MSL Technical Guide 20

Introduction to Temperature Control

Introduction

One of the most common reasons for measuring temperature is to control it. This applies not only to industrial processes but also to many laboratory tests and experiments. To the uninitiated, control systems seem to exhibit confusing and sometimes peculiar behaviour. The purpose of this guide is to help users understand the behaviour and to obtain the best performance from temperature controllers.

The technical guide is in two parts: the first is a mainly descriptive tutorial section with an overview of the various types of temperature controllers and simple methods for tuning them, and the second part is more mathematical and focuses on understanding proportional and PID controllers. It also provides some advice on design of control systems and instrumentation to enable well-behaved controlled systems.

Throughout the guide we describe the controlled system as 'the plant'. The plant is simply a generic label, and could represent a calibration bath, oven, furnace, refrigerator, or coolstore. The general principles given here can also be applied to control of other parameters such as level, flow, force, etc.

Part 1: Types of Controllers

All temperature controllers employ negative feedback. That is, if the plant temperature is observed to rise (the positive movement), the controller acts to oppose that movement (the negative response), by reducing the heat to the plant, and *vice versa*. A block diagram of a typical temperature control system, highlighting the feedback loop, is shown in Figure 1.

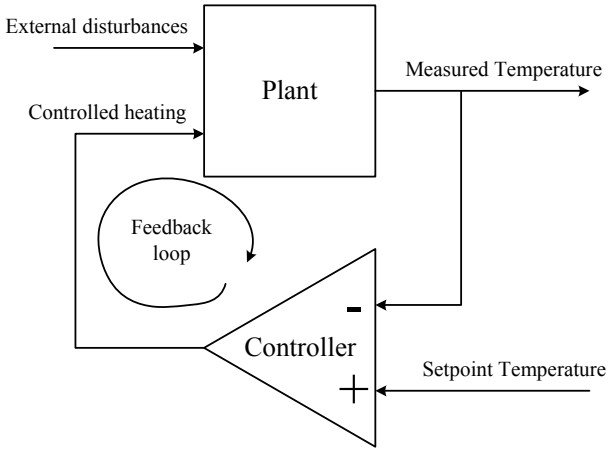

Figure 1. A simple block diagram of a temperature control system.

The purpose of the controller is to maintain the temperature of the plant close to the setpoint temperature and to minimise the effects of the external disturbances on the plant temperature.

There is a wide variety of controllers available. This is in part due to the number of different quantities that are controlled and the number of different sensors used for measuring the different quantities. However, the variety is also essential because some systems are very simple to control, or do not require especially accurate control, while others may exhibit particularly troublesome behaviour or require a very high level of control.

On/Off Controllers

On/Off controllers are the simplest controllers: they simply turn the heating off if the temperature is too high and turn it on if the temperature is too low. Generally, the setpoint is the only adjustable parameter in an On/Off controller.

Figure 2 is a graph of the output characteristic of a typical On/Off controller. In this case the controller setpoint (target temperature) is 20 °C. This controller also has 2 °C of hysteresis. The hysteresis forces the control system to cycle above and below the setpoint. When the controller is heating the plant, it waits until the measured temperature is 21 °C before switching off. When the power is off, it waits until the measured temperature is below 19 °C before switching the power on.

Hysteresis may seem undesirable, but in systems where very large loads are being controlled, hysteresis is necessary. The hysteresis prevents the system from switching too frequently ('chattering') and potentially destroying large switches or valves. In some commercial controllers the amount of hysteresis is adjustable or programmable. In many low-cost On/Off controllers the hysteresis is fixed, typically at 1 °C or 2 °C (2 °F to 4 °F in the USA).

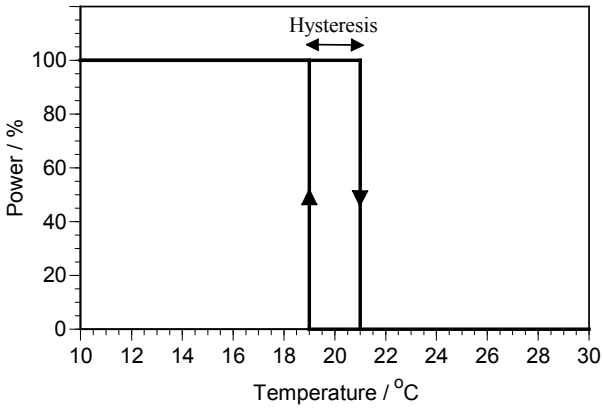

Figure 2. The typical characteristic of an On/Off controller with hysteresis.

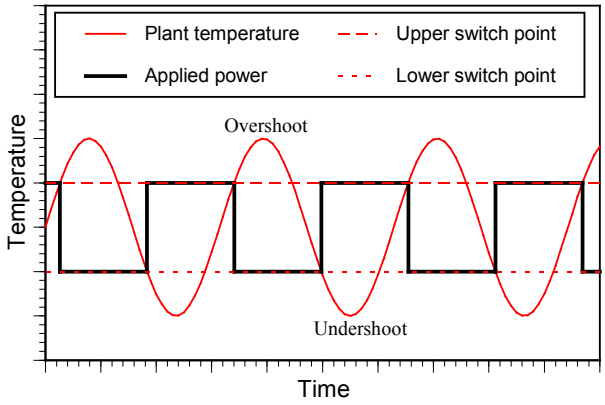

Figure 3. Typical system response to an On/Off controller showing overshoot and undershoot.

Figure 3 shows the typical response of a system controlled by an On/Off controller. Usually, because the process cannot respond rapidly to the controller, the temperature overshoots and undershoots by some fraction. In systems where the response is slow or there are long delays, the temperature range over which the process is controlled may be many times larger than the hysteresis of the controller.

Proportional Controllers

The next most sophisticated controller is a proportional controller, which adjusts the power supplied to the plant in direct proportion to the difference between the plant temperature and the setpoint temperature. That is,

$$\text{Power} = K \times \left(T_{\text{setpoint}} - T_{\text{plant}} \right), \qquad (1)$$

where K is called the proportional gain. Figure 4 shows the power-temperature characteristic for a proportional controller with a setpoint at 60 °C.

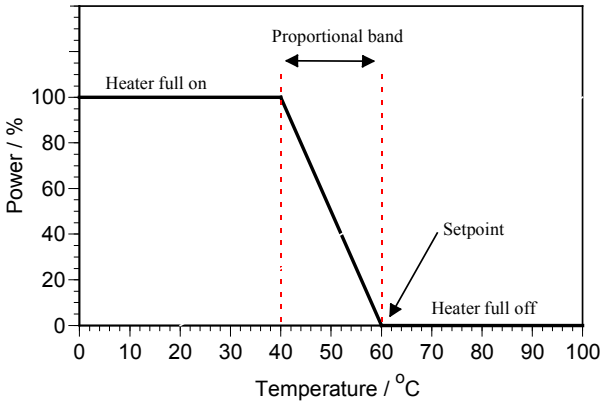

Figure 4. The typical operating characteristic of a proportional controller.

