PID CONTROL

Araki M.
Kyoto University, Japan

Keywords: feedback control, proportional, integral, derivative, reaction curve, process with self-regulation, integrating process, process model, steady-state error, overshoot, decay ratio, rise time, settling time, gain margin, phase margin, action mode, ultimate sensitivity test, step response test, Ziegler–Nichols' tuning methods, Chien–Hrones–Reswick's tuning method, modulus-optimum tuning method, symmetrical-optimum tuning method, anti-windup, two-degrees-of-freedom controller

Contents

Summary

The PID controller, which consists of proportional, integral and derivative elements, is widely used in feedback control of industrial processes. In applying PID controllers, engineers must design the control system: that is, they must first decide which action mode to choose and then adjust the parameters of the controller so that their control problems are solved appropriately. To that end, they need to know the characteristics of the process. As the basis for the design procedure, they must have certain criteria to evaluate the performance of the control system. The basic knowledge about those topics is summarized in this article.

1. Introduction

"PID" is an acronym for "proportional, integral, and derivative." A PID controller is a controller that includes elements with those three functions. In the literature on PID controllers, acronyms are also used at the element level: the proportional element is referred to as the "P element," the integral element as the "I element," and the derivative element as the "D element." The PID controller was first placed on the market in 1939 and has remained the most widely used controller in process control until today. An investigation performed in 1989 in Japan indicated that more than 90% of the controllers used in process industries are PID controllers and advanced versions of the PID controller.

"PID control" is the method of feedback control that uses the PID controller as the main tool. The basic structure of conventional feedback control systems is shown in Figure 1, using a block diagram representation. In this figure, the process is the object to be controlled. The purpose of control is to make the process variable y follow the set-point value r. To achieve this purpose, the manipulated variable u is changed at the command of the controller. As an example of processes, consider a heating tank in which some liquid is heated to a desired temperature by burning fuel gas. The process variable y is the temperature of the liquid, and the manipulated variable u is the flow of the fuel gas. The "disturbance" is any factor, other than the manipulated variable, that influences the process variable. Figure 1 assumes that only one disturbance is added to the manipulated variable. In some applications, however, a major disturbance enters the process in a different way, or plural disturbances need to be considered. The error e is defined by $e = r - y$. The compensator $C(s)$ is the computational rule that determines the manipulated variable u based on its input data, which is the error e in the case of Figure 1. The last thing to notice about Figure 1 is that the process variable y is assumed to be measured by the detector, which is not shown explicitly here, with sufficient accuracy instantaneously that the input to the controller can be regarded as being exactly equal to y.

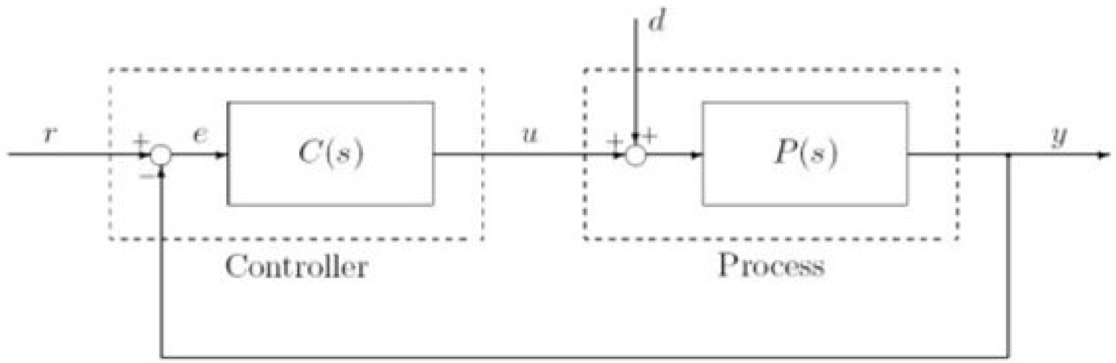

Figure 1. Conventional feedback control system

Early PID control systems had exactly the structure of Figure 1, where the PID controller is used as the compensator $C(s)$. When used in this way, the three elements of the PID controller produce outputs with the following nature:

- P element: proportional to the error at the instant t, which is the "present" error.
- I element: proportional to the integral of the error up to the instant t, which can be interpreted as the accumulation of the "past" error.
- D element: proportional to the derivative of the error at the instant t, which can be interpreted as the prediction of the "future" error.

Thus, the PID controller can be understood as a controller that takes the present, the past, and the future of the error into consideration. After digital implementation was introduced, a certain change of the structure of the control system was proposed and has been adopted in many applications. But that change does not influence the essential part of the analysis and design of PID controllers. So we will proceed based on the structure of Figure 1 up to Section 6, where the new structure is introduced.

The transfer function *C(s)* of the PID controller is

$$C(s) = K_P \left\{ 1 + \frac{1}{T_I s} + T_D \; D(s) \right\} \tag{1}$$

provided that all the three elements are kept in action. Here, K_P, T_I and T_D are positive parameters, which are respectively referred to as “proportional gain,” “integral time,” and “derivative time,” and as a whole, as “PID parameters.” *D(s)* is the transfer function given by

$$D(s) = \frac{s}{1 + \left(T_D / \gamma\right) s} \tag{2}$$

and is called the “approximate derivative.” The approximate derivative *D(s)* is used in place of the pure derivative *s*, because the latter is impossible to realize physically. In (2), γ is a positive parameter, which is referred to as “derivative gain.” The response of the approximate derivative approaches that of the pure derivative as γ increases. It must be noted, however, that the detection noise, which has strong components in the high frequency region in general, is superposed to the detected signal in most cases, and that choosing a large value of γ increases the amplification of the detection noise, and consequently causes malfunction of the controller. This means that the pure derivative is not the ideal element to use in a practical situation. It is usual practice to use a fixed value of γ, which is typically chosen as 10 for most applications. However, it is possible to use γ as a design parameter for the purpose of, for instance, compensating for a “zero” of the transfer function of the process.

In applying PID controllers, engineers must “design” the control system. In other words, they must first decide which element(s) to keep in action and then adjust the parameters so that their control problems are solved appropriately. To that end, they need to know the characteristics of the process. As the basis for this design procedure, they must have certain criteria to evaluate the performance of the control system. Those topics will be treated in the following four sections. (See *Elements of Control Systems*.)

2. Process Models

Define the unit step function $f_{step}(t)$ by

$$f_{step}(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases} \tag{3}$$

The response $y_{u,\ step}(t)$ of the process variable to the unit-step manipulated variable $u(t) = f_{step}(t)$ directly added to the process at rest is called the “step response” or “indicial response” of the process. The term “reaction curve” is also used, essentially with the same meaning but focusing on the graphical representation. If the step response converges to a finite value K when $t \to \infty$, as exemplified in Figure 2, the process is said to be "with self-regulation" and K is called "stationary gain." If the step response diverges when $t \to \infty$, the process is said to be "without self-regulation." If a process is without self-regulation and its step response approaches a straight line with the slope R, as exemplified in Figure 3, it is said to be "with a single integrator" or simply "integrating." It has been observed that step responses of many processes to which PID controllers are applied have monotonically increasing characteristics as shown in Figures 2 and 3, so most traditional design methods for PID controllers have been developed implicitly assuming this property. However, there exist some processes that exhibit oscillatory responses to step inputs. This topic will be treated later (see Section 6.3).

