Control Theory

Dhruva Hegde

1 Basics of control systems

What is a control system?
A control system is a meaningful interconnection of devices that produces desirable output with respect to a given input, which has the ability to modify its behaviour with respect to requirements.

Measures of a control system

1. Speed : Transient analysis
2. Accuracy : Steady state analysis
3. Stability : Various time domain and frequency domain techniques

Open Loop Control System
Control action does not depend on actual output. It is basically just a direct function mapping value of input to output.

While designing an open loop system is easier and cheaper, it is not always reliable and is more prone to disturbances.

Closed Loop Control System
Control action depends on the actual output. This is achieved using the concept of feedback.

A closed loop system is more difficult to design and can be expensive, but it is more robust.

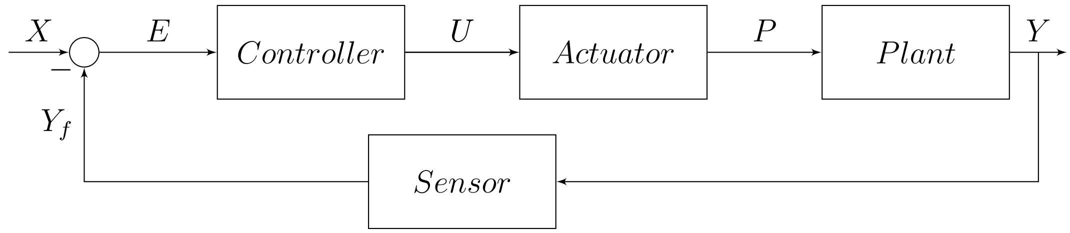
Feedback is the process of measuring the actual output of the system and feeding it back to the system for better performance.

For Control Systems, **Negative Feedback** is used. In negative feedback, the actual output signal is subtracted from the input signal to obtain the difference (or error). This will give information about how far the actual output is when compared to the desired output, using which the control system can adjust its behaviour.

It is worth noting that **Positive Feedback** is where the output is added to the input to produce a sum term, which is not very useful in control theory as it causes instability. More will be discussed about this later.

1.1 Control Systems terminology

Plant: The system that is to be controlled.
Actuator: The device that drives the plant to provide required output.
Controller: The device that uses input and feedback signal to control the actuator in such a way that desired output is produced.
Sensor: The device that measures the actual output and feeds it back.



The input signal, X is just the desired output value.
The difference between input signal and actual output measured by the sensor (also called feedback signal) is called error signal, E.
The controller provides the control signal, U in order to control the actuator.
The actuator provides the actuating/power signal, P to the plant to give actual output Y.

For simplicity, all the 3 blocks in the forward path are together represented as G and the feedback block is represented as H (unless specified otherwise).

1.2 Impulse Response and Transfer Function

Impulse Response of a system is defined as the output of the system when unit impulse function is provided as input.

The impulse response model of a control system is defined under relaxed conditions i.e under the assumption that the initial states of the system are 0 (meaning the system has no initial condition).
This model will be able to provide the output of the system for any given input. Hence, the impulse response model is basically an input-output relation.

If $r(t)$ is the given input such that $r(t) = 0 \; \forall \; t < 0$ and the impulse response of the system is $g(t)$, then the output $y(t)$ is given by,

$$y(t) = \int_{\tau=0}^{t} r(\tau)h(t-\tau)d\tau = r(t) * g(t)$$

The Laplace transform of the impulse response of a system is called its Transfer Function.
By taking the Laplace transform on the above equation,

$$Y(s) = R(s)G(s)$$

$$\implies G(s) = \frac{Y(s)}{R(s)}$$

Transfer Function of a system is defined as the ratio of the Laplace Transform of the output to the Laplace Transform of the input, computed under zero initial conditions.

From the definition, it can be noted that for transfer function to exist, the system must be a **Linear and Time-Invariant (LTI) System** and the input signal must be a **Causal Signal**.

The differential equation model is obtained by using basic laws such as Newton's Laws and Kirchhoff's Laws to obtain equations that describe how the components interact with each other when connected in specific ways. This model can be used to find the transfer function as well.

If a system input-output relation is defined by the differential equation,

$$\frac{d^2y(t)}{dt^2} + a_1\frac{dy(t)}{dt} + a_2y(t) = b_or(t)$$

then by taken Laplace transform on both sides and manipulating,

$$\frac{Y(s)}{R(s)} = \frac{b_o}{s^2 + a_1s + a_2}$$

which is its Transfer function.

Note that the impulse response or the differential equation represent the behaviour of the system in time domain whereas the transfer function represents the behaviour of the system in frequency domain.

The transfer function is expressed as ratio of 2 polynomials. $G(s) = \frac{n(s)}{d(s)}$

Poles of Transfer Function:
The poles are the values of s at which the transfer function tends to infinity. This corresponds to system being able to give an output when input is zero.
$\implies$ *poles are solutions to the equation* : $d(s) = 0$

Zeros of Transfer Function:
The zeros are the values of s at which the transfer function becomes zero. This corresponds to system not giving any output despite non-zero input.
$\implies$ *zeros are solutions to the equation* : $n(s) = 0$

Open Loop Transfer Function:
OLTF is the transfer function of the system prior to providing feedback.

Closed Loop Transfer Function:
CLTF is the transfer function obtained after simplifying the system with feedback.

Characteristic Equation:
The denominator polynomial of the CLTF equated to 0 gives the characteristic equation of the system.
The denominator polynomial that determines the behaviour of the system,

while the numerator polynomial only determines amplitudes.

Type of a system:
The number of poles at the origin in the OLTF of a system is called its type.

Order of a system:
The highest degree of 's' in the characteristic equation (or denominator polynomial of CLTF) is called its order.

Examples:

$$G(s) = \frac{s+1}{s(s+5)}$$

$\implies$ Type 1, Order 2

$$G(s) = \frac{2}{(s^3 - s + 6)}$$

$\implies$ Type 0, Order 3

A transfer function can be decomposed such that it can be expressed in Pole-Zero form as follows.

$$G(s) = \frac{(s+z_o)(s+z_1)....(s+z_n)}{(s+p_o)(s+p_1)....(s+p_m)}$$

If the highest power of the denominator polynomial is greater than or equal to the highest power of the numerator polynomial, then the transfer function is "Proper", otherwise it is "Improper".

Improper transfer functions are not physically realizable because a pure differentiator can't be physically realized.

If the highest power of the denominator polynomial is greater than the highest power of the numerator polynomial, then the transfer function is "Strictly Proper".

The value of the transfer function at $s = 0$ is called the DC gain of the system.

Loading Effect on Transfer Function:

Note that while deriving transfer function, it is assumed that there is no loading effect i.e no power is drawn at the output of the system.
This assumption must be satisfied even while deriving transfer functions for each component in a control system. If one component is acting as a load on another component, then:

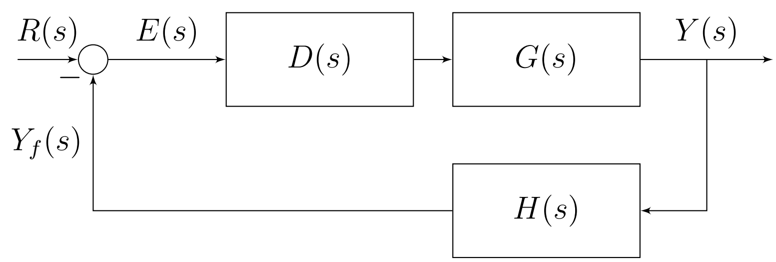
- Transfer function of each component cannot be determined separately.
- Transfer function of both components combined should be determined.
- Both components are put in the same block in the representation.

This problem is commonly encountered in cascaded electrical circuits where impedance matching has not been achieved. The overall transfer function of the cascaded circuit will not be equal to the product of the individual transfer functions. This is why impedance matching is extremely important in circuit design.

2 Principles of Feedback

As mentioned earlier, feedback is the process of measuring the actual output of the plant using a sensor and comparing it with the input (desired output) and controlling the plant accordingly.

A standard negative feedback system with open loop transfer function $G(s)$ and feedback element $H(s)$ with input $R(s)$ and output $Y(s)$ is represented as,
(each of the blocks and signals are expressed in terms of Laplace variable)



It can be algebraically derived that the overall transfer function of the system will be

$$T(s) = \frac{Y(s)}{R(s)} = \frac{D(s)G(s)}{1 + D(s)G(s)H(s)}$$

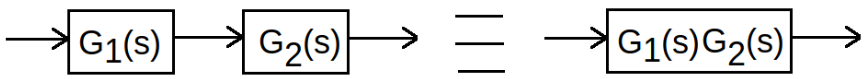
This will be the closed loop transfer function of the system.

Usually, $H(s)$ is taken to be unity and D(s)G(s) together is considered to be G(s). Then the transfer function will be reduced to, $T(s) = \frac{G(s)}{1+G(s)}$.
In this case, if the OLTF is expressed as $G(s) = \frac{n(s)}{d(s)}$, then the CLTF will be $G(s) = \frac{n(s)}{d(s)+n(s)}$.

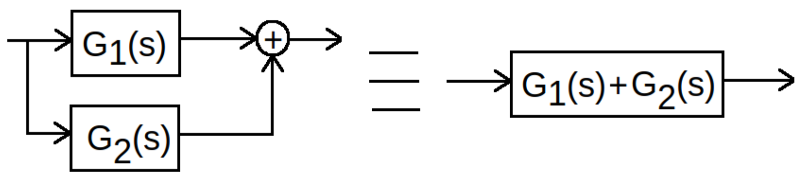
2.1 Block Diagram algebra

The most straightforward way to represent a control system is by using functional block diagrams.
Each block represents the transfer function of the particular subsystem.
The overall transfer function of the system can be found by using block diagram reduction techniques.

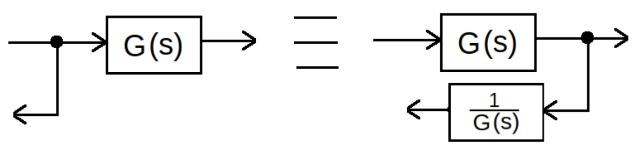
- Blocks in cascade



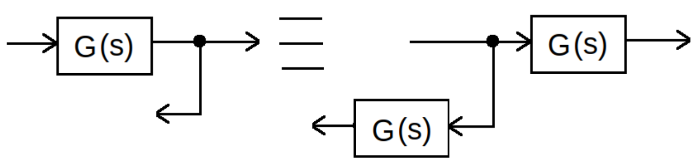
- Blocks in parallel



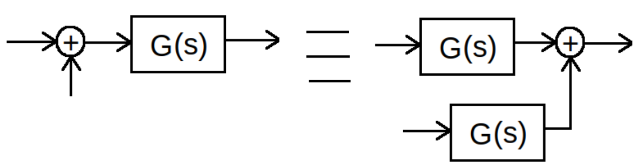
- Moving take-off point present before a block



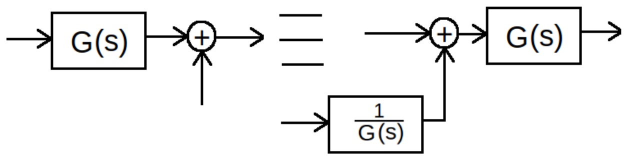
- Moving take-off point present after block



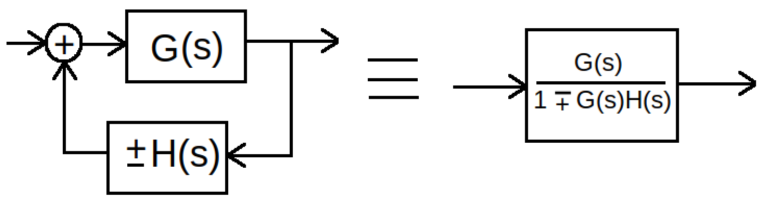
- Moving summing point present after block



- Moving summing point present before block



- Feedback loop



Also note that neighbouring take-off points can be interchanged and neighbouring summing points can be interchanged.

The block diagram technique is intuitive and can be used to solve simple systems. However, if the system becomes increasingly complex (i.e more blocks and connections), then the process of simplification using block diagram reduction becomes cumbersome.

2.2 Signal Flow Graph

The Signal Flow Graph is a graphical representation of a system, consisting of nodes and branches.
It provides a simpler way to obtain the overall transfer function of a system when the system consists of more elements.

Nodes represent system variables.
Branches are interconnections between nodes.

Forward Path is any path (of branches) that connects from the input node to the output note (there can be one or more forward paths in a system).

Note that any node can be considered an output node. It depends on the requirements of the analysis.

However, input nodes are only those nodes which have no incoming branches. Meaning, the nodes that have outgoing branches only can be input nodes.

Loop is a path which originates and terminates at the same node without repetition of any node in the middle.
Non-touching loops are a group of loops that have no nodes in common.

2.2.1 Mason's Gain formula

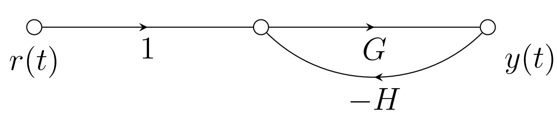
For a signal flow graph with N forward paths, the overall transfer function is given by,

$$\boxed{T(s) = \frac{\sum_{k=1}^{N} P_k \Delta_k}{\Delta}}$$

P_k is the k^{th} forward path gain
Δ_k is the value of Δ obtained by removing all loops that are touching the k^{th} forward path
Δ = 1 - [sum of all individual loop gains] + [sum of gain products of 2 non-touching loops] - [sum of gain products of 3 non-touching loops] + ...

- Δ is constant for a given SFG. This means, Δ is independent of the selection of the output node.
- Self loops immediately next to input node are useless and can be removed.
- If there are multiple inputs and the transfer function with respect to any one has to be found, then perform the analysis by making all the other inputs zero.
- If the SFG does not consist of any input node and A/B has to be found, then create a dummy input node I, find A/I and B/I and divide them.

The typical feedback structure, when converted to SFG is shown.



2.3 Sensitivity analysis

Sensitivity is defined as the change observed in a certain quantity due to variation (or change) that occurs in another quantity.
In control systems analysis, sensitivity is measured as change in the overall transfer function of the system with changes in some part of the system (like the parameter variations in the plant or the sensor).

$$S_{\beta}^{\alpha} = \frac{\Delta\alpha/\alpha}{\Delta\beta/\beta}$$

where α is the quantity under observation and β is the parameter varying.

For an Open Loop System, $T(s) = G(s)$. Hence, $S_G^T = 1$.
This means that an OL system is 100% sensitive to changes in system parameters.

