Stiffness[1]

What is a stiff system? Definitions of the term "stiffness" abound, and none is completely satisfactory. To illustrate the concepts we start with a simple example involving radioactive decay.

An Example. Suppose we have three elements (or isotopes) A, B, and C. A decays into B at a rate k_1, B decays into C at a rate k_2. C is stable. In this particular example we assume that B is highly unstable and therefore

$$k_2 \gg k_1. \tag{1}$$

Physically, the reciprocals of the rate constants indicate the amount of time during which the amount of a species is reduced by a factor e. For the sake of illustration let's assume that

$$k_1 = 10^{-9}[\text{sec}^{-1}] \quad \text{and} \quad k_2 = 10^{9}[\text{sec}^{-1}]. \tag{2}$$

Thus the rate constant k_1 corresponds to about 30 years. These constants occur in realistic problems involving the disposal of radioactive waste. Our time interval of interest would be about 10,000 years, i.e., 3×10^{11} seconds. Suppose a, b, and c denote the relative abundances of A, B, and C, respectively, and we start out with a sample of pure A. Thus we obtain the following (autonomous, linear, constant coefficient, homogeneous) initial value problem:

$$\begin{pmatrix} a' \\ b' \\ c' \end{pmatrix} = \begin{pmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \begin{pmatrix} a(0) \\ b(0) \\ c(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{3}$$

Clearly the eigenvalues of A are $-k_1$, $-k_2$, and 0. The corresponding eigenvectors are:

$$c_1 = \begin{pmatrix} 1 \\ \dfrac{k_1}{k_2 - k_1} \\ \dfrac{k_2}{k_1 - k_2} \end{pmatrix}, \quad c_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \quad c_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{4}$$

The general solution of the differential equation is given by

$$\begin{pmatrix} a(t) \\ b(t) \\ c(t) \end{pmatrix} = \gamma_1 e^{-k_1 t} \begin{pmatrix} 1 \\ \dfrac{k_1}{k_2 - k_1} \\ \dfrac{k_2}{k_1 - k_2} \end{pmatrix} + \gamma_2 e^{-k_2 t} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} + \gamma_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{5}$$

where the γ_i are determined by the initial conditions. In our case:

$$\gamma_1 = \gamma_3 = 1 \quad \text{and} \quad \gamma_2 = \frac{k_1}{k_1 - k_2}. \tag{6}$$

It is obvious that the amount of A will be slowly *decreasing*, and the amount of C slowly *increasing*. More interesting is the concentration of B:

$$b(t) = \frac{k_1}{k_2 - k_1} e^{-k_1 t} + \frac{k_1}{k_1 - k_2} e^{-k_2 t} = \frac{k_1}{k_2 - k_1} \left(e^{-k_1 t} - e^{-k_2 t} \right). \tag{7}$$

[1] by Peter Alfeld. TEX processing date July 16, 2014

Thus b(t) starts out at 0, reaches a maximum of approximately

$$\frac{k_1}{k_2 - k_1} \approx 10^{-18} \tag{8}$$

after about

$$\frac{\ln \frac{k2}{k_1}}{k_2} \approx 4 \times 10^{-8} \text{seconds} \tag{9}$$

and slowly decays thereafter. Loosely speaking, the period prior to reaching the maximum is called the *transient phase*, the period after reaching the maximum is called the *steady state phase*.

Now consider the step-size requirements on the Forward and Backward Euler Methods, imposed by local accuracy and stability, before and after reaching the maximum. Suppose for accuracy we require that the local truncation error (in b) be no larger than (the somewhat arbitrary) value 10^{-24}.

For the Backward Euler method, absolute stability presents no limitation. For Euler's method, we must have that

$$-k_2 h \in (-2, 0) \quad \Longrightarrow \quad h < 2 \times 10^{-9}. \tag{10}$$

The leading term of the local truncation error for the Euler methods is $\pm\frac{1}{2}h^2 b''(t)$, where

$$b''(t) = \frac{k_1}{k_2 - k_1}\left(k_1^2 e^{-k_1 t} - k_2^2 e^{-k_2 t}\right). \tag{11}$$

In the transient phase the k_2 term is dominant, in the steady state phase, the k_1 term dominates. Ignoring the subdominant terms altogether, we obtain

$$b''(t) \approx \begin{cases} 10^{0} & \text{in the transient phase} \\ 10^{-36} & \text{in the steady state phase} \end{cases}. \tag{12}$$

Thus the maximum allowable values of h would be

$$h_{\max} \approx \begin{cases} 10^{-12} & \text{in the transient phase} \\ 10^{6} & \text{in the steady state phase} \end{cases}. \tag{13}$$

(Of course in reality the maximum allowable step-size starts out at a suitably small value and then gradually increases as the integration proceeds.)