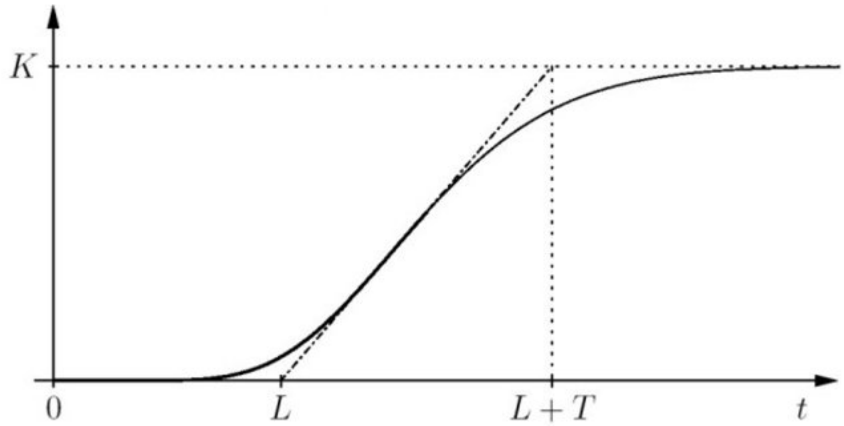

Figure 2. Step response of a process with self-regulation

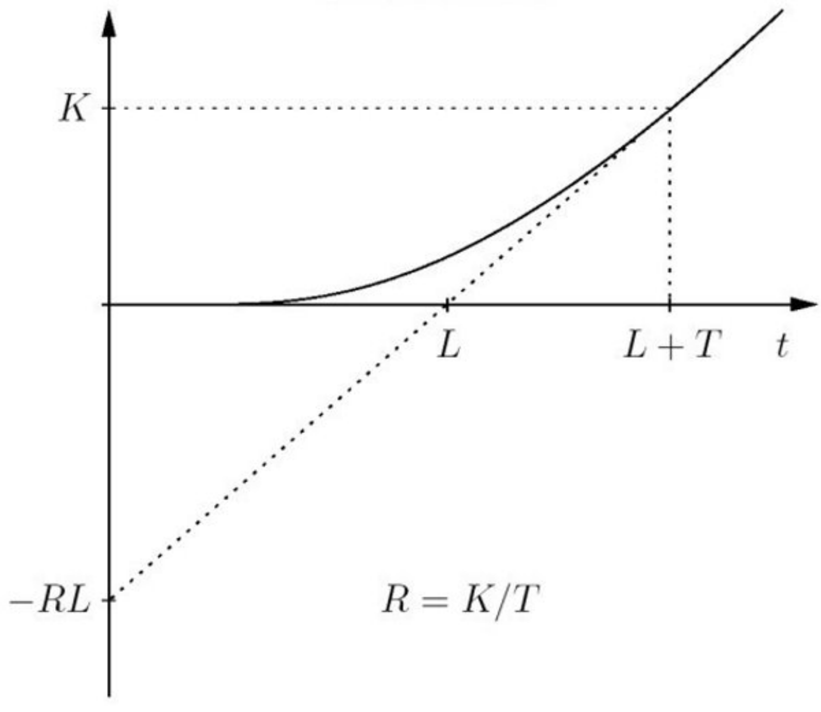

Figure 3. Step response of an integrating process

A more basic assumption employed in the design methods explained in the following is "linearity." "Linearity" means that, if the responses of the process variable to inputs $u_1(t)$ and $u_2(t)$ are, respectively, $y_1(t)$ and $y_2(t)$, then its response to the summed-up input $u_1(t)+u_2(t)$ becomes $y_1(t)+y_2(t)$, all under the condition that the process is at rest at the initial instant. In systems theory, it is generally expected that linearity approximately holds true in a small range of variables, while the approximation error increases as the range increases. This expectation is met in some processes but upset in others. There are processes, for instance, such that the response to the negative step is largely different from the inverse of the response to the positive step. In spite of such reality, the linearity assumption has been employed widely, first because it is difficult to establish a practically tractable general method without this assumption, second because experience shows that the designed results work approximately well for many processes, and third because the results obtained from desk work are in any case insufficient, so that trial-and-error adjustment at actual processes is always needed, and the nonlinear property can be considered in that procedure.

Under the above assumptions, the following transfer functions can be used to model the process. For a process with self-regulation,

$$P(s) = \frac{K}{1+Ts}e^{-Ls} \tag{4}$$

is the simplest model. This model is referred to as the "first-order-lag + pure-delay" model, because $K/(1+Ts)$ is the transfer function of the first-order-lag element whose stationary gain is K and time constant is T, and e^{-Ls} is that of the pure delay whose delay time is L. The simplest model for an integrating process is

$$P(s) = \frac{R}{s}e^{-Ls} = \frac{K}{Ts}e^{-Ls} \qquad R = \frac{K}{T} \tag{5}$$

This model is referred to as the "integrator + pure-delay" model. The parameters K and T of the second expression are redundant by one and so there is no way, mathematically speaking, to determine them uniquely. However, this expression is sometimes used, with understanding that the parameter T is the time constant of the process, first in order to make the denominator Ts dimensionless so that the time scale of the reaction curve is standardized, and second in order to make the equation giving the steepest slope of the reaction curve the same as that for the "first-order-lag + pure-delay" model. The latter makes the turning formulae of PID parameters applicable without confusion (see Section 5.3).

The above two models have long been used as the basis of design methods for PID control systems, because their parameters can easily be determined from simple tests (see Section 5.2), and the designed results are very often sufficient as the initial values to start the trial-and-error adjustment procedure. But recently there has been a move to make full use of the capability of modern computers and sensing systems for adjusting the controller as exactly as possible, based on the initial test or on-line data. For that purpose, the above models are too simple, so more sophisticated models are considered (see Section 6.3).

3. Performance Evaluation of PID Control Systems

PID control systems are evaluated, in normal practice, by their "unit-step set-point response" and "unit-step disturbance response." The "unit-step set-point response" is the response $y_{r,\ step}(t)$ of the process variable to the unit-step set-point value $r(t) = f_{step}(t)$ added to the control system at rest, keeping the disturbance $d(t) = 0$. The "unit-step disturbance response" is the response $y_{d,\ step}(t)$ of the process variable to the unit-step disturbance $d(t) = f_{step}(t)$ added to the control system at rest, keeping the set-point value $r(t) = 0$. The most important property required for those responses is that they converge to constant values $y_{r,\ \infty}$ and $y_{d,\ \infty}$, respectively. Under the linearity assumption, this property is equivalent to stability of the feedback control system. Stability is the condition that must be guaranteed by any means, and the evaluation of control systems only becomes meaningful under this condition.

Assuming stability, the "steady-state errors" $\varepsilon_{r,\ \infty}$ and $\varepsilon_{d,\ \infty}$ to the unit-step set-point input and to the unit-step disturbance input are defined respectively as follows, and their sizes are adopted as evaluation items for the steady-state performance:

$$\varepsilon_{r,\ \infty} = \lim_{t\to\infty}\left\{ 1 - y_{r,\ step}(t) \right\} = 1 - y_{r,\ \infty} \quad \text{and} \quad \varepsilon_{d,\ \infty} = \lim_{t\to\infty}\left\{ 0 - y_{d,\ step}(t) \right\} = -y_{d,\ \infty} \tag{6}$$

