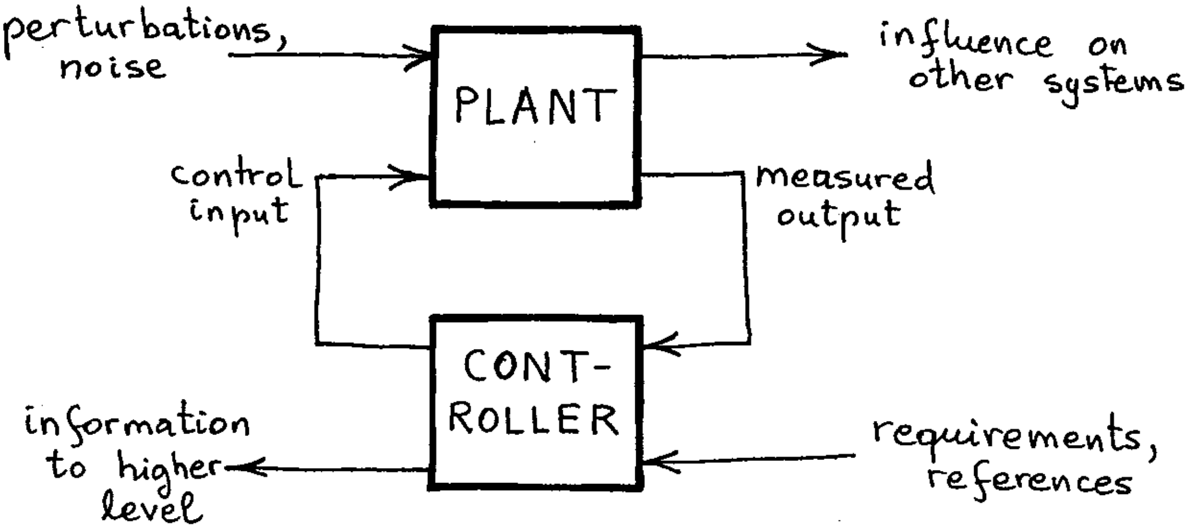
Introduction to control theory

WEEK 2

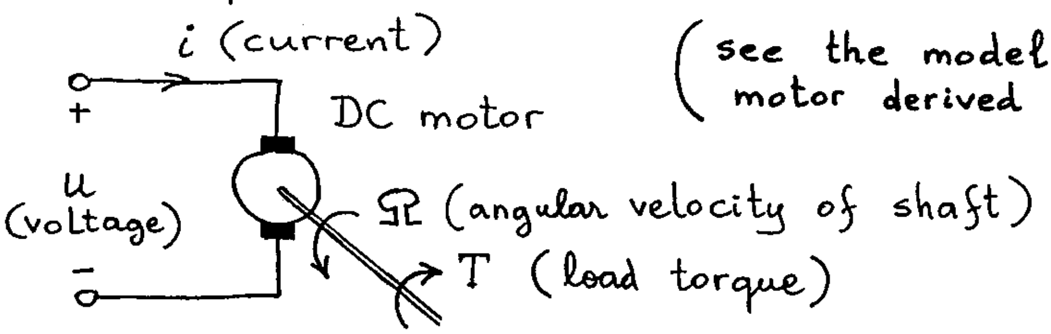
What is a control system

General structure (feedback connection):



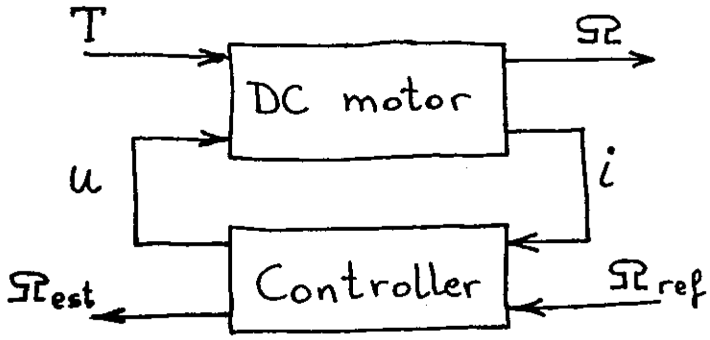
Each signal shown may have many components

Example 1.



