6 Stiff ODE Initial Value Problems

This chapter deals with *stiff* initial value problems for ODEs

$$\dot{x} = F(x),\ x(0) = x_0\,.$$

The discretization of such problems is known to involve the solution of nonlinear systems per each discretization step—in one way or the other.

In Section 6.1, the contractivity theory for linear ODEs is revisited in terms of *affine similarity*. Based on an affine similar convergence analysis for a simplified Newton method in *function space*, a *nonlinear contractivity* theory for stiff ODE problems is derived in Section 6.2, which is quite different from the theory given in usual textbooks on the topic. The key idea is to replace the Picard iteration in function space, known as a tool to show uniqueness in nonstiff initial value problems, by a simplified Newton iteration in function space to characterize stiff initial value problems. From this point of view, *linearly implicit* one-step methods appear as direct realizations of the simplified Newton iteration in function space. In Section 6.3, exactly the same theoretical characterization is shown to apply also to *implicit* one-step methods, which require the solution of a nonlinear system by some finite dimensional Newton-type method at each discretization step.

Finally, in a deliberately longer Section 6.4, we discuss a class of algorithms called *pseudo-transient continuation* algorithms, whereby steady state problems are solved via stiff integration. The latter type of algorithm is particularly useful, when the Jacobian matrix is singular due to hidden dynamical invariants (such as mass conservation). The affine similar theoretical characterization permits the derivation of an *adaptive (pseudo-)time step strategy* and an accuracy matching strategy for a residual based inexact Newton algorithm.

6.1 Affine Similar Linear Contractivity

For the time being, consider a *linear* ODE system of the kind

$$\dot{x} = Ax,\ x(0) = x_0\,. \tag{6.1}$$

P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer Series in Computational Mathematics 35,
DOI 10.1007/978-3-642-23899-4_6, © Springer-Verlag Berlin Heidelberg 2011

Formally, system (6.1) can be solved in terms of the matrix exponential

$$x(t) = \exp(At)x_0 \,.$$

In view of *affine similarity* as discussed in Section 1.2.2, we start from the (possibly complex) Jordan decomposition

$$A = TJT^{-1} \,,$$

wherein J is the *Jordan canonical form* consisting of elementary Jordan blocks for each separate eigenvalue $\lambda(A)$. Then the (possibly complex) transformation

$$z := T^{-1}(x - \hat{x})$$

has been shown in Section 1.2.2 to generate an affine similar coordinate frame.

In what follows we will have to work with norms $\|\cdot\|$ induced by certain inner products $(\cdot,\cdot)$. For simplicity, we may think of the Euclidean norm $\|\cdot\|$ induced by the (possibly complex) Euclidean inner product $(u,v) = u^*v$ with u^* the adjoint. If we phrase our subsequent theoretical statements in terms of the *canonical norm*

$$|u| := \|T^{-1}u\| \,, \tag{6.2}$$

induced by the *canonical inner product*

$$\langle u, v\rangle = (T^{-1}u, T^{-1}v) \,,$$

then such statements will automatically meet the requirement of affine similarity. In this setting, we may define some constant $\mu = \mu(A)$, allowed to be *positive, zero,* or *negative*, such that

$$\langle u, Au\rangle \leq \mu(A)|u|^2 \,. \tag{6.3}$$

This definition is obviously equivalent to

$$(\bar{u}, J\bar{u}) \leq \mu(A)\|\bar{u}\|^2 \,, \tag{6.4}$$

wherein $\bar{u} = T^{-1}u$. Assuming that the quantity μ is chosen best possible, it can be shown to satisfy

$$\mu(A) = \max_{u\neq 0} \frac{\langle u, Au\rangle}{|u|^2} \geq \max_i \Re\lambda_i(A) + \epsilon \;, \quad \epsilon \geq 0 \,. \tag{6.5}$$

Herein $\epsilon = 0$ and equality holds, if the eigenvalue defining $\mu(A)$ is simple. It is an easy task to show that

$$\mu(BAB^{-1}) = \mu(A) \tag{6.6}$$

for any nonsingular matrix B, which confirms that this quantity is indeed *affine similar*. In the canonical norm we may obtain the estimate

$$|x(t)| \leq \exp(\mu t)|x_0| .$$

Whenever

$$\mu \leq 0 \tag{6.7}$$

holds, then

$$|x(t)| \leq |x(0)| ,$$

which means that the linear dynamical system (6.1) is *contractive.*

For computational reasons, the Euclidean product (possibly in a scaled variant) is preferred to the canonical product. Suppose that we therefore replace the above definition (6.3) in terms of the canonical inner product by the analogous definition

$$\nu(A) = \max_{u \neq 0} \frac{(u, Au)}{\|u\|^2} \tag{6.8}$$

in terms of the Euclidean product. The thus defined quantity can be expressed as

$$\nu(A) = \lambda_{\max}\left(\tfrac{1}{2}(A + A^T)\right) ,$$

where $\lambda_{\max}$ is the maximum (real) eigenvalue of the symmetric part of the matrix A. Upon comparison with (6.4) we immediately observe that

$$\mu(A) = \nu(J) = \lambda_{\max}\left(\tfrac{1}{2}(J + J^T)\right) ,$$

which directly leads to the above result (6.5)—see, e.g., [71, Section 3.2]. From this we see that the quantities $\nu(A)$ and $\mu(A)$ may be rather different—in fact, unless A is symmetric, not even the signs may be the same. Moreover, in contrast to (6.6), we now have the undesirable property

$$\nu(BAB^{-1}) \neq \nu(A) ,$$

i.e., this quantity is *not affine similar.* Consequently, contractivity in the canonical norm $|\cdot|$ does *not* imply contractivity in the original norm $\|\cdot\|$. Whenever a relation of the kind

$$|u| \leq |v|$$

is transformed back to the original norm, we can only prove that

$$\|u\| \leq \operatorname{cond}(T)\|v\|$$

in terms of the condition number $\operatorname{cond}(T) = \|T^{-1}\| \cdot \|T\| \geq 1$, which here arises as an unavoidable *geometric distortion* factor. This distortion factor also indicates possible *ill-conditioning* of the Jordan decomposition as a whole—which may affect the theoretical presentation in terms of canonical inner products and norms. Nevertheless, we will stick to a formal notation in terms of the canonical norm $|\cdot|$ below to make the underlying structure transparent.

6.2 Nonstiff versus Stiff Initial Value Problems

Reliable numerical algorithms are, in one way or the other, appropriate implementations of uniqueness theorems of the underlying analytic problem. The most popular uniqueness theorem for ODEs is the well-known Picard-Lindelöf theorem: it is based on the *Picard fixed point iteration* in function space (Section 6.2.1) and characterizes the growth of the solution by means of the Lipschitz constant of the right-hand side—a characterization known to be appropriate for *nonstiff* ODEs, but inappropriate for *stiff* ODEs. As will be shown in this section, an associated uniqueness theorem for stiff ODEs can be derived on the basis of a *simplified Newton iteration* in function space, wherein the above Lipschitz constant is circumvented by virtue of a one-sided linear contractivity constant. As a natural spin-off, this theory produces some common *nonlinear contractivity* concept both for ODEs (Section 6.2.2) and for implicit one-step discretizations (see Section 6.3 below).

6.2.1 Picard iteration versus Newton iteration

Consider again the *nonlinear* initial value problem

$$\dot{x} = F(x), \ x(0) = x_0 \,.$$

For the subsequent presentation, its equivalent formulation as a Volterra operator equation (of the second kind) is preferable:

$$G(x, \tau) := x(\tau) - x_0 - \int\limits_{t=0}^{\tau} F(x(t))dt = 0 \,. \tag{6.9}$$

This equation defines a *homotopy* in terms of the interval length $\tau \geq 0$. Let Γ denote some neighborhood of the graph of a solution of (6.9). Then Peano's *existence* theorem requires that

$$L_0 := \sup_{\Gamma} \|F(x)\| < \infty$$

in terms of some pointwise norm $\| \cdot \|$ in $\mathbb{R}^n$.

In order to prove *uniqueness*, the standard approach is to construct the so-called Picard iteration

$$x^{i+1}(\tau) = x_0 + \int\limits_{t=0}^{\tau} F(x^i(t))dt \tag{6.10}$$

to be started with $x^0(t) \equiv x_0$. From this fixed point iteration, one immediately derives

$$\|x^{i+1}(\tau) - x^i(\tau)\| \leq \int\limits_{t=0}^{\tau} \|F(x^i(t)) - F(x^{i-1}(t))\| dt\,.$$

Hence, in order to study contraction, the most natural theoretical characterization is in terms of the Lipschitz constant L_1 defined by

$$\|F(u) - F(v)\| \leq L_1 \|u - v\|\,.$$

With this definition, the sequence $\{x^i\}$ can be shown to converge to some solution x^* such that

$$\|x^*(\tau) - x_0\| \leq L_0 \tau \varphi(L_1 \tau)$$

with

$$\varphi(s) := \begin{cases} (\exp(s) - 1)/s & s \neq 0 \\ 1 & s = 0\,. \end{cases} \tag{6.11}$$

Moreover, x^* is unique in Γ. This is the main result of the well-known Picard-Lindelöf theorem—ignoring for simplicity any distinction between local and global Lipschitz constants.

