,\\ \Lambda_{\ell-1} - \mathcal{P}_{\Delta t}^T\Lambda_\ell &= 0, && \ell = 2,\dots,L,\\ \Lambda_L - Y_L &= G_L. &&\end{aligned}$$

Solving this recurrence using the same techniques as in Lemma 4.3, we deduce for $\ell = 1,\dots,L-1$ that

$$\Lambda_\ell = (\mathcal{P}_{\Delta t}^T)^{L-\ell}\Lambda_L \qquad \text{and} \qquad Y_\ell = \mathcal{P}^\ell F_0 - \alpha^{-1}\sum_{k=1}^{\ell}\mathcal{P}_{\Delta t}^{\ell-k}\mathcal{R}_{\Delta t}(\mathcal{P}_{\Delta t}^T)^{L-k}\Lambda_L, \tag{A.3}$$

where Λ_L satisfies the reduced system

$$(I + \alpha^{-1}\mathcal{U})\Lambda_L = \mathcal{P}_{\Delta t}^L F_0 + G_L,$$

where $\mathcal{U}$ is the same matrix as in (4.21). Therefore, we have by Lemma 4.3 and inequality (4.23)

$$\|\Lambda_L\| \le \|\mathcal{P}_{\Delta t}^L F_0 + G_L\| \le (1+C_1\Delta T)^L\|F_0\| + \|G_L\| \le e^{C_1T}\|F_0\| + \|G_L\|. \tag{A.4}$$

We now take norms on the equations in (A.3) and use the estimates for $\|\mathcal{P}_{\Delta t}\|$ and $\|\mathcal{R}_{\Delta t}\|$ in (4.23) to get

$$\begin{aligned}\|\Lambda_\ell\| &\le (1+C_1\Delta T)^{L-\ell}\|\Lambda_L\|,\\ \|Y_\ell\| &\le (1+C_1\Delta T)^\ell\|F_0\| + \alpha^{-1}C_2\Delta T\sum_{k=1}^{\ell}(1+C_1\Delta T)^{2L-\ell-k}\|\Lambda_L\|\\ &\le (1+C_1\Delta T)^\ell\|F_0\| + \alpha^{-1}C_2T(1+C_1\Delta T)^{2L-\ell}\|\Lambda_L\|,\end{aligned}$$

where we used $(1+C_1\Delta T)^{-k} \le 1$ and $\ell\Delta T \le T$ to obtain the last inequality. The definition of the norm $\|\cdot\|_*$ finally gives

$$\begin{aligned}\|X\|_*^2 &= \Delta T\sum_{\ell=0}^{L}\|Y_\ell\|^2 + \alpha^{-2}\Delta T\sum_{\ell=1}^{L}\|\Lambda_\ell\|^2\\ &\le \Delta T\|F_0\|^2 + 2\Delta T\sum_{\ell=1}^{L}(1+C_1\Delta T)^{2\ell}\|F_0\|^2\\ &\quad + 2(C_2T)^2\alpha^{-2}\Delta T\sum_{\ell=1}^{L}(1+C_1\Delta T)^{4L-2\ell}\|\Lambda_L\|^2 + \alpha^{-2}\Delta T\sum_{\ell=1}^{L}(1+C_1\Delta T)^{2L-2\ell}\|\Lambda_L\|^2\\ &\le \left(1 + \frac{2(e^{2C_1T}-1)}{\Delta T}\right)\Delta T\|F_0\|^2 + \alpha^{-2}\Delta T(1+2(C_2T)^2e^{2C_1T})\frac{e^{2C_1T}-1}{(1+C_2\Delta T)^2-1}\|\Lambda_L\|^2.\end{aligned}$$

Combining the above with (A.4), we deduce that

$$\|X\|_* \le \frac{\gamma_2(1+\alpha^{-1})}{\sqrt{\Delta T}}\|E\|_*$$

for a constant $\gamma_2 > 0$ that depends on C_1, C_2 and T, but not on Δt and ΔT. This implies

$$\|\mathcal{M}_{\Delta t}^{-1}(\mathcal{I} - \Pi)\|_* \leq \frac{\gamma_2(1 + \alpha^{-1})}{\sqrt{\Delta T}}.$$

Together with the result $\|\mathcal{M}_{\Delta t}^{-1}\Pi\|_* \leq \frac{c_{\mathcal{M}^{-1}}}{\Delta T}(1+\alpha^{-1})$ from Theorem 4.5, we deduce that $\|\mathcal{M}_{\Delta t}^{-1}\|_* \leq \frac{\gamma_3(1+\alpha^{-1})}{\Delta T}$ for some constant $\gamma_3 \geq 0$, which finally allows us to conclude that

$$\|\mathcal{M}_{\Delta t}\|_* \|\mathcal{M}_{\Delta t}^{-1}\|_* \leq \frac{\gamma(1 + \alpha^{-1})^2}{\Delta T}$$

with $\gamma = \gamma_1\gamma_3$. □

Acknowledgment. We thank the anonymous referees for their valuable suggestions, which greatly improved our paper.