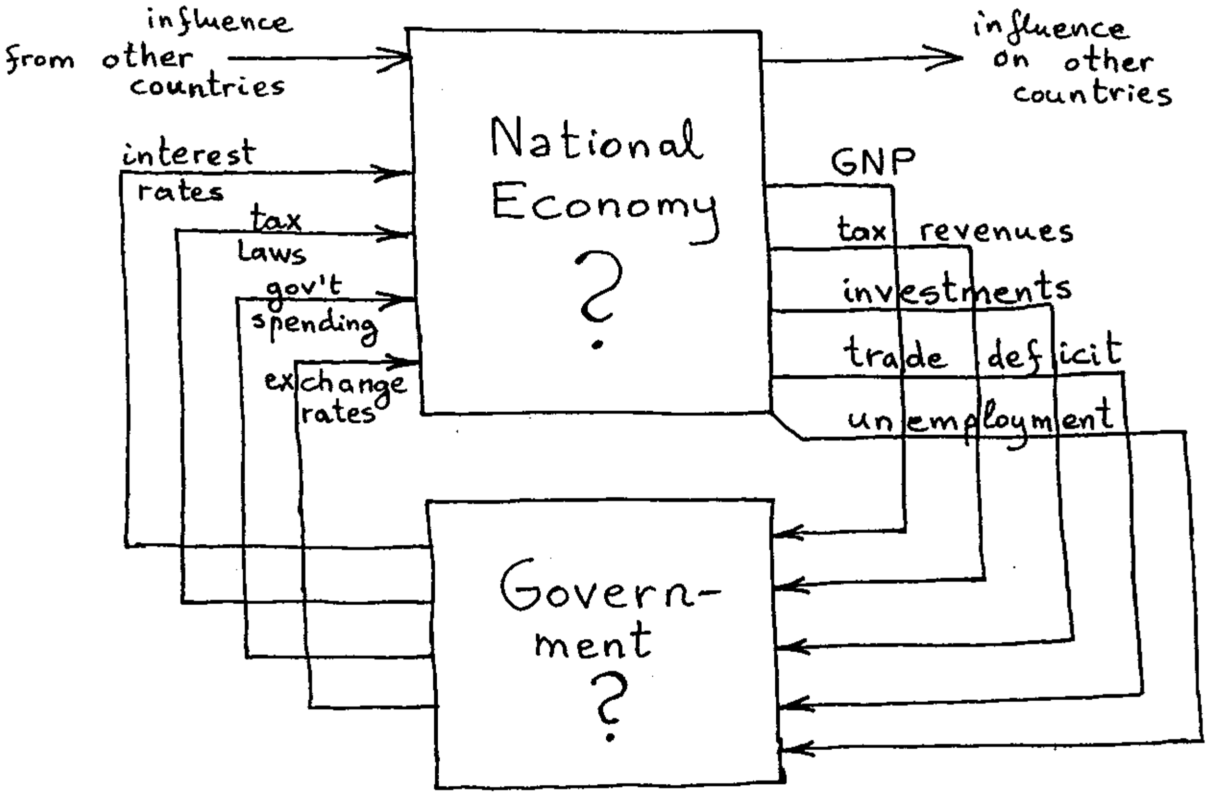
(see the model of the DC motor derived in WEEK 1)

Control objective: by the proper choice of u, Ω should be as close as possible to Ω_{ref}, irrespective of T.



Note that Ω is not measured directly, since it is easier to measure i and estimate Ω by $\Omega_{est} = (u - Ri)/k$, R and k are known.

Example 2.



No good mathematical model and no good control algorithm are known.

Example 3.

Control objectives:

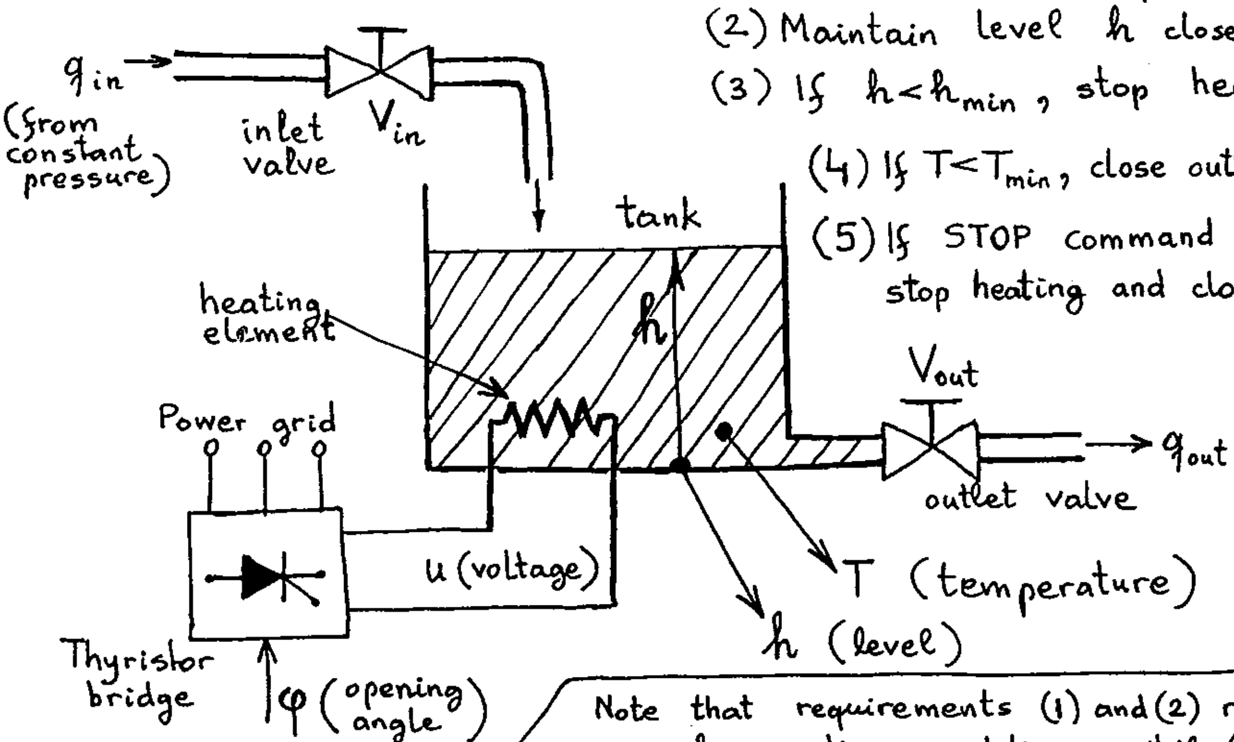
(1) Maintain temperature in tank, T close to imposed T_{ref}.

