Discrete adjoint computations for relaxation Runge-Kutta methods

Mario J. Bencomo*, Jesse Chan

Rice University, Houston TX

Abstract

Relaxation Runge-Kutta methods reproduce a fully discrete dissipation (or conservation) of entropy for entropy stable semi-discretizations of nonlinear conservation laws. In this paper, we derive the discrete adjoint of relaxation Runge-Kutta schemes, which are applicable to discretize-then-optimize approaches for optimal control problems. Furthermore, we prove that the derived discrete relaxation Runge-Kutta adjoint preserves time-symmetry when applied to linear skew-symmetric systems of ODEs. Numerical experiments verify these theoretical results while demonstrating the importance of appropriately treating the relaxation parameter when computing the discrete adjoint.

Keywords: relaxation Runge-Kutta method, discrete adjoint, time-symmetry, entropy conservation, entropy stability

1. Introduction

The *relaxation Runge-Kutta* method was first introduced by [1, 2] for stability of time discretizations of ordinary differential equations (ODEs) with respect to a given inner-product norm, or in more general a convex entropy functional. Recently, this relaxation approach has been generalized to multistep time integrators, [3]. We are interested in the application of these relaxation methods to optimal control problems, in hopes of leveraging stability properties especially for entropy stable semi-discretizations of nonlinear partial differential equations (PDEs).

In this paper we present novel adjoint computations and properties of the relaxation Runge-Kutta method. The discrete linearization and adjoint of the relaxation Runge-Kutta method is derived by using a matrix representation, similar to [4], and by using implicit differentiation in order to address the dependency of the relaxation parameter on the solution at current time steps. Numerical experiments presented here highlight the importance of proper linearization.

*Corresponding author
Email address: mjb6@rice.edu (Mario J. Bencomo)

We also prove, and demonstrate numerically, time-symmetry properties of the relaxation Runge-Kutta method when applied to linear skew-symmetric ODE systems, for general explicit and diagonally implicit Runge-Kutta methods.

Adjoint computations, by which we mean computations associated with the adjoint state method, are an efficient way to compute gradients of objective functionals in PDE-constrained optimization problems; see [5, 6] for an overview of optimal control problems and the adjoint state method. There are two main approaches associated with adjoint computations: one can either derive the adjoint state equations for the continuous problem and then discretize (referred to as the *optimize-then-discretize* approach) or alternatively discretize the state equations first and then compute the adjoint (referred to as the *discretize-then-optimize*). The advent of *automatic/algorithmic differentiation* (AD) has accelerated the advancement and utility of PDE-constrained optimization by essentially automating the adjoint state method and computation of sensitivities in a discretize-then-optimize approach, [7, 8]. However, regardless of the convenience of AD, one must exercise caution since a discretize-then-optimize approach may produce a discretization that is inconsistent with the continuous optimization problem, e.g., [9].

Issues with the discretize-then-optimize approach are dependent on the choice of discretization of the state equations, and much research has gone into analyzing this approach for different numerical schemes. Previous work on the discrete adjoint of Runge-Kutta methods showed that a discretize-then-optimize approach is indeed consistent, and moreover, that the adjoint of a Runge-Kutta method is yet another Runge-Kutta scheme of same order, [10, 11, 12]. In [13], the author examines the links between symplectic Runge-Kutta methods and applications into the computation of sensitivities and adjoints. See also [14] for a more general paper on the discrete differentiation and convergence of iterative solvers.

The relaxation Runge-Kutta method can be viewed as an adaptive time step method, making the step-size dependent on the solution at previous time steps. Previous work related to discrete adjoints of generic adaptive time stepping methods argues that taking the variable step-size into account in the linearization produces "non-physical" effects in the sensitivity and adjoint computations, [15, 16]. The authors also argue that the resulting discretize-then-optimize approach is inconsistent. In this paper, however, the opposite is true, and a proper linearization of relaxation Runge-Kutta methods is not only consistent but also necessary for accuracy.

The paper is outlined as follows: In section 2.1, we present some notation, along with the standard Runge-Kutta method, as well as its discrete linearization and adjoint. Next, in section 2.2, we discuss the relaxation Runge-Kutta method, derive its discrete linearization and adjoint, and discuss its time-symmetry property. In section 3, the numerical experiments and results section, we demonstrate the importance of proper linearization for relaxation Runge-Kutta methods. We also verify the time-symmetry property of these relaxation methods on a skew-symmetric linear problem. Results are summarized in the conclusion, section 4. Detailed derivations of discrete linearization and adjoint

formulas are given in the appendix.

2. Theory

Consider the following optimal control problem:

$$\min_{\mathbf{u}} \ \mathcal{C}(\mathbf{y}, \mathbf{u}) \tag{1a}$$

$$\text{s.t.} \ \ \mathcal{E}(\mathbf{y}, \mathbf{u}) = 0 \tag{1b}$$

where $\mathbf{y}$ and $\mathbf{u}$ denote the vectors of state and control variables respectively. The state equation 1b specifies a system of first-order initial value problem (IVP), potentially the semi-discretization of some PDE, of the form:

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}, \mathbf{u}, t), \quad 0 < t \leq T \tag{2a}$$

$$\mathbf{y}(0) = \mathbf{y}_{\text{init}}(\mathbf{u}), \tag{2b}$$

with $\mathbf{y}, \mathbf{y}_{\text{init}}, \mathbf{f}(\mathbf{y}, \mathbf{u}, t) \in \mathbb{R}^N$. Following a discretize-then-optimize approach, the continuous optimal control problem 1 is replaced by the following discrete optimization problem:

$$\min_{\mathbf{u}} \ \mathsf{C}(\mathbf{y}, \mathbf{u}) \tag{3a}$$

$$\text{s.t.} \ \ \mathbf{E}(\mathbf{y}, \mathbf{u}) = \mathbf{0} \tag{3b}$$

where C and $\mathbf{E}$ denote the discretized cost and state-equation operators respectively. Throughout this paper, we will use Sans Serif font to denote discretized quantities. The equality constraint 3b corresponds to the discretization of IVP 2 by some time stepping scheme, which in turn informs the discretization of the state and control vectors; we give more details in section 2.1 when discussing the Runge-Kutta method.

If the mapping $\mathbf{y} \mapsto \mathbf{E}(\mathbf{y}, \mathbf{u})$ is invertible, then we can use the equality constraint 3b to express the state variable as a function of the control variable. In other words,

$$\mathbf{y} = \mathbf{y}(\mathbf{u}) := \mathbf{E}^{-1}(\mathbf{0}, \mathbf{u}).$$

which allows us to reformulate 3 as an uncontrained optimization problem with *reduced* cost function

$$\widetilde{\mathsf{C}}(\mathbf{u}) := \mathsf{C}(\mathbf{y}(\mathbf{u}), \mathbf{u}).$$

Using implicit differentiation, one can show that the gradient of the reduced cost function is given by

$$\nabla \widetilde{\mathsf{C}}(\mathbf{u}) = \nabla_{\mathbf{u}} \mathsf{C}(\mathbf{y}, \mathbf{u}) - \left(\frac{\partial \mathbf{E}}{\partial \mathbf{u}}(\mathbf{y}, \mathbf{u}) \right)^{\top} \boldsymbol{\lambda} \tag{4}$$

where $\mathbf{y}$ must satisfy the state equation 3b, while the adjoint-state (or co-state)

vector $\boldsymbol{\lambda}$ satisfies what is known as the *adjoint equation*,

$$\left(\frac{\partial \mathbf{E}}{\partial \mathbf{y}}(\mathbf{y},\mathbf{u})\right)^{\top} \boldsymbol{\lambda} = \nabla_{\mathbf{y}} \mathsf{C}(\mathbf{y},\mathbf{u}). \tag{5}$$

Equations 4 and 5 can also be derived from standard optimization theory via Lagrange multipliers.

Equations 4 and 5 show that gradient computations of the cost function hinge on the linearization and adjoint of the state-equation operator, i.e., the choice of time integrator. Before we derive the linearization and adjoint of the relaxation Runge-Kutta method, we present the well understood updates for standard Runge-Kutta. For the majority of the paper, we drop the control vector $\mathbf{u}$ and simply focus on the state-equation operator $\mathbf{E}(\mathbf{y})$ (its linearization and adjoint) associated with discretizations of IVP 2.

2.1. Linearizations and adjoints of Runge-Kutta methods

A generic s-stage *Runge-Kutta* (RK) method, specified by its coefficients

$$\mathbf{A}_s := \begin{pmatrix} a_{11}, & a_{12} & \cdots & a_{1s} \\ a_{21}, & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \vdots \\ a_{s1} & a_{s2} & \cdots & a_{ss} \end{pmatrix}, \quad \mathbf{b}_s := \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{pmatrix}, \quad \mathbf{c}_s := \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_s \end{pmatrix},$$

applied to IVP 2, yields the following time-stepping formulas:

$$\mathbf{y}_k = \mathbf{y}_{k-1} + \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i}, \tag{6a}$$

$$\mathbf{Y}_{k,i} = \mathbf{y}_{k-1} + \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{F}_{k,j}, \quad \text{for} \quad i = 1, ..., s, \tag{6b}$$

where

$$\mathbf{F}_{k,i} := \mathbf{f}(\mathbf{Y}_{k,i}, t_{k-1} + c_i \Delta t).$$

and $\mathbf{y}_k$ is an approximation to the solution at time $t_k = t_{k-1} + \Delta t$. The $\mathbf{Y}_{k,i}$ are referred to as the RK internal stages. Assuming the RK method takes a total of K steps, we define vector $\overline{\mathbf{y}}$ to be the concatenation of all $\mathbf{y}_k$ and $\mathbf{Y}_{k,i}$. Let

$$\mathbf{Y}_k := \begin{pmatrix} \mathbf{Y}_{k,1} \\ \vdots \\ \mathbf{Y}_{k,s} \end{pmatrix} \in \mathbb{R}^{sN},$$

then

$$\overline{\mathbf{y}} := \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{Y}_1 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{Y}_K \\ \mathbf{y}_K \end{pmatrix} \in \mathbb{R}^{\overline{N}} \tag{7}$$

where $\overline{N} := N + N(s+1)K$. Throughout the remainder of this paper we will use an overline and bold Sans Serif font to denote vectors of dimension $\overline{N}$, whose components are grouped and denoted in a similar fashion to what was presented for $\overline{\mathbf{y}}$ in equation 7. For example, $\overline{\mathbf{w}} \in \mathbb{R}^{\overline{N}}$ is interpreted as

$$\overline{\mathbf{w}} = \begin{pmatrix} \mathbf{w}_0 \\ \mathbf{W}_1 \\ \mathbf{w}_1 \\ \vdots \\ \mathbf{W}_K \\ \mathbf{w}_K \end{pmatrix}$$

with $\mathbf{w}_k \in \mathbb{R}^N$ and

$$\mathbf{W}_k = \begin{pmatrix} \mathbf{W}_{k,1} \\ \vdots \\ \mathbf{W}_{k,s} \end{pmatrix} \in \mathbb{R}^{sN}.$$

It can be shown that the discrete state equation 3b, associated with an RK discretization, is of the form

$$\mathbf{E}(\overline{\mathbf{y}}) := \mathbf{L}\overline{\mathbf{y}} - \mathbf{N}(\overline{\mathbf{y}}) - \boldsymbol{\chi}_0 \mathbf{y}_{\text{init}} = \overline{\mathbf{0}} \tag{8}$$

where $\mathbf{L}$ is a lower unit-triangular matrix, and $\mathbf{N}$ is a block lower-triangular (potentially nonlinear) operator acting on $\overline{\mathbf{y}}$ that involves the evaluation of the right-hand-side function $\mathbf{f}$ at internal stages. The matrix $\boldsymbol{\chi}_k \in \mathbb{R}^{\overline{N} \times N}$ is defined such that $\boldsymbol{\chi}_k^\top \overline{\mathbf{y}} = \mathbf{y}_k$. Thus, for $\mathbf{v} \in \mathbb{R}^N$, we have that $\overline{\mathbf{v}} = \boldsymbol{\chi}_k \mathbf{v}$ is a vector of length $\overline{N}$ with $\mathbf{v}_k = \mathbf{v}$ and zeros everywhere else. In other words, the term $\boldsymbol{\chi}_0 \mathbf{y}_{\text{init}}$ accounts for the initial condition. We refer to the appendix for more details concerning this "matrix representation" of the RK method, which borrows notation from [4], and is used to derive the time-stepping formulas for the discrete linearization and adjoint of RK and its relaxation variant. Note that the inverse mapping of $\mathbf{E}$ is guaranteed to exist (barring any issues from the potentially nonlinear function $\mathbf{f}$) since $\mathbf{E}$ will be a block lower-triangular operator. In fact, time-stepping formulas 6 specify this inverse map, solving the discrete state equation by forward substitution.

We provide the linearization of standard RK formulas 6 as a point of reference for our discussion of adjoint computations for relaxation RK, which are

derived by computing the Jacobian of the discrete state operator $\mathbf{E}(\overline{\mathbf{y}})$ in equation 8 with respect to $\overline{\mathbf{y}}$. The linearized RK time-stepping formulas are

$$\boldsymbol{\delta}_k = \boldsymbol{\delta}_{k-1} + \Delta t \sum_{i=1}^{s} b_i \mathbf{J}_{k,i} \boldsymbol{\Delta}_{k,i} + \mathbf{w}_k \tag{9a}$$

$$\boldsymbol{\Delta}_{k,i} = \boldsymbol{\delta}_{k-1} + \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{J}_{k,j} \boldsymbol{\Delta}_{k,j} + \mathbf{W}_{k,i}, \quad 1, ..., s, \tag{9b}$$

where

$$\mathbf{J}_{k,i} := \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\mathbf{Y}_{k,i}, t_{k-1} + c_i \Delta t).$$

We refer to $\boldsymbol{\delta}_k$ and $\boldsymbol{\Delta}_k$ as the *linearized RK approximation and linearized RK internal stages*, respectively. These equations represent the solution to the following block lower-triangular system via forward substitution,

$$\mathbf{E}'(\overline{\mathbf{y}})\overline{\boldsymbol{\delta}} = \overline{\mathbf{w}}$$

for some given right-hand-side vector $\overline{\mathbf{w}} \in \mathbb{R}^{\overline{N}}$, which can be used to incorporate initial conditions and/or source terms. Note that $\mathbf{J}_{k,i}$ is the Jacobian matrix of $\mathbf{f}$ evaluated at the i-th internal stage $\mathbf{Y}_{k,i}$. In the case that $\mathbf{f}(\mathbf{y}(t), t)$ is linear in $\mathbf{y}$ we can expect $\mathbf{E}'(\overline{\mathbf{y}})\overline{\mathbf{y}} = \mathbf{E}(\overline{\mathbf{y}})$.

