THE CROSS DIRECTIONAL CONTROL OF WEB FORMING PROCESSES

A thesis submitted by
STEPHEN RICHARD DUNCAN
for the degree of
DOCTOR OF PHILOSOPHY
in the University of London
1989

Industrial Systems Group,
Department of Electrical Engineering,
Imperial College of Science, Technology and Medicine,
Exhibition Road,
London SW7 2BZ.
ABSTRACT

A web forming process is a 2-dimensional control problem. Although schemes which control the web at a single point are common, the problem of controlling the web across the whole of its width has only recently been considered. These cross-directional control schemes consist of an array of identical actuators arranged across the web, usually at an early stage in the process (e.g., the headbox slice in a paper mill) and a device for measuring the variations across the web, positioned further downstream. A number of control schemes have been developed which are based on a multivariable lumped parameter model of the response between the actuators and a finite number of measurements across the web. In reality, the actuators have a response which is continuous in both time and space, so the system is more accurately described by a distributed parameter model. This research sets up such a model and uses it to create a general procedure for the design of control systems for this class of process. Further analysis of the model also determines:

- the range of disturbances which can be controlled by a given physical system;
- the optimal response and spacing of the actuators;
- the optimal measurement and signal processing strategy.

In addition, the thesis describes an online procedure for estimating the response of the actuators, which includes a technique for minimizing the variance of the estimate and for reducing the sensitivity of the estimate to model order.
ACKNOWLEDGEMENTS

I wish to thank Professor Greyham Bryant of the Industrial Systems Group at Imperial College for his guidance and support during my postgraduate studies. His lively interest in my work and his friendship have made this research most enjoyable.

I would like to acknowledge the help and friendship of all of the other members of the Industrial Systems Group, in particular, Constantine Goulimis, Mike Poon, Lam Yeung and Julian Allwood and also Caroline Brice, who typed most of this manuscript.

I owe a great deal to the team at ICI Wilson led by Simon Pain, and more recently by John Groom. In particular, I would like to acknowledge the contribution of one member of the team, Mike Clifford, whose thorough understanding of web forming processes and perceptive questioning led to the development of a number of the ideas described in this thesis.

Finally I would like to thank my parents for their support and enthusiasm, not only during the period of this research, but throughout my education and also my wife and daughter for their help and encouragement, particularly during the writing of the thesis.
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>3</td>
</tr>
<tr>
<td>Contents</td>
<td>4</td>
</tr>
<tr>
<td>List of Figures</td>
<td>6</td>
</tr>
<tr>
<td><strong>Chapter 1 INTRODUCTION</strong></td>
<td>9</td>
</tr>
<tr>
<td>1.1 Web Forming Processes</td>
<td>9</td>
</tr>
<tr>
<td>1.2 Analysis of the Variations of Web Properties</td>
<td>11</td>
</tr>
<tr>
<td>1.3 Practical Cross Directional Control Schemes</td>
<td>21</td>
</tr>
<tr>
<td>1.4 Aims of the Research</td>
<td>36</td>
</tr>
<tr>
<td><strong>Chapter 2 A DISTRIBUTED PARAMETER MODEL OF THE PLANT</strong></td>
<td>39</td>
</tr>
<tr>
<td>2.1 Assumptions of the Model</td>
<td>39</td>
</tr>
<tr>
<td>2.2 The Distributed Parameter Model</td>
<td>41</td>
</tr>
<tr>
<td>2.3 Sampled Data Model</td>
<td>53</td>
</tr>
<tr>
<td><strong>Chapter 3 THE SPATIAL CONTROLLABILITY OF A CROSS DIRECTIONAL CONTROL SCHEME</strong></td>
<td>58</td>
</tr>
<tr>
<td>3.1 Spatial Controllability</td>
<td>59</td>
</tr>
<tr>
<td>3.2 Choosing the Optimal Set Points</td>
<td>64</td>
</tr>
<tr>
<td>3.3 The Spatial Frequency Response of an Array of Actuators</td>
<td>68</td>
</tr>
<tr>
<td>3.4 The Design of an Array of Actuators</td>
<td>114</td>
</tr>
<tr>
<td>3.5 The Response of a Finite Array of Actuators</td>
<td>117</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2.1</td>
<td>The variations from mean level of the density of pulp in a sheet of paper</td>
<td>12</td>
</tr>
<tr>
<td>1.2.2</td>
<td>The pulp density variations of figure 1.2.1 separated into components</td>
<td>14</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Two stands of a four-high tandem mill for cold rolling of metal sheet</td>
<td>23</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Schematic diagram of a shapemeter</td>
<td>25</td>
</tr>
<tr>
<td>1.3.3</td>
<td>The change in shape due to a unit change in output of a coolant spray</td>
<td>26</td>
</tr>
<tr>
<td>1.3.4</td>
<td>The layout of a Fourdriner paper machine</td>
<td>28</td>
</tr>
<tr>
<td>1.3.5</td>
<td>The layout of blown film extrusion process</td>
<td>34</td>
</tr>
<tr>
<td>3.3.1</td>
<td>The shape of the Gaussian actuator response</td>
<td>71</td>
</tr>
<tr>
<td>3.3.2</td>
<td>A bandlimited function and its periodic extension</td>
<td>76</td>
</tr>
<tr>
<td>3.3.3</td>
<td>The frequency response of an infinite array of actuators</td>
<td>81</td>
</tr>
<tr>
<td>3.3.4</td>
<td>The product of a cosine wave and a Gaussian function with a large standard deviation, $\sigma$, in the spatial and frequency domains</td>
<td>84</td>
</tr>
<tr>
<td>3.3.5</td>
<td>The response of an infinite array of actuators, each of which has a Gaussian response, when matching cosine waves</td>
<td>88</td>
</tr>
<tr>
<td>3.3.6</td>
<td>An example of an actuator frequency response for which $B(\omega) = 0$ for some $\omega$</td>
<td>90</td>
</tr>
<tr>
<td>3.3.7</td>
<td>The effect of aliasing</td>
<td>93</td>
</tr>
<tr>
<td>3.3.8</td>
<td>The effect of aliasing in the spatial frequency domain</td>
<td>94</td>
</tr>
<tr>
<td>3.3.9</td>
<td>The effect of aliasing when matching cosine responses of different frequencies</td>
<td>95</td>
</tr>
<tr>
<td>3.3.9</td>
<td>continued</td>
<td>96</td>
</tr>
<tr>
<td>3.3.9</td>
<td>continued</td>
<td>97</td>
</tr>
<tr>
<td>3.3.10</td>
<td>The effect of aliasing when matching sine responses of different frequencies</td>
<td>99</td>
</tr>
<tr>
<td>3.3.11</td>
<td>The actuator set points for matching a cosine wave at frequency $\omega_1 = \frac{1}{2}\omega_3$</td>
<td>100</td>
</tr>
<tr>
<td>3.3.12</td>
<td>The effect of aliasing in the spatial domain</td>
<td>104</td>
</tr>
<tr>
<td>3.3.13</td>
<td>The frequency response on an infinite array of actuators with different actuator bandwidths, when matching cosine waves</td>
<td>106</td>
</tr>
</tbody>
</table>
3.3.14 The size of the actuator set points required to generate the frequency responses shown in figure 3.3.13 ................................................................. 108

3.3.15 The frequency response on an infinite array of actuators with different actuator bandwidths, when matching sine waves ........................................ 110

3.3.16 The size of the actuator set points required to generate the frequency responses shown in figure 3.3.15 ................................................................. 111

3.3.17 The frequency response on an infinite array of actuators with different spacings between actuators, when matching cosine waves ........................................ 112

3.3.18 The size of the actuator set points required to generate the frequency responses shown in figure 3.3.17 ................................................................. 113

3.4.1 The optimal shape of the response of an actuator ........................................ 116

3.5.1 The partial Fourier series of the periodic function ........................................ 119

3.5.2 The frequency response of a finite array containing 2N+1 actuators ............ 120

3.5.3 The effect of sampling a bandlimited signal ............................................. 123

3.5.4 The effect of sampling a bandlimited signal of finite extent .......................... 125

3.5.5 The frequency response on a finite array of actuators with different actuator bandwidths, when matching cosine waves ........................................ 131

3.5.6 The size of the actuator set points required to generate the frequency responses shown in figure 3.5.5 ................................................................. 132

3.5.7 The frequency response on a finite array of actuators with different actuator bandwidths, when matching sine waves ........................................ 134

3.5.8 The size of the actuator set points required to generate the frequency responses shown in figure 3.5.7 ................................................................. 135

4.1.1 The frequency response of an ideal low pass filter with bandwidth, \( \omega_c \) ........ 139

4.1.2 The frequency response of a more practical low pass filter with a nominal bandwidth of \( \omega_c \) and an cut-off frequency at \( \omega_c' \) ........................................... 140

4.1.3 The aliasing problem ............................................................................. 143

4.1.4 Block averaging ................................................................................... 144
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1.5</td>
<td>The effect of block averaging</td>
<td>145</td>
</tr>
<tr>
<td>4.1.6</td>
<td>Block averaging in the spatial domain</td>
<td>147</td>
</tr>
<tr>
<td>4.1.7</td>
<td>The cross direction and machine direction components are indistinguishable in the profile signal from a scanning gauge</td>
<td>149</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Block diagram of the feedback control system</td>
<td>160</td>
</tr>
<tr>
<td>4.3.2</td>
<td>The open loop transfer function between the error signal and the response of the actuators</td>
<td>161</td>
</tr>
<tr>
<td>4.3.3</td>
<td>The closed loop return difference of the system illustrated in figure 4.3.1</td>
<td>163</td>
</tr>
<tr>
<td>4.5.1</td>
<td>An example of a true actuator frequency response and a nominal frequency response</td>
<td>178</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Block diagram representing the failure of the rth actuator</td>
<td>189</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Block diagram representing the case where the set point of the rth actuator reaches its physical limits</td>
<td>192</td>
</tr>
<tr>
<td>5.1.1</td>
<td>The addition of a perturbation to the actuator set points for the purpose of actuator response estimation</td>
<td>208</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

1.1 Web Forming Processes

Web forming is a term used to describe a collection of industrial processes which produce a flat expanse of material whose thickness is significantly less than its length or width. These expanses of material are known collectively as webs. In addition, there is a transport system which moves the web through the process in a continuous or semi-continuous mode. The materials that are most commonly formed into webs include:

- metals;
- paper;
- plastics.

Depending upon the particular industrial process, the webs are referred to as sheets, films, plates or foils and in this study, we will use all of these terms synonymously.

We will be interested in the control of properties of the web such as thickness, tension, moisture, etc. Because these properties can vary over the whole extent of the web, the control problem is two dimensional. In early control systems, variations in web properties were measured by a fixed gauge which looked at only the small fraction of the web which passed under it. Controllers designed to maintain this measured value of the property at a predetermined level are now well established (see for example, [Bryant, 1973], [Dumont, 1986]). However, such systems measure only a fraction of the web, so the rest of it remained effectively uncontrolled. More recently, the demand for high quality materials has led to the development of cross directional control schemes (this term will be defined more fully in the next section) where the properties of the webs are measured by allowing the gauge to scan across the sheet or by having an array of measuring devices positioned across the the web in a line perpendicular to the direction of its travel. These two measurement techniques give some indication of the variations of the properties over the whole web (although neither technique covers the web completely). A cross directional
control scheme uses these measurements in combination with an actuator for changing the
property over the width of the web. There are a number of different forms of actuator
which are used in cross directional control, but in this study, we will concentrate on
actuators which consist of an array of identical, individual actuators arranged in a single
line across the web. This is the most common form of actuator for high resolution cross
directional control.