(2) Maintain level h close to h_{ref}.

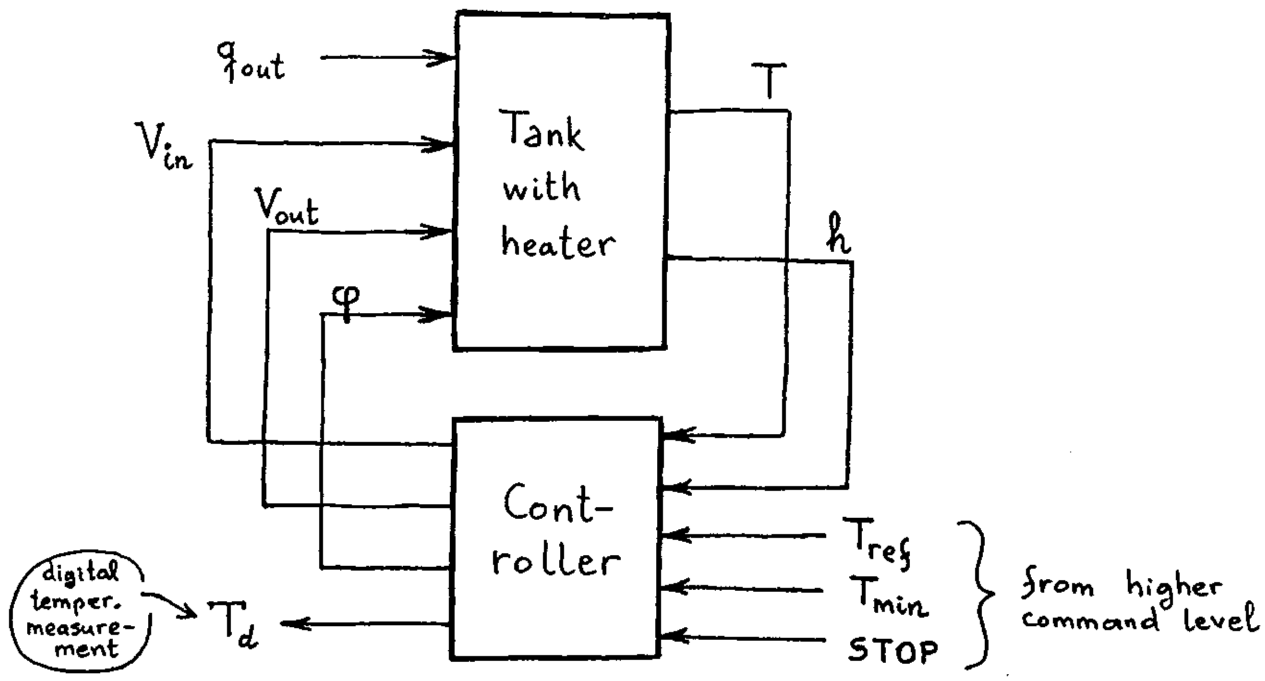
(3) If $h<h_{min}$, stop heating

(4) If $T<T_{min}$, close outlet valve.

(5) If STOP command received, stop heating and close outlet valve.



Note that requirements (1) and (2) refer to normal operating conditions, while (3), (4) and (5) refer to emergency or limit conditions.



