Basics of Linear Control Theory for Mechanical Engineering

ir. J.G. Gruijters

August the 18th, 2018

Abstract

This article is made to show the control theory and how it is applied to mechanical systems. This article starts with a very simple mass-spring-damper system. The equations of motion and the Laplace transformation to the frequency domain are shown. With the equations of motion in the frequency domain, controllers are added to a simple mechanical system, including a method of how to set the controllers in the field. Although controllers can be set it in the field, it is common practice and often necessary that the controller is designed by an engineer, to optimise the behaviour or to make an instable mechanical system stable. The controller design for simple systems is possible by using the Bode plot and the Bode stability criterion. For more complex (non-minimum phase) systems the Nyquist criterion should be used to check stability. At last some filters are shown which can be used to optimise the control loop or suppress measurement noise or disturbances.

Drives, like electrical and hydraulic drives, are used to move parts and masses on operator or control sequence command. The 'on command' part is done by a control system, which sends signals to the electric system or hydraulic valves, which in turn control the drives.

The control part is sometimes underexposed by the mechanical engineer. This is however often possible, as the control is straight forward on-off technology. This means that the system is either activated or not activated. Fortunately, the control becomes more interesting for other systems, for instance for frequency drive, proportional and servo valve control.

This paper aims to explain the simple control theory based on simple (theoretical) systems. This knowledge can then also be used for more complex systems.

1 Equations of motion

When looking at mechanical systems, it is logical to start with the mechanics: the equations of motion. The equations of motion describe the motion response of the mechanical system to forces. This might sound difficult, but Newton's second law ($F = m\,a$) is an example of an equation of motion. The equations of motion describe the behaviour of a system using a mathematical representation. This representation is in the form of (a) differential equation(s), meaning that there are (time) derivatives of the variables present in the equation. The most obvious example of time derivatives are position, speed and acceleration, which leads back to Newton's second law.

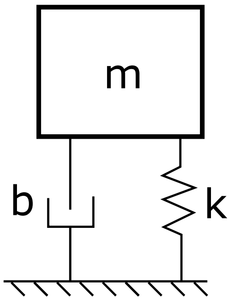

Figure 1: A simple mechanical system

One of the more simple mechanical systems is a mass-spring-damper system, which is shown in figure 1. It does not matter whether the theory of Newton is used or another method, like the method of Lagrange. The resulting equations of motion should be the same, as the system is the same. For the example of figure 1, the equations of motion become:

$$m\,\ddot{y} + b\,\dot{y} + k\,y = 0 \tag{1}$$

In equation 1 the $\ddot{y}$ is the acceleration, $\dot{y}$ the velocity and y is the position, where position $y = 0$ refers to the spring in rest position (spring delivers no force in $y = 0$).

1.1 Time response

Now that the equation of motion is known, it is possible to predict the system behaviour. One possibility is a time simulation and see how the system varies over time. For the example of equation 1 the boundary condition is for instance a start position, velocity and acceleration, assuming the mass, spring stiffness and damping coefficient are known. In Figure 2 examples of a response are shown, where only the damping is varied between the different responses.

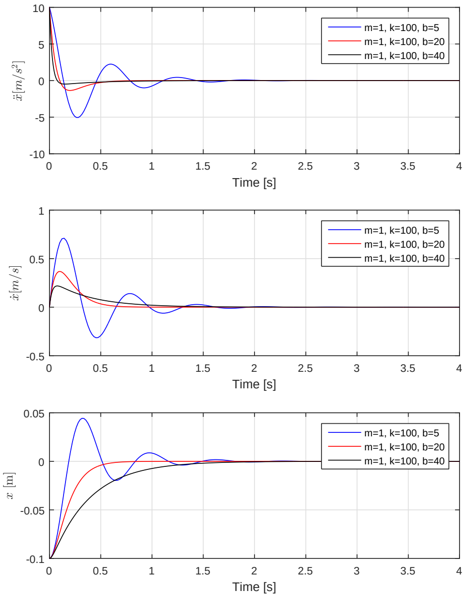

Figure 2: Possible time responses of the simple mass spring damper system. The mass is 1kg, the spring is 100N/m and the damping is varied between 40, 20 and 5 Ns/m

Of the three responses shown in Figure 2, only the blue shows a periodic response until it is damped out. This periodic response should be familiar for all mechanical engineers: the natural frequency or resonance frequency. In mechanics this type of periodic response, with an overshoot and damping out over time, is called an underdamped system.

The other two responses (red and black lines) show no periodic responses, because the damping is too large to allow the system to start resonating. The red line requires less time to go to the position of 0meter, while the black line requires more time (see the bottom graph of figure 2). The difference is only caused by a different damping coefficient. The difference can be explained when the analytical solution of equation 1 is found. This is done by applying:

$$\begin{aligned} y &= e^{\lambda t} \\ \dot{y} &= \lambda e^{\lambda t} \\ \ddot{y} &= \lambda^2 e^{\lambda t} \end{aligned} \tag{2}$$

When equation 2 is used in equation 1, the following is found:

$$m\lambda^2 + b\lambda + k = 0 \tag{3}$$

When the quadratic formula is used, the following two answers are found:

$$\lambda_1 = \frac{-b + \sqrt{b^2 - 4\,m\,k}}{2\,m} \tag{4}$$

$$\lambda_2 = \frac{-b - \sqrt{b^2 - 4\,m\,k}}{2\,m} \tag{5}$$

These values can then be used to calculate the general solution to the differential equation.

The λ_1 and λ_2 are both dependent on the square root of a negative, zero or positive number. This means that λ_1 and λ_2 are either both a complex number[1] or both a real number. Using the Euler's formula $e^{i\,\varphi} = cos(\varphi) + i\,sin(\varphi)$, the complex power to Euler's number (e) can be converted to the sum of a real and imaginairy number, making it possible to plot the real and complex numbers on the complex plane. This is shown in figure 3. Using the possible solutions of λ_1 and λ_2, the following three different cases are distinguished:

[1] These complex numbers exist of a real and imaginary part, where $i = \sqrt{-1}$ denotes the imaginary part.

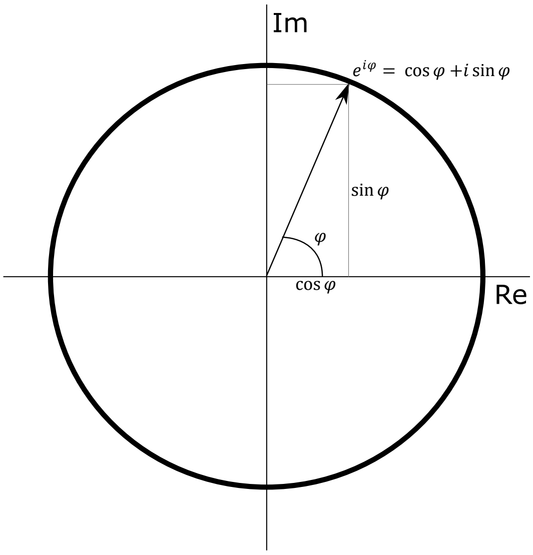

Figure 3: The complex plane, with the vertical axis being the imaginairy axis and the horizontal the real axis. Euler's formula for complex numbers is shown as well in this picture.

- Underdamping: λ_1 and λ_2 are both complex numbers, as $b^2 < 4\,m\,k$. This leads to a general solution of $y = e^{\lambda\,t}\,(C_1\,cos(\omega\,t) + C_2\,sin(\omega\,t))$ for $\lambda = k \pm i\,\omega$.

- Critical damping: λ_1 and λ_2 are equal, $\lambda_{1,2} = \frac{-b}{2\,m}$ as $b^2 = 4\,m\,k$, which leads to the general solution of $y = (C_1 + C_2\,t)\,e^{\lambda\,t}$

- Overdamping: λ_1 and λ_2 are both real numbers, as $b^2 > 4\,m\,k$, which leads to the general solution of $y = C_1\,e^{\lambda_1\,t} + C_2\,e^{\lambda_2\,t}$

The different solutions to the differential equation will thus result in different behaviour. An underdamped system will show the overshoot and vibrating solution. The critical damping, which is the lowest damping without overshoot, requires the least amount of time to reach the steady state of the system. Overdamped systems will also not show any overshoot, but it takes more time to get to the steady state.

So for all equations of motion of mechanical systems, the critical damping can be calculated. When the actual damping is divided by the critical damping, the damping ratio ζ is obtained:

$$\zeta = \frac{b_{actual}}{b_{critical}} \tag{6}$$

The resonance frequency, or the natural frequency, will be calculated. The natural frequency of an undamped system can be calculated as follows:

$$\omega_n = \sqrt{\frac{k}{m}} \tag{7}$$

Most real mechanical systems however have some sort of damping. This damping will change the natural frequency. The damped natural frequency is a function of the undamped natural frequency. The damped natural frequency of an underdamped ($0 < \zeta < 1$) system is:

$$\omega_d = \omega_n\,\sqrt{1 - \zeta^2} \tag{8}$$

For the overdamped case ($\zeta > 1$) the damped natural frequency is:

$$\omega_d = \omega_n\,\sqrt{\zeta^2 - 1} \tag{9}$$

1.2 Frequency response

A lot can be learned from the time domain response of a mechanical system. The time domain shows the response of a mechanical system to a force. When the force is for instance controlled by hydraulics, the influence of the force is known. But to obtain a control loop and predict the behaviour dependent on the controller and mechanical system, the time domain is only of limited use. The method to get this description for linear systems is to make the transfer to the frequency domain, by using the Laplace transformation.

The frequency domain is an abstract idea. It is the transfer from the time on the horizontal axis as in figure 2 to the frequency. The basic idea is that the response in the time domain exists of the sum of several frequencies, each with their own amplitude. A good graphic representation is shown in figure 4. The transformation to the frequency domain is done using the Laplace transformation.

To comprehend: the transition to the frequency domain basically means that the time signal can be seen as the superposition (i.e. sum of) periodic functions with separate frequencies and amplitudes.The frequencies and amplitudes are then displayed in the graph as shown in figure 4.

For the readability of this article, the Laplace transformation itself will not be explained. Instead the Appendix A shows a table with some basic transformations to the frequency domain, which are sufficient for the purpose of this article. For this article the first two rows in the left column are important.

When applying the Laplace transformation to the simple mechanical system, it means that the derivative is exchanged with an s, while the second derivative is exchanged by the s^2. So the end result is:

$$m\,Y(s)\,s^2 + b\,Y(s)\,s + k\,Y(s) = 0 \tag{10}$$

It might be immediately clear that with this equation, the quadratic formula can be used which leads to the same results as shown for the time domain, but now the variable is s instead of λ.

Up till now the differential equation was equal to zero, so no input was defined or used. Now it is time to add this extra force, as shown in figure 5. The equation of motion and the Laplace transformation are then:

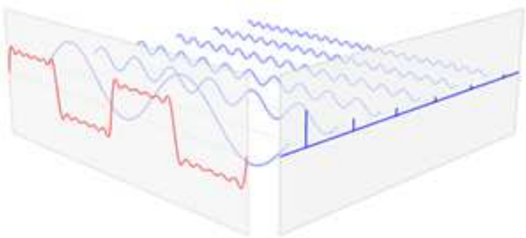

Figure 4: A graphic representation of the time domain (on the left in red), which can be split into periodic signals with several frequencies and amplitudes. In the frequency domain the frequencies are divided over the horizontal axis from low frequency on the left to high frequency on the right side, while the amplitude is shown on the vertical axis. See the animated file at ***this website***.

