The Nyquist Stability Test

1.0 Introduction

With negative feedback, the closed-loop transfer function $A(s)$ approaches the reciprocal of the feedback gain, $f$, as the magnitude of $a(s)f(s)$ increases. This behavior is particularly advantageous if $f$ is more linear and drifts less than the forward gain $a(s)$, as is often (but not always) the case. A satisfactory reduction in sensitivity to $a(s)$ (the only fundamental benefit of negative feedback, as argued earlier) thus depends on a sufficiently large loop transmission magnitude. However, the onset of instability usually bounds the amount of gain permitted in practice, and therefore bounds the performance of feedback systems.

In introducing stability criteria, we noted that a system $H(s)$ is stable in the BIBO sense if all of the poles of $H(s)$ are in the open left half plane. Now, the poles of $H(s)$ are the roots of $P(s) = 1 + a(s)f(s)$ or, equivalently, the values of $s$ for which $a(s)f(s) = -1$ (purely real, negative, and of unit magnitude). Thus, if some value $s$ is a pole of the closed-loop system, the magnitude of $a(s)f(s)$ has to be unity there, and the corresponding phase angle must be an odd multiple of 180 degrees. Gain margin considers the special case where $a(s)f(s)$ equals $a(j\omega)f(j\omega)$ and is purely real and negative (i.e., where the phase angle is an odd multiple of 180 degrees), and evaluates the magnitude of $a(j\omega)f(j\omega)$ there. If the magnitude is precisely unity where the phase shift is 180 degrees, we must conclude that there is at least one pole exactly on the $j\omega$-axis. Why? Certainly, we have found a pole (because we have simultaneously satisfied the necessary and sufficient magnitude and phase conditions). Furthermore, we are employing steady-state sinusoidal excitation ($s = j\omega$), meaning that the pole discovered must have a zero real part. We then make the assumption (and it is only that) that if the magnitude is less than unity, the real part of the pole(s) is negative, so that the system is stable. The factor by which we must increase the loop gain to obtain the unity magnitude condition is what we define as the gain margin. Conversely, phase margin identifies where the magnitude of $a(j\omega)f(j\omega)$ is unity, and considers the phase lag there, relative to 180 degrees. If the phase lag is precisely 180 degrees where the magnitude is unity, we continue to conclude that there is at least one pole on the $j\omega$-axis. If the phase lag is not as large as that, we assume that the real part of the pole(s) is again negative, implying stability. In contrast with the binary nature of the BIBO stability criterion, gain and phase margin permit us to define a continuum of relative degrees of stability. Furthermore, understanding the implications of gain and phase margin suggest important compensation methods (e.g., lead or lag compensation) to move stability in a desired direction. It’s difficult to see how the BIBO criterion could inspire such insights.

Note that gain and phase margin evaluate stability by examining the behavior of the loop transmission. The vast simplification that results cannot be overemphasized. Determination of the loop transmission is usually straightforward, while that of the closed-loop transfer function requires identification of the forward path (not always trivial, contrary to one’s initial impression) and an additional mathematical step (i.e., taking the ratio of $a(s)$ to $1 + a(s)f(s)$); hence, any method that can determine stability from the loop transmission alone saves a significant amount of labor.
2.0 The Nyquist Stability Criterion

We now introduce the stability test from which come gain and phase margin: the Nyquist test. As we shall see, the Nyquist stability test is not limited to situations where the transfer functions are rational. Thus it can be used to evaluate the stability of distributed systems (e.g., those involving transmission lines) or systems that contain pure delays ($e^{sT}$ sorts of things). Second, the method can be applied directly to measured frequency response data, obviating the necessity for a mathematical model of the system under study. This latter consideration is in contrast to the explicit representation in the form of a rational transfer function that a more direct approach (such as root-finding) generally requires.

We’ll develop the stability criterion in two steps. Cauchy’s Residue Theorem will be used to prove a theorem known as the Principle of the Argument (“argument” being the mathematician’s other word for “phase angle”). Then the Principle will be applied to the specific problem of feedback systems to develop the stability criterion.

Understanding the derivation that follows assumes familiarity with complex analysis. However, successful application of the stability test does not require any understanding of the proof, and those without the relevant background (or interest) may skip the intervening details if desired and proceed directly to the statement of the stability criterion. This proof is offered simply as a service to those who would like to know in deeper detail where these stability criteria come from.