REFERENCES

[1] W. Agboh, O. Grainger, D. Ruprecht, and M. Dogar, *Parareal with a learned coarse model for robotic manipulation*, Computing And Visualization In Science, 23 (2020), https://doi.org/10.1007/s00791-020-00327-0.

[2] Z. Belhachmi and D. Gilliocq-Hirtz, *Coupling parareal and adaptive control in optical flow estimation with application in movie's restoration*, in International Conference on Computer Vision and Image Analysis Applications, 2015, pp. 1–6, https://doi.org/10.1109/ICCVIA.2015.7351873.

[3] A. Bouillon, G. Samaey, and K. Meerbergen, *Diagonalization-based preconditioners and generalized convergence bounds for paraOpt*, SIAM Journal on Scientific Computing, 46 (2024), pp. S317–S345, https://doi.org/10.1137/23M1571423.

[4] T. Buvoli and M. Minion, *Imex runge-kutta parareal for non-diffusive equations*, in Parallel-in-time integration methods, B. Ong, J. Schroder, J. Shipton, and S. Friedhoff, eds., vol. 356 of Springer Proceedings in Mathematics & Statistics, 2021, pp. 95–127, https://doi.org/10.1007/978-3-030-75933-9_5. 9th Parallel-in-Time Workshop (PinT), Jun 08-12, 2020.

[5] J. J. Caceres Silva, B. Barán, and C. E. Schaerer, *Implementation of a distributed parallel in time scheme using petsc for a parabolic optimal control problem*, in 2014 Federated Conference on Computer Science and Information Systems, 2014, pp. 577–586, https://doi.org/10.15439/2014F340.

[6] G. Čaklović, T. Lunet, S. Götschel, and D. Ruprecht, *Improving efficiency of parallel across the method spectral deferred corrections*, SIAM Journal on Scientific Computing, 47 (2025), pp. A430–A453.

[7] L. D'Amore and R. Cacciapuoti, *Model reduction in space and time for ab initio decomposition of 4d variational data assimilation problems*, Applied Numerical Mathematics, 160 (2021), pp. 242–264, https://doi.org/10.1016/j.apnum.2020.10.003.

[8] A. L. Dontchev, W. W. Hager, and V. M. Veliov, *Second-order runge–kutta approximations in control constrained optimal control*, SIAM Journal on Numerical Analysis, 38 (2000), pp. 202–226, https://doi.org/10.1137/S0036142999351765.

[9] M. Fisher and S. Gürol, *Parallelization in the time dimension of four-dimensional variational data assimilation*, Quarterly Journal of the Royal Meteorological Society, 143 (2017), pp. 1136–1147, https://doi.org/https://doi.org/10.1002/qj.2997.

[10] S. Friedhoff and B. S. Southworth, *On "optimal" h-independent convergence of parareal and multigrid-reduction-in-time using runge-kutta time integration*, Numerical Linear Algebra With Applications, 28 (2021), https://doi.org/10.1002/nla.2301.

[11] M. J. Gander, *50 years of time parallel time integration*, in Multiple Shooting and Time Domain Decomposition Methods, T. Carraro, M. Geiger, S. Körkel, and R. Rannacher, eds., Springer International Publishing, Cham, 2015, pp. 69–113, https://doi.org/https://doi.org/10.1007/978-3-319-23321-5_3.

[12] M. J. Gander and E. Hairer, *Analysis for parareal algorithms applied to hamiltonian differential equations*, Journal Of Computational And Applied Mathematics, 259 (2014), pp. 2–13, https://doi.org/10.1016/j.cam.2013.01.011. 16th International Congress on Computational and Applied Mathematics (ICCAM), Ghent, BELGIUM, JUL 09-13, 2012.

[13] M. J. Gander, L. Halpern, J. Rannou, and J. Ryan, *A direct time parallel solver by diagonalization for the wave equation*, SIAM Journal on Scientific Computing, 41 (2019), pp. A220–A245.

[14] M. J. Gander, F. Kwok, and J. Salomon, *ParaOpt: A parareal algorithm for optimality systems*, SIAM Journal on Scientific Computing, 42 (2020), pp. A2773–A2802, https://doi.org/10.1137/19M1292291.

[15] M. J. Gander, T. Lunet, D. Ruprecht, and R. Speck, *A unified analysis framework for iterative parallel-in-time algorithms*, SIAM Journal On Scientific Computing, 45 (2023), pp. A2275–A2303, https://doi.org/10.1137/22M1487163.

[16] W. W. Hager, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM Journal on Numerical Analysis, 13 (1976), pp. 449–472, https://doi.org/10.1137/0713040.

[17] W. W. Hager, *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numerische Mathematik, 87 (2000), pp. 247–282, https://doi.org/https://doi.org/10.1007/s002110000178.

[18] J. Hahne, B. S. Southworth, and S. Friedhoff, *Asynchronous truncated multigrid-reduction-in-time*, SIAM Journal On Scientific Computing, 45 (2023), pp. S281–S306, https://doi.org/10.1137/21M1433149.