Although the number of cross directional control schemes that have been implemented has
increased dramatically in the last decade, particularly in the paper and board industry
[Smith, 1985] the literature on the subject is remarkably sparse. Most references simply
describe the implementation of a control scheme for a specific application. Apart from the
introductory work of Wilhelm [Wilhelm, 1984] and the computer simulation work of
Lindeborg [Lindeborg, 1986] there appears to have been no attempt to model the
underlying process and to use this model to design a better control scheme. Wilhelm is
forced to conclude that:

"At present there is no single method of analysis which provides the breadth of
insight needed to properly engineer a [cross directional] control algorithm and predict
its results" [Wilhelm, 1984]

This study develops a model of the process and a method of analysis which overcomes this
problem and leads not only to the design of a general optimal control algorithm but also to a
specification for the design of the plant (ie a specification for the design of actuators and
measuring devices). In particular, the innovative ideas that have come out of this work
include:

• using a distributed parameter approach to the modelling of the system;
• analysing the spatial controllability of the plant in terms of the Fourier
components of the cross directional variations;
• developing a design specification for the optimal shape of the actuator's spatial
response; the correct actuator separation and the best filtering and sampling
strategy for the measurement system;
• creating a multivariable lumped parameter model of the system which completely specifies the response of the system;
• developing a procedure for designing a controller for this model, which reduces the multivariable design problem to a single-input single-output design;
• investigating the robustness, integrity and implementation of the controller;
• designing an algorithm for estimating the spatial response of the actuators while the plant is running under closed loop control.

Taken together, these ideas make a major advance in the systematic design of cross directional control schemes.

The next section of this chapter examines the cross directional control problem more rigorously and sets up a framework for the rest of the study. Section 1.3 describes some industrial web forming processes in more detail, with particular reference to cross directional control and also provides a survey of the literature that applies to each process. Section 1.4 outlines the contents of the other chapters in the thesis.

1.2 Analysis of the Variations of Web Properties

The previous section has identified a number materials that can be formed into webs. As mentioned above, because variations can occur over any part of the sheet, controlling the web properties problem is a two dimensional problem. This is illustrated in figure 1.2.1 which shows the variations of the density of pulp (known as basis weight, as described in the next section) from the mean level in a sheet of paper which is produced by a machine with no control (from Wilhelm, 1984).

In order to specify the variation of the web property at a particular point we will define Cartesian coordinates in two directions. These directions are referred to as:

(i) the machine direction (MD) which is defined along the direction of movement
Figure 1.2.1 The variations from mean level of the density of pulp in a sheet of paper produced by an uncontrolled paper machine. The variations are plotted vertically against the cross direction (CD) and machine direction (MD) axes of the sheet (from [Wilhelm, 1984]).
of the web;

(ii) the cross direction (CD) which is perpendicular to the direction of movement of the sheet;

and are marked on figure 1.2.1.

The value of the web property at any point in the web will be written as

\[ P(x,t) \]  

1.2.1

where \( x \) is the coordinate in the cross direction and \( t \) (time) is the coordinate in the machine direction. We have used time as the machine direction coordinate because a fixed measuring device will see variations in the machine direction evolving in time as the web moves through the plant.

It is convenient to express the value of the web property, \( P(x,t) \), in four components (following the approach of [Wilhelm, 1984])

\[ P(x,t) = P_{\text{avg}} + P_{CD}(x) + P_{MD}(t) + P_{R}(x,t) \]  

1.2.2

where

- \( P_{\text{avg}} \) = the average value over the whole web;
- \( P_{CD}(x) \) = the pure cross directional component (which is a function of \( x \) only);
- \( P_{MD}(t) \) = the pure machine direction component (which is a function of \( t \) only);
- \( P_{R}(x,t) \) = the residual component.

Because we have included the \( P_{\text{avg}} \) term in equation 1.2.2, we can, without loss of generality, assume that mean level over the whole of the web of each of the other three components in the equation, is equal to zero. This particular decomposition is not unique, but it has been chosen because it simplifies the definition of cross directional and machine directional actuators, as will be seen below. The components of \( P(x,t) \) corresponding to the web in figure 1.2.1 are illustrated individually in figure 1.2.2. Note that the web variations in figure 1.2.1 are plotted relative to the mean level of the web so the \( P_{\text{avg}} \) component has been removed from the figure to improve clarity.
Figure 1.2.2 The pulp density variations of figure 1.2.1 separated into the machine direction component, $P_{MD}$, cross direction component, $P_{CD}$ and random component, $P_R$ (from [Wilhelm, 1984]).
It is important to stress that \( P(x,t) \) describes the value of the web property over an uncontrolled web. We will have some target value of this web property for the whole web and usually, this value will be constant over the whole web, so we will refer to it as a constant \( P_{\text{target}} \), although the analysis can be readily extended to include a target value which is a function of position. The aim of any control system is to make the value of the web property equal to this target value \( P_{\text{target}} \) and this is equivalent to ensuring that the average value of the property is equal to \( P_{\text{target}} \), while simultaneously removing the cross directional, machine directional and residual components as far as possible. Any deviations of the uncontrolled web property from the desired value are given by

\[
P(x,t) - P_{\text{target}} = (P_{\text{avg}} - P_{\text{target}}) + P_{\text{CD}}(x) + P_{\text{MD}}(t) + P_{\text{R}}(x,t)
\]

1.2.3

The first term can be viewed as an offset in the mean level of the web property. The second term, \( P_{\text{CD}}(x) \), usually arises from an asymmetry of the plant in the cross direction, for example, an incorrectly aligned headbox slice in a paper machine (as described in more detail in section 1.3). Because this term is independent of time we can consider it as a steady state disturbance. It can be seen from figure 1.2.2 that \( P_{\text{MD}}(t) \) has the same value across the width of the web. A typical cause of this type of disturbance would be an eccentric roll in a steel hot-mill which would create thickness variations across the whole web [Bryant, 1973].

As pointed out by Wilhelm [Wilhelm, 1984], it is a common misconception that the residual component, \( P_{\text{R}}(x,t) \), is completely uncontrollable. Although the residual component in figure 1.2.2 appears to contain a lot of high frequency components, it is likely that it also contains some components which are varying slowly in either the cross direction, the machine direction or both. Usually these high frequency components will be beyond the bandwidth of the controller, but it is possible that some of the low frequency components can be removed.

As mentioned in the previous section, early web control systems measured a property at a fixed point in the cross direction. An example of this would be a gauge which sits at a
fixed point across the web and measures the profile as the web passes underneath it. If the gauge is positioned at point $x = x_1$ in the cross direction, then this gauge measures

$$P(x_1,t) = P_{\text{avg}} + P_{\text{CD}}(x_1) + P_{\text{MD}}(t) + P_{\text{R}}(x_1,t)$$

1.2.3

Because the cross directional component, $P_{\text{CD}}(x_1)$ is independent of time, the gauge sees variations in only the pure machine direction component and the residual component at $x = x_1$. Schemes designed to control these variations are commonplace (see [Bryant, 1973] [Dumont, 1986]). These schemes attempt to control $P(x_1,t)$ to the nominal value $P_{\text{target}}$ by removing the machine direction and residual variations and correcting the fixed value of the property which is determined by $P_{\text{avg}}$ and $P_{\text{CD}}(x_1)$. Note that if $P_{\text{CD}}(x_1)$ is not zero (for example, if there is a "high spot" in the web at $x = x_1$) then the control of the mean level of the property of the web will be offset. This can have serious consequences because the web may not achieve the required specification (for example, in metal rolling, the mean thickness gauge over the whole sheet may be too large which means that excess metal has been used, resulting in an increased cost of production). A characteristic of all of these control schemes is that they apply changes to the web which are purely machine directional, ie they change the web by an equal amount across its whole width and they do not affect either the cross direction or residual components. For this reason, we will refer to these schemes as machine directional controllers.

The problem with this approach is that the gauge looks at only a small fraction of web and leaves the rest uncontrolled. In this study we aim to control the variations in web properties via an array of individual, identical actuators positioned in a line across the web (ie in the cross direction). A description of practical actuators of this form will be given in the next section, but for the moment, we will assume that each actuator can affect the given web property in both the cross and machine directions. The cross directional effect of an actuator will be referred to as its spatial response, while the machine directional response will be referred to as the time response. We will also assume that the responses generated by the actuators are linear, so that response of the whole array is equal to the sum of the responses of the individual actuators. If a two dimensional measure of the variations which
is continuous in both the cross and machine directions were available, then it would be possible to control all three of the components of the variations in equation 1.2.2 (ie \(P_{\text{CD}}(x), P_{\text{MD}}(t)\) and \(P_{\text{R}}(x,t)\)). The fraction of each component that can be controlled is determined by the spatial and time response of the actuators. (For example, we would not expect that a very narrow cross directional disturbance could be controlled by an array containing actuators which have a wide spatial response). The concept of controllability, particularly spatial controllability, will play an important role in this study (see chapter 3).

In reality, it is not possible to obtain a continuous two dimensional measurement, so practical systems have been restricted to controlling variations of the sheet from measurements which have been sampled in either the cross direction or the machine direction or both. We will start by taking continuous samples across the web at fixed time intervals. It will be seen later that this procedure is closely related to practical measurement systems. If the samples are taken at times \(t = T, 2T, 3T, \ldots\) (where \(T\) is the interval between samples) then the value of the web property that is measured is given by

\[
P(x,nT) = P_{\text{avg}} + P_{\text{CD}}(x) + P_{\text{MD}}(nT) + P_{\text{R}}(x,nT)
\]

We now choose to measure the web property relative to the mean level of the sample, \(P(nT)\) (for reasons that will become clear below) and we will define a new measurement, \(\Delta P(x,nT)\), as

\[
\Delta P(x,nT) = P(x,nT) - \bar{P}(nT)
\]

where \(\bar{P}(nT)\) is the average of equation 1.2.4 taken over the \(x\) dimension and is given by

\[
\bar{P}(nT) = P_{\text{avg}} + \bar{P}_{\text{CD}} + \bar{P}_{\text{MD}}(nT) + \bar{P}_{\text{R}}(nT)
\]

It is important to note that \(\bar{P}(nT)\) is the average of the cross directional sample taken at time \(t=nT\), which is not the same as the mean level of the property over the whole web, \(P_{\text{avg}}\). However, by definition above (see equation 1.2.2), the machine direction component is constant over the width of the web, so \(\bar{P}_{\text{MD}}(nT)\) is equal to \(P_{\text{MD}}(nT)\). Further, we have assumed that the cross directional component has zero mean over the whole web, so \(\bar{P}_{\text{CD}}\) is equal to zero. The mean level of the residual component across the web at one particular
instant need not be zero, even though its average value is zero over all of the web, but, without loss of generality, we can consider $PR(nT)$ as a constant offset across the web and incorporate it into the machine direction component, $P_{MD}(nT)$. This allows us to take $PR(nT)$ as zero and equation 1.2.6 reduces to

$$P(nT) = P_{avg} + P_{MD}(nT)$$

1.2.7

This means that the reading taken relative to $P(nT)$ becomes

$$\Delta P(x,nT) = P(x,nT) - \bar{P}(nT)$$

1.2.8

$$= P_{avg} + P_{CD}(x) + P_{MD}(nT) + P_{R}(x,nT) - P_{avg} - P_{MD}(nT)$$

1.2.9

$$= P_{CD}(x) + P_{R}(x,nT)$$

1.2.10

so by taking the measurement relative to the mean level of the property across the web, we are left with only the cross directional and residual components.

We require an actuator for controlling these two components of the web variations and in this study, we will restrict our attention to the array of identical individual actuators, that was described above. Although other forms of actuators are feasible (see, for example, the control of thickness in a hot metal mill through the use of jacks and screws, as described in the next section) these actuators tend to have a lower spatial controllable bandwidth than an actuator array, so they will not be considered here. It will become apparent that we need the actuator to redistribute the property across the web without affecting either the mean level or the machine direction component. Although this may seem a very restrictive requirement, in practice, many arrays of actuators operate in this manner. For example, in a paper machine, liquid pulp is projected from a container known as the headbox, through a longitudinal slit. The local width of the slit is adjusted by bending one of the sides of the slit (referred to as the slice) via an array of screws (This process will be described in more detail in the next section). If a single screw is adjusted so that the amount of pulp (referred to as basis weight) is increased adjacent to the screw, its effect is to draw pulp from the regions on either side of screw, reducing the basis weight of the web in these positions [Wilhelm, 1984]. This means that, even though the local change of basis weight may be
large, the net change over the whole web is very small.