Okay, ready or not, here’s Cauchy’s Residue Theorem:

If a function $h(s)$ is analytic on and within a simply connected region bounded by a closed contour $C$, except at a finite number of poles within $C$, then

$$\frac{1}{2\pi i} \oint_C h(s) ds = \sum \text{residues of } h(s) \text{ in the poles of } h(s) \text{ enclosed by } C .$$  \hspace{1cm} (EQ 1)

Cauchy’s theorem thus tells us that there is a relationship between the value of a contour integral, and the poles that reside within the contour. If, for example, we pick such a contour to enclose the right half plane, carrying out the integration can reveal the presence of one or more poles within that contour, allowing us to deduce something about the system’s stability. This fundamental idea may be clear, but the details might not be.

You may recall (or not) that if $s_k$ is a pole of $h(s)$, then the residue of $h(s)$ in this pole is given by some messy looking thing:

$$\text{Res} s_k = \frac{1}{(m-1)!} \frac{d^{m-1}}{ds^{m-1}} \{ (s - s_k)^m h(s) \} \bigg|_{s = s_k} ,$$  \hspace{1cm} (EQ 2)

where $m$ is the multiplicity of the pole at $s_k$ (i.e., the number of poles of the particular value $s_k$).
Now let us express \( h(s) \) indirectly in terms of some other function \( g(s) \) (you’ll just have to trust me that the reason for this seemingly random maneuver will become clear eventually):

\[
h(s) = \frac{g'(s)}{g(s)} = \frac{d}{ds} \left[ \log g(s) \right], \tag{EQ 3}
\]

where

\[
g(s) = \frac{\prod_{i=1}^{\alpha} (s - z_i)^{m_i}}{\prod_{i=1}^{\beta} (s - p_i)^{n_i}}. \tag{EQ 4}
\]

Here \( m_i \) is the multiplicity of the \( i \)th zero, \( n_i \) is the multiplicity of the \( i \)th pole, \( \alpha \) is the number of distinct zeros, and \( \beta \) is the number of distinct poles.

With \( g(s) \) defined as above, we may take the log of both sides of the equation:

\[
\log g(s) = \sum_{i=1}^{\alpha} m_i \cdot \log(s - z_i) - \sum_{i=1}^{\beta} n_i \cdot \log(s - p_i), \tag{EQ 5}
\]

so that

\[
h(s) = \frac{d}{ds} \left[ \log g(s) \right] = \sum_{i=1}^{\alpha} \frac{m_i}{s - z_i} - \sum_{i=1}^{\beta} \frac{n_i}{s - p_i}. \tag{EQ 6}
\]

Let’s examine this last equation. It’s clear that the poles of \( h(s) \) are all the poles and zeros of \( g(s) \) (just look at the equation and see what makes it blow up). Furthermore, for a zero of \( g(s) \) (i.e., any \( z_i \)), we see that the residue (i.e., the numerator coefficient in the sums on the right-hand equation) of \( h(s) \) in \( z_i \) is \( m_i \). Similarly, for a pole of \( g(s) \) (i.e., any \( p_i \)), the residue of \( h(s) \) in \( p_i \) is \(-n_i\). Therefore we can evaluate the contour integral of \( h(s) \) around any closed path that contains the poles and zeros of \( g(s) \) to get

\[
\frac{1}{2\pi j} \oint_C h(s) ds = \sum_{i=1}^{\alpha} m_i - \sum_{i=1}^{\beta} n_i = Z - P, \tag{EQ 7}
\]

where \( Z \) is the total number of zeros of \( g(s) \) and \( P \) is the total number of poles of \( g(s) \).

As unlikely as it may seem, we’re almost there. To complete the derivation, we will evaluate the contour integral another way. Consider

\[
\frac{1}{2\pi j} \oint_C h(s) ds = \frac{1}{2\pi j} \oint_C \frac{d}{ds} \left[ \log g(s) \right] ds = \frac{1}{2\pi j} \oint_C d[\log g(s)]. \tag{EQ 8}
\]
Now \( \log[g(s)] = \log |g(s)| + j\angle[g(s)] \), where \( \angle[g(s)] \) is the phase angle of \( g(s) \). Note also that, since zeros and poles of \( g(s) \) are poles of \( h(s) \), we could not have applied the residue theorem to \( h(s) \) if any zeros or poles of \( g(s) \) were on \( C \). Therefore, we can proceed with the assurance that \( \log |g(s)| \) in fact exists everywhere on \( C \) and we may therefore evaluate the contour integral as follows:

\[
\frac{1}{2\pi j} \oint_C d[g(s)] = \frac{1}{2\pi j} \{ \log |g(s)| + j \cdot \angle[g(s)] \} \bigg|_{s_1}^{s_2} .
\]  

(EQ 9)

Now, since \( C \) is a closed contour, \( \log |g(s_1)| = \log |g(s_2)| \). So, the integral reduces to:

\[
\frac{1}{2\pi j} \oint_C d[g(s)] = \frac{1}{2\pi} \{ \angle[g(s_2)] - \angle[g(s_1)] \} .
\]  

(EQ 10)

Because \( C \) is a closed contour, the net change in angle will be a multiple of \( 2\pi \), so we can write

\[
\frac{1}{2\pi j} \oint_C h(s) ds = \frac{1}{2\pi j} \oint_C d[g(s)] = N ,
\]  

(EQ 11)

where \( N \) is the number of encirclements of the origin by \( g(s) \) as \( s \) traverses \( C \) once.