The adjoint RK formulas follow from solving

$$\mathbf{E}'(\overline{\mathbf{y}})^\top \overline{\boldsymbol{\lambda}} = \overline{\mathbf{w}},$$

a block upper-triangular system, by back substitution:

$$\boldsymbol{\lambda}_{k-1} = \boldsymbol{\lambda}_k + \sum_{i=1}^{s} \boldsymbol{\Lambda}_{k,i} + \mathbf{w}_{k-1}, \tag{10a}$$

$$\boldsymbol{\Lambda}_{k,i} = \Delta t\, \mathbf{J}_{k,i}^\top \Big(b_i \boldsymbol{\lambda}_k + \sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} \Big) + \mathbf{W}_{k,i}, \quad i = 1, ..., s. \tag{10b}$$

In the adjoint formulas 10, right-hand-side vector $\overline{\mathbf{w}} \in \mathbb{R}^{\overline{N}}$ incorporates source terms as well as a final time condition. We refer to $\boldsymbol{\lambda}_k$ and $\boldsymbol{\Lambda}_k$ as the *adjoint RK approximation* and *adjoint RK internal stages*, respectively. Note that the adjoint update formula uses $\boldsymbol{\lambda}_k$ to update $\boldsymbol{\lambda}_{k-1}$, in other words, we are marching backwards in time. We also point out that this presentation of the adjoint RK method is based on the derivation from the matrix representation, unlike other presentations that reformulate the equations above to resemble an RK method; see [10, 12].

2.2. Relaxation RK methods

Let $\eta : \mathbb{R}^N \to \mathbb{R}$ denote the *entropy* function (smooth and convex) associated with IVP 2, where the time evolution of the entropy is given by

$$\frac{d\eta}{dt}(\mathbf{y}(t)) = \nabla\eta(\mathbf{y}(t))^\top \mathbf{f}(\mathbf{y}(t), t).$$

IVP 2 is said to be *entropy dissipative* if

$$\nabla\eta(\mathbf{y}(t))^\top \mathbf{f}(\mathbf{y}(t), t) \leq 0$$

or *entropy conservative* if

$$\nabla\eta(\mathbf{y}(t))^\top \mathbf{f}(\mathbf{y}(t), t) = 0.$$

In the discrete setting, we wish to ensure

$$\begin{aligned} \eta(\mathbf{y}_{k+1}) &\leq \eta(\mathbf{y}_k), \quad \text{(for entropy dissipative)}, \\ \text{or} \quad \eta(\mathbf{y}_k) &= \eta(\mathbf{y}_0), \quad \text{(for entropy conservative)}. \end{aligned}$$

The relaxation RK method achieves discrete entropy conservation/dissipation by modifying the update step as follows:

$$\mathbf{y}_k = \mathbf{y}_{k-1} + \gamma_k \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i} \tag{11}$$

where γ_k, the *relaxation parameter*, is the non-zero root (closest to one) of the nonlinear scalar function $r_k(\,\cdot\,; \overline{\mathbf{y}})$,

$$r_k(\gamma; \overline{\mathbf{y}}) := \eta\left(\mathbf{y}_{k-1} + \gamma\Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i}\right) - \eta(\mathbf{y}_{k-1}) - \gamma\Delta t \sum_{i=1}^{s} b_i \nabla\eta(\mathbf{Y}_{k,i})^\top \mathbf{F}_{k,i}. \tag{12}$$

Internal stages $\mathbf{Y}_{k,i}$ are calculated as in equation 6b. After computing γ_k, there are two options for determining the solution at the next time step:

i. *Incremental direction technique* (IDT) method: $\mathbf{y}_k \approx \mathbf{y}(t_{k-1} + \Delta t)$
ii. *Relaxation RK* (RRK) method: $\mathbf{y}_k \approx \mathbf{y}(t_{k-1} + \gamma_k \Delta t)$

Implementation wise, both IDT and RRK require the solution to a scalar root problem at each time step, though RRK resembles more of an adaptive time-stepping scheme given that $t_k = t_{k-1} + \gamma_k \Delta t$ at the moment the RRK approximation is updated. In terms of accuracy, it was shown in [2] that RRK methods preserve accuracy of the underlying RK scheme. On the other hand, IDT schemes are order $p - 1$ accurate for an underlying method of order p. For the purposes of a clean presentation, we first derive the linearized and adjoint formulas for the simpler IDT method, and then extend to RRK with some minor modifications.

2.2.1. Discrete linearization and adjoint of IDT

The state-equation operator, $\mathbf{E}$, associated with IDT is modified in accordance to equation 11 by adding a dependency on the relaxation parameter vector $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_K)^\top \in \mathbb{R}^K$. Assuming IDT takes K steps, where K is the number of steps taken by the underlying RK method, we have

$$\mathbf{E}(\overline{\mathbf{y}}, \boldsymbol{\gamma}) := \mathbf{L}\overline{\mathbf{y}} - \mathbf{N}(\overline{\mathbf{y}}, \boldsymbol{\gamma}) - \boldsymbol{\chi}_0 \mathbf{y}_{\text{init}}.$$

The relaxation parameter γ_k is defined by the solution to a root equation at each time step. This root equation can be written in vector form as

$$\mathbf{r}(\boldsymbol{\gamma}; \overline{\mathbf{y}}) := \begin{pmatrix} r_1(\gamma_1; \overline{\mathbf{y}}) \\ r_2(\gamma_2; \overline{\mathbf{y}}) \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \end{pmatrix}.$$

Note that the equation above implicitly defines the relaxation parameter vector as a function of $\overline{\mathbf{y}}$, i.e., $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\overline{\mathbf{y}})$. Using this, we denote the *reduced* state-equation operator by

$$\widetilde{\mathbf{E}}(\overline{\mathbf{y}}) := \mathbf{E}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}})).$$

A proper linearization of the IDT method will require the Jacobian of the reduced state-equation operator,

$$\widetilde{\mathbf{E}}'(\overline{\mathbf{y}}) = \frac{\partial \mathbf{E}}{\partial \overline{\mathbf{y}}}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}})) + \frac{\partial \mathbf{E}}{\partial \boldsymbol{\gamma}}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}}))\, \boldsymbol{\gamma}'(\overline{\mathbf{y}}).$$

Unfortunately, one cannot directly compute $\boldsymbol{\gamma}'$ since the relaxation parameters are computed numerically using some iterative root-solving algorithm. One could bypass this issue by simply taking $\partial \mathbf{E}/\partial \overline{\mathbf{y}}$ as the Jacobian, ignoring the second term above, essentially viewing the relaxation parameter as a constant in the linearization, as suggested by [15, 16]. This approach would result in linearized and adjoint time-stepping formulas that are almost identical to RK, equations 9a and 10a (and equation 10b for the adjoint internal stages), but with weights $\mathbf{b}_s \mapsto \gamma_k \mathbf{b}_s$. We will show in our numerical results that ignoring the relaxation parameter will have negative consequences.

For a proper linearization, we compute $\boldsymbol{\gamma}'$ via implicit differentiation. Note that γ_k is dependent on $\mathbf{y}_{k-1}$ and $\mathbf{Y}_k$ only, thus it suffices to compute partial derivatives with respect to these variables. In particular,

$$(\nabla_y \gamma_k)^\top := \frac{\partial \gamma_k}{\partial \mathbf{y}_{k-1}}(\overline{\mathbf{y}}) = -\left(\frac{\partial r_k}{\partial \gamma}\right)^{-1} \frac{\partial r_k}{\partial \mathbf{y}_{k-1}}\bigg|_{(\boldsymbol{\gamma}(\overline{\mathbf{y}}); \overline{\mathbf{y}})} \in \mathbb{R}^{1 \times N} \tag{13a}$$

$$(\nabla_Y \gamma_{k,i})^\top := \frac{\partial \gamma_k}{\partial \mathbf{Y}_{k,i}}(\overline{\mathbf{y}}) = -\left(\frac{\partial r_k}{\partial \gamma}\right)^{-1} \frac{\partial r_k}{\partial \mathbf{Y}_{k,i}}\bigg|_{(\boldsymbol{\gamma}(\overline{\mathbf{y}}); \overline{\mathbf{y}})} \in \mathbb{R}^{1 \times N}, \tag{13b}$$

$$(\nabla_Y \gamma_k)^\top = \left((\nabla_Y \gamma_{k,1})^\top, ..., (\nabla_Y \gamma_{k,s})^\top\right) \in \mathbb{R}^{1 \times sN}, \tag{13c}$$

where r_k, again, is defined in 12 and

$$\frac{\partial r_k}{\partial \gamma}(\boldsymbol{\gamma}(\overline{\mathbf{y}});\overline{\mathbf{y}}) = \Delta t \sum_{i=1}^{s} b_i \Big(\nabla\eta(\mathbf{y}_k) - \nabla\eta(\mathbf{Y}_{k,i})\Big)^\top \mathbf{F}_{k,i} \tag{14a}$$

$$\frac{\partial r_k}{\partial \mathbf{y}_{k-1}}(\boldsymbol{\gamma}(\overline{\mathbf{y}});\overline{\mathbf{y}}) = \nabla\eta(\mathbf{y}_k)^\top - \nabla\eta(\mathbf{y}_{k-1})^\top \tag{14b}$$

$$\frac{\partial r_k}{\partial \mathbf{Y}_{k,j}}(\boldsymbol{\gamma}(\overline{\mathbf{y}});\overline{\mathbf{y}}) = \gamma_k b_j \Delta t \left\{ \Big(\nabla\eta(\mathbf{y}_k) - \nabla\eta(\mathbf{Y}_{k,j})\Big)^\top \mathbf{J}_{k,j} - \mathbf{F}_{k,j}^\top \nabla^2\eta(\mathbf{Y}_{k,j}) \right\}, \quad j = 1, ..., s. \tag{14c}$$

We provide the remaining details in the appendix, and simply state the resulting time-stepping formulas in the following lemmas.

Lemma 1. *The linearized IDT time-stepping formulas are*

$$\boldsymbol{\delta}_k = \boldsymbol{\delta}_{k-1} + \gamma_k \Delta t \sum_{i=1}^{s} b_i \mathbf{J}_{k,i} \boldsymbol{\Delta}_{k,i} + \rho_k \, \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i} + \mathbf{w}_k \tag{15}$$

where

$$\rho_k = (\nabla_y \gamma_k)^\top \boldsymbol{\delta}_{k-1} + (\nabla_Y \gamma_k)^\top \boldsymbol{\Delta}_k$$

and the computation of the internal stages $\boldsymbol{\Delta}_k$ is the same as for RK (see equation 9b).

Lemma 2. *The adjoint IDT time-stepping formulas are*

$$\boldsymbol{\lambda}_{k-1} = \boldsymbol{\lambda}_k + \sum_{i=1}^{s} \boldsymbol{\Lambda}_{k,i} + \xi_k \nabla_y \gamma_k + \mathbf{w}_{k-1} \tag{16}$$

where

$$\xi_k = \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i}^\top \boldsymbol{\lambda}_k$$

$$\boldsymbol{\Lambda}_{k,i} = \Delta t \, \mathbf{J}_{k,i}^\top \Big(\gamma_k b_i \boldsymbol{\lambda}_k + \sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j}\Big) + \xi_k \nabla_Y \gamma_{k,i} + \mathbf{W}_{k,i}. \tag{17}$$

As we show in the appendix section, linearized and adjoint IDT formulas define solutions to the following matrix systems, respectively:

$$\left(\frac{\partial \mathbf{E}}{\partial \overline{\mathbf{y}}} + \frac{\partial \mathbf{E}}{\partial \boldsymbol{\gamma}} \boldsymbol{\gamma}'\right) \overline{\boldsymbol{\delta}} = \overline{\mathbf{w}}, \quad \text{(linearized)}$$

$$\left(\frac{\partial \mathbf{E}}{\partial \overline{\mathbf{y}}}^\top + (\boldsymbol{\gamma}')^\top \frac{\partial \mathbf{E}}{\partial \boldsymbol{\gamma}}^\top\right) \overline{\boldsymbol{\lambda}} = \overline{\mathbf{w}}, \quad \text{(adjoint)}.$$

Terms associated with the linearization of γ_k in the linearized/adjoint IDT up-

date formulas above, and in Lemmas 1 and 2, are highlighted using red text. We see that taking into account the relaxation parameter in the linearization results in having to compute gradients $\nabla_y \gamma_k$ and $\nabla_Y \gamma_k$ as well as scalars ρ_k and ξ_k at each time step. These gradients in turn require RK approximations at two time steps ($\mathbf{y}_{k-1}$ and $\mathbf{y}_k$), internal stages $\mathbf{Y}_k$, and the evaluation of the right-hand-side function $\mathbf{f}$ and its Jacobian, as well as the gradient and Hessian of the entropy function.

2.2.2. Time-symmetry of IDT

Before moving on to the discrete RRK adjoint, we discuss the special case where

$$\mathbf{f}(\mathbf{y}(t), t) = \mathbf{S}\mathbf{y}(t)$$

for a skew-symmetric matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ in IVP 2. It can be shown that IVP 2 is entropy/energy conservative with respect to the square entropy

$$\eta(\mathbf{y}(t)) = \frac{1}{2}\|\mathbf{y}(t)\|^2.$$

Of course, given that $\mathbf{f}(\mathbf{y})$ is linear in this case, it follows that linearization of the continuous state equation coincides with the original problem, ignoring initial conditions. This is not obvious but still true for the IDT formulas, even though the relaxation parameter is nonlinear with respect to $\overline{\mathbf{y}}$. In other words, discretization by an IDT method commutes with linearization in this special case.

Theorem 1. *Suppose we were to apply IDT and the linearized IDT to IVP 2 with $\mathbf{f}(\mathbf{y}(t), t) = \mathbf{S}\mathbf{y}(t)$, where $\mathbf{S}$ is skew-symmetric and the entropy function is $\eta(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|^2$. It follows that IDT updates are equivalent to linearized IDT updates. In particular, if linearized IDT is applied with $\boldsymbol{\delta}_0 = \mathbf{y}_{\text{init}}$, $\mathbf{w}_k = \mathbf{0}$ and $\mathbf{W}_k = \mathbf{0}$, then $\boldsymbol{\delta}_k = \mathbf{y}_k$ for all time steps.*

Proof. Before proving equivalence between IDT and linearized IDT, we first make some observations. In this linear case, we have

$$\mathbf{F}_{k,i} = \mathbf{S}\mathbf{Y}_{k,i}, \quad \text{and} \quad \mathbf{J}_{k,i} = \mathbf{S}.$$

Moreover, the root function for computing the relaxation parameter is quadratic in γ,

$$r_k(\gamma; \overline{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y}_{k-1} + \gamma \mathbf{d}_k\|^2 - \frac{1}{2}\|\mathbf{y}_{k-1}\|^2 - \gamma \mathbf{e}_k$$

where

$$\mathbf{d}_k := \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i} = \Delta t \sum_{i=1}^{s} b_i \mathbf{S}\mathbf{Y}_{k,i}, \tag{18}$$

$$\mathbf{e}_k := \Delta t \sum_{i=1}^{s} b_i \nabla \eta(\mathbf{Y}_{k,i})^\top \mathbf{F}_{k,i}.$$

The vectors $\mathbf{d}_k$ and $\mathbf{e}_k$ are based on notation from [2] and represent the search direction (based on a projection interpretation of relaxation methods) and the entropy production at the current time step, respectively. Note that $\mathbf{e}_k = 0$ since $\nabla\eta(\mathbf{y}) = \mathbf{y}$ and

$$\nabla\eta(\mathbf{Y}_{k,i})^\top \mathbf{F}_{k,i} = \mathbf{Y}_{k,i}^\top \mathbf{S}\mathbf{Y}_{k,i} = 0$$

by skew-symmetry of $\mathbf{S}$.