In general, in a control system, <u>signals</u> can represent

- voltage
- current
- position (angle)
- velocity (angular)
- temperature
- pressure
- flow rate
- light intensity

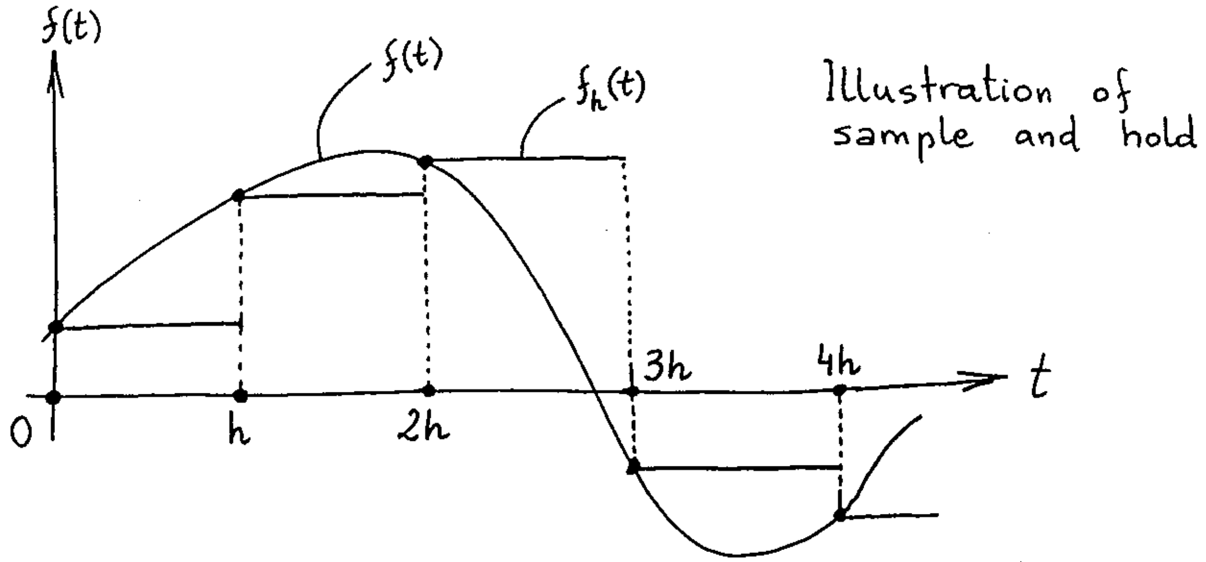
etc, etc. Another classification of signals:

- analog (in ex. 3: q_{out}, φ, T, h)
- digital (in ex. 3: T_{ref}, T_{min}, T_d)
- binary (in ex: 3: V_{in}, V_{out}, STOP)

Signals can be:

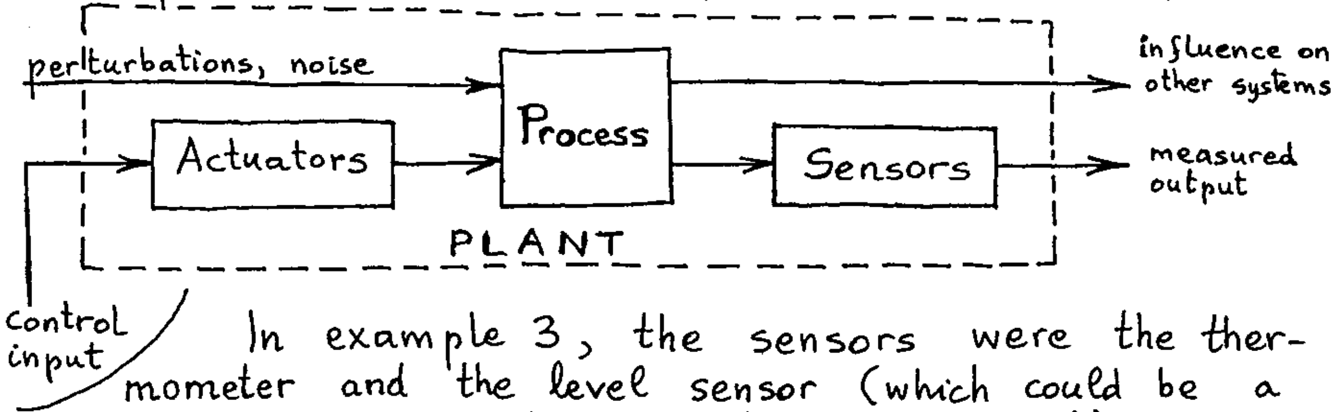
- continuous-time (functions defined on $[0,\infty)$)
- discrete-time (functions defined on $\{0,1,2,3,\ldots\}$)

A continuous-time signal $f(t)$ can be transformed into a discrete-time signal $g(n)$ by <u>sampling</u>: $g(n) = f(nh)$.
A discrete-time signal $g(n)$ can be transformed into a continuous-time signal in many ways, the simplest being by <u>hold</u>: $f_h(t) = g(n)$ for $t \in [nh, (n+1)h)$.