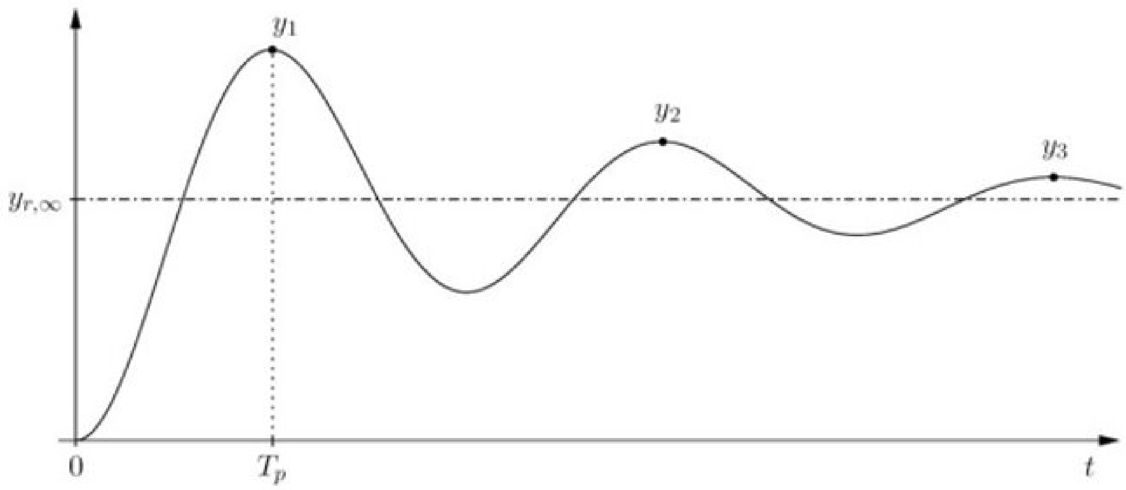

Figure 4. Oscillatory set-point response

The next property to be assessed is the oscillatory nature of the responses. First, oscillation prevents the process variable from settling to the set-point value quickly. Second, for the process variable to become considerably larger than the set-point value can lead to harmful effects on the product as well as on the process. Third, for feedback control systems with the structure of Figure 1, oscillatory responses usually imply small stability margins. Thus the responses need to be not too oscillatory. For evaluation of this property, the unit-step set-point response $y_{r,\ step}(t)$ is used. Assume that $y_{r,\ step}(t)$ has the peaks as illustrated in Figure 4, and let the height of the $i-th$ peak be y_i. The "overshoot value" *A* and the "decay ratio" Γ are defined as follows where $y_{\max}$ is the maximum of y_i's.

$$A = \frac{y_{\max} - y_{r,\ \infty}}{y_{r,\ \infty}} \times 100 \quad \text{and} \quad \Gamma = \frac{y_2 - y_{r,\ \infty}}{y_1 - y_{r,\ \infty}} \times 100 \tag{7}$$

These equations are used to evaluate the wave form of the response. A general guideline is to make *A* not larger than 20% and Γ not larger than 4%. In some problems in which overshooting is harmful, it is necessary to make $A = 0$ (i.e. to make the response free from overshoot).

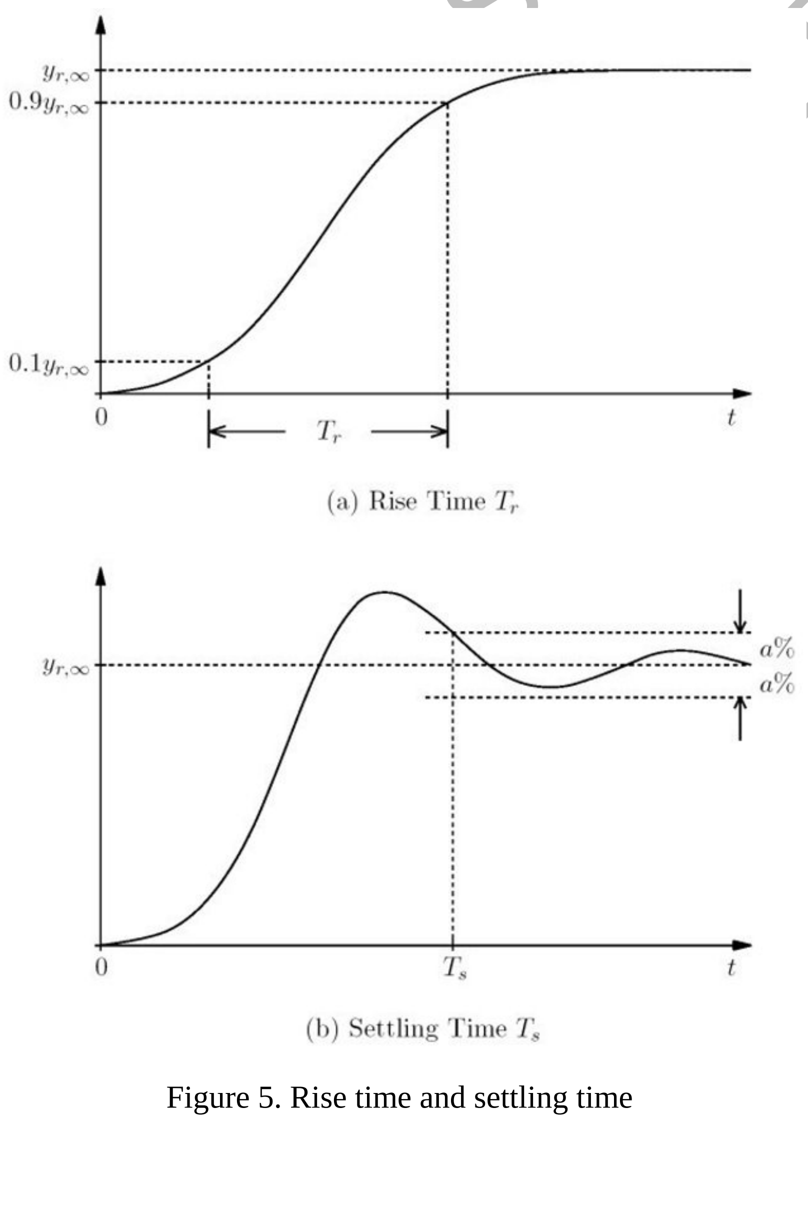

Figure 5. Rise time and settling time

Another important property is the speed of the response. For a unit-step set-point response $y_{r,\ step}(t)$, the "rise time" T_r is defined as the time which it takes for $y_{r,\ step}(t)$ to pass from 10% to 90% of $y_{r,\ \infty}$, and the "settling time" T_s as the minimum time for which the next inequality is met, where the frequently-used value of *a* is 2%.

$$\frac{\left|\, y_{r,\ step}(t) - y_{r,\ \infty}\,\right|}{y_{r,\ \infty}} \times 100 \ \leq \ a \quad \text{for} \quad t \geq T_s \tag{8}$$

The rise time and the settling time are illustrated in Figure 5. These two values are common evaluation items for speed, and their attainable values vary considerably, from seconds to hours, depending on the process.

The preceding two groups of values, about the waveform and the speed of the unit-step set-point response $y_{r,\ step}(t)$ have conventionally been adopted to evaluate the transient-state performance. But to derive a design method analytically, more sophisticated performance indices are needed. *A* representative and successful example of such indices is the ITAE (integrated time-weighted absolute error), which is defined by:

$$\int_0^{\infty} t\left|y_{r,\,step}(t) - y_{r,\,\infty}\,\right| dt \ \text{ and } \ \int_0^{\infty} t\left|y_{d,\,step}(t) - y_{d,\,\infty}\,\right| dt \tag{9}$$

The ITAE can totally evaluate the transient-state performance, and has been widely used as a common basis for discussing PID design methods. Another performance index that should be mentioned is the ISE (integrated squared error), defined by:

$$\int_0^{\infty}\left|y_{r,\,step}(t) - y_{r,\,\infty}\,\right|^2 dt \ \text{ and } \ \int_0^{\infty}\left|y_{d,\,step}(t) - y_{d,\,\infty}\,\right|^2 dt \tag{10}$$

The control method that minimizes the ISE is called "optimal linear quadratic control (LQ)." It is important not only as a design method but also as a theoretical basis for understanding the relationship between classical and modern control strategies.