Since $r_k(\gamma_k; \bar{\mathbf{y}}) = 0$ is nothing more than a quadratic root problem we can come up with an explicit formula for γ_k, the non-zero root of this quadratic function:

$$\gamma_k := -\frac{2\mathbf{y}_{k-1}^\top \mathbf{d}_k}{\|\mathbf{d}_k\|^2}, \tag{19}$$

which is consistent with what is reported in [1]. Furthermore, the gradients with respect to $\mathbf{y}_{k-1}$ and $\mathbf{Y}_{k,i}$ are given by

$$\nabla_y \gamma_k = -\frac{2}{\|\mathbf{d}_k\|^2}\mathbf{d}_k, \tag{20}$$

$$\nabla_Y \gamma_{k,i} = -\frac{2b_i\Delta t}{\|\mathbf{d}_k\|^2}\mathbf{S}^\top \Big(\underbrace{\mathbf{y}_{k-1} - \frac{2\mathbf{y}_{k-1}^\top \mathbf{d}_k}{\|\mathbf{d}_k\|^2}\mathbf{d}_k}_{\mathbf{y}_k = \mathbf{y}_{k-1} + \gamma_k \mathbf{d}_k} \Big) = -\frac{2b_i\Delta t}{\|\mathbf{d}_k\|^2}\mathbf{S}^\top \mathbf{y}_k. \tag{21}$$

We now prove that the k-th step of the linearized IDT method is equivalent to the k-th IDT step, assuming $\boldsymbol{\delta}_{k-1} = \mathbf{y}_{k-1}$. First note that the internal stages coincide, $\boldsymbol{\Delta}_k = \mathbf{Y}_k$, since their update formulas are identical, see equations 6b and 9b. This follows from $\boldsymbol{\delta}_{k-1} = \mathbf{y}_{k-1}$, $\mathbf{J}_{k,j} = \mathbf{S}$ and $\mathbf{F}_{k,j} = \mathbf{S}\mathbf{Y}_{k,j}$. Next we look at the update step for linearized IDT, equation 15. In particular,

$$\nabla_y \gamma_k^\top \boldsymbol{\delta}_{k-1} = -\frac{2}{\|\mathbf{d}_k\|^2}\mathbf{d}_k^\top \mathbf{y}_{k-1} = \gamma_k$$

and

$$\begin{aligned}
\nabla_Y \gamma_k^\top \boldsymbol{\Delta}_k &= -\frac{2}{\|\mathbf{d}_k\|^2}\mathbf{y}_k^\top \left(\Delta t \sum_{i=1}^{s} b_i \mathbf{S}\mathbf{Y}_{k,i} \right) \\
&= -\frac{2}{\|\mathbf{d}_k\|^2}\mathbf{y}_k^\top \mathbf{d}_k \\
&= -\frac{2}{\|\mathbf{d}_k\|^2}(\mathbf{y}_{k-1} + \gamma_k \mathbf{d}_k)^\top \mathbf{d}_k \\
&= -\gamma_k,
\end{aligned}$$

which implies that

$$\rho_k = \nabla_y \gamma_k^\top \boldsymbol{\delta}_{k-1} + \nabla_Y \gamma_k^\top \boldsymbol{\Delta}_k = 0.$$

Using $\boldsymbol{\delta}_{k-1} = \mathbf{y}_{k-1}$, $\boldsymbol{\Delta}_{k,i} = \mathbf{Y}_{k,i}$, and $\rho_k = 0$, we can conclude that the k-th step is given by

$$\begin{aligned}\boldsymbol{\delta}_k &= \boldsymbol{\delta}_{k-1} + \gamma_k \underbrace{\Delta t \sum_{i=1}^{s} b_i \mathbf{S}\boldsymbol{\Delta}_{k,i}}_{\mathbf{d}_k} + \cancel{\rho_k} \Delta t \sum_{i=1}^{s} b_i \mathbf{S}\mathbf{Y}_{k,i} \\ &= \mathbf{y}_{k-1} + \gamma_k \mathbf{d}_k \\ &= \mathbf{y}_k.\end{aligned}$$

Assuming $\boldsymbol{\delta}_0 = \mathbf{y}_0 = \mathbf{y}_{\text{init}}$, we can conclude by induction that $\boldsymbol{\delta}_k = \mathbf{y}_k$ and $\boldsymbol{\Delta}_k = \mathbf{Y}_k$ for all time steps. □

Building off of the equivalence of the forward and linearized continuous systems, and similarly for IDT and linearized IDT, we discuss an interesting relationship with their respective adjoints. The adjoint of IVP 2, again with $\mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{y}$, is given by

$$-\lambda'(t) = \mathbf{S}^\top \lambda(t), \quad 0 < t < T \tag{22a}$$

$$\lambda(T) = \boldsymbol{\lambda}_{\text{final}} \tag{22b}$$

along with the following adjoint condition

$$\boldsymbol{\lambda}_{\text{final}}^\top \mathbf{y}(T) = \lambda(0)^\top \mathbf{y}_{\text{init}}. \tag{23}$$

For skew-symmetric $\mathbf{S}$, system 22 is essentially the original problem but with a final time condition. In particular, if $\boldsymbol{\lambda}_{\text{final}} = \mathbf{y}(T)$, then system 22 is equivalent to the forward problem in reverse time, i.e., $\lambda(t) = \mathbf{y}(t)$ for all time $0 \le t \le T$. Note that condition 23 is automatically satisfied here since the problem is norm conservative,

$$\boldsymbol{\lambda}_{\text{final}}^\top \mathbf{y}(T) = \|\mathbf{y}(T)\|^2 = \|\mathbf{y}(0)\|^2 = \lambda(0)^\top \mathbf{y}_{\text{init}}.$$

We refer to this equivalency between the forward and adjoint systems as a *time-symmetry* property of the continuous problem. In Theorem 2 we prove that IDT preserves a similar time-symmetry with its adjoint.

Theorem 2. *Assume the underlying RK method is explicit or diagonally implicit. Suppose we were to apply IDT to IVP 2 with* $\mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{y}$*, where* $\mathbf{S}$ *is skew-symmetric and the entropy function is* $\eta(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|^2$*. It follows that the IDT method preserves time-symmetry. In particular, if adjoint IDT is applied with* $\boldsymbol{\lambda}_K = \mathbf{y}_K$ *(assuming IDT takes* K *steps),* $\mathbf{w}_k = \mathbf{0}$ *and* $\mathbf{W}_k = \mathbf{0}$*, then* $\boldsymbol{\lambda}_k = \mathbf{y}_k$ *for all time steps, up to machine precision.*

Proof. This proof makes use of some simplifications derived in the first half of the proof for Theorem 1, based on the linearity of the problem. In particular, we make use of $\mathbf{F}_{k,i} = \mathbf{S}\mathbf{Y}_{k,i}$, $\mathbf{J}_{k,i} = \mathbf{S}$, $\mathbf{d}_k$ as defined by equation 18, the explicit formula for γ_k given by equation 19, and gradients of γ_k (equations 20 and 21).

Consider the k-th step (in reverse time) of the adjoint IDT algorithm, assuming $\boldsymbol{\lambda}_k = \mathbf{y}_k$. Note that

$$\xi_k = \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i}^{\top} \boldsymbol{\lambda}_k = \mathbf{d}_k^{\top} \mathbf{y}_k.$$

The internal stages are given by

$$\begin{aligned}
\boldsymbol{\Lambda}_{k,i} &= \Delta t \mathbf{J}_{k,i}^{\top} \left(\gamma_k b_i \boldsymbol{\lambda}_k + \sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} \right) + \xi_k \Big(\nabla_Y \gamma_{k,i} \Big) \\
&= \Delta t \mathbf{S}^{\top} \left(\gamma_k b_i \mathbf{y}_k + \sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} \right) + \left(\mathbf{d}_k^{\top} \mathbf{y}_k \right) \left(-\frac{2\Delta t \; b_i}{\|\mathbf{d}_k\|^2} \mathbf{S}^{\top} \mathbf{y}_k \right) \\
&= \Delta t \mathbf{S}^{\top} \left(\sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} \right) + \gamma_k b_i \Delta t \mathbf{S}^{\top} \mathbf{y}_k - \gamma_k b_i \Delta t \mathbf{S}^{T} \mathbf{y}_k \\
&= \Delta t \mathbf{S}^{\top} \left(\sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} \right),
\end{aligned}$$

where we made use of

$$\begin{aligned}
-\frac{2}{\|\mathbf{d}_k\|^2} \mathbf{d}_k^{\top} \mathbf{y}_k &= -\frac{2}{\|\mathbf{d}_k\|^2} \mathbf{d}_k^{\top} (\mathbf{y}_{k-1} + \gamma_k \mathbf{d}_k) \\
&= \gamma_k - 2\gamma_k \\
&= -\gamma_k,
\end{aligned}$$

Recall that the RK coefficient matrix $\mathbf{A}_s$ is lower triangular for explicit or diagonally implicit RK method. Thus, we have the following set of equations for the internal stages,

$$\left(\mathbf{I} - a_{ii} \Delta t \mathbf{S}^{\top} \right) \boldsymbol{\Lambda}_{k,i} = \Delta t \mathbf{S}^{\top} \sum_{j>i} a_{ji} \boldsymbol{\Lambda}_{k,j}, \quad i = 1, ..., s.$$

It is easy to see, that for $i = s$,

$$\left(\mathbf{I} - a_{ss} \Delta t \mathbf{S}^{\top} \right) \boldsymbol{\Lambda}_{k,s} = \mathbf{0} \quad \Longrightarrow \quad \boldsymbol{\Lambda}_{k,s} = \mathbf{0}.$$

Moreover, by induction, it can be shown that $\boldsymbol{\Lambda}_{k,i} = \mathbf{0}$ for all $i = 1, ..., s$.

Since the internal stages zero-out, the k-th adjoint update step reduces to

$$\begin{aligned}\boldsymbol{\lambda}_{k-1} &= \boldsymbol{\lambda}_k + \sum_{i=1}^{s} \cancel{\boldsymbol{\Lambda}_{k,i}} + \xi_k \nabla_y \gamma_k \\ &= \boldsymbol{\lambda}_k + \left(\mathbf{d}_k^\top \mathbf{y}_k\right)\left(-\frac{2}{\|\mathbf{d}_k\|^2}\mathbf{d}_k\right) \\ &= \mathbf{y}_k - \gamma_k \mathbf{d}_k \\ &= \mathbf{y}_{k-1}.\end{aligned}$$

Assuming $\boldsymbol{\lambda}_K = \mathbf{y}_K$, we can conclude by induction that $\boldsymbol{\lambda}_k = \mathbf{y}_k$ for all subsequent time steps. □

Again, we emphasize that our notion of *adjoint* in this paper is based the transpose of the matrix-form of the linearized state-equations induced by the time-stepping scheme. Moreover, by *time-symmetry* we refer to the relationship between the forward and adjoint continuous system as well as the forward and adjoint IDT time-stepping formulas. In other words, for IDT, time-symmetry refers to the adjoint time-step as reversing the corresponding forward time-step. The use of adjoint and time-symmetry should not be equated with the standard definitions used in the literature for symplectic/geometric numerical integrators; see [17, 18] for other definitions of time-reversibility, time-symmetry, and adjoints of time-stepping methods.

2.2.3. Discrete linearization and adjoint of RRK

Recall that RRK yields the update $\mathbf{y}_k \approx \mathbf{y}(t_{k-1} + \gamma_k \Delta t)$. For this reason RRK can be interpreted as an adaptive time-stepping scheme. Consequently, special care is taken in order to ensure that at the end of the time loop we end up with an approximation at the desired final time. Suppose that at time step $K-1$ we have $t_{K-1} + \Delta t > T$ and $t_K = t_{K-1} + \gamma_K \Delta t > T$. In other words, the RRK time loop terminates after K steps. One could apply some continuation method to interpolate an approximation at $t = T$. This interpolation step will need to be entropy stable in some sense and be accounted for in the linearization and subsequently adjoint of the RRK method.

As an interpolation-free alternative, we take an IDT final step but with a corrected step size of $\Delta t^* := T - t_{K-1}$, resulting in $t_K = T$. In other words,

$$\begin{aligned}\mathbf{y}_K &= \mathbf{y}_{K-1} + \gamma_K \Delta t^* \sum_{i=1}^{s} b_i \mathbf{F}_{K,i}, \\ \mathbf{Y}_{K,i} &= \mathbf{y}_{K-1} + \Delta t^* \sum_{j=1}^{s} a_{ij} \mathbf{F}_{K,j}, \quad i = 1, ..., s,\end{aligned}$$

where $\mathbf{y}_K \approx \mathbf{y}(T)$. This approach is similar to what is considered in [15, 16] for

generic adaptive time-stepping methods. Note that

$$\Delta t^* = T - t_0 - \Delta t \sum_{\ell=1}^{K-1} \gamma_\ell, \tag{24}$$

which implies that Δt^* is dependent on γ_ℓ (and thus $\mathbf{y}_{\ell-1}$ and $\mathbf{Y}_\ell$) for $\ell = 1, ..., K-1$. Moreover, since Δt^* is the step size used in r_K this implies γ_K is dependent on $(\mathbf{y}_{\ell-1}, \mathbf{Y}_\ell)$ for $\ell = 1, ..., K$. These dependencies must be taken into account for a proper linearization. Again, we provide the derivation on the appendix and summarize the results here.

Lemma 3. *Assuming RRK takes K steps, and that the last step is taken as an IDT step, with $\Delta t^* = T - t_{K-1}$, then the linearized RRK approximations and internal stages, $\boldsymbol{\delta}_k$ and $\boldsymbol{\Delta}_k$ for $k = 1, ..., K-1$, are as given by the linearized IDT computations in equations 15 and 9b, respectively. The linearized RRK update formula for the final step is given by*

$$\boldsymbol{\delta}_K = \boldsymbol{\delta}_{K-1} + \gamma_K \Delta t^* \sum_{i=1}^{s} b_i \mathbf{J}_{K,i} \boldsymbol{\Delta}_{K,i} + \rho_K \Delta t^* \sum_{i=1}^{s} b_i \mathbf{F}_{K,i} + \mathbf{w}_K$$

where

$$\boldsymbol{\Delta}_{K,i} = \boldsymbol{\delta}_{K-1} + \Delta t^* \sum_{j=1}^{s} a_{ij} \mathbf{J}_{K,j} \boldsymbol{\Delta}_{K,j} - \rho_* \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{F}_{K,j} + \mathbf{W}_{K,i},$$

$$\rho_* = \sum_{k=1}^{K-1} \rho_k.$$

Lemma 4. *Assuming RRK takes K steps, and that the last step is taken as an IDT step with $\Delta t^* = T - t_{K-1}$, then the adjoint RRK approximations and internal stages for the first adjoint step, $\boldsymbol{\lambda}_{K-1}$ and $\boldsymbol{\Lambda}_K$, are given by the adjoint IDT formulas, equations 16 and 17 respectively, though with $\Delta t \mapsto \Delta t^*$. The adjoint RRK update formulas for $k = K-1, ..., 1$ are given by*

$$\boldsymbol{\lambda}_{k-1} = \boldsymbol{\lambda}_k + \sum_{i=1}^{s} \boldsymbol{\Lambda}_{k,i} + (\xi_k - \xi_*) \nabla_y \gamma_k + \mathbf{w}_{k-1}$$

where

$$\boldsymbol{\Lambda}_{k,i} = \Delta t \, \mathbf{J}_{k,i}^\top \Big(\gamma_k b_i \boldsymbol{\lambda}_k + \sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} \Big) + (\xi_k - \xi_*) \nabla_Y \gamma_{k,i} + \mathbf{W}_{k,i},$$

$$\xi_* = \Delta t \sum_{i=1}^{s} \sum_{j=1}^{s} a_{ji} \mathbf{F}_{K,i}^\top \boldsymbol{\Lambda}_{K,j}.$$

In relation to the matrix-form, linearized and adjoint IDT formulas define

solutions to the following matrix systems, respectively:

$$\left(\frac{\partial \mathbf{E}}{\partial \overline{\mathbf{y}}} + \frac{\partial \mathbf{E}}{\partial \boldsymbol{\gamma}}\boldsymbol{\gamma}' + \frac{\partial \mathbf{E}}{\partial \Delta t^*}\frac{d\Delta t^*}{d\overline{\mathbf{y}}}\right)\overline{\boldsymbol{\delta}} = \overline{\mathbf{w}}, \quad \text{(linearized)}$$

$$\left(\frac{\partial \mathbf{E}}{\partial \overline{\mathbf{y}}}^\top + (\boldsymbol{\gamma}')^\top \frac{\partial \mathbf{E}}{\partial \boldsymbol{\gamma}}^\top + \left(\frac{d\Delta t^*}{d\overline{\mathbf{y}}}\right)^\top \frac{\partial \mathbf{E}}{\partial \Delta t^*}^\top\right)\overline{\boldsymbol{\lambda}} = \overline{\mathbf{w}}, \quad \text{(adjoint)}.$$

Again, terms associated with linearization with respect to the relaxation parameter are in red. In blue we have terms related to linearization with respect to the final step size Δt^*. We note that accounting for Δt^* in the linearization requires computing the scalar quantity ρ_*, which can be done efficiently by simply accumulating the ρ_k scalars. This scalar shows up in the internal stage computations for the last time-step. For the adjoint RRK method, we see an expected structure that is complementary to linearized RRK. In particular, the final step is as given by the adjoint IDT formula (with $\Delta t \mapsto \Delta t^*$). The scalar ξ_*, computed from the final step, shows up as a correction term for the remaining time steps.

