19 Runge–Kutta Methods

In this chapter, we introduce Runge–Kutta (RK) approximation methods for initial value problems (IVPs). These are single-step methods that have multiple *stages*, and the form and function of these methods are rather distinct from those we have previously defined. For simplicity, we will largely neglect the convergence theory of RK methods, though it would not present any new technical difficulties, only tedious calculations.

We begin by recalling that we are trying to approximate the solution to (18.1). To achieve this, as before, we introduce a discretization of the time interval $[0, T]$ via the following procedure. Let $K \in \mathbb{N}$, $\tau = \frac{T}{K}$, and $t_k = \tau k$ for $k = 0, \ldots, K$. As before, we will produce a sequence $\{\boldsymbol{w}^k\}_{k=0}^K$ such that $\boldsymbol{u}(t_k) \approx \boldsymbol{w}^k$.

The main motivation behind these methods can be explained by taking a second look at Taylor's method (18.6), where the slope approximation function is given by

$$\boldsymbol{G}_{TM}(t, s, \boldsymbol{v}_1) = \boldsymbol{f}(t, \boldsymbol{v}_1) + \frac{s}{2}\left[\partial_t \boldsymbol{f}(t, \boldsymbol{v}_1) + D_{\boldsymbol{u}} \boldsymbol{f}(t, \boldsymbol{v}_1)\boldsymbol{f}(t, \boldsymbol{v}_1)\right].$$

Clearly, this approximation comes from an approximation of $\boldsymbol{u}'(t+s)$ via a second-order Taylor expansion. As we saw in Theorem 18.11, this method is convergent with rate $p = 2$. While this is a perfectly acceptable method, it has one major drawback. Namely, it requires knowledge not only of the slope function $\boldsymbol{f}$ but also of its partial derivatives with respect to time, $\partial_t \boldsymbol{f}$, and $\boldsymbol{u}$, $D_{\boldsymbol{u}} \boldsymbol{f}$. In practice, these functions may not be available, or they may be very difficult to compute.

If we are to allow dealing with derivatives of the slope function, the next logical step may be trying to devise a method, via a Taylor expansion, that is consistent to order at least $p = 3$. The procedure is clear. We do a Taylor expansion of $\boldsymbol{u}(t+s)$ about the point t and use (18.1) to arrive at

$$\begin{aligned}
\boldsymbol{u}(t+s) &= \boldsymbol{u}(t) + s\boldsymbol{u}'(t) + \frac{s^2}{2}\boldsymbol{u}''(t) + \frac{s^3}{6}\boldsymbol{u}''(t) + \mathcal{O}(|s|^4) \\
&\approx \boldsymbol{u}(t) + s\left\{\boldsymbol{f} + \frac{s}{2}\left[\partial_t \boldsymbol{f} + D_{\boldsymbol{u}} \boldsymbol{f}\boldsymbol{f}\right]\right. \\
&\quad \left. + \frac{s^2}{6}\left[\partial_t^2 \boldsymbol{f} + D_{\boldsymbol{u}} \partial_t \boldsymbol{f} \cdot \boldsymbol{f} + D_{\boldsymbol{u}} \boldsymbol{f}\left(\partial_t \boldsymbol{f} + D_{\boldsymbol{u}} \boldsymbol{f}\boldsymbol{f}\right) + \left(\partial_t D_{\boldsymbol{u}} \boldsymbol{f} + D_{\boldsymbol{u}}^2 \boldsymbol{f} \cdot \boldsymbol{f}\right)\boldsymbol{f}\right]\right\},
\end{aligned}$$

where, for simplicity, we have suppressed the arguments of the slope function and its derivatives. Hopefully, the reader appreciates the difficulties we have encountered. Not only will such a procedure require evaluating many derivatives of the slope function, but also the number of terms that is involved grows extremely large.

The idea behind RK methods is that, instead of differentiating the slope function, we evaluate it at a special collection of points in $[t_k, t_{k+1}] \times \bar{\Omega}$ called *stages*, so

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

that, for instance, in the case of a two-stage RK method we have

$$\boldsymbol{\kappa}_1 = \boldsymbol{f}(t_k, \boldsymbol{w}^k), \qquad \boldsymbol{\kappa}_2 = \boldsymbol{f}(t_k + c\tau, \boldsymbol{w}^k + a\tau\boldsymbol{\kappa}_1);$$

then the approximation at t_{k+1} is given by a *linear combination* of the stages:

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau\,(b_1\boldsymbol{\kappa}_1 + b_2\boldsymbol{\kappa}_2).$$

By properly choosing the coefficients a, c, b_1, and b_2, a method that is consistent to order at least $p = 2$ (like Taylor's method) can be obtained.

19.1 Simple Two-Stage Methods

Before we give a general definition of RK methods, let us elaborate on the previous idea for a simple, scalar, autonomous problem. We leave it to the reader to show that the following fits the general definition given later.

Theorem 19.1 (RK2)**.** *Let $T > 0$ be given. Consider the general two-stage explicit RK method, defined by*

$$\xi^k = w^k + a\tau f(w^k), \quad w^{k+1} = w^k + \tau\big[b_1 f(w^k) + b_2 f(\xi^k)\big],$$

for approximating the solution to the scalar autonomous IVP

$$u'(t) = f(u(t)), \quad t \in [0, T], \quad u(0) = u_0.$$

Assume that $f \in \mathcal{F}^2(S)$ and, therefore, $u \in C^3([0, T])$. If the coefficients satisfy $b_1 + b_2 = 1$, $b_1, b_2 \geq 0$, and $ab_2 = \frac{1}{2}$, then the method is consistent of order $p = 2$ and the method is convergent to second order.

Proof. Let us first show consistency. By Taylor's Theorem, for some ζ between $u(t-\tau)$ and $u(t-\tau) + a\tau f(u(t-\tau))$,

$$\begin{aligned} f\big(u(t-\tau) + a\tau f(u(t-\tau))\big) &= f(u(t-\tau)) + a\tau f(u(t-\tau))f'(u(t-\tau)) \\ &\quad + \frac{1}{2}\big(a\tau f(u(t-\tau))\big)^2 f''(\zeta). \end{aligned}$$

Upon setting $b_1 + b_2 = 1$ and $ab_2 = \frac{1}{2}$, the LTE satisfies

$$\begin{aligned} \mathcal{E}[u](t,\tau) &= \frac{u(t) - u(t-\tau)}{\tau} \\ &\quad - \big[b_1 f(u(t-\tau)) - b_2 f\big(u(t-\tau) + a\tau f(u(t-\tau))\big)\big] \\ &= \frac{u(t) - u(t-\tau)}{\tau} - b_1 f(u(t-\tau)) - b_2 f(u(t-\tau)) \\ &\quad - \tau a b_2 f(u(t-\tau))f'(u(t-\tau)) - \tau^2 \frac{a^2 b_2}{2} f^2(u(t-\tau))f''(\zeta) \\ &= \frac{u(t) - u(t-\tau)}{\tau} - f(u(t-\tau)) \\ &\quad - \frac{\tau}{2} f(u(t-\tau))\, f'(u(t-\tau)) - \tau^2 \frac{a^2 b_2}{2} f^2(u(t-\tau))f''(\zeta). \end{aligned}$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