If the sampling step h tends to zero, then $f_h(t)$ tends to $f(t)$ (if f is continuous). If the controller is digital (e.g. a microprocessor running a program written following a control algorithm) then the measured output of the plant has to be sampled and converted from analog to digital. At the output of the controller, the digital signal must be converted back to analog and (by hold or by other methods) must be transformed into a continuous-time signal. In this module, we do not deal with the problems of analog/digital conversion and with sample/hold: <u>all signals we consider are continuous-time and analog.</u>

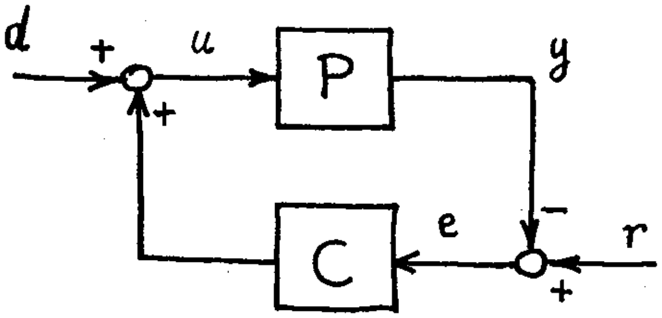
Often (but not always) the plant can be decomposed as follows (compare withe figure on p.1):



In example 3, the sensors were the thermometer and the level sensor (which could be a pressure sensor at the bottom of the tank). The actuators were the two valves and the thyristor bridge.

The standard feedback connection of two LTI systems

We consider only the single input-single output (SISO) case.



The feedback connection of two linear SISO systems with transfer functions P and C

It is clear from the block diagram that

$$\begin{bmatrix} \hat{d} \\ \hat{r} \end{bmatrix} = \begin{bmatrix} 1 & -C \\ P & 1 \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{e} \end{bmatrix}$$

(assuming zero initial conditions)

whence

$$\begin{bmatrix} \hat{u} \\ \hat{e} \end{bmatrix} = \begin{bmatrix} 1 & -C \\ P & 1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{d} \\ \hat{r} \end{bmatrix},$$

Note that we have here a 2×2 matrix of transfer functions.

$$\begin{bmatrix} 1 & -C \\ P & 1 \end{bmatrix}^{-1} = \begin{bmatrix} (1+CP)^{-1} & C(1+PC)^{-1} \\ -P(1+CP)^{-1} & (1+PC)^{-1} \end{bmatrix}.$$

The feedback connection of P and C is called <u>stable</u> if the transfer function matrix from d, r to u, e is stable, i.e., if all entries of the 2×2 matrix above are stable. Note that two entries are equal, so that we have three different entries to check. Stability of the feedback connection means that if the inputs d, r have finite energy, then all signals in the diagram have finite energy (see WEEK 1).

In terms of state space systems, the interpretation of the stability concept introduced on p. 5 is the following: If P and C are the transfer functions of a minimal plant and a minimal controller (which is usually the case), then also the feedback system is minimal. Hence, if its transfer function $\begin{bmatrix} 1 & -C \\ P & 1 \end{bmatrix}^{-1}$ is stable, then also the system is stable (see p. 2 of WEEK 1).

Intuitive interpretation of the diagram:

P = the transfer function of the plant
C = the transfer function of the controller
d = disturbance acting on the plant
u = the input signal of the plant
y = the output signal of the plant
r = the reference signal
e = the tracking error, $e = r - y$

Standard terminology:

$P \cdot C$ = the loop gain
$(1 + P \cdot C)^{-1}$ = the sensitivity, also denoted S

The goal:

We would like to have good tracking in spite of the disturbances, i.e., to make e small. Since $\hat{e} = -PS\hat{d} + S\hat{r}$, this can be achieved by making S as small as possible, without destroying the stability of the closed-loop system.

Proposition. The feedback connection of P and C is stable iff the following two conditions hold:

(1) $(1+PC)^{-1}$ is stable,

(2) there is no unstable pole-zero cancellation in the product PC. (Anderson & Gevers, 1981)

We say that there is <u>pole-zero cancellation</u> at z in the product PC if z is a pole of P (or of C) and a zero of C (or of P). Example: if

$$P(s) = \frac{s}{(s+1)(s-2)}, \quad C(s) = \frac{5(s+1)}{s^2+1},$$

stable

then there is a pole-zero cancellation at $z=-1$ in PC. A <u>pole-zero cancellation is stable</u> if $z \in \mathbb{C}_-$ and unstable otherwise. <u>Example</u>: if

$$P(s) = \frac{2s}{s^2+1}, \qquad C(s) = \frac{s^2+1}{(s+2)(s+100)},$$

then there are unstable pole-zero cancellations at $z=i$ and at $z=-i$. Hence, the feedback connection of this P and C cannot be stable.

<u>Example</u>: take

$$P(s) = \frac{3s}{s-1}, \qquad C(s) = \frac{1}{s-1}.$$

Obviously, there is no unstable pole-zero cancellation in PC (in fact, there is no pole-zero cancellation at all). We compute

$$\left(1+P(s)C(s)\right)^{-1} = \frac{s^2-2s+1}{s^2+s+1}$$

which is stable, so that the feedback connection is stable. Note that P and C are not stable.

