Stability and Conditioning in Numerical Analysis

D. Trigiante

University of Florence

trigiant@unifi.it

September, 2005.
"... It is by looking into the same problem from different points of view that one arrives to a complete insight of it. "

Euler
Stability vs. Conditioning

The terms *stability* and *conditioning* are used with a variety of meanings in Numerical Analysis. They have in common the general concept of the response of a set of computations to perturbations arising from

- the data,
- the specific arithmetic used on computers.

They are not synonymous.
In Mathematics the notion of stability derives from the *homonymous notion in mechanics*.

It regards the behavior of the motion of a system when it is moved away from the equilibrium. Three ingredients enters in the definition, i.e.

- the existence of a reference solution, i.e. the equilibrium;
- the perturbation of the initial status (the initial conditions);
- the duration of the motion, which is supposed to be infinite.
Stability

The behaviors around the equilibrium may be many:

- stability (marginal stability), uniform stability,
- asymptotic stability (AS), uniform AS,
- contractivity,
- orbital stability,
- instability,
- ...
General Perturbations

More general perturbations are also considered in the qualitative theory of dynamical systems.

The concept of stability under perturbation of the whole system, often called total stability, is also considered.

This kind of perturbation is frequent in Numerical Analysis where the source of errors, due to the computer arithmetic, can be seen as perturbation of the whole set of computations.

In other words, a numerical algorithm is not only perturbed by the errors in the data, but also with respect to the errors arising in the process of computations.
Conditioning

Many problems, however, do not last for a long (in principle, infinite) time, and/or do not have an equilibrium.

The above concept of stability do not apply, as it stands.

The non numerical analysis distinguishes such problems in well posed and ill-posed, according whether the solution depends continuously on data or not.
This is not enough for the Numerical Analysis purposes, where a more refined distinction is needed.
The numerical analysts would like to know if such dependence, although continuous, may result disastrous for the error growth.

This requires the notion of Conditioning.
Often, in the N.A. literature, two or more concepts are superimposed. For example, concepts such as discretization (in the case where the original problem is continuous), the stability of the algorithms and the ability of the arithmetic system implemented on the computers to perform operations with acceptable relative errors are often strictly tight together.

We almost completely agree with the following statements taken from Lax and Richmyer (written in 1956!).
We shall not be concerned with rounding errors..., but it will be evident to the reader that there is an intimate connection between stability and practicality of the equations from the point of view of growth and the amplification of rounding errors...

(Some authors) define stability in terms of the growth of rounding errors. However we have a slight preference for the (our) definition, (i.e. independent on the rounding errors) because it emphasizes that stability still has to be considered, even if the rounding errors are negligible, unless, of course, the initial data are chosen with diabolic care so as to be exactly free of those components that would be amplified if they were present.
A similar concept was expressed in the same years by Dahlquist.

"The most fundamental is the distinction between instability in the underlying mathematical problem and instability in an algorithm for the (exact or approximate) treatment of the problem".

We shall then avoid, when unnecessarily, to consider rounding errors.
Classes of Problems

- Asymptotically stable Problems (AS),
- Marginally Stable Problems,
- Unstable Problems,
- Boundary Value Problems,
Algorithms and Asymptotic Stability

More often than it can be thought, *numerical algorithms can be considered as discrete dynamical systems around critical points*. (equilibria).

The spaces where such dynamics are placed may vary considerably, ranging from $\mathbb{R}$, or $\mathbb{R}^N$ to the space of $N \times N$ real or complex matrices.
Example 1, trivial

\[ y_{n+1} = \alpha y_n, \]

where \( y \in \mathbb{R} \) or \( \mathbb{C} \)
\( \alpha \) is real or complex.

The equilibrium is at the origin and it is

- AS for \( |\alpha| < 1 \),
- stable for \( |\alpha| = 1 \)
- unstable for \( |\alpha| > 1 \).
Example 2, less trivial

\[ y_{n+1} = Ay_n, \]

where \( y \in \mathbb{R}^N \) and \( A = \alpha I_N - \beta K_N, \)

\[ K_N = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 1 & 0 \end{pmatrix}_{N \times N}. \]

The equilibrium is the origin in \( \mathbb{R}^N \) and it is AS for \( |\alpha| < 1 \), stable for \( |\alpha| = 1 \) and unstable for \( |\alpha| > 1 \).