A similar term arises in the analysis of *one-step discretization* methods for ODE initial value problems. Let $p \geq 1$ denote the consistency order of such a method and τ a selected uniform stepsize assumed to be sufficiently small. Then the discretization error between the continuous solution x and the discrete solution x_τ at some final point $T = n\tau$ can be represented in the form (see, e.g., the ODE textbook by E. Hairer and G. Wanner [114]):

$$\|x_\tau(T) - x(T)\| \leq C_p \cdot \tau^p \cdot T \cdot \varphi(\bar{L}_1 T)\,.$$

For *explicit* one-step methods, the coefficient C_p just depends on some bound in terms of higher derivatives of F. The above discrete Lipschitz constant $\bar{L}_1 \geq L_1$ is an analog of the continuous Lipschitz constant L_1, this time for the increment function of the one-step method. In order to assure that the notion of a consistency order p is meaningful at all, a restriction of the kind

$$L_1 \tau \leq C, \quad C = \mathcal{O}(1) \tag{6.12}$$

will be needed. Consequently, this characterization is appropriate only for *nonstiff* discretization methods.

Historical Note. Originally, it had first been thought that the use of *implicit* discretization methods would be the essential item to overcome the observed difficulties in the numerical integration of what have been called stiff ODEs—see, for instance, the early fundamental paper by G. Dahlquist [47]. For implicit one-step methods, the above coefficient C_p is bounded only, if the discrete solution can be locally continued over each discretization step of length τ. This aspect will be studied in detail in the subsequent Section 6.3. In the next stage of the development of stiff integration, however, it was recognized that the solution method for the thus arising algebraic equations is equally important: the early paper of W. Liniger and R.A. Willoughby [145] pointed out that any fixed point iteration based only on F-evaluations for the algebraic equations would again bring in restriction (6.12), whereas a *Newton-like iteration* could, in principle, avoid the restriction. Much later, so-called semi-implicit or linearly-implicit discretization methods (such as Rosenbrock methods, W-methods, or extrapolation methods) were constructed that only apply one single Newton-like iteration per discretization step. Therefore the present essence of insight seems to be that nonstiff integration is characterized by sampling of F only, whereas stiff integration requires the additional sampling of $F'(x)$ or an approximation.

With these preparations, a natural approach towards a uniqueness theorem covering stiff ODEs as well will be to replace the Picard iteration (6.10) by a Newton iteration. For the *ordinary* Newton method we would obtain

$$G'(x^i)\Delta x^i = -G(x^i), \quad x^{i+1} = x^i + \Delta x^i$$

or, in more explicit notation

$$\Delta x^i(\tau) - \int\limits_{t=0}^{\tau} F'(x^i(t))\Delta x^i(t)dt = -\left[x^i(\tau) - x_0 - \int\limits_{t=0}^{\tau} F(x^i(t))dt\right] . \tag{6.13}$$

However, the above iteration requires global sampling of the Jacobian $F'(x)$ rather than just pointwise sampling as in numerical stiff integration. Therefore, the *simplified* Newton method will be chosen instead: we just have to replace

$$G'(x^i) \to G'(x^0), \;\; x^0(t) \equiv x_0$$

or, equivalently,

$$F'(x^i(t)) \to F'(x_0) =: A\,.$$

The corresponding replacement in (6.13) then leads to

$$x^{i+1}(\tau) - A\int_0^\tau x^{i+1}(t)dt = x_0 + \int_0^\tau [F(x^i(t)) - Ax^i(t)]dt\,. \tag{6.14}$$

Note that this may be interpreted as a Picard iteration associated with the formally modified ODE

$$\dot{x} - Ax = F(x) - Ax, \; x(0) = x_0\,,$$

which is the basis for linearly-implicit one-step methods.

6.2.2 Newton-type uniqueness theorems

The above simplified Newton-iteration (6.14) is now exploited with the aim of proving uniqueness theorems for ODE IVP's that cover stiff ODEs as well. In order to guarantee *affine similarity*, we will define coordinates $x \in \mathbb{R}^n$ in such a way that any required (possibly approximate) Jacobian A is already in its Jordan canonical form J. Consequently, the selected vector norm $\|\cdot\|$ is identical to canonical norm as defined in (6.2)—see also the discussion in Section 6.1. For the time being, we assume that we have an exact initial Jacobian

$$F'(x_0) = A = J\,.$$

A discussion of the case of an approximate Jacobian will follow subsequently.

Theorem 6.1 *In the above notation let $F \in C^1(D)$, $D \subseteq \mathbb{R}^n$. For the Jacobian $A := F'(x_0)$ assume a one-sided Lipschitz condition of the form*

$$(u, Ju) \le \mu \|u\|^2\,,$$

where $(\cdot,\cdot)$ denotes the inner product that induces the norm $\|\cdot\|$. In this norm, assume that

$$\|F(x_0)\| \le L_0 \qquad for \quad x_0 \in D$$

$$\|(F'(x) - F'(x_0))u\| \le L_2 \|x - x_0\| \|u\| \qquad for \quad x, x_0, u \in D\,. \tag{6.15}$$

Then, for D sufficiently large, existence and uniqueness of the solution of the ODE IVP is guaranteed in $[0, \tau]$ such that

$$\tau \ \textit{unbounded}\ ,\ \textit{if}\ \mu\bar{\tau} \le -1\,,$$

$$\tau \le \bar{\tau}\Psi(\mu\bar{\tau})\ \ ,\ \textit{if}\ \mu\bar{\tau} > -1$$

with $\bar{\tau} := (2L_0L_2)^{-1/2}$ and

$$\Psi(s) := \begin{cases} \ln(1+s)/s & s \neq 0 \\ \quad 1 & s = 0\,. \end{cases}$$

Proof. Upon performing the variation of constants, we rewrite (6.14) as

$$\Delta x^i(\tau) = \int\limits_{t=0}^{\tau} \exp(A(\tau - t))\big(F(x^i(t)) - \frac{d}{dt}x^i(t)\big)dt\,, \tag{6.16}$$

where $\exp(At)$ denotes the matrix exponential. Within this proof let $|\cdot|$ denote the standard C^0-norm:

$$|u| := \max_{t \in [0,\tau]} \|u(t)\|\,.$$

In order to study convergence, we set the initial guess $x^0(t) \equiv x_0$ and apply Theorem 2.5 from Section 2.1.2, which essentially requires that

$$|\Delta x^0| \leq \alpha\,, \tag{6.17}$$

$$|G'(x^0)^{-1}(G'(x) - G'(x^0))| \leq \omega |x - x^0|\,, \tag{6.18}$$

$$\alpha\omega \leq \tfrac{1}{2}\,. \tag{6.19}$$

The rest of the assumptions holds for D sufficiently large. The task is now to derive upper bounds α, ω and to assure (6.19). With x^0 as set above, the first Newton correction satisfies—compare (6.16)

$$\Delta x^0(\tau) = \int\limits_{t=0}^{\tau} \exp(A(\tau - t))F(x_0)dt\,.$$

Hence

$$\begin{aligned}\|\Delta x^0(\tau)\| &\leq \int\limits_{s=0}^{\tau} \|\exp(As)F(x_0)\|ds \leq \\ &\leq L_0 \int\limits_{s=0}^{\tau} \exp(\mu s)ds = L_0\tau\varphi(\mu\tau) =: \alpha(\tau)\end{aligned}$$

with φ as introduced in (6.11). In order to derive $\omega(\tau)$, we introduce the operator norm in (6.18) by

$$\begin{aligned}&z := G'(y^0)^{-1}(G'(x^0 + w) - G'(x^0))u\,, \\ &|z| \leq \omega \cdot |u| \cdot |w|\,.\end{aligned}$$

Once more by variation of constants, we obtain

$$\|z(\tau)\| \leq \int\limits_{t=0}^{\tau} \|\exp(A(\tau - t))[F'(x_0 + w) - F'(x_0)]u\|dt\,,$$

which, similar as above, yields

$$|z| \leq L_2 \cdot \tau \cdot \varphi(\mu\tau) \cdot |u| \cdot |w|\,.$$

Hence, a natural definition is

$$\omega(\tau) := L_2\tau\varphi(\mu\tau)\,.$$

Insertion into the Kantorovich condition (6.19) produces

$$(\tau\varphi(\mu\tau))^2 \leq (2L_0L_2)^{-1} =: \bar{\tau}^2$$

or, equivalently,

$$\tau\varphi(\mu\tau) \leq \bar{\tau}\,.$$

Since μ may have either sign, the main statements of the theorem are an immediate consequence. □

A graphical representation of the above monotone decreasing function Ψ is given below in Figure 6.3.