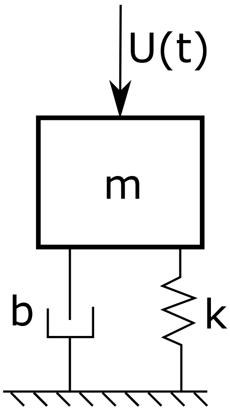

Figure 5: A simple mechanical system, now including a force to control the position of the mass

$$m\,\ddot{y} + b\,\dot{y} + k\,y = U(t) \tag{11}$$
$$m\,Y(s)\,s^2 + b\,Y(s)\,s + k\,Y(s) = U(s) \tag{12}$$

With the equation as in equation 12, the transfer function from input to output is defined:

$$\frac{Y(s)}{U(s)} = H(s) \tag{13}$$
$$H(s) = \frac{1}{m\,s^2 + b\,s + k} \tag{14}$$

The transfer function describes the linear behaviour of the input forces ($U(t)$ in this case) to the output (the position of the mass for this example). This transfer function is used by software and control engineers to model the mechanical system and check the system stability.

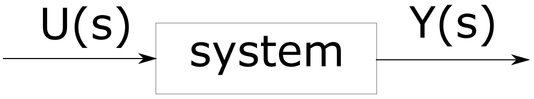

Figure 6: The input U(s) and output Y(s) to a system. The U(s) are for instance forces, the Y(s) are for instance positions.

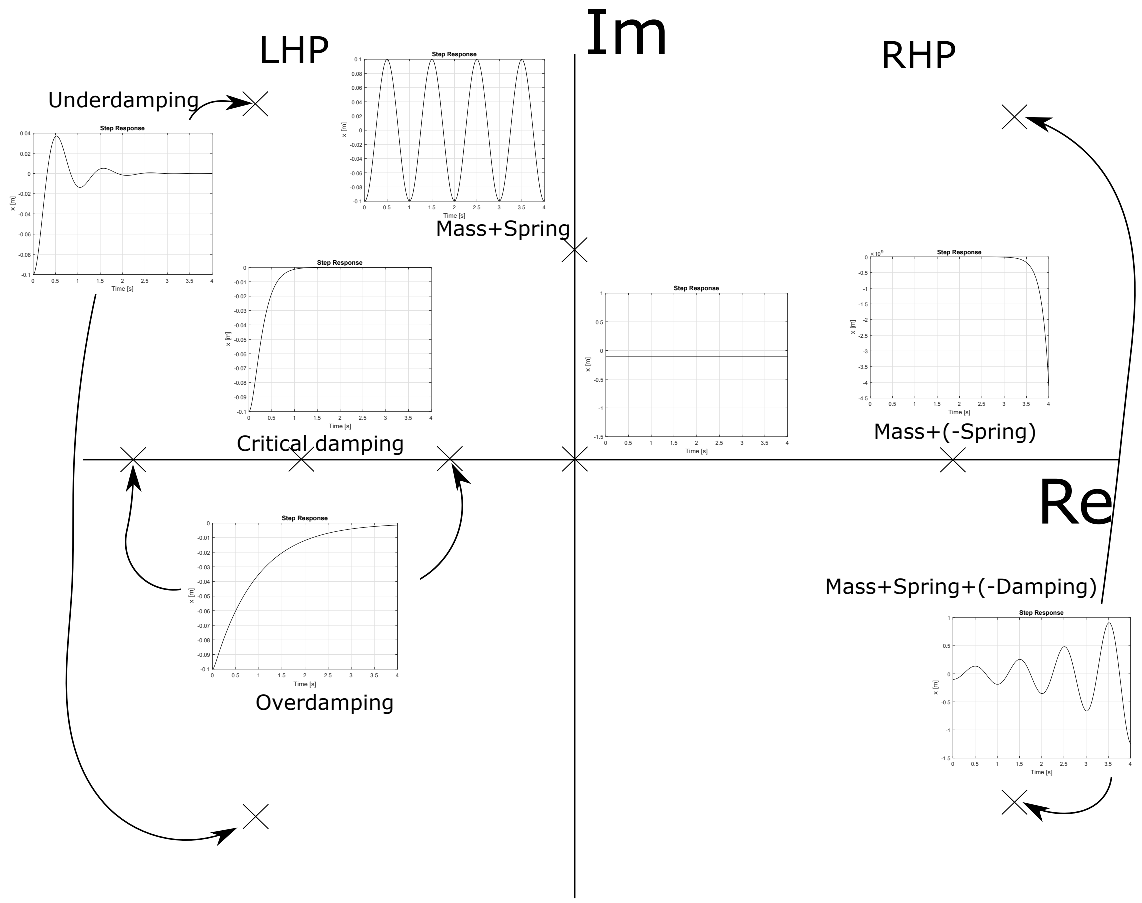

Figure 7: All different system responses to an impulse (step-response) for different poles. All responses are created based on the simple mass spring damper system as described in this article. The complex poles are the complex conjugates, meaning the two poles have the same real part, while the imaginairy parts have an opposite sign. This is also shown by the arrows.

2 Poles and zeros

The transfer function, as shown in the previous section, is typically a fraction, with a numerator and a denominator. For example equation 14 has a numerator which is equal to 1, while the denominator is $m\,s^2 + b\,s + k$.

2.1 Poles

A pole in control theory means that the denominator of the transfer function becomes equal to zero, while s is the variable of the equation. For the simple system as shown in the previous section, the two poles are calculated using the quadratic formula:

$$m\,s^2 + b\,s + k = 0 \tag{15}$$

$$s_{1,2} = \frac{-b \pm \sqrt{b^2 - 4\,m\,k}}{2\,m} \tag{16}$$

The influence of poles on the behaviour of the mechanical system, is huge: the poles show whether the mechanical system is stable. But before explaining when a system is stable or not, the definition of stability needs to be explained.

A system is called stable when the response to an impulse decays or remains equal. This sounds complicated, but in reality it is not. Consider a wheel of a car. When the car hits a sudden bump (which is the impulse in this case), the wheel will compress the spring, and afterward it will go back to the original position. As the wheel will go back to the original position, the response will decay, which means that the wheel is a stable mechanical system. The system would also be called stable when the wheel would keep vibrating around the original position with a constant amplitude. The wheel would be unstable when it would continue to move upwards, away from the original situation. This last run-away effect is often undesired in mechanical systems, as this would mean that the result of a small impulse is large on the future positions of the system, although for instance fighter jets use unstable mechanical systems for fast maneuvering, while the sensors, controller and actuators require to compensate for the instability.

When comparing equation 16 to equations 4 and 5, the similarity is well visible. Unfortunately, the poles for an undamped system are complex (meaning a real and imaginairy part, resulting in the complex conjugates, which have the same real part, but an opposite sign for the imaginairy part), so checking whether the response is stable in the time domain is not that easy. For the frequency domain, this is easier, as the position of the poles in the complex plane directly shows the stability. The different responses are shown in figure 7. Important to learn from figure 7, is that poles on the right half plane (RHP) result in instable behaviour, while poles on the imaginairy axis will result in a response which remains equal (thus stable), while poles in the left half plane (LHP) are stable and the response to an impulse decays over time.

A pole in the right half plane (RHP) results thus in unstable behaviour. In figure 7 the poles in the RHP are obtained by either making the spring or the damping negative. This would be quite difficult to obtain for the simple mass-spring-damper system, as this system is by nature stable. A pencil standing on the tip for instance is an unstable system by nature, and has a pole in the RHP (One can investigate by creating the equations of motions themselves).

2.2 Zeros

Poles are defined as the values for s where the denominator is zero. Zeros are the values for s where the numerator is zero. In the case of the simple mass-spring-damper system, the system does not have any zeros, as there is no value for s where the numerator is zero. It is however possible to artificially add a zero to the system, for example at $s = -1$, meaning that the transfer function becomes:

$$H(s) = \frac{s+1}{m\,s^2 + b\,s + k} \tag{17}$$

The effect is shown in figure 8. Adding the zero changes the system responses, as it changes the contributions of the different responses of the general solution, but it is hard to translate the zero to a physical parameter the time domain. The zeros are not bound to the LHP for stability reasons, so the zero can be placed in the RHP as well. The main effects of the zero at for example $s = -1$ is:

- Faster response
- More overshoot

In figure 8 multiple responses are shown. The added zero is shifted in different directions, even to the right half plane of the complex plane. The different responses show a correlation with the position of the added zero, but all zeros show a much higher initial response and a high overshoot.

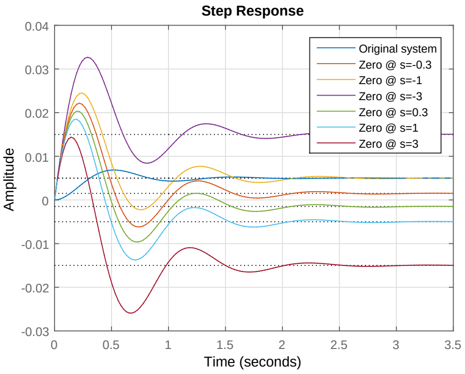

Figure 8: The blue line shows the step response of the underdamped simple mass-spring-damper system ($m = 5kg,\ \ k = 200\frac{N}{m},\ \ b = \sqrt{4\,m\,k}\,0.3$). For the other lines, one zero is added, but the position of the zero is changed.

3 Adding a controller and a control loop

This section describes the commonly used control loops for mechanical systems and the PID controller. An overview of the control loops is shown in figure 9. The block diagram is related to algebra using Mason's rule [1].

3.1 Types of control loops

Up till now, only the mechanical system is described, but in the end the mechanical system needs to be actuated and controlled. This means that a controller needs to be added, which controls the actuation force $U(s)$, which in turn applies this force to the mechanical system. First step is adding only a controller, but the controller needs user input in the form of a set point. This results in the most simple control loop, which is the open loop as shown at the top in figure 9.

The open loop controller uses the set point to determine the force, which is then applied to the mechanical system (the mass-spring-damper in this case). When the set point and the achieved position are not equal, the system does nothing to achieve the set point. This means that this type of controller does not respond to changes in the environment and the force is only dependent on user input.

Normally the environment has some influence (unpredictability in friction, wind forces, etc.), which means that the obtained accuracy with an open loop is not sufficient. Therefore a measurement is used, to calculate the error from the set point. By adding the measurement, the control loop is closed, as the effect of the controller is measured and used again. This is thus called a closed loop controller. In the case of the simple mass-spring-damper system, the height y is measured and subtracted from the set point, as is shown in the middle of figure 9. The resulting error is fed to the controller, which means that the controller will respond to the error instead of the set point. Using a measurement and feeding the resulting error between the set point and the measurement to the controller, is called a feedback loop. Using a feedback loop results in more accurate control of the system, which adapts to changes in the environment.

As stated before, the environment has some influences, for example the wind. When the set point itself changes over time, the error will increase, which results in an actuation force on the mechanical system and leads to a desired acceleration and velocity of the system. With only a feedback loop, the error will need to become larger to make the controller respond, but this always means that the controller is running behind the set point, as it requires an error to be able to respond. When the velocities, accelerations and/or (Coulomb) friction is well known, these can be added to the controller response, meaning that the feedback loop only has to respond to small errors, leading to even more accurate control and smaller errors. The well-known forces which are added to the controller response are called a feed forward controller, which is shown at the bottom of figure 9.

3.2 The PID feedback controller

In figure 9, the square representing the controller is left empty. This is because there are many types of controllers, although for feedback loops in linear systems the PID-controller is the most common. The PID controller means:

- P: Proportional control
- I: Integral control
- D: Derivative control

All these three control parts in the controller will be discussed separately below.