Not uniformly with respect to the dimension!

In the case of PDE, this makes the difference.
Example 3, time scales

\[ y_{n+1} = \begin{pmatrix} 0.9 \\ 10^{-5} \end{pmatrix} y_n \]

The origin is A.S., but a generic solution approaches the equilibrium \textit{with two different modes, one slow and the other fast}. 
Example 4, splittings

Iterative methods for the solution of linear systems

\[ Ax = b \]

fit very well in the framework of AS dynamical systems.

The splitting techniques are nothing but strategies to transform the solution \( \bar{y} = A^{-1}b \) into an AS equilibrium point of an appropriate dynamical system.

What is usually less emphasized is that, even in the so called direct methods, the AS plays a central role, especially when dealing with structured matrices.
What’s wrong with marginal stability?

We have said that AS appears often in Numerical Analysis, although problems which are *naturally* marginally stable or even *unstable* appear as well. *Naturally* means that they are *unstable by themselves* and not because of previous wrong choices, for example a wrong discretization. In principle, small perturbations may lead them to be unstable. In the real world, however, there are very important marginally stable systems which are far to be unstable and only perturbations of improbable nature (!) would turn them from *stability to instability*. In other words, *many marginally stable problems need to be safely computed*. 
What’s wrong with unstable problems?

Unstable solution may be computed within the limit imposed by the computer arithmetic.

Computer arithmetic is able to represent relatively well numbers in a certain fixed range. When the numbers become large such representation become poorer and poorer until it may not have a single digit in common with the represented number.

The use of large numbers is then unsafe on the computers.
Examples of unstable problems

- The power method to find the maximum eigenvalue of a matrix is an unstable problem. In this case the required information is only partial, i.e. the grow rate of the norm of a vector, instead of the norm of the vector itself. The algorithm is designed in order to avoid large numbers.

- The Miller’s problem provides an example of unstable problems. It arose in the fifties in computing recursively the values of the Bessel functions. It is worth to mention it because of the peculiar aspect of the devised stable algorithm.
Miller’s problem

In the simplest form the problem is

\[ y_{n+2} - 102y_{n+1} + 200y_n = 0, \quad y_0 = \sqrt{3}, y_1 = 2y_0, n < N_1. \]

\[ y_n = c_1 2^n + c_2 100^n. \]

Approximate solution.
Let \( N > N_1 \), \( y_N = 0 \)

\[ y_n = y_0 2^n \left( \frac{2}{100} \right)^{N-n} - 1 \]

\[ \left( \frac{2}{100} \right)^N - 1 \]
Conditioning

The notion of stability is not enough. A more general concept is needed.
Let the discrete solution be $y_0, y_1, \ldots, y_N, y_i \in \mathbb{R}^s$.

We shall consider the following two parameters:

$$
k_d(N) = \frac{\max_{i=0}^N \|y_i\|}{\|y_0\|}, \quad \gamma_d(N) = \frac{1}{N\|y_0\|} \sum_{i=0}^N \|y_i\|. \quad (1)
$$

They may provide two different types of information, especially in the case of AS.
Consider for example $y_n = \alpha^n y_0, \quad n < N$. If we are interested in knowing the maximum value reached by the solution, we immediately obtain

$$\max ||y_n|| = k_d ||y_0||,$$

with $k_d < 1$ in case of asymptotic stability and $k_d$ growing exponentially with $N$ in the case of instability. We may be interested in knowing how fast the solution returns to the equilibrium in case of A.S. Such information is provided by $\gamma_d$, which, in this case becomes:

$$\gamma_d = \frac{1 - |\alpha|^{N+1}}{N (1 - |\alpha|)}.$$

$\gamma_d \in (0, 1)$ in case of AS.
Information measure

\( \gamma_d \) is a sort of *information measure* in the sense that its smallness informs us that we are using large values of \( N \), i.e. we are computing terms already too small (smaller, for example, than the machine precision).