On the other hand, using Taylor's Theorem, the exact solution must satisfy

$$\begin{aligned} u(t) &= u(t-\tau) + \tau u'(t-\tau) + \frac{\tau^2}{2}u''(t-\tau) + \frac{\tau^3}{6}u'''(\sigma) \\ &= u(t-\tau) + \tau f(u(t-\tau)) + \frac{\tau^2}{2}f'(u(t-\tau))f(u(t-\tau)) + \frac{\tau^3}{6}u'''(\sigma) \end{aligned}$$

for some $\sigma \in (t-\tau, t)$. Comparing the expansions,

$$\mathcal{E}[u](t,\tau) = \frac{\tau^2}{6}u'''(\sigma) - \tau^2\frac{a^2 b_2}{2}f^2(u(t-\tau))f''(\zeta)$$

provided that $f \in C^2((-\infty,\infty))$, $u \in C^3([0,T])$, $b_1 + b_2 = 1$, and $ab_2 = \frac{1}{2}$. There is some $C > 0$ such that

$$|\mathcal{E}[u](t,\tau)| \le C\tau^2$$

for any $\tau \in (0,T]$ and $t \in [\tau, T]$. The proof for this is subtle and relies on the fact that u is bounded over $[0,T]$.

We will only show convergence in the case that $a = 1/2$, $b_1 = 0$, and $b_2 = 1$, leaving the general case to the reader as an exercise; see Problem 19.1. With this simplification, the method reads

$$w^{k+1} = w^k + \tau f\left(w^k + \frac{\tau}{2}f(w^k)\right).$$

The exact solution satisfies

$$u(t_{k+1}) = u(t_k) + \tau f\left(u(t_k) + \frac{\tau}{2}f(u(t_k))\right) + \tau\mathcal{E}^{k+1}[u].$$

Therefore,

$$e^{k+1} = e^k + \tau f\left(u(t_k) + \frac{\tau}{2}f(u(t_k))\right) - \tau f\left(w^k + \frac{\tau}{2}f(w^k)\right) + \tau\mathcal{E}^{k+1}[u].$$

Taking absolute values and using the triangle inequality and the Lipschitz continuity of the slope function f, we have

$$\begin{aligned} |e^{k+1}| &\le |e^k| + \tau L\left|e^k + \frac{\tau}{2}(f(u(t_k)) - f(w^k))\right| + \tau\left|\mathcal{E}^{k+1}[u]\right| \\ &\le (1+\tau L)|e^k| + \frac{\tau^2 L^2}{2}|e^k| + C\tau^3 \\ &= \left(1 + \tau L + \frac{\tau^2 L^2}{2}\right)|e^k| + C\tau^3. \end{aligned}$$

Using the discrete Grönwall inequality from Lemma 18.3,

$$|e^k| \le \frac{C\tau^2}{L + \frac{\tau L^2}{2}}\left[\left(1 + \tau L + \frac{\tau^2 L^2}{2}\right)^k - 1\right].$$

Now, since $\tau L > 0$,

$$1 + \tau L + \frac{\tau^2 L^2}{2} < e^{\tau L};$$

therefore, for any $m = 1, \ldots, K$,

$$(1+\tau L)^m < e^{m\tau L} \le e^{K\tau L} = e^{TL},$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

where we used that $K\tau = T$. It follows that, for all $k = 0, \ldots, K$,

$$|e^k| \leq \frac{C}{L}\left[e^{TL} - 1\right]\tau^2. \qquad \square$$

19.2 General Definition and Basic Properties

We now embark upon the study of general RK methods and their properties. We begin with their definition.

Definition 19.2 (RK)**.** Let $r \in \mathbb{N}$. A general r**-stage Runge–Kutta method** (RK method)[1] is a recursive algorithm for generating an approximation $\{\boldsymbol{w}^k\}_{k=0}^K$ to the solution of (18.1), via $\boldsymbol{w}^0 = \boldsymbol{u}_0$ and, for $k = 0, \ldots, K-1$,

$$\boldsymbol{\xi}_i = \boldsymbol{w}^k + \tau \sum_{j=1}^{r} a_{i,j} \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j), \quad i = 1, \ldots, r, \tag{19.1}$$

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau \sum_{j=1}^{r} b_j \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j). \tag{19.2}$$

Here, $a_{i,j} \in \mathbb{R}$ and $b_j, c_j \in [0,1]$ for $i, j = 1, \ldots, r$. An RK method is completely determined by its weights $\mathsf{A} = [a_{i,j}]_{i,j=1}^r \in \mathbb{R}^{r\times r}$, $\boldsymbol{b} = [b_i]_{i=1}^r \in \mathbb{R}^r$, and $\boldsymbol{c} = [c_i]_{i=1}^r \in \mathbb{R}^r$, which are often expressed in **tableau** form

$$\begin{array}{c|c} \boldsymbol{c} & \mathsf{A} \\ \hline & \boldsymbol{b}^{\mathsf{T}} \end{array},$$

which is commonly referred to as the **Butcher tableau**[2] of the method. The RK method is called **explicit** (ERK) if and only if $a_{i,j} = 0$ for all $i \leq j$, and is called **implicit** (IRK) otherwise. The RK method is called **diagonally implicit** (DIRK) if and only if $a_{i,j} = 0$ for all $i < j$.

Remark 19.3 (equivalent definition)**.** As we mentioned above, there is an alternate, but equivalent, definition of the general r-stage RK method. Let us write out this equivalent form. First, define

$$\boldsymbol{\kappa}_j = \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j), \quad j = 1, \ldots, r.$$

Then, from (19.2),

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau \sum_{j=1}^{r} b_j \boldsymbol{\kappa}_j, \tag{19.3}$$

where, from (19.1),

$$\boldsymbol{\kappa}_i = \boldsymbol{f}\left(t_k + c_i\tau, \boldsymbol{w}^k + \tau \sum_{j=1}^{r} a_{i,j}\boldsymbol{\kappa}_j\right), \quad i = 1, \ldots, r. \tag{19.4}$$

[1] Named in honor of the German mathematicians Carl David Tolmé Runge (1856–1927) and Martin Wilhelm Kutta (1867–1944).
[2] Named in honor of the New Zealand mathematician John Charles Butcher (1933–).

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

The following theorem describes constraints on the weights of an RK method, so that the method satisfies certain consistency requirements.