Remark 6.1 Upon using the same characterizing quantities $\mu, \bar{\tau}$ as above, an improved theorem has been shown by W. Walter [195] using differential inequalities—compare his book [194]. Since this theorem is nonconstructive and does not make a difference for the discretizations to be treated in Section 6.3, it is omitted here (details can be found in the paper [66]).

Nonlinear contractivity. As can be seen in Fig. 6.1 below, the above theorem (as well as the one in [66]) comes up with a pole at $s = -1$, which reflects the condition

$$\mu\bar{\tau} \leq -1$$

for *global boundedness* of the solution. This condition may be rewritten as

$$\mu + \sqrt{2L_0L_2} \leq 0\,. \tag{6.20}$$

As $L_2 = 0$ in the linear case, the above condition is immediately recognized as a direct generalization of the linear contractivity condition (6.7). In other words: the above pole represents *global nonlinear contractivity*, involving local contractivity via μ and the part from the nonlinearity in well-separated form.

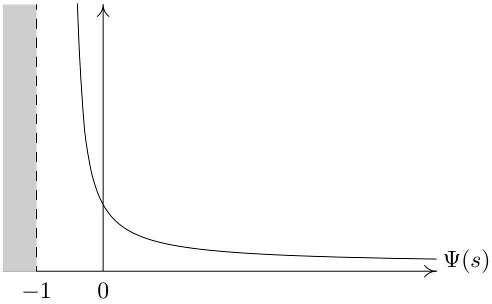

Fig. 6.1. Nonlinear contractivity: function Ψ as defined in Theorem 6.1 .

If, instead of the *exact* Jacobian $A = F'(x_0)$, an approximation error

$$\delta A = A - F'(x_0)$$

must be taken into account, then a modification of the above theorem will be necessary. In view of an affine similar presentation we again assume that

$$A = J\,,$$

which means that the approximate Jacobian is already in Jordan canonical form and the norm $\|\cdot\|$ is identical to the canonical norm.

Theorem 6.2 *Notation and assumptions as in Theorem* 6.1. *In addition, let the Jacobian approximation error be bounded as*

$$\|\delta A\| \leq \delta_0 \ , \ \ \delta_0 \geq 0 \, .$$

Then the results of Theorem 6.1 *hold with* $\bar{\tau}$ *replaced by*

$$\hat{\tau} := \bar{\tau}/(1 + \delta_o \bar{\tau}) \, .$$

Proof. For the proof we apply Theorem 2.6, the convergence theorem for Newton-like iterations, with $G'(x^0)$ now replaced by $M(x^0)$, which means replacing $F'(x_0)$ by $A \neq F'(x_0)$. With μ now associated with the Jacobian *approximation* A , the estimates $\alpha(\tau), \omega(\tau)$ carry over. In addition, the assumptions (6.17) up to (6.19) must be extended by

$$|M(x^0)^{-1}(G'(x^0) - M(x^0))| \leq \bar{\delta}_0 < 1 \, . \tag{6.21}$$

Upon defining

$$z := M(x^0)^{-1}(G'(x^0) - M(x^0))u$$

a similar estimate as in the proof of Theorem 6.1 leads to

$$\|z(\tau)\| \leq \int\limits_{t=0}^{\tau} \| \exp(A(\tau - t)) \cdot \delta A \cdot u\| dt \leq \delta_0 \tau \varphi(\mu\tau)|u| \, .$$

Hence, the above condition (6.21) shows up with the specification

$$\bar{\delta}_0 := \delta_0 \tau \varphi(\mu\tau) \, .$$

Insertion into the modified Kantorovich condition (2.23)

$$\frac{\alpha\omega}{(1 - \bar{\delta}_0)^2} \leq \tfrac{1}{2}$$

then yields

$$\tau\varphi(\mu\tau) \leq \bar{\tau}/(1 + \delta_0 \bar{\tau}) =: \hat{\tau} \, .$$

Note that condition (6.21) is automatically satisfied, since

$$\bar{\delta}_0 = \delta_0 \tau \varphi(\mu\tau) \leq \delta_0 \bar{\tau}/(1 + \delta_0 \bar{\tau}) < 1 \, ,$$

which completes the proof. □

Finally, we want to emphasize that all above results also hold, if the norm $\|\cdot\|$ is not identified with the canonical norm $|\cdot|$, but allowed to be a general vector norm. However, as already worked out at the end of Section 6.1, this would include a tacit deterioration of all results, since then the one-sided Lipschitz constant μ may be rather off scale, if not even nonnegative—compare again the definitions (6.3) and (6.8).

Remark 6.2 The experienced reader will be interested to know whether these results carry over to differential-algebraic equations (DAEs) as well. Unfortunately, this causes some difficulty, which can already be seen in the simple separable DAE of the form

$$y' = f(y,z)\,, \quad 0 = g(y,z)\,.$$

In this case, the differential part f and the variable y suggest affine similarity, whereas the equality constrained part g would require some affine covariance or contravariance. For this reason, a common affine invariant theoretical framework is hard to get, if at all possible. Up to now, more subtle estimation techniques use a characterization in terms of perturbation parameters ϵ, which by construction do not allow for any affine invariance concept.

6.3 Uniqueness Theorems for Implicit One-step Methods

A natural requirement for any discretization π of the above ODE initial value problem will be that it passes basic symmetries of the underlying continuous problem on to the generated discrete problem. In particular, we will require that the diagram in Figure 6.2 commutes, which ensures that $y_\pi = Bx_\pi$ holds whenever $y = Bx$. Among the discretization methods satisfying this requirement, we will restrict our attention to implicit and linearly implicit one-step methods, also called Runge-Kutta methods.

$$\begin{tikzcd}
\boxed{x} \arrow[r, shift left, "B"] \arrow[d, "\pi"] & \boxed{y} \arrow[l, shift left, "B^{-1}"] \arrow[d, "\pi"] \\
\boxed{x_\pi} \arrow[r, shift left, "B"] & \boxed{y_\pi} \arrow[l, shift left, "B^{-1}"]
\end{tikzcd}$$

Fig. 6.2. Affine invariance under discretization π

As an extension of Section 6.1, any affine similar discretization of linear ODEs can also be treated in affine similar terms. This idea directly leads to G. Dahlquist's linear scalar model equation [47]

$$\dot{x} = \lambda x, \quad x(0) = 1\,.$$

Therefore, *linear contractivity* of implicit discretizations as well as of linearly implicit discretizations can be treated just as described in usual numerical ODE textbooks—see, e.g., [114, 115].

Things are different with respect to *nonlinear contractivity.* First, recall from Section 6.2.1 that the simplified Newton iteration (6.14) for the continuous ODE problem may also be regarded as a Picard iteration (6.10) for the ODE

$$\dot{x} - Ax = F(x) - Ax\,.$$

This ODE is the starting point for linearly implicit one-step discretizations (such as Rosenbrock methods, W-methods, or extrapolation methods), which just discretize the above left hand side implicitly and the above right hand side explicitly. Therefore, *linearly implicit discretizations* may be interpreted as *direct realizations of the simplified Newton iteration in function space.* Of course, they should also observe the local timestep restrictions as worked out for the continuous problem in Section 6.2.2. For the special case of the linearly implicit Euler discretization we refer to the residual analysis given in the subsequent Section 6.4.

Here we concentrate on *implicit* one-step discretizations. In such discretizations the discrete system comprises a nonlinear algebraic system, which again brings up the question of local continuation. We will be interested to see, in which way some kind of nonlinear contractivity is inherited from the continuous initial value problem to various implicit one-step methods. In order to permit a comparison with Section 6.2.2, we will again assume that the coordinates have been chosen such that the local Jacobian matrix $A \approx F'(x_0)$ is already in Jordan canonical form—which implies that the canonical product and norm are identical to the Euclidean product and norm. To start with, we exemplify the formalism at a few simple cases.

Implicit Euler discretization. In each step of this discretization, we must solve the n algebraic equations

$$G(x, \tau) := x - x_0 - \tau F(x) = 0\,, \tag{6.22}$$

which represent a *homotopy* in $\mathbb{R}^n$ with embedding in terms of the stepsize τ—say $\tau \geq 0$. The *Newton-like iteration* for solving this system is

$$(I - \tau A)\Delta x_i = -(x_i - x_0 - \tau F(x_i)), \quad x_{i+1} = x_i + \Delta x_i\,, \tag{6.23}$$

where $\delta A := A - F'(x_0) \neq 0$ will be assumed.