For practical reasons there is a limit to the amount of power that a controller or its heaters can supply. Also, in most cases they cannot supply cooling. Therefore, at temperatures well above or well below the setpoint, the controller is either 'full on' or 'full off'. The remaining region where the controller produces a signal proportional to the temperature difference is called the proportional band or just PB. In Figure 4, the PB is 20 °C. Often, the PB is expressed as a percentage of the full-scale range of the controller. In this case, if the controller works over the range 0 °C to 200 °C, the PB is 10 %. The actual gain of the controller, K, as given in Equation (1), depends on the maximum power of the heater:

$$K = \frac{\text{Maximum heater power (W)}}{\text{Proportional band (°C)}}. \qquad (2)$$

If, for example, the maximum heater power is 1 kW, then the gain of the controller of Figure 4 is 50 W/°C; the smaller the proportional band, the higher the gain of the controller.

One problem with proportional controllers is that the plant generally does not settle at the setpoint (see Figure 5). This happens because for the controller to supply power there must be a difference between the plant temperature and the setpoint (see Equation (1)).

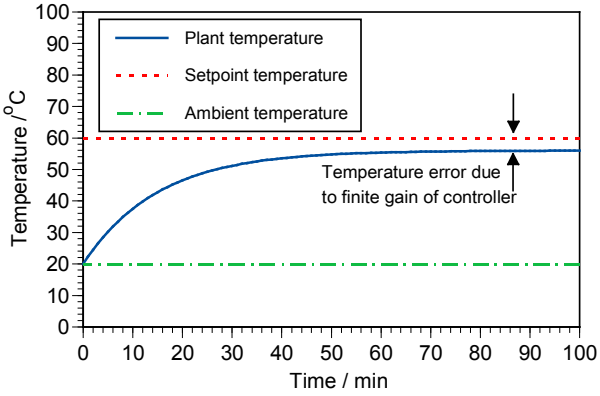

Figure 5. The settling response of a proportional control system showing the setpoint error.

We can model the setpoint error in Figure 5 by assuming the power lost by the plant depends on its temperature:

$$\text{Power lost} = \frac{T_{\text{plant}} - T_{\text{ambient}}}{R}. \qquad (3)$$

The R in Equation (3) is called a thermal resistance. The heat flowing through the resistance is proportional to the temperature difference across it. (This is similar to Ohm's law, where a current through an electrical resistance is proportional to the voltage across it.) If, for example, the plant loses an additional 5 W for every 1 °C rise in plant temperature, the thermal resistance is 0.2 °C/W.

Since the power supplied by the controller must be equal to the power lost by the plant, we can combine Equations (1) and (3) to determine the plant temperature:

$$T_{\text{plant}} = \frac{RK}{1+RK} T_{\text{setpoint}} + \frac{1}{1+RK} T_{\text{ambient}}. \qquad (4)$$

This equation explains several important features of proportional control systems, all of which depend on the

term 1 + *RK*, called the loop gain of the system. *RK* is the product of the thermal resistance and the controller gain. Using the values of *R* (0.2 °C/W) and *K* (50 W/°C) from the examples above, we would obtain a loop gain of 11. Note that the units for *R* and *K* cancel, so that the loop gain is simply a number.

The second term of Equation (4) shows that the plant temperature depends on the ambient temperature, but the effect is reduced by the loop gain. If, for example, the loop gain is 10, and the ambient temperature changes by 20 °C, then the plant temperature changes only 2 °C. In all control systems, a high loop gain helps suppress external disturbances such as changes in power supply voltage (which affects heater power), and the flow rate of goods flowing through a furnace or oven, as well as changes in the thermal environment of the plant.

If we rewrite Equation (4) in a slightly different form,

$$T_{\text{plant}} = T_{\text{setpoint}} + \frac{T_{\text{ambient}} - T_{\text{setpoint}}}{1+RK}, \qquad (5)$$

we can see that the plant temperature is close to the setpoint, but there is a small error called the setpoint error, that depends on the difference between the ambient temperature and the setpoint. The error is reduced by the loop gain; therefore, the higher the loop gain, the closer the plant is to the setpoint.

For the best accuracy and immunity to external disturbances, the loop gain should be as high as practical. Unfortunately there are limits; if the gain is too high the system will become unstable and oscillate. For industrial control systems it may be difficult to obtain a loop gain above 5, but for well-designed laboratory systems, the loop gain could be 1000 or more. Part 2 of this guide explains the stability problem in detail and explains methods for increasing the loop gain.

For many industrial systems where the loop gain may not be very high, the large difference between the setpoint and the actual plant temperature can be annoying and can mislead the plant operators. To overcome this problem, many proportional controllers have a 'reset' adjustment. The reset is usually programmable in digital controllers, but is commonly a recessed screw adjustment on the front panel of analogue controllers. The adjustment allows the operator to offset the setpoint until the plant temperature and the indicated setpoint are the same. Operators can then see at a glance whether the plant is at the correct temperature or not.

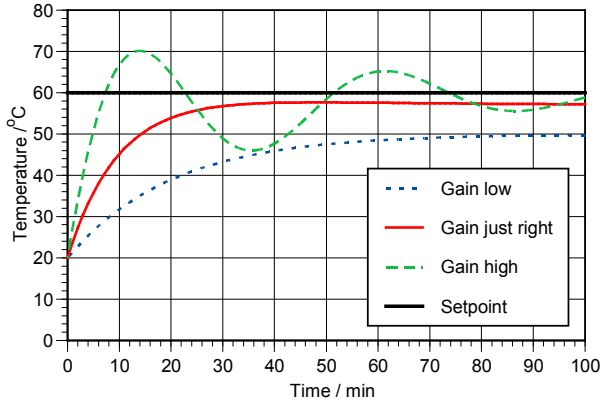

Figure 6. A proportional controller with different gain settings.

The behaviour of a proportional control system depends quite strongly on the amount of gain applied, as shown in Figure 6. At low gain (large proportional band) the setpoint error is large and the settling time can be long. With high gains (small proportional band) the setpoint error is small but the system becomes unstable and tends to oscillate. With a good gain setting, there is a small setpoint error and, typically, a small overshoot beyond the final settling point. In the gain-just-right curve of Figure 6, the overshoot is just detectable.