Theorem 19.4 (properties of weights)**.** *Assume that* $\boldsymbol{f} \in \mathcal{F}^1(S)$*. Consider the general* r*-stage RK method given by the weights* $\mathsf{A} = [a_{i,j}]_{i,j=1}^r \in \mathbb{R}^{r\times r}$*,* $\boldsymbol{b} = [b_i]_{i=1}^r \in [0,1]^r$*, and* $\boldsymbol{c} = [c_i]_{i=1}^r \in [0,1]^r$*. Let* $\mathbf{1} = [1]_{i=1}^r \in \mathbb{R}^r$*.*

1. *For the method to be at least first order, it is necessary that*
$$\boldsymbol{b}^\intercal \mathbf{1} = 1.$$
2. *For the* j*th RK stage* $\boldsymbol{\xi}_j$ *to be at least a first-order approximation of* $\boldsymbol{u}(t_k + c_j\tau)$*, it is necessary that*
$$\mathsf{A}\mathbf{1} = \boldsymbol{c}. \tag{19.5}$$
3. *Suppose that* $\boldsymbol{f} \in \mathcal{F}^2(S)$ *and* (19.5) *holds. For the method to be at least second order, it is necessary that*
$$\boldsymbol{b}^\intercal \boldsymbol{c} = \frac{1}{2}.$$
4. *Suppose that* $\boldsymbol{f} \in \mathcal{F}^3(S)$ *and* (19.5) *holds. For the method to be at least third order, it is necessary that*
$$\boldsymbol{b}^\intercal \mathsf{A}\boldsymbol{c} = \frac{1}{6}.$$

Proof. We sketch the proof and leave the details to the reader; see Problem 19.2. To prove the result, use the general r-step method to approximate the solution to the linear scalar problem $u'(t) = u(t)$, $u(0) = 1$, whose exact solution is $u(t) = e^t$. At time $t = \tau$, the solution may be expressed as

$$u(\tau) = 1 + \tau + \frac{\tau^2}{2} + \frac{\tau^3}{6} + \frac{\tau^4}{24}e^\eta$$

for some $\eta \in (0,\tau)$. For the RK stages, assume that the matrix $\mathsf{I} - \tau\mathsf{A}$ is invertible — this will always will be the case provided that τ is sufficiently small — and

$$(\mathsf{I} - \tau\mathsf{A})^{-1} = \mathsf{I} + \tau\mathsf{A} + \tau^2\mathsf{A}^2 + \tau^3\mathsf{A}^3 + \cdots.$$

It is possible to show that the vector of stages satisfies

$$\boldsymbol{\xi} = (\mathsf{I} - \tau\mathsf{A})^{-1}\mathbf{1}.$$

Solve explicitly for w^1 and compare the result with the expansion $u(\tau)$ above. □

The following expressions for the ERK and IRK approximations, respectively, of $u' = \lambda u$ will be needed in subsequent chapters.

Theorem 19.5 (amplification factor I)**.** *Applying an* r*-stage explicit RK method to approximate the solution of the differential equation* $u'(t) = \lambda u(t)$*,* $u(0) = u_0$*, one obtains*

$$w^{k+1} = g(\lambda\tau)w^k, \quad g(z) = \sum_{j=0}^{r} \beta_j z^j, \quad w^0 = u_0,$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

where $\beta_j \in \mathbb{R}$, $j = 0, \dots, r$. If the method is consistent to exactly order r, then

$$g(z) = \sum_{j=0}^{r} \frac{z^j}{j!},$$

i.e., $\beta_j = \frac{1}{j!}$ *for* $j = 0, \dots, r$.

Proof. See Problem 19.3. □

Theorem 19.6 (amplification factor II)**.** *Applying an r-stage implicit RK method to approximate the solution of the differential equation $u'(t) = \lambda u(t)$, $u(0) = u_0$, one obtains*

$$w^{k+1} = g(\lambda\tau)w^k, \quad g(z) = \frac{p_1(z)}{p_2(z)}, \quad w^0 = u_0,$$

where $p_1, p_2 \in \mathbb{P}_r$ and $p_2 \not\equiv 0$. In particular, g is the rational polynomial

$$g(z) = 1 + z\boldsymbol{b}^{\mathsf{T}}(\mathsf{I} - z\mathsf{A})^{-1}\mathbf{1} = \frac{\det(\mathsf{I} - z\mathsf{A} + z\mathbf{1}\,\boldsymbol{b}^{\mathsf{T}})}{\det(\mathsf{I} - z\mathsf{A})},$$

where $\mathbf{1} = [1]_{i=1}^{r} \in \mathbb{R}^r$.

Proof. See Problem 19.4. □

Remark 19.7 (amplification factor)**.** The function g (a polynomial in the ERK case and a rational function in the IRK case) that appears in Theorems 19.5 and 19.6 is called the *linear amplification factor* or just the *amplification factor*.

Remark 19.8 (LTE)**.** The *consistency error* for any explicit RK method is defined in a straightforward way, since the RK stages can be computed explicitly in terms of the approximations. Specifically,

$$\tau\boldsymbol{\mathcal{E}}[\boldsymbol{u}](t,s) = \boldsymbol{u}(t) - \boldsymbol{u}(t-s) - s\sum_{i=1}^{r} b_i \boldsymbol{f}(t - \tau + c_i\tau, \boldsymbol{\xi}_{e,i}),$$

where $\boldsymbol{\xi}_{e,1} = \boldsymbol{u}(t-\tau)$ and, for $i = 2, \dots, r$,

$$\boldsymbol{\xi}_{e,i} = \boldsymbol{u}(t-\tau) + \tau\sum_{j=1}^{i-1} a_{i,j}\boldsymbol{f}(t - \tau + c_j\tau, \boldsymbol{\xi}_{e,j}).$$

Notice that, in the end, the $\boldsymbol{\xi}_{e,i}$ can be completely eliminated, which is a key insight. For implicit RK methods, the situation is a bit more complicated.

We have already encountered the following, but in a slightly simpler form, in Theorem 19.1.

Theorem 19.9 (two-stage RK method)**.** *Suppose that $\boldsymbol{f} \in \mathcal{F}^2(S)$, so that $\boldsymbol{u} \in C^3\big([0,T];\mathbb{R}^d\big)$ is a classical solution to the IVP* (18.1)*. Consider an explicit two-stage RK method given by the tableau*

$$\begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & a_{2,1} & 0 \\ \hline & b_1 & b_2 \end{array}.$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

The method is consistent to order $p = 2$ if and only if

$$b_1 + b_2 = 1, \quad a_{2,1} = c_2, \quad b_2 c_2 = \frac{1}{2}.$$

Proof. For simplicity of notation, we will give a proof in the scalar case ($d = 1$).

We begin with a two-dimensional Taylor expansion of $f \in C^2(S)$. Set $\boldsymbol{q} = [c_2, a_{2,1} f(t-\tau, u(t-\tau))]^\mathsf{T}$. Then

$$\begin{aligned} f\big(t-\tau+c_2\tau, u(t-\tau) + a_{2,1}\tau f(t-\tau, u(t-\tau))\big) \\ = f(t-\tau, u(t-\tau)) + \tau q_1 \partial_t f(t-\tau, u(t-\tau)) \\ + \tau q_2 \partial_u f(t-\tau, u(t-\tau)) + \frac{\tau^2}{2} \boldsymbol{q}^\mathsf{T} \mathsf{H}_f(\eta, \gamma) \boldsymbol{q}, \end{aligned} \tag{19.6}$$

where H_f is the 2×2 Hessian matrix of second derivatives of f, η is some number between $t-\tau$ and $t-\tau+c_2\tau$, and γ is some number between $u(t-\tau)$ and $u(t-\tau) + a_{2,1}\tau f(t-\tau, u(t-\tau))$. For $t \in [\tau, T]$, it follows that, since the solution u is twice continuously differentiable,

$$|\boldsymbol{q}^\mathsf{T} \mathsf{H}_f(\eta, \gamma) \boldsymbol{q}| \le C,$$

where $C > 0$ is independent of t.