Theorem 6.3 *Assumptions and notation as in Theorems* 6.1 *and* 6.2 *above. Then the Newton-like iteration* (6.23) *for the implicit Euler discretization converges to a unique solution for all stepsizes*

$$\tau \text{ unbounded}\ , \text{ if } \mu\hat{\tau} \leq -1\,,$$
$$\tau \leq \hat{\tau}\Psi_D(\mu\hat{\tau})\ , \text{ if } \mu\hat{\tau} > -1\,,$$

where

$$\Psi_D(s) := (1+s)^{-1}\,.$$

Proof. Once more, Theorem 2.6 is applied, here to the finite-dimensional homotopy (6.22). The Jacobian approximation $A \approx F'(x_0)$ leads to the approximation

$$I - \tau A =: M(x_0) \approx F'(x_0)\,,$$

which is used in the definition of the affine covariant Lipschitz constant

$$\|M(x_0)^{-1}(F'(u) - F'(v))\| \leq \omega(\tau)\|u - v\|\,,$$

the first correction bound

$$\|\Delta x_0\| = \|M(x_0)^{-1}F(x_0)\| \leq \alpha(\tau)$$

and the approximation measure

$$\|M(x_0)^{-1}(M(x_0) - F'(x_0))\| \leq \bar{\delta}_0(\tau) < 1\,.$$

With these definitions, the modified Kantorovich condition here reads

$$\frac{\alpha\omega}{(1-\bar{\delta}_o)^2} \leq \tfrac{1}{2}\,. \tag{6.24}$$

Upon using similar techniques as in the proof of Theorem 6.2 above, we come up with the estimates:

$$\alpha(\tau) := \tau L_0/(1-\mu\tau), \quad \omega(\tau) := hL_2/(1-\mu\tau), \quad \bar{\delta}_0(\tau) := \tau\delta_0/(1-\mu\tau),$$

where the quantities L_0, L_2, δ_0 are the same as in Section 6.2.2. Insertion into condition (6.24) then yields, for $\mu\tau < 1$:

$$\frac{\tau}{1-\mu\tau} \leq \hat{\tau}$$

or, equivalently,

$$\tau \leq \hat{\tau}/(1+\mu\hat{\tau})\,.$$

This is the main statement of the theorem. Finally, note that for $\mu > 0$

$$\mu\tau \leq \mu\hat{\tau}/(1+\mu\hat{\tau}) < 1\,,$$

which assures the above requirement. The case $\mu \leq 0$ is trivial. □

The intriguing similarity of Theorems 6.2 and 6.3 is illustrated in Figure 6.3.

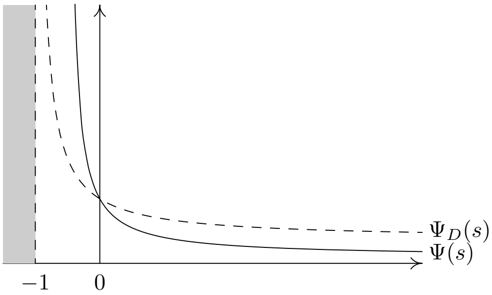

Fig. 6.3. Nonlinear contractivity inherited: function Ψ (continuous case) versus function Ψ_D (discrete case).

Implicit trapezoidal rule. This discretization requires the solution of the n in general nonlinear equations

$$G(x) := x - x_0 - \tfrac{1}{2}\tau\left(F(x) + F(x_0)\right) = 0\,. \tag{6.25}$$

Standard Newton-like iteration leads to the steplength restriction

$$\begin{aligned} &\text{a)}\quad \tau \text{ unbounded, if } \mu\bar{\tau} \le -\sqrt{2}\,,\\ &\text{b)}\quad \tau \le \bar{\tau}\sqrt{2}\Psi_D\left(\frac{\mu\bar{\tau}}{\sqrt{2}}\right)\,. \end{aligned} \tag{6.26}$$

Observe that the pole of Ψ at $s = -1$ is *not preserved* here, so that nonlinear contractivity is not correctly inherited from the continuous case.

Implicit midpoint rule. This discretization leads to the n equations

$$G(x) := x - x_0 - \tau F\left(\frac{x + x_0}{2}\right) = 0\,, \tag{6.27}$$

which, along similar lines of derivation, yields the stepsize bounds

$$\begin{aligned} &\text{a)}\quad \tau \text{ unbounded, if } \mu\bar{\tau} \le -1\,,\\ &\text{b)}\quad \tau \le 2\bar{\tau}\Psi_D(\mu\bar{\tau})\,. \end{aligned} \tag{6.28}$$

Here the pole is correctly preserved. Moreover, less restrictive bounds appear.

Summarizing, the implicit trapezoidal rule and the implicit midpoint rule have the same *linear* contractivity properties, but different *nonlinear* contractivity properties. From the nonlinear contractivity point of view, the implicit midpoint rule is clearly preferable. Both proofs are just along the lines of the proof for the implicit Euler method and therefore left as Exercise 6.1. Of course, one would really like to characterize the whole subclass of those one-step methods that preserve the pole exactly—a question left to future research.

6.4 Pseudo-transient Continuation for Steady State Problems

In this section we consider the case that the solution of a nonlinear system $F(x) = 0$ can be interpreted as *steady state* of a *dynamical system* of the kind

$$\dot{x} = F(x)\,. \tag{6.29}$$

Already from mere geometrical insight it is clear that such an approach will only work, if the fixed point of the dynamical system is *attractive* in a sufficiently large neighborhood. As an example, stiff integration towards a *hyperbolic* fixed point might come close to the fixed point for a while and run away afterwards. Exceptions will be possible only for a measure zero set of starting points x_0.

Dynamical invariants. This type of invariants occurs rather frequently in dynamical systems causing *singular* Jacobian matrices $F'(x)$ for all arguments x—which prohibits the application of standard Newton methods.

Example: mass conservation. Suppose the above ODE (6.29) describes some reaction kinetic model. Then mass conservation shows up as

$$e^T x(t) = e^T x_0\,,$$

where $e^T = (1, \ldots, 1)$. This implies

$$e^T \dot{x} = e^T F(x) \equiv 0, \quad x \in D \subset \mathbb{R}^n, \quad F(x) \neq 0\,.$$

By differentiation with respect to x we obtain

$$e^T F'(x) F(x) \equiv 0, \quad F(x) \neq 0$$

and hence every Jacobian has a zero eigenvalue with left eigenvector e. If we define the orthogonal projectors

$$P^\perp := \frac{1}{n} e e^T, \quad P = I - P^\perp\,,$$

then we can write equivalently

$$P^\perp F'(x) = 0\,. \tag{6.30}$$

Of course, naive application of any standard Newton method would fail in this situation. In this *special* case, a modification is possible that makes the Newton methods nevertheless work—see, e.g., Exercise 6.3.

In the *general* case, however, more than one dynamical invariant exists, most of them unspecified or even unknown, so that (6.30) holds again, now for an *unknown* projector P such that

$$P^\perp \dot{x} = P^\perp F(x) = 0 \quad \Longrightarrow \quad P^\perp F'(x) = 0\,. \tag{6.31}$$

Clearly, Newton methods cannot be modified to work without full knowledge about all dynamical invariants and are therefore bound to fail.

Fixed point iterations. In contrast to the other affine invariance classes of nonlinear problems, affine similarity also holds for *fixed point* iterations

$$\Delta x^k = x^{k+1} - x^k = \alpha F(x^k)$$

with a parameter α to be adapted. From equation (6.31) we see that such an iteration automatically realizes

$$P^{\perp} \Delta x = 0\,.$$

Pseudo-transient continuation. The same property can be shown to hold for any linear combination of Newton and fixed point iteration. A popular technique is the so-called pseudo-transient continuation method

$$(I - \tau A)\, \Delta x = F(x_0), \quad x(\tau) = x_0 + \tau \Delta x \tag{6.32}$$

with timestep τ to be adapted and $A = F'(x_0)$ or a Jacobian approximation. The above iteration is just a special stiff discretization of the ODE (6.29), known as the *linearly implicit Euler discretization.*

Of course, in order to obtain the solution, we may directly solve the time dependent system (6.29) by any numerical stiff integrator up to the steady state. In what follows, however, we want to restrict our attention to the simple case of the linearly implicit Euler discretization.

6.4.1 Exact pseudo-transient continuation

We now want to study an iterative method for the numerical solution of the nonlinear System $F(x) = 0$ based on the linearly implicit Euler discretization (6.32). Throughout this section we will assume that we can evaluate an *exact Jacobian* $A = F'(x)$ and solve the linear system (6.32) by *direct* elimination techniques.