To tune a proportional controller, gradually increase the gain until the system starts to 'ring' or 'hunt' (show signs of oscillation). The ringing will show on a chart recorder or logger as a trace with an oscillating pattern, as shown in the gain-high curve in Figure 6. Now turn the gain down about 50%: the oscillations should die away quickly. You can make fine adjustments to the gain depending on the amount of overshoot you can tolerate. Check the response of the system by changing the setpoint or restarting the system from cold.

PID Controllers

PID controllers are very similar to proportional controllers, except that they have three control parameters:

***Proportional gain, K*:** This parameter is exactly the same as the gain in the proportional controller.

***Integral time* constant, τ_I:** The integral term in the PID controller continually sums the difference between the plant temperature and the setpoint:

$$\text{Power} = \frac{K}{\tau_I}\int (T_{\text{setpoint}} - T_{\text{plant}})dt, \qquad (6)$$

where *K* is the proportional gain and *t* is time. The longer the two temperatures are different, the greater the power that is applied. The power stops increasing only when the two temperatures are the same. In this way, the integral part of the controller action eliminates the setpoint error, and provides stability against long-term external disturbances. For this reason it is also sometimes called 'automatic reset'. The rate of increase in the power depends on the integral time constant. Note that a small time constant, τ_I, corresponds to a high gain.

Differential time constant, τ_D: The differential term in the PID controller provides feedback that depends on the rate of change of the plant temperature:

$$\text{Power} = -K\tau_D \frac{dT_{\text{plant}}}{dt}. \qquad (7)$$

In effect, this term slows the plant and prevents rapid changes. It can be thought of as 'anticipating' the plant behaviour, or as damping unwanted rapid movement, much like a shock absorber. The benefit of this term is that it enables the controller to operate at higher loop gains without becoming unstable.

The overall equation for the PID controller output is:

$$\text{Power} = K\left[(T_{\text{setpoint}} - T_{\text{plant}}) + \frac{1}{\tau_I}\int (T_{\text{setpoint}} - T_{\text{plant}})dt - \tau_D \frac{dT_{\text{plant}}}{dt}\right]. \qquad (8)$$

Because there are now three parameters instead of just one, PID controllers are more difficult to tune than

proportional controllers. Indeed, there are dozens of different tuning schemes, each working slightly better or worse depending on the type of system that is controlled; there is no perfect tuning scheme. The simple scheme we give here, called the Ziegler-Nichols method, is probably the best known. It works quite well on a large number of systems and does not require extensive system testing or measurement beforehand.

The tuning method is very similar to the method used for tuning proportional controllers (see Table 1). First, turn the derivative and integral actions off, so that the controller is purely a proportional controller. Slowly turn up the gain (reduce the proportional band) until the system just begins to oscillate; that is, the ringing seen in the gain-high curve of Figure 6 persists for a long time. Record the gain setting: call this number K_0. Also record the period of the oscillations: call this τ_0. The three control parameters are then set according to the following table. Note that the controller can also be set up as a pure proportional controller or as a proportional plus integral controller.

Table 1. Tuning constants for a PID controller, according to the Ziegler-Nichols method.

Controller	K	τ_I	τ_D
P	0.50 K_0	-	-
PI	0.45 K_0	0.8 τ_0	-
PID	0.60 K_0	0.5 τ_0	0.12 τ_0

PID controllers are not a cure-all for control problems. Indeed, there are some systems where PID controllers provide practically no advantage over proportional controllers. We discuss the limitations of PID controllers in the Part 2 of this guide.

Intelligent Controllers

Microprocessors and computers have enabled the development of a very wide range of intelligent controllers, many of which may simply be lines of code in the software of a large process control system. Most intelligent controllers are simply digital versions of PID controllers. However, they often have additional useful features.

One of the most useful intelligent controllers are auto-tuning PID controllers. These are PID systems with additional software that automatically adjusts the three PID control parameters to achieve the best performance. There are a variety of systems using different tuning algorithms and different criteria for 'best' performance.

For systems that are difficult to control, adaptive controllers can be useful. These controllers build up a computer model of the plant, and then use the model to determine a suitable control algorithm. Both adaptive and auto-tune controllers monitor the response of the plant to the controller signals, and in some cases 'twiddle' the control signals to investigate the plant response. In some controllers, especially some of the earliest examples, the twiddling can cause unwanted excursions in temperature, and it pays to keep a close watch on the plant when the controllers are first installed.

Because of the large variety of more exotic controllers, and their relatively infrequent use, advice on their application is beyond the scope of this guide. For information on the suitability of controllers for different applications you should consult the manufacturers or their agents, many of whom will have detailed operating instructions, applications information, and often experience of specific industries.

Part 2: Stability and Tuning

Control systems depend on negative feedback for their effectiveness. Unfortunately it is quite easy to build systems where positive feedback occurs. That is, the controller reinforces unwanted behaviour instead of opposing it. This leads to plants that may be driven to extremes of temperature or burst into wild oscillations.

The purpose of this section is to describe how this unwanted behaviour happens and how to design and tune PID systems to avoid it. We begin first with introductions to the use of electrical analogues to describe thermal systems, and to the frequency response of systems. While we have attempted to keep the mathematics to a minimum, some maths is essential. There are also some concepts and abstract ideas that may be unfamiliar. Hopefully there is enough description in the text to convey the ideas. If you are not familiar with the concepts, it may take a couple of readings before it all makes sense.

Electrical Analogue Circuits

In Part 1, we modelled the heat lost from the plant using a thermal resistance. This model followed the similarity of Ohm's law for electricity and Fourier's law for heat. Ohm's law is

$$V = I\,R, \tag{9}$$

or voltage is equal to current multiplied by electrical resistance. Fourier's law is

$$\Delta T = q\,R, \tag{10}$$

or temperature difference is equal to heat flow multiplied by thermal resistance.

The similarity of the two laws means we can build electrical analogue models of thermal systems. Electrical analogues are convenient because they can be analysed using simple electrical calculations. Additionally, most engineers, scientists, etc., are more familiar with electrical systems than thermal systems. Table 2 shows the set of analogous quantities used in this guide.

Figure 7 shows a simple thermal system and its electrical analogue model. In the model, the current source, resistor and capacitor respectively represent the heater, the thermal insulation around the tank, and the heat capacity of the tank and its contents.

Table 2. Analogous thermal and electrical quantities.