Using the expansion (19.6) above, the local truncation error, which is defined as usual, may be expressed as

$$\begin{aligned} \tau \mathcal{E}[u](t,\tau) &= u(t) - u(t-\tau) - \tau b_1 f(t-\tau, u(t-\tau)) \\ &\quad - \tau b_2 f\big(t-\tau+c_2\tau, u(t-\tau) + \tau a_{2,1} f(t-\tau, u(t-\tau))\big) \\ &= u(t) - u(t-\tau) - \tau(b_1+b_2) f(t-\tau, u(t-\tau)) \\ &\quad - \tau^2 b_2 c_2 \partial_t f(t-\tau, u(t-\tau)) \\ &\quad - \tau^2 b_2 a_{2,1} \partial_u f(t-\tau, u(t-\tau)) f(t-\tau, u(t-\tau)) \\ &\quad - \tau^3 \frac{b_2}{2} \boldsymbol{q}^\mathsf{T} \mathsf{H}_f(\eta, \gamma) \boldsymbol{q}. \end{aligned} \tag{19.7}$$

On the other hand, applying Taylor's Theorem to the solution, we have, for some $\beta \in (t-\tau, t)$,

$$\begin{aligned} 0 &= u(t) - u(t-\tau) - \tau u'(t-\tau) - \frac{\tau^2}{2} u''(t-\tau) - \frac{\tau^3}{6} u'''(\beta) \\ &= u(t) - u(t-\tau) - \tau f(t-\tau, u(t-\tau)) - \frac{\tau^2}{2} \partial_t f(t-\tau, u(t-\tau)) \\ &\quad - \frac{\tau^2}{2} \partial_u f(t-\tau, u(t-\tau)) f(t-\tau, u(t-\tau)) - \frac{\tau^3}{6} u'''(\beta). \end{aligned} \tag{19.8}$$

($\Longrightarrow$) Suppose that $b_1 + b_2 = 1$, $a_{2,1} = c_2$, $b_2 c_2 = \frac{1}{2}$ in (19.7). Then, combining (19.7) and (19.8),

$$\tau \mathcal{E}[u](t,\tau) = \frac{\tau^3}{6} u'''(\beta) - \tau^3 \frac{b_2}{2} \boldsymbol{q}^\mathsf{T} \mathsf{H}_f(\eta, \gamma) \boldsymbol{q}$$

and the method is clearly consistent to order $p = 2$.

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

($\Longleftarrow$) On the other hand, according to Theorem 19.4, for the method to be consistent to exactly order $p = 2$, it must be that $b_1 + b_2 = 1$, $a_{2,1} = c_2$, $b_2 c_2 = \frac{1}{2}$. Otherwise, the method would be of order $p = 1$, or, perhaps, inconsistent. □

The following three explicit two-stage RK methods are consistent to order $p = 2$ and conform to Theorem 19.9.

Example 19.1 Midpoint method:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}.$$

Example 19.2 Heun's method:[3]

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

Example 19.3 Ralston's method:[4]

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}.$$

Theorem 19.10 (three-stage ERK methods)**.** *Suppose that* $\boldsymbol{f} \in \mathcal{F}^3(S)$*, so that* $\boldsymbol{u} \in C^4\left([0,T];\mathbb{R}^d\right)$ *is a classical solution to the IVP* (18.1)*. Consider an explicit three-stage RK method given by the tableau*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & a_{2,1} & 0 & 0 \\ c_3 & a_{3,1} & a_{3,2} & 0 \\ \hline & b_1 & b_2 & b_3 \end{array}.$$

The method is consistent to order $p = 3$ *if and only if*

$$b_1 + b_2 + b_3 = 1, \quad b_2 c_2 + b_3 c_3 = \frac{1}{2}, \quad b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3}, \quad b_3 a_{3,2} c_2 = \frac{1}{6}.$$

Proof. The proof can be found in [12]; see also [47]. □

[3] Named in honor of the German mathematician Karl Heun (1859–1929).
[4] Named in honor of the American mathematician Anthony Ralston (1930–).

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

Example 19.4 The following three-stage explicit RK method is consistent to exactly order $p = 3$:

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

This method is called the classical RK method.

Example 19.5 The following four-stage explicit RK method is consistent to exactly order $p = 4$:

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

For a proof of the consistency, see [12]. The usual approach, which can be quite tedious, is to use Taylor expansions to prove the result.

19.3 Collocation Methods

In this section we introduce collocation methods, which form a very general class of numerical methods for differential equations. We show that these are related to RK methods, in certain cases.

Definition 19.11 (RK collocation method)**.** Let $\boldsymbol{f} \in C\left(S; \mathbb{R}^d\right)$. Suppose that the so-called **collocation points** satisfy

$$0 \leq c_1 < c_2 < \cdots < c_r \leq 1.$$

Let $\boldsymbol{w}^k \in \mathbb{R}^d$ be given. Assume that $\boldsymbol{p}_k \in [\mathbb{P}_r]^d$ satisfies, if possible,

$$\boldsymbol{p}_k(t_k) = \boldsymbol{w}^k, \quad \boldsymbol{p}_k'(t_k + c_j\tau) = \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{p}_k(t_k + c_j\tau)) \tag{19.9}$$

for $j = 1, \ldots, r$. Define $\boldsymbol{w}^{k+1} = \boldsymbol{p}_k(t_{k+1})$, for $k = 0, \ldots, K-1$, with $\boldsymbol{w}^0 = \boldsymbol{u}_0$. This algorithm for producing the approximation sequence $\left\{\boldsymbol{w}^k\right\}_{k=0}^K \subset \mathbb{R}^d$ is called a **Runge–Kutta collocation method**.

Remark 19.12 (existence)**.** The previous definition only makes sense if we can find a vector-valued polynomial

$$\boldsymbol{p}_k(t) = \sum_{j=0}^{r} \boldsymbol{a}_j t^j, \quad \boldsymbol{a}_j \in \mathbb{R}^d, \quad j = 0, \ldots, r$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

that satisfies (19.9). If so, we say that the implicit r-stage RK collocation method is *well defined*. In fact, it may be the case that such a polynomial will not exist or will not be uniquely determined unless $\tau > 0$ is sufficiently small.

Theorem 19.13 (collocation)**.** *Let $\{c_j\}_{j=1}^r \subset [0,1]$ be a set of distinct collocation points. Suppose that a unique polynomial $\boldsymbol{p}_k \in [\mathbb{P}_r]^d$ satisfying* (19.9) *exists. Define*

$$\begin{aligned}
\boldsymbol{\xi}_i &= \boldsymbol{p}_k(t_k + c_i\tau), \quad i = 1, \dots, r,\\
L_j(t) &= \prod_{\substack{i=1\\ i\neq j}}^{r} \frac{(t - c_i)}{(c_j - c_i)}, \quad j = 1, \dots, r,\\
a_{i,j} &= \int_0^{c_i} L_j(s)\mathrm{d}s, \quad b_j = \int_0^1 L_j(s)\mathrm{d}s, \quad i, j = 1, \dots, r.
\end{aligned}$$

Then the collocation method of Definition 19.11 is a standard implicit RK method, as in Definition 19.2, with the weights $\mathsf{A} = [a_{i,j}]$, $\boldsymbol{b} = [b_j]$, and $\boldsymbol{c} = [c_j]$, the last weights being precisely the collocation points.