Unfortunately, it is not always possible to guarantee that the net effect across the web will be exactly zero, but we can always ensure that the average of the set points of the actuators in the array is constant, so that if the set point of one actuator is increased, the set points of all of the other actuators must be decreased slightly in order to keep the mean set point level the same. Fixing the set point average ensures that the effect of any control action is to redistribute the property in the cross direction, without producing any effect in the machine direction. Using the example above, if the average of the set points of the headbox screws remain constant, the mean headbox gap remains the same at all control actions, so the net flow of pulp will be unchanged. Further, if the dynamics of the mechanisms for adjusting the screws are all the same, then the mean headbox gap will not be affected by the transient response of the control actions, as will be shown in section 2.3. By adjusting the actuator set points so that the net flow of material is constant, we can ensure that $P_{\text{avg}}$ or $P_{\text{MD}(t)}$ are not affected by the control actions of the array.

We can now see that by:

i) measuring the variations of the web relative to the mean level across the web at each sample;

ii) using an array of actuators with a fixed average level of the set points, to redistribute the property across the web;

we only measure and control the cross directional and residual components, $P_{\text{CD}}(x)$ and $P_{\text{R}}(x,nT)$, at each sample. If the array of actuators is used to regulate both of these components as close as possible to zero, then the cross directional and residual variations of the web property can be reduced to a minimum. For the purposes of this study, we will define such a regulator as a cross directional controller (although this is a slight misnomer because both the residual and cross directional variations are regulated).

We can see that a cross directional controller is not influenced by the effects of a machine
direction controller. This is because machine direction controllers can change only $P_{\text{avg}}$ and $P_{\text{MD}}(t)$, neither of which contributes to $\Delta P(x,nT)$, the measurement used by the cross directional controller. Also, the fixed gauge of the machine directional control scheme will measure (from equation 1.2.3)

$$P(x_1,t) = P_{\text{avg}} + P_{\text{CD}}(x_1) + P_{\text{MD}}(t) + P_{R}(x_1,t)$$ 1.2.11

where $x=x_1$ is the position of the gauge. If a cross directional controller is in operation, both $P_{\text{CD}}(x_1)$ and $P_{R}(x_1,t)$ will be as close as possible to zero (within the bandwidth of the spatial and temporal response of the actuator array). This means that the machine directional control scheme is reduced to measuring and regulating $P_{\text{avg}}$ and $P_{\text{MD}}(t)$. If this controller removes the machine direction variations as far as possible within the bandwidth of its response and regulates the average level to the target value, $P_{\text{target}}$, then with the cross directional controller removing the cross directional and residual components, the web property will be as close as possible to $P_{\text{target}}$ over the whole extent of the web.

We have now taken the two dimensional control problem of regulating a property of the web and split it into two separate, independent controllers, each working in only one dimension:

i) a cross directional regulator which takes measurements relative to the instantaneous mean level across the web and an array of actuators which redistributes the property across the web, without affecting its mean level;

ii) a machine direction controller which measures the web from a fixed gauge and applies control actions which have the same effect across the width of the web.

We were able to make this simplification because we sampled the web property at fixed time intervals. The price that is paid for the simplification is that we have no control over the web in between samples. However, by relating the sample time to the time response of the actuator array, we can ensure that any variations between samples are outside the
temporal controllable bandwidth, so the sampling does not degrade the performance of the system. Sampling the web means that we have taken a system which had continuous variations in both directions and reduced it to a system which is continuous in the cross direction but discrete in the machine direction. In practice, it is not possible to handle continuous, cross directional measurements so samples are taken across the web as well, reducing the system to one which is discrete in both directions. As with the machine direction samples, it is important that the cross direction samples are taken at intervals which are related to the response of the actuator array (see section 4.1).

It was mentioned earlier that machine direction control schemes are well established on web forming processes, so in the rest of this study, will assume that such a system is in place on the plant and concentrate on the design of the cross directional control system.

1.3 Practical Cross Directional Control Schemes

Section 1.1 mentioned a number of materials which can be formed into webs (eg metals, plastics paper etc). It is impossible, within the confines of this study, to give a complete description of all web forming processes. Instead, this section gives a brief description of a number of these processes and then concentrates on the cross directional control schemes which are implemented on these plants. In each case we will focus on schemes which consist of an array of identical, individual actuators combined with a device for measuring variations of a property across the sheet.

Metals

We start by describing the production of metal sheet. The most common metals that are formed into sheets are steel and aluminium, although there are also plants using other materials such as copper and brass. Metal sheet is produced by rolling, where a slab or
thick sheet of metal is passed between two rollers whose separation is less than the thickness of the incoming metal with the gap between the rolls being maintained by applying compressive forces which hold the rolls together. It is usual to reduce the thickness of the metal in a number of stages by passing it through a succession of rolls (known as stands) until the desired thickness is achieved.

There are two types of rolling process: hot rolling, where the metal is preheated before rolling and cold rolling where the metal is rolled without any heating. Hot rolling usually starts by taking a heated slab of metal and reducing its thickness before coiling the metal. Cold rolling takes the product of a hot mill and reduces the thickness still further. The deformation of the metal that occurs in the gap between the two rollers (referred to as the roll gap) is the key to the uniformity of the rolled metal, but the mechanics of this deformation are fundamentally different for the two processes. For this reason, we will consider them separately.

**Cold Rolling.** In cold rolling, the reduction in thickness of the metal is caused by the high compressive stresses in the region of contact between the metal and the rollers (referred to as work rolls). This induces very large stresses in the rollers which causes them to bend along their length. In order to reduce the deflection of the work rolls, they are held between two larger rolls (the backup rolls) in a substantial rectangular frame as in figure 1.3.1. The necessary compressive forces which are required to deform the metal are applied to the top or bottom backup rolls by screws or rams. In addition to the bending, the work rolls are also deformed as a result of changes in diameter caused by thermal expansion and roll compression. The thermal expansion is due to the work rolls absorbing some of the heat that is generated during the reduction of the thickness of the metal, while the roll compression is caused by the reaction of the metal on the roller. The thermal effects and the compression are unlikely to be uniform across the roll gap and when combined with the work roll bending, there will be a cross directional variation in the width of the roll gap. If the profile of the incoming strip does not match the profile of the roll gap, the
Figure 1.3.1 Two stands of a four-high tandem mill for cold rolling of metal sheet (from [Bryant, Edwards and McClure, 1973]).
reduction of the metal will not be uniform across the strip and this will eventually produce
distortions in the strip. Initially, these distortions will be absorbed by the internal tensions
that develop in the strip. The cross directional distribution of tension is known as the
"shape" of the metal and these tensions remain in the strip after rolling.

There are a number of methods for measuring the shape of a sheet. Most of the modern
methods are based on the principle of a segmented roller which is placed in the path of the
sheet in a position which causes the sheet to be deflected (see figure 1.3.2). The deflection
of each segment or the force required to hold a segment in a fixed position is related to the
tension in the section of the strip which is passing over this segment [Roberts, 1978].

Given a measurement of the shape variation, it is possible to bend and tilt the rolls to
generate a roll gap which matches the profile of the incoming strip. A number of
successful control schemes have been designed on this basis [Spooner and Bryant, 1973]
[Ringwood and Grimble, 1986] [Roberts, 1978]. The problem with using these
"mechanical adjustments" to the roll gap is that they can induce only low frequency changes
across the strip (for example a tilt on the work roll produces a wedge variation while
bending the roll produce a variation which has a parabolic form). A more interesting
approach has been suggested by Bryant and Spooner [Bryant and Spooner, 1973]. They
point out that coolant sprays on the work rolls can affect the cross directional temperature
variations (referred to as thermal camber) of the rolls which in turn, changes the roll gap in
the cross direction. The response of a spray is illustrated in figure 1.3.3 which shows the
effect on the shape of the metal when a unit change is applied to the level of a spray. It can
be seen that the spray has a response which is continuous across the strip and also overlaps
the response of the adjacent sprays (shown as the dashed lines in the figure). Using an
array of sprays arranged across the work roll can generate quite complex responses which
gives the potential of a much wider spatial bandwidth for the control of shape compared to
the tilting and bending of rolls.
Figure 1.3.2 Schematic diagram of the shapemeter for measuring the cross directional tension variations in a sheet of metal (from [Roberts, 1978]).
Figure 1.3.3 The solid line shows the change in shape due to a unit change in the output of a coolant spray on the work roll. The dashed lines show the responses due to the adjacent sprays across the roll.
Hot Rolling. In hot rolling, although a mismatch between the profile of the incoming strip and the roll gap profile will induce cross directional tension variations in the strip, the strip tends to relax between successive stands of rolls, so the tension variations do not remain in the strip after rolling. There may be some sideways flow of material in the roll gap so that if the profile of the incoming strip is different from the profile of the roll gap, the thickness of the strip will be changed in the cross direction [Larke, 1963]. In order to produce sheet of uniform thickness, it is necessary to be able to change the roll gap profile which can be done by bending the rolls or by adjusting the thermal camber using coolant sprays as in cold rolling.

It is also important to control the cross directional distribution of temperature during hot rolling because differences in temperature result in non-uniform material properties across the sheet. The sheet temperature can be controlled by a bank of coolant sprays on the sheet, but unfortunately, the sheet profile and the sheet temperature interact via the thermal camber of the rolls, so the control of profile and the control of temperature cannot be considered independently.

Paper

Paper is formed by combining cellulose fibres (usually from plants) to form sheets. This is done by immersing the fibres in water and fibrillating them (which can be thought of as bruising or crushing) so that the fibre walls retain increasing amounts of water. This causes the fibres to swell and to become gelatinous. The bruising process roughens the fibres so that they form a stronger bond when brought into contact with each other. If the swollen fibres are deposited as a layer of pulp onto a sieve and the water content removed the gelatinised surfaces of the fibres act as a cement and form a sheet of paper.

This is the basic principle of papermaking which converts individual fibres into sheets.
Figure 1.3.4 The layout of a Fourdriner paper machine (from [Thomas, 1977]).
However, before this can take place, the raw material of the process (which is usually wood in Europe and North America) must be broken down to release the cellulose fibres. This process is known as pulping. After felling, the trees are cut into logs and their bark is removed and the debarked logs are then chipped into small fragments. The wood fragments are cooked in a chemical solution to dissolve the lignin (the material that binds the fibres together) creating a pulp. The pulp is then cleaned and bleached before being fibrillated. The degree of fibrillation determines the ultimate strength of the paper (for example, blotting paper is fibrillated for 30 minutes, while bank notes are fibrillated for 8 hours). The pulp is now ready to be made into a sheet of paper. Basically this is a simple process involving three steps: rolling out the pulp, squeezing out the water and drying the finished product. Although straightforward in principle, a paper-making machine is generally a large, complicated and very expensive piece of equipment.

Diagram 1.3.4 shows the major parts of a Fourdrinier paper machine, the most common type of machine in use. Papermaking begins at the wet end in the head box, when the pulp is projected onto a travelling "wire" under controlled conditions of speed and quantity. This "wire" is an endless travelling belt of wire mesh which is vibrated from side to side. This vibration has the effect of removing much of the water from the pulp as well as assisting in the intermeshing of the fibres. Most modern headboxes are pressurised, meaning that the air pad between the surface of the stock and the top of the headbox is kept under pressure. The pulp is projected onto the wire through a cross directional slit known as the slice and the width of the slice and the total head of pulp (ie the pressure of the air pad plus the stock level) determine the volume flow rate of pulp onto the wire. Suction boxes are placed under the wire belt to remove more moisture. At the end of the wire, the partially dried mass of pulp is lifted onto travelling belts of felt. Passing the pulp through pressure rolls while it is on the felts further reduces the moisture content.