Consider again the special case when $\mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is skew-symmetric. The same results we observed for IDT hold for RRK as well.

Theorem 3. *Suppose we were to apply RRK, and linearized RRK to IVP 2 with $\mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is skew-symmetric. It follows that RRK updates are equivalent to linearized RRK updates. In particular, if linearized RRK is applied with $\boldsymbol{\delta}_0 = \mathbf{y}_{\text{init}}$, $\mathbf{w}_k = \mathbf{0}$ and $\mathbf{W}_k = \mathbf{0}$, then $\boldsymbol{\delta}_k = \mathbf{y}_k$ for all time steps.*

Proof. Given that the first $K-1$ steps of RRK are algorithmically identical to IDT, it follows from theorem 1 that $(\boldsymbol{\delta}_{k-1}, \boldsymbol{\Delta}_k) = (\mathbf{y}_{k-1}, \mathbf{Y}_k)$ for $k = 1, ..., K-1$. Moreover, in the proof of theorem 1, we showed that $\rho_k = 0$ for $k = 1, ..., K-1$ ($\rho_K = 0$ can similarly be proven), thus the accumulated scalar ρ_* in RRK is also zero. From this it is easy to see that the proposition holds. □

Theorem 4. *Assume the underlying RK method is explicit or diagonally implicit. Suppose we were to apply RRK to IVP 2 with $\mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is skew-symmetric. It follows that the RRK method preserves time-symmetry. In particular, if adjoint RRK is applied with $\boldsymbol{\lambda}_K = \mathbf{y}_K$ (assuming RRK takes K steps), $\mathbf{w}_k = \mathbf{0}$ and $\mathbf{W}_k = \mathbf{0}$, then $\boldsymbol{\lambda}_k = \mathbf{y}_k$ for all time steps, up to machine precision.*

Proof. In adjoint RRK, we have to account for the auxiliary scalar ξ_*,

$$\xi_* = \Delta t \sum_{ij} a_{ji} \mathbf{F}_{K,i}^\top \boldsymbol{\Lambda}_{K,j}.$$

Just as in the IDT case, the first set of internal stages, $\boldsymbol{\Lambda}_{K,i}$ for $i = 1, ..., s$, are all zero, hence $\xi_* = 0$. This allows us to carry on and show that $\boldsymbol{\lambda}_{k-1} = \mathbf{y}_{k-1}$ for all time steps. □

3. Numerical experiments and results

This section presents numerical results that validate and highlight properties discussed in the previous sections. We restrict our focus to the more interesting and applicable RRK method, though analogous results can be generated for IDT as well. We mention in passing that a bisection algorithm, similar to that of [19], is used to solve the scalar root subproblem when computing relaxation parameters at each time step.

The previous section presented the discrete linearization and adjoint of RRK, taking into consideration the relaxation parameter, γ, and the corrected final step-size, Δt^*. We demonstrate various consequences of improper linearization, when $\boldsymbol{\gamma}$ and Δt^* are not considered in the linearization; we refer to these cases as the γ-constant or Δt^*-constant case respectively. Recall that Δt^* can be written in terms of γ_ℓ for $\ell = 1, ..., K-1$ (see equation 24), hence, in the γ-constant case Δt^* is also considered constant.

3.1. Mathematical models

We consider three mathematical models throughout our numerical tests which we present at this moment.

Nonlinear pendulum

For the nonlinear pendulum, we take the first-order form as done in [2],

$$\frac{d}{dt}\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} -\sin(y_2(t)) \\ y_1(t) \end{pmatrix}, \quad 0 < t \leq T \tag{25}$$

with initial condition

$$\mathbf{y}(0) = \mathbf{y}_{\text{init}} := \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}.$$

This problem is entropy conservative with respect to the following entropy function,

$$\eta(\mathbf{y}) = \frac{1}{2}y_1^2 - \cos(y_2).$$

1D compressible Euler

The 1D compressible Euler equations are given by

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x} = 0, \tag{26a}$$

$$\frac{\partial(\rho u)}{\partial t} + \frac{\partial(\rho u^2 + p)}{\partial x} = 0, \tag{26b}$$

$$\frac{\partial E}{\partial t} + \frac{\partial(u(E+p))}{\partial x} = 0, \tag{26c}$$

which we solve over $0 < t \leq 1.5$ and $-1 \leq x \leq 1$, with initial condition

$$\begin{aligned}
&\rho(x,0) = 1 + \tfrac{1}{2}\exp(-50(x-0.1)^2),\\
&u(x,0) = 0,\\
&p(x,0) = \rho(x,0)^\gamma,
\end{aligned}$$

assuming the pressure, p, satisfies the following constitutive relation:

$$p = (\gamma - 1)(E - \tfrac{1}{2}\rho u^2), \quad \gamma = 1.4.$$

We arrive at IVP 2 by discretizing in space the compressible Euler equations with an entropy stable DG scheme [20, 21]. In particular, $\mathbf{y}(t)$ represents semi-discretized quantities related to the variables $\rho, \rho u$ and E. The entropy function associated with this semi-discretized system corresponds to a discretization of the continuous total entropy,

$$\eta(\mathbf{y}(t)) \approx \int S(\rho, \rho u, E)\, dx$$

where S is the continuous entropy function

$$S(\rho, \rho u, E) := -\frac{\rho s}{\gamma - 1}, \quad s = \log\left(\frac{p}{\rho^\gamma}\right).$$

Linear skew-symmetric system

For experiments concerning time-symmetry of RRK, as detailed in theorem 4, we solve the following linear problem:

$$\mathbf{y}'(t) = \mathbf{S}\mathbf{y}(t), \quad 0 < t < T, \tag{27a}$$

$$\mathbf{y}(0) = \mathbf{y}_{\text{init}} \tag{27b}$$

where $\mathbf{S} \in \mathbb{R}^{N\times N}$ is a randomly generated skew-symmetric matrix and $\mathbf{y}_{\text{init}} \in \mathbb{R}^N$ is a randomly generated initial condition; we take $N = 10$. The final time is taken to be proportional to the Frobenius norm of $\mathbf{S}$, specifically,

$$T = 10 \cdot \|\mathbf{S}\|_F \approx 133.46.$$

As previously mentioned, this system is entropy conservative with respect to

$$\eta(\mathbf{y}(t)) = \frac{1}{2}\|\mathbf{y}(t)\|^2.$$

3.2. RK schemes

The following are the RK schemes used throughout our experiments:

- Heun's method, a 2-stage, second-order RK scheme which we refer to as RK2,

$$\mathbf{A}_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \quad \mathbf{c}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

- A 3-stage, third-order RK scheme (see [22]), which we refer to as RK3,

$$\mathbf{A}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1/4 & 1/4 & 0 \end{pmatrix}, \quad \mathbf{b}_3 = \begin{pmatrix} 1/6 \\ 1/6 \\ 2/3 \end{pmatrix}, \quad \mathbf{c}_3 = \begin{pmatrix} 0 \\ 1 \\ 1/2 \end{pmatrix}.$$

- The standard 4-stage, fourth-order RK scheme, referred to as RK4,

$$\mathbf{A}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{b}_4 = \begin{pmatrix} 1/6 \\ 1/3 \\ 1/3 \\ 1/6 \end{pmatrix}, \quad \mathbf{c}_4 = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \\ 1 \end{pmatrix}.$$

- A 3-stage, third-order DIRK scheme, referred to as DIRK3, with

$$\mathbf{A}_3 = \begin{pmatrix} \alpha & 0 & 0 \\ \tau_2 - \alpha & \alpha & 0 \\ b_1 & b_2 & \alpha \end{pmatrix}, \quad \mathbf{b}_3 = \begin{pmatrix} b_1 \\ b_2 \\ \alpha \end{pmatrix}, \quad \mathbf{c}_3 = \begin{pmatrix} \alpha \\ \tau_2 \\ 1 \end{pmatrix},$$

  where

$$\begin{aligned} \alpha &= 0.435866521508459 \\ \tau_2 &= (1+\alpha)/2 \\ b_1 &= -(6\alpha^2 - 16\alpha + 1)/4 \\ b_2 &= (6\alpha^2 - 20\alpha + 5)/4. \end{aligned}$$

  We refer the reader to [23, 24] for additional details on this DIRK scheme.

All relaxation variants of a specified RK scheme are simply denoted by an extra "R" in front of RK. For example, RRK4 and DIRRK3 refer to the relaxation variants of RK4 and DIRK3, respectively.

3.3. Verifying linearizations

The first set of numerical experiments verify our linearization RRK formulas, as well as highlight the importance of taking into consideration the relaxation parameter and the corrected final step-size in the linearization process. Let

$$\mathbf{E}(\overline{\mathbf{y}}) = \overline{\mathbf{w}}$$

denote an abstract system of time-stepping equations for a given right-hand-side vector $\overline{\mathbf{w}}$. Let $\mathbf{H}$ denote the inverse of $\mathbf{E}$, i.e., if $\overline{\mathbf{y}}$ is the solution to the equation above then $\overline{\mathbf{y}} = \mathbf{H}(\overline{\mathbf{w}})$. Essentially, $\mathbf{H}$ specifies the explicit update formulas of a given time-integration scheme. From implicit differentiation, it follows that the directional derivative of $\mathbf{H}$, evaluated at $\overline{\mathbf{w}}$ in direction $\overline{\mathbf{d}}$, is given by $\overline{\boldsymbol{\delta}}$,

$$\overline{\boldsymbol{\delta}} = \mathbf{H}'(\overline{\mathbf{w}})\overline{\mathbf{d}}.$$

In practice, we compute $\overline{\boldsymbol{\delta}}$ as the solution to

$$\mathbf{E}'(\overline{\mathbf{y}})\overline{\boldsymbol{\delta}} = \overline{\mathbf{d}}.$$

This motivates our work in computing the Jacobian $\mathbf{E}'$ for the different schemes.

To verify that our linearized code indeed computes derivatives we study the numerical convergence of a simple finite difference approximation to the directional derivative, similar to what is done in [25] under a different context. In particular, we check that

$$\left\| \frac{\mathbf{H}(\overline{\mathbf{w}} + h\overline{\mathbf{d}}) - \mathbf{H}(\overline{\mathbf{w}})}{h} - \mathbf{H}'(\overline{\mathbf{w}})\overline{\mathbf{d}} \right\| = \left\| \frac{\overline{\mathbf{y}}_h - \overline{\mathbf{y}}}{h} - \overline{\boldsymbol{\delta}} \right\| = \mathcal{O}(h) \tag{28}$$

as $h \to 0^+$, where $\overline{\mathbf{y}}$, $\overline{\mathbf{y}}_h$ and $\overline{\boldsymbol{\delta}}$ are solutions to the following systems:

$$\begin{aligned} \mathbf{E}(\overline{\mathbf{y}}) &= \overline{\mathbf{w}}, \\ \mathbf{E}(\overline{\mathbf{y}}_h) &= \overline{\mathbf{w}} + h\overline{\mathbf{d}}, \\ \mathbf{E}'(\overline{\mathbf{y}})\overline{\boldsymbol{\delta}} &= \overline{\mathbf{d}}. \end{aligned}$$

3.3.1. Nonlinear pendulum results

We use the nonlinear pendulum equations 25 as our first model for tests concerning proper linearization. When computing the finite difference error in equation 28, we take $\overline{\mathbf{w}} = \overline{\mathbf{0}}$ and $\overline{\mathbf{d}} = \boldsymbol{\chi}_0 \mathbf{d}_0$ where $\mathbf{d}_0$ is a random vector meant to represent a random initial condition.

Figure 1 plots finite difference error 28 for different choices of Jacobian $\mathbf{E}'$ representing linearized RRK formulas with proper linearization or for the γ-constant or Δt^*-constant cases. As expected, we observe that proper linearization achieves linear convergence while the other two cases do not, which is apparent for lower order RK schemes. Interestingly enough, the Δt^*-constant case yields significantly smaller errors than the $\boldsymbol{\gamma}$-constant case, though both fail to converge.

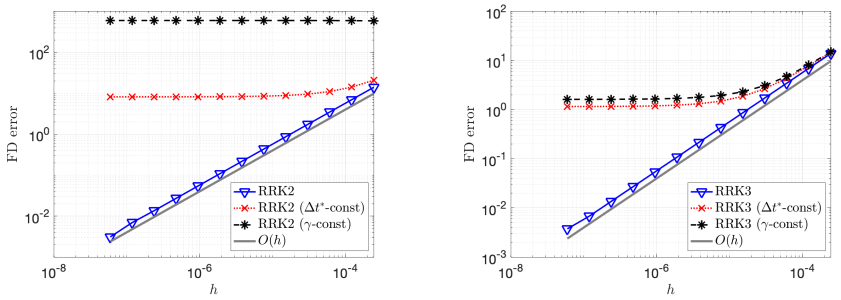
3.3.2. 1D Compressible Euler equations results

We run the same FD convergence test with proper and improper linearizations of RRK when solving the semi-discretization of the one-dimensional compressible Euler equations 26. For the step size, we use

$$\Delta t = \mathrm{CFL} \times \frac{h}{C_N}, \quad C_N = \frac{(N+1)^2}{2}$$

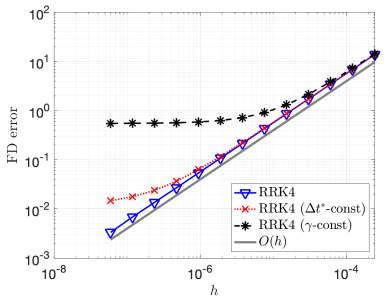
where $h = 1/16$ is the size of the DG element, $N = 3$ the polynomial order of the DG method, and CFL is a user-defined constant.