3.2.1 Proportional controller

The proportional controller is quite simple: Multiply the error $E(s)$ with the factor (gain) K_p, which then leads to an output force. In the case of the simple mass-spring-damper, the proportional controller means that the higher the error, the larger the actuator force U. Mathematical this is written as:

$$\frac{U(s)}{E(s)} = K_p \tag{18}$$

$$U(s) = E(s)\, K_p \tag{19}$$

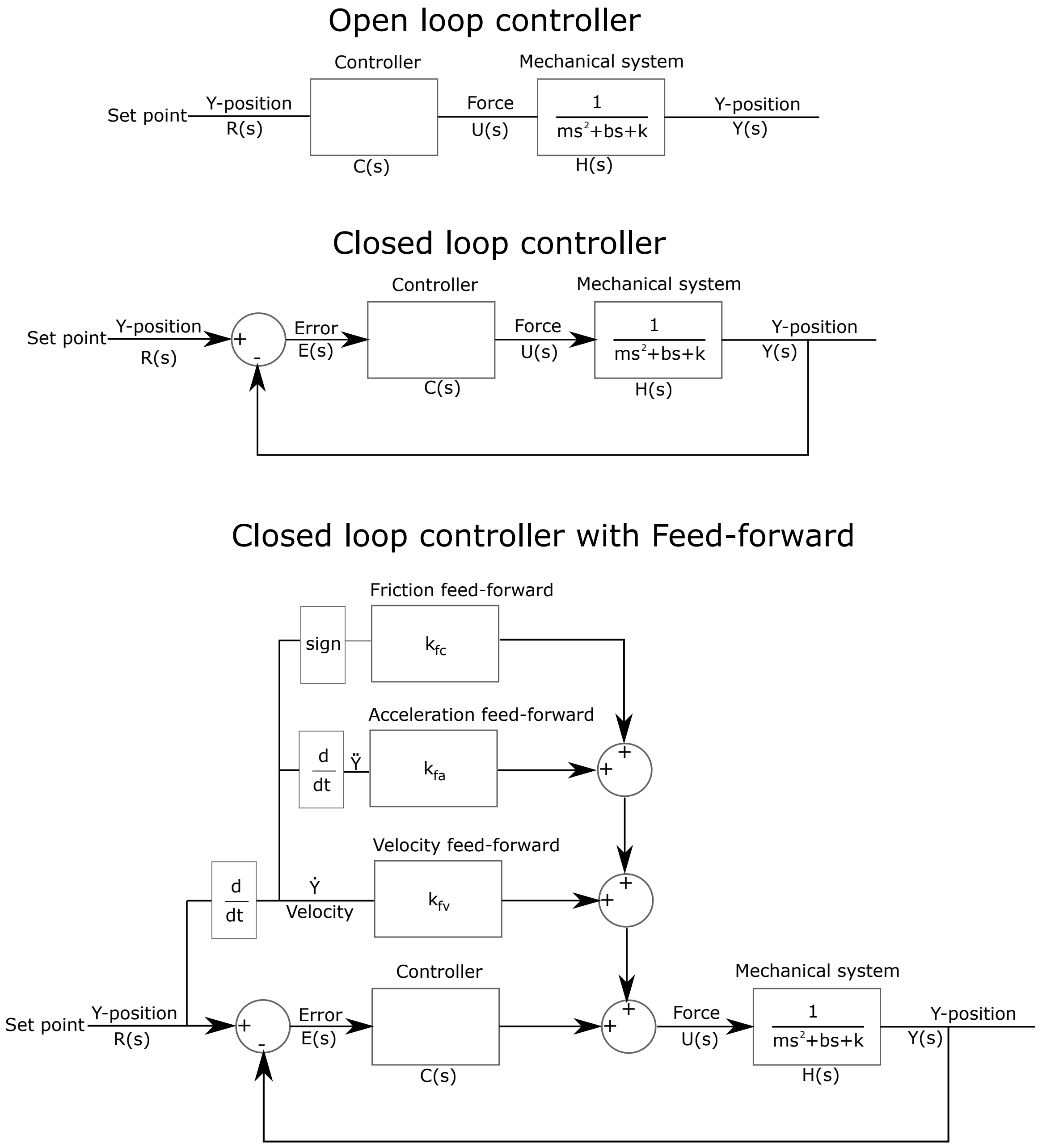

Figure 9: Possible control loops, from simple open loop, to feedback control and a feedback controller with feed-forward loops.

This does however mean that the controller requires an error to keep exerting a force on the mass. This means that if the set point is non-zero, the proportional controller requires an error in order to maintain an actuation force. The actuation force is required because of the spring force at a non-zero position. When looking at the simple mass-spring-damper system it means that the proportional controller will have a static error for any non-zero position. The magnitude of the static error, which is called the offset error, is dependent on the mechanical system (the spring stiffness), the gain of the proportional controller and the set point.

The static error can be calculated from the block diagram in the middle of figure 9 (feedback loop without feed-forward). The error $E(s)$ is the difference between the set point $R(s)$ and the actual or measured position $Y(s)$ ($E(s) = R(s) - Y(s)$). The actual position $Y(s)$ is the product of the mechanical system $H(s)$, the proportional gain K_p and the error $E(s)$. When the last equation is used to substitute $Y(s)$ in the first equation, the following is found:

$$E(s) = \frac{R(s)}{1 + H(s)\,K_p} \tag{20}$$

For a simple mechanical system as the mass-spring-damper, the force of the actuator is easy to calculate using static equations, meaning that the error is also easy to determine. For more complex systems this might be more difficult.

The gain of the proportional controller K_p can be chosen by the control engineer. A small gain will result in a slow reduction of the error and a large static error. A high gain can mean that the transient response (non-steady state response, the sinusoidal response of the system which fades over time) is no longer acceptable. The proportional controller can start a resonance-like behaviour. The actuator can keep adding energy to the system, meaning that the movements can actually grow until parts start to break down. So choosing the gain correctly is important.

When looking at the dynamics of $\frac{Y(s)}{R(s)}$[2], after applying the proportional controller $C(s)$, the following equations are obtained:

$$\frac{Y(s)}{R(s)} = \frac{H(s)\,C(s)}{1 + H(s)\,C(s)} \tag{21}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{1}{m\,s^2+b\,s+k}\,K_p}{1 + \frac{1}{m\,s^2+b\,s+k}\,K_p} \tag{22}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{K_p}{m\,s^2+b\,s+k}}{\frac{m\,s^2+b\,s+k+K_p}{m\,s^2+b\,s+k}} \tag{23}$$

$$\frac{Y(s)}{R(s)} = \frac{K_p}{m\,s^2 + b\,s + k + K_p} \tag{24}$$

When the denominator is examined more closely, it is seen that by adding the proportional controller, a ”spring-like” item is added to the dynamics. This means that the gain K_p is influencing the natural frequency of the controlled feedback system. If the gain K_p is made large, it decreases the static error, although the damping might be insufficient for a satisfactory response due to resonances.

3.2.2 Adding integral action

In order to reduce the static error to zero, it is possible to add an integral action to the propor-

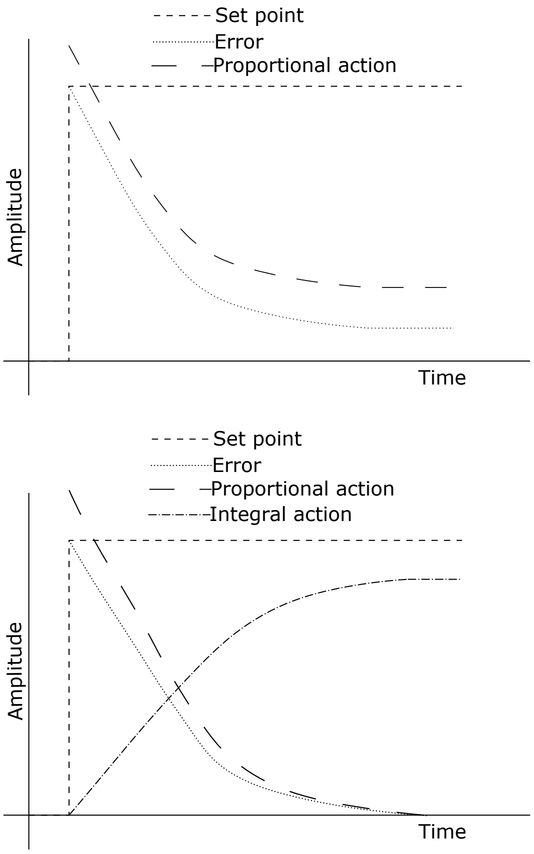

Figure 10: Above a sketch of the proportional action, with a static error. Below a controller with integral action is shown, without a static error.

[2] This is obtained by using the same equations as for the error calculation above: $E(s) = R(s) - Y(s)$ and $Y(s) = H(s)\,C(s)\,E(s)$

tional controller. In the time domain, this integral action is:

$$u(t) = K_p\, e \;+\; K_i \int_{t_0}^{t} e(\tau)\, d\tau \tag{25}$$

In the frequency domain the integral action is much more simple:

$$\frac{U(s)}{E(s)} = K_p + \frac{K_i}{s} \tag{26}$$

The integral action integrates the error over time. Note that the integral of the error represents the area below the graph. This means that once the simple mass-spring-damper system is given a set point, as shown in figure 10, the proportional controller decreases the error $E(s)$. The integral action takes longer to grow, but will eventually be large enough to cancel the static error and be able to keep the static error zero for a long period of time, even when the error $E(s)$ is zero.

Similar as for the proportional controller, the effect on the system is shown by looking at $\frac{Y(s)}{R(s)}$:

$$\frac{Y(s)}{R(s)} = \frac{H(s)\, C(s)}{1 + H(s)\, C(s)} \tag{27}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{1}{m\, s^2 + b\, s + k}\, \left(K_p + \frac{K_i}{s}\right)}{1 + \frac{1}{m\, s^2 + b\, s + k}\, \left(K_p + \frac{K_i}{s}\right)} \tag{28}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{K_p + \frac{K_i}{s}}{m\, s^2 + b\, s + k}}{1 + \frac{K_p + \frac{K_i}{s}}{m\, s^2 + b\, s + k}} \tag{29}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{K_p\, s + K_i}{m\, s^3 + b\, s^2 + k\, s}}{1 + \frac{K_p\, s + K_i}{m\, s^3 + b\, s^2 + k\, s}} \tag{30}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{K_p\, s + K_i}{m\, s^3 + b\, s^2 + k\, s}}{\frac{m\, s^3 + b\, s^2 + k\, s + K_p\, s + K_i}{m\, s^3 + b\, s^2 + k\, s}} \tag{31}$$

$$\frac{Y(s)}{R(s)} = \frac{K_p\, s + K_i}{m\, s^3 + b\, s^2 + k\, s + K_p\, s + K_i} \tag{32}$$

When looking at the denominator of the obtained equation, the second order equation is changed into an third order equation. This means that the dynamics of the entire system is changed by the controller. The system now has three poles, , which can be determined using the ***cubic function***, and one zero. With the PI-controller, two parameters of the denominator can be set by the controller.

3.2.3 Adding derivative action

The last part of the PID controller is the derivative action. The derivative action can speed up the system response, which was not yet done by the integral action. This also means that the derivative action is not required for a system, unless it is used to make the system respond quicker to set point changes. It can also make the loop more stable, which can lead to a possible increase of the proportional action.