For a Closed Loop System, $T(s) = \frac{D(s)G(s)}{1+D(s)G(s)H(s)}$.
Hence,

$$S_G^T = \frac{1}{1 + D(s)G(s)H(s)}$$

(considering H(s) as constant) The above relation indicates that by choosing appropriate value of controller $D(s)$, the sensitivity to changes in the plant parameters can be made negligible (at the cost of decreasing the overall gain).

It can also be deduced that,

$$S_H^T = \frac{-D(s)G(s)H(s)}{1 + D(s)G(s)H(s)}$$

(considering G(s) as constant) This relation however tells that the system will be highly sensitive to the changes in the sensor parameters, especially if the gain of the controller is high.

Therefore, it must be ensured that the sensor is extremely well designed and should not be susceptible to changes.
(it is easier to pose restrictions on the design on the sensor than the design of the plant)

2.4 Disturbances

Disturbances are the unwanted signals that might get added to some processed signals in the control system due to practical factors.

There are two types of Disturbances, namely Erratic High Frequency disturbances and Slow-varying disturbances.

Erratic disturbances (also called Noise) have unpredictable wave-forms and can't be analyzed using any standard modelling. Stochastic modelling has to be used (which is not covered here).

Slow-varying disturbances are easier to observe and understand, and though not entirely predictable, they can be modelled or approximated using standard signals and analysed.

2.4.1 Standard Disturbance Models

Impulse Signal
Used to model a sudden (almost instantaneous) jerk.

$$\delta(t) \longleftrightarrow 1$$

Step Signal
Used to model a sudden (almost instantaneous) change.

$$u(t) \longleftrightarrow \frac{1}{s}$$

Ramp Signal
Used to model linear change.

$$tu(t) \longleftrightarrow \frac{1}{s^2}$$

Parabolic Signal
Used to model quadratic change.

$$\frac{t^2}{2}u(t) \longleftrightarrow \frac{1}{s^3}$$

Pulse Signal

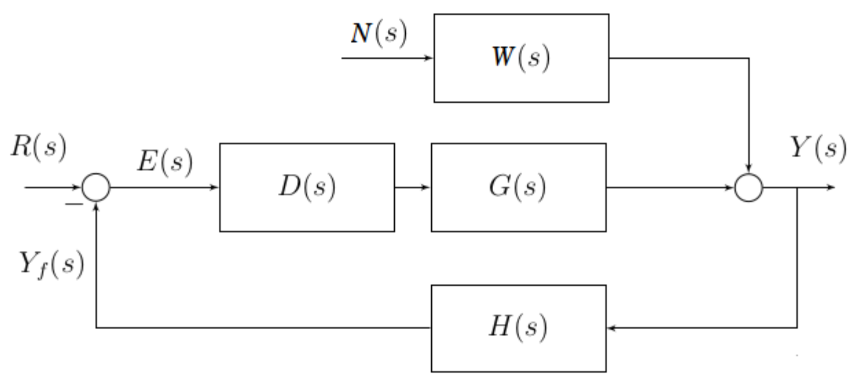
Used to model a sudden change which lasts only for a certain known period.

$$u(t) - u(t-T) \longleftrightarrow \frac{1-e^{-sT}}{s}$$

If a certain system gives desired response to these inputs, then it usually means the system can work accordingly for an arbitrary input. Hence, the above signals are also called as **Standard Test Signals**.

2.4.2 Disturbance rejection

External disturbance signal can be modelled into the system by adding a new input which adds to the output.



Here, the transfer function considering the disturbance signal as input is given by,

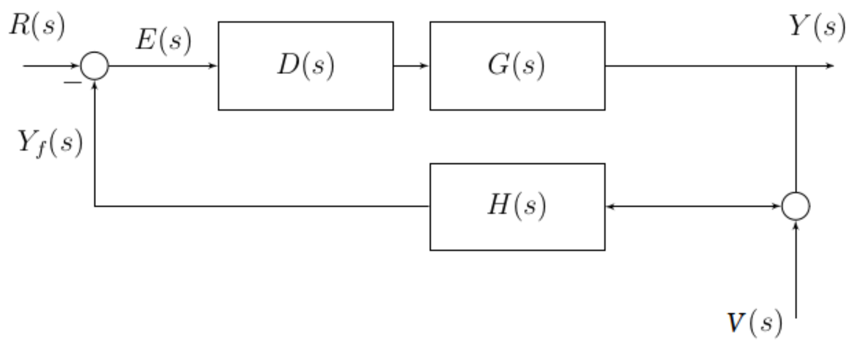
$$\frac{N(s)}{Y(s)} = \frac{W(s)}{1 + D(s)G(s)H(s)}$$

Hence the effect of disturbance can be effectively nullified by choosing appropriate value for the controller $D(s)$.

Such disturbances will directly add to the output and nothing can be done to compensate for it in the case of open loop control systems.

2.4.3 Noise filtering

The major source of noise in a system is the sensor itself. Hence, effect of noise is something that is added due to feedback which was absent in the open loop system.
This can be modelled into the block diagram as follows.



Here, the transfer function considering the noise signal as input is given by,

$$\frac{V(s)}{Y(s)} = \frac{-D(s)G(s)H(s)}{1 + D(s)G(s)H(s)}$$

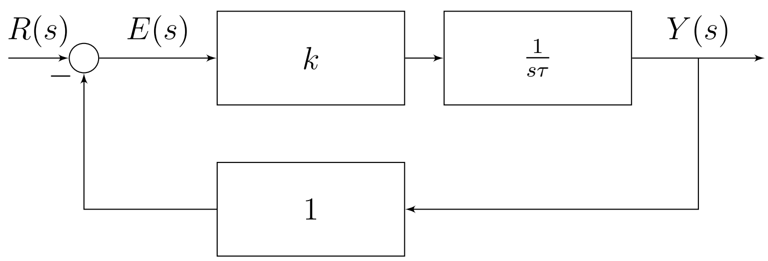
Hence the effect of noise becomes more significant as the controller gain is increased. This reinforces the intuition that adding a sensor and closing the loop will lead to noise effects.
To take care of this, a low pass filter must be used in order to avoid high frequency noise from entering the loop.

3 Analysis of Standard Systems

Most practical systems can be modelled using First and Second order systems. Hence, analysis and characterization of these two will be extremely useful.

3.1 First order systems

Consider an Open Loop transfer function, $G(s) = \frac{k}{s\tau}$ with unity negative feedback.



Closed Loop transfer function is given by,

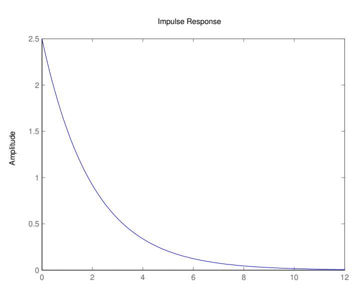
$$\boxed{T(s) = \frac{Y(s)}{R(s)} = \frac{k}{1 + s\tau}}$$

This is the standard 1^{st} order system transfer function.

k is referred to as **System gain** and τ is referred to as **Time constant**

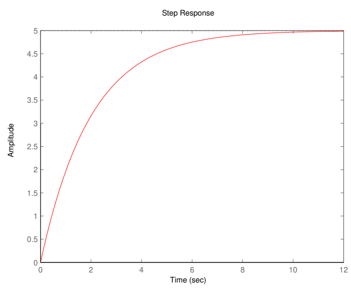
Impulse response is given by,

$$h(t) = \frac{k}{\tau} e^{-\frac{t}{\tau}} u(t)$$



Unit step response is given by,

$$s(t) = k[1 - e^{-\frac{t}{\tau}}]u(t)$$



Here, k is the steady state response of the system and $ke^{-\frac{t}{\tau}}$ is the transient response of the system.
This means, the unit step input is getting multiplied by a factor of k in steady state, hence it is called the system gain.

The time constant τ defines the speed of the system, since the transient response tells how fast the system will reach it's steady state.

Initial slope of step response is k/τ. This means, if the slope was maintained, then the system transient would die out at $t = \tau$.
However, since the slope is changing, the following observations can be made.

- at $t = \tau$, the output would have reached 63.2% it's steady state value.
- at $t = 3\tau$, the output would have reached 95.0% it's steady state value.
- at $t = 4\tau$, the output would have reached 98.2% it's steady state value.

Hence, 3τ and 4τ are called **Settling Time** with 5% and 2% tolerance band respectively.

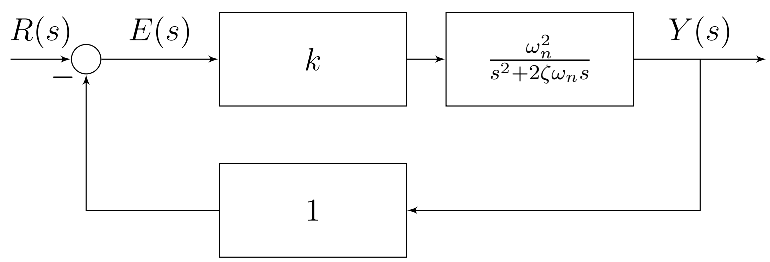
By the time $t = 5\tau$, the output would have reached more than 99% of it's steady state value and for all practical purposes, it can be said that steady state has been reached.

Hence, the system transients die faster if the time constant is lesser, which is why it defines the speed of the system.

(note that the time constant is defined only for stable systems, meaning such systems where the output actually settles down to some value and does not tend to infinity [more on stability later])

3.2 Second order systems

Consider an Open loop transfer function, $G(s) = \frac{k\omega_n^2}{s(s+2\zeta\omega_n)}$ with unity negative feedback.



Closed loop transfer function is given by,

$$\boxed{T(s) = \frac{Y(s)}{R(s)} = \frac{k\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}}$$

This is the standard 2^{nd} order system transfer function.

k is the system gain as usual.
ω_n is called **Natural Frequency**
ζ is called **Damping Ratio**

$s^2 + 2\zeta\omega_n s + \omega_n^2 = 0$ is the Characteristic equation.
Roots of the Characteristic equation are given by,
$s = -\zeta\omega_n \pm j\omega_n\sqrt{1-\zeta^2}$
$\omega_n\sqrt{1-\zeta^2} = \omega_d$, which is the **Damped Frequency**.

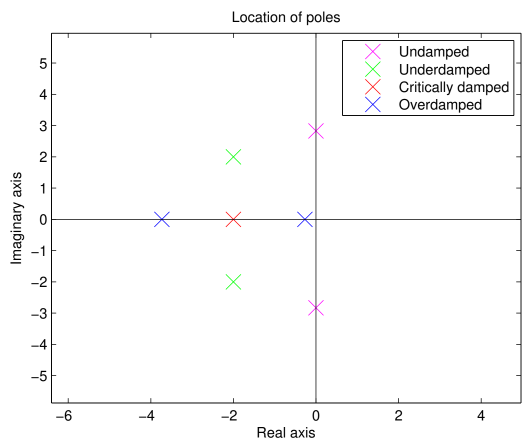
General expression for unit step response will be,

$$\boxed{s(t) = 1 - \frac{e^{-\zeta\omega_n t}}{\sqrt{1-\zeta^2}} sin(\omega_d t + \theta)} \quad where \;\; \theta = cos^{-1}\zeta$$

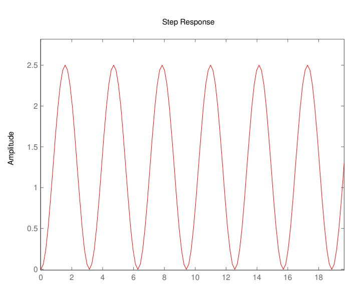
$cos\theta = \zeta$
$sin\theta = \sqrt{1-\zeta^2}$
$tan\theta = \frac{\sqrt{1-\zeta^2}}{\zeta}$

Time constant of a second order system is given by, $\tau = \frac{1}{\zeta\omega_n}$.

The response to a second order system can be classified to 4 different cases depending on the value of the damping ratio. Location of poles in the s-plane for each of the 4 cases:



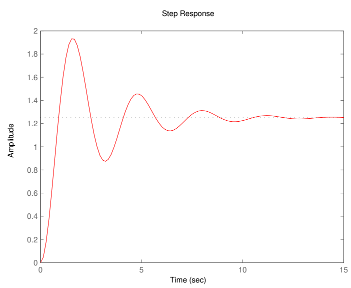
Undamped Case: $\zeta = 0$
There is no damping factor. Roots of CE are purely imaginary. Hence the step response will be oscillatory, which is not usually favoured.



Under-damped Case: $0 < \zeta < 1$
Damping factor causes oscillations to die out and settle to steady state value

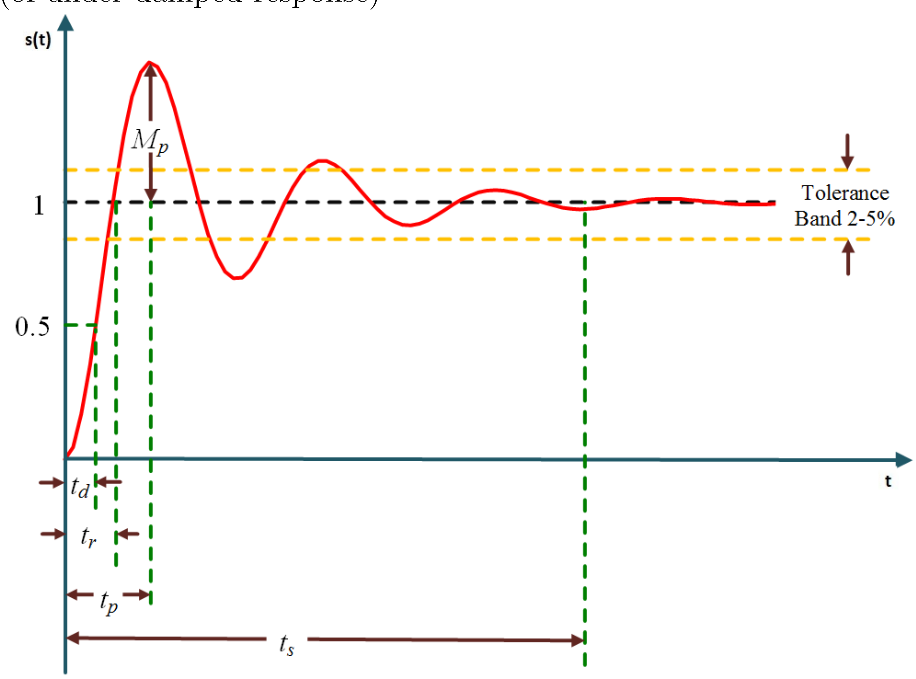
eventually.
This is the most commonly used configuration in second order systems.