The steady-state errors, the waveform (e.g. overshoot value and damping ratio) and the speed (e.g. rise time and settling time) are the major items of performance evaluation that can be observed explicitly, in simulation or by running the control system under a fixed condition. But there is another important item that must be considered in design. That is "robustness." "Robustness" means the strength of the control system against the modeling error and changes in the characteristics of the process. The simplest measure of robustness is the gain and the phase margins. A guideline in PID control system design is to ensure a 3–10dB gain margin and 20° phase margin. Here, it must be noted that 3dB and 20° are the endurable minimum values, and that when these values are chosen, control systems are usually very oscillatory and can easily become unstable. If satisfactory damping and robustness are expected, a phase margin of greater than 50°is recommended. The LQ strategy mentioned above, for instance, guarantees a phase margin of greater than 60° and an infinite gain margin. A more direct way of evaluating robustness is to repeat simulation for variety of plausible process models, or to run the system changing the operating conditions. In modern control theory, the sensitivity and complementary sensitivity functions are the major tools for evaluating robustness. (See *Stability Concepts*, *Optimal Linear Quadratic Control*, and *Robust Control*.)

4. Action Modes of PID Controllers

In application, engineers have freedom of using the three functional elements (P, I, and D) of the PID controller in whatever combination they consider most appropriate for their problems. The combination of element(s) used is called the “action mode” of the PID controller. Theoretically, there exist seven action modes. Among them, the five listed in Table 1 are important in practice.

Action mode	**Element(s) used**	**Transfer function** $C(s)$
Proportional (P)	P element only	$C(s)=K_P$
Integral (I) with $T_I=1$	I element only	$C(s)=\frac{K_P}{s}$
Proportional-Integral (PI)	P and I elements	$C(s)=K_P\left(1+\frac{1}{T_I s}\right)$
Proportional-Derivative (PD)	P and D elements	$C(s)=K_P\left\{1+T_D D(s)\right\}$
Proportional-Integral-Derivative (PID)	All 3 elements	$C(s)=K_P\left\{1+\frac{1}{T_I s}+T_D D(s)\right\}$

Table 1. Action modes of PID controllers

5. Design of PID Control Systems

“Design” is an engineering activity that often includes trial-and-error procedures at various levels. In designing PID control systems, the three basic tasks explained in the following must be carried out first, and then repeated if a satisfactory result is not obtained.

5.1. Selection of Action Mode

In selecting the action mode of PID controllers, the following two facts are important. First, in order to make the steady-state errors $\varepsilon_{r,\infty}$ and $\varepsilon_{d,\infty}$ to step inputs zero robustly, it is necessary and sufficient to include the *I* element in the compensator *C(s)*. Second, inclusion of the *I* element in *C(s)* makes the control system more likely to be oscillatory and, in the worst case, unstable. To be more specific, it makes the gain and the phase margins smaller. The details of the latter fact are as follows.

In order to make explanation simple, let us assume that the process is modeled by (4) or (5), as is very often the case in PID applications. By calculating the frequency response of the compensator *C(s)*, it is found that the phase shift of the return ratio *P(s)C(s)* of the control system shown in Figure 1 is increased by 90 ^{o} compared with that of the process *P(s)* at all frequencies if the controller *C(s)* is set in the *I* mode, and that it is increased by 90^{o} in the low frequency range (i.e. for $\omega << 1/T_I$) but remains the same or decreased in the high frequency range (i.e. for $\omega >> 1/T_I$) if the controller *C(s)* is set in the PI or the PID mode. An increase in the overall phase shift naturally causes deterioration of the phase margin, but when the process is with self-regulation (i.e. when *P(s)* is given by (4)), the deterioration can be canceled out by choosing to make the proportional gain K_P small enough. However, when the process is integrating (i.e. when *P(s)* in given by (5)), the above increase of the overall phase shift causes a serious situation. To be exact, if the controller *C(s)* is set in the *I* mode for the integrating process, the phase shift of the return ration *P(s)C(s)* becomes more than 180 ^{o} at all frequencies, and consequently it becomes impossible to stabilize the feedback control system. Thus, to select the *I* mode for an integrating process is prohibitory. The situation with the PI and the PID modes is a little milder. If the integral time T_I of the compensator *C(s)* in the PI or the PID mode is chosen to be large enough that the phase shift of the process *P(s)* around the frequency $\omega = 1/T_I$ is sufficiently small (note that “sufficiently small” in this case means “a little larger than 90°,” because the process *P(s)* is integrating), the overall phase shift can be made less than 180° in a certain interval. Then the phase margin can be made positive by choosing the proportional gain K_P appropriately, and consequently the feedback control system is stabilized. But even though stability is attained in that way, the phase shift of the return ratio in the low frequency range remains nearly 180°, and this fact causes the phenomenon that not only “increase” but also “decrease” in the proportional gain K_P reduces the phase margin, and as a result makes the control system oscillatory. This means that the permissible range of the proportional gain K_P is finite in both directions, and very often fairly narrow. As a result, tuning of the controller parameters becomes difficult. Specifically, it is necessary to retune the controller parameters carefully whenever the characteristics of the process changes, which is often caused by changes in operating conditions.

The following is a guideline for selection of the action mode. In the case of processes with self-regulation, the usual practice is to select PI or PID mode. This selection guarantees zero steady-state errors to step inputs. Inclusion of the D element (i.e.

selecting PID mode) improves the speed of the responses, and consequently serves to suppress the influence of the disturbance more strongly. However, the D element functions effectively only when the parameters are tuned appropriately. This means good maintenance is necessary to make the second choice meaningful.

In the case of integrating processes, deterioration of the oscillatory property caused by inclusion of the *I* element is serious, as explained above, so the usual practice is to select P or PD mode, but PI or PID mode should be selected if making the steady-state error $\varepsilon_{d,\infty}$ zero is a mandatory requirement. In this case, the initial tuning as well as the maintenance work must be done very carefully. Inclusion of the D element is useful, and in many cases necessary, to avoid undesirable oscillatory responses.

In general, selecting P mode makes the control system simple and the design and maintenance activities easy. So if high performance is not necessarily required, this is a practical choice.

5.2. Identification of Process Model Parameters

Two kinds of identification method for process characteristics have been adopted widely in PID control system design: the "ultimate sensitivity test" and "step response methods."

The "ultimate sensitivity test" is carried out, in its standard form, as follows. For the given process, construct the PID control system as in Figure 1 with the compensator *C(s)* in the P mode, and increase the proportional gain K_P gradually, starting with a very small value. Provided that the process has the characteristics as given by Eqs. (4) or (5), the feedback control system remains stable at first. Then at a certain stage it reaches the stability limit; that is, it exhibits a sinusoidal oscillation. The value K_{Pu} of the proportional gain at this stage is referred to as the "ultimate gain," and the period T_u of the exhibited sinusoid as the "ultimate period." The "ultimate angular frequency" ω_u is defined by $\omega_u = 2\pi/T_u$. By the Nyquist stability condition, it can be concluded that the Nyquist locus of the process transfer function *P(s)* crosses the negative real axis at $s = j\omega_u$ and the gain $\left|P\left(j\omega_u\right)\right|$ at that point is $1/K_{Pu}$. Thus, the ultimate sensitivity test gives accurate information about the frequency response $P(j\omega)$ of the process at the crossing point with the negative real axis, which is the most important part for stability of the feedback control system. On the other hand, this test gives no information at the other frequencies.