The feedback connection of P and C is called <u>w-stable</u> if $|P(\infty)\cdot C(\infty)| < 1$. If this condition is not satisfied then a very small change in P and C can cause all signals in the interconnection to tend to ∞ very fast. <u>Example</u>:

$$P(s) = \frac{5s}{s+1}, \qquad C(s) = -\frac{s+2}{s-1}.$$

We have $P(\infty) = 5$, $C(\infty) = -1$, so that $|P(\infty)C(\infty)| = 5$ and the feedback connection of P and C is not w-stable. It is easy to check that this feedback connection is stable, using the proposition on top of p. 3. Indeed, there are no pole-zero cancellations in PC and

$$(1+PC)^{-1} = -\frac{s^2-1}{4s^2+10s+1},$$

which is stable.

> When we design a feedback connection, we always require that it should be stable and w-stable.

Indeed, in any engineering application, we can not allow any signal to grow without bound. A transfer function G is called <u>strictly proper</u> if $G(\infty) = 0$. Equivalently, the degree of the numerator < the degree of the denominator. If P or C are strictly proper, then obviously the feedback connection is w-stable (as in the 3 examples on p. 7).

The transfer function from r to e (or from d to u) in the feedback connection on p.1 is

$$S = (1 + PC)^{-1} .$$

This transfer function is called the <u>sensitivity</u> of the feedback system. The stability of S is one condition for the stability of the feedback system, see the proposition on top of p. 7.

If

$$PC = \frac{n}{d}$$

where n and d are polynomials, then

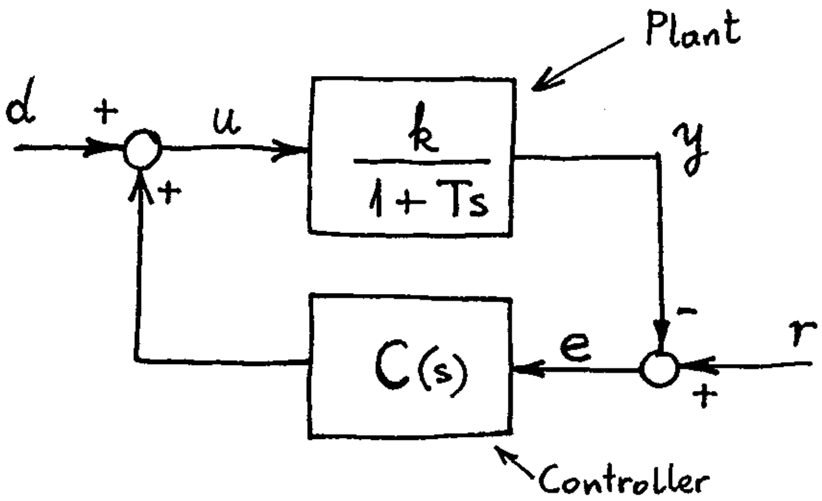
$$S = \left(1 + \frac{n}{d}\right)^{-1} = \frac{d}{d+n} .$$

Thus, we see that <u>the poles of PC become zeros of S</u>. This can be seen in all the examples we have given. For reasons that will be explained later, it is usually required (in addition to stability and w-stability) that $|S(i\omega)|$ should be small for small ω, and $S(i\omega)$ should be close to 1 for large values of ω.

Another frequently used terminology in control is to call PC the <u>loop gain</u> of the feedback system. An equivalent reformulation of the requirements on $S(i\omega)$ stated above is that $|P(i\omega)C(i\omega)|$ should be large for small ω, and it should be close to zero for large values of ω. This is achieved by the proper choice of the controller C.

Extended example: (intended for a study group)

CONTROL OF A FIRST ORDER SYSTEM



We want that this control system should be stable and the output y of the plant should be close to the reference signal r (i.e., e should be small). The signal d is a perturbation. Since the plant is strictly proper, w-stability is not a problem. Usually, T and k are positive, but this need not be the case. The most important class of reference signals are steps (i.e., we would like a constant y).

● Proportional control: This is one possible type of control, not the best but very simple: $C(s) = K$ (constant gain).

We examine the performance of the closed-loop system. We compute its sensitivity :

$$S(s) = \left(1 + \frac{kK}{1+Ts}\right)^{-1} = \frac{1+Ts}{1+kK+Ts}.$$

This is the transfer function from r to e. For <u>stability</u> (since there are no pole-zero cancellations), we need that S should be stable. Its pole is at $p_0 = -\frac{1+kK}{T}$ and the stability condition is $p_0 < 0$. In particular, if $T > 0$ (stable plant) then we need $1 + kK > 0$ (otherwise, we need $1+kK<0$, of course). The transfer function from r to y is