Notes:

1. The example is characterized by the presence of two distinct time scales. Action on the fast time scale (the transient phase) is over rapidly, and the remainder of the action proceeds on the slow time scale.
2. However, stability for the explicit method is governed by the fast time scale.
3. In the transient phase, the step-size for both methods is determined by the action on the fast time scale.
4. In the steady phase the implicit method is unencumbered by stability considerations, whereas the explicit method is reduced to a step-size roughly 10^{15} times smaller than that called for by local accuracy. In other words, we can afford 10^{15} times the effort per step for the implicit method than that for the explicit method.

Definition of Stiffness. For our purposes, we consider a problem stiff if, for an explicit method, stability imposes more stringent step-size restrictions than local accuracy. The weakness of this definition is that it is dependent upon the method. This however, is not serious. The typical interval of absolute stability of an explicit method has a length of 1 within a factor 10, or so.

A typical definition of stiffness you might find in the literature is based on the ratio of the smallest and largest negative real parts of the eigenvalue. This is usually coupled with the assumption that no real parts are positive. The larger that ratio, the stiffer the problem. This gives rise to ridiculous consequences, for example, a problem with eigenvalues -2,-1,0 would be infinitely stiffer than a problem with eigenvalues $-10^9, -1, -10^{-9}$.

Note that according to our concept of stiffness, our example problem is not stiff in the transient phase.

Now what? We saw that an explicit linear multistep method must have a bounded region of absolute stability. We therefore investigate *implicit* linear multistep methods. To match the behavior of the true solution of the test equation

$$y' = \lambda y \tag{14}$$

ideally the region of absolute stability should be the open left half plane. It is practical to relax the requirement slightly:

Definition 1. *A linear multistep method with first and second characteristic polynomial ρ and σ is A-stable if*

$$\mathbf{Re} h\lambda < 0 \quad \Longrightarrow \quad h\lambda \in \mathcal{A}(\rho, \sigma) \tag{15}$$

where $\mathcal{A}(\rho, \sigma)$ is the region of absolute stability.

Remark 2. *Thus we allow for the region of absolute stability to contain parts of the right half plane.*

Remark 3. *The above definition is restricted to linear multistep methods because those are the only methods we know. But it could be extended in an obvious way to other types of methods.*

Now we ask for the existence and specification of A-stable methods, and experience

Setback 1. The following is due to Dahlquist:

Theorem 4. *An explicit linear multistep method cannot be A-stable. The order of an A-stable implicit linear multistep method cannot exceed 2. The second order A-stable implicit linear multistep method with the smallest error constant is the Trapezoidal Rule.*

Setback 2. Thus we seem to be restricted to low order methods. Moreover, one can argue that A-stability is not enough. The numerical solution of the test equation (14) by the Trapezoidal Rule

$$y_{n+1} - y_n = \frac{h}{2}\left(f_n + f_{n+1}\right) \tag{16}$$

is given by

$$y_{n+1} = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} y_n. \tag{17}$$

Now consider the case that λ is real and negative, and tends to $-\infty$. Then the analytic solution tends to zero faster and faster, whereas the numerical solution oscillates, and decays slower and slower.

This gives rise to the more stringent stability concept due to Ehle:

Definition 5. *A one-step numerical method is L-stable if it is A-stable and, in addition, when applied to the test equation (14), it yields*

$$y_{n+1} = R(h\lambda)y_n \quad \textit{where} \quad |R(h\lambda)| \longrightarrow 0 \quad \textit{as} \quad \mathbf{Re}h\lambda \longrightarrow -\infty. \tag{18}$$

The only L-stable and convergent[2] linear multistep method is the Backward Euler method

$$y_{n+1} - y_n = hf_{n+1}. \tag{19}$$

Thus we seem to have ended up at the conclusion that the Backward Euler method is the only acceptable method for stiff problems. The situation is not quite that bleak, however.