Proof. Suppose that $\boldsymbol{p}_k \in [\mathbb{P}_r]^d$ satisfies (19.9). Consider the unique Lagrange interpolating polynomial of degree at most $r-1$, $\boldsymbol{\rho} \in [\mathbb{P}_r]^d$ such that

$$\boldsymbol{\rho}(t_k + c_j\tau) = \boldsymbol{p}_k'(t_k + c_j\tau) = \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{p}_k(t_k + c_j\tau)) = \boldsymbol{\nu}_j, \quad j = 1, \dots, r.$$

Theorem 9.11 guarantees that

$$\boldsymbol{\rho}(t) = \sum_{j=1}^{r} L_j\left(\frac{t - t_k}{\tau}\right)\boldsymbol{\nu}_j.$$

Observe that $\boldsymbol{p}_k' \in [\mathbb{P}_{r-1}]^d$ and, in fact,

$$\boldsymbol{p}_k'(t_k + c_j\tau) = \boldsymbol{\rho}(t_k + c_j\tau), \quad j = 1, \dots, r.$$

Therefore, $\boldsymbol{p}_k' \equiv \boldsymbol{\rho}$, since these polynomials (of degree at most $r-1$) agree at r points. By (19.9),

$$\boldsymbol{p}_k'(t) = \sum_{j=1}^{r} L_j\left(\frac{t - t_k}{\tau}\right)\boldsymbol{f}(t_k + c_j\tau, \boldsymbol{p}_k(t_k + c_j\tau)).$$

Integrating the last expression and using the condition $\boldsymbol{p}_k(t_k) = \boldsymbol{w}^k$, we observe that

$$\begin{aligned}
\boldsymbol{p}_k(t) &= \boldsymbol{w}^k + \int_{t_k}^{t} \sum_{j=1}^{r} L_j\left(\frac{s - t_k}{\tau}\right)\boldsymbol{f}(t_k + c_j\tau, \boldsymbol{p}_k(t_k + c_j\tau))\mathrm{d}s\\
&= \boldsymbol{w}^k + \tau \sum_{j=1}^{r} \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j) \int_0^{\frac{t-t_k}{\tau}} L_j(s)\mathrm{d}s.
\end{aligned} \tag{19.10}$$

Setting $t = t_k + c_i\tau$ in (19.10), we have

$$\boldsymbol{\xi}_i = \boldsymbol{w}^k + \tau \sum_{j=1}^{r} a_{i,j}\boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j).$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

Setting $t = t_{k+1}$ in (19.10), we find

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau \sum_{j=1}^{r} b_j \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j),$$

and the proof is finished. □

For collocation methods, everything is determined by picking the collocation points. These are usually chosen as the roots of certain orthogonal polynomials.

Theorem 19.14 (collocation order)**.** *Suppose that* $\{c_j\}_{j=1}^r \subset [0,1]$ *is the set of r distinct collocation points that determine the RK collocation method of Definition 19.11. Define*

$$q(t) = \prod_{n=1}^{r}(t - c_n) \in \mathbb{P}_r.$$

If, for some $m \in \{1, \dots, r\}$,

$$\int_0^1 q(s)p(s)\mathrm{d}s = 0, \quad \forall p \in \mathbb{P}_{m-1},$$

but there is some $\tilde{p} \in \mathbb{P}_m$ *such that*

$$\int_0^1 q(s)\tilde{p}(s)\mathrm{d}s \neq 0,$$

then the RK collocation method is consistent to exactly order $p = r + m$.

Proof. Consider the simple quadrature rule,

$$Q_1^{(0,1)}[f] = \sum_{j=1}^{r} b_j f(c_j),$$

with quadrature nodes $\{c_j\}_{j=1}^r$ and quadrature weights $\{b_j\}_{j=1}^r$. Suppose that the quadrature weights b_j are chosen to satisfy the exactness condition (14.7); namely,

$$b_j = \int_0^1 L_j(s)\mathrm{d}s, \quad L_j(s) = \prod_{\substack{k=1 \\ k\neq j}}^{r} \frac{s - c_k}{c_j - c_k}, \quad j = 1, \dots, r.$$

Appealing to Corollary 14.47, the quadrature rule is consistent of order exactly $r + m - 1$. In other words,

$$\int_0^1 \psi(s)\mathrm{d}s = \sum_{j=1}^{r} b_j \psi(c_j), \quad \forall \psi \in \mathbb{P}_{r+m-1}.$$

By the Alekseev–Gröbner Lemma, stated in Theorem 17.16,

$$\boldsymbol{p}_k(t_{k+1}) - \boldsymbol{u}(t_{k+1}) = \int_{t_k}^{t_{k+1}} D_{\boldsymbol{v}}\boldsymbol{U}(s, \boldsymbol{p}_k(s), t_{k+1})\boldsymbol{g}(s, \boldsymbol{p}_k(s))\mathrm{d}s,$$

where $\boldsymbol{p}_k$ is the vector-valued polynomial in the definition of the collocation method, assuming that $\boldsymbol{p}_k(t_k) = \boldsymbol{u}(t_k)$ and that the deviation, $\boldsymbol{g}$, in the Alekseev–Gröbner Lemma satisfies

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

j	$\tilde{P}_j(t)$
0	1
1	$2t-1$
2	$6t^2-6t+1$
3	$20t^3-30t^2+12t-1$
4	$70t^4-140t^3+90t^2-20t+1$
5	$252t^5-630t^4+560t^3-210t^2+30t-1$

Table 19.1 The first six transformed Legendre polynomials.

$$\boldsymbol{g}(s,\boldsymbol{p}_k(s)) = \boldsymbol{p}_k'(s) - \boldsymbol{f}(s,\boldsymbol{p}_k(s)) = \tilde{\boldsymbol{g}}(s).$$

Observe that the collocation rule requires that

$$\boldsymbol{p}_k'(t_k + c_j\tau) = \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{p}_k(t_k + c_j\tau)), \quad j = 1,\dots,r,$$

which implies that $\tilde{\boldsymbol{g}}$ vanishes at the collocation points:

$$\tilde{\boldsymbol{g}}(t_k + c_j\tau) = \boldsymbol{0}, \quad j = 1,\dots,r.$$

Applying the quadrature rule,

$$\begin{aligned}\boldsymbol{p}_k(t_{k+1}) - \boldsymbol{u}(t_{k+1}) &= \int_{t_k}^{t_{k+1}} D_{\boldsymbol{v}}\boldsymbol{U}(s,\boldsymbol{p}_k(s),t_{k+1})\boldsymbol{g}(s,\boldsymbol{p}_k(s))\mathrm{d}s \\ &= \sum_{j=1}^{r} b_j D_{\boldsymbol{v}}\boldsymbol{U}(t_k+\tau c_j, \boldsymbol{p}_k(t_k+\tau c_j), t_{k+1})\tilde{\boldsymbol{g}}(t_k+\tau c_j) + \boldsymbol{E}_Q \\ &= \boldsymbol{E}_Q,\end{aligned}$$

where $\boldsymbol{E}_Q$ denotes the quadrature error. If we assume that

$$\hat{\boldsymbol{g}}(\cdot) = D_{\boldsymbol{v}}\boldsymbol{U}(\cdot, p_k(\cdot), t_{k+1})\tilde{\boldsymbol{g}}(\cdot) \in C^{r+m}([t_k, t_{k+1}]; \mathbb{R}^d),$$

then, by Theorem 14.18,

$$\|\boldsymbol{E}_Q\|_2 \le C\tau^{r+m+1}\left\|\hat{\boldsymbol{g}}^{(r+m)}\right\|_{L^\infty(t_k,t_{k+1};\mathbb{R}^d)}$$

for some constant $C > 0$. Hence,

$$\left\|\boldsymbol{w}^{k+1} - \boldsymbol{u}(t_{k+1})\right\|_2 = \|\boldsymbol{p}_k(t_{k+1}) - \boldsymbol{u}(t_{k+1})\|_2 \le \tilde{C}\tau^{r+m+1},$$

where we assume that $\boldsymbol{w}^k = \boldsymbol{u}(t_k)$. We leave it to the reader as an exercise to prove that the local truncation error must be of order $r+m$; see Problem 19.6. □

Definition 19.15 (transformed Legendre polynomials)**.** By $\left\{\tilde{P}_j\right\}_{j\in\mathbb{N}_0}$ we denote the set of **transformed Legendre polynomials**,[5] which have the property that

$$\int_0^1 \tilde{P}_i(s)\tilde{P}_j(s)\mathrm{d}s = \frac{1}{2j+1}\delta_{i,j}.$$

The first few transformed Legendre polynomials are given in Table 19.1.