Thermal quantity	Electrical quantity
Temperature difference (ºC)	Voltage difference (V)
Heat flow (W)	Current (A)
Thermal resistance (ºC/W)	Electrical resistance (Ω)
Heat capacity (J/ºC)	Electrical capacitance (F)

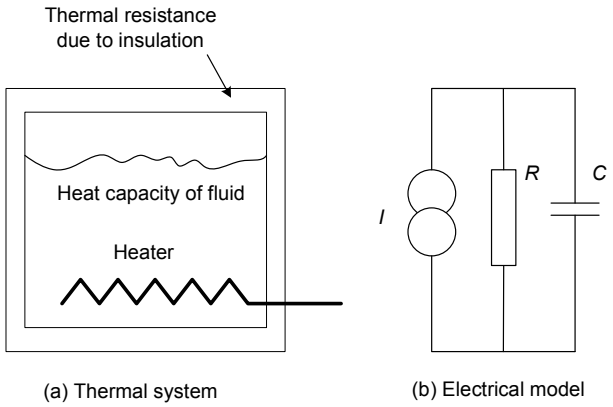

Figure 7. A simple example of a thermal system, a heated tank full of fluid, and its electrical analogue model.

With the real thermal system, the heater slowly heats the tank and its contents, but as the tank temperature increases, the losses through the insulation increase. Eventually the system reaches its maximum temperature where the heater is only able to replace the heat lost through the insulation and no more.

In the model, the current from the current source slowly charges the capacitor. When the voltage across the capacitor reaches the point where the current through the resistor balances that provided by the source, the voltage stops increasing.

In both the thermal system and the electrical model, the losses through the resistance increase as the temperature or voltage increases. As the losses increase, the tank temperature (or capacitor voltage) increase more slowly. As we will see shortly (Figure 8), the slow settling of the system to its final temperature or voltage is characterised by the 'time constant' of the system. The value of the time constant is given by the product of the thermal resistance and the heat capacity, and has units of seconds. The time constant is approximately the time taken for the temperature to change to within about 37 % of the final value, following a step change in the heat or current. To calculate the time constant we need to know how to calculate the thermal resistance and heat capacity.

The thermal resistance of a layer of insulation is given by

$$R = \frac{d}{\sigma A}, \qquad (11)$$

where σ is the thermal conductivity of the insulating material (W.°C^{-1}.m^{-1}), d is the thickness of the insulating layer (m), and A is the total area of the layer (m^2). The heat capacity of an object is given by

$$C = c_p m \qquad (12)$$

where c_p is the specific heat for the material (J.°C^{-1}.kg^{-1}), and m is the mass of material (kg).

To illustrate these calculations, suppose the tank in Figure 7 is surrounded by a layer of polystyrene, 10 mm thick with total surface area of 10 m^2, and the tank contains 0.8 m^3 (800 kg) of water. Polystyrene has a thermal conductivity of 0.08 W.m^{-1}.°C^{-1}, so the thermal resistance is

$$R = \frac{0.01}{0.08 \times 10} = 0.0125 \ \text{°C.W}^{-1}.$$

The specific heat of the water is approximately 4200 J.kg^{-1}.°C^{-1}. Hence, the heat capacity of the tank (ignoring the tank itself) is approximately

$$C = 4200 \times 800 = 3.36 \text{ x } 10^6 \text{ J.°C}^{-1}.$$

The time constant of the system, in seconds, is given by the product RC and is 42,000 seconds (11.7 hours).

The maximum temperature rise of the tank occurs when the heat input by the heater is balanced exactly by the losses through the insulation. This temperature difference is given by the heater power times the thermal resistance, which is Fourier's law (Equation (10)). If the tank is heated by a 5 kW heater, then the maximum temperature difference

$$\Delta T = q\,R = 5000 \text{ W} \times 0.0125 \text{ °C.W}^{-1} = 62.5 \text{ °C}.$$

With the final temperature and the time constant determined, we can now calculate the heating curve. The increasing losses through the insulation or resistance mean that the temperature changes quickly initially and then slows as it approaches the final temperature. This behaviour is described by the exponential function

$$T = T_{\text{ambient}} + qR\left[1 - \exp(-t / RC)\right], \qquad (13)$$

where t is the time in seconds. If the initial temperature of the tank is equal to the ambient temperature, 20 °C, and the heater is full on, the temperature slowly rises according to

$$T = 82.5\,\text{°C} - 62.5\exp(-t/42{,}000)\,\text{°C}.$$

After a long time, the temperature will settle at 82.5 ºC. The heating curve is plotted in Figure 8.

The time constants of the various objects in a controlled system are the key properties required to understand the system behaviour. In most thermal systems there are many distinct components, all with an associated time constant. The plant itself usually has the largest time constant, which may range from minutes to days. Other time constants include those due to the heaters and temperature sensors. If the heaters or thermometers are in special pockets or thermowells, their response may be characterised by two or more time constants.

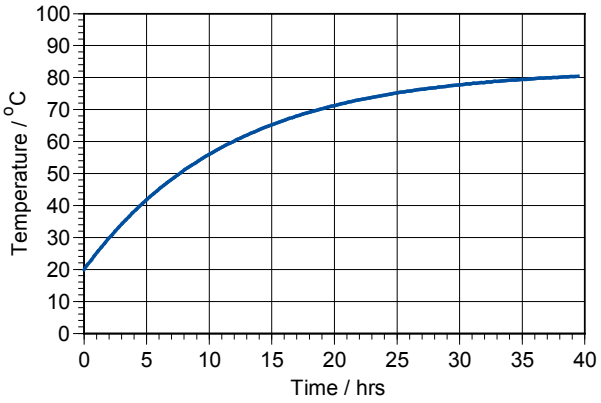

Figure 8. The heating curve for the system of Figure 7 using the numerical values given in the text.

Frequency Response of RC Networks

When the loop gain of a controlled system is too high, the system becomes unstable and tends to oscillate at a specific frequency. In order to understand why this happens, it is necessary to understand the frequency response of *RC* networks.

Figure 9 shows the response of a large object (e.g., the tank of Figure 7) to pulsed heating at two different frequencies. In both cases the rates of cooling and heating are the same, but because the pulses are much shorter in one case, the changes in temperature are much smaller.

Another, less obvious, feature of Figure 9 is that the peak temperature occurs after the middle of the heating pulse. That is, there appears to be a delay between the peak of the heating and the peak of the response to the heating.

Both of these features are characteristics of electrical and thermal capacitances. Both can be modelled as a frequency-dependent impedance, Z_C,

$$Z_C = \frac{1}{j2\pi fC}, \qquad (14)$$

where C is the capacitance, f is the frequency, and j represents a 90° phase shift. The temperature change caused by heat flowing into this thermal impedance is the same as for Fourier's law; i.e., $\Delta T = q\, Z_c$. Equation (14) captures the two main characteristics of capacitors in a single equation. Firstly, the impedance falls in direct proportion to frequency. This means that the temperature changes caused by a sequence of heating pulses fall in direct proportion to the frequency of the pulses.