Figure 2 demonstrates the FD convergence, or lack of, for different choice of CFL constant. As before, linear convergence is achieved clearly under proper linearization. Larger step sizes reveal that indeed the improper linearizations



(a) Using RK2.

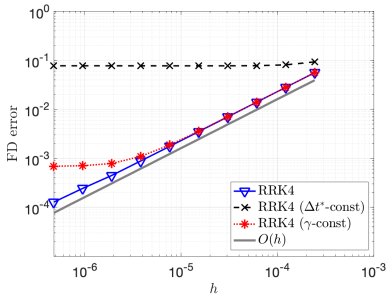
(b) Using RK3.



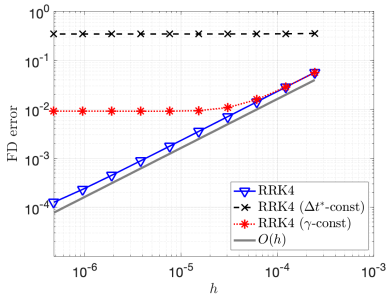
(c) Using RK4.

Figure 1: Convergence of the finite difference error 28 for RRK with nonlinear pendulum model; $\Delta t = 0.1$ and $T = 200$.

fail to converge. Unlike the nonlinear pendulum example, however, the Δt^*-constant case yielded larger errors than the γ-constant case.



(a) Using RK4, CFL = 1.



(b) Using RK4, CFL = 1.5.

Figure 2: Convergence of the finite difference error 28 for RRK with 1D compressible Euler model.

3.4. Consistency of discrete RRK adjoints

As mentioned in the introduction, the discretize-then-optimize approach yields a discrete adjoint equation 5 for gradient computations of cost functions. An optimize-then-discretize approach would have resulted in a continuous analog, with some continuous adjoint equation. The discrete adjoint equation is said to be *consistent* if its solution converges to the solution of the continuous adjoint equation.

Depending on the choice of time-integrator, the resulting adjoint equation may or may not be consistent. For RK time-integrators, it has been shown that the discrete adjoint equation correspond to an RK discretization (of same order) of the continuous adjoint equation, [12]. Concerning adaptive time-integrators, [15, 16] argue that taking into consideration the adaptive step-size in the linearization process can result in inconsistent adjoint scheme. We present here a convergence study of the discrete RRK adjoint to address these concerns and verify (at least numerically) consistency.

Consider control problem 1 with cost function

$$\mathcal{C}(\mathbf{u}) = \frac{1}{2}\|\mathbf{y}(T)\|^2,$$

subject to

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} &= \begin{pmatrix} -\sin(y_2(t)) \\ y_1(t) \end{pmatrix}, \quad 0 < t \leq T, \\ \mathbf{y}(0) &= \mathbf{u}, \end{aligned}$$

where the initial condition is the control variable. The discretized optimal control problem is given by 3 with

$$\begin{aligned} \mathsf{C}(\overline{\mathbf{y}}, \mathbf{u}) &= \frac{1}{2}\|\mathbf{y}_K\|^2 = \frac{1}{2}\overline{\mathbf{y}}^\top \boldsymbol{\chi}_K \boldsymbol{\chi}_K^\top \overline{\mathbf{y}}, \\ \mathbf{E}(\overline{\mathbf{y}}, \mathbf{u}) &= \mathbf{L} - \mathbf{N}(\overline{\mathbf{y}}) - \boldsymbol{\chi}_0 \mathbf{u}, \end{aligned}$$

assuming the state equation has been discretized by an RK/RRK scheme, taking K steps to arrive at the final time. The gradient of the reduced, discrete cost function is

$$\nabla \widetilde{\mathsf{C}}(\mathbf{u}) = \boldsymbol{\chi}_0^\top \overline{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_0$$

where $\overline{\boldsymbol{\lambda}}$ is the solution to the discrete adjoint equation,

$$\left(\frac{\partial \mathbf{E}}{\partial \overline{\mathbf{y}}}(\overline{\mathbf{y}}, \mathbf{u})\right)^\top \overline{\boldsymbol{\lambda}} = \boldsymbol{\chi}_K \boldsymbol{\chi}_K^\top \overline{\mathbf{y}}. \tag{29}$$

Note that $\overline{\mathbf{y}}$ in the discrete adjoint problem corresponds to the solution of the forward problem, that is, the state equation $\mathbf{E}(\overline{\mathbf{y}}, \mathbf{u}) = \overline{\mathbf{0}}$. Moreover, the right-hand-side of the adjoint equation here simply dictates the final time condition, specified by $\mathbf{y}_K$.

In the numerical experiments presented here, we examine the convergence of

$$\text{forward error} = \frac{\|\mathbf{y}_K - \mathbf{y}_{\text{ref}}(T)\|}{\|\mathbf{y}_{\text{ref}}(T)\|}, \quad \text{adjoint error} = \frac{\|\boldsymbol{\lambda}_0 - \lambda_{\text{ref}}(0)\|}{\|\lambda_{\text{ref}}(0)\|}$$

as $\Delta t \to 0$, where $\mathbf{y}_{\text{ref}}$ and λ_{ref} are computed using RK4 (and its adjoint) with a small step-size of $\Delta t = 10^{-5}$. Given that RK4 and its adjoint are consistent discretizations of the continuous state and adjoint state equations, we can expect that our reference solution will be sufficiently accurate for these tests.

Figure 3 shows the well expected convergence of RK and RRK schemes of different order for the forward problem. In the adjoint problem, figure 4, we observe that adjoint RRK, along with its improper linearization variants, is indeed consistent. Moreover, optimal convergence rates are achieved by adjoint RRK with proper linearization and the γ-constant case. Interestingly enough, we see that we lose an order of convergence in the Δt^*-constant case.

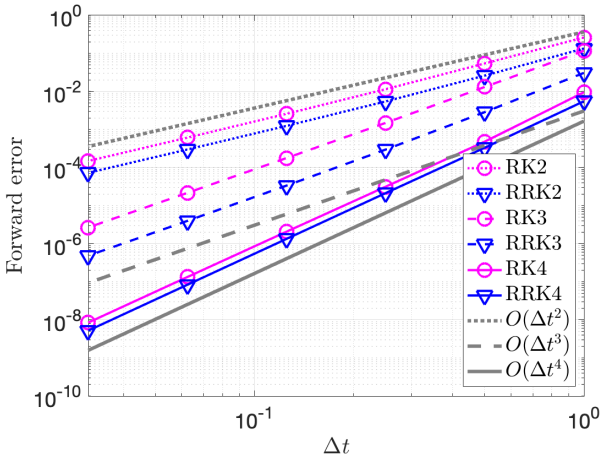
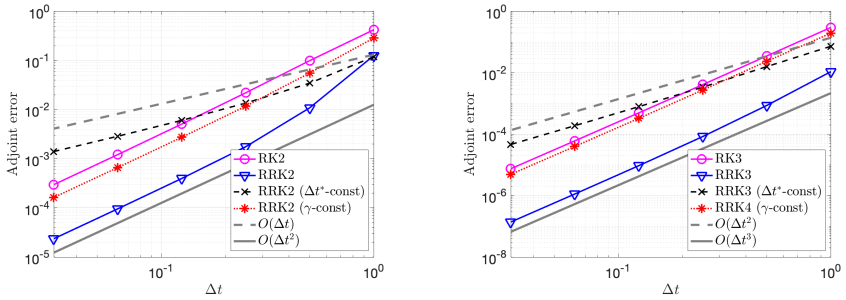


Figure 3: Convergence of RK and RRK discretizations of nonlinear pendulum problem; $T = 2$.

3.5. Qualitative behavior of adjoint solutions

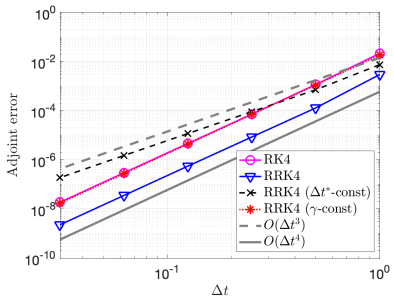
The adjoint convergence plots in figure 4 not only demonstrate that the adjoint RRK method (with proper linearization) is consistent but is also the most accurate out of all of the other methods, including standard RK. We explore this in these next set of experiments, using again the nonlinear pendulum as our state equations with a much larger final time of $T = 200$.

In figure 5 and 6a, we plot the norm of the adjoint solution as a function or time. Note that the plots are presented with the time axis in reverse, in spirit with the back-propagation of adjoint numerical methods. We see that all of the methods for computing the adjoint solution yield consistent results in the case where we use both RK4 and a small step-size of $\Delta t = 0.1$, figure 5c. However,



(a) Using RK2.

(b) Using RK3.



(c) Using RK4.

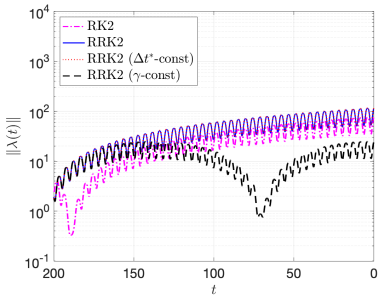
Figure 4: Convergence of RK, and improperly linearized RRK, discretizations of the adjoint nonlinear pendulum problem; $T = 2$.

when using RK2, we begin to see standard RK2 and RRK2 (γ-constant) deteriorate, figure 5a. Standard RK2 yields significant differences from the onset but eventually recovers a similar growth in the adjoint solution. Conversely, RRK2 (γ-constant) yields reasonably accurate results as it begins stepping backwards in time, but soon after becomes erratic. RRK with proper linearization and with Δt^*-constant produce accurate results for all three choices of RK schemes, at least for the smaller step-size, $\Delta t = 0.1$.

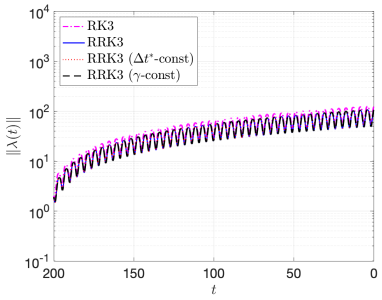
For the larger step-size, figure 6a, we can observe more significant differences between RRK with proper linearization and the Δt^*-constant case, which becomes more apparent as we progress backwards in time. Moreover, the γ-constant case results in a substantial growth of the adjoint solution by the time we arrive at the initial time. The standard adjoint RK4 method yields highly inaccurate results, in part due to the numerical dissipation observed in its forward solution, see figure 6b.

3.6. Preservation of time-symmetry

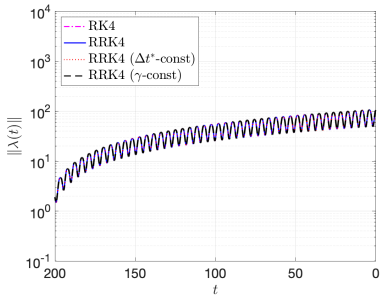
The next set of results verify the time-symmetry property of RRK schemes for model skew-symmetric problems 27, as detailed in theorem 4. To showcase time-symmetry (or the lack thereof) we solve the skew-symmetric problem and use the numerical solution at the final time as the final-time condition for the



(a) Using RK2.

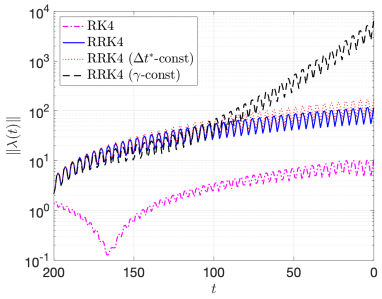


(b) Using RK3.

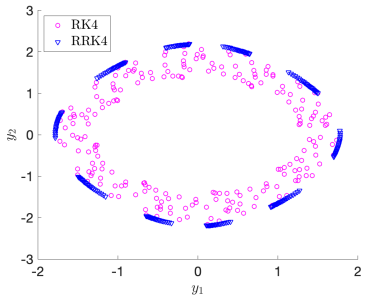


(c) Using RK4.

Figure 5: Norm of adjoint solutions to the nonlinear pendulum model; $\Delta t = 0.1$ and $T = 200$.



(a) Norm of adjoint solutions.



(b) Phase plot of forward solutions.

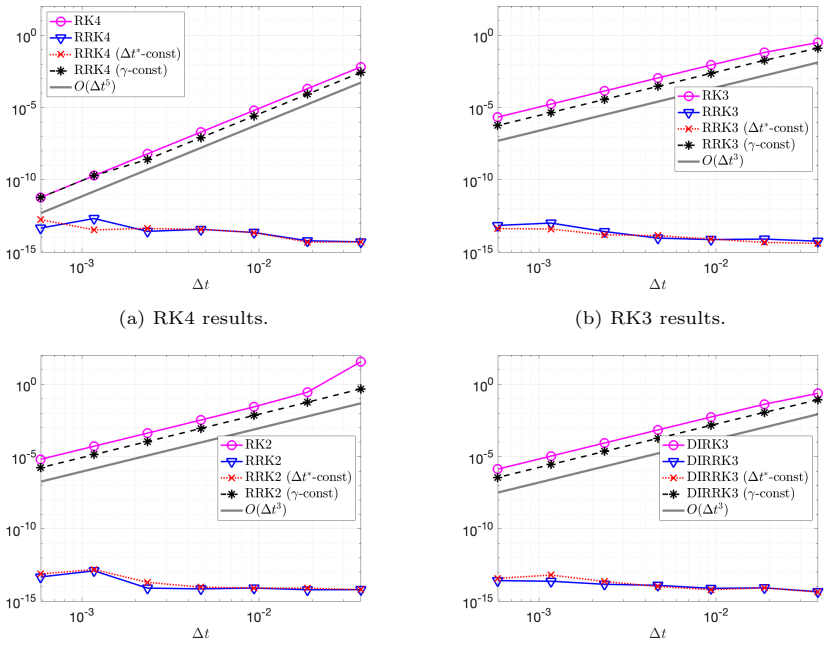
Figure 6: Norm of adjoint solutions and scattered plot of forward solution to the nonlinear pendulum model, using RK4; $\Delta t = 0.9$ and $T = 200$.

adjoint solution. If a scheme is time symmetric then the initial condition and the numerical adjoint solution should coincide up to machine precision. Figure 7 plots the error

$$\frac{\|\boldsymbol{\lambda}_0 - \mathbf{y}_{\text{init}}\|}{\|\mathbf{y}_{\text{init}}\|}$$

versus Δt.

Numerical results demonstrate the ability for RRK to preserve time-symmetry in its adjoint, as observed by the small errors for both explicit and implicit RK schemes. Time-symmetry is violated in both the standard RK and RRK with $\boldsymbol{\gamma}$-constant cases. Interestingly enough, numerical results show that RK4 and RK2 (e.g., even order RK schemes) converge at a rate higher than anticipated (fifth and third order convergence respectively) while RK3 and DIRK3 maintain their third order convergence. The authors are unaware as to why RK4 and RK2 exhibit this super convergence behavior. It is also observed that the Δt^*-constant case demonstrates time-symmetry, which can be explain by the fact that the terms appearing in the proper linearization (specifically scalar ρ_* and ξ_* in Lemma 3 and 4) that would otherwise be missing in the Δt^*-constant case actually zero-out for this linear problem. In other words, proper linearization and the Δt^*-constant case are equivalent for this linear problem with square entropy.



(a) RK4 results.

(b) RK3 results.

(c) RK2 results.

(d) DIRK3 results.