The dry end of the process starts with a drying chamber where the pulp moves over a series of very large, rotating, heated cylinders. These cylinders remove most of the
remaining water. Near the end of the drying chamber the paper is sometimes passed through a bath of size which gives it a high gloss finish. After the drying chamber, the paper is passed through a series of calender rolls. These heated and pressured rolls have an effect similar to ironing the paper and also reduces the thickness of the final product.

All stages of papermaking, from pulping through to finishing, require a high level of control (see [Dumont, 1986] and [McGill, 1980]). We consider the cross directional control of three properties:

(i) basis weight;
(ii) moisture;
(iii) caliper.

Basis weight (which is also known as grammage or substance) is the weight per unit area of the finished paper. There is a technical problem with measuring basis weight because paper is hygroscopic (i.e., dry paper absorbs moisture) so the value of the basis weight depends on the moisture in the surrounding atmosphere. [McGill, 1980]. Usually the measurements are provided by a scanning absorption gauge (using either β or infrared radiation) which identifies the local cross directional variations in the density of the fibres in the pulp.

Cross directional variations in basis weight can be controlled by adjusting the gap at the headbox. This is done by bending the slice lip using a set of screws arranged across the headbox. Tightening one of the screws deforms the slice lip which locally narrows the gap through which the pulp is passed. This has the effect of redistributing pulp away from the point of action of the screw. The set of screws represent an array of actuators and originally the screws were operated manually, but recently there have been a number of schemes reported which have closed the loop between the scanning gauge and the screws by using automatic control ([Chen et al, 1986], [Wilkinson and Hering, 1983], [Siler, 1984], [McFarlin, 1983], [Walker, 1987], [Karlsson et al, 1985], [Wahren, 1986] [Wilhelm and Fjeld, 1983], [Goumand et al, 1982], [Hansson et al, 1982]). Improvements of up to 60%
in the cross direction variations of basis weight have been achieved [Wilkinson and Hering, 1983].

Controlling the moisture content of the paper is important because the amount of water in the paper affects sheet properties in particular the strength and stiffness of the final sheet. In addition, cross directional variations in the moisture content of the sheet can generate unequal dimensional stability across the sheet because the drying process causes differential shrinkage. [McGill, 1980].

There are a number of methods for measuring the cross directional moisture variations including measuring differences in resistance or capacitance of the sheet. Modern measuring devices tend to use infrared absorption techniques to gauge the water content of the sheet. Using infrared absorption also allows the measurement of moisture content and basis weight to be combined. This is done by alternating the infrared beam between a frequency that is absorbed by the water and a frequency that is absorbed by the "solid" pulp and then deducing the ratio of pulp to water from the difference between the two levels of absorption. This form of "dual" gauge is usually mounted on a scanning frame [McGill, 1980], [Siler, 1984].

The moisture content of the sheet can be changed either by drying or by rewetting [McFarlin, 1983]. Drying is applied by a bank of infrared heaters arranged across the sheet and applying heat through an individual heater will create a local change in the moisture content [Roth, 1986], [Lindeborg, 1986], [McFarlin, 1983], [Higham, 1984]. The sheet can be rewet locally by a bank of water sprays [McFarlin, 1983], [Edholm and Johansson, 1982], [McGill, 1980]. These sprays have to be positioned below the sheet so that excess water does not sit on the sheet and damage the surface. Whether heaters or sprays are used, both systems can be thought of as an array of identical actuators arranged across the sheet.
The caliper of a sheet is its thickness. For quality reasons it is important to have a uniform caliper over the whole sheet. The caliper is controlled by the calendering rolls (see figure 1.3.4). In general, the higher the pressure between the rolls the thinner the sheet. The sheet caliper can be changed locally by varying the diameter of one of the calendering rolls through thermal effects [McFarlin, 1983], [Fjeld and Hickey, 1981], [Grott, 1984], [Haverinen, 1983], [Hilden and Randle, 1984]. This is done using an array of air showers positioned along the length of the roll. Applying air via one of the showers heats or cools a small area of the roll producing a local change in the diameter of the roll and hence affects the sheet caliper. This is very similar to the control of metal sheet thickness by spraying the rolls, as described above.

The caliper variations are usually measured using a scanning β radiation absorption gauge, although this does require that the basis weight is constant over the width of the sheet, otherwise the gauge cannot distinguish between caliper variations and basis weight variations [McGill, 1980].

Plastics

There are a number of polymers which are commonly manufactured into films or sheets. These include [Briston, 1974]:

- polyethylene (polythene);
- polypropylene
- polyvinyl chloride (PVC)
- polystyrene;
- polyester.

If the thickness of the polymer is less than 250µm then the web is referred to as film, while if the thickness is greater than 250µm, it is known as a sheet. For the purposes of this study, this distinction will not be made and the terms film and sheet will be used
Plastic film is produced by plasticising the raw material (which is usually in either granule or powder form) and then forcing the plasticised product through a die. The plastic then needs to be cooled so that it solidifies to form a sheet. The process of plasticising the material and forcing it through a die is known as extrusion and there are two forms of this process:

(i) slit die extrusion, where the plasticised polymer is forced through a linear die gap to produce flat film;

(ii) blow extrusion, where the polymer is forced through an annular die resulting in tubular film.

The method of plasticising the polymer is common to both processes. The design of extrusion systems to deliver polymer at the die which is uniform in temperature, plasticity, flow rate etc. is extremely complex (see [Michaeli, 1984]) and will not be considered in this study. We will assume that there is a uniform flow of polymer through the die.

Following extrusion through a slit die, the polymer is solidified by passing it around a chilled roller (or rollers). At this stage, the polymer is in the form of a film, but there is usually some extra processing applied, for example, coating the surface of the film or orienting the film by stretching while applying heat. This latter process helps to improve the strength and barrier properties of the film.

In blown film extrusion, the annular die produces a tube of polymer which is then inflated with compressed air into a bubble (see figure 1.3.5). The pinch rolls usually draw off the film at a faster rate than it is being fed through the die. This drawing combined with the increase in diameter resulting from the inflation, has the effect of orienting the film in both directions so that, unlike slit die extrusion, it is not necessary to post process the film to provide the required orientation. Having passed through the pinch rollers, the tube of film
Figure 1.3.5  The layout of blown film extrusion process (from [Briston, 1983]).
will have been flattened and the edges can be trimmed off to produce two sheets which are wound up separately. Although figure 1.3.5 shows the bubble being blown upwards, it is possible for the bubble to be blown in other directions (eg downwards). The bubble pressure is maintained by the pinch rolls at the top of the process.

The cross directional control of plastic film is restricted to the control of thickness (cross directional thickness variations are also referred to as profile variations). The thickness of the film is usually measured by a scanning absorption gauge which uses either infrared radiation or X-rays. In early cross directional control schemes, thickness variations were corrected by making local adjustments to the width of the die gap (this applies to both slit die extrusion and blown film extrusion). These adjustments were made by bending a choker bar which was placed in the orifice of the die. Bending this bar effectively changed the die gap and locally restricted the flow of the plastic. The choker bar was bent through the use of an array of bolts which were either mechanically screwed in against the bar, or heated so that the bolts expanded, causing the bar to bend [Rudd, 1978],[Michaeli, 1984].

More recently, the thickness variations have been controlled by direct thermal effects. An array of heaters is attached to the die, so that the end of the heaters can apply heat to a local area of the die. Raising the temperature of the die lip in the region around a heater decreases the viscosity of the polymer to flow more easily in this region and to produce a local thickening of the material [Fulton et al, 1982], [Michaeli, 1984], [Feistkorn and Sensen, 1987], [Feistkorn and Sensen, 1988].

For a slit die, we can consider the cross directional thickness controller as an array of actuators, whether the change in thickness is made via a set of bolts which change the die gap width or through the use of thermal effects from an array of heaters. For the purpose of this study we will restrict our attention to such a linear array of actuators. However, an annular die in a blown film extrusion process can also be considered as an array of actuators, the only difference being that the array is circular in this case (rather than linear).
This difference is important because it means that the die is continuous so that there are no edge effects. This will be commented upon at greater length in chapter 3.

1.4 Aims of the Research

The previous section has given a brief description of some practical cross directional control schemes for a variety of properties in different materials. As mentioned earlier (section 1.2), in this study we will assume that a machine direction control system is in place and we will concentrate on the design of cross directional control schemes. In particular, we will focus on systems which consist of an array of identical actuators arranged across the strip and a measurement technique which takes measurements at a finite number of points across the strip (either by sampling a continuous scanning measurement or from a fixed array of measuring devices). From the previous section it can be seen that this is a common arrangement for practical cross directional control schemes.

In order to simplify the discussion, the analysis of the study will be stated in terms of a system for controlling the thickness of the web. This allows us to describe a specific variable (thickness) rather than the more abstract concept of a variation in web property. It should be emphasised that this restriction is applied for simplicity of exposition and that the analysis is applicable to any of the variables described in section 1.3 (eg thickness, tension, temperature in metals, basis weight, moisture, caliper in paper, thickness in plastics etc).

Because practical cross directional control schemes consist of a finite number of actuators and measuring points, this has lead to the development of a simple multivariable lumped parameter model of response of the actuator. As described in the previous section, this approach has been used as the basis of a number of successful control designs, particularly in the paper industry. While these approaches undoubtedly yield systems which work in
practice, as was mentioned in section 1.1, there does not exist any background analysis which can be applied to the performance and optimal design of the systems. As an example, given an array of actuators with a known response, there is no reported analysis of the spatial bandwidth of cross directional web variations that can be controlled. Without this knowledge it is impossible to determine whether a particular design of actuator array will be able to achieve a given specification on the cross directional variations of the given web property. The converse question arises in plant design. If there is a required specification for the cross directional variations, how should we design:

(i) the shape of the response of the actuators;
(ii) the separation between the actuators;
(iii) the spacing between the measurements;

in order to ensure that this specification is met?

In this study we do not start from the multivariable lumped parameter model of the response of the actuator array, but instead develop a distributed parameter model of the response. Because the true response of the array is continuous in both space and time (see section 1.2) this model gives a fundamental description of the response. This model is developed in chapter 2.

In chapter 3, this model is then used to determine the cross directional (or spatial) controllability of the array. This is done by analysing the Fourier components of the cross directional variations and by building up a spatial frequency response of the actuator array from the response of a single actuator. We can then differentiate between the spatial components of a particular variation which can be controlled and those which cannot. From this analysis, we can determine the optimal shape of the response of the actuators and the correct spacing between actuators.

Once the frequency range of spatially controllable components has been determined, we can develop a method for processing the cross directional measurement of the web variations as
well as determining the optimal spacing between the discrete measurements. We now have
the optimal shape of actuator response, the optimal separation of actuators and measuring
points so we can determine an "optimal", multivariable, lumped parameter model
describing the response between the finite number of actuators and the finite number of
measuring points. Chapter 4 describes a general design procedure for a class of systems
described by this model. The procedure reduces the multivariable design problem to a
single input - single output design. The controller designs reported in section 1.3 can be
considered as special cases of this general design procedure. The "robustness" to errors in
the model of the plant response, the performance of the system when an actuator fails and
the implementation of the controller are also considered in some detail.

It will be seen that the design of the control system requires a description of the spatial
response of the actuators. In the past this has been obtained from a physical knowledge of
the plant or from offline experiments, because it has been considered too difficult to
estimate the responses of the actuators while control system is in operation due to the
effects of the feedback. Chapter 5 describes a procedure for determining the responses of
the actuators online while the system is running under closed loop control, which means
that the actuator array can be monitored continuously.
CHAPTER 2: A DISTRIBUTED PARAMETER MODEL OF THE PLANT

In this chapter, we aim to establish a distributed parameter model for the cross directional control system. This model will relate the changes in actuator set points to the observed profile changes measured at the output of the plant. The model will form the basis of the analysis of the response of the system.