When designing a second order system, the designer has to ensure the system meets the transient and steady state specifications required. The transient specifications are generally analysed from step response of the system in time domain. Frequency domain specifications can also be useful in design.



3.2.1 Time Domain Specifications

(of under-damped response)



- Delay time (t_d) is the time required for the response to reach 50% of it's steady state value.
- Rise time (t_r) is the time required for the response to reach the steady state value for the first time.
- Peak time (t_p) is the time required for the response to reach it's maximum value.
- Maximum peak overshoot (m_p) is the maximum value the response reaches, it is also the maximum deviation from the steady state value. (it is measured in terms of percentage)
- Settling time (t_s) is the time required for the response to reach 95% of 98% of it's final value (depending on which is appropriate for the specific application)

These terms can be derived from the typical step response formulae using the definitions.

$$t_r = \frac{\pi - \theta}{\omega_d}$$
$$t_p = \frac{\pi}{\omega_d}$$
$$m_p = e^{\frac{-\zeta\pi}{\sqrt{1-\zeta^2}}}$$
$$t_s\ (for\ 95\%) = \frac{3}{\zeta\omega_n}$$
$$t_s\ (for\ 98\%) = \frac{4}{\zeta\omega_n}$$

Number of oscillations completed before settling (N) is given by,
$N = t_s \frac{\omega_d}{2\pi}$

Peak overshoots occur at π/ω_d, $3\pi/\omega_d$, $5\pi/\omega_d$, ...
Peak undershoots occur at $2\pi/\omega_d$, $4\pi/\omega_d$, $6\pi/\omega_d$, ...
Similarly, the peak values at the corresponding overshoots and undershoots are governed by, $m_p = e^{\frac{-n\zeta\pi}{\sqrt{1-\zeta^2}}}$ where n is the integer before π/ω_d term.

If the input is scaled by a factor, the output will also be scaled by that factor and the peak overshoot will be same.

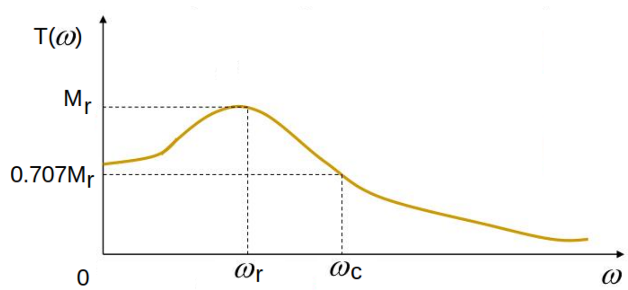
It can be noted that m_p is a function of ζ and others are all functions of ζ and ω_n. Peak overshoot and rise time are usually chosen as design specifications for satisfying the transient response required.

Some patterns/conclusion that can be useful:

- $m_p \propto 1/\zeta$
- $m_p \propto 1/t_p$
- $\zeta \propto 1/\theta$
- $5\% < m_p < 25\%$ (preferred range in general)
- $0.4 < \zeta < 0.7$ (preferred range in general)

3.2.2 Frequency Domain Specifications

(of under-damped response)



- Resonant peak (M_r) is the peak magnitude of the frequency response of an under-damped second order system.
- Resonant frequency (ω_r) is the frequency at which the resonant peak occurs.
- Cut-off frequency (ω_c) is the frequency at which the magnitude of the frequency response is $\frac{1}{\sqrt{2}}$ (or 0.707) times it's resonant peak.

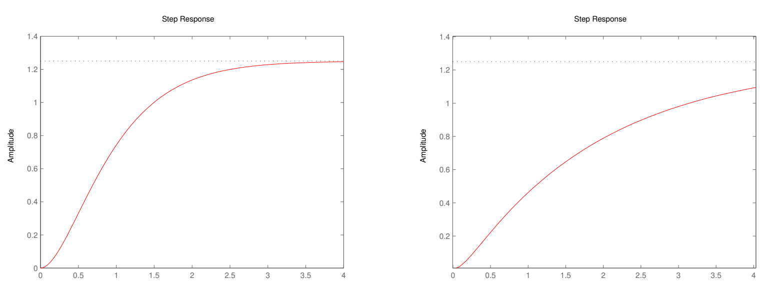
$$M_r = \frac{k}{2\zeta\sqrt{1-\zeta^2}}$$
$$\omega_r = \omega_n\sqrt{1-2\zeta^2}$$

From the relation, it is obvious that M_r exists only when ω_r is real and positive i.e $\sqrt{1-2\zeta^2} > 0$.
$\zeta < 0.707 \implies M_r > 1$
$\zeta = 0.707 \implies M_r = 1$
$\zeta > 0.707 \implies M_r$ does not exist

M_r is a function of ζ and ω_c is a function of ζ and ω_n. Hence, they directly related to the time domain specifications and can be used as design specifications if frequency domain analysis is being carried out.

Critically Damped Case: $\zeta = 1$
Limiting case of damping, meaning the oscillations are effectively killed and the response is similar to that of first order system.

Over-damped Case: $\zeta > 1$
Damping factor greater than 1 causes the system response to be slow, hence this configuration is not used.



For a second order control system to be stable, the damping ratio has to be positive. If it is negative, the system is certainly unstable.

3.3 Dominant Poles

First and second order systems are well studied and majority of the practical systems can be modelled using them. However, it is not out of the question that a designer might have to deal with higher order systems. In analysis of systems of order 2 or more than 2, the concept of dominant poles becomes very useful.

Consider a system has 2 poles at $s = -s_1$ and $s = -s_2$. The location of these poles determine the response of the system.

$G(s) = \frac{k}{(s+s_1)(s+s_2)}$
$\implies G(s) = \frac{k_o}{(1+s/s_1)(1+s/s_2)}$ in time constant form.
$\implies G(s) = \frac{A}{s+s_1} + \frac{B}{s+s_2}$ using partial fractions.
$\therefore g(t) = Ae^{-t/s_1} + Be^{-t/s_2}$ by Inverse Laplace transform.

The time constants of the terms are $1/s_1$ and $1/s_2$. If the values of s_1 and s_2 are such that $s_2 \geq 5s_1$, it means that the term with time constant $1/s_1$ will decay very slowly when compared to the term with time constant $1/s_2$ and hence, the fast decaying pole is considered to be **Insignificant pole** whereas the slow decaying pole is considered to be **Dominant pole**.

In case of higher order systems where the dominant poles are complex conjugates, the effective time constant is given by,

$$\tau = \frac{-1}{Real\ [Dominant\ Pole]}$$

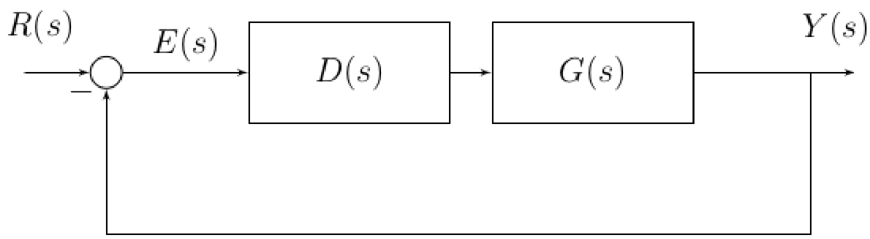
If the dominant pole criterion is satisfied, it effectively reduces higher order systems to lower order systems because insignificant poles can be neglected (given the transfer function is expressed in pole-zero form).

3.4 Steady State Error

The accuracy of a control system is given by it's steady state response as mentioned earlier. The closer the steady state response is to the desired output, the more accurate the system is. Which is why the accuracy of a system can be defined in terms of steady state error.

For a general control system, the steady state error is defined as the difference between the steady state value of the actual output and the desired output.
$\implies e_{ss} = y_{ss} - r_{ss}$

Steady State Error for unity feedback systems
If the system is a unity feedback system, it can be observed that the steady state error is same as the final value of the error signal obtained (which is fed to the controller).



$$E(s) = \frac{R(s)}{1 + D(s)G(s)}$$

Using Final Value Theorem,

$$e_{ss} = \lim_{s \to 0} \frac{sR(s)}{1 + D(s)G(s)}$$

The following table gives steady state error for different types of inputs against different types of systems.

	Step input	Ramp input	Parabolic input
Type 0 system	$\frac{1}{1+K_p}$	∞	∞
Type 1 system	0	$\frac{1}{K_v}$	∞
Type 2 system	0	0	$\frac{1}{K_a}$

Positional Error Constant

$$K_p = \lim_{s \to 0} G(s)$$

Velocity Error Constant

$$K_v = \lim_{s \to 0} sG(s)$$

Acceleration Error Constant

$$K_a = \lim_{s \to 0} s^2 G(s)$$

Note that all these constants are DC gains of the particular systems. Hence, by designing the appropriate controller which can manipulate the overall DC gain, it is possible to reduce the steady state error using feedback.

Steady State Error for non-unity feedback systems

For non-unity feedback systems ($H(s) \neq 1$), the error signal obtained (as input to controller) is not the steady state error. It is more suitably called actuating error.

To find the steady state error, there are 3 methods.

- Find final value of the output (using final value theorem) and find the difference of it with desired output.
- Convert the non-unity feedback system to equivalent unity feedback system by manipulating the block diagram.
  This will give $G'(s) = \frac{D(s)G(s)}{1+D(s)G(s)H(s)-D(s)G(s)}$

  $$\implies e_{ss} = \lim_{s \to 0} \; sR(s)\left[\frac{1 + D(s)G(s)H(s) - D(s)G(s)}{1 + D(s)G(s)H(s)}\right]$$

- Use the formulae:

  $$e_{ss} = \lim_{s \to 0} \; sR(s)[1 - T(s)]$$

  where $T(s)$ is the closed loop transfer function

Note that the steady state error is defined only for stable systems.

4 Stability Analysis

Any control system being designed has to be designed under the condition that it must satisfy stability. Meaning, an unstable system which has all other desired qualities is still unacceptable.

Stability is generally defined in the following two ways.

Bounded Input Bounded Input (BIBO) Stability:
A system is said to be BIBO stable if for a bounded input, the output always turns out to be bounded.
This means for any input (command or disturbance), if it is absolutely summable, the output must also be absolutely summable.
For LTI systems, it can also be stated that BIBO stability means the impulse response must be absolutely summable.

Zero Input Stability:
A system is said to be Zero-input stable if all the states of the system are bounded (or finite) when the input to the system is zero.

For most systems, both the above definitions will coincide, meaning for this class of systems, analysis can be directly done for stability.

4.1 Stability analysis in s-plane

The closed loop stability of any system is analysed using its transfer function, mainly the characteristic equation. The location of poles of the transfer function i.e the solutions to the characteristic equation in the s-plane will determine the stability of the system.
Consider a system with closed loop transfer function as $T(s) = \frac{(s+z_o)(s+z_1)....(s+z_m)}{(s+p_o)(s+p_1)....(s+p_n)}$ where $n > m$.
The partial fraction expansion of the system can be expressed as,

$$T(s) = \frac{A_o}{s+p_o} + \frac{A_1}{s+p_1} + ... + \frac{A_n}{s+p_n}$$

By taking inverse Laplace transform to obtain the impulse response,

$$h(t) = A_o e^{-p_o t} + A_1 e^{-p_1 t} + ... + A_n e^{-p_n t}$$

It can be noted that the impulse response will be absolutely summable only if $p_o, p_1, ...p_n$ are positive.
If even one of them is negative, the impulse response will eventually tend to infinity and it will be an unstable system.
The unique case in which the impulse response is neither dying nor growing, but stays as a non-zero constant or oscillates in a fixed range (in steady state) implies the system is "Marginally stable".

Therefore, it can be concluded that the system will be stable if and only if all of its poles lie on the left half of s-plane.

4.1.1 Routh-Hurwitz Criteria

Routh-Hurwitz is a method that can provide both absolute as well as relative stability of a system by analysing it's characteristic equation.
This method also points out the general location of the poles but not the exact points.
(note that the same technique can also be used to gather information about the zeros, though it is of no use for stability analysis)

Routh-Hurwitz method needs the construction of the Routh table, which is done using the characteristic equation.

Let the equation under consideration be,
$P(s) = a_0 s^n + a_1 s^{n-1} + a_2 s^{n-2} + ...a_{n-1}s + a_n$
For convenience, assume that $a_n \neq 0$ and $a_0 > 0$ (so as to maintain uniformity in sign and to make sure $s = 0$ factor is not present).

The characteristic equation directly gives the coefficients of first two rows of the Routh table. The rest of the table is filled according to the pattern shown.

Routh Table

s^n	a_0	a_2	a_4	$\dots$
s^{n-1}	a_1	a_3	a_5	$\dots$
s^{n-2}	b_0	b_1	b_2	$\dots$
s^{n-3}	c_0	c_1	c_2	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s^0	h_0	0	0	0

where,

$$b_0 = \frac{a_1 a_2 - a_3 a_0}{a_1}\ ; \qquad b_1 = \frac{a_1 a_4 - a_5 a_0}{a_1}$$

$$c_0 = \frac{b_0 a_3 - a_1 b_1}{b_0}\ ; \qquad c_1 = \frac{b_0 a_5 - a_1 b_2}{b_0}$$

and so on...

Once the Routh Table is constructed, the sign changes in the first column of the table are to be noted.

For the system to be stable, all coefficients in the first column of the Routh table must be strictly positive.

If there are one or more sign changes, then the system is unstable.
The number of sign changes in the first row of the Routh table is equal to the number of roots (of the equation) that lie in the right half of s-plane.

Cases with coefficients turning out to be 0-
1. If one coefficient in the first column is zero, with its corresponding row coefficients being non-zero, it means there is at least one root on the right hand side of s-plane.
Replace zero with some small positive number ϵ and perform normal analysis.

2. If row/s of zeros occurs, it means symmetric roots about the origin are present.
Obtain **Auxiliary Equation** which is constructed by using the coefficients of the row above the row of zeros.
Then differentiate this equation (w.r.t s) to obtain new coefficients for the row of zeros and perform the usual analysis. Note that power of Auxiliary

equation must be even since it will give number of roots that are symmetric about the origin.

- Number of sign changes above AE = number of roots on RHP
- Number of sign changes below AE = number of symmetric roots on RHP (each such root on RHP will have a corresponding symmetric root in LHP)

If there is only a single row of zeros, with no sign change, it means the system has simple poles on the $j\omega$ axis and is marginally stable.
If there are multiple rows of zeros, with no sign change, it means the system has repeated poles on the $j\omega$ axis and is unstable.