We may now state the Principle of the Argument in a form of more direct use to us: Given a function \( g(s) \) and a closed contour \( C \) such that \( g(s) \) is analytic on and within \( C \) (except, possibly, at a finite number of poles of \( g(s) \) within \( C \)) and which does not vanish on \( C \), then

\[
N = Z - P ,
\]  

(EQ 12)

where \( N \) is the number of positive encirclements of the origin by \( g(s) \) as \( s \) traverses \( C \) in the positive direction, \( Z \) is the number of zeros of \( g(s) \) within \( C \), and \( P \) is the number of poles of \( g(s) \) within \( C \).

It is customary to measure positive angles in the counterclockwise direction. However, it’s not necessary to be too fussy about this; all that is really required is that the positive direction for \( g(s) \) be the same as the positive direction for the traversal of \( C \) by \( s \). As long as you remain consistent in your definitions throughout the application of the test, it doesn’t matter which convention you select.

From the Principle of the Argument, it’s only a short hop to the Nyquist stability criterion. Since we’re interested in whether there are any right half-plane roots of the function \( 1 + a(s)f(s) \), set \( g(s) = 1 + a(s)f(s) \) and choose the contour \( C \) to be the semicircular contour (sometimes called a “D” contour because of its shape) that encloses the right half \( s \)-plane shown below:
From the Principle of the Argument, \( Z = N + P \), where

- \( Z \) is the number of zeros (roots) of \( 1 + a(s)f(s) \) in the right half plane;
- \( P \) is the number of poles of \( 1 + a(s)f(s) \) in the right half plane; and
- \( N \) is the number of counterclockwise encirclements of the origin by \( 1 + a(s)f(s) \) as \( s \) traverses \( C \).

To apply this principle to our stability problem, note that the poles of \( g(s) \) are the poles of \( a(s)f(s) \). Also, instead of plotting \( g(s) = 1 + a(s)f(s) \) and determining encirclements of the origin, one may instead simply plot \( a(s)f(s) \) and determine encirclements of \(-1 + j0\). So we may state the Nyquist stability criterion as follows:

**Let \( a(s)f(s) \) have only poles and zeros in the s-plane. The number of roots, \( Z \), of \( 1 + a(s)f(s) \) which lie in the right half plane is given by \( Z = N + P \)** where

- \( P \) is the number of poles of \( a(s)f(s) \) in the right half plane; and
- \( N \) is the number of CCW encirclements of \(-1 + j0\) by \( a(s)f(s) \) as \( s \) traverses the contour \( C \) which encloses the right half plane.

For stability, one must have \( Z = 0 \).

### 2.1 Special considerations

Quite frequently, the function \( g(s) \) will have poles on the imaginary axis. Since, strictly speaking, these are not allowed, modifications must be made to the contour \( C \) as shown in the following figure:
Each detour shown is a semicircle of radius \( \rho \). The behavior of \( g(s) \) is examined as this radius goes to zero. For a pole at the origin, for example, let \( s = \rho e^{j\theta} \). As \( C \) is traversed in the vicinity of the origin, the change in the angle of \( s \) is \(-\pi\). If the pole is of multiplicity \( m \) then, near the origin,

\[
g(s) \propto \frac{1}{s^m}, \tag{EQ 13}
\]

and the change in angle of \( g(s) \) is therefore just \(-m\pi\).

Similarly, one may find what happens to the angle of \( g(s) \) in the vicinity of a pole at \( s = j\omega_k \) by letting \( s = j\omega_k + \rho e^{j\theta} \).

Another consideration is the question of essential singularities such as \( e^{-sT} \) (which, you may recall, is the transfer function of a pure time delay of \( T \) seconds). This is a singularity at infinity and since the basic contour \( C \) is the limit of a finite semicircle as its radius tends to infinity, the essential singularity is never actually enclosed by \( C \). So the condition that \( C \) enclose only poles and zeros is not violated.