The derivative action takes the rate of change of the error into account. So if the error is growing in time, the derivative action will extrapolate the error growth and increase the response of the system, to counter the growing error. In the time domain, the PID controller would become:

$$u(t) = K_p\, e \;+\; K_i \int_{t_0}^{t} e(\tau)\, d\tau + K_d \frac{d}{dt} e(t) \tag{33}$$

In the frequency domain, the PID controller is simpler:

$$\frac{U(s)}{E(s)} = K_p + \frac{K_i}{s} + K_d\, s \tag{34}$$

The derivative action has one major disadvantage: when the error is small, but it contains noise, the derivative action will magnify the noise. Noise is by nature random and rapid changing, which is exactly to what the derivative action will respond. So when the error signal includes significant noise, the derivative action must be made small or zero. When looking at the $\frac{Y(s)}{R(s)}$ term, the following is found:

$$\frac{Y(s)}{R(s)} = \frac{H(s)\,C(s)}{1 + H(s)\,C(s)} \tag{35}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{1}{m\,s^2+b\,s+k}\,\left(K_p + \frac{K_i}{s} + K_d\,s\right)}{1 + \frac{1}{m\,s^2+b\,s+k}\,\left(K_p + \frac{K_i}{s} + K_d\,s\right)} \tag{36}$$

$$\frac{Y(s)}{R(s)} = \frac{\frac{K_p+\frac{K_i}{s}+K_d\,s}{m\,s^2+b\,s+k}}{1 + \frac{K_p+\frac{K_i}{s}+K_d\,s}{m\,s^2+b\,s+k}} \tag{37}$$

$$\frac{Y(s)}{R(s)} = \frac{K_p + \frac{K_i}{s} + K_d\,s}{m\,s^2 + b\,s + k + K_p + \frac{K_i}{s} + K_d\,s} \tag{38}$$

$$\frac{Y(s)}{R(s)} = \frac{K_p\,s + K_i + K_d\,s^2}{m\,s^3 + b\,s^2 + k\,s + K_p\,s + K_i + K_d\,s^2} \tag{39}$$

$$\frac{Y(s)}{R(s)} = \frac{K_d\,s^2 + K_p\,s + K_i}{m\,s^3 + s^2\,(b + K_d) + s\,(k + K_p) + K_i} \tag{40}$$

The denominator is a third order polynomial, similar to the PI-controller. This means that the system has three poles, which can be determined using the ***cubic function***. The difference with the PI-controller is that the denominator of the PID controller has three terms that are influenced by the controller (one not multiplied by s, one multiplied with s and one multiplied with s^2). The number of zeros has increased to two.

3.2.4 Effects of the PID controller

The effects of the different controller actions on the system response are shown in figure 11. The set point is shown as a red line. The set point increases twice within these 30 seconds: at $t = 0$ the set point makes a step to 0.5, and at $t = 15$ the set point makes a step to 1. The used parameters for this graph is:

- $m = 5kg$
- $b = 18,97\frac{N\,s}{m}$
- $k = 200\frac{N}{m}$

When only a proportional controller is used for the simple system, the system will respond as shown by the blue line. The static offset of the proportional controller is clearly visible. The proportional gain in this graph is set to 500.

When an integral action is added to the proportional controller, the black system response is obtained. The static error is gone now a PI-controller is used, although there is some overshoot by the system. The used gain for the proportional controller is 500, and the gain of the integral action is 400.

The magnenta line shows the response when the derivative action is added to the PI-controller. The static error is still gone, but with the derivative action the system initial response is faster and the system does not show any overshoot. The used gain for the proportional controller is 500, the gain of the integral action is 400, and the gain for the derivative action is 100.

3.3 Tuning of the PID controller in the field

When using the PID controller, the PID controller must also be tuned, to make sure that the system functions properly and efficiently. In the next chapter the theoretical tuning of the controller is shown. The final tuning is often done in the field, while the system is active or being commissioned, or sometimes a controller needs to be set completely from scratch. There are a few rules of thumb to set the PID controller in the field. In this subsection two variants are shown: one with only a PID controller and one with a PID controller with feed forward.

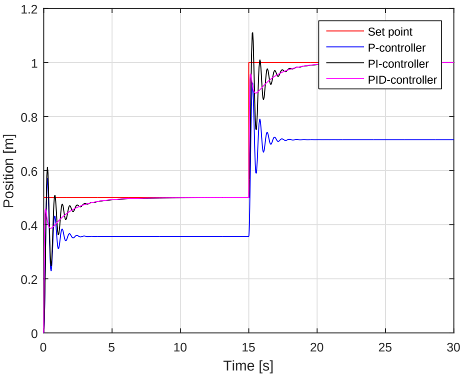

Figure 11: The system response of the simple mass-spring-damper system to a set point. The response of the system with a P-controller, PI-controller and a PID-controller is shown.

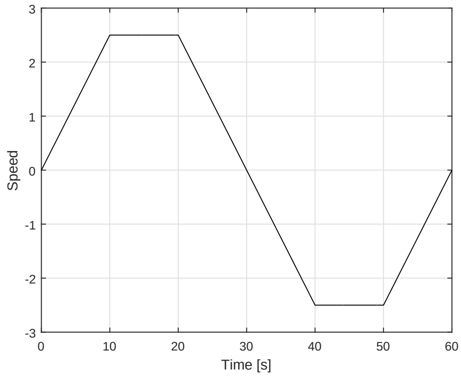

Figure 12: An example of a motion profile to tune the feed forward loop. Important is to have accelerations, decelerations and constant speed in forward as well as backward motion.

3.3.1 PID controller without feed forward

For only a PID controller in a system, the following steps are a rule of thumb to obtain proper control:

1. Set all gains (K_p, K_i and K_d) to zero;
2. Increase the proportional gain K_p until the response shows oscillations. Set the K_p factor to approximately half of this value, to lower the overshoot;
3. Increase the integral action gain K_i until the static error is zero within a reasonable time. Note that a too high K_i factor will result in instability;
4. If required, increase the derivative gain K_d until the response reaches the set point in an acceptable time period. Note that a too large K_d factor results in overshoot and excessive responses.

3.3.2 PID controller with feed forward

When a feed forward loop is used, the PID controller can first be set as in the previous section, while all factors of the feed forward are zero. To tune the feed forward, a motion profile is required. It is important that the motion profile has a part of constant acceleration, a part with a constant speed in forward and backward direction and a part of constant deceleration. Best is to divide the time equally over constant acceleration, constant speed and constant deceleration. An example of such a profile is shown in figure 12. It is assumed that the feed forward as shown at the bottom of figure 9 is used, so including a friction feed forward, acceleration feed forward and velocity feed forward:

1. Set the integral action gain K_i of the PID controller to zero;
2. If a friction feed forward is used, start with tuning the friction gain K_{fc} to minimise the error between deceleration of the forward motion and acceleration in backward movement;
3. Minimise the error during constant velocity by tuning the velocity feed forward K_{fv};
4. As K_{fc} and K_{fv} interact with each other, the setting can be checked by changing the velocity set point. Both K_{fc} and K_{fv} should still have their optimum values;
5. Minimise the error during acceleration and deceleration by changing K_{fa};
6. Reset the I-action of the PID controller.

4 Controller design

Now that the PID controller is explained and practical rules of thumb are explained to set a PID controller in the field, the next step is to design the controller to stabilise the system if it is unstable and optimise the system response. This means that the theory below is applicable to unstable mechanical systems (e.g. a pencil standing on its tip), which was not shown up till now in this article.

As a first step, the sensitivity function is shown, too show in what region the feedback control is actually beneficial. Then the next step is to define the frequency response function, which shows the relation between the frequency response variable s and sinusoidal signals. At last two methods of displaying the frequency response, including their own stability criteria in order to make or keep the system stable and optimalise the performance.

4.1 Sensitivity

Before the actual stability criteria can be shown, it is important to know in what region the feedback loop increases the performance, also in the case when disturbances and sensor noise are present in the system (and these will be present in a real system). The disturbance $W(s)$ and sensor noise $V(s)$ are shown in figure 13.

Note that the sensor in figure 13 also has a 'box', as the sensor normally returns a voltage or current, which needs to be translated to a physical quantity (in this case the position). Whenever the sensor is fast and accurate enough, the transfer function becomes constant to the *volts/unit of output*. When the frequencies of the controller and the sensor are within a factor 5 to 10 of each other, the transfer function is not constant anymore and should be incorporated correctly in the control loop. In the calculation below the transfer function of the sensor will be neglected, to show the base case.

To know where the closed loop control is beneficial, the output of the closed loop is calculated:

$$Y_{cl} = H\,W + H\,C\,(R - (Y_{cl} + V)) \tag{41}$$

$$Y_{cl}\,(1 + C\,H) = H\,W + C\,H\,R - C\,H\,V \tag{42}$$

$$Y_{cl} = \frac{H}{1 + C\,H}\,W - \frac{C\,H}{1 + C\,H}\,R - \frac{C\,H}{1 + C\,H}\,V \tag{43}$$

This means that the closed loop error becomes:

$$E_{cl} = R - Y_{cl} \tag{44}$$

$$E_{cl} = R - \left(\frac{C\,H}{1 + C\,H}\,R + \frac{H}{1 + C\,H}\,W \ldots - \frac{C\,H}{1 + C\,H}\,V\right) \tag{45}$$

$$E_{cl} = \frac{1}{1 + C\,H}\,R + \frac{H}{1 + C\,H}\,W - \frac{C\,H}{1 + C\,H}\,V \tag{46}$$

The last equation above shows three terms which make up the closed loop error. The closed loop error needs to be as small as possible, to make sure that the output is following the reference signal.

Looking at the first term, the factor behind the fraction is the reference R, which means that this term is due to the system response to a reference signal. The fraction $\frac{1}{1+C\,H}$ is called the ***sensitivity function***. To obtain the highest performance, so the lowest error and an output which follows the reference signal directly, the sensitivity function should be zero. This means that the sensitivity function shows in what frequency region the feedback is actually useful, which is where the sensitivity function is small (smaller than 1).

The second term contains the disturbance W. This means that when the fraction $\frac{H}{1+C\,H}$ is small, the disturbance has only little influence on the error. The term $\frac{1}{1+C\,H}\,H$ can be seen as the transfer function of the mechanical system H multiplied with the sensitivity function. This means that the sensitivity function needs to be small for a high disturbance rejection. This is the same as for the first term, although the frequencies of the changes in the reference R are normally much lower than the frequencies of the disturbance W. This means that it is beneficial if the sensitivity function is small for low and high frequencies. As the second term describes the load disturbance sensitivity, the fraction $\frac{H}{1+C\,H}$ is in the literature mentioned as the ***load disturbance sensitivity function***.

The last term of the equation is the error due to sensor noise V. The fraction $\frac{C\,H}{1+C\,H}$ is called the ***complementary sensitivity function***. Important to note is that the sum of the sensitivity function and the complementary sensitivity function is defined as 1, as $\frac{1}{1+C\,H} + \frac{C\,H}{1+C\,H} = 1$. To keep the error due to the sensor noise small, the complementary sensitivity function needs to be small. But a small complementary sensitivity function means a large sensitivity function, since the sum of these two is equal to 1. This means that the designer of the controller needs to verify where he would like to have a robust system with good output performance and how much sensor noise error is acceptable in the process.