In section 2.1 we set out some assumptions about the plant which are required for setting up the model. Section 2.2 develops the distributed parameter model and sets up a simple (but realistic) example which will be used in future chapters. A typical way of operating a cross directional control system is to take a measurement of the cross directional profile variations across the strip and then use this as the basis of a calculation of new actuator set points. The set points remain fixed until a new cross directional profile is obtained. This "sample and hold" mode of control is modelled in section 2.3.

The analysis of distributed parameter systems is a major area of research and we will be using a number of standard results in section 2.2. Extensive (but rather old) surveys of the field can be found in [Wang, 1968] and [Robinson, 1971]. Recent references are listed in a survey by [Kubrusly and Malebranche, 1985], although they concentrate on only a small section of the whole field of research. Good introductions to the field (together with most of the results required for section 2.2) can be found in [Curtain and Pritchard, 1978], [Curtain and Pritchard, 1977] and [Banks, 1983].

2.1 Assumptions of the Model

This section contains a list of the assumptions about the cross directional control system that are necessary for setting up an "ideal" model of a web forming process. This model was devised as a vehicle for the development of an analysis and design procedure which is,
as far as we know, applicable to all web forming processes. Specific applications may require some changes to this model (and some of the more common of these will be discussed at specific points in the study) but these changes will affect the details and not the underlying philosophy of the development.

Assumption 2.1.1: Actuators The thickness is to be controlled by a group of actuators placed in a line across the web. The array of actuators is assumed to have the following properties.

(a) The time and space responses of the actuators are linear and each actuator has an identical time response.

(b) The centres of the actuators are equally spaced.

(c) Each of the actuators has an identical, continuous cross directional spatial response which is symmetrical about its point of application. This means that if a unit change is applied to the set point of any actuator, the shape of the resultant profile change is the same with respect to the point of application.

(d) The spatial response of each actuator tends asymptotically to zero with distance from the point of application.

(e) The actuator set points are adjusted so that the net flow of material through the actuator array is fixed, so that the machine directional and cross directional controllers can be decoupled (as discussed in section 1.2).

Assumption 2.1.2: Measuring The forms of measuring systems that are encountered in practice are:

(i) a slowly scanning measuring device which either generates an analogue thickness signal or samples the thickness at a finite number of positions
across the sheet (common in plastic film and paper production as described in section 1.3)

(ii) a fixed finite number of measuring devices, spaced equally across the sheet, all measuring continuously (common in measuring shape in metal rolling).

In this study, we will consider an ideal model of the measuring process which will be taken to have the following properties.

(a) The measuring device takes an instantaneous cross directional profile reading which is continuous across the whole of the sheet.

(b) The device measures the cross directional profile as the deviation from the mean thickness across the web, $\bar{P}(nT)$, as described in equation 1.2.10.

(c) The measuring device will be downstream of the actuator array. This means that there will be a transport delay between a change in an actuator set point being applied and the effect of the change appearing at the measuring device.

Although these assumptions are not always fulfilled in practice, they do allow us to develop a useful theory which can be readily modified to accommodate practical measuring devices.

2.2 The Distributed Parameter Model

This section creates this form of model for the response of the system to changes in the set point of actuators in the array. If we know the impulse responses of the actuators and their set points changes, then we can predict the changes in the profile. We will require a number of concepts from the analysis of vector spaces. For reference, these are set out in Appendix A.

Assume that width of the sheet is $2l$, and that the cross directional variable is denoted by $x \in [-L, L]$. The cross directional profile at the measuring point will be the profile variation across the sheet which is a function defined on $[-L, L]$ that is continually varying in time. We
will be interested in modelling the response due to the actuators over the time interval 
$t \in [t_0, t_1]$ and we will denote this response as 
$$y(x, t) : \mathbb{R} \times [t_0, t_1]$$ 2.2.1

where the expression to the right of the colon indicates the spaces over which the function is defined.

For convenience, we will take the number of actuators to be odd and assume that there are 
$2N+1$ actuators spaced equally across the sheet. The impulse response of the actuators (ie the change in profile caused by a unit set point change on an actuator) will be denoted by 
$$b_k(x) : [-L, L] \quad \text{for all } k = (-N, -(N-1), ..., -1, 0, 1, ..., (N-1), N)$$ 2.2.2

Here we have assumed that the spatial shape of this response has no dynamics, so that 
$b_k(x)$ is independent of time. The extension of the analysis to include responses which are functions of time will be discussed in section 2.3.

Assume that each $b_k(x)$ is sufficiently smooth so that it is square integrable over the region 
$x \in [-L, L]$ (ie $b_k(x) \in L^2[-L, L]$ for all $k = (-N, ..., 0, ..., N)$). This assumption is reasonable, since it is very unlikely that practical actuator response will have discontinuities that are so severe that the responses do not have finite square integrals. In fact, practical actuator responses are likely to be smooth in the sense that $b_k(x) \in C[a, b]$, (where $C[., .]$ represents the space of continuous functions) or $b_k(x) \in C^\infty[a, b]$, (where $C^\infty[., .]$ represents the space of functions whose higher order differentials are all continuous).

The inputs to the system will be taken to be the set points of the actuators in the array. These will be functions of time, which will be denoted by 
$$u_k(t) : [t_0, t_1] \quad \text{for all } k = (-N, -(N-1), ..., -1, 0, 1, ..., (N-1), N)$$ 2.2.3

At the moment we will not make any assumptions about the form of $u_k(t)$, except to note that when the plant is operated in a sample and hold mode, there be discontinuities in the time course of the set points.
We now assume that the time dynamics of the profile response due to the actuator array are described by a differential equation with respect to time, so the response at any point $x \in [-L,L]$ across the web can be written as

$$
\frac{d^n y(x,t)}{dt^n} + a_{n-1} \frac{d^{n-1} y(x,t)}{dt^{n-1}} + \ldots + a_1 \frac{dy(x,t)}{dt} + a_0 y(x,t) = \sum_{k=-N}^{N} b_k(x) u_k(t-\tau_d)
$$

2.2.4

where $\tau_d$ is the transport delay between the actuators and the measuring point. The right hand side of the equation describes the shape of the combined response of all of the actuators, while the left hand side describes the dynamics of the transient response of the actuators. Note that it is assumed that all actuators have the same dynamics (see Assumption 2.1.1a) which are taken to be of nth order.

In order to reduce the amount of notation in the subsequent discussion, we will drop the variable $x$ from the functions $y(x,t)$ and $b_k(x)$.

Equation 2.2.4 can now be written in a simpler form as

$$
\dot{q}(t) = \mathcal{A} q(t) + \sum_{k=-N}^{N} b_k u_k(t-\tau_d)
$$

2.2.5

where

$$
q^T(t) = \begin{bmatrix} y(x,t) & \frac{dy(x,t)}{dt} & \ldots & \frac{d^{n-1} y(x,t)}{dt^{n-1}} \end{bmatrix}
$$

2.2.6

and

$$
\mathcal{A} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 1 & \ldots & 0 \\
0 & 0 & 0 & \ldots & 1 \\
-a_0 & -a_1 & \ldots & -a_{n-2} & -a_{n-1} \end{bmatrix}
$$

2.2.7

Note that $q(t)$ contains a finite number ($n$) of infinite dimensional functions, so $q(t)$ is of infinite dimensions.
Equation 2.2.5 is a distributed parameter equation. In order to solve 2.2.5, over the time interval $t \in [t_0, t_1]$, we need the initial conditions on $y(t)$ and its derivatives which we will write as

$$q(t_0) = q_0$$  \hspace{1cm} 2.2.8

We are now in a position to define both the distributed parameter problem and the solution that we seek.

**Definition 2.2.1** The distributed parameter model of the system is described by

$$\dot{q}(t) = \mathcal{A}q(t) + \sum_{k=1}^{N} b_k u_k(t-\tau_k) \quad \text{for} \quad t \in [t_0, t_1] \quad \text{with} \quad q(t_0) = q_0$$  \hspace{1cm} 2.2.9

(with $q(t)$ and $\mathcal{A}$ defined in 2.2.6 and 2.2.7). For the initial given initial conditions, $q(t_0) = q_0$, we seek a unique solution for $q(t)$ on the interval, $t \in [t_0, t_1]$.

We have assumed that each of the responses of the actuators, $b_k$ is in $L^2[-1,1]$ for $t \in [t_0, t_1]$, but we also need to ensure that $y(x,t)$ is in $L^2[-1,1]$ for the whole of the time interval, $t \in [t_0, t_1]$. It can be readily shown [Curtain and Pritchard, 1977, section 8.5] that provided each of the individual elements of the vector of initial conditions, $q_0$, is in $L^2[-1,1]$, then all of the elements of $q(t)$ (and hence $y(t)$, which is the first element of $q(t)$) are in $L^2[-1,1]$ for $t \in [t_0, t_1]$. This requirement on the vector of initial condition can be viewed as ensuring that

$$y(t_0) \in L^2[-1,1], \quad \frac{dy(t)}{dt} \bigg|_{t=t_0} \in L^2[-1,1], \quad \ldots, \quad \frac{d^{(n-1)}y(t)}{dt^{(n-1)}} \bigg|_{t=t_0} \in L^2[-1,1]$$  \hspace{1cm} 2.2.10

(using the definition of $q(t_0)$ in equation 2.2.8). For all realistic systems, this is a reasonable restriction, because it is extremely unlikely that the spatial response of the actuator array or any of the differentials of the response with respect to time, are going to be sufficiently unsmooth that they are not square integrable.

Although this Definition 2.2.1 is rather abstract, it fits in with the standard distributed
parameter problem (see, for example [Curtain and Pritchard, 1977]) for which there is a known solution. Before presenting this solution, it is worth pointing out two important differences between the standard problem and the problem developed for this study.

(i) Usually, partial differential equations are used for modelling. This leads to the $A$ matrix containing operators which are partial derivatives with respect to the space over which the equations are defined. For the problem considered here $y(t)$ is defined on $x \in [-1, 1]$, so $A$ would contain operators such as \[ \frac{\partial}{\partial x}, \frac{\partial^2}{\partial x^2}, \text{etc} \]

The system in this study contains only derivatives with respect to time which means that $A$ contains only constants (see equation 2.2.7). This is a reflection of the relatively simple dynamics of web forming processes. As we will see below, this leads to a major simplification, because $A$ is bounded.

(ii) A standard distributed parameter system usually defines conditions which must be satisfied on the boundary of the domain of the system (ie at $x = -L$ and $x = L$ for the system in this problem). This means that a set of eigenfunction solutions can be found [Curtain and Pritchard, 1977] and any general solution can be expressed as an infinite Fourier Series in terms of the eigenfunctions. Because the eigenfunctions are “discrete”, a lumped parameter approximation to the true system can be generated by truncating the Fourier series at a suitable point [Kubrusly and Malebranche, 1985]. In the problem in this study, there are no natural boundary conditions (since we do not specify anything about the profile error at the edge of the web), so we do not have a set of discrete eigenfunctions. Instead there is a continuum of solutions and we have to use the Fourier transform rather than Fourier series (see section 3.3).
With these two differences in mind, we appeal to standard distributed parameter theory to derive a solution to the system equation given in Definition 2.2.1. Equation 2.2.9 is an inhomogenous differential equation. We start by considering only the homogenous part of the equation

$$\dot{q}(t) = A q(t) \text{ with } q(t_0) = q_0 \text{ for all } t \in [t_0, t_1]$$

Equation 2.2.11

Once this has been solved, the solution to the inhomogenous problem can be obtained.