[5] Named in honor of the French mathematician Adrien-Marie Legendre (1752–1833).

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

Corollary 19.16 (Gauss–Legendre–RK)**.** *Let the collocation points $c_1, \ldots, c_r$ be precisely the zeros of the transformed Legendre polynomial $\tilde{P}_r \in \mathbb{P}_r$. According to Theorem 11.10, these lie in the open interval $(0, 1)$. Then the corresponding collocation method of Definition 19.11 is consistent to order exactly $p = 2r$.*

Proof. In this case, $q \equiv C_r \tilde{P}_r$, where $0 \neq C_r \in \mathbb{R}$. Since the $\tilde{P}_i$ form an orthogonal basis for $\mathbb{P}_r$, for any $j \in \{0, 1, 2, \ldots, r\}$, we can express

$$t^j = \sum_{m=0}^{j} \beta_{j,m} \tilde{P}_m(t)$$

for some constants $\beta_{j,1}, \ldots, \beta_{j,j}$. Therefore,

$$\int_0^1 q(s) s^j \mathrm{d}s = C_r \sum_{m=0}^{j} \beta_{j,m} \int_0^1 \tilde{P}_r(s) \tilde{P}_m(s) \mathrm{d}s = 0,$$

provided that $j \leq r - 1$. By Theorem 19.14, the method is exactly of order $p = r + r = 2r$. □

Definition 19.17 (Gauss–Legendre–RK method)**.** The implicit r-stage RK methods constructed as collocation methods whose collocation points are the zeros of the transformed Legendre polynomial $\tilde{P}_r$ are called **Gauss–Legendre–Runge–Kutta methods**.[6]

The following Gauss–Legendre–RK methods are collocation methods constructed using Corollary 19.16; see Table 19.1.

Example 19.6 *The midpoint rule*: Suppose that $r = 1$. The transformed Legendre polynomial of order one is

$$\tilde{P}_1(t) = 2t - 1 \implies c_1 = \frac{1}{2}.$$

The corresponding Gauss–Legendre IRK method is given by

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

and is of order $2r = 2$. For a scalar autonomous system, $u' = f(u)$, the method can be expressed as

$$w^{k+1} = w^k + \tau f\left(w^k + \frac{\tau}{2}\kappa_1\right), \quad \kappa_1 = f\left(w^k + \frac{\tau}{2}\kappa_1\right). \tag{19.11}$$

It is a simple exercise to show that this is equivalent to the midpoint rule,

$$w^{k+1} = w^k + \tau f\left(\frac{w^{k+1} + w^k}{2}\right). \tag{19.12}$$

[6] Named in honor of the German mathematician and physicist Johann Carl Friedrich Gauss (1777–1855) and the French mathematician Adrien-Marie Legendre (1752–1833).

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

But let us write this another way. Define

$$\tilde{w}^{k+\frac{1}{2}} = \frac{w^{k+1} + w^k}{2}.$$

Then we can express the midpoint rule as

$$\tilde{w}^{k+\frac{1}{2}} = w^k + \frac{\tau}{2} f\left(\tilde{w}^{k+\frac{1}{2}}\right), \quad w^{k+1} = 2\tilde{w}^{k+\frac{1}{2}} - w^k. \tag{19.13}$$

Still another, equivalent, way of writing this method is as follows:

$$\tilde{w}^{k+\frac{1}{2}} = w^k + \frac{\tau}{2} f\left(\tilde{w}^{k+\frac{1}{2}}\right), \quad w^{k+1} = \tilde{w}^{k+\frac{1}{2}} + \frac{\tau}{2} f\left(\tilde{w}^{k+\frac{1}{2}}\right). \tag{19.14}$$

Observe that method (19.13) shows that the midpoint rule is essentially a backward (implicit) Euler method with half the time step size followed by an extrapolation. Method (19.14) expresses the midpoint rule as a half-step-size backward Euler method followed by a half-step-size forward (explicit) Euler method.

In either case, the simple modification of a backward Euler method, which is only first-order accurate, leads to a second-order accurate method; for more insight on this, see [11].

Example 19.7 Suppose that $r = 2$. The transformed Legendre polynomial of order two is

$$\tilde{P}_2(t) = 6t^2 - 6t + 1 \implies c_1 = \frac{1}{2} - \frac{\sqrt{3}}{6},\ c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}.$$

The Gauss–Legendre IRK method is given by

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

and is of order $2r = 4$.

Example 19.8 Suppose that $r = 3$. The transformed Legendre polynomial of order three is

$$\tilde{P}_3(t) = 20t^3 - 30t^2 + 12t - 1 \implies c_1 = \frac{1}{2} - \frac{\sqrt{15}}{10},\ c_2 = \frac{1}{2},\ c_3 = \frac{1}{2} + \frac{\sqrt{15}}{10}.$$

The Gauss–Legendre IRK method is given by

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

and is of order $2r = 6$.

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

Example 19.9 Not all IRK methods are of collocation type. Consider, for example, the methods given by the tables

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}.$$

For both methods, the necessary conditions of Theorem 19.4 are satisfied. The method on the left, which is consistent to exactly order $p = 3$, is not of collocation type. (How do we know this?) The method on the right is of collocation type. One can check that the collocation points $c_1 = \frac{1}{3}$ and $c_2 = 1$ completely determine the other weights. The method on the right is consistent to exactly order $p = 3$. To see this, observe that

$$\int_0^1 \left(s - \frac{1}{3}\right)(s-1)s^j \mathrm{d}s = 0$$

only for $j = 0$. Thus, invoking Theorem 19.14, we have $m = 1$ and $p = r + m = 3$.

19.4 Dissipative Methods

In this section, we demonstrate how some IRK methods are particularly well suited to approximate dissipative equations, as defined in Section 17.4. We follow the notation introduced there and consider IVPs posed on $\mathbb{C}^d$, with $d \in \mathbb{N}$. The reader will easily verify that all the numerical methods, and theory for them, we have constructed so far extend to this case without difficulty.

One may wish to have a numerical method that preserves the dissipation property presented in Theorem 17.18. To achieve this, we begin with a definition.