There is one other sensitivity function mentioned in the literature, which is the ***noise sensitivity function***. This sensitivity function is

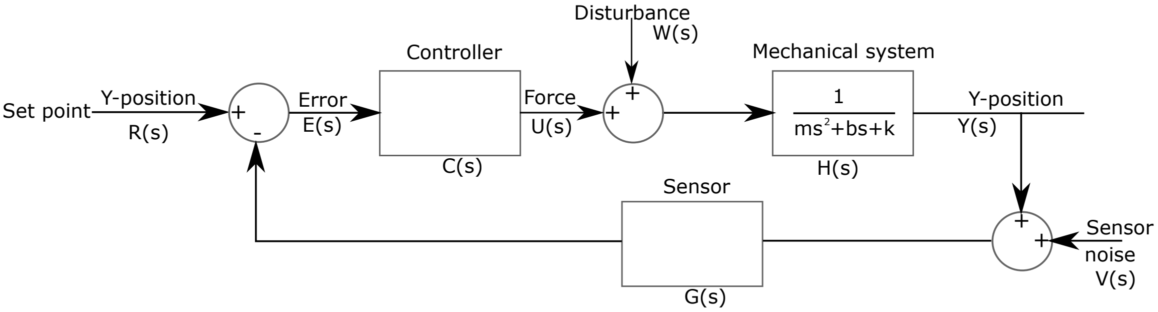

Figure 13: A control loop including disturbance and sensor noise

found when the influence of the sensor noise V on the actuator force U is investigated[3]. This term is has the fraction $\frac{C}{1+C\,H}$.

4.2 Frequency response function

For linear systems it is common to express their response to a sinusoidal signal. The sinusoidal actuation signal has a known amplitude, and in the case of the simple mass-spring-damper system the external force on the mass will be used as actuation.

The system response will have the same sinusoidal response, although the amplitude, in the general case, will be different and the sinusoidal signal will have a phase shift compared to the input signal. Both the system amplitude and the phase shift vary with the frequency.

To obtain the magnitude and phase of the system response to a sine, Euler's formula is taken as starting point:

$$A\,e^{i\,\phi} = A\,(cos(\phi) + i\;sin(\phi)) \tag{47}$$

$$A\,e^{-i\,\phi} = A\,(cos(\phi) - i\;sin(\phi)) \tag{48}$$

$$A\,cos(\phi) = \frac{A}{2}\,(e^{i\,\phi} + e^{-i\,\phi}) \tag{49}$$

Now the input signal, so $u(t)$, is described as a sinusoidal signal. As shown above the cosine has two parts which are superimposed:

$$u(t) = e^{i\,\omega\,t} \tag{50}$$

$$u(t) = e^{-i\,\omega\,t} \tag{51}$$

The transfer function, as shown in equation 13, then leads to:

$$y(t) = H(i\,\omega)\,e^{i\,\omega\,t} \tag{52}$$

$$y(t) = H(-i\,\omega)\,e^{-i\,\omega\,t} \tag{53}$$

Superposition then leads to the sum of these two, to create the response to the sinusoidal signal:

$$y(t) = \frac{A}{2}\,\left(H(i\,\omega)\,e^{i\,\omega\,t} + H(-i\,\omega)\,e^{-i\,\omega\,t}\right) \tag{54}$$

The transfer function is a complex number, dependent on the frequency, to transfer the input to a certain output. This complex number can be described as a magnitude and a phase, as shown in figure 3. This means that the transfer function can be written as:

$$H(i\,\omega) = M(i\,\omega)\,e^{i\,\varphi(\omega)} \tag{55}$$

For simplicity this can also be written as $H = M\,e^{i\,\varphi}$. This representation of the transfer function is used in the function for the output $y(t)$:

$$y(t) = \frac{A}{2}\,M\,\left(e^{i\,(\omega\,t+\varphi)} + e^{-i\,(\omega\,t+\varphi)}\right) \tag{56}$$

$$y(t) = A\,M\,cos(\omega\,t + \varphi) \tag{57}$$

where

$$M = |H(i\omega)|\,,\; \varphi = \angle H(i\omega) \tag{58}$$

The conclusion of this derivation is that when the system can be represented by a transfer function $H(s)$ with a sinusoidal input with magnitude A, the output $y(t)$ will be a sinusoidal signal at the same frequency as the input with magnitude AM and will be shifted in phase by the angle φ.

[3] In figure 13, start with $U =$ and do this similar as done for $Y =$ as shown above.

This derivation is normally shortened by assuming $s = i\,\omega$[4], which has the same result as the derivation above.

4.3 Bode diagram

The Bode diagram is a way to represent a transfer function of a linear time-invariant continuous system[5]. The Bode diagram shows the magnitude of the transfer function $|H(i\omega)|$ in the top graph, and the angle of the transfer function $\angle H(i\omega)$ in the bottom graph. On the X-axis the frequency is shown on a logarithmic scale, to be able to show the low frequencies as well as the high frequencies. The magnitude of the transfer function is displaced in the decibel (dB) scale, which makes the magnitude also a logarithmic scale. The angle is displaced in degrees (°). Examples will be given below.

The Bode diagram and its stability criterion is to be used for linear, time-invariant minimum phase systems only. A linear system means that the differential equation is linear, so there are no multiplications of the differential variable with itself or with its higher or lower differentials. Time-invariant means that the system is not dependent on time directly (the differential equation is not dependent on t). A minimum phase system is a stable system (so all poles in the left-half plane as shown in figure 7) which has an inverse which is causal and stable. In practice this means that a minimum phase systems is a stable system which has no zeros in the right-half plane.

4.3.1 Influence of poles, zeros and resonances on a Bode plot

As stated above, all poles should be in the left-half plane (or on the axis) of the complex plane. Each pole or zero can be related to a frequency by looking at the frequency response function, as shown in paragraph 4.2.

The real poles and zeros (so poles and zeros without an imaginairy part) become 'active' when the frequency ω in rad/s is equal to the pole or zero (i.e. $s = -3$ means a frequency of $-3rad/s$).

Imaginairy poles and zeros (so poles and zeros without a real part) become 'active' when the frequency $i\omega$ in rad/s is equal to the pole or zero (i.e. $s = -3i$ means a frequency of $-3rad/s$). Keep in mind that undamped second order systems have resonances with a magnitude going to infinity in the Bode plot, as there is no damping in the system.

The frequency of complex poles and zeros are less easy to translate to a Bode plot, as the damping is also of influence and there might be resonances in the system. The frequency of the complex pole or zero is dependent on the real and imaginairy part of the pole or zero (in analogy with the eigen value $\lambda = k \pm i\,\omega$ as seen in paragraph 1.1), but also of the real part, as the frequency is now the damped natural frequency. This frequency for mechanical systems can be calculated using the damping ratio ζ as shown in paragraph 1.1.

The Bode plot top graph shows the magnitude of the transfer function on a logarithmic scale, the bottom graph the phase on a linear scale. The frequency X-axis is also a logarithmic scale. An example is shown in figure 14.

The top graph, which shows the magnitude, changes the gradient at a frequency of a pole or a zero. A pole will result in a gradient which

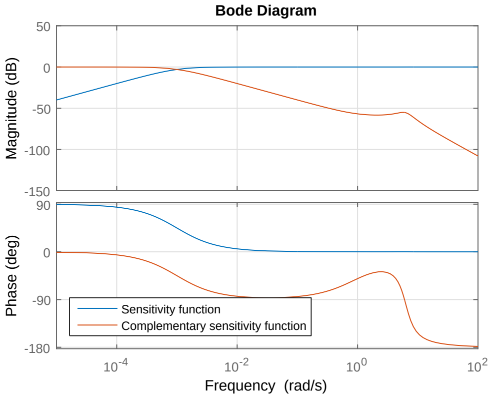

Figure 14: A Bode plot example of the sensitivity and complementary sensitivity function of the example above.

[4] The magnitude of the transfer function $|H(s)|$ is then replaced by $|H(i\omega)|$. The same is done for the angle, resulting in $\angle H(s) \rightarrow \angle H(i\omega)$.

[5] A linear time-invariant continuous system is a linear system, which means that the differential equation is linear, that does not change its behaviour over time.

is -1 (or $-20dB$ per decade) compared to the gradient as before the pole. A zero will increase the gradient by $+1$ (or $+20dB$ per decade) at the frequency of the zero.

The bottom graph, which shows the phase, will have a phase shift. A pole will change the shift by $-90°$ at the frequency of the pole, while a zero will shift the phase by $+90°$.

Both the phase shift as the magnitude change is smooth in a Bode diagram, so there are no sharp edges, but smooth bends in the lines.

A resonance, as seen for second order mechanical systems, has two poles (complex conjugates or two poles on the same position) and will show as a large peak in the magnitude, and a phase shift of $-180°$.

4.3.2 Sensitivity function

The sensitivity function and complementary sensitivity function are plotted in figure 14. The sensitivity function is smaller than $0dB$ until $6rad/s$, which means that the feedback loop is beneficial below the $6rad/s$ for this system, as a small sensitivity function means that disturbances will only result in small errors.

The complementary sensitivity function is approximately 1 until a frequency of $0.001rad/s$, which means that sensor noise far above the $0.001rad/s$ will lead to small errors in the feedback loop. The higher the frequency of the sensor noise, the less impact on the feedback loop. Sensor noise of $0.001rad/s$ or lower will result in errors in the feedback loop.

In the previous paragraph the use of poles and zeros was explained. Below the poles and zeros of the system as shown in figure 14 will be shown and explained, just as an example.

The sensitivity function as plotted in figure 14 has three poles:

- at -0.0010;
- at $-1.8969 + 6.0364\,i$ and at $-1.8969 - 6.0364\,i$.

And the sensitivity function has three zeros:

- at 0.0;
- at $-1.8974 + 6.0332\,i$ and at $-1.8974 - 6.0332\,i$.

For low frequencies, there is only one zero at 0.0, which is active from a frequency of $\omega = 0rad/s$, which results in a gradient of $+1$ or $+20dB$ per decade and a phase of $+90°$. At the frequency of $10^{-3}rad/s$, the real pole becomes active and the gradient goes to 0 and a phase of $0°$. The complex poles and zeros have a similar frequency, as their complex numbers are close together. This means that these will (almost) cancel each other.

The complementary sensitivity function as plotted in figure 14 has six poles:

- at 0.0;
- at -0.0010;
- at $-1.8969 + 6.0364\,i$ and at $-1.8969 - 6.0364\,i$;
- at $-1.8974 + 6.0332\,i$ and at $-1.8974 - 6.0332\,i$.

And the complementary sensitivity function has four zeros:

- at 0.0;
- at -1;
- at $-1.8974 + 6.0332\,i$ and at $-1.8974 - 6.0332\,i$.

What is clear is that two complex poles and zeros at $-1.8974+6.0332i$ and $-1.8974-6.0332i$ cancel each other. Also the pole and zero at the origin 0.0 cancel each other. This means that the gradient at low frequencies is 0 and the phase is $0°$. At a frequency of $10^{-3}rad/s$ the pole at -0.0010 becomes active, which results in a gradient of -1 or $-20dB$ per decade and a phase of $-90°$. At a frequency of $1rad/s$ the zero at -1 becomes active, which means that the gradient goes back to 0 and the phase goes back to $0°$. At a frequency of approximately $6rad/s$ the two poles at $-1.8969+6.0364i$ and $-1.8969+6.0364i$ become active, which means that the gradient goes to -2 or $-40dB$ per decade and a phase of $-180°$.

4.3.3 System response and stability criterion

To investigate the stability of a closed loop linear time-invariant continuous system, the open loop (so in most cases $C(s)\,H(s)$) is examined. The Bode stability criterion has criteria for the open loop, and if the criteria are met, the closed loop transfer function is stable. Note that for the Bode stability criterion the system may not have any poles or zeros outside the left half plane of the complex plane, as discussed in the beginning of this paragraph.