The ultimate gain and the ultimate period are characteristic quantities of the process that give the condition of the stability limit. In most cases, they are directly used to determine the controller parameters (see Section 5.3). But they can also be used to estimate the parameters of the transfer function model of the process. For the model (5) of an integrating process, its parameters are given by the formulae:

$$L = \frac{T_u}{4}, \quad R = \frac{2\pi}{K_{Pu} T_u} \tag{11}$$

which are obtained from the Nyquist criterion. To determine the parameters of the model (4) and of the sophisticated models explained later, the number of data (i.e. two) obtained from the standard ultimate sensitivity test is too few, and some more data are needed. Such data, for instance, can be obtained by a modified ultimate sensitivity test in which the compensator *C(s)* is set in the *I* mode with $T_I = 1$. By the modified test, two more data are obtained and it becomes possible to estimate the parameters of the model (4), and of the models given by the first equations of (16) and (17) (see Section 6.3).

In step response methods, the "step response test" (i.e. to record the reaction curve) is carried out for the actual process first. Here, attention should be drawn to the point that the process must be operated under the open-loop condition in this test, so that the step response of the process without feedback is recorded. Then the parameters of a transfer function model are determined, so that its step response best fits that of the actual process. The simplest way is to assume model (4) or (5) and to determine the parameters graphically as illustrated in Figures 2 or 3. In the case of a process with self-regulation, the stationary gain *K* is determined from the reaction curve at large *t*. Then the tangent is drawn at the point of inflection, and *T* and *L* of model (4) are determined as illustrated in Figure 2. In the case of an integrating process, the tangent to the reaction curve at large *t* is drawn, and *R* and *L* of model (5) are determined as in Figure 3. As stated before, the parameters *K* and *T* of the second expression of (5) are redundant by one, so one of them can be chosen at the engineer's convenience. An alternative for this choice is to select *T* so that the reaction curve exhibits significant change during the initial time interval of the length *T*. While the above method has traditionally been used for the design of PID control systems, several modern alternatives can be used for the determination of process model parameters. The most popular is the "least squared-error method," which minimizes the integrated squared error between the responses of the model and the actual process.

5.3. Tuning of PID Parameters

"Tuning" is the engineering work to adjust the parameters of the controller so that the control system exhibits desired property. Two tuning methods were proposed by Ziegler and Nichols in 1942 and have been widely utilized either in the original form or in modified forms. One of them, referred to as Ziegler–Nichols' ultimate sensitivity method, is to determine the parameters as given in Table 2 using the data K_{Pu} and T_u obtained from the ultimate sensitivity test. The other, referred to as Ziegler–Nichols' step response method, is to assume the model (5) and to determine the parameters of the PID controller as given in Table 3 using the parameters *R* and *L* of (5) which are determined from the step response test as in Figure 3.

Action mode	K_P	T_I	T_D
P	$0.5K_{Pu}$		

PI	$0.45K_{Pu}$	$0.833\,T_u$	
PID	$0.6K_{Pu}$	$0.5\,T_u$	$0.125\,T_u$

Table 2. Ziegler-Nichols' ultimate sensitivity method

Action mode	K_P	T_I	T_D
P	$1/RL$		
PI	$0.9/RL$	**3.33L**	
PID	$1.2/RL$	**2L**	*0.5L*

Table 3. Ziegler-Nichols' step response method ($RL \neq 0$)

It must be noted that the Ziegler–Nichols' step response method can be also used for a process with self-regulation, by considering a virtual integrating process that approximates the actual process for small *t*. To be exact, suppose the case where the actual process is given by Eq. (4). Construct the reaction curve of the virtual integrating process by extending the actual reaction curve from the point of inflection with a straight line of the gradient *K*/*T*. (Note, as illustrated in Figure 2, that this straight line is the tangent to the actual reaction curve at the point of inflection.) Then determine the PID parameters by applying the formulae of Table 3 to the virtual integrating process. This method can be stated, in other words, as follows: "Equate *R* of Table 3 to *K*/*T* with *K* and *T* being the parameters of (4), and apply the formulae of Table 3." This way of application is utilized for the purpose of starting feedback control from the very first start-up of the process. Frequency-domain stability analysis tells that the above way of applying the Ziegler–Nichols' step response method to processes with self-regulation tends to set the parameters on the safe side, in the sense that the actual gain and phase margins become larger than the values expected in the case of integrating processes.

Response to be optimized	**Overshoot**	**Action mode**	K_P	T_I	T_D
Set-point response	Zero	P	$0.3T/KL$		
		PI	$0.35T/KL$	1.17T	
		PID	$0.6T/KL$	T	0.5L
	20%	P	$0.7T/KL$		
		PI	$0.6T/KL$	T	

		PID	$0.95T/KL$	1.36T	0.47L
Disturbance response	Zero	P	$0.3T/KL$		
		PI	$0.6T/KL$	***4L***	
		PID	$0.95T/KL$	2.38L	0.42L
	20%	P	$0.7T/KL$		
		PI	$0.7T/KL$	2.33L	
		PID	$1.2T/KL$	***2L***	0.417L

Table 4. Chien–Hrones–Reswick's tuning method
K, *T* and *L* are the parameters of model (4)
$0.05 \le L/T \le 1.0$

Several more tuning methods were proposed, and have been used, after Ziegler and Nichols. One of the most popular among them is Chien–Hrones–Reswick's tuning method. They assumed the model (4) and derived the tuning rules given in Table 4 via analogue computer simulation. The formulae of Table 4 are applicable in a limited range of the value of *L*/*T*, as is understandable since they are empirical formulae derived from simulation results. It is not easy to indicate sharply in which range they are applicable, but experience shows that $0.05 \le L/T \le 1.0$ is the safe zone. Another distinctive feature of Chien–Hrones–Reswick's tuning method is that four tables are given, depending upon which response (i.e. set-point response or disturbance response) is to be optimized, and whether an overshoot is allowed or not. Consequently, in applying this method engineers must decide which table to use. Specifically, choice about which response to optimize is difficult, because optimizing the disturbance response usually brings about a very poor set-point response and vice versa. This situation is illustrated in Figure 6, in which the dotted lines are the responses of the control system optimally tuned for the set-point input and the continuous lines optimally tuned for the disturbance input. The process is assumed to be:

$$P(s) = \frac{1}{1+s} e^{-0.2s} \tag{12}$$

Note that the disturbance response of the dotted line is markedly slow compared with that of the real line. and that on the other hand, the overshoot value of the set-point response of the real line is more than 50%. This problem can be solved by employing the two-degree-of-freedom structure (see Section 6.2).