In the case of AS, by posing \( R = -\log_{10}(\|\alpha\|) \), we have

\[
\gamma_d(N) \simeq \frac{1 - 10^{-R(N+1)}}{RN}
\]

which, if \( RN \) is large, becomes \( \gamma_d(N) \simeq (RN)^{-1} \).

The quantity \( R \) is already known in NA and is usually called *asymptotic rate of convergence*. It is useful because its inverse measures approximatively the number of iterations needed to approximate the equilibrium within a precision of \( 10^{-1} \).
\( \gamma_d \) and machine precision

If the machine precision is \( 10^{-m} \) then the number of iteration necessary to reach the machine precision is \( N^* \approx mR^{-1} \). One has then

\[
\gamma_d \approx \frac{N^*}{N} \frac{1}{m}
\]

It is useless to use \( N > N^* \). It follows that

\( \gamma_d \) should be of the order of \( m^{-1} \)

The ratio

\[
\sigma_d = \frac{k_d}{\gamma_d}
\]

will be called stiffness ratio. In the example \( \sigma_d \approx m \).
Stiffness

Both $k_d$ and $\gamma_d$ measure the sensitivity of the solution with respect to change of the initial condition, although such measure may change considerably. The problems will be classified as follows:

- well conditioned if $k_d$ and $\gamma_d$ are small and of the same order;
- stiff if $\sigma_d$ is large (in the example $> m$);
- ill conditioned if both parameters are large.

The definition of stiffness is *unusual in this context*. We will show later that it fits with the usual definition in the context of numerical methods for ODE.
Example of stiff problem

Figure 1: There is no need to use smaller value of $\gamma_d$. 

$\alpha = 0.1$

$\gamma_d = 1/5$

$\gamma_d = 1/40$
Entropy and information content

Let associate to \( y_1, y_2, \ldots, y_N \in (0, 1) \) the frequencies \( p_1, p_2, \ldots, p_N \). This can be obtained as usual by dividing the interval \((0, 1)\) in subintervals and counting the numbers of points in each of them. The entropy is given by

\[
H(p) = \sum p_i \log(1/p_i).
\]

\( H(p) \) measures the expected values of information of the distribution.

*It is maximum when the distribution is uniform*, i.e. when in each subinterval fall the same number of points, for example one.

The maximum value is attained at \( N^* + 1 \).
Optimal mesh

To the distribution on the $y$ axis, it will correspond a mesh on the $n$ axis.
Reduction of stiffness

Stiffness may be reduced by reducing $N$. Consider the multiscale example. The two modes are $0.9^n$ and $10^{-5n}$. After the fastest mode has become less than the machine precision (say for $n = \tilde{n}$), one may compute the slower mode not at each step but every $\mu$ steps, i.e. one may define a new variable $z_s = y^{\tilde{n} + s\mu}$. In the new variable the slower mode equation becomes

$$z_{s+1} = (0.9)^{\mu} z_s$$

It is then enough to choose $\mu$ such that $-\mu \log_{10} 0.9 \approx 5$ to have comparable stiffness ratio for the two modes.
Time scale change a stiff problem to a non stiff one.
Ill conditioned problem.

As example of ill conditioned problem consider the case of

\[ y_{n+1} = Ay_n, \]

where \( y \in \mathbb{R}^N \) and \( A = \alpha I_N - \beta K_N \).

The solution is

\[
y_n = \alpha^n \sum_{i=0}^{N-1} \binom{n}{i} \left( \frac{\alpha}{\beta} \right)^i K^i y_0.
\]

Even if \( |\alpha| < 1 \), the solution may become large for large \( N \). In this case both parameters \( k_d \) and \( \gamma_d \) are large.

This is not the case if \( \left| \frac{\alpha}{\beta} \right| < 1 \). Note that \( z_1 = \frac{\alpha}{\beta} \) is the root of
the polynomial associated to \( A \).
$k_d = 4.4 \times 10^4$, $\gamma_d = 6 \times 10^3$
More general problems

The same approach of using two conditioning parameters can be used for more general problems. In the case of boundary value problems, the parameters $k_d$ and $\gamma_d$ can be defined as well. In other words, the notion of conditioning continues to hold, as well as the notion of stiffness.