To check the Bode stability criterion, the open loop transfer function ($C(s)\,H(s)$) is plotted in a Bode plot. The stability criterion is:
If at the gain crossover frequency, which is the frequency at which the magnitude of the transfer function is 1 *(which is* $0dB$*), the corresponding phase is not be equal to or lower than* $-180°$*, the feedback system is stable.*
This means it is also important that after the phase crossover frequency (which is the frequency where the phase is $-180°$) the gain does not reach $0dB$ anymore, as that type of system is too complex for the Bode stability criterion. The same applies to the phase margin: to use the Bode stability criterion the phase should not cross the $-180°$ multiple times.

This means that a stable system has a phase margin at the gain crossover frequency, as the phase needs to have a sufficient clearance from the $-180°$ at the gain frequency. As a rule of thumb the phase margin should be at least 30°.

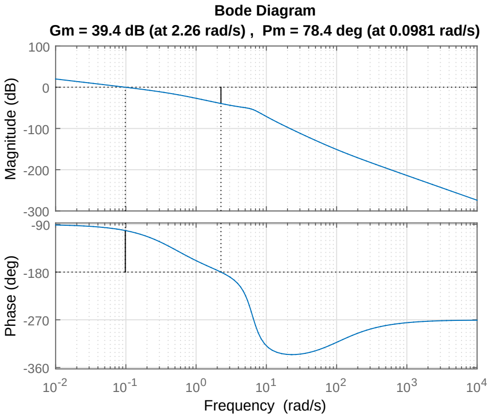

Figure 15: A Bode plot with the phase and gain margin plotted. To be able to show the gain as well as the phase margins, one extra pole at $-1/2$ is added to the controller and the proportional gain of the controller is increased to 100.

Similarly, it means that a stable system has a gain margin at the phase crossover frequency, as the magnitude (=gain) needs to be below the $0dB$ line.

Both the gain and phase margins are shown in the Bode plot of the open loop in figure 15. The system (with the added pole) now has one zero at -100 and four poles:

- at $0, 0$;
- at -0.5;
- at $-1.8974 + 6.0332\,i$ and at $-1.8974 - 6.0332\,i$.

Using the rules above, the Bode plot can be explained by the zero and poles.

4.4 Nyquist diagram

The Nyquist diagram and Nyquist stability criterion give a more broad stability criterion than the Bode stability criterion. The Bode stability criterion still has the requirement that the system is a minimum phase system, but for more complex mechanical systems this is not always the case.

The Nyquist criterion makes it possible to study non-minimum phase systems, systems with multiple resonances or systems which cross the $0dB$ line or $-180°$ multiple times. The Nyquist criterion uses the open loop frequency response to examine the stability of the closed loop poles, similar as the Bode criterion.

The Nyquist plot draws the transfer function in the complex plane. An example of a Nyquist plot is shown in figure 17. A Nyquist plot shows the position of the transfer function at each frequency, which means that each dot on the line represents a frequency. It is common to show the negative frequencies as well in a Nyquist plot, which will be explained later on.

Similar to the Bode diagram, for a minimum phase system the phase margin at $0dB$ (point where the line crosses the unit circle around the origin) and the gain margin at $-180°$ are important. The phase margin is shown clearly in figure 17. The gain margin is slightly more hidden, as shown by the $1/(Gain\ margin)$. To give more clarity about the thought behind the Nyquist plot, the *argument principle* will be explained first.

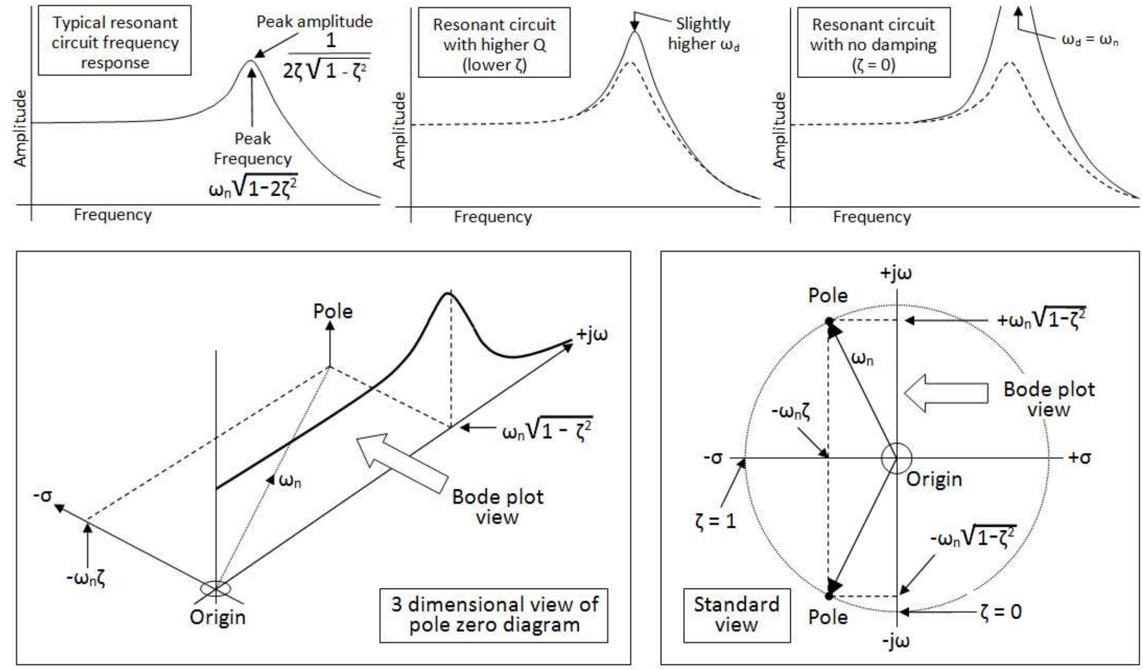
Typical resonant circuit frequency response
Peak amplitude
$\frac{1}{2\zeta\sqrt{1-\zeta^2}}$
Peak Frequency
$\omega_n\sqrt{1-2\zeta^2}$
Amplitude
Frequency
Resonant circuit with higher Q (lower ζ)
Slightly higher ω_d
Amplitude
Frequency
Resonant circuit with no damping (ζ = 0)
$\omega_d = \omega_n$
Amplitude
Frequency
Pole
+jω
$\omega_n\sqrt{1-\zeta^2}$
ω_n
-σ
$-\omega_n\zeta$
Bode plot view
Origin
3 dimensional view of pole zero diagram
+jω
Pole
$+\omega_n\sqrt{1-\zeta^2}$
ω_n
Bode plot view
$-\omega_n\zeta$
-σ
+σ
Origin
ζ = 1
$-\omega_n\sqrt{1-\zeta^2}$
Standard view
Pole
-jω
ζ = 0

Figure 16: A plot to give more clarity about the poles and the Bode diagram.

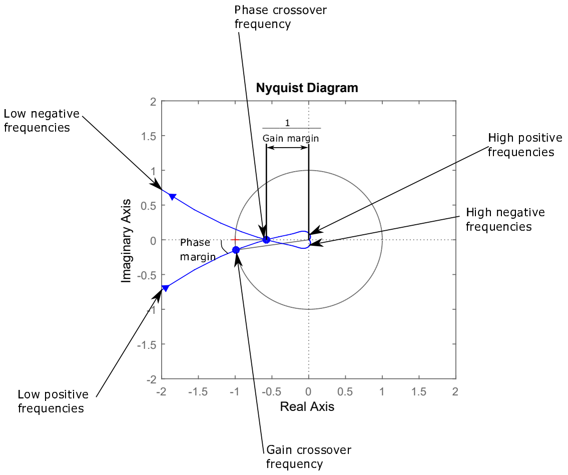

Figure 17: An explanation of the Nyquist plot, including the gain en phase margins. To show the gain as well as the phase margins, one extra pole at $-1/2$ is added to the controller and the proportional gain is increased to 5000.

4.4.1 The argument principle

The argument principle is also referred to as the contour evaluation. For the argument principle the closed loop response will be evaluated (so $C(s)H(s)/(1+C(s)H(s))$), as the closed loop system needs to be stable (and the controller needs to stabalise the mechanical system when the system itself is unstable, like a pencil standing on its tip).

For stability, no poles are allowed in the RHP, which means that for the closed loop system all solutions to $(1 + C(s)\,H(s)) = 0$ should have a negative real part. Now assume the following [2]:

$$C(s) = \frac{N_C}{D_C} \tag{59}$$

$$H(s) = \frac{N_H}{D_H} \tag{60}$$

Then the conclusion for $(1 + C(s)\,H(s)) = 0$ is:

$$1 + C(s)\,H(s) = 1 + \frac{N_C\,N_H}{D_C\,D_H} \tag{61}$$

$$= \frac{D_C\,D_H + N_C\,N_H}{D_C\,D_H} \tag{62}$$

The poles of $(1+C(s)\,H(s))=0$ are equal to the poles of the open loop system $(C(s)\;H(s))=0$ and are all known, as both equal $D_C\;D_H=0$.

The poles of the closed loop $C(s)\;H(s)/(1+C(s)\;H(s))$ however are unknown, as this would mean $\frac{D_C\;D_H+N_C\;N_H}{D_C\;D_H}=0$ and thus $D_C\;D_H+N_C\;N_H=0$ as long as $D_C\;D_H\neq0$, which is not known from the open loop. This can be solved by mathematics alone, but can be done more easily, namely by making a Nyquist plot, which is based on a contour plot.

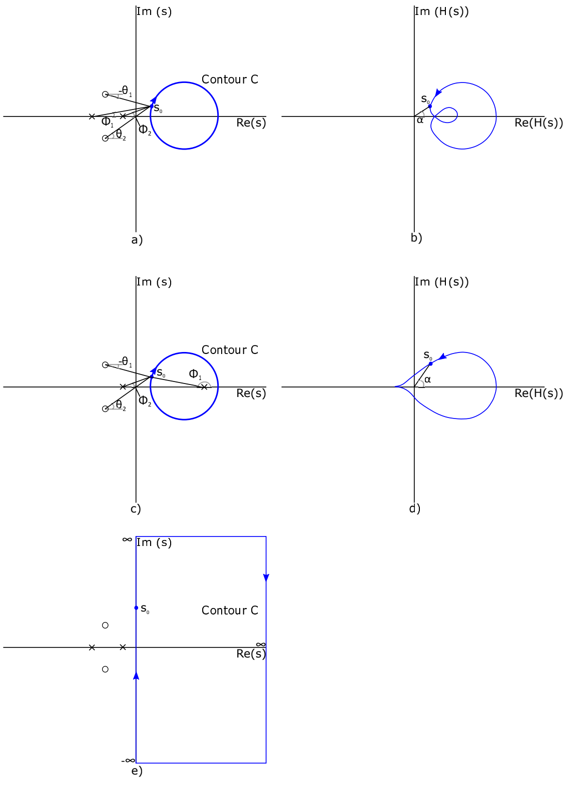

Figure 18: The reasoning behind the argument principle: when examining a transfer function, an encirclement of the origin in the right plot is of influence of the system behaviour. As any pole or zero in the right half plane (left graphs) will result in an encirclement of the origin (right graph), both will result in either a clockwise encirclements (zero) or counterclockwise encirclement (pole) of the origin.

The contour of the contour plot is a part of the Right Half Plane (RHP) as shown in figure 18 a), c) and e). For this example and for simplicity, the examined system has two poles and two zeros. The angle α as shown in figure 18 b) and d) is:

$$\alpha=\theta_1+\theta_2-(\phi_1+\phi_2)\tag{63}$$

If all poles and zeros are in the left half plane, as shown in parts a) and b) of figure 18, the angle and thus the phase of the transfer function, will be between $-90°$ and $90°$ for each pole or zero.