[19] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer-Verlag Berlin Heidelberg, 1993.

[20] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Springer-Verlag Berlin Heidelberg, 1996.

[21] F. Kwok, *On the time-domain decomposition of parabolic optimal control problems*, in Domain Decomposition Methods In Science And Engineering XXIII, C. Lee, X. Cai, D. Keyes, H. Kim, A. Kla-

wonn, E. Park, and O. Widlund, eds., vol. 116 of Lecture Notes in Computational Science and Engineering, KAIST Math Res Stat; Natl Inst Math Sci; Korean Federat Sci & Technol Soc; KISTI Supercomputing Ctr; A3 Foresight Program; Nvidia; Jeju Convent & Visitors Bur, 2017, pp. 55–67, https://doi.org/10.1007/978-3-319-52389-7_5. 23rd International Conference on Domain Decomposition Methods, SOUTH KOREA, JUL 06-10, 2015.

[22] A. Lachapelle, J. Salomon, and G. Turinici, *COMPUTATION OF MEAN FIELD EQUILIBRIA IN ECONOMICS*, MATHEMATICAL MODELS & METHODS IN APPLIED SCIENCES, 20 (2010), pp. 567–588, https://doi.org/{10.1142/S0218202510004349}.

[23] A. Lapin and A. Romanenko, *Udzawa-type iterative method with parareal preconditioner for a parabolic optimal control problem*, in 11TH International conference on mesh methods for boundry-value problems and applications, vol. 158 of IOP Conference Series-Materials Science and Engineering, Kazan Fed Univ; Keldysh Inst Appl Math; Tatarstan Acad Sci; Lomonosov Moscow State Univ, 2016, https://doi.org/10.1088/1757-899X/158/1/012059. 11th International Conference on Mesh Methods for Boundry-Value Problems and Applications, Kazan, RUSSIA, OCT 20-25, 2016.

[24] J.-L. Lions, Y. Maday, and G. Turinici, *Résolution d'edp par un schéma en temps "pararéel"*, Comptes Rendus de l'Académie des Sciences - Series I - Mathematics, 332 (2001), pp. 661–668, https://doi.org/https://doi.org/10.1016/S0764-4442(00)01793-6.

[25] Y. Maday, J. Salomon, and G. Turinici, *Monotonic parareal control for quantum systems*, SIAM Journal On Numerical Analysis, 45 (2007), pp. 2468–2482, https://doi.org/10.1137/050647086.

[26] Y. Maday and G. Turinici, *A parallel in time approach for quantum control: the parareal algorithm*, in Proceedings of the 41st IEEE Conference on Decision and Control, 2002., vol. 1, 2002, pp. 62–66 vol.1, https://doi.org/10.1109/CDC.2002.1184468.

[27] Y. Maday and G. Turinici, *Parallel in time algorithms for quantum control: Parareal time discretization scheme*, International Journal Of Quantum Chemistry, 93 (2003), pp. 223–228, https://doi.org/10.1002/qua.10554.

[28] T. P. Mathew, M. Sarkis, and C. E. Schaerer, *Analysis of block parareal preconditioners for parabolic optimal control problems*, SIAM Journal On Scientific Computing, 32 (2010), pp. 1180–1200, https://doi.org/10.1137/080717481.

[29] B. W. Ong and J. B. Schroder, *Applications of time parallelization*, Computing and Visualization in Science, 23 (2020), p. 11, https://doi.org/10.1007/s00791-020-00331-4.

[30] V. Rao and A. Sandu, *A time-parallel approach to strong-constraint four-dimensional variational data assimilation*, Journal Of Computational Physics, 313 (2016), pp. 583–593, https://doi.org/10.1016/j.jcp.2016.02.040.

[31] M. K. Riahi, J. Salomon, S. J. Glaser, and D. Sugny, *Fully efficient time-parallelized quantum optimal control algorithm*, Physical Review A, 93 (2016), https://doi.org/{10.1103/PhysRevA.93.043410}.

[32] B. S. Southworth, *Necessary conditions and tight two-level convergence bounds for parareal and multigrid reduction in time*, SIAM Journal On Matrix Analysis And Applications, 40 (2019), pp. 564–608, https://doi.org/10.1137/18M1226208.

[33] E. Trélat, *Contrôle optimal: théorie & applications*, Vuibert, Paris, 2nd ed., 2008.

[34] S.-L. Wu and T.-Z. Huang, *A fast second-order parareal solver for fractional optimal control problems*, Journal Of Vibration And Control, 24 (2018), pp. 3418–3433, https://doi.org/10.1177/1077546317705557.

[35] X. Yue, S. Shu, X. Xu, W. Bu, and K. Pan, *Parallel-in-time multigrid for space-time finite element approximations of two-dimensional space-fractional diffusion equations*, Computers & Mathematics With Applications, 78 (2019), pp. 3471–3484, https://doi.org/10.1016/j.camwa.2019.05.017.