If $y_n \in \mathbb{R}^N$ is a solution of a discrete BVP depending on some boundary condition $\eta \in \mathbb{R}^N$ we can define

$$k_d = \frac{1}{\|\eta\|} \max_{i=1}^{N} \|y_i\|, \quad \gamma_d = \frac{1}{N\|\eta\|} \sum_{i=1}^{N} \|y_i\|, \quad \sigma_d = \frac{k_d}{\gamma_d},$$

where $\| \cdot \|$ is any norm in $\mathbb{R}^N$. 
Let $A$ be a real $N \times N$ matrix. Consider the following sequence,

$$
A_0 = A, \quad A_N = I_N
$$

$$
A_{j+1} = L_j A_j U_j;
$$

where

$$
L_j = \begin{pmatrix}
I_{j-1} & 1 \\
& -a_j^{-1} s_j & I_{N-j}
\end{pmatrix}, \quad U_j = \begin{pmatrix}
I_{j-1} & a_j & t_j^T \\
& & I_{N-j}
\end{pmatrix},
$$

Rodi – Stability and Conditioning. – p.35/73
\[ A_j = \begin{pmatrix} L_j^{-1} & 1 \\ Q_{N-j} \end{pmatrix}, \]

with \( s_j \) and \( t_j \) appropriately defined vectors. The final result is \( I_N \).

This example differs from the others for the following reasons:

1. the underlying space is the space of real \( N \times N \) matrices;

2. the motion is not autonomous, since the factor matrices \( L_j \) and \( U_j \) change at each step.

The problem here is in the possibility that the entries of product matrices \( L_j \) and \( U_j \) may become large and then not well represented in the computer arithmetic.
Grow factors

The dynamic is not uniquely defined in the sense that there is a certain freedom in the choice of both $L_j$ and $U_j$. One is able to maintain small the entries of $L_j$ by the pivoting strategy, but still the entries of $A_j$ and consequently $U_j$ may grow. Parameters which monitor such grow, called grow factors, have been defined by Wilkinson and more recently by Amodio and Mazzia. When such parameters grow exponentially with $N$ the problem is said unstable. The parameter $k_d$ now becomes:

$$k_d = \max_{j=1}^N \frac{\|A_j\|_{\infty}}{\|A\|_{\infty}},$$

which coincides with the grow factor used by Amodio and Mazzia.
Tridiagonal Toeplitz matrices

Let

\[ A = \alpha I + \beta K + \gamma K^T, \]

with \( \alpha, \beta, \gamma \) real.

\( A \) is a tridiagonal \( N \times N \) Toeplitz matrix.

Consider the problem \( Ay = b \equiv y_0 E_1 \), where \( E_1 \equiv (1, 0, \ldots, 0) \)
is the first unit vector in \( \mathbb{R}^N \). It is equivalent to solve the following discrete BVP, where \( y_i \) are the entries of the vector \( y \),

\[
\begin{align*}
\gamma y_{n+1} + \alpha y_n + \beta y_{n-1} &= 0, \quad n = 1, 2, \ldots, N \\
y_0 &= -1/\beta, \quad y_{N+1} = 0.
\end{align*}
\]
The solution of the above problem can be expressed by means of the roots of the polynomial

\[ p(z) = \gamma z^2 + \alpha z + \beta, \]

Let \( z_1, z_2 \) be such roots, the solution is \( y_n = c_1 z^n + c_2 z^n \). After imposing the boundary condition, we obtain,

\[ y_n = y_0 z_1^n \frac{\left( \frac{z_1}{z_2} \right)^{N-n}}{\left( \frac{z_1}{z_2} \right)^N} - 1 \]

1. in order to the solution being bounded, the two roots need to be distinct. Actually they should have distinct moduli in order to prevent the denominator becoming too small;

2. the solution is essentially generate by the root of minimal modulus \( z_1 \);

3. if \( |z_2| > 1 \) and \( |z_1| < 1 \) the solution is bounded with respect to \( N \).

Even when \( z_1 > 1 \), small perturbations of initial data, will cause perturbation growing as \( z_1^n \) and not as the faster mode \( z_2^n \).
Both $k_d$ and $\gamma_d$ behaves as $z_1^N$.