Finally, branch points such as \( 1/\sqrt{s} \) (which often arise in distributed systems, for example) require a detour (just as in the case of poles on the imaginary axis) so that they are not enclosed by \( C \).

### 2.2 How to Apply the Nyquist Stability Test

In most applications, \( g(s) \) is usually known up to some multiplicative gain factor \( k \), and we are generally interested in the range of \( k \) values for which \( 1 + g(s) \) has no roots in the open right half plane. To avoid having to re-plot \( g(s) \) for each value of \( k \), we can normalize the function \( 1 + g(s) \) by dividing through by \( k \). Note that the roots are unchanged by this
maneuver. Now, instead of looking at encirclements of \(-1 + j0\), we look for encirclements of \(-1/k + j0\). So, we can state the following recipe for applying the test:

1) Normalize \(a(s)f(s)\) by dividing by \(k\).
2) Let \(s\) vary along \(C\), and plot the resulting map of \(a(s)f(s)/k\) in the Nyquist plane.
3) Count the number \(N\) of positive encirclements of \(-1/k\) in the Nyquist plane.
4) The number \(Z\) of right half plane roots of \(1 + a(s)f(s)\) is \(N + P\), where \(P\) is the number of right half plane poles of \(a(s)f(s)\). For stability, of course, we require that \(Z = 0\).

An example:

Consider the system whose open-loop transfer function is given by

\[
g(s) = \frac{k}{s(s + a)}
\]  
(EQ 14)

so that

\[
g(s) k = \frac{1}{s(s + a)}
\]  
(EQ 15)

The \(s\)-plane contour and the mapping onto the Nyquist plane are:

So, how does all this mapping business work? It’s much less complicated than it initially appears. Here’s the recipe:

1) Pick a value of \(s\) that is on \(C\).
2) Plug this value into the expression for \(g(s)\). The result will be some complex number.
3) Plot this complex number in the Nyquist plane (think polar, i.e., magnitude and phase).
4) Repeat for all values of $s$ on $C$. The resulting curve is what we mean by a map of $g(s)$ onto the Nyquist plane.
5) Count positive encirclements ($N$) and determine if $Z$ is zero.

With this recipe in mind, let us investigate the relationship between the two plots. The positive imaginary axis of the $s$-plane maps into the curve marked $\omega > 0$ in the Nyquist plane, while the negative imaginary axis of the $s$-plane maps into the curve marked $\omega < 0$ in the Nyquist plane. The infinite semicircle $C_2$ in the $s$-plane maps into the origin of the Nyquist plane (since, for large $s$, $g(s)$ has a small magnitude), and the infinitely small semicircle $C_1$ in the $s$-plane maps into the infinite semicircle in the Nyquist plane (since, for a small value of $s$, $g(s)$ has a large magnitude).

You may have noticed that the resulting map is symmetric about the real axis (except for the direction of the arrows). This property is in fact a general one, so that you can cut your labor in half by plotting the map as $s$ traverses just the positive or negative half of the contour $C$.

Now we have the map, so what happens next? Well, to evaluate the stability of this system via the Nyquist test, note that there are two possible general locations of the $-1/k + j0$ point (marked $A$ and $B$ in the diagram). Then we can readily generate the following table:

**Table 1:**

<table>
<thead>
<tr>
<th>Region</th>
<th>Range of $k$</th>
<th>$P$</th>
<th>$N$</th>
<th>$Z$</th>
<th>Stable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$0 &lt; k &lt; \infty$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>$B$</td>
<td>$-\infty &lt; k &lt; 0$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>no</td>
</tr>
</tbody>
</table>

One last point needs to be discussed: What is the inside of the contours involved? This question is by no means frivolous since we have to count encirclements. Luckily, it’s easy to decide this issue: As we traverse $C$ in the $s$-plane and look to the left, we see the inside of the region enclosed by $C$. Correspondingly, as we traverse the map in the Nyquist plane, the interior of the map is located similarly by looking to the left. Those of you with a background in complex analysis may recognize this as a consequence of the conformal mapping property of functions of a complex variable, namely, that angles are preserved. The rest of you are making that funny grimacing face. Cut it out.

### 2.3 Counting Encirclements -- Examples

To reinforce the foregoing developments, here are a few random Nyquist contours, with the number of CCW encirclements indicated in the various regions of the map:
Note the tremendous care and artistic ability evident in the contours above.

Seriously, though, the precise shape of the contour is not as important as simply preserving the number of encirclements. In fact, it is frequently impractical to draw a contour exactly because of the large range of magnitudes typically spanned by $a(s)f(s)$. Fortunately, gross distortions are allowed as long as they don’t alter fundamentally the topology of the contour.