For the moment, let us consider the case where $q(t)$ is finite dimensional and $A$ is bounded. From Definition A7 in appendix A, this means that $A \in L(X)$ with $X$ being the Banach space over which the transformation is defined. When $A$ is bounded, the solution of 2.2.11 is [Curtain and Pritchard, 1977, section 8.1]

$$q(t) = e^{A(t-t_0)} q_0 \text{ for all } t \in [t_0, t_1]$$

Equation 2.2.12

This equation describes the evolution of the system from time, $t_0$, to a later time, $t$, when there is no forcing term. The evolution from $t_0$ to $t$ is determined by the state transition matrix $e^{A(t-t_0)}$. For infinite dimensional systems with more general $A$ matrices (which are possibly unbounded), the exponential matrix does not necessarily exist, so we cannot use this solution. Therefore we have to use the more abstract concept of a semigroup to describe the evolution of a system. We will see that the exponential matrix is a semigroup when the $A$ matrix is bounded. Because semigroup theory is well established, most of the proofs of the theories will be referred to standard texts.

Definition 2.2.2: Semigroup If $T(t)$ is an operator from $R^+$ to $L(X)$, where $X$ is a Banach space, and satisfies

(i) $T(t+s) = T(t) T(s) \quad t > s \geq 0$

(ii) $T(0) = I$

(iii) $\| T(t) q - q \|_X \to 0$ as $t \to 0$ for all $q \in X$

then $T(t)$ is said to be a strongly continuous semigroup.
It should be noted that the exponential function, \( e^{At} \), satisfies the definition of a strongly continuous semigroup [Banks, 1983, chapter 3].

**Definition 2.2.3: Infinitesimal Generator of a Semigroup**

The infinitesimal generator, \( \mathcal{A} \), of a strongly continuous semigroup, \( \mathcal{T}(t) \), is given by

\[
\mathcal{A}q = \lim_{t \to 0^+} \frac{1}{t} (\mathcal{T}(t)q - q) \quad 2.2.13
\]

whenever the limit exists. The domain of \( \mathcal{A} \) (which we will write as \( D(\mathcal{A}) \)) is the set of elements, \( q \in X \), for which the limit exists.

If \( \mathcal{A} \) is bounded, so that \( e^{At} \) exists see Definition A9) then

\[
\lim_{t \to 0^+} \frac{1}{t} (e^{At}q - q) = \lim_{t \to 0^+} \frac{1}{t} (q + t\mathcal{A}q + \frac{1}{2!}t^2\mathcal{A}^2q + \frac{1}{3!}t^3\mathcal{A}^3q + \ldots - q) = \mathcal{A}q \quad 2.2.14
\]

\[
= \lim_{t \to 0^+} (\mathcal{A}q + \frac{1}{2!}t\mathcal{A}^2q + \frac{1}{3!}t^2\mathcal{A}^3q + \ldots ) = \mathcal{A}q \quad 2.2.15
\]

This means that if \( \mathcal{A} \) is bounded, \( e^{At} \) exists and \( \mathcal{A} \) is the infinitesimal generator of the strongly continuous semigroup, \( e^{At} \).

Having defined a semigroup and its infinitesimal generator we can use these concepts to solve the homogenous equation given in equation 2.2.11.

**Theorem 2.2.1** If \( \mathcal{T}(t) \) is a strongly continuous semigroup with an infinitesimal generator, \( \mathcal{A} \), then the solution of the homogenous equation

\[
\dot{q}(t) = \mathcal{A}q(t) \quad \text{with} \quad q(t_0) = q_0 \in D(\mathcal{A}) \quad \text{for all} \quad t \in [t_0, t_1] \quad 2.2.17
\]

is given by

\[
q(t) = \mathcal{T}(t-t_0)q_0 \quad 2.2.18
\]

This solution is unique for each choice of \( q_0 \in D(\mathcal{A}) \).
Proof: See [Curtain and Pritchard, 1977, section 8.6].

This gives us a solution to the homogeneous evolution expression in equation 2.2.11, although we are not in a position to use it because we need an expression for the semigroup associated with the transformation, \( \mathcal{A} \). Definition 2.2.3 shows us how to obtain an infinitesimal generator from a given semigroup, but we require the reverse: given an infinitesimal generator, \( \mathcal{A} \), what is the corresponding semigroup, \( T(t) \)? For general distributed parameter systems, finding the semigroup is the most difficult part of solving the problem, but for the system being considered in this study, the following theorem gives us a major simplification.

**Theorem 2.2.2** The linear transformation, \( \mathcal{A} \), given in equation 2.2.7, is bounded (ie \( \mathcal{A} \in \mathcal{L}(X) \)).

**Proof** For the system in equation 2.2.7, the Banach space, \( X \), over which the transformation, \( \mathcal{A} \), is described, is the set of \( q(t) \) defined in equation 2.2.6. Each \( q(t) \) is a vector containing \( n \) infinite dimensional functions which are defined on \( x \in [\alpha, \beta] \) and \( t \in [t_0, t_1] \). We will specify a norm on this Banach space as

\[
\| q(t) \|_X = \sqrt{\sum_{j=1}^{n} \langle q(t)_j, q(t)_j \rangle}
\]

where \( \{q(t)_j\} \) is the \( j \)th element of the vector and the inner product is defined as the product integral on \( x \in [-\beta, \alpha] \) (see equation A3). It is readily seen that this definition satisfies the criteria for a norm as given in Definition A1 and because we have used the inner product to define the norm, this means that \( X \) is a Hilbert space (see Definition A4).

From Definition A6, we know that the transformation, \( \mathcal{A} \), is bounded if

\[
\| \mathcal{A} q(t) \|_X \leq k \| q(t) \|_X
\]

for some finite constant, \( k \). From the definition of the norm we can write the norm of \( \mathcal{A} q(t) \) as
\[ \| \mathcal{A} q(t) \|_X = \sum_{j=1}^{n} < \{ \mathcal{A} q(t) \}_j, \{ \mathcal{A} q(t) \}_j > \]  

Looking at the definition of \( \mathcal{A} \) in equation 2.2.7, we can see that the transformation is a matrix containing constants and that

\[ \{ \mathcal{A} q(t) \}_j = \{ q(t) \}_{j+1} \quad \text{for } j=1, ..., n-1 \]

\[ \{ \mathcal{A} q(t) \}_n = -a_0 \{ q(t) \}_1 - a_1 \{ q(t) \}_2 - ... - a_{(n-1)} \{ q(t) \}_n \]  

For \( j=1, ..., (n-1) \), this means that

\[ < \{ \mathcal{A} q(t) \}_j, \{ \mathcal{A} q(t) \}_j > = < \{ q(t) \}_{j+1}, \{ q(t) \}_{j+1} > \]

For the nth term of the vector \( \mathcal{A} q(t) \), using the definitions of the inner product and the norm (Definitions A1 and A2), we can show that

\[ < \{ \mathcal{A} q(t) \}_n, \{ \mathcal{A} q(t) \}_n > \leq |a_0| < \{ q(t) \}_1, \{ q(t) \}_1 > + |a_1| < \{ q(t) \}_2, \{ q(t) \}_2 > + ... + |a_{(n-1)}| < \{ q(t) \}_n, \{ q(t) \}_n > \]

Combining equations 2.2.23 and 2.2.24 leads to

\[ \| \mathcal{A} q(t) \|_X \leq [1+|a_0|] < \{ q(t) \}_1, \{ q(t) \}_1 > + [1+|a_1|] < \{ q(t) \}_2, \{ q(t) \}_2 > + ... + [1+|a_{(n-1)}|] < \{ q(t) \}_n, \{ q(t) \}_n > \]

\[ \leq [1+k'] \| q(t) \|_X \]

where \( k' \) is the maximum magnitude of the coefficients of the dynamics, ie

\[ k' = \max_{j=1,(n-1)} |a_j| \]

Provided all of these coefficients are finite, then comparing equations 2.2.20 and 2.2.26 we see that the transformation, \( \mathcal{A} \), is bounded and \( \mathcal{A} \in L(X) \).

Because \( \mathcal{A} \in L(X) \), this means that the exponential transformation, \( e^{\mathcal{A} t} \) (as described in Definition A9) exists and, as shown above, this means that \( e^{\mathcal{A} t} \) is a strongly continuous semigroup with infinitesimal generator, \( \mathcal{A} \). The solution of the homogenous evolution equation
\[ \begin{align*}
\dot{q}(t) &= Aq(t) \quad \text{with} \quad q(t_0) = q_0 \in \mathcal{D}(A) \quad \text{for} \quad t \in [t_0, t_1] \\
\text{becomes} \quad q(t) &= e^{A(t-t_0)} q_0 \quad \text{provided that} \quad A \in \mathcal{L}(X) \\
\text{In obtaining this solution we have not placed any restrictions on the dimensions of} \quad q(t) \quad \text{and if we compare equations} \quad 2.2.12 \quad \text{and} \quad 2.2.29 \quad \text{we see that the solution for} \quad q(t) \quad \text{with infinite dimensions is the same as the solution for finite dimensional} \quad q(t), \quad \text{provided that} \quad A \in \mathcal{L}(X). \\
\text{Having solved the homogenous equation, we can return to the solution of the inhomogenous equation.}
\end{align*} \]

**Theorem 2.2.3** The solution of the inhomogenous equation

\[ \begin{align*}
\dot{q}(t) &= Aq(t) + \sum_{k=-N}^{N} b_k u_k(t) \quad \text{with} \quad q(t_0) = q_0 \in \mathcal{D}(A) \quad \text{for} \quad t \in [t_0, t_1] \\
\text{is given by} \quad q(t) &= e^{A(t-t_0)} q_0 + \int_{t_0}^{t} e^{A(t-s)} \sum_{k=-N}^{N} b_k u_k(s) \, ds \\
&= e^{A(t-t_0)} q_0 + \sum_{k=-N}^{N} \int_{t_0}^{t} e^{A(t-s)} b_k u_k(s) \, ds \quad \text{provided all of the} \quad u_k(t) \quad \text{are continuously differentiable (ie} \quad u_k(t) \in C^1[t_0, t_1] \quad \text{for all} \quad k = \{-N, \ldots, N\}).
\end{align*} \]

**Proof** see [Curtain and Pritchard, 1977, section 8.7]

In practice it is usual to operate the system in a "sample and hold" mode so \( u_k(t) \) will have discontinuities at the sample points, and \( u_k(t) \) is not in \( C^1[t_0, t_1] \). However, if the following conditions hold:

(i) \( e^{A(t-s)} b_k u_k(s) \in \mathcal{D}(A) \) for almost all \( t > s \in [t_0, t_1] \).
then equation 2.2.32 is a unique solution of the inhomogenous equation (almost everywhere) [Banks, 1983, chapter 3]. These three conditions are very weak and are reasonable for bounded $\mathcal{A}$ and smooth actuator responses, $b_k$. In the rest of this study we will assume that these conditions hold so that the continuous response of the array of actuators is given by equation 2.2.32.

Example 2.2.1 We now consider a simple example system which will be used as an illustration throughout this study. We will take the actuator dynamics to be first order, so the system equation for the response of the actuators becomes (from 2.2.4)

$$\frac{dy(t)}{dt} + a y(t) = \sum_{k=-N}^{N} b_k u_k(t-S)$$

with $y(t_0) = y_0$ for $t \in [t_0, t_1]$ 2.2.33

Writing $y(t) = q(t)$, this can be put in the form of equation 2.2.5 as

$$\dot{q}(t) = -a q(t) + \sum_{k=-N}^{N} b_k u_k(t-S)$$

with $q(t_0) = y_0$ for $t \in [t_0, t_1]$ 2.2.34

The transformation $\mathcal{A}$ is now simply multiplication by a scalar, $-a$. If $a$ is finite, this transformation is bounded, so from Theorem 2.2.3, the solution becomes

$$y(t) = e^{-a(t-t_0)} y_0 + \sum_{k=-N}^{N} \int_{t_0}^{t} e^{-a(t-s)} b_k u_k(s-S) \, ds$$

for $t \in [t_0, t_1]$ 2.2.35

provided $u(s)$ satisfies the three conditions above.