Rising from the Ashes. The key to success is relaxing the requirement of Absolute Stability. Usually the imaginary part of an eigenvalue is not that large relative to the real part. (If it is we have a "highly oscillatory" problem, which forms a special class and requires special techniques.) Thus the eigenvalues will usually lie in a wedge emanating from the origin, and symmetric about the real line. More formally, we define:

Definition 6. *A numerical method is said to be $A(\alpha)$-stable if its region of absolute stability contains the infinite wedge*

$$W_\alpha = \{h\lambda | -\alpha < \pi - \arg h\lambda < \alpha\}\,; \tag{20}$$

it is $A(0)$-stable if it is $A(\alpha)$-stable for some (sufficiently small) $\alpha > 0$; it is A_0-stable if it contains all negative real $h\lambda$.

To select a suitable A(α)-stable method we return to the concept of L-stability. We would like the numerical solution to converge to zero the faster the larger the negative real part of $h\lambda$. Recall that the behavior of the numerical solution is governed by the roots of the stability polynomial

$$\pi(h\lambda, r) = \rho(r) - h\lambda\sigma(r). \tag{21}$$

To satisfy our requirement, the roots should tend to zero as $h\lambda$ tends to $-\infty$. Thus we require:

$$\sigma(r) = \beta_k r^k \tag{22}$$

and determine ρ by the requirement that the order of the method is as high as possible, i.e.,

$$p = k. \tag{23}$$

The methods so defined are the *Backward Differentiation Methods*, listed in the following Table:

k	β_k	α_6	α_5	α_4	α_3	α_2	α_1	α_0	α
1	1						1	-1	90°
2	$\frac{2}{3}$					1	-$\frac{4}{3}$	$\frac{1}{3}$	90°
3	$\frac{6}{11}$				1	-$\frac{18}{11}$	$\frac{9}{11}$	-$\frac{2}{11}$	88°
4	$\frac{12}{25}$			1	-$\frac{48}{25}$	$\frac{36}{25}$	-$\frac{16}{25}$	$\frac{3}{25}$	73°
5	$\frac{60}{137}$		1	-$\frac{300}{137}$	$\frac{300}{137}$	-$\frac{200}{137}$	$\frac{75}{137}$	-$\frac{12}{137}$	51°
6	$\frac{60}{147}$	1	-$\frac{360}{147}$	$\frac{450}{147}$	-$\frac{400}{147}$	$\frac{225}{147}$	-$\frac{72}{147}$	$\frac{10}{147}$	18°

[2] Can you think of an L-stable and non-convergent linear multistep method?

Final Drawback. However, it turns out that Backward Differentiation methods are not zero-stable if

$$k > 6. \tag{24}$$

Solution of the Nonlinear System. Running an implicit linear multistep method requires the solution of a nonlinear system at each step. It is tempting to consider the standard corrector iterations:

$$y_{n+k}^{[m+1]} = -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \sum_{j=0}^{k-1} \beta_j f_{n+j} + h\beta_k f\left(x_{n+k}, y_{n+k}^{[m]}\right). \tag{25}$$

This is a fixed point iteration which requires for its convergence that

$$\|h\beta_k f_y()\| < 1 \tag{26}$$

in a suitable domain. A necessary condition for this to hold is that for *all eigenvalues* λ of f_y

$$|h\beta_k \lambda| < 1 \tag{27}$$

which is not going to be true for a stiff problem (since if it was we might as well use an explicit method, and, according to our definition, the problem wouldn't be stiff).

Thus we have to use a form of Newton's method which requires knowledge of the Jacobian f_y or an estimate of it. However, we will always have a good starting vector, obtained e.g., by using an explicit linear multistep method.

Moreover, since the Jacobian has to be present anyway we might as well exploit it for other purposes, e.g., for improving stability properties. This leads to Skeel and Kong's *blended linear multistep methods".* These are pretty ingenious but they do not seem to be known or used as much as they deserve. For details see: Skeel, R.D., and Kong, A.K., *Blended Linear Multistep Methods*, ACM Trans. Math. Softw. 3,4, (Dec. 1977), 326–345.