Figure 7: Time-symmetry results.

4. Conclusion

We have presented the discrete linearization and adjoint of relaxation RK methods. Our approach is based on implicit differentiation and a global matrix representation of the time-stepping equations. Even though the relaxation parameter is a nonlinear function of the state variables, we are able to prove that the relaxation RK method is equivalent to its linearization when applied to a skew-symmetric linear problem. Moreover, the relaxation method is proven to be *time-symmetric* for explicit and diagonally implicit RK schemes on skew-symmetric linear problems, in the sense that the adjoint time-step reverses the forward time-step. Numerical results also demonstrate the importance of proper linearization, in particular, the importance of taking into account the relaxation parameter, and the corrected final step-size for our implementation of RRK. Numerical results also show that the discrete RRK adjoint is not only consistent (with optimal convergence), but is more accurate in computing adjoint solutions.

Appendix

In this appendix we present in detail derivations of linearization and adjoint computations for RK and its relaxation variant using a matrix representation. We use an approach and notation similar to [4]. The plan is to interpret time-stepping algorithms as solutions to global matrix-vector systems and use the Jacobians of said systems to deduce a time stepping scheme for the discrete linearization and adjoint.

We first recap and introduce some notation.

- N is the dimension of the state vector $\mathbf{y}(t)$ in IVP 2;
- K is the total number of time steps taken by a given time-stepping scheme;
- s is the number of internal stages for an RK method;
- $\mathsf{A}_s, \mathsf{b}_s, \mathsf{c}_s$ are the coefficients of a specified s-stage RK method;
- The concatenation of vectors indexed by internal stages is denoted by simply removing the internal stage index, e.g.,

$$\mathbf{Y}_k := \begin{pmatrix} \mathbf{Y}_{k,1} \\ \vdots \\ \mathbf{Y}_{k,s} \end{pmatrix} \in \mathbb{R}^{sN};$$

- Vectors of size $\overline{N} := N + N(s+1)K$ are denoted using bold font and an

overline and have components denoted as follows:

$$\overline{\mathbf{y}} := \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{Y}_1 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{Y}_K \\ \mathbf{y}_K \end{pmatrix},$$

with $\mathbf{y}_k \in \mathbb{R}^N$ and $\mathbf{Y}_k \in \mathbb{R}^{sN}$;

- Matrix $\boldsymbol{\chi}_k \in \mathbb{R}^{\overline{N} \times N}$ is defined such that $\boldsymbol{\chi}_k^\top \overline{\mathbf{y}} = \mathbf{y}_k$, extracting the k-th step vector. Moreover, for a given $\mathbf{v} \in \mathbb{R}^N$, then $\overline{\mathbf{v}} := \boldsymbol{\chi}_k \mathbf{v}$ is vector of length $\overline{N}$ with zero entries everywhere except at $\mathbf{v}_k = \mathbf{v}$;
- $\mathbf{I}_M$ denotes the $M \times M$ identity matrix;
- $\mathbf{0}_M$ is the zero vector of dimension M;
- $\mathbf{0}_{M_1 \times M_2}$ is the $M_1 \times M_2$ zero matrix;
- $\otimes$ denotes the Kronecker product.

RK Matrix-representation

A single step of the RK method, as specified by equations 6, can be written in matrix form as

$$\begin{pmatrix} -\mathbf{C} & \mathbf{I}_{sN} & \\ -\mathbf{I}_N & & \mathbf{I}_N \end{pmatrix} \begin{pmatrix} \mathbf{y}_{k-1} \\ \mathbf{Y}_k \\ \mathbf{y}_k \end{pmatrix} - \begin{pmatrix} \mathbf{A}\mathbf{F}_k \\ \mathbf{B}^\top \mathbf{F}_k \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{sN} \\ \mathbf{0}_N \end{pmatrix}$$

where

- $\mathbf{A} := \Delta t \mathbf{A}_s \otimes \mathbf{I}_N \in \mathbb{R}^{sN \times sN}$,
- $\mathbf{B} := \Delta t \mathbf{b}_s \otimes \mathbf{I}_N \in \mathbb{R}^{sN \times N}$,
- $\mathbf{C} := \mathbf{1}_s \otimes \mathbf{I}_N \in \mathbb{R}^{sN \times N}$, with $\mathbf{1}_s := (1, ..., 1)^\top \in \mathbb{R}^s$,
- $\mathbf{F}_k$ is the concatenation of the $\mathbf{F}_{k,i} := \mathbf{f}(\mathbf{Y}_{k,i}, t_{k-1} + c_i \Delta t)$, and subsequently can be viewed as a vector valued function of $\mathbf{Y}_k$.

The RK method as a whole can be represented as a concatenation of the matrix systems above, resulting in a global system of time-stepping equations:

$$\mathbf{E}(\overline{\mathbf{y}}) := \mathbf{L}\overline{\mathbf{y}} - \mathbf{N}(\overline{\mathbf{y}}) - \boldsymbol{\chi}_0 \mathbf{y}_{\text{init}} = \overline{\mathbf{0}}$$

with

$$\mathbf{L} := \begin{pmatrix} \mathbf{I}_N & & & & \\ -\mathbf{C} & \mathbf{I}_{sN} & & & \\ -\mathbf{I}_N & & \mathbf{I}_N & & \\ & & -\mathbf{C} & \mathbf{I}_{sN} & \\ & & -\mathbf{I}_N & & \mathbf{I}_N \\ & & & & \ddots \end{pmatrix}, \quad \mathbf{N}(\overline{\mathbf{y}}) := \begin{pmatrix} \mathbf{0}_N \\ \hline \mathbf{A}\mathbf{F}_1 \\ \mathbf{B}^\top \mathbf{F}_1 \\ \hline \mathbf{A}\mathbf{F}_2 \\ \mathbf{B}^\top \mathbf{F}_2 \\ \hline \vdots \end{pmatrix}.$$

We make some remarks and observations:

- Boxes are meant to help visually separate blocks associated with different time steps in both $\mathbf{L}$ and $\mathbf{N}(\overline{\mathbf{y}})$.
- The unboxed $\mathbf{I}_N$ in the top left corner of $\mathbf{L}$ is related to the enforcement of the initial condition.
- $\mathbf{L} \in \mathbb{R}^{\overline{N} \times \overline{N}}$ is lower unit triangular, though not quite block diagonal due to some slight overlap in columns.
- Given the repeating block structure of matrices presented here, we only write out the blocks associated the initial/final conditions and two subsequent time steps. Dots indicate a repeating pattern with the understanding that the block structure repeats K times with appropriate indexing when relevant.
- $\mathbf{N} : \mathbb{R}^{\overline{N}} \to \mathbb{R}^{\overline{N}}$ is block lower triangular in the sense that the k-th $N(s+1)$ block of the output does not depend on the j-th $N(s+1)$ block of the input, for $j > k$. For example, the $N(s+1)$ block of $\mathbf{N}(\overline{\mathbf{y}})$ associated with the k-th time step, is

  $$\begin{pmatrix} \mathbf{A}\mathbf{F}_k \\ \mathbf{B}^\top \mathbf{F}_k \end{pmatrix}$$

  which only depends on $\mathbf{Y}_k$, and not on $\overline{\mathbf{y}}$ at later time steps.

Linearized RK formulas 9 are derived by computing the Jacobian, $\mathbf{E}'$, and interpreting the solution to linear system $\mathbf{E}'(\overline{\mathbf{y}})\overline{\boldsymbol{\delta}} = \overline{\mathbf{w}}$ (via forward substitution) as a time-stepping algorithm. The Jacobian of the global time-stepping equation operator $\mathbf{E}(\overline{\mathbf{y}})$ is given by

$$\mathbf{E}'(\overline{\mathbf{y}}) = \mathbf{L} - \mathbf{N}'(\overline{\mathbf{y}})$$

with

$$\mathbf{N}'(\overline{\mathbf{y}}) = \begin{pmatrix} \mathbf{0}_{N\times N} & & & & \\ & \mathbf{A}\mathbf{J}_1 & & & \\ & \mathbf{B}^\top \mathbf{J}_1 & \mathbf{0}_{N\times N} & & \\ & & & \mathbf{A}\mathbf{J}_2 & \\ & & & \mathbf{B}^\top \mathbf{J}_2 & \mathbf{0}_{N\times N} \\ & & & & \ddots \end{pmatrix}$$

$$\mathbf{J}_k := \mathrm{diag}(\mathbf{J}_{k,1}, \cdots, \mathbf{J}_{k,s}) \in \mathbb{R}^{sN \times sN},$$
$$\mathbf{J}_{k,i} := \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\mathbf{Y}_{k,i}, t_{k-1} + c_i \Delta t) \in \mathbb{R}^{N \times N}.$$

We see that the block lower triangular structure of the operator $\mathbf{N}$ yields a proper block lower triangular Jacobian $\mathbf{N}'$. Putting it together, we have

$$\mathbf{E}'(\overline{\mathbf{y}}) = \begin{pmatrix} \mathbf{I}_N & & & & \\ -\mathbf{C} & \mathbf{I}_{sN} - \mathbf{A}\mathbf{J}_1 & & & \\ -\mathbf{I}_N & -\mathbf{B}^\top \mathbf{J}_1 & \mathbf{I}_N & & \\ & & -\mathbf{C} & \mathbf{I}_{sN} - \mathbf{A}\mathbf{J}_2 & \\ & & -\mathbf{I}_N & -\mathbf{B}^\top \mathbf{J}_2 & \mathbf{I}_N \\ & & & & \ddots \end{pmatrix}$$

In particular, each time step is associated with solving the following system for $\boldsymbol{\Delta}_k$ and $\boldsymbol{\delta}_k$, with $\boldsymbol{\delta}_{k-1}$ given by the previous time step:

$$\begin{pmatrix} -\mathbf{C} & \mathbf{I}_{sN} - \mathbf{A}\mathbf{J}_k & \\ -\mathbf{I}_N & -\mathbf{B}^\top \mathbf{J}_k & \mathbf{I}_N \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_{k-1} \\ \boldsymbol{\Delta}_k \\ \boldsymbol{\delta}_k \end{pmatrix} = \begin{pmatrix} \mathbf{W}_k \\ \mathbf{w}_k \end{pmatrix},$$

$$\implies \quad \begin{cases} \boldsymbol{\Delta}_{k,i} = \boldsymbol{\delta}_{k-1} + \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{J}_{k,j} \boldsymbol{\Delta}_{k,j} + \mathbf{W}_{k,i}, \quad i = 1, ..., s, \\ \\ \boldsymbol{\delta}_k = \boldsymbol{\delta}_{k-1} + \Delta t \sum_{i=1}^{s} b_i \mathbf{J}_{k,i} \boldsymbol{\Delta}_{k,i} + \mathbf{w}_k. \end{cases}$$

Adjoint RK formulas 10 are derived in a similar fashion, but with the transpose of $\mathbf{E}'$, which results in a block upper triangular matrix,

$$\mathbf{E}'(\overline{\mathbf{y}})^\top = \begin{pmatrix} \ddots & & & & \\ & \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N & \\ & & \mathbf{I}_{sN} - \mathbf{J}_{K-1}^\top \mathbf{A}^\top & -\mathbf{J}_{K-1}^\top \mathbf{B} & \\ & & & \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N \\ & & & & \mathbf{I}_{sN} - \mathbf{J}_K^\top \mathbf{A}^\top & -\mathbf{J}_K^\top \mathbf{B} \\ & & & & & \mathbf{I}_N \end{pmatrix}.$$

Analogous to $\mathbf{L}$, we see a repeating block structure with overlapping columns, though with an identity block at the lower right corner, associated with the final time condition. We interpret the solution to linear system $\mathbf{E}'(\overline{\mathbf{y}})^\top \overline{\boldsymbol{\lambda}} = \overline{\mathbf{w}}$ (via back substitution) as a time-stepping algorithm running in reverse time. Each time step is associated with solving the following system for $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\lambda}_{k-1}$, with

$\boldsymbol{\lambda}_k$ given by the previous adjoint time step:

$$\begin{pmatrix} \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N \\ & \mathbf{I}_{sN} - \mathbf{J}_k^\top \mathbf{A}^\top & -\mathbf{J}_k^\top \mathbf{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_{k-1} \\ \boldsymbol{\Lambda}_k \\ \boldsymbol{\lambda}_k \end{pmatrix} = \begin{pmatrix} \mathbf{w}_{k-1} \\ \mathbf{W}_k \end{pmatrix},$$

$$\Longrightarrow \quad \begin{cases} \boldsymbol{\lambda}_{k-1} = \boldsymbol{\lambda}_k + \sum_{i=1}^{s} \boldsymbol{\Lambda}_{k,i} + \mathbf{w}_{k-1}, \\ \boldsymbol{\Lambda}_{k,i} = b_i \Delta t \mathbf{J}_{k,i}^\top \boldsymbol{\lambda}_k + \Delta t \sum_{j=1}^{s} a_{ji} \mathbf{J}_{k,i}^\top \boldsymbol{\Lambda}_{k,j} + \mathbf{W}_{k,i}, \quad i = 1, ..., s. \end{cases}$$

IDT matrix representation

The matrix representation of the IDT method is very similar to what we derived for RK,

$$\mathbf{E}(\overline{\mathbf{y}}, \boldsymbol{\gamma}) := \mathbf{L}\overline{\mathbf{y}} - \mathbf{N}(\overline{\mathbf{y}}, \boldsymbol{\gamma}) - \boldsymbol{\chi}_0 \mathbf{y}_{\text{init}} = \overline{\mathbf{0}}$$

where the relaxation parameters $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_k)$ appear on the (nonlinear) term $\mathbf{N}$ only, i.e.,

$$\mathbf{N}(\overline{\mathbf{y}}, \boldsymbol{\gamma}) := \begin{pmatrix} \mathbf{0}_N \\ \hline \mathbf{A}\mathbf{F}_1 \\ \hline \gamma_1 \mathbf{B}^\top \mathbf{F}_1 \\ \hline \mathbf{A}\mathbf{F}_2 \\ \hline \gamma_2 \mathbf{B}^\top \mathbf{F}_2 \\ \hline \vdots \end{pmatrix}. \tag{30}$$

Recall that γ_k is defined as the positive root near 1 (for Δt small enough) of the root function $r(\gamma; \mathbf{y}_{k-1}, \mathbf{Y}_k)$, equation 12. In other words the relaxation parameters depend implicitly on $\overline{\mathbf{y}}$, i.e., $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\overline{\mathbf{y}})$. Let $\widetilde{\mathbf{E}}(\overline{\mathbf{y}})$ denote the reduced state-equation operator, that is,

$$\widetilde{\mathbf{E}}(\overline{\mathbf{y}}) := \mathbf{E}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}})).$$

The Jacobian is given by

$$\widetilde{\mathbf{E}}'(\overline{\mathbf{y}}) = \mathbf{L} - \underbrace{\left(\frac{\partial \mathbf{N}}{\partial \overline{\mathbf{y}}}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}})) + \frac{\partial \mathbf{N}}{\partial \boldsymbol{\gamma}}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}})) \boldsymbol{\gamma}'(\overline{\mathbf{y}}) \right)}_{\widetilde{\mathbf{N}}'(\overline{\mathbf{y}})}.$$

In particular,

$$\widetilde{\mathbf{N}}'(\bar{\mathbf{y}}) = \begin{pmatrix} \mathbf{0}_{N\times N} & & & & \\ & \mathbf{A}\mathbf{J}_1 & & & \\ & \boldsymbol{\Gamma}_{y,1} \quad \gamma_1\mathbf{B}^\top\mathbf{J}_1 + \boldsymbol{\Gamma}_{Y,1} \quad \mathbf{0}_{N\times N} & & \\ & & & \mathbf{A}\mathbf{J}_2 & \\ & & & \boldsymbol{\Gamma}_{y,2} \quad \gamma_2\mathbf{B}^\top\mathbf{J}_2 + \boldsymbol{\Gamma}_{Y,2} \quad \mathbf{0}_{N\times N} & \\ & & & & \ddots \end{pmatrix}$$

where

$$\boldsymbol{\Gamma}_{y,k} := \mathbf{B}^\top\mathbf{F}_k(\nabla_y\gamma_k)^\top, \quad \boldsymbol{\Gamma}_{Y,k} := \mathbf{B}^\top\mathbf{F}_k(\nabla_Y\gamma_k)^\top$$

are associated with the term $\frac{\partial \mathbf{N}}{\partial \bar{\mathbf{y}}}\boldsymbol{\gamma}'$, which we highlight using red text. Again, gradient terms in $\boldsymbol{\Gamma}_{y,k}$ and $\boldsymbol{\Gamma}_{Y,k}$ correspond to gradients of γ_k with respect to $\mathbf{y}_{k-1}$ and $\mathbf{Y}_k$ respectively, and are computed via implicit differentiation; see equations 13 and 14.