The Routh-Hurwitz criteria can be reduced to simple short-cuts for second and third order characteristic equations.

Conditions for stability of second order system:

$$as^2 + bs + c = 0$$

$a > 0,\ b > 0,\ c > 0$

Conditions for stability of third order system:

$$as^3 + bs^2 + cs + d = 0$$

$a > 0,\ b > 0,\ c > 0,\ d > 0,\ bc > ad$
(note that if $bc = ad$, it means the system is marginally stable)

Conditional Stability

The most important utility of the Routh-Hurwitz criteria is to determine the range of values of the system gain K for which the closed loop system will be stable.

In general, if a system with open loop transfer function $G(s)$ is closed using unity feedback and controller gain K. The characteristic equation of the closed loop system will be $1 + KG(s) = 0$.
Using Routh-Hurwitz criteria, the constraint equations on K can be found.

The value of K where the system is marginally stable is K_m.
The conclusion for range of values of K can be-

- Stable for $0 < K < K_m$; Unstable for $K > K_m$
- Unstable for $0 < K < K_m$; Stable for $K > K_m$

The value of frequency ω at $K = K_m$ is the frequency of oscillation ω_n.

Relative Stability

So far, Routh-Hurwitz criteria has been used to evaluate absolute stability by analysing position of poles with respect to the $j\omega$ axis i.e $s = 0$ line. The same criteria can also be used to evaluate relative stability of a system with respect to any vertical line in the s-plane.

To find stability of a system with respect to the line $s = -a$, let $s + a = s_o$. Hence, $s = s_o - a$. Substitute $s_o - a$ for s in the characteristic equation of the system and apply Routh-Hurwitz criteria for the new equation in s_o-plane.

This technique can be used to find the number of roots (poles or zeroes from the equation) that lie on the $s = -a$ line, between $s = 0$ and $s = -a$ lines, RHS of $s = -a$ line, LHS of $s = -a$ line.

4.1.2 Root Locus

The Root Locus technique provides a graphical method of plotting the locus of the roots (poles and zeros) of a closed loop transfer function in the s-plane as a given system parameter (typically gain E) is varied over the complete range of values (0 to ∞).

The Root Locus plot is extensively used in control systems design because once the sketch is obtained, the roots corresponding to a particular value of the system parameter can be obtained or the value of a desired root location can be determined.

Consider a system with open loop transfer function $G(s)$, is closed using feedback element $H(s)$ with controller gain K. Then the closed loop transfer function will be $T(s) = \frac{KG(s)}{1+KG(s)H(s)}$. Hence the characteristic equation of the closed loop system will be $1 + KG(s)H(s) = 0$. $\implies KG(s)H(s) = -1$

From the above result, two conditions can be obtained for any point s to lie on the root locus.

Magnitude Criterion:

$$|KG(s)H(s)| = 1$$

Angle Criterion:

$$\angle G(s)H(s) = \pm(2q + 1)180^o$$

It can be noted that since K can take any value, every point in the s-plane inherently satisfies the magnitude criterion. Hence, the magnitude criterion is used to find the value of K given a particular value of s.

Since K does not contribute any angle, angle criterion is used to check if any given point $s = s_1$ lies on the root locus plot of the given closed loop system.

The construction of the root locus plot for any closed loop system for stability and design purpose is done using the following steps/rules.

Root Locus Construction Rules

1. The root locus is symmetrical about the σ axis ($j\omega = 0$).

2. As K increases from 0 to ∞, each branch of the root locus originates from an open loop pole with $K = 0$ and terminates at either an open loop zero or infinity with $K = \infty$. The number of branches terminating at ∞ is equal to the difference between open loop poles and zeros $(n - m)$.

3. A point on the real axis lies on the root locus if the number of open loop poles and zeros on the real axis to the right of this point is odd.

4. The $(n - m)$ branches that terminate at ∞ do so along straight line asymptotes whose angles are given by, $\phi_A = \frac{(2q+1)180^o}{n-m}$; $q = 0, 1, 2, ...n - m - 1$.

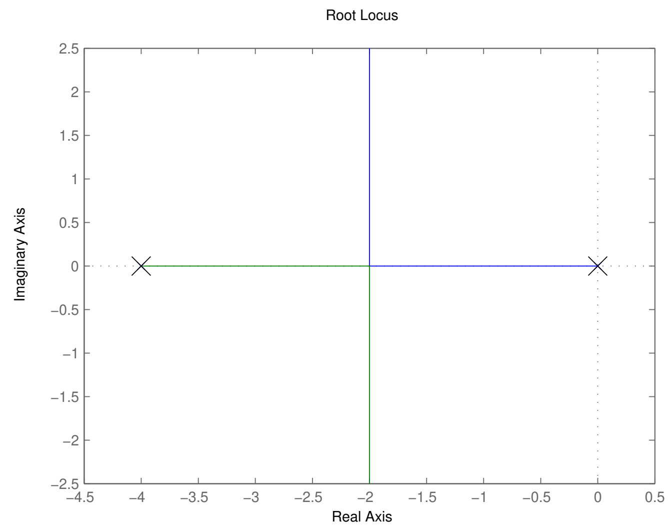
5. The asymptotes cross the real axis at a point known as centroid, determined by the relationship, $C = (\Sigma\text{Real[poles]} - \Sigma\text{Real[zeros]})/(n - m)$.

6. Break away points are those points at which the root locus of two different branches meet and break away towards ∞, hence they are the points where multiple roots occur. The break away points of the root locus are the solutions to $dK/ds = 0$.
   Break away points are points on the real axis only. Meaning, if the solutions to $dK/ds = 0$ are complex, then there are no points at which the closed loop poles meet for breaking away.

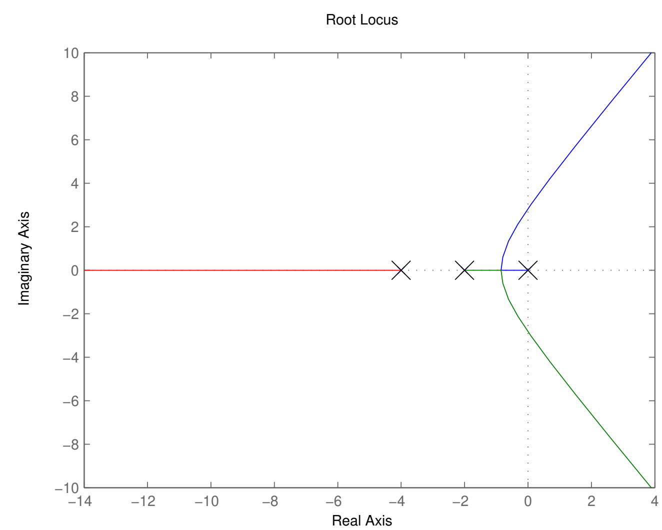
7. In case of existence of complex break poles or zeros, the break away directions are given by,
   Angle of departure from an open loop pole side: $\phi_d = \pm 180^o + \phi$;
   Angle of arrival from an open loop zero side: $\phi_a = \pm 180^o - \phi$;
   where $\phi = \Sigma\theta_z - \Sigma\theta_p$ (i.e net angle contributed at the open loop root by all other open loop roots)

8. The intersection of root locus branches with the $j\omega$ axis can be determined by using Routh-Hurwitz criteria. At $K = K_m$ where the root locus branches intersect the $j\omega$ axis, a row of 0s will occur in the Routh table. The row above this row gives auxiliary equation, whose roots will give the points of intersection.

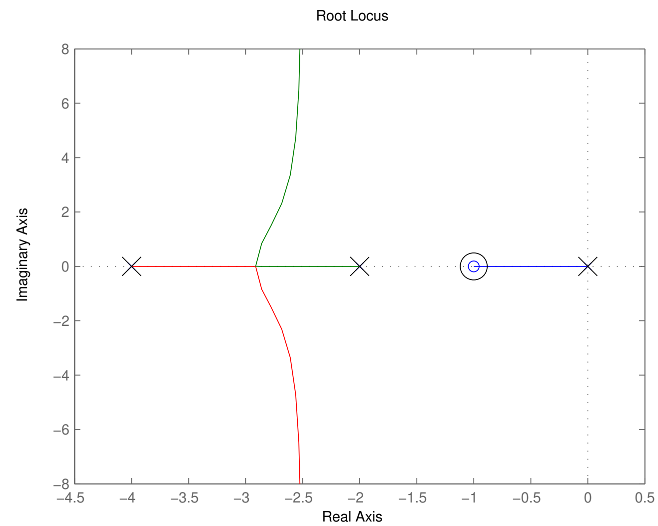
Consider a system with open loop transfer function $G(s) = \frac{k}{s(s+4)}$. The root locus plot will be as indicated in the following figure.



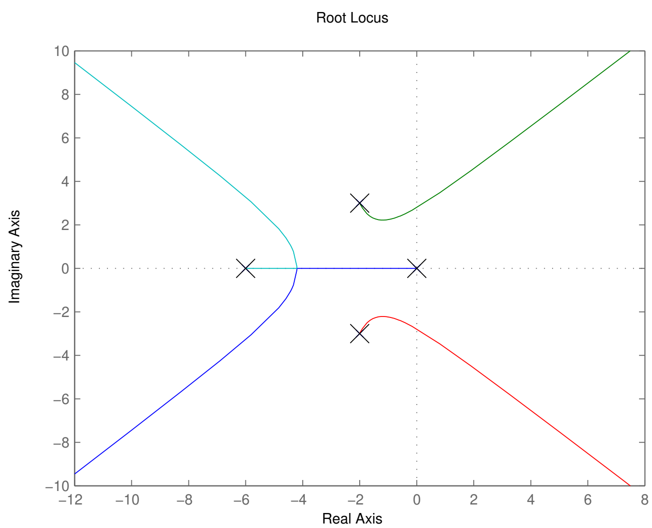
The addition of a pole to an open loop system will drive the system towards instability if the loop is closed. This is illustrated in the root locus plot of $G(s) = \frac{k}{s(s+2)(s+4)}$.



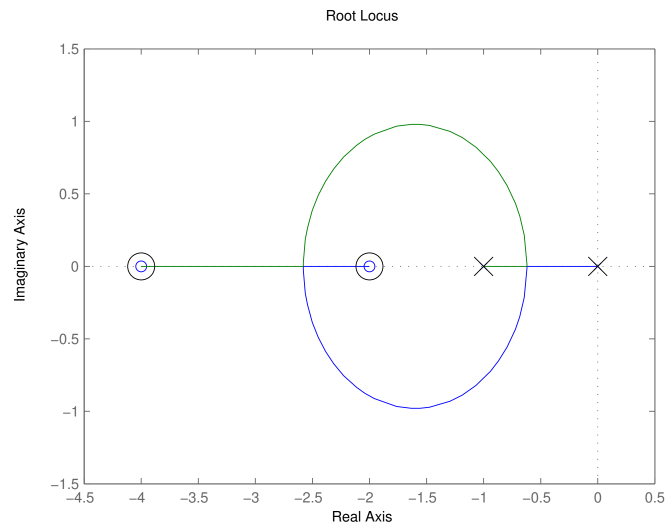
The addition of a zero to an open loop system will increase the relative stability of the closed loop system. This is illustrated in the root locus plot of $G(s) = \frac{k(s+1)}{s(s+2)(s+4)}$.



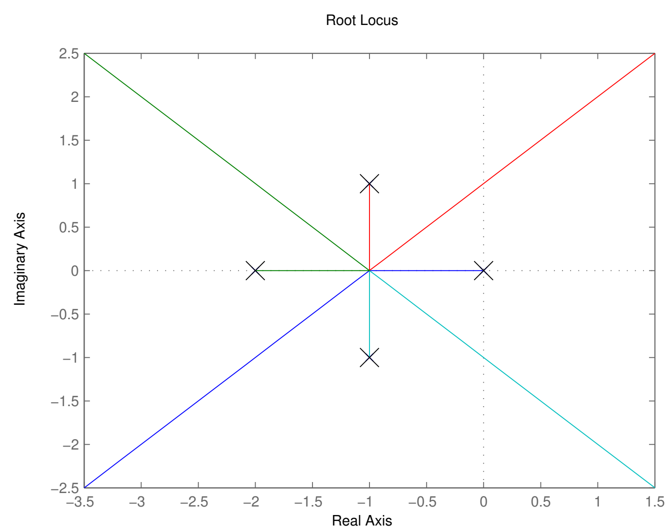
The root locus plot of a system with complex poles is shown.
$G(s) = \frac{k}{s(s+6)(s^2+4s+13}$



If the root locus of a system traces the path of a circle, then the break away points will be of the form $a + \sqrt{b}$ where a is the center of the circle and $\sqrt{b}$ is it's radius. $G(s) = \frac{k(s+2)(s+4)}{s(s+1)}$



The root locus plot of a system with symmetric poles will have break away points at the centroid. $G(s) = \frac{k}{s(s+2)(s^2+2s+2)}$



Transportation Lag
The transportation lag is the delay between the time an input signal is applied to a system and the time the system reacts to that input signal. Meaning if input is r(t), the output can be represented as y(t-T).
In the Laplace domain, the transfer function representing the transportation lag will be: $\frac{Y(s)}{R(s)} = e^{-sT}$.
To incorporate the exponential term in the root locus, it can be approximated as $e^{-sT} = 1 - sT$ using first order approximation of Taylor's series expansion.

If transportation lag occurs in a system, then by using the above approximation, it can be inferred that the effective characteristic equation will be 1 - G(s)H(s) = 0 i.e it has to be modelled as a positive feedback system. To analyse such systems in the root locus, the magnitude criterion will remain the same but angle criterion will change to
$\angle G(s)H(s) = \pm 2q(180^o) \quad q = 0, 1, 2, ...$
. The root locus plot derived from this angle criterion is called **Complementary Root Locus** and is used for positive feedback systems.

The rules that change in the complementary root locus from the original are-

- A point on the real axis lies on the complementary root locus if the number of open loop poles and zeros on the real axis to the right of this point is even.

- The $(n - m)$ branches that terminate at ∞ do so along straight line asymptotes whose angles are given by, $\phi_A = \frac{(2q)180^o}{n-m}; \quad q = 0, 1, 2, ...n - m - 1$.

- In case of existence of complex break poles or zeros, the break away directions are given by,
  Angle of departure from an open loop pole side: $\phi_p = \phi$;
  Angle of arrival from an open loop zero side: $\phi_z = -\phi$;
  where $\phi = \Sigma\theta_z - \Sigma\theta_p$ (i.e net angle contributed at the open loop root by all other open loop roots)

The rest of the rules are same as usual.