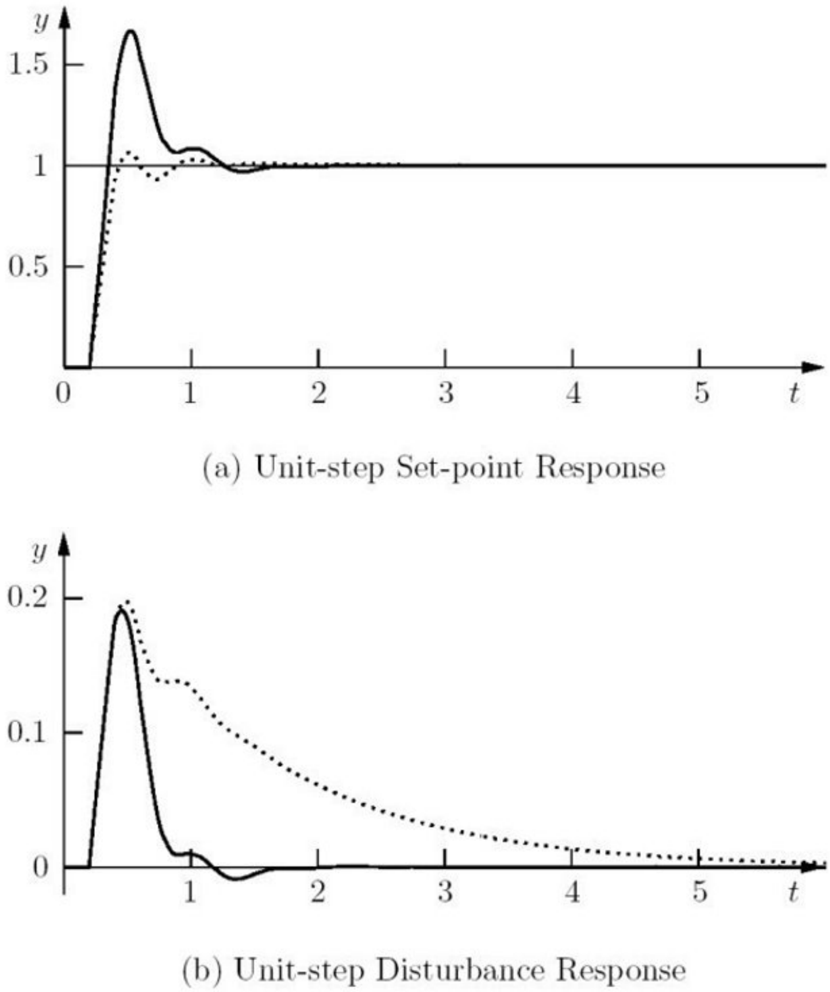

Figure 6. Response of the conventional PID control system

The above three methods determine the PID parameters using empirical formulae that are constructed based on the unit-step set-point response and/or the unit-step disturbance response. The three methods, as well as several other tuning methods developed on the same principle, are often referred to as “classical” tuning methods. There are many other tuning methods that determine the PID parameters based on different principles. Some of them will be introduced briefly later (see Section 6.4).

6. Advanced Topics

6.1. Windup of the Integral Element and Anti-Windup Mechanism

For various physical reasons, the manipulated variable is subject to “saturation”: that is, it can only take values within certain limits: $u_{\min} \le u(t) \le u_{max}$. For instance, in the case of a heating tank, the flow of fuel gas can only take values between 0 and the maximum value determined by the section area of the supplying tube and the pressure difference. When the controller gives a command beyond the limits, the feedback mechanism stops functioning normally, and the error is kept larger than the value expected in normal operation. As a result, the output of the *I* element becomes very large, provided that it is in action. This phenomenon is called “windup” of the *I* element.

Once windup occurs, it takes a long time to bring back the *I* element to the normal state and, as a result, unfavorable results are observed. For instance, if a large step input, which forces the controller output to go beyond the limits, is added as the set-point value, an excessively large overshoot of the process variable is brought about because of windup. “Anti-windup” is the mechanism to avoid windup of the *I* element. Many researches on anti-windup mechanisms have been reported. In the following, a simple but often effective strategy is introduced.

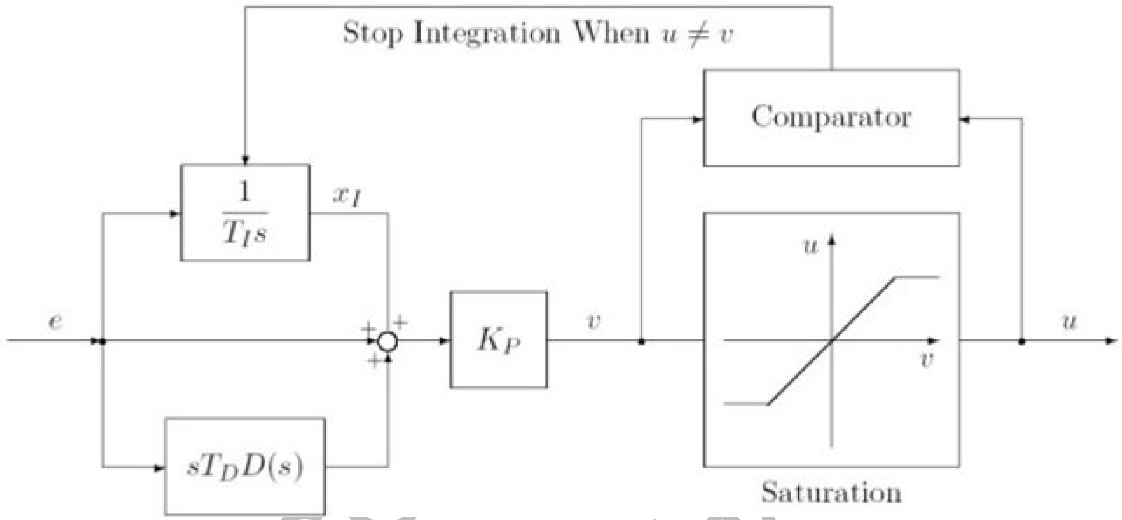

Figure 7. Anti-windup mechanism

The modern controller is mostly implemented using digital components, and the anti-windup mechanism is usually constructed in such a framework. Unfortunately the details of the digital implementation are not included in this article, so the explanation in the following will become rather conceptual. When the PID controller is implemented digitally, the three functional elements (i.e. the *I*, the P and the D elements) are often implemented as shown in the left half of Figure 7, so that the integration can be treated separately. The saturation element located in the right half of Figure 7 is very often also equipped, in order to avoid adding an excessive input signal to the actuator. The anti-windup mechanism can be constructed making use of those structures. The input v and the output u of the saturation element are compared. If the two signals are different, i.e. if $u \neq v$, integration is stopped (which means the output x_I of the *I* element is kept the same) in order to prevent the windup phenomenon from occurring. Even when the saturation element is not used in the digital controller, the same mechanism(in principle) can be constructed by measuring the actual output of the actuator. Several modifications of the above anti-windup mechanism are available, as well as other mechanisms based on different principles.

6.2. Two-Degree-of-Freedom PID Controllers

The closed-loop transfer functions $G_{yr1}(s)$ and $G_{yd1}(s)$ of the control system of Figure 1, respectively from *r* to *y* and from *d* to *y*, are given by

$$G_{yr1}(s) = \frac{P(s)C(s)}{1+P(s)C(s)}, \quad G_{yd1}(s) = \frac{P(s)}{1+P(s)C(s)} \tag{13}$$

These equations show that the two closed-loop transfer functions are related by $P(s)\{1-G_{yr1}(s)\} = G_{yd1}(s)$ and cannot be changed separately. This fact causes the problem of tuning as illustrated in Figure 6: that is, optimizing the disturbance response brings about a very poor set-point response and vice versa (see Section 5.3).

The above problem is solved by implementing another compensator $C_f(s)$ in addition to *C(s)* as shown in Figure 8. The closed-loop transfer functions $G_{yr2}(s)$ and $G_{yd2}(s)$ of this control system are

$$G_{yr2}(s) = \frac{P(s)\{C(s)+C_f(s)\}}{1+P(s)C(s)}, \quad G_{yd2}(s) = \frac{P(s)}{1+P(s)C(s)} \tag{14}$$

The existence of $C_f(s)$ in the numerator of the first equation indicates that the two closed-loop transfer functions can be changed separately by adjusting the two compensators *C(s)* and $C_f(s)$. The controller with the structure of Figure 8 is referred to as a "two-degree-of-freedom controller," *C(s)* of Figure 8 as its "main compensator" and $C_f(s)$ as its "feedforward compensator." A "two-degree-of-freedom PID controller" is the controller of this type, with *C(s)* being given by (1) and $C_f(s)$ given by

$$C_f(s) = -K_P\,\{\alpha + \beta\,T_D D(s)\} \tag{15}$$

The parameters K_P, T_I and T_D are referred to as "basic PID parameters" and α and β as "two-degree-of-freedom parameters." If the two-degree-of-freedom PID controller is applied to the process of Eq. (12), the responses shown in Figure 9 are attained, where K_P, T_I and T_D are set to the values obtained from the "optimize the disturbance response, 20% overshoot" formula of Table 4, and α and β are, respectively, set to: $\alpha = 0.63$ and $\beta = 0.70$. Note that the disturbance response is exactly the same with the real line of Figure 5, and that the set-point response exhibits little overshoot, whereas the settling time is approximately the same with that of the dotted line of Figure 5.