### 3.0 Relationship to Gain and Phase Margin Stability Measures

As we have seen, the Nyquist test for stability involves counting the number of positive encirclements of the point $-1 + j0$ by the map onto the Nyquist plane of the function $a(s)f(s)$ as $s$ varies along a particular contour that comprises the entire imaginary axis and a semicircle of asymptotically infinite radius in the right half $s$-plane. For stability, we require that there be no positive encirclements of $-1 + j0$. As stated, then, the Nyquist test provides a binary answer to the stability question. However, we could imagine talking about degrees of stability, perhaps measured by the proximity of the Nyquist contour to the $-1 + j0$ point.

Refer to the following figure to develop this idea further:
Conventional measures include gain and phase margin, as shown. We know that gain margin is the factor by which the loop transmission magnitude could increase before a tangency with the \(-1 + j0\) point would occur. Similarly, the phase margin is the amount of negative phase shift the loop transmission could tolerate before such a tangency would occur. Just as a reminder, recall that these quantities are easily computed as follows:

1) To calculate the gain margin, find the frequency at which the phase shift of \(a(j\omega)f(j\omega)\) is \(-180^\circ\). Call this frequency \(\omega_\pi\). Then the gain margin is simply:

\[
gain \ margin = \frac{1}{|a(j\omega_\pi)f(j\omega_\pi)|} \quad (\text{EQ 16})
\]

2) To calculate the phase margin, find the frequency at which the magnitude of \(a(j\omega)f(j\omega)\) is unity. Call this frequency \(\omega_c\) (known as the crossover frequency). Then the phase margin (in degrees) is simply:
Because of the ease with which these two measures of relative stability are calculated (or obtained from actual frequency response measurements), they are often used in lieu of performing an actual Nyquist test. In fact, most engineers dispense with a calculation of gain margin altogether and compute only the phase margin. However, note that gain and phase margin are computed using the behavior of \( a(s)/f(s) \) at only two frequencies, while the Nyquist test determines stability by taking into account the behavior of \( a(s)/f(s) \) at all frequencies. Hence, it should not be surprising that occasionally situations arise (e.g., say if \( a(s)/f(s) \) oscillates about -1 so that there are multiple crossover frequencies) that cannot be adequately examined using only the gain and phase margin ideas. In such cases, it is self-evidently inappropriate to use gain and phase margin concepts, and one must perform an actual Nyquist test. The key idea, then, is that the gain and phase margin tests are only a subset of the more general Nyquist test. This important point is frequently overlooked by practicing engineers, who are often unaware of the limited nature of gain and phase margin as stability measures. This confusion persists because the stability of the commonest systems happens to be well determined by gain and phase margin, and this success encourages many to make inappropriate generalizations.

Okay, now that we’ve derived a new set of stability criteria that enable us to quantify degrees of relative stability, what values are acceptable in design? Unfortunately, there are no universally correct answers, but we can offer a few guidelines: One must choose a gain margin large enough to accommodate all anticipated variations in the magnitude of the loop transmission without endangering stability. The more variation in \( a_0f_0 \) that one anticipates, the greater the required gain margin. In most cases, a minimum gain margin of about 3-5 is satisfactory.

Similarly, one must choose a phase margin large enough to accommodate all anticipated variations in the phase shift of the loop transmission. Typically, a minimum phase margin of 30°-60° is acceptable, with the lower end of the range associated with substantial overshoot and ringing in the step response, and significant peaking in the frequency response. Note that this range is quite approximate and will vary according to application. For example, overshoot might be tolerable in an amplifier, but unacceptable in the landing controls of an aircraft.

As a parting note, it is worthwhile to point out that gain and phase margin are easily read off of Bode diagrams (in fact, it is precisely this ease that encourages many designers to use gain and phase margin as stability measures). As a service to you, the ever-eager consumer, here’s a picture that shows how to find gain and phase margin using the recipe already given:

\[
\text{phasemargin} = 180^\circ + \angle a(j\omega_c)f(j\omega_c) .
\]
4.0 Summary

We have presented several techniques for determining the stability of closed-loop systems through examination of loop transmission behavior. Gain and phase margin are most commonly used in practice, but are actually a subset of the more general Nyquist test. Gain and phase margin and the Nyquist test may operate on measured frequency response data, while root locus techniques require a rational transfer function for $a(s)f(s)$. All of these tests are relatively simple to apply, and allow one to evaluate rapidly the stability as well as the efficacy of proposed compensation techniques, all without having to determine directly the actual roots of $P(s)$. 