Chapter 1

On the notion of the "size" of a system and its applications

1.1 Introduction

The need to define a notion of the size of a system arises naturally in several control-related problems where one is interested to describe fundamental system properties using simple models and/or a small number of parameters. Typical applications of this notion include the modeling/approximation of LTI systems and the use of feedback to attenuate the effects of external disturbances and modeling errors on the plant output.

For example, in an approximation (or model order reduction) problem, given a transfer function $G(s)$, it is desired to find a transfer function $\hat{G}(s)$ of lower order, such that the difference $G(s) - \hat{G}(s)$ is small in some sense. In other words, if u is any "possible" input signal we would like to find $\hat{G}(s)$ such that the error signal

$$e = y - \hat{y} = G(s)[u] - \hat{G}(s)[u]$$

is small. Similarly, in a disturbance attenuation problem the plant output can be expressed as

$$y = G_r(s)[r] + G_d(s)[d]$$

where r, d are the reference input and disturbance respectively, $G_r(s)$ is the transfer function from r to y and $G_d(s)$ is the transfer function from d to y. Since in a closed loop system both $G_r(s)$ and $G_d(s)$ depend on the designer-selected controller transfer function, it is desirable to perform such a design as to make the contribution of the term $G_d(s)[d]$ as small as possible.

The objective of this note is to give a formulation of these problems and provide the basic insight and guidelines for their solution. Some proofs of statements are also included for reasons of completeness, although their knowledge is not required at this stage.

1.2 Generalities

We begin our discussion by establishing certain fundamental properties of systems associated with the notion of their size. In order to clarify the concepts behind this development, let us first consider the case of a real function, say f, mapping points of the real line into points of the real line (see Fig. 1.1) i.e.,

$$f : t \mapsto f(t) \ , \quad \text{or} \quad f : \mathbf{R} \mapsto \mathbf{R}$$

To such a function we can associate a number, say M_f, that is the maximum value of the function. A technical problem arises at this point since, in general, the maximum of a function may not exist. For example, the function $1 - e^{-t}$, $t \in \mathbf{R}$ does not have a maximum. To deal with such cases, we define the

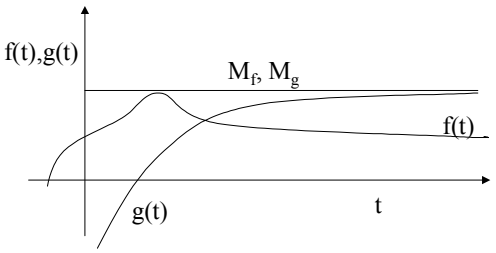

Figure 1.1: The functions f and g illustrating the maximum and supremum of a function.

supremum of a function f as the "least upper bound" of f, i.e., the smallest number M_f such that the following inequality holds

$$f(t) \leq M_f \ , \quad \forall t \in \mathbf{R}$$

We denote the least upper bound of f by $\sup_{t\in\mathbf{R}} f(t)$, or, simply, $\sup_t f(t)$ when there is no ambiguity on the definition of the domain of f. Fig. 1.1 illustrates the difference between the maximum and the supremum of a function.

1.2.1 Some Properties and Definitions

- When there exists t_* in the domain of f such that $f(t_*) = \sup_t f(t)$, then $\sup_t f(t) = \max_t f(t) = f(t_*)$. Notice that t_* may not be unique (e.g., consider the function $f(t) = \sin t$).
- For any real number $m < \sup_t f(t)$ there exists t_0 such that $|f(t_0)| > m$.
- A function f is called bounded if $\sup_t |f(t)|$ is finite.
- All properties and definition can be applied to functions whose domain is a subset of $\mathbf{R}$.

1.2.1 Example: Let $g(t) = 1 - e^{-t}$, $t \geq 0$. We claim that $\sup_{t\geq 0} |g(t)| = 1$. Indeed $1 > |g(t)|, \ \forall t \geq 0$. furthermore, let $m < 1$; if $m < 0$ then $|g(t)| > m$ for any $t \geq 0$. If $1 > m \geq 0$, let $t_0 = \ln(\frac{1}{1-m})$ which is finite since $1 > m$. It is straightforward to verify that for any $t > t_0$, $g(t) > m$ which establishes the above claim. $\bigtriangledown\bigtriangledown$

Next, consider the function f shown in Fig. 1.2, which, for simplicity, it is assumed to be such that $f(0) = 0$. The (linear) growth of the function f can be characterized by a single number, termed as the gain of f and denoted by $\gamma[f]$, and defined as follows

$$\gamma[f] = \sup_{t\in\mathbf{R}\neq 0} \left(\frac{|f(t)|}{|t|} \right) \tag{1.1}$$

In other words $\gamma[f]$ is the largest slope of lines connecting a point in the graph of the function to the origin i.e., it is the least upper bound of the function $f(t)/t$ ($t \neq 0$). Notice that not all functions have finite gains, as demonstrated by the following example:

1.2.2 Examples:

a. Let $f(t) = t^2$. Then $\gamma[f]$ is not defined since $f(t)/t = t$ is not bounded.

b. Let $f(t) = \sqrt{t}$. Then $\gamma[f]$ is not defined since $f(t)/t = 1/\sqrt{t}$ is not bounded.

c. Let $f(t) = t\sin(t)$. Then $\gamma[f] = \sup_{t\neq 0}(|f(t)|/|t|) = \sup_{t\neq 0}(|\sin(t)|) = 1$. $\bigtriangledown\bigtriangledown$

Notice that it is possible to modify the definition of a gain to apply to functions which have non-zero value at $t = 0$; this generalization, however, is beyond the scope of this note and is omitted.

Similar ideas can be extended to the case of systems or operators[1] mapping functions to functions. For example, consider a system described by $\hat{y}(s) = G(s)\hat{u}(s)$; the system can be viewed as an operator mapping

[1] e.g., the integral or the derivative are familiar operators.

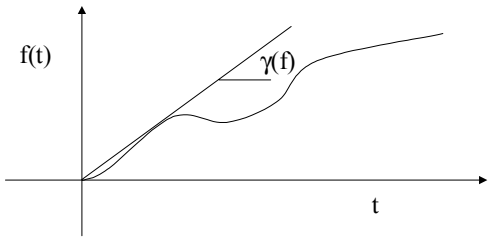

Figure 1.2: The "gain" of the function f.

the input $\hat{u}(s)$ to the output $\hat{y}(s)$. Taking inverse Laplace transforms we also have that the input time-function $u(t)$ is mapped to the output time-function $y(t)$. This situation is analogous to the previous case where a function f was mapping time points t to values $f(t)$. Thus, it is conceptually straightforward to generalize (1.1) in order to define the gain of a system. The only essential difference and difficulty of this generalization is associated with the appearance of whole functions in the numerator and denominator of (1.1), instead of numbers, requiring a generalization of the concept of the absolute value or "size" to the case of functions.

Before we proceed with this development, we need to introduce some notation and terminology.

1.2.2 Notation

- The input and output of a system are denoted by (usually) lower-case letters, e.g., u, y. Depending on the context, these can be considered as functions of time or frequency, as related by the Laplace transform. For example, we use u to denote the time-function $t \mapsto u(t)$ and $\hat{u}$ to denote its Laplace transform $s \mapsto \hat{u}(s)$ (assuming that it exists).
- Systems are denoted by upper-case letters. For example, we use G to denote a system mapping inputs u to outputs y. We then write $G : u \mapsto y$ and $y = G[u]$. A precise definition would also require a suitable definition of the function space where u, y take values, but we avoid this issue for simplicity. Also for simplicity, we use $G(s)$ to denote the transfer function associated with the system G (here all systems are LTI). Thus, depending on the context, we may use the following notation to define the relation between the input and the output of a system:

$$\begin{aligned} y &= G[u] \\ y(t) &= G[u](t) \\ \hat{y}(s) &= G(s)\hat{u}(s) \end{aligned}$$

1.2.3 Gain of a System

The gain of a system can be defined in an analogous way as the gain of a function. For example, letting G denote a system $u \mapsto y = G[u]$, its gain can be defined as

$$\gamma[G] \triangleq \sup_{u \neq 0} \left(\frac{size(y)}{size(u)} \right)$$

provided that the supremum in the right hand side is finite. Here, $size(u)$ denotes any suitable way to define the size of a function. For the latter, there are several alternatives, all of which aim to generalize the concept of the absolute value of a number to that of a function.

One possible approach is to define the size of a function as the supremum of its absolute value. This is commonly denoted by the symbol $\| \cdot \|_\infty$ (called the "infinity-norm") i.e.,

$$\|f\|_\infty \triangleq \sup_t |f(t)|$$

Observe that $\|\cdot\|_\infty$ enjoys all the fundamental properties of absolute values, namely

1. $\|f\|_\infty \geq 0$; $\|f\|_\infty = 0$ iff $f \equiv 0$.
2. $\|af\|_\infty = |a|\,\|f\|_\infty$; $\forall a \in \mathbf{R}$.
3. $\|f+g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$ (Triangle inequality).

Another useful definition of the size of the function is in terms of the square-root of its energy, i.e.,

$$\|f\|_2 \triangleq \left(\int_{-\infty}^{\infty} |f(t)|^2\,dt\right)^{1/2}$$

In this case, the size $\|\cdot\|_2$ (called the "two-norm") also satisfies the same three properties as $\|\cdot\|_\infty$.

In this framework, we may now define the gain of a system, induced by $\|\cdot\|_\infty$, as:

$$\gamma_\infty[G] \triangleq \sup_{\|u\|_\infty \neq 0} \left(\frac{\|y\|_\infty}{\|u\|_\infty}\right)$$

Similarly, the gain of a transfer function, induced by $\|\cdot\|_2$, is defined as

$$\gamma_2[G] \triangleq \sup_{\|u\|_2 \neq 0} \left(\frac{\|y\|_2}{\|u\|_2}\right)$$

Notice that, since G is a linear system,

$$\gamma_\infty[G] \triangleq \sup_{\|u\|_\infty = 1} \|y\|_\infty$$

In other words, $\gamma_\infty[G]$ is the least upper bound of the output time-function for any input function u, bounded by 1 and achieving the value 1 for some t (including ∞).[2] A similar interpretation can be given to the "two-gain" of a system.