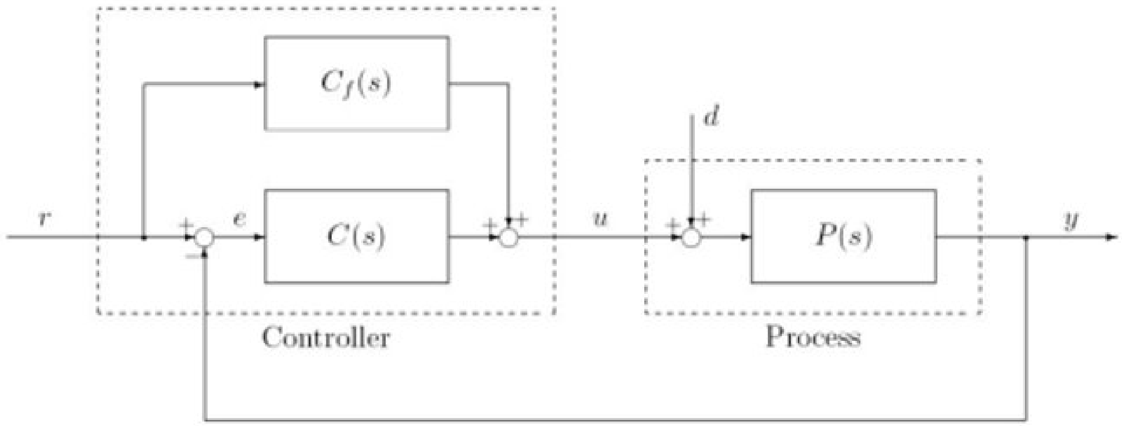

Figure 8. Two-degree-of-freedom control system

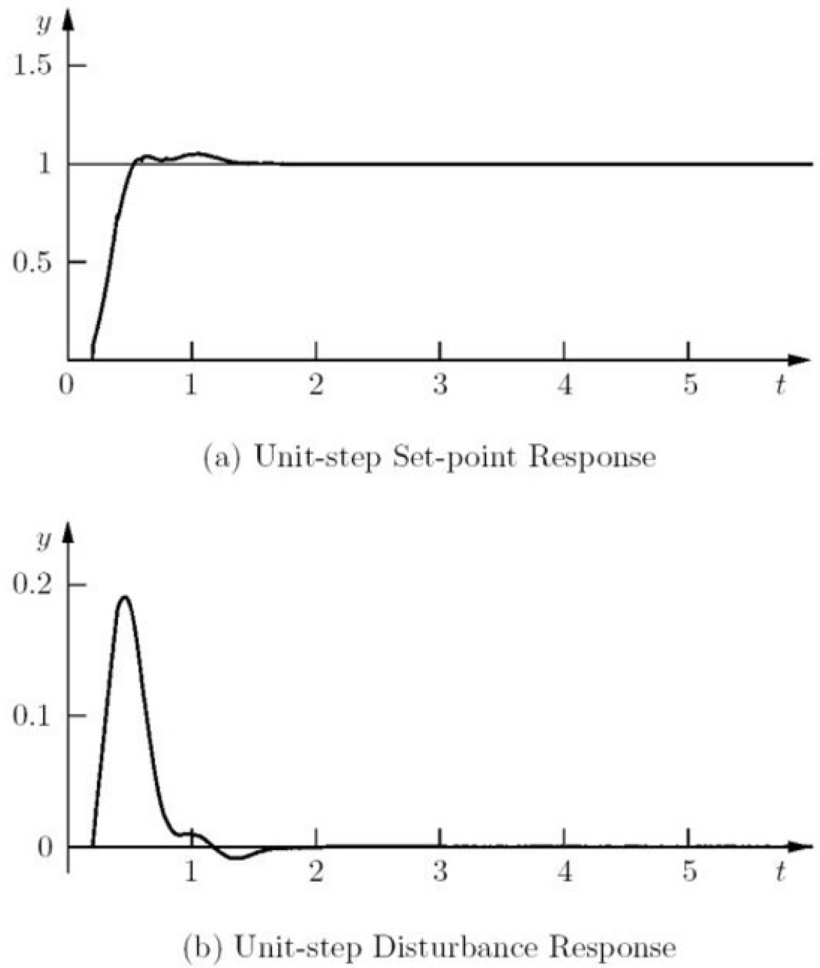

Figure 9. Responses of the two-degree-of-freedom PID control system

6.3. Sophisticated Models

As explained before, the transfer functions (4) and (5) have been and are used as representative process models. But they are not necessarily precise enough for the purpose of adjusting the controller accurately, making full use of the capability of modern computers and sensing systems, so more sophisticated models are expected. The following are examples of such models, for a process with self-regulation and for an integrating process:

$$P(s) = \frac{K}{(1+Ts)^n} e^{-Ls} \quad \text{and} \quad P(s) = \frac{K}{Ts(1+Ts)^n} e^{-Ls} \tag{16}$$

where *K*, *T*, and *L* are positive parameters and *n* is a positive integer parameter. Those parameters are to be determined so that the response of the model best fits that of the actual process. However, it is not easy to determine the integer parameter *n* based only on the measured date, and some theoretical or empirical knowledge is needed for its determination. The next transfer functions are also used specifically when the process exhibits an oscillatory response:

$$P(s) = \frac{K}{1+2\zeta\, Ts + T^2 s^2} e^{-Ls} \quad \text{and} \quad P(s) = \frac{K}{Ts(1+2\zeta\, Ts + T^2 s^2)} e^{-Ls} \tag{17}$$

where *K*, ζ, *T* and *L* are positive parameters with $0 < \zeta < 1$ (see *Elements of Control Systems*).

6.4. Other Tuning Methods for PID Parameters

The classical tuning methods explained in Section 5.3 have the following features:

- The process is assumed, implicitly (in the case of Ziegler–Nichols' ultimate sensitivity method) or explicitly (in the case of Ziegler–Nichols' step response method and Chien–Hrones–Reswick's method), to be modeled by the simple transfer function (4) or (5).
- The optimal values of the PID parameters are given by formulae of the process parameters that are determined directly and uniquely from experimental data.

The first feature is a weakness of these classical methods, in the sense that the applicable processes are limited, or in other words that the claimed "optimal" values are not necessarily, and are sometimes fairly far from, the true optimal in practical situations where the transfer function (4) or (5) is nothing but an approximation of the real process characteristics. Specifically, the problem is serious when the pure delay *L* of the process is very short or very long, where "very short" and "very long" roughly means outside the range $0.05 \le L/T \le 1.0$. (Note that this is the interval suggested as the effective range for Chien–Hrones–Reswick's method as noted in Table 4.) The second feature is a strength of the methods, in the sense that they can be applied straightforwardly. On the other hand, it can be interpreted as a weakness in the sense that there is no room to improve the results by making use of more detailed information about the process which is obtainable from theoretical study and accurate measurement.