They are bounded with respect to $N$ when $|z_1| < 1$. Considering that the vector $b$ is the first unit vector in $\mathbb{R}^N$, this implies that \textit{the vector $y$ is the first column of $A^{-1}$}. By using sequentially the vectors $E_i$ and applying similar arguments, we arrive to the conclusion that $|z_1| < 1$, $|z_2| > 1$ is a \textit{necessary and sufficient condition} to have $\|A^{-1}\|$ bounded uniformly with respect to $N$.

Let $k_d^i$ and $\gamma_d^i$ be the corresponding conditioning parameters, we have

$$K(A) = \|A\| \max_{1 \leq i \leq N} k_d^i.$$
When each $k_d^i$ is bounded with respect to $N$, the matrix is said well conditioned ($|z_1| < 1, |z_2| > 1$).

When either $|z_2| = 1, |z_1| < 1$, or $|z_1| = 1, |z_2| > 1$, $|A^{-1}|$ may grow linearly with $N$. We say that the matrix $A$ is weakly well conditioned.

Let

$$\gamma(A) = \min \gamma_d^{(i)}$$

$$\sigma(A) = \frac{k(A)}{\gamma(A)}$$

What kind of information may provide this parameter?

Let us summarize with a specific example.
Stiffness and Toeplitz matrices

$$A_N = \begin{pmatrix}
\alpha & 1 & 0 \\
\alpha & 1 & \vdots \\
\vdots & \ddots & 0 \\
0 & \ldots & c(\alpha) & \alpha
\end{pmatrix}_{N \times N}$$
More general Toeplitz matrices

Let $T_N$ be a generic banded Toeplitz matrix.

The matrix $T$ is *well conditioned* if the roots of the polynomial associated to the matrix are such that $k_1$ of them are inside and $k_2$ are outside the unit disk.

Further generalization to define *well conditioning in a region* can be done as follows.
Conditioning in a region

Consider the matrix $A_N$ in the previous example, and let  
$\alpha = a - \lambda$, with $a$ real and $\lambda$ complex.
By posing $T = A + \lambda I$, the problem becomes

$$(T - \lambda I)y = b$$

Now the roots of the polynomial $p(z) - \lambda z = \gamma z^2 + (a - \lambda)z + \beta$
depend on the complex parameter $\lambda$.
In correspondence of the values of $\lambda$ such that $\left(\frac{z_2}{z_1}\right)^N = 1$ the
solution does not exist
Boundary locus

Suppose now that we ask to the system to be well conditioned for all value of $\lambda$ in a prefixed region, for example $\mathbb{C}^-$. This means that we ask that for all $\lambda \in \mathbb{C}^-$ the two roots need to be one outside and one inside the unit circle.

The best way to check if the two roots remain in the same position with respect the unit disk is to monitor if they cross the unit circle for $\lambda \in \mathbb{C}^-$. This leads to consider the map

$$\lambda(\theta) = \frac{p(e^{i\theta})}{e^{i\theta}}$$

and the curve (boundary locus)

$$\Gamma = \{\lambda(\theta), 0 \leq \theta \leq 2\pi\}.$$
The curve $\Gamma$ is the locus of value of $\lambda$ where one or both the roots $z_1(\lambda), z_2(\lambda)$ cross the unit circle.

If such curve completely lies in $\mathbb{C}^+$, then the matrix $T - \lambda I$ will be well conditioned in $\mathbb{C}^-$. 
BC with eigenvalues

Typical boundary locus
Generalization

Let $T_N = A_N - qB_N$

- $A_N$ and $B_N$ two banded Toeplitz matrices;
- $q \in \mathcal{D} \subset \mathbb{C}$;
- $k + 1$ the maximum value of the corresponding bandwidths.
- $\rho(z)$ and $\sigma(z)$ the two polynomial of degree $k$;
- $\pi(z, \lambda) = \rho(z) - q\sigma(z)$
- $\Gamma = \{ q(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, 0 \leq \theta < 2\pi \}$
- $k_1$ is the number of lower diagonals and $k_2$ the number of upper diagonals of $T_N$. 
Theorem. \( T_N \) is well conditioned in a region \( \mathcal{D} \) of the complex plane, for all \( q \in \mathcal{D} \), iff one of the two conditions hold true:

a) \( \Gamma \cap \mathcal{D} = \emptyset \)

b) \( k_1 \) of the roots of the polynomial \( \pi(z, \lambda) \) are inside the unit disk and \( k_2 \) are outside.