4.2 Stability analysis using Frequency response

If a sinusoidal signal is applied as an input to a Linear Time-Invariant (LTI) system, then it produces the steady state output, which is also a sinusoidal signal. The input and output sinusoidal signals have the same frequency, but differ in amplitudes and phase angles.

Consider a system with transfer function T(s) with input $r(t) = Asin(\omega_o t+\theta)$ and output $y(t) = Bsin(\omega_o t+\phi)$ where $B = A|G(j\omega_o)|$ and $\phi = \theta+\angle G(j\omega_o)$. Here, $|G(j\omega_o)|$ is the magnitude of the frequency response and $\angle G(j\omega_o)$ is the angle of the frequency response.
$\implies G(\jmath\omega) = |G(j\omega)|\angle G(j\omega)$

Stability of a system can be analysed using its frequency response. One of the major advantages of this type of analysis in design is, there is no need to come up with an actual model for the system. The frequency response can be obtained by giving the system sample inputs and analysing the outputs.

Gain Cross-over Frequency is the frequency at which the magnitude of the transfer function is one. It is denoted by ω_{gc}.

Phase Cross-over Frequency is the frequency at which the phase of the transfer function is 180^o. It is denoted by ω_{pc}.

Gain Margin is the magnitude factor by which a stable system can be increased before making it marginally stable (or driving it towards instability). Mathematically, the gain margin (GM) is equal to the reciprocal of the magnitude of the Nyquist plot at the phase cross over frequency.
$\implies GM = \frac{1}{M_{pc}}$
Gain is usually expressed in decibels for convenience, hence $GM = 20\log_{10}\frac{1}{M_{pc}}$
Where, M_{pc} is the magnitude in normal scale at $\omega = \omega_{pc}$.

Phase Margin is the additional phase which can be added to a stable system before making it marginally stable (or driving it towards instability). Mathematically, the phase margin (PM) is equal to the sum of 180^o and the phase angle at the gain cross over frequency.
$\implies PM = 180^o + \Phi_{gc}$
Where, Φ_{gc} is the phase angle at $\omega = \omega_{gc}$.

Frequency domain stability criteria:

	GM & PM	ω_{gc} & ω_{pc}
Stable	$GM > 0dB$ & $PM > 0^o$	$\omega_{gc} < \omega_{pc}$
Marginally Stable	$GM = 0dB$ & $PM = 0^o$	$\omega_{gc} = \omega_{pc}$
Unstable	$GM < 0dB$ & $PM < 0^o$	$\omega_{gc} > \omega_{pc}$

(this criteria is for closed loop system, assuming open loop system is stable)

4.2.1 Polar Plot

The Polar plot is a plot, which can be drawn between the magnitude and the phase angle of $G(j\omega)H(j\omega)$ by varying ω from 0 to ∞.
Hence, it provides a mapping from the s-plane to G(s)H(s)-plane.

Steps to obtain polar plot:

1. Substitute, $s = j\omega$ in the open loop transfer function and write the expressions for magnitude and the phase of $G(j\omega)H(j\omega)$.

2. Find the starting magnitude and the phase of $G(j\omega)H(j\omega)$ by substituting $\omega = 0$. So, the polar plot starts with this magnitude and the phase angle.

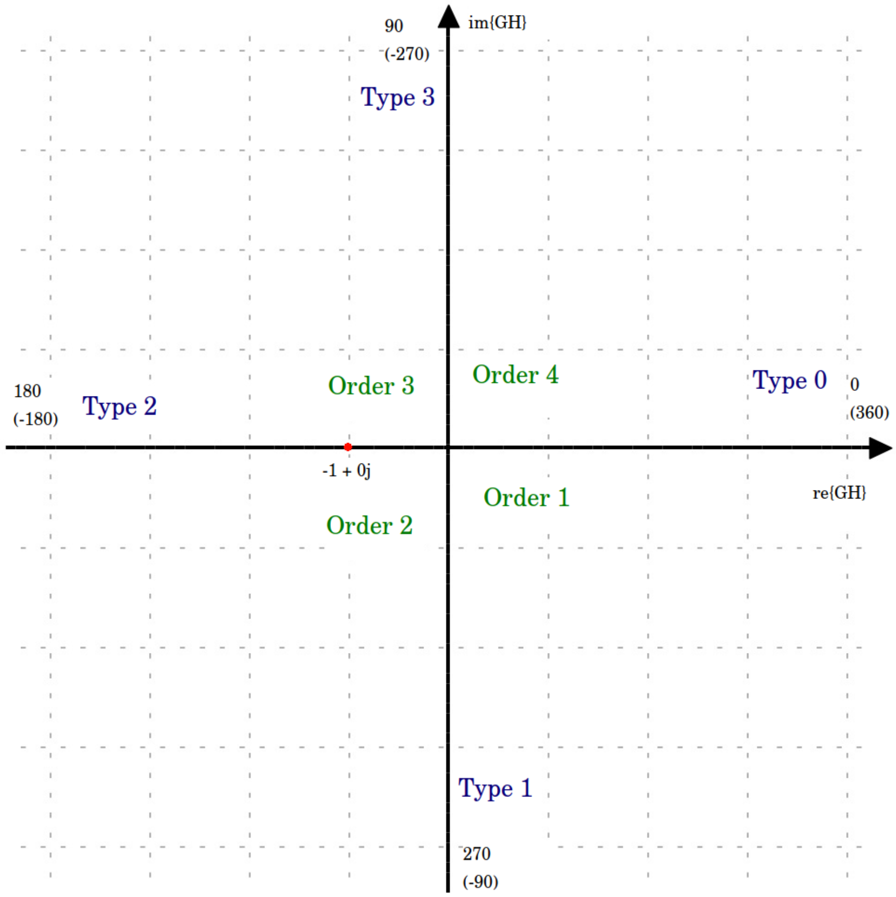
3. Find the ending magnitude and the phase of $G(j\omega)H(j\omega)$ by substituting $\omega = \infty$. So, the polar plot ends with this magnitude and the phase angle.

4. Check whether the polar plot intersects the real axis, by making the imaginary term of $G(j\omega)H(j\omega)$ equal to zero and find the value of ω.

5. Check whether the polar plot intersects the imaginary axis, by making real term of $G(j\omega)H(j\omega)$ equal to zero and find the value of ω.

6. For drawing polar plot more clearly, find the magnitude and phase of $G(j\omega)H(j\omega)$ by considering the other value of ω.

Standard polar plot structure for different types and orders of systems is shown.

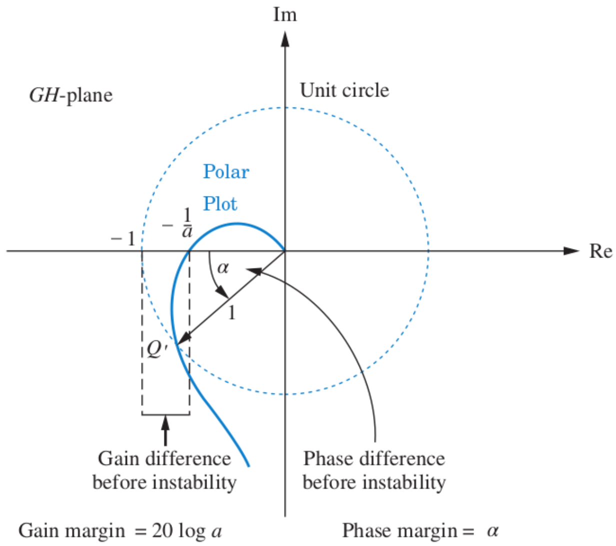


The blue words (type of system) indicate the starting direction and the green words (order of system) indicate the ending direction of the polar plot.

Polar plot of an open loop transfer function can be used for stability analysis of its corresponding closed loop system.

- If the plot encloses the critical point $(-1 + 0j)$, the system is unstable since $\omega_{gc} > \omega_{pc}$ i.e GM and PM are negative.
- If the plot intersects the critical point $(-1+0j)$, the system is marginally stable since $\omega_{gc} = \omega_{pc}$ i.e GM and PM are zero.
- If the plot does not enclose the critical point $(-1 + 0j)$, the system is stable since $\omega_{gc} < \omega_{pc}$ i.e GM and PM are positive.

The figure shows the polar plot of a stable system and illustrates when it can reach instability using gain and phase margins.



The polar plot is used to analyse stability of closed loop system in more detail the Nyquist stability criteria.

4.2.2 Nyquist Criterion

Nyquist plot is the continuation of polar plot for finding the stability of the closed loop control systems by varying ω from -∞ to ∞.

The Nyquist stability criterion works on the **Principle of Argument**.
Any closed contour defined in the s-plane will map to another closed contour in the $1 + G(s)$-plane, which is called as a plot.

- If the contour in the s-plane encircles a pole in the clockwise direction, the plot encircles the origin in anti-clockwise direction
- If the contour in the s-plane encircles a zero in the clockwise direction, the plot encircles the origin in clockwise direction.

Therefore, the number of anti-clockwise encirclements of the origin in $1 + G(s)$-plane is equal to the difference between number of poles and number of zeros encircled by the contour in s-plane.

This principle can be applied to analyse stability.
Consider a system with open loop transfer function $G(s)$, which is closed using unity negative feedback. The closed loop transfer function will be $T(s) = \frac{G(s)}{1+G(s)}$.
The stability of open loop system is known, because it is given by the poles of $G(s)$ that are on right half of s-plane, denoted by P. To analyse the stability of the closed loop system, it is necessary to get the number of poles of $T(s)$ present on the right half of s-plane which is same as the number of zeros of the characteristic equation $1 + G(s)$, denoted by Z.

Hence, a contour on the s-plane is defined such that it covers the entire right half of s-plane (starting from origin to $+j\infty$ (imaginary axis), extends to ∞ on positive real axis to $-j\infty$ (imaginary axis) and ends back at origin. This is called "Nyquist Contour".
Instead of taking $1+G(s)$-plane and considering anti-clockwise encirclements of $0 + 0j$ (origin) for mapping, it is more convenient to take $G(s)$-plane and consider anti-clockwise encirclements of $-1 + 0j$ instead (which is called "Critical Point"). When the Nyquist Contour is mapped to the $G(s)$-plane, the obtained graph is called "Nyquist Plot".

$$N = P - Z$$

where P is known and N is found using the Nyquist plot.
Stability is determined using the value of Z.

- $Z > 0 \implies$ closed loop system is unstable (or marginally stable)
- $Z = 0$ is zero $\implies$ closed loop system is stable

A few examples implying the Nyquist stability criterion are given.

(i) Consider an open loop transfer function $G(s) = \frac{5(s+1)}{(s-2)(s+3)}$ in unity negative feedback configuration.
There is 1 open loop pole on RHP ($s = 2$), hence $P = 1$. The characteristic equation will be $s^2 + 6s - 1 = 0$, which is unstable with 1 root on RHP, hence $Z = 1$. $N = Z - P = 0$, meaning the Nyquist plot must encircle the critical point 0 times.

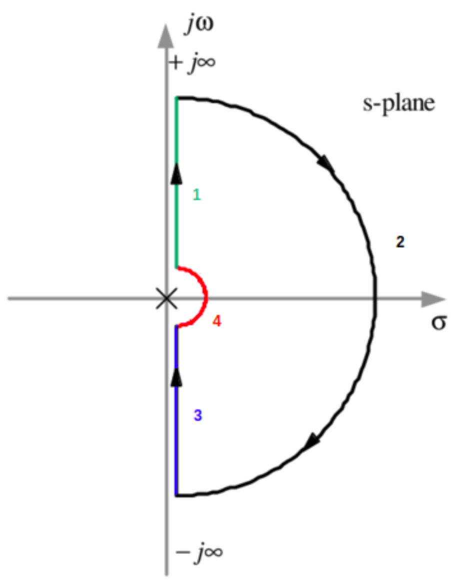
(ii) Next consider a different open loop transfer function $G(s) = \frac{4(s+2)}{(s-1)(s-2)}$ in unity negative feedback configuration.

There are 2 open loop poles on RHP ($s = 1, 2$), hence $P = 2$. The characteristic equation will be $s^2 + s + 10 = 0$, which is stable i.e no roots on RHP, hence $Z = 0$. $N = Z - P = -2$, meaning the Nyquist plot must encircle the critical point in clockwise direction 2 times.

(iii) Finally, consider another open loop transfer function $G(s) = \frac{10}{(s-1)(s+2)(s+3)}$ in unity negative feedback configuration.
There is 1 open loop pole on RHP ($s = 1$), hence $P = 1$. The characteristic equation will be $s^3 + 4s^2 + s + 4 = 0$, which is marginally stable with a pair of complex conjugate roots on $j\omega$ axis, hence the Nyquist plot should intersect the critical point twice.

The following steps are to be followed to obtain Nyquist plot from the Nyquist contour i.e from s-plane to $G(s)$-plane.

The Nyquist contour is divided to 4 parts.



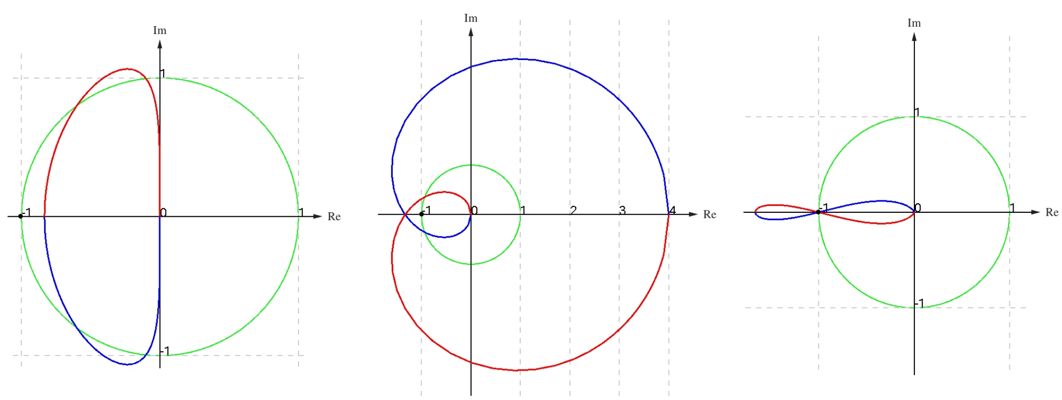
1. First part is mapped by obtaining the polar plot.
2. For the second part, substitute $s = lim_{s\to\infty} Re^{j\theta}$; θ goes from 90^o to -90^o. Usually this part maps to a single point in the G(s)-plane.