Many attempts have been made to make up for these weaknesses of the classical methods. First, tables and/or formulae giving the optimal PID parameters have been made for more detailed process models such as (16) and (17). Second, many theoretical considerations have been used to develop sophisticated methods that use, as the basis of tuning, the shape of the frequency response of the return ratio, poles (and zeros) of the closed-loop transfer function $G_{yr}(s)$, time-domain performance indices such as (9) and (10), or frequency-domain performance indices. In the following, the method referred to as the "modulus optimum" method will be introduced. This offers a solution for the case when the pure delay is practically zero and the process can be modeled by a rational transfer function. A representative example of such processes is the electrical drive. whose simplest model is:

$$P_{ED}(s) = \frac{K_{ED}}{T_{ED}s(1+T_{ED}s)} \tag{18}$$

The principle of the modulus optimum method is to tune the parameters of the controller so that the closed-loop transfer function $G_{yr}(s)$ from the setpoint value *r* to the process variable *y* satisfies the next conditions:

(a) $G_{yr}(0) = 1$

(b) $\dfrac{d^m |G_{yr}(j\omega)|}{d\omega^m} = 0$ at $\omega = 0$ for as many positive integers m (starting from 1) as possible.

The transfer functions satisfying the above conditions are called "modulus optimal." The second order and the third order modulus optimum transfer functions are given by:

$$G_{MO}(s) = \frac{1}{1+\sqrt{2}T_{MO}s + T_{MO}{}^2 s^2} \tag{19}$$

$$G_{MO}(s) = \frac{1}{(1+T_{MO}s)(1+T_{MO}s + T_{MO}{}^2 s^2)} \tag{20}$$

where T_{MO} is the time constant of the feedback control system designed by this tuning method.

In the practice of the modulus optimum method, the structure of the controller and the form of the compensator(s) *C(s)* (and $C_f(s)$ in the case of the two-degree-of-freedom controller) are chosen first. Then the order of the modulus optimum transfer function $G_{MO}(s)$ is determined so that it matches with the closed-loop transfer function $G_{yr}(s)$ derived from the given process and the chosen controller. Lastly the parameters of the controller and the parameters of $G_{MO}(s)$ are determined from the equation obtained by equating $G_{yr}(s)$ to $G_{MO}(s)$:

$$G_{yr}(s) = G_{MO}(s) \tag{21}$$

In the case of the one-degree-of-freedom control system (Figure 1), $G_{yr}(s)$ on the left-hand side of (21) must be equated to $G_{yr1}(s)$ given by (13), and in the case of the two-degree-of-freedom control system (Figure 8) to $G_{yr2}(s)$ given by (14). It sometimes happens that (21) has no solution for a fixed type of controller. In such a case, change of the type of controller must be considered. The situation will be explained in the following within the framework of PID controllers, using the example of the electrical drive given by (18).

First, let us study the case where the one-degree-of-freedom PID controller is chosen and *C(s)* is set in the P-mode. For this choice, the order of $G_{yr1}(s)$ becomes two, so $G_{MO}(s)$ of (19) must be used as the right-hand side of (21). In this case, the unknowns of the equation are K_P of *C(s)* and T_{MO} of $G_{MO}(s)$. (21) has the solution

$$K_P = \frac{1}{2K_{ED}}, \quad T_{MO} = \sqrt{2}T_{ED} \tag{22}$$

The feedback control system designed as above exhibits the zero steady-state error to the step set-point input, but non-zero steady-state errors to the step disturbance input and to the ramp set-point input. If the aim is to make the latter two zero too, the PID controller must be set in the PI- or PID-mode. The case of the PI-mode is studied in the following.

First, let us see if (21) has a solution for the one-degree-of-freedom PID control system. When *C(s)* is set in the PI-mode, the order of $G_{yr1}(s)$ becomes three and $G_{MO}(s)$ of (20) must be used as the right-hand side of (21). Then, by simple calculation, it is obtained that $G_{yr1}(s)$ always has a finite zero (i.e. the factor $T_I s + 1$ appears in its numerator), so it cannot be equal to $G_{MO}(s)$ which does not have a finite zero. In other words, (21) does not have a solution for this choice. This means that a change in the class of controller must be considered. Since the use of the PI action is the requirement from the design purpose, it is natural to try the two-degree-of-freedom controller. The transfer function *C(s)* and $C_f(s)$ of the two-degree-of-freedom PID controller in the PI action mode are, respectively, given by (1) with $T_D = 0$ and (15) with $\beta = 0$. For this controller, the order of $G_{yr2}(s)$ becomes three too, and $G_{MO}(s)$ of (20) is to be used as the right-hand side of (21). The unknowns of (21) in this case are α, K_P, T_I and T_{MO}. Calculation tells that (21) is satisfied for

$$\alpha = 1, \quad K_p = \frac{1}{2K_{ED}}, \quad T_I = 4T_{ED}, \quad T_{MO} = 2T_{ED} \tag{23}$$

For the above values, the return ratio has the symmetrical Bode diagram, and for that reason, the above tuning is called “symmetrical optimum.”

Glossary

Action mode:	The combination of the three (proportional, integral, and derivative) elements that are kept in action when the PID controller is applied.
Anti-windup:	A mechanism to prevent windup of the integral element.
PID:	Proportional, integral, and derivative.
PID control:	The feedback control method that uses the PID controller as its main tool.
PID controller:	A controller consisting of the proportional, integral and derivative elements.
Process:	A system that produces certain product(s) in their widest sense.
Process model:	A mathematical description that expresses the characteristics of a process.
Two – degree – of - freedom PID controller:	The modern type of PID controller, which can adjust two closed-loop transfer functions separately.
Tuning:	The engineering work to adjust the parameters of a PID controller so that the control system exhibits a desired property.
Windup:	The phenomenon that the output of the integral element becomes excessively large because of saturation of the manipulated variable.

Bibliography

Åström K.J. and Hägglund T. (1995). *PID Controllers: Theory, Design and Tuning*, 343pp. Instrument Society of America [This book provides detailed explanation on the most topics treated in the text except the classical tuning method of the two-degree-of-freedom PID controller.]

Taguchi H. and Araki M (2000). Two-degree-of-freedom PID controllers. *Proceedings of the IFAC Workshop on Digital Control: Past, Present and Future of PID Control*, pp. 91–96. Elsevier. [This paper provides a brief survey of the properties of the two-degree-of-freedom PID controller as well as its tuning method.]

Bennett S. (1993), *A History of Control Engineering 1930–1955*, pp.28–69. Peter Peregrinus. [This book contains a short history of early PID controllers.]

Biographical Sketch

Mituhiko Araki received his B.E., M.E., and Ph.D. degrees, all in electronic engineering, from Kyoto University, Japan, in 1966, 1968, and 1971. Since 1971 he has been with the Department of Electrical Engineering, Kyoto University, where he is currently a Professor. He visited Imperial College, London from 1973 to 1975, Santa Clara University, California in 1979, and Waterloo University, Ontario in 1982.

His research interests are in systems and control theory and its application to industrial and medical problems. As theoretical topics, he has been engaged in research on stability of composite systems, M-matrices, the Nyquist array method for design of multivariable controllers, state predictive controllers for plants with pure delays, multirate digital control systems, two-degree-of-freedom optimal controllers, and frequency responses of sampled-date systems. As industrial applications, he proposed two-degree-of-freedom PID controllers, studied their optimal tuning, and applied them to temperature controllers and servo-pulser controllers. In addition to the above, he has been engaged in researches on control of synchronous generators, control and scheduling of steel producing systems, and scheduling of elevators.
Since 1991 he has been engaged in co-operative topics with medical doctors such as the control of blood pressure of patients under surgery, the control of blood sugar after surgery, the control of intraocular pressure during eye surgery, and clinical stage classification. Currently he is placing more time and efforts on these medical applications, while he is still much interested in industrial applications of control theory and scheduling techniques.
He is a member of the IEEE and of many Japanese academic societies. He was a council member of IFAC, an associate editor of the *IEEE Transactions on Automatic Control*, the editor-in-chief of *Systems, Control and Information* (a Japanese journal), and the editor-in-chief of the *Transactions of the Society of Instrument and Control Engineers in Japan*. He is currently the editor for control system applications of *Automatica*.