If \( \partial \mathcal{D} \cap \Gamma \) have common points and if \( \Gamma \) is a Jordan curve, then the matrix is weakly well conditioned.
The definitions of critical points, stability, asymptotic stability and instability apply to this case as well, only considering that now the variable $t$ is continuous.

Moreover, the parameters $k_c, \gamma_c, \sigma_c$ can be defined in a way very similar to the discrete case.
Stiffness of continuous problems

\[ y' = \lambda y, \quad \lambda < 0, \quad y(0) = y_0, \quad t \in (0, T). \]

\[ k_c = 1; \quad \gamma_c = \frac{1 - e^{\lambda T}}{\lambda T} \approx \frac{1}{|\lambda T|}; \]

\[ \sigma_c = |\lambda T| = \frac{T}{T^*} \]

In the non scalar case, \( \sigma_c \) is the ratio of largest and smallest eigenvalues.
**Definition.** A *continuous problem will be said well-represented by a discrete problem if*

\[ k_d \approx k_c; \quad \gamma_d \approx \gamma_c \]

Let \( h = (h_1, h_2, \ldots, h_N) \), be the mesh. Both \( k_d \) and \( \gamma_d \) depend on \( h \).

The request may lead to define the optimal mesh.

This has been discussed by Francesca Mazzia in her talk.
Not well represented problem

\[
\frac{d}{dt} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 100 & 99 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad 0 \leq t \leq 1,
\]

\[y_1(0) = 1, \quad y_1(1) = e^{-1}.
\]

Let \( \mu \) be the unknown second component,

\[
y_1(t, \mu) = \frac{\mu + 1}{101} e^{100t} + \frac{100 - \mu}{101} e^{-t},
\]

\[
y_2(t, \mu) = 100 \frac{\mu + 1}{101} e^{100t} - \frac{100 - \mu}{101} e^{-t},
\]
The solution of the problem is

\[ y(t) = e^{-t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \]

which corresponds to \( \mu^* = -1 \). The problem is very well conditioned as BVP \( (k < 2) \) but very ill conditioned with respect to \( \mu \), \((shooting\ method)\).

Let

\[ e(\mu) = \max_{0 \leq t \leq 1} \| y(t, \mu) - y(t, -1) \|_\infty \]
Test equations

In the process of constructing numerical methods the notions of stability and asymptotic stability have played a central role. In fact the numerical methods have been essentially modeled on the two test equations

\[ y' = 0; \]
\[ y' = \lambda y, \quad Re(\lambda) < 0. \]
Perron Theorem

In the first case the origin is a (marginally) stable equilibrium, while it is AS in the second case. There is a fundamental difference between the two cases. The second case may be considered as a model representative of more difficult (non linear) equations. A general theorem of dynamical systems (Perron) support such model.

**Theorem.** Let

\[ y' = \lambda(y - \bar{y}) + g(y), \]

where \( g(0) = 0 \) and \( \lim_{y \to \bar{y}} \frac{|g(y)|}{\|y - \bar{y}\|} = 0. \)

If the \( \bar{y} \) is A.S. for the linear part, then it is AS for the complete equation.
It is not by chance that a theorem essentially similar to the above one was established independently by Ostrowski in the context of iterative procedures to find zeros of nonlinear equations (see Ortega). The linear test equation is then more representative. No similar general results are available in the case of marginal stability. Test equation $y' = 0$ has, however, played an important role essentially in proving the convergence of the methods.
Let $y_0, y_1, \ldots, y_{k_1}$ be the initial data and $y_{N+k_1}, y_{N+k_1+1}, \ldots, y_{N+k_1+k_2}$ be the final data. A LMM applied to the test equation with the above boundary conditions leads to the following discrete problem,

$$(A_N - qB_N)y_N = b,$$

where $A_N$ and $B_N$ are Toeplitz matrices having $k_1$ lower non-zero diagonals and $k_2$ upper non-zero diagonals. The case $k_2 = 0$ corresponds to the classical choice of using discrete IVPs.
The non zero entries of the matrix $A_N$ are the coefficients of polynomial $\rho$, while those of $B_N$ are the coefficients of the polynomial $\sigma$. To the matrix $C_N(q) = A_N - qB_N$ we may apply either the generalized root condition or the boundary locus condition.