Definition 19.18 (algebraic stability)**.** Assume that $\boldsymbol{f}\colon [0,T] \times \mathbb{C}^d \to \mathbb{C}^d$ is monotone with respect to $(\cdot,\cdot)$. Let $\boldsymbol{u}_0, \boldsymbol{v}_0 \in \mathbb{C}^d$. Assume that $\boldsymbol{f}$ is such that there are unique classical solutions on $[0,T]$ to the problems

$$\boldsymbol{u}'(t) = \boldsymbol{f}(t,\boldsymbol{u}(t)), \quad \boldsymbol{u}(0) = \boldsymbol{u}_0, \qquad \boldsymbol{v}'(t) = \boldsymbol{f}(t,\boldsymbol{v}(t)), \quad \boldsymbol{v}(0) = \boldsymbol{v}_0.$$

Let $\{\boldsymbol{w}^k\}_{k=0}^K$ and $\{\boldsymbol{z}^k\}_{k=0}^K$ be approximations to $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, obtained by the same numerical method; for example, some RK method. Notice that we must have $\boldsymbol{w}^0 = \boldsymbol{u}_0$ and $\boldsymbol{z}^0 = \boldsymbol{v}_0$. We say that the method is **dissipative** or **algebraically stable** if and only if, for any K and for all starting values $\boldsymbol{u}_0, \boldsymbol{v}_0 \in \mathbb{C}^d$,

$$\left\|\boldsymbol{w}^k - \boldsymbol{z}^k\right\| \leq \left\|\boldsymbol{w}^{k-1} - \boldsymbol{z}^{k-1}\right\| \leq \|\boldsymbol{u}_0 - \boldsymbol{v}_0\|$$

for all $k = 1, \ldots, K$.

In the same way that an RK method can be encoded in a Butcher tableau, its properties can be encoded in a particular matrix.

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

Definition 19.19 (M-matrix)**.** Suppose that the r-stage RK method is defined by the weights $\mathsf{A} = [a_{i,j}]_{i,j=1}^{r}$, $\boldsymbol{b} = [b_i]_{i=1}^{r}$, and $\boldsymbol{c} = [c_i]_{i=1}^{r}$. The **M-matrix** of the method is the matrix $\mathsf{M} = [m_{i,j}]_{i,j=1}^{r} \in \mathbb{R}^{r\times r}$ defined via

$$m_{i,j} = b_i a_{i,j} + b_j a_{j,i} - b_i b_j, \quad i,j = 1,\ldots,r.$$

The importance of the M-matrix of an RK method is in the dissipativity condition given in the following result.

Theorem 19.20 (dissipativity condition)**.** *Suppose that $\boldsymbol{f}\colon [0,T]\times\mathbb{C}^d \to \mathbb{C}^d$ is monotone with respect to $(\cdot,\cdot)$. Let $\boldsymbol{u}_0, \boldsymbol{v}_0 \in \mathbb{C}^d$. Assume that $\boldsymbol{f}$ is such that there are unique classical solutions on $[0,T]$ to the problems*

$$\boldsymbol{u}'(t) = \boldsymbol{f}(t,\boldsymbol{u}(t)), \quad \boldsymbol{u}(0) = \boldsymbol{u}_0, \qquad\qquad \boldsymbol{v}'(t) = \boldsymbol{f}(t,\boldsymbol{v}(t)), \quad \boldsymbol{v}(0) = \boldsymbol{v}_0.$$

Let the r-stage RK method be defined by the weights $\mathsf{A} = [a_{i,j}]_{i,j=1}^{r}$, $\boldsymbol{b} = [b_i]_{i=1}^{r}$, and $\boldsymbol{c} = [c_i]_{i=1}^{r}$. If its M-matrix is positive semi-definite and $b_j \geq 0$, $j = 1,\ldots,r$, then the RK method is dissipative.

Proof. Define, for $i,j = 1,\ldots,r$,

$$\boldsymbol{\rho}_j = \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\xi}_j), \quad \boldsymbol{\xi}_i = \boldsymbol{w}^k + \tau\sum_{j=1}^{r} a_{i,j}\boldsymbol{\rho}_j, \quad \boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \tau\sum_{j=1}^{r} b_j\boldsymbol{\rho}_j,$$

and

$$\boldsymbol{\sigma}_j = \boldsymbol{f}(t_k + c_j\tau, \boldsymbol{\zeta}_j), \quad \boldsymbol{\zeta}_i = \boldsymbol{z}^k + \tau\sum_{j=1}^{r} a_{i,j}\boldsymbol{\sigma}_j, \quad \boldsymbol{z}^{k+1} = \boldsymbol{z}^k + \tau\sum_{j=1}^{r} b_j\boldsymbol{\sigma}_j.$$

Then observe that

$$\left\|\boldsymbol{w}^{k+1} - \boldsymbol{z}^{k+1}\right\|^2 = \left\|\boldsymbol{w}^k - \boldsymbol{z}^k\right\|^2 + 2\tau\Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^{r} b_j\boldsymbol{d}_j\right)\right] + \tau^2\left\|\sum_{j=1}^{r} b_j\boldsymbol{d}_j\right\|^2,$$

where, for $j = 1,\ldots,r$,

$$\boldsymbol{d}_j = \boldsymbol{\rho}_j - \boldsymbol{\sigma}_j.$$

Thus, we will have proven the result if we can show that

$$2\tau\Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^{r} b_j\boldsymbol{d}_j\right)\right] + \tau^2\left\|\sum_{j=1}^{r} b_j\boldsymbol{d}_j\right\|^2 \leq 0.$$

Since, for $j = 1,\ldots,r$,

$$\boldsymbol{w}^k = \boldsymbol{\xi}_j - \tau\sum_{i=1}^{r} a_{j,i}\boldsymbol{\rho}_i, \quad \boldsymbol{z}^k = \boldsymbol{\zeta}_j - \tau\sum_{i=1}^{r} a_{j,i}\boldsymbol{\sigma}_i,$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

we have

$$\Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^{r} b_j \boldsymbol{d}_j\right)\right] = \sum_{j=1}^{r} b_j \Re\left[\left(\boldsymbol{\xi}_j - \boldsymbol{\zeta}_j - \tau \sum_{i=1}^{r} a_{j,i} \boldsymbol{d}_i, \boldsymbol{d}_j\right)\right]$$
$$= \sum_{j=1}^{r} b_j \Re\left[\left(\boldsymbol{\xi}_j - \boldsymbol{\zeta}_j, \boldsymbol{d}_j\right)\right] - \tau \sum_{j=1}^{r} \sum_{i=1}^{r} b_j a_{j,i} \Re\left[\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right)\right].$$

Since $\boldsymbol{f}$ is monotone, for each $j = 1, \ldots, r$,

$$\Re\left[\left(\boldsymbol{\xi}_j - \boldsymbol{\zeta}_j, \boldsymbol{d}_j\right)\right] \leq 0.$$

Since the weights b_j are all nonnegative,

$$\sum_{j=1}^{r} b_j \Re\left[\left(\boldsymbol{\xi}_j - \boldsymbol{\zeta}_j, \boldsymbol{d}_j\right)\right] \leq 0.$$

It then follows that

$$\Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^{r} b_j \boldsymbol{d}_j\right)\right] \leq -\tau \sum_{j=1}^{r} \sum_{i=1}^{r} b_j a_{j,i} \Re\left[\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right)\right],$$

or, equivalently, after swapping summation indices,

$$\Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^{r} b_j \boldsymbol{d}_j\right)\right] \leq -\tau \sum_{i=1}^{r} \sum_{j=1}^{r} b_i a_{i,j} \Re\left[\left(\boldsymbol{d}_j, \boldsymbol{d}_i\right)\right].$$