As worked out in detail in Section 6.1, the problem itself is invariant under affine similarity transformation, which would suggest some theoretical treatment in terms of canonical norms and inner products. Usual stiff integration focuses on the *accuracy of the solution* which naturally belongs to an affine covariant setting. For reasons of numerical realization, however, we need to study the convergence of the iteration in terms of its *residual* behavior—which leads to an *affine contravariant* setting. For that reason, we will need to replace the canonical norm $|\cdot|$ (see Section 6.1) by some Euclidean norm $\|\cdot\|$, possibly *scaled.* Accordingly $(\cdot,\cdot)$ will denote the Euclidean inner product, also possibly scaled.

Let $x(\tau)$ denote the *homotopy path* defined by (6.32) and starting at the point $x(0) = x_0$. Before we actually study the residual norm $\|F(x(\tau))\|$, the following auxiliary result will be helpful.

Lemma 6.4 *Notation as just introduced with* $A \approx F'(x_0)$. *Then the residual along the homotopy path* $x(\tau)$ *starting at* x_0 *satisfies*

$$F(x(\tau)) = (I-\tau A)^{-1}F(x_0) + \int_{\sigma=0}^{\tau} \left(F'(x(\sigma)) - A\right)(I-\sigma A)^{-2}F(x_0)d\sigma\,. \quad (6.33)$$

Proof. Taylor's expansion of the residual yields

$$\begin{aligned} F(x(\tau)) &= F(x_0) + \int_{\sigma=0}^{\tau} F'(x(\sigma))\dot{x}(\sigma)d\sigma \\ &= F(x_0) + A(x(\tau) - x_0) + \int_{\sigma=0}^{\tau} \left(F'(x(\sigma)) - A\right)\dot{x}(\sigma)d\sigma\,. \end{aligned}$$

Upon differentiating the homotopy (6.32) with respect to τ, we obtain

$$(I - \tau A)\dot{x} = F(x_0) + A(x(\tau) - x_0) = F(x_0) + \tau A(I - \tau A)^{-1}F(x_0)$$

and therefore

$$\dot{x}(\tau) = (I - \tau A)^{-2}F(x_0)\,,$$

which then readily leads to the result of the lemma. □

Discussion of Lipschitz conditions. With the above representation at hand, the question is now how to formulate first and second order Lipschitz conditions in view of theoretical estimates. The switch from the canonical norms in Sections 6.2 and 6.3 to the Euclidean norm here implies changes in all our definitions of first and second order Lipschitz constants below. Needless to mention that we are bound to lose the nice property of affine similarity in all our characterizing quantities. Instead all of our estimates will now depend on the scaling of the residual (to be carefully handled).

First order Lipschitz condition: linear contractivity. We may employ (6.8) to define some one-sided Lipschitz constant ν. Recall, however, that due to dynamical invariants, zero eigenvalues will occur in the Jacobian, which implies that $\nu \geq 0$—just apply the definition (6.8) again. Therefore, in order to take care of dynamical invariants, we will restrict our attention to iterative corrections in the subspace

$$S_P = \{u \in \mathbb{R}^n \mid P^{\perp}u = 0\}\,.$$

Then the inequality

$$(u, Au) \leq \nu\|u\|^2, \quad u \in S_P$$

is equivalent to

$$(Pu, (PAP)Pu) \leq \nu\|Pu\|^2\,.$$

With this modified definition, the case $\nu < 0$ may well happen even in the presence of dynamical invariants.

Since $\Delta x(\tau) \in S_P$, we may insert it into the above definition and obtain

$$\hat{\nu}(\tau) = \frac{(\Delta x, A\Delta x)}{\|\Delta x\|^2} \leq \nu \,. \tag{6.34}$$

If we multiply equation (6.32) by Δx from the left, we obtain

$$\begin{aligned} \|\Delta x\|^2 &= \tau(\Delta x, A\Delta x) + (\Delta x, F(x_0)) \\ &= \tau\hat{\nu}(\tau)\|\Delta x\|^2 + (\Delta x, F(x_0)) \\ &\leq \tau\hat{\nu}(\tau)\|\Delta x\|^2 + \|\Delta x\|\|F(x_0)\| \\ &\leq \tau\nu\|\Delta x\|^2 + \|\Delta x\|\|F(x_0)\| \,, \end{aligned}$$

which then leads to the estimates

$$\|\Delta x\| \leq \frac{\|F(x_0)\|}{1-\hat{\nu}\tau} \leq \frac{\|F(x_0)\|}{1-\nu\tau} \,. \tag{6.35}$$

Moreover, since

$$\Delta x(\tau) = F(x_0) + \mathcal{O}(\tau) \,,$$

we also have

$$\hat{\nu}(0) = \frac{(F(x_0), AF(x_0))}{\|F(x_0)\|^2} \leq \nu \,. \tag{6.36}$$

This quantity can be monitored even before the linear equation (6.32) is actually solved. It plays a key role in the residual reduction process, as shown in the following lemma.

Lemma 6.5 *Let* $\hat{\nu}(0) < 0$ *as defined in* (6.36)*. Then there exists some* $\tau^* > 0$ *such that*

$$\|F(x(\tau))\| < \|F(x_0)\| \quad \text{and} \quad \hat{\nu}(\tau) < 0 \quad \text{for all} \quad \tau \in [0, \tau^*[\,.$$

Proof. By differentiating the residual norm with respect to τ we obtain

$$\begin{aligned} \frac{d}{d\tau}\|F(x(\tau))\|^2|_{\tau=0} &= 2(F'(\cdot)^T F(\cdot), \dot{x}(\tau))|_{\tau=0} \\ &= 2(F(x_0), AF(x_0)) = 2\hat{\nu}(0)\|F(x_0)\|^2 < 0 \,. \end{aligned}$$

Since both $F(x(\tau))$ and the norm $\|\cdot\|$ are continuously differentiable, there exists some nonvoid interval w.r.t. τ, wherein the residual norm decreases—compare the previous Lemma 3.2. The proof of the statement for $\hat{\nu}(\tau)$ uses the same kind of argument. □

In other words: if at the given starting point x_0 the condition $\hat{\nu}(0) < 0$ is not satisfied, then the pseudo-transient continuation method based on residual reduction cannot be expected to work at all. Recall, however, the discussion at the end of Section 6.1 which pointed out that residual reduction is not coupled to canonical norm reduction.

Second order Lipschitz condition. Here we may start from the affine similar Lipschitz condition (6.15) and replace the canonical norm $|\cdot|$ therein by the Euclidean norm $\|\cdot\|$. Thus we take the fact into account that in the affine similar setting domain and image space transform in the same way.

Convergence analysis. With these preparations we are now ready to state our main result.

Theorem 6.6 *Notation as in the preceding Lemma* 6.4*, but with* $A = F'(x_0)$ *and partly* $L_0 = \|F(x_0)\|$*. Let dynamical invariants show up via the properties* $F(x) \in S_P$*. Assume the one-sided first order Lipschitz condition*

$$(u, Au) \leq \nu \|u\|^2 \quad for \quad u \in S_P, \quad \nu < 0$$

and the second order Lipschitz condition

$$\|\big(F'(x) - F'(x_0)\big)u\| \leq L_2 \|x - x_0\| \|u\| \,. \tag{6.37}$$

Then the following estimate holds

$$\|F(x(\tau))\| \leq \left(1 + \frac{\frac{1}{2}L_0 L_2 \tau^2}{1 - \nu\tau}\right) \frac{\|F(x_0)\|}{1 - \nu\tau} \,. \tag{6.38}$$

From this, residual monotonicity

$$\|F(x(\tau))\| \leq \|F(x_0)\|$$

is guaranteed for all $\tau \geq 0$ *satisfying the sufficient condition*

$$\nu + (\tfrac{1}{2}L_0 L_2 - \nu^2)\tau \leq 0 \,. \tag{6.39}$$

Moreover, if

$$L_0 L_2 > \nu^2 \,, \tag{6.40}$$

then the theoretically optimal pseudo-timestep is

$$\tau_{opt} = \frac{|\nu|}{L_0 L_2 - \nu^2} \tag{6.41}$$

leading to a residual reduction

$$\|F(x(\tau))\| \leq \left(1 - \frac{\frac{1}{2}\nu^2}{L_0 L_2}\right) \|F(x_0)\| < \|F(x_0)\| \,.$$