The Jacobian matrix for IDT is thus given by

$$\widetilde{\mathbf{E}}'(\bar{\mathbf{y}}) = \begin{pmatrix} \mathbf{I}_N & & & & \\ -\mathbf{C} & \mathbf{I}_{sN} - \mathbf{A}\mathbf{J}_1 & & & \\ -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,1} & -\gamma_1\mathbf{B}^\top\mathbf{J}_1 - \boldsymbol{\Gamma}_{Y,1} & \mathbf{I}_N & & \\ & & -\mathbf{C} & \mathbf{I}_{sN} - \mathbf{A}\mathbf{J}_2 & \\ & & -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,2} & -\gamma_2\mathbf{B}^\top\mathbf{J}_2 - \boldsymbol{\Gamma}_{Y,2} & \mathbf{I}_N \\ & & & & \ddots \end{pmatrix}.$$

Solving $\widetilde{\mathbf{E}}'(\bar{\mathbf{y}})\bar{\boldsymbol{\delta}} = \bar{\mathbf{w}}$ via forward substitution results in solving at each time step the following system for $\boldsymbol{\Delta}_k$ and $\boldsymbol{\delta}_k$, with $\boldsymbol{\delta}_{k-1}$ given by the previous time step, deriving the linearized IDT formulas in lemma 1:

$$\begin{pmatrix} -\mathbf{C} & \mathbf{I}_{sN} - \mathbf{A}\mathbf{J}_k & \\ -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,k} & -\gamma_k\mathbf{B}^\top\mathbf{J}_k - \boldsymbol{\Gamma}_{Y,k} & \mathbf{I}_N \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_{k-1} \\ \boldsymbol{\Delta}_k \\ \boldsymbol{\delta}_k \end{pmatrix} = \begin{pmatrix} \mathbf{W}_k \\ \mathbf{w}_k \end{pmatrix}$$

$$\implies \begin{cases} \boldsymbol{\Delta}_{k,i} = \boldsymbol{\delta}_{k-1} + \Delta t \sum_{j=1}^{s} a_{ij}\mathbf{J}_{k,j}\boldsymbol{\Delta}_{k,j} + \mathbf{W}_{k,i}, \quad i = 1, ..., s, \\ \boldsymbol{\delta}_k = \boldsymbol{\delta}_{k-1} + \gamma_k\Delta t \sum_{i=1}^{s} b_i\mathbf{J}_{k,i}\boldsymbol{\Delta}_{k,i} + \boldsymbol{\Gamma}_{y,k}\boldsymbol{\delta}_{k-1} + \boldsymbol{\Gamma}_{Y,k}\boldsymbol{\Delta}_k + \mathbf{w}_k. \end{cases}$$

Note that

$$\boldsymbol{\Gamma}_{y,k}\boldsymbol{\delta}_{k-1} + \boldsymbol{\Gamma}_{Y,k}\boldsymbol{\Delta}_k = \rho_k\Delta t \sum_{i=1}^{s} b_i\mathbf{F}_{k,i}$$

with scalar $\rho_k := \nabla_y\gamma_k^\top\boldsymbol{\delta}_{k-1} + \nabla_Y\gamma_k^\top\boldsymbol{\Delta}_k$.

The transpose of the Jacobian is given by

$$\widetilde{\mathbf{E}}'(\overline{\mathbf{y}})^\top = \begin{pmatrix} \ddots & & & \\ & \boxed{\begin{matrix} \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,K-1}^\top \\ & \mathbf{I}_{sN} - \mathbf{J}_{K-1}^\top \mathbf{A}^\top & -\gamma_{K-1}\mathbf{J}_{K-1}^\top \mathbf{B} - \boldsymbol{\Gamma}_{Y,K-1}^\top \end{matrix}} & & \\ & & \boxed{\begin{matrix} \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,K}^\top \\ & \mathbf{I}_{sN} - \mathbf{J}_K^\top \mathbf{A}^\top & -\gamma_K \mathbf{J}_K^\top \mathbf{B} - \boldsymbol{\Gamma}_{Y,K}^\top \\ & & \mathbf{I}_N \end{matrix}} \end{pmatrix}.$$

Solving $\widetilde{\mathbf{E}}'(\overline{\mathbf{y}})^\top \overline{\boldsymbol{\lambda}} = \overline{\mathbf{w}}$ via back substitution results in solving at each time step the following system for $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\lambda}_k - 1$, with $\boldsymbol{\lambda}_k$ given by the previous time step, deriving the adjoint IDT formulas in lemma 2:

$$\begin{pmatrix} \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,k}^\top \\ & \mathbf{I}_{sN} - \mathbf{J}_k^\top \mathbf{A}^\top & -\gamma_k \mathbf{J}_k^\top \mathbf{B} - \boldsymbol{\Gamma}_{Y,k}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_{k-1} \\ \boldsymbol{\Lambda}_k \\ \boldsymbol{\lambda}_k \end{pmatrix} = \begin{pmatrix} \mathbf{w}_{k-1} \\ \mathbf{W}_k \end{pmatrix}$$

$$\implies \begin{cases} \boldsymbol{\lambda}_{k-1} = \displaystyle\sum_{i=1}^{s} \boldsymbol{\Lambda}_{k,i} + \boldsymbol{\lambda}_k + \boldsymbol{\Gamma}_{y,k}^\top \boldsymbol{\lambda}_k + \mathbf{w}_{k-1}, \\ \boldsymbol{\Lambda}_{k,i} = \Delta t\, \mathbf{J}_{k,i}^\top \displaystyle\sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} + \gamma_k b_i \Delta t\, \mathbf{J}_{k,i}^\top \boldsymbol{\lambda}_k + \boldsymbol{\Gamma}_{Y,k}^\top \boldsymbol{\lambda}_k + \mathbf{W}_{k,i}, \quad i = 1, ..., s. \end{cases}$$

Note that

$$\boldsymbol{\Gamma}_{y,k}^\top \boldsymbol{\lambda}_k = (\nabla_y \gamma_k)\, \mathbf{F}_k^\top \mathbf{B} \boldsymbol{\lambda}_k = \xi_k (\nabla_y \gamma_k)$$
$$\boldsymbol{\Gamma}_{Y,k}^\top \boldsymbol{\lambda}_k = (\nabla_Y \gamma_k) \mathbf{F}_k^\top \mathbf{B} \boldsymbol{\lambda}_k = \xi_k (\nabla_Y \gamma_k)$$

with scalar

$$\xi_k := \mathbf{F}_k^\top \mathbf{B} \boldsymbol{\lambda}_k = \Delta t \sum_{i=1}^{s} b_i \mathbf{F}_{k,i}^\top \boldsymbol{\lambda}_k.$$

4.1. RRK matrix representation

The matrix representation for RRK is quite similar to what we derived for IDT:

$$\mathbf{E}(\overline{\mathbf{y}}, \boldsymbol{\gamma}, \Delta t^*) := \mathbf{L}\overline{\mathbf{y}} - \mathbf{N}(\overline{\mathbf{y}}, \boldsymbol{\gamma}, \Delta t^*) - \boldsymbol{\chi}_0 \mathbf{y}_{\text{init}} = \overline{\mathbf{0}},$$

where we have made the operator $\mathbf{N}$ dependent on the modified step size $\Delta t^* := T - t_{K-1}$ as well. Only the last $N(s+1)$ rows of $\mathbf{N}$, associated with the last time step, differ from what was presented in the IDT case (equation 30). These last last $N(s+1)$ rows of $\mathbf{N}$ are specified by

$$\begin{pmatrix} \mathbf{A}_* \mathbf{F}_K \\ \gamma_K \mathbf{B}_*^\top \mathbf{F}_K \end{pmatrix}$$

with

$$\mathbf{A}_* := \Delta t^* \mathbf{A}_s \otimes \mathbf{I}_N = \frac{\Delta t^*}{\Delta t}\mathbf{A},$$
$$\mathbf{B}_* := \Delta t^* \mathbf{b}_s \otimes \mathbf{I}_N = \frac{\Delta t^*}{\Delta t}\mathbf{B}.$$

Recall that in RRK we have $t_k = t_{k-1} + \gamma_k \Delta t$ for $k = 1, ..., K-1$, and hence

$$\Delta t^* = T - t_0 - \Delta t \sum_{\ell=1}^{K-1} \gamma_\ell$$

which makes Δt^* a function of $\mathbf{y}_{\ell-1}$ and $\mathbf{Y}_\ell$ for $\ell = 1, ..., K-1$, i.e., $\Delta t^* = \Delta t^*(\overline{\mathbf{y}})$. With this is mind, let $\widetilde{\mathbf{E}}(\overline{\mathbf{y}})$ denote the reduced state-equation operator,

$$\widetilde{\mathbf{E}}(\overline{\mathbf{y}}) := \mathbf{E}(\overline{\mathbf{y}}, \boldsymbol{\gamma}(\overline{\mathbf{y}}), \Delta t^*(\overline{\mathbf{y}})).$$

Given how $\mathbf{N}$ is modified in RRK, it follows that the Jacobian $\frac{d\widetilde{\mathbf{E}}}{d\overline{\mathbf{y}}}$ will coincide with what we derived for IDT except at the last $N(s+1)$ rows. In particular, computing

$$\frac{d}{d\overline{\mathbf{y}}}\Big(\mathbf{A}_*\mathbf{F}_K\Big) = \mathbf{A}_*\frac{d\mathbf{F}_K}{d\mathbf{y}} + \mathbf{A}\mathbf{F}_K\left(\frac{d}{d\overline{\mathbf{y}}}\frac{\Delta t^*}{\Delta t}\right),$$
$$\frac{d}{d\overline{\mathbf{y}}}\left(\gamma_K\mathbf{B}_*^\top\mathbf{F}_K\right) = \gamma_K\mathbf{B}_*^\top\frac{d\mathbf{F}_K}{d\overline{\mathbf{y}}} + \mathbf{B}_*^\top\mathbf{F}_K\left(\frac{d\gamma_K}{d\overline{\mathbf{y}}}\right) + \gamma_K\mathbf{B}^\top\mathbf{F}_K\left(\frac{d}{d\overline{\mathbf{y}}}\frac{\Delta t^*}{\Delta t}\right),$$

will require the derivatives of γ_K and Δt^* with respect to $\overline{\mathbf{y}}$.

The derivatives of Δt^* can be expressed in terms of derivatives of the relaxation parameters as follows:

$$\begin{aligned}\frac{d}{d\overline{\mathbf{y}}}\frac{\Delta t^*}{\Delta t}(\overline{\mathbf{y}}) &= -\left(\frac{\partial\gamma_1}{\partial\mathbf{y}_1},\ \frac{\partial\gamma_1}{\partial\mathbf{Y}_1},\ \cdots,\ \frac{\partial\gamma_{K-1}}{\partial\mathbf{y}_{K-1}},\ \frac{\partial\gamma_{K-1}}{\partial\mathbf{Y}_{K-1}},\ \mathbf{0}_{N(s+2)}^\top\right)\Big|_{\overline{\mathbf{y}}} \\ &= -\left(\nabla\gamma_1^\top,\ \cdots,\ \nabla\gamma_{K-1}^\top,\ \mathbf{0}_{N(s+2)}^\top\right),\end{aligned}$$

where

$$\nabla\gamma_k := \begin{pmatrix} \nabla_y\gamma_k \\ \nabla_Y\gamma_k \end{pmatrix}.$$

An added complication is that Δt^* is also the step size used in r_K, thus implying that γ_K is dependent on information from all previous time steps. As before, we can compute $\frac{d\gamma_K}{d\overline{\mathbf{y}}}$ via implicit differentiation, though we will have to

compute partial derivatives of γ_K with respect to $\mathbf{y}_{\ell-1}$ and $\mathbf{Y}_\ell$ for all $\ell = 1, ..., K$;

$$\frac{\partial \gamma_K}{\partial \mathbf{y}_{\ell-1}} = -\left(\frac{\partial r_K}{\partial \gamma}\right)^{-1} \frac{\partial r_K}{\partial \mathbf{y}_{\ell-1}},$$
$$\frac{\partial \gamma_K}{\partial \mathbf{Y}_\ell} = -\left(\frac{\partial r_K}{\partial \gamma}\right)^{-1} \frac{\partial r_K}{\partial \mathbf{Y}_\ell}.$$

The partial derivatives of r_K with respect to γ, $\mathbf{y}_{K-1}$ and $\mathbf{Y}_K$, evaluated at $(\gamma_K, \mathbf{y}_{K-1}, \mathbf{Y}_K)$ are as given in equation 14, with $k \mapsto K$ and $\Delta t \mapsto \Delta t^*$. Just as in equation 13, we use $\nabla_y \gamma_K$ and $\nabla_Y \gamma_K$ to denote the gradient of γ_K with respect to $\mathbf{y}_{K-1}$ and $\mathbf{Y}_K$ respectively. For the remaining $\ell = 1, ..., K-1$,

$$\begin{aligned}\frac{\partial r_K}{\partial \mathbf{y}_{\ell-1}}(\boldsymbol{\gamma}(\overline{\mathbf{y}}), \overline{\mathbf{y}}) &= \gamma_K \sum_{i=1}^{s} b_i \Big(\nabla\eta(\mathbf{y}_K) - \nabla\eta(\mathbf{Y}_{K,i})\Big)^\top \mathbf{F}_{K,i} \frac{\partial \Delta t^*}{\partial \mathbf{y}_{\ell-1}}(\overline{\mathbf{y}}) \\ &= -\gamma_K \left(\frac{\Delta t}{\Delta t^*} \frac{\partial r_K}{\partial \gamma}(\boldsymbol{\gamma}(\overline{\mathbf{y}}), \overline{\mathbf{y}})\right) (\nabla_y \gamma_\ell)^\top\end{aligned}$$

where we have used

$$\frac{\partial \Delta t^*}{\partial \mathbf{y}_{\ell-1}}(\overline{\mathbf{y}}) = -\Delta t \, (\nabla_y \gamma_\ell)^\top.$$