3. Third part is mapped by obtaining the inverse polar plot i.e mirror of polar plot with respect to real axis.

4. For the fourth part, substitute $s = lim_{s \to 0} Re^{j\theta}$; $\quad \theta$ goes from -90^o to 90^o. However, this is necessary only if there are open loop poles on the $j\omega$ axis.

Once the complete Nyquist plot is obtained, the number of anti-clockwise encirclements around the origin can be found and closed loop stability is determined using $Z = N - P$ from the criterion explained earlier.

The Nyquist plots corresponding to the transfer functions given in examples (i), (ii) and (iii) respectively are illustrated.



It can be verified that the number of anti-clockwise encirclements counted from each of the plots match with the value of N obtained through analysis.

For Nyquist plots that go to ∞ for some values of ω, in order to find the line asymptotic to the plot, split the transfer function to real and imaginary parts and then substitute of ω. This is because using the magnitude directly will not give information regarding which part is causing the transfer function to go to ∞.

For example, consider $G(s) = \frac{1}{s(s+1)(s+2)}$. $|G(j\omega)| \to \infty$ at $\omega = 0$.

$$G(\jmath\omega) = \frac{1}{j\omega(j\omega+1)(j\omega+2)} = \frac{-j(-j\omega+1)(-j\omega+2)}{\omega(j\omega+1)(j\omega+2)(-j\omega+1)(-j\omega+2)}$$

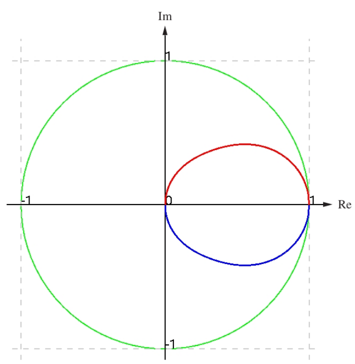
$$G(j\omega) = \frac{-3}{(\omega^2+1)(\omega^2+4)} + j\frac{-2+\omega^2}{(\omega^2+1)(\omega^2+4)}$$

$\implies$ $\mathrm{Re}[G(j\omega)] \to -3/4$ when $\omega \to 0$ and hence the plot is asymptotic to $\mathrm{Re}[G(j\omega)] = -3/4$ at $\omega = 0$.

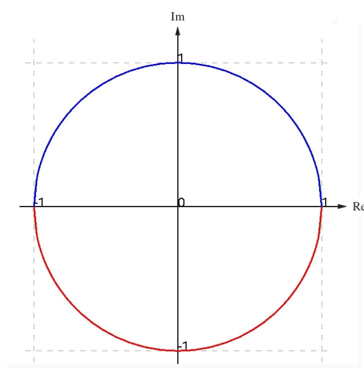
Observations regarding Nyquist plots of some standard systems:

Systems having neither poles nor zeros in the right side of s-plane are called "Minimum phase systems". Minimum phase systems will have Nyquist plot only on right side of GH-plane.

Systems with poles in left side of s-plane and zeros in right side of s-plane that are symmetric about $j\omega$ axis are called "All pass systems". All pass systems will have symmetric Nyquist plot with respect to both real and imaginary axes of GH-plane.



Minimum Phase System



All Pass System

A strictly proper transfer function implies the number of poles is greater than the number of zeros. Hence, the Nyquist plot of a system having a strictly proper transfer function will go to 0 as $\omega \to \infty$.

4.2.3 Bode Plot

Bode plot is another tool used to analyse stability of a closed loop system in the frequency domain. It actually consists of two plots, namely the Magnitude Bode plot and the Phase Bode plot.
Magnitude Bode plot is drawn between $20\log_{10}|G(j\omega)|$ (i.e magnitude in dB) against ω in log scale. Phase Bode plot is drawn between $\angle|G(j\omega)|$ (i.e phase in degrees) against ω in log scale.

The gain and phase margins of the closed loop system can be obtained by the two graphs and hence stability can be analysed.

Construction of Bode plot:
The magnitude and phase plot construction is made simple by taking different individual building blocks that appear in a system and analysing their effects.

1. Constant gain: $G(s) = K$
   $|G(j\omega)| = 20\log_{10} K = C\ dB$ $\rightarrow$ Constant line in magnitude plot
   $\angle G(j\omega) = \tan^{-1}(0/K) = 0^o$ $\rightarrow$ No effect in phase plot

2. Poles and Zeros at origin: $G(s) = s^{\pm n}$
   $|G(j\omega)| = 20\log_{10}(j\omega)^{\pm n} = \pm n\ 20\ dB$
   $\angle G(j\omega) = \pm n \tan^{-1}\omega/0 = \pm n\ 90^o$

First order pole	$-20\ dB/dec$	-90^o
Second order pole	$-40\ dB/dec$	-180^o
First order zero	$20\ dB/dec$	90^o
Second order zero	$40\ dB/dec$	180^o

   For a type 'n' system $G(s) = \frac{k}{s^n}$, Bode magnitude plot cuts the ω-axis at $\omega = k^{1/n}$

3. First order factors : $G(s) = (1 + s\tau)^{\pm n}$
   $$|G(j\omega)| = \pm n\ 20\log_{10}(1 + j\omega\tau)^{\pm n} = \begin{cases} \pm n\ 20\ dB/dec & : \omega\tau > 1 \\ 0 & : \omega\tau < 1 \end{cases}$$
   $$\angle G(j\omega) = \pm n \tan^{-1}\omega\tau = \begin{cases} \pm 0^o & : \omega\tau \rightarrow 0 \\ \pm n\ 45^o & : \omega\tau = 1 \\ \pm n\ 90^o & : \omega\tau \rightarrow \infty \end{cases}$$
   where $\omega = 1/\tau$ is Corner Frequency.

4. Second order factors : $G(s) = (s^2 + 2\zeta\omega_n s + \omega_n^2)^{\pm n}$

   $|G(j\omega)| = \pm 20\, n\, \log_{10}(\sqrt{(1-(\frac{\omega}{\omega_n})^2)^2 + (\frac{2\zeta\omega}{\omega_n})^2} = \begin{cases} \pm n\ 40\ dB/dec & : \omega > \omega_n \\ 0 & : \omega < \omega_n \end{cases}$

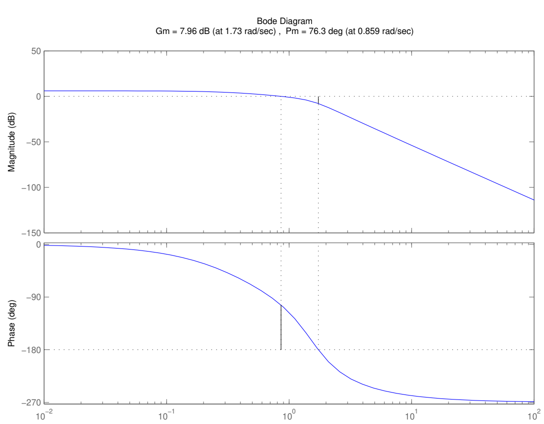
   $\angle G(j\omega) \approx \pm n \tan^{-1} \frac{\omega}{\omega_n} = \begin{cases} \pm 0^o & : \omega \to 0 \\ \pm n\ 90^o & : \omega = \omega_n \\ \pm n\ 180^o & : \omega \to \infty \end{cases}$

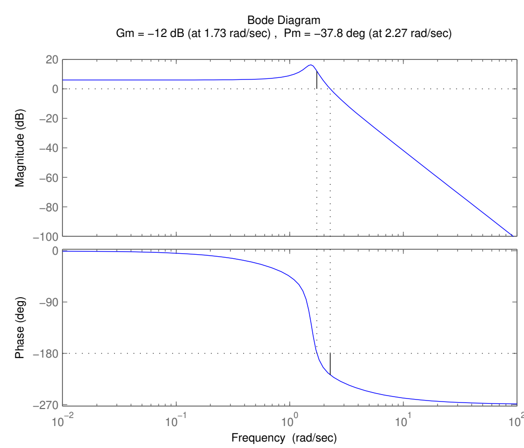
   where $\omega = \omega_n$ is Corner Frequency.

Note that the above methods are to obtain approximate Bode plot. The approximations will not hold at the corner frequencies and around it. If calculated, at the corner frequency, the magnitude will be $3n\ dB$ more or less than the approximated magnitude.
And also, Bode plot analysis is generally applicable only to Minimum phase systems and where the open loop system is known to be stable.

Stability analysis using Bode Plot:
The graphs below indicate the bode magnitude and phase plots of stable and unstable systems respectively.

- Stable systems will have positive values of GM & PM as $\omega_{gc} < \omega_{pc}$.
- Unstable systems will have negative values of GM & PM as $\omega_{gc} > \omega_{pc}$.





(for marginally stable systems, the crossover frequency will be same, hence there will be no margins i.e $GM = 0\ dB$ and $PM = 0^o$)

Obtaining transfer function from given Bode Plot:
For a minimum phase system, the transfer function can be uniquely determined from the magnitude curve alone.

The following steps are to be followed to estimate the minimum phase transfer function from a given magnitude Bode plot.

- Observe starting slope: Will give information regarding poles or zeros at origin.
- Observe change in slope at each corner frequency: Will give information about 1^{st} and 2^{nd} order factors. Express in time constant form.
- Find DC gain i.e the constant k using $Y = mX + C$ where Y is the magnitude (in dB) at the frequency X, m is the slope of the curve at the frequency X and C is the intercept on the magnitude axis. $\implies \quad C = 20\log_{10}(k)$.

Bode plots and Nyquist plots are extremely useful in designing of controllers and compensators using frequency domain representation.

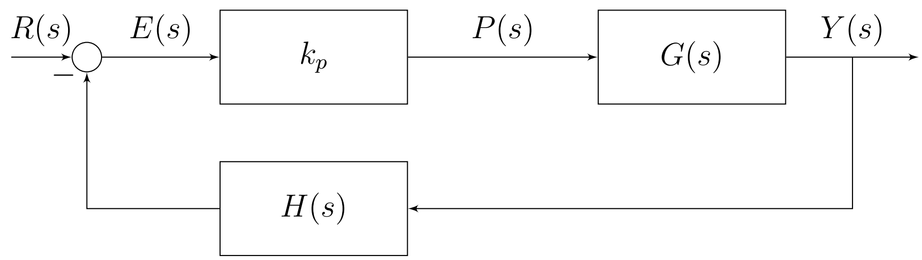
5 Industrial Systems

As studied in the introduction, the plant by itself will not give desired response. It is to be controlled by another system which gives signal to the plant based on command input and actual output (i.e error signal). These systems that modify the response of the plant to desired response based on the error signal are either called controllers or compensators.

5.1 Controllers

Controllers are systems that are cascaded before the plant in order to use the error signal and modify the action of the plant to satisfy design requirements such as steady state error, rise time, settling time, maximum peak overshoot, etc.
There are really 4 types of controllers that are used in practical systems, but theoretically there can be 6 types which are all discussed in detail.

5.1.1 Proportional Controller



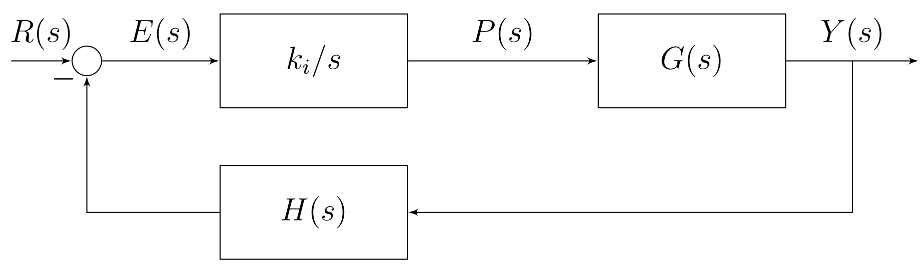
The simplest controller, where the plant response is modified by changing it's open loop gain.
Hence, the controller output is proportional to the error signal.

$$P(s) = k_p E(s)$$

$$p(t) = k_p e(t)$$

In majority of the practical systems, controlling the system gain alone will not be able to meet all design requirements. There will be a trade-off between speed and accuracy of the response.
When the error signal eventually reaches 0, then the control signal also becomes 0, which means there is nothing to drive the plant to stay in that state and steady state output of the system will not be same as desired output. In such a case, the gain can be increased to reduce the steady state error, but it will never go to 0. (note that for systems that don't need any power to stay at the steady state value, proportional control will work).
Also, increasing the gain to solve this problem can lead to instability of the closed loop system.

5.1.2 Integral Controller



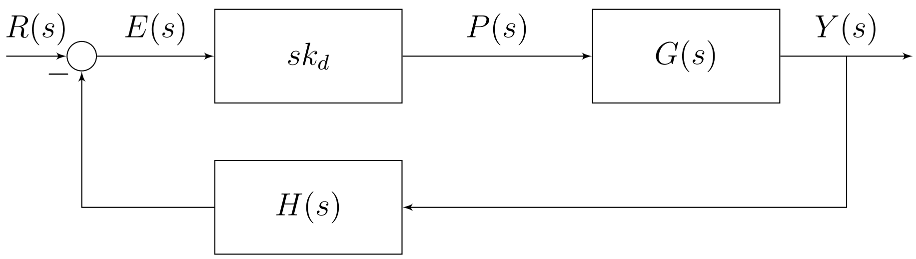
The integral controller is used to increase the type of a system, which will eliminate steady state error as it adds an additional pole. Hence, the controller output is proportional to the sum of accumulated error signals i.e integral of the error signal.

$$P(s) = \frac{k_i}{s}E(s)$$

$$p(t) = k_i \int_{\tau=0}^{t} e(\tau)d\tau$$

However, adding a pole to the system will adversely affect stability.

5.1.3 Derivative Controller



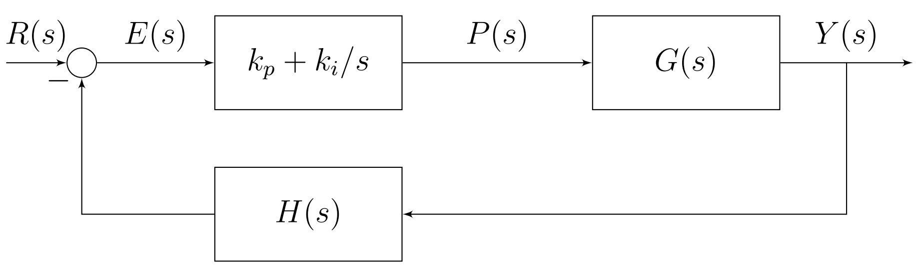
The differential controller is used to make the system transient faster and better since it adds a zero. Hence, the controller output is proportional to the derivative of the error signal.

$$P(s) = sk_dE(s)$$

$$p(t) = k_d\frac{de(t)}{dt}$$

However, adding a zero to the system will adversely affect steady state accuracy, and also a pure differentiator is not practically realizable.