Proof. We return to the preceding Lemma 6.4. Obviously, the first and the second right hand terms in equation (6.33) are independent. Upon recalling (6.35) for the first term, we immediately recognize that, in order to be able to prove residual reduction, we necessarily need the condition $\nu < 0$, which means $\nu = -|\nu|$ throughout the proof. For the second term we may estimate, again recalling (6.35),

$$\begin{aligned}
\int_{\sigma=0}^{\tau} &\| \left(F'(x(\sigma)) - F'(x_0)\right)(I - \sigma A)^{-2} F(x_0)\| d\sigma \\
&\leq L_2 \int_{\sigma=0}^{\tau} \|x(\sigma) - x_0\| \|(I - \sigma A)^{-2} F(x_0)\| d\sigma \\
&\leq L_2 \int_{\sigma=0}^{\tau} \frac{\sigma \|F(x_0)\|^2}{(1 - \sigma\nu)^3} d\sigma \\
&= \tfrac{1}{2} L_2 \|F(x_0)\|^2 \tau^2 (1 - \nu\tau)^{-2} .
\end{aligned}$$

Combination of the two estimates then directly confirms (6.38), which we here write as

$$\|F(x(\tau))\| \leq \alpha(\tau) \|F(x_0)\| ,$$

in terms of

$$\alpha(\tau) = \left(1 - \nu\tau + \tfrac{1}{2} L_0 L_2 \tau^2\right) / (1 - \nu\tau)^2 .$$

Upon requiring $\alpha(\tau) \leq 1$, we obtain the equivalent sufficient condition (6.39). Finally, in order to find the optimal residual reduction, a short calculation shows that

$$\dot{\alpha}(\tau) = \left(\nu + \frac{L_0 L_2 \tau}{1 - \nu\tau}\right) / (1 - \nu\tau)^2 .$$

An interior minimum can arise only for $\dot{\alpha}(\tau) = 0$, which is equivalent to (6.41) under the condition (6.40). Insertion of τ_{opt} into the expression for $\alpha(\tau)$ then completes the proof. □

From the above condition (6.39) we may conclude: if

$$\nu + \tfrac{1}{2}\sqrt{2 L_0 L_2} \leq 0 ,$$

then τ is *unbounded* for local continuation. Obviously, this is the residual oriented *nonlinear contractivity* condition to be compared with the error oriented relation (6.20). (The difference in the prefactor just indicates that there we needed to show uniqueness in addition.) If

$$\nu + \tfrac{1}{2}\sqrt{2 L_0 L_2} > 0 , \qquad (6.42)$$

then the pseudo-timestep is bounded according to

$$\tau \leq \frac{|\nu|}{\frac{1}{2}L_0 L_2 - |\nu|^2}\,.$$

Note that condition (6.40) is less restrictive than (6.42) so that either unbounded or bounded optimal timesteps may occur.

Adaptive (pseudo-)timestep strategy. In the spirit of the whole book we now want to derive an adaptive strategy based on the theoretical optimal pseudo-timestep (6.41), which we repeat for convenience

$$\tau_{opt} = \frac{|\nu|}{L_0 L_2 - \nu^2}\,.$$

The above expression can be rewritten in implicit form as

$$\tau_{opt} = \frac{|\nu|(1-\nu\tau_{opt})}{L_0 L_2}\,. \tag{6.43}$$

In passing we note that from this representation we roughly obtain

$$\tau_{opt} \sim \frac{1}{L_0} = \frac{1}{\|F(x_0)\|}\,,$$

which gives some justification for a quite popular *heuristic strategy*: new timesteps are proposed on the basis of successful old ones via

$$\tau_{new} = \frac{\|F(x_{old})\|}{\|F(x_{new})\|}\,\tau_{old}\,. \tag{6.44}$$

For reference see, e.g., the recent paper [133] by C.T. Kelley and D.E. Keyes, where also a whole class of further heuristics is mentioned. A different approach is taken by S.B. Hazra, V. Schulz, J. Brezillon, and N. Gauger in [116] where in a fluid dynamical problem no overall timestep exists; this approach is not treated here.

Here, however, we want to exploit the structure of (6.43) in a different way by rewriting it in the form

$$\|\Delta x(\tau_{opt})\| L_2 \tau_{opt} \leq \frac{L_0 L_2}{1-\nu\tau_{opt}} = |\nu|\,.$$

On this basis, we replace τ_{opt} by the upper bound

$$\bar{\tau}_{opt} = \frac{|\nu|}{L_2\|\Delta x(\tau)\|} \geq \tau_{opt}\,.$$

So we are left with the task of identifying cheap computational estimates $[\nu] \leq \nu < 0, [L_2] \leq L_2$ to replace the unknown theoretical quantities ν, L_2. Once this is achieved, we can compute the corresponding pseudo-timestep

$$[\tau_{opt}] = \frac{|[\nu]|}{[L_2]\|\Delta x(\tau)\|} \geq \bar{\tau}_{opt} \geq \tau_{opt}\,. \tag{6.45}$$

As for the estimation of ν, we may exploit (6.34) in a double way: First, whenever

$$\|\Delta x(\tau)\| \geq \|F(x_0)\|\,,$$

then we know that $\nu \geq [\nu] \geq 0$ is guaranteed, which means that we should terminate the iteration. Second, we may recognize that

$$[\nu]\tau = \hat{\nu}(\tau)\tau = \tau\frac{(\Delta x, A\Delta x)}{\|\Delta x\|^2} = \frac{(\Delta x, \Delta x - F(x_0))}{\|\Delta x\|^2} \leq \nu\tau \tag{6.46}$$

gives us a quite cheap estimation formula for ν. As for the estimation of L_2, we may rearrange terms in the proof of Theorem 6.6 to obtain

$$\begin{aligned} \|F(x(\tau)) - \Delta x(\tau)\| &\leq \int\limits_{\sigma=0}^{\tau} \|\left(F'(x(\sigma)) - F'(x_0)\right)(I - \sigma A)^{-2}F(x_0)\|d\sigma \\ &\leq L_2 \int\limits_{\sigma=0}^{\tau} \|x(\sigma) - x_0\|\|(I - \sigma A)^{-1}\Delta x(\sigma)\|d\sigma\,. \end{aligned}$$

If we approximate the integral by the trapezoidal rule, we arrive at

$$\begin{aligned} \|F(x(\tau)) - \Delta x(\tau)\| &\leq \tfrac{1}{2}L_2\|\Delta x(\tau)\|^2\tau^2/(1 - \nu\tau) + \mathcal{O}(\tau^4) \\ &\leq \tfrac{1}{2}L_2\|\Delta x(\tau)\|^2\tau^2 + \mathcal{O}(\tau^4)\,. \end{aligned}$$

Note that already the approximation term, ignoring the $\mathcal{O}(\tau^4)$ term, gives rise to the upper bound

$$\|F(x(\tau)) - \Delta x(\tau)\| \leq \tfrac{1}{2}L_2\|F(x_0)\|^2\tau^2/(1 - \nu\tau)^2\,,$$

which is the basis of the derivation of Theorem 6.6. Hence, we may well regard

$$[L_2] = \frac{2\|F(x(\tau)) - \Delta x\|}{\tau^2\|\Delta x\|^2} \leq L_2 + \mathcal{O}(\tau^2)$$

as a suitable computational estimate for L_2. Upon collecting all above estimates and inserting them into (6.45), we arrive at the following pseudo-timestep suggestion

$$[\tau_{opt}] = \frac{|(\Delta x(\tau), F(x_0) - \Delta x(\tau))|}{2\|\Delta x(\tau)\|\|F(x(\tau)) - \Delta x(\tau)\|}\,\tau\,.$$

On this basis, an *adaptive τ-strategy* can be realized in the usual two modes, a *correction* and a *prediction* strategy: If in the iterative step $x_0 \longrightarrow x(\tau)$ the residual norm does *not* decrease, then the actual step size τ is replaced by

$[\tau_{opt}] < \tau$; if the residual norm decreases successfully, then the next step is started with the trial value $[\tau_{opt}]$. Finally, note that the above strategy will just terminate, if the steady state to be computed is not attractive in the (residual) sense that $[\nu] \geq 0$. For $[\nu] \to 0^-$, the suggested stepsize behaves like $[\tau_{opt}] \to 0^+$ - as to be reasonably required.

6.4.2 Inexact pseudo-transient continuation

Suppose the linear system (6.32) is so large that we cannot but solve it *iteratively* $(i = 0, 1, \ldots)$:

$$(I - \tau A)\delta x_i = F(x_0) - r_i, \quad x_i(\tau) = x_0 + \tau\delta x_i\,. \tag{6.47}$$

Herein r_i represents the iterative linear residual, δx_i the corresponding inexact correction, and $x_i(\tau)$ the approximate homotopy path instead of the exact $x(\tau)$. To start the iteration, let $x_0(\tau) = x_0$ so that $\delta x_0 = 0$ and $r_0 = F(x_0)$.