The length of the vector in the Nyquist plot, is determined by the multiplication of the lengths of the zeros to that point on the contour, divided by the multiplication of the lengths of all poles to the point on the contour. For example, the length of the arrow in the Nyquist plot of a system with two zeros and three poles, is determined by:

$$R=\frac{V_1\;V_2}{V_3\;V_4\;V_5}\tag{64}$$

where R is the distance of the origin to the Nyquist plot at that point on the contour, V_1 and V_2 are the lengths of the zeros to the point on the contour and V_3, V_4 and V_5 are the lengths of the poles to the point on the contour. Note that the point of the contour on the imaginairy axis correspond to the positive and negative frequencies, as shown at the frequency response function in paragraph 4.2.

If there is a pole in the right half plane (which leads to unstable behaviour), the transfer function will encircle the origin one time counterclockwise, as shown in parts c) and d) of figure 18. A zero in the right half plane will result in a clockwise encirclement of the origin. This means that the encirclement of the origin is:

$$Z=N+P\tag{65}$$

where N is the number of clockwise encirclements, Z is the number of zeros of $1+C(s)\;H(s)$ in the RHP and P are the number of poles of $1+C(s)\;H(s)$ in the RHP. As stated before, the Z is not directly known from the open loop $C(s)\;H(s)$. The encirclements can be counted and the poles in the RHP are known from the open loop. This way the number of zeros of $1+C(s)\,H(s)$ in the RHP can be determined, and thus the amount of poles of the closed loop system in the RHP. This means that this way the stability of the closed loop system can be examined by only counting the open loop poles in the RHP and the amount of encirclements.

Important note: the contour is not allowed to pass through a pole or zero, so poles / zeros at the imaginairy axis need special attention. To be able to analyse systems with poles or zeros at the imaginairy axis (most commonly at $0 + 0i$), the a mathematical 'trick' is used: the contour goes around the pole with a half circle to the RHP, where the radius approaches zero. This way it is still possible to analyse a system with poles or zeros on the imaginairy axis.

4.4.2 Application to the Nyquist plot

Using the Nyquist criterion, the stability of a system is determined. This is why the open loop poles in the RHP and the encirclements of the $-1 + 0\,i$ point in the Nyquist plot need to be determined. Therefore, the contour as shown previously will contain the complete RHP, as shown in 18 e).

With the contour as shown in 18 e), the system must have Z open loop zeros in the RHP when the open loop encircles the origin N times in clockwise direction and the open loop system has P poles in the RHP ($Z = N + P$).

Because the contour also contains the negative imaginairy axis (contour goes from$-\infty i$ to $+\infty i$), the negative frequencies need to be included as well in the analysis. This is a difference with the Bode analysis, as the frequency range for the Bode analysis is from $s = 0$ up to $s = \infty i$. Note that the negative part is the complex conjugate of the positive imaginairy axis.

The closed loop system is to be stable[6], even for a mechanical system which is not stable. The response of the system is:

$$\frac{Y(s)}{R(s)} = \frac{C\,H}{1 + C\,H} \tag{66}$$

Therefore, the closed loop poles are the solution of $1 + C\,H = 0$, which determine the stability of the closed loop system. The contour argument is thus now applied to $1 + C\,H = 0$. For simplicity, the argument is now transferred to the open loop $C\,H$:

$$C\,H = -1 \tag{67}$$

Therefore, the open loop analysis will evaluate the encirclements of the -1 point on the real axis instead of the origin. Thus: if the transfer function $C\,H$ has any poles or zeros in the RHP, this leads to encirclements of the -1 point. As the Nyquist criterion examines the open loop system (Nyquist($C\,H$)), the encirclements of the $-1, 0\,i$ point is taken.

The Nyquist criterium for the *closed loop* system, states:

In order for the closed loop system to be stable the open loop zeros in the RHP Z need to be zero, hence $Z = 0$.

The main conclusion is then that for *closed loop* stability, the *open loop* Nyquist plot should have P **counterclockwise** encirclements of the -1 point. If a system has no poles in the RHP ($P = 0$), the -1 point should have no encirclements. The counterclockwise encirclements can be made by adjusting the controller $C(s)$, for instance by adding zeros to the controller.

One tricky part for counting the encirclements, are the poles at the origin[7]. A pole at the origin will make the Nyquist plot go to infinity for very small (i.e. approaching 0 for the positive as well as negative) frequencies. As the contour is a closed contour, i.e. all parts are connected, it is important to know how these frequencies connect to each other at infinity. A pole at the origin will always give a 180° clockwise encirclement at infinity. So 4 poles at the origin (for example $H(s) = \frac{1}{s^4}$) will create two complete clockwise encirclements of the -1 point at infinity (not well visible in the Nyquist graph, but important to know and calculate). These clockwise encirclements result in unstable behaviour, which means that these clockwise encirclements must be undone by the controller $C(s)$, which is to be designed by the engineer.

This means that for non-minimum phase systems it is possible to make these stable, but

[6] If the open loop is unstable, for instance due to an unstable plant (i.e. a pencil standing on its tip), the controller is to be made such that it stabilises the closed loop behaviour of the system.

[7] Note here that zeros of the open loop system at the origin cancel a pole at the origin, but with more zeros than poles at the origin the Nyquist plot does not go to infinity, thus it cannot have encirclements at infinity.

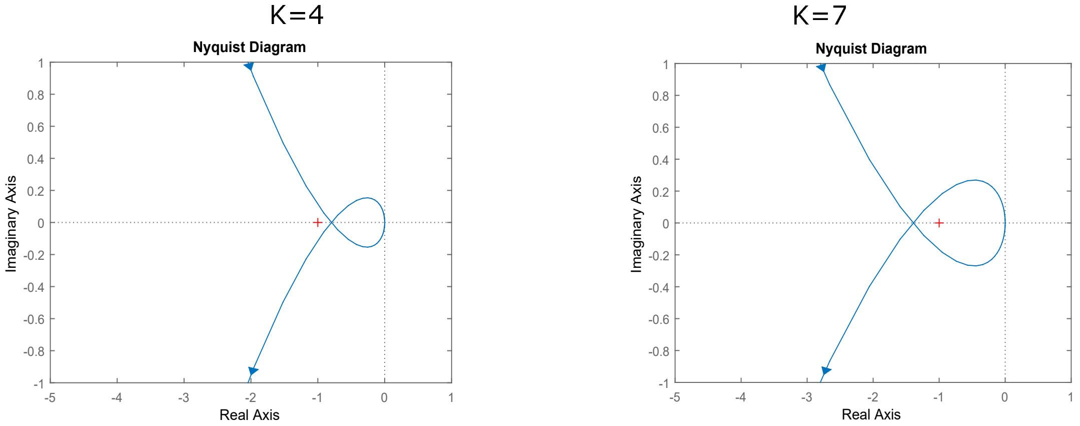

Figure 19: The Nyquist diagrams of the example for $K = 4$ and $K = 7$.

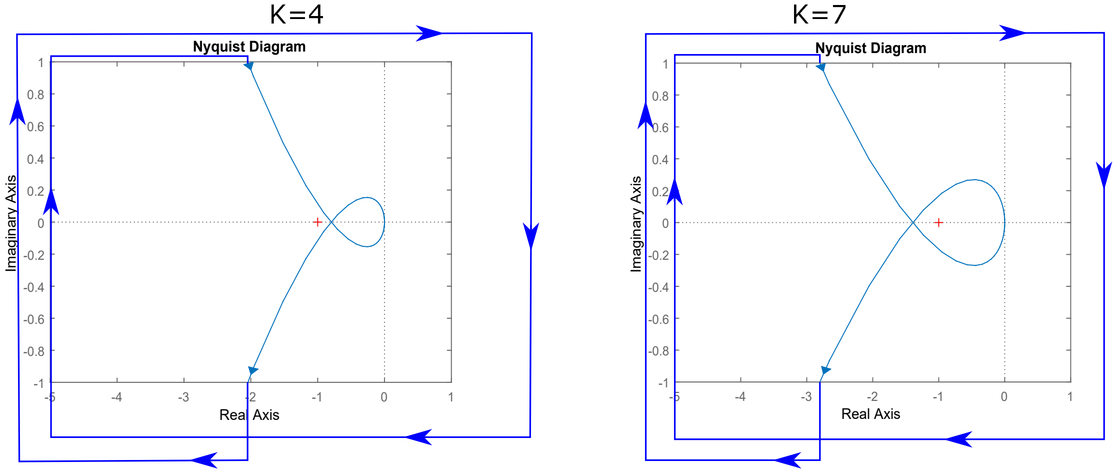

Figure 20: The Nyquist diagrams of the example for $K = 4$ and $K = 7$, now with the encirclements due to the poles at the drawn as well.

there is no set of rules which can be followed to stabalise an unstable mechanical system. This means that the engineer must checks the Nyquist plot and probably requires simulations to check the analysis before implementation in a real system.

4.4.3 An example for the Nyquist criterion

Up till now there was a lot of theory. An example will be used to get a better understanding of the theory. In this example a 'conditionally stable' system will be used to show the stability principle.

To start this example, the mechanical system is assumed to have the following transfer function:

$$H(s) = \frac{K\ (s+10)^2}{s^3} \tag{68}$$

where the factor K is used for the stability. The controller is neglected for now. The stability is now checked with the Nyquist theorem and the simple Nyquist plots are shown in figure 19.

When looking at equation 68, it is clear that the mechanical system has no poles in the RHP, as all three (3) poles are at the origin. This means that $P = 0$.

Now the number of encirclements of the $-1+oi$ point are examined. When examining at the Nyquist plots of figure 19, it becomes clear that the system connects the positive and negative frequencies at infinity. This means that these encirclements must be drawn as well, and only figure 19 is not sufficient to examine the stability of the system. The three poles at the origin create three (3) 180° clockwise rotations. These encirclements are drawn in figure 20. For the case of $K = 4$, the -1 point has 2 encirclements ($N_{K=4} = 2$). The case of $K = 7$ the -1 point has no nett encirclements ($N_{K=7} = 0$).

When the zeros in the RHP are then calculated using the Nyquist equation, the following results are found:

- $Z_{K=4} = N_{K=4} + P = 2 + 0 = 2$
- $Z_{K=7} = N_{K=7} + P = 0 + 0 = 0$

So even without noticing, the variant with $K = 4$ will have unstable behaviour. This can be checked by checking the closed loop poles (see equation 66), which are equal in this case to $1 + H(s)$:

$$1 + H(s) = 1 + \frac{K\ (s+10)^2}{s^3} = 0 \tag{69}$$

$$= \frac{s^3 + K\ (s+10)^2}{s^3} = 0 \tag{70}$$

and thus:

$$s^3 + K\ (s+10)^2 = 0 \tag{71}$$

When the closed loop poles are calculated for $K = 4$, two poles are in the RHP (the three poles are -4.778, $0.389 - 9.14\,i$, $0.389 + 9.14\,i$), meaning this system is unstable. When the same is done for the $K = 7$ case, there are no poles in the RHP (the poles are -5.34, $-0.83 - 11.42\,i$, $-0.83 - 11.42\,i$), meaning that this system is stable.

This example shows the added value of the Nyquist criterion, as it predicts the stability of the mechanical system, even when it is not a minimum phase system.

5 Types of filters

Within the control, several types of filters are used. Sometimes these are used in the controller, or for example the filters are used to filter noise from a measurement signal. This section will show the most basic and commonly used filters.