In this example, we shall assume that each of the actuators to has a spatial response which is the shape of a Gaussian curve. Following the assumptions of section 2.1, it be taken that the actuators are equally spaced with the centre line of the actuator in the middle of the array (corresponding to $k=0$) positioned at $x=0$. This means that the responses of the actuators are given by
\[ b_k = \exp\left[-\frac{(x-kd)^2}{2\sigma^2}\right] \quad \text{for} \quad k=\{-N, -(N-1), \ldots, -1, 0, 1, \ldots, (N-1), N\} \]

\( \sigma \) is the standard deviation and \( \xi \) is the distance between actuators. Note that this means that \( b_k \in L^2[\mathbb{L}] \) because a Gaussian curve has a finite square integral. This form of actuator response is typical of practical processes, particularly those with actuators which rely on some heating or cooling process, for example, the array of heaters described in section 1.3 for the control of thickness of plastic films.

Although the Gaussian spatial response obeys Assumption 2.1.1d which requires that the spatial response goes asymptotically to zero as the distance from the centre line of the actuator increases, the net response (over the whole web) is not zero because the average level of the response on \( x \in [-\mathbb{L}] \) is positive. This appears to contradict the requirement given in chapter 1 that a cross directional control scheme should redistribute material across the web without affecting the mean level of the web. However, it will be shown in the next section that, provided the measured profile is taken relative to the average profile and the set points of the actuators in the array are adjusted so that their average is fixed, an array of actuators with this type of response will operate as a cross directional controller.

In this example, \( y(t) \) defines the continuous response of the array of actuators. The actual measured profile will be given by the actuator response corrupted by any disturbances or noise processes in the system. Thus the measured response, \( z(t) \), is given by

\[ z(t) = y(t) + d + e(t) \quad 2.2.37 \]

where \( d \) defined on \( [-\mathbb{L}] \) is a steady state disturbance and \( e(t) \) defined on \( [t_0, t_1];[-\mathbb{L}] \) is a continuous noise process. We can identify \( d \) as the fixed cross directional error, \( P_{CD}(x) \) and \( e(t) \) with the residual disturbance, \( P_R(x,t) \) as defined in section 1.2.
2.3 Sampled Data Model

The model that we have developed for the actuator response is continuous in both the spatial and the time domain. The normal mode of operation in most plants is to take a reading of the sheet and then to use this profile to sample to calculate the new set points for the array of actuators. The actuator set points remain fixed until the next profile sample is taken and a new group of set points are generated. This sample and hold mode of operation produces responses which are discontinuous in the time domain, but continuous in the spatial domain. This section will set up the model of such a system. The techniques required for calculating the new set points at each sample interval will be examined in the next chapter.

Assume that the profile samples are taken at times \( t = 0, T, 2T, \ldots \) where \( T \) is the sample interval. From Assumption 2.1.2a the measuring device takes an instantaneous reading of the profile across the whole sheet at each sample interval. We will simplify the analysis by considering the first order system described in Example 2.2.1 which has a response on \( t \in [t_0, t_1] \) given by (from equation 2.2.35)

\[
y(t) = e^{-s(t-t_0)} y(t_0) + \sum_{k=-N}^{N} \int_{t_0}^{t} e^{-s(t-s)} b_k u_k(s-t_d) \, ds  \tag{2.3.1}
\]

If we make \( t_0 = nT \) and \( t_1 = (n+1)T \), then \( y[(n+1)T] \) is can be expressed as

\[
y[(n+1)T] = e^{-s[(n+1)T-t_0]} y(nT) + \sum_{k=-N}^{N} \int_{nT}^{(n+1)T} e^{-s[(n+1)T-s]} b_k u_k(s-t_d) \, ds \tag{2.3.2}
\]

\[
= e^{-sT} y(nT) + \sum_{k=-N}^{N} \int_{nT}^{(n+1)T} e^{-s[(n+1)T-s]} b_k u_k(s-t_d) \, ds \tag{2.3.3}
\]
\[ N (n+1)T-Td = e^{-aT} y(a, e^{-a(n+1)T+s'\tau_d} b_k u_k(s') \, ds') \quad 2.3.4 \]

where \( s' = s - \tau_d \). We will assume that \( \tau_d < T \), although the analysis is readily extended to the case where \( \tau_d \geq T \).

Because the system is operated in a sample and hold mode, the set points, \( u_k(t) \) are constant between sample intervals. So

\[ u_k(t) = u_k(nT) \quad \text{for} \quad nT \leq t \leq (n+1)T \quad 2.3.5 \]
\[ u_k(t) = u_k[(n-1)T] \quad \text{for} \quad (n-1)T \leq t < nT \quad 2.3.6 \]

This allows us to rewrite equation 2.3.4 as

\[ y[(n+1)T] = e^{-aT} y(nT) + \sum_{k=-N}^{(n+1)T-\tau_d} \left\{ \int_{nT-\tau_d}^{(n+1)T-\tau_d} e^{-a[(n+1)T-s'\tau_d]} \, ds' \cdot b_k u_k[(n-1)T] \right\} \quad 2.3.7 \]

\[ y[(n+1)T] = e^{-aT} y(nT) + \sum_{k=-N}^{(n+1)T-\tau_d} \left\{ \frac{1}{a} \left( e^{-aT} - e^{-aT} \right) b_k u_k[(n-1)T] \right\} + \frac{1}{a} \left( 1 - e^{-aT} \right) b_k u_k(nT) \quad 2.3.8 \]

It is important to note that the evolution of the profile from \( t = nT \) to \( t = (n+1)T \) depends upon the set points \( u_k[(n-1)T] \) as well as \( u_k[nT] \). This is a result of the transport delay, \( \tau_d \).

Equation 2.3.8 can be written in a more compact form by using the backward difference operator, \( q^{-1} \), where

\[ q^{-1} u(t) = u(t-T) \quad 2.3.9 \]

This allows us to write equation 2.3.8 as

\[ (1 + f_1 q^{-1}) y(nT) = \sum_{k=-N}^{N} (h_1 q^{-1} + h_2 q^{-1}) b_k u_k(nT) \quad 2.3.10 \]

where
We are now in a position to show why we can ensure that the array of actuators does not affect the mean web profile when the average of the set points is fixed, as postulated in section 1.2. Rearranging equation 2.3.10 as

\[ y(nT) = \frac{h_1 q^{-1} + h_2 q^{-2}}{1 + f_1 q^{-1}} \sum_{k=-N}^{N} b_k u_k(nT) \]

(where we have taken the dynamic term outside the summation because all of the actuators have the same dynamics) the average effect over the whole of the web is given by

\[ \bar{y}(nT) = \frac{1}{2L} \int_{-L}^{L} \frac{h_1 q^{-1} + h_2 q^{-2}}{1 + f_1 q^{-1}} \sum_{k=-N}^{N} b_k u_k(nT) \]

\[ = \frac{1}{2L} \frac{h_1 q^{-1} + h_2 q^{-2}}{1 + f_1 q^{-1}} \sum_{k=-N}^{N} u_k(nT) \int_{-L}^{L} b_k \, dx \]

If we ignore edge effects for the moment, then because we have assumed that the responses of all of the actuators is the same, we can denote the effect on the net profile due to a unit change in the set point of any of the actuators, as

\[ \bar{b} = \int_{-L}^{L} b_k \, dx \quad \text{for } k = \{-N, -(N-1), ..., (N-1), N\} \]

which leads to

\[ \bar{y}(nT) = \frac{1}{2L} \frac{h_1 q^{-1} + h_2 q^{-2}}{1 + f_1 q^{-1}} (2N+1) \bar{b} \sum_{k=-N}^{N} u_k(nT) \]

If we ensure that the average of the set points is fixed at all time samples (and hence for all times because the set points do not change between samples in a sample and hold system) so that
\[
\bar{u} = \frac{1}{2N+1} \sum_{k=-N}^{N} u_k(nT) \quad \text{for } n = 0, 1, 2, \ldots
\]

then
\[
\bar{y}(nT) = \frac{1}{2A} \frac{h_1 q^{-1} + h_2 q^{-2}}{1 + f_1 q^{-1}} \frac{(2N+1)^2 \bar{u}}{b}
\]

Because \( u \) does not change with time, the net effect of the actuator array is given by
\[
\bar{y}(nT) = \frac{(2N+1)^2}{2A} \bar{u} g(0)
\]

where \( g(0) \) is the dc gain of the dynamics. This shows us that \( \bar{y}(nT) \) is constant for all time samples, so we are justified in assuming that the actuator array does not change the mean level of the profile. This allows us to separate the cross directional control system from the machine direction controller as described in section 1.2.

If edge effects are included in the analysis, then we have to allow for the fact that the net profile effect due to a unit change in one the actuators, as defined in equation 2.3.17, will not be the same for all of the actuators in the array, because the responses of the actuators at the edges of the web will be truncated. In this case, if we ensure that the weighted average of the actuator set points is held fixed for all samples (rather than a simple average), where the weights in the average are equal to the net effect of a unit change on each of the actuators, then we can still be certain that the actuator array will not affect the mean profile.

The analysis above has concentrated on the case where there is no process noise so the expressions for the profile, \( y(t) \) describe the effects of only the actuators. Disturbance and noise terms can be included by writing
\[
z(nT) = y(nT) + d + e(nT)
\]

where \( e(nT) = e(t=nT) \). If we take \( z(nT) \) to be \( \Delta P(x,nT) \), the measured web thickness relative to the mean of the sample, as defined in equation 1.2.10, then \( d \) can be seen to be the steady state cross directional disturbance, \( P_{CD}(x) \), while \( e(nT) \) is the residual component at time \( t=nT \), \( P_R(x,nT) \).
Up to this point, we have assumed that the spatial responses of the actuators do not have any dynamic response. This allows us to use spatial responses, $b_k$, which are functions of $x$ only and independent of time, but this is unlikely to be true for a large class of practical plants. However, if the lengths of the transient of the dynamic response of the spatial responses are less than the time between samples, $T$, then the measurements do not "see" the dynamics, which means that the assumption is valid, when we are using a sampled data model. If the length of the transients is longer than the sample period, then it will be necessary to include extra terms in the response of the actuators which is given in equation 2.3.10. To illustrate this, we will simplify 2.3.10 by considering a plant which has $f_1$ and $h_2$ equal to zero. For the case where the spatial responses have no dynamics or where the lengths of the transients are less than the sample period, the response of the actuators is given by

$$y[(n+1)T] = \sum_{k=-N}^{N} h_1 b_k u_k(nT)$$

If the lengths of the transients are longer than $T$, then it is necessary to include extra terms into the response of the actuators

$$y[(n+1)T] = \sum_{k=-N}^{N} h_1 \left[ b_k^{(1)} u_k(nT) + b_k^{(2)} u_k((n-1)T) \right]$$

Here $b_k^{(1)}$ is the spatial response that is seen at the first sample following an impulse set point change and $b_k^{(2)}$ is the spatial response at the subsequent sample. Sufficient terms can be included to describe the full dynamic response. Although the subsequent analysis can be expanded to include this more general form, in order to simplify the notation in the rest of this study, it will be taken that $b_k$ is independent of time.
CHAPTER 3: THE SPATIAL CONTROLLABILITY OF A CROSS DIRECTIONAL CONTROL SCHEME

In this chapter we investigate the extent to which a particular cross directional profile variation can be controlled by an array of actuators. We will be concentrating on the controllability of the spatial variations (rather than the temporal controllability) of the system so we will assume that array is operating in "sample and hold" mode and has the simple dynamics described in the previous section. Section 3.1 considers the spatial controllability of the distributed parameter model of the array that was created in the previous chapter. In particular, we investigate whether it is possible to make the response of the array of actuators match a given profile and it will be seen that the array cannot generate an arbitrary spatial response. Given that some responses are unattainable, section 3.2 shows how to choose the actuator set points so that the response of the array matches the desired response as closely as possible in a least squares sense.

The definition of controllability developed in section 3.1 is rather abstract and is not immediately applicable in practice. Section 3.3 transfers the distributed parameter model to the spatial frequency domain and generates the spatial frequency response of the actuator array. This allows us to identify the uncontrollable frequency components of a web profile directly and to develop criteria for designing the separation and shape of response of the actuators in the array. These criteria are described in section 3.4. For technical reasons, the spatial frequency response of the array is based upon array and a web of infinite width. The results are modified for the practical case of arrays and webs with finite width, in section 3.5.