If we want to *minimize the residuals* within each iterative step, we are directly led to GMRES—see Section 1.4 and the notation therein. In terms of the Euclidean norm $\|\cdot\|$ we define the approximation quantities

$$\eta_i := \frac{\|r_i\|}{\|F(x_0)\|} < 1 \quad for \quad i = 1, 2, \ldots\,.$$

Recall that GMRES assures $\eta_{i+1} \leq \eta_i$, in the generic case even $\eta_{i+1} < \eta_i$. Moreover, due to the residual minimization property and $r_0 = F(x_0)$, we have

$$\|F(x_0) - r_i\|^2 = (1 - \eta_i^2)\|F(x_0)\|^2\,.$$

In the present context of pseudo-transient continuation, we may additionally observe that GMRES realizes the special structure

$$\delta x_i(\tau) = V_i z_i(\tau) \quad \text{and} \quad H_i(\tau) = (I_i, 0)^T + \tau\hat{H}_i\,.$$

Herein V_i is just the orthonormal basis of the Krylov space $\mathcal{K}_i(r_0, A)$ and $\hat{H}_i$ is a Hessenberg matrix like $H_i(\tau)$, but also independent of τ. On this basis, we see that dynamical invariants are correctly treated throughout the iteration. The proof of these properties is left as Exercise 6.5. The special structure permits computational savings when the same system is solved for different pseudo-timesteps τ.

Convergence analysis. As before, we first analyze the convergence behavior theoretically as a basis for the subsequent derivation of an adaptive algorithm, which here will have to include the matching of inner and outer iteration. For this purpose we need to modify Lemma 6.4.

Lemma 6.7 *Notation as in Lemma* 6.4 *with* $A \approx F'(x_0)$. *Then the residual along the approximate homotopy path* $x_i(\tau)$ *starting at* x_0 *satisfies*

$$F(x(\tau)) - r_i = (I - \tau A)^{-1}\left(F(x_0) - r_i\right) + \int\limits_{\sigma=0}^{\tau} \left(F'(x_i(\sigma)) - A\right)(I - \sigma A)^{-2}\left(F(x_0) - r_i\right) d\sigma\,.$$

Proof. The proof is just an elementary modification of the proof of Lemma 6.4. For example, if we differentiate the homotopy (6.47) with respect to τ, we now obtain

$$\dot{x}_i(\tau) = (I - \tau A)^{-2}\left(F(x_0) - r_i\right)\,.$$

Further details can be omitted. □

Theorem 6.8 *Notation as in the preceding Lemma* 6.7. *Let* $A = F'(x_0)$ *and partly* $\tilde{L}_0 = \sqrt{1-\eta_i^2}\|F(x_0)\|$. *Assume that dynamical invariants show up via the properties* $F(x) \in S_P$. *Then, with the Lipschitz conditions*

$$(u, Au) \le \nu\|u\|^2, \quad \nu < 0, \quad for \quad u \in S_P$$

and

$$\|\big(F'(x) - F'(x_0)\big)u\| \le L_2\|x - x_0\|\|u\|\,,$$

the estimates

$$\|F(x(\tau)) - r_i\| \le \left(1 + \frac{\frac{1}{2}\tilde{L}_0 L_2 \tau^2}{1 - \nu\tau}\right)\frac{\|F(x_0) - r_i\|}{1 - \nu\tau}$$

and

$$\|F(x(\tau))\| \le \left(\eta_i + \frac{\sqrt{1-\eta_i^2}}{1 - \nu\tau}\left(1 + \frac{\frac{1}{2}\tilde{L}_0 L_2 \tau^2}{1 - \nu\tau}\right)\right)\|F(x_0)\|$$

hold. Let

$$s(\eta_i) := \sqrt{\frac{1-\eta_i}{1+\eta_i}} > \tfrac{1}{2} \quad or,\ equivalently, \quad \eta_i < \tfrac{3}{5}\,. \qquad (6.48)$$

Then residual monotonicity

$$\|F(x(\tau))\| \le \|F(x_0)\|$$

is guaranteed for all $\tau \ge 0$ *satisfying the sufficient condition*

$$1 - s(\eta_i) + (2s(\eta_i) - 1)\nu\tau + \left(\tfrac{1}{2}\tilde{L}_0 L_2 - s(\eta_i)\nu^2\right)\tau^2 \le 0\,. \qquad (6.49)$$

Assume further that

$$\tfrac{4}{5}L_0L_2 > \nu^2 \tag{6.50}$$

and that `GMRES` *has been continued until*

$$\eta_i + \sqrt{1-\eta_i^2} < 1 + \frac{\frac{1}{2}\nu^2}{L_0L_2}\,. \tag{6.51}$$

Then the theoretically optimal pseudo-timestep is

$$\tau_{opt} = \frac{|\nu|}{\tilde{L}_0L_2 - \nu^2} \tag{6.52}$$

leading to the estimate

$$\|F(x(\tau)) - r_i\| \le \left(1 - \frac{\frac{1}{2}\nu^2}{\tilde{L}_0L_2}\right)\|F(x_0) - r_i\| < \|F(x_0) - r_i\|$$

and to the residual reduction

$$\|F(x(\tau))\| \le \left(\eta_i + \sqrt{1-\eta_i^2} - \frac{\frac{1}{2}\nu^2}{L_0L_2}\right)\|F(x_0)\| < \|F(x_0)\|\,. \tag{6.53}$$

Proof. We return to the preceding Lemma 6.7 and modify the proof of Theorem 6.6 carefully step by step. For example, the second order term may be estimated as

$$\int_{\sigma=0}^{\tau} \|\left(F'(x(\sigma)) - F'(x_0)\right)(I - \sigma A)^{-2}(F(x_0) - r_i)\|d\sigma$$
$$\le \tfrac{1}{2}L_2\|F(x_0) - r_i\|^2\tau^2(1-\nu\tau)^{-2}\,.$$

Combination of estimates then directly confirms

$$\|F(x_i(\tau)) - r_i\| \le \bar{\alpha}_i(\tau)\|F(x_0) - r_i\|$$

in terms of

$$\bar{\alpha}_i(\tau) = \left(1 - \nu\tau + \tfrac{1}{2}\tilde{L}_0L_2\tau^2\right)/(1-\nu\tau)^2\,,$$

from which we obtain

$$\|F(x_i(\tau))\| \le \alpha_i(\tau)\|F(x_0)\|$$

with

$$\alpha_i(\tau) = \eta_i + \sqrt{1-\eta_i^2}\,\bar{\alpha}_i(\tau)\,.$$

Upon requiring $\alpha(\tau) \le 1$, we have

$$\eta_i + \sqrt{1-\eta_i^2}\,\bar{\alpha}_i(\tau) \le 1\,,$$

which is equivalent to

$$\bar{\alpha}_i(\tau) \leq s(\eta_i) \leq 1\,. \tag{6.54}$$

From this, we immediately verify the sufficient condition (6.49). Note that $2s-1>0$, which is just condition (6.48), is necessary to have at least one negative term in the left hand side of (6.49).

Finally, in order to find the *optimal* residual reduction, a short calculation shows that

$$\dot{\alpha}(\tau) = \sqrt{1-\eta_i^2}\,\dot{\bar{\alpha}}(\tau) = \frac{\sqrt{1-\eta_i^2}}{(1-\nu\tau)^2}\left(\nu + \frac{\tilde{L}_0 L_2 \tau}{1-\nu\tau}\right)\,.$$

For the interior minimum we require $\dot{\bar{\alpha}}(\tau) = 0$, which is equivalent to (6.52) under the condition (6.50), where

$$\sqrt{1-\eta_i^2} \geq \sqrt{1-(\tfrac{3}{5})^2} = \tfrac{4}{5}$$

has been used. Insertion of τ_{opt} into the expression for $\alpha(\tau)$ then leads to

$$\|F(x_i(\tau)) - r_i\| \leq \left(1 - \frac{\frac{1}{2}\nu^2}{\tilde{L}_0 L_2}\right)\|F(x_0) - r_i\|$$

and eventually to (6.53). In order to assure an actual *residual reduction*, condition (6.54) must also hold for τ_{opt}, which confirms the necessary condition (6.51). Note that the scalar function $\eta_i + \sqrt{1-\eta_i^2}$ is monotonically increasing for $\eta_i < \frac{1}{2}\sqrt{2} \approx 0.7$, hence also for $\eta_i < \frac{3}{5} = 0.6$. Therefore GMRES may be just continued until the relation (6.51) is satisfied. This completes the proof. □

Adaptive (pseudo-)timestep strategy. We follow the line of the derivation for the exact pseudo-transient continuation in Section 6.4.1. For convenience, we repeat the expression

$$\tau_{opt} = \frac{|\nu|}{\tilde{L}_0 L_2 - \nu^2}\,,$$

which can be rewritten in implicit form as

$$\tau_{opt} = \frac{|\nu|(1-\nu\tau_{opt})}{\tilde{L}_0 L_2}\,.$$

Recall now that

$$\|\delta x_i(\tau)\| = \|(I-\tau A)^{-1}(F(x_0)-r_i)\| \leq \frac{\|F(x_0)-r_i\|}{1-\nu\tau} = \frac{\tilde{L}_0}{1-\nu\tau}\,, \tag{6.55}$$

which directly implies

$$\bar{\tau}_{opt} = \frac{|\nu|}{L_2\|\delta x_i(\tau)\|} \geq \tau_{opt} \,.$$

So we need to compute the pseudo-timestep

$$[\tau_{opt}] = \frac{|[\nu]|}{[L_2]\|\delta x_i(\tau)\|} \geq \bar{\tau}_{opt} \geq \tau_{opt} \tag{6.56}$$

in terms of the appropriate estimates of the unknown theoretical quantities ν, L_2.