$$G(s) = \frac{kK}{1+Ts} \cdot S(s) = \frac{kK}{1+kK} \cdot \frac{1}{1+\tau s}$$

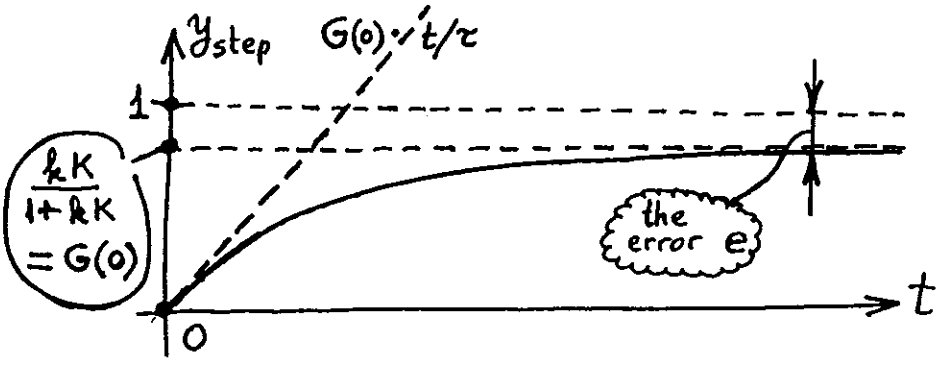
(note: $G(s) = 1 - S(s)$)

where

$$\tau = \frac{T}{1+kK}$$

is the <u>time-constant</u> of G. From this we can easily compute the step response of G (i.e., the function $y(t)$ when $r(t) = 1$):

$$y_{step}(t) = \frac{kK}{1+kK}\left(1 - e^{-\frac{t}{\tau}}\right).$$



As expected, for large t, y_{step} approaches $G(0)$.

The error $e_{step}(t) = 1 - y_{step}(t)$ converges to $1 - G(0) = S(0)$, i.e.,

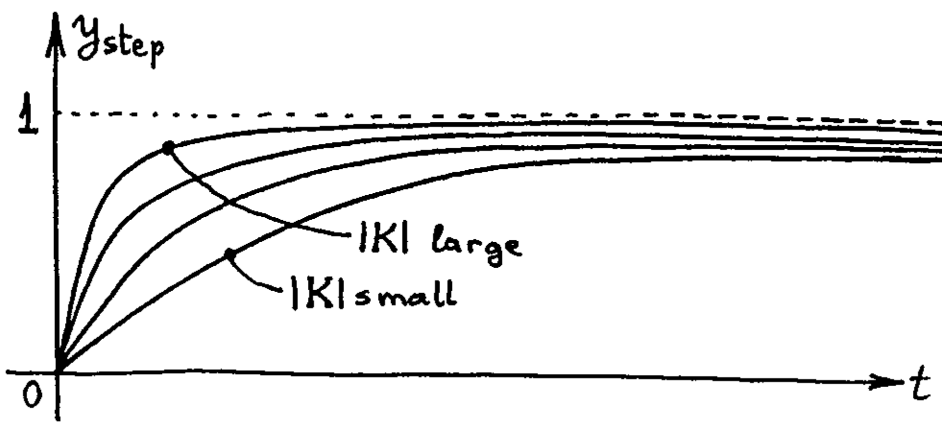
$$e_{step}(\infty) = \frac{1}{1+kK}.$$

The derivative of $y_{step}(t)$ at $t=0$ (its slope) is

$$\frac{d}{dt}\, y_{step}(0) = \frac{G(0)}{\tau} = \frac{1+kK}{T} \cdot G(0) = \frac{kK}{T}.$$

(In the graph on p. 2, we have shown the tangent to y_{step} at $t=0$, which is $y(t) = (t/\tau)G(0)$.) The smaller the time constant τ, the faster the step response $y_{step}(t)$ will approach its final value (which is $G(0)$).

Our aim is to make $e_{step}(\infty)$ small (small stationary error) and τ small (fast response of the control system). From the formulas, it is clear that both aims can be achieved by choosing $|K|$ large. The correct sign of K depends on the signs of T and k, for example, if $T>0$ and $k>0$ then also $K>0$. In the graphs below, we show various graphs of possible y_{step}, as $|K|$ increases.



However, there are also disadvantages with a large controller gain K:

(1) The control input $u(t)$ becomes large for small values of t. Indeed,

$$u(0) = K e(0) = K S(\infty) = K .$$

Often, it is not permissible to apply large u to the plant (something could get damaged).

(2) If the transfer function of the plant $k/(1+Ts)$ is just an approximation, and in reality there are more poles present (which is almost always the case), then a high gain K can lead to instability. (This phenomenon can be better understood after studying the Nyquist test for stability of feedback systems.)

Thus, a sensible compromise should be found between the need to increase K and the need to avoid the two problems above.

