Lecture 1

Introduction[1]

This first lecture has two parts. The first part gives an introduction and overview of the course, starting from two examples of modern control, a DVD-reader and car dynamics. The second part of the lecture is a brief review of linear input-output models in continuous time used in the basic course. The concepts of signal norm and system gain are introduced.

1.1 First example — a DVD player

The appearance of cheap sensors, actuators and computing devices opens new application areas for feedback control all the time, even in mass produced consumer products. The control technology is mostly hidden to the user, but still critical for operation and performance. A prime example of this is positioning of the pick-up head in a storage device as a DVD or CD-rom, where the speed of data recovery is directly correlated to the control performance.

A DVD (Digital Versatile Disk) is a data disk of the same physical size as a CD. Its use is mostly for video, but also for computer software as a large CD-ROM disk. The storage technology is in principle the same as for the CD, but improved. A CD holds about 650 megabytes of data whereas a DVD holds 4.7 gigabytes (for single layer, single side).

Figure 1.1 The right picture shows pits forming tracks on the DVD surface.

The disk surface is reflective, so that laser light is reflected back. Data bits are represented by pits of different lengths in *tracks* on the disk. These pits make the laser beam interfere destructively with itself, and therefore the pits look black to the laser.

The surface velocity is constant (about 3.5 m/s), meaning that the disc should rotate at different speeds depending on the current reading position. The challenge of the control problem is related to the fact that only 0.022 μm deviations from the

[1]Written by A. Rantzer with contributions by K.J. Åström, B. Lincoln and B. Wittenmark

bit-track can be accepted. At the same time, a disk is always slightly asymmetric, causing it to oscillate up to 100 μm per rotation, and the rotation speed is up to 23 Hz (for *single* speed). The tracking controller must compensate for this oscillation.

A typical DVD player has a *pick-up-head* consisting of a laser, an astigmatic lens, and a light detector with four fields – see Figure 1.2. The lens is mounted on springs in the axial (focus) and radial direction, and can be moved by electromagnets. This way, the laser spot can be moved very fast in a small range (a few hundred tracks sideways). The lens and laser are mounted on the *sledge*, which can move over the whole disk (in radial direction), but with much less precision and speed.

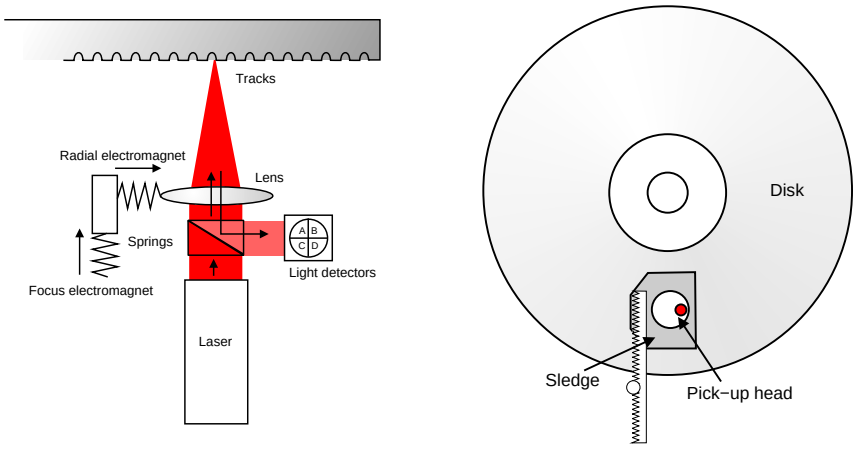


Figure 1.2 The pick-up-head has two electromagnets for fast positioning of the lens (left). Larger radial movements are taken care of by the sledge (right).

Four light detectors are available to estimate the focus error and radial error of the lens. Measurements are taken with a sampling frequency of 40 kHz and the DVD standard specifies that the speed of control (cross-over frequency) must be at least 2.4 kHz.

It turns out that most of the main topics of this course are relevant for the solution of the DVD control problem and we have therefore chosen to use it as a demonstrator. Both the focus control and the disc tracking will be treated in a case study in lecture 5.

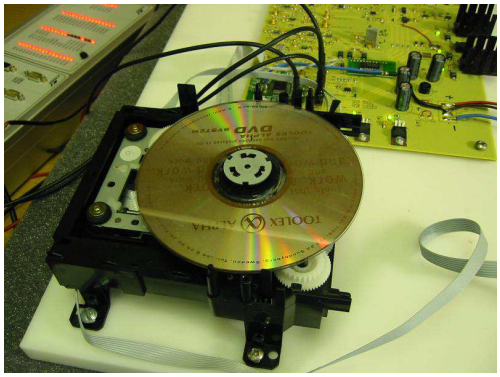


Figure 1.3 The DVD reader used in lecture 5.

Example: Midranging control

As an example towards multivariable control we will look at the pick-up-head of the DVD-player. The pick-up-head can be moved radially by two different actua-

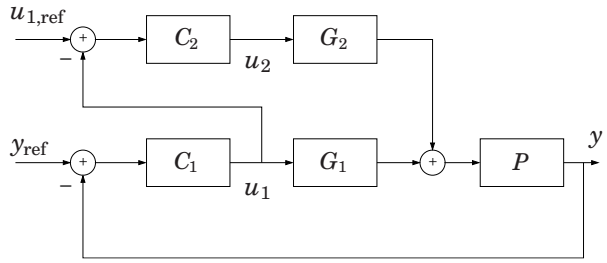


Figure 1.4 Mid-ranging controller: Used to control a process P with one output and two actuators (G_1, G_2).

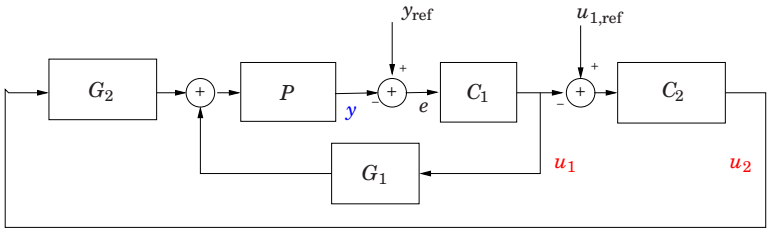


Figure 1.5 Alternative block diagram for the midrange controller in Fig. 1.4.

tors; the sledge on which it is mounted (large radial motion but relatively slow) and by one of the electromagnets (limited range but fast actuation), see Fig. 1.2. For the radial positioning of the head we can view it as a system with one output and two control signals. Similar situations appear in aerospace control where we may have many control surfaces ("rudders"), in car dynamics (braking on different wheels) and in process industry where it is common to place a large, slow valve in parallel with a fast, small valve. One control structure often used for this kind of actuation is *mid-ranging control*, see Fig.1.4

The control error $e = y_{\text{ref}} - y$ is used as input to the "fast controller"/"fast actuator", $C_1(s)$ and $G_1(s)$, respectively. This corresponds to the control of the electromagnet in the DVD-player case. As the electromagnet can only move the pick-up-head a small distance we would like to have it in the middle of its operating range to be able to react fast in both positive and negative directions during normal operation. By comparing u_1 and $u_{1,\text{ref}}$, typically $u_{1,\text{ref}} = 50\%$, and using this deviation as input to the slow controller circuit $G_2(s)C_2(s)$, which in our case corresponds to control of the sledge, the motion of the sledge will add an offset to the output of G_1 (takes care of large but slow variations). This means that after a fast position compensation using G_1C_1, the control signal u_1 can be brought back to 50% of $u_{1,max}$, its middle position, by moving the sledge, and the electromagnet G_1 can be ready to react on fast changes in both directions (should not be saturated).

Remark: Mid-ranging control can be seen as a dual to cascade control where you have two outputs and one control input [Hagander *et. al.*]

Similar tuning rules as for the cascade controller applies for the mindrange controller, see Fig. 1.5.

- First tune the fast inner loop, then the slower outer loop
- Controllers have separate time scales to avoid interaction

Warning: When you have two or more actuators in parallel, do NOT use parallel integral action! Exercise: Explain why this will cause problems with e.g. drift and pole-zero cancellations.

1.2 Second example — Control of car dynamics

A modern car contains numerous micro-processors devoted to feedback control. For example, feedback from oxygen sensors in the exhaust gas are needed for proper operation of the engine and catalyzer. This is essential for fuel efficiency and to reduce the emission of polluting exhaust gases.

Other feedback loops are used to improve safety, by controlling the brakes to prevent wheel-locks and to prevent skidding on slippery roads. A simplified model for car dynamics is given by the state space description

$$\begin{bmatrix} \dot{V} \\ \dot{r} \end{bmatrix} = A \begin{bmatrix} V \\ r \end{bmatrix} + \begin{bmatrix} 0 \\ b_1 \end{bmatrix} (u_1 + u_2 - u_3 - u_4) + \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} \delta$$

where V is lateral speed and r is angular velocity. There are five control signals, the steering angle δ and the brake forces u_1, u_2, u_3 and u_4 on the four wheels.

Figure 1.6 A modern car relies on feedback control for comfort, safety and fuel efficiency. The left picture shows a test-car used in a research project together with DaimlerChrysler.

The state is generally not available for direct measurement. Even if the angular velocity of each wheel can be measured, there is always some discrepancy between the rotational speed and the speed over ground. Hence the velocity of the car must be estimated based on information from several sources and the remaining uncertainty must be taken into account in the control algorithms.

A typical sampling frequency for speed measurements is a few milliseconds. This may sound fast enough compared to typical car dynamics, but when the purpose is to prevent wheel-lock or accidents, a delay of a few milliseconds can in fact be a severe obstacle for proper control performance.

1.3 Course overview

The objective of the course is that the students should learn the basic principles for control of systems with multiple inputs and outputs. A schematic picture of such a system is given below.

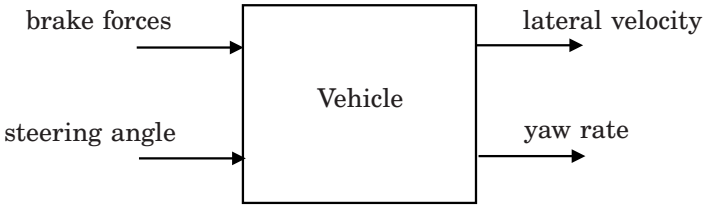
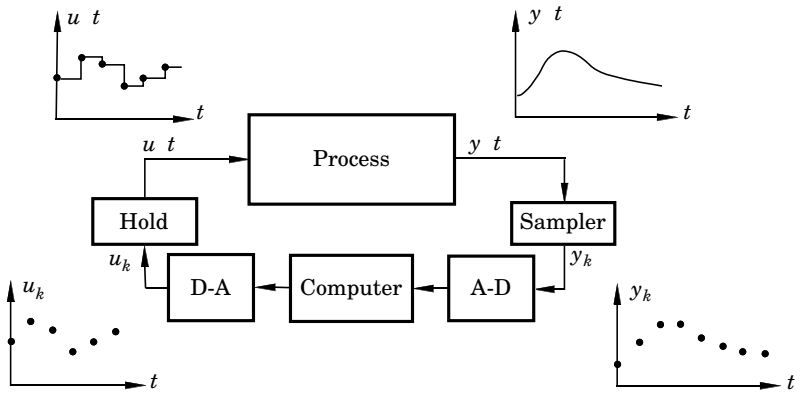


Figure 1.7 Input-output diagram for car dynamics control.



The control signal $u(t)$ is determined using measurements of $y(t)$ to achieve desired behaviour of the process. For the DVD player, $y(t)$ would be a vector of four variables, representing intensities in the four light detectors in Figure 1.2, while $u(t)$ would correspond to the two electromagnets.

It is important to note that the dynamics of a real process is never known exactly. Neither is it possible to precisely state the "true design objectives". It is therefore necessary to maintain a broader perspective on the engineering design problem, see Figure 1.8.

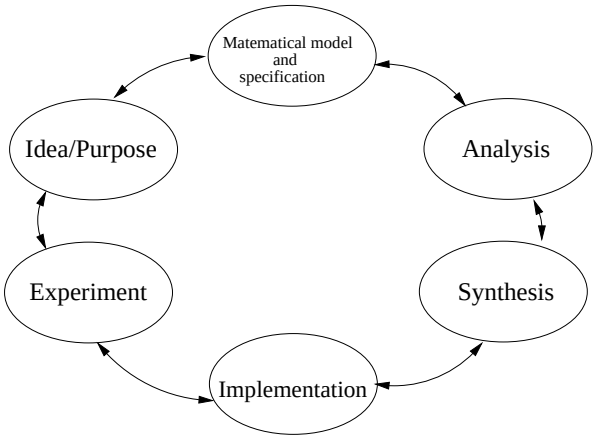


Figure 1.8 Schematic overview of the design process

Everything starts with an idea about the purpose of the control task. In simple cases, it is possible to directly come up with a solution proposal that can be tested experimentally and be accepted, possibly after minor modifications. However, in a vast number of applications costs and time can be reduced by analyzing or simulating a mathematical model before trying real experiments. The purpose of the diagram is to illustrate this methodology. Note that the arrows point in two direc-

tions. Failure in the experimental phase could not only require reimplementation, but also new analysis, more accurate models, or even redefinition of the control purpose.

Imagine stepping through the diagram in order to design a controller for car dynamics as in the previous example. Suppose that a controller has been synthesized based on the given two state model. Implementing a controller on a prototype car is costly, so a second step would typically involve computer simulation. For this purpose a more complex and accurate car model is needed, a model that is less transparent from a synthesis perspective but better suited to reveal the deficiencies of a proposed controller. If the simulations fail, a reason could be that the two state model was too simple and that additional features need to be taken into account in the synthesis phase. After a sequence of attempts, one could hope to find a solution ready for experimental tests. Alternatively, persisting failures could be an indication that the original goal was overly optimistic and impossible to achieve.

The main focus of this course is on the analysis/synthesis phase of the diagram in Figure 1.8 with particular emphasis on linear multi-input-multi-output systems. The outline and main topics of the course are the following

- Design of scalar controllers
- Stability and Robustness
- Fundamental system limitations
- Multi–input-multi–output systems
- Control design (LQ and LQG)
- Synthesis by convex optimization

Related courses on real-time control and implementation aspects, modelling, system identification and nonlinear control can be found on
`http://www.control.lth.se/education`.

During the first five lectures we look at the basic control loops and see how outputs and control signals are affected by reference values and disturbances, similar to what was done in the basic course. We build on material from that course, but make a deeper study of robustness and performance evaluation in controller design. The first lab exercise is aimed to give practical training in scalar controller design by frequency domain loop shaping.

After this, we start with multivariable systems, look at poles, zeros, observability, controllability, realizations etc. and discuss how the previous ides can be applied to systems with several inputs and outputs. There is a short intermezzo were we look at fundamental limitations in controller design and we also look how some multivariable control problems can be transfered to simpler control problems (decentralized and decoupled control). The second lab exercise will deal with multivariable control of a system where the fundamental limitations play an important role.

The second part of the course, lectures 9-14, continues along the lines of the textbook and bring in the subject of optimization for controller design and synthesis. The theory of linear quadratic (LQ) optimal control and Kalman filtering is a cornerstone of modern control. It clarifies fundamental relationships between measurement accuracy, control authority and achievable performance. Multivariable systems also fit in very nicely. Computer tools and new optimization algorithms have come to play an increasingly important role. Some recent research results developed at the department are taught in this section, in particular on control synthesis based on convex optimization.

Finally, the course is concluded by a lab exercise devoted to crane control. Most of the main topics in the course are relevant for a successful solution to this problem.

1.4 Linear input-output maps

In this section, we will review some different ways of specifying the input-output relationship of a finite-dimensional linear time-invariant system. This is a system that can be described by a state space equation

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

The differential equation has the solution formula

$$y(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t)$$

Note that the formula remains valid for multivariable systems, i.e. when both $u(t)$ and $y(t)$ are vector valued.

The map from u to y is linear provided that $x(0) = 0$. Introducing the *impulse response* $g(t)$ as

$$g(t) = \int_0^t Ce^{A(t-\tau)}B\delta(\tau)d\tau + D\delta(t) = Ce^{At}B + D\delta(t)$$

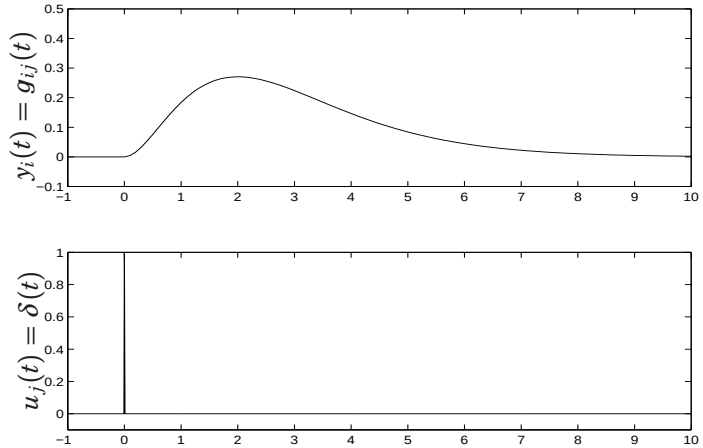
the input-output map can be written as a convolution

$$y(t) = \int_0^t g(t-\tau)u(\tau)d\tau = [g * u](t)$$

In frequency domain, the convolution becomes multiplication

$$Y(s) = G(s)U(s)$$

and the Laplace transform of the impulse response is equal to the *transfer function* $G(s) = C(sI - A)^{-1}B + D$. For multivariable systems, both $g(t)$ and $G(s)$ are matrices. The term impulse response is of course motivated by the fact that the matrix element $g_{ij}(t)$ is the value of output i obtained when input j is an impulse (Dirac function) at $t = 0$. This is sometimes used to determine $g(t)$ experimentally.



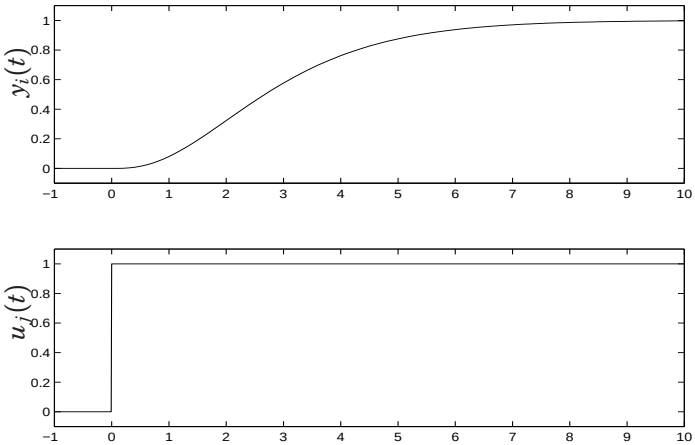
A more common experiment in process industry is the *step response*. Assuming that the (possibly vector valued) input is a step

$$u(t) = \begin{cases} 0 & t < 0 \\ u_0 & t \geq 0 \end{cases}$$

the output becomes

$$y(t) = \int_0^t g(t-s)u_0 ds = \left(\int_0^t g(\tau) d\tau \right) u_0$$

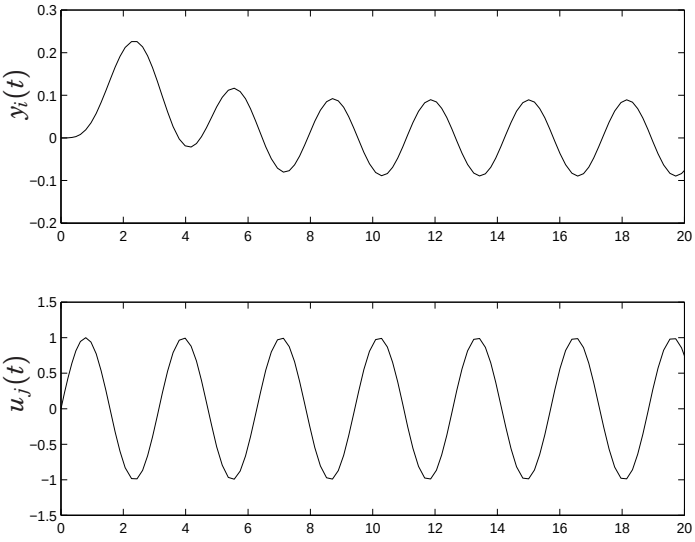
Accordingly, the Laplace transform of the step response is $G(s)u_0/s$.



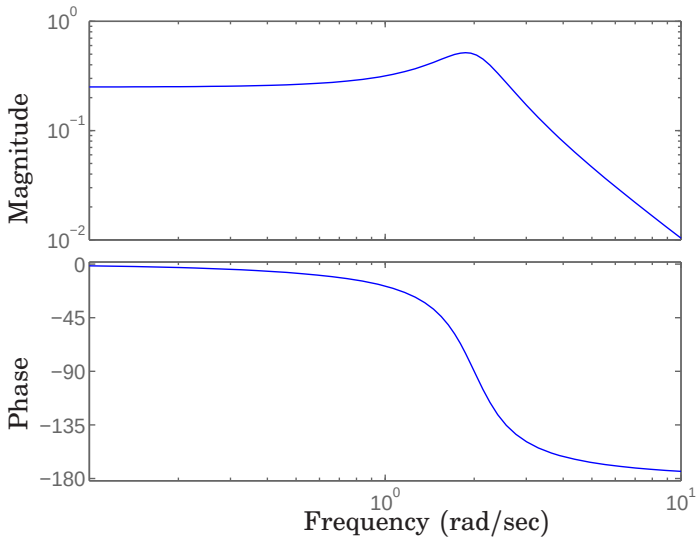
The main use of the Laplace transform is however to characterize the *frequency response*. The input $u(t) = u_0 \sin \omega t$ gives

$$y(t) = \int_0^t g(\tau)u(t-\tau)d\tau = \mathrm{Im}\left[\int_0^t g(\tau)e^{-i\omega\tau}d\tau \cdot e^{i\omega t}u_0 \right]$$

The integral approaches $G(i\omega)$ as $t \to \infty$, so after a transient, also the output becomes sinusoidal and $y(t) = \mathrm{Im}\left(G(i\omega)e^{i\omega t}\right)u_0$. To summarize, a linear time-invariant system always gives a sinusoidal response to a sinusoidal input. For a scalar system, the gain and phase shifts are determined by the amplitude and phase of the complex number $G(i\omega)$.



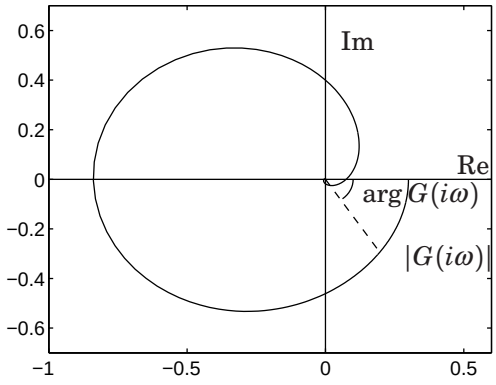
There are several ways to graphically illustrate the transfer function $G(i\omega)$. One is to plot the amplitude and phase separately versus the frequency. This is called the *Bode diagram*:



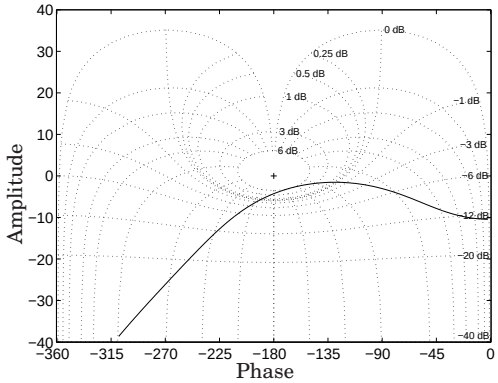
It should be noted that each additional factor in the transfer function contributes additively to the Bode plots:

$$\log|G_1G_2G_3| = \log|G_1| + \log|G_2| + \log|G_3|$$
$$\arg G_1G_2G_3 = \arg G_1 + \arg G_2 + \arg G_3$$

The *Nyquist diagram* is obtained by plotting $G(i\omega)$ directly in the complex plane for different values of ω:



If instead $\log G(i\omega) = \log|G(i\omega)| + i\arg G(i\omega)$ is plotted in the complex plane, the *Nichols diagram* is obtained:



The level curves of $|G/(1+G)|$ and $\arg G/(1+G)$ are plotted as dotted lines to support use of the diagram in controller design.

1.5 Signal norm and system gain

In order to efficiently analyze and optimize dynamical systems, it is useful to have mathematical notions that measure the size of a signal and the gain of a system. This is the reason for the following definitions.

The size of a signal $y(t) \in \mathbf{R}^n$ can be measured by the *L_2-norm*, defined as

$$\|y\|_2 := \sqrt{\int_0^\infty |y(t)|^2 dt}$$

According to a theorem known as Parseval's formula, the same norm can be defined in frequency domain as

$$\|y\|_2 = \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} |\mathcal{L}y(i\omega)|^2 d\omega}$$

For a system $\mathcal{S}$ with input u, output $\mathcal{S}(u)$ and zero initial state, the *L_2-gain* is defined as the largest possible fraction between the input norm and the output norm

$$\|\mathcal{S}\| := \sup_u \frac{\|\mathcal{S}(u)\|}{\|u\|}$$

The system is called *input-output stable* (or *L_2-stable*) if its L_2-gain is finite. For example, a time delay does not change the signal norm, so it has gain one. However, an integrator has infinite gain, since an input $u(t)$ that is identically zero for $t \geq 1$, can give an output $y(t)$ that is a nonzero constant for $t \geq 1$. Hence, the fraction $\|y\|_2/\|u\|_2$ can be arbitrarily large.

More generally, the L_2-gain of a system can be obtained as the maximum amplitude in the Bode diagram:

THEOREM 1.1
A stable system with transfer function $G(s)$ has the L_2-gain

$$\|G\|_\infty := \sup_\omega |G(i\omega)|$$

□

Remark. For multivariable systems the $|G(i\omega)|$ should be interpreted as the matrix norm (the largest singular value) of $G(i\omega)$. This case will be studied more carefully later.

Proof. Let y be the output corresponding to the input u. Then

$$\|y\|^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\mathcal{L}y(i\omega)|^2 d\omega \leq \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(i\omega)|^2 \cdot |\mathcal{L}u(i\omega)|^2 d\omega \leq \|G\|_\infty^2 \|u\|^2$$

The inequality is arbitrarily tight when $u(t)$ is a sinusoid near the maximizing frequency. □

Example 1

a. For a time delay $G(s) = e^{-sT}$ we have $|G(i\omega)| \equiv 1$.

b. For an integrator $|G(i\omega)| = |\frac{1}{i\omega}| = \frac{1}{\omega}$ which is unbounded $\omega = 0$.

c. The Bode diagram plotted in the previous section has a peak magnitude about 0.5 at the frequency 2 rad/sec. Hence, the L_2-gain of the corresponding system is smaller than one and the highest gain is obtained for an input sinusoid of this frequency. □