As for the estimation of ν, we exploit (6.55). Whenever

$$\|\delta x_i(\tau)\| \geq \|F(x_0) - r_i\| \,,$$

then we know that $\nu \geq 0$ is guaranteed and the iteration must be terminated. Moreover, the relation

$$[\nu]\tau = \tau\frac{(\delta x_i, A\delta x_i)}{\|\delta x_i\|^2} = \frac{(\delta x_i, \delta x_i - F(x_0) + r_i)}{\|\delta x_i\|^2} \leq \nu\tau$$

supplies an estimation formula for ν. As for the estimation of L_2, we revisit Lemma 6.7 to obtain

$$\begin{aligned}
&\|F(x_i(\tau)) - r_i - \delta x_i(\tau)\| \\
&\leq \int_{\sigma=0}^{\tau} \|\left(F'(x_i(\sigma)) - F'(x_0)\right)(I - \sigma A)^{-2}\left(F(x_0) - r_i\right)\|d\sigma \\
&\leq L_2 \int_{\sigma=0}^{\tau} \|x_i(\sigma) - x_0\|\|(I - \sigma A)^{-1}\delta x_i(\sigma)\|d\sigma \\
&\leq \tfrac{1}{2}L_2\tau^2\frac{\tilde{L}_0^2}{(1-\nu\tau)^2} \,.
\end{aligned}$$

If we approximate the above integral by the trapezoidal rule (*before* using the final estimate), we arrive at

$$\begin{aligned}
\|F(x_i(\tau)) - r_i - \delta x_i(\tau)\| &\leq \tfrac{1}{2}L_2\|\delta x_i(\tau)\|^2\tau^2/(1-\nu\tau) + \mathcal{O}(\tau^4) \\
&\leq \tfrac{1}{2}L_2\tau^2\|\delta x_i(\tau)\|^2 + \mathcal{O}(\tau^4) \,.
\end{aligned}$$

Already the first right hand term gives rise to the above upper bound—compare (6.55). Hence, as in Section 6.4.1, we will pick

$$[L_2] = \frac{2\|F(x_i(\tau)) - r_i - \delta x_i(\tau)\|}{\tau^2\|\delta x_i(\tau)\|^2} \leq L_2 + \mathcal{O}(\tau^2)$$

as computational estimate for L_2. Upon inserting the two derived estimates into (6.56), we arrive at the pseudo-timestep estimate

$$[\tau_{opt}] = \frac{|(\delta x_i(\tau), F(x_0) - r_i - \delta x_i(\tau))|}{2\|\delta x_i(\tau)\|\|F(x_i(\tau)) - r_i - \delta x_i(\tau)\|}\,\tau\,.$$

On this basis, an *adaptive τ-strategy* can again be realized as in the case of the *exact* pseudo-transient continuation method.

Finally, we want to mention that the iterative version of the pseudo-transient continuation method still works in the case of *unbounded* timestep. To see this, just rewrite (6.47) in the form

$$\left(\frac{1}{\tau}I - A\right)(x_i(\tau) - x_0) = F(x_0) - r_i\,.$$

Herein $\tau \to \infty$ is possible leaving $x_i(\tau) - x_0$ well-defined even in the presence of singular Jacobian A caused by *dynamical invariants*: This is due to the fact that `GMRES` (like any Krylov solver) keeps the nullspace components of the solution unchanged, so that $P^{\perp}(x_i(\infty) - x_0) = 0$ is guaranteed throughout the iteration.

Preconditioning. If we multiply the nonlinear system by means of some nonsingular matrix M from the left as

$$M\dot{x} = MF(x) = 0\,,$$

then `GMRES` will have to work on the preconditioned residuals Mr_i and adaptivity must be based on norms $\|M \cdot \|$. Note that it is totally unclear, whether such a transformation leads to the necessary linear contractivity result $\nu(MA) < 0$ for the preconditioned system with $A \approx F'(x^0)$.

Preconditioning from the right will just influence the convergence speed of `GMRES` without changing the above derived adaptivity devices.

Matrix-free realization. Sometimes the inexact pseudo-continuation method is realized in a matrix-free variant using the first order approximation

$$A\delta x \approx F(x + \delta x) - F(x)\,.$$

A numerically stable realization will use *automatic differentiation* as suggested by A. Griewank [112].

Exercises

Exercise 6.1 Prove the results (6.26) for the implicit trapezoidal rule (6.25) and (6.28) for the implicit midpoint rule (6.27).

Exercise 6.2 Consider the linearly implicit Euler (LIE) discretization for the ODE system $y' = f(y)$, which reads (for $k = 0, 1, \ldots$)

$$y_{k+1} = y_k + (I - \tau A)^{-1} f(y_k)\,,$$

where $A = f_y(y_k)$. This scheme is usually monitored to run in some 'neighborhood' of the implicit Euler (IE) discretization

$$F(y) = y - y_k - \tau f(y) = 0\,.$$

For this purpose the LIE is interpreted as the first iterate of IE and local contraction within that IE scheme is required. Most LIE codes realize this requirement via an error oriented criterion introduced in [66]. Here we want to look at a residual based variant due to [120].

a) On the basis of the residual based Newton-Mysovskikh theorem derive a computational monitor that is cheap to evaluate.

b) Of which order $O(\tau^s)$ is this contraction factor? Derive an adaptive stepsize procedure on that basis.

   Hint: Interpret the method as a continuation method with embedding parameter τ.

c) Compare the error oriented and the residual based variant in terms of computational amount per discretization step.

d) *Optional:* Implement the two variants within an adaptive integrator (like `LIMEX`) and compare them at several ODE examples.

Exercise 6.3 Consider the system of n nonlinear differential equations (with time variable t)

$$\dot{x} = F(x)\,,\ x(0) = x^0$$

modeling some process $x(t)$. Assume that there exists a *dynamical invariant* (such as mass conservation) of the form

$$e^T x(t) = e^T x^0\,,\ e^T = (1, \ldots, 1) \in \mathbb{R}^n\,.$$

In many cases, one is only interested in a steady state solution $x^* = x(\infty)$ defined by

$$F(x^*) = 0\,.$$

Since, in general, x^* will depend on the initial value x^0, uniqueness of the solution is not guaranteed.

a) Show that the Jacobian $F'(x^*)$ is singular, which makes a naive application of Newton methods impossible.

b) As a remedy, consider the iterative method

$$\begin{bmatrix} F'(x^k) \\ e^T \end{bmatrix} \Delta x^k = -\begin{bmatrix} F(x^k) \\ 0 \end{bmatrix}, \qquad x^{k+1} := x^k + \Delta x^k$$

started by some initial guess x^0. Show that the thus produced iterates satisfy

$$e^T x^k = e^T x^0 \,.$$

What kind of restriction is necessary for the choice of x^0?

c) Develop a program to treat the described problem type—test examples may come from chemical kinetics, where mass conservation is often tacitly assumed without explicitly stating it.

Exercise 6.4 Consider the pseudo-transient continuation method with *approximate* Jacobian $A \approx F'(x^0)$. Upon using the notation of Section 6.4.1 and, in addition,

$$\|(A - F'(x^0))u\| \leq \delta|\nu|\|u\| \,, \quad \delta < 1 \,,$$

prove a variant of Theorem 6.6, containing results on the residual descent and the optimal pseudo-timestep.

Check: For $\nu < 0$, the optimal timestep τ_{opt} comes out to be

$$\tau_{opt} = \frac{(1-\delta)|\nu|}{L_0 L_2 - (1-\delta)\nu^2}$$

in the terms defined—assuming, of course, that the denominator is positive.

Exercise 6.5 How can the iterative linear solver `GMRES` be optimally adapted to pseudo-transient continuation? Design a special version, which saves computing time and storage.