5.1.4 PI Controller

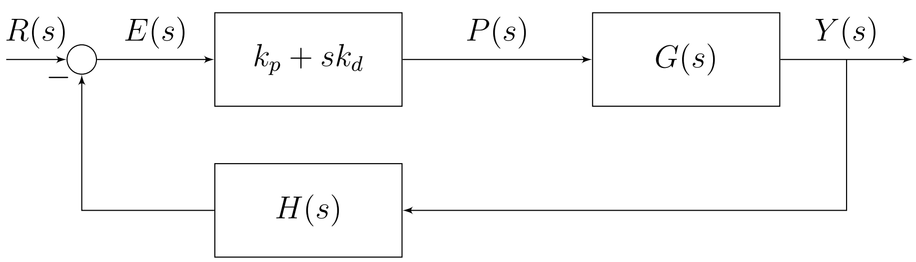


As discussed, a drawback of P controller is that if the error signal is 0, then the control signal is 0 and there is no input to the plant to drive the system. To avoid this, the control signal is made to be dependent on the accumulation of error signal along with only the error signal, so the plant will never get zero input. This will cause the plant to be driven exactly to the required steady state and also ensure it has the energy to maintain it.
Hence, PI controller will effectively improve the steady state response of the plant and is used when P controller alone would give some steady state error for any value of the gain.

$$P(s) = \left(k_p + \frac{k_i}{s}\right)E(s)$$

$$p(t) = k_p e(t) + k_i \int_{\tau=0}^{t} e(\tau)d\tau$$

5.1.5 PD Controller

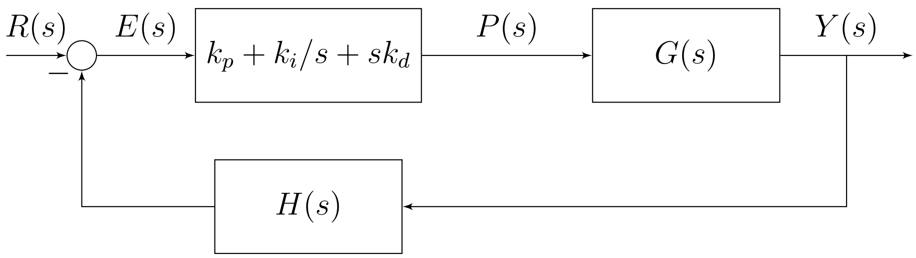


P controller does not take into account the changes in the error signal which should actually have an impact on the plant. For example, if the error is decreasing too fast, then the plant needs to slow down to avoid overshoot of the steady state value. Hence PD controller is used, to provide better transient response. Note that PD controller also prevents any potential instability issues due to high gain of P controller.
It should be noted that due to derivative action, high frequency or high amplitude signals can be introduced into the system, which have to be filtered out or limited.

$$P(s) = \big(k_p + sk_d\big)E(s)$$

$$p(t) = k_p e(t) + k_d \frac{de(t)}{dt}$$

5.1.6 PID Controller



$$P(s) = \left(k_p + \frac{k_i}{s} + sk_d\right) E(s)$$

$$p(t) = k_p e(t) + k_i \int_{\tau=0}^{t} e(\tau) d\tau + k_d \frac{de(t)}{dt}$$

The PID controller is the optimal controller which generates control signal proportional to the error, its integral and its derivative. This controller can almost always be used to ensure both steady state and transient response requirements are met since the derivative part ensures gives higher speed and the integral part ensures low steady state error, while stability is also maintained.

Effect of increasing k_p, k_i and k_d in PID controller:

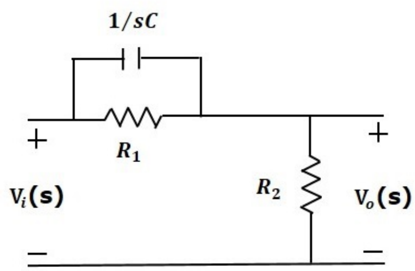
Parameter	Peak overshoot	Settling time	Steady state error
k_p	Increases	Low impact	Decreases
k_i	Increases	Increases	Eliminated
k_d	Decreases	Decreases	No impact

5.2 Compensators

Compensators are more practical systems used to modify the behaviour of the plant to satisfy the design specifications. Compensators are systems that add extra poles and zeros in appropriate points in s-plane to get the desired behaviour.

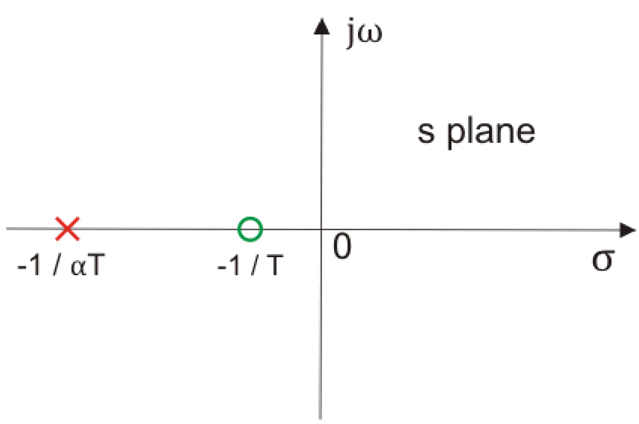
5.2.1 Lead Compensator

The lead compensator is an electrical network which produces an output having phase lead for the input applied. The lead compensator circuit in the 's' domain is shown.

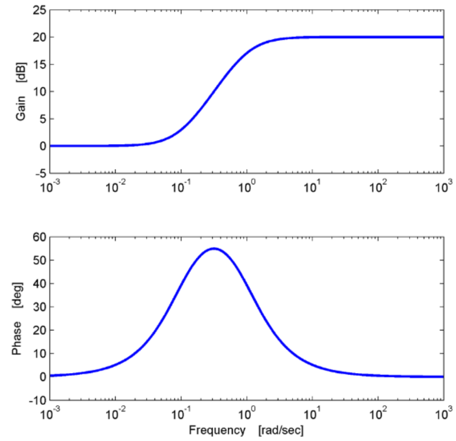


$$\boxed{G_c(s) = \frac{\alpha[1+Ts]}{1+\alpha Ts}} \qquad \alpha = \frac{R_2}{R_1+R_2}; \quad T = R_1C$$

Pole at $s = -1/\alpha T$ and Zero at $s = -1/T$ Since $\alpha < 1$, the zero is closer to the origin than the pole.



The lead compensator provides slope of +20 dB/decade for the magnitude response and a positive phase for the phase response of the uncompensated system.



Maximum phase lead provided by the lead compensator is given by

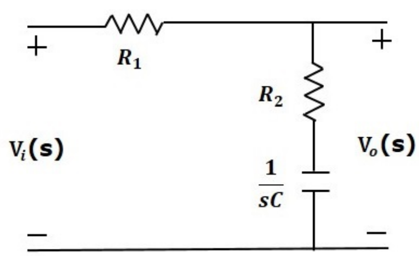
$$\theta_m = \sin^{-1}\left[\frac{1-\alpha}{1+\alpha}\right]$$

The frequency at which maximum phase occurs is given by the geometric mean of the corner frequencies.

$$\omega_m = \frac{1}{T\sqrt{\alpha}}$$

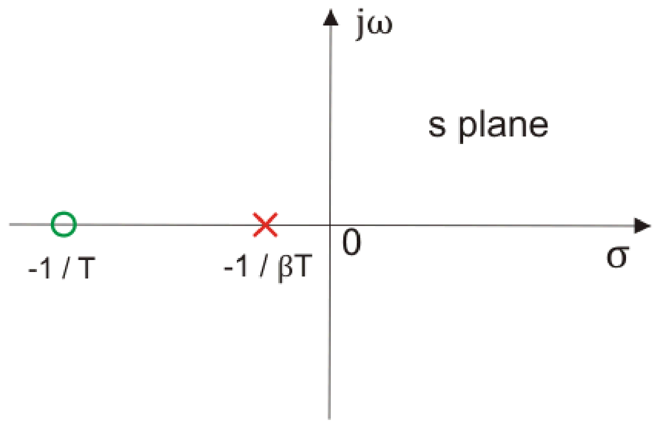
5.2.2 Lag Compensator

The lag compensator is an electrical network which produces an output having the phase lag for the input applied. The lag compensator circuit in the 's' domain is shown.

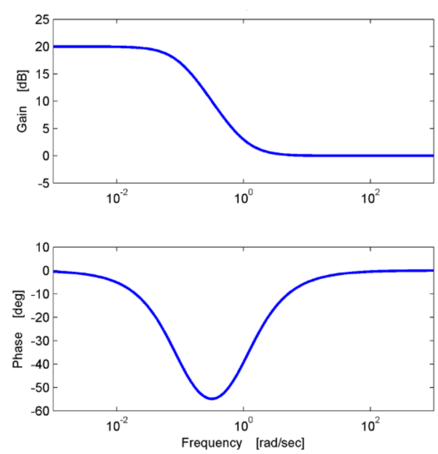


$$\boxed{G_c(s) = \frac{[1+Ts]}{1+\beta Ts}} \qquad \beta = \frac{R_1+R_2}{R_2}; \quad T = R_2C$$

Pole at $s = -1/\beta T$ and Zero at $s = -1/T$ Since $\beta > 1$, the pole is closer to the origin than the zero.



The lag compensator provides slope of -20 dB/decade for the magnitude response and a negative phase for the phase response of the uncompensated system.



Maximum phase lag provided by the lag compensator is given by

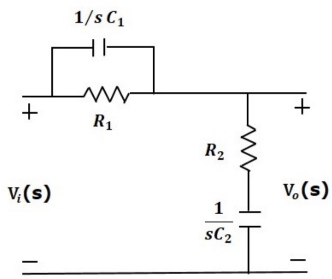
$$\theta_m = \sin^{-1}\left[\frac{1-\beta}{1+\beta}\right]$$

The frequency at which minimum phase occurs is given by the geometric mean of the corner frequencies.

$$\omega_m = \frac{1}{T\sqrt{\beta}}$$

5.2.3 Lag-Lead Compensator

Lag-Lead compensator is an electrical network which produces phase lag at one frequency region and phase lead at other frequency region. It is a combination of both the lag and the lead compensators. The lag-lead compensator circuit in the 's' domain is shown.

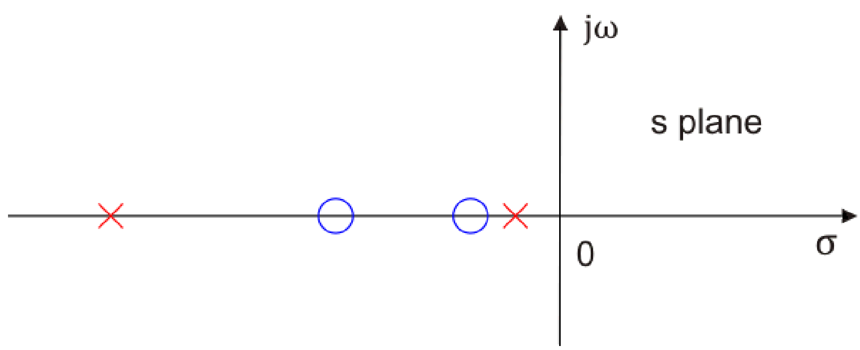


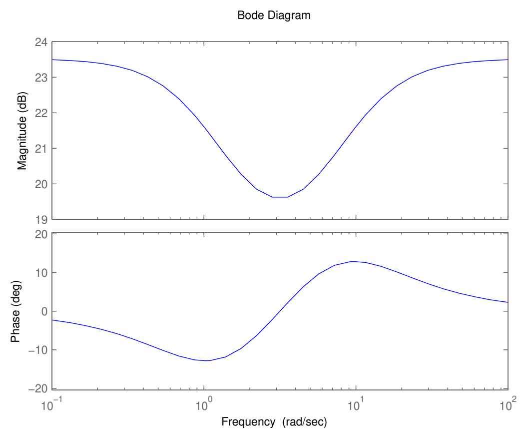
$$\boxed{G_c(s) = \frac{\alpha[1+T_1 s][1+T_2 s]}{[1+\alpha T_1 s][1+\beta T_2 s]}}$$

$$\alpha = \frac{R_2}{R_1+R_2}; \quad T_1 = R_1 C; \qquad \beta = \frac{R_1+R_2}{R_2}; \quad T_2 = R_2 C$$

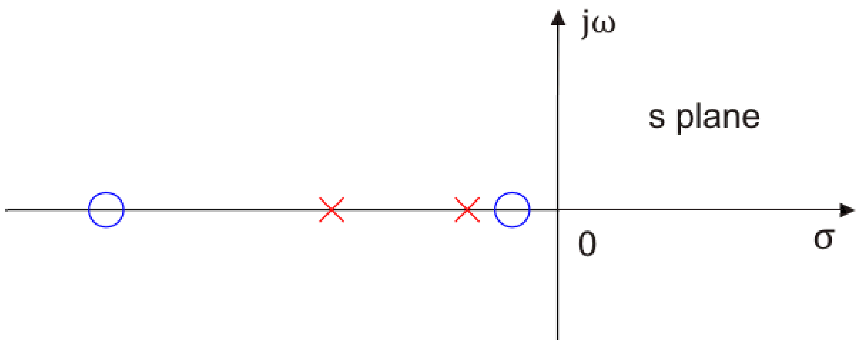
The lag-lead compensator can work both ways i.e provide phase lag first and then phase lead or vice versa.

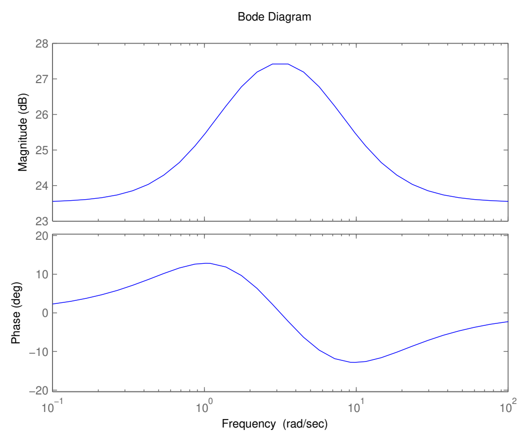
If the lag compensator pair is closer to origin, then the compensator behaves like a band stop filter as shown in the magnitude Bode plot.