Thus,

$$2\tau \Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^{r} b_j \boldsymbol{d}_j\right)\right] + \tau^2 \left\|\sum_{j=1}^{r} b_j \boldsymbol{d}_j\right\|^2$$
$$\leq -\tau^2 \sum_{j=1}^{r} \sum_{i=1}^{r} b_j a_{j,i} \Re\left[\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right)\right] - \tau^2 \sum_{i=1}^{r} \sum_{j=1}^{r} b_i a_{i,j} \Re\left[\left(\boldsymbol{d}_j, \boldsymbol{d}_i\right)\right] + \tau^2 \left\|\sum_{j=1}^{r} b_j \boldsymbol{d}_j\right\|^2$$
$$= -\tau^2 \sum_{i=1}^{r} \sum_{j=1}^{r} m_{i,j} \Re\left[\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right)\right].$$

Since M is symmetric positive semi-definite, there is an orthogonal matrix Q and a diagonal matrix $\mathsf{D} = \mathrm{diag}[\lambda_1, \ldots, \lambda_r]$, with nonnegative diagonal entries, $\lambda_j \geq 0$, such that

$$\mathsf{M} = \mathsf{Q}\mathsf{D}\mathsf{Q}^{\mathsf{T}}.$$

In terms of coordinates,

$$m_{i,j} = \sum_{k=1}^{r} \lambda_k q_{i,k} q_{j,k}.$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

Thus,

$$\begin{aligned}
2\tau\Re\left[\left(\boldsymbol{w}^k - \boldsymbol{z}^k, \sum_{j=1}^r b_j \boldsymbol{d}_j\right)\right] + \tau^2 \left\|\sum_{j=1}^r b_j \boldsymbol{d}_j\right\|^2 \\
&\leq -\tau^2 \sum_{i=1}^r \sum_{j=1}^r \sum_{k=1}^r \lambda_k q_{i,k} q_{j,k} \Re\left[(\boldsymbol{d}_i, \boldsymbol{d}_j)\right] \\
&= -\tau^2 \sum_{k=1}^r \lambda_k \Re\left[\left(\sum_{i=1}^r q_{i,k} \boldsymbol{d}_i, \sum_{j=1}^r q_{j,k} \boldsymbol{d}_j\right)\right] \\
&= -\tau^2 \sum_{k=1}^r \lambda_k \left\|\sum_{j=1}^r w_{j,k} \boldsymbol{d}_j\right\|^2 \\
&\leq 0.
\end{aligned}$$

The result is proved:

$$\left\|\boldsymbol{w}^{k+1} - \boldsymbol{z}^{k+1}\right\|^2 \leq \left\|\boldsymbol{w}^k - \boldsymbol{z}^k\right\|^2. \qquad \square$$

We will conclude this section by proving that the Gauss–Legendre IRK methods are dissipative. To do this, we need some definitions.

Definition 19.21 (type)**.** Let $q \in \mathbb{N}$. An r-stage RK method defined by the weights $\mathsf{A} = [a_{i,j}]_{i,j=1}^r$, $\boldsymbol{b} = [b_i]_{i=1}^r$, and $\boldsymbol{c} = [c_i]_{i=1}^r$ is said to be of **type** $B(q)$ if and only if

$$\sum_{i=1}^r b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, q.$$

The method is said to be of **type** $C(q)$ if and only if

$$\sum_{j=1}^r a_{i,j} c_j^{k-1} = \frac{c_i^k}{k}, \quad i = 1, \ldots, r, \quad k = 1, \ldots, q.$$

Theorem 19.22 (dissipativity criterion)**.** *Consider an r-stage RK method defined by the weights $\mathsf{A} = [a_{i,j}]_{i,j=1}^r$, $\boldsymbol{b} = [b_i]_{i=1}^r$, and $\boldsymbol{c} = [c_i]_{i=1}^r$. Assume that the entries of $\boldsymbol{c} = [c_i]_{i=1}^r$ are distinct. If the method is of type $B(2r)$ and of type $C(r)$, then the M-matrix for this method is the zero matrix.*

Proof. Let $\mathsf{M} = [m_{i,j}] \in \mathbb{R}^{r\times r}$ denote the M-matrix for the method. Define the matrix $\mathsf{N} = [n_{i,j}] \in \mathbb{R}^{r\times r}$ via

$$n_{k,m} = \sum_{i=1}^r \sum_{j=1}^r c_i^{k-1} m_{i,j} c_j^{m-1} = \sum_{i=1}^r \sum_{j=1}^r c_i^{k-1} \left(b_i a_{i,j} + b_j a_{j,i} - b_i b_j\right) c_j^{m-1} = 0,$$

where $k, m = 1, \ldots, r$ and we used the fact that the method is of type $B(2r)$ and type $C(r)$. The details are left to the reader as an exercise; see Problem 19.9. In conclusion, $\mathsf{N} = \mathsf{O}$, the zero matrix. But observe that

$$\mathsf{N} = \mathsf{V}^\intercal \mathsf{M} \mathsf{V},$$

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press

where $\mathsf{V} = [c_i^{j-1}]$ is a variant of the Vandermonde matrix, which is nonsingular provided that the entries of $\boldsymbol{c} = [c_i]_{i=1}^r$ are distinct (Theorem 9.4). It follows that $\mathsf{M} = \mathsf{O}$. □

Theorem 19.23 (dissipativity)**.** *All Gauss–Legendre RK methods are dissipative.*

Proof. The idea is to show that all Gauss–Legendre RK methods are of type $B(2r)$ and type $C(r)$, thus implying that their M-matrices are $r \times r$ zero matrices. Finally, one needs to show that the weights b_j are all nonnegative. The result then follows from Theorem 19.20. The details are left to the reader as an exercise; see Problem 19.10. □

Problems

19.1 Let $T > 0$ be given. Consider the general two-stage explicit RK method, defined by

$$\xi^k = w^k + a\tau f(w^k), \quad w^{k+1} = w^k + \tau(b_1 f(w^k) + b_2 f(\xi^k)),$$

for approximating the solutions to the autonomous IVP

$$u'(t) = f(u(t)), \quad t \in [0, T], \quad u(0) = u_0.$$

Assume that $f \in \mathcal{F}^2(S)$ and the coefficients satisfy $b_1 + b_2 = 1$ and $ab_2 = \frac{1}{2}$. Prove that the method is convergent to second order.

19.2 Provide all the details for the proof of Theorem 19.4.

19.3 Prove Theorem 19.5.

19.4 Prove Theorem 19.6.

19.5 Consider the implicit RK methods given by the tableaux

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}.$$

a) Show that the necessary conditions of Theorem 19.4 are all satisfied.

b) Show that one of the methods is a collocation method and that the other is not. For the one that is a collocation method, find its order of consistency.

19.6 Complete the proof of Theorem 19.14.

19.7 Complete the proof of Theorem 17.18.

19.8 Show that no explicit RK method can be dissipative.

19.9 Complete the proof of Theorem 19.22.

19.10 Complete the proof of Theorem 19.23.

https://doi.org/10.1017/9781108942607.025 Published online by Cambridge University Press