The above formulation establishes the theoretical framework of our study. It is not very useful, however, for the computation and practical application of these concepts. For the latter, we need to express the gains of systems in terms of easier-to-compute quantities.

1.2.3 Proposition: *Let $G(s)$ be the transfer function of a causal, BIBO stable, finite dimensional LTI system G*

$$\dot{x} = Ax + bu\ ;\quad y = cx$$

with impulse response $g(t) = ce^{At}b$ i.e., $G(s) = \mathcal{L}\{g(t)\}$. Then, $\gamma_\infty[G]$ exists, is finite and is given by

$$\gamma_\infty[G] = \int_0^\infty |g(t)|\,dt$$

$\triangledown\triangledown$

Proof: Let u be any arbitrary input such that $\|u\|_\infty = 1$. Then,

$$y(t) = \int_{-\infty}^{t} g(t-\tau)u(\tau)\,d\tau$$

Taking absolute values,

$$\begin{aligned} |y(t)| &\leq \int_{-\infty}^{t} |g(t-\tau)||u(\tau)|\,d\tau \\ &\leq \int_{-\infty}^{t} |g(t-\tau)|\,d\tau \\ &\leq \int_0^\infty |g(t)|\,dt \end{aligned}$$

[2] In this development, it is assumed that initial conditions are zero, although this is not necessary.

Hence, $\gamma_\infty[G] \leq \int_{-\infty}^{\infty} |g(t)|\, dt$. To complete the proof we must show that we can construct an input such that $\|u\| = 1$ and $\|y\| = \gamma_\infty[G]$. For this, let $T > 0$ be a fixed number and define $u_T(t) = sign[g(T-t)],\ t \geq 0$. (Notice that $g(t) = 0$ for $t < 0$). Obviously, $\|u_T\| = 1$. Evaluating the output at T,

$$y(T) = \int_0^T |g(T-\tau)|\, d\tau = \int_0^T |g(t)|\, dt$$

which by the definition of the gain of G must satisfy,

$$\int_0^T |g(t)|\, dt \leq \gamma_\infty[G]$$

Letting $T \to \infty$ the result follows. □□

1.2.4 Example: Let $G(s) = \frac{1}{s+2}$ corresponding to the impulse response $g(t) = e^{-2t};\ t \geq 0$. Then $\gamma_\infty[G] = 1/2$. Hence, for any bounded input such that $|u(t)| \leq M$, $|y(t)| \leq M/2$. ▽▽

Although the gain γ_∞ is an intuitively appealing way of characterizing the size of a system, it leads to excessive computational requirements when used for analysis and design purposes. Furthermore, it is hard to relate to other useful properties of signals, such as, energy content, power spectrum etc. For this reason, the γ_2-gain of a system has been a more popular way to measure the size of a system. Using Parseval's theorem, it can be shown that:

1.2.5 Proposition: *Let $G(s)$ be the transfer function of a causal, LTI system G and suppose that $G(s)$ is analytic and bounded in the right-half complex plane. Then, $\gamma_2[G]$ exists, is finite and is given by*

$$\gamma_2[G] = \sup_{Re[s] \geq 0} |G(s)| = \sup_{w \in \mathbf{R}} |G(jw)|$$

▽▽

Notice that the analyticity of $G(s)$ [3] implies that $G(s)$ has no poles in the right-half complex plane i.e., it represents a BIBO stable system. This definition of the system gain is indeed quite convenient and useful and is used throughout the rest of our discussion. Observe that it only takes a Bode plot to determine the "size" of a system since the γ_2-gain is simply the peak in the magnitude plot of the associated frequency response.

1.2.6 Examples:

a. Let $G(s) = \frac{s+1}{s+2}e^{-0.5s}$. Then $\gamma_2[G] = 1$.

b. Let $G(s) = \frac{s+2}{s+1}$. Then $\gamma_2[G] = 2$.

c. Let $G(s) = \frac{1}{s+10}$. Then $\gamma_2[G] = 0.1$.

d. Let $G(s) = \frac{w_n^2}{s^2+2\zeta w_n s+w_n^2}$, $\zeta < 1/\sqrt{2}$. Then $\gamma_2[G] = M_{pw} = \frac{1}{2\zeta\sqrt{1-\zeta^2}}$. ▽▽

We conclude this section with some useful properties of the γ_2-gain of a system.

1. $\gamma_2[G] \leq \gamma_\infty[G]$.
2. $\gamma_2[G+H] \leq \gamma_2[G] + \gamma_2[H]$.
3. $\gamma_2[aG] = |a|\gamma_2[G]$, $\forall a \in \mathbf{R}$.
4. $\gamma_2[GH] \leq \gamma_2[G]\gamma_2[H]$.
5. If $y = G[u] + H[r]$ then
$$\|y\|_2 \leq \gamma_2[G]\|u\|_2 + \gamma_2[H]\|r\|_2$$

[3] For the transfer functions considered in this course, analyticity in the right-half plane is guaranteed by the condition that the roots of the denominator are in the open left-half plane.

6. $\int_0^t |y(t)|^2\,dt \leq \gamma_2^2[G]\int_0^t |u(t)|^2\,dt$, where $y(s) = G(s)u(s)$, $G(s)$ is causal and any initial conditions are assumed to be zero.

7. Consider an exponentially stable system $\dot{x} = Ax + bu$, $y = cx + du$ with initial conditions $x(0) = x_0$ and let $G(s) = c(sI - A)^{-1} + d$.[4]

   Then $y(s) = G(s)u(s) + c(sI - A)^{-1}x_0$ and, therefore,

$$\int_0^t |y^2(t)|\,dt \leq \left(\gamma_2[G]\int_0^t |u(t)|^2\,dt + K\right)^2$$

   where K is a finite constant such that $(\int_0^t |ce^{At}x_0|^2\,dt)^{1/2} \leq K$. (Prove this property!)

As a final remark, notice that the previous development has the minor shortcoming that it only admits energy signals. This is of course inconvenient since in applications, the various signals may not be energy signals e.g., a step function. The problem can be circumvented thanks to a causality assumption and the properties 6 and 7. That is, causality allows us to consider truncated signals [5] at any finite time t. Under some mild assumptions, satisfied by all signals met in practice, the truncated part of a signal u is an energy signal. On the other hand, it is worthwhile to mention that for a periodic input u, admitting a Fourier series expansion

$$u(t) = \sum_{n=-\infty}^{\infty} c_n e^{jnw_0 t}$$

the output y is also periodic, with a Fourier series expansion

$$y(t) = \sum_{n=-\infty}^{\infty} G(jnw_0)c_n e^{jnw_0 t}$$

and, by Parseval's theorem, the energy of y over one period is less than or equal to the energy of u over one period multiplied by $\gamma_2^2[G]$.

1.3 Model Order Reduction and Approximations of Transfer Functions

Employing the notion of a gain of a transfer function we may now give a precise formulation and interpretation of the model order reduction problem. This problem can be stated in a various ways e.g.,

- Given a system with transfer function $G(s)$ of order n, find a system with transfer function $\hat{G}(s)$ of order $\hat{n} < n$ such that $\gamma_2[G - \hat{G}]$ is minimum.
- Given a system with transfer function $G(s)$ of order n and a positive number ϵ, find a system with transfer function $\hat{G}(s)$ of minimum order $\hat{n}$ such that $\gamma_2[G - \hat{G}] \leq \epsilon$.
- Given a system with transfer function $G(s)$ of order n, find a system with transfer function $\hat{G}(s)$ of minimum order $\hat{n}$ such that $\gamma_2[G - \hat{G}]$ is small and give an upper bound of $\gamma_2[G - \hat{G}]$.

At this stage, we only consider the last of these problems as the other two require some more involved mathematical tools, extending beyond the scope of the class. A solution of this problem (not necessarily the "best" or the only one) can be obtained in a straightforward and constructive way as follows:

1. Write the partial fraction expansion of $G(s)$

$$G(s) = \sum_i G_i(s)$$

[4] The eigenvalues of A are in the open left-half plane.

[5] The truncation of a signal u at t is defined as the signal $u_t(\tau) = u(\tau)$ for all $\tau \leq t$ and 0 otherwise.

2. Sketch the Bode plots of each individual $G_i(s)$.

3. Select a threshold for the desired maximum error $\gamma_2[G-\hat{G}]$, say ϵ, and group the G_i's in two terms:
   i. $\Delta(s)=\sum_{j=1}^{m} G_j(s)$, the transfer functions satisfying (a) $\gamma_2[G_j]\leq \epsilon/m$ and (b) $G_j(s)$ is BIBO stable;
   ii. $\hat{G}(s)$, the sum of the rest of the G_i's.

4. Sketch the Bode plot of $\Delta(s)$ and find $\gamma_2[\Delta]$. If too small, repeat the previous step increasing m by 1 and transfering the smallest term in $\hat{G}(s)$ to $\Delta(s)$. If too large, repeat the previous step decreasing m by 1 and transfering the largest term in $\Delta(s)$ to $\hat{G}(s)$.

The outcome of this procedure is a transfer function $\hat{G}(s)$ of order $\leq n$ and a transfer function $\Delta(s)$ such that

$$G(s)=\hat{G}(s)+\Delta(s)$$
$$\gamma_2[\Delta]\leq\epsilon$$

Intuitively speaking, the original transfer function is decomposed into two parts, a low order approximation and a small perturbation part. The smallness of the perturbation is simply expressed as the worst-case energy of the error between the actual output $y=G[u]$ and the approximate one $y_{app}=\hat{G}[u]$ for all possible inputs of unit energy i.e.,

$$y-y_{app}=G[u]-\hat{G}[u]=\Delta[u]$$
$$\|y-y_{app}\|_2\leq\gamma_2[\Delta]\ ;\quad \|u\|_2=1$$

Notice that the first iteration of the above method will always produce $\Delta(s)$ such that $\gamma_2[\Delta]\leq\epsilon$ (see property 2 of the γ_2-gain). It is possible however, that the actual $\gamma_2[\Delta]$ is a lot smaller than ϵ and hence, the order of $\hat{G}(s)$ may be further reduced. On the other hand, there is always the possibility that none of the G_i's has a small gain, implying that a reduction of the model order is not possible.