Similarly,

$$\begin{aligned}\frac{\partial r_K}{\partial \mathbf{Y}_{\ell,j}}(\boldsymbol{\gamma}(\overline{\mathbf{y}}), \overline{\mathbf{y}}) &= \gamma_K \sum_{i=1}^{s} b_i \Big(\nabla\eta(\mathbf{y}_K) - \nabla\eta(\mathbf{Y}_{K,i})\Big)^\top \mathbf{F}_{K,i} \frac{\partial \Delta t^*}{\partial \mathbf{Y}_{\ell,j}}(\overline{\mathbf{y}}) \\ &= -\gamma_K \left(\frac{\Delta t}{\Delta t^*} \frac{\partial r_K}{\partial \gamma}(\boldsymbol{\gamma}(\overline{\mathbf{y}}), \overline{\mathbf{y}})\right) (\nabla_Y \gamma_{\ell,j})^\top.\end{aligned}$$

Thus,

$$\frac{\partial \gamma_K}{\partial \mathbf{y}_{\ell-1}}(\overline{\mathbf{y}}) = \gamma_K \frac{\Delta t}{\Delta t^*} (\nabla_y \gamma_\ell)^\top,$$
$$\frac{\partial \gamma_K}{\partial \mathbf{Y}_\ell}(\overline{\mathbf{y}}) = \gamma_K \frac{\Delta t}{\Delta t^*} (\nabla_Y \gamma_\ell)^\top.$$

Putting it all together, we have

$$\begin{aligned}\frac{d\gamma_K}{d\overline{\mathbf{y}}}(\overline{\mathbf{y}}) &= \left(\gamma_K \tfrac{\Delta t}{\Delta t^*} \nabla\gamma_1^\top, \cdots, \gamma_K \tfrac{\Delta t}{\Delta t^*} \nabla\gamma_{K-1}^\top, \nabla\gamma_K^\top, \mathbf{0}_N^\top\right) \\ &= -\gamma_K \frac{\Delta t}{\Delta t^*} \left(\frac{d}{d\overline{\mathbf{y}}} \frac{\Delta t^*}{\Delta t}(\overline{\mathbf{y}})\right) + \left(\mathbf{0}_{N(s+1)(K-1)}^\top, \ \nabla\gamma_K^\top, \ \mathbf{0}_N^\top\right).\end{aligned}$$

In summary,

$$\frac{d}{d\mathbf{y}}\Big(\mathbf{A}_*\mathbf{F}_K\Big)\Big|_{\overline{\mathbf{y}}} = -\left(\mathbf{A}\mathbf{F}_K\nabla\gamma_1^\top,\ \cdots,\ \mathbf{A}\mathbf{F}_K\nabla\gamma_{K-1}^\top,\ \mathbf{0}_{sN\times N},\ -\mathbf{A}_*\mathbf{J}_K,\ \mathbf{0}_{sN\times N}\right)$$

$$\frac{d}{d\overline{\mathbf{y}}}\Big(\gamma_K\mathbf{B}_*^\top\mathbf{F}_K\Big)\Big|_{\overline{\mathbf{y}}} = \left(\mathbf{0}_{N(s+1)(K-1)}^\top,\ \mathbf{\Gamma}_{y,K},\ \mathbf{\Gamma}_{Y,K},\ \mathbf{0}_N^\top\right)$$

with

$$\mathbf{\Gamma}_{y,K} = \mathbf{B}_*^\top\mathbf{F}_K(\nabla_y\gamma_K)^\top,\quad \mathbf{\Gamma}_{Y,K} = \mathbf{B}_*^\top\mathbf{F}_K(\nabla_Y\gamma_K)^\top.$$

We jump forward to interpreting the solution of $\widetilde{\mathbf{E}}'(\overline{\mathbf{y}})\overline{\boldsymbol{\delta}} = \overline{\mathbf{w}}$ via forward substitution as a time stepping scheme. Again, the first $K-1$ steps are as given by IDT. The last step, as shown in lemma 1, is derived from the solution of the following system for $\boldsymbol{\Delta}_K$ and $\boldsymbol{\delta}_K$, with $(\boldsymbol{\delta}_{\ell-1}, \boldsymbol{\Delta}_\ell)$ for $\ell = 1, ..., K-1$ given by the the previous time steps:

$$\begin{pmatrix} \mathbf{A}\mathbf{F}_K\nabla\gamma_1^\top & \cdots & \mathbf{A}\mathbf{F}_K\nabla\gamma_{K-1}^\top & -\mathbf{C} & \mathbf{I}_{sN}-\mathbf{A}_*\mathbf{J}_K & \\ & & & -\mathbf{I}_N-\mathbf{\Gamma}_{y,K} & -\gamma_K\mathbf{B}_*^\top\mathbf{J}_K-\mathbf{\Gamma}_{Y,K} & \mathbf{I}_N \end{pmatrix} \begin{pmatrix} \begin{pmatrix}\boldsymbol{\delta}_0\\ \boldsymbol{\Delta}_1\end{pmatrix} \\ \vdots \\ \begin{pmatrix}\boldsymbol{\delta}_{K-2}\\ \boldsymbol{\Delta}_{K-1}\end{pmatrix} \\ \boldsymbol{\delta}_{K-1} \\ \boldsymbol{\Delta}_K \\ \boldsymbol{\delta}_K \end{pmatrix} = \begin{pmatrix} \mathbf{W}_K \\ \mathbf{w}_K \end{pmatrix}$$

$$\Longrightarrow \begin{cases} \boldsymbol{\Delta}_{K,i} = \boldsymbol{\delta}_{K-1} + \Delta t^* \sum_{j=1}^{s} a_{ij}\mathbf{J}_{K,j}\boldsymbol{\Delta}_{K,j} - \rho_*\Delta t\sum_{j=1}^{s} a_{ij}\mathbf{F}_{K,j} + \mathbf{W}_{K,i}, \quad i = 1, ..., s, \\ \boldsymbol{\delta}_K = \delta_{K-1} + \gamma_K\Delta t^*\sum_{i=1}^{s} b_i\mathbf{J}_{K,i}\boldsymbol{\Delta}_{K,i} + \mathbf{\Gamma}_{y,K}\boldsymbol{\delta}_{K-1} + \mathbf{\Gamma}_{Y,K}\boldsymbol{\Delta}_K + \mathbf{w}_K \end{cases}$$

with scalar

$$\rho_* := \sum_{\ell=1}^{K-1}\Big((\nabla_y\gamma_\ell)^\top\boldsymbol{\delta}_{\ell-1} + (\nabla_Y\gamma_\ell)^\top\boldsymbol{\Delta}_\ell\Big).$$

Similar to before,

$$\mathbf{\Gamma}_{y,K}\boldsymbol{\delta}_{K-1} + \mathbf{\Gamma}_{Y,K}\boldsymbol{\Delta}_K = \rho_K\Delta t^*\sum_{i=1}^{s} b_i\mathbf{F}_{K,i},$$

with scalar $\rho_K := (\nabla_y\gamma_K)^\top\boldsymbol{\delta}_{K-1} + (\nabla_Y\gamma_K)^\top\boldsymbol{\Delta}_K$.

The transpose of the Jacobian is

$$\left(\frac{d\widetilde{\mathbf{E}}}{d\overline{\mathbf{y}}}(\overline{\mathbf{y}})\right)^\top = \begin{pmatrix} \ddots & & & & \vdots \\ & \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N-\mathbf{\Gamma}_{y,K-1}^\top & \nabla_y\gamma_{K-1}\mathbf{F}_K^\top\mathbf{A}^\top \\ & & \mathbf{I}_{sN}-\mathbf{J}_{K-1}^\top\mathbf{A}_*^\top & -\gamma_{K-1}\mathbf{J}_{K-1}^\top\mathbf{B}_*-\mathbf{\Gamma}_{Y,K-1}^\top & \nabla_Y\gamma_{K-1}\mathbf{F}_K^\top\mathbf{A}^\top \\ & & & \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N-\mathbf{\Gamma}_{y,K}^\top \\ & & & & \mathbf{I}_{sN}-\mathbf{J}_K^\top\mathbf{A}_*^\top & -\gamma_K\mathbf{J}_K^\top\mathbf{B}_*-\mathbf{\Gamma}_{Y,K}^\top \\ & & & & & \mathbf{I}_N \end{pmatrix}$$

We now interpret the solution of $\widetilde{\mathbf{E}}'(\overline{\mathbf{y}})^\top \overline{\boldsymbol{\lambda}} = \overline{\mathbf{w}}$ via back substitution as a time stepping scheme. Again, the last step (or first step in reverse-time) is as given by the adjoint IDT formulas in lemma 2, but with $\Delta t \mapsto \Delta t^*$. The remaining $K-1$ steps, as given in lemma 4, are derived from the solution to the following systems for $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\lambda}_{k-1}$, with $\boldsymbol{\lambda}_k$ given by previous time step and $\boldsymbol{\Lambda}_K$ given by the last step,

$$\begin{pmatrix} \mathbf{I}_N & -\mathbf{C}^\top & -\mathbf{I}_N - \boldsymbol{\Gamma}_{y,k}^\top & \mathbf{0}_{N\times P} & \nabla_y \gamma_k^\top \mathbf{F}_K^\top \mathbf{A}^\top \\ & \mathbf{I}_{sN} - \mathbf{J}_k^\top \mathbf{A}^\top & -\gamma_k \mathbf{J}_k^\top \mathbf{B} - \boldsymbol{\Gamma}_{Y,k}^\top & \mathbf{0}_{sN\times P} & \nabla_Y \gamma_k^\top \mathbf{F}_K^\top \mathbf{A}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_{k-1} \\ \boldsymbol{\Lambda}_k \\ \boldsymbol{\lambda}_k \\ \vdots \\ \boldsymbol{\Lambda}_K \end{pmatrix} = \begin{pmatrix} \mathbf{w}_{k-1} \\ \mathbf{W}_k \end{pmatrix}$$

where $P = N(s+1)((K-1)-k)$, which gives

$$\implies \begin{cases} \boldsymbol{\lambda}_{k-1} = \sum_{i=1}^{s} \boldsymbol{\Lambda}_{k,i} + \boldsymbol{\lambda}_k + \boldsymbol{\Gamma}_{y,k}^\top \boldsymbol{\lambda}_k + \xi_* \nabla_y \gamma_k^\top + \mathbf{w}_{k-1}, \\ \boldsymbol{\Lambda}_{k,i} = \Delta t\, \mathbf{J}_{k,i}^\top \sum_{j=1}^{s} a_{ji} \boldsymbol{\Lambda}_{k,j} + \gamma_k b_i \Delta t\, \mathbf{J}_{k,i}^\top \boldsymbol{\lambda}_k + \boldsymbol{\Gamma}_{Y,k}^\top \boldsymbol{\lambda}_k - \xi_* \nabla_Y \gamma_k^\top + \mathbf{W}_{k,i}, \quad i = 1, ..., s, \end{cases}$$

with scalar

$$\xi_* := \Delta t \sum_{i=1}^{s} \sum_{j=1}^{s} a_{ji} \mathbf{F}_{K,i} \boldsymbol{\Lambda}_{K,j} = \mathbf{F}_K^\top \mathbf{A}^\top \boldsymbol{\Lambda}_K.$$

References

[1] D. I. Ketcheson, Relaxation Runge–Kutta methods: Conservation and stability for inner-product norms, SIAM Journal on Numerical Analysis 57 (6) (2019) 2850–2870.

[2] H. Ranocha, M. Sayyari, L. Dalcin, M. Parsani, D. I. Ketcheson, Relaxation Runge–Kutta methods: Fully discrete explicit entropy-stable schemes for the compressible Euler and Navier–Stokes equations, SIAM Journal on Scientific Computing 42 (2) (2020) A612–A638.

[3] H. Ranocha, L. Lóczi, D. I. Ketcheson, General relaxation methods for initial-value problems with application to multistep schemes, Numerische Mathematik 146 (4) (2020) 875–906.

[4] K. Rothauge, The discrete adjoint method for high-order time-stepping methods, Ph.D. thesis, University of British Columbia (2016).

[5] M. D. Gunzburger, Perspectives in flow control and optimization, SIAM, 2002.

[6] H. Antil, D. Leykekhman, A brief introduction to PDE-constrained optimization, in: Frontiers in PDE-constrained optimization, Springer, 2018, pp. 3–40.

[7] A. Griewank, A. Walther, Evaluating derivatives: principles and techniques of algorithmic differentiation, SIAM, 2008.

[8] A. Griewank, A mathematical view of automatic differentiation, Acta Numerica 12 (2003) 321–398.

[9] Z. Sirkes, E. Tziperman, Finite difference of adjoint or adjoint of finite difference?, Monthly weather review 125 (12) (1997) 3373–3378.

[10] W. W. Hager, Runge–Kutta methods in optimal control and the transformed adjoint system, Numerische Mathematik 87 (2) (2000) 247–282.

[11] A. Walther, Automatic differentiation of explicit Runge-Kutta methods for optimal control, Computational Optimization and Applications 36 (1) (2007) 83–108.

[12] A. Sandu, On the properties of Runge–Kutta discrete adjoints, in: International Conference on Computational Science, Springer, 2006, pp. 550–557.

[13] J. M. Sanz-Serna, Symplectic Runge–Kutta schemes for adjoint equations, automatic differentiation, optimal control, and more, SIAM review 58 (1) (2016) 3–33.

[14] A. Griewank, C. Bischof, G. Corliss, A. Carle, K. Williamson, Derivative convergence for iterative equation solvers, Optimization methods and software 2 (3-4) (1993) 321–355.

[15] P. Eberhard, C. Bischof, Automatic differentiation of numerical integration algorithms, Mathematics of Computation 68 (226) (1999) 717–731.

[16] M. Alexe, A. Sandu, On the discrete adjoints of adaptive time stepping algorithms, Journal of Computational and Applied Mathematics 233 (4) (2009) 1005–1020.

[17] E. Hairer, C. Lubich, G. Wanner, Geometric numerical integration: structure-preserving algorithms for ordinary differential equations, Vol. 31, Springer Science & Business Media, 2006.

[18] D. M. Hernandez, E. Bertschinger, Time-symmetric integration in astrophysics, Monthly Notices of the Royal Astronomical Society 475 (4) (2018) 5570–5584.

[19] H. Ranocha, D. I. Ketcheson, ConvexRelaxationRungeKutta. Relaxation Runge–Kutta methods for convex functionals, `https://github.com/ranocha/ConvexRelaxationRungeKutta` (05 2019). `doi:10.5281/zenodo.3066518`.

[20] J. Chan, On discretely entropy conservative and entropy stable discontinuous Galerkin methods, Journal of Computational Physics 362 (2018) 346–374.

[21] J. Chan, C. Taylor, Explicit Jacobian matrix formulas for entropy stable summation-by-parts schemes, arXiv preprint arXiv:2006.07504 (2020).

[22] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, Journal of computational physics 77 (2) (1988) 439–471.

[23] P.-O. Persson, High-order Navier–Stokes simulations using a sparse line-based discontinuous Galerkin method, in: 50th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition, 2012, p. 456.

[24] R. Alexander, Diagonally implicit Runge–Kutta methods for stiff ODE's, SIAM Journal on Numerical Analysis 14 (6) (1977) 1006–1021.

[25] L. C. Wilcox, G. Stadler, T. Bui-Thanh, O. Ghattas, Discretely exact derivatives for hyperbolic PDE-constrained optimization problems discretized by the discontinuous Galerkin method, Journal of Scientific Computing 63 (1) (2015) 138–162.