The results that are used in section 3.1 for describing the controllability of distributed parameter systems are well established (see, for example [Curtain and Pritchard, 1977] or [Banks, 1983]), but they do not appear to have been applied to cross directional control schemes. Similarly, the use of the least squares criterion for approximating functions is
also well known [Davis, 1975], but again it has not been used in connection with distributed parameter models of cross directional control schemes. The use of the spatial frequency domain as a tool for analysing the response of an array of actuators in a cross directional control scheme is completely new. Although the problem of positioning actuators for general distributed parameter systems has received considerable attention (see the survey paper by [Kubrusly and Malebranche, 1985]) these results have been applied to systems defined by partial differential equations and are not applicable to the systems being considered here. Therefore the approach to the design of the shape of the actuator responses and the spacing of the actuators in the array is also new. Similar criteria have been investigated by a number of authors [Hadra, 1985], [Karlsson et al, 1985], [Lindeborg, 1986] but their results have been based on simulations or heuristics. The analysis presented in this study is both rigorous and provides valuable insight together with new engineering design methods.

3.1 Spatial Controllability

In this section we investigate the controllability of the distributed parameter system. In order to restrict the notational complexity in the subsequent discussion, we will consider the first order, continuous time model of the actuator array given in example 2.2.1.

\[ y(t) = e^{-at}y_0 + \sum_{k=-N}^{N} \int_{t_0}^{t} e^{-s(t-t_0)} b_k u_k(s) \, ds \quad \text{for all } t \in [t_0, t_1] \quad 3.1.1 \]

where \( y(t_0) = y_0 \in L^2[-L,L], b_k \in L^2[-L,L] \) and \( u_k(t) \in L^1[t_0,t_1] \). For simplicity we have taken \( \tau_d = 0 \).

**Definition 3.1.1 Controllability.** The system in equation 3.1.1 is controllable if and only if, for any pair of functions \( (y_0,y_1) \), both in \( L^2[-L,L] \), there exists a set of actuator set points, \( u_k(t) \in L^1[t_0,t_1] \) which drives the system from \( y_0 \) to \( y_1 \) over some finite time interval \( t \in [t_0,t_1] \) [Curtain and Pritchard, 1977, section 11.1].
Associated with this definition is the concept of a Controllable Subspace.

**Definition 3.1.2 Controllable Subspace.** The set of all functions, \( y(t_0) \), which can be steered to the origin \( y(t_1) = 0 \), in some finite time interval \( t \in [t_0, t_1] \).

Usually, for a distributed parameter system, it is necessary to introduce the concept of approximate controllability which defines a system as being approximately controllable if there exists a set of \( u_k(t) \in L^1[t_0, t_1] \) which will drive the system from an initial state, \( y_0 \), to a state which is arbitrarily close to the final state, \( y_1 \), within a finite time. However, because we have shown that \( A \) is bounded for the systems in this study, it is sufficient to use the definition of controllability given above [Banks, 1983, section 5.1]. It should be noted that this definition given is the same as the definition of controllability for finite dimensional systems.

**Theorem 3.1.1** The system 3.1.1 is controllable if and only if

\[
\langle h, e^{-at} b_k \rangle = 0 \quad \text{for} \quad k = -N, \ldots, N
\]

for all \( t \in [t_0, t_1] \) and for all \( u_k \in L^1[t_0, t_1] \), implies that \( h = 0 \), where \( h \) is an arbitrary function defined on \( x \in [-L, L] \).

**Proof** [Curtain and Pritchard, 1977, section 14.1]

Note that \( e^{-at} \) is a scalar and can be taken outside the inner product, so the theorem can be simplified to say that the system is controllable if and only if \( \langle h, b_k \rangle = 0 \) implies that \( h = 0 \).

Some insight into this result can be obtained by noting that if \( y_0 = 0 \), then from equation 3.1.1, \( y(t) \) must be a linear combination of the actuator responses, \( b_k \). (The coefficients of this linear combination will be determined by the function \( u_k(t) \) on \( t \in [t_0, t_1] \) and \( e^{-at} \). This
means that any point which can be reached from \( y_0 = 0 \) can be expressed as a linear combination of \( b_k \)'s, so the controllable subspace (see Definition 3.1.1) must be spanned by the \( b_k \)'s. If the system is controllable, the controllable subspace spans the whole of \( L^2[-1,1] \) [Curtain and Pritchard, 1977], so the only function, \( h \), which can satisfy \( \langle h, b_k \rangle = 0 \) for \( k = -N, ..., N \) is \( h = 0 \).

However, we can turn this argument around to prove that the system in equation 3.1.1 cannot be controllable. This is because \( L^2[-1,1] \) is a space with infinite dimensions, so it cannot be spanned by a finite set of functions, \( b_k, k = -N, ..., N \). Thus, the controllable subspace of the system cannot contain the whole of \( L^2[-1,1] \). In particular, this means that if \( y(t_0) \) is in the controllable subspace, then all future \( y(t) \) will remain in the controllable subspace. This means that the response of the array will not be able to match exactly an arbitrary function in \( L^2[-1,1] \). The best that can be done is to drive the system to a point as close as possible to the desired function. Thus, if \( z_1 \) is the desired end point, the system is driven to a point \( y_1^* \) at time \( t = t_1 \), where \( y_1^* \) minimises

\[
\min_{y_1} \| z_1 - y_1 \|_2 \tag{3.1.3}
\]

subject to \( y_1 \) being in the controllable subspace. Because we are dealing with functions in \( L^2[-1,1] \) the norm in this minimisation will be taken as the \( L^2 \) norm given in equation A3 in appendix A.

We can now generate a control strategy that can be applied to the sample and hold system developed in section 2.3. Suppose that all of the actuator set points are zero and that we have just taken a sample of the profile and it is required that the response of the actuators should have a shape given by the function \( z_1 \in L^2[-1,1] \), by the time that the next sample is taken. The required response of the actuators \( z_1 \), will be determined by the control law, but if a "one-step ahead" controller is being used [Goodwin and Sin, 1984], then \( z_1 \) would be chosen to cancel the error in the profile sample that has just been taken. This will be discussed further in chapter 4.
As pointed out above, unless $z_1$ happens to lie in the controllable subspace, it will not be possible to drive the response of the actuator array to $z_1$, so we choose the optimal end point as $y_1^*$ which minimises expression 3.1.3 above. Since we are in a sample and hold mode of operation, we require the fixed actuator set points which will make the actuators' response equal to $y_1^*$ at the next sample interval. Without loss of generality, we will take $t_0 = 0$ and $t_1 = T$ (where $T$ is the time between samples). We have assumed that all of the actuator set points are zero at $t_0 = 0$, so, using equation 3.1.1, we require the set of fixed set points, $u_k$, $k=-N,\ldots, N$, which satisfy:

$$y_1^* = \sum_{k=-N}^{N} b_k u_k \int_{0}^{T} e^{-as} \, ds$$

$$= \frac{1}{a} \left(1 - e^{-aT}\right) \sum_{k=-N}^{N} b_k u_k$$

where $u_k \in \mathbb{R}$; $y_1^*(T)$ and $b_k \in L^2[0,T]$. The term

$$\frac{1}{a} \left(1 - e^{-aT}\right)$$

is a scalar which is determined by the dynamics of the actuators and the sampling interval. The term simply scales the size of the actuator set points to compensate for the dynamic response of the system over the period $[0,T]$ and in this chapter we are mainly interested in the spatial response, so we can scale the responses to ensure that this factor is unity. Equation 3.1.5 then becomes

$$y_1^* = \sum_{k=-N}^{N} b_k u_k$$

We require the value of $y_1^*$ that minimises

$$\min_{y_1} \frac{1}{2} \left\lVert z_1 - y_1 \right\rVert_2$$

so we can substitute in to this expression for $y_1^*$ to give

$$\min_{u_k} \frac{1}{2} \left\lVert z_1 - \sum_{k=-N}^{N} b_k u_k \right\rVert_2$$

The solution to this minimisation problem is the choice of actuator set points which will
make the response of the actuator array as close as possible (in a 2 norm sense) to the desired response, at the next sample time.

There are two issues that arise from this minimisation problem.

(i) Given a $z_1$ and a set of individual actuator responses, $b_k; k=\{-N,...,N\}$, how do we choose the optimal set points $u_k^*$ to minimise 3.1.8? The solution to this problem will be given in section 3.2.

(ii) As pointed out above, the response of the actuators will only match the required response exactly, when $z_1$ lies in the controllable subspace. We will be interested in determining which values of $z_1$ lie in the controllable subspace, and for those $z_1$ which are not completely controllable, we would like to know the size of the residual between the desired response and the achieved response. In principle, these details come directly from the analysis of section 3.2, but a more practical analysis is obtained by considering the response of the actuator array when $z_1$ is a sinusoid on $x \in [-L]$. By choosing sinusoids of different frequencies and phases, we can build up a spatial frequency response of the actuator array. Because any general $z_1$ can be expressed as a sum of sinusoids, the spatial frequency response gives us a practical way of deducing the response of the array, as well as valuable insight into the behaviour of the system and the optimal design of the actuators and measuring device. This approach will be developed in section 3.3.
3.2 Choosing the Optimal Set Points

Section 3.1 set up the optimisation problem for choosing the best set points for generating an actuator response which matches the required response as closely as possible at the next sample point. This problem is

\[ \min_{u_k \in \mathbb{R}} \| z_1 - \sum_{k=-N}^{N} b_k u_k \|_2 \]  

3.2.1

where \( z_1 \in L^2[-L, L] \) and \( b_k \in L^2[-L, L] \) for \( k=-N, \ldots, N \). From appendix A, we know that \( L^2[-L, L] \) is a Hilbert space and that the norm is defined from an inner product on the space given by

\[ \| z \|_2 = \langle z, z \rangle^{1/2} \]  

3.2.2

\[ = \left[ \int_{-L}^{L} z^*(x) z(x) \, dx \right]^{1/2} \]  

3.2.3

Because we are dealing with real functions, \( z^*(x) = z(x) \), so 3.2.3 can be rewritten as

\[ \| z \|_2^2 = \int_{-L}^{L} | z(x) |^2 \, dx \]  

3.2.4

The minimisation problem now becomes

\[ \min_{u_k \in \mathbb{R}} \left[ \int_{-L}^{L} | z_1(x) - \sum_{k=-N}^{N} b_k(x) u_k |^2 \, dx \right]^{1/2} \]  

3.2.5

Essentially, the set of \( u_k \)'s which give the optimal solution, are the coefficients of the linear combination of \( b_k(x) \) for \( k=-N, \ldots, N \), which is closest (in a 2 norm sense) to \( z_1(x) \). This is a standard problem in least squares approximation theory and the solution is well known to be (see, for example, [Davis, 1975, chapter 8]) the set of \( u_k \)'s which satisfy the 2N+1 simultaneous equations
\[ u_N < b_N, b_N > + u_{(N-1)} < b_{(N-1)}, b_N > + \ldots + u_N < b_N, b_N > = < z_1, b_N > \]
\[ \vdots \quad \vdots \quad \vdots \quad \vdots \]
\[ u_N < b_N, b_N > + u_{-(N-1)} < b_{-(N-1)}, b_N > + \ldots + u_N < b_N, b_N > = < z_1, b_N > \]

This set of simultaneous equations can be expressed more simply by writing

\[ \Gamma = \begin{bmatrix} < b_N, b_N > & < b_{(N-1)}, b_N > & \ldots & < b_N, b_N > \\
\vdots & \vdots & \ddots & \vdots \\
< b_N, b_N > & < b_{(N-1)}, b_N > & \ldots & < b_N, b_N > \end{bmatrix} \]

where \( \Gamma \in \mathbb{R}^{(2N+1) \times (2N+1)} \). