Control System Design
Lecture notes for ME 155A

Karl Johan Åström

Department of Mechanical & Environmental
Engineering
University of California
Santa Barbara

Department of Automatic Control
Lund Institute of Technology
Box 118
SE-221 00 LUND
Sweden

Contents

1

Introduction

1.1 Introduction

Control systems are ubiquitous. They appear in our homes, in cars, in industry and in systems for communication and transport, just to give a few examples. Control is increasingly becoming mission critical, processes will fail if the control does not work. Control has been important for design of experimental equipment and instrumentation used in basic sciences and will be even more so in the future. Principles of control also have an impact on such diverse fields as economics, biology, and medicine.

Control, like many other branches of engineering science, has developed in the same pattern as natural science. Although there are strong similarities between natural science and engineering science it is important to realize that there are some fundamental differences. The inspiration for natural science is to understand phenomena in nature. This has led to a strong emphasis on analysis and isolation of specific phenomena, so called reductionism. A key goal of natural science is to find basic laws that describe nature. The inspiration of engineering science is to understand, invent, and design man-made technical systems. This places much more emphasis on interaction and design. Interaction is a key feature of practically all man made systems. It is therefore essential to replace reductionism with a holistic systems approach. The technical systems are now becoming so complex that they pose challenges comparable to the natural systems. A fundamental goal of engineering science is to find system principles that make it possible to effectively deal with complex systems. Feedback, which is at the heart of automatic control, is an example of such a principle.

A simple form of feedback consists of two dynamical systems connected in a closed loop which creates an interaction between the systems. Simple

causal reasoning about such a system is difficult because, the first system influences the second and the second system influences the first, leading to a circular argument. This makes reasoning based on cause and effect difficult and it is necessary to analyze the system as a whole. A consequence of this is that the behavior of a feedback system is often counterintuitive. To understand feedback systems it is therefore necessary to resort to formal methods based on mathematics.

Feedback has many advantages. It is possible to create linear behavior out of nonlinear components. Feedback can make a system very resilient towards external influences. The total system can be made very insensitive to external disturbances and to variations in its individual components. Feedback has one major disadvantage, it may create instability, which is intrinsically a dynamic phenomenon. To understand feedback systems it is therefore necessary to have a good insight into dynamics.

The wide applicability of control has many advantages. Since control can be used in so many different fields, it is a very good vehicle for technology transfer. Ideas invented in one field can be applied to another technical field.

Control is inherently multidisciplinary. A typical control system contains sensors, actuators, computers and software. Analysis of design of control systems require domain knowledge about the particular process to be controlled, knowledge of the techniques of control and specific technology used in sensors and actuators. Controllers are typically implemented using digital computers. Knowledge about real time computing and software is therefore also essential. Sensors and actuators are often often connected by communication networks. This implies that knowledge about communication is also important. In the future we can see a convergence of the technologies of control, computing and communication.

Team work is essential in control because of the wide range of technologies and techniques involved. Education in control has proven to be an excellent background when working with complex engineering systems. The interdisciplinary nature of control has created some difficulties for educators. Education and research in engineering grew out of specific technologies such as mining, building of roads and dams, construction of machines, generation and transmission of electricity, and industrial use of chemistry. This led to an organization of engineering schools based on departments of mining, civil engineering, mechanical engineering, electrical engineering, and chemical engineering etc. This served very well in the end of the 19th century and the beginning of the 20th century. The situation changed significantly with the advent of fields like control, that cut cross traditional department boundaries. Industry has adapted quickly to the new demands but academia has not.

There are many reasons why an engineer should know control. First

of all because practically all engineers will use control, and some will design control systems. But the most important reason is that control is an essential element of practically all engineering systems. It happens too often that systems perform poorly because they are designed from purely static analysis with no consideration of dynamics and control. This can be avoided by engineers being aware of control even if they are not specialists. Control can also give designers extra degrees of freedom. It is in fact a very powerful tool for designers of all systems. Cars are typical examples. The stringent requirements on emission were solved by controlling the combustion engines. Other examples are anti-lock braking systems (ABS) and systems for traction control. Other reasons to study control is that there are many beautiful theoretical results and some really neat devices.

Control has for a long time been confined to engineering but it is increasingly clear that the ideas and concepts have a much wider use. The concepts of feedback and control are thus essential in understanding biological and economical systems. We illustrate this with a quote from the book Way Life Works : The Science Lover's Illustrated Guide to How Life Grows, Develops, Reproduces, and Gets Along – by Mahlon Hoagland, Bert Dodson.

> Feedback is a central feature of life: All organisms share the ability to sense how they are doing and to make changes in "mid-flight" if necessary. The process of feedback governs how we grow, respond to stress and challenge, and regulate factors such as body temperature, blood pressure and cholesterol level. This apparent purposefulness, largely unconscious, operates at every level - from the interaction of proteins in cells to the interaction of organisms in complex ecologies.

It is thus reasonable to claim that control not only makes our lives more comfortable it is also essential for our existence.

The rest of this chapter gives a brief history of the development of the field of control. The richness of the field is then illustrated by a number of examples from a wide range of applications ranging from industrial applications to biology.

1.2 A Brief History

Although there are early examples of the use of feedback in ancient history, the development of automatic control is strongly connected to the industrial revolution and the development of modern technology. When

new sources of power were discovered the need to control them immediately arose. When new production techniques were developed there were needs to keep them operating smoothly with high quality.

The Centrifugal Governor

The centrifugal governor is one of the most celebrated early feedback systems. It was first used to control the speed of windmills. Later around 1788 it was used by James Watt to control the velocity of steam engines. A textile mill is a typical application, one steam engine drives several spinning wheels and looms. The power from the steam engine is transmitted to the spinning wheels and looms via belt-drives. It is highly desirable to keep the speed of the steam engine constant because changes in speed may cause thread breaks and require adjustments of the looms. It was observed that engine speed changed with changes in the load, for example when a loom was connected to the drive belt. The centrifugal governor was introduced in order to keep the speed constant. Figure 1.1 shows a typical system. It consists of two balls hinged on a rotating shaft which is connected to the output shaft of the steam engine. When the speed increases, the balls swing out. This motion is connected to the valve which admits steam into the engine via mechanical links. The connection is made in such a way that steam flow increases when the velocity decreases. The system is a feedback system because changes in the velocity are fed back to the steam valve. The feedback is negative because the the steam supply is increased when the velocity decreases.

The improvement obtained when using a centrifugal governor is illustrated in Figure 1.2. The figure shows that the velocity drops when an additional loom is connected to the drive belt. The figure also shows that the velocity drop is significantly smaller when a centrifugal governor is used.

It is possible to change the characteristics of the governor by changing the mechanism that transmits the motion of the balls to the steam valve. To describe this we introduce the notion of gain of the governor. This is the ratio of the change in steam valve opening (Δu) to the change in the angle (Δv) of the velocity. See Figure 1.2 which shows how the velocity responds to changes in the load. The figure shows that the largest velocity error decreases with decreasing gain of the controller but also that there is a tendency for oscillations that increases with increasing gain. The centrifugal governor was a very successful device that drastically simplified the operation of steam driven textile mills.

The action of the basic centrifugal governor can crudely be describe with the equation

$$u = k(V_r - V) + b$$

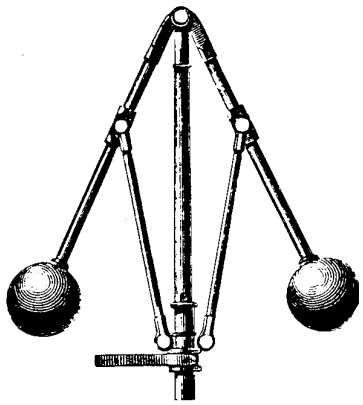


Figure 1.1 The centrifugal governor, which has been used to control the speed of engines since the beginning of the industrial revolution. When the axis spins faster the balls move away from the axis. The motion is transfered to the engine to reduce its power. The governor has also become an icon of the field of control.

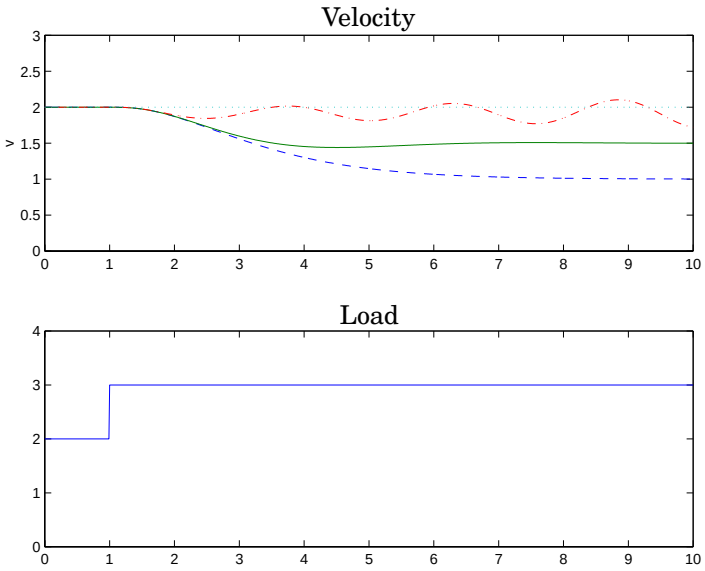


Figure 1.2 Response of the velocity to changes in the load of an engine controlled by a governor with different values of the gain, $k = 0$ (dashed), $k = 1$ (full), $k = 12$ (dash-dotted).

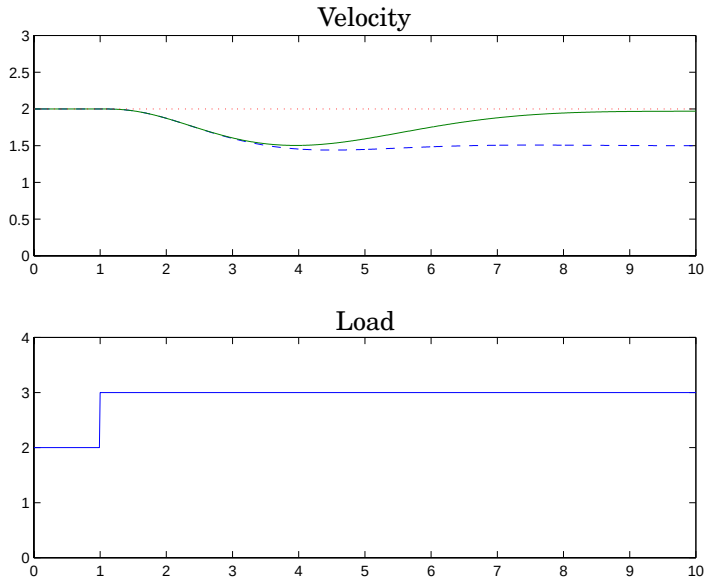


Figure 1.3 Response of velocity to changes in the load of an engine controlled by a governor having proportional control (dashed line) and PI control (full line).

where u is the opening of the steam valve, V_r is the desired speed, V the actual speed and k and b are constants. This is called a proportional controller because the control action u is proportional to the error. The parameter b is a bias term that was adjusted manually to make sure that the V was equal to V_r. Siemens made a clever invention which eliminated the bias adjustment. His governor can mathematically be described by the equation

$$u = k(V_r - V) + k_i \int_0^t (V_r(\tau) - V(\tau))d\tau \tag{1.1}$$

the bias term was thus replaced by a term proportional to the integral of past errors. A controller described by (1.1) has the amazing property that the velocity V is always equal to the desired velocity V_r in steady state. This is illustrated by Figure 1.3 which shows the behavior of an engine with control actions proportional to the integral of the error. In standard terminology the Siemens governor is called a PI controller, indicating that the control action is proportional to the error and the integral of the error. Integral action has some amazing properties that will be discussed further in Section 2.2.

The Emergence of Control

Apart from the centrifugal governor the early applications of control include, autopilots for ships and aircrafts, the electronic feedback amplifier and process control. The fundamental similarities between the different systems were not noticed. Control emerged around 1945 as a result of intensive military research in the period 1940-1945. During the Second World War it became apparent that science and technology could have a major impact on the war effort. Among the goals were development of radar and fire control systems. Control was a central feature element of these systems. In many countries groups of scientists and engineers were gathered in research institutes. It was realized that feedback was essential for such diverse systems as autopilots for ships and aircrafts, electronic amplifiers, systems for orientation of radar antennas and guns, and industrial production of uranium. Control was born out of this multidisciplinary effort.

The first manifestation of control was called servo-mechanism theory. This theory used block diagrams as a tool for abstraction. This clearly showed the similarity between the widely different systems. The mathematical tools were Laplace transforms and the theory of complex variables. Analog computers were used for simulation and controllers were implemented as analog computers. It is significant that one of the first books was edited by three persons, a mathematician, a physicist and and engineer from a control company.

One factor that strongly contributed to the emergence of control was that many of the results from the military efforts were disseminated very quickly. After the war it became apparent that control was very useful in practically all branches of engineering and the ideas spread like wild fire around the world. One result was that education in control was introduced as an essential element of engineering education practically all over the world. A characteristic feature of control is that it is not confined to a particular branch of engineering such as mechanical, electrical, chemical, aeronautical, and computer. Control appears in all disciplines, it is in fact the first systems science.

In the late 1950's there were organizational activities which created meeting places for researchers and practitioners of control. Conferences were organized and journals on control started to appear. The International Federation of Automatic Control (IFAC) was the key organization but the traditional engineering organization also introduced special interest groups on control. The first World Congress of IFAC in Moscow in 1960, where control engineers and scientists from all over the world met for the first time, was a landmark. Another effect was the industrialization of control which created companies specializing in control equipment.

The Second Wave

Any field could have been proud of the development that took place from 1940 to 1960, a second wave of development starting in the late 1950s. The driving forces were new challenging applications and a strong stimulation from mathematics. The major drivers for the development were the space race and the use of digital computers for industrial process control. Notice that Sputnik was launched in 1957 and the first computer installed to control an oil refinery started on-line control in 1959. There was a very vigorous development of both theory and practice of control. Digital computers replaced analog computers both for simulation and implementation of controllers. A number of subspecialities of control were also developed. It turned out that a wide range of mathematics fitted the control problems very well, and a tradition of a high respect for mathematical rigor emerged.

Today control is a well established field with a solid body of theory and very wide applications. Practically all controllers are today implemented in digital computers. There are also many challenges. Control is increasingly becoming mission critical which means that great attention has to be given to safety and reliability. The complexity of the control systems are also increasing substantially.

There are tremendous challenges in the future when we can visualize a convergence of control, computing and communication which will result in large interconnected systems. It is also reasonable to guess that a deeper understanding of control in the field of biology will be very exciting. In this sections we will present a number of applications together with some historical notes. The idea is to illustrates the broad applicability of control and the way it has impacted on many different fields.

1.3 Process Control

Many products that are essential for us in daily life are manufactured by the process industries. Typical examples are oil, petrochemical, paper, pharmaceuticals. These industries use continuous manufacturing processes. Process control is an essential part of these industries, the processes can not be run without control systems. We will start by discussing the centrifugal governor.

Advances in Process Control

Key issues in process control are to:

- Keep essential quality variables at specified values.

Figure 1.4 Microprocessor based single loop controller. By courtesy of ABB Industrial Systems.

- Minimize use of energy and raw material.
- Make rapid changes of production or grades.

The control systems are key factors to obtain these goals. There have been major advances since the centrifugal governor appeared.

In the centrifugal governors the actions of sensing, computing and actuation were executed by purely mechanical devices. A lot of ingenuity went into the designs which were often patented. Feedback actually had a crucial role in the design of the controllers. By using feedback it was possible to construct controllers that had a stable well defined behavior from components with a large variability. The technology of controlling engine speed by governors was applied to all types of engines. When electricity emerged in the end of the 19th century there was a similar need to control the speed of generator for hydroelectric generators. Since there was little communication between different fields ideas like integral action were reinvented several times. Major advances were made in the 1930s and 40s. Controllers then appeared as special devices separated from sensors and actuators. New industries , such as Fisher Control, Foxboro, Honeywell, Leeds and Northrup, Taylor Instruments, which supplied sensors, actuators, controller and complete system emerged. The controllers were implemented in different technologies, mechanic, pneumatic and electronic. Controllers for many loops were located in central control rooms. The technology of using digital computers for process control emerged in the 1960s. Computer control is the standard technology for process control.

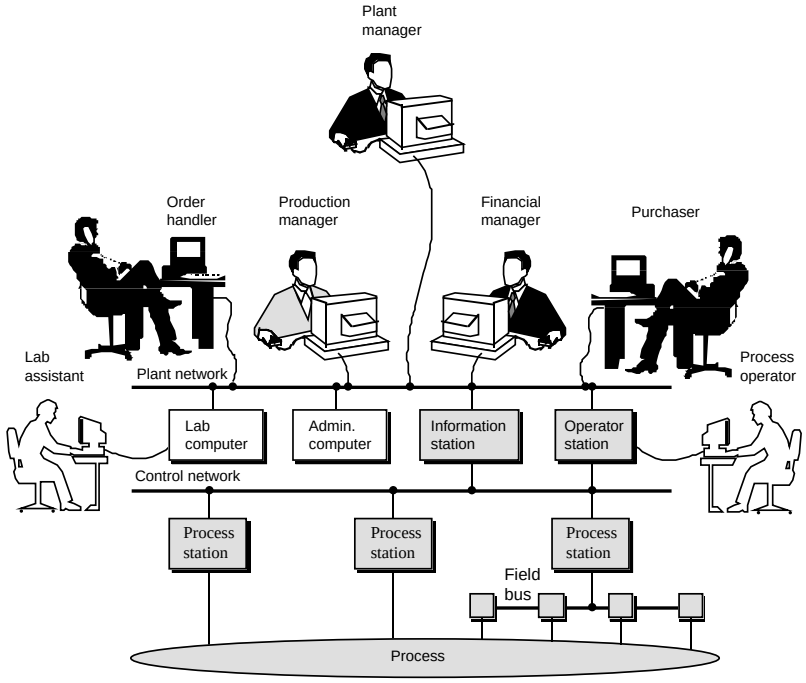


Figure 1.5 Modern industrial systems for process control, like the Advant OCS tie computers together and help create a common uniform computer environment supporting all industrial activities, from input to output, from top to bottom. (By courtesy of ABB Industrial System, Västerås, Sweden.).

New companies have often emerged when major technology changes occurred, there have also been many mergers. Today there are a few large companies that supply control world wide, ABB, Honeywell, Siemens and Toshiba.

The processes vary significantly in scale, a small process unit may have 20-100 control loops but a complete paper mill may have up to several thousand control loops. A wide variety of systems are used for control. There are microprocessor based controllers for single or multiple loops, see Figure 1.4, systems based on personal computers (PC), programmable logic controllers (PLCs) and distributed control systems consisting of many computers connected in a network, see Figure 1.5 Large plants may have 5000-10000 control loops, organized in an hierarchical structure. Most (about 90%) of the lower level control loops are still PID control, which is implemented in the digital computer.

The Role of Sensors

Sensors are key elements of control systems because they provide information about the process. The variables of primary interest in process control are closely related to product quality. Unfortunately it is very difficult to measure such variables on-line. Control is therefore often done indirectly by controlling secondary variables such as level, temperature and pressure. The on-line measurements are often augmented by laboratory analysis of samples of the product.

There are interesting synergies between development of sensors and control. Feedback is often used in sensors. When new sensors become available there are opportunities to develop new control systems. One example is the emergence of sensors for on-line measurement of basis weight and moisture in the 1960s which led to emergence of two new companies Accuray and Measurex which specialized in control of paper machines. There is a similar window of opportunity today because new sensors for measuring composition and surface structure based on infrared and near infrared spectroscopy are available.

1.4 Manufacturing

Process control is continuous manufacturing. Feedback has also had a major impact on manufacturing of discrete parts. Numerically controlled machine tools developed at the Control Systems Laboratory at MIT in the 1950s was a first step where control was used to improve precision of mechanical machining. Welding is highly automated using control and vision systems. Machines for manufacturing systems based on machining with lasers and electrical arcs depend heavily on use of control.

Large manufacturing operations are made on transfer lines, where the parts are moved along a line to stations which perform the operations. The individual stations do operations such as drilling, machining and polishing. A line for making car engines has 10 to 20 machines, which are separated by buffers. Each machine has around 10 stations. The there are simple continuous control systems in each station for tasks such as positioning. A complete transfer line has a few hundred feedback loops. The major control problem is, however, the safe control of the whole operation, for example, to determine when a part should be moved, when a drilling operation should start, when a tool should be removed. This is a discrete control problem where the information comes from on-off signals such as limit switches. The control actions are also discrete, to start a motor, to start drilling, to stop drilling, to change a tool. The systems are very complex because of the large number of signals involved. A single

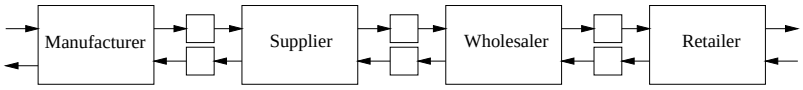


Figure 1.6 Schematic diagram of a simple supply chain consisting of manufacturing, suppliers, wholesaler and retailer, with delays between the different operations.

station in a transfer line may have 5 to 10 000 discrete inputs and output and 10 to 20 continuous control loops. Logic control was originally done by relay systems. When microprocessors were developed in the 1970s the relays were replaced by programmable logic controllers (PLCs). Design of the discrete control system is typically done in an ad hoc manner based on past experience. A general solution is still missing.

Supply Chains

There are also other uses of control in manufacturing, namely control of the complete business operation and complete supply chains. Manufacturing and distribution involve large flows of materials. Raw materials have to be supplied to the manufacturer, the products have to be supplied to the consumers. This is illustrated in Figure 1.6 which shows a simple supply chain. The diagram shows only a few levels, in practice a system may contain thousands of outlets and storages and it may deal with a very large number of products. Several important functions in the system like quality control, sales, and production management are not shown in the figure.

Manufacturing facilities and sales operations have traditionally been quite inflexible in the sense that it is difficult to change products and production rates. In such a system it is necessary to have many buffer storages to match production rates to fluctuating sales and supplies of raw materials and parts from sub-contractors. Production rate is largely determined in an open loop manner based on prediction of sales. An elaborate bookkeeping system is also required to keep inventories of parts.

A different system is obtained if the production time can be reduced and if the production can be made so flexible that products can be changed rapidly. It is then possible to produce a part when an order is obtained. Such a system may be regarded as a closed loop system where production is determined by feedback from sales. An advantage of such a system is that inventories can be vastly reduced. Administration of the system is also greatly simplified because it is no longer necessary to keep track of many products and parts in storage. A system of this type contributed significantly to the efficiency of Walmart. They relied on human sensors in the form of managers who analyzed the sales when the shops were

Figure 1.7 Remote robot surgery using the ZEUS system from Computer Motion Inc. The doctors on the left are in New York and the patient and the robots are in Strasbourg, France. Courtesy of Computer Motion Inc. Goleta

closed every evening to make new orders.

1.5 Robotics

The origin of the industrial robot is a patent application for a device called Programmed Article Transfer submitted in 1956 by the engineer George Devol. The robotics industry was created when Devol met Joseph Engelberger and they founded the company Unimation. The first industrial robot, Unimate, was developed by in 1961. A major breakthrough occurred in 1964 when General Motors ordered 66 machines Unimate robots from Unimation. In 1998 there were about 720 000 robots installed. The majority of them, 412 000 are in Japan. Robots are used for a wide range of tasks: welding, painting, grinding, assembly and transfer of parts in a production line or between production lines. Robots that are used extensively in manufacturing of cars, and electronics. There are emerging applications in the food industry and in packaging. Robots for vacuuming and lawn moving as well as more advanced service robots are also appearing.

Robots are also started to be used used in medical applications. This is illustrated in Figure 1.7 which shows robots that are used for an endoscopic operation. One advantage of the system is that it permits doctors to work much more ergonomically. Another advantage is that operations can be done remotely. The system in the figure is from a robot operation, where the patient was in Strasbourg, France and the doctors in New York.

Another aspect of robots

The word robota is a Slavic word that means work, often meaning of slave work. It came into prominence in a novel from 1918 and a play from 1921 by the check author Karel Capek called *Rossum's Universal Robots*. The play is about robot workers who revolt and kill their human ruler. Another literary aspect of robot is the famous robot trilogy by Isaac Asimov from 1940 who introduced the three laws of robotics

First Law: A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

Second Law: A robot must obey orders given it by human beings, except where such orders would conflict with the First Law.

Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Asimovs book captured the imagination of many and there is clear evidence that Engelberger was inspired by Asimovs writing. Other examples about imaginative speculations about robots are found in Arthur Clarkes book *2001 A Space Odyssey* where the robot HAL takes over the operation of a space ship and R2- D2 in the Star Wars series.

There are currently some very interesting developments in robotics. Particularly in Japan there is much research on humanoid and animaloid robots. There are very advanced humanoid robots in research laboratories, there are also robots that mimic snakes, cats, birds and fish. A robot dog AIBO and a service robot have been commercialized by Sony.

Design Issues - Task Based Control and Automomy

Design of robots is a typical multidisciplinary task which is a mixture of mechanical engineering, electronics, computer science, artificial intelligence, and control. It is a typical example of the challenges posed by new industries. Control systems are essential parts of a robot.

The control problems for industrial robots are servo problems, typically to control position of an arm or the force exerted by a tool. There are also other forms of feedback based on force sensors and vision.

Humanoid robots require a much more advanced task based control. It is necessary to provide them with functions for obstacle avoidance, path planning, navigation and map making. Since the robots have to be highly autonomous they must also have capabilities for learning, reasoning and decision making. Development of systems with such capabilities is a challenging research task.

1.6 Power

The power industry started to develop in the end of the 19the century and accelerated rapidly in the 20th century. Availability of power has improved quality of life tremendously. In 2000 the total amount of electric energy generated in the world was about about 15 000 TWh. It is expected to grow by a factor of 3 in 60 years where the main growth is outside the OECD countries.

Control is an essential element in all systems for generation and transmission of electricity. Many central ideas in control were developed in this context. When electricity emerged in the end of the 19th century the generators were typically driven by water turbines. Since alternating current (AC) was the preferred means for transmission there was immediately a need to control the speed of the generators to maintain constant frequency. Derivative and integral control appeared early as did stability critiera. The development paralleled the work on centrifugal governors but it was done independently and it had a stronger engineering flavor. One of the earliest books on control, with the title *Die Regelung der Kraftmaschinen*, was published by Tolle as early as 1905. It was discovered that the performance of hydroelectric power stations was severely limited by dynamics of the water duct. The power decreased rapidly initially when the valve was opened and then increased slowly. This property made the systems difficult to control. It is an example of what is now call a non-minimum phase dynamics.

As the demand for electricity grew many generators were connected in a network. These networks became larger and larger as more generators and consumers were connected. Figure 1.8, which is a schematic picture of the network for the Scandinavian countries, is an example of a network of moderate size. Sweden, Norway and Finland has much hydroelectric power, Sweden and Finland has nuclear power and Denmark has wind and thermal power. In Sweden the hydroelectric power is generated in the north, but most of the consumption is in the south. Power thus has to be transmitted over long lines. The system in the different countries are connected via AC and DC lines. There are also connections to Germany and Poland.

It is difficult to store energy and it is therefore necessary that production and consumption are well balanced. This is a difficult problem because consumption can change rapidly in a way that is difficult to predict. Generators for AC can only deliver power if the generators are synchronized to the voltage variations in the network. This means that the rotors of all generators in a network must line up. Synchronism is lost when the angles deviate too much.

Matching production and consumption is a simple regulation problem

Figure 1.8 The Nordel power grid which supplies electric energy for the Scandinavian countries. The squares represent hydroelectric stations and the triangles represent thermal stations, both nuclear and conventional, and circles denote transformers. The lines denote major power lines. Only the major components are shown in the system.

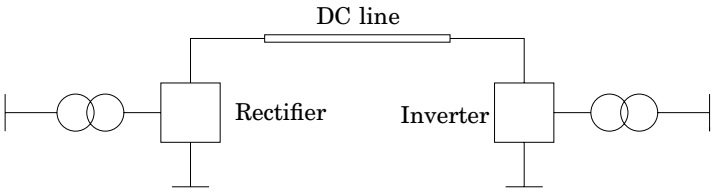


Figure 1.9 Schematic diagram of an HVDC transmission link. The system is fed from the right by AC which is converted to DC by the rectifier and transmitted over a DC line to the inverter which converts it to AC.

for one generator and one consumer, but it is a more difficult problem in a highly distributed system with long distances between consumption and generation. To have a reliable system it is highly desirable to avoid transmission of information over long distances. Control should therefore be done locally at each station based on the information available at the station. Several interesting control principles have been developed to do this. Control of each generator must be based on information that is locally available. Because of reliability requirements it is not possible to rely on information that is transmitted over wide distances.

High Voltage DC Transmission Systems

Even if most of the electricity is transmitted as AC there are some situation where it is advantageous to transmit direct current (DC). One example is when electric power has to be transmitted using under water cables. DC links can also be used to connect two asynchronous power grids. The systems for transmission of high voltage DC (HVDC systems) have many interesting properties. A schematic picture of such a system is shown in Figure 1.9. The AC is rectified to generate DC which is transmitted and converted to AC by the inverter. The rectifier and the inverter consists of semiconductor switches that permit very rapid changes of the power. It is possible to switch the direction of 600MW in fractions of a second. Control of such a system in a safe precise way is a great challenge.

Control of Networked Power Systems

An interconnected power system of the type shown in Figure 1.8 is a complicated system. The behavior of such systems is not easy to predict. Problems when interconnecting two systems were encountered by Edison. He found that it was not possible to connect two turbine driven generators when both generators had controllers with integral action. The system will drift so that one generator takes all the load. This was one of the

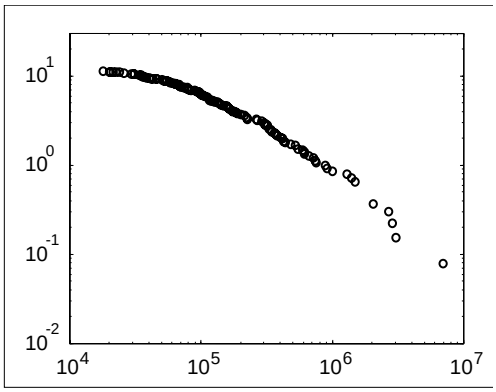


Figure 1.10 Power outages in the US 1984-97. The horizontal axis shows the number of persons N affected by the outages and the vertical axis shows the yearly frequency of outages that influence more than N persons. Notice that the scales on both axes are logarithmic.

first observations that problems may occur when several regulators are connected to an integrated system.

Edisons observation led to interesting developments of control theory. In current practice one large generator in the network controls the frequency using a controller with integral action. The other generators use proportional control. The amount of power delivered by each generator is set by the gain of the proportional controller. Each generator has separate voltage control.

There have been many other surprises in interconnected systems. In the Nordel system it has been observed that a moderated increase of power load in the north could result in large oscillations in the power transmission between Sweden and Denmark in the south. Oscillations have been observed when modern trains with switched power electronics have put in operation. An understanding of such phenomena and solutions require knowledge about dynamics and control.

Safety and Reliability

The power systems are generally very reliable. Customers will have power even when generators and lines fail. This is achieved by good engineering of the system based on redundancies. Networked generators contribute significantly to the reliability of the system because it is possible for a large number of generators to take up the load if one generator fails. The drawback is however that there may be massive failures in the system which also has occurred. This is illustrated in Figure 1.10 which shows the statistics of power failures.

1.7 Aeronautics

Control has often emerged jointly with new technology. It has often been an enabler but in some cases it has had a much more profound impact. This has been the case in aeronautics and astronautics as will be discussed in this section.

Emergence of Flight

The fact that the ideas of control has contribute to development of new technology is very nicely illustrated by the following quote from a lecture by Wilbur Wright to the Western Society of Engineers in 1901:

> " Men already know how to construct wings or airplanes, which when driven through the air at sufficient speed, will not only sustain the weight of the wings themselves, but also that of the engine, and of the engineer as well. Men also know how to build engines and screws of sufficient lightness and power to drive these planes at sustaining speed ... Inability to balance and steer still confronts students of the flying problem. ... When this one feature has been worked out, the age of flying will have arrived, for all other difficulties are of minor importance."

The Wright brothers thus realized that control was a key issue to enable flight. They resolved compromise between stability and maneuverability by building an airplane, Kitty Hawk, that was unstable but maneuverable. The pioneering flight was in 1905. Kitty Hawk had a rudder in the front of the airplane, which made the plane very maneuverable. A disadvantage was the necessity for the pilot to keep adjusting the rudder to fly the plane. If the pilot let go of the stick the plane would crash. Other early aviators tried to build stable airplanes. These would have been easier to fly, but because of their poor maneuverability they could not be brought up into the air. By using their insight and skillful experiments the Wright brothers made the first successful flight with Kitty Hawk in 1905. The fact that this plane was unstable was a strong impetus for the development of autopilots based on feedback.

Autopilots

Since it was quite tiresome to fly an unstable aircraft, there was strong motivation to find a mechanism that would stabilize an aircraft. Such a device, invented by Sperry, was based on the concept of feedback. Sperry used a gyro-stabilized pendulum to provide an indication of the vertical. He then arranged a feedback mechanism that would pull the stick to make the plane go up if it was pointing down and vice versa. The Sperry

Figure 1.11 Picture from Sperry's contest in Paris. Sperry's son is at the stick and the mechanic walks on the wings to introduce disturbances. Notice the proximity to the ground.

autopilot is the first use of feedback in aeronautical engineering. Sperry won a prize in a competition for the safest airplane in Paris in 1912. Figure 1.11 is a picture from the event. The autopilot is a good example of how feedback can be used to stabilize an unstable system.

Autonomous Systems

Fully automatic flight including take off and landing is a development that naturally follows autopilots. It is quite surprising that this was done as early as 1947, see Figure 1.12. The flight was manually supervised but the complete flight was done without manual interaction.

Autonomous flight is a challenging problem because it requires automatic handling of a wide variety of tasks, landing, flying to the normal flight altitude, navigation, approaching the airfield, and landing. This requires a combination of continuous control and logic, so called hybrid control. In the flight by Robert E. Lee the logic required was provided by an IBM computer with punch cards. The theory of hybrid systems is still in its infancy. There are many similarities with the problem of designing humanoid robots. They require facilities for navigation, map making, adaptation, learning, reasoning. There are many unsolved research prob-

Robot-Piloted Plane Makes Safe Crossing of Atlantic

No Hand on Controls From Newfoundland to Oxfordshire—Take-Off, Flight and Landing Are Fully Automatic

By ANTHONY LEVIERO

Special to THE NEW YORK TIMES.

WASHINGTON, Sept. 22 — A Douglas C-54 Skymaster with a mechanical brain landed without human aid near London today after a robot directed hop from Newfoundland.

The revolutionary flight across the Atlantic, effected by the push of the button, was hailed by Air Force spokesmen said. Two ships somewhere in the Atlantic furnished bearings to the Skymaster's brain. She had 3,700 gallons of fuel aboard.

Great Britain several weeks ago had asked as a favor that the Air Force send the Skymaster there to make demonstration flights for

Figure 1.12 Excerpt from article in New York Times on September 23, 1947, describing the first fully automatic transatlantic flight.

lems in this field.

The use of autonomous systems is rapidly increasing. The Boeing 777 and the Airbus are modern examples of systems with a high degree of autonomy. There have also been extensive developments of Unmanned Air Vehicles (UAV).

Integrated Process and Control Design

Flight control is a good illustration of the value of integrated process and control design. The Wright brothers succeeded where others failed, because they made an unstable airplane that was maneuverable. Aircrafts that were both stable and maneuverable were built later. There are still substantial advantages in having unstable aircrafts that rely on a control system for stabilization. Modern fighters obtain their performance in this way. A schematic picture of two modern jet fighters are shown in Figure 1.13. The positions of the center of mass CM and the center of pressure are key elements. To be stable the center of pressure must be behind of the center of mass. The center of pressure of an aircraft shifts backwards when a plane goes supersonic. If the plane is stable at sub-sonic speeds it becomes even more stable at supersonic speeds. Very large forces and large control surfaces are then required to maneuver the airplane. A more balanced design is obtained by placing the center of pressure in front of

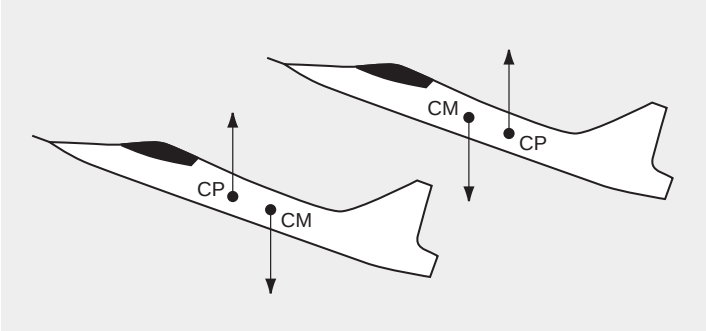


Figure 1.13 Schematic digram of two aircrafts. The aircraft above is stable because it has the center of pressure behind the center of mass. The aircraft below is unstable.

the center of mass at sub-sonic speeds. Such a plane will however by unstable at sub-sonic speeds, i.e. at take off and landing. This imposes severe constraints on the safety and robustness of the control system, but the aircraft will have superior performance. When the automatic control system becomes a critical part of the process it may also become mission critical which means that the system will fail if the controls fails. This induces strong demands on the reliability of the control system.

The development of aeronautical and aerospace engineering has often gone hand in hand with the development of feedback control. It was realized that control has to be considered up front in the design at the same time as, structures, engines, aerodynamics. A very interesting illustration of this is the recent development of high performance military aircrafts. Most aircrafts built today are designed to be stable.

Control is also mission critical for rockets and satellites.

1.8 Electronics and Communication

Electronics emerged in the in 1906 with the invention of the audion, a prototype of the vacuum tube, by Le De Forest. This was the start of the revolutionary development of the electronics industry which has had a big impact on the way we live. Control has had a major impact on this industry as well. The first application of feedback in electronics was a patent on vacuum tube amplifiers by the rocket pioneer Robert Goddard in 1912, but the most influential development is undoubtedly the negative feedback amplifier. Before dealing with this we will, however, briefly

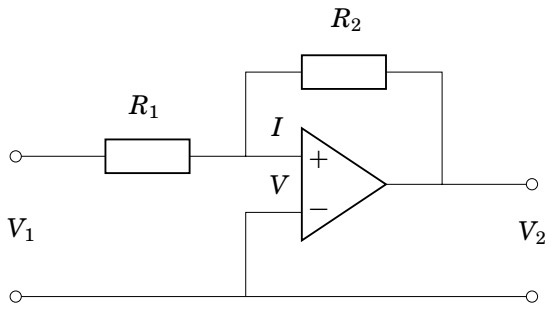


Figure 1.14 Schematic diagram of Armstrongs super-regenerative receiver.

discuss an application of positive feedback.

The Super-regenerative Amplifier - Positive Feedback

Vacuum tubes were expensive and it was highly desirable to use as few amplifiers as possible. In 1915 Armstrong suggested to use positive feedback to obtain a high amplification. A schematic of Armstrongs amplifier is shown in Figure 1.14.

Assume that the current I into the amplifier is zero, then the current through the resistors R_1 and R_2 are the same. It then follows from Ohms Law that

$$\frac{V_1 - V}{R_1} = \frac{V - V_2}{R_2} \tag{1.2}$$

Let G be the open loop gain of the amplifier, it then follows that

$$V_2 = GV \tag{1.3}$$

Eliminating the variable V between Equations (1.2) and (1.3) gives the following equation for the ratio of the output and input voltages

$$\frac{V_2}{V_1} = G\frac{R_1}{R_1 + R_2 - GR_2} \tag{1.4}$$

This equation gives the gain of the amplifier with positive feedback. The gain is very large if the resistors are chosen so that $R_1 + R_2 - GR_2$ is small. Assume for example that $R_1 = 100k\Omega$, $R_1 = 24k\Omega$, and $G = 5$. The formula above shows that the gain of the feedback system is 25. With $R_2 = 24.5k\Omega$ the gain will be 50, and with $R_1 = 25k\Omega$ the gain is infinite.

Regeneration was the word used for feedback at the time, and Armstrongs amplifier was called a super-regenerative amplifier because it obtained high gain by positive feedback. Armstrongs invention made it

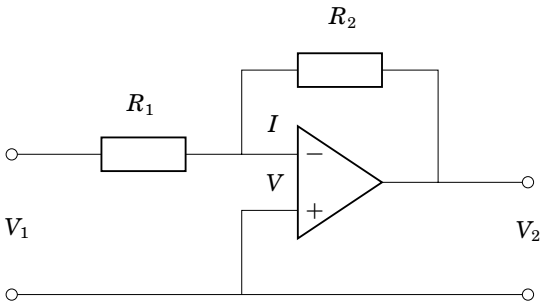


Figure 1.15 An amplifier with negative feedback.

possible to build inexpensive amplifiers with very high gain. The amplifiers were, however, extremely sensitive and they could easily start to oscillate. Prices of vacuum tubes also dropped and the interest in the amplifier dwindled. Next we will discuss another use of feedback in an amplifier which still has profound consequences.

The Negative Feedback Amplifier

When telephone communications were developed, amplifiers were used to compensate for signal attenuation in long lines. The vacuum tube was a component that could be used to build amplifiers. Distortion caused by the nonlinear characteristics of the tube amplifier together with amplifier drift were obstacles that prevented development of line amplifiers for a long time. A major breakthrough was Black's invention of the feedback amplifier in 1927. Black used negative feedback which reduces the gain but makes the amplifier very insensitive to variations in tube characteristics. Black's invention made it possible to build stable amplifiers with linear characteristics despite nonlinearities of the vacuum tube amplifier.

A schematic diagram of a feedback amplifier is shown in Figure 1.15. Assume that the current I into the amplifier is zero, the current through the resistors R_1 and R_2 are then the same and it follows from Ohms Law that Equation (1.2) holds. Let the gain of the amplifier be G it follows that

$$V_2 = -GV \tag{1.5}$$

Eliminating the variable V between Equations (1.2) and (1.5) gives the following equation for the ratio of the output and input voltages

$$\frac{V_2}{V_1} = -\frac{R_2}{R_1}\frac{1}{1+\frac{1}{G}(1+\frac{R_2}{R_1})} \tag{1.6}$$

This equation gives the gain of the amplifier with negative feedback. Since the gain G is a very large number, typically of the order of 10^5 or 10^8, it follows from this equation that the input-output property of the amplifier is essentially determined by resistors R_1 and R_2. These are passive components which are very stable. The properties of the active components appear in parameter G. Even if G changes significantly, the input-output gain remains constant. Also notice that the relation between V_{out} and V_{in} is very close to linear even if the relation between V_{out} and V, equation 1.5, is strongly nonlinear. □

Like many clever ideas the idea of the feedback amplifier seems almost trivial when it is described. It took, however, six years of hard work for Black to come up with it. The invention and development of the feedback amplifier was a key step in the development of long distance communications. The following quote from the presentation of the IEEE Lamme Medal to Black in 1957 gives a perspective on the importance of Black's invention of the feedback amplifier:

> It is no exaggeration to say that without Black's invention, the present long-distance telephone and television networks which cover our entire country and the transoceanic telephone cables would not exist. The application of Black's principle of negative feedback has not been limited to telecommunications. Many of the industrial and military amplifiers would not be possible except for its use. ... Thus, the entire explosive extension of the area of control, both electrical and mechanical, grew out of an understanding of the feedback principle. The principle also sheds light in psychology and physiology on the nature of the mechanisms that control the operation of animals, including humans, that is, on how the brain and senses operate.

It is interesting to observe that while Armstrong used positive feedback Black was using negative feedback.

Feedback quickly became an indispensable companion of electronics and communication. The applications are abundant. Today we find interesting use of feedback in power control in system for cellular telephony. A handset must naturally use enough power to transmit so that it can be heard by the nearest station, using too much power will increase interference with other handsets, necessitating use of even more power. Keeping the power at the correct level gives a large pay-off because the batteries will last longer.

Hewlett's Stabilized Oscillator

Many control problems are solved by linear systems. There are, however, problems where nonlinearities are essential. One of them is the design of

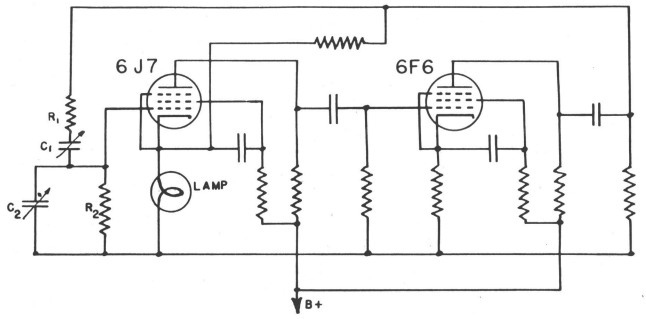


Figure 1.16 Circuit diagram of William Hewletts oscillator that gives a stable oscillation through nonlinear feedback using a lamp.

an oscillator that gives a signal with a constant amplitude. This problem was solved very elegantly by William Hewlett in his PhD thesis at Stanford University in 1939. Hewlett simply introduced a nonlinear element in the form of a lamp in the circuit, see Figure 1.16. Hewletts oscillator was the beginning of a very successful company HP, that Hewlett founded with David Packard.

CD Players and Optical Memories

The CD player is an interesting device which critically depends on several high performance control loops. The yearly sales of CD players was about 100 million units in the year 2000. That makes it one of the most common control systems.

A schematic picture of the main components of a CD player is shown in Figure 1.17. The information on the disc is read by light from a laser diode that is reflected by the disc to an optical assembly with photo transistors. Processing of the signals from the transistors gives the radial track error and a a focus signal. The laser diode, the optical assembly and the actuator are mounted on a sled which can be moved precisely over a small range. The sled is moved by another servo which permits large motions.

A block diagram of the major servos are shown in Figure 1.18. There are three critical servo loops. The focus servo concentrates the laser spot in the disc information layer. The tracking servo positions the laser spot on the track. The sled servo moves the sled so that the tracking system is in operating range. The tracking and the sled servos are also used to switch between tracks. The servos are all based on error feedback since only the error signal is available from the sensors. The major disturbance is due to misalignment of the track. In a CD player this is due to an

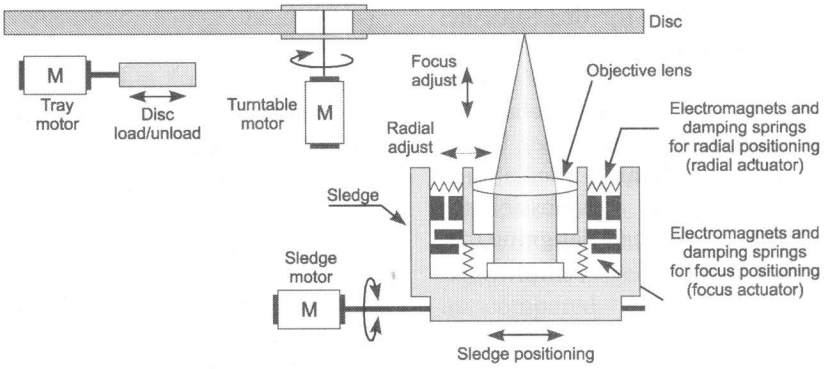


Figure 1.17 Schematic picture of a CD player.

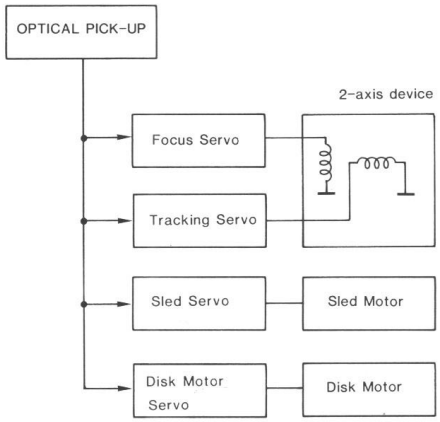


Figure 1.18 Block diagram of the major servos in a CD player.

off set of the center both due to manufacturing variations and errors in centering of the CD. The disturbance is approximately sinusoidal. The tracking accuracy of the systems is quite remarkable. For a typical CD player, which stores 650 M-bytes, the variations in a track are typically $200\,\mu m$, the track width is $1.6\,\mu m$ and the tracking accuracy is $0.1\mu m$. The tracking speed varies from 1.2 to $1.4m/s$. The servos used in the Digital Versatile Disc (DVD) and in optical memories are very similar to the servos in a CD but the precision is higher. For a Digital Versatile Disc (DVD) the variations in a track are typically $100\,\mu m$, the track width is

1.6 μm and the tracking accuracy is $0.022\mu m$. The tracking speed varies in the range $3.5m/s$. The quality of the major servo loops have a direct impact on the performanc of the storage system. A better tracking servo permits a higher storage density in an optical drive. An improved servo for switching tracks is immediately reflected in the search time in an optical memory.

1.9 Automotive

Cars are increasingly being provided with more and more control systems. The possibility of doing this are due to availability of cheap microprocessors and sensors and good control technology. The automotive industry has also been a strong driving force for the development of micro-controllers and sensors.

Reducing Emissions

California introduced a standard that required a substantial reduction of emissions for internal combustion engines. To achieve this it was necessary to introduce feedback in the engine based on measurement of the oxygen in the exhaust. The following quote from a plenary lecture by William E. Powers a former vice president of Ford at the 1999 World Congress of IFAC is illuminating.

> The automobiles of the 1990s are at least 10 times cleaner and twice as fuel efficient as the vehicles of the 1970s. These advancements were due in large part to *distributed microprocessor-based control systems.* Furthermore the resultant vehicles are safer, more comfortable and more maneuverable.

Cruise and Traction Control

Most new cars are provided with a system for cruise control that keeps the speed constant. This is a common control system that practically everyone comes in contact with. More advanced systems, called adaptive cruise control, are now appearing. The system is called adaptive even if it is a regular servo system that keeps the distance to the car in front at a constant value. The distance is measured by radar. In this context it is interesting to note that in 1995 Dickmann's modified a Mercedes to make it fully autonomous. The system had vision sensors and could make automatic lane changes. The car has been tested with human supervision at high speed in Germany and France.

Systems for anti-lock braking and traction control have also been introduced. These systems were used in an unexpected fashion in the Mercedes

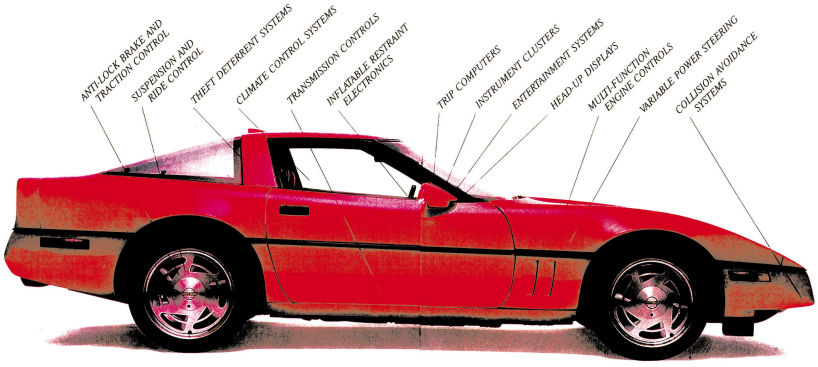


Figure 1.19 Some of the control systems in a car.

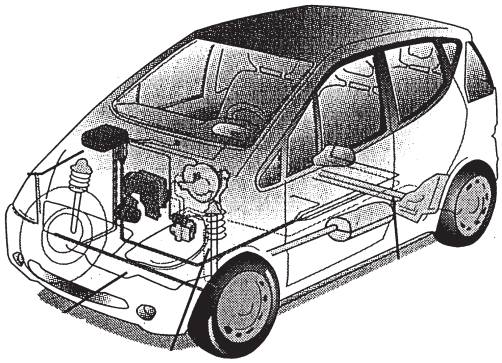


Figure 1.20 The Mercedes A-class is a small car where control helped to solve a serious problem.

A class, see Figure 1.20. This is a small car which achieves a high degree of safety through a thick floor that can be deformed in case of an accident. A consequence of this is that the center of gravity is high. When the car was introduced it was discovered that the car had a tendency to fall over in sharp turns. This difficult problem was solved by providing the car with the most sophisticated traction control system available in the company at the time. Together with minor changes of tires a severe difficulty was overcome.

Technology Drivers

The automotive industry is an important driver for technology because of the large number of produced parts and hard requirements for low cost. Several interesting developments took place when computers started to be used for computer control of engines. To save costs the microcomputer and input-output devices that connect sensors and actuators were merged on one chip, so called micro controllers. These devices made computer control cost effective in many other fields. The total number of micro-controllers for embedded systems now far exceed the number of microprocessors manufactured each year. The automotive applications also required new sensors and actuators. New accelerometers and gyros based on MEMS devices were developed, electric actuators based on new magnetic materials have also been developed. The sensors and actuators use feedback internally to obtain robustness and performance.

Autonomous Driving

There have been several attempts at developing autonomous vehicles. In 1995 Dickmanns demonstrated a fully autonomous Mercedes Benz with vision sensors. Under human supervision the car drove autonomously on the Autobahn from Munich to Copenhagen. Experiments with autonomous driving have also been done in California in the Path program.

1.10 Computing

There has been a strong symbiosis between control and computing. Computing devices are integral parts of a controller and computing and simulation are used extensively in design and validation of a control system.

Analog Computing

Early controllers, such as the centrifugal governor, were implemented using mechanical devices. Integral action was implemented using the ball and disc integrator invented by Lord Kelvin. In the process industries analog computing was instead done using pneumatic devices. The key elements were pneumatic amplifiers, restrictions and volumes. Feedback was used extensively to obtain linear behavior from the nonlinear devices.

The early development of control was severely hampered by the lack of computing, many clever graphical methods were developed to obtain insight and understanding using modest computing. The situation is summarized very clearly in the following quote from Vannevar Bush from 1923.

> "Engineering can proceed no faster than the mathematical analysis on which it is based. Formal mathematics is frequently inadequate for numerous problems pressing for solution, and in the absence of radically new mathematics, a mechanical solution offers the most promising and powerful attack wherever a solution in graphical form is adequate for the purpose. This is usually the case in engineering problems."

Bush later built the the first mechanical differential analyzer. The key elements were the ball and disc integrator and the torque amplifier. It could be used to integrate a handful of differential equations. This computer was used by Ziegler and Nichols to devise tuning rules for PID controllers.

Bush's work laid the foundation for analog computing which developed rapidly when cheap electronic amplifiers became available. This coincided with the emergence of control and analog computing became the standard tool for simulation of control systems. The analog computers were however large expensive systems that required a large staff to maintain. The use of analog computing was thus limited to persons with access to these rare resources. Analog computing was also used to implement the controllers in the servomechanism era and through the 1960s. Controllers were also implemented as small dedicated analog computers.

Even if computer control is the dominating technology for implementing controllers there are still niches where analog computing is used extensively. One area is micro-mechanical systems (MEMS) where mechanics and control are integrated on the same chip. Analog computing is also used for systems with extremely fast response time.

Computer Control

When digital computers became available they were first used for computing and simulation. The early computers were large and expensive and not suitable to be embedded in controllers. The first computer controlled systems was installed by TRW at the Port Arthur refinery in Texas in 1959. This initiated a development which started slowly and accelerated rapidly with the advances in computing. Today practically all controllers are implemented as computer controlled systems.

Real Time Computing

Use of computers for in control systems imposes demands on the architecture of systems and software, because of the requirement of fast response to external events. It turns out that the features that were useful for control were also useful for other purposes. There are in particular severe requirements on the operating system to provide rapid response to

external events. It is also necessary to make sure that the system operates without interruptions. Special real time operating systems therefore emerged. When implementing control systems there is a clear need to understand both control algorithms and software. It is also necessary to have systems with a high degree of reliability. The luxury of restarting the computer (CTRL+ALT+DEL) if something strange happens is not a feasible solution for computer control.

Simulation

Simulation is an indispensable tool for the control engineer. Even if systems can be designed based on relatively simple models it is essential to verify that the system works in a wide range of operations. This is typically done by simulating the closed loop system. To be reliable the simulation requires a high fidelity model of the system, sometimes parts of the real system is actually interfaced with the simulator, so called hardware in the loop simulation. If a controller is build using a dedicated computer it can also be verified against the simulation. Simulation can also be used for many other purposes, to explore different systems configurations, for operator training and for diagnostics. Because of the advances in computers and software simulation is now easily available at every engineers desk top. Development of a suitable model for the process requires a major effort.

The Internet

The Internet was designed to be an extremely robust communication network. It achieves robustness by being distributed and by using feedback. The key function of the system is to transmit messages from a sender to a receiver. The system has a large number of nodes connected with links. At each node there are routers that receives messages and sends them out to links. The routers have buffers that can store messages. It is desirable to operate the system to exploit capacity by maximizing throughput subject to the constraint that all users are treated fairly. There are large variations in traffic and in the lengths of the messages. Routers and links can also fail so the system may also be changing. The Internet depends critically on feedback to deal with uncertainty and variations. All control is decentralized

The routers receive messages on the incoming lines and distributes them to the outgoing lines. Messages are stored in a buffer if the router can not handle the incoming traffic. In the simplest scheme a router that has a full buffer will simply drop the incoming messages. Information about congestion is propagated to the sender indirectly through the lost messages.

Flow control is done by the senders. When a message is received by a receiver it sends an acknowledgment to the sender. The sender can detect lost messages because the messages are tagged. A very simple algorithm for traffic control was proposed by Jacobson 1990. This algorithm is called Additive Increase Multiplicative Decrease (AIMD) works as follows. The transmission rate is increased by additively as long as no messages are lost. When a message is lost the transmission rate is reduced by a factor of 2.

The Internet is a nice illustration that a very large distributed system can be controlled effectively by reasonably simple control schemes. There are many variations of the basic scheme described above. Many technical details have also been omitted. One drawback with the current scheme is that the control mechanism creates large variations in traffic which is undesirable. There are many proposals for modifications of the system.

1.11 Mathematics

There has always been a strong interplay between mathematics and control. This has undoubtedly contributed to the success of control. The theory of governors were developed by James Clarke Maxwell in a paper from 1868, about 100 years after James Watt had developed the governor. Maxwell realized that stability was connected to the algebraic problem of determining if an algebraic equation has roots in the right half plane. Maxwell turned to the mathematician Routh to get help. An analogous situation occurred in the development of water turbines where Stodola turned to the mathematician Hurwitz for assistance.

Mathematics played a major role in the development of servomechanism theory in the 1940s. There were a number of outstanding mathematicians at the Radiation Laboratory at MIT, there were also outstanding mathematicians at Bell Laboratories at the time. An interesting perspective can be obtained by comparing the major advances in the theory of feedback amplifiers with the meager advances in process control and to speculate what could have happened if there had been mathematicians working on the process control problems.

There was a very fruitful interaction between mathematics and control in the 1960s in connection with the space race when optimal control was developed. The Russian mathematician Pontryagin and his coworkers developed the maximum principle, which can be viewed as an extension of the Euler-Lagrange theory in calculus of variations. The American mathematician Richard Bellman developed dynamic programming, which can be regarded as an extension of the Hamilton-Jacobi theory in calculus of variations. In both cases the developments were directly inspired by the

control problems.

In the US the eminent topologist and past President of the American Mathematical Society argued for a strong effort in applied mathematics. With support form the Office of Naval Research he established a research center devoted to nonlinear ordinary differential equations and dynamics at Princeton. The center was later moved to RIAS and later to Brown University where it became the Lefschetz Center for Dynamical Systems. Later centers were also created at the University of Minnesota.

The Russian activity was centered in Moscow at the famous Institute of Automatic Control and Telemechanics and in smaller institutes in St Petersburg and Sverdlovsk. In Peking a strong center was established under the supervision of the Academy of Sciences. A very strong center INRIA was also created in Paris under Professor Jean Jacques Lions.

A wide range of mathematics such as dynamical systems, differential geometry and algebra has been important for the development of control theory after 1960.

Numerical Mathematics

A standard computational problem is to solve ordinary differential equations by time-stepping methods. As solutions often vary significantly over the range of integration, efficient computation requires step length control. This is done by estimating the local error and adjusting the step length to make this error sufficiently small. Most solvers for differential equations have carried out the correction in a simplistic fashion, and the adjustment has often been mixed with other algorithmic elements.

Recently a drastic improvement has been made by viewing step length adjustment as a feedback problem. Substantial improvements of performance can be obtained by replacing the heuristic schemes for step length adjustment, that were used traditionally, with a scheme based on a PID controller. These advantages are achieved without incurring additional computational costs. This has resulted in more reliable software, as well as software with much better structure. Knowledge of basic control schemes has thus proven very beneficial. It is likely that the same idea can be applied to other numerical problems.

1.12 Physics

Feedback has always had a central role in scientific instruments. An early example is the development of the mass spectrometer. In a paper from 1935 by Nier it is observed that the deflection of the ions depend on both the magnetic and the electric fields. Instead of keeping both fields constant

Nier let the magnetic field fluctuate and the electric field was controlled to keep the ratio of the fields constant. The feedback was implemented using vacuum tube amplifiers. The scheme was crucial for the development of mass spectroscopy.

Another example is the work by the Dutch Engineer van der Meer. He invented a clever way to use feedback to maintain a high density and good quality of the beam of a particle accelerator. The scheme, called stochastic cooling, was awarded the Nobel prize in Physics in 1984. The method was essential for the successful experiments in CERN when the existence of the the particles W and Z was first demonstrated. Another use of feedback called repetitive control was developed by Nakano for particle accelerators. The key idea was to obtain very precise control by exploiting the fact the particles move in circular orbits.

The atomic force microscope is a more recent example. The key idea is to move a narrow tip on a cantilever beam across the surface and to register the forces on the tip. Such systems rely on feedback systems for precise motion and precise measurement of the forces.

Adaptive Optics

A severe problem in astronomy is that turbulence in the atmosphere blurs images in telescopes because of variations in diffraction of light in the atmosphere. The blur is of the order of an arc-second in a good telescope. One way to eliminate the blur is to move the telescope outside the Earths atmosphere as is done with the Hubble telescope. Another way is to use feedback to eliminate the effects of the variations in a telescope on the Earth. This is the idea of adaptive optics. A schematic picture of a system for adaptive optics is shown in Figure 1.21. The reference signal is a bright star or an artificial laser beam projected into the atmosphere. The actuator which is shown simply as a box in the figure is obtained by reflecting the light on a mirror that can be deformed selectively. The mirror can have from 13 to 1000 elements. The error signal is formed by analyzing the shape of the distorted wave form from the reference. This signal is sent to the controller which adjusts the deformable mirror. The light from the observed star is compensated because it is also reflected in the deformable mirror before it is sent to the detector. The wave lengths used for observation and control are often different. Since diffraction in the atmosphere changes quite rapidly the response time of the control system must be of the order of milliseconds.

In normal feedback terminology adaptive optics is a regular feedback system, feedback is used to compensate for variations in diffraction in the atmosphere. The word adaptive is used in because it can also be said that the system adapts for variations in the atmosphere. In the control community the word adaptive is often used with a different meaning.

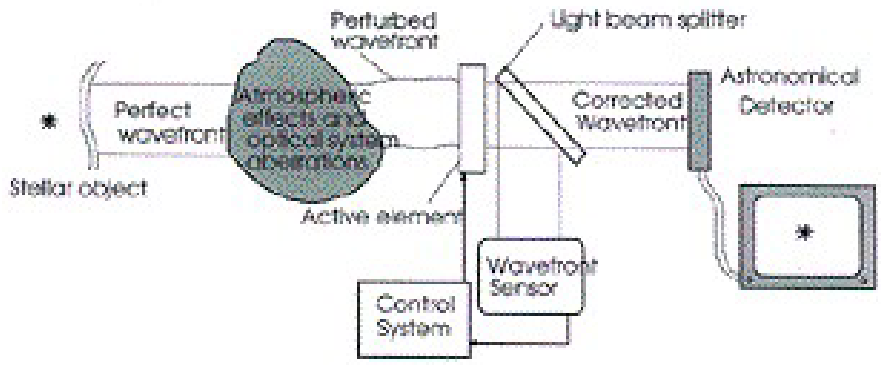


Figure 1.21 Schematic diagram of a system for adaptive optics.

Quantum Systems

Control of quantum systems is currently receiving a lot of interest. Molecular dynamics is a very spectacular application. The idea is to use modulated laser light to break up bonds in molecules to obtain ions which can react with other ions to form new molecules. This is done by tailoring the laser pulses so that they will break specific bonds between atoms. This is precision surgery at the molecular level, quite different from the methods used in conventional chemistry.

1.13 Biology

It was mentioned already in Section 1.1 that feedback is an essential mechanism in biology. Here are a few examples.

The Pupillary Reflex

The human eye has an effective system to control the amount of light that is let into the eye. The light intensity is measured and the pupil opening is adjusted. This control systems is easily accessible for experimentation. Extensive investigations have been made by changing the light intensity and measuring the pupil opening. The results show that it is a very effective feedback system.

Human Posture

The human body hasmany feedback systems. They allow us to stand upright, to walk, jump and balance on ropes. They also adjust the sensitivity

of our eyes and ears, enabling us to see and hear over a wide range of intensity levels. They maintain a constant body temperature and a delicate balance of chemical substances in our body. As an illustration we will discuss the system that allows us to stand upright. The key features of the system are known although several details are poorly understood. The primary sensors are the semicircular canals located in the mastoid bone close to the ear. The sensors consist of toroidal canals that are filled with liquid. Neurons connected to hairs in the canals give signals related to the motion of the head. The major actuators are muscles in feet, legs, knees, hips and arms. There are also sensory neurons in the feet and the muscles. There is local feedback from pressure sensors in the feet and sensors in the muscles to the actuating muscles. This loop has a reaction time of about 20 ms. The interconnection between sensors and actuators is made in the spinal cord. These interconnections are responsible for fast feedback like reflexes. The reaction time is of the order of 100 ms. There is also a high level feedback loop that receives information from the vestibular system, which gives information about the position and orientation of our body parts in space. The sensory information is processed in the cerebellum and transmitted to the muscle neurons in the spinal cord. This feedback has a reaction time of about 250 ms.

This system for control of posture illustrates that feedback can be used to stabilize an unstable system. It also shows that there are very reliable biological feedback systems which are essential to everyday life. A particularly interesting feature is that the system has learning capabilities. Think about a child learning to stand up and walk or learning to bicycle. These functions are far superior to those of any technical system.

A Simple Experiment

A simple experiment on one of the systems in the body can be executed manually with very modest equipment. Take a book with text and hold it in front of you. Move the text sideways back and forth and increase the speed of motion until the text is blurred. Next hold the text in front of you and move the head instead. Notice the difference in the speeds when the text gets blurred. You will observe that higher speeds are possible when you move your head. The reason for this is that when you move the text, the information about the motion comes via the processing of the image at your retina, but when you move your head the information comes from the semicircular canals. The feedback from the visual processing is much slower because it uses higher functions in the brain.

There are many other nice control systems in the human body. Top performing athletes such as tennis players have interesting abilities for very advanced motion control that involves much interaction with vision. Humans also have interesting learning abilities.

1.14 Summary

This chapter has given glimpses of how control and feedback have influenced the development of technology and how control is used. The examples show that control has emerged concurrently with new technology, that it has had a major influence on technology and that it sometimes has been an enabler. A large number of control systems ranging from small micro devices to large global systems for generation and distribution of electricity and large communication systems have also been given. The wide range of uses of control also point to some difficulties. The ideas of control and feedback are abstract which makes them less obvious. It is therefore common that the ideas are neglected in favor of hardware which is much easier to talk about.

2

Feedback

2.1 Introduction

Feedback is a powerful idea, which is used extensively in natural and technical systems. The principle of feedback is very simple: base correcting actions on the difference between desired and actual performance. In engineering feedback has been rediscovered and patented many times in many different contexts. Use of feedback has often resulted in vast improvements in system capability, sometimes they have even be revolutionary as discussed in Chapter 1. The reason for this is that feedback has some truly remarkable properties. In this chapter we will discuss some of the properties of feedback that can be understood intuitively. The benefits of feedback can often be obtained using simple forms of feedback such as on-off control and PID control, which are discussed in Section 2.2. Particular attention is given to integral action which is has truly remarkable properties. Feedback systems may appear complicated because they involve many different subsystems and many different technologies. To reduce the complexity it is necessary to have abstractions that makes it possible to have an overview of the systems. Section 2.3 presents different ways to describe feedback systems. The block diagram is the most important representation because it is a uniform description that can be adapted to many different purposes. The remarkable properties of feedback are presented in Section 2.4. It is shown that feedback can reduce the effects of disturbances and process variations, it can create well defined relations between variables, it makes it possible to modify the properties of a system, e.g. stabilize a unstable system. The discussion is based on block diagrams and simple static mathematical models. The major drawback is that feedback can create instability. This is discussed briefly in Section 2.5. To understand stability it is necessary to have knowledge of

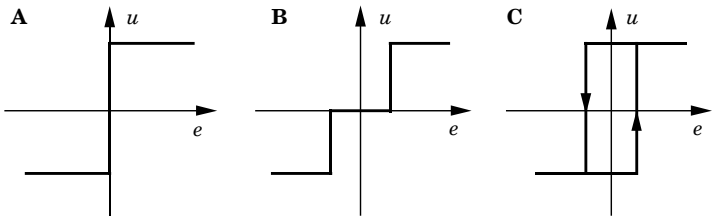


Figure 2.1 Controller characteristics for ideal on-off control (A), and modifications with dead zone (B) and hysteresis (C).

dynamical systems which is the topic of Chapter 3. Feedback systems are also called closed loop systems. Section 2.6 compares closed loop systems with their opposite, open loop systems. Feedback is reactive because actions are based on deviations. Feedforward, which is proactive, is the opposite of feedback. Since feedback and feedforward have complementary properties they are often combined.

2.2 Simple Forms of Feedback

Some properties of feedback were illustrated in Chapter 1. Many of the nice properties of feedback can be obtained with simple controllers. In this section we will discuss some simple forms of feedback namely, on-off control, proportional control and PID control.

On-Off Control

The feedback can be arranged in many different ways. A simple feedback mechanism can be described as follows:

$$u = \begin{cases} u_{\max}, \text{if } e > 0 \\ u_{\min}, \text{if } e < 0 \end{cases} \tag{2.1}$$

where $e = r - y$ is the control error which is the difference between the reference signal and the output of the system. Figure 2.1 shows the relation between error and control. This control law implies that maximum corrective action is always used. This type of feedback is called *on-off control*. It is simple and there are no parameters to choose. On-off control often succeeds in keeping the process variable close to the reference, but it will typically result in a system where the variables oscillate. Notice that in (2.1) the control variable is not defined when the error is zero. It is common to have some modifications either by introducing hysteresis or a dead zone (see Figure 2.1).

PID Control

The reason why on-off control often gives rise to oscillations is that the system over reacts since a small change in the error will make the manipulated variable change over the full range. This effect is avoided in proportional control where the characteristic of the controller is proportional to the control error for small errors. This can be achieved by making the control signal proportional to the error

$$u = k(r - y) = ke \tag{2.2}$$

where k is the controller gain.

Integral Action Proportional control has has the drawback that the process variable often deviates from its reference value. This can be avoided by making the control action proportional to the integral of the error

$$u(t) = k_i \int_0^t e(\tau)d\tau \tag{2.3}$$

where k_i is the integral gain. This control form is called integral control. It follows from this equation that if there is a steady state where the control signal and the error are constant, i.e. $u(t) = u_0$ and $e(t) = e_0$ respectively then

$$u_0 = k_i e_0 t$$

This equation is a contradiction unless $e_0 = 0$. It has thus demonstrated that there will be no steady state error with a controller that has integral action. Notice that the argument also holds for any process and any controller that has integral action. The catch is that there may not always be a steady state because the system may be oscillating. This amazing property which we call the *Magic of Integral Control* has been rediscovered many times. It is one of the properties that have strongly contributed to the wide applicability of PID control.

Derivative Action An additional refinement is to provide the controller with an anticipative ability by using a prediction of the error. A simple prediction is given by the linear extrapolation

$$e(t + T_d) \approx e(t) + T_d \frac{de(t)}{dt},$$

which predicts the error T_d time units ahead, see Figure 2.2. Combining

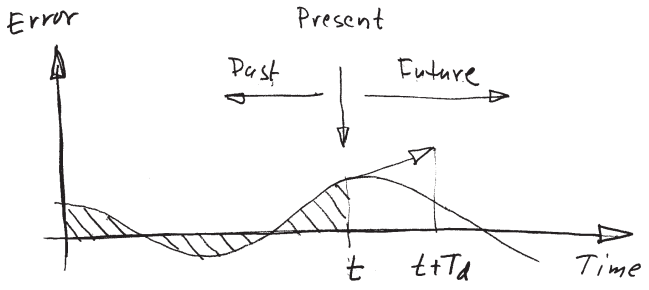


Figure 2.2 A PID controller takes control action based on past, present and future control errors.

proportional, integral and derivative control we obtain a controller that can be expressed mathematically as follows:

$$\begin{aligned} u(t) &= ke(t) + k_i \int_0^t e(\tau)d\tau + k_d \frac{de(t)}{dt} \\ &= k\left(e(t) + \frac{1}{T_i} \int_0^t e(\tau)d\tau + T_d \frac{de(t)}{dt} \right) \end{aligned} \tag{2.4}$$

The control action is thus a sum of three terms representing the past by the integral of the error (the I-term), the present (the P-term) and the future by a linear extrapolation of the error (the D-term). The term $e + T_d de/dt$ is a linear prediction of the error T_d time units in the future. Notice that the controller can be parameterized in different ways. The second parameterization is commonly used in industry. The parameters of the controller are called: are proportional gain k, integral time T_i, and derivative time T_d.

The PID controller is very useful. It is capabable of solving a wide range of control problems. The PI controller is the most common controller. It is quoted that about 90% of all control problems can be solved by PID control, many of these controllers are actually PI controller because derivative action is not so common. There are more advanced controllers which differ from the PID controller by using more sophisticated methods for prediction.

2.3 Representation of Feedback Systems

Feedback systems are often large and complex. It is therefore a major challenge to understand, analyze and design them. This is illustrated by the fact that the idea of feedback was developed independently in many different application areas. It took a long time before it was found that the systems were based on the same idea. The similarities became apparent when proper abstractions were made. In this section we will develop some ideas that are used to describe feedback systems. The descriptions we are looking for should capture the essential features of the systems and hide unnecessary details. They should be applicable to many different systems.

Schematic Diagrams

In all branches of engineering, it is common practice to use some graphical description of systems. They can range from stylistic pictures to drastically simplified standard symbols. These pictures make it possible to get an overall view of the system and to identify the physical components. Examples of such diagrams are shown in Figure 2.3

Block Diagrams

The schematic diagrams are useful because they give an overall picture of a system. They show the different physical processes and their interconnection, and they indicate variables that can be manipulated and signals that can be measured.

A special graphical representation called *block diagrams* has been developed in control engineering. The purpose of block diagrams is to emphasize the information flow and to hide technological details of the system. It is natural to look for such representations in control because of its multidisciplinary nature. In a block diagram, different process elements are shown as boxes. Each box has inputs denoted by lines with arrows pointing toward the box and outputs denoted by lines with arrows going out of the box. The inputs denote the variables that influence a process and the outputs denote some consequences of the inputs that are relevant to the feedback system.

Figure 2.4 illustrates how the principle of information hiding is used to derive an abstract representation of a system. The upper part of the picture shows a photo of a physical system which a small desk-top process in a control laboratory. It consists of two tanks, a pump that pumps water to the tanks, sensors, and a computer which implements the control algorithm and provides the user interface. The purpose of the system is to maintain a specified level in the lower tank. To do so, it is necessary to measure the level. The level can be influenced by changing the speed of the motor that pumps water into the upper tank. The voltage to the

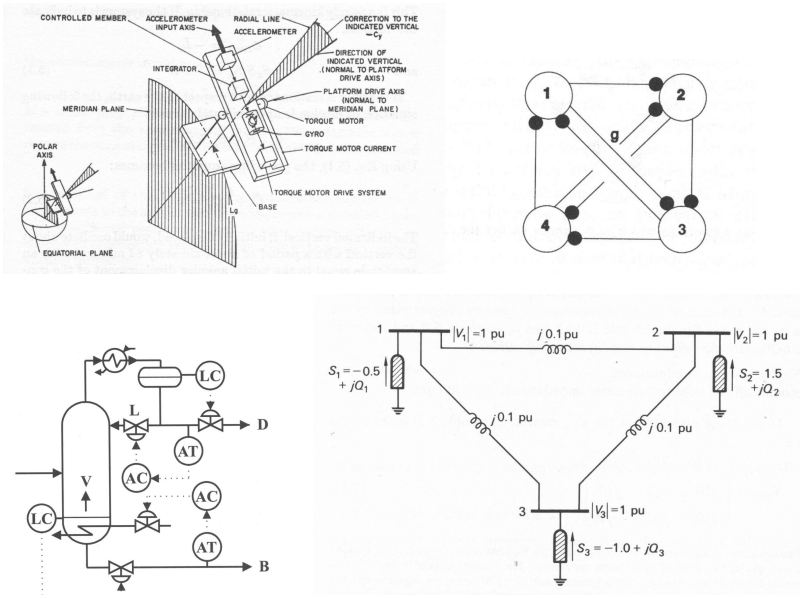


Figure 2.3 Examples of schematic descriptions: a schematic picture of an inertial navigation system (upper left), a neuron network for respiratory control (upper right), a process and instrumentation diagram (lower left) and a power system (lower right).

amplifier that drives the pump is selected as the control variable. The controller receives information about the desired level in the tank and the actual tank level. This is accomplished using an AD converter to convert the analog signal to a number in the computer. The control algorithm in the computer then computes a numerical value of the control variable. This is converted to a voltage using a DA converter. The DA converter is connected to an amplifier for the motor that drives the pump.

The first step in making a block diagram is to identify the important signals: the control variable, the measured signals, disturbances and goals. Information hiding is illustrated in the figure by covering systems by a cloth as shown in the lower part of Figure 2.4. The block diagram is simply a stylized picture of the systems hidden by the cloth.

In Figure 2.4, we have chosen to represent the system by two blocks only. This granularity is often sufficient. It is easy to show more details simply by introducing more subsystems, as indicated in Figure 2.5 where we show the drive amplifier, motor, pump, and tanks, the sensors with electronics, the AD converter, the computer and the DA converter. The

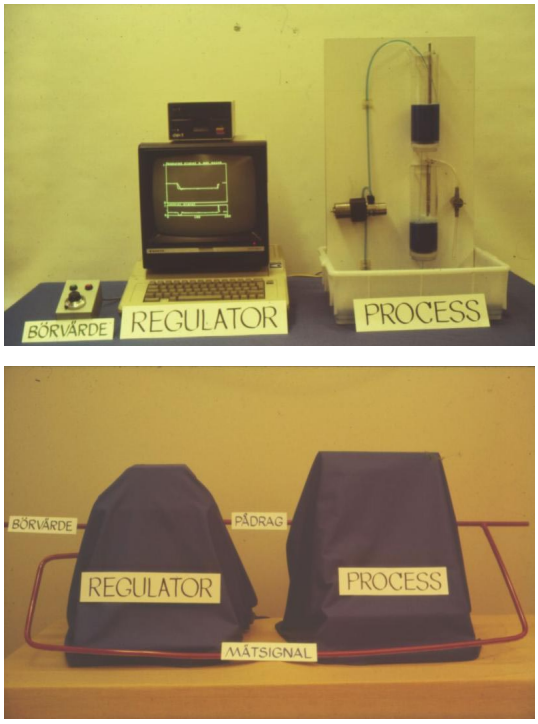


Figure 2.4 Illustrates the process of information hiding used to obtain a block diagram. The top figure is a picture of the physical system, the middle figure is obtained by hiding many details about the system and the bottom figure is the block diagram.

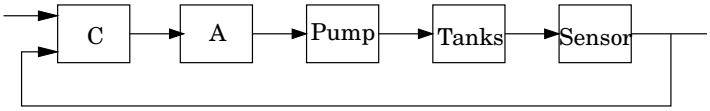


Figure 2.5 A more detailed block diagram of the system in Figure 2.4 showing controller C, amplifier A, pump, tanks and sensor.

detail chosen depends on the aspects of the system we are interested in and the taste of the person doing the investigation. Remember that parsimony is a trademark of good engineering. Very powerful tools for design, analysis and simulation were developed when the block diagrams were complemented with descriptions of the blocks in terms of transfer functions.

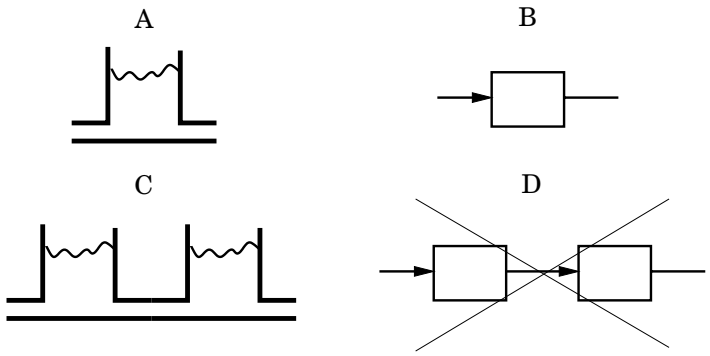


Figure 2.6 A simple hydraulic system with an inflow and a free outflow is shown in A. The block diagram representation of the system is shown in B. The system obtained by connecting two hydraulic systems is shown in C. This system cannot be represented by the series connection of the block diagrams in B.

Causality

The arrows in a block diagram indicate causality because the output of a block is caused by the input. To use the block diagram representation, it is therefore necessary that a system can be partitioned into subsystems with causal dependence. Great care must be exercised when using block diagrams for detailed physical modeling as is illustrated in Figure 2.6. The tank system in Figure 2.6B is a cascade combination of the two tanks shown in Figure 2.6B. It cannot be represented by cascading the block diagram representations because the level in the second tank influences the flow between the tanks and thus also the level in the first tank. When using block diagrams it is therefore necessary to choose blocks to represent units which can be represented by causal interactions. We can thus conclude that even if block diagrams are useful for control they also have serious limitation. In particular they are not useful for serious physical modeling which has to be dealt with by other tools which permit bidirectional connections.

Examples

An important consequence of using block diagrams is that they clearly show that control systems from widely different domains have common features because their block diagrams are identical. This observation was one of the key factors that contributed to the emergence of the discipline of automatic control in the 1940s. We will illustrate this by showing the block diagrams of some of the systems discussed in Chapter 1.

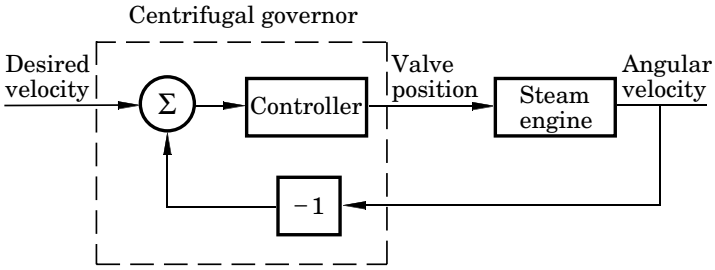


Figure 2.7 Block diagram of a steam engine with a centrifugal governor.

EXAMPLE 2.1—A STEAM ENGINE WITH A CENTRIFUGAL GOVERNOR

The steam engine with the centrifugal governor in Example 2.1 can be represented with the block diagram shown in Figure 2.7. In this block diagram we have chosen to represent the steam engine with one block. This block has two inputs: the position of the steam valve and the load torque of the systems that the engine is driving. The system has one output which is engine speed. The controller is a box with two inputs: the engine speed and desired engine speed. The output of the controller is the steam valve position. There is some two-way interaction between the controller and the valve position but with appropriate gearing and heavy balls in the governor it may be assumed that the force exerted by the valve on the governor is negligible. □

EXAMPLE 2.2—AN AIRCRAFT STABILIZER

To develop a block diagram for an airplane with the Sperry stabilizer, we first introduce suitable variables and describe the system briefly. The pitch angle that describes how the airplane is pointing is an important variable and is measured by the gyro-stabilized pendulum. The pitch angle is influenced by changing the rudder. We choose to represent the airplane by one box whose input is the rudder angle and whose output is the pitch angle. There is another input representing the forces on the airplane from wind gusts. The stabilizer attempts to keep the pitch angle small by appropriate changes in the rudder. This is accomplished by wires that connect the rudder to the gyro-stabilized pendulum. There is also a mechanism enabling the pilot to choose a desired value of the pitch angle if he wants the airplane to ascend or descend. In the block diagram we represent the controller with one block where the difference between desired and actual pitch angles is the input and the rudder angle is the output. Figure 2.8 shows the block diagram obtained.

□

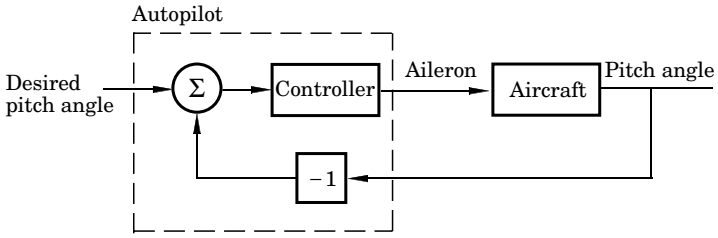


Figure 2.8 Block diagram of an airplane with the Sperry autopilot.

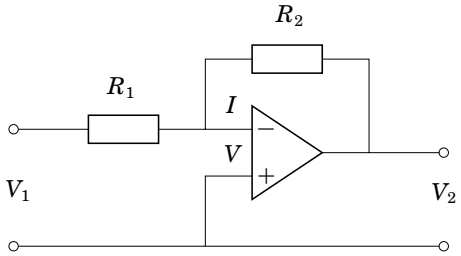


Figure 2.9 A feedback amplifier.

Even if block diagrams are simple, it is not always entirely trivial to obtain them. It happens frequently that individual physical components to not necessarily correspond to specific blocks and that it may be necessary to use mathematics to obtain the block. We illustrate this by an example.

EXAMPLE 2.3—A FEEDBACK AMPLIFIER

An electronic amplifier with negative feedback was discussed in Section 1.8. A schematic diagram of the amplifier is shown in Figure 2.9. To develop a block diagram we first decide to represent the pure amplifier as one block. This has input V and output V_2. The input-output relation is

$$V_2 = -GV$$

where G is the gain of the amplifier and the negative sign indicates negative feedback. If the current I into the amplifier is negligible the current through resistors R_1 and R_2 are the same and we get

$$\frac{V_1 - V}{R_1} = \frac{V - V_2}{R_2}$$

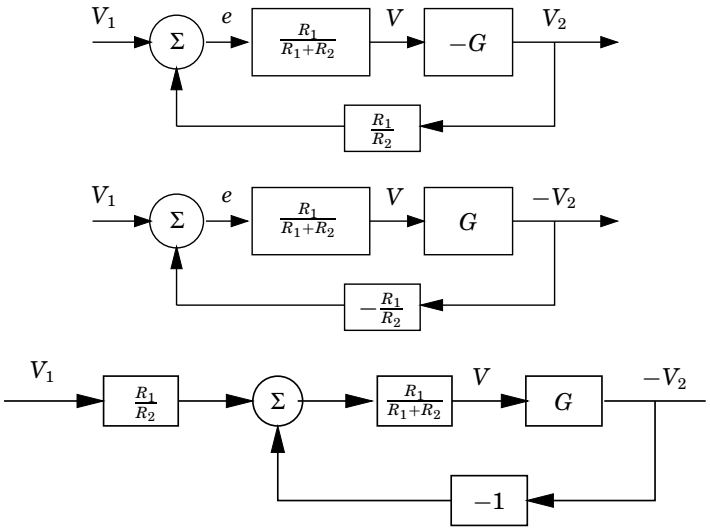


Figure 2.10 Three block diagrams of the feedback amplifier in Figure 2.9.

Solving this equation for the input voltage V to the amplifier we get

$$V = \frac{R_2 V_1 + R_1 V_2}{R_1 + R_2} = \frac{R_2}{R_1 + R_2}\left(V_1 + \frac{R_1}{R_2} V_2\right)$$

This equation can be represented by one block with gain $R_2/(R_1 + R_2)$ and the input $V_1 + R_1 V_2/R_1$ and we obtain the block diagram shown in Figure 2.10. The lower representation where the process has positive gain and the feedback gain is negative has become the standard of representing feedback systems. Notice that the individual resistors do not appear as individual blocks, they actually appear in various combinations in different blocks. This is one of the difficulties in drawing block diagrams. Also notice that the diagrams can be drawn in many different ways. The middle diagram in Figure 2.10 is obtained by viewing $-V_2$ as the output of the amplifier. This is the standard convention where the process gain is positive and the feedback gain is negative. The lowest diagram in Figure 2.10 is yet another version, where the ratio R_1/R_2 is brought outside the loop. In all three diagrams the gain around the loop is $R_2 G/(R_1 + R_2)$, this is one of the invariants of a feedback system. □

A Generic Control System with Error Feedback

Although the centrifugal governor, the autopilot and the feedback amplifier in Examples 2.1, 2.2 and 2.3 represent very different physical sys-

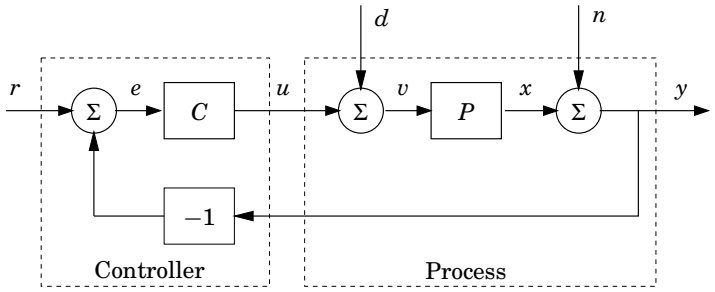


Figure 2.11 Generic control system with error feedback.

tems, their block diagrams are identical apart from the labeling of blocks and signals, compare Figures 2.7, 2.8 and 2.10. This illustrates the universality of control. A generic representation of the systems is shown in Figure 2.11. The system has two blocks. One block P represents the process and the other C represents the controller. Notice negative sign of the feedback. The signal r is the reference signal which represents the desired behavior of the process variable x.

Disturbances are an important aspect of control systems. In fact if there were no disturbances there is no reason to use feedback. In Figure 2.11 there are two types of disturbances, labeled d and n. The disturbance labeled d is called a load disturbance and the disturbance labeled n is called measurement noise. Load disturbances drive the system away from its desired behavior. In Figure 2.11 it is assumed that there is only one disturbance that enters at the system input. This is called an input disturbance. In practice there may be many different disturbances that enter the system in many different ways. Measurement noise corrupts the information about the process variable obtained from the measurements. In Figure 2.11 it is assumed that the measured signal y is the sum of the process variable x and measurement noise. In practice the measurement noise may appear in many other ways.

The system in Figure 2.11 is said to have error feedback, because the control actions are based on the error which is the difference between the reference r and the output y. In some cases like a CD player there is no explicit information about the reference signal because the only information available is the error signal. In such case the system shown in Figure 2.11 is the only possibility but if the reference signal is available there are other alternatives that may give better performance.

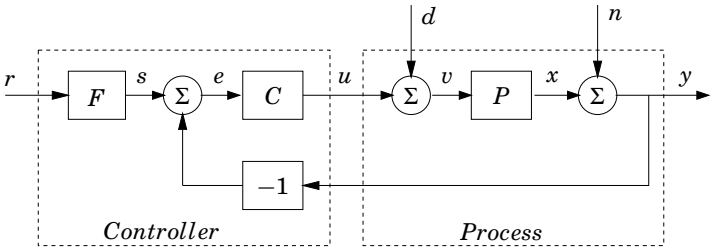


Figure 2.12 Block diagram of a generic feedback system with two degrees of freedom.

A Generic Control Loop Two Degrees of Freedom

In Figure 2.11 the control actions are based on the error e. When both the reference signal r and the measured output y are available it is possible to obtain improved control. Such a system is shown in Figure 2.12. This system is similar to the one in Figure 2.11 but the controller now has two blocks, the feedback C and the feedforward block F. This means that the signal path from y to u is different from that from r to u. Such controllers are said to have two degrees of freedom. The extra freedom gives substantial advantages.

A Qualitative Understanding of Feedback Systems - Cherchez l'erreur

The block diagram is very useful to get an overview of a system. It allows you to see the wood in spite of the trees. A simple way to understand how a system works is to assume that the feedback works so well that the error is zero. We illustrate this with a few examples.

EXAMPLE 2.4—THE GENERIC CONTROL SYSTEM

Consider the system in Figure 2.12. If the error is zero we find that the output y is equal to the signal s, which is the output of the block F. □

EXAMPLE 2.5—THE FEEDBACK AMPLIFIER

Consider the block diagram of the feedback amplifier in the top of Figure 2.10. If the system works well the error is close to zero which means that

$$V_2 = \frac{R_1}{R_2}$$

□

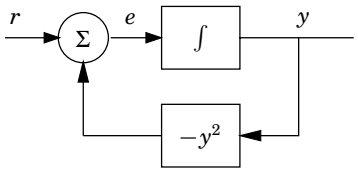


Figure 2.13 Block diagram of a nonlinear system.

EXAMPLE 2.6—A NONLINEAR SYSTEM
Consider a nonlinear system with the block diagram shown in Figure 2.13. Assume that r is positive and the error is zero, it follows that $r = y^2$, which implies that $y = \sqrt{r}$. The output is thus the positive square root of the input. Notice, that since the feedback loop contains an integrator the error will always be zero if a steady state exists. □

This example illustrates the fact that if a component for generating a function is available, it is easy to generate the inverse function by an amplifier with feedback. This idea has been much applied in analog computing and in instrumentation.

2.4 Properties of Feedback

To understand the behavior of a feedback system it is necessary to describe the behavior of the process and the controller. A full understanding requires a description of the dynamic properties of the process and the controller. Some properties can however be understood by describing the behaviors by static models. This will be developed in this Section.

Mathematical Models

A block diagram gives an overview of a system but more details are required to obtain a more complete description of a system. In particular it is necessary to describe the behavior of each individual block. This requires mathematical models. A function

$$y = f(u)$$

is a simple way to characterized the relation between input and output in a block. This is called a *static model* because a change in the input gives an instantaneous change of the output. The function can be represented by a graph as in Figure 2.14. For small perturbations around a given operating point, the curve can be approximated by its tangent. The slope

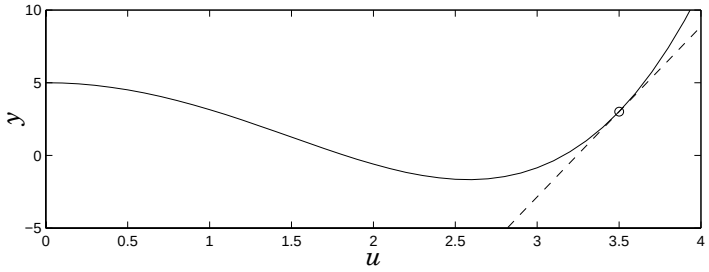


Figure 2.14 Static input-output function for a system and the linear approximation around the operating point $u_0 = 3.5$. The slope of the curve at the operating point is the gain of the system.

of the tangent is called the gain of the system at the operating point. Assume that the control variable at the operating point has the value u_0. The corresponding value of the output is then $y_0 = f(u_0)$. For values of the control variable close to the operating point the system can approximately be described by the model

$$y - y_0 = f(u) - f(u_0) \approx f'(u_0)(u - u_0)$$

where f' denotes the derivative of f. The approximation which is obtained by making a Taylor series expansion of the function and neglecting higher order terms is called linearization. The constant $K = f_u(u_0)$ is called the gain of the system at the operating point. To simplify notation it is customary to choose the variables so that they are zero at the operating point of interest. The system can then be described by

$$y = Ku$$

which we call a linear static model. Static models of control systems have severe limitations, because many of the properties of feedback systems rely on dynamic effects. A major part of this book will be devoted to this beginning with next chapter which deals with dynamics. Control is thus intimately connected with dynamics.

Static Analysis

We will start by a very simplistic analysis of a system with error feedback. Consider the system in Figure 2.11. Assume that the variables r, d and n are constants and that the process and the controller can be described by linear static models. Let k_p be the process gain and k_c the controller gain.

The following equations are obtained for the process and the controller.

$$\begin{aligned} y &= x + n \\ x &= k_p(u + d) \\ u &= k_c(r - y) \end{aligned} \tag{2.5}$$

Solving these equations for y and u gives

$$\begin{aligned} x &= \frac{k_p k_c}{1 + k_p k_c} r + \frac{k_p}{1 + k_p k_c} d - \frac{k_p k_c}{1 + k_p k_c} n \\ y &= \frac{k_p k_c}{1 + k_p k_c} r + \frac{k_p}{1 + k_p k_c} d + \frac{1}{1 + k_p k_c} n \\ u &= \frac{k_c}{1 + k_p k_c} r - \frac{k_p k_c}{1 + k_p k_c} d - \frac{k_c}{1 + k_p k_c} n \end{aligned} \tag{2.6}$$

The product $L = k_p k_c$ is called the loop gain. It is the total gain around the feedback loop. It is an important system property which is dimension-free.

Several interesting conclusions can be drawn from (2.6). First we observe that since the equation is linear we can discuss the effects of reference values r, load disturbances d and measurement noise n separately. It follows from (2.6) that the output will be very close to the reference value if the loop gain $L = k_p k_c$ is large. It also follows from (2.6) that the effect of the load disturbances will be small if the controller gain is large. We will now take a closer look at some of the responses.

Assuming that $r = 0$ and $n = 0$ we find that the process variable is given by

$$x = \frac{k_p}{1 + k_p k_c} d$$

The influence of the load disturbance on the output can be reduced significantly by having a controller with high gain.

If the disturbances are zero, i.e. $d = n = 0$ the response to reference signals is given by

$$x = \frac{k_p k_c}{1 + k_p k_c} r \tag{2.7}$$

By having a controller with high gain the process variable will be very close to the reference value. For example if $k_p k_c = 100$ the deviation between x and r is less than 1%

The properties of the process are seldom constant. To investigate the effects of process variations we differentiate (2.7) with respect to the process gain k_p. This gives

$$\frac{dx}{dk_p} = \frac{k_c}{(1 + k_p k_c)^2} r = \frac{k_c}{1 + k_p k_c} \frac{1}{1 + k_p k_c} r = \frac{x}{k_p} \frac{1}{1 + k_p k_c}$$

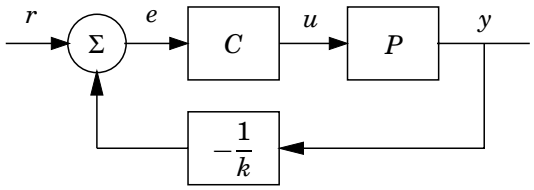


Figure 2.15 This feedback system has the input output relation $y = kr$ even if the process P is highly nonlinear.

Hence

$$\frac{d\log x}{d\log k_p} = \frac{dx/x}{dk_p/k_p} = \frac{1}{1+k_pk_c} \tag{2.8}$$

The relative variation in the process variable caused by process variations will thus be very small if the loop gain is high. For example if the loop gain is $k_pk_c = 100$ it follows from (2.8) that a 10% variation in the process gives only a variation of 0.1% in the relation between the process variable and the reference. This was the idea used by Black when he invented the feedback amplifier, (2.8) was actually part of Black's patent application.

The simple analysis above captures several properties of feedback. The analysis can however be misleading because it does not consider the dynamics of the process and the controller. The most serious drawback is that most systems will become unstable when the loop gain is large. Another factor is that the trade-off between reduction of load disturbances and injection of measurement noise is not represented well. In practice the load disturbances are often dominated by components having low frequencies while measurement noise often has high frequencies.

Using Feedback to Obtain Linear Behavior

Another nice feature of feedback is that it can be used to obtain precise linear relations between the reference and the output even if the process is nonlinear and time-variable. Consider the system in Figure 2.15. Assume that the process P is modeled as a nonlinear static system

$$y = f(u)$$

Let the controller C be a proportional controller with gain k, i.e.

$$u = k(r - y)$$

Eliminating u between these equations we find that the closed loop system is then described by

$$y + \frac{1}{k}f^{-1}(y) = r$$

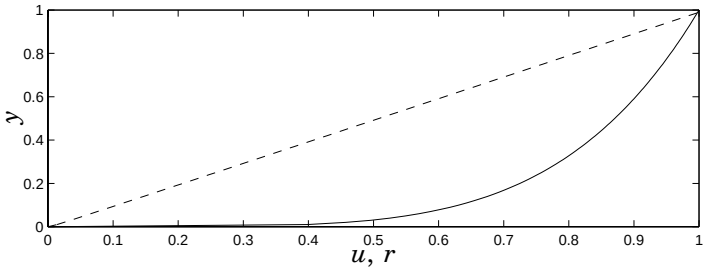


Figure 2.16 Input output relations for the process $y = u^5$ (solid lines) and for the feedback system in Figure 2.15 (dashed lines) when the controller gain k is 100.

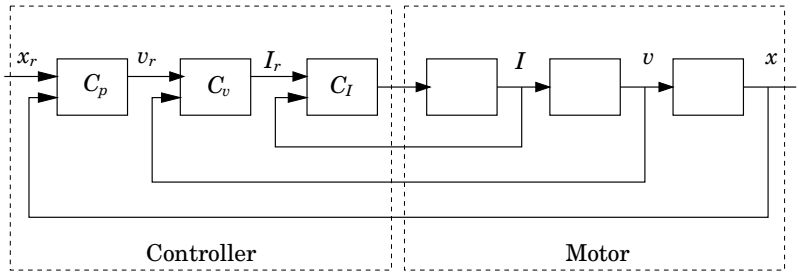


Figure 2.17 An hierarchical control system for a motor drive.

where f^{-1} is the inverse of the function f. If the controller gain k is large we have $y \approx r$, i.e. a linear relation. This is illustrated in Figure 2.16 which shows that strongly nonlinear relations can be made linear by feedback. Because of (2.8) the linear relation is also very insensitive to variations in the nonlinearity.

The scheme in Figure 2.15 is routinely used in many practical control systems. This is also very useful in order to structure complex systems. We illustrate the idea with a few examples.

EXAMPLE 2.7—ELECTRICAL MOTOR DRIVES

An electrical motor drive consists of a motor with a drive amplifier and possible also gear boxes. The primary control signal is the input voltage to the drive amplifier and the primary output is the motor angle. It is customary to have a control system with the hierarchical structure indicated in Figure 2.17. The controller thus consist of three cascaded controllers. The inner loop is a current controller C_I, based on measurement of the current through the rotor. Since the drive torque of the motor is propor-

tional to current the inner loop is essentially controlling motor torque. The next loop with the controller C_v controls the motor velocity by adjusting the reference value I_r to the current controller. The velocity signal is either created by differentiating the angle measurements or it is obtained from a special sensor. The outermost loop controls the position by adjusting the reference value of the velocity loop. In a configuration like the one in Figure 2.17 the innermost loop is typically the fastest and the outermost loop is the slowest. □

EXAMPLE 2.8—PROCESS CONTROL
A large industrial process such as a refinery or a paper mill has thousands of control loops. Most variables are primarily influenced by valves, which have nonlinear characteristics. The model of the system can be simplified by using local feedback, because it can be assumed that the flows can be regarded as the inputs to the system. It is then sufficient to find the relations between flows and the interesting physical quantities. □

We have shown how feedback can be used to obtain linear behavior. By a small modification of the scheme in Figure 2.15 it is also possible to use feedback to obtain well defined nonlinear relations. It can also be used to obtain well defined dynamic relations. We illustrate this by a few examples.

EXAMPLE 2.9—FLIGHT CONTROL
An airplane is primarily influence by changing rudders and engine speed. The relation between these actuators and the motion of the airplane is quite complicated and it also changes with flight conditions such as speed and height. Consider for example the rolling motion of the airplane. By providing the aircraft with sensors and feedback it is possible to make the relation between the sideways stick motion and the roll angle look like an integrator which is easy to control manually. The output remains constant if the input is zero and the rate of change of the output is proportional to the input. □

EXAMPLE 2.10—A TRAINER FOR SHIPS STEERING
Large ships like tankers are difficult to maneuver. It is possible to have a small boat and use feedback to make it behave like a large tanker. This is very convenient for training situations because if the helmsman does mistakes it is easy to switch off the control system thus giving a small boat that can easily be maneuvered to recover to a safe situation. □

EXAMPLE 2.11—COMPLIANCE CONTROL
Consider a tool for grinding that is mounted on an industrial robot. The robot can apply a force in an arbitrary direction, if the robot is also provided with force sensors it is possible to arrange a feedback so that the force is a function of the position of the tool and the its orientation. In this way it is possible to constrain the tool to be orthogonal to a specified surface. When forcing the tool away from the surface it will exert a force that drives it to the surface. □

EXAMPLE 2.12—HAPTICS
There are special types of joy sticks which are provided with motors so that the joy stick can generate forces that give additional information to the user. With these joy sticks it is possible to use feedback to generate forces when the cursor approaches given regions. In this way it is possible to simulate things like pushing on soft objects. □

The possibility of using feedback to modify the behavior of systems is an idea that has very wide applicability.

2.5 Stability

Static analysis of feedback systems is based on the assumption that a control action is immediately manifested by a change in system output. This is a strong simplification because systems are typically dynamic which means that changes in the input do not immediately give rise to changes in the output. A major effort of control theory is actually devoted to finding appropriate ways to describe dynamical phenomena.

One consequence of the systems being dynamic is that feedback may give rise to oscillations. The risk for instability is the major drawback of using feedback. The static analysis resulting in (2.6) has indicated that it is advantageous to have controllers with high gain. It is a common practical experience that feedback systems will oscillate if the feedback gain is too high. After a perturbation the control error typically has one of the following behaviors:

S1: The error may decrease monotonously

U1: The error may grow monotonously

S2: The error may be oscillatory with decaying amplitude

U2: The error may be oscillatory with increasing amplitude

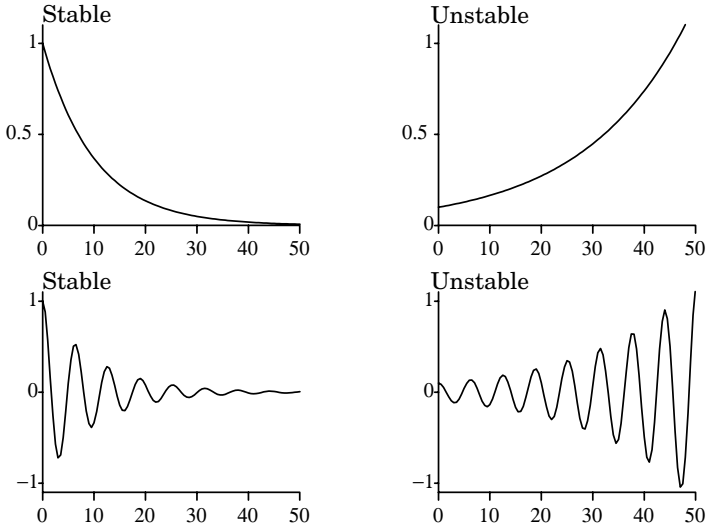


Figure 2.18 Illustration of different system behaviors used to define stability.

This is illustrated in Figure 2.18, the behaviors S1 and S2 are called stable the ones labeled U1 and U2 are unstable. There are also intermediate situations where the error remains constant or oscillates with constant amplitude.

In a paper from 1868 Maxwell made the important observation that stability was related to properties of the roots of an algebraic equation. In particular he established the following relations between the behaviors in Figure 2.18 and the properties of the roots.

U1: Corresponds to real roots with positive real part

S1: Corresponds to real roots with negative real part

U2: Corresponds to complex roots with positive real part

S2: Corresponds to complex roots with negative real part

Stability analysis of a feedback system can thus be done simply by investigating if the roots of an algebraic equation have negative real parts. This will be discussed in detail in Chapter 3.

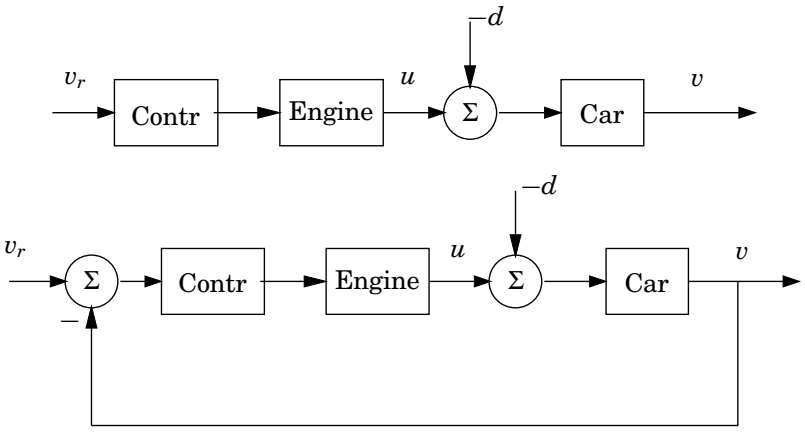


Figure 2.19 Open and closed loop systems for cruise control. The disturbance is the slope of the road.

2.6 Open and Closed Loop Systems

Using block diagrams a feedback system can be defined as a system where there is a loop in the block diagram. For that reason feedback systems are also called closed loop systems. The opposite of feedback systems are open loop systems, which have block diagrams which simply are a sequence of connected blocks.

Open loop systems are easy to understand because we can explain them by causal reasoning. Consider two blocks connected in series. The blocks can be investigated one at a time. The output of one block will be the input to the other and we can use cause and effect reasoning. This becomes complicated when the blocks are connected in a feedback loop because the interactions between the blocks. To understand feedback systems it is therefore necessary to use abstractions and mathematical reasoning. Open and closed loop systems have very different properties as is illustrated by the following example.

EXAMPLE 2.13—CRUISE CONTROL
Consider a system for cruise control. The system to be controlled is a car. The major disturbance is the slope of the road. The input to the system is the fuel flow into the engine. Block diagrams an open and closed loop systems is shown in Figure 2.19. The reference signal is fed to the controller which influences the fuel flow. To analyze the differences between the open and closed loop systems we will use simple static models. The

car can be described by

$$v = k_p(u - d) \tag{2.9}$$

where v is the velocity, u is the force generated by the car engine and d is the gravity force due to the slope of the road. The parameter k_p is a constant that relates the velocity to the total force acting on the car. With open loop control we have the following relation between the force of the car and the reference signal.

$$u = k_c v_r$$

where k_c is the controller gain and v_r is the desired velocity. Eliminating u between the equations and introducing the velocity error we find

$$e = v_r - v = (1 - k_p k_c)v_r + k_p d \tag{2.10}$$

The error is equal to zero if the slope is zero and if $k_c = 1/k_p$.

For á closed loop system with error feedback the controller can be described by

$$u = k_c(v_r - v)$$

Combining this with the process model given by (2.9) gives the following equation for the velocity error

$$e = v_r - v = \frac{k_p k_c}{1 + k_p k_c} v_r + \frac{1}{1 + k_p k_c} d \tag{2.11}$$

Some interesting conclusions can be made by comparing Equations (2.10) and (2.11). First consider the situation when there are no disturbances, $d = 0$. With open loop control, (2.10), the velocity error can be made small by matching the controller gain to be the inverse of the process gain. With feedback control, (2.11), the velocity error will be small if the controller gain is large, no matching is required.

Next we will investigate the effect of disturbances. For the open loop system the effect of a disturbance on the velocity is proportional to the process gain k_p, but for the closed loop system the effect is proportional to $k_p/(1 + k_p k_c)$. We thus find that with closed loop control the disturbances are reduced by the factor

$$\frac{1}{1 + k_p k_c}$$

compared with open loop control. Notice that this factor is the same as the factor encountered in (2.6). □

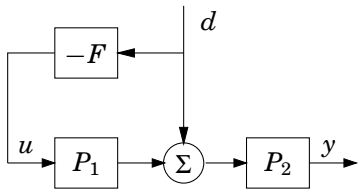


Figure 2.20 Illustration of the concept of feedforward.

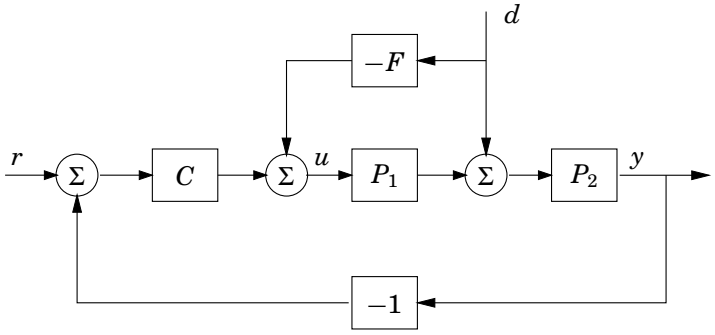


Figure 2.21 Control system that combines feedback and feedforward from a measurable disturbance.

2.7 Feedforward

Despite all the nice properties of feedback, it also has some limitations. When using feedback, an error must occur before the control system will act. This means that there must be an error before corrective actions are taken. Feedback is thus reactive. In some circumstances it is possible to measure a disturbance before it enters the system. It then appears natural to use the information about the disturbance to take corrective action before the disturbance has influenced the system. This way of controlling a system is called feedforward. The concept of feedforward is illustrated in Figure 2.20. The effect of the measured signal is reduced by measuring it and generating a control signal that counteracts it. Since feedforward attempts to match two signals it requires good process models, otherwise the corrections may have the wrong size or it may be badly timed. Feedforward is often combined with feedback as is illustrated in Figure 2.21. Such a system combines the good features of feedback and feedforward. The system with a controller having two degrees of freedom shown in Figure 2.12 can also be interpreted as a system that used feedforward to

Table 2.1 Properties of Feedback and Feedforward.

Feedback	*Feedforward*
Closed loop	Open loop
Market driven	Planning
Reactive	Pro-active
Robust to modeling errors	Sensitive to modeling errors
Risk for instability	No risk for instability

obtain a good response to reference signals.

The ideas of feedback and feedforward are very general and appear in many different fields. We illustrate this with an example from economics.

EXAMPLE 2.14—FEEDBACK AND FEEDFORWARD IN ECONOMICS
In economics feedforward corresponds to a planned economy while feedback corresponds to a market economy. In this case the discussion of feedback versus feedforward has risen to ideological heights. In business administration a pure feedforward strategy corresponds to running a company based on extensive strategic planning while a feedback strategy corresponds to a pure reactive approach. □

The empirical of control systems indicate that it is often advantageous to combine feedback and feedforward. Feedforward is particularly useful when disturbances can be measured or predicted. A typical example is in process control where disturbances in one process may be due to processes upstream. It is an educated guess that the experience in control that it is useful to combine feedback and feedforward can also be extended to other fields. The correct balance of the approaches requires insight and understanding of their properties. A summary of the properties of feedback and feedforward are given in the Table 2.1.

2.8 Summary

The idea of feedback has been discussed in this section. It has been shown that feedback is a powerful concept that has played a major role in the development of many areas of engineering, e.g. process control, telecommunications, instrumentation and computer engineering. This widespread use of feedback derives from its interesting properties.

- Feedback can reduce effects of disturbances

- Feedback can make a system insensitive to process variations
- Feedback can make a system follow reference values faithfully
- Feedback can create well defined relations between variables in a system
- Feedback can create desired behavior of a system
- Feedback can stabilize an unstable system
- Feedback can create instabilities

The first six features are the main reasons why control is so useful and so widely used, we summarize them in the catch phrase *the magic of feedback*. The main drawback of feedback is that it may cause instabilities. This is one reason why it is necessary to have some knowledge of control. Some special forms of feedback, on-off control and PID control have also been discussed. Most benefits of feedback can actually be obtained by PI or PID control. Integral action is very useful because it will give zero steady state error whenever there is a steady state. Different ways to describe feedback systems have been introduced, including the notion of block diagrams which is a powerful method of information hiding. Mathematical models in terms of static models have also been introduced and used to derive some properties of feedback systems. A deeper analysis requires other tools that take dynamics into account. This is necessary to describe stability and instability. The chapter ended with a discussion of another control principle, feedforward, which has properties that are complementary to feedback.

3

Dynamics

3.1 Introduction

From the perspective of control a dynamical system is such that the effects of actions do not occur immediately. Typical examples are: The velocity of a car does not change immediately when the gas pedal is pushed. The temperature in a room does not rise immediately when an air conditioner is switched on. Dynamical systems are also common in daily life. An headache does not vanish immediately when an aspirin is taken. Knowledge of school children do not improve immediately after an increase of a school budget. Training in sports does not immediately improve results. Increased funding for a development project does not increase revenues in the short term.

Dynamics is a key element of control because both processes and controllers are dynamical systems. Concepts, ideas and theories of dynamics are part of the foundation of control theory. Dynamics is also a topic of its own that is closely tied to the development of natural science and mathematics. There has been an amazing development due to contributions from intellectual giants like Newton, Euler, Lagrange and Poincare.

Dynamics is a very rich field that is partly highly technical. In this section we have collected a number of results that are relevant for understanding the basic ideas of control. The chapter is organized in separate sections which can be read independently. For a first time reader we recommend to read this section section-wise as they are needed for the other chapters of the book. To make this possible there is a bit of overlap between the different sections. in connection with the other chapters. There is a bit of overlap so that the different sections can be read independently.

Section 3.2 gives an overview of dynamics and how it is used in control which has inherited ideas both from mechanics and from electrical

engineering. It also introduces the standard models that are discussed in the following sections. In Section 3.3 we introduce a model for dynamics in terms of linear, time-invariant differential equations. This material is sufficient for analysis and design of simple control systems of the type discussed in the beginning of Chapter 5. The concepts of transfer function, poles and zeros are also introduced in Section 3.3. Another view of linear, time-invariant systems is given in Section 3.4 introduces the Laplace transform which is a good formalism for linear systems. This also gives another view on transfer functions. Combining the block diagrams introduced in Chapter 2 with the transfer functions gives a simple way to model and analyze feedback systems. The material Section 3.4 gives a good theoretical base for Chapters 4 and 5. Frequency response is yet another useful way of describing dynamics that provides additional insight. The key idea is to investigate how sine waves are propagating through a dynamical system. This is one of the contributions from electrical engineering discussed in Section 3.5. This section together with Section 3.4 gives the basis for reading Chapters 4, 5, 6 and 7 of the book.

Section 3.6 presents the idea of state models which has its origin in Newtonian mechanics. The problems of control have added richness by the necessity to include the effect of external inputs and the information obtained from sensors. In Section 3.6 we also discuss how to obtain models from physics and how nonlinear systems can be approximated by linear systems, so called linearization. In

The main part of this chapter deals with linear time invariant systems. We will frequently only consider systems with one input and one output. This is true for Sections 3.3, 3.4 and 3.5. The state models in Section 3.5 can however be nonlinear and have many inputs and outputs.

3.2 Two Views on Dynamics

Dynamical systems can be viewed from two different ways: the internal view or the external views. The internal view which attempts to describe the internal workings of the system originates from classical mechanics. The prototype problem was the problem to describe the motion of the planets. For this problem it was natural to give a complete characterization of the motion of all planets. The other view on dynamics originated in electrical engineering. The prototype problem was to describe electronic amplifiers. It was natural to view an amplifier as a device that transforms input voltages to output voltages and disregard the internal detail of the amplifier. This resulted in the input-output view of systems. The two different views have been amalgamated in control theory. Models based on the internal view are called internal descriptions, state models or white

box models. The external view is associated with names such as external descriptions, input-output models or black box models. In this book we will mostly use the words state models and input-output models.

The Heritage of Mechanics

Dynamics originated in the attempts to describe planetary motion. The basis was detailed observations of the planets by Tycho Brahe and the results of Kepler who found empirically that the orbits could be well described by ellipses. Newton embarked on an ambitious program to try to explain why the planets move in ellipses and he found that the motion could be explained by his law of gravitation and the formula that force equals mass times acceleration. In the process he also invented calculus and differential equations. Newtons results was the first example of the idea of reductionism, i.e. that seemingly complicated natural phenomena can be explained by simple physical laws. This became the paradigm of natural science for many centuries.

One of the triumphs of Newton's mechanics was the observation that the motion of the planets could be predicted based on the current positions and velocities of all planets. It was not necessary to know the past motion. The state of a dynamical system is a collection of variables that characterize the motion of a system completely for the purpose of predicting future motion. For a system of planets the state is simply the positions and the velocities of the planets. A mathematical model simply gives the rate of change of the state as a function of the state itself, i.e. a differential equation.

$$\frac{dx}{dt} = f(x) \tag{3.1}$$

This is illustrated in Figure 3.1 for a system with two state variables. The particular system represented in the figure is the van der Pol equation

$$\begin{aligned} \frac{dx_1}{dt} &= x_1 - x_1^3 - x_2 \\ \frac{dx_2}{dt} &= x_1 \end{aligned}$$

which is a model of an electronic oscillator. The model (3.1) gives the velocity of the state vector for each value of the state. These are represented by the arrows in the figure. The figure gives a strong intuitive representation of the equation as a vector field or a flow. Systems of second order can be represented in this way. It is unfortunately difficult to visualize equations of higher order in this way.

The ideas of dynamics and state have had a profound influence on philosophy where it inspired the idea of predestination. If the state of a

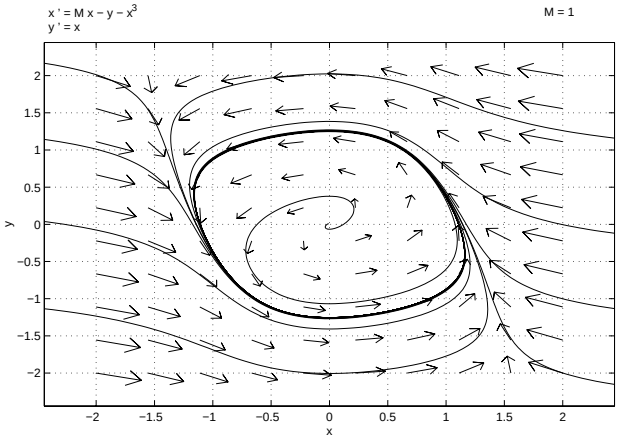


Figure 3.1 Illustration of a state model. A state model gives the rate of change of the state as a function of the state. The velocity of the state are denoted by arrows.

natural system is known ant some time its future development is complete determined. The vital development of dynamics has continued in the 20th century. One of the interesting outcomes is chaos theory. It was discovered that there are simple dynamical systems that are extremely sensitive to initial conditions, small perturbations may lead to drastic changes in the behavior of the system. The behavior of the system could also be extremely complicated. The emergence of chaos also resolved the problem of determinism, even if the solution is uniquely determined by the initial conditions it is in practice impossible to make predictions because of the sensitivity of initial conditions.

The Heritage of Electrical Engineering

A very different view of dynamics emerged from electrical engineering. The prototype problem was design of electronic amplifiers. Since an amplifier is a device for amplification of signals it is natural to focus on the input-output behavior. A system was considered as a device that transformed inputs to outputs, see Figure 3.2. Conceptually an input-output model can be viewed as a giant table of inputs and outputs. The input-output view is particularly useful for the special class of linear systems. To define linearity we let $(u_1,\ y_1)$ och $(u_2,\ y_2)$ denote two input-output pairs, and a and b be real numbers. A system is linear if $(au_1+bu_2,\ ay_1+ay_2)$ is also an input-output pair (superposition). A nice property of control problems is that they can often be modeled by linear, time-invariant systems.

Time invariance is another concept. It means that the behavior of the

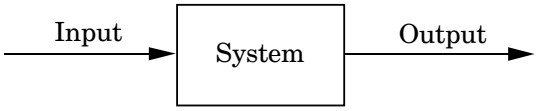


Figure 3.2 Illustration of the input-output view of a dynamical system.

system at one time is equivalent to the behavior at another time. It can be expressed as follows. Let $(u,\ y)$ be an input-output pair and let u_t denote the signal obtained by shifting the signal u, t units forward. A system is called time-invariant if $(u_t,\ y_t)$ is also an input-output pair. This view point has been very useful, particularly for linear, time-invariant systems, whose input output relation can be described by

$$y(t) = \int_0^t g(t-\tau)u(\tau)d\tau. \tag{3.2}$$

where g is the impulse response of the system. If the input u is a unit step the output becomes

$$y(t) = h(t) = \int_0^t g(t-\tau)d\tau = \int_0^t g(\tau)u(\tau)d\tau \tag{3.3}$$

The function h is called the step response of the system. Notice that the impulse response is the derivative of the step response.

Another possibility to describe a linear, time-invariant system is to represent a system by its response to sinusoidal signals, this is called frequency response. A rich powerful theory with many concepts and strong, useful results have emerged. The results are based on the theory of complex variables and Laplace transforms. The input-output view lends it naturally to experimental determination of system dynamics, where a system is characterized by recording its response to a particular input, e.g. a step.

The words input-output models, external descriptions, black boxes are synonyms for input-output descriptions.

The Control View

When control emerged in the 1940s the approach to dynamics was strongly influenced by the Electrical Engineering view. The second wave of developments starting in the late 1950s was inspired by the mechanics and the two different views were merged. Systems like planets are autonomous and cannot easily be influenced from the outside. Much of the classical

development of dynamical systems therefore focused on autonomous systems. In control it is of course essential that systems can have external influences. The emergence of space flight is a typical example where precise control of the orbit is essential. Information also plays an important role in control because it is essential to know the information about a system that is provided by available sensors. The models from mechanics were thus modified to include external control forces and sensors. In control the model given by (3.4) is thus replaced by

$$\begin{aligned} \frac{dx}{dt} &= f(x,u) \\ y &= g(x,u) \end{aligned} \tag{3.4}$$

where u is a vector of control signal and y a vector of measurements. This viewpoint has added to the richness of the classical problems and led to new important concepts. For example it is natural to ask if all points in the state space can be reached (reachability) and if the measurement contains enough information to reconstruct the state.

The input-output approach was also strengthened by using ideas from functional analysis to deal with nonlinear systems. Relations between the state view and the input output view were also established. Current control theory presents a rich view of dynamics based on good classical traditions.

The importance of disturbances and model uncertainty are critical elements of control because these are the main reasons for using feedback. To model disturbances and model uncertainty is therefore essential. One approach is to describe a model by a nominal system and some characterization of the model uncertainty. The dual views on dynamics is essential in this context. State models are very convenient to describe a nominal model but uncertainties are easier to describe using frequency response.

Standard Models

Standard models are very useful for structuring our knowledge. It also simplifies problem solving. Learn the standard models, transform the problem to a standard form and you are on familiar grounds. We will discuss four standard forms

- Ordinary differential equations
- Transfer functions
- Frequency responses
- State equations

The first two standard forms are primarily used for linear time-invariant systems. The state equations also apply to nonlinear systems.

3.3 Ordinary Differential Equations

Consider the following description of a linear time-invariant dynamical system

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + \ldots + a_n y = b_1 \frac{d^{n-1} u}{dt^{n-1}} + b_2 \frac{d^{n-2} u}{dt^{n-2}} + \ldots + b_n u, \qquad (3.5)$$

where u is the input and y the output. The system is of order n order, where n is the highest derivative of y. The ordinary differential equations is a standard topic in mathematics. In mathematics it is common practice to have $b_n = 1$ and $b_1 = b_2 = \ldots = b_{n-1} = 0$ in (3.5). The form (3.5) adds richness and is much more relevant to control. The equation is sometimes called a controlled differential equation.

The Homogeneous Equation

If the input u to the system (3.5) is zero, we obtain the equation

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \ldots + a_n y = 0, \qquad (3.6)$$

which is called the homogeneous equation associated with equation (3.5). The characteristic polynomial of Equations (3.5) and (3.6) is

$$A(s) = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n \qquad (3.7)$$

The roots of the characteristic equation determine the properties of the solution. If $A(\alpha) = 0$, then $y(t) = Ce^{\alpha t}$ is a solution to Equation (3.6).

If the characteristic equation has distinct roots α_k the solution is

$$y(t) = \sum_{k=1}^{n} C_k e^{\alpha_k t}, \qquad (3.8)$$

where C_k are arbitrary constants. The Equation (3.6) thus has n free parameters.

Roots of the Characteristic Equation give Insight

A real root $s = \alpha$ correspond to ordinary exponential functions $e^{\alpha t}$. These are monotone functions that decrease if α is negative and increase if α is positive as is shown in Figure 3.3. Notice that the linear approximations shown in dashed lines change by one unit for one unit of αt. Complex roots $s = \sigma \pm i\omega$ correspond to the time functions.

$$e^{\sigma t} \sin \omega t, \qquad e^{\sigma t} \cos \omega t$$

which have oscillatory behavior, see Figure 3.4. The distance between zero crossings is π/ω and corresponding amplitude change is $e^{\sigma \pi/\omega}$.

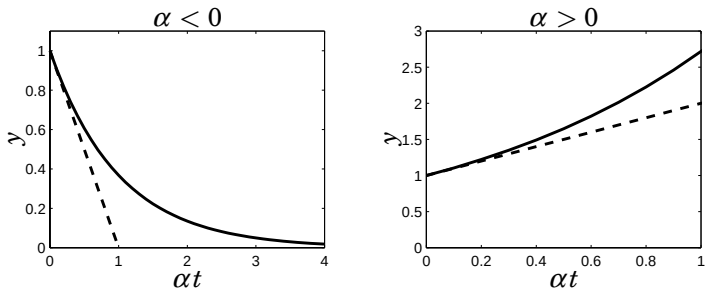


Figure 3.3 The exponential function $y(t) = e^{\alpha t}$. The linear approximations of of the functions for small αt are shown in dashed lines. The parameter $T = 1/\alpha$ is the time constant of the system.

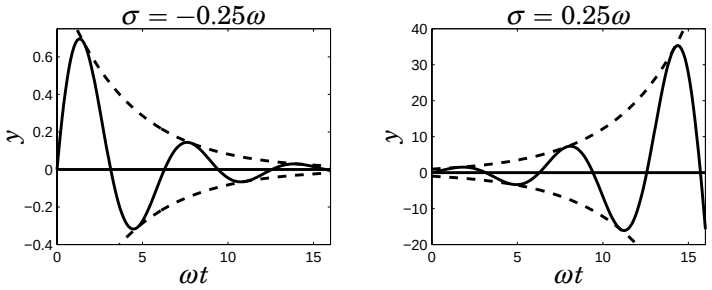


Figure 3.4 The exponential function $y(t) = e^{\sigma t} \sin \omega t$. The linear approximations of of the functions for small αt are shown in dashed lines. The dashed line corresponds to a first order system with time constant $T = 1/\sigma$. The distance between zero crossings is π/ω.

Multiple Roots

When there are multiple roots the solution to Equation (3.6) has the form

$$y(t) = \sum_{k=1}^{n} C_k(t) e^{\alpha_k t}, \tag{3.9}$$

Where $C_k(t)$ is a polynomial with degree less than the multiplicity of the root α_k. The solution (3.9) thus has n free parameters.

The Inhomogeneous Equation – A Special Case

The equation

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \ldots + a_n y = u(t) \tag{3.10}$$

has the solution

$$y(t) = \sum_{k=1}^{n} C_{k-1}(t)e^{\alpha_k t} + \int_0^t h(t-\tau)u(\tau)d\tau, \tag{3.11}$$

where f is the solution to the homogeneous equation

$$\frac{d^n f}{dt^n} + a_1 \frac{d^{n-1} f}{dt^{n-1}} + \ldots + a_n f = 0$$

with initial conditions

$$f(0) = 0, \quad f'(0) = 0, \ldots, \; f^{(n-2)}(0) = 0, \quad f^{(n-1)}(0) = 1. \tag{3.12}$$

The solution (3.11) is thus a sum of two terms, the general solution to the homogeneous equation and a particular solution which depends on the input u. The solution has n free parameters which can be determined from initial conditions.

The Inhomogeneous Equation - The General Case

The Equation (3.5) has the solution

$$y(t) = \sum_{k=1}^{n} C_{k-1}(t)e^{\alpha_k t} + \int_0^t g(t-\tau)u(\tau)d\tau, \tag{3.13}$$

where the function g, called the *impulse response*, is given by

$$g(t) = b_1 f^{(n-1)}(t) + b_2 f^{(n-2)}(t) + \; \ldots \; + b_n f(t). \tag{3.14}$$

The solution is thus the sum of two terms, the general solution to the homogeneous equation and a particular solution. The general solution to the homogeneous equation does not depend on the input and the particular solution depends on the input.

Notice that the impulse response has the form

$$g(t) = \sum_{k=1}^{n} c_k(t)e^{\alpha_k t}. \tag{3.15}$$

It thus has the same form as the general solution to the homogeneous equation (3.9). The coefficients c_k are given by the conditions (3.12).

The impulse response is also called the weighting function because the second term of (3.13) can be interpreted as a weighted sum of past inputs.

The Step Response

Consider (3.13) and assume that all initial conditions are zero. The output is then given by

$$y(t) = \int_0^t g(t-\tau)u(\tau)d\tau, \tag{3.16}$$

If the input is constant $u(t) = 1$ we get

$$y(t) = \int_0^t g(t-\tau)d\tau = \int_0^t g(\tau)d\tau = H(t), \tag{3.17}$$

The function H is called the unit step response or the step response for short. It follows from the above equation that

$$g(t) = \frac{dh(t)}{dt} \tag{3.18}$$

The step response can easily be determined experimentally by waiting for the system to come to rest and applying a constant input. In process engineering the experiment is called a bump test. The impulse response can then be determined by differentiating the step response.

Stability

The solution of system is described by the ordinary differential equation (3.5) is given by (3.9). The solution is stable if all solutions go to zero. A system is thus stable if the real parts of all α_i are negative, or equivalently that all the roots of the characteristic polynomial (3.7) have negative real parts.

Stability can be determined simply by finding the roots of the characteristic polynomial of a system. This is easily done in Matlab.

The Routh-Hurwitz Stability Criterion

When control started to be used for steam engines and electric generators computational tools were not available and it was it was a major effort to find roots of an algebraic equation. Much intellectual activity was devoted to the problem of investigating if an algebraic equation have all its roots in the left half plane without solving the equation resulting in the Routh-Hurwitz criterion. Some simple special cases of this criterion are given below.

- The polynomial $A(s) = s + a_1$ has its zero in the left half plane if $a_1 > 0$.
- The polynomial $A(s) = s^2 + a_1 s + a_2$ has all its zeros in the left half plane if all coefficients are positive.

- The polynomial $A(s) = s^3 + a_2 s^2 + a_3$ has all its zeros in the left half plane if all coefficients are positive and if $a_1 a_2 > a - 3$.

Transfer Functions, Poles and Zeros

The model (3.5) is characterized by two polynomials

$$A(s) = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_{n-1} s + a_n$$
$$B(s) = b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_{n-1} s + b_n$$

The rational function

$$G(s) = \frac{B(s)}{A(s)} \tag{3.19}$$

is called the transfer function of the system.

Consider a system described by (3.5) assume that the input and the output have constant values u_0 and y_0 respectively. It then follows from (3.5) that

$$a_n y_0 = b_n u_0$$

which implies that

$$\frac{y_0}{u_0} = \frac{b_n}{a_n} = G(0)$$

The number $G(0)$ is called the static gain of the system because it tells the ratio of the output and the input under steady state condition. If the input is constant $u = u_0$ and the system is stable then the output will reach the steady state value $y_0 = G(0)u_0$. The transfer function can thus be viewed as a generalization of the concept of gain.

Notice the symmetry between y and u. The *inverse system* is obtained by reversing the roles of input and output. The transfer function of the system is $\dfrac{B(s)}{A(s)}$ and the inverse system has the transfer function $\dfrac{A(s)}{B(s)}$.

The roots of $A(s)$ are called poles of the system. The roots of $B(s)$ are called zeros of the system. The poles of the system are the roots of the characteristic equation, they characterize the general solution to to the homogeneous equation and the impulse response. A pole $s = \lambda$ corresponds to the component $e^{\lambda t}$ of the solution, also called a mode. If $A(\alpha) = 0$, then $y(t) = e^{\alpha t}$ is a solution to the homogeneous equation (3.6). Differentiation gives

$$\frac{d^k y}{dt^k} = \alpha^k y(t)$$

and we find

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \ldots + a_n y = A(\alpha) y(t) = 0$$

The modes thus correspond to the terms of the solution to the homogeneous equation (3.6) and the terms of the impulse response (3.15) and the step response.

If $s = \beta$ is a zero of $B(s)$ and $u(t) = Ce^{\beta t}$, then it follows that

$$b_1 \frac{d^{n-1}u}{dt^{n-1}} + b_2 \frac{d^{n-2}u}{dt^{n-2}} \ldots + b_n u = B(\beta) C e^{\beta t} = 0.$$

A zero of $B(s)$ at $s = \beta$ blocks the transmission of the signal $u(t) = Ce^{\beta t}$.

3.4 Laplace Transforms

The Laplace transform is very convenient for dealing with linear time-invariant system. The reason is that it simplifies manipulations of linear systems to pure algebra. It also a natural way to introduce transfer functions and it also opens the road for using the powerful tools of the theory of complex variables. The Laplace transform is an essential element of the language of control.

The Laplace Transform

Consider a function f defined on $0 \leq t < \infty$ and a real number $\sigma > 0$. Assume that f grows slower than $e^{\sigma t}$ for large t. The Laplace transform $F = \mathcal{L}f$ of f is defined as

$$\mathcal{L}f = F(s) = \int_0^\infty e^{-st} f(t) dt$$

We will illustrate computation of Laplace transforms with a few examples

Transforms of Simple Function The transform of the function $f_1(t) = e^{-at}$ is given by

$$F_1(s) = \int_0^\infty e^{-(s+a)t} dt = -\frac{1}{s+a} e^{-st} \Big|_0^\infty = \frac{1}{s+a}$$

Differentiating the above equation we find that the transform of the function $f_2(t) = te\ -at$ is

$$F_2(s) = \frac{1}{(s+a)^2}$$

Repeated differentiation shows that the transform of the function $f_3(t) = t^n e^{-at}$ is

$$F_3(s) = \frac{(n-1)!}{(s+a)^n}$$

Setting $a = 0$ in f_1 we find that the transform of the unit step function $f_4(t) = 1$ is

$$F_4(s) = \frac{1}{s}$$

Similarly we find by setting $a = 0$ in f_3 that the transform of $f_5 = t^n$ is

$$F_5(s) = \frac{n!}{s^{n+1}}$$

Setting $a = ib$ in f_1 we find that the transform of $f(t) = e^{-ibt} = \cos bt - i \sin bt$ is

$$F(s) = \frac{1}{s+ib} = \frac{s-ib}{s^2+b^2} = \frac{s}{s^2+b^2} - i\frac{b}{s^2+b^2}$$

Separating real and imaginary parts we find that the transform of $f_6(t) = \sin bt$ and $f_7(t) = \cos bt$ are

$$F_6(t) = \frac{b}{s^2+b^2}, \quad F_7(t) = \frac{s}{s^2+b^2}$$

Proceeding in this way it is possible to build up tables of transforms that are useful for hand calculations.

Properties of Laplace Transforms The Laplace transform also has many useful properties. First we observe that the transform is linear because

$$\begin{aligned}\mathcal{L}(af+bg) = aF(s) + bF(s) &= a\int_0^\infty e^{-st} f(t)dt + b\int_0^\infty e^{-st} g(t)dt \\ &= \int_0^\infty e^{-st}(af(t)+bg(t))dt = a\mathcal{L}f + b\mathcal{L}g\end{aligned}$$

Next we will calculate the transform of the derivative of a function, i.e. $f'(t) = \frac{df(t)}{dt}$. We have

$$\mathcal{L}\frac{df}{dt} = \int_0^\infty e^{-st} f'(t)dt = e^{-st} f(t)\Big|_0^\infty + s\int_0^\infty e^{-st} f(t)dt = -f(0) + s\mathcal{L}f$$

where the second equality is obtained by integration by parts. This formula is very useful because it implies that differentiation of a time function corresponds to multiplication of the transform by s provided that the

initial value $f(0)$ is zero. We will consider the transform of an integral

$$\begin{aligned}\mathcal{L}\int_0^t f(\tau)d\tau &= \int_0^\infty e^{-st}\int_0^t f(\tau)d\tau \\ &= -\frac{e^{-st}}{s}\int_0^t e^{-s\tau}f'(\tau)d\tau\Big|_0^\infty + \int_0^\infty \frac{e^{-s\tau}}{s}f(\tau)d\tau \\ &= \frac{1}{s}\int_0^\infty e^{-s\tau}f(\tau)d\tau = \mathcal{L}f\end{aligned}$$

The relation between the input u and the output y of a linear time-invariant system is given by the convolution integral

$$y(t) = \int_0^\infty g(t-\tau)u(\tau)d\tau$$

see (3.18). We will now consider the Laplace transform of such an expression. We have

$$\begin{aligned}Y(s) &= \int_0^\infty e^{-st}y(t)dt = \int_0^\infty e^{-st}\int_0^\infty g(t-\tau)u(\tau)d\tau dt \\ &= \int_0^\infty \int_0^t e^{-s(t-\tau)}e^{-s\tau}g(t-\tau)u(\tau)d\tau dt \\ &= \int_0^\infty e^{-s\tau}u(\tau)d\tau \int_0^\infty e^{-st}g(t)dt = G(s)U(s)\end{aligned}$$

The description of a linear time-invariant systems thus becomes very simple when working with Laplace transforms.

Next we will consider the effect of a time shift. Let the number a be positive and let the function f_a be a time shift of the function f, i.e.

$$f_a(t) = \begin{cases} 0 & \text{for PI } t < 0 \\ f(t-a) & \text{for } t \geq 0 \end{cases}$$

The Laplace transform of f_a is given by

$$\begin{aligned}F_a(s) &= \int_0^\infty e^{-st}f(t-a)dt = \int_a^\infty e^{-st}f(t-a)dt \\ &= \int_a^\infty e^{-as}e^{-s(t-a)}f(t-a)dt = e^{-as}\int_0^\infty e^{-st}f(t)dt = e^{-as}F(s)\end{aligned} \tag{3.20}$$

Delaying a signal by a time units thus correspond to multiplication of its Laplace transform by e^{-as}.

The behavior of time functions for small arguments is governed by the behavior of the Laplace transform for large arguments. This is expressed by the so called initial value theorem.

$$\lim_{s\to\infty} sF(s) = \lim_{s\to\infty} \int_0^\infty se^{-st} f(t)dt = \lim_{s\to\infty} \int_0^\infty e^{-v} f(\frac{v}{s})dv = f(0)$$

This holds provided that the limit exists.

The converse is also true which means that the behavior of time functions for large arguments is governed by the behavior of the Laplace transform for small arguments. Final value theorem. Hence

$$\lim_{s\to 0} sF(s) = \lim_{s\to 0} \int_0^\infty se^{-st} f(t)dt = \lim_{s\to 0} \int_0^\infty e^{-v} f(\frac{v}{s})dv = f(\infty)$$

These properties are very useful for qualitative assessment of a time functions and Laplace transforms.

Linear Differential Equations

The differentiation property $\mathcal{L}\frac{df}{dt} = s\mathcal{L}f - f(0)$ makes the Laplace transform very convenient for dealing with linear differential equations. Consider for example the system

$$\frac{dy}{dt} = ay + bu$$

Taking Laplace transforms of both sides give

$$sY(s) - y(0) = aY(s) + bU(s)$$

Solving this linear equation for $Y(s)$ gives

$$Y(s) = \frac{y(0)}{s-a} + \frac{b}{s-a}U(s)$$

Transforming back to time function we find

$$y(t) = e^{at}y(0) + b\int_0^t e^{a(t-\tau)}u(\tau]d\tau$$

To convert the transforms to time functions we have used the fact that the transform

$$\frac{1}{s-a}$$

corresponds to the time function e^{at} and we have also used the rule for transforms of convolution.

Inverse Transforms

A simple way to find time functions corresponding to a rational Laplace transform. Write $F(s)$ in a partial fraction expansion

$$F(s) = \frac{B(s)}{A(s)} = \frac{B(s)}{(s-\alpha_1)(s-\alpha_2)\dots(s-\alpha_n)} = \frac{C_1}{s-\alpha_1} + \frac{C_2}{s-\alpha_2} + \dots + \frac{C_n}{s-\alpha_n}$$
$$C_k = \lim_{s\to\alpha_k}(s-\alpha_k)F(s) = \frac{B(\alpha_k)}{(\alpha_k-\alpha_1)\dots(\alpha_k-\alpha_{k-1})(s-\alpha_{k+1})\dots(\alpha_k-\alpha_n)}$$

The time function corresponding to the transform is

$$f(t) = C_1 e^{\alpha_1 t} + C_2 e^{\alpha_2 t} + \dots + C_n e^{\alpha_n t}$$

Parameters α_k give shape and numbers C_k give magnitudes.

Notice that α_k may be complex numbers. With multiple roots the constants C_k are instead polynomials.

The Transfer Function

The transfer function of an LTI system was introduced in Section 3.3 when dealing with differential equations. Using Laplace transforms it can also be defined as follows. Consider an LTI system with input u and output y. The transfer function is the ratio of the transform of the output and the input where *the Laplace transforms are calculated under the assumption that all initial values are zero.*

$$G(s) = \frac{Y(s)}{U(s)} = \frac{\mathcal{L}y}{\mathcal{L}u}$$

The fact that all initial values are assumed to be zero has some consequences that will be discussed later.

EXAMPLE 3.1—LINEAR TIME-INVARIANT SYSTEMS

Consider a system described by the ordinary differential equation (3.5), i.e

$$\frac{d^n y}{dt^n} + a_1\frac{d^{n-1}y}{dt^{n-1}} + \dots + a_n y = b_1\frac{d^{n-1}u}{dt^{n-1}} + b_2\frac{d^{n-2}u}{dt^{n-2}} + \dots + b_n u,$$

Taking Laplace transforms under the assumption that all initial values are zero we get.

$$\begin{aligned}(s^n + a_1 s^{n-1} + a_2 s^{n-2} + \dots + a_{n-1}s + a_n)Y(s)\\ = (b_1 s^{n-1} + b_2 s^{n-2} + \dots + b_{n-1}s + b_n)U(s)\end{aligned}$$

The transfer function of the system is thus given by

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_{n-1} s + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_{n-1} s + a_n} = \frac{B(s)}{A(s)} \quad (3.21)$$

□

EXAMPLE 3.2—A TIME DELAY
Consider a system which simply delays the input T time units. It follows from 3.20 that the input output relation is

$$Y(s) = e^{-sT} U(s)$$

The transfer function of a time delay is thus

$$G(s) = \frac{Y(s)}{U(s)} = e^{-sT}$$

□

It is also possible to calculate the transfer functions for systems described by partial differential equations.

EXAMPLE 3.3—THE HEAT EQUATION

$$G(s) = e^{-\sqrt{sT}}$$

$$G(s) = \frac{1}{\cosh \sqrt{sT}}$$

□

Transfer functions and Laplace transforms are ideal to deal with block diagrams for linear time-invariant systems. We have already shown that a block is simply characterized by

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

The transform of the output of a block is simply the product of the transfer function of the block and the transform of the input system. Algebraically this is equivalent to multiplication with a constant. This makes it easy to find relations between the signals that appear in a block diagram. The combination of block diagrams and transfer functions is a very nice combination because they make it possible both to obtain an overview of a system and to guide the derivation of equations for the system. This is one of the reasons why block diagrams are so widely used in control.

Notice that it also follows from the above equation that signals and systems have the same representations. In the formula we can thus consider g as the input and u as the transfer function.

To illustrate the idea we will consider an example.

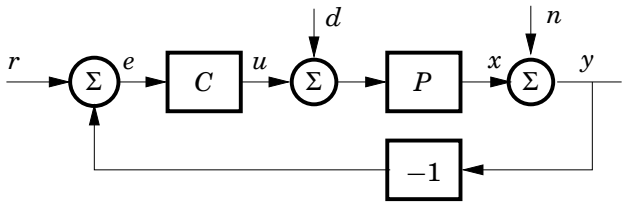


Figure 3.5 Block diagram of a feedback system.

EXAMPLE 3.4—RELATIONS BETWEEN SIGNALS IN A BLOCK DIAGRAM
Consider the system in Figure 3.5. The system has two blocks representing the process P and the controller C. There are three external signals, the reference r, the load disturbance d and the measurement noise n. A typical problem is to find out how the error e related to the signals r d and n? Introduce Laplace transforms and transfer functions. To obtain the desired relation we simply trace the signals around the loop. Starting with the signal e and tracing backwards in the loop we find that e is the difference between r and y, hence $E = R - Y$. The signal y in turn is the sum of n and the output of the block P, hence $Y = N + P(D + V)$. Finally the signal v is the output of the controller which is given by $V = PE$. Combining the results we get

$$E = R - \big(N + P(D + CE)\big)$$

With a little practice this equation can be written directly. Solving for E gives

$$E = \frac{1}{1+PC}R - \frac{1}{1+PC}N - \frac{P}{1+PC}D$$

□

Notice *the form of the equations* and the use of *superposition*.

Simulating LTI Systems

Linear time-invariant systems can be conveniently simulated using Matlab. For example a system with the transfer function

$$G(s) = \frac{5s+2}{s^2+3s+2}$$

is introduced in matlab as

```
G=tf([5 2],[1 3 2])
```

The command `step(G)` gives the step response of the system.

Transfer Functions

The transfer function of a linear system is defined as

$$G(s) = \frac{Y(s)}{U(s)} = \frac{\mathcal{L}y}{\mathcal{L}u} \tag{3.22}$$

where $U(s) = \mathcal{L}u$ is the Laplace transform of the input u and $Y(s) = \mathcal{L}y$ is the Laplace transform of the output y. The Laplace transforms are computed under the assumption that all initial conditions are zero.

Circuit Analysis

Laplace transforms are very useful for circuit analysis. A resistor is described by the algebraic equation

$$V = RI$$

but inductors and capacitors are describe by the linear differential equations

$$CV = \int_0^t I(\tau]d\tau$$
$$L\frac{dI}{dt} = V$$

Taking Laplace transforms we get

$$\mathcal{L}V = RI$$
$$\mathcal{L}V = \frac{1}{sC}\mathcal{L}I$$
$$\mathcal{L}V = sL\mathcal{L}I$$

The transformed equations for all components thus look identical, the transformed voltage $\mathcal{L}V$ is a generalized impedance Z multiplied by the transformed current $\mathcal{L}I$. The impedance is

$Z(s) = R$ for a resistor

$Z(s) = \frac{1}{sC}$ for a capacitor

$Z(s) = sL$ for an inductor

Operating with the transforms we can thus pretend that all elements of a circuit is a resistor which means that circuit analysis is reduced to pure algebra. This is just another illustration of the fact that differential equations are transformed to algebraic equations. We illustrate the procedure by an example.

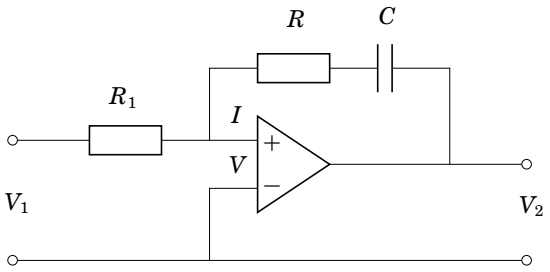


Figure 3.6 Schematic diagram of an electric circuit.

EXAMPLE 3.5—OPERATIONAL AMPLIFIERS
Consider the electric circuit shown in Figure 3.6. Assume that the problem is to find the relation between the input voltage V_1 and the output voltage V_2. Assuming that the gain of the amplifier is very high, say around 10^6, then the voltage V is negligible and the current I_0 is zero. The currents I_1 and I_2 then are the same which gives

$$\frac{\mathcal{L}V_1}{Z_1(s)} = -\frac{\mathcal{L}V_2}{Z_2(s)}$$

It now remains to determine the generalized impedances Z_1 and Z_2. The impedance Z_2 is a regular resistor. To determine Z_1 we use the simple rule for combining resistors which gives

$$Z_1(s) = R + \frac{1}{sC}$$

Hence

$$\frac{\mathcal{L}V_2}{\mathcal{L}V_2} = -\frac{Z_1(s)}{Z_2(s)} = -\frac{R}{R_2} - \frac{1}{R_2 C s}$$

Converting to the time domain we find

$$V_2(t) = -\frac{R}{R_2} - \frac{1}{R_2 C}\int_0^t V_1(\tau)d\tau$$

The circuit is thus a PI controller. □

3.5 Frequency Response

The idea of frequency response is to characterize linear time-invariant systems by their response to sinusoidal signals. The idea goes back to

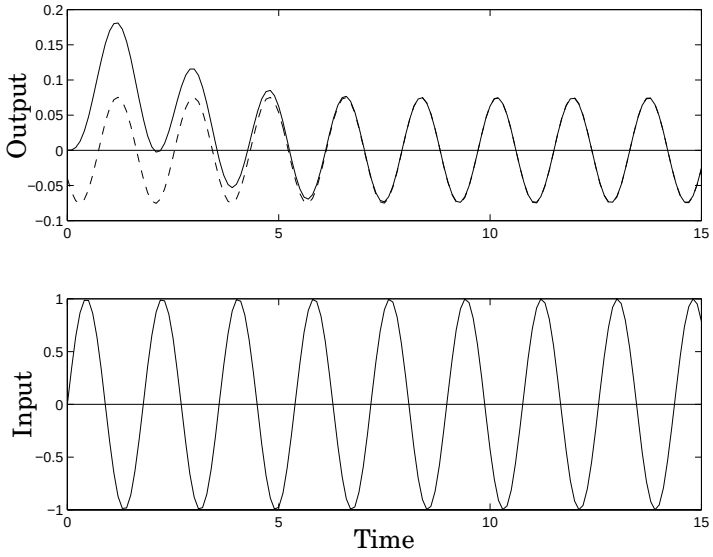


Figure 3.7 Response of a linear time-invariant system to a sinusoidal input (full lines). The dashed line shows the steady state output calculated from (3.23).

Fourier, who introduced the method to investigate propagation of heat in metals. Figure 3.7 shows the response of a linear time-invariant system to a sinusoidal input. The figure indicates that after a transient the output is a sinusoid with the same frequency as the input. The steady state response to a sinusoidal input of a stable linear system is in fact given by $G(i\omega)$. Hence if the input is

$$u(t) = a \sin \omega t = a\Im e^{i\omega t}$$

the output is

$$y(t) = a|G(i\omega)| \sin\left(\omega t + \arg G(i\omega)\right) = a\Im e^{i\omega t} G(i\omega) \tag{3.23}$$

The dashed line in Figure 3.7 shows the output calculated by this formula. It follows from this equation that the transfer function G has the interesting property that its value for $s = i\omega$ describes the steady state response to sinusoidal signals. The function $G(i\omega)$ is therefore called the frequency response. The argument of the function is frequency ω and the function takes complex values. The magnitude gives the magnitude of the steady state output for a unit amplitude sinusoidal input and the argument gives the phase shift between the input and the output. Notice that the system must be stable for the steady state output to exist.

The frequency response can be determined experimentally by analyzing how a system responds to sinusoidal signals. It is possible to make very accurate measurements by using correlation techniques.

To derive the formula we will first calculate the response of a system with the transfer function $G(s)$ to the signal e^{at} where all poles of the transfer function have the property $\Re p_k < \alpha < a$. The Laplace transform of the output is

$$Y(s) = G(s)\frac{1}{s-a}$$

Making a partial fraction expansion we get

$$Y(s) = \frac{G(a)}{s-a} + \sum \frac{R_k}{(s-p_k)(p_k-a)}$$

It follows the output has the property

$$\left|y(t) - G(a)e^{at}\right| < ce^{-\alpha t}$$

where c is a constant. Asymptotically we thus find that the output approaches $G(a)e^{at}$. Setting $a = ib$ we find that the response to the input

$$u(t) = e^{ibt} = \cos bt + i \sin bt$$

will approach

$$\begin{aligned} y(t) &= G(ib)e^{ibt} = |G(ib)|e^{i(b+\arg G(ib))} \\ &= |G(ib)|\cos\left(bt + \arg\arg G(ib)\right) + i|G(ib)|\sin\left(bt + \arg\arg G(ib)\right) \end{aligned}$$

Separation of real and imaginary parts give the result.

Nyquist Plots

The response of a system to sinusoids is given by the the frequency response $G(i\omega)$. This function can be represented graphically by plotting the magnitude and phase of $G(i\omega)$ for all frequencies, see Figure 3.8. The magnitude $a = |G(i\omega)|$ represents the amplitude of the output and the angle $\phi = \arg G(i\omega)$ represents the phase shift. The phase shift is typically negative which implies that the output will lag the input. The angle ψ in the figure is therefore called phase lag. One reason why the Nyquist curve is important is that it gives a totally new way of looking at stability of a feedback system. Consider the feedback system in Figure 3.9. To investigate stability of a the system we have to derive the characteristic equation of the closed loop system and determine if all its roots are in the left half plane. Even if it easy to determine the roots of the equation

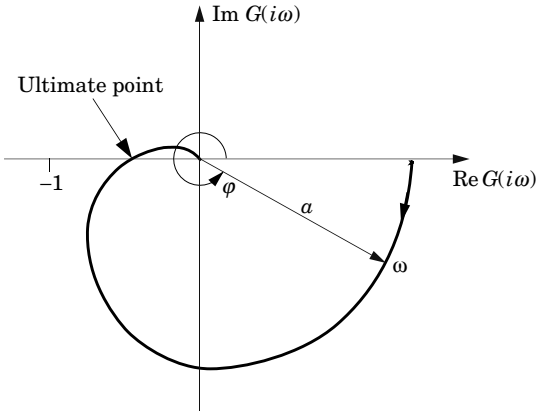


Figure 3.8 The Nyquist plot of a transfer function $G(i\omega)$.

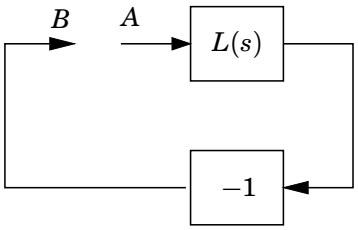


Figure 3.9 Block diagram of a simple feedback system.

numerically it is not easy to determine how the roots are influenced by the properties of the controller. It is for example not easy to see how to modify the controller if the closed loop system is stable. We have also defined stability as a binary property, a system is either stable or unstable. In practice it is useful to be able to talk about degrees of stability. All of these issues are addressed by Nyquist's stability criterion. This result has a strong intuitive component which we will discuss first. There is also some beautiful mathematics associated with it that will be discussed in a separate section.

Consider the feedback system in Figure 3.9. Let the transfer functions of the process and the controller be $P(s)$ and $C(s)$ respectively. Introduce the loop transfer function

$$L(s) = P(s)C(s) \tag{3.24}$$

To get insight into the problem of stability we will start by investigating

the conditions for oscillation. For that purpose we cut the feedback loop as indicated in the figure and we inject a sinusoid at point A. In steady state the signal at point B will also be a sinusoid with the same frequency. It seems reasonable that an oscillation can be maintained if the signal at B has the same amplitude and phase as the injected signal because we could then connect A to B. Tracing signals around the loop we find that the condition that the signal at B is identical to the signal at A is that

$$L(i\omega_0) = -1 \tag{3.25}$$

which we call the condition for oscillation. This condition means that the Nyquist curve of $L(i\omega)$ intersects the negative real axis at the point -1. Intuitively it seems reasonable that the system would be stable if the Nyquist curve intersects to the right of the point -1 as indicated in Figure 3.9. This is essentially true, but there are several subtleties that are revealed by the proper theory.

Stability Margins

In practice it is not enough to require that the system is stable. There must also be some margins of stability. There are many ways to express this. Many of the criteria are based on Nyquist's stability criterion. They are based on the fact that it is easy to see the effects of changes of the gain and the phase of the controller in the Nyquist diagram of the loop transfer function $L(s)$. An increase of controller gain simply expands the Nyquist curve radially. An increase of the phase of the controller twists the Nyquist curve clockwise, see Figure 3.10. The gain margin g_m tells how much the controller gain can be increased before reaching the stability limit. Let ω_{180} be the smallest frequency where the phase lag of the loop transfer function $L(s)$ is 180°. The gain margin is defined as

$$g_m = \frac{1}{|L(i\omega_{180})|} \tag{3.26}$$

The stability margin is a closely related concept which is defined as

$$s_m = 1 + |L(i\omega_{180})| = 1 - \frac{1}{g_m} \tag{3.27}$$

A nice feature of the stability margin is that it is a number between 0 and 1. Values close to zero imply a small margin.

The phase margin φ_m is the amount of phase lag required to reach the stability limit. Let ω_{gc} denote the lowest frequency where the loop transfer function $L(s)$ has unit magnitude. The phase margin is then given by

$$\varphi_m = \pi + \arg L(i\omega_{gc}) \tag{3.28}$$

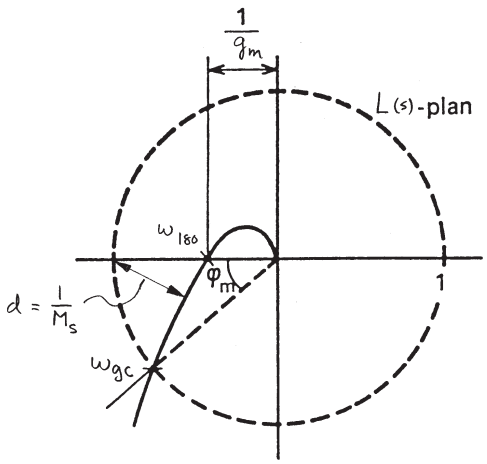


Figure 3.10 Nyquist curve of the loop transfer function L with indication of gain, phase and stability margins.

The margins have simple geometric interpretations in the Nyquist diagram of the loop transfer function as is shown in Figure 3.10. The stability margin s_m is the distance between the critical point and the intersection of the Nyquist curve with the negative real axis.

One possibility to characterize the stability margin with a single number is to choose the shortest distance d to the critical point. This is also shown in Figure 3.10.

Reasonable values of the margins are phase margin $\varphi_m = 30^\circ - 60^\circ$, gain margin $g_m = 2 - 5$, stability margin $s_m = 0.5 - 0.8$, and shortest distance to the critical point $d = 0.5 - 0.8$.

The gain and phase margins were originally conceived for the case when the Nyquist curve only intersects the unit circle and the negative real axis once. For more complicated systems there may be many intersections and it is then necessary to consider the intersections that are closest to the critical point. For more complicated systems there is also another number that is highly relevant namely the delay margin. The delay margin is defined as the smallest time delay required to make the system unstable. For loop transfer functions that decay quickly the delay margin is closely related to the phase margin but for systems where the amplitude ratio of the loop transfer function has several peaks at high frequencies the delay margin is a much more relevant measure.

Nyquist's Stability Theorem*

We will now prove the Nyquist stability theorem. This will require more results from the theory of complex variables than in many other parts of the book. Since precision is needed we will also use a more mathematical style of presentation. We will start by proving a key theorem about functions of complex variables.

THEOREM 3.1—PRINCIPLE OF VARIATION OF THE ARGUMENT
Let D be a closed region in the complex plane and let Γ be the boundary of the region. Assume the function f is analytic in D and on Γ except at a finite number of poles and zeros, then

$$w_n = \frac{1}{2\pi}\Delta_\Gamma \arg f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f'(z)}{f(z)}dz = N - P$$

where N is the number of zeros and P the number of poles in D. Poles and zeros of multiplicity m are counted m times. The number w_n is called the winding number and $\Delta_\Gamma \arg f(z)$ is the variation of the argument of the function f as the curve Γ is traversed in the positive direction.

PROOF 3.1
Assume that $z = a$ is a zero of multiplicity m. In the neighborhood of $z = a$ we have

$$f(z) = (z-a)^m g(z)$$

where the function g is analytic and different form zero. We have

$$\frac{f'(z)}{f(z)} = \frac{m}{z-a} + \frac{g'(z)}{g(z)}$$

The second term is analytic at $z = a$. The function f'/f thus has a single pole at $z = a$ with the residue m. The sum of the residues at the zeros of the function is N. Similarly we find that the sum of the residues of the poles of is $-P$. Furthermore we have

$$\frac{d}{dz}\log f(z) = \frac{f'(z)}{f(z)}$$

which implies that

$$\int_\Gamma \frac{f'(z)}{f(z)}dz = \Delta_\Gamma \log f(z)$$

where Δ_Γ denotes the variation along the contour Γ. We have

$$\log f(z) = \log|f(z)| + i\arg f(z)$$

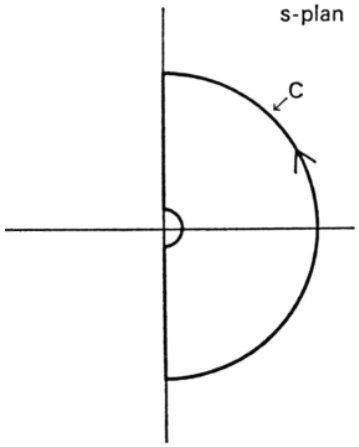


Figure 3.11 Contour Γ used to prove Nyquist's stability theorem.

Since the variation of $|f(z)|$ around a closed contour is zero we have

$$\Delta_\Gamma \log f(z) = i\Delta_\Gamma \arg f(z)$$

and the theorem is proven.

REMARK 3.1
The number w_n is called the winding number.

REMARK 3.2
The theorem is useful to determine the number of poles and zeros of an function of complex variables in a given region. To use the result we must determine the winding number. One way to do this is to investigate how the curve Γ is transformed under the map f. The variation of the argument is the number of times the map of Γ winds around the origin in the f-plane. This explains why the variation of the argument is also called the winding number. □

We will now use the Theorem 1 to prove Nyquist's stability theorem. For that purpose we introduce a contour that encloses the right half plane. For that purpose we choose the contour shown in Figure 3.11. The contour consists of a small half circle to the right of the origin, the imaginary axis and a large half circle to the right with with the imaginary axis as a diameter. To illustrate the contour we have shown it drawn with a small radius r and a large radius R. The Nyquist curve is normally the map

of the positive imaginary axis. We call the contour Γ the full Nyquist contour.

Consider a closed loop system with the loop transfer function $L(s)$. The closed loop poles are the zeros of the function

$$f(s) = 1 + L(s)$$

To find the number of zeros in the right half plane we thus have to investigate the winding number of the function $f = 1 + L$ as s moves along the contour Γ. The winding number can conveniently be determined from the Nyquist plot. A direct application of the Theorem 1 gives.

THEOREM 3.2—NYQUIST'S STABILITY THEOREM
Consider a simple closed loop system with the loop transfer function $L(s)$. Assume that the loop transfer function does not have any poles in the region enclosed by Γ and that the winding number of the function $1+L(s)$ is zero. Then the closed loop characteristic equation has not zeros in the right half plane.

We illustrate Nyquist's theorem by an examples.

EXAMPLE 3.6—A SIMPLE CASE
Consider a closed loop system with the loop transfer function

$$L(s) = \frac{k}{s((s+1)^2}$$

Figure 3.12 shows the image of the contour Γ under the map L. The Nyquist curve intersects the imaginary axis for $\omega = 1$ the intersection is at $-k/2$. It follows from Figure 3.12 that the winding number is zero if $k < 2$ and 2 if $k > 2$. We can thus conclude that the closed loop system is stable if $k < 2$ and that the closed loop system has two roots in the right half plane if $k > 2$. □

By using Nyquist's theorem it was possible to resolve a problem that had puzzled the engineers working with feedback amplifiers. The following quote by Nyquist gives an interesting perspective.

> Mr. Black proposed a negative feedback repeater and proved by tests that it possessed the advantages which he had predicted for it. In particular, its gain was constant to a high degree, and it was linear enough so that spurious signals caused by the interaction of the various channels could be kept within permissible limits. For best results, the feedback factor, the quantity usually known as $\mu\beta$ (the loop transfer function), had to be numerically much larger than unity. The possibility of stability with a feedback factor greater than unity was

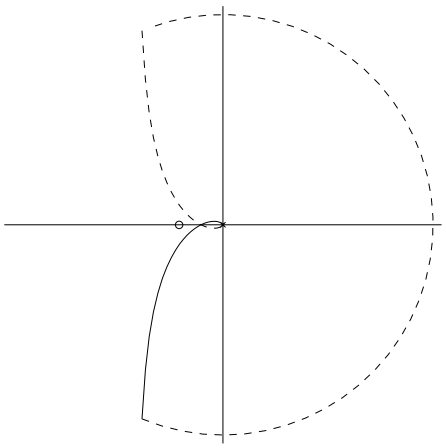


Figure 3.12 Map of the contour Γ under the map $L(s) = \frac{k}{s((s+1)^2}$. The curve is drawn for $k < 2$. The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semi circle at the origin in dashed lines.

puzzling. Granted that the factor is negative it was not obvious how that would help. If the factor was -10, the effect of one round trip around the feedback loop is to change the magnitude of the current from, say 1 to -10. After a second trip around the loop the current becomes 100, and so forth. The totality looks much like a divergent series and it was not clear how such a succession of ever-increasing components could add to something finite and so stable as experience had shown. The missing part in this argument is that the numbers that describe the successive components 1, -10, 100, and so on, represent the steady state, whereas at any finite time many of the components have not yet reached steady state and some of them, which are destined to become very large, have barely reached perceptible magnitude. My calculations were principally concerned with replacing the indefinite diverging series referred to by a series which gives the actual value attained at a specific time t. The series thus obtained is convergent instead of divergent and, moreover, converges to values in agreement with the experimental findings.

This explains how I came to undertake the work. It should perhaps be explained also how it come to be so detailed. In the course of the calculations, the facts with which the term conditional stability have come to be associated, became apparent. One aspect of this was that it is possible to have a feedback loop which is stable and can be made unstable by by increasing the loop loss. this seemed a very surprising

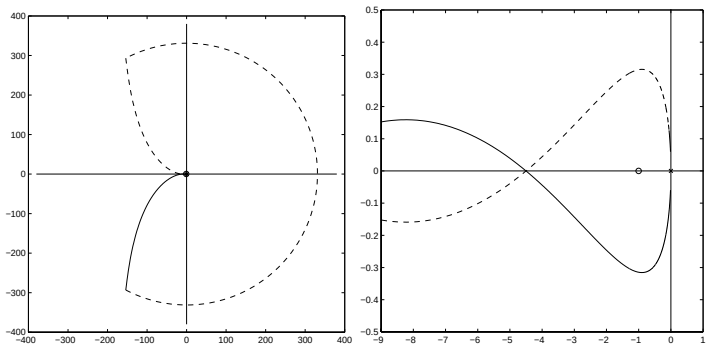


Figure 3.13 Map of the contour Γ under the map $L(s) = \frac{3(s+1)^2}{s(s+6)^2}$, see (3.29), which is a conditionally stable system. The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semicircle at the origin in dashed lines. The plot on the right is an enlargement of the area around the origin of the plot on the left.

> result and appeared to require that all the steps be examined and set forth in full detail.

This quote clearly illustrate the difficulty in understanding feedback by simple qualitative reasoning. We will illustrate the issue of conditional stability by an example.

EXAMPLE 3.7—CONDITIONAL STABILITY
Consider a feedback system with the loop transfer function

$$L(s) = \frac{3(s+1)^2}{s(s+6)^2} \tag{3.29}$$

The Nyquist plot of the loop transfer function is shown in Figure 3.13 The figure shows that the Nyquist curve intersects the negative real axis at a point close to -5. The naive argument would then indicate that the system would be unstable. The winding number is however zero and stability follows from Nyquist's theorem. □

Notice that Nyquist's theorem does not hold if the loop transfer function has a pole in the right half plane. There are extensions of the Nyquist theorem to cover this case but it is simpler to invoke Theorem 1 directly. We illustrate this by two examples.

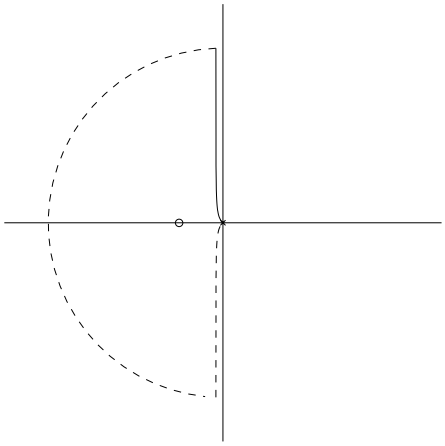


Figure 3.14 Map of the contour Γ under the map $L(s) = \frac{k}{s(s-1)(s+5)}$. The curve on the right shows the region around the origin in larger scale. The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semi circle at the origin in dashed lines.

EXAMPLE 3.8—LOOP TRANSFER FUNCTION WITH RHP POLE
Consider a feedback system with the loop transfer function

$$L(s) = \frac{k}{s(s-1)(s+5)}$$

This transfer function has a pole at $s = 1$ in the right half plane. This violates one of the assumptions for Nyquist's theorem to be valid. The Nyquist curve of the loop transfer function is shown in Figure 3.14. Traversing the contour Γ in clockwise we find that the winding number is 1. Applying Theorem 1 we find that

$$N - P = 1$$

Since the loop transfer function has a pole in the right half plane we have $P = 1$ and we get $N = 2$. The characteristic equation thus has two roots in the right half plane. □

EXAMPLE 3.9—THE INVERTED PENDULUM
Consider a closed loop system for stabilization of an inverted pendulum with a PD controller. The loop transfer function is

$$L(s) = \frac{s+2}{s^2-1} \tag{3.30}$$

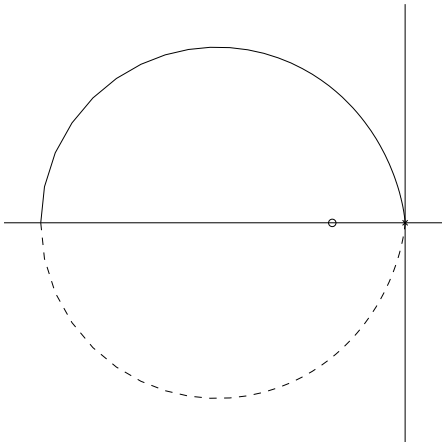


Figure 3.15 Map of the contour Γ under the map $L(s) = \frac{s+2}{s^2-1}$ given by (3.30). The map of the positive imaginary axis is shown in full lines, the map of the negative imaginary axis and the small semi circle at the origin in dashed lines.

This transfer function has one pole at $s = 1$ in the right half plane. The Nyquist curve of the loop transfer function is shown in Figure 3.15. Traversing the contour Γ in clockwise we find that the winding number is -1. Applying Theorem 1 we find that

$$N - P = -1$$

Since the loop transfer function has a pole in the right half plane we have $P = 1$ and we get $N = 0$. The characteristic equation thus has no roots in the right half plane and the closed loop system is stable. □

Bode Plots

The Nyquist curve is one way to represent the frequency response $G(i\omega)$. Another useful representation was proposed by Bode who represented it by two curves, the gain curve and the phase curve. The gain curve gives the value of $G(i\omega)$ as a function of ω and the phase curve gives $\arg G(i\omega)$ as a function of ω. The curves are plotted as shown below with logarithmic scales for frequency and magnitude and linear scale for phase, see Figure 3.16 An useful feature of the Bode plot is that both the gain curve and the phase curve can be approximated by straight lines, see Figure 3.16 where the approximation is shown in dashed lines. This fact was particularly useful when computing tools were not easily accessible.

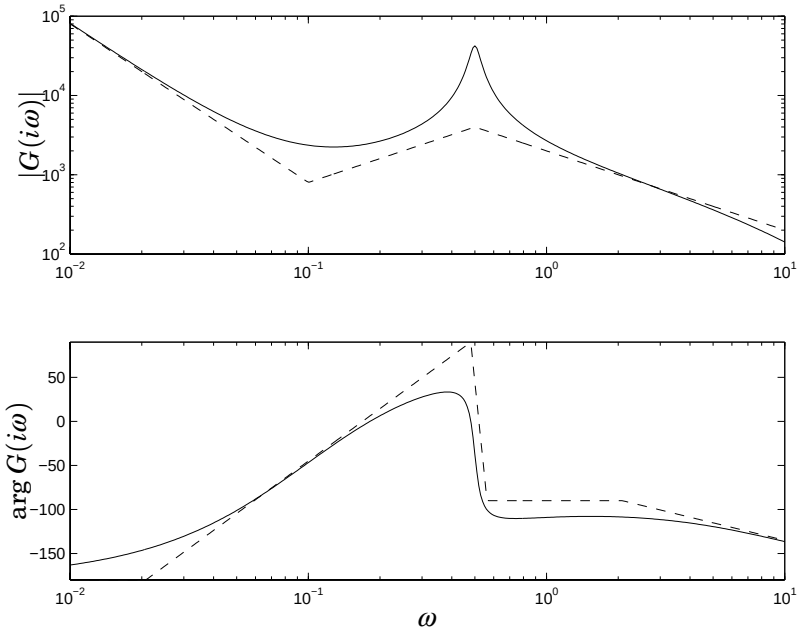


Figure 3.16 Bode diagram of a frequency response. The top plot is the gain curve and bottom plot is the phase curve. The dashed lines show straight line approximations of the curves.

The fact that logarithmic scales were used also simplified the plotting. We illustrate Bode plots with a few examples.

It is easy to sketch Bode plots because with the right scales they have linear asymptotes. This is useful in order to get a quick estimate of the behavior of a system. It is also a good way to check numerical calculations.

Consider first a transfer function which is a polynomial $G(s) = B(s)/A(s)$. We have

$$\log G(s) = \log B(s) - \log A(s)$$

Since a polynomial is a product of terms of the type :

$$s, \quad s + a, \quad s^2 + 2\zeta as + a^2$$

it suffices to be able to sketch Bode diagrams for these terms. The Bode plot of a complex system is then obtained by composition.

EXAMPLE 3.10—BODE PLOT OF A DIFFERENTIATOR
Consider the transfer function

$$G(s) = s$$

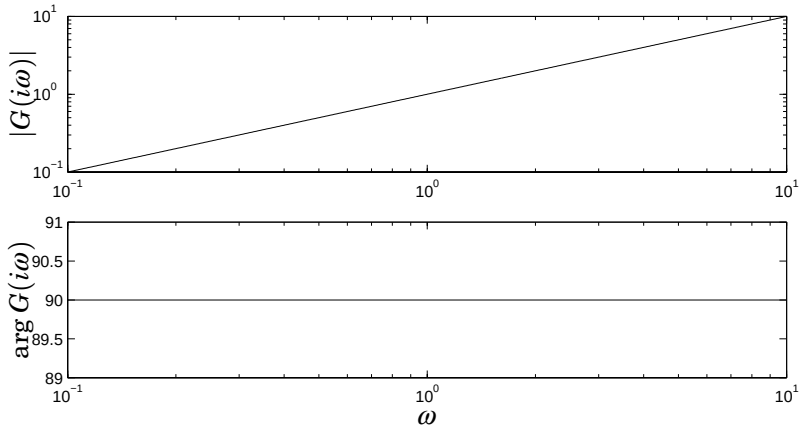


Figure 3.17 Bode plot of a differentiator.

We have $G(i\omega) = i\omega$ which implies

$$\log|G(i\omega)| = \log\omega$$
$$\arg G(i\omega) = \pi/2$$

The gain curve is thus a straight line with slope 1 and the phase curve is a constant at 90°.. The Bode plot is shown in Figure 3.17 □

EXAMPLE 3.11—BODE PLOT OF AN INTEGRATOR
Consider the transfer function

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

We have $G(i\omega) = 1/i\omega$ which implies

$$\log|G(i\omega)| = -\log\omega$$
$$\arg G(i\omega) = -\pi/2$$

The gain curve is thus a straight line with slope -1 and the phase curve is a constant at $-90°$. The Bode plot is shown in Figure 3.18 □

Compare the Bode plots for the differentiator in Figure 3.17 and the integrator in Figure 3.18. The sign of the phase is reversed and the gain curve is mirror imaged in the horizontal axis. This is a consequence of the property of the logarithm.

$$log\frac{1}{G} = -logG = -log|G| - i\arg G$$

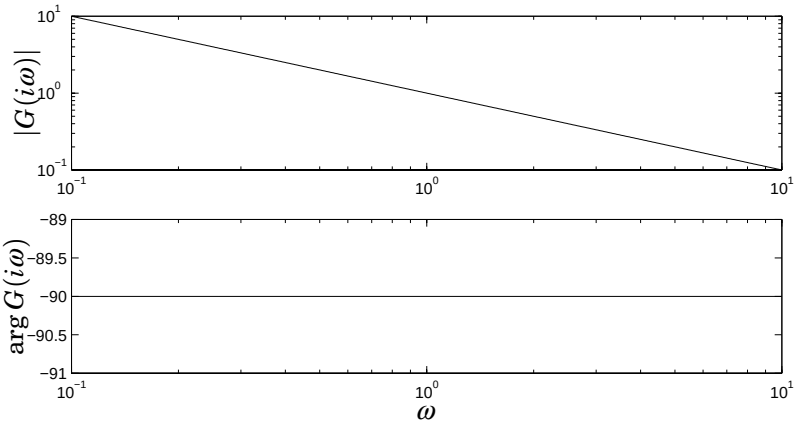


Figure 3.18 Bode plot of an integrator.

EXAMPLE 3.12—BODE PLOT OF A FIRST ORDER FACTOR
Consider the transfer function

$$G(s) = s + a$$

We have

$$G(i\omega) = a + i\omega$$

and it follows that

$$|G(i\omega)| = \sqrt{\omega^2 + a^2}, \quad \arg G(i\omega) = \arctan \omega/a$$

Hence

$$\log |G(i\omega)| = \frac{1}{2} \log (\omega^2 + a^2), \quad \arg G(i\omega) = \arctan \omega/a$$

The Bode Plot is shown in Figure 3.19. Both the gain curve and the phase curve can be approximated by straight lines if proper scales are chosen

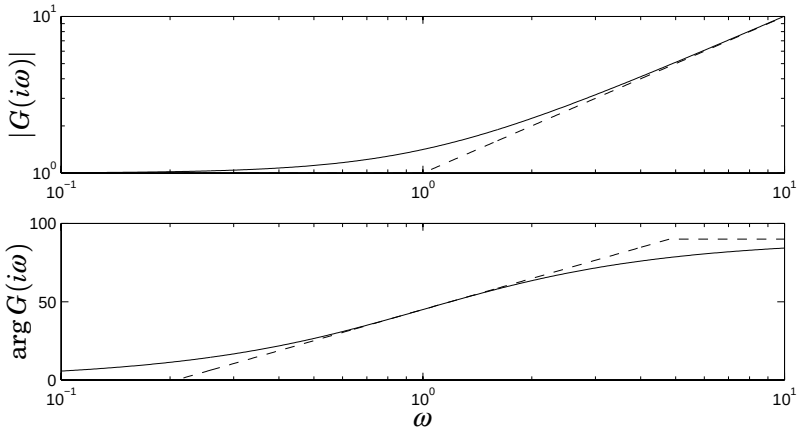


Figure 3.19 Bode plot of a first order factor. The dashed lines show the piece-wise linear approximations of the curves.

and we obtain the following approximations.

$$\log|G(i\omega)| \approx \begin{cases} \log a & \text{if } \omega << a, \\ \log a + \log\sqrt{2} & \text{if } \omega = a, \\ \log\omega & \text{if } \omega >> a \end{cases},$$

$$\arg G(i\omega) \approx \begin{cases} 0 & \text{if } \omega << a, \\ \dfrac{\pi}{4} + \dfrac{1}{2}\log\dfrac{\omega}{a} & \text{if } \omega \approx a, \\ \dfrac{\pi}{2} & \text{if } \omega >> a \end{cases},$$

Notice that a first order system behaves like an integrator for high frequencies. Compare with the Bode plot in Figure 3.18. □

EXAMPLE 3.13—BODE PLOT OF A SECOND ORDER SYSTEM
Consider the transfer function

$$G(s) = s^2 + 2a\zeta s + a^2$$

We have

$$G(i\omega) = a^2 - \omega^2 + 2i\zeta a\omega$$

Hence

$$\log|G(i\omega)| = \frac{1}{2}\log\left(\omega^4 + 2a^2\omega^2(2\zeta^2 - 1) + a^4\right)$$
$$\arg G(i\omega) = \arctan 2\zeta a\omega/(a^2 - \omega^2)$$

Notice that the smallest value of the magnitude $\min_\omega |G(i\omega)| = 1/2\zeta$ is obtained for $\omega = a$ The gain is thus constant for small ω. It has an asymptote with zero slope for low frequencies. For large values of ω the gain is proportional to ω^2, which means that the gain curve has an asymptote with slope 2. The phase is zero for low frequencies and approaches 180° for large frequencies. The curves can be approximated with the following piece-wise linear expressions

$$\log|G(i\omega)| \approx \begin{cases} 2\log a & \text{if } \omega << a, \\ 2\log a + \log 2\zeta & \text{if } \omega = a, \\ 2\log\omega & \text{if } \omega >> a \end{cases},$$

$$\arg G(i\omega) \approx \begin{cases} 0 & \text{if } \omega << a, \\ \dfrac{\pi}{2} + \dfrac{\omega - a}{a\zeta} & \text{if } \omega = a, \\ \pi & \text{if } \omega >> a \end{cases},$$

The Bode Plot is shown in Figure 3.20, the piece-wise linear approximations are shown in dashed lines. □

Sketching a Bode Plot

It is easy to sketch the asymptotes of the gain curves of a Bode plot. This is often done in order to get a quick overview of the frequency response. The following procedure can be used

- Factor the numerator and denominator of the transfer functions.
- The poles and zeros are called break points because they correspond to the points where the asymptotes change direction.
- Determine break points sort them in increasing frequency
- Start with low frequencies
- Draw the low frequency asymptote
- Go over all break points and note the slope changes

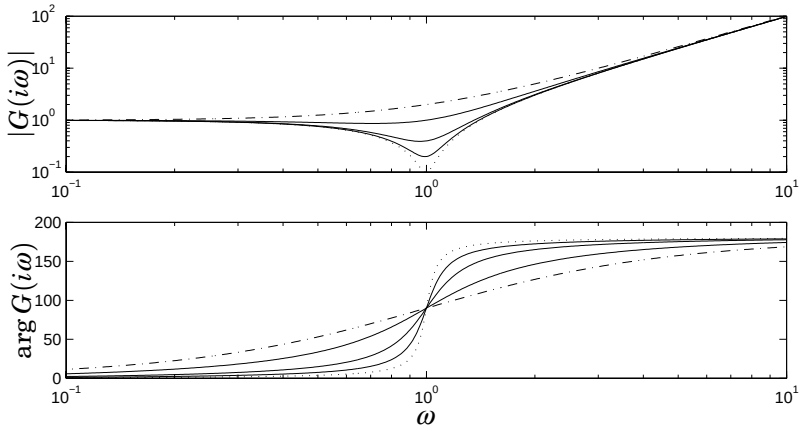


Figure 3.20 Bode plot of a second order factor with $\zeta = 0.05$ (dotted), 0.1, 0.2, 0.5 and 1.0 (dash-dotted). The dashed lines show the piece-wise linear approximations of the curves.

- A crude sketch of the phase curve is obtained by using the relation that, for systems with no RHP poles or zeros, one unit slope corresponds to a phase of 90°

We illustrate the procedure with the transfer function

$$G(s) = \frac{200(s+1)}{s(s+10)(s+200)} = \frac{1+s}{10s(1+0.1s)(1+0.01s)}$$

The break points are 0.01, 0.1, 1. For low frequencies the transfer function can be approximated by

$$G(s) \approx \frac{1}{10s}$$

Following the procedure we get

- The low frequencies the system behaves like an integrator with gain 0.1. The low frequency asymptote thus has slope -1 and it crosses the axis of unit gain at $\omega = 0.1$.
- The first break point occurs at $\omega = 0.01$. This break point corresponds to a pole which means that the slope decreases by one unit to -2 at that frequency.
- The next break point is at $\omega = 0.1$ this is also a pole which means that the slope decreases to -3.

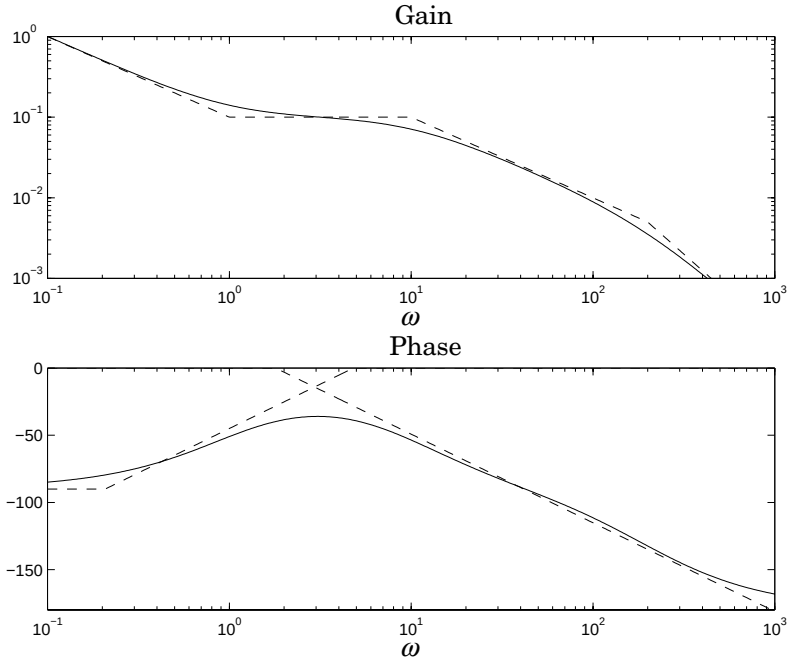


Figure 3.21 Illustrates how the asymptotes of the gain curve of the Bode plot can be sketched. The dashed curves show the asymptotes and the full lines the complete plot.

- The next break point is at $\omega = 1$, since this is a zero the slope increases by one unit to -2.

Figure 3.21 shows the asymptotes of the gain curve and the complete Bode plot.

Gain and Phase Margins

The gain and phase margins can easily be found from the Bode plot of the loop transfer function. Recall that the gain margin tells how much the gain has to be increased for the system to reach instability. To determine the gain margin we first find the frequency ω_{pc} where the phase is $-180°$. This frequency is called the phase crossover frequency. The gain margin is the inverse of the gain at that frequency. The phase margin tells how the phase lag required for the system to reach instability. To determine the phase margin we first determine the frequency ω_{gc} where the gain of the loop transfer function is one. This frequency is called the gain crossover frequency. The phase margin is the phase of the loop transfer function

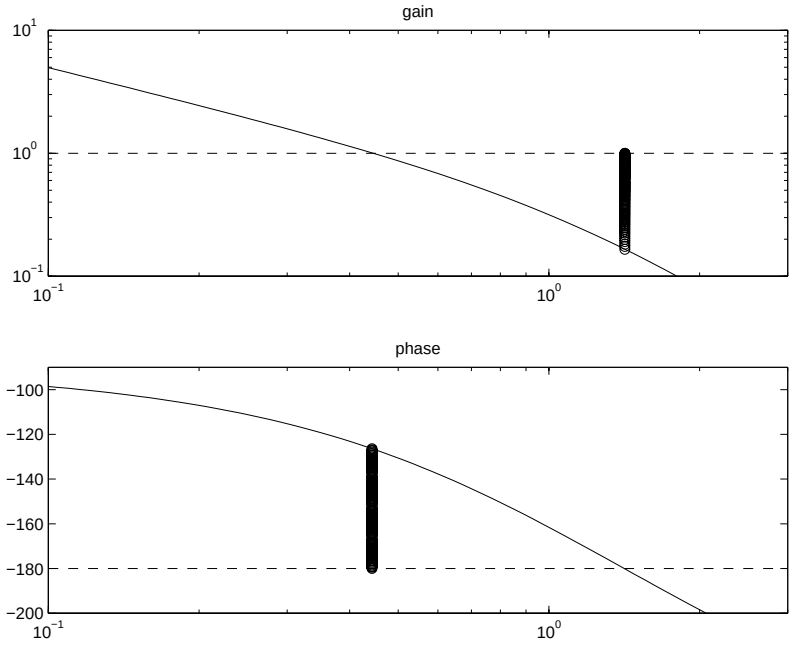


Figure 3.22 Finding gain and phase margins from the Bode plot of the loop transfer function.

at that frequency plus 180°. Figure 3.22 illustrates how the margins are found in the Bode plot of the loop transfer function.

Bode's Relations

Analyzing the Bode plots in the examples we find that there appears to be a relation between the gain curve and the phase curve. Consider e.g. the curves for the differentiator in Figure 3.17 and the integrator in Figure 3.18. For the differentiator the slope is +1 and the phase is constant $pi/2$ radians. For the integrator the slope is -1 and the phase is $-pi/2$. Bode investigated the relations between the curves and found that there was a unique relation between amplitude and phase for many systems. In particular he found the following relations for system with no

poles and zeros in the right half plane.

$$\begin{aligned}
\arg G(i\omega_0) &= \frac{2\omega_0}{\pi}\int_0^\infty \frac{\log|G(i\omega)| - \log|G(i\omega_0)|}{\omega^2-\omega_0^2}d\omega \\
&= \frac{1}{\pi}\int_0^\infty \frac{d\log|G(i\omega)|}{d\log\omega}\log\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right|d \\
&\approx \frac{\pi}{2}\frac{d\log|G(i\omega)|}{d\log\omega} \\
\frac{\log|G(i\omega)|}{\log|G(i\omega_0)|} &= -\frac{2\omega_0^2}{\pi}\int_0^\infty \frac{\omega^{-1}\arg G(i\omega) - \omega_0^{-1}\arg G(i\omega_0)}{\omega^2-\omega_0^2}d\omega \\
&= -\frac{2\omega_0^2}{\pi}\int_0^\infty \frac{d(\omega^{-1}\arg G(i\omega))}{d\omega}\log\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right|d\omega
\end{aligned} \tag{3.31}$$

The formula for the phase tells that the phase is a weighted average of the logarithmic derivative of the gain, approximatively

$$\arg G(i\omega) \approx \frac{\pi}{2}\frac{d\log|G(i\omega)|}{d\log\omega} \tag{3.32}$$

This formula implies that a slope of $+1$ corresponds to a phase of $\pi/2$, which holds exactly for the differentiator, see Figure 3.17. The exact formula (3.31) says that the differentiated slope should be weighted by the kernel

$$\int_0^\infty \log\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right|d\omega = \frac{\pi^2}{2}$$

Figure 3.23 is a plot of the kernel.

Minimum Phase and Non-minimum Phase

Bode's relations hold for systems that do not have poles and zeros in the left half plane. Such systems are called minimum phase systems. One nice property of these systems is that the phase curve is uniquely given by the gain curve. These systems are also relatively easy to control. Other systems have larger phase lag, i.e. more negative phase. These systems are said to be non-minimum phase, because they have more phase lag than the equivalent minimum phase systems. Systems which do not have minimum phase are more difficult to control. Before proceeding we will give some examples.

EXAMPLE 3.14—A TIME DELAY
The transfer function of a time delay of T units is

$$G(s) = e^{-sT}$$

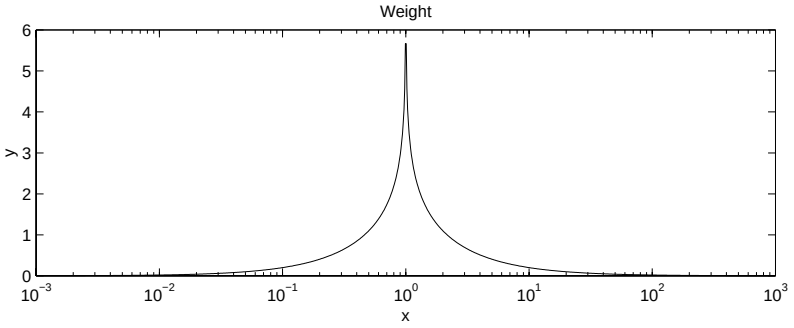


Figure 3.23 The weighting kernel in Bodes formula for computing the phase from the gain.

This transfer function has the property

$$|G(i\omega)| = 1, \quad \arg G(i\omega) = -\omega T$$

Notice that the gain is one. The minimum phase system which has unit gain has the transfer function $G(s) = 1$. The time delay thus has an additional phase lag of ωT. Notice that the phase lag increases with increasing frequency. Figure 3.24 □

It seems intuitively reasonable that it is not possible to obtain a fast response of a system with time delay. We will later show that this is indeed the case. □

Next we will consider a system with a zero in the right half plane

EXAMPLE 3.15—SYSTEM WITH A RHP ZERO
Consider a system with the transfer function

$$G(s) = \frac{a - s}{a + s}$$

This transfer function has the property

$$|G(i\omega)| = 1, \quad \arg G(i\omega) = -2 \arctan \frac{\omega}{a}$$

Notice that the gain is one. The minimum phase system which has unit gain has the transfer function $G(s) = 1$. In Figure 3.25 we show the Bode plot of the transfer function. The Bode plot resembles the Bode plot for a time delay which is not surprising because the exponential function e^{-sT}

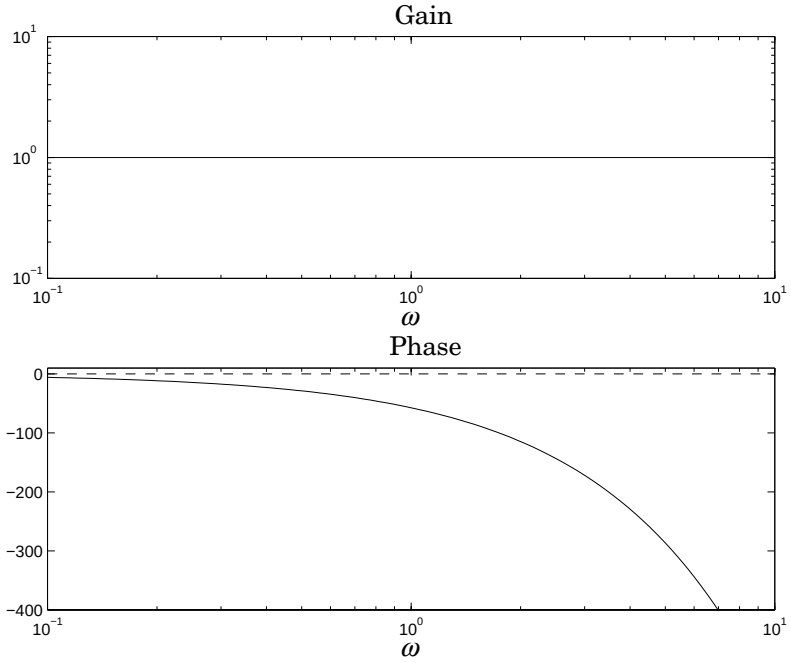


Figure 3.24 Bode plot of a time delay which has the transfer function $G(s) = e^{-s}$.

can be approximated by

$$e^{-sT} = \frac{1 - sT/2}{1 + sT/2}$$

The largest phase lag of a system with a zero in the RHP is however pi. □

We will later show that the presence of a zero in the right half plane severely limits the performance that can be achieved. We can get an intuitive feel for this by considering the step response of a system with a right half plane zero. Consider a system with the transfer function $G(s)$ that has a zero at $s = -\alpha$ in the right half plane. Let h be the step response of the system. The Laplace transform of the step response is given by

$$H(s) = \frac{G(s)}{s} = \int_0^t e^{-st} h(t) dt$$

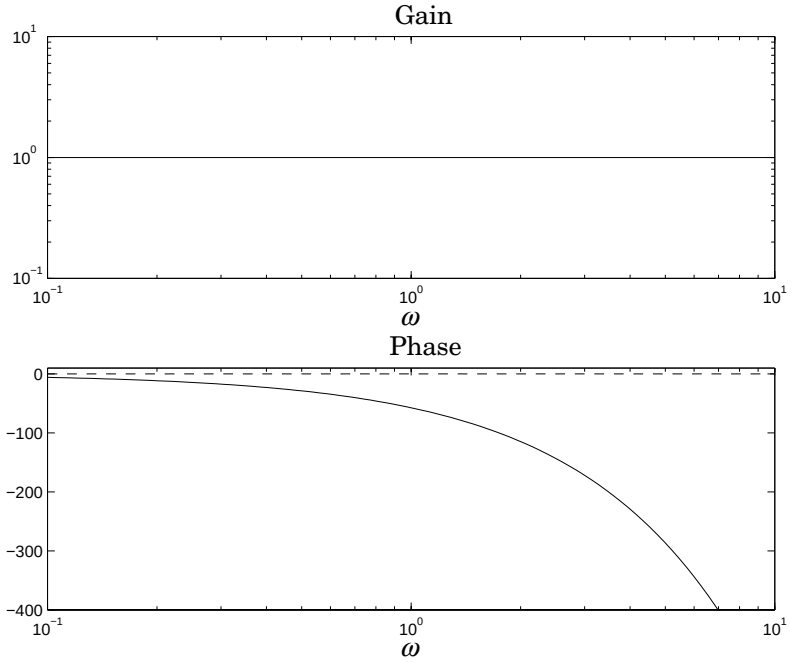


Figure 3.25 Bode plot of a the transfer function $G(s) = \dfrac{a-s}{a+s}$.

Since $G(\alpha)$ is zero we have

$$0 = \int_0^t e^{-\alpha t} h(t) dt$$

Since $e^{-\alpha t}$ is positive it follows that the step response $h(t)$ must be negative for some t. This is illustrated in Figure 3.26 which shows the step response of a system having a zero in the right half plane. Notice that the output goes in the wrong direction initially. This is sometimes referred to as inverse response. It seems intuitively clear that such systems are difficult to control fast. This is indeed the case as will be shown in Chapter 5. We have thus found that systems with time delays and zeros in the right half plane have similar properties. Next we will consider a system with a right half plane pole.

EXAMPLE 3.16—SYSTEM WITH A RHP POLE
Consider a system with the transfer function

$$G(s) = \frac{s+a}{s-a}$$

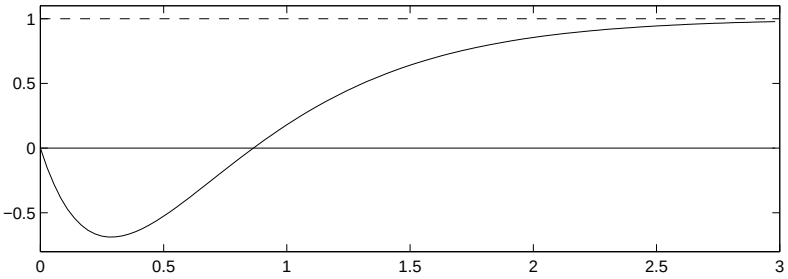


Figure 3.26 Step response of a system with a zero in the right half plane. The system has the transfer function $G(s) = \dfrac{6(-s+1)}{s^2+5s+6}$.

This transfer function has the property

$$|G(i\omega)| = 1, \quad \arg G(i\omega) = -2\arctan\frac{a}{\omega}$$

Notice that the gain is one. The minimum phase system which has unit gain has the transfer function $G(s) = 1$. In Figure 3.27 we show the Bode plot of the transfer function. □

Comparing the Bode plots for systems with a right half plane pole and a right half plane zero we find that the additional phase lag appears at high frequencies for a system with a right half plane zero and at low frequencies for a system with a right half plane pole. This means that there are significant differences between the systems. When there is a right half plane pole high frequencies must be avoided by making the system slow. When there is a right half plane zero low frequencies must be avoided and it is necessary to control these systems rapidly. This will be discussed more in Chapter 5.

It is a severe limitation to have poles and zeros in the right half plane. Dynamics of this type should be avoided by redesign of the system. The zeros of a system can also be changed by moving sensors or by introducing additional sensors. Unfortunately systems which are non-minimum phase are not uncommon i real life. We end this section by giving a few examples.

EXAMPLE 3.17—HYDRO ELECTRIC POWER GENERATION
The transfer function from tube opening to electric power in a hydroelectric power station has the form

$$\frac{P(s)}{A(s)} = \frac{P_0}{A_0}\frac{1-2sT}{1+sT}$$

where T is the time it takes sound to pass along the tube. □

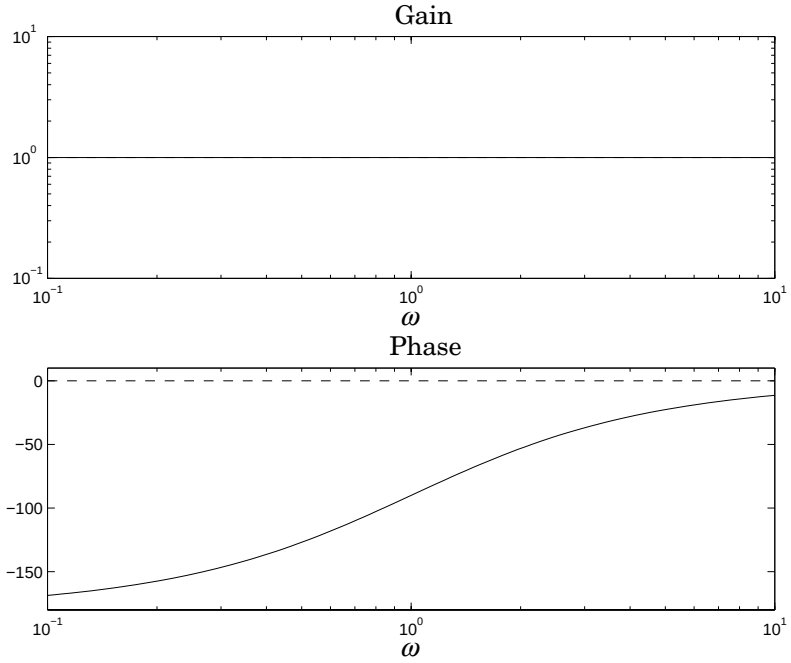


Figure 3.27 Bode plot of a the transfer function $G(s) = \dfrac{s+a}{s-a}$ which has a pole in the right half plane.

EXAMPLE 3.18—LEVEL CONTROL IN STEAM GENERATORS

Consider the problem of controlling the water level in a steam generator. The major disturbance is the variation of steam taken from the unit. When more steam is fed to the turbine the pressure drops. There is typically a mixture of steam and water under the water level. When pressure drops the steam bubbles expand and the level increases momentarily. After some time the level will decrease because of the mass removed from the system. □

EXAMPLE 3.19—FLIGHT CONTROL

The transfer function from elevon to height in an airplane is non-minimum phase. When the elevon is raised there will be a force that pushes the rear of the airplane down. This causes a rotation which gives an increase of the angle of attack and an increase of the lift. Initially the aircraft will however loose height. The Wright brothers understood this and used control surfaces in the front of the aircraft to avoid the effect. □

EXAMPLE 3.20—BACKING A CAR
Consider backing a car close to a curb. The transfer function from steering angle to distance from the curve is non-minimum phase. This is a mechanism that is similar to the aircraft. □

EXAMPLE 3.21—REVENUE FROM DEVELOPMENT
The relation between revenue development effort in a new product development is a non-minimum phase system. This means that such a system is very difficult to control tightly. □

3.6 State Models

The state is a collection of variables that summarize the past of a system for the purpose of prediction the future. For an engineering system the state is composed of the variables required to account for storage of mass, momentum and energy. An key issue in modeling is to decide how accurate storage has to be represented. The state variables are gathered in a vector, the state vector x. The control variables are represented by another vector u and the measured signal by the vector y. A system can then be represented by the model

$$\begin{aligned} \frac{dx}{dt} &= f(x,u) \\ y &= g(x,u) \end{aligned} \tag{3.33}$$

The dimension of the state vector is called the order of the system. The system is called time-invariant because the functions f and g do not depend explicitly on time t. It is possible to have more general time-varying systems where the functions do depend on time. The model thus consists of two functions. The function f gives the velocity of the state vector as a function of state x, control u and time t and the function g gives the measured values as functions of state x, control u and time t. The function f is called the velocity function and the function g is called the sensor function or the measurement function. A system is called linear if the functions f and g are linear in x and u. A linear system can thus be represented by

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

where A, B, C and D are constant varying matrices. Such a system is said to be linear and time-invariant, or LTI for short. The matrix A is

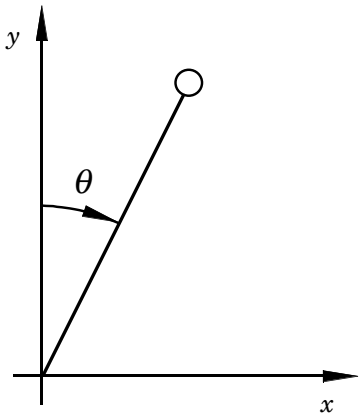


Figure 3.28 An inverted pendulum. *The picture should be mirrored.*

called the dynamics matrix, the matrix B is called the control matrix, the matrix C is called the sensor matrix and the matrix D is called the direct term. Frequently systems will not have a direct term indicating that the control signal does not influence the output directly. We will illustrate by a few examples.

EXAMPLE 3.22—THE DOUBLE INTEGRATOR
Consider a system described by

$$\begin{aligned} \frac{dx}{dt} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 & 1 \end{pmatrix} u \\ y &= \begin{pmatrix} 1 & 0 \end{pmatrix} x \end{aligned} \tag{3.34}$$

This is a linear time-invariant system of second order with no direct term. □

EXAMPLE 3.23—THE INVERTED PENDULUM
Consider the inverted pendulum in Figure 3.28. The state variables are the angle $\theta = x_1$ and the angular velocity $d\theta/dt = x_2$, the control variable is the acceleration ug of the pivot, and the output is the angle θ.

Newtons law of conservation of angular momentum becomes

$$J\frac{d^2\theta}{dt^2} = mgl\sin\theta + mul\cos\theta$$

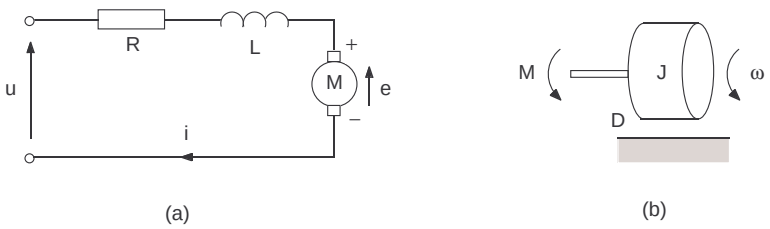


Figure 3.29 Schematic diagram of an electric motor.

Introducing $x_1 = \theta$ and $x_2 = d\theta/dt$ the state equations become

$$\frac{dx}{dt} = \begin{pmatrix} x_2 \\ \frac{mgl}{J}\sin x_1 + \frac{mlu}{J}\cos x_1 \end{pmatrix}$$
$$y = x_1$$

It is convenient to normalize the equation by choosing $\sqrt{J/mgl}$ as the unit of time. The equation then becomes

$$\frac{dx}{dt} = \begin{pmatrix} x_2 \\ \sin x_1 + u\cos x_1 \end{pmatrix} \tag{3.35}$$
$$y = x_1$$

This is a nonlinear time-invariant system of second order. □

EXAMPLE 3.24—AN ELECTRIC MOTOR

A schematic picture of an electric motor is shown in Figure 3.29 Energy stored is stored in the capacitor, and the inductor and momentum is stored in the rotor. Three state variables are needed if we are only interested in motor speed. Storage can be represented by the current I through the rotor, the voltage V across the capacitor and the angular velocity ω of the rotor. The control signal is the voltage E applied to the motor. A momentum balance for the rotor gives

$$J\frac{d\omega}{dt} + D\omega = kI$$

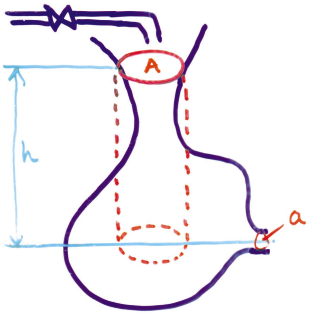


Figure 3.30 A schematic picture of a water tank.

and Kirchoffs laws for the electric circuit gives

$$E = RI + L\frac{dI}{dt} + V - k\frac{d\omega}{dt}$$
$$I = C\frac{dV}{dt}$$

Introducing the state variables $x_1 = \omega$, $x_2 = V$, $x_3 = I$ and the control variable $u = E$ the equations for the motor can be written as

$$\frac{dx}{dt} = \begin{pmatrix} -\frac{D}{J} & 0 & \frac{k}{J} \\ 0 & 0 & \frac{1}{C} \\ -\frac{kD}{JL} & -\frac{1}{L} & \frac{k^2}{JL} - \frac{R}{L} \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ \frac{1}{L} \end{pmatrix} uy = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} x \qquad (3.36)$$

This is a linear time-invariant system with three state variables and one input. □

EXAMPLE 3.25—THE WATER TANK
Consider a tank with water where the input is the inflow and there is free outflow, see Figure 3.30 Assuming that the density is constant a mass balance for the tank gives

$$\frac{dV}{dt} = q_{in} - q_{out}$$

The outflow is given by

$$q_{out} = a\sqrt{2gh}$$

There are several possible choices of state variables. One possibility is to characterize the storage of water by the height of the tank. We have the following relation between height h and volume

$$V = \int_0^h A(x)dx$$

Simplifying the equations we find that the tank can be described by

$$\frac{dh}{dt} = \frac{1}{A(h)}(q_{in} - a\sqrt{2gh})$$
$$q_{out} = a\sqrt{2gh}$$

The tank is thus a nonlinear system of first order. □

Equilibria

To investigate a system we will first determine the equilibria. Consider the system given by (3.33) which is assumed to be time-invariant. Let the control signal be constant $u = u_0$. The equilibria are states x_0 such that the $dx/dt = 0$. Hence

$$f(x_0, u_0) = 0$$

Notice that there may be several equilibria.

For second order systems the state equations can be visualized by plotting the velocities for all points in the state space. This graph is called the phase plane shows the behavior qualitative. The equilibria corresponds to points where the velocity is zero. We illustrate this with an example.

EXAMPLE 3.26—THE PHASE PLANE
Consider the

$$\frac{dx}{dt} = \begin{pmatrix} x_2 - x_2^3 \\ -x_1 - x_2^2 \end{pmatrix}$$

The equilibria are given by

$$x_2 - x_2^3 = 0$$
$$x_1 - x_2^2 = 0$$

There are three equilibria:

$$x_1 = -1 \quad x_2 = -1$$
$$x_1 = -1 \quad x_2 = 1$$
$$x_1 = 0 \quad x_2 = 0$$

The phase plane is shown in Figure 3.31. The phase plane is a good visualization of solutions for second order systems. It also illustrates that nonlinear systems can be interpreted as a vector field or a flow. □

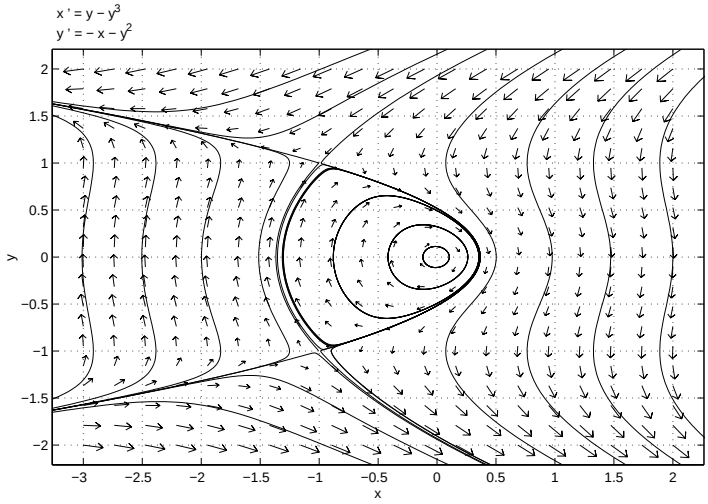


Figure 3.31 Phase plane for the second order system $dx_1/dt = x_2 - x_2^3$,$dx_2/dt = -x_1 - x_2^2$.

Linearization

Nonlinear systems are unfortunately difficult. It is fortunate that many aspects of control can be understood from linear models. This is particularly true for regulation problems where it is intended to keep variables close to specified values. When deviations are small the nonlinearities can be approximated by linear functions. With efficient control the deviations are small and the approximation works even better. In this section we will show how nonlinear dynamics systems are approximated. We will start with an example that shows how static systems are approximated.

EXAMPLE 3.27—LINEARIZATION OF STATIC SYSTEM
Consider the system

$$y = g(u)$$

A Taylor series expansion around $u = u_0$ gives

$$y = g(u_0) + g'(u_0)(u - u_0) + \ldots$$

The linearized model is

$$y - y_0 = g'(u_0)(u - u_0)$$

The linearized model thus replaces the nonlinear curve by its tangent at the operating point. □

Linearization of dynamic systems is done in the same way. We start by determining the appropriate equilibria. The nonlinear systems are then approximated using Taylor series expansions. Consider the system

$$\begin{aligned}\frac{dx}{dt} &= f(x,u)\\ y &= g(x,u)\end{aligned}$$

Consider small deviations from the equilibrium!

$$x = x_0 + \delta x, \quad u = u_0 + \delta u, \quad y = y_0 + \delta y$$

Make a series expansion of the differential equation and neglect terms of second and higher order. This gives

$$\begin{aligned}\frac{dx}{dt} &= f(x_0+\delta x, u_0+\delta u) \approx f(x_0,u_0) + \frac{\partial f}{\partial x}(x_0,u_0)\delta x + \frac{\partial f}{\partial u}(x_0,u_0)\delta u\\ y &= g(x_0+\delta x, u_0+\delta u) \approx y_0 + \frac{\partial g}{\partial x}(x_0,u_0)\delta x + \frac{\partial g}{\partial u}(x_0,u_0)\delta u\end{aligned}$$

We have $f(x_0,u_0) = 0$ because x_0 is an equilibrium and we find the following approximation for small deviations around the equilibrium.

$$\begin{aligned}\frac{d(x-x_0)}{dt} &= A(x-x_0) + B(u-u_0)\\ y - y_0 &= C(x-x_0) + D(u-u_0)\end{aligned}$$

where

$$A = \frac{\partial f}{\partial x}(x_0,u_0) \qquad B = \frac{\partial f}{\partial u}(x_0,u_0)$$
$$C = \frac{\partial g}{\partial x}(x_0,u_0) \qquad D = \frac{\partial g}{\partial u}(x_0,u_0)$$

The linearized equation is thus a linear time-invariant system, compare with (3.37). It is common practice to relabel variables and simply let x, y and u denote deviations from the equilibrium.

We illustrate with a few examples

EXAMPLE 3.28—LINEARIZATION OF THE WATER TANK

$$\begin{aligned}\frac{dh}{dt} &= \frac{1}{A(h)}(q_{in} - a\sqrt{2gh})\\ q_{out} &= a\sqrt{2gh}\end{aligned}$$

To determine the equilibrium we assume that the inflow is constant $qin = q_0$. It follows that

$$q_{out} = q_{in} = q_0 = a\sqrt{2gh_0}$$
$$h_0 = \frac{q_0^2}{2ga^2}$$

Let A_0 be the cross section A at level h_0, introduce the deviations. The linearized equations are

$$\frac{d\delta h}{dt} = -\frac{a\sqrt{2gh_0}}{2A_0h_0}\delta h + \frac{1}{A_0}\delta q_{in}$$
$$\delta q_{out} = \frac{a\sqrt{2gh_0}}{h_0}\delta h = \frac{q_0}{h_0}\delta h$$

The parameter

$$T = \frac{2A_0h_0}{q_0} = 2 \times \frac{\text{Total water volume } [m^3]}{\text{Flow rate } [\text{m}^3/\text{s}]}$$

is called the time constant of the system. Notice that $T/2$ is the time it takes to fill the volume A_0h_0 with the steady state flow rate q_0 □

EXAMPLE 3.29—LINEARIZATION OF THE INVERTED PENDULUM
Consider the inverted pendulum in Example 3.23 which is described by (3.35). If the control signal is zero the equilibria are given by

$$x_2 = 0$$
$$\sin x_1 = 0$$

i.e. $x_2 = \theta/dt$ and $x_1 = \theta = 0$ and $x_1 = \theta = \pi$. The first equilibrium corresponds to the pendulum standing upright and the second to the pendulum hanging straight down. We have

$$\frac{\partial f(x,0)}{\partial x} = \begin{pmatrix} 0 & 1 \\ \cos x_1 - u\sin x_1 & 0 \end{pmatrix}, \quad \frac{\partial f}{\partial u} = \begin{pmatrix} 0 \\ \cos x_1 \end{pmatrix},$$

Evaluating the derivatives at the upper equilibrium $u = 0$, $x_1 = 0$ and $x_2 = 0$ we get

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \end{pmatrix}.$$

For the equilibrium when then pendulum is hanging down, $u = 0$, $x_1 = \pi$ and $x_2 = 0$ we have instead

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -1 \end{pmatrix}.$$

□

3.7 Linear Time-Invariant Systems

The model

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \tag{3.37}$$

is one of the standard models in control. In this section we will present an in depth treatment. Let us first recall that x is the state vector, u the control, y the measurement. The model is nice because it can represent systems with many inputs and many outputs in a very compact form. Because of the advances in numeric linear algebra there are also much powerful software for making computations. Before going into details we will present some useful results about matrix functions. It is assumed that the reader is familiar with the basic properties of matrices.

Matrix Functions

Some basic facts about matrix functions are summarized in this section. Let A be a square matrix, since it is possible to compute powers of matrices we can define a matrix polynomial as follows

$$f(A) = a_0 I + a_1 A + \ldots + a_n A^n$$

Similarly if the function $f(x)$ has a converging series expansion we can also define the following matrix function

$$f(A) = a_0 I + a_1 A + \ldots + a_n A^n + \ldots$$

The matrix exponential is a nice useful example which can be defined as

$$e^{At} = I + At + \frac{1}{2}(At)^2 + \ldots + \frac{1}{n!}A^n t^n + \ldots$$

Differentiating this expression we find

$$\begin{aligned} \frac{de^{At}}{dt} &= A + A^2 t + \frac{1}{2}A^3 t^2 + \ldots + \frac{1}{(n-1)!}A^n t^{n-1} + \ldots \\ &= A(= I + At + \frac{1}{2}(At)^2 + \ldots + \frac{1}{n!}A^n t^n + \ldots) = Ae^{At} \end{aligned}$$

The matrix exponential thus has the property

$$\frac{de^{At}}{dt} = Ae^{At} = e^{At}A \tag{3.38}$$

Matrix functions do however have other interesting properties. One result is the following.

THEOREM 3.3—CAYLEY-HAMILTON
Let the $n \times n$ matrix A have the characteristic equation

$$\det(\lambda I - A) = \lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} \ldots + a_n = 0$$

then it follows that

$$\det(\lambda I - A) = A^n + a_1A^{n-1} + a_2A^{n-2} \ldots + a_nI = 0$$

A matrix satisfies its characteristic equation.

PROOF 3.2
If a matrix has distinct eigenvalues it can be diagonalized and we have $A = T^{-1}\Lambda T$. This implies that

$$A^2 = T^{-1}\Lambda T T^{-1}\Lambda T = T^{-1}\Lambda^2 T$$
$$A^3 = T^{-1}\Lambda T A^2 = T^{-1}\Lambda T T^{-1}\Lambda^2 T = T^{-1}\Lambda^3 T$$

and that $A^n = T^{-1}\Lambda^n T$. Since λ_i is an eigenvalue it follows that

$$\lambda_i^n + a_1\lambda_i^{n-1} + a_2\lambda_i^{n-2} \ldots + a_n = 0$$

Hence

$$\Lambda_i^n + a_1\Lambda_i^{n-1} + a_2\Lambda_i^{n-2} \ldots + a_nI = 0$$

Multiplying by T^{-1} from the left and T from the right and using the relation $A^k = T^{-1}\Lambda^k T$ now gives

$$A^n + a_1A^{n-1} + a_2A^{n-2} \ldots + a_nI = 0$$

□

The result can actually be sharpened. The minimal polynomial of a matrix is the polynomial of lowest degree such that $g(A) = 0$. The characteristic polynomial is generically the minimal polynomial. For matrices with common eigenvalues the minimal polynomial may, however, be different from the characteristic polynomial. The matrices

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

have the minimal polynomials

$$g_1(\lambda) = \lambda - 1, \qquad g_2(\lambda) = (\lambda - 1)^2$$

A matrix function can thus be written as

$$f(A) = c_0I + c_1A + \ldots + c_{k-1}A^{k-1}$$

where k is the degree of the minimal polynomial.

Solving the Equations

Using the matrix exponential the solution to (3.37) can be written as

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau \tag{3.39}$$

To prove this we differentiate both sides and use the property 3.38) of the matrix exponential. This gives

$$\frac{dx}{dt} = Ae^{At}x(0) + \int_0^t Ae^{A(t-\tau)}Bu(\tau)d\tau + Bu(t) = Ax + Bu$$

which prove the result. Notice that the calculation is essentially the same as for proving the result for a first order equation.

Input-Output Relations

It follows from Equations (3.37) and (3.39) that the input output relation is given by

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

Taking the Laplace transform of (3.37) under the assumption that $x(0) = 0$ gives

$$\begin{aligned} sX(s) &= AX(s) + BU(s) \\ Y(s) &= CX(s) + DU(s) \end{aligned}$$

Solving the first equation for $X(s)$ and inserting in the second gives

$$\begin{aligned} X(s) &= [sI - A]^{-1}BU(s) \\ Y(s) &= \big(C[sI - A]^{-1}B + D\big)U(s) \end{aligned}$$

The transfer function is thus

$$G(s) = C[sI - A]^{-1}B + D \tag{3.40}$$

we illustrate this with an example.

EXAMPLE 3.30—TRANSFER FUNCTION OF INVERTED PENDULUM
The linearized model of the pendulum in the upright position is characterized by the matrices

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad D = 0.$$

The characteristic polynomial of the dynamics matrix A is

$$\det(sI - A) = \det \begin{pmatrix} s & -1 \\ -1 & s \end{pmatrix} = s^2 - 1$$

Hence

$$(sI - A)^{-1} = \frac{1}{s^2 - 1} \det \begin{pmatrix} s & 1 \\ 1 & s \end{pmatrix}$$

The transfer function is thus

$$G(s) = C[sI - A]^{-1}B = \frac{1}{s^2 - 1} \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} s & 1 \\ 1 & s \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{s^2 - 1}$$

□

Transfer function and impulse response remain invariant with coordinate transformations.

$$\tilde{g}(t) = \tilde{C}e^{\tilde{A}t}\tilde{B} = CT^{-1}e^{TAT^{-1}t}TB = Ce^{At}B = g(t)$$

and

$$\begin{aligned} \tilde{G}(s) &= \tilde{C}(sI - \tilde{A})^{-1}\tilde{B} = CT^{-1}(sI - TAT^{-1})^{-1}TB \\ XS &= C(sI - A)^{-1}B = G(s) \end{aligned}$$

Consider the system

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx \end{aligned}$$

To find the input output relation we can differentiate the output and we

obtain

$$\begin{aligned} y &= Cx \\ \frac{dy}{dt} &= C\frac{dx}{dt} = CAx + CBu \\ \frac{d^2y}{dt^2} &= CA\frac{dx}{dt} + CB\frac{du}{dt} = CA^2x + CABu + CB\frac{du}{dt} \\ &\vdots \\ \frac{d^ny}{dt^n} &= CA^nx + CA^{n-1}Bu + CA^{n-2}B\frac{du}{dt} + \ldots + CB\frac{d^{n-1}u}{dt^{n-1}} \end{aligned}$$

Let a_k be the coefficients of the characteristic equation. Multiplying the first equation by a_n, the second by a_{n-1} etc we find that the input-output relation can be written as.

$$\frac{d^ny}{dt^n} + a_1\frac{d^{n-1}y}{dt^{n-1}} + \ldots + a_ny = B_1\frac{d^{n-1}u}{dt^{n-1}} + B_2\frac{d^{n-2}u}{dt^{n-2}} + \ldots + B_nu,$$

where the matrices B_k are given by.

$$\begin{aligned} B_1 &= CB \\ B_2 &= CAB + a_1CB \\ B_3 &= CA^2B + a_1CAB + a_2CB \\ &\vdots \\ B_n &= CA^{n-1}B + a_1CA^{n-1}B + \ldots + a_{n-1}CB \end{aligned}$$

Coordinate Changes

The components of the input vector u and the output vector y are unique physical signals, but the state variables depend on the coordinate system chosen to represent the state. The elements of the matrices A, B and C also depend on the coordinate system. The consequences of changing coordinate system will now be investigated. Introduce new coordinates z by the transformation $z = Tx$, where T is a regular matrix. It follows from (3.37) that

$$\begin{aligned} \frac{dz}{dt} &= T(Ax + Bu) = TAT^{-1}z + TBu = \tilde{A}z + \tilde{B}u \\ y &= Cx + DU = CT^{-1}z + Du = \tilde{C}z + Du \end{aligned}$$

The transformed system has the same form as (3.37) but the matrices A, B and C are different

$$\tilde{A} = TAT^{-1}, \qquad \tilde{B} = TB, \qquad \tilde{C} = CT^{-1}, \qquad \tilde{D} = D \tag{3.41}$$

It is interesting to investigate if there are special coordinate systems that gives systems of special structure.

The Diagonal Form Some matrices can be transformed to diagonal form, one broad class is matrices with distinct eigenvalues. For such matrices it is possible to find a matrix T such that the matrix TAT^{-1} is a diagonal i.e.

$$TAT^{-1} = \Lambda == \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}$$

The transformed system then becomes

$$\begin{aligned} \frac{dz}{dt} &= \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} z + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} u \\ y &= \begin{pmatrix} \gamma_1 & \gamma_2 & \dots & \gamma_n \end{pmatrix} z + Du \end{aligned} \tag{3.42}$$

The transfer function of the system is

$$G(s) = \sum_{i=1}^{n} \frac{\beta_i \gamma_i}{s - \lambda_i} + D$$

Notice appearance of eigenvalues of matrix A in the denominator.

Reachable Canonical Form Consider a system described by the n-th order differential equation

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + \ldots + a_n y = b_1 \frac{d^{n-1} u}{dt^{n-1}} + \ldots + b_n u$$

To find a representation in terms of state model we first take Laplace transforms

$$Y(s) = \frac{b_1 s^{n-1} + \ldots + b_1 s + b_n}{s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n} U(s) = \frac{b_1 s^{n-1} + \ldots + b_1 s + b_n}{A(s)} U(s)$$

Introduce the state variables

$$\begin{aligned} X_1(s) &= \frac{s^{n-1}}{A(s)}U(s) \\ X_2(s) &= \frac{s^{n-2}}{A(s)}U(s) = \frac{1}{s}X_1(s) \\ X_3(s) &= \frac{s^{n-2}}{A(s)}U(s) = \frac{1}{s^2}X_1(s) = \frac{1}{s}X_2(s) \\ &\vdots \\ X_n(s) &= \frac{1}{A(s)}U(s) = \frac{1}{s^{n-1}}X_1(s) = \frac{1}{s}X_{n-1}(s) \end{aligned} \tag{3.43}$$

Hence

$$\begin{aligned} (s^n + a_1 s^{n-1} + \ldots + a_{n-1}s + a_n)X_1(s) &= s^{n-1}U(s) \\ sX_1(s) + a_1X_1(s) + a_2\frac{1}{s}X_1(s) + \ldots + a_n\frac{1}{s^{n-1}X_1(s)} &= U(s) \\ sX_1(s) + a_1X_2(s) + a_2X_2(s) + \ldots + a_nX_n(s) &= U(s) \end{aligned}$$

Consider the equation for $X_1(s)$, dividing by s^{n-1} we get

$$sX_1(s) + a_1X_2(s) + a_2X_2(s) + \ldots + a_nX_n(s) = U(s)$$

Conversion to time domain gives

$$\frac{dx_1}{dt} = -a_1x_1 - a_2x_2 - \ldots - a_nx_n + u$$

(3.43) also implies that

$$\begin{aligned} X_2(s) &= \frac{1}{s}X_1(s) \\ X_3(s) &= \frac{1}{s}X_2(s) \\ &\vdots \\ X_n(s) &= \frac{1}{s}X_{n-1} \end{aligned}$$

Transforming back to the time domain gives

$$\begin{aligned}\frac{dx_2}{dt} &= x_1\\ \frac{dx_3}{dt} &= x_2\\ &\vdots\\ \frac{dx_n}{dt} &= x_{n-1}\end{aligned}$$

With the chosen state variables the output is given by

$$Y(s) = b_1X_1(s) + b_2X_2(s) + \ldots + b_nX_n(s)$$

Collecting the parts we find that the equation can be written as

$$\begin{aligned}\frac{dz}{dt} &= \begin{pmatrix} -a_1 & -a_2 & \ldots & a_{n-1} & -a_n \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & 1 & 0 \end{pmatrix} z + \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} u \\ y &= \begin{pmatrix} b_1 & b_2 & \ldots & b_{n-1} & b_n \end{pmatrix} z + Du\end{aligned} \tag{3.44}$$

The system has the characteristic polynomial

$$D_n(s) = \det \begin{pmatrix} s+a_1 & a_2 & \ldots & a_{n-1} & a_n \\ -1 & s & & 0 & 0 \\ 0 & -1 & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & -1 & s \end{pmatrix}$$

Expanding the determinant by the last row we find that the following recursive equation for the polynomial $D_n(s)$.

$$D_n(s) = sD_{n-1}(s) + a_n$$

It follows from this equation that

$$D_n(s) = s^n + a_1s^{n-1} + \ldots + a_{n-1}s + a_n$$

Transfer function

$$G(s) = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n} + D$$

The numerator of the transfer function $G(s)$ is the characteristic polynomial of the matrix A. This form is called the reachable canonical for for reasons that will be explained later in this Section.

Observable Canonical Form The reachable canonical form is not the only way to represent the transfer function

$$G(s) = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n}$$

another representation is obtained by the following recursive procedure. Introduce the Laplace transform X_1 of first state variable as

$$X_1 = Y = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n} U$$

then

$$(s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n) X_1 = (b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n) U$$

Dividing by s^{n-1} and rearranging the terms we get

$$sX_1 = -a_1 X_1 + b_1 U + X_2$$

where

$$\begin{aligned} s^{n-1} X_2 = &-(a_2 s^{n-2} + a_3 s^{n-3} + \ldots + a_n) X_1 \\ &+ (b_2 s^{n-2} + b_3 s^{n-3} + \ldots + b_n) U \end{aligned}$$

Dividing by s^{n-2} we get

$$sX_2 = -a_2 X_2 + b_2 U + X_3$$

where

$$s^{n-2} X_3 = -(a_3 s^{n-3} + a_4^{n-4} \ldots + a_n) X_1 + (b_3 s^{n-3} + \ldots + b_n) U$$

Dividing by s^{n-3} gives

$$sX_3 = -a_3 X_1 ? b_3 U + X_4$$

Proceeding in this we we finally obtain

$$X_n = -a_n X_1 + b_1 U$$

Collecting the different parts and converting to the time domain we find that the system can be written as

$$\begin{aligned} \frac{dz}{dt} &= \begin{pmatrix} -a_1 & 1 & 0 & \dots & 0 \\ -a_2 & 0 & 1 & & 0 \\ \vdots & & & & \\ -a_{n-1} & 0 & 0 & & 1 \\ -a_n & 0 & 0 & & 0 \end{pmatrix} z + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix} u \\ y &= \begin{pmatrix} 1 & 0 & 0 \dots & 0 \end{pmatrix} z + Du \end{aligned} \tag{3.45}$$

Transfer function

$$G(s) = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n} + D$$

The numerator of the transfer function $G(s)$ is the characteristic polynomial of the matrix A.

Consider a system described by the n-th order differential equation

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + \ldots + a_n y = b_1 \frac{d^{n-1} u}{dt^{n-1}} + \ldots + b_n u$$

Reachability

We will now disregard the measurements and focus on the evolution of the state which is given by

$$\frac{sx}{dt} = Ax + Bu$$

where the system is assumed to be or order n. A fundamental question is if it is possible to find control signals so that any point in the state space can be reached. For simplicity we assume that the initial state of the system is zero, the state of the system is then given by

$$x(t) = \int_0^t e^{A(t-\tau)} B u(\tau) d\tau = \int_0^t e^{A(\tau)} B u(t-\tau) d\tau$$

It follows from the theory of matrix functions that

$$e^{A\tau} = I\alpha_0(s) + A\alpha_1(s) + \ldots + A^{n-1}\alpha_{n-1}(s)$$

and we find that

$$x(t) = B\int_0^t \alpha_0(\tau)u(t-\tau)d\tau + AB\int_0^t \alpha_1(\tau)u(t-\tau)d\tau + \\ \ldots + A^{n-1}B\int_0^t \alpha_{n-1}(\tau)u(t-\tau)d\tau$$

The right hand is thus composed of a linear combination of the columns of the matrix.

$$W_r = \begin{pmatrix} B & AB & \ldots & A^{n-1}B \end{pmatrix}$$

To reach all points in the state space it must thus be required that there are n linear independent columns of the matrix W_c. The matrix is therefor called the reachability matrix. We illustrate by an example.

EXAMPLE 3.31—REACHABILITY OF THE INVERTED PENDULUM
The linearized model of the inverted pendulum is derived in Example 3.29. The dynamics matrix and the control matrix are

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

The reachability matrix is

$$W_r = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.46}$$

This matrix has full rank and we can conclude that the system is reachable. □

Next we will consider a the system in (3.44), i.e

$$\frac{dz}{dt} = \begin{pmatrix} -a_1 & -a_2 & \ldots & a_{n-1} & -a_n \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & 1 & 0 \end{pmatrix} z + \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} u = \tilde{A}z + \tilde{B}u$$

The inverse of the reachability matrix is

$$\tilde{W}_r^{-1} = \begin{pmatrix} 1 & a_1 & a_2 & \ldots & a_n \\ 0 & 1 & a_1 1 & \ldots & a_{n-1} \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \tag{3.47}$$

To show this we consider the product

$$\begin{pmatrix} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{n-1}B \end{pmatrix} W_r^{-1} = \begin{pmatrix} w_0 & w_1 & \cdots & w_{n-1} \end{pmatrix}$$

where

$$\begin{aligned} w_0 &= \tilde{B} \\ w_1 &= a_1\tilde{B} + \tilde{A}\tilde{B} \\ &\vdots \\ w_{n-1} &= a_{n-1}B + a_{n-2}\tilde{A}B + \cdots + \tilde{A}^{n-1}B \end{aligned}$$

The vectors w_k satisfy the relation

$$w_k = a_k + \tilde{w}_{k-1}$$

Iterating this relation we find that

$$\begin{pmatrix} w_0 & w_1 & \cdots & w_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

which shows that the matrix (3.47) is indeed the inverse of $\tilde{W}_r$.

Systems That are Not Reachable It is useful of have an intuitive understanding of the mechanisms that make a system unreachable. An example of such a system is given in Figure 3.32. The system consists of two identical systems with the same input. The intuition can also be demonstrated analytically. We demonstrate this by a simple example.

EXAMPLE 3.32—NON-REACHABLE SYSTEM
Assume that the systems in Figure 3.32 are of first order. The complete system is then described by

$$\begin{aligned} \frac{dx_1}{dt} &= -x_1 + u \\ \frac{dx_2}{dt} &= -x_2 + u \end{aligned}$$

The reachability matrix is

$$W_r = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$$

This matrix is singular and the system is not reachable. □

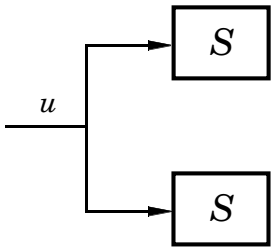


Figure 3.32 A non-reachable system.

Coordinate Changes

It is interesting to investigate how the reachability matrix transforms when the coordinates are changed. Consider the system in (3.37). Assume that the coordinates are changed to $z = Tx$. It follows from (3.41) that the dynamics matrix and the control matrix for the transformed system are

$$\begin{aligned}\tilde{A} &= TAT^{-1}\\ \tilde{B} &= TB\end{aligned}$$

The reachability matrix for the transformed system then becomes

$$\tilde{W}_r = \begin{pmatrix} \tilde{B} & \tilde{A}\tilde{B} & \dots & \tilde{A}^{n-1}\tilde{B} \end{pmatrix} =$$

We have

$$\begin{aligned}\tilde{A}\tilde{B} &= TAT^{-1}TB = TAB\\ \tilde{A}^2\tilde{B} &= (TAT^{-1})^2TB = TAT^{-1}TAT^{-1}TB = TA^2B\\ &\vdots\\ \tilde{A}^n\tilde{B} &= TA^nB\end{aligned}$$

and we find that the reachability matrix for the transformed system has the property

$$\tilde{W}_r = \begin{pmatrix} \tilde{B} & \tilde{A}\tilde{B} & \dots & \tilde{A}^{n-1}\tilde{B} \end{pmatrix} = T\begin{pmatrix} B & AB & \dots & A^{n-1}B \end{pmatrix} = TW_r$$

This formula is very useful for finding the transformation matrix T.

Observability

When discussing reachability we neglected the output and focused on the state. We will now discuss a related problem where we will neglect the input and instead focus on the output. Consider the system

$$\begin{aligned} \frac{dx}{dt} &= Ax \\ y &= Cx \end{aligned} \tag{3.48}$$

We will now investigate if it is possible to determine the state from observations of the output. This is clearly a problem of significant practical interest, because it will tell if the sensors are sufficient.

The output itself gives the projection of the state on vectors that are rows of the matrix C. The problem can clearly be solved if the matrix C is invertible. If the matrix is not invertible we can take derivatives of the output to obtain.

$$\frac{dy}{dt} = C\frac{sc}{dt} = CAx$$

From then derivative of the output we thus get the projections of the state on vectors which are rows of the matrix CA. Proceeding in this way we get

$$\begin{pmatrix} y \\ \frac{dy}{dt} \\ \frac{d^2y}{dt^2} \\ \vdots \\ \frac{d^{n-1}y}{dt^{n-1}} \end{pmatrix} = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{pmatrix} x$$

We thus find that the state can be determined if the matrix

$$W_o = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{pmatrix} \tag{3.49}$$

has n independent rows. Notice that because of the Cayley-Hamilton equation it is not worth while to continue and take derivatives higher than d^{n-1}/dt^{n-1}. The matrix W_o is called the observability matrix. A system is called observable if the observability matrix has full rank. We illustrate with an example.

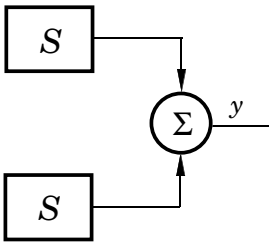


Figure 3.33 A non-observable system.

EXAMPLE 3.33—OBSERVABILITY OF THE INVERTED PENDULUM
The linearized model of inverted pendulum around the upright position is described by (3.41). The matrices A and C are

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

The observability matrix is

$$W_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which has full rank. It is thus possible to compute the state from a measurement of the angle.

A Non-observable System

It is useful to have an understanding of the mechanisms that make a system unobservable. Such a system is shown in Figure 3.33. Next we will consider the system in (3.45) on observable canonical form, i.e.

$$\frac{dz}{dt} = \begin{pmatrix} -a_1 & 1 & 0 & \dots & 0 \\ -a_2 & 0 & 1 & & 0 \\ \vdots & & & & \\ -a_{n-1} & 0 & 0 & & 1 \\ -a_n & 0 & 0 & & 0 \end{pmatrix} z + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix} u$$
$$y = \begin{pmatrix} 1 & 0 & 0 \dots & 0 \end{pmatrix} z + Du$$

A straight forward but tedious calculation shows that the inverse of the

observability matrix has a simple form. It is given by

$$W_o^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ a_1 & 1 & 0 & \dots & 0 \\ a_2 & a_1 & 1 & \dots & 0 \\ \vdots & & & & \\ a_{n-1} & a_{n-2} & a_{n-3} & \dots & 1 \end{pmatrix}$$

This matrix is always invertible. The system is composed of two identical systems whose outputs are added. It seems intuitively clear that it is not possible to deduce the states from the output. This can also be seen formally.

Coordinate Changes

It is interesting to investigate how the observability matrix transforms when the coordinates are changed. Consider the system in (3.37). Assume that the coordinates are changed to $z = Tx$. It follows from (3.41) that the dynamics matrix and the output matrix are given by

$$\begin{aligned} \tilde{A} &= TAT^{-1} \\ \tilde{C} &= CT^{-1} \end{aligned}$$

The observability matrix for the transformed system then becomes

$$\tilde{W}_o = \begin{pmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \tilde{C}\tilde{A}^2 \\ \vdots \\ \tilde{C}\tilde{A}^{n-1} \end{pmatrix}$$

We have

$$\begin{aligned} \tilde{C}\tilde{A} &= CT^{-1}TAT^{-1} = CAT^{-1} \\ \tilde{C}\tilde{A}^2 &= CT^{-1}(TAT^{-1})^2 = CT^{-1}TAT^{-1}TAT^{-1} = CA^2T^{-1} \\ &\vdots \\ \tilde{C}\tilde{A}^n &= CA^nT^{-1} \end{aligned}$$

and we find that the observability matrix for the transformed system has the property

$$\tilde{W}_o = \begin{pmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \tilde{C}\tilde{A}^2 \\ \vdots \\ \tilde{C}\tilde{A}^{n-1} \end{pmatrix} T^{-1} = W_o T^{-1}$$

This formula is very useful for finding the transformation matrix T.

Kalman's Decomposition

The concepts of reachability and observability make it possible understand the structure of a linear system. We first observe that the reachable states form a linear subspace spanned by the columns of the reachability matrix. By introducing coordinates that span that space the equations for a linear system can be written as

$$\frac{d}{dt}\begin{pmatrix} x_c \\ x_{\bar{c}} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} x_c \\ x_{\bar{c}} \end{pmatrix} + \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u$$

where the states x_c are reachable and $x_{\bar{c}}$ are non-reachable. Similarly we find that the non-observable or quiet states are the null space of the observability matrix. We can thus introduce coordinates so that the system can be written as

$$\frac{d}{dt}\begin{pmatrix} x_o \\ x_{\bar{o}} \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} x_o \\ x_{\bar{0}} \end{pmatrix}$$
$$y = \begin{pmatrix} C_1 & 0 \end{pmatrix}\begin{pmatrix} x_o \\ x_{\bar{o}} \end{pmatrix}$$

where the states x_o are observable and $x_{\bar{o}}$ not observable (quiet) Combining the representations we find that a linear system can be transformed to the form

$$\frac{dx}{dt} = \begin{pmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & A_{43} & A_{44} \end{pmatrix} x + \begin{pmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{pmatrix} u$$
$$y = \begin{pmatrix} C_1 & 0 & C_2 & 0 \end{pmatrix} x$$

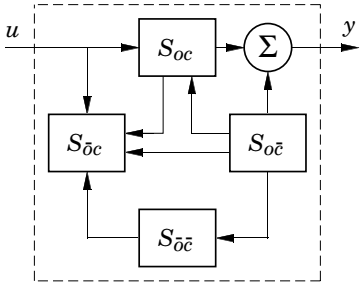


Figure 3.34 Kalman's decomposition of a system.

where the state vector has been partitioned as

$$x = \begin{pmatrix} x_{ro} \\ x_{r\bar{o}} \\ x_{\bar{r}o} \\ x_{\bar{r}\bar{o}} \end{pmatrix}^T$$

A linear system can thus be decomposed into four subsystems.

- S_{ro} reachable and observable
- $S_{r\bar{o}}$ reachable not observable
- $S_{\bar{r}o}$ not reachable observable
- $S_{\bar{r}\bar{o}}$ not reachable not observable

This decomposition is illustrated in Figure 3.34. By tracing the arrows in the diagram we find that the input influences the systems S_{oc} and $S_{\bar{o}c}$ and that the output is influenced by S_{oc} and $S_{o\bar{c}}$. The system $S_{\bar{o}\bar{c}}$ is neither connected to the input nor the output.

The transfer function of the system is

$$G(s) = C_1(sI - A_{11})^{-1}B_1 \tag{3.50}$$

It is thus uniquely given by the subsystem S_{ro}.

The Cancellation Problem Kalman's decomposition resolves one of the longstanding problems in control namely the problem of cancellation of poles and zeros. To illustrate the problem we will consider a system described by the equation.

EXAMPLE 3.34—CANCELLATION OF POLES AND ZEROS

$$\frac{dy}{dt} - y = \frac{du}{dt} - u \tag{3.51}$$

Integrating this system we find that

$$y(t) = u(t) + ce^t$$

where c is a constant. The transfer function of the system is

$$\frac{Y(s)}{U(s)} = \frac{s-1}{s-1} = 1$$

Since s is a complex variable the cancellation is clearly permissible and we find that the transfer function is $G(s) = 1$ and we have seemingly obtained a contradiction because the system is not equivalent to the system

$$y(t) = u(t)$$

The problem is easily resolved by using the Kalman representation. In this particular case the system has two subsystems S_{ro} and $S_{\bar{r}o}$. The system S_{ro} is a static system with transfer function $G(s) = 1$ and the subsystem $S_{\bar{r}o}$ which is observable but non reachable has the dynamics.

$$\frac{dx}{dt} = x$$

□

Notice that cancellations typically appear when using Laplace transforms because of the assumption that all initial values are zero. The consequences are particularly serious when factors like $s - 1$ are cancelled because they correspond to exponentially growing signals. In the early development of control cancellations were avoided by ad hoc rules forbidding cancellation of factors with zeros in the right half plane. Kalman's decomposition gives a very clear picture of what happens when poles and zeros are cancelled.

3.8 Summary

This chapter has summarized some properties of dynamical systems that are useful for control. Both input-output descriptions and state descriptions are given. Much of the terminology that is useful for control has also been introduced.

4

Simple Control Systems

4.1 Introduction

In this chapter we will give simple examples of analysis and design of control systems. We will start in Sections 4.2 and 4.3 with two systems that can be handled using only knowledge of differential equations. Section 4.2 deals with design of a cruise controller for a car. In Section 4.3 we discuss the dynamics of a bicycle, many of its nice properties are due to a purely mechanical feedback which has emerged as a result of trial and error over a long period of time. Section 3.3 is a suitable preparation for Sections 4.2 and 4.3. Differential equations are cumbersome for more complicated problems and better tools are needed. Efficient methods for working with linear systems can be developed based on a basic knowledge of Laplace transforms and transfer functions. Coupled with block diagrams this gives a very efficient way to deal with linear systems. The block diagram gives the overview and the behavior of the individual blocks are described by transfer functions. The Laplace transforms make it easy to manipulate the system formally and to derive relations between different signals. This is one of the standard methods for working with control systems. It is exploited in Section 4.4, which gives a systematic way of designing PI controllers for first order systems. This section also contains material required to develop an intuitive picture of the properties of second order systems. Section 4.5 deals with design of PI and PID controllers for second order systems. A proper background for Sections 4.4 and 4.5 is Section 3.4. Section 4.6 deals with the design problem for systems of arbitrary order. This section which requires more mathematical maturity can be omitted in a first reading. For the interested reader it gives, however, important insight into the design problem and the structure of stabilizing controllers. Section 4.6 summarizes the chapter and

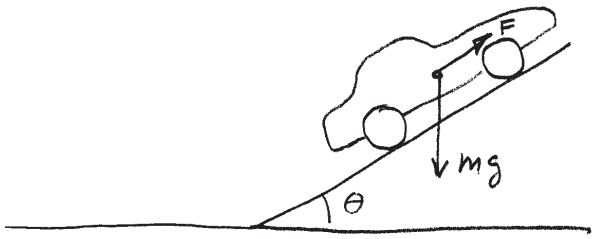


Figure 4.1 Schematic diagram of a car on a sloping road.

outlines some important issues that should be considered.

4.2 Cruise Control

The purpose of cruise control is to keep the velocity of a car constant. The driver drives the car at the desired speed, the cruise control system is activated by pushing a button and the system then keeps the speed constant. The major disturbance comes from changes of the slope of the road which generates forces on the car due to gravity. There are also disturbances due to air and rolling resistance. The cruise control system measures the difference between the desired and the actual velocity and generates a feedback signal which attempts to keep the error small in spite of changes in the slope of the road. The feedback signal is sent to an actuator which influences the throttle and thus the force generated by the engine.

We will start by developing a mathematical model of the system. The mathematical model should tell how the velocity of the car is influenced by the throttle and the slope of the road. A schematic picture is shown in Figure 4.1

Modeling

We will model the system by a momentum balance. The major part of the momentum is the product of the velocity v and the mass m of the car. There are also momenta stored in the engine, in terms of the rotation of the crank shaft and the velocities of the cylinders, but these are much smaller than mv. Let θ denote the slope of the road, the momentum balance can be

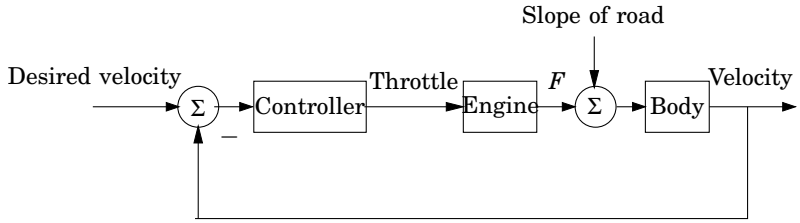


Figure 4.2 Block diagram of a car with cruise control.

written as

$$m\frac{dv}{dt} + cv = F - mg\theta \tag{4.1}$$

where the term cv describes the momentum loss due to air resistance and rolling and F is the force generated by the engine. The retarding force due to the slope of the road should similarly be proportional to the sine of the angle but we have approximated $\sin\theta \approx \theta$. The consequence of the approximations will be discussed later. It is also assumed that the force F developed by the engine is proportional to the signal u sent to the throttle. Introducing parameters for a particular car, an Audi in fourth gear, the model becomes

$$\frac{dv}{dt} + 0.02v = u - 10\theta \tag{4.2}$$

where the control signal is normalized to be in the interval $0 \leq u \leq 1$, where $u = 1$ corresponds to full throttle. The model implies that with full throttle in fourth gear the car cannot climb a road that is steeper than 10%, and that the maximum speed in 4th gear on a horizontal road is $v = 1/0.02 = 50$ m/s (180 km/hour).

Since it is desirable that the controller should be able to maintain constant speed during stationary conditions it is natural to choose a controller with integral action. A PI controller is a reasonable choice. Such a controller can be described by

$$u = k(v_r - v) + k_i \int_0^t (v_r - v(\tau))d\tau \tag{4.3}$$

A block diagram of the system is shown in Figure 4.2. To understand how the cruise control system works we will derive the equations for the closed loop systems described by Equations (4.2) and (4.3) Since the effect of the slope on the velocity is of primary interest we will derive an equation that tells how the velocity error $e = v_r - v$ depends on the slope of the road.

Assuming that v_r is constant we find that

$$\frac{dv}{dt} = -\frac{de}{dt}, \quad \frac{d^2v}{dt^2} = -\frac{d^2e}{dt^2}$$

It is convenient to differentiate (4.3) to avoid dealing both with integrals and derivatives. Differentiating the process model (4.2) the term du/dt can be eliminated and we find the following equation that describes the closed loop system

$$\frac{d^2e}{dt^2} + (0.02 + k)\frac{de}{dt} + k_i e = 10\frac{d\theta}{dt} \tag{4.4}$$

We can first observe that if θ and e are constant the error is zero. This is no surprise since the controller has integral action, see the discussion about the integral action Section 2.2.

To understand the effects of the controller parameters k and k_i we can make an analogy between (4.4) and the differential equation for a mass-spring-damper system

$$M\frac{d^2x}{dt^2} + D\frac{dx}{dt} + Kx = 0$$

We can thus conclude that parameter k influences damping and that k_i influences stiffness.

The closed loop system (4.4) is of second order and it has the characteristic polynomial

$$s^2 + (0.02 + k)s + k_i \tag{4.5}$$

We can immediately conclude that the roots of this polynomial can be given arbitrary values by choosing the controller parameters properly. To find reasonable values we compare the characteristic polynomial with the characteristic polynomial of the normalized second order polynomial

$$s^2 + 2\zeta\omega_0 s + \omega_0^2 \tag{4.6}$$

where ζ denotes relative damping and ω_0 is the undamped natural frequency. The parameter ω_0 gives response speed, and ζ determines the shape of the response. Comparing the coefficients of the closed loop characteristic polynomial (4.5) with the standard second order polynomial (4.6) we find that the controller parameters are given by

$$\begin{aligned} k &= 2\zeta\omega_0 - 0.02 \\ k_i &= \omega_0^2 \end{aligned} \tag{4.7}$$

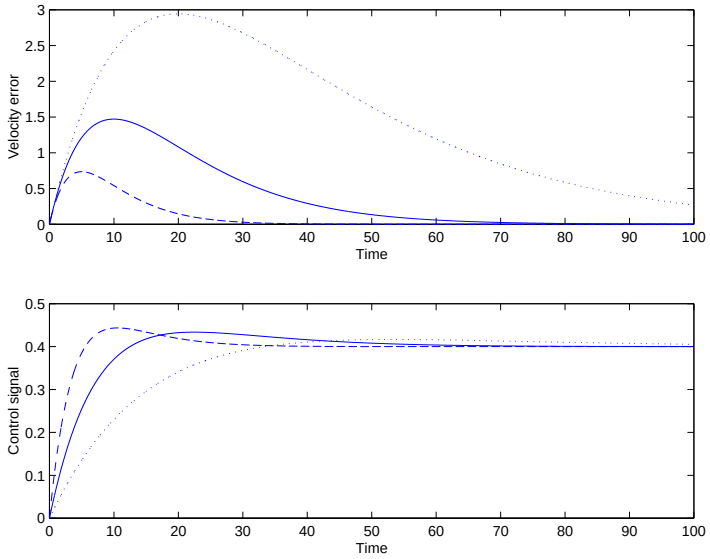


Figure 4.3 Simulation of a car with cruise control for a step change in the slope of the road. The controllers are designed with relative damping $\zeta = 1$ and $\omega_0 = 0.05$ (dotted), $\omega_0 = 0.1$ (full) and $\omega_0 = 0.2$ (dashed).

Since it is desirable that a cruise control system should respond to changes in the slope in a smooth manner without oscillations it is natural to choose $\zeta = 1$, which corresponds to critical damping. Then there is only one parameter ω_0 that has to be determined. The selection of this parameter is a compromise between response speed and control actions. This is illustrated in Figure 4.3 which shows the velocity error and the control signal for a simulation where the slope of the road suddenly changes by 4%. Notice that the largest velocity error decreases with increasing ω_0, but also that the control signal increases more rapidly. In the simple model (4.1) it was assumed that the force responded instantaneously to the throttle. For rapid changes there may be additional dynamics that has to be accounted for. There are also physical limitations to the rate of change of the force. These limitations, which are not accounted for in the simple model (4.1), limit the admissible value of ω_0. Figure 4.3 shows the velocity error and the control signal for a few values of ω_0. A reasonable choice of ω_0 is in the range of 0.1 to 0.2. The performance of the cruise control system can be evaluated by comparing the behaviors of cars with and without cruise control. This is done in Figure 4.4 which shows the velocity error when the slope of the road is suddenly increased by 4%. Notice the drastic

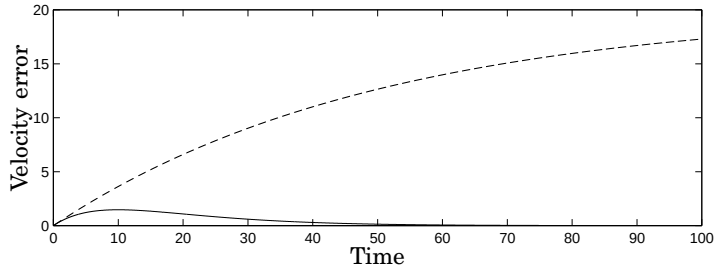


Figure 4.4 Simulation of a car with (solid line) and without cruise control (dashed line) for a step change of 4% in the slope of the road. The controller is designed for $\omega_0 = 0.1$ and $\zeta = 1$.

difference between the open and closed loop systems.

With the chosen parameters $\omega_0 = 0.2$ and $\zeta = 1$ we have $2\zeta\omega_0 = 0.2$ and it follows from (4.7) that the parameter $c = 0.02$ has little influence on the behavior of the closed loop system since it is an order of magnitude smaller than $2\zeta\omega_0$. Therefore it is not necessary to have a very precise value of this parameter. This is an illustration of an important and surprising property of feedback, namely that feedback systems can be designed based on simplified models. This will be discussed extensively in Chapter 5.

A cruise control system contains much more than an implementation of the PI controller given by (4.3). The human-machine interface is particularly important because the driver must be able to activate and deactivate the system and to change the desired velocity. There is also logic for deactivating the system when braking, accelerating or shifting gear.

4.3 Bicycle Dynamics

The bicycle is an ingenious device for recreation and transportation, which has evolved over a long period of time. It is a very effective vehicle that is extremely maneverable. Feedback is essential for understanding how the bicycle really works. In the bicycle there is no explicit control system with sensing and actuation, instead control is accomplished by clever mechanical design of the front fork which creates a feedback that under certain conditions stabilizes the bicycle. It is worth mentioning that the literature on bicycles is full of mistakes or misleading statements. We quote from the book Bicycling Science by Whitt and Wilson:

> The scientific literature (Timoshenko, Young, DenHartog et. al.) shows

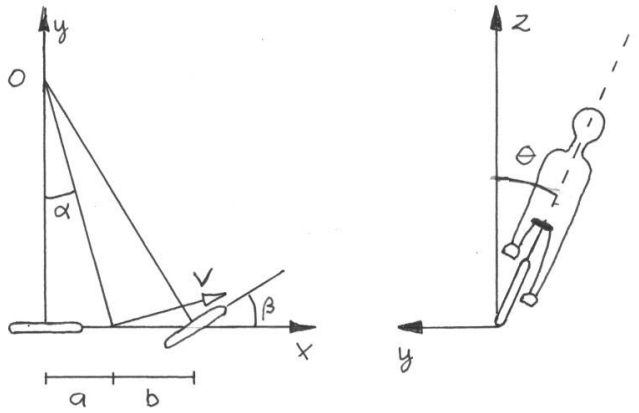


Figure 4.5 Schematic picture of a bicycle. The top view is shown on the left and the rear view on the right.

> often complete disagreement even about fundamentals. One advocates that a high center of mass improves stability, another concludes that a low center of mass is desirable.

We start by developing a simple modeling that clearly shows that feedback is an essential aspect of a bicycle.

Modeling

A detailed model of the bicycle is quite complicated. We will derive a simplified model that captures many relevant balancing properties of the bicycle. To understand how a bicycle works it is necessary to consider the system consisting of the bicycle and the rider. The rider can influence the bicycle in two ways by exerting a torque on the handle bar and by leaning. We will neglect the lean and consider the rider as a rigid body, firmly attached to the bicycle. A schematic picture of the bicycle is shown in Figure 4.5. To describe the dynamics we must account for the tilt of the bicycle. We introduce a coordinate system fixed to the bicycle with the x-axis through the contact points of the wheels with the ground, the y-axis horizontal and the z-axis vertical, as shown in Figure 4.5. Let m be the mass of the bicycle and the rider, J the moment of inertia of the bicycle and the rider with respect to the x-axis. Furthermore let l be the distance from the x-axis to the center of mass of bicycle and rider, θ the tilt angle and F the component of the force acting on rider and the bicycle.

A momentum balance around the x-axis gives

$$J\frac{d^2\theta}{dt^2} = mgl\sin\theta + Fl\cos\theta \tag{4.8}$$

The force F has two components, a centripetal force and an inertia force due to the acceleration of the coordinate system. The force can be determined from kinematic relations, see Figure 4.5. To describe these we introduce the steering angle β, and the forward velocity V_0. Furthermore the distance between the contact point of the front and rear wheel is b and the distance between the contact point of the rear wheel and the projection of the center of mass of bicycle and rider is a. To simplify the equations it is assumed that the angles β and θ are so small that sines and tangent are equal to the angle and cosine is equal to the one. Viewed from the top as shown in Figure 4.5 the bicycle has its center of rotation at a distance b/θ from the rear wheel. The centripetal force is

$$F_c = \frac{mV_0^2}{b}\beta$$

The y-component of the velocity of the center of mass is

$$V_y = V_0\alpha = \frac{aV_0}{b}\beta$$

where a is the distance from the contact point of the back wheel to the projection of the center of mass. The inertial force due to the acceleration of the coordinate system is thus

$$F_i = \frac{amV_0}{b}\frac{d\beta}{dt}$$

Inserting the total force $F = F_c + F_i$ into (4.8) we find that the bicycle can be described by

$$J\frac{d^2\theta}{dt^2} = mgl\theta + \frac{amV_0l}{b}\frac{d\beta}{dt} + \frac{mV_0^2l}{b}\beta \tag{4.9}$$

This equation has the characteristic equation

$$Js^2 - mgl = 0$$

which has the roots

$$s = \pm\sqrt{\frac{mgl}{J}}$$

The system is unstable, because the characteristic equation has one root in the right half plane. We may therefore believe that the rider must actively stabilize the bicycle all the time.

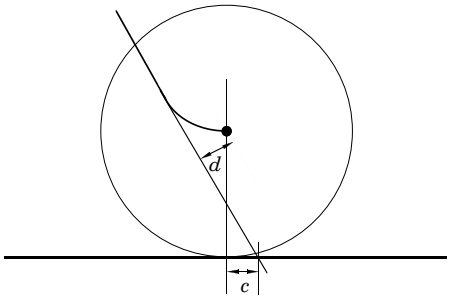


Figure 4.6 Schematic picture of the front fork.

The Front Fork

The bicycle has a front fork of rather intriguing design, see Figure 4.6. The front fork is angled and shaped so that the contact point of the wheel with the road is behind the axis of rotation of the front wheel assembly. The distance c is called the trail. The effect of this is that there will be a torque on the front wheel assembly when the bicycle is tilted. Because of the elasticity of the wheel there will be a compliance that also exerts a torque. The driver will also exert a torque on the front wheel assembly. Let T be the torque applied on the front fork by the driver. A static torque balance for the front fork assembly gives

$$\beta = -k_1\theta + k_2 T \tag{4.10}$$

Strictly speaking we should have a differential equation, for simplicity we will use the static equation.

Taking the action of the front fork into account we find that the bicycle is described by the Equations 4.9 and 4.10. A block diagram of representation of the system is shown in Figure 4.7. The figure shows clearly that the bicycle with the front fork is a feedback system. The front wheel angle β influences the tilt angle θ as described by (4.9) and the tilt angle influences the front wheel angle as described by (4.10). We will now investigate the consequences of the feedback created by the front fork. Inserting the expression (4.10) for steering angle β in the momentum balance (4.9) we get

$$J\frac{d^2\theta}{dt^2} + \frac{amV_0lk_1}{b}\frac{d\theta}{dt} + \Big(\frac{mV_0^2lk_1}{b} - mgl\Big)\theta = \frac{amV_0lk_2}{b}\frac{dT}{dt} + \frac{mV_0^2k_2l}{b}T \tag{4.11}$$

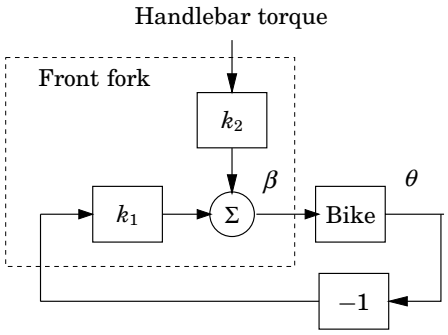


Figure 4.7 Block diagram of a bicycle with the front fork.

The characteristic equation of this system is

$$Js^2 + \frac{amV_0lk_1}{b}s + \left(\frac{mV_0^2lk_1}{b} - mgl\right) = 0$$

This equation is stable if

$$V_0 > V_c = \sqrt{\frac{gb}{k_1}} \tag{4.12}$$

We can thus conclude that because of the feedback created by the design of the front fork the bicycle will be stable provided that the velocity is sufficiently large. The velocity V_c is called the critical velocity.

Useful information about bicycle dynamics can be obtained by driving it with constant speed V_0 in a circle with radius r_0. To determine the numerical values of the essential parameters a torque wrench can be used to measure the torque the driver exerts on the handle bar. In steady state conditions the centripetal force must be balanced by the gravity. Assuming that the bicycle moves counter clockwise the lean angle is

$$\theta_0 = -\frac{V_0^2}{r_0 g}$$

It then follows from (4.11) that the torque required is given by

$$T_0 = \frac{bgl - V_0^2lk_1)}{k_2lr_0g} = \frac{k_1(V_c^2 - V_0^2)}{k_2r_0g}$$

This means that no torque is required if the bicycle is driven at the critical velocity and that the torque changes sign at the critical velocity.

Rear-wheel Steering

The analysis performed shows that feedback analysis gives substantial insight into behavior of bicycles. Feedback analysis can also indicate that a proposed system may have substantial disadvantages that are not apparent from static analysis. It is therefore essential to consider feedback and dynamics at an early stage of design. We illustrate this with a bicycle example. There are advantages in having rear-wheel steering on recumbent bicycles because the design of the drive is simpler. Again we quote from Whitt and Wilson Bicycling Science:

> The U.S. Department of Transportation commissioned the construction of a safe motorcycle with this configuration (rear-wheel steering). It turned out to be safe in an unexpected way: No one could ride it.

The reason for this is that a bicycle with rear-wheel steering has dynamics which makes it very difficult to ride. This will be discussed in Sections 5.9. Let it suffice to mention that it is essential to consider dynamics and control at an early stage of the design process. This is probable the most important reason why all engineers should have a basic knowledge about control.

4.4 Control of First Order Systems

We will now develop a systematic procedure for finding controllers for simple systems. To do this we will be using the formalism based on Laplace transforms and transfer functions which is developed in Section 3.4. This simplifies the calculations required substantially. In this section we will consider systems whose dynamics are of first order differential equations. Many systems can be approximately described by such equations. The approximation is reasonable if the storage of mass, momentum and energy can be captured by one state variable. Typical examples are

- Velocity of car on the road
- Control of velocity of rotating system
- Electric systems where energy is essentially stored in one component
- Incompressible fluid flow in a pipe
- Level control of a tank
- Pressure control in a gas tank
- Temperature in a body with essentially uniform temperature distribution e.g., a vessel filled with a mixture of steam and water.

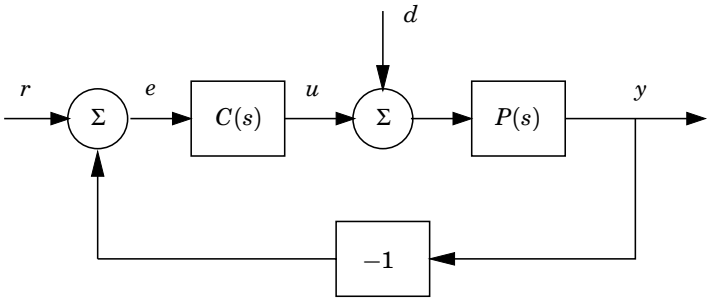


Figure 4.8 Block diagram of a first order system with a PI controller.

A linear model of a first order system can be described by the transfer function

$$P(s) = \frac{b}{s+a} \tag{4.13}$$

The system thus has two parameters. These parameters can be determined from physical consideration or from a step response test on the system. A step test will also reveal if it is reasonable to model a system by a first order model.

To have no steady state error a controller must have integral action. It is therefore natural to use a PI controller which has the transfer function

$$C(s) = k + \frac{k_i}{s} \tag{4.14}$$

A block diagram of the system is shown in Figure 4.8. The loop transfer function of the system is

$$L(s) = P(s)C(s) = \frac{kbs + k_i b}{s(s+a)} = \frac{n_L(s)}{d_L(s)} \tag{4.15}$$

The transfer function of the closed system from reference r to output y is given by

$$\frac{Y(s)}{R(s)} = \frac{P(s)C(s)}{1 + P(s)C(s)} = \frac{n_L(s)}{d_L(s) + n_L(s)} = \frac{b(ks + k_i)}{s^2 + (a + bk)s + bk_i}$$

The closed loop system is of second order and its characteristic polynomial is

$$d_L(s) + n_L(s) = s^2 + (a + bk)s + bk_i. \tag{4.16}$$

The poles of the closed loop system can be given arbitrary values by choosing the parameters k and k_i properly. Intuition about the effects of the

parameters can be obtained from the mass-spring-damper analogy as was done in Section 4.2 and we find that integral gain k_i corresponds to stiffness and that proportional gain k corresponds to damping.

It is convenient to re-parameterize the problem so that the characteristic polynomial becomes

$$s^2 + 2\zeta\omega_0 s + \omega_0^2 \tag{4.17}$$

Identifying the coefficients of s in the polynomials (4.16) and (4.17) we find that the controller parameters are given by

$$\begin{aligned} k &= \frac{2\zeta\omega_0 - a}{b} \\ k_i &= \frac{\omega_0^2}{b} \end{aligned} \tag{4.18}$$

Since the design method is based on choosing the poles of the closed loop system it is called pole placement. Instead of choosing the controller parameters k and k_i we now select ζ and ω_0. These parameters have a good physical interpretation. The parameter ω_0 determines the speed of response and ζ determines the shape of the response. Controllers often have parameters that can be tuned manually. For a PI controller it is customary to use the parameters k and k_i. When a PI controller is used for a particular system, where the model is known, it is much more practical to use other parameters. If the model can be approximated by a first order model it is very convenient to have ω_0 and ζ as parameters. We call this performance related parameters because they are related directly to the properties of the closed loop system.

If the parameters ω_0 and ζ are known the controller parameters are given by (4.18). We will now discuss how to choose these parameters.

Behavior of Second Order Systems

We will first consider a second order system with the transfer function

$$G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}. \tag{4.19}$$

This is a normalized transfer function of a second order system without zeros. The step responses of systems with different values of ζ are shown in Figure 4.9 The figure shows that parameter ω_0 essentially gives a time scaling. The response is faster if ω_0 is larger. The shape of the response

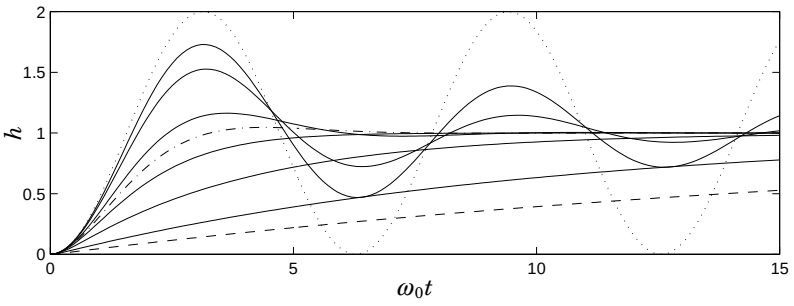


Figure 4.9 Step responses h for the system (4.19) with the transfer function $G(s) = \frac{\omega_0^2}{s^2+2\zeta\omega_0 s+\omega_0^2}$ for $\zeta = 0$ (dotted), 0.1, 0.2, 0.5, 0.707 (dash dotted), 1, 2, 5 and 10 (dashed).

is determined by ζ. The step responses have an overshoot of

$$M = \begin{cases} e^{-\frac{\pi\zeta}{\sqrt{1-\zeta^2}}} & \text{for } |\zeta| < 1 \\ 1 & \text{for} \zeta \geq 1 \end{cases}$$

For $\zeta < 1$ the maximum overshoot occurs at

$$t_{max} = \frac{2\pi}{\omega_0\sqrt{1-\zeta^2}}$$

There is always an overshoot if $\zeta < 1$. The maximum decreases and is shifted to the right when ζ increases and it becomes infinite for $\zeta = 1$ when the overshoot disappears. In most cases it is desirable to have a moderate overshoot which means that the parameter ζ should be in the range of 0.5 to 1. The value $\zeta = 1$ gives no overshoot.

Behavior of Second Order Systems with Zeros

We will now consider a system with the transfer function

$$G(s) = \frac{\omega_0}{\beta}\frac{s+\beta\omega_0}{s^2+2\zeta\omega_0 s+\omega_0^2} \tag{4.20}$$

Notice that the transfer function has been parameterized so that the steady state gain $G(0)$ is one. Step responses for this transfer function for different values of β are shown in Figure 4.10. The figure shows that

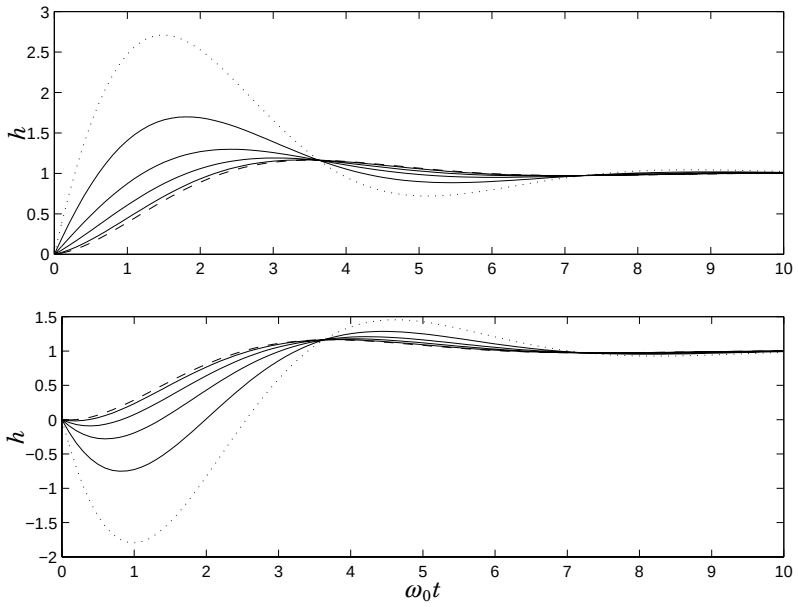


Figure 4.10 Step responses h for the system (4.19) with the transfer function $G(s) = \frac{\omega_0(s+\beta\omega_0)}{\beta(s^2+2\zeta\omega_0 s+\omega_0^2)}$ for $\omega_0 = 1$ and $\zeta = 0.5$. The values for $\beta = 0.25$ (dotted), 0.5 1, 2, 5 and 10 (dashed), are shown in the upper plot and $\beta = -0.25$, -0.5 -1, -2, -5 and -10 (dashed) in the lower plot.

the zero introduces overshoot for positive β and an undershoot for negative β. Notice that the effect of β is most pronounced if β is small. The effect of the zero is small if $|\beta| > 5$. Intuitively it it appears that systems with negative values of β, where the output goes in the wrong direction initially, are difficult to control. This is indeed the case as will be discussed later. Systems with this type of behavior are said to have inverse response. The behavior in the figures can be understood analytically. The transfer function $G(s)$ can be written as

$$G(s) = \frac{\omega_0}{\beta}\frac{s+\beta\omega_0}{s^2+2\zeta\omega_0 s+\omega_0^2} = \frac{\omega_0^2}{s^2+2\zeta\omega_0 s+\omega_0^2} + \frac{1}{\beta}\frac{s\omega_0}{s^2+2\zeta\omega_0 s+\omega_0^2}$$

Let $h_0(t)$ be the step response of the transfer function

$$G_0(s) = \frac{\omega_0^2}{s^2+2\zeta\omega_0 s+\omega_0^2}$$

It follows from (4.20) that the step response of $G(s)$ is

$$h(t) = h_0(t) + \frac{1}{\beta\omega_0}\frac{dh_0(t)}{dt} \tag{4.21}$$

It follows from this equation that all step responses for different values of β go through the point where dh_0/dt is zero. The overshoot will increase for positive β and decrease for negative β. It also follows that the effect of the zero is small if $|\beta|$ is large. The largest magnitude of dh/dt is approximately $0.4\omega_0/2.7$, which implies that the largest value of the second term is approximately $0.4/\beta$. The term is thus less than 8% if $|\beta|$ is larger than 5.

Notice in Figure 4.10 that the step response goes in the wrong direction initially when β is negative. This phenomena is called inverse response, can also be seen from (4.21). When β is negative the transfer function (4.20) has a zero in the right half plane. Such are difficult to control and they are called non-minimum phase system, see Section 3.5. Several physical systems have this property, for example level dynamics in steam generators (Example 3.18, hydro-electric power stations (Example 3.17), pitch dynamics of an aircraft (Example 3.19) and vehicles with rear wheel steering.

The Servo Problem

Having developed insight into the behavior of second order systems with zeros we will return to the problem of PI control of first order systems. We will discuss selection of controller parameters for the servo problem where the main concern is that the output should follow the reference signal well. The loop transfer function of the system is given by (4.15) and the transfer function from reference r to output y is

$$G_{yr} = \frac{Y(s)}{R(s)} = \frac{P(s)C(s)}{1 + P(s)C(s)} = \frac{n_L(s)}{n_D(s) + n_L(s)} = \frac{(a + bk)s + bk_i}{s^2 + (a + bk)s + bk_i}$$

Choosing control parameters to give the characteristic polynomial (4.17) we find as before that the controller parameters are given by (4.18) and the transfer function above becomes

$$\frac{Y(s)}{R(s)} = \frac{(a + bk)s + bk_i}{s^2 + (a + bk)s + bk_i} = \frac{2\zeta\omega_0 s + \omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{4.22}$$

Comparing this transfer function with the transfer function (4.20) we find that

$$\beta = 2\zeta$$

This implies that parameter β is in the range of 1 to 2 for reasonable choices of ζ. Comparing with Figure 4.10 shows that the system has a significant overshoot. This can be avoided by a simple modification of the controller.

Avoiding the Overshoot - Systems with two degrees of freedom

The controller used in Figure 4.8 is based on error feedback. The control signal is related to the reference and the output in the following way

$$u(t) = k(r(t) - y(t)) + k_i \int_0^t (r(\tau) - y(\tau))d\tau \tag{4.23}$$

The reason for the overshoot is that the controller reacts quite violently on a step change in the reference. By changing the controller to

$$u(t) = -ky(t) + k_i \int_0^t (r(\tau) - y(\tau))d\tau \tag{4.24}$$

we obtain a controller that is reacting much less violent to changes in the reference signal. Taking Laplace transforms of this controller we get

$$U(s) = -kY(s) + \frac{k_i}{s}(R(s) - Y(s)) \tag{4.25}$$

Combining this equation with the equation (4.13) which describes the process we find that

$$\frac{Y(s)}{R(s)} = \frac{bk_i}{s^2 + (a + bk)s + bk_i} = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{4.26}$$

and we obtain a transfer function from reference r to output y which does not have a zero, compare with (4.20).

The controller given by (4.23) is said to have error feedback because all control actions are based on the error $e = r - y$. The controller given by (4.24) is said to have two degrees of freedom (2DOF) because the signal path from reference r to control u is different from the signal path from output y to control u. Figure 4.11 shows block diagrams of the systems. The transfer function (4.26) is the standard transfer function for a second order system without zeros, its step responses are shown in Figure 4.9.

The Regulation Problem

It will now be investigated how the parameters ω_0 and ζ should be chosen for the regulation problem. In this problem the main concern is reduction

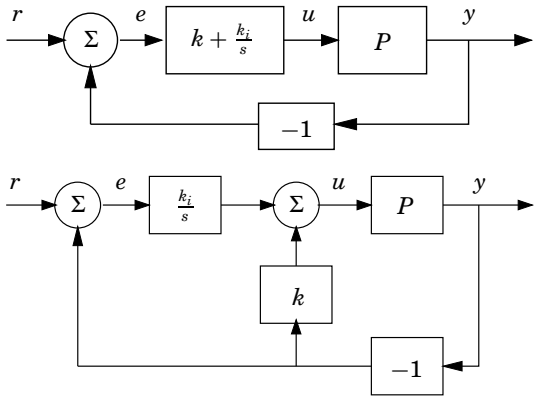


Figure 4.11 Block diagrams of a system with a conventional PI controller (above) and a PI controller having two degrees of freedom (below).

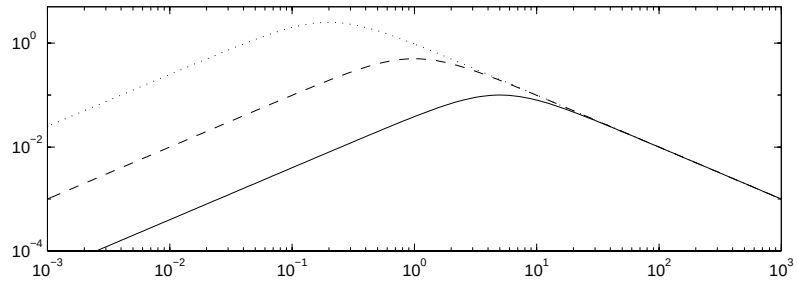


Figure 4.12 Gain curves for the transfer function from load disturbance to process output for $b = 1$, $\zeta = 1$ and $\omega_0 = 0.2$ dotted, $\omega_0 = 1.0$, dashed and $\omega_0 = 5$ full.

of load disturbances. Consider the system in Figure 4.8, the transfer function from load disturbance d to output y is

$$G_{yd}(s) = \frac{Y(s)}{D(s)} = \frac{P(s)}{1 + P(s)C(s)} = \frac{s}{s^2 + (a + bk)s + bk_i}$$
$$= \frac{bs}{s^2 + 2\zeta\omega_0 s + \omega_0^2} = \frac{b}{\omega_0}\frac{\omega_0 s}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

We will first consider the effect of parameter ω_0. Figure 4.12 shows the gain curves of the Bode diagram for different values of ω_0. The figure shows that disturbances of high and low frequencies are reduced significantly and that the disturbance reduction is smallest for frequencies

around ω_0, they may actually be amplified. The figure also shows that the disturbance rejection at low frequencies is drastically influenced by the parameter ω_0 but that the reduction of high frequency disturbances is virtually independent of ω_0. It is easy to make analytical estimates because we have

$$G_{yd}(s) \approx \frac{bs}{\omega_0^2} = \frac{s}{bk_i}$$

for small s, where the second equality follows from (4.18). It follows from this equation that it is highly desirable to have a large value of ω_0. A large value of ω_0 means that the control signal has to change rapidly. The largest permissible value of ω_0 is typically determined by how quickly the control signal can be changed, dynamics that was neglected in the simple model (4.13) and possible saturations. The integrated error for a unit step disturbance in the load disturbance is

$$IE = \int_0^\infty e(t)dt = \lim_{s\to 0} E(s) = \lim_{s\to 0} G_{yd}\frac{1}{s} = \frac{b}{\omega_0^2} = \frac{1}{bk_i}$$

The largest value of $|G_{yd}(i\omega)|$ is

$$\max|G_{yd}(i\omega)| = |G_{yd}(i\omega_0)| = \frac{b}{2\zeta\omega_0}$$

The closed loop system obtained with PI control of a first order system is of second order. Before proceeding we will investigate the behavior of second order systems.

4.5 Control of Second Order Systems

We will now discuss control of systems whose dynamics can approximately be described by differential equations of second order. Such an approximation is reasonable if the storage of mass, momentum and energy can be captured by two state variables. Typical examples are

- Position of car on the road
- Motion control systems
- Stabilization of satellites
- Electric systems where energy is stored in two elements
- Levels in two connected tanks
- Pressure in two connected vessels

- Simple bicycle models

The general transfer function for a process of second order is

$$P(s) = \frac{b_1 s + b_2}{s^2 + a_1 s + a_2} \tag{4.27}$$

In some cases we will consider the special case when $b_1 = 0$.

PD control

We will first design a PD control of the process

$$P(s) = \frac{b}{s^2 + a_1 s + a_2}$$

A PD controller with error feedback has the transfer function

$$C(s) = k + k_d s$$

The loop transfer function is

$$L(s) = P(s)C(s) = \frac{bk_d s + bk}{s^2 + a_1 s + a_2} = \frac{n_L(s)}{d_L(s)}$$

The closed loop transfer function from reference to output is

$$\begin{aligned}\frac{Y(s)}{R(s)} &= \frac{PC}{1 + PC} = \frac{n_L(s)}{n_D(s) + n_L(s)} = \frac{b(k_d s + k)}{s^2 + a_1 s + a_2 + b(k_d s + k)} \\ &= \frac{b(k_d s + k)}{s^2 + (a_1 + bk_d)s + a_2 + bk}\end{aligned}$$

The closed loop system is of second order and the controller has two parameters. The characteristic polynomial of the closed loop system is

$$s^2 + (a_1 + bk_d)s + a_2 + bk \tag{4.28}$$

Matching this with the standard polynomial

$$s^2 + 2\zeta\omega_0 s + \omega_0^2$$

we get

$$\begin{aligned}k &= \frac{\omega_0^2 - a_2}{b} \\ k_d &= \frac{2\zeta\omega_0 - a_1}{b}\end{aligned} \tag{4.29}$$

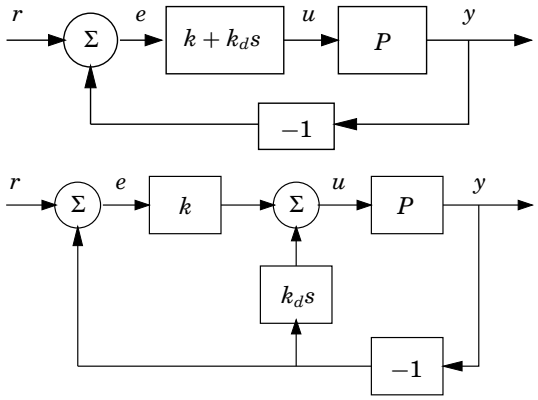


Figure 4.13 Block diagrams of system with PD control based on error feedback (above) and with a PD controller with two degrees of freedom (below). Compare with Figure 4.11.

The closed loop transfer function from reference to output becomes

$$\frac{Y(s)}{R(s)} = \frac{PC}{1+PC} = \frac{(2\zeta\omega_0 - a_1)s + \omega_0^2 - a_2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

Notice that there will be a steady state error unless $a_2 = 0$. The steady state error is small if $|a_2| << \omega_0^2$. Also notice that the zero in the numerator may cause overshoot. To avoid this the controller based on error feedback can be replaced with the following controller

$$U(s) = k(R(s) - Y(s)) - k_d s Y(s) \tag{4.30}$$

which has two degrees of freedom. The transfer function from reference to output for the closed loop system then becomes

$$\frac{Y(s)}{R(s)} = \frac{\omega_0^2 - a_2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

Notice that this transfer function does not have a zero. Block diagrams for the system with error feedback and with two degrees of freedom are shown in Figure 4.13.

PI Control

Next we will investigate what can be achieved with PI control of the process given by (4.27). Let the PI controller have the transfer function

$$C(s) = k + \frac{k_i}{s} = \frac{ks + k_i}{s}$$

The loop transfer function becomes

$$L(s) = P(s)C(s) = \frac{(ks + k_i)(b_1 s + b_2)}{s^3 + a_1 s^2 + a_2 s} = \frac{n_L(s)}{d_L(s)}$$

The characteristic polynomial is

$$n_L(s) + d_L(s) = s^3 + (a_1 + kb_1)s^2 + (a_2 + kb_2 + k_i b_1) + b_2 k_i$$

Identifying the coefficients of this equation with the desired characteristic polynomial

$$(s^2+2\zeta\omega_0 s+\omega_0^2)(s+\alpha\omega_0) = s^3+(\alpha+2\zeta)\omega_0 s^2+(1+2\alpha\zeta)\omega_0^2 s+\alpha\omega_0^3 \quad (4.31)$$

we obtain

$$\begin{aligned} a_1 + b_1 k &= (\alpha + 2\zeta)\omega_0 \\ a_2 + b_1 k_i + b_2 k &= (1 + 2\alpha\zeta)\omega_0^2 \\ b_2 k_i &= \alpha\omega_0^3 \end{aligned}$$

Since there are three equations and only two unknowns the problem cannot be solved in general. To have a solution we can let ω_0 be a free parameter. If $b_1 = 0$ and $b_2 \neq 0$ the equation then has the solution

$$\begin{aligned} \omega_0 &= \frac{a_1}{\alpha + 2\zeta} \\ k &= \frac{(1 + 2\alpha\zeta)\omega_0^2 - a_2}{b_2} \\ k_i &= \frac{\alpha\omega_0^3}{b_2} \end{aligned} \quad (4.32)$$

The parameter ω_0 which determines the response time is thus uniquely given by the process dynamics. When $b_1 \neq 0$ the parameter ω_0 is instead the real solution to the equation

$$\alpha b_1^2 \omega_0^3 - (1 + 2\alpha\zeta)b_1 b_2 \omega_0^2 + (\alpha + 2\zeta)b_2^2 \omega_0 + a_2 b_1 b_2 - a_1 b_2^2 = 0.$$

and the controller parameters are given by

$$\begin{aligned} k &= \frac{(\alpha + 2\zeta)\omega_0 - a_1}{b_1} \\ k_i &= \frac{\alpha\omega_0^3}{b_2} \end{aligned}$$

In both cases we find that with PI control of a second order system there is only one choice of ω_0 that is possible. The performance of the closed loop system is thus severely restricted when a PI controller is used.

PID Control

Assume that the process is characterized by the second-order model

$$P(s) = \frac{b_1 s + b_2}{s^2 + a_1 s + a_2} \tag{4.33}$$

This model has four parameters. It has two poles that may be real or complex, and it has one zero. This model captures many processes, oscillatory systems, and systems with right half-plane zeros. The right half-plane zero can also be used as an approximation of a time delay. Let controller be

$$U(s) = k(bR(s) - Y(s)) + \frac{k_i}{s}(R(s) - Y(s)) + k_d s(cR(s) - Y(s))$$

The loop transfer function is

$$L(s) = \frac{(k_d s^2 + ks + k_i)(b_1 s + b_2)}{s(s^2 + a_1 s + a_2)} = \frac{n_L(s)}{d_L(s)}$$

The closed-loop system is of third order with the characteristic polynomial

$$\begin{aligned} d_L(s) + n_L(s) &= s(s^2 + a_1 s + a_2) + (b_1 s + b_2)(k_d s^2 + ks + k_i) \\ &= (1 + b_1 k)s^3 + (a_1 + b_1 k + b_2 k_d)s^2 + (a_2 + b_1 k_i + b_2 k)s + b_2 k_i \\ &= (1 + b_1 k)\Big(s^3 + \frac{a_1 + b_1 k + b_2 k_d}{1 + b_1 k} s^2 + \frac{a_2 + b_1 k_i + b_2 k}{1 + b_1 k} s + \frac{b_2 k_i}{1 + b_1 k}\Big) \end{aligned}$$

A suitable closed-loop characteristic equation of a third-order system is

$$(s + \alpha\omega_0)(s^2 + 2\zeta\omega_0 s + \omega_0^2)$$

Equating coefficients of equal power in s in this polynomial with the normalized characteristic polynomial gives

$$\begin{aligned} \frac{a_1 + b_2 k_d + b_1 k}{1 + b_1 k_d} &= (\alpha + 2\zeta)\omega_0 \\ \frac{a_2 + b_2 k + b_1 k_i}{1 + b_1 k_d} &= (1 + 2\alpha\zeta)\omega_0^2 \\ \frac{b_2 k_i}{1 + b_1 k_d} &= \alpha\omega_0^3 \end{aligned}$$

This is a set of linear equations in the controller parameters. The solution is straightforward but tedious and is given by

$$
\begin{aligned}
k &= \frac{a_2 b_2^2 - a_2 b_1 b_2(\alpha + 2\zeta)\omega_0 - (b_2 - a_1 b_1)(b_2(1 + 2\alpha\zeta)\omega_0^2 + \alpha b_1 \omega_0^3)}{b_2^3 - b_1 b_2^2(\alpha + 2\zeta)\omega_0 + b_1^2 b_2(1 + 2\alpha\zeta)\omega_0^2 - \alpha b_1^3 \omega_0^3} \\
k_i &= \frac{(-a_1 b_1 b_2 + a_2 b_1^2 + b_2^2)\alpha\omega_0^3}{b_2^3 - b_1 b_2^2(\alpha + 2\zeta)\omega_0 + b_1^2 b_2(1 + 2\alpha\zeta)\omega_0^2 - \alpha b_1^3 \omega_0^3} \\
k_d &= \frac{-a_1 b_2^2 + a_2 b_1 b_2 + b_2^2(\alpha + 2\zeta)\omega_0 - b_1 b_2 \omega_0^2(1 + 2\alpha\zeta) + b_1^2 \alpha\omega_0^3}{b_2^3 - b_1 b_2^2(\alpha + 2\zeta)\omega_0 + b_1^2 b_2(1 + 2\alpha\zeta)\omega_0^2 - \alpha b_1^3 \omega_0^3}
\end{aligned}
\tag{4.34}
$$

The transfer function from set point to process output is

$$
G_{yr}(s) = \frac{(b_1 s + b_2)(c k_d s^2 + b k s + k_i)}{(s + \alpha\omega_0)(s^2 + 2\zeta\omega_0 s + \omega_0^2)}
$$

The parameters b and c have a strong influence on shape of the transient response of this transfer function.

The transfer function from load disturbance to process output is

$$
G_{yd} = \frac{b_1 s^2 + b_2 s}{(s + \alpha\omega_0)(s^2 + 2\zeta\omega_0 s + \omega_0^2)}
$$

These formulas are useful because many processes can be approximately described by the transfer function (4.27). We illustrate this with an example.

EXAMPLE 4.1—OSCILLATORY SYSTEM WITH RHP ZERO
Consider a system with the transfer function

$$
P(s) = \frac{1 - s}{s^2 + 1}
$$

This system has one right half-plane zero and two undamped complex poles. The process is difficult to control.

$$
s^3 + 2s^2 + 2s + 1.
$$

(4.34) gives a PID controller with the parameters $k = 0$, $k_i = 1/3$, and $k_d = 2/3$. Notice that the proportional gain is zero. □

We will give an example that illustrates that there are situations where a PID controller can be much better than a PI controller.

EXAMPLE 4.2—PID CAN BE MUCH BETTER THAN PI
Consider a process described by

$$P(s) = \frac{k_v}{s(1+sT)} e^{-sT_d} \tag{4.35}$$

where the time delay T_d is much smaller than the time constant T. Since the time constant T is small it can be neglected and the design can be based on the second order model

$$P(s) \approx \frac{k_v}{s(1+sT)} \tag{4.36}$$

A PI controller for this system can be obtained from Equation (4.32) and we find that a closed loop system with the characteristic polynomial (4.31) can be obtained by choosing the parameter ω_0 equal to $1/(\alpha+2\zeta)T$. Since $T_d << T$ it follows that $\omega_0 T_d << 1$ and it is reasonable to neglect the time delay.

If the approximation (4.36) it is possible to find a PID controller that gives the closed loop characteristic polynomial with arbitrarily large values of ω_0. Since the real system is described by (4.35) the parameter ω_0 must be chosen so that the approximation (4.36) is valid. This requires that the product $\omega_0 T_d$ is not too large. It can be demonstrated that the approximation is reasonable if $\omega_0 T_d$ is smaller than 0.2.

Summarizing we find that it is possible to obtain the characteristic polynomial (4.31) with both PI and PID control. With PI control the parameter ω_0 must be chosen as $1/(\alpha+2\zeta)T$. With PID control the parameter instead can be chosen so that the product $\omega_0 T_d < 1$ is small, e.g. 0.2 or less. With PI control the response speed is thus determined by T and with PID control it is determined by T_d. The differences can be very significant. Assume for example that $T = 100$, $T_d = 1$, $\alpha = 1$ and $\zeta = 0.5$. Then we find that with $\omega_0 = 0.005$ with PI control and $\omega_0 = 0.1$ with PID control. This corresponds to a factor of 200 in response time. This will also be reflected in a much better disturbance attenuation with PID control. □

4.6 Control of Systems of High Order*

The method for control design used in the previous sections can be characterized in the following way. Choose a controller of given complexity, PD, PI or PID and determine the controller parameters so that the closed loop

characteristic polynomial is equal to a specified polynomial. This technique is called pole placement because the design is focused on achieving a closed loop system with specified poles. The zeros of the transfer function from reference to output can to some extent be influenced by choosing a controller with two degrees of freedom. We also observed that the complexity of the controller reflected the complexity of the process. A PI controller is sufficient for a first order system but a PID controller was required for a second order system. Choosing a controller of too low order imposed restrictions on the achievable closed loop poles. In this section we will generalize the results to systems of arbitrary order. This section also requires more mathematical preparation than the rest of the book.

Consider a system given by the block diagram in Figure 4.8. Let the process have the transfer function

$$P(s) = \frac{Y(s)}{U(s)} = \frac{b(s)}{a(s)} = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n}{s^n + a_1 s^{n-1} + \ldots + a_n} \tag{4.37}$$

where $a(s)$ and $b(s)$ are polynomials. A general controller can be described by

$$f(s)U(s) = -g(s)Y(s) + h(s)R(s) \tag{4.38}$$

where $f(s)$, $g(s)$ and $h(s)$ are polynomials. The controller given by (4.38) is a general controller with two degrees of freedom. The transfer function from measurement signal y to control signal u is $-g(s)/f(s)$ and the transfer function from reference signal r to control signal u is $h(s)/f(s)$. For a system with error feedback we have $g(s) = h(s)$. Elimination of $U(s)$ between Equations (4.37) and (4.38) gives

$$\big(a(s)f(s) + b(s)g(s)\big)Y(s) = b(s)h(s)R(s) + b(s)f(s)D(s) \tag{4.39}$$

The closed loop has the characteristic polynomial

$$c(s) = a(s)f(s) + b(s)g(s) \tag{4.40}$$

Notice that this only depends on the polynomials $f(s)$ and $g(s)$. The design problem can be stated as follows: Given the polynomials $a(s)$, $b(s)$ and $c(s)$ find the polynomials $f(s)$ and $g(s)$ which satisfies (4.40). This is a well known mathematical problem. It will be shown in the next section that the equation always has a solution if the polynomials $a(s)$ and $b(s)$ do not have any common factors. If one solution exists there are also infinitely many solutions. This is useful because it makes it possible to introduce additional constraints. We may for example require that the controller should have integral action.

A Naive Solution

To obtain the solution to the design problem the equation (4.40) must be solved. A simple direct way of doing this is to introduce polynomials f and g with arbitrary coefficients, writing equating coefficients of equal powers of s, and solving the equations. This procedure is illustrated by an example.

EXAMPLE 4.3—GENERAL POLE PLACEMENT
Consider a process with the transfer function

$$P(s) = \frac{1}{(s+1)^2}$$

Find a controller that gives a closed loop system with the characteristic polynomial

$$(s^2 + as + a^2)(s + a)$$

(4.40) becomes

$$(s+1)^2 f + g = (s^2 + as + a^2)(s + a) = s^3 + 2as^2 + 2a^2 s + a^3$$

One solution is

$$\begin{aligned} f &= 1 \\ g &= s^3 + (2a - 1)s^2 + (2a^2 - 2)s + a^3 - 1 \end{aligned}$$

but there are also other solutions e.g.

$$\begin{aligned} f &= s + 2a - 2 \\ g &= (2a^2 - 4a + 3)s + a^3 - 2a + 2 \end{aligned}$$

□

The Diophantine Equation

The naive solution of (4.40) hides many interesting aspects of the problem. The equation (4.40) is a classical equation which has been studied extensively in mathematics. To discuss this equation we will use more mathematics than in most parts of the book. We will also change to a more formal style of presentation. This is a nice illustration of the fact that control is a field where many branches of mathematics are useful.

We will start by observing that polynomials belong to a mathematical object called a ring. This means that they can be multiplied and added,

and that there are units: the zero polynomial for addition and the polynomial 1 for multiplication. Division of polynomials does not always give a polynomial, but quotient and remainders are defined. Integers are other objects that also is a ring. To develop some insight we will first explore two examples.

EXAMPLE 4.4—AN EQUATION IN INTEGERS
Consider the following equation

$$3x + 2y = 1,$$

where x and y are integers. By inspection we find that $x = 1$ and $y = -1$ is a solution. We also find that if we have a solution other solutions can be obtained by adding 2 to x and subtracting 3 from y. The equation thus has infinitely many solutions. □

EXAMPLE 4.5—AN EQUATION IN INTEGERS
Consider the equation

$$6x + 4y = 1,$$

where x and y are integers. This equation cannot have a solution because the left hand side is an even number and the right hand side is an odd number. □

EXAMPLE 4.6—AN EQUATION IN INTEGERS
Consider the equation

$$6x + 4y = 2,$$

where x and y are integers. Dividing the right hand side by 2 we obtain the equation in Example 4.4 □

These examples tell most about the (4.40) when a, b, f, g and c belong to a ring. To be precise we have the following result.

THEOREM 4.1—EUCLID'S ALGORITHM
Let a, b, and c be polynomials with real coefficients. Then the equation

$$ax + by = c \tag{4.41}$$

has a solution if and only if the greatest common factor of a and b divides c. If the equation has a solution x_0 and y_0 then $x = x_0 - bn$ and $y = y_0 + an$, where n is an arbitrary integer, is also a solution.

PROOF 4.1
We will first determine the largest common divisor of the polynomials a and b by a recursive procedure. Assume that the degree of a is greater than or equal to the degree of b. Let $a^0 = a$ and $b^0 = b$. Iterate the equations

$$a^{n+1} = b^n$$
$$b^{n+1} = a^n \text{ mod } \mathrm{b^n}$$

until $b^{n+1} = 0$. The greatest common divisor is then b^n. If a and b are co-prime we have $b^n = 1$. Backtracking we find that

$$ax + by = b^n$$

where the polynomials x and y can be found by keeping track of the quotients and the remainders in the iterations. When a and b are co-prime we have

$$ax + by = 1$$

and the result is obtained by multiplying x and y by c. When a and b have a common factor it must be required that the largest common divisor of a and b is also a factor of c. Dividing the equation with this divisor we are back to the case when a and b are co-prime. □

Since the proof has only used addition, multiplication, quotients and remainders it follows that the results holds for any ring.

An Algorithm

The following is a convenient way of organizing the recursive computations. With this method we also obtain the minimum degree solution to the homogeneous equation.

$$\begin{aligned} ax + by &= 1 \\ au + bv &= 0 \end{aligned} \tag{4.42}$$

where g is the greatest common divisor of a and b and u and v are the minimal degree solutions to the homogeneous equation These equations can be written as

$$\begin{pmatrix} x & y \\ u & v \end{pmatrix} \begin{pmatrix} a & 1 & 0 \\ b & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & x & y \\ 0 & u & v \end{pmatrix}$$

The solution to Equation (4.42) can be obtained by transforming the matrix

$$A^0 = \begin{pmatrix} a & 1 & 0 \\ b & 0 & 1 \end{pmatrix}$$

by row operations to a matrix where the 21 element is zero. This can be done recursively as follows. Assume that $\deg a$ is greater than or equal to $\deg b$, exchange the rows if this is not the case. Form the following recursion.

$$A^{n+1} = \begin{pmatrix} A_{21}^n & A_{22}^n & A_{23}^n \\ r^n & A_{12}^n - q^n A_{22}^n & A_{13}^n - q^n A_{23}^n \end{pmatrix}$$

where $q^n = A_{11}^n$ div A_{21}^n and $r_n = A_{11}^n$ div A_{21}^n. Proceed until $A_{21}^{n+1} = 0$. It follows from Euclid's algorithm that A_{11}^n is the greatest common divisor of a and b and that a and b are co-prime if $A_{11}^n = 1$. The equation (4.41) then has a solution if A_{11}^n is a factor of c.

System Theoretic Consequences

The following result is an immediate consequence of Euclid's algorithm, Theorem 4.1.

THEOREM 4.2—CONTROLLER PARAMETERIZATION
Consider a system with a rational transfer function $P = b/a$. Let $C_0 = g_0/f_0$ be a controller which gives a closed loop system with the characteristic polynomial c. Then all controllers which give a closed loop system with the characteristic polynomial c are given by

$$C = \frac{g_0 + qa}{f_0 - qb}$$

where q is an arbitrary polynomial.

PROOF 4.2
The loop transfer function obtained with the controller C is

$$L = PC = \frac{b(g_0 + qa)}{a(f_0 - qb)}$$

we have

$$1 + L = \frac{a(f_0 - qb) + b(g_0 + qa)}{a(f_0 - qb)} = \frac{af_0 + bg_0}{a(f_0 - qb)} = \frac{c}{a(f_0 - qb)}$$

which shows that the characteristic polynomial is c. Let $C = g/f$ be any controller that gives the characteristic polynomial c it follows that

$$af + bg = c$$

and it follows from Theorem 4.1 that $f = f_0 - bq$ and $g = g_0 + aq$. □

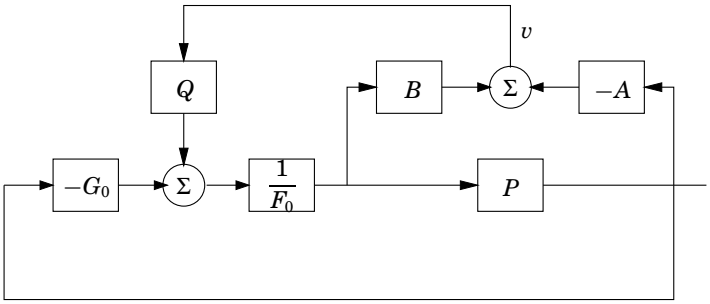


Figure 4.14 Block diagram that illustrates the Youla-Kučera parameterization theorem. If $C = G_0/F_0$ stabilizes the system $P = B/A$, then the controller shown in the block diagram also stabilizes the system for all stable rational functions Q.

This theorem is useful because it characterizes all controllers that give specified closed loop poles. Since the theorem tells that there are many solutions we may ask if there are some solutions that are particularly useful. It is natural to look for simple solutions. It follows from Theorem 4.2 that there is one controller where $\deg f < \deg b$, i.e. a controller of lowest order, and another where $\deg g < \deg a$, a controller with highest pole excess.

Youla-Kučera Parameterization

Theorem 4.2 characterizes all controllers that give a closed loop system with a given characteristic polynomial. We will now derive a related result that characterizes all stabilizing controllers. To start with we will introduce another representation of a transfer function.

DEFINITION 4.1—STABLE RATIONAL FUNCTIONS
Let $a(s)$ be a polynomial with all zeros in the left half plane and $b(s)$ an arbitrary polynomial. The rational function $b(s)/a(s)$ is called a stable rational function.

□

Stable rational functions are also a ring. This means that Theorem 4.1 also holds for rational functions. A fractional representation of a transfer function P is

$$P = \frac{B}{A}$$

where A and B are stable rational transfer functions. We have the following result.

THEOREM 4.3—YOULA-KUČERA REPRESENTATION
Consider a process with the transfer function $P = B/A$, where A and B are stable rational functions that are co-prime, let $C_0 = G_0/F_0$ be a fractional representation of a controller that stabilizes P, all stabilizing controllers are then given by

$$C = \frac{G_0 + QA}{F_0 - QB} \tag{4.43}$$

where Q is an arbitrary stable rational transfer function.

PROOF 4.3
The loop transfer function obtained with the controller C is

$$L = PC = \frac{B(G_0 + QA)}{A(F_0 - QB)}$$

we have

$$1 + L = \frac{A(F_0 - QB) + B(G_0 + QA)}{A(F_0 - QB)} = \frac{AF_0 + BG_0}{A(F_0 - QB)}$$

Since the rational function $AF_0 + BG_0$ has all its zeros in the left half plane the closed loop system is stable. Let $C = G/F$ be any controller that stabilizes the closed loop system it follows that

$$AF + BG = C$$

is a stable rational function with all its zeros in the left half plane. Hence

$$\frac{A}{C}F + \frac{B}{C}G = 1$$

and it follows from Theorem 4.1 that

$$F = F_0 - \frac{B}{C}Q = F_0 - B\bar{Q}$$
$$G = G_0 - \frac{A}{C}Q = G_0 - A\bar{Q}$$

where Q is a stable rational function because C has all its zeros in the left half plane. □

It follows from Equation (4.43) that the control law can be written as

$$\frac{U}{Y} = -\frac{G}{F} = -\frac{G_0 + QA}{F_0 - QB}$$

or

$$F_0 U = -G_0 Y + Q(BU - AY)$$

The Youla-Kučera parameterization theorem can then be illustrated by the block diagram in Figure 4.14. Notice that the signal v is zero. It therefore seems intuitively reasonable that a feedback based on this signal cannot make the system unstable.

4.7 Summary

In this section we started by investigating some simple control systems. A systematic method for analysis and design was developed. The closed loop system was first represented by a block diagram. The behavior of each block was represented by a transfer function. The relations between the Laplace transforms of all signals could be derived by simple algebraic manipulations of the transfer functions of the blocks. An interesting feature of using Laplace transforms is that systems and signals are represented in the same way. The analysis gave good insight into the behavior of simple control systems and how its properties were influenced by the poles and zeros of the closed loop system. The results can also be developed using differential equations but it is much simpler to use Laplace transforms and transfer functions. This is also the standard language of the field of control.

To design a controller we selected a controller with given structure, PI or PID. The parameters of the controller were then chosen to obtain a closed loop system with specified poles, or equivalently specified roots of the characteristic equation. This design method was called pole placement. The design methods were worked out in detail for first and second order systems but we also briefly discussed the general case. To find suitable closed loop poles we found that it was convenient to introduce standard parameters to describe the closed loop poles. Results that guide the intuition of choosing the closed loop poles were also developed.

The analysis was based on simplified models of the dynamics of the process. The example on cruise control in Section 4.2 indicated that it was not necessary to know some parameters accurately. One of the amazing properties of control systems is that they can often be designed based on simple models. This will be justified in the next chapter.

5

Feedback Fundamentals

5.1 Introduction

Fundamental properties of feedback systems will be investigated in this Chapter. We begin in Section 5.2 by discussing the basic feedback loop and typical requirements. This includes the ability to follow reference signals, effects of load disturbances and measurement noise and the effects of process variations. It turns out that these properties can be captured by a set of six transfer functions, called the Gang of Six. These transfer functions are introduced in Section 5.3. For systems where the feedback is restricted to operate on the error signal the properties are characterized by a subset of four transfer functions, called the Gang of Four. Properties of systems with error feedback and the more general feedback configuration with two degrees of freedom are also discussed in Section 5.3. It is shown that it is important to consider all transfer functions of the Gang of Six when evaluating a control system. Another interesting observation is that for systems with two degrees of freedom the problem of response to load disturbances can be treated separately. This gives a natural separation of the design problem into a design of a feedback and a feedforward system. The feedback handles process uncertainties and disturbances and the feedforward gives the desired response to reference signals.

Attenuation of disturbances are discussed in Section 5.4 where it is demonstrated that process disturbances can be attenuated by feedback but that feedback also feeds measurement noise into the system. It turns out that the sensitivity function which belongs to the Gang of Four gives a nice characterization of disturbance attenuation. The effects of process variations are discussed in Section 5.5. It is shown that their effects are well described by the sensitivity function and the complementary sensitivity function. The analysis also gives a good explanation for the fact that

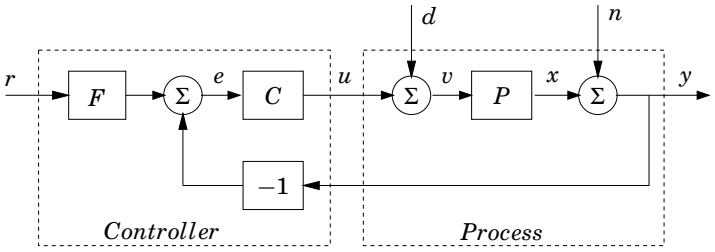


Figure 5.1 Block diagram of a basic feedback loop.

control systems can be designed based on simplified models. When discussing process variations it is natural to investigate when two processes are similar from the point of view of control. This important nontrivial problem is discussed in Section 5.6. Section 5.7 is devoted to a detailed treatment of the sensitivity functions. This leads to a deeper understanding of attenuation of disturbances and effects of process variations. A fundamental result of Bode which gives insight into fundamental limitations of feedback is also derived. This result shows that disturbances of some frequencies can be attenuated only if disturbances of other frequencies are amplified. Tracking of reference signals are investigated in Section 5.8. Particular emphasis is given to precise tracking of low frequency signals. Because of the richness of control systems the emphasis on different issues varies from field to field. This is illustrated in Section 5.10 where we discuss the classical problem of design of feedback amplifiers.

5.2 The Basic Feedback Loop

A block diagram of a basic feedback loop is shown in Figure 5.1. The system loop is composed of two components, the process P and the controller. The controller has two blocks the feedback block C and the feedforward block F. There are two disturbances acting on the process, the load disturbance d and the measurement noise n. The load disturbance represents disturbances that drive the process away from its desired behavior. The process variable x is the real physical variable that we want to control. Control is based on the measured signal y, where the measurements are corrupted by measurement noise n. Information about the process variable x is thus distorted by the measurement noise. The process is influenced by the controller via the control variable u. The process is thus a system with three inputs and one output. The inputs are: the control variable

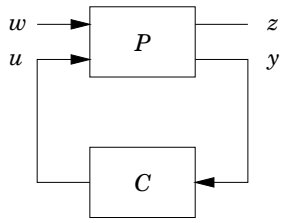


Figure 5.2 An abstract representation of the system in Figure 5.1. The input u represents the control signal and the input w represents the reference r, the load disturbance d and the measurement noise n. The output y is the measured variables and z are internal variables that are of interest.

u, the load disturbance d and the measurement noise n. The output is the measured signal. The controller is a system with two inputs and one output. The inputs are the measured signal y and the reference signal r and the output is the control signal u. Note that the control signal u is an input to the process and the output of the controller and that the measured signal is the output of the process and an input to the controller. In Figure 5.1 the load disturbance was assumed to act on the process input. This is a simplification, in reality the disturbance can enter the process in many different ways. To avoid making the presentation unnecessarily complicated we will use the simple representation in Figure 5.1. This captures the essence and it can easily be modified if it is known precisely how disturbances enter the system.

More Abstract Representations

The block diagrams themselves are substantial abstractions but higher abstractions are sometimes useful. The system in Figure 5.1 can be represented by only two blocks as shown in Figure 5.2. There are two types of inputs, the control u, which can be manipulated and the disturbances $w = (r, d, n)$, which represents external influences on the closed loop systems. The outputs are also of two types the measured signal y and other interesting signals $z = (e, v, x)$. The representation in Figure 5.2 allows many control variables and many measured variables, but it shows less of the system structure than Figure 5.1. This representation can be used even when there are many input signals and many output signals. Representation with a higher level of abstraction are useful for the development of theory because they make it possible to focus on fundamentals and to solve general problems with a wide range of applications. Care must, however, be exercised to maintain the coupling to the real world control problems we intend to solve.

Disturbances

Attenuation of load disturbances is often a primary goal for control. This is particularly the case when controlling processes that run in steady state. Load disturbances are typically dominated by low frequencies. Consider for example the cruise control system for a car, where the disturbances are the gravity forces caused by changes of the slope of the road. These disturbances vary slowly because the slope changes slowly when you drive along a road. Step signals or ramp signals are commonly used as prototypes for load disturbances disturbances.

Measurement noise corrupts the information about the process variable that the sensors delivers. Measurement noise typically has high frequencies. The average value of the noise is typically zero. If this was not the case the sensor will give very misleading information about the process and it would not be possible to control it well. There may also be dynamics in the sensor. Several sensors are often used. A common situation is that very accurate values may be obtained with sensors with slow dynamics and that rapid but less accurate information can be obtained from other sensors.

Actuation

The process is influenced by actuators which typically are valves, motors, that are driven electrically, pneumatically, or hydraulically. There are often local feedback loops and the control signals can also be the reference variables for these loops. A typical case is a flow loop where a valve is controlled by measuring the flow. If the feedback loop for controlling the flow is fast we can consider the set point of this loop which is the flow as the control variable. In such cases the use of local feedback loops can thus simplify the system significantly. When the dynamics of the actuators is significant it is convenient to lump them with the dynamics of the process. There are cases where the dynamics of the actuator dominates process dynamics.

Design Issues

Many issues have to be considered in analysis and design of control systems. Basic requirements are

- Stability
- Ability to follow reference signals
- Reduction of effects of load disturbances
- Reduction of effects of measurement noise
- Reduction of effects of model uncertainties

The possibility of instabilities is the primary drawback of feedback. Avoiding instability is thus a primary goal. It is also desirable that the process variable follows the reference signal faithfully. The system should also be able to reduce the effect of load disturbances. Measurement noise is injected into the system by the feedback. This is unavoidable but it is essential that not too much noise is injected. It must also be considered that the models used to design the control systems are inaccurate. The properties of the process may also change. The control system should be able to cope with moderate changes. The focus on different abilities vary with the application. In process control the major emphasis is often on attenuation of load disturbances, while the ability to follow reference signals is the primary concern in motion control systems.

5.3 The Gang of Six

The feedback loop in Figure 5.1 is influenced by three external signals, the reference r, the load disturbance d and the measurement noise n. There are at least three signals x, y and u that are of great interest for control. This means that there are nine relations between the input and the output signals. Since the system is linear these relations can be expressed in terms of the transfer functions. Let X, Y, U, D, N R be the Laplace transforms of x, y, u, d, n r, respectively. The following relations are obtained from the block diagram in Figure 5.1

$$\begin{aligned} X &= \frac{P}{1+PC}D - \frac{PC}{1+PC}N + \frac{PCF}{1+PC}R \\ Y &= \frac{P}{1+PC}D + \frac{1}{1+PC}N + \frac{PCF}{1+PC}R \\ U &= -\frac{PC}{1+PC}D - \frac{C}{1+PC}N + \frac{CF}{1+PC}R. \end{aligned} \tag{5.1}$$

To simplify notations we have dropped the arguments of all Laplace transforms. There are several interesting conclusions we can draw from these equations. First we can observe that several transfer functions are the same and that all relations are given by the following set of six transfer functions which we call the Gang of Six.

$$\begin{matrix} \frac{PCF}{1+PC} & \frac{PC}{1+PC} & \frac{P}{1+PC} \\ \frac{CF}{1+PC} & \frac{C}{1+PC} & \frac{1}{1+PC}, \end{matrix} \tag{5.2}$$

The transfer functions in the first column give the response of process variable and control signal to the set point. The second column gives the same signals in the case of pure error feedback when $F = 1$. The transfer function $P/(1 + PC)$ in the third column tells how the process variable reacts to load disturbances the transfer function $C/(1 + PC)$ gives the response of the control signal to measurement noise.

Notice that only four transfer functions are required to describe how the system reacts to load disturbance and the measurement noise and that two additional transfer functions are required to describe how the system responds to set point changes.

The special case when $F = 1$ is called a system with (pure) error feedback. In this case all control actions are based on feedback from the error only. In this case the system is completely characterized by four transfer functions, namely the four rightmost transfer functions in (5.2), i.e.

$$\begin{aligned} &\frac{PC}{1+PC}, && \text{the complementary sensitivity function} \\ &\frac{P}{1+PC}, && \text{the load disturbance sensitivity function} \\ &\frac{C}{1+PC}, && \text{the noise sensitivity function} \\ &\frac{1}{1+PC}, && \text{the sensitivity function} \end{aligned} \tag{5.3}$$

These transfer functions and their equivalent systems are called the Gang of Four. The transfer functions have many interesting properties that will be discussed in then following. A good insight into these properties are essential for understanding feedback systems. The load disturbance sensitivity function is sometimes called the input sensitivity function and the noise sensitivity function is sometimes called the output sensitivity function.

Systems with Two Degrees of Freedom

The controller in Figure 5.1 is said to have two degrees of freedom because the controller has two blocks, the feedback block C which is part of the closed loop and the feedforward block F which is outside the loop. Using such a controller gives a very nice separation of the control problem because the feedback controller can be designed to deal with disturbances and process uncertainties and the feedforward will handle the response to reference signals. Design of the feedback only considers the gang of four and the feedforward deals with the two remaining transfer functions in the gang of six. For a system with error feedback it is necessary to make a compromise. The controller C thus has to deal with all aspects of the

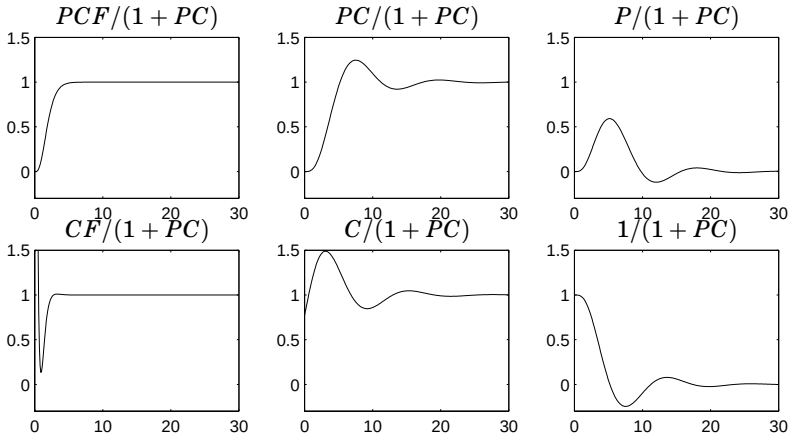


Figure 5.3 Step responses of the Gang of Six for PI control $k = 0.775$, $T_i = 2.05$ of the process $P(s) = (s+1)^{-4}$. The feedforward is designed to give the transfer function $(0.5s+1)^{-4}$ from reference r to output y.

problem.

To describe the system properly it is thus necessary to show the response of all six transfer functions. The transfer functions can be represented in different ways, by their step responses and frequency responses, see Figures 5.3 and 5.4.

Figures 5.3 and 5.4 give useful insight into the properties of the closed loop system. The time responses in Figure 5.3 show that the feedforward gives a substantial improvement of the response speed. The settling time is substantially shorter, 4 s versus 25 s, and there is no overshoot. This is also reflected in the frequency responses in Figure 5.4 which shows that the transfer function with feedforward has higher bandwidth and that it has no resonance peak.

The transfer functions $CF/(1+PC)$ and $-C/(1+PC)$ represent the signal transmission from reference to control and from measurement noise to control. The time responses in Figure 5.3 show that the reduction in response time by feedforward requires a substantial control effort. The initial value of the control signal is out of scale in Figure 5.3 but the frequency response in 5.4 shows that the high frequency gain of $PCF/(1+PC)$ is 16, which can be compared with the value 0.78 for the transfer function $C/(1+PC)$. The fast response thus requires significantly larger control signals.

There are many other interesting conclusions that can be drawn from Figures 5.3 and 5.4. Consider for example the response of the output to

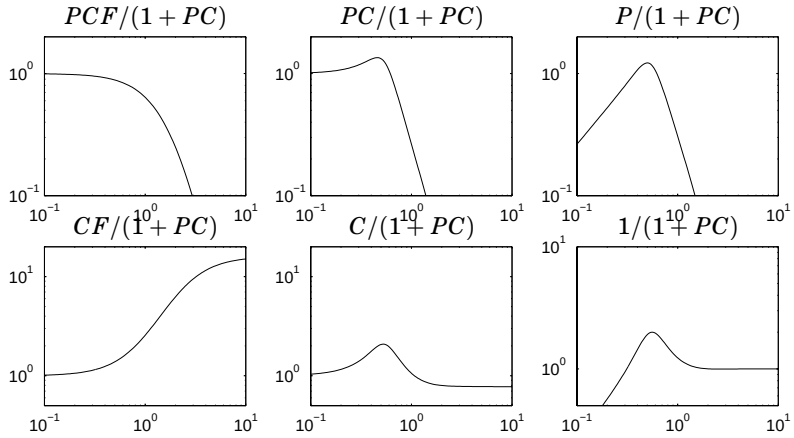


Figure 5.4 Gain curves of frequency responses of the Gang of Six for PI control $k = 0.775$, $T_i = 2.05$ of the process $P(s) = (s+1)^{-4}$ where the feedforward has been designed to give the transfer function $(0.5s+1)^{-4}$ from reference to output.

load disturbances expressed by the transfer function $P/(1 + PC)$. The frequency response has a pronounced peak 1.22 at $\omega_{max} = 0.5$ the corresponding time function has its maximum 0.59 at $t_{max} = 5.2$. Notice that the peaks are of the same magnitude and that the product of $\omega_{max} t_{max} = 2.6$.

The step responses can also be represented by two simulations of the process. The complete system is first simulated with the full two-degree-of-freedom structure. The simulation begins with a step in the reference signal, when the system has settled to equilibrium a step in the load disturbance is then given. The process output and the control signals are recorded. The simulation is then repeated with a system without feedforward, i.e. $F = 1$. The response to the reference signal will be different but the response to the load disturbance will be the same as in the first simulation. The procedure is illustrated in Figure 5.5.

A Remark

The fact that 6 relations are required to capture properties of the basic feedback loop is often neglected in literature. Most papers on control only show the response of the process variable to set point changes. Such a curve gives only partial information about the behavior of the system. To get a more complete representation of the system all six responses should be given. We illustrate the importance of this by an example.

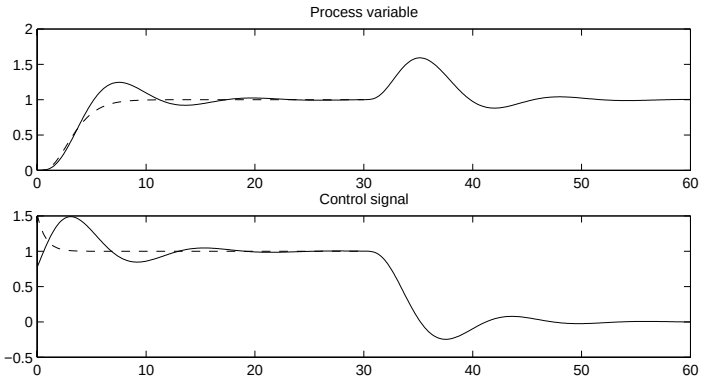


Figure 5.5 Representation of properties of a basic feedback loop by step responses in the reference at time 0, and at the process input at time 30. The dashed full lines show the response for a system with error feedback $F = 1$, and the dashed lines show responses for a system having two degrees of freedom.

EXAMPLE 5.1—ASSESSMENT OF A CONTROL SYSTEM
A process with the transfer function

$$P(s) = \frac{1}{(s+1)(s+0.02)}$$

is controlled using error feedback with a controller having the transfer function

$$C(s) = \frac{50s+1}{50s}$$

The loop transfer function is

$$L(s) = \frac{1}{s(s+1)}$$

Figure 5.6 shows that the responses to a reference signal look quite reasonable. Based on these responses we could be tempted to conclude that the closed loop system is well designed. The step response settles in about 10 s and the overshoot is moderate.

To explore the system further we will calculate the transfer functions of the Gang of Six, we have

$$\frac{P(s)C(s)}{1+P(s)C(s)} = \frac{1}{s^2+s+1} \qquad \frac{P(s)}{1+P(s)C(s)} = \frac{s}{(s+0.02)(s^2+s+1)}$$

$$\frac{C(s)}{1+P(s)C(s)} = \frac{(s+0.02)(s+1)}{s^2+s+1} \qquad \frac{1}{1+P(s)C(s)} = \frac{s(s+1)}{s^2+s+1}$$

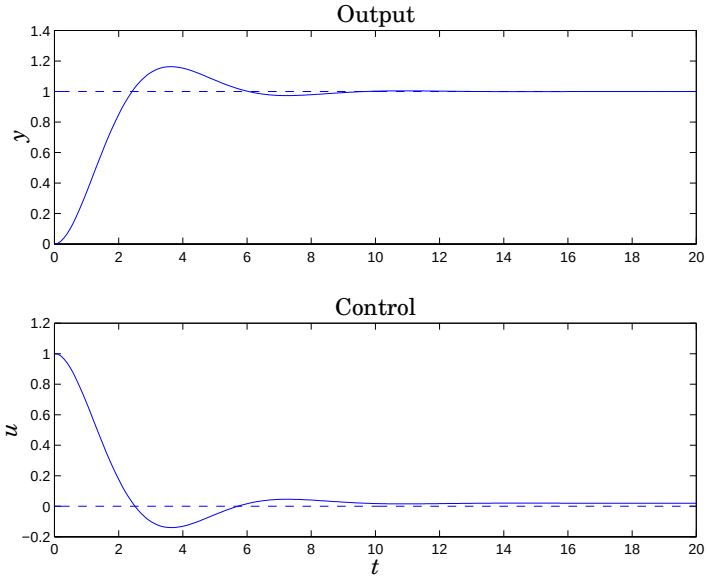


Figure 5.6 Response of output y and control u to a step in reference r.

The responses of y and u to the reference r are given by

$$Y(s) = \frac{1}{s^2 + s + 1} R(s), \quad U(s) = \frac{(s+1)(s+0.02)}{s^2 + s + 1} R(s)$$

and the responses of y and u to the load disturbance d are given by

$$Y(s) = \frac{s}{(s+0.02)(s^2 + s + 1)} D(s), \quad U(s) = -\frac{1}{s^2 + s + 1} D(s)$$

Notice that the process pole $s = 0.02$ is cancelled by a controller zero. This implies that the loop transfer function is of second order even if the closed loop system itself is of third order. The characteristic equation of the closed loop system is

$$(s + 0.02)(s^2 + s + 1) = 0$$

where the the pole $s = -0.02$ corresponds the process pole that is canceled by the controller zero. The presence of the slow pole $s = -0.02$ which appears in the response to load disturbances implies that the output decays very slowly, at the rate of $e^{-0.02t}$. The controller will not respond to the

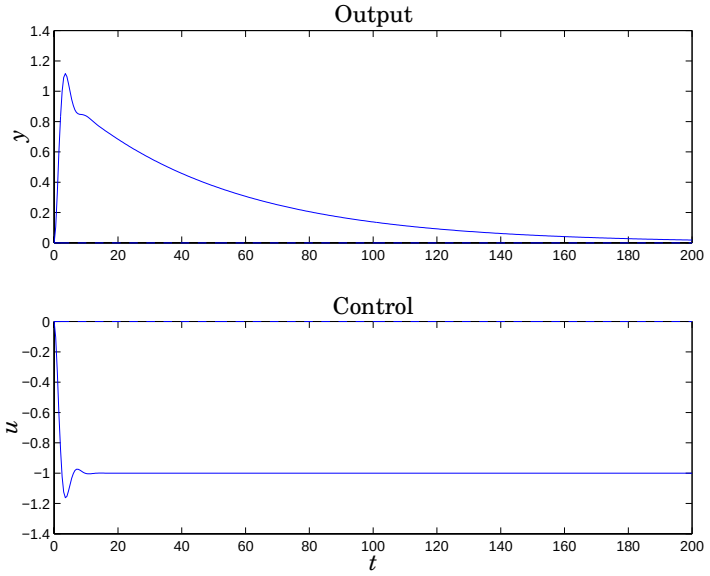


Figure 5.7 Response of output y and control u to a step in the load disturbance. Notice the very slow decay of the mode $e^{-0.02t}$. The control signal does not respond to this mode because the controller has a zero $s = -0.02$.

signal $e^{-0.02t}$ because the zero $s = -0.02$ will block the transmission of this signal. This is clearly seen in Figure 5.7, which shows the response of the output and the control signals to a step change in the load disturbance. Notice that it takes about 200 s for the disturbance to settle. This can be compared with the step response in Figure 5.6 which settles in about 10s. □

The behavior illustrated in the example is typical when there are cancellations of poles and zeros in the transfer functions of the process and the controller. The canceled factors do not appear in the loop transfer function and the sensitivity functions. The canceled modes are not visible unless they are excited. The effects are even more drastic than shown in the example if the canceled modes are unstable. This has been known among control engineers for a long time and a there has been a design rule that cancellation of slow or unstable modes should be avoided. Another view of cancellations is given in Section 3.7.

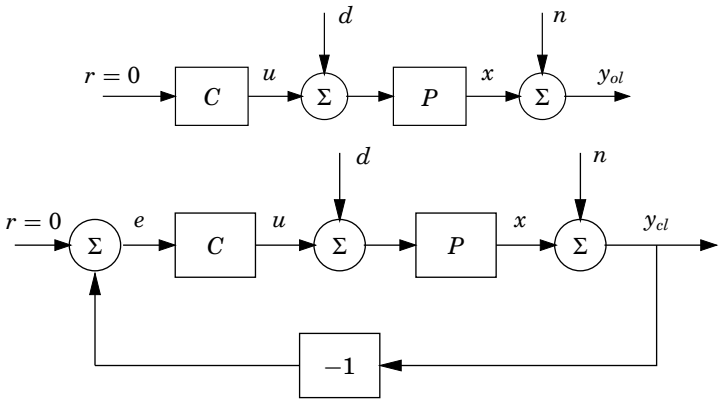


Figure 5.8 Open and closed loop systems subject to the same disturbances.

5.4 Disturbance Attenuation

The attenuation of disturbances will now be discussed. For that purpose we will compare an open loop system and a closed loop system subject to the disturbances as is illustrated in Figure 5.8. Let the transfer function of the process be $P(s)$ and let the Laplace transforms of the load disturbance and the measurement noise be $D(s)$ and $N(s)$ respectively. The output of the open loop system is

$$Y_{ol} = P(s)D(s) + N(s) \tag{5.4}$$

and the output of the closed loop system is

$$Y_{cl} = \frac{P(s)D(s) + N(s)}{1 + P(s)C(s)} = S(s)(P(s)D(s) + N(s)) \tag{5.5}$$

where $S(s)$ is the sensitivity function, which belongs to the Gang of Four. We thus obtain the following interesting result

$$Y_{cl}(s) = S(s)Y_{ol}(s) \tag{5.6}$$

The sensitivity function will thus directly show the effect of feedback on the output. The disturbance attenuation can be visualized graphically by the gain curve of the Bode plot of $S(s)$. The lowest frequency where the sensitivity function has the magnitude 1 is called the sensitivity crossover frequency and denoted by ω_{sc}. The maximum sensitivity

$$M_s = \max_{\omega} |S(i\omega)| = \max_{\omega} \left| \frac{1}{1 + P(i\omega)C(i\omega)} \right| \tag{5.7}$$

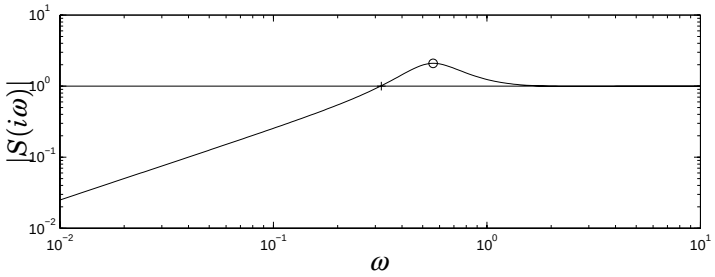


Figure 5.9 Gain curve of the sensitivity function for PI control ($k = 0.8$, $k_i = 0.4$) of process with the transfer function $P(s) = (s + 1)^{-4}$. The sensitivity crossover frequency is indicated by + and the maximum sensitivity by *o*.

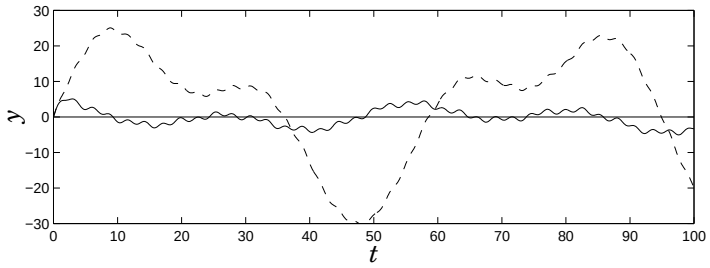


Figure 5.10 Outputs of process with control (full line) and without control (dashed line).

is an important variable which gives the largest amplification of the disturbances. The maximum occurs at the frequency ω_{ms}.

A quick overview of how disturbances are influenced by feedback is obtained from the gain curve of the Bode plot of the sensitivity function. An example is given in Figure 5.9. The figure shows that the sensitivity crossover frequency is 0.32 and that the maximum sensitivity 2.1 occurs at $\omega_{ms} = 0.56$. Feedback will thus reduce disturbances with frequencies less than 0.32 rad/s, but it will amplify disturbances with higher frequencies. The largest amplification is 2.1.

If a record of the disturbance is available and a controller has been designed the output obtained under closed loop with the same disturbance can be visualized by sending the recorded output through a filter with the transfer function $S(s)$. Figure 5.10 shows the output of the system with and without control.

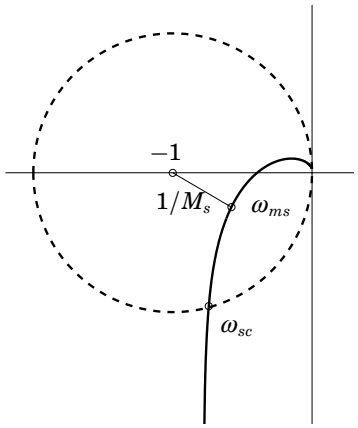


Figure 5.11 Nyquist curve of loop transfer function showing graphical interpretation of maximum sensitivity. The sensitivity crossover frequency ω_{sc} and the frequency ω_{ms} where the sensitivity has its largest value are indicated in the figure. All points inside the dashed circle have sensitivities greater than 1.

The sensitivity function can be written as

$$S(s) = \frac{1}{1 + P(s)C(s)} = \frac{1}{1 + L(s)}. \tag{5.8}$$

Since it only depends on the loop transfer function it can be visualized graphically in the Nyquist plot of the loop transfer function. This is illustrated in Figure 5.11. The complex number $1 + L(i\omega)$ can be represented as the vector from the point -1 to the point $L(i\omega)$ on the Nyquist curve. The sensitivity is thus less than one for all points outside a circle with radius 1 and center at -1. Disturbances of these frequencies are attenuated by the feedback. If a control system has been designed based on a given model it is straight forward to estimated the potential disturbance reduction simply by recording a typical output and filtering it through the sensitivity function.

Slow Load Disturbances

Load disturbances typically have low frequencies. To estimate their effects on the process variable it is then natural to approximate the transfer function from load disturbances to process output for small s, i.e.

$$G_{xd}(s) = \frac{P(s)}{1 + P(s)C(s)} \approx c_0 + c_1 s + c_2 s^2 + \cdots \tag{5.9}$$

The coefficients c_k are called stiffness coefficients. This means that the process variable for slowly varying load disturbances d is given by

$$x(t) = c_0 d(t) + c_1 \frac{dd(t)}{dt} + c_2 \frac{d^2 d(t)}{dt^2} + \cdots$$

For example if the load disturbance is $d(t) = v_0 t$ we get

$$x(t) = c_0 v_0 t + c_1 v_0$$

If the controller has integral action we have $c_0 = 0$ and $x(t) = c_1 v_0$.

5.5 Process Variations

Control systems are designed based on simplified models of the processes. Process dynamics will often change during operation. The sensitivity of a closed loop system to variations in process dynamics is therefore a fundamental issue.

Risk for Instability

Instability is the main drawback of feedback. It is therefore of interest to investigate if process variations can cause instability. The sensitivity functions give a useful insight. Figure 5.11 shows that the largest sensitivity is the inverse of the shortest distance from the point -1 to the Nyquist curve.

The complementary sensitivity function also gives insight into allowable process variations. Consider a feedback system with a process P and a controller C. We will investigate how much the process can be perturbed without causing instability. The Nyquist curve of the loop transfer function is shown in Figure 5.12. If the process is changed from P to $P + \Delta P$ the loop transfer function changes from PC to $PC + C\Delta P$ as illustrated in the figure. The distance from the critical point -1 to the point L is $|1 + L|$. This means that the perturbed Nyquist curve will not reach the critical point -1 provided that

$$|C\Delta P| < |1 + L|$$

This condition must be valid for all points on the Nyquist curve. The condition for stability can be written as

$$\frac{|\Delta P(i\omega)|}{|P(i\omega)|} < \frac{1}{|T(i\omega)|} \tag{5.10}$$

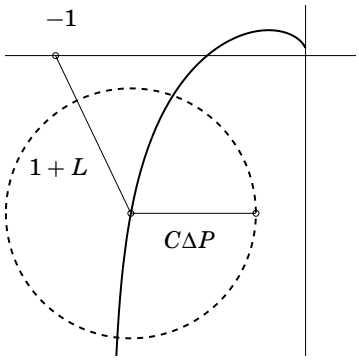


Figure 5.12 Nyquist curve of a nominal loop transfer function and its uncertainty caused by process variations ΔP.

A technical condition, namely that the perturbation ΔP is a stable transfer function, must also be required. If this does not hold the encirclement condition required by Nyquist's stability condition is not satisfied. Also notice that the condition (5.10) is conservative because it follows from Figure 5.12 that the critical perturbation is in the direction towards the critical point -1. Larger perturbations can be permitted in the other directions.

This formula (5.10) is one of the reasons why feedback systems work so well in practice. The mathematical models used to design control system are often strongly simplified. There may be model errors and the properties of a process may change during operation. Equation (5.10) implies that the closed loop system will at least be stable for substantial variations in the process dynamics.

It follows from (5.10) that the variations can be large for those frequencies where T is small and that smaller variations are allowed for frequencies where T is large. A conservative estimate of permissible process variations that will not cause instability is given by

$$\frac{|\Delta P(i\omega)|}{|P(i\omega)|} < \frac{1}{M_t}$$

where M_t is the largest value of the complementary sensitivity

$$M_t = \max_{\omega} |T(i\omega)| = \max_{\omega} \left| \frac{P(i\omega)C(i\omega)}{1 + P(i\omega)C(i\omega)} \right| \tag{5.11}$$

The value of M_t is influenced by the design of the controller. For example if $M_t = 2$ gain variations of 50% and phase variations of 30° are permitted

without making the closed loop system unstable. The fact that the closed loop system is robust to process variations is one of the reason why control has been so successful and that control systems for complex processes can indeed be designed using simple models. This is illustrated by an example.

EXAMPLE 5.2—MODEL UNCERTAINTY
Consider a process with the transfer function

$$P(s) = \frac{1}{(s+1)^4}$$

A PI controller with the parameters $k = 0.775$ and $T_i = 2.05$ gives a closed loop system with $M_s = 2.00$ and $M_t = 1.35$. The complementary sensitivity has its maximum for $\omega_{mt} = 0.46$. Figure 5.13 shows the Nyquist curve of the transfer function of the process and the uncertainty bounds $\Delta P = |P|/|T|$ for a few frequencies. The figure shows that

- Large uncertainties are permitted for low frequencies, $T(0) = 1$.
- The smallest relative error $|\Delta P/P|$ occurs for $\omega = 0.46$.
- For $\omega = 1$ we have $|T(i\omega)| = 0.26$ which means that the stability requirement is $|\Delta P/P| < 3.8$
- For $\omega = 2$ we have $|T(i\omega)| = 0.032$ which means that the stability requirement is $|\Delta P/P| < 31$

□

The situation illustrated in the figure is typical for many processes, moderately small uncertainties are only required around the gain crossover frequencies, but large uncertainties can be permitted at higher and lower frequencies. A consequence of this is also that a simple model that describes the process dynamics well around the crossover frequency is sufficient for design. Systems with many resonance peaks are an exception to this rule because the process transfer function for such systems may have large gains also for higher frequencies.

Variations in Closed Loop Transfer Function

So far we have investigated the risk for instability. The effects of small variation in process dynamics on the closed loop transfer function will now be investigated. To do this we will analyze the system in Figure 5.1. For simplicity we will assume that $F = 1$ and that the disturbances d and n are zero. The transfer function from reference to output is given by

$$\frac{Y}{R} = \frac{PC}{1+PC} = T \tag{5.12}$$

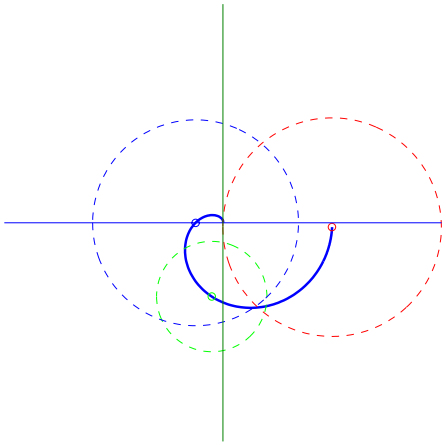


Figure 5.13 Nyquist curve of a nominal process transfer function $P(s) = (s+1)^{-4}$ shown in full lines. The circles show the uncertainty regions $|\Delta P| = 1/|T|$ obtained for a PI controller with $k = 0.775$ and $T_i = 2.05$ for $\omega = 0$, 0.46 and 1.

Compare with (5.2). The transfer function T which belongs to the Gang of Four is called the complementary sensitivity function. Differentiating (5.12) we get

$$\frac{dT}{dP} = \frac{C}{(1+PC)^2} = \frac{PC}{(1+PC)(1+PC)P} = S\frac{T}{P}$$

Hence

$$\frac{d\log T}{d\log P} = \frac{dT}{dP}\frac{P}{T} = S \tag{5.13}$$

This equation is the reason for calling S the sensitivity function. The relative error in the closed loop transfer function T will thus be small if the sensitivity is small. This is one of the very useful properties of feedback. For example this property was exploited by Black at Bell labs to build the feedback amplifiers that made it possible to use telephones over large distances.

A small value of the sensitivity function thus means that disturbances are attenuated and that the effect of process perturbations also are negligible. A plot of the magnitude of the complementary sensitivity function as in Figure 5.9 is a good way to determine the frequencies where model precision is essential.

Constraints on Design

Constraints on the maximum sensitivities M_s and M_t are important to ensure that closed loop system is insensitive to process variations. Typical constraints are that the sensitivities are in the range of 1.1 to 2. This has implications for design of control systems which are illustrated by an example.

EXAMPLE 5.3—SENSITIVITIES CONSTRAIN CLOSED LOOP POLES

PI control of a first order system was discussed in Section 4.4 where it was shown that the closed loop system was of second order and that the closed loop poles could be placed arbitrarily by proper choice of the controller parameters. The process and the controller are characterized by

$$Y(s) = \frac{b}{s+a}U(s)$$
$$U(s) = -kY(s) + \frac{k_i}{s}(R(s) - Y(s))$$

where U, Y and R are the Laplace transforms of the process input, output and the reference signal. The closed loop characteristic polynomial is

$$s^2 + (a + bk)s + bk_i$$

requiring this to be equal to

$$s^2 + 2\zeta\omega_0 s + \omega_0^2$$

we find that the controller parameters are given by

$$k = \frac{2\zeta\omega_0 - 1}{b}$$
$$k_i = \frac{\omega_0^2}{b}$$

and there are no apparent constraints on the choice of parameters ζ and ω_0. Calculating the sensitivity functions we get

$$S(s) = \frac{s(s+a)}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$
$$T(s) = \frac{(2\zeta\omega_0 - a)s + \omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

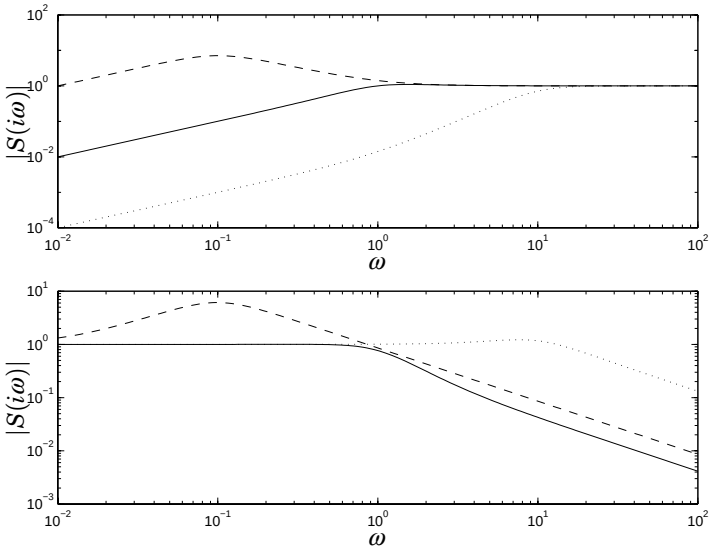


Figure 5.14 Magnitude curve for bode plots of the sensitivity function (above) and the complementary sensitivity function (below) for $\zeta = 0.7$, $a = 1$ and $\omega_0/a = 0.1$ (dashed), 1 (solid) and 10 (dotted).

Figure 5.14 shows clearly that the sensitivities will be large if the parameter ω_0 is chosen smaller than a. The equation for controller gain also gives an indication that small values of ω_0 are not desirable because proportional gain then becomes negative which means that the feedback is positive. □

Sensitivities and Relative Damping

For simple low order control systems we have based design criteria on the patterns of the poles and zeros of the complementary transfer function. To relate the general results on robustness to the analysis of the simple controllers it is of interest to find the relations between the sensitivities and relative damping. The complementary sensitivity function for a standard second order system is given by

$$T(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

This implies that the sensitivity function is given by

$$S(s) = 1 - T(s) = \frac{s(s + 2\zeta\omega_0)}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

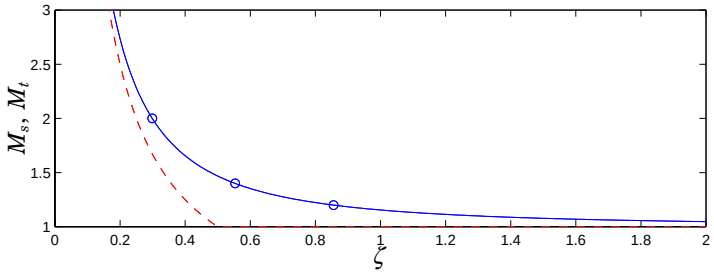


Figure 5.15 Maximum sensitivities M_s (full line) and M_t (dashed line) as functions of relative damping for $T(s) = \frac{\omega_0^2}{s^2+2\zeta\omega_0 s+\omega_0^2}$ and $S(s) = \frac{s(s+2\zeta\omega_0)}{s^2+2\zeta\omega_0 s+\omega_0^2}$.

Straight forward but tedious calculations give.

$$
\begin{aligned}
M_s &= \sqrt{\frac{8\zeta^2+1+(4\zeta^2+1)\sqrt{8\zeta^2+1}}{8\zeta^2+1+(4\zeta^2-1)\sqrt{8\zeta^2+1}}} \\
w_{ms} &= \frac{1+\sqrt{8\zeta^2+1}}{2}\omega_0 \\
M_t &= \begin{cases} 1/(2\zeta\sqrt{1-\zeta^2}) & \text{if } \zeta \le \sqrt{2}/2 \\ 1 & \text{if } \zeta > \sqrt{2}/2 \end{cases} \\
\omega_{mt} &= \begin{cases} \omega_0\sqrt{1-2\zeta^2} & \text{if } \zeta \le \sqrt{2}/2 \\ 0 & \text{if } \zeta > \sqrt{2}/2 \end{cases}
\end{aligned}
\tag{5.14}
$$

The relation between the sensitivities and relative damping are shown in Figure 5.15. The values $\zeta = 0.3$, 0.5 and 0.7 correspond to the maximum sensitivities $M_s = 1.99$, 1.47 and 1.28 respectively.

5.6 When are Two Processes Similar?

A fundamental issue is to determine when two processes are close. This seemingly innocent problem is not as simple as it may appear. When discussing the effects of uncertainty of the process on stability in Section 5.5 we used the quantity

$$
\delta(P_1, P_2) = \max_{\omega} |P_1(i\omega) - P_2(i\omega)| \tag{5.15}
$$

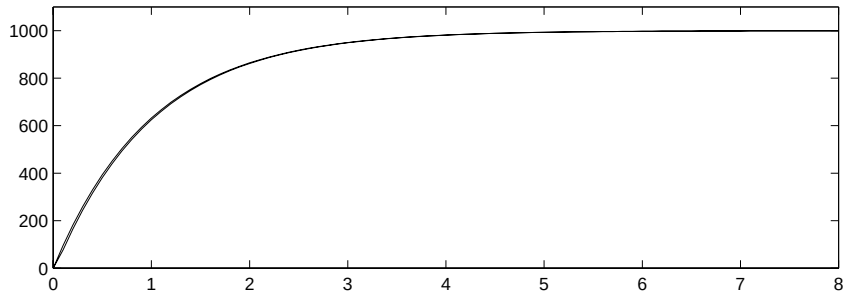


Figure 5.16 Step responses for systems with the transfer functions $P_1(s) = 1000/(s+1)$ and $P_2(s) = 10^7/((s+1)(s+100)^2)$.

as a measure of closeness of two processes. In addition the transfer functions P_1 and P_2 were assumed to be stable. This means conceptually that we compare the outputs of two systems subject to the same input. This may appear as a natural way to compare two systems but there are complications. Two systems that have similar open loop behaviors may have drastically different behavior in closed loop and systems with very different open loop behavior may have similar closed loop behavior. We illustrate this by two examples.

EXAMPLE 5.4—SIMILAR IN OPEN LOOP BUT DIFFERENT IN CLOSED LOOP
Systems with the transfer functions

$$P_1(s) = \frac{1000}{s+1}, \qquad P_2(s) = \frac{1000a^2}{(s+1)(s+a)^2}$$

have very similar open loop responses for large values of a. This is illustrated in Figure 5.16 which shows the step responses of for $a = 100$. The differences between the step responses are barely noticeable in the figure. The transfer functions from reference values to output for closed loop systems obtained with error feedback with $C = 1$ are

$$T_1 = \frac{1000}{s+1001}, \qquad T_2 = \frac{10^7}{(s-287)(s^2+86s+34879)}$$

The closed loop systems are very different because the system T_1 is stable and T_2 is unstable. □

EXAMPLE 5.5—DIFFERENT IN OPEN LOOP BUT SIMILAR IN CLOSED LOOP
Systems with the transfer functions

$$P_1(s) = \frac{1000}{s+1}, \qquad P_2(s) = \frac{1000}{s-1}$$

have very different open loop properties because one system is unstable and the other is stable. The transfer functions from reference values to output for closed loop systems obtained with error feedback with $C = 1$ are

$$T_1(s) = \frac{1000}{s+1001} \qquad T_2(s) = \frac{1000}{s+999}$$

which are very close. □

These examples show clearly that to compare two systems by investigating their open loop properties may be strongly misleading from the point of view of feedback control. Inspired by the examples we will instead compare the properties of the closed loop systems obtained when two processes P_1 and P_2 are controlled by the same controller C. To do this it will be assumed that the closed loop systems obtained are stable. The difference between the closed loop transfer functions is

$$\delta(P_1, P_2) = \Big| \frac{P_1 C}{1 + P_1 C} - \frac{P_2 C}{1 + P_2 C} \Big| = \Big| \frac{(P_1 - P_2)C}{(1 + P_1 C)(1 + P_2 C)} \Big| \tag{5.16}$$

This is a natural way to express the closeness of the systems P_1 and P_2, when they are controlled by C. It can be verified that δ is a proper norm in the mathematical sense. There is one difficulty from a practical point of view because the norm depends on the feedback C. The norm has some interesting properties.

Assume that the controller C has high gain at low frequencies. For low frequencies we have

$$\delta(P_1, P_2) \approx \frac{P_1 - P_2}{P_1 P_2 C}$$

If C is large it means that δ can be small even if the difference $P_1 - P_2$ is large. For frequencies where the maximum sensitivity is large we have

$$\delta(P_1, P_2) \approx M_{s1} M_{s2} |C(P_1 - P_2)|$$

For frequencies where P_1 and P_2 have small gains, typically for high frequencies, we have

$$\delta(P_1, P_2) \approx |C(P_1 - P_2)|$$

This equation shows clearly the disadvantage of having controllers with large gain at high frequencies. The sensitivity to modeling error for high frequencies can thus be reduced substantially by a controller whose gain goes to zero rapidly for high frequencies. This has been known empirically for a long time and it is called high frequency roll off.

5.7 The Sensitivity Functions

We have seen that the sensitivity function S and the complementary sensitivity function T tell much about the feedback loop. We have also seen from Equations (5.6) and (5.13) that it is advantageous to have a small value of the sensitivity function and it follows from (5.10) that a small value of the complementary sensitivity allows large process uncertainty. Since

$$S(s) = \frac{1}{1 + P(s)C(s)} \text{ and } T(s) = \frac{P(s)C(s)}{1 + P(s)C(s)}$$

it follows that

$$S(s) + T(s) = 1 \tag{5.17}$$

This means that S and T cannot be made small simultaneously. The loop transfer function L is typically large for small values of s and it goes to zero as s goes to infinity. This means that S is typically small for small s and close to 1 for large. The complementary sensitivity function is close to 1 for small s and it goes to 0 as s goes to infinity.

A basic problem is to investigate if S can be made small over a large frequency range. We will start by investigating an example.

EXAMPLE 5.6—SYSTEM THAT ADMITS SMALL SENSITIVITIES
Consider a closed loop system consisting of a first order process and a proportional controller. Let the loop transfer function

$$L(s) = P(s)C(s) = \frac{k}{s+1}$$

where parameter k is the controller gain. The sensitivity function is

$$S(s) = \frac{s+1}{s+1+k}$$

and we have

$$|S(i\omega)| = \sqrt{\frac{1+\omega^2}{1+2k+k^2+\omega^2}}$$

This implies that $|S(i\omega)| < 1$ for all finite frequencies and that the sensitivity can be made arbitrary small for any finite frequency by making k sufficiently large. □

The system in Example 5.6 is unfortunately an exception. The key feature of the system is that the Nyquist curve of the process lies in the fourth quadrant. Systems whose Nyquist curves are in the first and fourth quadrant are called positive real. For such systems the Nyquist curve never enters the region shown in Figure 5.11 where the sensitivity is greater than one.

For typical control systems there are unfortunately severe constraints on the sensitivity function. Bode has shown that if the loop transfer has poles p_k in the right half plane and if it goes to zero faster than $1/s$ for large s the sensitivity function satisfies the following integral

$$\int_0^\infty \log |S(i\omega)| d\omega = \int_0^\infty \log \frac{1}{|1 + L(i\omega)|} d\omega = \pi \sum \text{Re } p_k \qquad (5.18)$$

This equation shows that if the sensitivity function is made smaller for some frequencies it must increase at other frequencies. This means that if disturbance attenuation is improved in one frequency range it will be worse in other. This has been been called the water bed effect.

Equation (5.18) implies that there are fundamental limitations to what can be achieved by control and that control design can be viewed as a redistribution of disturbance attenuation over different frequencies.

For a loop transfer function without poles in the right half plane (5.18) reduces to

$$\int_0^\infty \log |S(i\omega)| d\omega = 0$$

This formula can be given a nice geometric interpretation as shown in Figure 5.17 which shows $\log |S(i\omega)|$ as a function of ω. The area over the horizontal axis must be equal to the area under the axis.

Derivation of Bode's Formula*

This is a technical section which requires some knowledge of the theory of complex variables, in particular contour integration. Assume that the loop transfer function has distinct poles at $s = p_k$ in the right half plane and that $L(s)$ goes to zero faster than $1/s$ for large values of s. Consider the integral of the logarithm of the sensitivity function $S(s) = 1/(1 + L(s))$ over the contour shown in Figure 5.18. The contour encloses the right half plane except the points $s = p_k$ where the loop transfer function $L(s) = P(s)C(s)$ has poles and the sensitivity function $S(s)$ has zeros. The direction of the contour is counter clockwise.

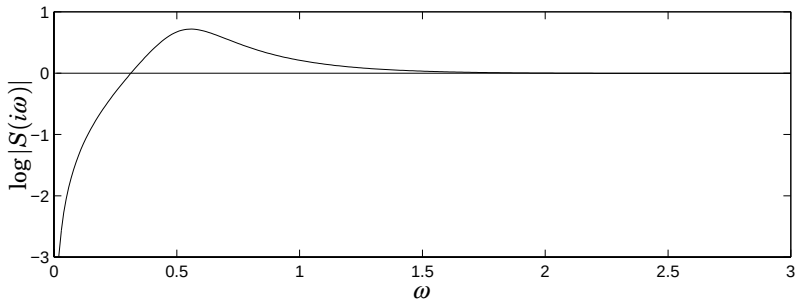


Figure 5.17 Geometric interpretation of Bode's integral formula (5.18).

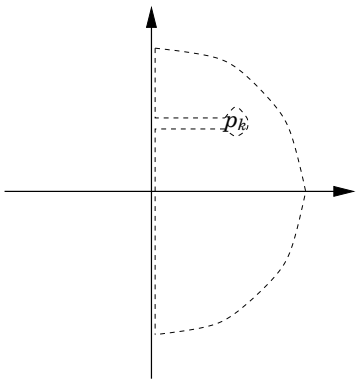


Figure 5.18 Contour used to prove Bode's theorem.

$$\int_\Gamma \log(S(s))ds = \int_{i\omega}^{-i\omega} \log(S(s))ds + \int_R \log(S(s))ds + \sum_k \int_\gamma \log(S(s))ds$$
$$= I_1 + I_2 + I_3 = 0$$

where R is a large semi circle on the right and γ_k is the contour starting on the imaginary axis at $s = \operatorname{Im} p_k$ and a small circle enclosing the pole p_k. The integral is zero because the function $\log S(s)$ is regular inside the contour. We have

$$I_1 = -i \int_{-iR}^{iR} \log(S(i\omega))d\omega = -2i \int_0^{iR} \log(|S(i\omega)|)d\omega$$

because the real part of $\log S(i\omega)$ is an even function and the imaginary

part is an odd function. Furthermore we have

$$I_2 = \int_R \log(S(s))ds = \int_R \log(1 + L(s))ds \approx \int_R L(s)ds$$

Since $L(s)$ goes to zero faster than $1/s$ for large s the integral goes to zero when the radius of the circle goes to infinity. Next we consider the integral I_3, for this purpose we split the contour into three parts X_+, γ and X_- as indicated in Figure 5.18. We have

$$\int_\gamma \log(S(s))ds = \int_{X+} \log(S(s))ds + \int_\gamma \log(S(s))ds + \int_{X-} \log(S(s))ds$$

The contour γ is a small circle with radius r around the pole p_k. The magnitude of the integrand is of the order $\log r$ and the length of the path is $2\pi r$. The integral thus goes to zero as the radius r goes to zero. Furthermore we have

$$\begin{aligned}\int_{X_+} \log(S(s))ds + \int_{X_-} \log(S(s))ds \\ = \int_{X_+} \big(\log(S(s)) - \log(S(s - 2\pi i)\big)ds = 2\pi p_k\end{aligned}$$

Letting the small circles go to zero and the large circle go to infinity and adding the contributions from all right half plane poles p_k gives

$$I_1 + I_2 + I_3 = -2i \int_0^{iR} \log(|S(i\omega)|)d\omega + \sum_k 2\pi p_k = 0.$$

which is Bode's formula (5.18).

5.8 Reference Signals

The response of output y and control u to reference r for the systems in Figure 5.1 having two degrees of freedom is given by the transfer functions

$$\begin{aligned}G_{yr} &= \frac{PCF}{1 + PC} = FT \\ G_{ur} &= \frac{CF}{1 + PC}\end{aligned}$$

First we can observe that if $F = 1$ then the response to reference signals is given by T. In many cases the transfer function T gives a satisfactory

response but in some cases it may be desirable to modify the response. If the feedback controller C has been chosen to deal with disturbances and process uncertainty it is straight forward to find a feedforward transfer function that gives the desired response. If the desired response from reference r to output y is characterized by the transfer function M the transfer function F is simply given by

$$F = \frac{M}{T} = \frac{(1+PC)M}{PC} \tag{5.19}$$

The transfer function F has to be stable and it therefore follows that all right half plane zeros of C and P must be zeros of M. Non-minimum phase properties of the process and the controller therefore impose restrictions on the response to reference signals. The transfer function given by (5.19) can also be complicated so it may be useful to approximate the transfer function.

Tracking of Slowly Varying Reference Signals

In applications such as motion control and robotics it may be highly desirable to have very high precision in following slowly varying reference signals. To investigate this problem we will consider a system with error feedback. Neglecting disturbances it follows that

$$E(s) = S(s)R(s)$$

To investigate the effects of slowly varying reference signals we make a Taylor series expansion of the sensitivity function

$$S(s) = e_0 + e_1 s + e_2 s^2 + \dots.$$

The coefficients e_k are called error coefficients. The output generated by slowly varying inputs is thus given by

$$y(t) = r(t) - e_0 r(t) - e_1 \frac{dr(t)}{dt} - e_2 \frac{d^2 r(t)}{dt^2} + \dots \tag{5.20}$$

Notice that the sensitivity function is given by

$$S(s) = \frac{1}{1+P(s)C(s)}$$

The coefficient e_0 is thus zero if $P(s)C(s) \approx 1/s$ for small s, i.e. if the process or the controller has integral action.

EXAMPLE 5.7—TRACKING A RAMP SIGNALS
Consider for example a ramp input

$$r(t) = v_0 t.$$

It follows from (5.20) that the output is given by

$$y(t) = v_0 t - e_0 v_0 t - e_1 v_0.$$

The error grows linearly if $e_0 \neq 0$. If $e_0 = 0$ there is a constant error which is equal to $e_1 v_0$ in the steady state. □

The example shows that a system where the loop transfer function has an integrator there will be a constant steady state error when tracking a ramp signal. The error can be eliminated by using feedforward as is illustrated in the next example.

EXAMPLE 5.8—REDUCING TRACKING ERROR BY FEEDFORWARD
Consider the problem in Example 5.7. Assume that $e_0 = 0$. Introducing the feedforward transfer function

$$F = 1 + f_1 s \tag{5.21}$$

we find that the transfer function from reference to output becomes

$$\begin{aligned} G_{yr}(s) &= F(s)T(s) = F(s)\big(1 - S(s)\big) \\ &= (1 + f_1 s)(1 - e_0 - e_1 s - e_2 s^2 - \ldots) \\ &= 1 - e_0 + (f_1(1 - e_0) - e_1)s + (e_2 - f_1 e_2)s^2 + \ldots \end{aligned}$$

If the controller has integral action it follows that $e_0 = 0$. It then follows that the tracking error is zero if $f_1 = e_1$. The compensator (5.21) implies that the feedforward compensator predicts the output. Notice that two coefficients have to be matched. □

The error in tracking ramps can also be eliminated by introducing an additional integrator in the controller as is illustrated in the next example.

EXAMPLE 5.9—REDUCING TRACKING ERROR BY FEEDBACK
Choosing $F = 1$ and a controller which gives a loop transfer function with two integrators we have

$$L(s) \approx \frac{k}{s^2}$$

for small s. This implies that $e_0 = e_1 = 0$ and $e_2 = 1/k$ and it follows from (5.20) that there will be no steady state tracking error. There is, however,

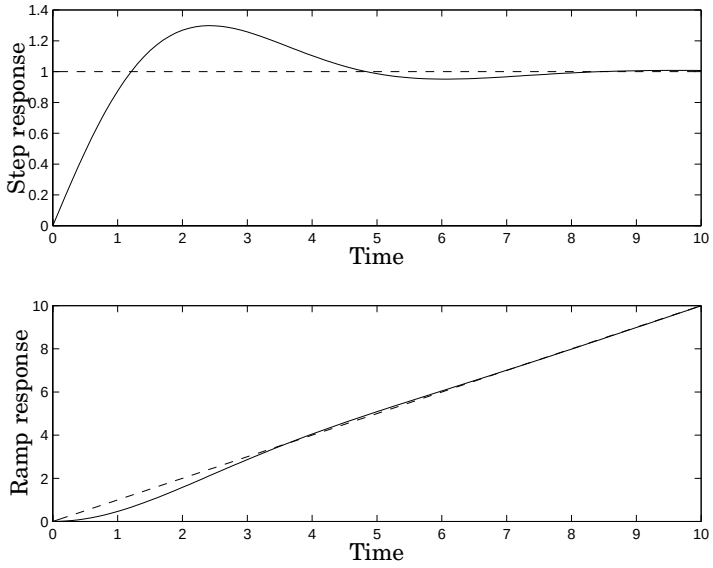


Figure 5.19 Step (above) and ramp (below) responses for a system with error feedback having $e_0 = e_1 = 0$.

one disadvantage with a loop transfer function having two integrators because the response to step signals will have a substantial overshoot. The error in the step response is given by

$$E(s) = S(s)\frac{1}{s}$$

The integral of the error is

$$\frac{E(s)}{s} = S(s)\frac{1}{s^2}$$

Using the final value theorem we find that

$$\lim_{t\to\infty}\int_0^t e(\tau)d\tau = \lim_{s\to 0}\frac{sS(s)}{s^2} = 0$$

Since the integral of the error for a step in the reference is zero it means that the error must have an overshoot. This is illustrated in Figure 5.19. This is avoided if feedforward is used. □

The figure indicates that an attempt to obtain a controller that gives good responses to step and ramp inputs is a difficult compromise if the

controller is linear and time invariant. In this case it is possible to resolve the compromise by using adaptive controllers that adapt their behavior to the properties of the input signal.

Constraints on the sensitivities will, in general, give restrictions on the closed loop poles that can be chosen. This implies that when controllers are designed using pole placements it is necessary to check afterwards that the sensitivities have reasonable values. This does in fact apply to all design methods that do not introduce explicit constraints on the sensitivity functions.

5.9 Fundamental Limitations

In any field it is important to be aware of fundamental limitations. In this section we will discuss these for the basic feedback loop. We will discuss how quickly a system can respond to changes in the reference signal. Some of the factors that limit the performance are

- Measurement noise
- Actuator saturation
- Process dynamics

Measurement Noise and Saturations

It seems intuitively reasonable that fast response requires a controller with high gain. When the controller has high gain measurement noise is also amplified and fed into the system. This will result in variations in the control signal and in the process variable. It is essential that the fluctuations in the control signal are not so large that they cause the actuator to saturate. Since measurement noise typically has high frequencies the high frequency gain M_c of the controller is thus an important quantity. Measurement noise and actuator saturation thus gives a bound on the high frequency gain of the controller and therefore also on the response speed.

There are many sources of measurement noise, it can caused by the physics of the sensor, in can be electronic. In computer controlled systems it is also caused by the resolution of the analog to digital converter. Consider for example a computer controlled system with 12 bit AD and DA converters. Since 12 bits correspond to 4096 it follows that if the high frequency gain of the controller is $M_c = 4096$ one bit conversion error will make the control signal change over the full range. To have a reasonable system we may require that the fluctuations in the control signal due to measurement noise cannot be larger than 5% of the signal span. This

means that the high frequency gain of the controller must be restricted to 200.

Dynamics Limitations

The limitations caused by noise and saturations seem quite obvious. It turns out that there may also be severe limitations due to the dynamical properties of the system. This means that there are systems that are inherently difficult or even impossible to control. It is very important for designers of any system to be aware of this. Since systems are often designed from static considerations the difficulties do not show up because they are dynamic in nature. A brief summary of dynamic elements that cause difficulties are summarized briefly.

It seems intuitively clear that time delay cause limitations in the response speed. A system clearly cannot respond in times that are shorter than the time delay. It follows from

$$e^{-sT_d} \approx \frac{1 - sT_d/2}{1 + sT_d/2} = \frac{s - 2/T_d}{s + 2/T_d} = \frac{s - z}{s + z} \tag{5.22}$$

that a zero in the right half plane z can be approximated with a time delay $T_d = 2/z$ and we may thus expect that zeros in the right half plane also cause limitations. Notice that a small zero corresponds to a long time delay.

Intuitively it also seems reasonable that instabilities will cause limitations. We can expect that a fast controller is required to control an unstable system.

Summarizing we can thus expect that time delays and poles and zeros in the right half plane give limitations. To give some quantitative results we will characterize the closed loop system by the gain crossover frequency ω_{gc}. This is the smallest frequency where the loop transfer function has unit magnitude, i.e. $|L(i\omega_{gc})|$. This parameter is approximately inversely proportional to the response time of a system. The dynamic elements that cause limitations are time delays and poles and zeros in the right half plane. The key observations are:

- A right half plane zero z limits the response speed. A simple rule of thumb is
$$\omega_{gc} < 0.5z \tag{5.23}$$
Slow RHP zeros are thus particularly bad.
- A time delay T_d limits the response speed. A simple rule of thumb is
$$\omega_{gc} T_d < 0.4 \tag{5.24}$$

- A right half plane pole p requires high gain crossover frequency. A simple rule of thumb is
$$\omega_{gc} > 2p \tag{5.25}$$
Fast unstable poles require a high crossover frequency.

- Systems with a right half plane pole p and a right half plane zero z cannot be controlled unless the pole and the zero are well separated. A simple rule of thumb is
$$p > 6z \tag{5.26}$$

- A system with a a right half plane pole and a time delay T_d cannot be controlled unless the product pT_d is sufficiently small. A simple rule of thumb is
$$pT_d < 0.16 \tag{5.27}$$

We illustrate this with a few examples.

EXAMPLE 5.10—BALANCING AN INVERTED PENDULUM
Consider the situation when we attempt to balance a pole manually. An inverted pendulum is an example of an unstable system. With manual balancing there is a neural delay which is about $T_d = 0.04$ s. The transfer function from horizontal position of the pivot to the angle is

$$G(s) = \frac{s^2}{s^2 - \frac{g}{\ell}}$$

where $g = 9.8\ m/s^2$is the acceleration of gravity and ℓ is the length of the pendulum. The system has a pole $p = \sqrt{g/\ell}$. The inequality (5.27) gives

$$0.04\sqrt{g/\ell} = 0.16$$

Hence, $\ell = 0.6$ m. Investigate the shortest pole you can balance. □

EXAMPLE 5.11—BICYCLE WITH REAR WHEEL STEERING
The dynamics of a bicycle was derived in Section 4.3. To obtain the model for a bicycle with rear wheel steering we can simply change the sign of the velocity. It then follows from (4.9) that the transfer function from steering angle β to tilt angle θ is

$$P(s) = \frac{mV_0\ell}{b}\frac{Js^2 - mgl}{-as + V_0}$$

Notice that the transfer function depends strongly on the forward velocity of the bicycle. The system thus has a right half plane pole at $p = \sqrt{mg\ell/J}$ and a right half plane zero at $z = V_0/a$, and it can be suspected that the system is difficult to control. The location of the pole does not depend on velocity but the the position of the zero changes significantly with velocity. At low velocities the zero is at the origin. For $V_0 = a\sqrt{mg\ell/J}$ the pole and the zero are at the same location and for higher velocities the zero is to the right of the pole. To draw some quantitative conclusions we introduce the numerical values $m = 70$ kg, $\ell = 1.2$ m, $a = 0.7$, $J = 120$ kgm^2 and $V = 5$ m/s, give $z = V/a = 7.14$ rad/s and $p = \omega_0 = 2.6$ rad/s we find that $p = 2.6$. With $V_0 = 5$ m/s we get $z = 7.1$, and $p/z = 2.7$. To have a situation where the system can be controlled it follows from (5.26) that to have $z/p = 6$ the velocity must be increased to 11 m/s. We can thus conclude that if the speed of the bicycle can be increased to about 10 m/s so rapidly that we do not loose balance it can indeed be ridden. □

The bicycle example illustrates clearly that it is useful to assess the fundamental dynamical limitations of a system at an early stage in the design. If this had been done the it could quickly have been concluded that the study of rear wheel steered motor bikes in 4.3 was not necessary.

Remedies

Having understood factors that cause fundamental limitations it is interesting to know how they should be overcome. Here are a few suggestions.

Problems with sensor noise are best approached by finding the roots of the noise and trying to eliminate them. Increasing the resolution of a converter is one example. Actuation problems can be dealt with in a similar manner. Limitations caused by rate saturation can be reduced by replacing the actuator.

Problems that are caused by time delays and RHP zeros can be approached by moving sensors to different places. It can also be beneficial to add sensors. Recall that the zeros depend on how inputs and outputs are coupled to the states of a system. A system where all states are measured has no zeros.

Poles are inherent properties of a system, they can only be modified by redesign of the system.

Redesign of the process is the final remedy. Since static analysis can never reveal the fundamental limitations it is very important to make an assessment of the dynamics of a system at an early stage of the design. This is one of the main reasons why all system designers should have a basic knowledge of control.

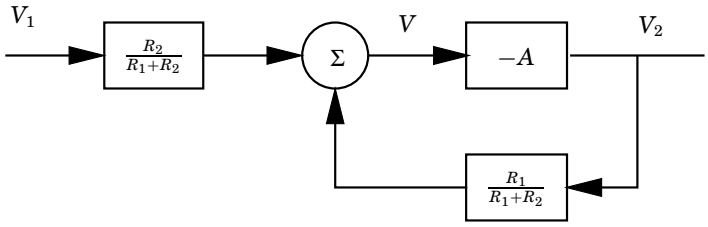


Figure 5.20 Block diagram of the feedback amplifier in Figure 2.9. Compare with Figure 2.10

5.10 Electronic Amplifiers

There are many variations on the prototype problem discussed in Section 5.2. To illustrate this we will discuss electronic amplifiers. Examples of such amplifiers have been given several times earlier, see Section 1.8 and Example 2.3.

The key issues in amplifier design are gain, noise and process variations. The purpose of an electronic amplifier is to provide a high gain and a highly linear input output characteristics. The main disturbance is electrical noise which typically has high frequencies. There are variations in the components that create nonlinearities and slow drift that are caused by temperature variations. A nice property of feedback amplifiers that differ from many other processes is that many extra signals are available internally.

The difficulty of finding a natural block diagram representation of a simple feedback amplifier was discussed in Example 2.3. Some alternative block diagram representations were given in Figure 2.10. In particular we noted the difficulty that there was not a one to one correspondence between the components and the blocks. We will start by showing yet another representation. In this diagram we have kept the negative gain of the feedback loop in the forward path and the standard -1 block has been replaced by a feedback. It is customary to use diagrams of this type when dealing with feedback amplifiers. The generic version of the diagram is shown in Figure 5.21. The block A represents the open loop amplifier, block F the feedback and block H the feedforward. The blocks F and H are represented by passive components.

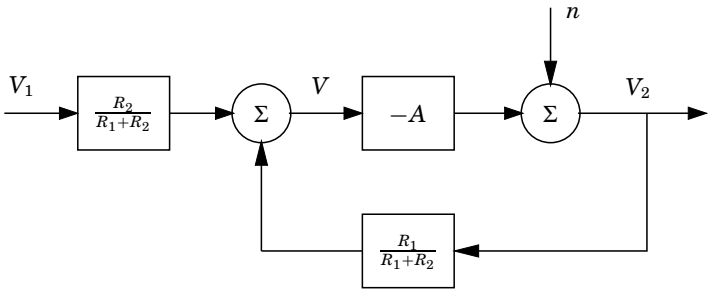


Figure 5.21 Generic block diagram of a feedback amplifier.

For the circuit in Figure 5.20 we have

$$F = \frac{R_1}{R_1 + R_2}$$
$$H = \frac{R_2}{R_1 + R_2}$$

notice that both F and H are less than one.

The Gain Bandwidth Product

The input-output relation for the system in Figure 5.21 is given by

$$\frac{\mathcal{L}V_2}{\mathcal{L}V_1} = -G$$

where

$$G = \frac{AH}{1 + AF} \tag{5.28}$$

The transfer function of an operational amplifier can be approximated by

$$A(s) = \frac{b}{s + a}$$

The amplifier has gain b/a and bandwidth a. The gain bandwidth product is b. Typical numbers for a simple operational amplifier that is often used to implement control systems, LM 741, are $a = 50$ Hz, $b = 1$ MHz. Other amplifiers may have gain bandwidth products that are several orders of magnitude larger.

Furthermore we have

$$F = \frac{R_1}{R_1 + R_2}, \quad H = \frac{R_2}{R_1 + R_2}$$

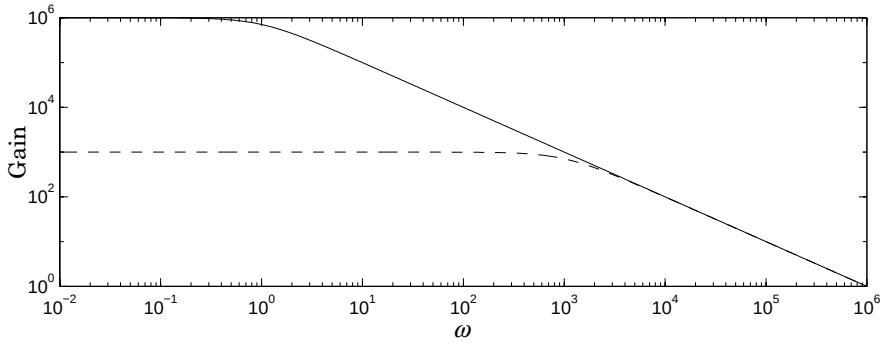


Figure 5.22 Gain curves of the open loop amplifier (full lines) and the feedback amplifier (dashed)

Combining this with the expression for $A(s)$ in (5.28) gives

$$G = \frac{bR_2}{(R_1 + R_2)(s + a) + bR_1} \approx \frac{bR_2}{R_2 s + bR_1}$$

where the approximation is obtained from the inequalities $b >> a$ and $R_2 >> R_1$. The closed loop system thus has gain $R_2 R_1$ and bandwidth $\omega_0 = bR_1/R_2$ and it follows that the gain bandwidth product is constant

$$Gain \times Bandwidth = b$$

Notice that feedback does not change the gain bandwidth product. The effect of feedback is simply to decrease the gain and increase the bandwidth. This is illustrated in Figure 5.22 which shows the gain curves of the open and closed loop systems. Also notice that the sensitivity of the system is

$$S = \frac{1}{1 + AF} = \frac{(R_1 + R_2)(s + a)}{(R_1 + R_2)(s + a) + bR_1} \approx \frac{R_2(s + a)}{R_2 s + bR_1}$$

The high open loop gain of the amplifier is traded off for high bandwidth and low sensitivity. This is some times expressed by saying that gain is the hard currency of feedback amplifiers which can be traded for sensitivity and linearity.

Sensitivity

It follows from (5.28) that

$$\log G = \log AH - \log(1 + AF)$$

Differentiating this expression we find that

$$\frac{d\log G}{d\log A} = \frac{1}{1+AF}$$
$$\frac{d\log G}{d\log F} = -\frac{AF}{1+AF}$$
$$\frac{d\log G}{d\log H} = 1$$

The loop transfer function is normally large which implies that it is only the sensitivity with respect the amplifier that is small. This is, however, the important active part where there are significant variations. The transfer functions F and H typically represent passive components that are much more stable than the amplifiers.

Signal to Noise Ratio

The ratio between signal and noise is an important parameter for an amplifier. Noise is represented by the signals n_1 and n_2 in Figure 5.21. Noise entering at the amplifier input is more critical than noise at the amplifier output. For an open loop system the output voltage is given by

$$V_{ol} = N_2 - A(N_1 + HV_1)$$

For a system with feedback the output voltage is instead given by

$$V_{cl} = \frac{1}{1+AF}\big(N_2 - A(N_1 + HV_1)\big) = \frac{1}{1+AF}V_{ol}$$

The signals will be smaller for a system with feedback but the signal to noise ratio does not change.

5.11 Summary

Having got insight into some fundamental properties of the feedback loop we are in a position to discuss how to formulate specifications on a control system. It was mentioned in Section 5.2 that requirements on a control system should include stability of the closed loop system, robustness to model uncertainty, attenuation of measurement noise, injection of measurement noise ability to follow reference signals. From the results given in this section we also know that these properties are captured by six

transfer functions called the Gang of Six. The specifications can thus be expressed in terms of these transfer functions.

Stability and robustness to process uncertainties can be expressed by the sensitivity function and the complementary sensitivity function

$$S = \frac{1}{1+PC}, \qquad T = \frac{PC}{1+PC}.$$

Load disturbance attenuation is described by the transfer function from load disturbances to process output

$$G_{yd} = \frac{P}{1+PC} = PS.$$

The effect of measurement noise is be captured by the transfer function

$$-G_{un} = \frac{C}{1+PC} = CS,$$

which describes how measurement noise influences the control signal. The response to set point changes is described by the transfer functions

$$G_{yr} = \frac{FPC}{1+PC} = FT, \qquad G_{ur} = \frac{FC}{1+PC} = FCS$$

Compare with (5.1). A significant advantage with controller structure with two degrees of freedom is that the problem of set point response can be decoupled from the response to load disturbances and measurement noise. The design procedure can then be divided into two independent steps.

- First design the feedback controller C that reduces the effects of load disturbances and the sensitivity to process variations without introducing too much measurement noise into the system
- Then design the feedforward F to give the desired response to set points.

6

PID Control

6.1 Introduction

The PID controller is the most common form of feedback. It was an essential element of early governors and it became the standard tool when process control emerged in the 1940s. In process control today, more than 95% of the control loops are of PID type, most loops are actually PI control. PID controllers are today found in all areas where control is used. The controllers come in many different forms. There are stand-alone systems in boxes for one or a few loops, which are manufactured by the hundred thousands yearly. PID control is an important ingredient of a distributed control system. The controllers are also embedded in many special-purpose control systems. PID control is often combined with logic, sequential functions, selectors, and simple function blocks to build the complicated automation systems used for energy production, transportation, and manufacturing. Many sophisticated control strategies, such as model predictive control, are also organized hierarchically. PID control is used at the lowest level; the multivariable controller gives the setpoints to the controllers at the lower level. The PID controller can thus be said to be the "bread and butter" of control engineering. It is an important component in every control engineer's tool box.

PID controllers have survived many changes in technology, from mechanics and pneumatics to microprocessors via electronic tubes, transistors, integrated circuits. The microprocessor has had a dramatic influence on the PID controller. Practically all PID controllers made today are based on microprocessors. This has given opportunities to provide additional features like automatic tuning, gain scheduling, and continuous adaptation.

6.2 The Algorithm

We will start by summarizing the key features of the PID controller. The "textbook" version of the PID algorithm is described by:

$$u(t) = K\left(e(t) + \frac{1}{T_i}\int_0^t e(\tau)d\tau + T_d\frac{de(t)}{dt}\right) \tag{6.1}$$

where y is the measured process variable, r the reference variable, u is the control signal and e is the control error ($e = y_{sp} - y$). The reference variable is often called the set point. The control signal is thus a sum of three terms: the P-term (which is proportional to the error), the I-term (which is proportional to the integral of the error), and the D-term (which is proportional to the derivative of the error). The controller parameters are proportional gain K, integral time T_i, and derivative time T_d. The integral, proportional and derivative part can be interpreted as control actions based on the past, the present and the future as is illustrated in Figure 2.2. The derivative part can also be interpreted as prediction by linear extrapolation as is illustrated in Figure 2.2. The action of the different terms can be illustrated by the following figures which show the response to step changes in the reference value in a typical case.

Effects of Proportional, Integral and Derivative Action

Proportional control is illustrated in Figure 6.1. The controller is given by (6.1) with $T_i = \infty$ and $T_d = 0$. The figure shows that there is always a steady state error in proportional control. The error will decrease with increasing gain, but the tendency towards oscillation will also increase.

Figure 6.2 illustrates the effects of adding integral. It follows from (6.1) that the strength of integral action increases with decreasing integral time T_i. The figure shows that the steady state error disappears when integral action is used. Compare with the discussion of the "magic of integral action" in Section Section 2.2. The tendency for oscillation also increases with decreasing T_i. The properties of derivative action are illustrated in Figure 6.3.

Figure 6.3 illustrates the effects of adding derivative action. The parameters K and T_i are chosen so that the closed-loop system is oscillatory. Damping increases with increasing derivative time, but decreases again when derivative time becomes too large. Recall that derivative action can be interpreted as providing prediction by linear extrapolation over the time T_d. Using this interpretation it is easy to understand that derivative action does not help if the prediction time T_d is too large. In Figure 6.3 the period of oscillation is about 6 s for the system without derivative

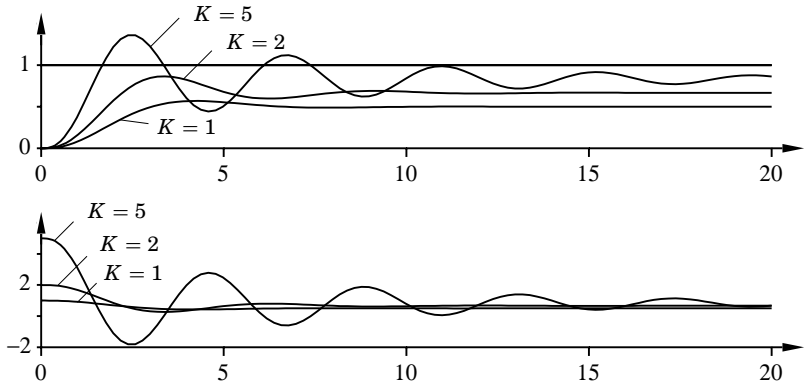


Figure 6.1 Simulation of a closed-loop system with proportional control. The process transfer function is $P(s) = 1/(s+1)^3$.

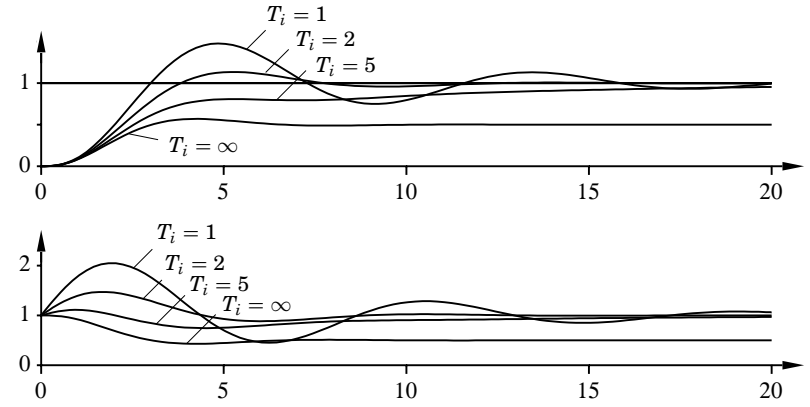


Figure 6.2 Simulation of a closed-loop system with proportional and integral control. The process transfer function is $P(s) = 1/(s+1)^3$, and the controller gain is $K = 1$.

action. Derivative actions ceases to be effective when T_d is larger than a 1 s (one sixth of the period). Also notice that the period of oscillation increases when derivative time is increased.

A Perspective

There is much more to PID than is revealed by (6.1). A faithful implementation of the equation will actually not result in a good controller. To obtain a good PID controller it is also necessary to consider

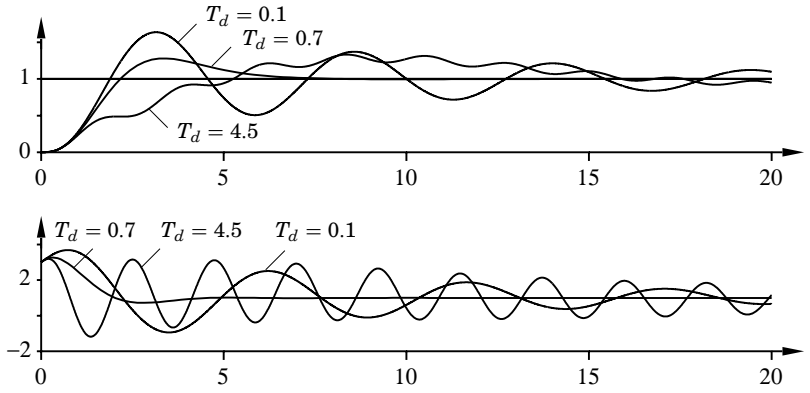


Figure 6.3 Simulation of a closed-loop system with proportional, integral and derivative control. The process transfer function is $P(s) = 1/(s+1)^3$, the controller gain is $K = 3$, and the integral time is $T_i = 2$.

- Noise filtering and high frequency roll-off
- Set point weighting and 2 DOF
- Windup
- Tuning
- Computer implementation

In the case of the PID controller these issues emerged organically as the technology developed but they are actually important in the implementation of all controllers. Many of these questions are closely related to fundamental properties of feedback, some of them have been discussed earlier in the book.

6.3 Filtering and Set Point Weighting

Filtering

Differentiation is always sensitive to noise. This is clearly seen from the transfer function $G(s) = s$ of a differentiator which goes to infinity for large s. The following example is also illuminating.

EXAMPLE 6.1—DIFFERENTIATION AMPLIFIES HIGH FREQUENCY NOISE
Consider the signal

$$y(t) = \sin t + n(t) = \sin t + a_n \sin \omega_n t$$

where the noise is sinusoidal noise with frequency ω. The derivative of the signal is

$$\frac{dy(t)}{dt} = \cos t + n(t) = \cos t + a_n \omega \cos \omega_n t$$

The signal to noise ratio for the original signal is $1/a_n$ but the signal to noise ratio of the differentiated signal is ω/a_n. This ratio can be arbitrarily high if ω is large. □

In a practical controller with derivative action it is therefor necessary to limit the high frequency gain of the derivative term. This can be done by implementing the derivative term as

$$D = -\frac{sKT_d}{1 + sT_d/N} Y \tag{6.2}$$

instead of $D = sT_dY$. The approximation given by (6.2) can be interpreted as the ideal derivative sT_d filtered by a first-order system with the time constant T_d/N. The approximation acts as a derivative for low-frequency signal components. The gain, however, is limited to KN. This means that high-frequency measurement noise is amplified at most by a factor KN. Typical values of N are 8 to 20.

Further limitation of the high-frequency gain

The transfer function from measurement y to controller output u of a PID controller with the approximate derivative is

$$C(s) = -K\left(1 + \frac{1}{sT_i} + \frac{sT_d}{1 + sT_d/N}\right)$$

This controller has constant gain

$$\lim_{s\to\infty} C(s) = -K(1 + N)$$

at high frequencies. It follows from the discussion on robustness against process variations in Section 5.5 that it is highly desirable to roll-off the

controller gain at high frequencies. This can be achieved by additional low pass filtering of the control signal by

$$F(s) = \frac{1}{(1 + sT_f)^n}$$

where T_f is the filter time constant and n is the order of the filter. The choice of T_f is a compromise between filtering capacity and performance. The value of T_f can be coupled to the controller time constants in the same way as for the derivative filter above. If the derivative time is used, $T_f = T_d/N$ is a suitable choice. If the controller is only PI, $T_f = T_i/N$ may be suitable.

The controller can also be implemented as

$$C(s) = -K\left(1 + \frac{1}{sT_i} + sT_d\right)\frac{1}{(1 + sT_d/N)^2} \tag{6.3}$$

This structure has the advantage that we can develop the design methods for an ideal PID controller and use an iterative design procedure. The controller is first designed for the process $P(s)$. The design gives the controller parameter T_d. An ideal controller for the process $P(s)/(1+sT_d/N)^2$ is then designed giving a new value of T_d etc. Such a procedure will also give a clear picture of the trade-off between performance and filtering.

Set Point Weighting

When using the control law given by (6.1) it follows that a step change in the reference signal will result in an impulse in the control signal. This is often highly undesirable therefor derivative action is frequently not applied to the reference signal. This problem can be avoided by filtering the reference value before feeding it to the controller. Another possibility is to let proportional action act only on part of the reference signal. This is called set point weighting. A PID controller given by (6.1) then becomes

$$u(t) = K\left(br(t) - y(t) + \frac{1}{T_i}\int_0^t e(\tau)d\tau + T_d\left(c\frac{dr(t)}{dt} - \frac{dy(t)}{dt}\right)\right) \tag{6.4}$$

where b and c are additional parameter. The integral term must be based on error feedback to ensure the desired steady state. The controller given by (6.4) has a structure with two degrees of freedom because the signal path from y to u is different from that from r to u. The transfer function from r to u is

$$\frac{U(s)}{R(s)} = C_r(s) = K\left(b + \frac{1}{sT_i} + csT_d\right) \tag{6.5}$$

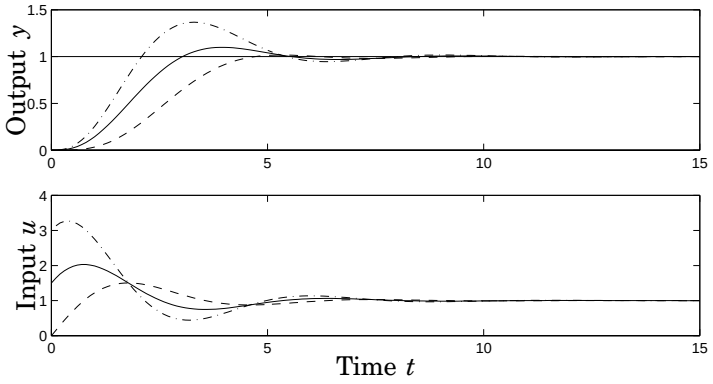


Figure 6.4 Response to a step in the reference for systems with different set point weights $b = 0$ dashed, $b = 0.5$ full and $b = 1.0$ dash dotted. The process has the transfer function $P(s) = 1/(s+1)^3$ and the controller parameters are $k = 3$, $k_i = 1.5$ and $k_d = 1.5$.

and the transfer function from y to u is

$$\frac{U(s)}{Y(s)} = C_y(s) = K\left(1 + \frac{1}{sT_i} + sT_d\right) \tag{6.6}$$

Set point weighting is thus a special case of controllers having two degrees of freedom.

The system obtained with the controller (6.4) respond to load disturbances and measurement noise in the same way as the controller (6.1) . The response to reference values can be modified by the parameters b and c. This is illustrated in Figure 6.4, which shows the response of a PID controller to setpoint changes, load disturbances, and measurement errors for different values of b. The figure shows clearly the effect of changing b. The overshoot for setpoint changes is smallest for $b = 0$, which is the case where the reference is only introduced in the integral term, and increases with increasing b.

The parameter c is normally zero to avoid large transients in the control signal due to sudden changes in the setpoint.

6.4 Different Parameterizations

The PID algorithm given by Equation (6.1) can be represented by the

transfer function

$$G(s) = K\left(1 + \frac{1}{sT_i} + sT_d\right) \tag{6.7}$$

A slightly different version is most common in commercial controllers. This controller is described by

$$G'(s) = K'\left(1 + \frac{1}{sT_i'}\right)(1 + sT_d') = K'\left(1 + \frac{T_d'}{T_i'} + \frac{1}{sT_i'} + sT_d'\right) \tag{6.8}$$

The controller given by Equation (6.7) is called non-interacting, and the one given by Equation (6.8) interacting. The interacting controller Equation (6.8) can always be represented as a non-interacting controller whose coefficients are given by

$$\begin{aligned} K &= K'\frac{T_i' + T_d'}{T_i'} \\ T_i &= T_i' + T_d' \\ T_d &= \frac{T_i'T_d'}{T_i' + T_d'} \end{aligned} \tag{6.9}$$

An interacting controller of the form Equation (6.8) that corresponds to a non-interacting controller can be found only if

$$T_i \geq 4T_d$$

The parameters are then given by

$$\begin{aligned} K' &= \frac{K}{2}\left(1 + \sqrt{1 - 4T_d/T_i}\right) \\ T_i' &= \frac{T_i}{2}\left(1 + \sqrt{1 - 4T_d/T_i}\right) \\ T_d' &= \frac{T_i}{2}\left(1 - \sqrt{1 - 4T_d/T_i}\right) \end{aligned} \tag{6.10}$$

The non-interacting controller given by Equation (6.7) is more general, and we will use that in the future. It is, however, sometimes claimed that the interacting controller is easier to tune manually.

It is important to keep in mind that different controllers may have different structures when working with PID controllers. If a controller is replaced by another type of controller, the controller parameters may have to be changed. The interacting and the non-interacting forms differ only

when both the I and the D parts of the controller are used. If we only use the controller as a P, PI, or PD controller, the two forms are equivalent. Yet another representation of the PID algorithm is given by

$$G''(s) = k + \frac{k_i}{s} + sk_d \tag{6.11}$$

The parameters are related to the parameters of standard form through

$$k = K \qquad k_i = \frac{K}{T_i} \qquad k_d = KT_d$$

The representation Equation (6.11) is equivalent to the standard form, but the parameter values are quite different. This may cause great difficulties for anyone who is not aware of the differences, particularly if parameter $1/k_i$ is called integral time and k_d derivative time. It is even more confusing if k_i is called integration time. The form given by Equation (6.11) is often useful in analytical calculations because the parameters appear linearly. The representation also has the advantage that it is possible to obtain pure proportional, integral, or derivative action by finite values of the parameters.

The PIPD Controller

The controller with set point weighting can also be represented by the block diagram in Figure 6.5. To see this we introduce the transfer functions of the blocks

$$G_{PI}(s) = k' + \frac{k_i'}{s}$$
$$G_{PD}(s) = 1 + k_d's$$

Notice that the proportional gain of the PD controller must be one in order to have zero steady state error. The input-output relation of the complete controller is

$$U(s) = k'R(s) + \frac{k_i'}{s}(R(s) - Y(s)) - (k' + k_d'h_i')Y(s) - k'k_d'sY(s)$$

Which shows that the controller is thus identical to the controller given

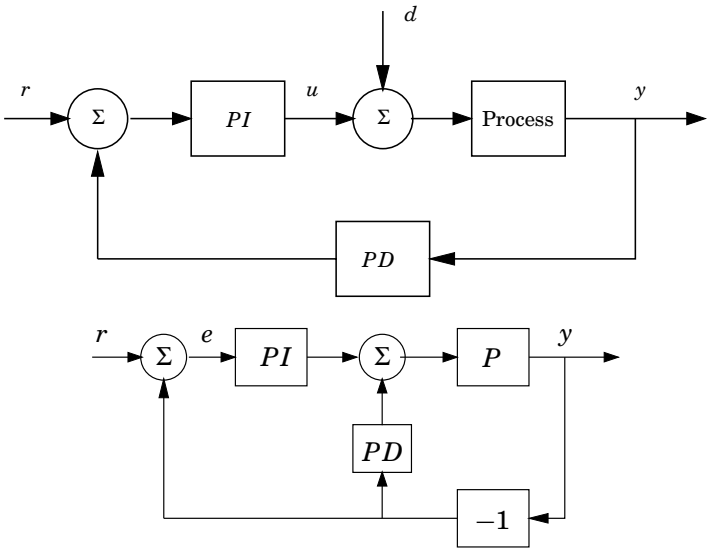


Figure 6.5 Block diagram of a PI-PD controller. This controller is equivalent to a conventional PID controller with set point weighting.

by (6.4). The parameters are related in the following way

$$
\begin{aligned}
k &= k' + k'_d k'_i \\
k_i &= k'_i \\
k_d &= k' k'_d \\
T_i &= \frac{k' + k'_d k'_i}{k'_i} = \frac{k'}{k'_i} \\
T_d &= \frac{k' k'_d}{k'_i} \\
b &= \frac{k'}{k' + k'_d h'_i} \\
c &= 0
\end{aligned}
$$

Following the same pattern the controller with $b = 0$ and $c = 0$ is sometimes called an I-PD controller, and the controller with $b = 1$ and $c = 0$ is called a PI-D controller.

Notice, however, that the representation given by (6.4) is much better suitable for tuning, because the parameters k, T_i and T_d can first be determined to deal with load disturbances, measurement noise and process

uncertainty. When this is done the response to set points can be adjusted by choosing the parameters b and c. The controller parameters appear in a much more complicated way in the PIPD controller.

6.5 Windup

Although many aspects of a control system can be understood based on linear theory, some nonlinear effects must be accounted for in practically all controllers. Windup is such a phenomena, which is caused by the interaction of integral action and saturations. All actuators have limitations: a motor has limited speed, a valve cannot be more than fully opened or fully closed, etc. For a control system with a wide range of operating conditions, it may happen that the control variable reaches the actuator limits. When this happens the feedback loop is broken and the system runs as an open loop because the actuator will remain at its limit independently of the process output. If a controller with integrating action is used, the error will continue to be integrated. This means that the integral term may become very large or, colloquially, it "winds up". It is then required that the error has opposite sign for a long period before things return to normal. The consequence is that any controller with integral action may give large transients when the actuator saturates. We will illustrate this by an example.

EXAMPLE 6.2—ILLUSTRATION OF INTEGRATOR WINDUP
The wind-up phenomenon is illustrated in Figure 6.6, which shows control of an integrating process with a PI controller. The initial setpoint change is so large that the actuator saturates at the high limit. The integral term increases initially because the error is positive; it reaches its largest value at time $t = 10$ when the error goes through zero. The output remains saturated at this point because of the large value of the integral term. It does not leave the saturation limit until the error has been negative for a sufficiently long time to let the integral part come down to a small level. Notice that the control signal bounces between its limits several times. The net effect is a large overshoot and a damped oscillation where the control signal flips from one extreme to the other as in relay oscillation. The output finally comes so close to the setpoint that the actuator does not saturate. The system then behaves linearly and settles. □

The example show integrator windup which is generated by a change in the reference value. Windup may also be caused by large disturbances or equipment malfunctions. It can also occur in many other situations.

The phenomenon of windup was well known to manufacturers of analog controllers who invented several tricks to avoid it. They were described

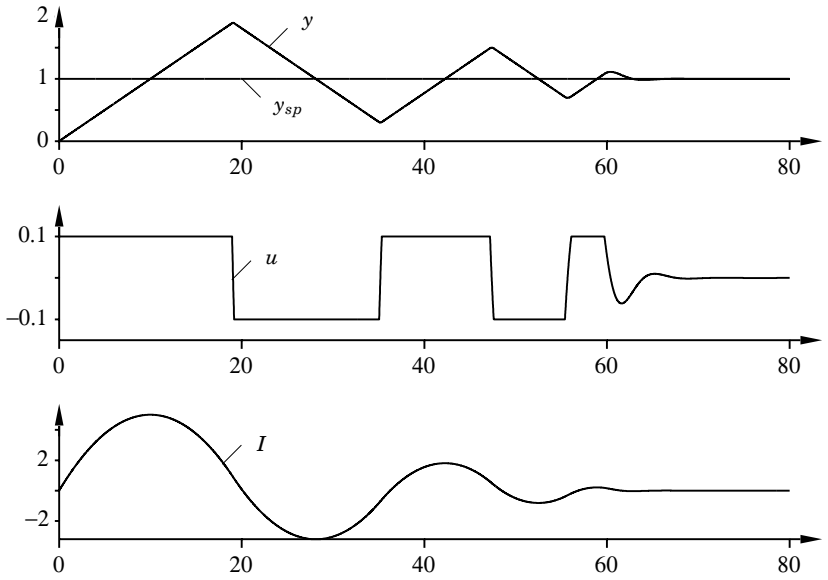


Figure 6.6 Illustration of integrator windup. The diagrams show process output y, setpoint y_{sp}, control signal u, and integral part I.

under labels like preloading, batch unit, etc. Although the problem was well understood, there were often restrictions caused by the analog technology. The ideas were often kept as trade secrets and not much spoken about. The problem of windup was rediscovered when controllers were implemented digitally and several methods to avoid windup were presented in the literature. In the following section we describe some of the methods used to avoid windup.

Setpoint Limitation

One attempt to avoid integrator windup is to introduce limiters on the setpoint variations so that the controller output never reaches the actuator limits. This frequently leads to conservative bounds and poor performance. Furthermore, it does not avoid windup caused by disturbances.

Incremental Algorithms

In the early phases of feedback control, integral action was integrated with the actuator by having a motor drive the valve directly. In this case windup is handled automatically because integration stops when the valve stops. When controllers were implemented by analog techniques,

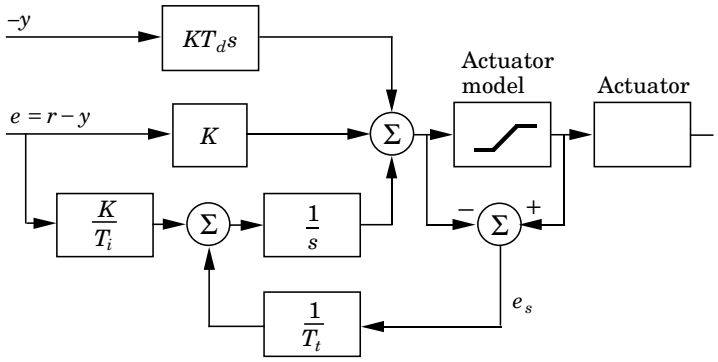


Figure 6.7 Controller with anti-windup where the actuator output is estimated from a mathematical model.

and later with computers, many manufacturers used a configuration that was an analog of the old mechanical design. This led to the so-called velocity algorithms. A velocity algorithm first computes the rate of change of the control signal which is then fed to an integrator. In some cases this integrator is a motor directly connected to the actuator. In other cases the integrator is implemented internally in the controller. With this approach it is easy to avoid windup by inhibiting integration whenever the output saturates. This method is equivalent to back-calculation, which is described below. If the actuator output is not measured, a model that computes the saturated output can be used. It is also easy to limit the rate of change of the control signal.

Back-Calculation and Tracking

Back-calculation works as follows: When the output saturates, the integral term in the controller is recomputed so that its new value gives an output at the saturation limit. It is advantageous not to reset the integrator instantaneously but dynamically with a time constant T_t.

Figure 6.7 shows a block diagram of a PID controller with anti-windup based on back-calculation. The system has an extra feedback path that is generated by measuring the actual actuator output and forming an error signal (e_s) as the difference between the output of the controller (v) and the actuator output (u). Signal e_s is fed to the input of the integrator through gain $1/T_t$. The signal is zero when there is no saturation. Thus, it will not have any effect on the normal operation when the actuator does not saturate. When the actuator saturates, the signal e_s is different from zero. The normal feedback path around the process is broken because the

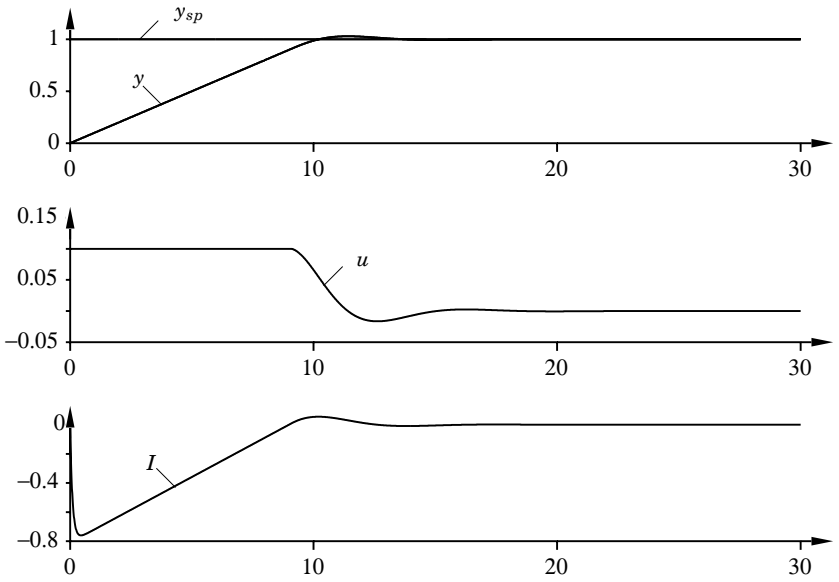


Figure 6.8 Controller with anti-windup applied to the system of Figure 6.6. The diagrams show process output y, setpoint y_{sp}, control signal u, and integral part I.

process input remains constant. There is, however, a feedback path around the integrator. Because of this, the integrator output is driven towards a value such that the integrator input becomes zero. The integrator input is

$$\frac{1}{T_t}\, e_s + \frac{K}{T_i}\, e$$

where e is the control error. Hence,

$$e_s = -\frac{K T_t}{T_i}\, e$$

in steady state. Since $e_s = u - v$, it follows that

$$v = u_{\lim} + \frac{K T_t}{T_i}\, e$$

where u_{lim} is the saturating value of the control variable. This means that the signal v settles on a value slightly out side the saturation limit and the control signal can react as soon as the error changes time. This prevents

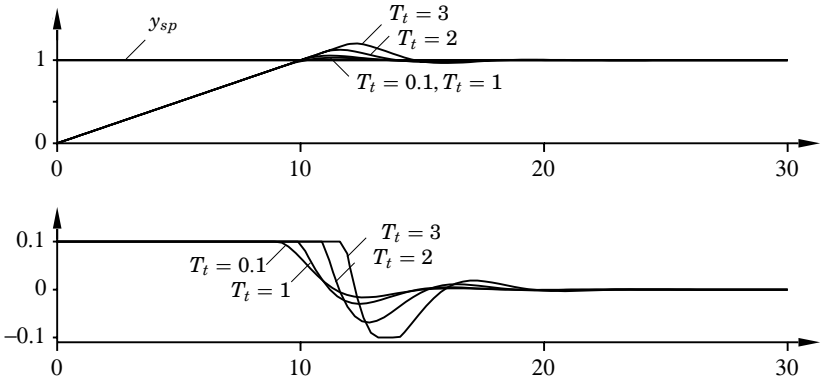


Figure 6.9 The step response of the system in Figure 6.6 for different values of the tracking time constant T_t. The upper curve shows process output y and setpoint y_{sp}, and the lower curve shows control signal u.

the integrator from winding up. The rate at which the controller output is reset is governed by the feedback gain, $1/T_t$, where T_t can be interpreted as the time constant, which determines how quickly the integral is reset. We call this the tracking time constant.

It frequently happens that the actuator output cannot be measured. The anti-windup scheme just described can be used by incorporating a mathematical model of the saturating actuator, as is illustrated in Figure 6.7.

Figure 6.8 shows what happens when a controller with anti-windup is applied to the system simulated in Figure 6.6. Notice that the output of the integrator is quickly reset to a value such that the controller output is at the saturation limit, and the integral has a negative value during the initial phase when the actuator is saturated. This behavior is drastically different from that in Figure 6.6, where the integral has a positive value during the initial transient. Also notice the drastic improvement in performance compared to the ordinary PI controller used in Figure 6.6.

The effect of changing the values of the tracking time constant is illustrated in Figure 6.9. From this figure, it may thus seem advantageous to always choose a very small value of the time constant because the integrator is then reset quickly. However, some care must be exercised when introducing anti-windup in systems with derivative action. If the time constant is chosen too small, spurious errors can cause saturation of the output, which accidentally resets the integrator. The tracking time constant T_t should be larger than T_d and smaller than T_i. A rule of thumb

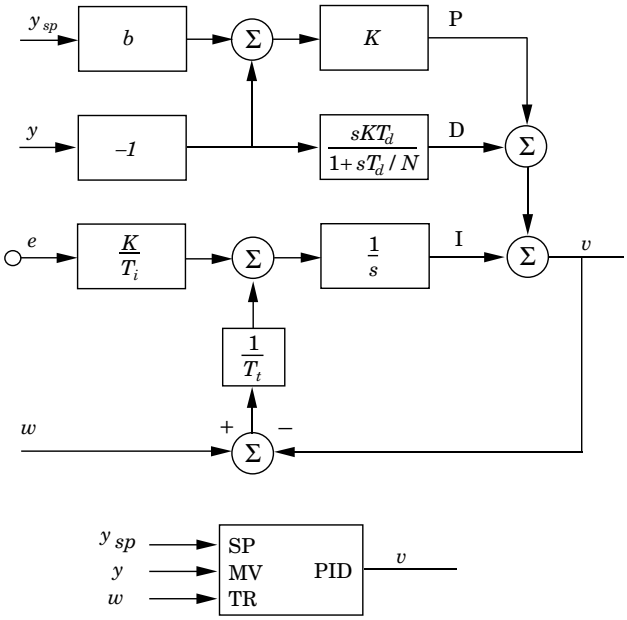


Figure 6.10 Block diagram and simplified representation of PID module with tracking signal.

that has been suggested is to choose $T_t = \sqrt{T_i T_d}$.

Controllers with a Tracking Mode

A controller with back-calculation can be interpreted as having two modes: the normal *control mode*, when it operates like an ordinary controller, and a *tracking mode*, when the controller is tracking so that it matches given inputs and outputs. Since a controller with tracking can operate in two modes, we may expect that it is necessary to have a logical signal for mode switching. However, this is not necessary, because tracking is automatically inhibited when the tracking signal w is equal to the controller output. This can be used with great advantage when building up complex systems with selectors and cascade control.

Figure 6.10 shows a PID module with a tracking signal. The module has three inputs: the setpoint, the measured output, and a tracking signal. The new input TR is called a tracking signal because the controller output will follow this signal. Notice that tracking is inhibited when $w = v$. Using the module the system shown in Figure 6.7 can be presented as shown in Figure 6.11.

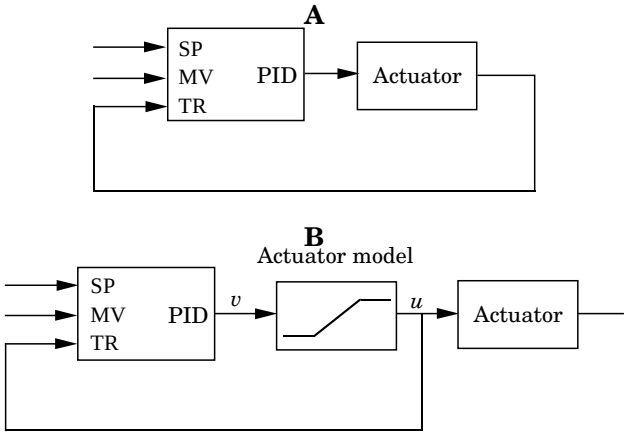


Figure 6.11 Representation of the controllers with anti-windup in Figure 6.7 using the basic control module with tracking shown in Figure 6.10.

6.6 Tuning

All general methods for control design can be applied to PID control. A number of special methods that are tailor-made for PID control have also been developed, these methods are often called tuning methods. Irrespective of the method used it is essential to always consider the key elements of control, load disturbances, sensor noise, process uncertainty and reference signals.

The most well known tuning methods are those developed by Ziegler and Nichols. They have had a major influence on the practice of PID control for more than half a century. The methods are based on characterization of process dynamics by a few parameters and simple equations for the controller parameters. It is surprising that the methods are so widely referenced because they give moderately good tuning only in restricted situations. Plausible explanations may be the simplicity of the methods and the fact that they can be used for simple student exercises in basic control courses.

The Step Response Method

One tuning method presented by Ziegler and Nichols is based on a process information in the form of the open-loop step response obtained from a bump test. This method can be viewed as a traditional method based on modeling and control where a very simple process model is used. The step response is characterized by only two parameters a and L, as shown in

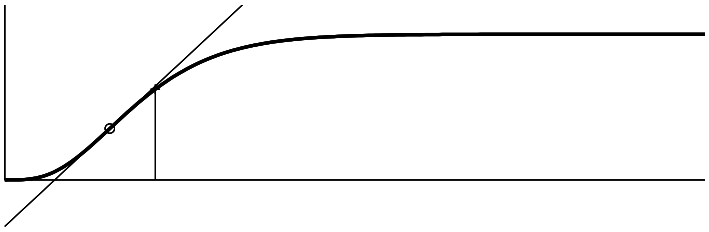


Figure 6.12 Characterization of a step response in the Ziegler-Nichols step response method.

Table 6.1 PID controller parameters obtained for the Ziegler-Nichols step response method.

Controller	K	T_i	T_d	T_p
P	1/a			4L
PI	0.9/a	3L		5.7L
PID	1.2/a	2L	L/2	3.4L

Figure 6.12.

The point where the slope of the step response has its maximum is first determined, and the tangent at this point is drawn. The intersections between the tangent and the coordinate axes give the parameters a and L. The controller parameters are then obtained from Table 6.1. An estimate of the period T_p of the closed-loop system is also given in the table.

The Frequency Response Method

A second method developed by Ziegler and Nichols is based on a simple characterization of the the frequency response of the process process dynamics. The design is based on knowledge of only one point on the Nyquist curve of the process transfer function $P(s)$, namely the point where the Nyquist curve intersects the negative real axis. This point can be characterized by two parameters the frequency ω_{180} and the gain at that frequency $k_{180} = |P(i\omega_{180})|$. For historical reasons the point has been called the ultimate point and characterized by the parameters $K_u = 1/K_{180}$ and $T_u = 2\pi/\omega_{180}$, which are called the *ultimate gain* and the *ultimate period*. These parameters can be determined in the following way. Connect a controller to the process, set the parameters so that control action is proportional, i.e., $T_i = \infty$ and $T_d = 0$. Increase the gain slowly until the

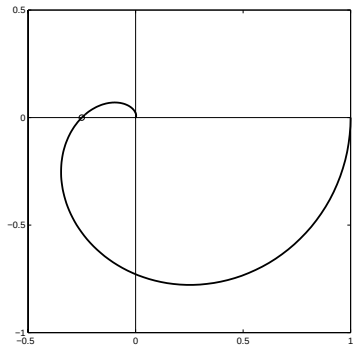


Figure 6.13 Characterization of a step response in the Ziegler-Nichols step response method.

Table 6.2 Controller parameters for the Ziegler-Nichols frequency response method.

Controller	K	T_i	T_d	T_p
P	$0.5K_u$			T_u
PI	$0.4K_u$	$0.8T_u$		$1.4T_u$
PID	$0.6K_u$	$0.5T_u$	$0.125T_u$	$0.85T_u$

process starts to oscillate. The gain when this occurs is K_u and the period of the oscillation is T_u. The parameters of the controller are then given by Table 6.2. An estimate of the period T_p of the dominant dynamics of the closed-loop system is also given in the table.

The frequency response method can be viewed as an empirical tuning procedure where the controller parameters are obtained by direct experiments on the process combined with some simple rules. For a proportional controller the rule is simply to increase the gain until the process oscillates and then reduce it by 50%.

Assessment of the Ziegler Nichols Methods

The Ziegler-Nichols tuning rules were developed to give closed loop systems with good attenuation of load disturbances. The methods were based on extensive simulations. The design criterion was quarter amplitude decay ratio, which means that the amplitude of an oscillation should be reduced by a factor of four over a whole period. This corresponds to closed loop poles with a relative damping of about $\zeta = 0.2$, which is too small.

Controllers designed by the Ziegler-Nichols rules thus inherently give closed loop systems with poor robustness. It also turns out that it is not sufficient to characterize process dynamics by two parameters only. The methods developed by Ziegler and Nichols have been been very popular in spite of these drawbacks. Practically all manufacturers of controller have used the rules with some modifications in recommendations for controller tuning. One reason for the popularity of the rules is that they are simple and easy to explain. The tuning rules give ball park figures. Final tuning is then done by trial and error. Another (bad) reason is that the rules lend themselves very well to simple exercises for control education.

With the insight into controller design that has developed over the years it is possible to develop improved tuning rules that are almost as simple as the Zigler-Nichols rules. These rules are developed by starting with a solid design method that gives robust controllers with effective disturbance attenuation. We illustrate with some rules where the process is characterized by three parameters.

An Improved Step Response Method

This method characterizes the unit step response by three parameters K, L and T for stable processes and $Kv = K/T$ and L for integrating processes. This parameterization matches the transfer functions

$$P_1(s) = \frac{k_p}{1+sT}e^{-sL}$$
$$P_2(s) = \frac{k_v}{s}e^{-sL}$$

The transfer function $P_1(s)$, which is called a first order system with time delay or a KLT model. Parameter L is determined from the intercept of the tangent with largest slope with the time axis as was described in Figure 6.12. Parameter T is also determined as shown in the figure as the difference between the time when the step response reaches 63% of its steady state value. Parameter k_p is the static gain of the system. The parameter k_v is the largest slope of the unit step response. Parameter L is called the apparent time delay and parameter T the apparent time constant or the apparent lag. The adverb apparent is added to indicate that parameters are based on approximations. The parameter

$$\tau = \frac{L}{L+T}$$

is called the relative time delay. This parameter is a good indicator of process dynamics.

To obtain improved tuning rules we use a design method that maximizes integral gain subject to the robustness constraint that the maximum sensitivity is less than $M_s = 1.4$. The procedure has been applied to a a large test batch representing many different processes. One tuning rule is

$$K = \begin{cases} 0.3\dfrac{T}{K_v L} & \text{for} + L < 2T \\ 0.15K_p & \text{for } 2T < L \end{cases}$$
$$T_i = \begin{cases} 8L & \text{for } L < 0.1T \\ 0.8T & \text{for } 0.1T < L < 2T \\ 0.4L & \text{for } 2T < L \end{cases} \tag{6.12}$$

The properties of the improved tuning rules are illustrated by applying them to systems with the transfer functions

$$P_1(s) = \frac{1}{(s+1)(0.2s+1)}$$
$$P_2(s) = \frac{1}{(s+1)^4}$$
$$P_3(s) = \frac{1}{0.05s+1)^2} e^{-1.2s}$$

The process $P_1(s)$ has lag dominated dynamics, process $P_3(s)$ has delay dominated dynamics and process $P_2(s)$ has balanced dynamics.

Figure 6.14 shows the response to a step change in the reference at time zero and a step change in a load disturbance at the process input for PI control of the process $P_1(s)$. The dashed lines show the responses obtained by the Ziegler-Nichols step response method and the full line shows the response obtained with the improved rule which restricted the maximum sensitivity to 1.4. The oscillatory responses to obtained by the Ziegler-Nichols method are clearly visible in the figure which reflects the design choice of quarter amplitude damping. The response to load disturbances obtained by the Ziegler-Nichols method comes at a price of poor sensitivity. There is also a very large overshoot in the response to reference values. Figure 6.15 shows the corresponding responses for the system $P_4(s)$. The oscillatory character obtained with Ziegler Nichols tuning is clearly visible. Figure 6.16 shows the response for a process that is delay dominated. The figure shows that Ziegler-Nichols tuning performs very poorly for this process. Overall we find that the improved tuning rules work for a wide range of processes and that they give robust systems with good responses.

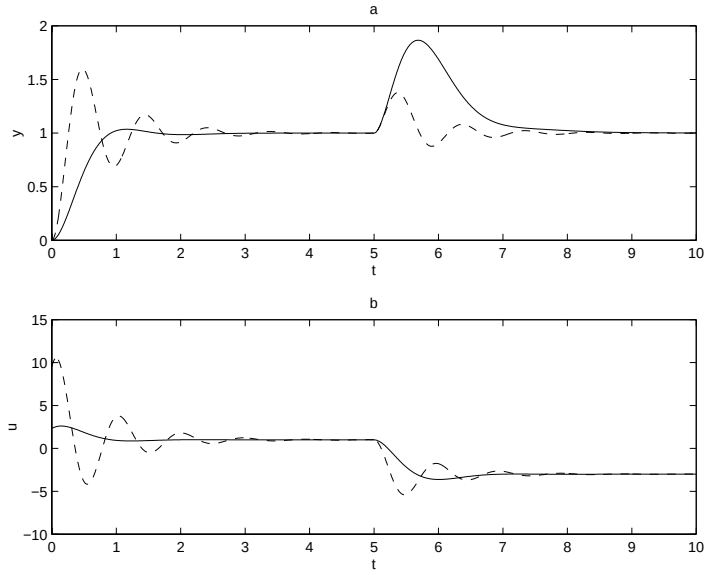


Figure 6.14 Behavior of closed loop systems with PI controllers designed by the Ziegler-Nichols rule (dashed) and the improved tuning rules (solid). The process has lag dominated dynamics with the transfer function $P(s) = \frac{1}{(s+1)(0.2s+1)}$.

6.7 Computer Implementation

Most controllers are nowadays implemented in computers. In this section we will discuss many practical issues related to computer implementation.

Sampling

When the controller is implemented in a computer, the analog inputs are read and the outputs are set with a certain sampling period. This is a drawback compared to the analog implementations, since the sampling introduces dead-time in the control loop.

When a digital computer is used to implement a control law, the ideal sequence of operation is the following.

1. Wait for clock interrupt
2. Read analog input
3. Compute control signal
4. Set analog output

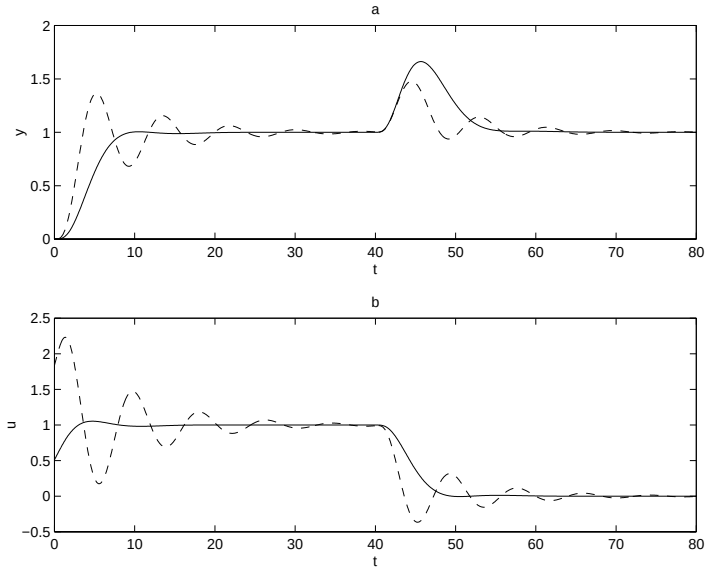


Figure 6.15 Behavior of closed loop systems with PI controllers designed by the Ziegler-Nichols rule (dashed) and the improved tuning rules (solid). The process has balanced dynamics with the transfer function $P(s) = \frac{1}{(s+1)^4}$.

5. Update controller variables
6. Go to 1

With this implementation, the delay is minimized. If the analog input is read with a sampling period h, the average delay of the measurement signal is $h/2$. The computation time is often short compared to the sampling period. This means that the total delay is about $h/2$. However, most controllers and instrument systems do not organize the calculation in this way. Therefore, the delays introduced because of the sampling is often several sampling periods.

Aliasing

The sampling mechanism introduces some unexpected phenomena, which must be taken into account in a good digital implementation of a PID controller. To explain these, consider the signals

$$s(t) = \cos(n\omega_s t \pm \omega t)$$

and

$$s_a(t) = \cos(\omega t)$$

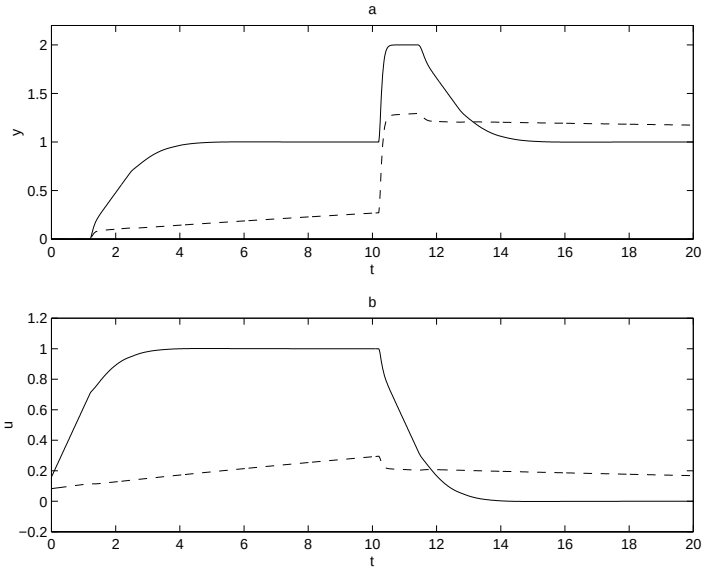


Figure 6.16 Behavior of closed loop systems with PI controllers designed by the Ziegler-Nichols rule (dashed) and the improved tuning rules (solid). The process has delay dominated dynamics with the transfer function $P(s) = \frac{1}{0.05s+1)^2} e^{-1.2s}$.

where $\omega_s = 2\pi/h$ [rad/s] is the sampling frequency. Well-known formulas for the cosine function imply that the values of the signals at the sampling instants $[kh, k = 0, 1, 2, ...]$ have the property

$$s(kh) = \cos(nkh\omega_s \pm \omega kh) = \cos(\omega kh) = s_a(\omega kh)$$

The signals s and s_a thus have the same values at the sampling instants. This means that there is no way to separate the signals if only their values at the sampling instants are known. Signal s_a is, therefore, called an *alias* of signal s. This is illustrated in Figure 6.17. A consequence of the aliasing effect is that a high-frequency disturbance after sampling may appear as a low-frequency signal. In Figure 6.17 the sampling period is 1 s and the sinusoidal disturbance has a period of 6/5 s. After sampling, the disturbance appear as a sinusoid with the frequency

$$f_a = 1 - \frac{5}{6} = 1/6 \text{ Hz}$$

This low-frequency signal with time period 6 s is seen in the figure.

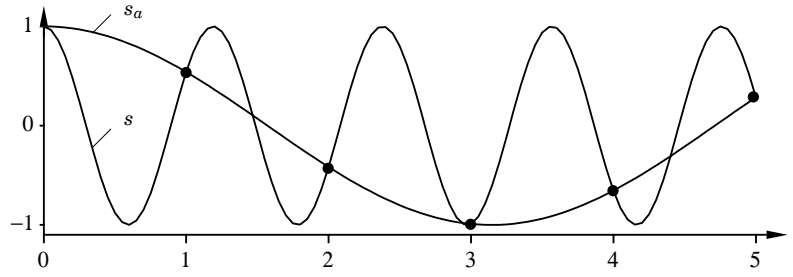


Figure 6.17 Illustration of the aliasing effect. The diagram shows signal s and its alias s_a.

Prefiltering

The aliasing effect can create significant difficulties if proper precautions are not taken. High frequencies, which in analog controllers normally are effectively eliminated by low-pass filtering, may, because of aliasing, appear as low-frequency signals in the bandwidth of the sampled control system. To avoid these difficulties, an analog prefilter (which effectively eliminates all signal components with frequencies above half the sampling frequency) should be introduced. Such a filter is called an anti-aliasing filter. A second-order Butterworth filter is a common anti-aliasing filter. Higher-order filters are also used in critical applications. The selection of the filter bandwidth is illustrated by the following example.

EXAMPLE 6.3—SELECTION OF PREFILTER BANDWIDTH

Assume it is desired that the prefilter attenuate signals by a factor of 16 at half the sampling frequency. If the filter bandwidth is ω_b and the sampling frequency is ω_s, we get

$$(\omega_s/2\omega_b)^2 = 16$$

Hence,

$$\omega_b = \frac{1}{8}\,\omega_s$$

□

Notice that the dynamics of the prefilter is often significant. It should be accounted for in the control design by combining it with the process dynamics.

Discretization

To implement a continuous-time control law, such as a PID controller in a digital computer, it is necessary to approximate the derivatives and the integral that appear in the control law. A few different ways to do this are presented below.

Proportional Action The proportional term is

$$P = K(by_{sp} - y)$$

This term is implemented simply by replacing the continuous variables with their sampled versions. Hence,

$$P(t_k) = K\left(by_{sp}(t_k) - y(t_k)\right) \tag{6.13}$$

where $\{t_k\}$ denotes the sampling instants, i.e., the times when the computer reads the analog input.

Integral Action The integral term is given by

$$I(t) = \frac{K}{T_i}\int_0^t e(s)ds$$

It follows that

$$\frac{dI}{dt} = \frac{K}{T_i}e \tag{6.14}$$

The derivative is approximated by a forward difference gives

$$\frac{I(t_{k+1}) - I(t_k)}{h} = \frac{K}{T_i}e(t_k)$$

This leads to the following recursive equation for the integral term

$$I(t_{k+1}) = I(t_k) + \frac{Kh}{T_i}e(t_k) \tag{6.15}$$

Derivative Action The derivative term is given by Equation (6.2), i.e.

$$\frac{T_d}{N}\frac{dD}{dt} + D = -KT_d\frac{dy}{dt} \tag{6.16}$$

This equation can be approximated in the same way as the integral term. In this case we approximate the derivatives by a backward difference.

$$\frac{T_d}{N}\frac{D(t_k) - D(t_{k-1})}{h} + D(t_k) = -KT_d\frac{y(t_k) - y(t_{k-1})}{h}$$

This can be rewritten as

$$D(t_k) = \frac{T_d}{T_d + Nh} D(t_{k-1}) - \frac{KT_dN}{T_d + Nh} (y(t_k) - y(t_{k-1})) \tag{6.17}$$

The advantage by using a backward difference is that the parameter $T_d/(T_d + Nh)$ is in the range of 0 to 1 for all values of the parameters. This guarantees that the difference equation is stable.

Summarizing we find that the PID controller can be approximated by

$$\begin{aligned} p(t_k) &= k * (br(t_k) - y(t_k)) \\ e(t_k) &= r(t_k) - y(t_k) \\ d(t_k) &= \frac{T_d}{T_d + Nh} \Big(d(t_{k-1}) - kN\big(y(t_k) - y(t_{k-1})\big)\Big) \\ u(t_k) &= p(t_k) + i(t_k) + d(t_k) \\ i(t_{k+1}) &= i(t_k) + \frac{kh}{T_i} e(t_k) \end{aligned}$$

Velocity Algorithms

The algorithms described so far are called positional algorithms because the output of the algorithms is the control variable. In certain cases the control system is arranged in such a way that the control signal is driven directly by an integrator, e.g., a motor. It is then natural to arrange the algorithm in such a way that it gives the velocity of the control variable. The control variable is then obtained by integrating its velocity. An algorithm of this type is called a velocity algorithm. A block diagram of a velocity algorithm for a PID controller is shown in Figure 6.18.

Velocity algorithms were commonly used in many early controllers that were built around motors. In several cases, the structure was retained by the manufacturers when technology was changed in order to maintain functional compatibility with older equipment. Another reason is that many practical issues, like wind-up protection and bumpless parameter changes, are easy to implement using the velocity algorithm. This is discussed further in Sections 6.5 and 6.7. In digital implementations velocity algorithms are also called incremental algorithms.

Incremental algorithm

The incremental form of the PID algorithm is obtained by computing the time differences of the controller output and adding the increments.

$$\Delta u(t_k) = u(t_k) - u(t_{k-1}) = \Delta P(t_k) + \Delta I(t_k) + \Delta D(t_k)$$

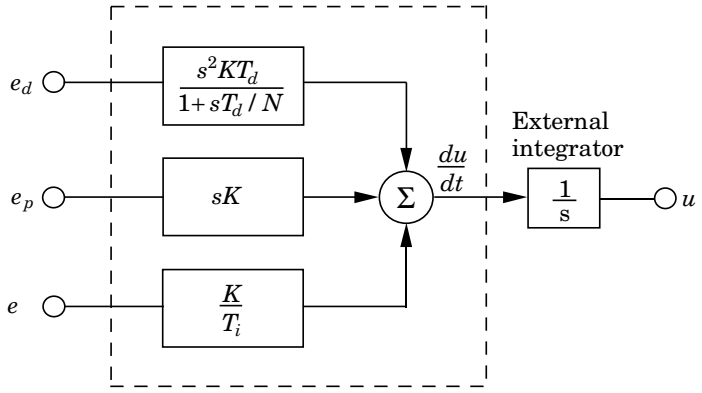


Figure 6.18 Block diagram of a PID algorithm in velocity form.

In some cases integration is performed externally. This is natural when a stepper motor is used. The output of the controller should then represent the increments of the control signal, and the motor implements the integrator. The increments of the proportional part, the integral part, and the derivative part are easily calculated from Equations 6.13, 6.15 and 6.17:

$$\Delta P(t_k) = P(t_k) - P(t_{k-1}) = K\,(by_{sp}(t_k) - y(t_k) - by_{sp}(t_{k-1}) + y(t_{k-1}))$$
$$\Delta I(t_k) = I(t_k) - I(t_{k-1}) = b_{i1}\,e(t_k) + b_{i2}\,e(t_{k-1})$$
$$\Delta D(t_k) = D(t_k) - D(t_{k-1}) = a_d \Delta D(t_{k-1}) - b_d\,(y(t_k) - 2y(t_{k-1}) + y(t_{k-2}))$$

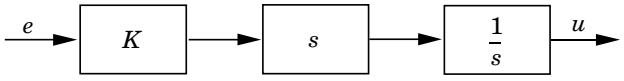
One advantage with the incremental algorithm is that most of the computations are done using increments only. Short word-length calculations can often be used. It is only in the final stage where the increments are added that precision is needed.

Velocity algorithms for controllers without integral action

A velocity algorithm cannot be used directly for a controller without integral action, because such a controller cannot keep the stationary value. This can be understood from the block diagram in Figure 6.19A, which shows a proportional controller in velocity form. Stationarity can be obtained for any value of the control error e, since the output from the derivation block is zero for any constant input. The problem can be avoided with the modification shown in Figure 6.19B. Here, stationarity is only obtained when $u = Ke + u_b$.

If a sampled PID controller is used, a simple version of the method illustrated in figure 6.19B is obtained by implementing the P controller

A



B

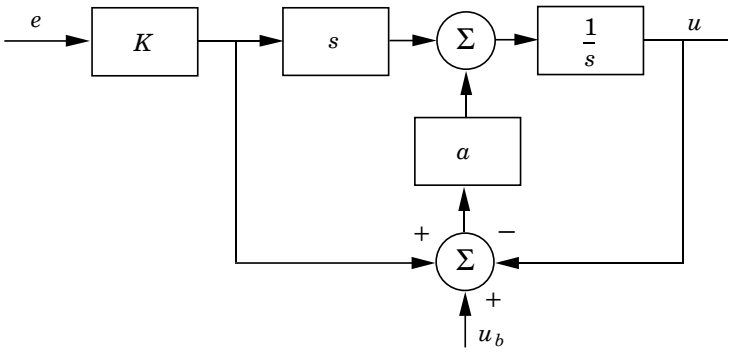


Figure 6.19 Illustrates the difficulty with a proportional controller in velocity form (A) and a way to avoid it (B).

as

$$\Delta u(t) = u(t) - u(t-h) = Ke(t) + u_b - u(t-h)$$

where h is the sampling period.

Feedforward control

In feedforward control, the control signal is composed of two terms,

$$u = u_{FB} + u_{FF}$$

Here u_{FB} is the feedback component and u_{FF} is the feedforward component, either from a measurable disturbance or from the setpoint.

To avoid integrator windup, it is important that the anti-windup mechanism acts on the final control signal u, and not only on the feedback component u_{FB}.

Unfortunately, many of the block-oriented instrument systems available today have the anti-windup mechanisms inside the feedback controller blocks, without any possibility to add feedforward signals to these blocks. Hence, the feedforward signals must be added after the controller blocks. This may lead to windup. Because of this, several tricks, like feeding the feedforward signal through high-pass filters, are used to reduces

the windup problem. These strategies do, however, lead to a less effective feedforward.

Incremental algorithms are efficient for feedforward implementation. By first adding *the increments* of the feedback and feedforward components,

$$\Delta u = \Delta u_{FB} + \Delta u_{FF}$$

and then forming the control signal as

$$u(t) = u(t-h) + \Delta u(t)$$

windup is avoided. This requires that the feedback control blocks have inputs for feedforward signals.

Operational Aspects

Practically all controllers can be run in two modes: manual or automatic. In manual mode the controller output is manipulated directly by the operator, typically by pushing buttons that increase or decrease the controller output. A controller may also operate in combination with other controllers, such as in a cascade or ratio connection, or with nonlinear elements, such as multipliers and selectors. This gives rise to more operational modes. The controllers also have parameters that can be adjusted in operation. When there are changes of modes and parameters, it is essential to avoid switching transients. The way the mode switchings and the parameter changes are made depends on the structure chosen for the controller.

Bumpless Transfer Between Manual and Automatic

Since the controller is a dynamic system, it is necessary to make sure that the state of the system is correct when switching the controller between manual and automatic mode. When the system is in manual mode, the control algorithm produces a control signal that may be different from the manually generated control signal. It is necessary to make sure that the two outputs coincide at the time of switching. This is called *bumpless transfer*.

Bumpless transfer is easy to obtain for a controller in incremental form. This is shown in Figure 6.20. The integrator is provided with a switch so that the signals are either chosen from the manual or the automatic increments. Since the switching only influences the increments there will not be any large transients.

A similar mechanism can be used in the series, or interacting, implementation of a PID controller shown in Figure 6.22. In this case there will be a switching transient if the output of the PD part is not zero at the switching instant.

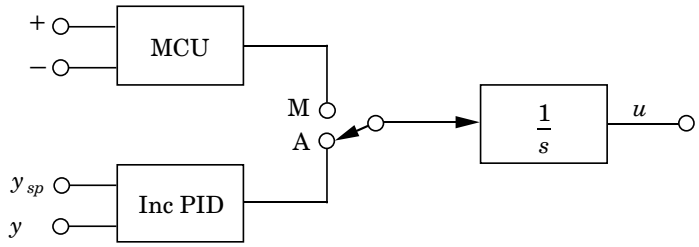


Figure 6.20 Bumpless transfer in a controller with incremental output. MCU stands for manual control unit.

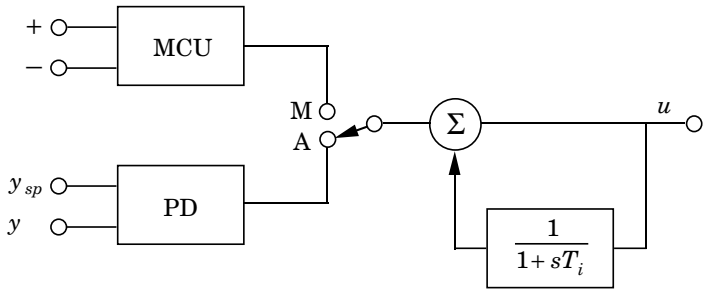


Figure 6.21 Bumpless transfer in a PID controller with a special series implementation.

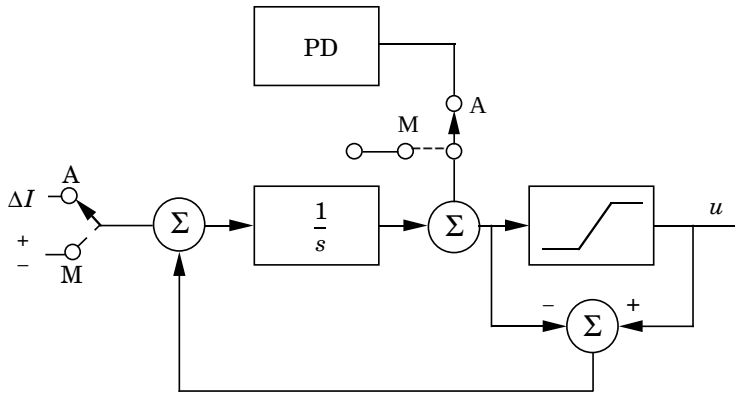


Figure 6.22 A PID controller where one integrator is used both to obtain integral action in automatic mode and to sum the incremental commands in manual mode.

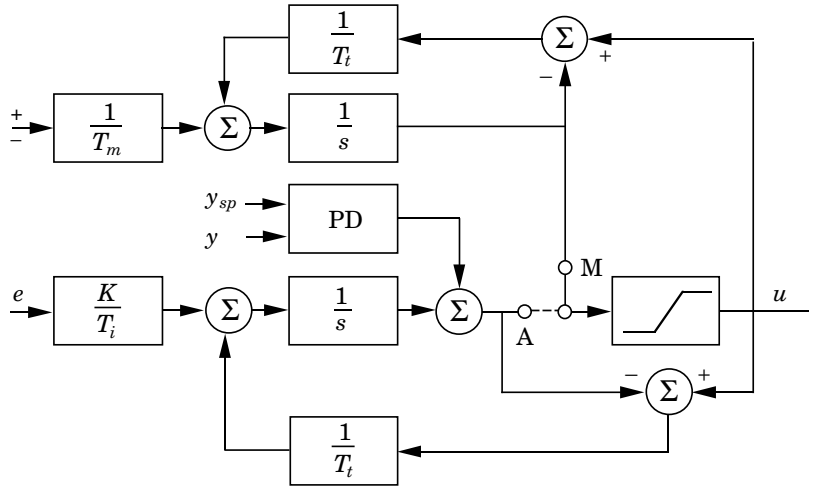


Figure 6.23 PID controller with parallel implementation that switches smoothly between manual and automatic control.

For controllers with parallel implementation, the integrator of the PID controller can be used to add up the changes in manual mode. The controller shown in Figure 6.22 is such a system. This system gives a smooth transition between manual and automatic mode provided that the switch is made when the output of the PD block is zero. If this is not the case, there will be a switching transient.

It is also possible to use a separate integrator to add the incremental changes from the manual control device. To avoid switching transients in such a system, it is necessary to make sure that the integrator in the PID controller is reset to a proper value when the controller is in manual mode. Similarly, the integrator associated with manual control must be reset to a proper value when the controller is in automatic mode. This can be realized with the circuit shown in Figure 6.23. With this system the switch between manual and automatic is smooth even if the control error or its derivative is different from zero at the switching instant. When the controller operates in manual mode, as is shown in Figure 6.23, the feedback from the output v of the PID controller tracks the output u. With efficient tracking the signal v will thus be close to u at all times. There is a similar tracking mechanism that ensures that the integrator in the manual control circuit tracks the controller output.

Bumpless Parameter Changes

A controller is a dynamical system. A change of the parameters of a dynamical system will naturally result in changes of its output. Changes in the output can be avoided, in some cases, by a simultaneous change of the state of the system. The changes in the output will also depend on the chosen realization. With a PID controller it is natural to require that there be no drastic changes in the output if the parameters are changed when the error is zero. This will hold for all incremental algorithms because the output of an incremental algorithm is zero when the input is zero, irrespective of the parameter values. For a position algorithm it depends, however, on the implementation.

Assume that the state is chosen as

$$x_I = \int^t e(\tau)d\tau$$

when implementing the algorithm. The integral term is then

$$I = \frac{K}{T_i}x_I$$

Any change of K or T_i will then result in a change of I. To avoid bumps when the parameters are changed, it is essential that the state be chosen as

$$x_I = \int^t \frac{K(\tau)}{T_i(\tau)}\, e(\tau)d\tau$$

when implementing the integral term.

With sensible precautions, it is easy to ensure bumpless parameter changes if parameters are changed when the error is zero. There is, however, one case where special precautions have to be taken, namely, if setpoint weighting is used. To have bumpless parameter changes in such a case it is necessary that the quantity $P + I$ is invariant to parameter changes. This means that when parameters are changed, the state I should be changed as follows

$$I_{\text{new}} = I_{\text{old}} + K_{\text{old}}(b_{\text{old}}\, y_{sp} - y) - K_{\text{new}}(b_{\text{new}}\, y_{sp} - y)$$

To build automation systems it is useful to have suitable modules. Figure 6.24 shows the block diagram for a manual control module. It has two inputs, a tracking input and an input for the manual control commands. The system has two parameters, the time constant T_m for the manual control input and the reset time constant T_t. In digital implementations

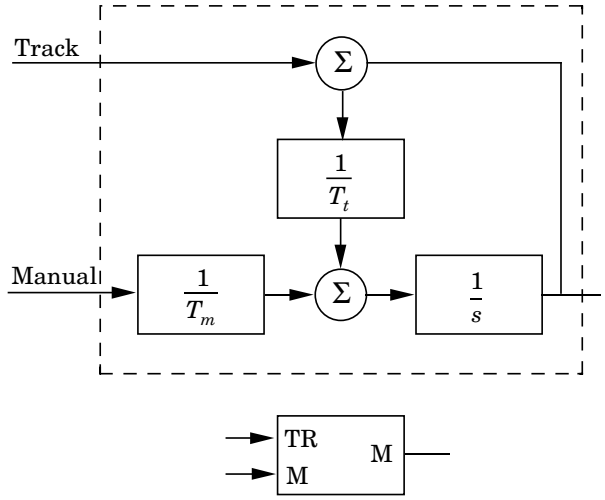


Figure 6.24 Manual control module.

it is convenient to add a feature so that the command signal accelerates as long as one of the increase-decrease buttons are pushed. Using the module for PID control and the manual control module in Figure 6.24, it is straightforward to construct a complete controller. Figure 6.25 shows a PID controller with internal or external setpoints via increase/decrease buttons and manual automatic mode. Notice that the system only has two switches.

Computer Code

As an illustration, the following is a computer code for a PID algorithm. The controller handles both anti-windup and bumpless transfer.

```
"Compute controller coefficients
bi=K*h/Ti                    "integral gain
ad=(2*Td-N*h)/(2*Td+N*h)
bd=2*K*N*Td/(2*Td+N*h)       "derivative gain
a0=h/Tt

"Bumpless parameter changes
I=I+Kold*(bold*ysp-y)-Knew*(bnew*ysp-y)

"Control algorithm
r=adin(ch1)                  "read setpoint from ch1
y=adin(ch2)                  "read process variable from ch2
```

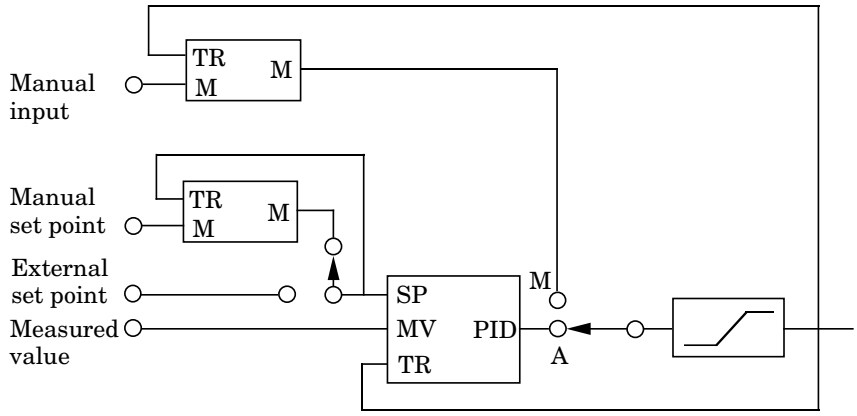


Figure 6.25 A reasonable complete PID controller with anti-windup, automatic-manual mode, and manual and external setpoint.

```
P=K*(b*ysp-y)                "compute proportional part
D=ad*D-bd*(y-yold)           "update derivative part
v=P+I+D                      "compute temporary output
u=sat(v,ulow,uhigh)          "simulate actuator saturation
daout(ch1)                   "set analog output ch1
I=I+bi*(ysp-y)+ao*(u-v)      "update integral
yold=y                       "update old process output
```

The computation of the coefficients should be done only when the controller parameters are changed. Precomputation of the coefficients ad, ao, bd, and bi saves computer time in the main loop. The main program must be called once every sampling period. The program has three states: yold, I, and D. One state variable can be eliminated at the cost of a less readable code. Notice that the code includes derivation of the process output only, proportional action on part of the error only ($b \neq 1$), and anti-windup.

6.8 Summary

In this Section we have given a detailed treatment of the PID controller, which is the most common way controller. A number of practical issues have been discussed. Simple controllers like the PI and PID controller are naturally not suitable for all processes. The PID controller is suitable for processes with almost monotone step responses provided that the

requirements are not too stringent. The quantity

$$m = \frac{\int_0^t g(t)dt}{\int_0^t |g(t)|dt}$$

where $g(t)$ is the impulse response can be used as a measure of monotonicity. PID control is not suitable for processes that are highly oscillatory or when the requirements are extreme.

The PI controller has no phase advance. This means that the PI controller will not work for systems which have phase lag of 180° or more. The double integrator is a typical example. Controller with derivative action can provide phase advance up to about 50°. Simple processes can be characterized by the relative time delay τ introduced in the Ziegler-Nichols tuning procedure. PI control is often satisfactory for processes that are lag dominated, i.e. when τ close to one. Derivative action is typically beneficial for processes with small relative delay τ.

7

Specifications

7.1 Introduction

In this chapter we will discuss how the properties of a control system can be specified. This is important for control design because it gives the goals. It is also important for users of control so that they know how to specify, evaluate and test a system so that they know it will have the desired properties. Specifications on a control systems typically include: stability of the closed loop system, robustness to model uncertainty, attenuation of measurement noise, injection of measurement noise, and ability to follow reference signals. From the results of Chapter 5 it follows that these properties are captured by six transfer functions called the Gang of Six. The specifications can be expressed in terms of these transfer functions. Essential features of the transfer functions can be expressed in terms of their poles and zeros or features of time and frequency responses.

7.2 Stability and Robustness to Process Variations

Stability and robustness to process uncertainties can be expressed in terms of the loop transfer function $L = PC$, the sensitivity function and the complementary sensitivity function

$$S = \frac{1}{1+PC} = \frac{1}{1+L}, \qquad T = \frac{PC}{1+PC} = \frac{L}{1+L}.$$

Since both S and T are functions of the loop transfer function specifications on the sensitivities can also be expressed in terms of specifications on the loop transfer function L. Many of the criteria are based on Nyquist's

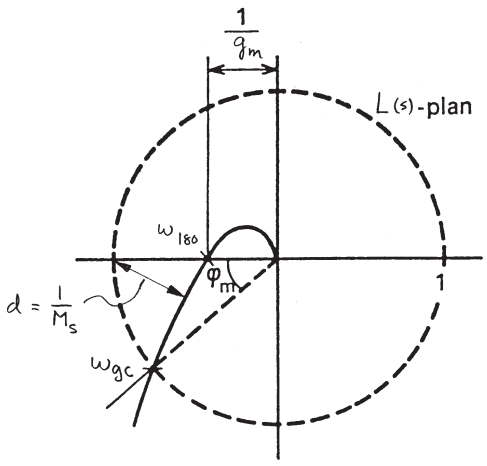


Figure 7.1 Nyquist curve of the loop transfer function L with indication of gain, phase and stability margins.

stability criterion, see Figure 7.1. Common criteria are the maximum values of the sensitivity functions, i.e.

$$M_s = \max_{\omega} |S(i\omega)|, \quad M_t = \max_{\omega} |T(i\omega)|$$

Recall that the number $1/M_s$ is the shortest distance of the Nyquist curve of the loop transfer function to the critical point, see Figure 7.1. Also recall that the closed loop system will remain stable for process perturbations ΔP provided that

$$\frac{|\Delta P(i\omega)|}{|P(i\omega)|} \leq \frac{1}{|T(i\omega)|},$$

see Section 5.5. The largest value M_t of the complementary sensitivity function T is therefore a simple measure of robustness to process variations.

Typical values of the maximum sensitivities are in the range of 1 to 2. Values close to one are more conservative and values close to 2 correspond to more aggressive controllers.

Gain and Phase Margins

The gain margin g_m and the phase margin φ_m are classical stability criteria. Although they can be replaced by the maximum sensitivities it is useful to know about them because they are still often used practically.

The gain margin tells how much the gain has to be increased before the closed loop system becomes unstable and the phase margin tells how much the phase lag has to be increased to make the closed loop system unstable.

The gain margin can be defined as follows. Let ω_{180} be the lowest frequency where the phase lag of the loop transfer function $L(s)$ is 180°. The gain margin is then

$$g_m = \frac{1}{|L(i\omega_{180})|} \tag{7.1}$$

The phase margin can be defined as follows. Let ω_{gc} denote gain crossover frequency, i.e. the lowest frequency where the loop transfer function $L(s)$ has unit magnitude. The phase margin is then given by

$$\varphi_m = \pi + \arg L(i\omega_{gc}) \tag{7.2}$$

The margins have simple geometric interpretations in the Nyquist diagram of the loop transfer function as is shown in Figure 7.1. Notice that an increase of controller gain simply expands the Nyquist curve radially. An increase of the phase of the controller twists the Nyquist curve clockwise, see Figure 7.1.

Reasonable values of the margins are phase margin $\varphi_m = 30^\circ - 60^\circ$, gain margin $g_m = 2 - 5$. Since it is necessary to specify both margins to have a guarantee of a reasonable robustness the margins g_m and φ_m can be replaced by a single stability margin, defined as the shortest distance of the Nyquist curve to the critical point -1, this distance is the inverse of the maximum sensitivity M_s. It follows from Figure 7.1 that both the gain margin and the phase margin must be specified in order to ensure that the Nyquist curve is far from the critical point. It is possible to have a system with a good gain margin and a poor phase margin and vice versa. It is also possible to have a system with good gain and phase margins which has a poor stability margin. The Nyquist curve of the loop transfer function of such a system is shown in Figure 7.2. This system has infinite gain margin, a phase margin of 70° which looks very reassuring, but the maximum sensitivity is $M_s = 3.7$ which is much too high. Since it is necessary to specify both the gain margin and the phase margin to endure robustness of a system it is advantageous to replace them by a single number. A simple analysis of the Nyquist curve shows that the following inequalities hold.

$$\begin{aligned} g_m &\geq \frac{M_s}{M_s - 1} \\ \phi_m &\geq 2 \arcsin \frac{1}{2M_s} \end{aligned} \tag{7.3}$$

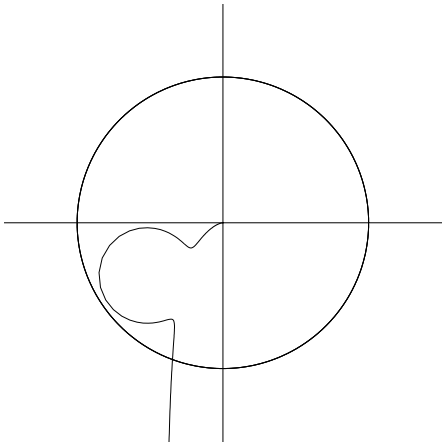


Figure 7.2 Nyquist curve of the loop transfer function for a system with good gain and phase margins but with high sensitivity and poor robustness. The loop transfer function is $L(s) = \frac{0.38(s^2+0.1s+0.55)}{s(s+1)(s^2+0.06s+0.5}$.

A controller with $M_s = 2$ thus has a gain margin of at least 2 and a phase margin of at least 30°. With $M_s = 1.4$ the margins are $g_m \geq 3.5$ and $\phi_m \geq 45°$.

Delay Margin

The gain and phase margins were originally developed for the case when the Nyquist curve only intersects the unit circle and the negative real axis once. For more complicated systems there may be many intersections and it is more complicated to find suitable concepts that capture the idea of a stability margin. One illustration if given in Figure 7.3. In this case the Nyquist curve has a large loop and the Nyquist curve intersects the circle $|L| = 1$ in three points corresponding to the frequencies 0.21, 0.88 and 1.1. If there are variations in the time delay the Nyquist curve can easily enclose the critical point. In the figure it is shown what happens when the time delay is increased from 3 to 4.5 s. This increase corresponds to a phase lag of 0.3 rad at the crossover frequency 0.21 rad/s, the phase lag is however 1.6 rad at the frequency 1.1 rad/s which is marked A in the figures. Notice that the point A becomes very close to the critical point. A good measure of the stability margin in this case is the delay margin which is the smallest time delay required to make the system unstable. For loop transfer functions that decay quickly the delay margin is closely related to the phase margin but for systems where the amplitude ratio of

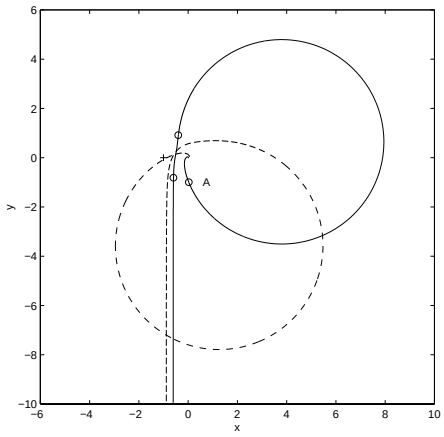


Figure 7.3 Nyquist curve of the loop transfer function $L(s) = \frac{0.2}{s(s^2+0.025s+1)}e^{-3s}$.

the loop transfer function has several peaks at high frequencies the delay margin is a much more relevant measure.

7.3 Disturbances

In the standard system, Figure 5.1, we have used in this book there are two types of disturbances, the load disturbances that drive the system away from its desired behavior and the measurement noise that corrupts the information about the process obtained by the sensors.

Response to Load Disturbances

The response of the process variable to a load disturbance is given by the transfer function

$$G_{xd} = \frac{P}{1+PC} = PS = \frac{T}{C} \tag{7.4}$$

Since load disturbances typically have low frequencies it is natural that the specifications should emphasize the behavior of the transfer function at low frequencies. The loop transfer function $L = PC$ is typically large for small s and we have the approximation

$$G_{xd} = \frac{T}{C} \approx \frac{1}{C}. \tag{7.5}$$

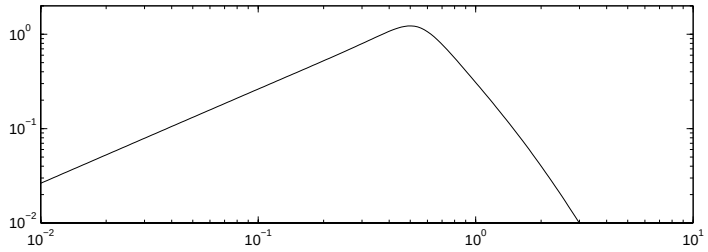


Figure 7.4 Typical gain curve for the transfer function G_{xd} from load disturbance to process output. The gain curve is shown in full lines and the transfer function k_i/s in dotted lines and the process transfer function in full lines.

If $P(0) \neq 0$ and the controller with integral action control we have the following approximation for small s

$$G_{xd} \approx \frac{s}{k_i}.$$

Since load disturbances typically have low frequencies this equation implies that integral gain k_i is a good measure of load disturbance attenuation. Figure 7.4 shows the magnitude curve of the transfer function (7.4) for a PI control of the process $P = (s+1)^{-4}$. The transfer function G_{xd} has typically the form shown in Figure 7.4. The curve can typically be characterized by the low frequency asymptote (k_i), the peak (M_{xd}), the frequency (ω_{xd})where the peak occurs and the high frequency roll-off. It follows from (7.4) that the high frequency behavior is essentially determined by the process and the maximum sensitivity.

Attenuation of load disturbances can also be characterized in the time domain by showing the time response due to a representative disturbance. This is illustrated in 7.5 which shows the response of the process output to a unit step disturbance at the process input. The figure shows maximum error e_{max}, the steady state error e_{ss}, the error of the open loop system e_{ol}, the time to maximum t_{max} and the settling time t_s.

Measurement Noise

An inevitable consequence of using feedback is that measurement noise is fed into the system. Measurement noise thus causes control actions which in turn generate variations in the process variable. It is important to keep these variations of the control signal at reasonable levels. A typical requirement is that the variations are only a fraction of the span of the control signal. The variations in the control variable are also detrimental

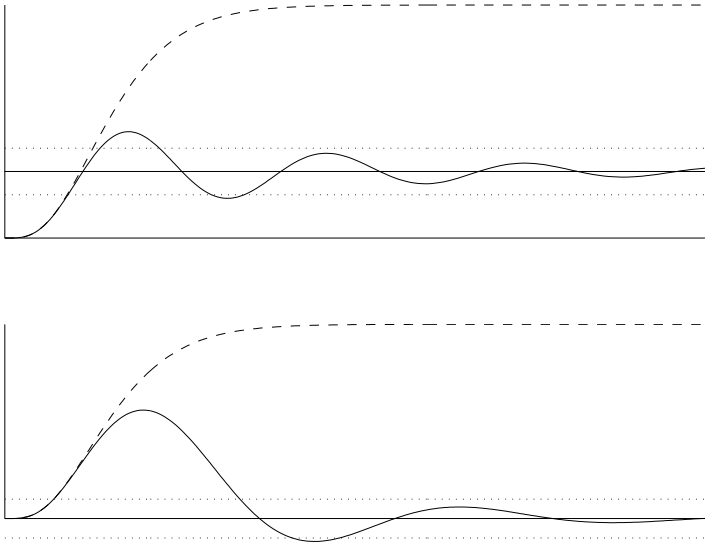


Figure 7.5 Errors due to a unit step load disturbance at the process input and some features used to characterize attenuation of load disturbances. The curves show the open-loop error (dashed lines) and the error (full lines) obtained using a controller without integral action (upper) and with integral action (lower).

by themselves because they cause wear of actuators. Since measurement noise typically has high frequencies the high high frequency gain of the controller is a relevant measure. Notice however that the low frequency gain of the controller is not essential since measurement noise is high frequency. Frequencies above the gain crossover frequency will be regarded as high.

To get a feel for the orders of magnitude consider an analog system where the signal levels are 10V. A measurement noise of 1 mV then saturates the input if the gain is 10^4. If it is only permitted that measurement noise gives control signals of 1V the high frequency gain of the controller must be less than 10^3.

As an other illustration we consider a digital control system with 12 bit AD- and DA-converters. A change of the input of one bit saturates the DA-converter if the gain is 4096. Assume that we permit one bit to give a variation of 0.4% of the output range. The high frequency gain of the controller must then be less than 500. With converters having lower resolution the high frequency gain would be even lower.

High precision analog systems with signal ranges of 1 to 10^4 have been

designed. For digital systems the signal ranges are limited by the sensors and the actuators. Special system architectures with sensors and actuators having multiple signal ranges are used in order to obtain systems with a very high signal resolution. In these cases it is possible to have signal ranges up to 1 to 10^6.

The effects of measurement noise can be evaluated by the transfer function from measurement noise to the control signal, i.e.,

$$G_{un} = \frac{C}{1+PC} = CS = \frac{T}{P}. \tag{7.6}$$

Recall that P and C and are the transfer functions of the process and the controller, and that S is the sensitivity function. Notice that when $L = PC$ is large we have approximately $G_{un} \approx 1/C$. Since measurement noise typically has high frequencies and since the sensitivity function is one for high frequencies we find that the response to measurement noise is essentially determined by the high frequency behavior of the transfer function C. A simple measure is given by

$$M_c = \max_{\omega \geq \omega_{gc}} |G_{un}(i\omega)| \leq M_s \max_{\omega \geq \omega_{gc}} |C(i\omega)|$$

where M_c is called the maximum high frequency gain of the controller. When there is severe measurement noise it is advantageous to make sure that the transfer function C goes to zero for large frequencies. This is called high frequency roll-off.

7.4 Reference Signals

The response to set point changes is described by the transfer functions

$$G_{yr} = \frac{FPC}{1+PC} = FT, \qquad G_{ur} = \frac{FC}{1+PC} = FCS$$

Compare with (5.1). A significant advantage with controller structure with two degrees of freedom is that the problem of set point response can be decoupled from the response to load disturbances and measurement noise. The design procedure can then be divided into two independent steps.

- First design the feedback controller C that reduces the effects of load disturbances and the sensitivity to process variations without introducing too much measurement noise into the system

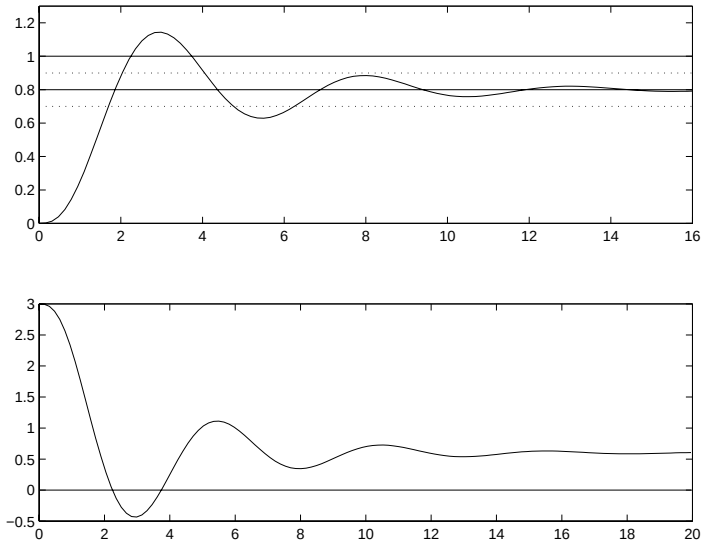


Figure 7.6 Specifications on reference following based on the time response to a unit step in the reference.

- Then design the feedforward F to give the desired response to set points.

Specifications on reference following are typically expressed in the time domain. They may include requirements on rise time, settling time, decay ratio, overshoot, and steady-state offset for step changes in reference. These quantities are defined as follows, see 7.6. These quantities are defined in different ways and there are also different standards.

- The *rise time* t_r is either defined as the inverse of the largest slope of the step response or the time it takes the step to pass from 10% to 90% of its steady state value.

- The *settling time* t_s is the time it takes before the step response remains within p percent of its steady state value. The value $p = 2$ % is commonly used.

- The *delay time* is the time required for the step response to reach 50 % of its steady state value for the first time.

- The *decay ratio* d is the ratio between two consecutive maxima of the error for a step change in reference or load. The value $d = 1/4$, which

is called quarter amplitude damping, has been used traditionally. This value is, however, too high as will be shown later.

- The *overshoot* o is the ratio between the difference between the first peak and the steady state value and the steady state value of the step response. In industrial control applications it is common to specify an overshoot of 8%–10%. In many situations it is desirable, however, to have an over-damped response with no overshoot.
- The *steady-state error* e_{ss} is the value of control error e in steady state. With integral action in the controller, the steady-state error is always zero.

Classical specifications have been strongly focused on the behavior of the process output. It is however important to also consider the control signal. Analogous quantities can be defined for the control signal. The overshoot of the control signal is of particular importance, see Figure 7.4.

Step signals are often used as reference inputs In motion control systems it is often more relevant to consider responses to ramp signals or jerk signals. Specifications are often given in terms of the the value of the first non-vanishing error coefficient.

Tracking Slowly Varying Signals - Error Coefficients

Step signals is one prototype of reference signals. There are however situations when other signals are more appropriate. One example is when the reference signal has constant rate of change, i.e.

$$r(t) = v_0 t$$

The corresponding Laplace transform is $R(s) = v_0/s^2$.

For a system with error feedback the error $e = r - y$ has the Laplace transform

$$E(s) = S(s)V(s) = S(s)\frac{v_0}{s^2} \tag{7.7}$$

The steady state error obtained depends on the properties of the sensitivity function at the origin. If $S(0) = e_0$ the steady state tracking error is asymptotically $e(t) = v_0 e_0 t$. To have a constant tracking error it must be required that $S(0) = 0$. With $S(s) \approx e_1 s$ for small s we find that the steady state error is $e(t) = v_0 e_1$ as t goes to infinity. To have zero steady state error for a ramp signal the function $S(s)$ must go to zero faster than s for small s. If $S(s) \approx e_2 s^2$ for small s we find that the error is asymptotically zero. Since

$$L(s) = \frac{1}{S(s)} - 1$$

it follows that the condition $S(s) \approx e_2 s^2$ implies that $L(s) \approx s^{-2}$ for small s. This implies that there are two integrations in the loop. Continuing this reasoning we find that in order to have zero steady state error when tracking the signal

$$r(t) = \frac{t^2}{2}$$

it is necessary that $s(s) \approx e_3 s^3$ for small s. This implies that there are three integrals in the loop.

The coefficients of the Taylor series expansion of the sensitivity $s(s)$ function for small s,

$$S(s) = e_0 + e_1 s + e_2 s^2 + \ldots + e_n s^n + \ldots \tag{7.8}$$

are thus useful to express the steady state error in tracking low frequency signals. The coefficients e_k are called error coefficients. The first non vanishing error coefficient is the one that is of most interest, this is often called the error coefficient.

7.5 Specifications Based on Optimization

The properties of the transfer functions can also be based on integral criteria. Let $e(t)$ be the error caused by reference values or disturbances and let $u(t)$ be the corresponding control signal. The following criteria are commonly used to express the performance of a control system.

$$\begin{aligned} IE &= \int_0^\infty e(t)dt \\ IAE &= \int_0^\infty |e(t)|dt \\ ITAE &= \int_0^\infty t|e(t)|dt \\ IQ &= \int_0^\infty e^2(t)dt \\ WQ &= \int_0^\infty (e^2(t) + \rho u^2(t))dt \end{aligned}$$

They are called, IE integrated error, IAE integrated absolute error, ITAE integrated time multiplies absolute error, integrated quadratic error and WQ weighted quadratic error. The criterion WQ makes it possible to trade the error against the control effort required to reduce the error.

7.6 Properties of Simple Systems

It is useful to have a good knowledge of properties of simple dynamical systems. In this section we have summarize such data for easy reference.

First Order Systems

Consider a system where the transfer function from reference to output is

$$G(s) = \frac{a}{s+a} \tag{7.9}$$

The step and impulse responses of the system are

$$h(t) = 1 - e^{-at} = 1 - e^{-t/T}$$
$$g(g) = ae^{-at} = \frac{1}{T}e^{-t/T}$$

where the parameter T is the time constant of the system. Simple calculations give the properties of the step response shown in Table 7.1. The 2% settling time of the system is 4 time constants. The step and impulse responses are monotone. The velocity constant e_1 is also equal to the time constant T. This means that there will be a constant tracking error of $e_1 v = v_0 T$ when the input signal is a ramp $r = v_0 t$.

This system (7.9) can be interpreted as a feedback system with the loop transfer function

$$L(s) = \frac{a}{s} = \frac{1}{sT}$$

This system has a gain crossover frequency $\omega_{gc} = a$. The Nyquist curve is the negative imaginary axis, which implies that the phase margin is 90°. Simple calculation gives the results shown in Table 7.1. The load disturbance response of a first order system typically has the form

$$G_{xd} = \frac{s}{s+a}$$

The step response of this transfer function is

$$h_{xd} = e^{-at}$$

The maximum thus occurs when the disturbance is applies and the settling time is $4T$. The frequency response decays monotonically for increasing frequency. The largest value of the gain is a zero frequency.

Some characteristics of the disturbance response are given in Table 7.2.

Table 7.1 Properties of the response to reference values for the first order system $G_{xr} = a/(s+a)$.

Propety	*Value*
Rise time	$T_r = 1/a = T$
Delay time	$T_d = 0.69/a = 0.69T$
Settling time (2%)	$T_s = 4/a = 4T$
Overshoot	$o = 0$
Error coefficients	$e_0 = 0$, $e_1 = 1/a = T$
Bandwidth	$\omega_b = a$
Resonance peak	$\omega_r = 0$
Sensitivities	$M_s = M_t = 1$
Gain margin	$g_m = \infty$
Phase margin	$\varphi_m = 90^\circ$
Crossover frequency	$\omega_{gc} = a$
Sensitivity frequency	$\omega_{sc} = \infty$

Table 7.2 Properties of the response to disturbances for the first order system $G_{xd} = s/(s+a)$.

Property	*Value*
Peak time	$T_p = 0$
Max error	$e_{max} = 1$
Settling time	$T_s = 4T$
Error coefficient $e_1 = T$	
Largest norm	$\|\|G_{xd}\|\| = 1$
Integrated error	$IE = 1/a = T$
Integrated absolute error	$IAE = 1/a = T$

Second Oder System without Zeros

Consider a second order system with the transfer function

$$G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{7.10}$$

The system has two poles, they are complex if $\zeta < 1$ and real if $\zeta > 1$. The step response of the system is

$$h(t) = \begin{cases} 1 - \dfrac{e^{-\zeta\omega_0 t}}{\sqrt{1-\zeta^2}}\sin(\omega_d t + \phi) & \text{for } |\zeta| < 1 \\ 1 - (1+\omega_0 t)e^{-\omega_0 t} & \text{for} \zeta = 1 \\ 1 - \left(\cosh\omega_d t + \dfrac{\zeta}{\sqrt{\zeta^2-1}}\sinh\omega_d t\right)e^{-\zeta\omega_d t} & \text{for } |\zeta| > 1 \end{cases}$$

where $\omega_d = \omega_0\sqrt{|1-\zeta^2|}$ and $\phi = \arccos\zeta$. When $\zeta < 1$ the step response is a damped oscillation, with frequency $\omega_d = \omega_0\sqrt{1-\zeta^2}$. Notice that the step response is enclosed by the envelopes

$$e^{-\zeta\omega_0 t} \leq h(t) \leq 1 - e^{-\zeta\omega_0 t}$$

This means that the system settles like a first order system with time constant $T = \frac{1}{\zeta\omega_0}$. The 2% settling time is thus $T_s \approx \frac{4}{\zeta\omega_0}$. Step responses for different values of ζ are shown in Figure 4.9.

The maximum of the step response occurs approximately at $T_p \approx \pi/\omega_d$, i.e. half a period of the oscillation. The overshoot depends on the damping. The largest overshoot is 100% for $\zeta = 0$. Some properties of the step response are summarized in Table 7.3.

The system (7.10) can be interpreted as a feedback system with the loop transfer function

$$L(s) = \frac{\omega_0^2}{s(s+2\zeta\omega_0)}$$

This means that we can compute quantities such as sensitivity functions and stability margins. These quantities are summarized in Table 7.3.

Second Oder System with Zeros

The response to load disturbances for a second order system with integral action can have the form

$$G(s) = \frac{\omega_0 s}{s^2 + 2\zeta\omega_0 s + \omega_0^2}$$

The frequency response has a maximum $1/(2\zeta)$ at $\omega = \omega_0$. The step response of the transfer function is

$$h(t) = \frac{e^{-\zeta\omega_0 t}}{sqrt1-\zeta^2} sin\omega_d t$$

Table 7.3 Properties of the response to reference values of a second order system.

Property	*Value*
Rise time	$T_r = \omega_0 e^{\phi/\tan\phi} \approx 2.2 T_d$
Delay time	T_d
Peak time	$T_p \approx \pi/\omega_D = T_d/2$
Settling time (2%)	$T_s \approx 4/(\zeta\omega_0)$
Overshoot	$o = e^{-\pi\zeta/\sqrt{1-\zeta^2}}$
Error coefficients	$e_0 = 0,\ e_1 = 2\zeta/\omega_0$
Bandwidth	$\omega_b = \omega_0\sqrt{1-2\zeta^2+\sqrt{(1-2\zeta^2)^2+1}}$
Maximum sensitivity	$M_s = \sqrt{\frac{8\zeta^2+1+(4\zeta\sqrt{2}+1)\sqrt{8\zeta^2+1}}{8\zeta^2+1+(4\zeta\sqrt{2}-1)\sqrt{8\zeta^2+1}}}$
Frequency	$w_{ms} = \frac{1+\sqrt{8\zeta^2+1}}{2}\omega_0$
Max. comp. sensitivity	$M_t = \begin{cases} 1/(2\zeta\sqrt{1-\zeta^2}) & \text{if } \zeta \le \sqrt{2}/2 \\ 1 & \text{if } \zeta \le \sqrt{2}/2 \end{cases}$
Frequency	$\omega_{mt} = \begin{cases} \omega_0\sqrt{1-2\zeta^2} & \text{if } \zeta \le \sqrt{2}/2 \\ 1 & \text{if } \zeta \le \sqrt{2}/2 \end{cases}$
Gain margin	$g_m = \infty$
Phase margin	$\varphi_m = 90^\circ - \arctan\omega_c/(2\zeta\omega_0)$
Crossover frequency	$\omega_{gc} = \omega_0\sqrt{\sqrt{4\zeta^4+1}-2\zeta^2}$
Sensitivity frequency	$\omega_{sc} = \omega_0/\sqrt{2}$

This could typically represent the response to a step in the load disturbance. Figure 7.7 shows the step response for different values of ζ. The step response has its maximum

$$\max_t h(t) = \omega_0 e^{-\zeta//sqrt1-\zeta^2} \tag{7.11}$$

for

$$t = t_m = \frac{\arccos\zeta}{\omega_0}$$

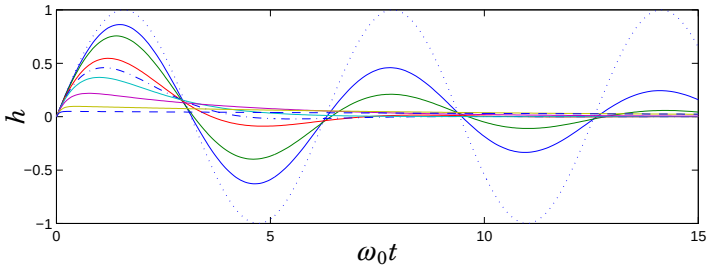


Figure 7.7 Step responses of the transfer function (7.11) for $\zeta = 0$ (dotted), 0.1, 0.2, 0.5, 0.7 (dash-dotted), 1, 2, 5, 10 (dashed).

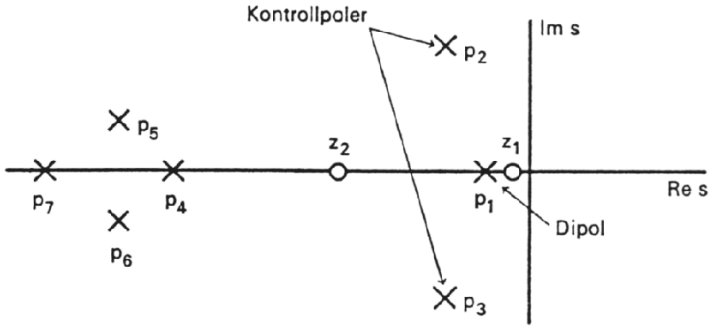


Figure 7.8 Typical configuration of poles and zeros for a transfer function describing the response to reference signals.

Systems of Higher Order

7.7 Poles and Zeros

Specifications can also be expressed in terms of the poles and zeros of the transfer functions. The transfer function from reference value to the output of a system typically has the pole zero configuration shown in Figure 7.8. The behavior of a system is characterized by the poles and zeros with the largest real parts. In the figure the behavior is dominated by a complex pole pair p_1 and p_2 and real poles and zeros. The dominant poles are often characterized by the relative damping ζ and the distance from the origin ω_0. Robustness is determined by the relative damping and the response speed is inversely proportional to ω_0.

- Dominant poles
- Zeros
- Dipoles

7.8 Relations Between Specifications

A good intuition about the different specifications can be obtained by investigating the relations between specifications for simple systems as is given in Tables 7.1, 7.2 and 7.3.

The Rise Time Bandwidth Product

Consider a transfer function $G(s)$ for a stable system with $G(0) \neq 0$. We will derive a relation between the rise time and the bandwidth of a system. We define the rise time by the largest slope of the step response.

$$T_r = \frac{G(0)}{\max_t g(t)} \tag{7.12}$$

where g is the impulse response of G, and let the bandwidth be defined as

$$\omega_b = \frac{\int_0^\infty |G(i\omega)|}{\pi G(0)} \tag{7.13}$$

This implies that the bandwidth for the system $G(s) = 1/(s+1)$ is equal to 1, i.e. the frequency where the gain has dropped by a factor of $1/\sqrt{2}$. The impulse response g is related to the transfer function G by

$$g(t) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{st}G(s)ds = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t}G(i\omega)d\omega$$

Hence

$$\max_t g(t) \leq \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|e^{i\omega t}G(i\omega)\right| d\omega = \frac{1}{\pi}\int_0^\infty |G(i\omega)|d\omega$$

Equations (7.12) and (7.13) now give

$$T_r\omega_b \geq 1$$

This simple calculation indicates that the product of rise time and bandwidth is approximately constant. For most systems the product is around 2.

7.9 Summary

It is important for both users and designers of control systems to understand the role of specifications. The important message is that it is necessary to have specifications that cover properties of the Gang of Six, otherwise there is really no guarantee that the system will work well. This important fact is largely neglected in much of the literature and in control practice. Some practical ways of giving reasonable specifications are summarized.

8

Feedforward Design

8.1 Introduction

Feedforward is a powerful technique that complements feedback. It can be used both to reduce the effect of measurable disturbances and to improve set-point responses. Uses of feedforward was already discussed in connection with systems having two degrees of freedom in Section 6.3. We will now give a systematic treatment of feedforward and also discuss design of model-following systems.

8.2 Disturbance attenuation

Disturbances can be eliminated by feedback. With a feedback system it is, however, necessary that there be an error before the controller can take actions to eliminate disturbances. In some situations it is possible to measure disturbances before they have influenced the processes. It is then natural to try to eliminate the effects of the disturbances before they have created control errors. This control paradigm is called *feedforward*. The principle is illustrated in Figure 8.1.

In Figure 8.1 process transfer function P is composed of two factors, $P = P_1P_2$. A measured disturbance d enters at the input of process section P_2. The measured disturbance is fed to the process input via the feedforward transfer function G_{ff}.

The transfer function from load disturbance to process output is

$$\frac{Y(s)}{D(s)} = \frac{P_2(1 - P_1G_{ff})}{1 + PC} = P_2(1 - P_1G_{ff})S \tag{8.1}$$

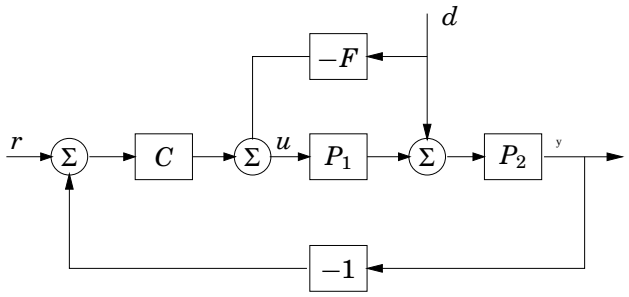


Figure 8.1 Block diagram of a system where measured disturbance d is reduced by a combination of feedback and feedforward.

where $S = 1/(1 + PC)$ is the sensitivity function. This equation shows that there are two ways of reducing the disturbance. We can try to make $1 - P_1 G_{ff}$ small by a proper choice of the feedforward transfer function G_{ff} or we can make the loop transfer function PC large by feedback. Feedforward and feedback can also be combined.

Notice that with feedforward we are trying to make the difference between two terms small but with feedback we simply multiply with a small number. An immediate consequence is that feedforward is more sensitive than feedback. With feedback there is risk of instability, there is no such risk with feedforward. Feedback and feedforward are therefore complementary and it is useful to combine them.

An ideal feedforward compensator is given by

$$G_{ff} = P_1^{-1} = \frac{P_{yd}}{P_{yu}}. \tag{8.2}$$

where P_{yd} is the transfer function from d to y and $P_{yu} = P$ is the transfer function from u to y.

The ideal feedforward compensator is formed by taking the inverse of the process dynamics P_1. This inverse is often not realizable, but approximations have to be used. This problem is treated in the next section.

Feedforward is most effective when the disturbance d enters early in the process. This occurs when most of the dynamics are in process section P_2. When $P_1 = 1$, and therefore $P_2 = P$, the ideal feedforward compensator is realizable and the effects of the disturbance can be eliminated from the process output y. On the other hand, when the dynamics enter late in the process, so that $P_1 \approx P$, the effects of the disturbance are seen in the process output y at the same time as they are seen in the feedforward signal. In this case, there is no advantage of using feedforward compared to feedback.

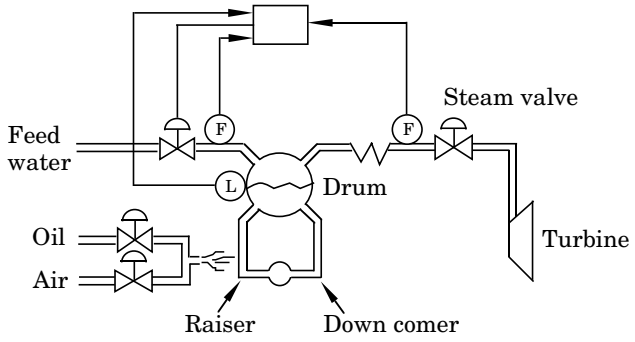


Figure 8.2 Schematic diagram of a drum boiler with level control.

Applications

In many process control applications there are several processes in series. In such cases it is often easy to measure disturbances and use feedforward. Typical applications of feedforward control are: drum-level control in steam boilers, control of distillation columns and rolling mills. An application of combined feedback and feedforward control follows.

EXAMPLE 8.1—DRUM LEVEL CONTROL

A simplified diagram of a steam boiler is shown in Figure 8.2. The water in the raiser is heated by the burners. The steam generated in the raiser, which is lighter than the water, rises toward the drum. This causes a circulation around the loop consisting of the raisers, the drum, and the down comers. The steam is separated from the water in the drum. The steam flow to the turbine is controlled by the steam valve.

It is important to keep the water level in the drum constant. Too low a water level gives insufficient cooling of the raisers, and there is a risk of burning. With too high a water level, water may move into the turbines, which may cause damage. There is a control system for keeping the level constant. The control problem is difficult because of the so-called *shrink and swell effect*. It can be explained as follows: Assume that the system is in equilibrium with a constant drum level. If the steam flow is increased by opening the turbine valve, the pressure in the drum will drop. The decreased pressure causes generation of extra bubbles in the drum and in the raisers. As a result the drum level will initially increase. Since more steam is taken out of the drum, the drum level will of course finally decrease. This phenomena, which is called the *shrink and swell effect*, causes severe difficulties in the control of the drum level. Mathematically it also gives rise to right half plane zero in the transfer function.

The problem can be solved by introducing the control strategy shown in Figure 8.2. It consists of a combination of feedback and feedforward. There is a feedback from the drum level to the controller, but there is also a feedforward from the difference between steam flow and feed-water flow so that the feedwater flow is quickly matched to the steam flow. □

8.3 System inverses

It follows from (8.2) that the feedforward compensator contains an inverse of the process model P_1. A key issue in design of feedforward compensators is thus to find inverse dynamics. It is easy to compute the inverse formally. There are, however, severe fundamental problems in system inversion that are illustrated by the following examples.

EXAMPLE 8.2—INVERSE OF KLT SYSTEM
The system

$$P(s) = \frac{1}{1+sT}e^{-sL}$$

has the formal inverse.

$$P^{-1}(s) = (1+sT)e^{sL}$$

This system is not a causal dynamical system because the term e^{sL} represents a prediction. The term $(1+sT)$ requires and ideal derivative which also is problematic as was discussed in Section 6.3. Implementation of feedforward thus requires approximations. □

EXAMPLE 8.3—INVERSE OF SYSTEM WITH RHP ZERO
The system

$$P(s) = \frac{s-1}{s+1}$$

has the inverse

$$P^{-1}(s) = \frac{s+2}{s-1}$$

Notice that this inverse is an unstable system. □

Use of system inverses and feedforward often gives rise to pole-zero cancellations. The canceled poles and zeros must be stable and sufficiently fast otherwise there will be signals in the system that will grow exponentially or decay very slowly.

To design feedforward we thus have to compute approximate system inverses with suitable properties. Since feedforward is frequently combined

with feedback we can take that into account when finding approximate inverses.

Let $P^\dagger$ denote the approximate inverse of the transfer function P. A common approximation in process control is to neglect all dynamics and simply take the inverse of the static gain, i.e.

$$P^\dagger(s) = P(0)^{-1}$$

A number of results on more accurate system inverses have been derived in system theory. Some of these will be shown here. Note that it follows from (8.2) that the inverse transfer function only has to be small for those frequencies where the sensitivity function is large.

EXAMPLE 8.4—APPROXIMATE INVERSE OF KLT SYSTEM
The system

$$P(s) = \frac{1}{1+sT}e^{-sL}$$

has the approximate inverse.

$$P^\dagger(s) = \frac{1+sT}{1+sT/N}$$

where N gives the frequency range where inversion is valid. □

EXAMPLE 8.5—APPROXIMATE INVERSE OF SYSTEM WITH RHP ZERO
The system

$$P(s) = \frac{s-1}{s+2}$$

has the inverse

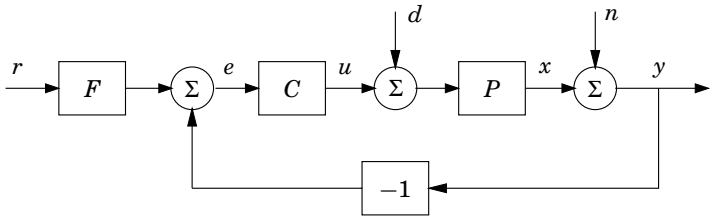
$$P^\dagger(s) = \frac{s+2}{s+1}$$

Notice that the unstable zero in P gives rise to a pole in $P^\dagger$ which is the mirror image of the unstable zero. □

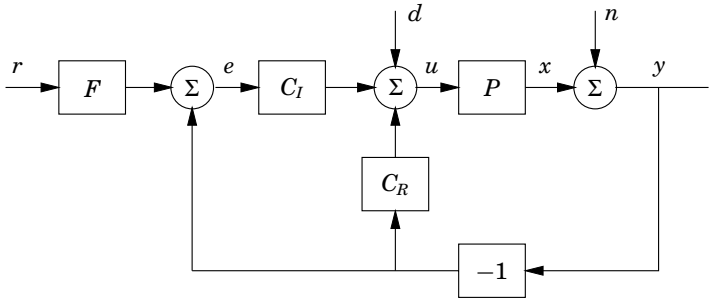
8.4 Response to Reference Inputs

Feedforward can be used very effectively to improve the set point response of the system. This was has been discussed in connection with systems having two degrees of freedom in Section 5.3. Here we will go a little deeper into the design of feedforward compensators.

a



b



c

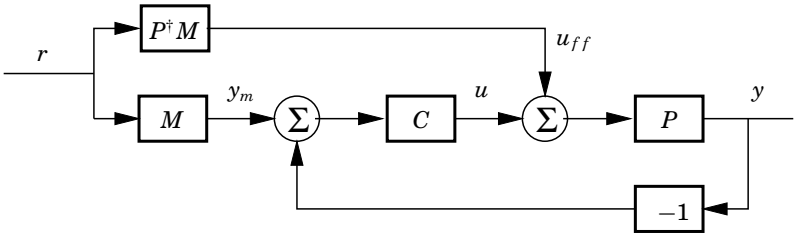


Figure 8.3 Block diagram of three system with two degrees of freedom.

We will start with a system having two degrees of freedom. It is assumed that a feedback controller which gives good rejection of disturbances and good robustness has been designed an we will consider the problem of designing feedforward compensators that give good response to set point changes. There are many possible configurations of a system with two degrees of freedom. Three examples are shown in in Figure 8.3. The block diagram in the top of the figure (a) is the system that was discussed in the previous chapters. The process has the transfer function P

and the feedback controller has the transfer function C. The system in the middle (b) is a slight modification where the controller transfer function is written as

$$C = C_I + C_R$$

where C_I is the term of the controller function that contains integral action and C_R is the rest of the controller transfer function. For the PID controller

$$C = k + \frac{k_i}{s} + k_d s$$

we have

$$C_I = \frac{k_i}{s}$$
$$C_R = k + k_d s$$

The block diagram at the bottom of the figure (c) is yet another configuration. In this system there is an explicit feedforward signal u_{ff} that generates an input to the process that gives the ideal response y_m to a reference input r. This input is generated by the transfer function M. The feedback operates on the error $y_m - y$. It can be shown that all configurations are equivalent if all systems are linear and the transfer functions are chosen properly.

There are however some differences between the systems from a practical point of view. Assume for example that the transfer function M that gives the ideal response to reference inputs is such that the transfer function $P^{-1}M$ is stable with the same number of poles and zeros. It then follows that the scheme c) in Figure 8.3 will give the ideal response to reference signals for all controllers C that stabilize the system. There is thus a clean separation between feedback and feedforward. The scheme a) in Figure 8.3 will give the ideal response if

$$P^{-1}M + CM = CF$$

which implies that in a) the transfer function F must be chosen as

$$F = M\frac{1 + PC}{PC}$$

The transfer function F thus depends on C. This means that if the feedback controller C is changed it is necessary to also change the feedforward F. Notice that the feedforward F must cancel zeros in PC that are not zeros of M. The corresponding equation for scheme b) is

$$F = M\frac{1 + PC}{PC_I}$$

The transfer function $C_I = k_i/s$ does not have any zeros. With this scheme the feedforward F must cancel zeros in P that are not zeros of M.

Notice that in all cases it is necessary to have an invertible process transfer function P. Approximate inverses must be used if this transfer function has RHP zeros or time delays. In scheme c) the process inverse appears in the combination $P^{\dagger}M$. This transfer function will not contain derivatives if the pole excess of P is not larger than the pole excess of M. For example, if

$$P(s) = \frac{1}{s(s+1)}$$
$$M(s) = \frac{1}{s^2+s+1}$$

then

$$P^{-1}(s)M(s) = \frac{s(s+1)}{s^2+s+1}$$

Notice also that in this case the steady state gain of the transfer function $P^{-1}(s)M(s)$ is zero.

8.5 Summary

Design of feedforward has been discussed in this chapter. Feedforward can be used to reduce the effect of measurable disturbances. Design of feedforward is essentially a matter of finding inverse process models. Different techniques to do this have been discussed. The major part of the chapter has been devoted to set point response. A structure with two degrees of freedom has been used. This gives a clean separation of regulation and set point response and of feedback and feedforward. It has been assumed that the feedback controller has been designed. A simple way to modify the set point response is to use set point weighting. If the desired results cannot be obtained by zero set point weighting a full fledged two degree of freedom can be used. This makes is possible to make a complete separation between load disturbance response and set point response. The crucial design issue is to decide the achievable response speed. For systems with monotone set point responses the notion of neutral feedforward has been proposed. Many other variants have also been discussed. Finally it has been demonstrated that very fast set point responses can be obtained by using nonlinear methods.

Special care must be taken when implementing feedforward control, otherwise integrator windup may occur.

9

State Feedback

9.1 Introduction

The state of a dynamical system is a collection of variables that permits prediction of the future development of a system. It is therefore very natural to base control on the state. This will be explored in this chapter. It will be assumed that the system to be controlled is described by a state model. Furthermore it is assumed that the system has one control variable. The technique which will be developed may be viewed as a prototype of an analytical design method. The feedback control will be developed step by step using one single idea, the positioning of closed loop poles in desired locations.

The case when all the state variables are measured is first discussed in Section 9.2. It is shown that if the system is reachable then it is always possible to find a feedback so that the closed loop system has prescribed poles. The controller does not have integral action. This can however be provided by a hack using state augmentation.

In Section 9.3 we consider the problem of determining the states from observations of inputs and outputs. Conditions for doing this are established and practical ways to do this are also developed. In particular it is shown that the state can be generated from a dynamical system driven by the inputs and outputs of the process. Such a system is called an observer. The observer can be constructed in such a way that its state approaches the true states with dynamics having prescribed poles. It will also be shown that the problem of finding an observer with prescribed dynamics is mathematically equivalent to the problem of finding a state feedback.

In Section 9.4 it is shown that the results of Sections 9.2 and 9.3 can be combined to give a controller based on measurements of the process output only. The conditions required are simply that the system is reachable and

observable. This result is important because the controller has a very interesting structure. The controller contains a mathematical model of the system to be controlled. This is called the internal model principle. The solution to the pole placement problem also illustrates that the notions of reachability and observability are essential. The result gives a good interpretation of dynamic feedback. It shows that the dynamics of the controller arises from the need to reconstruct the state of the system.

The controller obtained in Section 9.4 is compared with a PID controller in Section 9.5. The comparison shows that one difference is that the prediction by derivative action in the PID controller is replaced by a better prediction using a mathematical model in the controller in the controller with output feedback. The comparison also shows that the controller with output feedback does not have integral action. The lack of integral action is due to assumptions made when modeling the system. Once this is understood it is easy to modify the model so that integral action is obtained. The modification requires modeling of disturbances which is discussed in Section 9.6. This shows that integral action is obtained when there are constant disturbances. The theory also make it possible to deal with other types of disturbances in a similar way. The approach is based on the idea of state augmentation. This means that the model of the system is augmented by mathematical models that describe disturbances acting on the system.

Following reference signals is discussed in Section 9.7. It is shown that following of reference signals can be completely decoupled from disturbance attenuation by a design which gives a controller having two degrees of freedom. Finally in Section 9.8 we give an example that illustrates the design technique.

The details of the designs in this chapter are carried out for systems with one input and one output. It turns out that the structure of the controller and the forms of the equations are exactly the same for systems with many inputs and many outputs. There are also many other design techniques that give controllers with the same structure. A characteristic feature of a controller with state feedback and an observer is that the complexity of the controller is given by the complexity of the system to be controlled. The controller actually contains a model of the system, the internal model principle.

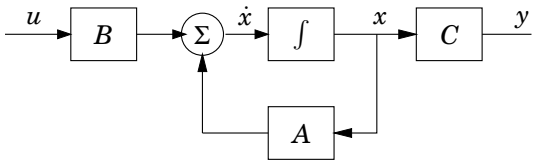


Figure 9.1 Block diagram of the process described by the state model in Equation (9.1).

9.2 State Feedback

Consider a system described by the linear differential equation

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{9.1}$$

A block diagram of the system is shown in Figure 9.1. The output is the variable that we are interested in controlling. To begin with it is assumed that all components of the state vector are measured. Since the state at time t contains all information necessary to predict the future behavior of the system, the most general time invariant control law is function of the state, i.e.

$$u(t) = f(x)$$

If the feedback is restricted to be a linear, it can be written as

$$u = -Lx + L_r r \tag{9.2}$$

where r is the reference value. The negative sign is simply a convention to indicate that negative feedback is the normal situation. The closed loop system obtained when the feedback (9.1) is applied to the system (9.2) is given by

$$\frac{dx}{dt} = (A - BL)x + BL_r r \tag{9.3}$$

It will be attempted to determine the feedback gain L so that the closed loop system has the characteristic polynomial

$$p(s) = s^n + p_1 s^{n-1} + \ldots + p_{n-1}s + p_n \tag{9.4}$$

This control problem is called the pole assignment problem or the pole placement problem.

The transfer function from reference r to output y of the closed loop system is

$$G_{yr}(s) = C(sI - A + BL)^{-1}BL_r \tag{9.5}$$

Requiring that the static gain from reference r to output y should be equal to one gives

$$L_r = \frac{1}{C(-A + BL)^{-1}B} = \frac{1 + LA^{-1}B}{CA^{-1}B} \tag{9.6}$$

The static gain given by (9.6) critically depends on the parameter values. We express this by saying that the system is calibrated. This is different from a system with integral action where the static gain does not depend on the parameter values.

Examples

We will start by considering a few examples that give insight into the nature of the problem.

EXAMPLE 9.1—THE DOUBLE INTEGRATOR
The double integrator is described by

$$\begin{aligned} \frac{dx}{dt} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \\ y &= \begin{pmatrix} 1 & 0 \end{pmatrix} x \end{aligned}$$

Introducing the feedback

$$u = -l_1 x_1 - l_2 x_2 + L_r r$$

the closed loop system becomes

$$\begin{aligned} \frac{dx}{dt} &= \begin{pmatrix} 0 & 1 \\ -l_1 & -l_2 \end{pmatrix} x + \begin{pmatrix} 0 \\ L_r \end{pmatrix} r \\ y &= \begin{pmatrix} 1 & 0 \end{pmatrix} x \end{aligned} \tag{9.7}$$

The closed loop system has the characteristic polynomial

$$det \begin{pmatrix} s & -1 \\ l_1 & s + l_2 \end{pmatrix} = s^2 + l_2 s + l_1$$

Assume it is desired to have a feedback that gives a closed loop system with the characteristic polynomial

$$p(s) = s^2 + 2\zeta\omega_0 s + \omega_0^2$$

Comparing this with the characteristic polynomial of the closed loop system we fin the following values of the We find that the feedback gains should be chosen as

$$l_1 = \omega_0^2, \quad l_2 = 2\zeta\omega_0$$

The closed loop system which is given by Equation (9.7) has the transfer function

$$\begin{aligned} G_{yr}(s) &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} s & -1 \\ l_1 & s + l_2 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ L_r \end{pmatrix} \\ &= \frac{L_r}{s^2 + l_2 s + l_1} = \frac{L_r}{s^2 + 2\zeta_0\omega_0 s + \omega_0^2} \end{aligned}$$

To have unit steady state gain the parameter L_r must be equal to $l_1 = \omega_0^2$. The control law can thus be written as

$$u = l_1(r - x_1) - l_2 x_2 = \omega_0^2(r - x_1) - 2\zeta_0\omega_0 x_2$$

□

In the next example we will encounter some difficulties.

EXAMPLE 9.2—AN UNREACHABLE SYSTEM
Consider the system

$$\begin{aligned} \frac{dx}{dt} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u \\ y = Cx &= \begin{pmatrix} 1 & 0 \end{pmatrix} x \end{aligned}$$

with the control law

$$u = -l_1 x_1 - l_2 x_2 + L_r r$$

The closed loop system is

$$\frac{dx}{dt} = \begin{pmatrix} -l_1 & 1 - l_2 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} L_r \\ 0 \end{pmatrix} r$$

This system has the characteristic polynomial

$$\det \begin{pmatrix} s + l_1 & -1 + l_2 \\ 0 & s \end{pmatrix} = s^2 + l_1 s = s(s + l_1)$$

This polynomial has zeros at $s = 0$ and $s = -l_1$. One closed loop pole is thus always equal to $s = 0$ and it is not possible to obtain an arbitrary characteristic polynomial. □

This example shows that the pole placement problem cannot be solved. An analysis of the equation describing the system shows that the state x_2 is not reachable. It is thus clear that some conditions on the system are required. The reachable and observable canonical forms have the property that the parameters of the system are the coefficients of the characteristic equation. It is therefore natural to consider systems on these forms when solving the pole placement problem. In the next example we investigate the case when the system is in reachable canonical form.

EXAMPLE 9.3—SYSTEM IN REACHABLE CANONICAL FORM
Consider a system in reachable canonical form, i.e,

$$\begin{aligned}\frac{dz}{dt} = \tilde{A}z + \tilde{B}u &= \begin{pmatrix} -a_1 & -a_2 & \dots & -a_{n-1} & -a_n \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} z + \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} u \\ y = \tilde{C}z &= \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix} z \end{aligned} \tag{9.8}$$

The open loop system has the characteristic polynomial

$$D_n(s) = \det \begin{pmatrix} s + a_1 & a_2 & \dots & a_{n-1} & a_n \\ -1 & s & & 0 & 0 \\ 0 & -1 & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & -1 & s \end{pmatrix}$$

Expanding the determinant by the last row we find that the following recursive equation for the determinant.

$$D_n(s) = sD_{n-1}(s) + a_n$$

It follows from this equation that

$$D_n(s) = s^n + a_1 s^{n-1} + \ldots + a_{n-1}s + a_n$$

A useful property of the system described by (9.8) is thus that the coefficients of the characteristic polynomial appear in the first row. Since the all elements of the B-matrix except the first row are zero it follows that the state feedback only changes the first row of the A-matrix. It is

thus straight forward to see how the closed loop poles are changed by the feedback. Introduce the control law

$$u = -\tilde{L}z + L_r r = -\tilde{l}_1 z_1 - \tilde{l}_2 z_2 - \ldots - \tilde{l}_n z_n + L_r r \tag{9.9}$$

The closed loop system then becomes

$$\begin{aligned}
\frac{dz}{dt} &= \begin{pmatrix} -a_1 - \tilde{l}_1 & -a_2 - \tilde{l}_2 & \ldots & -a_{n-1} - \tilde{l}_{n-1} & -a_n - \tilde{l}_n \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & 1 & 0 \end{pmatrix} z + \begin{pmatrix} L_r \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} r \\
y &= \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix} z
\end{aligned} \tag{9.10}$$

The feedback thus changes the elements of the first row of the A matrix, which corresponds to the parameters of the characteristic equation. The closed loop system thus has the characteristic polynomial

$$s^n + (a_l + \tilde{l}_1)s^{n-1} + (a_2 + \tilde{l}_2)s^{n-2} + \ldots + (a_{n-1} + \tilde{l}_{n-1})s + a_n + \tilde{l}_n$$

Requiring this polynomial to be equal to the desired closed loop polynomial (9.4) we find that the controller gains should be chosen as

$$\begin{aligned}
\tilde{l}_1 &= p_1 - a_1 \\
\tilde{l}_2 &= p_2 - a_2 \\
&\vdots \\
\tilde{l}_n &= p_n - a_n
\end{aligned}$$

This feedback simply replace the parameters a_i in the system (9.10) by p_i. The feedback gain for a system in reachable canonical form is thus

$$\tilde{L} = \begin{pmatrix} p_1 - a_1 & p_2 - a_2 & \cdots & p_n - a_n \end{pmatrix} \tag{9.11}$$

The system (9.10) has the following transfer function from reference to output from

$$G_{yr}(s) = \frac{(b_1 s^{n-1} + b_2 s^{n-2} \cdots + b_{n-1}s + b_n)L_r}{s^n + p_1 s^{n-1} + p_2 s^{n-2} + \ldots + p_{n-1}s + p_n}$$

Notice that the system has the same zeros as the open loop system. To have unit steady state gain the parameter L_r should be chosen as

$$L_r = \frac{a_n + \tilde{l}_n}{b_n} = \frac{p_n}{b_n} \tag{9.12}$$

Notice that it is essential to know the precise values of parameters a_n and b_n in order to obtain the correct steady state gain. The steady state gain is thus obtained by precise calibration. This is very different from obtaining the correct steady state value by integral action. We thus find that it is easy to solve the pole placement problem when the system has the structure given by (9.8). □

The General Case

To solve the problem in the general case, we simply change coordinates so that the system is in reachable canonical form. Consider the system (9.1). Change the coordinates by a linear transformation

$$z = Tx$$

so that the transformed system is in observable canonical form (9.8). For such a system the feedback is given by (9.9) where the coefficients are given by (9.11). Transforming back to the original coordinates gives the feedback

$$u = -\tilde{L}z + L_r r = -\tilde{L}Tx + L_r r$$

It now remains to find the transformation. To do this we observe that the reachability matrices have the property.

$$\tilde{W}_r = \begin{pmatrix} \tilde{B} & \tilde{A}\tilde{B} & \dots & \tilde{A}^{n-1}\tilde{B} \end{pmatrix} = T \begin{pmatrix} B & AB & \dots & A^{n-1}B \end{pmatrix} = TW_r$$

The transformation matrix is thus given by

$$T = \tilde{W}_r W_r^{-1} \tag{9.13}$$

and the feedback gain can be written as

$$L = \tilde{L}T = \tilde{L}\tilde{W}_r W_r^{-1} \tag{9.14}$$

Notice that the matrix $\tilde{W}_r$ is given by (3.47). The feedforward gain L_r is given by Equation (9.12).

The results obtained can be summarized as follows.

THEOREM 9.1—POLE-PLACEMENT BY STATE FEEDBACK
Consider the system given by Equation (9.1)

$$\begin{aligned}\frac{dx}{dt} &= Ax + Bu \\ y &= Cx\end{aligned}$$

with one input and one output which has the transfer function

$$G(s) = \frac{(b_1 s^{n-1} + b_2 s^{n-2} \cdots + b_{n-1}s + b_n)L_r}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_{n-1}s + a_n}$$

If the system is reachable there exits a feedback

$$u = -Lx + L_r r$$

that gives a closed loop system with the characteristic polynomial

$$p(s) = s^n + p_1 s^{n-1} + \ldots + p_{n-1}s + p_n$$

The feedback gain is given by

$$\begin{aligned}L &= \tilde{L}T = \begin{pmatrix} p_1 - a_1 & p_2 - a_2 & \ldots & p_n - a_n \end{pmatrix} \tilde{W}_r W_r^{-1} \\ L_r &= \frac{p_n}{a_n}\end{aligned}$$

where a_i are the coefficients of the characteristic polynomial of the matrix A and the matrices W_r and $\tilde{W}_r$ are given by

$$\begin{aligned}W_r &= \begin{pmatrix} B & AB & \ldots & A^{n-1} \end{pmatrix} \\ \tilde{W}_r &= \begin{pmatrix} 1 & a_1 & a_2 & \ldots & a_{n-1} \\ 0 & 1 & a_1 & \ldots & a_{n-2} \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}^{-1}\end{aligned}$$

□

REMARK 9.1—A MATHEMATICAL INTERPRETATION
Notice that the pole-placement problem can be formulated abstractly as the following algebraic problem. Given an $n \times n$ matrix A and an $n \times 1$ matrix B, find a $1 \times n$ matrix L such that the matrix $A - BL$ has prescribed eigenvalues.

Computing the Feedback Gain

We have thus obtained a solution to the problem and the feedback has been described by a closed form solution.

For simple problems is is easy to solve the problem by the following simple procedure: Introduce the elements l_i of L as unknown variables. Compute the characteristic polynomial

$$\det(sI - A + BL)$$

Equate coefficients of equal powers of s to the coefficients of the desired characteristic polynomial

$$p(s) = s^n + p_1 s^{n-1} + \ldots + p_{n-1} + p_n$$

This gives a system of linear equations to determine l_i. The equations can always be solved if the system is observable. Example 9.1 is typical illustrations.

For systems of higher order it is more convenient to use Equation 9.14, this can also be used for numeric computations. However, for large systems this is not sound numerically, because it involves computation of the characteristic polynomial of a matrix and computations of high powers of matrices. Both operations lead to loss of numerical accuracy. For this reason there are other methods that are better numerically. In Matlab the state feedback can be computed by the procedures `acker` or `place`.

```
ACKER  Pole placement gain selection using Ackermann's formula.

   K = ACKER(A,B,P)  calculates the feedback gain matrix K such that
   the single input system
           .
           x = Ax + Bu

   with a feedback law of  u = -Kx  has closed loop poles at the
   values specified in vector P, i.e.,  P = eig(A-B*K).

   Note: This algorithm uses Ackermann's formula.  This method
   is NOT numerically reliable and starts to break down rapidly
   for problems of order greater than 10, or for weakly controllable
   systems.  A warning message is printed if the nonzero closed-loop
   poles are greater than 10\% from the desired locations specified
   in P.

   See also  PLACE.

PLACE  Pole placement technique
```

```
K = PLACE(A,B,P)  computes a state-feedback matrix K such that
the eigenvalues of  A-B*K  are those specified in vector P.
No eigenvalue should have a multiplicity greater than the
number of inputs.

[K,PREC,MESSAGE] = PLACE(A,B,P)  returns PREC, an estimate of how
closely the eigenvalues of A-B*K match the specified locations P
(PREC measures the number of accurate decimal digits in the actual
closed-loop poles).  If some nonzero closed-loop pole is more than
10\% off from the desired location, MESSAGE contains a warning
message.

See also  ACKER.
```

Notice that the program `acker` does not give the static gain L_r, compare (9.6), and that the notation is different from the one we use in the book. This is easily dealt with by writing a new Matlab function `sfb` which computes the gain matrices for the state feedback problem.

```
function [L Lr]=sfb(A,B,C,p)
% Compute gains L  Lr for state feedback
%   A, B and C are the matrices in the state model
%   p is a vector of desired closed loop poles
L=acker(A,B,p);
Lr=1/(C*((-A+B*L)\B);
```

This program also computes the gain L_r.

Integral Action

The controller (9.2) does not have integral action. The correct steady state response to reference values was obtained by a proper choice of the gain L_r, i.e. a calibrated procedure. Compare with (9.6). This means that the controller is not useful practically. One way to ensure that the output will equal the reference in steady state is enforce integral action. One way to do this is to introduce the integral of the error as an extra state, i.e.

$$x_{n+1} = \int (y - r)dt$$

Differentiating this equation gives

$$\frac{dx_{n+1}}{dt} = y - r = Cx - r$$

Augmenting the state by considering x_{n+1} as an extra state variable we

find that the augmented system can be described by the equation

$$\begin{aligned}\frac{d\xi}{dt} &= \frac{d}{dt}\begin{pmatrix} x \\ x_{n+1} \end{pmatrix} \\ &= \begin{pmatrix} A & 0 \\ C & 0 \end{pmatrix}\begin{pmatrix} x \\ x_{n+1} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} u + \begin{pmatrix} 0 & -I \end{pmatrix} r \\ &= \bar{A}\xi + \bar{B}r \end{aligned} \tag{9.15}$$

This equation has the same form as the original system (9.1). The reachability matrix of the augmented system is

$$W_r = \begin{pmatrix} B & AB & \dots & A^nB \\ 0 & CB & \dots & CA^{n-1}B \end{pmatrix}$$

To find the conditions for W_r to be of full rank the matrix will be transformed by making column operations. Let a_k be the coefficients of the characteristic polynomial of the matrix A. Multiplying the first column by a_n, the second by a_{n-1} and the (n-1)th column by a_1 and adding to the last column the matrix W_r it follows from the Cayley-Hamilton theorem that the transformed matrix becomes

$$W_r = \begin{pmatrix} B & AB & \dots & 0 \\ 0 & CB & \dots & b_n \end{pmatrix}$$

where

$$b_n = C(A^{n-1}B + a_1A^{n-2}B + \ldots + a_{n-1}B) \tag{9.16}$$

Notice that b_n can be identified with a coefficient of the transfer function of the original system

$$G(s) = \frac{b_1s^{n-1} + b_2s^{n-2} + \ldots + b_n}{s^n + a_1s^{n-1} + \ldots + a_n}$$

Notice that $b_n = a_nG(0) = -a_nCA^{-1}B$. Compare with Section 3.7.

The state feedback for the augmented system (9.15) is

$$u = -\bar{L}\xi = -\begin{pmatrix} L & L_I \end{pmatrix}\begin{pmatrix} x \\ x_{n+1} \end{pmatrix} + L_r r \tag{9.17}$$

and the closed loop system becomes

$$\frac{d}{dt}\begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} A - BL & -BL_I \\ C & 0 \end{pmatrix}\begin{pmatrix} x \\ x_{n+1} \end{pmatrix} + \begin{pmatrix} BL_r \\ -1 \end{pmatrix} r$$

If the reference is constant r_0 and the closed loop system is stable it follows that

$$y_0 = Cx_0 = r_0$$

The steady state output is thus equal to the reference for all values of the gain L_r. This is no surprise because the control law (9.17) can be written as

$$u(t) = -Lx(t) - L_I \int_0^t (y(\tau) - r(\tau))d\tau + L_r r(t) = L_r r(t) + L_I \int_0^t e(\tau)d\tau - Lx(t) \tag{9.18}$$

and it clearly has integral action. Compare with Equation (2.4). This comparison also shows that the term $-Lx(t)$ is a generalization of derivative action.

Summary

It has been found that the control problem is simple if all states are measured. The most general feedback is a static function from the state space to space of controls. A particularly simple case is when the feedback is restricted to be linear, because it can then be described as a matrix or a vector in the case of systems with only one control variable. A method of determining the feedback gain in such a way that the closed loop system has prescribed poles has been given. This can always be done if the system is reachable. A method of obtaining integral action was also introduced. A comparison with the PID controller showed that state feedback can be interpreted as a PID controller where the derivative is replaced by a better prediction based on the state of the system.

9.3 Observers

In Section 9.2 it was shown that the pole it was possible to find a feedback that gives desired closed loop poles provided that the system is reachable and that all states were measured. It is highly unrealistic to assume that all states are measured. In this section we will investigate how the state can be estimated by using the mathematical model and a few measurements. Before reading this section it is recommended to refresh the material on observability in Section 3.7. In that section it was shown that the state could be computed from the output if the the system is observable. In this Section we will develop some practical methods for determining the state of the system from inputs and outputs. It will be shown that the computation of the states can be done by dynamical systems. Such systems will be called observers.

Consider a system described by

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{9.19}$$

where x is the state, u the input, and y the measured output. The problem of determining the state of the system from its inputs and outputs will be considered. It will be assumed that there is only one measured signal, i.e. that the signal y is a scalar and that C is a vector.

Observers Based on Differentiation

An observer based on differentiation will first be given. The construction is an extension of the derivation of the criterion for observability in Section 3.7.

First observe that the output equation

$$y = Cx$$

gives the projection of the state on the vector C. Differentiation of this equation gives

$$\frac{dy}{dt} = C\frac{dx}{dt} = CAx + CBu$$

The derivative of the output together with CBu thus gives the projection of the state vector on the vector CA. Proceeding in this way and taking higher derivatives give the projections of the state vector on the vectors $C, CA, \ldots, CA^{n-1}$. If these vectors are linearly independent, the projections of the state on n linearly independent vectors are obtained and the state can thus be determined. Carrying out the details, we get

$$\begin{aligned} y &= Cx \\ \frac{dy}{dt} &= Cdxdt = CAx + CBu \\ \frac{d^2y}{dt^2} &= CA\frac{dx}{dt} + CB\frac{du}{dt} = CA^2x + CABu + CB\frac{du}{dt} \\ &\vdots \\ \frac{d^{n-1}y}{dt^{n-1}} &= CA^{n-1}x + CA^{n-2}Bu + CA^{n-3}B\frac{du}{dt} + \ldots + CB\frac{d^{n-2}u}{dt^{n-2}} \end{aligned}$$

This equation can be written in matrix form as

$$\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} x = \begin{pmatrix} y \\ \frac{dy}{dt} - CBu \\ \vdots \\ \frac{d^{n-1}y}{dt^{n-1}} - CA^{n-2}Bu - CA^{n-3}B\frac{du}{dt} - \ldots - CB\frac{d^{n-2}u}{dt^{n-2}} \end{pmatrix}$$

Notice that the matrix on the left-hand side is the observability matrix W_o. If the system is observable, the equation can be solved to give

$$x = W_o^{-1} \begin{pmatrix} y \\ \frac{dy}{dt} \\ \vdots \\ \frac{d^{n-1}}{dt^{n-1}} \end{pmatrix} - W_o^{-1} \begin{pmatrix} 0 & 0 & \cdots & 0 \\ CB & 0 & \cdots & 0 \\ \vdots & & & \\ CA^{n-2}B & CA^{n-3}B & \cdots & CB \end{pmatrix} \begin{pmatrix} u \\ \frac{du}{dt} \\ \vdots \\ \frac{d^{n-2}}{dt^{n-2}} \end{pmatrix} \tag{9.20}$$

This is an exact expression for the state. The state is obtained by differentiating inputs and outputs. Notice that it has been derived under the assumption that there is no measurement noise. Differentiation can give very large errors when there is measurement noise and the method is therefore not very practical particularly when derivatives of high order appear.

Using a Dynamical Systems to Observe the State

For a system governed by Equation (9.19), it can be attempted to determine the state simply by simulating the equations with the correct input. An estimate of the state is then given by

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu \tag{9.21}$$

To find the properties of this estimate, introduce the estimation error

$$\tilde{x} = x - \hat{x}$$

It follows from (9.19) and (9.21) that

$$\frac{d\tilde{x}}{dt} = A\tilde{x}$$

If matrix A has all its eigenvalues in the left half plane, the error $\tilde{x}$ will thus go to zero. Equation (9.21) is thus a dynamical system whose output converges to the state of the system (9.19).

The observer given by (9.21) uses only the process input u, the measured signal does not appear in the equation. It must also be required that the system is stable. We will therefore attempt to modify the observer so that the output is used and that it will work for unstable systems. Consider the following

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + K(y - C\hat{x}) \tag{9.22}$$

observer. This can be considered as a generalization of (9.21). Feedback from the measured output is provided by adding the term $K(y - C\hat{x})$. Notice that $C\hat{x} = \hat{y}$ is the output that is predicted by the observer. To investigate the observer (9.22), form the error

$$\tilde{x} = x - \hat{x}$$

It follows from (9.19) and (9.22) that

$$\frac{d\tilde{x}}{dt} = (A - KC)\tilde{x}$$

If the matrix K can be chosen in such a way that the matrix $A - KC$ has eigenvalues with negative real parts, error $\tilde{x}$ will go to zero. The convergence rate is determined by an appropriate selection of the eigenvalues.

The problem of determining the matrix K such that $A - KC$ has prescribed eigenvalues is very similar to the pole placement problem that was solved in Section 3.7. In fact, if we observe that the eigenvalues of the matrix and its transpose are the same, we find that could determine K such that $A^T - C^T K^T$ has given eigenvalues. First we notice that the problem can be solved if the matrix

$$\begin{pmatrix} C^T & A^T C^T & \dots & A^{(n-1)T} C^T \end{pmatrix}$$

is invertible. Notice that this matrix is the transpose of the observability matrix for the system (9.19).

$$W_o = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}$$

of the system. Assume it is desired that the characteristic polynomial of the matrix $A - KC$ is

$$p(s) = s^n + p_1 s^{n-1} + \ldots + p_n$$

It follows from Remark 9.1 of Theorem 9.1 that the solution is given by

$$K^T = \begin{pmatrix} p_1 - a_1 & p_2 - a_2 & \dots & p_n - a_n \end{pmatrix} \tilde{W}_o^T W_o^{-T}$$

where W_o is the observability matrix and $\tilde{W}_o$ is the observability matrix of the system of the system

$$\frac{dz}{dt} = \begin{pmatrix} -a_1 & 1 & 0 & \dots & 0 \\ -a_2 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ -a_{n-1} & 0 & 0 & \dots & 1 \\ -a_n & 0 & 0 & \dots & 0 \end{pmatrix} z + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix} u$$
$$y = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \end{pmatrix}$$

which is the observable canonical form of the system (9.19). Transposing the formula for K we obtain

$$K = W_o^{-1} \tilde{W}_o \begin{pmatrix} p_1 - a_1 \\ p_2 - a_2 \\ \vdots \\ p_n - a_n \end{pmatrix}$$

The result is summarized by the following theorem.

THEOREM 9.2—OBSERVER DESIGN BY POLE PLACEMENT
Consider the system given by

$$\frac{dx}{dt} = Ax + Bu$$
$$y = Cx$$

where output y is a scalar. Assume that the system is observable. The dynamical system

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + K(y - C\hat{x})$$

with K chosen as

$$K = W_o^{-1} \tilde{W}_o \begin{pmatrix} p_1 - a_1 \\ p_2 - a_2 \\ \vdots \\ p_n - a_n \end{pmatrix} \tag{9.23}$$

where the matrices W_o and $\tilde{W}_o$ are given by

$$W_o = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}, \quad \tilde{W}_o^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ a_1 & 1 & 0 & \dots & 0 \\ a_2 & a_1 & 1 & \dots & 0 \\ \vdots & & & & \\ a_{n-1} & a_{n-2} & a_{n-3} & \dots & 1 \end{pmatrix}$$

Then the observer error $\tilde{x} = x - \hat{x}$ is governed by a differential equation having the characteristic polynomial

$$p(s) = s^n + p_1 s^{n-1} + \ldots + p_n$$

□

REMARK 9.2
The dynamical system (9.22) is called an observer for (the states of the) system (9.19) because it will generate an approximation of the states of the system from its inputs and outputs.

REMARK 9.3
The theorem can be derived by transforming the system to observable canonical form and solving the problem for a system in this form.

REMARK 9.4
Notice that we have given two observers, one based on pure differentiation (9.20) and another described by the differential equation (9.22). There are also other forms of observers.

Interpretation of the Observer

The observer is a dynamical system whose inputs are process input u and process output y. The rate of change of the estimate is composed of two terms. One term $A\hat{x} + Bu$ is the rate of change computed from the model with $\hat{x}$ substituted for x. The other term $K(y - \hat{y})$ is proportional to the difference $e = y - \hat{y}$ between measured output y and its estimate $\hat{y} = C\hat{x}$. The estimator gain K is a matrix that tells how the error e is weighted and distributed among the states. The observer thus combines measurements with a dynamical model of the system. A block diagram of the observer is shown in Figure 9.2.

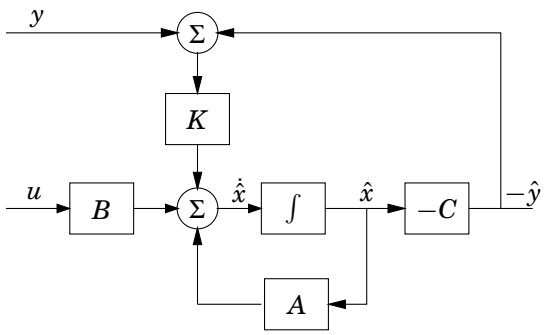


Figure 9.2 Block diagram of the observer. Notice that the observer contains a copy of the process.

Duality

Notice the similarity between the problems of finding a state feedback and finding the observer. The key is that both of these problems are equivalent to the same algebraic problem. In pole placement it is attempted to find L so that $A - BL$ has given eigenvalues. For the observer design it is instead attempted to find K so that $A - KC$ has given eigenvalues. The following equivalence can be established between the problems

$$A \leftrightarrow A^T$$
$$B \leftrightarrow C^T$$
$$L \leftrightarrow K^T$$
$$W_r \leftrightarrow W_o^T$$

The similarity between design of state feedback and observers also means that the same computer code can be used for both problems. To avoid mistakes it is however convenient to have a special code for computing the observer gain. This can be done by the following Matlab program.

```
function K=obs(A,C,p)
% Compute observer gain K
%   A and C are the matrices in the state model
%   p is a vector of desired observer poles
K=acker(A',B',p)';
```

Computing the Observer Gain

The observer gain can be computed in several different ways. For simple problems it is convenient to introduce the elements of K as unknown parameters, determine the characteristic polynomial of the observer $\det(A - KC)$

and identify it with the desired characteristic polynomial. Another alternative is to use the fact that the observer gain can be obtained by inspection if the system is in observable canonical form. In the general case the observer gain is then obtained by transformation to the canonical form. There are also reliable numerical algorithms. They are identical to the algorithms for computing the state feedback. The procedures are illustrated by a few examples.

EXAMPLE 9.4—THE DOUBLE INTEGRATOR
The double integrator is described by

$$\frac{dx}{dt} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u$$
$$y = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

The observability matrix is

$$W_o = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

i.e. the identity matrix. The system is thus observable and the problem can be solved. We have

$$A - KC = \begin{pmatrix} -k_1 & 1 \\ -k_2 & 0 \end{pmatrix}$$

It has the characteristic polynomial

$$\det A - KC = \det \begin{pmatrix} s + k_1 & -1 \\ -k_2 & s \end{pmatrix} = s^2 + k_1 s + k_2$$

Assume that it is desired to have an observer with the characteristic polynomial

$$s^2 + p_1 s + p_2 = s^2 + 2\zeta\omega s + \omega^2$$

The observer gains should be chosen as

$$k_1 = p_1 = 2\zeta\omega$$
$$k_2 = p_2 = \omega^2$$

The observer is then

$$\frac{d\hat{x}}{dt} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \hat{x} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u + \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} (y - \hat{x}_1)$$

□

9.4 Output Feedback

In this section we will consider the same system as in the previous sections, i.e. the nth order system described by

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{9.24}$$

where only the output is measured. As before it will be assumed that u and y are scalars. It is also assumed that the system is reachable and observable. In Section 9.2 we had found a feedback

$$u = -Lx + L_r r$$

for the case that all states could be measured and in Section 9.3 we have presented developed an observer that can generate estimates of the state $\hat{x}$ based on inputs and outputs. In this section we will combine the ideas of these sections to find an feedback which gives desired closed loop poles for systems where only outputs are available for feedback.

If all states are not measurable, it seems reasonable to try the feedback

$$u = -L\hat{x} + L_r r \tag{9.25}$$

where $\hat{x}$ is the output of an observer of the state (9.22) ,i.e.

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + K(y - C\hat{x}) \tag{9.26}$$

Since the system (9.24) and the observer (9.26) both are of order n, the closed loop system is thus of order $2n$. The states of the system are x and $\hat{x}$. The evolution of the states is described by equations (9.24), (9.25)(9.26). To analyze the closed loop system, the state variable $\hat{x}$ is replace by

$$\tilde{x} = x - \hat{x} \tag{9.27}$$

Subtraction of (9.24) from (9.24) gives

$$\frac{d\tilde{x}}{dt} = Ax - A\hat{x} - K(y - C\hat{x}) = A\tilde{x} - KC\tilde{x} = (A - KC)\tilde{x}$$

Introducing u from (9.25) into this equation and using (9.27) to eliminate $\hat{x}$ gives

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu = Ax - BL\hat{x} + BL_r r = Ax - BL(x - \tilde{x}) + BL_r r \\ &= (A - BL)x + BL\tilde{x} + BL_r r \end{aligned}$$

The closed loop system is thus governed by

$$\frac{d}{dt}\begin{pmatrix} x \\ \tilde{x} \end{pmatrix} = \begin{pmatrix} A - BL & BL \\ 0 & A - KC \end{pmatrix}\begin{pmatrix} x \\ \tilde{x} \end{pmatrix} + \begin{pmatrix} BL_r \\ 0 \end{pmatrix} r \tag{9.28}$$

Since the matrix on the right-hand side is block diagonal, we find that the characteristic polynomial of the closed loop system is

$$\det(sI - A + BL)\det(sI - A + KC)$$

This polynomial is a product of two terms, where the first is the characteristic polynomial of the closed loop system obtained with state feedback and the other is the characteristic polynomial of the observer error. The feedback (9.25) that was motivated heuristically thus provides a very neat solution to the pole placement problem. The result is summarized as follows.

THEOREM 9.3—POLE PLACEMENT BY OUTPUT FEEDBACK
Consider the system

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx \end{aligned}$$

The controller described by

$$\begin{aligned} u &= -L\hat{x} + L_r r \\ \frac{d\hat{x}}{dt} &= A\hat{x} + Bu + K(y - C\hat{x}) \end{aligned}$$

gives a closed loop system with the characteristic polynomial

$$\det(sI - A + BL)\det(sI - A + KC)$$

This polynomial can be assigned arbitrary roots if the system is observable and reachable. □

REMARK 9.5
Notice that the characteristic polynomial is of order $2n$ and that it can naturally be separated into two factors, one $\det(sI - A + BL)$ associated with the state feedback and the other $\det(sI - A + KC)$ with the observer.

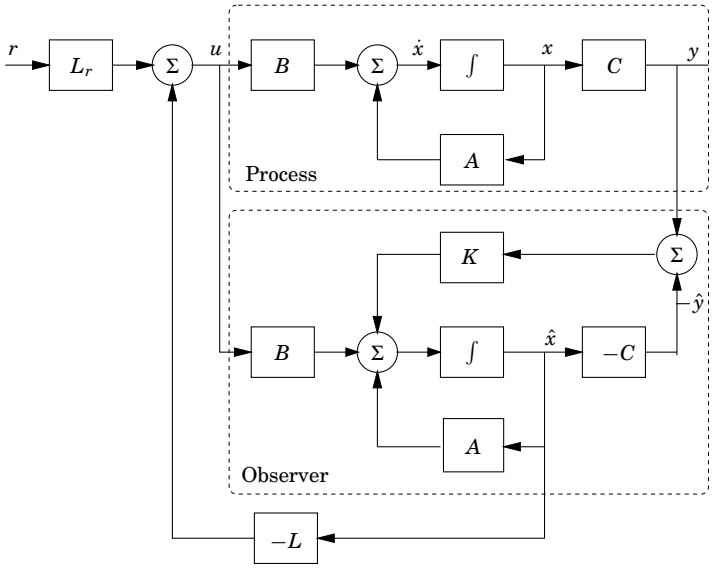


Figure 9.3 Block diagram of a controller which combines state feedback with an observer.

REMARK 9.6
The controller has a strong intuitive appeal. It can be thought of as composed of two parts, one state feedback and one observer. The feedback gain L can be computed as if all state variables can be measured.

The Internal Model Principle

A block diagram of the controller is shown in Figure 9.3. Notice that the controller contains a dynamical model of the plant. This is called the internal model principle. Notice that the dynamics of the controller is due to the observer. The controller can be viewed as a dynamical system with input y and output u.

$$\frac{d\hat{x}}{dt} = (A - BL - KC)\hat{x} + Ky$$
$$u = -L\hat{x} + L_r r$$

The controller has the transfer function

$$C(s) = L[sI - A + BL + KC]^{-1}K \tag{9.29}$$

This can be compared with the controller obtained by the frequency response method. Notice that the order of the controller is equal to the order

of the system. A complex model thus gives a complex controller. Also notice that it follows from (9.29) that the transfer function of the controller has the property that $C(s)$ goes to zero at least as fast as s^{-1} for large s. The approach thus results in controllers which have high frequency roll-off. Compare with the discussion of frequency response in Section 5.5 where it was found that controllers with high frequency roll-off were desirable in order to make systems less sensitive to model uncertainty at high frequencies.

Properties of the Closed Loop System

To get insight into the properties of the closed loop system we will calculate the transfer functions of the gang of four. For this purpose we introduce a load disturbance d at the process input and measurement noise n at the process output, see the block diagram in Figure 9.3. The closed loop system is then described by the equations

$$\begin{aligned}\frac{dx}{dt} &= Ax + B(u + d)\\ u &= -L\hat{x} + L_r r = -Lx + L\tilde{x} + L_r r\\ \frac{d\hat{x}}{dt} &= A\hat{x} + Bu + K(Cx + n - C\hat{x})\end{aligned}$$

Introducing the error $\tilde{x} = x - \hat{x}$ this equation can be written as

$$\begin{aligned}\frac{dx}{dt} &= (A - BL)x + BL\hat{x} + Bd + BL_r r = A_s x + BL\hat{x} + Bd + BL_r r\\ \frac{d\tilde{x}}{dt} &= (A - KC)\tilde{x} + Bd - Kn = A_o\tilde{x} + Bd - Kn\\ u &= -Lx + L\tilde{x} + L_r r\\ y &= Cx + n\end{aligned}$$

Notice that this equation has a triangular structure since the equation for $\tilde{x}$ does not depend on x. Also notice that the reference signal does not influence $\tilde{x}$. This makes a lot of sense because it would be highly undesirable to have a system where reference inputs generate observer errors.

EXAMPLE 9.5—OUTPUT FEEDBACK OF DOUBLE INTEGRATOR
Consider a system described by

$$\begin{aligned}\frac{dx}{dt} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u\\ y &= \begin{pmatrix} 1 & 0 \end{pmatrix} x\end{aligned}$$

A state feedback was determined in Example 9.1 and and observer was computed in Example 9.4. Assume that it is desired to have a closed loop system with the characteristic polynomial

$$(s^2 + 2\zeta_c\omega_c s + \omega_c^2)(s^2 + 2\zeta_o\omega_o s + \omega_o^2)$$

Assuming that the first factor is chosen as the characteristic polynomial associated with the state feedback, i.e. $A - BL$, and the second with the observer, i.e. $A - KC$, it follows from the results of the previous examples that the controller

$$\begin{aligned} \frac{d\hat{x}}{dt} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \hat{x} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u + \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} (y - \hat{x}_1) \\ u &= -l_1\hat{x}_1 - l_2\hat{x}_2 + L_r r \end{aligned}$$

gives the desired closed loop polynomial, provided that

$$\begin{aligned} l_1 &= \omega_c^2, & k_1 &= 2\zeta_o\omega_o \\ l_2 &= 2\zeta_c\omega_c, & k_2 &= \omega_o^2 \end{aligned}$$

It follows from (9.29) that the transfer function from y to u of the controller is

$$\begin{aligned} C(s) &= L[sI - A + BL + KC]^{-1}K \\ &= \frac{\omega_c\omega_o(2(\zeta_c\omega_o + \zeta_o\omega_c)s + \omega_c\omega_o)}{s^2 + 2s(\zeta_c\omega_c + \zeta_o\omega_o) + \omega_c^2 + \omega_o^2 + 4\zeta_1\zeta_2\omega_1\omega_2} \end{aligned}$$

Notice that this transfer function is invariant to permutations of the indices 1 and 2. This means that the controller is the same if the state feedback was designed to give the characteristic polynomial

$$s^2 + 2\zeta_2\omega_2 s + \omega_2$$

and the observer was designed to give the characteristic polynomial

$$s^2 + 2\zeta_1\omega_1 s + \omega_1^2$$

□

It can be shown that the observation about the association of poles to state feedback and observers is true in general and we can draw the following important conclusion: *Although it is convenient to split the design problem into two parts, design of a state feedback and design of an observer,*

the controller transfer function is uniquely given by all the specified closed loop poles. It does not matter if a pole is allocated by the state feedback or the observer, all assignments will give the same transfer function from y to u! Transfer functions between other signal pairs do however depend on which poles are associated with the observer and the state feedback. If reference values are introduced as described by (9.25) the transfer function from reference r to output y is uniquely determined by the poles assigned to the state feedback.

9.5 Comparison with PID Control

The controller obtained by combining an observer with state feedback will now be compared with a PID controller. This gives a perspective on what has been done and it also reveals a significant drawback of the controller which is caused by assumptions made when modeling the system. With this insight it is possible to make a simple modification which gives a much more practical controller.

The input-output relation for an ideal PID controller can be written as

$$u = k(br - y) + k_i \int_o^t (r(\tau) - y(\tau))d\tau - T_d \frac{dy}{dt}$$

in Section 2.2 the derivative term was interpreted as a prediction of the output based on linear extrapolation, see Figure 2.2. From Equation (9.20) it also follows that the derivative term can be interpreted as an estimate of one state variable. The first three terms of the PID controller can thus be interpreted as a controller with feedback from two states y and dy/dt and the reference signal r.

The controller obtained by feedback from the observed states has the form

$$u = -L\hat{x} + L_r r = -l_1\hat{x}_1 - l_2\hat{x}_2 - \ldots - l_n\hat{x}_n + L_r r \qquad (9.30)$$

The differences between this controller and a PID controller are that there are more terms and there is no integral action in (9.30). The estimates $\hat{x}_i$ in (9.30) are filtered through the estimator.

We can thus conclude that a PD controller can be interpreted as a controller with feedback from two states where the first state is the output and the second state is an estimate of the derivative of the output. We can also conclude that the controller based on feedback from estimated states lacks integral action.

A Calibrated System

The controller based on state feedback achieves the correct steady state response to reference signals by careful calibration of the gain L_r and that it lacks the nice property of integral control. It is then natural to ask why the the beautiful theory of state feedback and observes does not automatically give controllers with integral action. This is a consequence of the assumptions made when deriving the analytical design method which we will now investigate.

When using an analytical design method, we postulate criteria and specifications, and the controller is then a consequence of the assumptions. In this case the problem is the model (9.1). This model assumes implicitly that the system is perfectly calibrated in the sense that the output is zero when the input is zero. In practice it is very difficult to obtain such a model. Consider, for example, a process control problem where the output is temperature and the control variable is a large rusty valve. The model (9.1) then implies that we know exactly how to position the valve to get a specified outlet temperature–indeed, a highly unrealistic assumption.

Having understood the difficulty it is not too difficult to change the model. By modifying the model to

$$\begin{aligned} \frac{dx}{dt} &= Ax + B(u + v) \\ y &= Cx \end{aligned} \tag{9.31}$$

where v is an unknown constant we can can capture the idea that the model is no longer perfectly calibrated. This model is called a model with an input disturbance. Another possibility is to use the model

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu \\ y &= Cx + v \end{aligned}$$

where v is an unknown constant. This is a model with an output disturbance. It will now be shown that a straight forward design of an output feedback for this system does indeed give integral action

Integral Action by Disturbance Estimation

Both disturbance models will produce controllers with integral action. We will start by investigating the case of an input disturbance. This is a little more convenient for us because it fits the control goal of finding a controller that drives the state to zero.

The model with an input disturbance can conveniently be brought into the framework of state feedback. To do this, we first observe that v is an

unknown constant which can be described by

$$\frac{dv}{dt} = 0$$

To bring the system into the standard format we simply introduce the disturbance v as an extra state variable. The state of the system is thus

$$z = \begin{pmatrix} x \\ v \end{pmatrix}$$

This is also called state augmentation. Using the augmented state the model (9.31) can be written as

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} x \\ v \end{pmatrix} &= \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} u \\ y &= \begin{pmatrix} C & 0 \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix} \end{aligned} \tag{9.32}$$

Notice that the disturbance state is not reachable. If the disturbance can be measured, the state feedback is then

$$u = -\tilde{L}z + L_r r = -L_x x - L_v v + L_r r \tag{9.33}$$

h The disturbance state v is not reachable. The the effect of the disturbance on the system can, however, be eliminated by choosing $L_v = 1$. If the disturbance v is known the The control law above can be interpreted as a combination of feedback from the system state and feedforward from a measured disturbance. It is not realistic to assume that the disturbance can be measured and we will instead replace the states by estimates. The feedback law then becomes

$$u = -L_x \hat{z} + L_r r = -L_x \hat{x} - \hat{v} + L_r r$$

This means that feedback is based on estimates of the state and the disturbance.

Observability of Augmented Model

Before attempting to estimate the state we will investigate if the system is observable. To do this we form the observability matrix

$$W_o = \begin{pmatrix} C & 0 \\ CA & CB \\ CA^2 & CAB \\ \vdots & \\ CA^{n-1} & CA^{n-2}B \\ CA^n & CA^{n-1}B \end{pmatrix}$$

The first n rows of this matrix are linearly independent if the original system is observable. Let $a_1, a_2, \ldots, a_n$ be the coefficients of the characteristic polynomial of A. To find out if the last row is linearly independent, we multiply the rows of the matrix with $a_n, a_{n-1}, a_{n-2}, \ldots, a_1$ and add to the last row. It follows from Cayley-Hamilton Theorem that

$$\bar{W}_o = \begin{pmatrix} C & 0 \\ CA & CB \\ CA^2 & CAB \\ \vdots & \\ CA^{n-1} & CA^{n-2}B \\ 0 & b_n \end{pmatrix}$$

where

$$b_n = CA^{n-1}B + a_1 CA^{n-2}B + \ldots + a_{n-2}CAB + a_{n-1}CB = a_n G(0)$$

The last row is linearly independent of the first n rows if the parameter b_n is different from zero. We can thus conclude that the state can be observed if the original system is observable and if parameter b_n is different from zero. Notice that the condition for observability of the input disturbance is the same as the condition (9.16) for reachability of the augmented system used to introduce integral action in Section 9.2.

Observer and Control Law

The observer of the system (9.32) is given by

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} &= \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} u + \begin{pmatrix} K_x \\ K_v \end{pmatrix}(y - C\hat{x}) \\ &= \begin{pmatrix} A - K_x C & B \\ -K_v & 0 \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} u + \begin{pmatrix} K_x \\ K_v \end{pmatrix} y \end{aligned}$$

where matrix K can be chosen to give arbitrary eigenvalues to the matrix

$$\begin{pmatrix} A - K_x C & B \\ -K_v & 0 \end{pmatrix}$$

The controller can be written as

$$\begin{aligned} u &= -L_x\hat{x} - \hat{v} + L_r r \\ \frac{d}{dt}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} &= \begin{pmatrix} A - BL_x - K_x C & 0 \\ -K_v C & 0 \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} K_x \\ K_v \end{pmatrix} y + \begin{pmatrix} B \\ 0 \end{pmatrix} r \end{aligned}$$

Notice that the system matrix of this equation has zero as an eigenvalue, which implies that the controller has integral action. The transfer function of the controller is given by

$$\begin{aligned} C(s) &= \begin{pmatrix} L_x & 1 \end{pmatrix} \begin{pmatrix} sI - A + BL_x + K_xC & 0 \\ K_v & s \end{pmatrix}^{-1} \begin{pmatrix} K_x \\ K_v \end{pmatrix} \\ &= \frac{1}{s}K_v + (L_x - \frac{1}{s}K_vC)(sI - A + BL_x + K_xC)^{-1}K_x \end{aligned}$$

The controller has integral action and the integral gain is

$$K_i = K_v\big(1 + C(A - BL_x - K_xC)^{-1}K_x\big)$$

We thus find that integral action can be obtained by introducing a constant but unknown disturbance at the process input. This state is be observable under very natural conditions. The control law then obtained by the feedback from the estimated states. In this way integral action can thus be interpreted as an estimate from an estimated disturbance. The observer gives an explicit estimate of the disturbance. Notice that the augmented system (9.32) is not reachable because the disturbance states are unreachable, the effect of the disturbance on the state can however be reduced.

EXAMPLE 9.6—THE DOUBLE INTEGRATOR
Consider the double integrator. The following model is obtained by augmenting the model by an input disturbance

$$\frac{dx}{dt} = \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} u$$

The observer becomes

$$\begin{aligned} \frac{d\hat{x}}{dt} &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \hat{x} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} u + \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix} (y - \hat{x}_1) \\ &= \begin{pmatrix} -k_1 & 1 & 0 \\ -k_2 & 0 & 1 \\ -k_3 & 0 & 0 \end{pmatrix} \hat{x} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} u + \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix} y \end{aligned} \tag{9.34}$$

and the feedback law is

$$u = -l_1\hat{x}_1 - l_2\hat{x}_2 - \hat{x}_3 + L_r r \tag{9.35}$$

The characteristic polynomial of the observer is

$$s^3 + k_1 s^2 + k_2 s + k_3$$

Requiring this to be equal to

$$(s + \alpha\omega)(s^2 + 2\zeta\omega s + \omega^2)$$

gives the following expressions for the components of the observer gain K

$$\begin{aligned} k_1 &= (\alpha + 2\zeta)\omega \\ k_2 &= (1 + 2\alpha\zeta)\omega^2 \\ k_3 &= \alpha\omega^3 \end{aligned}$$

Eliminating u between Equations (9.34) and (9.35) gives the following expression for the controller

$$\begin{aligned} u &= -l_1\hat{x}_1 - l_2\hat{x}_2 - \hat{x}_3 \\ \frac{d\hat{x}}{dt} &= \begin{pmatrix} -k_1 & 1 & 0 \\ -k_2 - l_1 & -l_2 & 0 \\ -k_3 & 0 & 0 \end{pmatrix} \hat{x} + \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix} y \end{aligned}$$

Straightforward calculations give the following transfer function

$$C(s) = \frac{s^2(k_1 l_1 + k_2 l_2 + k_3) + s(k_2 l_1 + k_3 l_2) + l_1 k_3}{s(s^2 + s(k_1 + l_2) + k_1 l_2 + k_2 + l_1)} \tag{9.36}$$

The transfer function has two zeros and three poles, where one pole is at the origin. □

9.6 Disturbance Models

The results will now be generalized. Before doing this we will discuss the problem formulation that has been used. When a pole placement controller is designed the chosen closed loop characteristic polynomial determines the way the state goes to zero. It seems intuitively reasonable that the controller also would react well to disturbances of short duration provided that the disturbances are so widely spaced that the state goes to zero between them. The modification of the controller done in Section 9.5 shows that integral action is obtained when an input disturbance of constant but unknown amplitude is introduced in the model. In this section it the results for constant disturbances will be generalized to different types of disturbances.

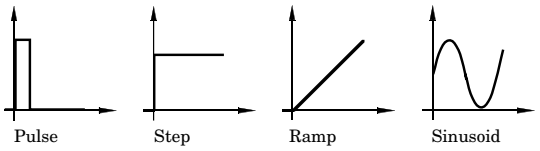


Figure 9.4 Classical disturbance models.

Examples of Disturbances

A suitable characterization of disturbances will first be given. To do this we will first give some simple example of disturbances and their mathematical models.

Some prototype disturbances like pulses, steps, ramps and sinusoids, see Figure 9.4 , have traditionally been used to model disturbances. These disturbances can all be described by simple dynamical systems as is illustrated by the following examples.

EXAMPLE 9.7—STEP DISTURBANCE
Consider a constant disturbance $v = c$,where a is a constant. Taking the derivative gives

$$\frac{dv}{dt} = 0, \quad v(0) = c$$

□

EXAMPLE 9.8—RAMP DISTURBANCE
Consider a disturbance that is affine, i.e. $v = c_1 + c_2 t$ Differentiating twice gives

$$\frac{d^2v}{dt^2} = 0$$

The disturbance is thus represented by a second order differential equation. It can also be represented by

$$\begin{aligned} \frac{d\xi}{dt} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \xi, \quad \xi(0) = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \xi \\ v &= \begin{pmatrix} 1 & 0 \end{pmatrix} \xi \end{aligned}$$

□

EXAMPLE 9.9—SINUSOIDAL DISTURBANCE
Consider a sinusoidal disturbance with known frequency ω, i.e

$$v(t) = c_1 \sin \omega t + c_2 \cos \omega t$$

Differentiation twice gives

$$\frac{d^2v}{dt^2} = -\omega^2 a \sin \omega t = -\omega^2 v$$

This differential equation can also be written as

$$\begin{aligned} \frac{d\xi}{dt} &= \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \xi, \quad \xi(0) = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \xi \\ v &= \begin{pmatrix} 1 & 0 \end{pmatrix} \xi \end{aligned}$$

□

These examples show that simple disturbances can be represented as differential equations with initial conditions. In all examples the shape of the disturbance is given by the differential equation and the amplitude is determined by the initial conditions.

A general approach is to model disturbances as

$$\begin{aligned} \frac{d\xi}{dt} &= A_v \xi \\ v &= C_v \xi \end{aligned} \tag{9.37}$$

It is natural that eigenvalues of A_ξ are on the imaginary axis or in the right half plane, because this will represent disturbances that do not go to zero as time goes to infinity.

Assuming that the disturbances act on the process input the system can be described by

$$\begin{aligned} \frac{dx}{dt} &= Ax + B(u + v) \\ \frac{d\xi}{dt} &= A_v \xi \\ v &= C_v \xi \end{aligned} \tag{9.38}$$

This system can be brought to standard form by introducing the augmented state

$$z = \begin{pmatrix} x \\ \xi \end{pmatrix}$$

The system equations then become

$$\frac{dz}{dt} = \begin{pmatrix} A & B \\ 0 & A_v \end{pmatrix} z + \begin{pmatrix} B \\ 0 \end{pmatrix} u$$

This system is in standard form. Assuming that all states are measurable it is natural to use the feedback

$$u = -Lx - v + L_r r$$

Notice that the disturbance state is not reachable. This is natural because the disturbance model (9.37) is not influenced by the control signal. In spite of this the effect of the disturbances on the process can be eliminated by proper choice of the gain L_v. Combining the control law with the feedback law we obtain the following closed loop system

$$\frac{dx}{dt} = Ax + B(u + v) = (A - BL)x + B(1 - L_v)v + L_r r$$

and it is clear that the disturbance can be eliminated by choosing $L_v = 1$. The term $L_\xi \xi$ can be interpreted as a feedforward from a measured disturbances.

It is not realistic to assume that the disturbance can be measured. In this case we can use an observer to obtain an estimate of the disturbance. The control law then becomes

$$\begin{aligned} u &= -L\hat{x} - \hat{v} + L_r r \\ \frac{d}{dt}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} &= \begin{pmatrix} A & B \\ 0 & A_v \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} u + \begin{pmatrix} K_x \\ K_v \end{pmatrix}(y - C\hat{x}) \end{aligned} \tag{9.39}$$

Eliminating the control signal u the controller becomes

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} &= \begin{pmatrix} A - BL - K_x C & 0 \\ -K_v C & A_v \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} BL_r \\ 0 \end{pmatrix} r + \begin{pmatrix} K_x \\ K_v \end{pmatrix} y \\ u &= -L\hat{x} - \hat{v} + L_r r \end{aligned}$$

The transfer function from measured signal y to control u is

$$\begin{aligned} C(s) &= C_v(sI - A_v)^{-1}K_v \\ &\quad + \big(L - C_v(sI - A_v)^{-1}K_v\big)(sI - A + BL + K_v C)^{-1}K_x \end{aligned}$$

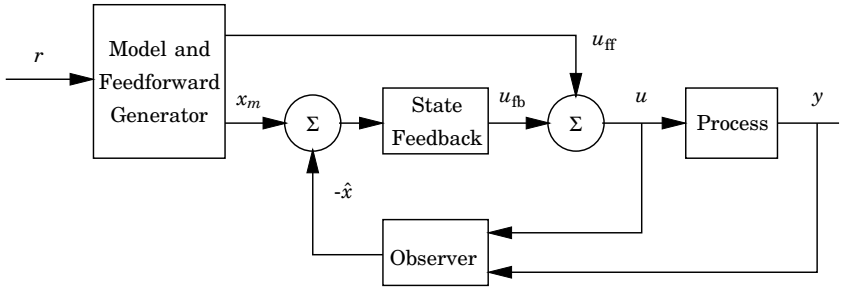


Figure 9.5 Block diagram of a controller based on a structure with two degrees of freedom. The controller consists of a command signal generator, state feedback and an observer.

9.7 Reference Signals

So far we have only introduced the reference signals in a very primitive way by adding it to the state feedback. When discussing simple control system in Sections 4.4 and 8.4 we found that response to command signals could be completely decoupled from disturbance rejection by using a controller structure having two degrees of freedom. In this section we will present such a structure for a system with state feedback.

A block diagram of the system is shown in Figure 9.5. Notice that the system admits independent specification of response to command signals and disturbances. The response to command signals is given by the response model which can be represented by a transfer function M. The response to disturbances is given by the state feedback and the observer.

The system in Figure 9.5 is a natural version of the two-degree of freedom structure for systems with state feedback. To get some insight into the behavior of the system let us reason through what happens when the command signal is changed. To fix the ideas let us assume that the system is in equilibrium with the observer state equal to the process state. When the command signal is changed a feedforward signal is generated. This signal has the property that the process output gives the desired output when the feedforward signal is applied. The process state naturally changes in response to the feedforward signal. The observer will track the state perfectly because the initial state was correct. This means that the estimated state will be equal to the desired model state which implies the feedback signal $L(x_m - \hat{x})$ is zero. If there are some disturbances or some modeling errors the feedback signal will be different from zero and attempt to correct the situation.

Details

Having described the key idea we will now consider the details. Let the process be described by (9.1) and let $P(s)$ denote the process transfer function, Furthermore let $M(s)$ denote the transfer function defining the ideal response to the reference signal. The feedforward signal is then given by

$$U_{ff}(s) = M(s)P^{\dagger}(s)R(s) \tag{9.40}$$

where $P^{\dagger}(s)$ is an inverse or a generalized inverse of $P(s)$. Approximate process models can often be used to determine $P^{\dagger}(s)$, see Section 8.3. To generate the feedforward signals a system with the transfer function $M(s)P^{-1}(s)$ must be realized. Notice that this transfer function can be implemented even if P^{-1} is not realizable. Also notice that since the feedforward signal is most useful for the rapid response it is possible to make several approximations, compare with Section 8.3.

The ideal response of the states to the signal u_{ff} can be generated from the process model, (9.1) hence

$$\begin{aligned} \frac{dx_m}{dt} &= Ax_m + Bu_{\text{ff}} \\ y_m &= Cx_m \end{aligned} \tag{9.41}$$

The control signal then becomes

$$u = u_{\text{ff}} + L(x_m - \hat{x}) \tag{9.42}$$

and the complete controller is given by

$$\begin{aligned} u &= u_{\text{ff}} + L(x_m - \hat{x}) \\ \frac{dx_m}{dt} &= Ax_m + Bu_{\text{ff}} \\ y_m &= Cx_m \\ \frac{d\hat{x}}{dt} &= A\hat{x} + Bu + K(y - C\hat{x}) \end{aligned}$$

The dynamics required to generate u_{ff} from the reference input is also requred. This dynamics which has the transfer function $P^{\dagger}M$ can often be generated from the states x_m as will be illustrated by the examples in the following.

This controller given above has the state vectors x_m and $\hat{x}$. Replacing

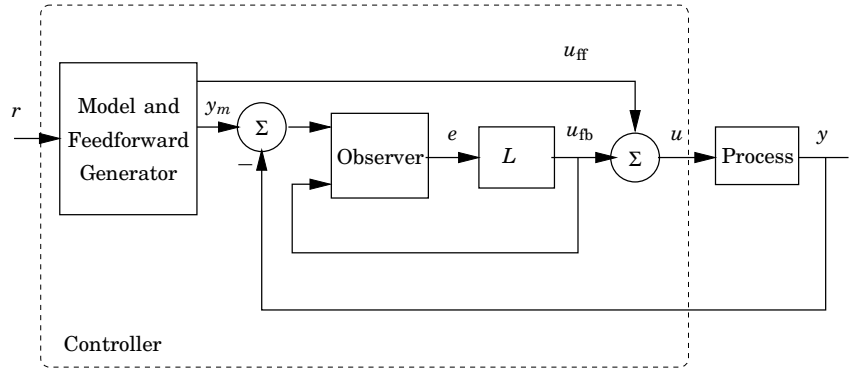


Figure 9.6 Block diagram of a controller based on a structure with two degrees of freedom. The controller consists of a command signal generator, state feedback and an observer which estimates the error $e = x_m - \hat{x}$, between the ideal states and the estimates states.

the state $\hat{x}$ by error $e = x_m - \hat{x}$ gives the following controller

$$\begin{aligned} u &= u_{\text{ff}} + Le \\ \frac{dx_m}{dt} &= Ax_m + Bu_{\text{ff}} \\ y_m &= Cx_m \\ \frac{de}{dt} &= (A - KC)e + B(u_{\text{ff}} - u) + K(y_m - y) \\ &= (A - KC - BL)e + K(y_m - y) \end{aligned} \tag{9.43}$$

This equation for e is driven by the output error $y - y_m$. The model and feedforward generator only have to provide the feedforward signal u_{ff} and the ideal output y_m. A block diagram of the system is shown in Figure 9.6. Notice that this system has the same structure as the feedforward controller in Figure 8.3C. The controller has integral actions the state feedback or the observer is designed so that the transfer function

$$C(s) = L(sI - A + KC + BL)^{-1}K$$

has integral action. Since integral action is in the error path, the steady state error will always be zero if a steady state exists.

In some situations like in path following for robots or for numerically controlled machine tools it may be natural to generate the desired behavior of the state in other ways instead of generating it in real time as we have described here.

We will illustrate with an example.

EXAMPLE 9.10—THE DOUBLE INTEGRATOR
Consider the double integrator which has the transfer function

$$P(s) = \frac{1}{s^2}$$

Assume that it is desired to have a response characterized by the transfer function

$$M(s) = \frac{\omega_m^2}{s^2 + 2\zeta_m\omega_m + \omega_m^2}$$

The transfer function which generates the feedforward signal u_{ff} is given by

$$G_{ff}(s) = M(s)P^{-1}(s) = \frac{s^2\omega_m^2}{s^2 + 2\zeta_m\omega_m + \omega_m^2}$$

Notice that this transfer function is realizable without using derivatives even if $P^{-1}(s)$ is not realizable. To find a realization of the transfer function we rewrite it as

$$M(s)P^{-1}(s) = \frac{s^2\omega_m^2}{s^2 + 2\zeta_m\omega_m + \omega_m^2} = \omega_m^2\left(1 - \frac{2\zeta_m\omega_m s + \omega_m^2}{s^2 + 2\zeta_m\omega_m + \omega_m^2}\right)$$

This transfer function can be realized as

$$\begin{aligned}
\frac{dz_1}{dt} &= -2\zeta_m\omega_m z_1 - \omega_m z_2 + \omega_m r \\
\frac{dz_2}{dt} &= \omega_m z_1 \\
y_m &= \omega_m z_2 \\
u_{\text{ff}} &= \omega_m^2(r - 2\zeta_m z_1 - z_2)
\end{aligned}$$

Notice that the same dynamical system generates both the desired model output y_m and the feedforward signal u_{ff}. Also notice that the gain increases as the square of ω_m. Fast responses can thus be obtained but large control signals are required.

Combining the feedforward generator with the determination of an output feedback in Example 9.5 we can obtain a complete controller based on command following, state feedback and an observer for the double integrator. The controller is a dynamical system of fifth order which is

described by the following equations by

$$\begin{aligned}
u &= u_{\text{ff}} + u_{\text{fb}} \\
u_{\text{ff}} &= \omega_m^2(r - 2\zeta_m z_1 - z_2) \\
u_{\text{fb}} &= l_1 e_1 + l_2 e_2 + e_3 \\
y_m &= \omega_m z_2 \\
\frac{dz_1}{dt} &= -2\zeta_m \omega_m z_1 + \omega_m z_2 + \omega_m r \\
\frac{dz_2}{dt} &= -\omega_m z_1 \\
\frac{de_1}{dt} &= e_2 + k_1(y_m - y - e_1) \\
\frac{de_2}{dt} &= e_3 + u + k_2(y_m - y - e_1) \\
\frac{de_3}{dt} &= k_3(y - e_1)
\end{aligned}$$

The states z_1 and z_2 represent the feedforward generator, the states $e_1 = x_{m1} - \hat{x}_1$ and $e_2 = x_{m2} - \hat{x}_2$ represent the components of the deviations of the estimated state $\hat{x}$ from its ideal values x_m and the state e_3 is an estimate of the load disturbance.

This controller can be represented as

$$U(s) = G_{ff}(s)R(s) + G_{fb}(s)(M(s)R(s) - Y(s))$$

where G_{ff} is given by Equation (9.36) and

$$\begin{aligned}
G_{ff}(s) &= \frac{s^2\omega_m^2}{s^2 + 2\zeta_m\omega_m + \omega_m^2} \\
M(s) &= \frac{\omega_m^2}{s^2 + 2\zeta_m\omega_m + \omega_m^2} \\
G_{fb}(s) &= \frac{(k_1l_1 + k_2l_2 + k_3)s^2 + (k_2l_1 + l_2)s + k_3l_1}{s(s^2 + (k_1 + l_1)s + k_1l_2 + k_2 + l_1)}
\end{aligned}$$

The loop transfer function is

$$L(s) = \frac{(k_1l_1 + k_2l_2 + k_3)s^2 + (k_2l_1 + l_2)s + k_3l_1}{s^3(s^2 + (k_1 + l_1)s + k_1l_2 + k_2 + l_1)}$$

□

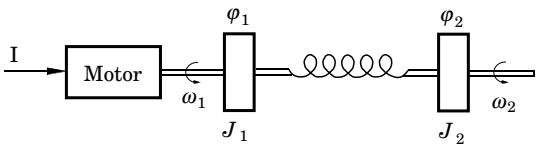


Figure 9.7 Schematic diagram of the system

9.8 An Example

The design procedure will be illustrated by an example.

The Process

Consider the system shown in Figure 9.7. It consists of a motor that drives two wheels connected by a spring. The input signal is the motor current I and the output is the angle of the second wheel φ_2. It is assumed that friction can be neglected. The torque constant of the motor is k_I, the moments of inertia are J_1 and J_2 and the damping in the spring is k_d.

The system is a representative example of control of systems with mechanical resonances. Such systems are common in many different contexts, industrial drive systems, robots with flexible arms, disk drives and optical memories. If the requirements on the control system are modest it is possible to use a PID controller but the achievable performance is limited by the controller. The PID controller will work quite well as long as the rotors move in essentially the same way. When the system is driven in such a way that the angles φ_1 and φ_2 starts to deviate substantially the PID controller is not working well and superior performance can be obtained by using a more complex controller.

The equations of motion of the system are given by momentum balances for the rotors

$$\begin{aligned} J_1\frac{d\omega_1}{dt} &= k_I I + k(\varphi_2 - \varphi_1) + k_d(\omega_2 - \omega_1) \\ J_2\frac{d\omega_1}{dt} &= k(\varphi_1 - \varphi_2) + k_d(\omega_1 - \omega_2) \\ \frac{d\varphi_1}{dt} &= \omega_1 \\ \frac{d\varphi_2}{dt} &= \omega_2 \end{aligned} \tag{9.44}$$

Choose the state variables as

$$\begin{aligned} x_1 &= \varphi_1, \quad x_2 = -\varphi_2 \\ x_3 &= \omega_1/\omega_0, \quad x_4 = \omega_2/\omega_0 \end{aligned}$$

where $\omega_0 = \sqrt{k(J_1+J_2)/(J_1J_2)}$, which is the undamped natural frequency of the system when the control signal is zero. The state equations of the system are

$$\frac{1}{\omega_0}\frac{dx}{dt} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\alpha_1 & \alpha_1 & -\beta_1 & \beta_1 \\ \alpha_2 & -\alpha_2 & \beta_2 & -\beta_2 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ \gamma_1 \\ 0 \end{pmatrix} u + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \gamma_2 \end{pmatrix} d$$

where the parameters are given by

$$\alpha_1 = \frac{J_2}{J_1+J_2}, \quad \alpha_2 = \frac{J_1}{J_1+J_2} \quad \beta_1 = \frac{k_d}{\omega_0 J_1}, \quad \beta_2 = \frac{k_d}{\omega_0 J_2}$$
$$\gamma_1 = \frac{k_I}{\omega_0^2 J_1}, \quad \gamma_2 = \frac{1}{\omega_0 J_2}$$

The parameters have the following numerical values $J_1 = 10/9$, $J_2 = 10$, $k = 1$, $k_d = 0.1$, and $k_I = 5$, which gives $\omega_0 = 1$.

It is convenient to use Matlab to for analysis and design and we will simply describe the procedure through Matlab scripts. First we create the model by the following script.

```
%The robot arm model
%Parameters
J1=10/9;J2=10;k=1;kd=0.1;ki=5;
w0=sqrt(k*(J1+J2)/J1/J2);a1=J2/(J1+J2);a2=J1/(J1+J2);
b1=kd/(w0^2*J1);b2=kd/(w0^2*J2);g1=ki/(w0^2*J1);g2=1/(w0^2*J2);
%System matrices
A=[0 0 1 0;0 0 0 1;-a1 a1 -b1 b1;a2 -a2 b2 -b2]*w0;
B1=[0;0;g1;0];B2=[0;0;0;g2];C1=[1 0 0 0];C2=[0 1 0 0];D1=0;
B=[B1 B2];C=[C1;C2];D=zeros(2,2);
%SISO system where output is angle of first rotor
sys1=ss(A,B1,C1,D1);
%SISO system where output is angle of second rotor
sys2=ss(A,B2,C2,D1);
%Complete system
sys=ss(A,B,C,D);
%Compute transfer function from current to angle
%of second wheel
[num,den]=ss2tf(A,B1,C2,0);P=tf(num,den);
```

As a first assessment of the system we compute its poles, which is done in the following Matlab dialog:

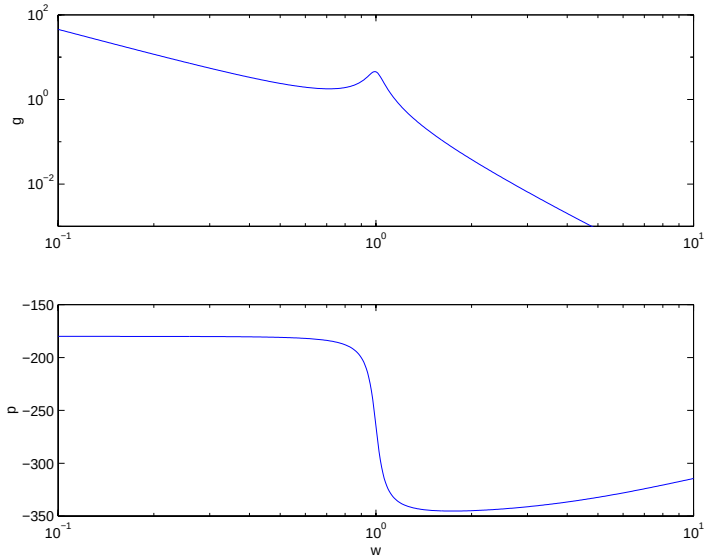


Figure 9.8 Bode plot for the open loop transfer function from current to angle of the second rotor.

```
 robarmdata
 eig(sys.A)
ans =
  -0.0500 + 0.9987i
  -0.0500 - 0.9987i
   0.0000 + 0.0000i
   0.0000 - 0.0000i
```

The system thus has two poles at the origin corresponding to the rigid body motion and two complex poles corresponding to the relative motion of the wheels. The relative damping of the complex poles is $\zeta = 0.05$ which means that there is very little damping. The transfer function from motor current to the angle of the second rotor is

$$P(s) = \frac{0.045s + 0.45}{s^4 + 0.1s^3 + s^2}$$

The Bode diagram of this transfer function is shown in Figure 9.8.

PD Control

For low frequencies process dynamics can be approximated by the transfer

function

$$P(s) \approx \frac{0.45}{s^2}$$

Such a system can conveniently be controlled by a a PD controller. Requiring that the closed loop system should have the characteristic polynomial

$$s^2 + 2\zeta_1\omega_1 s + \omega_1^2$$

we find that the PD controller becomes

$$C(s) = \omega_1^2 + 2\zeta_1\omega_1 s = k + k_d s = k(1 + sT_d)$$

We add high frequency roll-off by modifying the controller to

$$C(s) = k\frac{1 + sT_d}{(1 + sT_d/N)^2} \tag{9.45}$$

where N is a parameter that is chosen empirically, a good choice is typically in the range of 4 to 20. The approximate model is very poor for frequencies close to 1 rad/s because of the phase lag caused by the resonant poles. Notice in Figure 9.8 that the phase drops very rapidly for frequencies around 1 rad/s. The frequency ω_1 should thus be chosen well below 1 rad/s. Using the standard rule of thumb we choose $\omega_1 = 0.2$. It is essential to have high-frequency roll-off in the controller to make sure that the loop gain drops rapidly after the crossover frequency.

To evaluate the controller we compute the maximum sensitivity, $M_s = 2.2$, which occurs at $\omega = 0.93$, i.e. close to the resonance. To reduce the sensitivity we reduce the parameter ω_1 to 0.15 and the maximum sensitivity reduces to $M_s = 1.6$ at $\omega = 0.13$. The frequency responses of the gang of four, see Figure 9.9. The figure shows that the sensitivity functions are reasonable, the maximum sensitivities are $M_s = 1.6$ and $M_t = 1.8$. The disturbance rejection is quite poor, the largest value of the transfer function G_{xd} is 60. There are no problems with measurement noise because the largest value of the transfer function G_{un} is only 0.16. The resonance peak is noticeable in several of the frequency responses, because of the high frequency roll-off of the controller the peak is kept at a low level for the chosen parameters. It is essential that the controller has high frequency roll-off, without this the controller with the parameters in Figure 9.9 actually gives an unstable closed loop system. The choice of the parameter ω_1 is sensitive, increasing the value to $\omega_1 = 0.2867$ makes the closed loop system unstable. The step responses of the gang of four are shown in Figure 9.10. The time responses look quite reasonable, they represent what can be achieved with PD control. The overshoot indicated

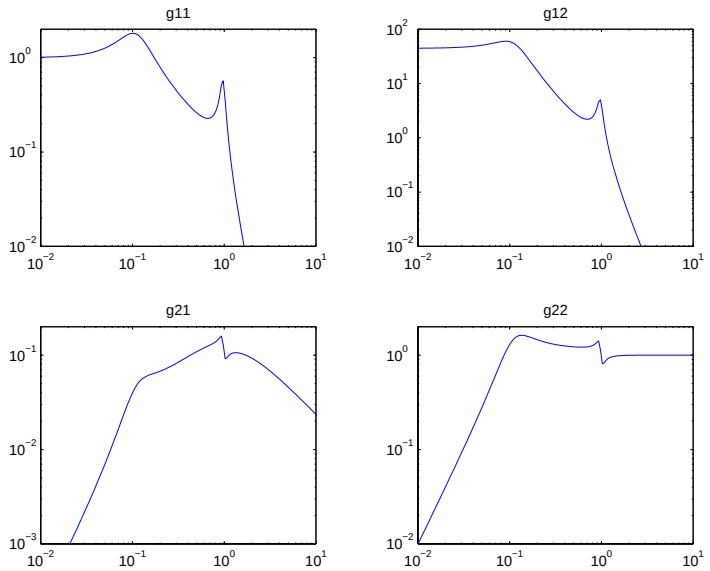


Figure 9.9 Gain curve of the frequency responses of the gang of four for PD control of the system. The controller is given by (9.45) with parameters $\omega_1 = 0.15$, $\zeta_1 = 0.707$ and $N = 10$.

by the response of $P(s)C(s)/(1 + P(s)C(s))$ can be reduced by set point weighting. The load disturbance response can be improved at low frequencies by introducing integral action. The peak in the response to load disturbances can be reduced a little by further tuning of the parameters. Notice that the high frequency resonant mode is traceable in the graphs although the amplitude is small. The mode becomes very visible if the parameter ω_1 is increased towards 0.3.

Design of State Feedback

We will first design a state feedback that gives a closed loop system with poles in -2 -1 and $-1 \pm i$. This choice implies that the oscillatory poles are damped and that the integrating poles are move well into the left half plane. The design is executed with the following Matlab script.

```
%Design of State feedback for robotarm
%Get process model
robarmdata;B=B1;C=C2;
%Desired closed loop poles
P=[-1 -2 -1+i -1-i];
```

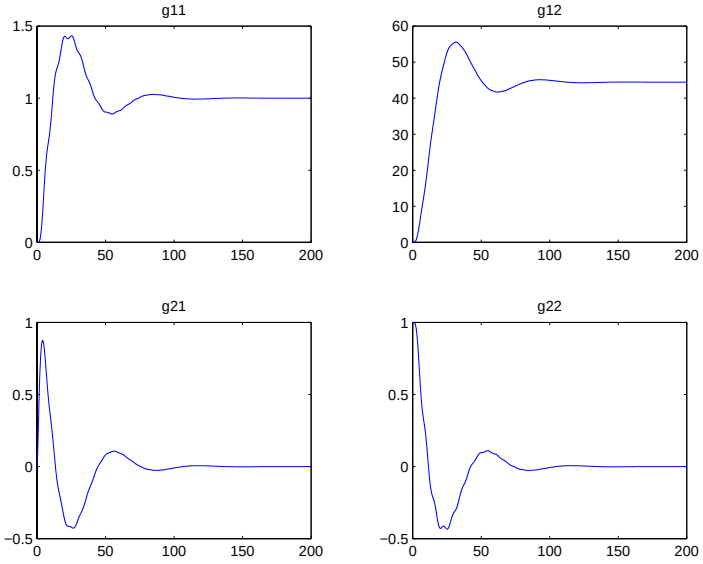


Figure 9.10 Time responses of the gang of four for PD control of the system. The controller is given by (9.45) with parameters $\omega_1 = 0.15$, $\zeta_1 = 0.707$ and $N = 10$.

```
L=acker(A,B,P);Acl=A-B*L;
Lr=-1/(C*inv(Acl)*B);
disp('Feedback gain L=');disp(L)
disp('Refererence gain Lr=');disp(Lr)
%Check the results
n=size(A);disp('Specified closed loop poles')
disp(P);
disp('Closed loop poles obtained')
disp(eig(Acl)');
%Compute Closed loop transfer function
[num,den]=ss2tf(Acl,B*Lr,C,0);Gcl=tf(num,den);
disp('Closed loop transfer function:')
Gcl
```

We find that the control law is

$$u = -l_1 x_1 - l_2 x_2 - l_3 x_3 - l_4 x_4 + L_r x_r$$

where $l_1 = 1.78$ $l_2 = 7.10$ $l_3 = 1.09$ and $l_4 = 20.24$ and $L_r = 8.89$. The

closed loop transfer function is

$$G(s) = \frac{0.4s + 4}{s^4 + 5s^3 + 10s^2 + 20s + 4}$$

Notice that when using Matlab a few small terms appear because of inaccuracies in the computations.

To find the stiffness of the system we calculate the transfer function from a a disturbance torque on the second rotor to the deflection of the rotor. This can be done by the Matlab script

```
%Computation of impedance
%This script computes the transfer function
%from a disturbance on the second rotor to
%the angle of that rotor
%Compute state feedback
robarmsfb;
%Compute the transfer function
A=sys2.A;B=sys2.B;C=sys2.C;
[num,den]=ss2tf(Acl,B,C,0);Gcl=tf(num,den);
disp('Closed loop impedance function:')
Gcl
%Compute stiffness at second rotor
ks=den(5)/num(4);
disp('Stiffness=')
disp(ks)
```

Executing the script we find that the transfer function from a torque on the second mass to deflection is

$$G(s) = \frac{0.1s^2 + 0.499s + 0.894}{s^4 + 5s^3 + 10s^2 + 10s + 4}$$

The steady state gain of this transfer function is $G(0) = 0.894/4 = 0.22$. The stiffness is thus $1/G(0) = 4.5$ which can be compared with the stiffness of the uncontrolled system which is 1. The stiffness can be increased by choosing faster closed loop poles.

Simulation of the Closed Loop System

To illustrate the properties of the closed loop system further we will simulate the system. A Simulink block diagram of the system is shown in Figure 9.11. In the simulation we will first apply a unit step in the reference signal and we will then apply a disturbance torque of 2 units at the second rotor. The results of the simulation are shown in Figure 9.12. The simulation is executed using the following Matlab script.

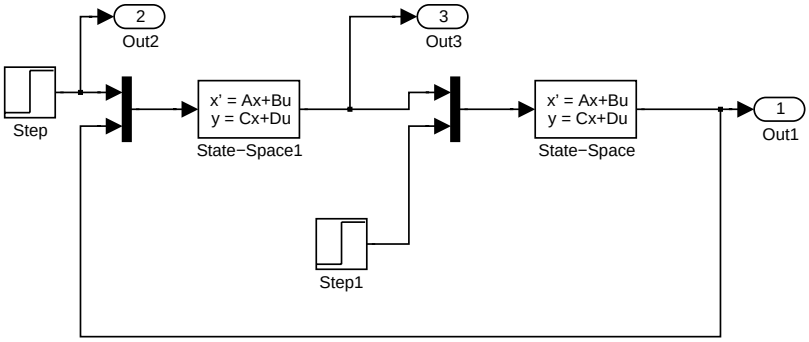


Figure 9.11 Simulink block diagram of the system with a controller.

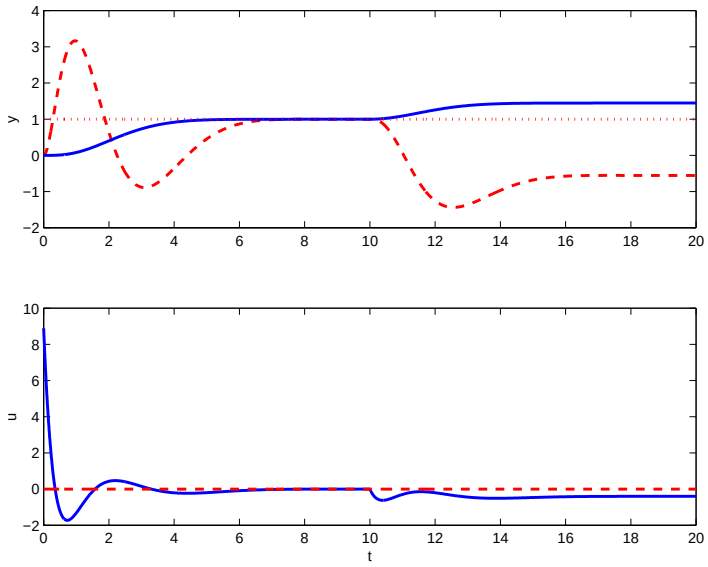


Figure 9.12 Simulation of the closed loop system with state feedback. The angle φ_1 of the first rotor is shown in dashed curves, and the motion of the second rotor in solid lines.

```
%Simlation of robot arm with state feedback
robarmsfb;       %Execute controller design
%Set procees parameters
Ap=A;Bp=[B B2];Cp=eye(n);Dp=zeros(4,2);
```

```
%Set controller parameters
Ac=0;Bc=zeros(1,5);Cc=0;Dc=[Lr -L];
%Execute simulation
r=1;d=2;ts=20;[t,x,y]=sim('robarmblk',[0 ts]);
subplot(2,1,1);
plot(t,y(:,1),'r--',t,y(:,2),'b',t,ones(size(t)),'r:','linew',2);
ylabel('y')
subplot(2,1,2);plot(t,y(:,6),t,zeros(size(t)),'r--','linew',2);
ylabel('u')
```

The simulation shows that the second rotor responds very nicely to the reference signal. To achieve the fast response the first rotor has a large overshoot. This is necessary because the only way to exert a force on the second rotor is to turn the first rotor so that a large force is obtained. To execute effective control it is necessary to coordinate the motion of both rotors accurate. This is done very efficiently by the controller based on state feedback. To obtain the response to a unit step command shown in the figure the initial value of the control signal is 8.9. The value is proportional to the magnitude of the step. To judge if the value is reasonable it it necessary to know the numerical values of typical steps and the signal level when the control signal saturates. If the a larger value of the control signal can be permitted the specified closed loop poles can be chose to be faster. If the value is too large the desired closed loop poles should have smaller magnitude.

The response to a torque disturbance at the second rotor shows that there is a steady state error. The reason for this is that the controller does not have integral action. Also notice that the angle φ_1 of the first rotor is negative. This is necessary because the only way to exert a torque on the second rotor is by turning the first rotor.

Controller with Integral Action

To avoid the steady state error for a load disturbance we will introduce integral action as described in Section 9.2. This gives the control law

$$u = -l_1 x_1 - l_2 x_2 - l_3 x_3 - l_4 x_4 - l_5 \int_0^t (x_2 - r)dt + L_r r$$

With a controller having integral action the closed loop system is of fifth order and it is necessary to specify five poles. They are chosen as -3, -2, -1, and $-1 \pm i$. The design is executed with the following Matlab script.

```
%Design of State feedback for robotarm
%Get process model
robarmdata;A=sys2.A;B=sys1.B;C=sys2.C;
```

```
%Form augmented system
A=[A;C];A=[A,zeros(5,1)];B=[B;0];C=[C 0];
%Desired closed loop poles
P=[-0.5 -1 -2 -1+i -1-i];
L=acker(A,B,P);Acl=A-B*L;
Lr=-1/(C*inv(Acl)*B);
disp('Feedback gain L=');disp(L)
disp('Refererence gain Lr=');disp(Lr)
%Check the results
n=size(A);disp('Specified closed loop poles')
disp(P);
disp('Closed loop poles obtained')
disp(eig(Acl)');
%Compute Closed loop transfer function
%using the same value of Lr as for regular
%state feedback Lr=8.89
Ba=B*8.89+[zeros(4,1);-1]
[num,den]=ss2tf(Acl,Ba,C,0);Gcl=tf(num,den);
disp('Closed loop transfer function:')
Gcl
```

Executing the script we find that the following controller parameters

$$l_1 = 3.61, \quad l_2 = 47.95, \quad l_3 = 1.53, \quad l_4 = 59.98, \quad l_5 = 17.78$$

The transfer function from reference to angle of the second rotor for the closed loop system is

$$G(s) = \frac{0.4001s^2 + 4.801s + 8}{s^5 + 7s^4 + 20s^3 + 30s^2 + 24s + 8}$$

The transfer function from torque on the second rotor to its displacement can be computed using the following Matlab script.

```
%Computation of impedance
%This script computes the transfer function
%from a disturbance on the second rotor to
%the angle of that rotor for a state feedback
%with integral action
%Compute state feedback
robarmsfbi;
%Compute the transfer function
B=[sys2.B;0];C=[sys2.C 0];
[num,den]=ss2tf(Acl,B,C,0);Gcl=tf(num,den);
```

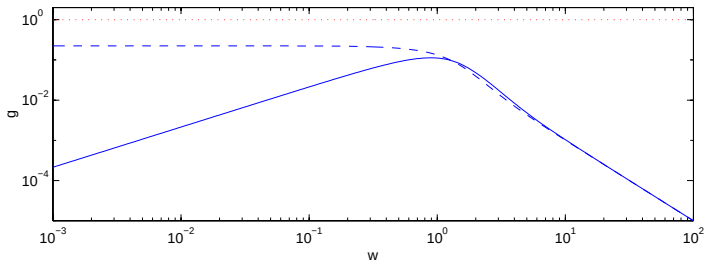


Figure 9.13 Gain curves for the open loop system (dotted), the system with state feedback (dashed) and the system with state feedback and integral action (full).

```
disp('Closed loop impedance function:')
Gcl
%Compute stiffness at second rotor
ks=den(6)/num(6);
```

The transfer function is

$$G(s) = \frac{0.1s^3 + 0.699s^2 + 1.713s}{s^5 + 7s^4 + 20s^3 + 30s^2 + 24s + 8}$$

In Figure 9.13 we show the gain curves of the transfer functions relating angular deviation to disturbance torque for the open loop system and for the system with state feedback and state feedback with integral action. The figure shows clearly that the deflection caused by a given torque can be reduced by feedback. With a regular state feedback the gain is reduced by more than a factor of XXX. By introducing integral action in the controller the deviation can be reduce more at low frequencies.

Simulation of the Controller with Integral Action

The simulink block shown in Figure 9.11 can also be used to simulate the controller with integral action. The simulation is executed with the script

```
%Simlation of robot arm with state feedback
%having integral action
%Get procees and controller parameters
robarmsfbi;
Ap=A;Bp=[B B2];Cp=eye(4);Dp=zeros(4,2);
%Set controller parameters
Lr=8.8;
Ac=0;Bc=[1 0 -1 0 0];Cc=L(5);Dc=[Lr -L(1:4)];
%Execute simulation
```

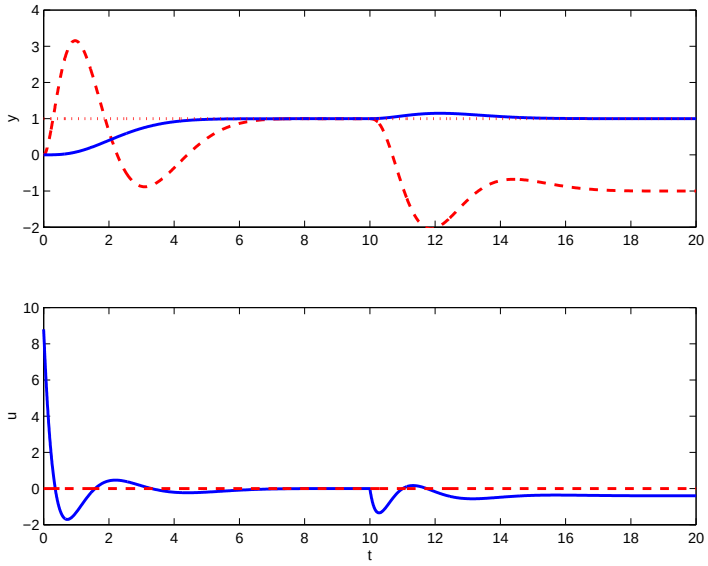


Figure 9.14 Simulation of the closed loop system with state feedback having integral action. The angle φ_1 of the first rotor is shown in dashed curves, and the motion of the second rotor in solid lines.

```
r=1;d=2;ts=20;[t,x,y]=sim('robarmblk',[0 ts]);
subplot(2,1,1);
plot(t,y(:,1),'r--',t,y(:,2),'b',t,ones(size(t)),'r:','linew',2);
ylabel('y')
axis([0 20 -2 4])
subplot(2,1,2);plot(t,y(:,6),t,zeros(size(t)),'r--','linew',2);
ylabel('u')
axis([0 20 -2 10])
```

The result of the simulation is shown in Figure 9.14. Compare this with the simulation of the controller without integral action in Figure 9.14. The responses to reference signals are quite comparable but the responses to a disturbance torque are different. The controller with integral action gives zero steady state error when a torque is applied. The first rotor does however have a substantial deviation. This is necessary to create a torque to oppose the disturbance.

9.9 Summary

In this chapter we have presented a systematic method for design of a controller. The controller has an interesting structure, it can be thought of as composed of three subsystems: a system that generates the desired output and a feedforward signals from the reference value, an estimator and a feedback from estimated states. This structure has the property that the response to reference signals can be decoupled from the response to disturbances. The details are carried out only for systems with one input and one output but it turns out that the structure of the controller is the same for systems with many inputs and many outputs. The equations for the controller have the same form, the only difference is that the feedback gain L and the observer gain K are matrices instead of vectors for the single-input single-output case. There are also many other design methods that give controllers with the same structure but the gains K and L are computed differently. The analysis also gives an interesting interpretation of integral action as a disturbance estimator. This admits generalizations to many other types of disturbances.

Index