The second feature of (14) is that the capacitor causes a 90° phase shift. This is described by the imaginary number $j = \sqrt{-1}$, which is used by engineers to represent a 90° phase shift. Two successive 90° phase shifts are given by $j^2 = -1$, which means that the signal is inverted. Four successive 90° phase shifts are given by $j^4 = 1$, which means that the signal has been rotated through 360° and is back in phase with the original signal.

Figures 10 and 11 show the amplitude and phase response of a resistor and capacitor in parallel, as in Figure 7(b). The temperature change across the combination of the two impedances is

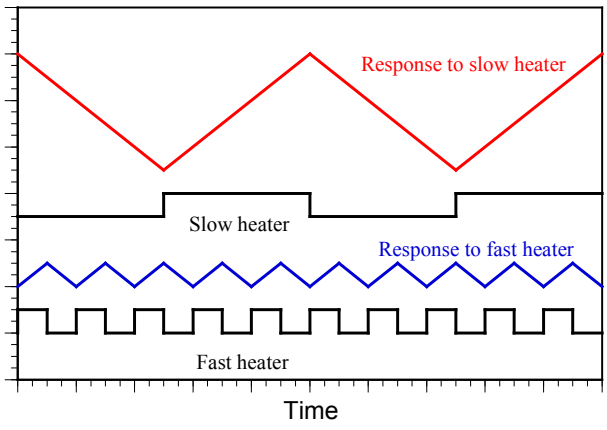

Figure 9. Heat capacity response to fast and slow heaters: the higher the frequency, the smaller the temperature changes.

$$\Delta T = q \times \left(\frac{1}{R} + \frac{1}{Z_c}\right)^{-1} = q\frac{R}{1 + j2\pi fRC}. \qquad (15)$$

This equation contains information on both the amplitude and phase response. We can also give separate equations for the amplitude and phase. The amplitude is

$$|\Delta T| = q \times R\left(\frac{1}{1 + 4\pi^2 f^2 R^2 C^2}\right)^{1/2} \qquad (16)$$

and is plotted in Figure 10. At very low frequencies, the resistor has the lowest impedance so the behaviour of the system is like the resistor by itself, and the observed temperature changes are independent of frequency. At very high frequencies, the capacitor has the lowest impedance so it dominates the behaviour of the system, and the response falls with increasing frequency.

The resistor and capacitor have the same impedance at the frequency

$$f_0 = \frac{1}{2\pi RC}, \qquad (17)$$

and in the region near this frequency we see gradual change between the resistor and capacitor behaviour. Note that the *RC* product in the denominator of Equation (17) is the time constant.

Both systems in Figure 7 are actually low-pass filters. In the thermal system and at frequencies below f_0, the temperature changes are in direct proportion to changes in heat flow and independent of frequency. At frequencies above f_0, the changes in temperature get much smaller as the frequency increases; in effect they are attenuated by the ratio f_0/f. On a log-log graph, the behaviour is approximated well by the straight lines shown in Figure 10. The two lines meet at f_0, which is called the cut-off frequency of the filter.

The right hand axis of Figure 10 gives the relative frequency response in decibels, dB:

$$\text{response in dB} = 20\log_{10}\left[\frac{R(f)}{R(0)}\right], \qquad (18)$$

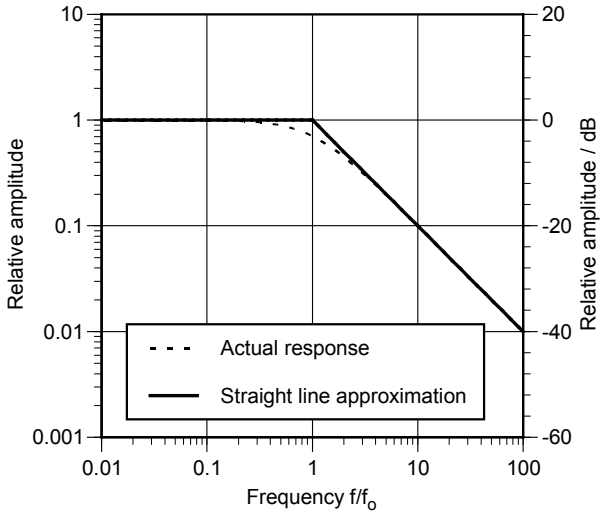

Figure 10. The amplitude response of the simple *RC* system.

where $R(f)$ is the response of the system at frequency f, and $R(0)$ is the response at dc (0 Hz). The decibel scale is commonly used where signals change over a wide range and log graphs are used. Another advantage is that addition of responses in decibels is equivalent to multiplication of the absolute responses, and we will do this later. Note that the slope of the line at frequencies above f_0 is –20 dB per decade. One decade is a 10:1 range of frequencies

The phase response of the simple *RC* system is

$$\phi = -\arctan(f/f_0) \qquad (19)$$

and this is plotted in Figure 11. The phase response of the system shows similar trends to the amplitude response. At very low frequencies the behaviour is dominated by the resistor and there is no phase shift. As the frequency increases, the phase shift increases progressively, until at high frequencies the behaviour is dominated by the capacitor and the signal is phase shifted by –90°. Note that the range of frequencies over which the phase changes is about two decades ($0.1f_0$ to $10f_0$), and the straight lines in Figure 11 give a good approximation to the actual response. At f_0 the phase shift is exactly –45°.

Figure 12 shows an electrical analogue model of a complete temperature-controlled system. This system has three objects with distinct heat capacities: the plant itself, the heater, and the thermometer. For each object with a distinct heat capacity there is a separate heat-transfer process characterised by a distinct *RC* time constant. For example, when the controller turns the heater on, it may take several minutes for the heaters to warm up and start transferring heat to the plant. Similarly, the thermometer may take up to a minute or two to respond to changes in the plant temperature. Each of the time constants will exhibit the same amplitude and phase responses as plotted Figures 10 and 11.