If the lead compensator pair is closer to origin, then the compensator behaves like a band pass filter as shown in the magnitude Bode plot.





6 State Space Analysis

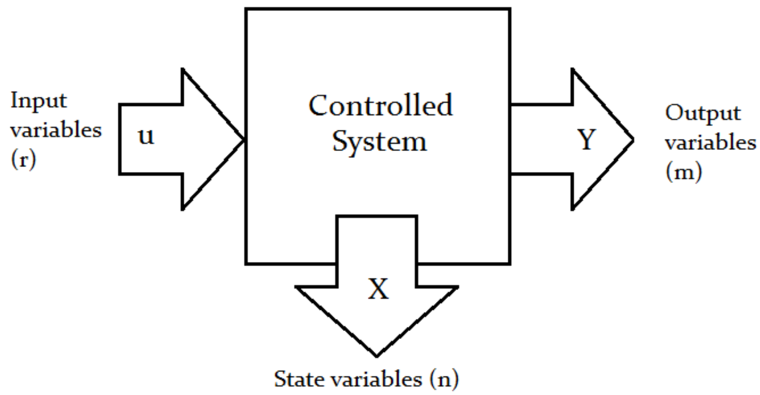
Disadvantages of classical control theory:

1. Applicable only to linear time invariant systems.
2. Applicable only if initial conditions are zero.
3. Design of systems with multiple inputs and multiple outputs becomes cumbersome.
4. Does not consider internal states of the system, considers only input and output.
5. Higher order system analysis becomes cumbersome.

Advantages of modern control theory:

1. Applicable to linear and non-linear systems.
2. Applicable to time invariant and time varying systems.
3. Applicable even if initial conditions are non-zero.
4. Design of systems with multiple inputs and multiple outputs is easier.
5. Any higher order system can be represented using first order equations only.
6. Considers internal states of the system, along with input and output.

The state of a system corresponds to the minimum set of variables such that the knowledge of these variables at time $t = t_o$ and input at time $t \geq t_o$ completely determines the behaviour of the system for any time $t \geq t_o$.
These variables are called "State Variables". Since a system will have multiple state variables (say 'n'), then the vector consisting of all the n state variables is called the "State Vector". The vector space consisting of all possible states of a system is called it's "State Space".



The above system consists of r input variables $(u_1,\ u_2,\ ....u_r)$, m output variables $(y_1,\ y_2,\ ....y_m)$ and n state variables $(x_1,\ x_2,\ ....x_n)$ all being functions of time.

The state variable representation of such a system is given by n first order differential equations.
$\dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t), t)$ for time varying systems
$\dot{\vec{x}}(t) = f(\vec{x}(t), \vec{u}(t))$ for time invariant systems
where $\dot{\vec{x}}(t)$ is the vector of first derivatives of the state variables.

The outputs are also represented as functions of the states and inputs.
$\vec{y}(t) = f(\vec{x}(t), \vec{u}(t), t)$ for time varying systems
$\vec{y}(t) = f(\vec{x}(t), \vec{u}(t))$ for time invariant systems

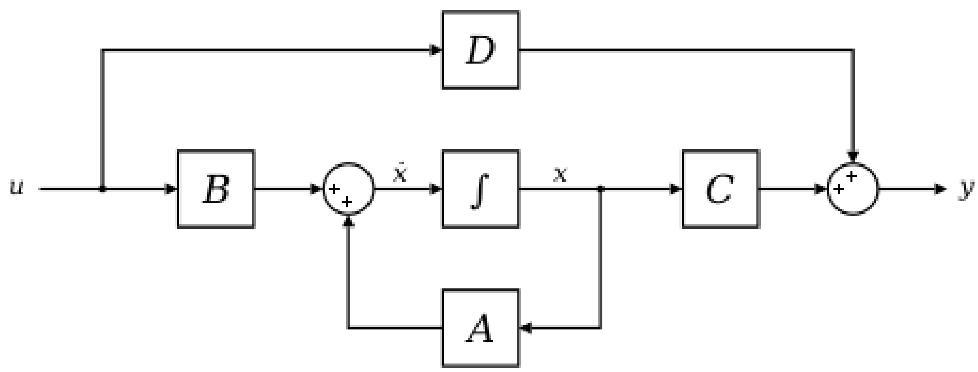
For linear time invariant systems, due to property of superposition and homogeneity, it can be found that the state equations and output equations can be represented as follows.

$$\dot{\vec{x}}(t) = A\vec{x}(t) + B\vec{u}(t)$$

$$\vec{y}(t) = C\vec{x}(t) + D\vec{u}(t)$$

$A \rightarrow$ System matrix; $\quad B \rightarrow$ Input matrix;
$C \rightarrow$ Output matrix; $\quad D \rightarrow$ Transmission matrix;

Block diagram representation



Selection of state variables

- The number of state variables must be minimized.
- The state variables must be linearly independent.
- In a practical system, the number of state variables should be equal to the number of independent energy storage elements.
- The number of state variables will be equal to the order of the differential equations and equal to the number of integrators used.

6.1 State Model Formation

State model of a system can be obtained from general models such as differential equation, transfer function, state diagram (block diagram or signal flow graph).

6.1.1 State model from differential equations

Most physical systems can be expressed in terms of differential equations as mentioned earlier. The state model can be obtained using these equations. Consider an SISO system defined by a second order differential equation

$$\frac{d^2y(t)}{dt^2} + a\frac{dy(t)}{dt} + by(t) = ku(t)$$

where $y(t)$ is the output and $u(t)$ is the input.
Since order of the equation is 2, this system has to be defined using 2 state

variables.
Take $y(t) = x_1(t)$, then
$\frac{dy(t)}{dt} = \dot{x_1}(t) = x_2(t)$ and
$\frac{d^2y(t)}{dt^2} = \dot{x_2}(t) = ku(t) - ax_2(t) - bx_1(t)$
where $x_1(t)$ and $x_2(t)$ are the state variables.

From this, the state equations and output equations can be expressed in matrix form to obtain the 4 defining matrices.
$\begin{bmatrix} \dot{x_1}(t) \\ \dot{x_2}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ k \end{bmatrix} u(t)$
$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + 0u(t)$
Same concept can be applied to MIMO systems of higher order. Any order differential equation with any number of inputs and outputs can be represented in the same manner.

6.1.2 State model from transfer function

Transfer function of a system will be of the form $G(s) = Y(s)/U(s)$.
This can be expressed as $G(s) = \frac{Y(s)X(s)}{X(s)U(s)}$
where $\frac{Y(s)}{X(s)} = n(s)$ i.e numerator polynomial and $\frac{X(s)}{U(s)} = \frac{1}{d(s)}$ i.e denominator polynomial.
$Y(s) = n(s)X(s)$ and $U(s) = d(s)X(s)$.
By taking inverse Laplace transform of the above two expressions, differential equations are obtained, and that can be used to obtain state model from the method explained earlier.

For example,

$$G(s) = \frac{Y(s)}{U(s)} = \frac{s+l}{s^2+cs+d}$$

$$\implies Y(s) = (s+l)X(s); \quad (s^2+cs+d)X(s) = U(s)$$

$$y(t) = \frac{dx(t)}{dt} + px(t)$$

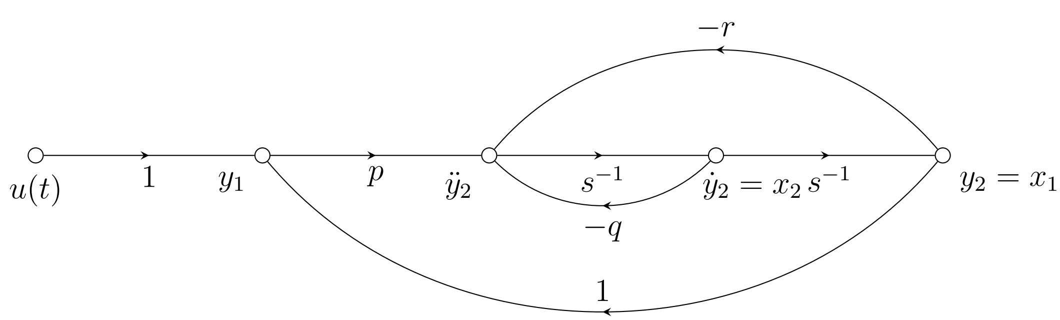
$$\frac{dx^2(t)}{dt^2} + c\frac{dx(t)}{dt} + dx(t) = u(t)$$

The state variables are $x_1(t) = x(t)$ and $x_2(t) = \frac{dx(t)}{dt}$

These equations expressed in matrix will give the 4 defining matrices.
$$\begin{bmatrix} \dot{x_1}(t) \\ \dot{x_2}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -d & -c \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$
$$y(t) = \begin{bmatrix} l & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + 0u(t)$$
For MIMO systems, a unique transfer function will be defined for each combination of input and output. Similar exercise for each of those transfer functions provides overall state model.

6.1.3 State model from state diagram

Using the signal flow graph or block diagram, the transfer function can be found, which can be used to obtain the state model as explained.
Another technique would be to assign state variables in the diagram and get the state equations. Here, the number of integrators will give the number of state variables. The outputs of integrators are assigned as state variables and equations are obtained from incoming branches at each node.



In the above state diagram (SFG) there are 2 integrators, whose outputs are taken to be the state variables.
State equations:
$\dot{x_1}(t) = x_2(t)$
$\dot{x_2}(t) = (p - q)x_1(t) - rx_2(t) + pu(t)$

Output equation:
$y_2(t) = x_1(t)$

The above equations expressed in matrix will give the 4 defining matrices.
$\begin{bmatrix} \dot{x_1}(t) \\ \dot{x_2}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ (p-q) & -r \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ p \end{bmatrix} u(t)$
$y_2(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + 0u(t)$

6.2 Transfer Function from State Model

Consider a system with the state equations $\dot{x} = Ax + Bu$ and output equation being $y = Cx + Du$ (where y is a single output; since different outputs will have different transfer functions).
Taking Laplace Transform, $sX(s) = AX(s) + Bu(s)$ and $Y(s) = CX(s) = Du(s)$.
$\implies X(s) = [SI - A]^{-1}Bu(s)$

$$\therefore \boxed{\frac{Y(s)}{U(s)} = C[SI - A]^{-1}B + D}$$

Characteristic equation of the system will be given by $|SI - A| = 0$ since the determinant of that matrix is to be divided why finding the inverse. $|SI - A| = 0$ will give the locations of the poles, which determine the stability and characteristic behaviour of the system.

6.3 Controllability and Observability

Controllabability of a system:
A system is said to be completely state controllable if it is possible to transfer the system from any initial state $x(t_o)$ to any desired state $x(t)$ in specified finite time by a control vector.

A general n^{th} order MIMO LTI system with state equations defined by $\dot{x} = Ax + Bu$ and output equations defined by $y = Cx + Du$ is controllable if and only if the rank of the matrix Q_c is n i.e determinant of the matrix Q_c is non-zero.

$$Q_c = [B : AB : .... : A^{n-1}B]$$

Observability of a system:
A system is said to be completely state observable if every state $x(t_o)$ can be completely identified by measurements of the outputs $y(t)$ over a finite time interval.

A general n^{th} order MIMO LTI system with state equations defined by $\dot{x} = Ax + Bu$ and output equations defined by $y = Cx + Du$ is observable if and only if the rank of the matrix Q_o is n i.e determinant of the matrix Q_o is non-zero.

$$Q_o = [C^T : A^T C^T : .... : (A^T)^{n-1} C^T]$$

6.4 Solution to State Equations

The state equations are expressed as $\dot{x(t)} = Ax(t) + Bu(t)$.
The overall response of the system is given by the sum of free response and forced response.
Free response is due to the initial conditions of the system and forced response is due to the input provided.

Free Response (or zero input response)

$$x(t) = L^{-1}([SI - A]^{-1})x(0)$$

where $x(0)$ gives the initial conditions
Forced Response (or zero state response)

$$x(t) = L^{-1}([SI - A]^{-1})Bu(t)$$

Overall Response

$$\therefore x(t) = L^{-1}([SI - A]^{-1})[x(0) + Bu(t)]$$

$L^{-1}([SI - A]^{-1}) = e^{At} = \phi(t)$
$\phi(t)$ is called **State transition matrix**.

- $\phi(0) = I$
- $\phi'(0) = AI$
- $\phi(t_1 + t_2) = \phi(t_1)\phi(t_2)$
- $\phi^{-1}(t) = \phi(-t)$

6.4.1 State space transformations

A state space transformation can be obtained using a linear transformation which links the old state vector $\vec{x}(t)$ with the new vector $\vec{X}(t)$.
$\vec{X}(t) = T\vec{x}(t)$.
Applying this to the state equations and output equations, a new equivalent set of equations are obtained.

$$\dot{\vec{X}}(t) = A_1\vec{X}(t) + B_1\vec{u}(t)$$

$$\vec{y}(t) = C_1\vec{x}(t) + D\vec{u}(t)$$

Where $A_1 = T^{-1}AT$; $B_1 = T^{-1}B$; $C_1 = CT$

By properly choosing matrix T, it is possible to obtain mathematical descriptions of the given system characterized by matrices A_1, B_1, C_1 and D which have particularly simple structures.

For each choice of matrix T, a different but equivalent mathematical description of the given system is obtained. All these different mathematical models maintain the basic physical properties of the given dynamic system: stability, controllability and observability.

If the matrix T can be selected such that $A_1 = T^{-1}AT$ is a diagonal matrix, then T is represented as S and is called the diagonalizing matrix and A_1 is now written as Λ.

6.4.2 Eigenvalues and Eigenvectors

For any system matrix A, there exists some vectors $\bar{x}$ that satisfies $A\bar{x} = \lambda\bar{x}$ where λ is a scaling factor i.e multiplying the vector with the matrix will give the scaled version of the same vector.
These vectors are called **Eigenvectors** and their respective scaling factors are called **Eigenvalues**.

This gives $(A - \lambda I)\bar{x} = \bar{0}$.
$\implies |A - \lambda I| = 0$
The above equation gives the characteristic equation of the given matrix and it is used to find the eigenvalues of A.

The matrix Λ obtained by diagonalizing the system matrix A is the eigenvalue matrix which consists of the eigenvalues as its principle diagonal elements. The matrix that diagonalizes A i.e the matrix S is formed by placing the eigenvectors of A as columns of the matrix. $S = [\bar{x_1}\ \bar{x_2}\ ...\ \bar{x_n}]$

The eigenvalues and eigenvectors provide an alternate method to find the state transition matrix.

$$e^{At} = S \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{\lambda_n t} \end{bmatrix} S^{-1}$$