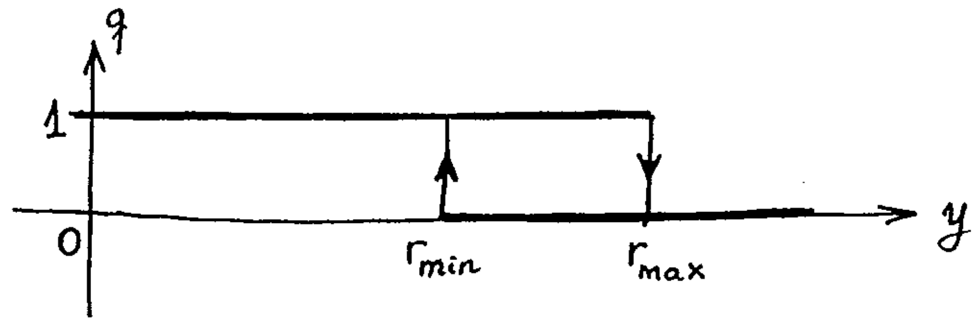
- <u>Hysteresis control</u>: This is most common in boilers, refrigerators, air conditioners, liquid level controllers and many other processes which are slow (large time constant) and where precision is not critical.

Instead of one reference signal there are two, r_{min} and r_{max}, and the controller maintains y between these two limits:

$$r_{min} \leqslant y(t) \leqslant r_{max} .$$

The controller output q is binary ("on" or "off", denoted 1 and 0) and the input u applied to the plant is either a constant u_0 (if $q=1$) or zero (if $q=0$). For example, u_0 might be the grid voltage, which is either applied to the heating element of a boiler (if $q=1$) or not (if $q=0$).

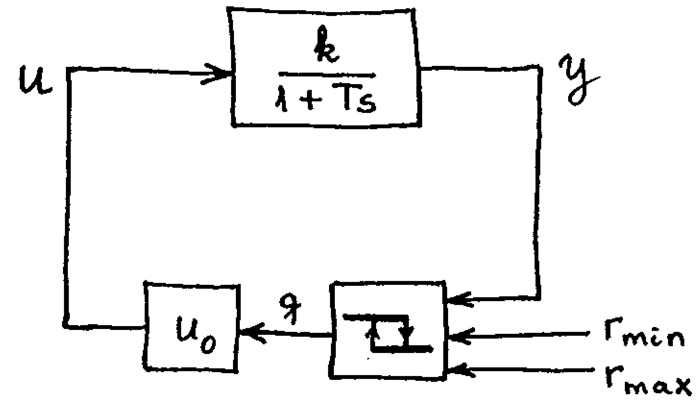
The controller switches into the state $q=1$ if $y < r_{min}$, and it switches into $q=0$ if $y > r_{max}$. If $r_{min} \leq y(t) \leq r_{max}$, then the value of q is maintained unchanged (could be 0 or 1). The graphic illustration of the dependence of q on y, r_{min} and r_{max} is as follows:



Since the path of the point (y,q) is different for increasing y and for decreasing y, this is called a hysteresis curve (as in physics, the dependence of the magnetic induction B on the magnetic intensity H, in a ferromagnetic material).

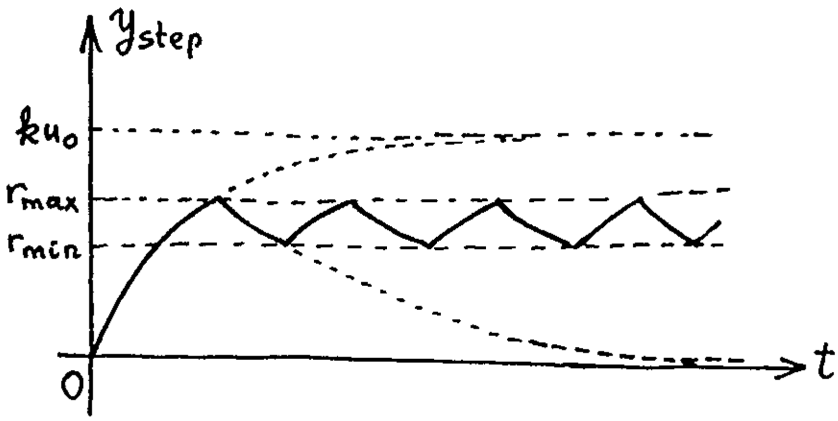
The dependence $u = u_0 \cdot q$ can be realized by a relay or by an electronic switch, such as a thyristor, a triac, etc.

The control system looks like this:



We assume $k > 0$, $T > 0$.

The step response of this system (i.e., r_{min} and r_{max} constant, initial value of y zero) is:



Note that y_{step} oscillates between r_{min} and r_{max}.

Obviously, the precision with which we can keep y_{step} close to a desired value increases as we make the interval $[r_{min}, r_{max}]$ small, i.e., $r_{max} - r_{min}$ small. The drawback of doing this is that the frequency of the switchings increases (it tends to ∞ as $r_{max} - r_{min}$ tends to zero) and too frequent switchings can destroy the switching element (relay, etc). Again, a sensible compromise must be found.