5.1 Lead-Lag filter

A lead or a Lag filter is used in robotics (for example satellite control) in analogue as well as digital control. The lead filter creates a lower response at low frequencies, while it has a higher response at higher frequencies. For the lag filter it is just the other way around. The transfer function of a lead-lag filter is:

$$C(s) = \frac{s+z}{s+p} \tag{72}$$

where the z is used to place the zero and the p to place the pole of the filter. When $z < p$ the filter is a lead filter, when $p < z$ the filter is a lag filter. The bode plots of these types of filters are shown in figure 21.

Often the lead and lag filters are used together to create a band pass filter. This filter amplifies a certain frequency range. This filter is described

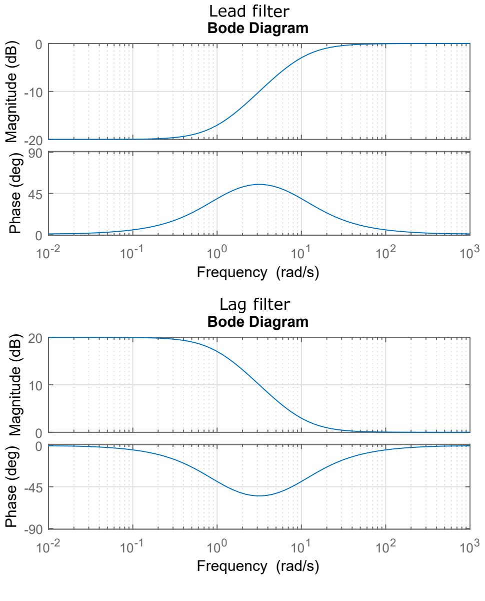

Figure 21: A Bode plot of the lead (top) and lag (bottom) filter. The poles and zeros are placed at $1 rad/s$ and $10 rad/s$.

by the transfer function:

$$C(s) = \frac{(s + z_1)\,(s + z_2)}{(s + p_1)\,(s + p_2)} \tag{73}$$

where $z_2 > p_2 > p_1 > z_1$. The bode plot of the band pass filter is shown in figure 21. When the same filter is used, but now for $p_2 > z_2 > z_1 > p_1$, the magnitude of a certain frequency range will be decreased.

An important note for the lead, lag and band pass filter is that these can influence the stability when the poles or zeros are close to the crossover frequency. The engineer should take this into account and keep the controller action in the correct bandwidth and check the stability after applying the filter.

5.2 Low pass filter

A low pass filter is used when the higher frequencies need to be reduced. The higher frequencies can be considered as noise in many cases, which immediately shows the common application for this filter: noise reduction.

The low pass filter is commonly used in a first or second order low pass filter. The low pass filter is defined using the cut-off frequency ω_c:

$$C(s) = \frac{1}{\tau\, s + 1} = \frac{1}{\frac{1}{\omega_c}\, s + 1} \tag{74}$$

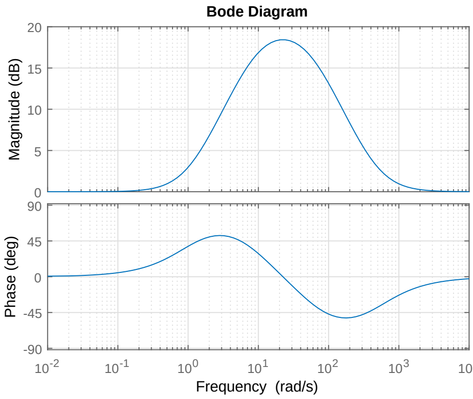

Figure 22: A Bode plot of a band pass filter. The poles and zeros are placed at $1rad/s$, $10rad/s$, $50rad/s$ and $500rad/s$.

where τ is the time constant, dependent on ω_c. The time constant τ is in electronics equal to RC, but the complete electronic scheme is beyond the scope of this article. An example of the first order low pass filter is shown in figure 23.

The second order low pass filter is defined using the Q-factor Q. The Q-factor is defined based on the damping ratio ζ:

$$Q = \frac{1}{2\,\zeta} \tag{75}$$

This means that at $Q = 1/2$ the filter will act as

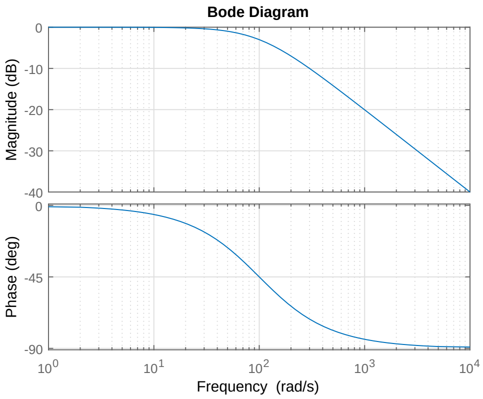

Figure 23: A Bode plot of a first order low pass filter. The cut-off frequency is at $100rad/s$.

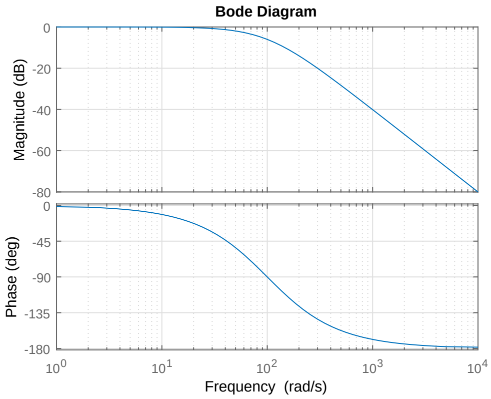

Figure 24: A Bode plot of a second order low pass filter. The cut-off frequency is at $100rad/s$ and $Q = 1/2$. To get the same magnitude as the first order low pass filter, this filter is multiplied with a gain of 100.

if it was critically damped, when $Q > 1/2$ it is underdamped and $Q < 1/2$ is is overdamped.

The transfer function of the second order low pass filter is:

$$C(s) = \frac{\omega_c}{s^2 + \frac{\omega_c}{Q}\, s + \omega_c^2} \tag{76}$$

where ω_c is the cut-off frequency of the filter. An example of the bode plot is shown in figure 24. Note that the slope is twice as high (-2 or -40dB/decade instead of -1 or -20dB/decade) with respect to the first order low pass filter, meaning that the higher frequency (and thus noise) is much more suppressed.

Both the first and second order low pass filters can be multiplied with a gain, to further increase the magnitude of the output.

5.3 High pass filter

In mechanics the high pass filters are seldom used, as they magnify measurement noise, but in some specific cases these filters can be useful. These filters suppress the low frequencies, while allowing the high frequencies to pass or to be magnified using a gain. Similar as the low pass filter, the high pass filter has a first and second order variant.

The first order high pass filter transfer function is:

$$C(s) = \frac{s}{\tau\, s + 1} = \frac{s}{\frac{1}{\omega_c}\, s + 1} \tag{77}$$

The second order high pass filter transfer function is:

$$C(s) = \frac{s^2}{s^2 + \frac{\omega_c}{Q}\, s + \omega_c^2} \tag{78}$$

The Bode plots are shown in figure 25 and figure 26 respectively.

5.4 Notch

The notch filter is a filter which suppresses along a narrow frequency band. This can for instance be used to suppress resonances of the mechanical system using this filter.

The notch has a general form, but often it is used in the standard form. The general definition of the transfer function of the notch filter is:

$$C(s) = \frac{s^2 + \omega_z^2}{s^2 + \frac{\omega_p}{Q}\, s + \omega_p^2} \tag{79}$$

where ω_z is the frequency of the zero (cut-off frequency) and ω_p denotes the frequency of the pole. Normally these frequencies are chosen equal $\omega_z = \omega_p$[8], which results in:

$$C(s) = \frac{s^2 + \omega_0^2}{s^2 + 2\, \omega_c\, s + \omega_0^2} \tag{80}$$

where ω_0 is the center of the rejected frequency

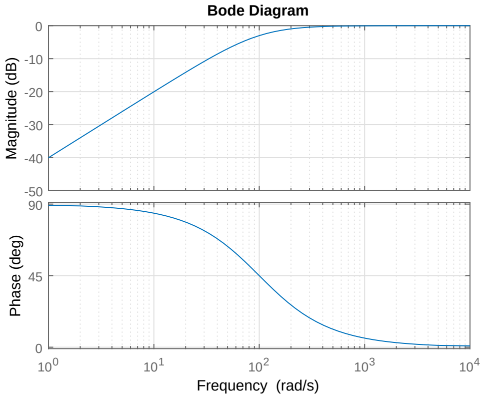

Figure 25: A Bode plot of a first order high pass filter. The cut-off frequency is at $100 rad/s$ and the transfer function is multiplied with a gain of $K = 1/100$, in order to obtain a similar plot as for the second order high pass filter.

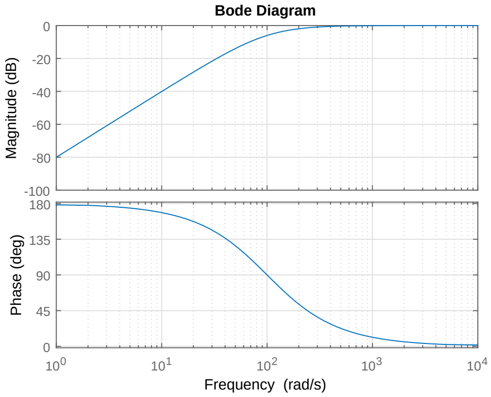

Figure 26: A Bode plot of a second order high pass filter. The cut-off frequency is at $100 rad/s$ and $Q = 1/2$.

[8] When $\omega_z > \omega_p$ the notch is called a low pass notch, while it is called a high pass notch is created when $\omega_p > \omega_z$

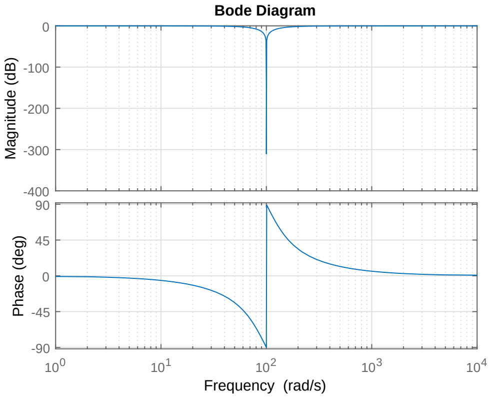

Figure 27: A Bode plot of a notch filter. The center frequency ω_0 is at $100 rad/s$ and the width of the rejected band ω_c is $50 rad/s$.

and ω_c is the width of the rejected band. An example is shown in figure 27.

References

[1] Gene F. Franklin, J. David Powell, and Abbas Emami Naeini. *Feedback Control of Dynamic Systems*. Pearson Prentice Hall, 2006.

[2] Norman S. Nise. *Regeltechniek voor Technici*. Wiley, 2005.

Appendices

A Laplace Transform table

Below is the Laplace Transform table, which shows the most important transformations of the time domain differential equation to the frequency domain. More comprehensive tables can be found on the internet.

f(t)	**F(s)**	**f(t)**	**F(s)**
$f'(t)$	$sF(s) - f(0)$	$sin(\omega t)$	$\frac{\omega}{s^2-\omega^2}$
$f^n(t)$	$s^nF(s) - s^{n-1}f(0) \cdots - f^{n-1}(0)$	$cos(\omega t)$	$\frac{s}{s^2-\omega^2}$
1	$\frac{1}{s}$	t^n	$\frac{n!}{s^{n+1}}$
$e^{a\,t}$	$F(s-a)$	$t\,e^{a\,t}$	$\frac{1}{(s-a)^2}$