It should also be mentioned that variations of the same idea may be used to further "refine" the approximation $\hat{G}$ or introduce some desirable properties to it. For example, it is often the case that we are interested in approximations that preserve the properties of the original system in some frequency range (typically low frequencies). This problem can be stated as a "weighted model reduction." Without getting too deep in the technical details, we note that having performed the previous model-order reduction, we may like to adjust our approximation $\hat{G}$ so that we obtain a better match at low frequencies. There are two easy ways to achieve this:

1. Let $\hat{G}_o(s)=\hat{G}(s)+\Delta(0)$. Then $G(s)=\hat{G}_o(s)+[\Delta(s)-\Delta(0)]$, implying that $G(0)=\hat{G}_o(0)$ and $\gamma_2[G-\hat{G}_o]\leq 2\epsilon$. This approach can be applied in general and produces appealing approximation error bounds, but has the drawback that the resulting approximation is usually bi-proper.

2. Let $\hat{G}_o(s)=\frac{G(0)}{\hat{G}(0)}\hat{G}(s)$. Then $G(s)=\hat{G}_o(s)+[\Delta(s)+(1-\frac{G(0)}{\hat{G}(0)})\hat{G}(s)]$, implying that $G(0)=\hat{G}_o(0)$ and $\gamma_2[G-\hat{G}_o]\leq\epsilon(1+\gamma_2[\hat{G}]/\hat{G}(0))$. In this case, the refinement preserves the roll-off properties of the original approximation but the approach may fail if $\hat{G}(0)$ is close to zero. In addition, a minor inconvenience comes from the dependence of the resulting error bounds on the system itself.

1.3.1 Examples:

a. (*Dominant-Parasitic Pole decomposition*) Let $G(s)=\frac{50}{(s+1)(s+50)}$. Then

$$G(s)=\hat{G}(s)+\Delta(s)=\frac{50/49}{s+1}-\frac{50/49}{s+50}$$

from which we obtain the magnitude plots shown in Fig. 1.3 and $\gamma_2[\Delta]=-33.8\,dB$. Notice that the pole closer to the jw-axis becomes part of the reduced order approximation —termed as a dominant pole— while the 'fast' pole becomes part of the perturbation —termed as a parasitic pole. This is usually the case unless the slow pole is "almost" canceled by a zero (see other examples below). This example depicts a case frequently encountered in practice, with the overall system being a cascade combination of a 'slow', dominant part and a 'fast' parasitic part, often due to actuator/sensor dynamics.

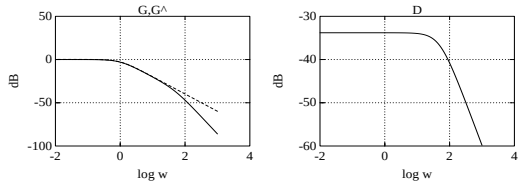

Figure 1.3: Magnitude Plots for Example 1.3.1.a.

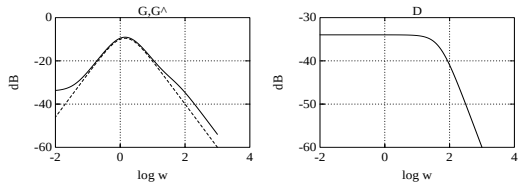

Figure 1.4: Magnitude Plots for Example 1.3.1.b.

b. Consider the transfer function

$$G(s) = \frac{2s^2 + 53s + 2}{(s+1)(s+2)(s+50)}$$

Decomposing $G(s)$ via partial fraction expansion, we get

$$G(s) = \hat{G}(s) + \Delta(s) = \left[\frac{2}{s+2} - \frac{1}{s+1}\right] + \frac{1}{s+50} = \frac{s}{(s+1)(s+2)} + \frac{1}{s+50}$$

with the magnitude plots shown in Fig. 1.4. Notice that the dominant part has zero DC gain while the parasitic part has DC gain 0.02. In other words, the steady-state of a step input response gives very little information about the reduced order model of $G(s)$. This is demonstrated in Fig. 1.5 where the step response of $G(s)$ and $\hat{G}(s)$ are shown; notice that although $\hat{y}$ yields a small bias in the steady state error (0.02), it does provide a good approximation of the transient response of y.

c. (*"Near" pole-zero cancellation*) In this example we demonstrate that the slow pole of the transfer function is not necessarily the one describing the dominant response. Such a case occurs when there is an approximate or exact pole zero cancellation. (Reminder: Right-half plane cancellations cause internal stability problems, invalidating any other results.) Let $G(s) = \frac{s+1.01}{(s+2)(s+1)}$. Taking the partial fraction expansion,

$$\hat{G}(s) = \frac{0.99}{s+2} \quad ; \quad \Delta(s) = \frac{0.01}{s+1}$$

with $\gamma_2[\Delta] = -40\,dB$ and magnitude plots as shown in Fig. 1.6. Observe that the slow pole has a small effect on the response of the system, due to its approximate cancellation by the zero (it is not dominant any more). This observation is also useful in several cases of feedback controller design e.g., "lag" compensators examined later in this class.

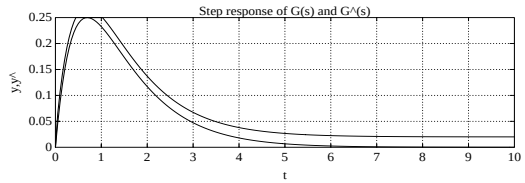

Figure 1.5: Step responses for Example 1.3.1.c.

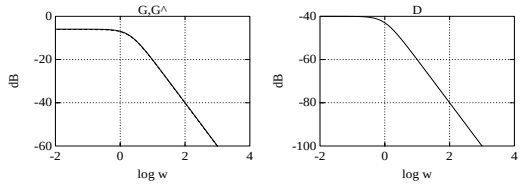

Figure 1.6: Magnitude Plots for Example 1.3.1.c.

Until this point we have considered the conceptually simpler case of the approximation of a transfer function by a low order model and an additive perturbation. However, in several occasions it may be more convenient to express the perturbation in a different form such as the multiplicative one:

$$G(s) = \hat{G}(s)[1 + \Delta(s)]$$

A basic difference from the additive case is that $\Delta(s)$ is now scaled by $\hat{G}(s)$ and is no longer required to be proper. This introduces an additional degree of difficulty since non-proper transfer functions do not have finite γ_2-gain. (Why?) Instead, Δ has poles in the open left-half plane and $\gamma_2[G\Delta]$ is small.

Although an analytical calculation of Δ may be difficult, the multiplicative form is particularly useful in cases where the frequency response of a transfer function is measured from real data. The uncertainty/ measurement error in the magnitude and phase of the frequency response measurements can be expressed as a multiplication by a complex number which is close to 1 when the error is small. That is,

$$G(jw) = |G(jw)|e^{j\phi(jw)} = |\hat{G}(jw)||\Delta'(jw)|e^{j\hat{\phi}(jw)+j\delta'(jw)} = \hat{G}(jw)\Delta'(jw)$$

where

- $G(jw)$ is the actual (measured) frequency response with magnitude $|G(jw)|$ and phase $\phi(jw)$;
- $\hat{G}(jw)$ is the approximate frequency response with magnitude $|\hat{G}(jw)|$ and phase $\hat{\phi}(jw)$; this is usually termed as the nominal model of the system;
- $\Delta'(jw)$ is the ratio $G(jw)/\hat{G}(jw)$ with magnitude $|\Delta'(jw)|$ and phase $\delta'(jw)$; in the ideal case where $G(s) = \hat{G}(s)$, $\Delta'(jw)$ should be equal to 1. The multiplicative perturbation Δ can now be defined simply as $\Delta' - 1$ and is commonly referred to as the multiplicative uncertainty.

The measurement and modeling process can be visualized by the graph in Fig. 1.7. In this figure, measurements of the magnitude of the frequency response are shown as small circles. These measurements describe a high order transfer function and are corrupted by noise and possibly affected by nonlinearities in the closed loop. They can be thought as defining an envelope for the magnitude of the frequency response. [6] Thus, a nominal model can be selected as to have a frequency response with magnitude anywhere inside or near the envelope as shown in Fig. 1.7. A similar procedure is then repeated for the phase. The outcome of this process is a function $\hat{G}(jw)$ describing the nominal frequency response of the system to be modeled.

Having extracted a nominal model from the data, the final step is to assess the properties of the modeling error (uncertainty). The frequency response of $\hat{G}(jw)$ is plotted against the real data and a figure for the worst case $\Delta'(jw)$ or $\Delta(jw) = \Delta'(jw) - 1$ is determined. The properties of $\Delta'(jw)$ can be described in various ways, depending on the amount of complexity that can be afforded in the design. Typically, one may describe $\Delta'(jw)$ with a magnitude and a phase envelope i.e.,

$$d_1(w) \leq |\Delta'(jw)| \leq d_2(w) \quad ; \quad p_1(w) \leq \delta'(jw) \leq p_2(w)$$

where d_i, p_i are some functions of frequency. On the other hand, $\Delta(jw)$ is often described by a proper transfer function $W(s)$ with poles and zeros in the open left-half plane such that $\gamma_2[W\Delta] \leq 1$. Alternatively, $\gamma_2[\Delta] \leq 1/\gamma_2[W]$, i.e., the magnitude of W^{-1} is an upper bound of the magnitude of Δ. An example of $W(s)$ is shown in Fig.1.8.

[6] When only a few measurements are available, there is a lot of freedom in choosing the size and shape of the envelope.

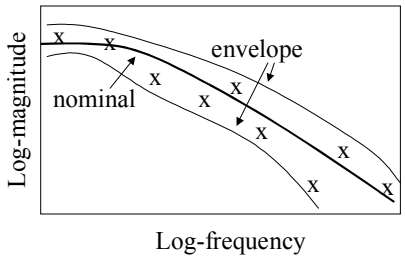

Figure 1.7: Frequency response measurements and the magnitude envelope.

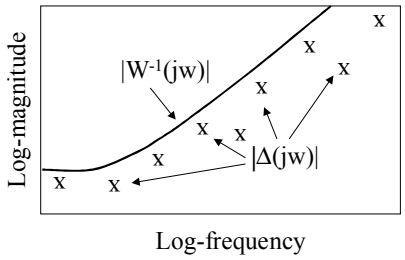

Figure 1.8: Characterization of the uncertainty: $|\Delta(jw)|$ and $W^{-1}(jw)$.

1.4 Disturbance attenuation: Loop and sensitivity transfer functions

Another application of the concept of the system gain is in the design of compensators to attenuate the effects of disturbances or sensor noise in the plant output. To formulate this problem, let us consider the general closed-loop system shown in Fig. 1.9, where $G(s)$ is the plant transfer function, $C(s), F(s)$ are the Cascade and Feedback compensator transfer functions, r is a reference or command input and d, n are disturbance signals. d is termed as output disturbance and usually describes the effect of low-pass external signals [7] affecting the plant output e.g., wind gusts on the trajectory of an airplane, bias signals on a servomotor etc. n is termed as sensor noise and describes noise signals that may affect the measurements of the plant output; such signals are usually high-pass.

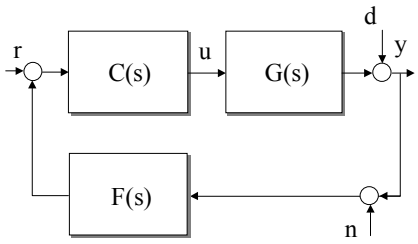

Figure 1.9: The closed-loop system.

Let us also define the transfer functions

- the loop transfer function $L(s) = G(s)C(s)F(s)$;
- the Sensitivity transfer function $S(s) = \dfrac{1}{1 + L(s)}$;

[7] A low-pass signal has energy in the low-frequency range.

- the Complementary Sensitivity transfer function $T(s) = \dfrac{L(s)}{1+L(s)}$;

It is easy to verify that $S(s)$ is the transfer function from the output disturbance d to the output while $-T(s)$ is the transfer function from the sensor noise n to the output. Thus, the output of the closed-loop system is described by

$$\hat{y}(s) = C(s)G(s)S(s)\hat{r}(s) + S(s)\hat{d}(s) - T(s)\hat{n}(s) \tag{1.1}$$

Needless to say, a similar treatment can be applied for perturbations entering at other points in the closed-loop system (e.g., input disturbances) or if the signal of interest is other than y (e.g. u). For reasons of simplicity, we focus our attention on eqn. (1.1) which, nevertheless, describes a fundamental trade-off in the design of control systems.

Further, at this point we assume that the sensitivity transfer function has all its poles in the open left-half plane and it has no right-half-plane cancellations. In fact, this assumption imposes a constraint on the possible selections for the compensator transfer functions and is usually referred to as:

"Internal Stability Condition:" C, F internally stabilize G;

meaning simply that the closed-loop system is exponentially stable. (i.e., its state-space description has no eigenvalues in the closed right-half plane.) The design of simple C and F satisfying this condition will be studied later in this class via Root-locus and Nyquist techniques. It is worth mentioning however, that the characteristic equation determining the stability properties of the closed-loop system is $1 + L(s) = 0$ which, in turn, depends only on the product $C(s)F(s)$ and not the individual choice of $C(s)$ and $F(s)$. This observation is used subsequently to enhance the tracking capabilities of the closed-loop system without affecting stability (see also eqn. (1.1)).

Under the internal stability condition, the disturbance attenuation problem is well-posed and can be formulated in various ways. For example,

a. Given certain characteristics of the energy spectrum of d and n, select C, F as to minimize the square root of the energy of $S[d] - T[n]$ in (1.1); or, in a simpler form,

b. Given certain characteristics of the energy spectrum of d and n, and a threshold μ, select C, F as to make $\|S[d]\|_2 + \|T[n]\|_2 \leq \mu(\|d\|_2 + \|n\|_2)$.

Here, we only consider the latter problem beginning with the much simpler case where $n = 0$.

1.4.1 Attenuation of Output Disturbances

Let us suppose that d is an energy signal with energy E_d and spectrum $\hat{d}(jw)$. From Parseval's theorem, it follows that the energy of the disturbance appearing in the output y is

$$\int_0^\infty |y_d(t)|^2\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |S(jw)|^2 |\hat{d}(jw)|^2\, dw$$

where, $y_d = S[d]$. [8]

From the last equation, it becomes apparent that an easy solution would be to make $|S(jw)|$ small for all w or, in other words, make $\gamma_2[S]$ small. Due to physical limitations however, this is not possible. The reason is that, in general, practical systems roll-off as the frequency increases i.e., in terms of Bode plots, their magnitude goes to zero as $w \to \infty$. Alternatively, we may say that the loop transfer function is strictly proper i.e., it has relative degree ≥ 1 where

$$\textit{relative degree of } G(s) \triangleq \textit{degree of denominator} - \textit{degree of numerator of } G(s)$$

[8] For non-zero initial conditions, property 7 of the previous section can be used.

This implies that $L(s)$ is strictly proper and $|L(jw)| \to 0$ as $w \to \infty$; hence $|S(jw)| \to 1$ as $w \to \infty$ and therefore $\gamma_2[S] \geq 1$.

On the other hand, typical output disturbances are low-pass signals such that $|\hat{d}(jw)|$ is small for $|w| > w_d$. Thus, intuitively we expect that we should only need to make $|S(jw)|$ small whenever $|\hat{d}(jw)|$ is large i.e., $|w| \leq w_d$. [9] The following results summarize some of the basic principles and limitations behind the design of compensators for the attenuation of output disturbances. In all cases we assume that the following condition holds

1.4.1 Condition:

a. *C, F internally stabilize G.*

b. *Initial conditions are zero.*

$\triangledown\triangledown$

Condition 1.4.1.b is not critical and can easily be removed by using property 7 of the previous section to find a correction term incorporating the effect of initial conditions. The former, however, is more fundamental in the sense that the compensator cannot be arbitrarily selected. Condition 1.4.1.a imposes certain limitations on the attenuation of output disturbances, depending both on the properties of $G(s)$ and the class of possible disturbances.

1.4.2 Lemma: *Under Conditions 1.4.1.a, b, suppose that d is a band-limited energy signal with magnitude spectrum $|\hat{d}(jw)|$ such that $\hat{d}(jw) = 0$ for $|w| > w_d$. Further, suppose that $|S(jw)| \leq \mu,\ \forall w \in [-w_d, w_d]$ and define $y_d = S[d]$. Then*

$$\|y_d\|_2 \leq \mu \|d\|_2$$

where both y_d and d are functions $\mathbf{R} \mapsto \mathbf{R}$. $\triangledown\triangledown$

Proof: Immediate from Parseval's theorem

$$\begin{aligned}
\int_{-\infty}^{\infty} |y_d(t)|^2\, dt &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |S(jw)|^2 |\hat{d}(jw)|^2\, dw \\
&= \frac{1}{2\pi}\int_{-w_d}^{w_d} |S(jw)|^2 |\hat{d}(jw)|^2\, dw \\
&\leq \sup_{w\in[-w_d,w_d]} [|S(jw)|^2] \frac{1}{2\pi}\int_{-w_d}^{w_d} |\hat{d}(jw)|^2\, dw \\
&\leq \mu^2 \int_{-\infty}^{\infty} |d(t)|^2\, dt
\end{aligned}$$

Taking square roots, the inequality of the lemma follows. $\square\square$

An alternative, more general form of the above Lemma which is also applicable to the case where the signals are zero for $t < 0$ can be stated as follows.

1.4.3 Lemma: *Under Conditions 1.4.1.a, b, suppose that d is an energy signal with magnitude spectrum $|\hat{d}(jw)|$ such that*

$$\frac{1}{2\pi}\int_{|w|>w_d} |\hat{d}(jw)|^2\, dw \leq \epsilon \|d\|_2^2$$

Further, suppose that $|S(jw)| \leq \mu,\ \forall w \in [-w_d, w_d]$ and define $y_d = S(s)d$. Then

$$\|y_d\|_2^2 \leq (\mu^2 + \gamma_2^2[S]\epsilon)\|d\|_2^2$$

$\triangledown\triangledown$

[9] Keep in mind that band-limited signals are necessarily non-causal.

That is, if most of the energy of d is contained in the interval $[-w_d, w_d]$ ($\epsilon \ll 1$) and $S(jw)$ is small in magnitude in the same interval, then the energy of y_d will be small provided that $|S(jw)|$ does not attain large values for any $w \in \mathbf{R}$. The implication of this lemma is that, since in practice d is not strictly band-limited, we are not only interested in making $S(jw)$ small inside a frequency range, but also in avoiding excessively large values of $S(jw)$ at any frequency.

Proof: From Parseval's theorem,

$$\begin{aligned}
\int_{-\infty}^{\infty} |y_d(t)|^2\, dt &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |S(jw)|^2 |\hat{d}(jw)|^2\, dw \\
&= \frac{1}{2\pi}\int_{-w_d}^{w_d} |S(jw)|^2 |\hat{d}(jw)|^2\, dw + \frac{1}{2\pi}\int_{|w|>w_d} |S(jw)|^2 |\hat{d}(jw)|^2\, dw \\
&\leq \sup_{w\in[-w_d,w_d]} [|S(jw)|^2] \frac{1}{2\pi}\int_{-w_d}^{w_d} |\hat{d}(jw)|^2\, dw + \gamma_2^2[S]\frac{1}{2\pi}\int_{|w|>w_d} |\hat{d}(jw)|^2\, dw \\
&\leq \mu^2 \|d\|_2^2 + \gamma_2^2[S]\epsilon\|d\|_2^2
\end{aligned}$$

□□

Notice that the above lemma is also applicable if d and y_d are signals which are zero for $t < 0$ in which case $\|\cdot\|_2$ is simply $(\int_0^\infty (\cdot)^2\, dt)^{1/2}$. Moreover, similar expressions can be developed for a wide class of not-necessarily-energy-signals by limiting our attention to finite intervals. However, since these expressions are more complicated without offering any additional insight, they are omitted from the present discussion.

1.4.2 Attenuation of Sensor Noise

Analogous results can be obtained in the case of sensor noise i.e., the case where

$$y_n = -T[n]$$

should be attenuated. Typically, the sensor noise is a signal with high frequency components, which indicates that $T(jw)$ should be made small for large w.

1.4.4 Lemma: *Under Conditions 1.4.1.a, b, suppose that n is an energy signal with magnitude spectrum $|\hat{n}(jw)|$ such that*

$$\frac{1}{2\pi}\int_{-w_n}^{w_n} |\hat{n}(jw)|^2\, dw \leq \epsilon\|n\|_2^2$$

Further, suppose that $|T(jw)| \leq \mu,\ \forall |w| > w_n$ and define $y_n = T(s)n$. Then

$$\|y_n\|_2^2 \leq (\mu^2 + \gamma_2^2[T]\epsilon)\|n\|_2^2$$

▽▽

Since $T(s) = L(s)/[1 + L(s)]$, the reduction of the effect of sensor noise requires that $L(jw)$ should be small. This requirement is of course the opposite of the previous one for the Sensitivity reduction ($L(jw)$ large) which is another manifestation of the fundamental limitation:

$$S(s) + T(s) = 1$$

$S(jw)$ and $T(jw)$ *cannot both be small at the same frequency.*

Fortunately, in most applications output disturbances d are low frequency signals while sensor noise n is high frequency and $w_d \ll w_n$. (If this is not the case, you either have to give up the attenuation of d or select a better sensor.)

Combining the previous Lemmas, it is straightforward to establish the following result for the mixed output-disturbance/sensor-noise case.

1.4.5 Corollary: *Under Conditions 1.4.1.a, b, suppose that*

a. *d, n are energy signals with magnitude spectrum $|\hat{d}(jw)|$, $|\hat{n}(jw)|$, respectively, such that*

$$\frac{1}{2\pi}\int_{|w|>w_d} |\hat{d}(jw)|^2\, dw \leq \epsilon \|d\|_2^2$$

$$\frac{1}{2\pi}\int_{-w_n}^{w_n} |\hat{n}(jw)|^2\, dw \leq \epsilon \|n\|_2^2$$

where $w_d \ll w_n$ and $\epsilon \ll 1$. [10]

b. $|S(jw)| \leq \mu_S,\ \forall w \in [-w_d, w_d]$ *and* $|T(jw)| \leq \mu_T,\ \forall |w| \geq w_n$

Then,

$$\|y_d + y_n\|_2 \leq \mu_S \|d\|_2 + \mu_T \|n\|_2 + \sqrt{\epsilon}\,(\gamma_2[S]\|d\|_2 + \gamma_2[T]\|n\|_2)$$

$\bigtriangledown\bigtriangledown$

Proof: Immediate, using the facts $\sqrt{|a|+|b|} \leq \sqrt{|a|} + \sqrt{|b|}$ and $\|x+y\|_2 \leq \|x\|_2 + \|y\|_2$ in the previous expressions. $\square\square$

The above Corollary shows how output-disturbance and sensor-noise attenuation specifications can be translated into Sensitivity and Complementary Sensitivity specifications. Typical examples of such specifications are shown in Fig.1.10, limiting both the magnitude of $S(jw)$ and $T(jw)$ in the respective frequency range of interest, as well as their maximum values.

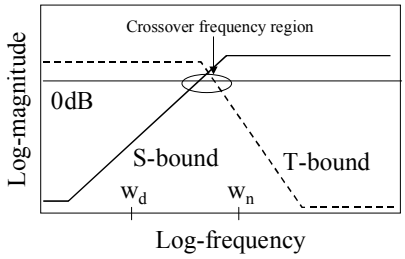

Figure 1.10: Sensitivity and Complementary Sensitivity Specifications.

In several cases it is convenient to use an alternative way to describe the specifications on $S(jw)$ and $T(jw)$, by means of two auxiliary transfer functions, say $W_S(s)$ and $W_T(s)$ i.e.,

$$\gamma_2[W_S S] \leq 1 \ \ ; \ \ \gamma_2[W_T T] \leq 1$$

These transfer functions act as frequency-dependent weights and they are large wherever the corresponding sensitivity function should be small. Achieving the above inequalities would mean that $|S(jw)|_{dB} \leq -|W_S(jw)|_{dB}$ and similarly for $T(s)$. For example, in order to reject constant output disturbances while keeping $S(jw)$ below 6.021 dB for frequencies $w \geq 1$, one may choose $W_S(s) = \frac{0.5(s+1)}{s}$. The magnitude plots for such a weight and the corresponding bound for $|S(jw)|$ are shown in Fig. 1.11. The reason behind this formulation is its compatibility with modern controller design techniques, e.g., H_∞, H_2. These provide unified and reliable computational tools to design controllers that satisfy the above sensitivity bounds, or come close in an optimal sense. The H_∞ tools that were developed in the late 80's and 90's have to a large extend bridge the old gap between classical techniques (Bode/Nyquist/root-locus) and "modern" ones (state-space/linear quadratic). They offer attractive state-space computational properties with frequency domain interpretations. What is even more important is their generality. They work for arbitrary system dimension and number of inputs/outputs. And when they fail, the reason is overly tight or infeasible specifications. The designer no longer needs to spend a lot of time in translating the problem to something more

[10] Only the cases where ϵ is small are of practical interest.

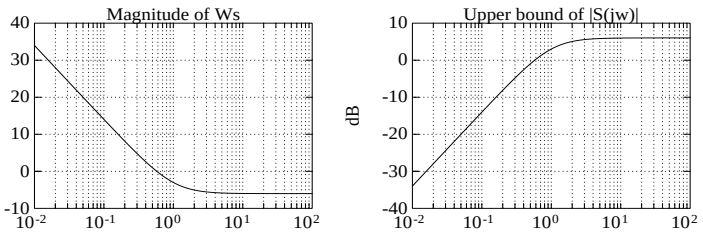

Figure 1.11: Sensitivity Weight-Specifications.

tractable. Instead, the main problem becomes the choice of the objectives and how that reflects practical perfromance concerns.

From a different point of view, one may translate the sensitivity specifications into specifications on the loop transfer function. For this, the following inequalities can be used as *sufficient* conditions to achieve the sensitivity specifications.

1. For $|S(jw)| \leq \mu_S \ll 1$ at some frequency, select $|L(jw)| \gg 1$ such that

$$|L(jw)| \geq \frac{1+\mu_S}{\mu_S}$$

2. For $|T(jw)| \leq \mu_T \ll 1$ at some frequency, select $|L(jw)| \ll 1$ and such that

$$|L(jw)| \leq \frac{\mu_T}{1+\mu_T}$$

Vice-versa, if $|L(jw)| > 1$ at some frequency we can always find $\mu_S > 0$ such that

$$|L(jw)| = \frac{1+\mu_S}{\mu_S}$$

Then, $|S(jw)| \leq \mu_S$ at the same frequency. Similarly, if $|L(jw)| < 1$ at some frequency we can always find $\mu_T > 0$ such that

$$|L(jw)| = \frac{\mu_T}{1+\mu_T}$$

Then, $|T(jw)| \leq \mu_T$ at the same frequency. (Verify these inequalities!)

This technique, called *loop-shaping*, has the advantage that one deals directly with $L(s) = G(s)C(s)F(s)$ which is a simple function of the compensator transfer functions. On the other hand, the drawbacks of this technique are that closed-loop stability must be treated separately and that it does not allow for an easy control over the maximum magnitude of $S(jw)$ and $T(jw)$. To avoid excessive sensitivity peaking, a good "rule of thumb" is

- *$L(jw)$ should roll-off with an approximate rate of -20 dB/decade around the crossover frequency (where $L(jw) \simeq 1$).*

Of course, the previously encountered fundamental limitations e.g. sufficient separation of w_d and w_n and issues related to internal stability are present in this case as well. Typical loop-shaping specifications are shown in Fig. 1.12.

At this point, it is also important to emphasize two fundamental limitations arising from the internal stability condition, in relation to right-half plane poles and zeros of $L(s)$. These limitations can be demonstrated using root-locus arguments or, for a more general interpretation, complex analytic arguments. They are stated below only as qualitative "rules of thumb."

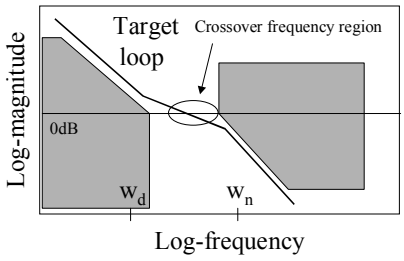

Figure 1.12: Loop-Shaping Specifications.

- Right-half plane poles force a "minimum bandwidth" i.e., large gains $L(jw)$ at certain frequencies are needed to stabilize the closed loop system. The implication of this fact is that right-half plane poles impose a lower bound on the range over which $T(jw)$ can be made small ($|w| > w_n$).
- Right-half plane zeros limit the "maximum achievable bandwidth" i.e. $L(jw)$ cannot be too large at certain frequencies if closed-loop stability is to be ensured. The implication of this fact is that right-half plane zeros restrict the frequencies over which $S(jw)$ can be made small.

1.4.6 Example: *Loop-shaping for disturbance attenuation*

Suppose that, for a plant $G(s)$, we would like to design —if possible— a controller such that output disturbances in the frequency range $[0, 1]$ are attenuated by at least $40\,dB$ and sensor noise in the frequency range $[10, \infty]$ is attenuated by at least $20\,dB$. We would also like the closed loop system to completely reject constant disturbances, assuming that $G(s)$ has no zeros at $s = 0$ (see also next subsection). Our problem is to describe the magnitude properties of the Loop transfer function, $|L(jw)|$ for which the above specifications are met, assuming that there exists $C(s)F(s)$ producing such an $L(s)$ and for which the closed loop is internally stable.

For the first specification, we need $|S(jw)| \leq 0.01$ for $w \in [0, 1]$. For the second, $|T(jw)| \leq 0.1$ for $w \in [10, \infty]$. The third specification, implies that $S(s)$ must have a zero at $s = 0$, that is $L(s)$ should have a pole at $s = 0$. Taking $\mu_S = 0.01$ and $\mu_T = 0.1$ we have the following conditions on $L(jw)$:

$$|L(jw)| > \frac{1.01}{0.01} = 101 = 40.0864\,dB; \quad w \in [0, 1]$$

$$|L(jw)| < \frac{0.1}{1.1} = 0.0909 = -20.8279\,dB; \quad w \in [10, \infty]$$

$$L(s) = \frac{1}{s}L'(s)$$

for some $L'(s)$. To find such an $L(s)$ observe first that its magnitude should drop more than $60\,dB$ in one decade, meaning roughly that $L(s)$ should have more than 3 poles in excess of zeros in the frequency interval [1,10]. Since the factor $1/s$, needed for the third specification accounts for -20 dB/dec roll-off we may try the following $L(s)$-candidate:

$$L(s) = \frac{(101)(1.8^3)}{s(s + 1.8)^3}$$

where the factor $101/s$ gives $L(s)$ the desired properties at low frequencies and the rest introduce the necessary roll-off for the high-frequency specifications, hopefully without distorting the low-frequency characteristics very much. Calculating the frequency response of $L(s)$ we find that it is approximately 4 dB below the first specification at $w = 1$ and meets the second one with approximately 4 dB margin. Hence, adjusting the gain constant, we obtain

$$L(s) = \frac{933.553}{s(s + 1.8)^3}$$

which has magnitude $40.5818\,dB$ at $w = 1$ and $-21.0127\,dB$ at $w = 10$, meeting all the specifications.

It should be emphasized that these computations are only indicative of the loop-shaping procedure. High accuracy is often unnecessary. Moreover, a target loop that meets the S and T specifications does **not** necessarily yield a stable closed-loop system. Here, we do expect problems since this target loop violates the $-20dB/dec$-slope rule around the crossover. Indeed, the poles of the corresponding $1/(1+L)$ are

$$-5.297 \pm j\,3.87 \;\;;\;\; 2.597 \pm j\,3.86$$

Consequently, additional factors should be introduced to stabilize such a closed loop without distorting the properties obtained so far. In fact, these factors only need to affect the middle-frequency range properties of $L(s)$, e.g., adjust the phase and roll-off rate of $L(s)$ around the crossover. Their design will be studied later in this class via Root-locus and Nyquist techniques (lead-lag compensation) and is feasible provided that there is a sufficiently wide middle-frequency range $[w_d, w_n]$. Of course, elementary lead-lag design techniques may not be very efficient for the compensator design. Typical problems that may appear include excessive peaks of the sensitivity functions or the need for a very wide middle-frequency range. Such problems may be partially solved —if at all possible— with several iterations and a very careful (and tedious) design. On the other hand, in such tight design problems it is recommended that one uses more advanced design techniques.

In conclusion, one may decompose the frequency spectrum in three parts, low, middle and high frequency range with, roughly, the following significance for each part:

- Low-frequency range: Performance spec's, Disturbance attenuation.
- High-frequency range: Sensor noise attenuation, modeling error effects/robustness (studied in a subsequent section).
- Mid-frequency range: Closed-loop stability specifications; stability margins, robustness.

$\triangledown\triangledown$

1.4.3 The Internal Model Principle

Other interesting interpretations of the general disturbance attenuation problem in terms of the sensitivity transfer functions can be obtained for periodic signals admitting a Fourier series expansion. Consider, for example, a disturbance d such that

$$d(t) = \sum_n a_n e^{jnw_0 t}$$

It follows that

$$y_d(t) = \sum_n S(jnw_0) a_n e^{jnw_0 t}$$

The last relationship shows that the contribution of the n-th component (complex exponential) can be attenuated if $S(jnw_0)$ is small. This observation can be generalized via Laplace transforms to what is also known as the *internal model principle*.

1.4.7 Lemma: *Under the internal stability condition (1.4.1.a), suppose that the disturbance d is modeled by a differential equation $\Lambda(s)d(s) = 0$ (expressed in terms of Laplace transforms). Further, suppose that $\Lambda(s)$ is a factor of the numerator of $S(s)$. Then $y_d \to 0$ as $t \to \infty$ exponentially fast, i.e., the disturbance is completely rejected from the output response asymptotically with time.* $\triangledown\triangledown$

Given that the loop transfer function $L(s) = G(s)C(s)F(s)$ is a ratio of two polynomials, a necessary and sufficient condition for the above lemma to hold is that, in addition to internal stability, $\Lambda(s)$ is a factor of the denominator of $L(s)$.

Proof: From the assumptions of the Lemma, $S(s) = S'(s)\Lambda(s)$ where $S'(s)$ is strictly proper with poles in the open left-half plane. Further, writing a state-space representation of $\Lambda(s)d(s) = 0$ we have that

$$d(s) = \frac{1}{\Lambda(s)} \sum_i \Lambda_i(s) d_{0i}$$

where $\Lambda_i(s)$ are polynomials of degree less than the degree of $\Lambda(s)$ and d_{0i} are the initial conditions for d. Hence,

$$y_d(s) = S'(s) \sum_i \Lambda_i(s) d_{0i}$$

with $S'(s)\Lambda_i(s)$ being strictly proper transfer functions with poles in the open left-half plane. It now follows that $y_d(t) \to 0$ as $t \to \infty$ exponentially fast, irrespective of the values of d_{0i} or the precise form of $L_i(s)$. □□

Thus, in all of the above cases, the attenuation of output disturbances is directly related to the magnitude of the Sensitivity transfer function at the frequencies where the disturbance may have energy. This can be achieved by making the loop transfer function large at the same frequencies:

$$|S(jw)| = \frac{1}{|L(jw)+1|} \leq \frac{1}{|L(jw)|-1} \; ; \quad |L(jw)| \gg 1$$

Obviously, if jw_0 is a pole of $L(s)$, the disturbance component associated with this frequency is completely eliminated from the output asymptotically as $t \to \infty$. Notice that it is exactly at this point where the internal stability assumption is needed. The reason is that for the statements to make sense, the outputs y_d as well as y and all internal signals must be bounded for all possible bounded inputs and all initial conditions. This requires that the poles of $S(s)$ are in the open left-half plane and there are no cancellations in the closed right-half plane.

To demonstrate the limitations imposed by such a condition, consider an example where $G(s)$ has a zero at $s = 0$ i.e., s is a factor of the numerator of $G(s)$. Then, constant disturbances cannot be rejected (in fact, they cannot even be attenuated) since that would require $C(s)F(s)$ to have a double pole at $s = 0$. Such a design however, would contain a pole-zero cancellation at $s = 0$, something that is not permitted by the internal stability condition.

A more subtle trade-off in the design of the compensator is imposed by the need to avoid large magnitudes of $S(jw)$ at all frequencies. In general, a reduction of the sensitivity in a frequency range implies an increase of the sensitivity in the rest of the frequencies. The fundamental limitations imposed by right-half plane poles and zeros or high roll-off rates are captured in the following results.

1.4.8 Theorem: *(Waterbed effect) Let M denote the maximum magnitude of $S(jw)$ in a frequency band $[w_1, w_2]$ and suppose that $L(s)$ has a zero at z_0 with $Rez_0 > 0$. Then there exist positive constants c_1, c_2, depending only on w_1, w_2 and z_0, such that*

$$c_1 \log M_1 + c_2 \log \gamma_2[S] \geq 0$$

1.4.9 Theorem: *(Bode's Integral Theorem/Area Formula) Suppose that the relative degree of $L(s)$ is ≥ 2 and let p_i denote the open right-half plane poles of $L(s)$ (including multiplicities). Also, suppose that $S(s)$ is BIBO stable. Then,*

$$\int_0^\infty \ln |S(jw)| \, dw = \pi \sum_i Re(p_i)$$

▽▽

Both theorems state that pushing the Sensitivity magnitude down inside a frequency interval, (e.g., as dictated by performance specifications) results in the Sensitivity magnitude popping up outside that frequency interval. The first result applies to non-minimum-phase systems only whereas the area formula applies in general, except for the relative degree assumption. In particular, the area formula does not by itself imply a peaking phenomenon, only an area conservation. However, one can infer a type of Sensitivity peaking from the area formula when another constraint is imposed, namely Loop bandwidth. Such a constraint, arising from sensor noise attenuation of robustness considerations is almost always present in practice and effectively requires that the area preservation should occur in a finite frequency range. The manner in which the "positive area" is distributed over the available frequencies is precisely what separates good controller design techniques from bad ones. And although good designs have limitations, bad designs do not!

Finally, the principles of output disturbance rejection can be employed in the design of a compensator to reject external disturbances that enter anywhere in the loop, as long as an internal model is available. All one has to do is to design the compensator to introduce the appropriate zeros in the transfer function from the disturbance to the output. Needless to say, given a plant transfer function, the internal stability condition imposes certain limitations on the disturbance models for which such a design is possible or, vice-versa, limitations on the plant transfer function given a model for the disturbance. For example, suppose that a sensor noise n, satisfying the internal model $\Lambda(s)n(s) = 0$ [11] is to be rejected at the plant output. Since the contribution of the sensor noise at the output is $y_n(s) = -T(s)n(s)$, we must require that $T(s)$ has zeros at the zeros of $\Lambda(s)$ and poles in the open left-half plane. The former condition means that the product $C(s)F(s)$ should have $\Lambda(s)$ as a factor in the numerator while the latter requires that the plant $G(s)$ has no poles at the zeros of $\Lambda(s)$.

1.4.4 Tracking of Reference Signals

The tracking problem can be thought as a different version of the disturbance attenuation problem, by considering the tracking error $e = r - y$ as the output of the closed-loop system and r as a disturbance. In its simplest form, the feedback compensator $F(s)$ is identically 1, which yields a transfer function $r \mapsto e$ being equal to the Sensitivity transfer function $S(s)$. Consequently, all the arguments of subsection 4.1 can be employed to produce analogous results. Of particular interest is the use of internal models to track a class of reference signals. For example, in order to track unit steps asymptotically with time, ($d = n = 0$), $L(s)$ must have a factor 's' in the denominator (type 1 system) while the asymptotic tracking of ramps would require a factor 's^2' in the denominator of $L(s)$ (type 2 system). Both can 'easily' be achieved by designing $C(s)$ as to have a pole at $s = 0$ of the appropriate multiplicity, assuming of course that the plant has no zeros at $s = 0$. Similarly, sinusoid inputs of frequency w_0 can be tracked asymptotically in time provided that $s^2 + w_0^2$ is not a factor of the plant numerator and that it is included in the denominator of $C(s)$.

If $F(s)$ is not identically 1, the tracking problem can be studied by considering the error transfer function

$$\frac{e(s)}{r(s)} = 1 - G(s)C(s)S(s)$$

and applying the same principles. One distinct advantage of the decomposition of the compensator as a cascade part $C(s)$ and a feedback part $F(s)$, is that undesirable zeros of the compensator which are required for stability and sensitivity properties, can be removed from the transfer function $r \mapsto y$. This follows immediately from the observation that any decomposition of $C(s)F(s)$, not involving right-half plane cancellations, still satisfies the internal stability condition and leaves the closed-loop poles unaffected. Observe that the zeros of the transfer function $G(s)C(s)S(s)$ are the zeros of $G(s)$, the zeros of $C(s)$ and the poles of $F(s)$. Hence, the zeros of $G(s)C(s)S(s)$ and can be altered by an appropriate decomposition of the product $C(s)F(s)$.

In order to demonstrate the effect of the zeros on the response of the system, consider the transfer function

$$\frac{N(s)}{s^2 + s + 1}$$

The step response of this transfer function is shown in Fig. 1.13 for several polynomials $N(s)$. Notice that for zeros which are "close" to the jw-axis, with respect to the poles, there is a rapid deterioration of the overshoot characteristics. Similar undesired behavior is exhibited when $N(s)$ has roots in the right-half plane, with the additional characteristic of an "undershoot." Such effects can be avoided by including any zeros of the compensator "near" the jw-axis in $F(s)$ and any poles of the compensator "near" the jw-axis in $C(s)$. For example, suppose that the desired shapes of $S(s), T(s)$ have been obtained for $C(s)F(s) = \frac{s-1}{s(s+1)}$. We may now select $C(s) = \frac{s+5}{s(s+1)}$ and $F(s) = \frac{s-1}{s+5}$ for which the zero at $s = 1$ does not appear in the transfer function $r \mapsto y$. Notice however, that the zeros of the plant always appear in the same transfer function unless they are explicitly cancelled by the poles of $C(s)$. (open left-half plane zeros *only*!)

[11] E.g., for $n(t)$ being a sinusoid of frequency w_0 rad/s, $\Lambda(s) = s^2 + w_0^2$

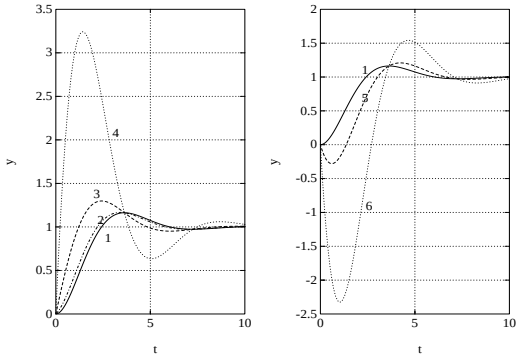

Figure 1.13: Step Response of $N(s)/(s^2+s+1)$; Legend: 1-$N(s)=1$, 2-$N(s)=0.2s+1$, 3-$N(s)=s+1$, 4-$N(s)=5s+1$, 5-$N(s)=-s+1$, 6-$N(s)=-5s+1$.

1.4.5 The Effects of Modelling Error

In order to get the complete picture of the controller design problem, we must also address the issue of the effects of modeling error on the closed-loop system. We have already discussed the model order reduction problem, that is the approximation of a complicated transfer function by a simpler one together with a (small) perturbation transfer function. The same ideas were extended to the modeling problem where a transfer function of a system is derived based on experimental data (typically, a frequency response).

Both these problems are parts of the overall controller design process. For example, one first obtains measurements of the frequency response of the system to be controlled which are used to develop a nominal-transfer-function-plus-uncertainty description of the system.[12] Since this nominal transfer function is usually too complicated for control purposes, one may then apply a model order reduction technique to obtain a simplified model of the system, capturing its essential dynamical behavior. Thus, the original system is eventually described by a ("simple") nominal transfer function and a perturbation transfer function that is usually expressed in either of the following two forms:

- *Additive Uncertainty:* $G(s)=\hat{G}(s)+\Delta(s)$;
- *Multiplicative Uncertainty:* $G(s)=\hat{G}(s)[1+\Delta(s)]$;

where $G(s)$ is the original system transfer function, $\hat{G}(s)$ is the simplified low-order nominal transfer function and $\Delta(s)$ is the uncertainty transfer function, possessing certain "smallness" properties as defined in Section 3. That is, for the additive case, $\gamma_2[\Delta]$ is small and for the multiplicative case $\gamma_2[W\Delta]$ is small where $W(s)$ is some transfer function acting as a frequency-dependent weight for $\Delta(s)$.

Next, the nominal model of the plant $\hat{G}(s)$ is used to design a controller achieving certain disturbance attenuation and tracking properties. For this problem we have described the basic characteristics of such a design in terms of the *nominal* Sensitivity and Complementary Sensitivity transfer functions or the nominal Loop transfer function and studied the fundamental trade-off's and limitations of such a design.

Notice that in the development so far, the transfer function of the plant was assumed to be known and was used in all the derivations. In practice however, and in the context of system modeling and model order reduction, only $\hat{G}(s)$, the nominal transfer function of the plant, is known. Consequently, the results of the previous section are applicable to the *nominal* closed-loop system (i.e., if $G(s)$ where equal to $\hat{G}(s)$), with the natural definitions of

- the nominal loop transfer function $\hat{L}(s)=\hat{G}(s)C(s)F(s)$;
- the nominal Sensitivity transfer function $\hat{S}(s)=\dfrac{1}{1+\hat{L}(s)}$;

[12] In some cases, this may be the outcome of a description of the system based on "first principles", e.g., Newton's law, Kirchoff's law, Mass-Energy Balances etc.

- the nominal Complementary Sensitivity transfer function $\hat{T}(s) = \dfrac{\hat{L}(s)}{1+\hat{L}(s)}$.

Since in practice $\hat{G}(s)$ is only an approximation of $G(s)$, one may naturally pose the question:

- "Suppose we design a controller $C(s)F(s)$ to meet certain stability/disturbance attenuation specifications for the approximate plant $\hat{G}(s)$. What can we say about the behavior of the *actual* closed-loop system if the *same* controller is used to control the actual plant?"

In other words, we would like to know whether a controller design based on the simplified nominal model of the plant guarantees a "desirable" behavior (stability/disturbance attenuation) for the actual closed-loop system. If this is not the case, the whole approximation process is futile and one would need the *exact* description of the system in order to perform a controller design. If, on the other hand, such guarantees can be established for a controller design and certain class of perturbations then this controller is referred to as "robust" with respect to that class of perturbations. For example, if the controller $C(s)F(s)$ guarantees that for any perturbation $\Delta(s)$ such that $\gamma_2[\Delta] < \epsilon$ the closed-loop system is stable, then we will say that $C(s)F(s)$ guarantees robust stability of the closed-loop with respect to the class of perturbations $\gamma_2[\Delta] < \epsilon$. Analogously, if the controller $C(s)F(s)$ guarantees that for any perturbation $\Delta(s)$ such that $\gamma_2[\Delta] < \epsilon$ the closed-loop system attenuates output disturbances by at least $20\,dB$, we will say that $C(s)F(s)$ guarantees robust performance of the closed-loop —in the sense of output disturbance attenuation— with respect to the class of perturbations $\gamma_2[\Delta] < \epsilon$.

The robustness problem is non-trivial and conceptually different than the disturbance attenuation problem. Although both are concerned with the closed-loop behavior in the presense of perturbations, the latter deals with exogenous signals which cannot destabilize the closed-loop system. In contrast, the robustness problem deals with state-dependent perturbations which *may* destabilize the closed loop system. For example, consider the system

$$\dot{x} = -x + r$$

which produces a bounded x for any bounded r. Suppose now that the input of this system is perturbed by a state-dependent perturbation of the form $u = \delta x$ where δ is some unknown constant i.e.,

$$\dot{x} = -x + r + u = -(1-\delta)x + r$$

One can immediately see that if $|\delta| < 1$ the perturbed system will be BIBO stable. However, if larger perturbations are allowed e.g., $|\delta| < 2$, then the closed-loop system may be unstable e.g., $\delta = 1.5$. In an analogous fashion, in the case of additive uncertainty, one may think of the output of Δ as an output disturbance $d = \Delta[u]$, although d is not an external signal but depends on the plant input u and, consequently, on the state-vector of the closed-loop system. Such perturbations (additive or multiplicative uncertainty) may destabilize the closed-loop system if they are allowed to be large enough.

In order to determine the stability and disturbance attenuation properties of an actual closed-loop system, let us consider the case of the additive uncertainty first. In this case the actual closed-loop system, shown in Fig.1.14, can be viewed as the nominal closed-loop system perturbed by an uncertainty $\Delta(s)$.

Furthermore, let us define the sensitivity-like transfer function $\hat{S}_a(s)$ by

$$\hat{S}_a(s) = \frac{C(s)F(s)}{1+\hat{L}(s)} = C(s)F(s)\hat{S}(s)$$

Then the transfer function from the uncertainty output to the plant input u is $-\hat{S}_a(s)$. Notice that $\hat{S}_a(s)$ depends only on the nominal plant $\hat{G}(s)$ and it can be evaluated independent of $\Delta(s)$.

Calculating the Sensitivity and Complementary Sensitivity transfer functions for the actual closed-loop system we have

$$S(s) = \frac{1}{1+[\hat{G}(s)+\Delta(s)]C(s)F(s)} = \hat{S}(s)\frac{1}{1+\Delta(s)\hat{S}_a(s)}$$

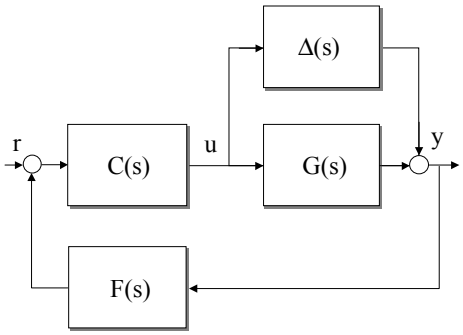

Figure 1.14: The actual closed-loop system in the case of additive uncertainty.

$$T(s) = \frac{[\hat{G}(s) + \Delta(s)]C(s)F(s)}{1 + [\hat{G}(s) + \Delta(s)]C(s)F(s)} = [\hat{T}(s) + \Delta(s)\hat{S}_a(s)]\frac{1}{1 + \Delta(s)\hat{S}_a(s)}$$

It follows that both sensitivities will be BIBO stable if $1 + \Delta(s)S_a(s) \neq 0$ for all s in the closed right-half plane. This observation can be stated precisely as follows.

1.4.10 Theorem: *Suppose that $C(s), F(s)$ internally stabilize $\hat{G}(s)$ and $\Delta(s)$ is analytic in the open right-half plane and bounded in the closed right-half plane. Further, suppose that*

$$\gamma_2[\Delta\hat{S}_a] < 1.$$

Then the actual closed-loop system is internally (and BIBO) stable. Furthermore,

$$\gamma_2[S] \leq \frac{\gamma_2[\hat{S}]}{1 - \gamma_2[\Delta\hat{S}_a]} \quad ; \quad \gamma_2[T] \leq \frac{\gamma_2[\hat{T}] + \gamma_2[\Delta S_a]}{1 - \gamma_2[\Delta\hat{S}_a]}$$

$\triangledown\triangledown$

Proof: (Outline) Condition 1.4.1.a and the assumption on $\Delta(s)$ imply that $\hat{S}_a(s), \hat{S}(s), \hat{T}(s)$ and $\Delta(s)$ have all their poles in the open left-half plane. Since the poles of $S(s)$, $T(s)$ or any other transfer function in the closed-loop are a subset of the poles of the above transfer functions and the roots of $1 + \Delta(s)\hat{S}_a(s) = 0$, stability follows from $1+\Delta(s)\hat{S}_a(s) \neq 0$ in the closed right-half plane. For this, it suffices that $\sup_{s\in RHP} |\Delta(s)\hat{S}_a(s)| < 1$. Since $\Delta(s)\hat{S}_a(s)$ is analytic and bounded in the RHP, by the maximum modulus theorem the supremum is achieved on the the jw-axis. Hence, it suffices that $\sup_{w\in\mathbf{R}} |\Delta(jw)\hat{S}_a(jw)| < 1$ which is precisely the condition stated in the Theorem. Further, the inequalities of the theorem are easily verified using the properties of the γ_2-gains and $\gamma_2[\frac{1}{1+H}] \leq \frac{1}{1-\gamma_2[H]}$ for $\gamma_2[H] < 1$ (verify this). $\square\square$

The above theorem describes the class of additive uncertainty perturbations for which a controller designed for the nominal plant, guarantees the stability of the actual (perturbed) plant. This class is precisely additive uncertainties whose magnitude of the frequency response is strictly below that of $1/S_a(s)$. In other words, in order to guarantee stability in the presence of an additive uncertainty $\Delta(s)$, we should design the controller $C(s)F(s)$ such that $|S_a(jw)|_{dB} < -|\Delta(jw)|_{dB}$. Notice that, if only $\gamma_2[\Delta]$ is available, $S_a(s)$ must satisfy $|S_a(jw)|_{dB} < -\gamma_2[\Delta]_{dB}$.

Further, the worst case performance of the actual closed-loop system in terms of the attenuation of external disturbances can be evaluated exactly as before, using the upper-bound estimates of $\gamma_2[S], \gamma_2[T]$ as given by the Theorem. Observe that these estimates depend only on the nominal Sensitivities and the size of the uncertainty and can easily be evaluated given an upper bound of the latter. Moreover, as $\gamma_2[\Delta] \to 0$, the actual disturbance attenuation approaches the nominal one.

Also notice the expected trade-off: In order to improve the output disturbance attenuation of the actual closed-loop system we must decrease $\hat{S}(s)$ without increasing $\gamma_2[\Delta\hat{S}_a]$. For the former we must increase $\hat{L}(s)$ by increasing $C(s)F(s)$. Hence, for the latter, we must require that any increase in $|\hat{L}(jw)|$ should

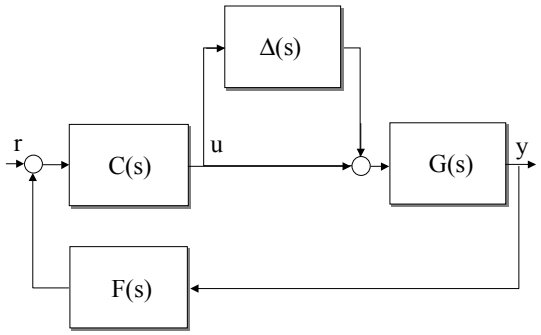

Figure 1.15: The actual closed-loop system in the case of multiplicative uncertainty.

be in a frequency range where $|\Delta(jw)|$ is small. More precisely, for large $|C(jw)F(jw)|$, $|S_a(jw)|$ tends to $1/|\hat{G}(jw)|$. Therefore, performance improvement is possible in the frequency range where $|\Delta(jw)|/|\hat{G}(jw)|$ is small (small modeling error). This can be stated simply as a rule of thumb:

- *Performance can be improved when a good model is available.*

Analogous statements can be made in the case of multiplicative uncertainty, for which the actual closed-loop system is shown in Fig. 1.15.

Loosely speaking, this type of uncertainty can be thought as an additive uncertainty $\hat{G}(s)\Delta(s)$ (modulo some technical details). As in the previous case, one can develop expressions for the actual Sensitivity and Complementary Sensitivity transfer functions in terms of their nominal values:

$$S(s) = \frac{1}{1+\hat{G}(s)[1+\Delta(s)]C(s)F(s)} = \hat{S}(s)\frac{1}{1+\Delta(s)\hat{T}(s)}$$

$$T(s) = \frac{\hat{G}(s)[1+\Delta(s)]C(s)F(s)}{1+\hat{G}(s)[1+\Delta(s)]C(s)F(s)} = [\hat{T}(s)+\Delta(s)\hat{T}(s)]\frac{1}{1+\Delta(s)\hat{T}(s)}$$

Observing that, in this case, the critical transfer function is the Complementary Sensitivity $T(s)$, we can state the following result.

1.4.11 Theorem: *Suppose that $C(s), F(s)$ internally stabilize $\hat{G}(s)$ and there exists $W(s)$ such that*

1. *$W(s)$ and $W(s)\Delta(s)$ are analytic in the open right-half plane;*
2. *$W(s)$ and $W(s)\Delta(s)$ are bounded in the closed right-half plane;*
3. *The zeros of $W(s)$ lie in the open left-half plane or at ∞;*
4. *$\gamma_2[W\Delta] \leq 1$.*

Further, suppose that [13]

$$\gamma_2[TW^{-1}] < 1.$$

Then the actual closed-loop system is internally (and BIBO) stable. Furthermore, $\gamma_2[T\Delta] \leq \gamma_2[TW^{-1}] < 1$ and

$$\gamma_2[S] \leq \frac{\gamma_2[\hat{S}]}{1-\gamma_2[TW^{-1}]} \quad ; \quad \gamma_2[T] \leq \frac{\gamma_2[\hat{T}]+\gamma_2[TW^{-1}]}{1-\gamma_2[TW^{-1}]}$$

$\triangledown\triangledown$

[13] This implies that relative degree of $W(s) \leq$ relative degree of $\hat{T}(s)$

Proof: Similar to the proof for the additive uncertainty; in this case however, notice the use of the stable minimum-phase weight $W(s)$, since $\Delta(s)$ may be improper. That is, γ_2-gains of the uncertainty are well-defined only when Δ is multiplied by either T or W. □□

The trade-off between the closed-loop stability requirement and performance improvement is quite apparent here. To ensure stability, $|T(jw)|$ must be small wherever $|W(jw)|$ is large (i.e., wherever $|\Delta(jw)|$ may be large). On the other hand, to ensure good disturbance attenuation $|S(jw)|$ must be small in the frequencies where the output disturbance has energy. Obviously, both cannot be achieved at the same frequency ($S + T = 1$), meaning simply that good output disturbance attenuation can be achieved for the frequencies for which a good model of the plant is available.

Thus, with either one of the two uncertainty descriptions, the closed-loop behavior (stability/performance) can be deduced from the properties of the nominal closed-loop system and some information on the size of the uncertainty. As a final remark, an interesting case that was not considered in the previous discussion is the effect of perturbations in the unstable modes of the plant. In our uncertainty descriptions, such perturbations would cause $\Delta(s)$ to have poles in the right-half plane and, consequently, "infinite" gain. This technical problem can be cirumvented by using somewhat more intricate analytical arguments, or in a more general manner, by employing different uncertainty descriptions (other than multiplicative and additive). Although the final results with this approach remain qualitatively the same, the details are beyond the scope of this note.