The overall frequency response of the system, caused by the combination of all of the different heat transfer processes, is obtained by multiplying together all of the individual amplitude responses and adding

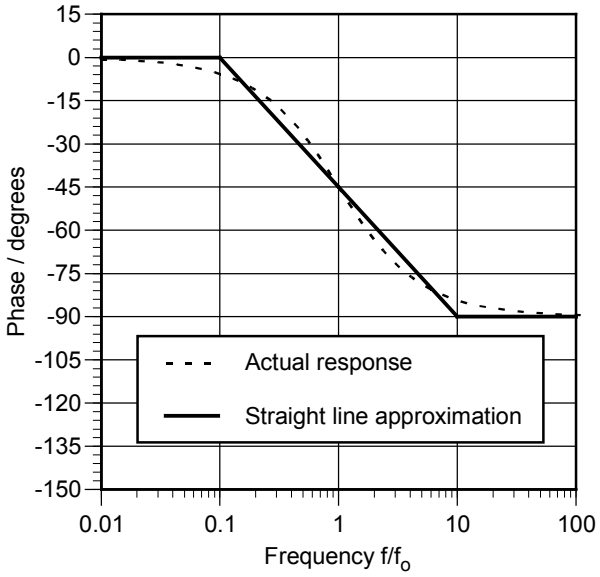

Figure 11. The phase response of a simple RC system.

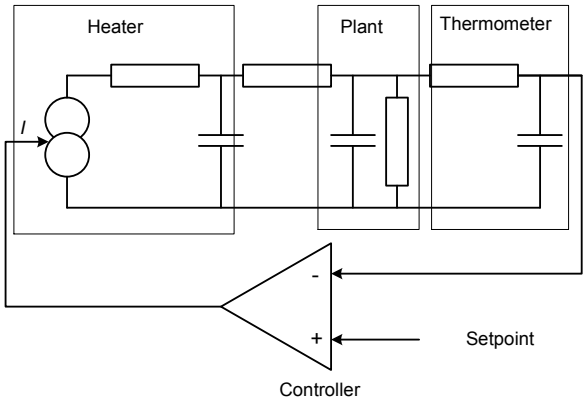

Figure 12. An electrical analogue model of a simple temperature-controlled plant.

together the individual phase responses. The overall amplitude response is plotted in Figure 13. Now instead of a single cut-off frequency, there are three frequencies (called poles) where the amplitude response changes and decreases faster with increasing frequency. The slope between the first and second pole is –20 dB/decade, between the second and third pole it is –40 dB/decade, and after the third pole, the slope of the line is –60 dB/decade.

We could also plot the overall phase response of the system. However, it is just as simple to infer the phase response from Figure 13. Remember that the phase shift at a single pole (Figure 11) is –45°, and that it gradually changes from 0° to –90° over the two decades either side of the pole. From Figure 13 we can infer that the phase shift at the first pole is –45°, as it was for the simple example. At the second pole, the full –90° phase shift of the first pole, due to the plant response, has taken effect, and –45° of the phase shift due to the heater response is present. At the third pole, both of the 90° phase shifts due to the heater and the plant have taken full effect, and there is a further –45° phase shift due to the thermometer response. We have noted these phase shifts at each of the poles in Figure 13. In general, the phase shift at the N^{th} pole is $-45° - (N-1) \times 90°$, approximately. Note that the approximation is not so good if the pole is within a decade of another. The behaviour shown in Figure 13 is typical of the complex frequency response seen in many temperature controlled systems.

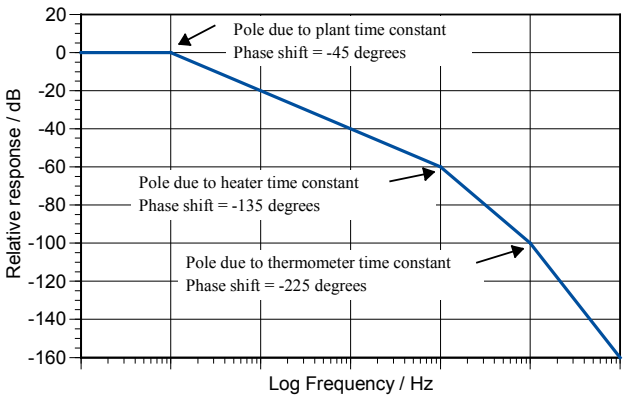

Figure 13. The frequency response of a complete system usually involves several time constants, with their corresponding cut-off frequencies (poles).

Proportional Controllers

In Part 1, we derived Equation (4) for the closed-loop behaviour of the proportional control system. In practice, Equation (4) applies only for static signals, i.e., at 0 Hz. If we want to consider the response at other frequencies then the thermal resistance R must be replaced by the frequency response of the plant $R(f)$. The frequency response of the controlled system is then

$$T_{\text{plant}} = \frac{R(f)K}{1+R(f)K}T_{\text{setpoint}} + \frac{1}{1+R(f)K}T_{\text{ambient}}. \qquad (20)$$

Note that the response $R(f)$ still has the same units as the thermal resistance, °C.W^{-1}, since it describes the temperature change for a given change in input power. The controller gain is, as before, in units of W.°C^{-1}, so that the product of the two, $R(f)K$, is still a dimensionless number. Note too, that $R(f)K$ has a magnitude and a phase. We discuss the magnitude first.

The dotted line in Figure 14 plots the magnitude of the open-loop gain, $|R(f)K|$, for a system with a similar frequency response to Figure 13. The open-loop gain is the product of the frequency responses of the plant and the controller. Note that the value of $|R(f)K|$ is different at different frequencies. At low frequencies it has the value of 1000 (60 dB). This means that the loop gain at these frequencies is 1001, and external disturbances at these frequencies are attenuated by a factor of 1001. At frequencies above 0.0001 Hz, the magnitude of $R(f)K$ falls slowly with frequency, so that the higher frequency disturbances are not so well attenuated. At frequencies well above 0.1 Hz, $R(f)K$ has a magnitude well below 1, so the controller has no significant effect on external disturbances.

The solid line in Figure 14 is the closed-loop response of the control system. This is calculated as $R(f)K/(1+R(f)K)$; it is the coefficient of T_{setpoint} in Equation (20). The closed-loop response tells us how well the controlled system tracks setpoint changes. Ideally the closed-loop gain should be equal to 1.0 at all frequencies. In the example, the closed-loop gain is close to 1.0 up to frequencies of about 0.1 Hz. Note that the closed-loop response has a pole at 0.1 Hz; that is, the controlled system is about 1000 times faster than the uncontrolled system, which has a cut-off frequency of 0.0001 Hz. This shows an additional benefit of controlled systems: the response times are also reduced by an amount proportional to the loop gain.

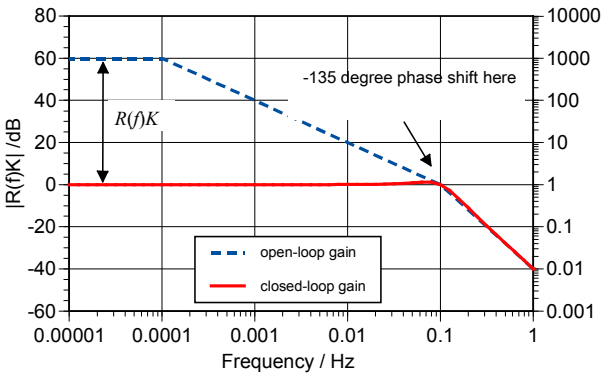

Figure 14. The open- and closed-loop response of a system with a proportional controller.

The solid line in Figure 14 shows a slight increase in the closed-loop gain where the line intersects the open-loop gain. This is the first sign of the oscillatory behaviour that occurs when the gain is too high. In the discussion so far we have considered the magnitude of $R(f)K$ and neglected the phase. As we did with Figure 13, we can infer values of the phase shifts by noting the positions of the poles. For the system in Figure 14, the phase shift of the open-loop gain where the two curves intersect is 135°. That is, the magnitude of $R(f)K$ is 1 with a phase shift of 135°. Now suppose the controller gain was higher, so that the frequency where $|R(f)K| = 1$ is higher, and the phase shift is very close to 180°; that is, the actual value of $R(f)K$ is nearly equal to -1. Then, in Equation (20), the loop gain, $1 + R(f)K$, is very close to zero, and instead of attenuating external influences, the controller amplifies them. At this frequency, the system also becomes very poor at tracking setpoint changes. We can now see that if the closed-loop gain intersects the open-loop gain at the frequency where the phase shift is close to 180°, then the system becomes unstable, and shows signs of oscillation at that frequency.

Figure 15 shows some closed-loop responses for the system of Figure 14 for different gain settings. With very low gain (the uppermost curve) the system is very stable. As the gain increases the 'peaking' in the frequency response becomes more and more apparent. The increasing peaks in the frequency response correspond to increasingly oscillatory behaviour in the time domain. If the controller gain is sufficiently high, and the phase of $R(f)K$ gets beyond 180º, the system becomes unconditionally unstable and bursts into wild oscillations.

To maintain a stable control system, we must therefore limit the controller gain to keep the open-loop phase shift where $|R(f)K| = 1$ well below 180°. Most commonly, the recommended gain setting is one where the 'phase margin' is 45°, i.e., the total phase where $|R(f)K| = 1$ is below 135°. This choice means that the closed-loop gain intersects at the second pole of the open loop response, as shown in Figure 14. If the closed-loop gain intersects the open-loop gain at frequencies above the second pole, then we can expect the excessive peaking seen in the lowest curve of Figure 15, or worse, complete instability.

If the open-loop and closed-loop gains intersect at the second pole, then the dc loop gain is very close to the ratio of the frequencies of the first and second poles. Therefore, to maximise the dc loop gain of a system, we

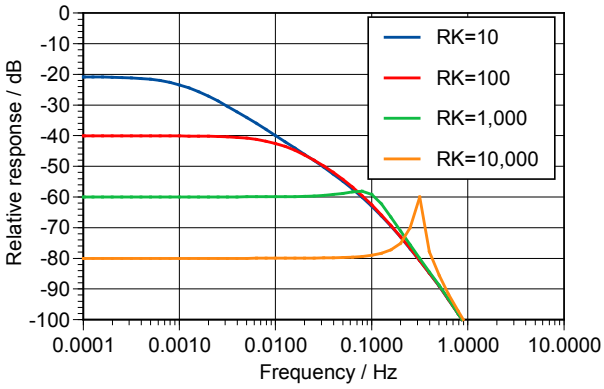

Figure 15. The closed-loop response of a controlled system for different values of dc loop gain. The curves have been offset to show the relationship with the open-loop response.

must make sure that all of the secondary components of the plant (heaters, thermometers, etc.) have a short time constant.

Additionally, if the loop gains intersect at the second pole, then the closed-loop response has its first pole at the second pole of the open-loop response. In general, this means that the fastest response obtainable from a proportional control system corresponds to the second longest time constant of the open-loop system. This is another good reason for making all of the secondary time constants in a plant as short as practicable.

PID Controllers

In the previous section we assumed that the gain of the controller is constant with frequency. PID controllers have a frequency response with two additional features, as shown in Figure 16.

The first of the features is the rising gain with decreasing frequency at low frequencies due to the action of the integrator. The higher controller gain leads to a higher low-frequency loop gain, and hence to better control at low frequencies. This is what gives PID systems good long-term stability and low setpoint error.

Note that at the frequencies where the integrator is active, the gain falls at –20 dB/decade, just as it did with a single-pole response. Indeed, the ideal integrator has a pole at 0 Hz so that the controller gain is infinite at dc. One of the problems with integrators is that they cause an additional –90° phase shift. We will consider the effect of the phase shift on stability shortly.

The second major feature of the PID frequency response is the rising gain at high frequencies. This is due to the action of the derivative term in the controller. The frequency at which the response begins to increase because of the derivative action is called a 'zero' (marked by f_D in Figure 16). Whereas poles are associated with a –20 dB/decade slope and –90° phase shift, zeros introduce +20 dB/decade change in amplitude response and +90° phase shift.

Figure 17 shows the combined response of a PD controller (no integral action) and the plant response. The overall open-loop response of the system is the product of the of the separate plant and controller responses (the sum when the responses are given in decibels). Note particularly, the effect of the zero of the derivative term is to compensate for the second pole in the plant response at 0.1 Hz. This is called a pole-zero cancellation. The benefit of the pole-zero cancellation is that it shifts the frequency at which the 180° phase shift

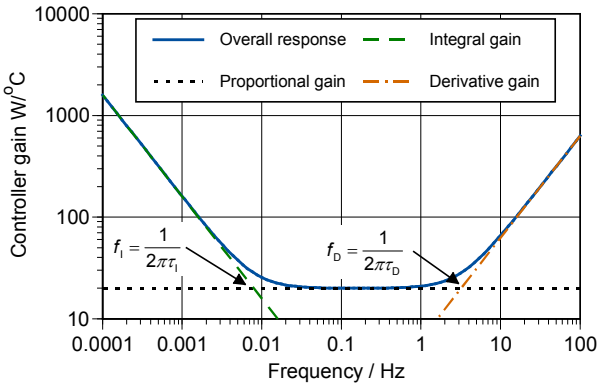

Figure 16. A typical frequency response for a PID controller.