The representative polynomial is now $\pi(z, q) = \rho(z) - q\sigma(z)$. The matrix $C_N$ will be well conditioned if $k_1$ roots will be inside and $k_2$ will be outside the unit disk. The convergence is ensured if $\Gamma$ is a Jordan curve (this will prevent having double roots on the unit circle) and $0 \in \Gamma$. 
Example: The midpoint method

The midpoint rule may generate two different discrete methods, the classical one where the additional condition is placed at the origin, and the BVM method where the additional condition is placed at the end. In the first case the matrix $C_N(q)$ is

$$C^{(1)}_N(q) = \begin{pmatrix} 1 & 1 \\ -q & 1 \\ -1 & -q \\ 1 & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}_{N \times N}$$
while in the second case it is

\[ C_N^{(2)}(q) = \begin{pmatrix} -q & 1 \\ -1 & -q & 1 \\ -1 & -q & 1 \\ & \ddots & \ddots & \ddots \end{pmatrix}_{N \times N} \]

The representing polynomial \( \pi(z, q) = z^2 - 2qz - 1 \) is the same in the two cases but \( C_N^{(1)}(q) \) has two lower diagonals, while \( C_N^{(2)}(q) \) has only one. Except for the values of \( q \) on the segment \( I = (-i, i) \), the roots of \( \pi(z, q) \) are always one outside and one inside the unit disk.

**The matrix** \( C_N^{(2)} \) **is well conditioned for** \( q \in \mathbb{C} \setminus I \), **while** \( C_N^{(1)} \) **is not.**
Classes of BVMs

The result in the previous example is not isolated.

The **BVM approach** permits to obtain classes of well-conditioned numerical methods ($A$-stable methods) and also classes of perfect stable methods of any order.
Conservative Problems

Consider the non-linear pendulum.

\[
\frac{d}{dt} \begin{pmatrix} x \\ y \end{pmatrix} = J \begin{pmatrix} \sin(x) \\ y \end{pmatrix}
\]

The energy \( H = \frac{1}{2}y^2 + 1 - \cos(x) \) is a constant of the motion.

When the trapezoidal method is applied we get

\[
\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} + \frac{h}{2} J \begin{pmatrix} \sin(x_n) + \sin(x_{n+1}) \\ y_n + y_{n+1} \end{pmatrix}
\]
BREATHING EFFECT

\[ y_{n+1}^2 + x_{n+1} \sin(x_{n+1}) + R_{n,n+1} = y_n^2 + x_n \sin(x_n), \]

where

\[ R_{n,n+1} = x_{n+1} \sin x_n - x_n \sin x_{n+1} \]

will be called *volatile hamiltonian*.

\[ y_{n+1}^2 + x_{n+1} \sin(x_{n+1}) + \sum_{j=0}^{n} R_{j,j+1} = y_0^2 + x_0 \sin(x_0), \]
Further conditions

- Simplecticity
- Conservation
- Ergodicity
- .....
Definitions

\[ y_{n+1} = f(y_n), \quad y_0 \text{ fixed} \]

where \( y, f \in \mathbb{R}^N \). The critical solutions (equilibria) are the solutions of \( y = f(y) \).

Definition. The critical solution \( \bar{y} \) is stable if \( \forall \epsilon > 0, \exists \delta > 0, \) such that \( \forall y_0 \in B(\bar{y}, \delta), y_n \in B(\bar{y}, \epsilon), \) for \( n > 0 \).

Definition. The critical solution \( \bar{y} \) is asymptotically stable (AS) if it is stable and, moreover, \( \lim_{n \to \infty} = \bar{y} \).

If \( \delta \) can be chosen independent on \( \epsilon \), the stability or the AS stability is global.

Definition. The critical solution \( \bar{y} \) is unstable if it is not stable.