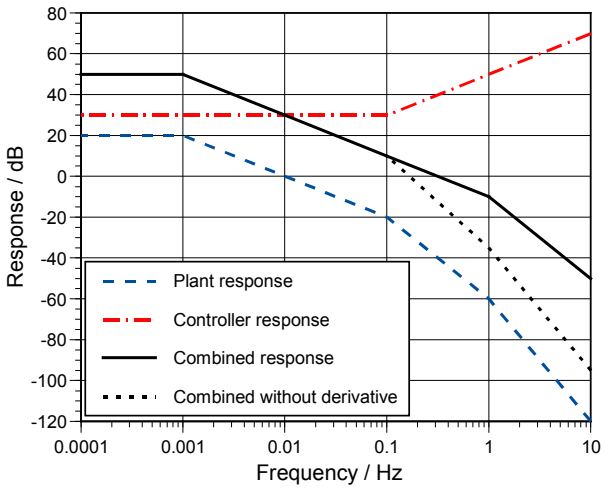

Figure 17. The benefit of the derivative action is to cancel one of the poles in the plant response so that the system has a greater phase margin.

occurs to higher frequencies, and therefore the system can operate with higher loop gain without becoming unstable

Figure 17 might suggest that the derivative time constant should be exactly equal to the second time constant in the plant response. However, such fine tuning is not usually required. The change in phase that occurs with poles and zeros occurs over a range of two decades (see Figure 11), so the positive phase shift due to the derivative action provides benefit even when the pole-zero cancellation is not exact.

We can now return to the issue of instability due to the integrator. The controller frequency response in Figure 16 actually has two zeros in it. One where the derivative and proportional gain intersect at f_D, and one where the integrator gain and proportional gain intersect at f_I. Each zero is associated with a +20 dB/decade change in the amplitude response, and a +90° phase shift. The +90° shift associated with the integrator zero compensates for the –90° phase shift due to the integrator pole at 0 Hz. So long as the integrator zero is more than 1 decade below the frequency where $|R(f)K| = 1$, the integrator can have no effect on the stability. If the integrator zero is within 1 decade of that frequency, then the –90° phase shift from the pole will not be completely compensated by the zero, and the phase margin will be reduced.

From this discussion it is clear that a PID controller can be tuned in many different ways, depending on the type of plant and on the operator's criteria. For example, if short-term stability is required, then the focus of attention should be on maximising the loop gain and application of the derivative term. Also, the integrator time constant τ_I should be long i.e., gain low (see Equation 6), so that the integrator does not introduce any phase shift near the frequency where $|R(f)K| = 1$.

If long-term stability is required then the integrator gain should be as high as practical. In this situation the derivative may not offer much advantage, so the tuning for a PI controller may be satisfactory. In either case, the Ziegler-Nichols tuning constants provide a good starting estimate of the best values for the various gains. Whatever tuning method is employed, the operator should test the system with various types of disturbance, and

both large and small setpoint changes, to ensure the closed-loop behaviour is satisfactory.

Major Causes of Instability

To get a very stable controlled system, there must be one dominant (long) time constant in the system and all the other time constants must be as short as practical. This enables the maximum loop gain to be achieved at low frequencies, and ensures that the closed-loop response extends to high frequencies. There are three common classes of systems where high loop gain can be very difficult to achieve.

Pure Time Delays

Probably the most troublesome cause of instabilities is delay, and there are many possible causes. For example, a pure delay occurs when a gas heater is used to heat a plant, where it takes many seconds to transport the hot air to the plant. Another example may be where the heated fluid in the tank takes a long time to circulate between the heater and the thermometer. Digital systems can also have long delays, where different parts of a large plant are being serviced by a single computer or microprocessor.

Pure delays are troublesome because they introduce large phase shifts. Consider a system with a sinusoidal disturbance at 1 Hz. That is, the signal repeats every 1 s. If the system has a 0.5 s delay somewhere within it, then the controller response will be exactly 180° out of phase with the disturbance: it will reinforce the behaviour and eventually burst into wild oscillations.

The phase shift caused by a pure delay is

$$\phi = -360 f \Delta t \,, \qquad (21)$$

where f is the frequency and Δt is the time delay. Figure 18 shows the phase shift caused by a 10 s delay as a function of frequency. At low frequencies the phase shift is small and has little effect on the stability of the control system. However, at frequencies above 0.5 Hz the phase shift is more than $-180°$ and changing very rapidly. In these conditions it is very easy to induce wild oscillations with relatively small increases in gain, and the 90º positive phase shift from the derivative action in a PID controller has little effect.

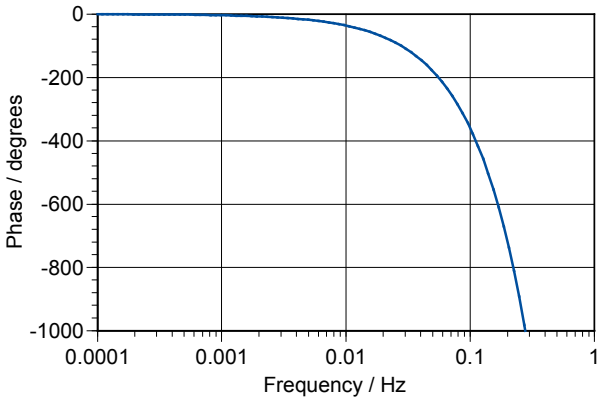

Figure 18. Phase shift versus frequency for a 10 second delay.

Complex Multi-Pole Systems

In the discussion of the stability of proportional and PID control systems above, the examples used were of relatively simple systems with a good separation between the poles in the frequency response. Plant responses are not always this simple. In cases where the plant frequency response has many poles, all near the same frequency, the phase can change very quickly with frequency. As with delay problems, it is very difficult to obtain high loop gain, and the derivative action on a PID controller will not have much effect. Although the cause is different, this type of behaviour is very similar to that of systems with delays.

Integrator Windup

All of the analysis above has assumed that the controller is always acting within the proportional band. However, it is quite common at start-up and with large setpoint changes that the heaters will saturate (be fully on or fully off). Under these conditions, controllers with integral action normally exhibit large overshoots due to a condition called integrator windup.

The integration term constantly integrates the difference between the setpoint temperature and the actual plant temperature, and, ideally, applies power in proportion to that difference. If the heater saturates (fully on say) then the controller cannot provide as much power as it requires, the plant temperature will increase more slowly than expected, and hence the integrator signal continues to grow faster than it should. The integrator windup causes the plant to oscillate about the setpoint until the controller action is entirely within the proportional band, at which point the system settles quickly.

Integrator wind up is not a simple problem to solve, and there are many proprietary techniques of varying utility. One of the simplest solutions to implement in digital systems is to switch the integrator off whenever the controller action is outside the proportional band.

Further Reading

K J Astom and T Haggland, *Advanced PID Control*, ISA, Research Triangle Park, NC USA, 2006

Prepared by D R White, June 2008.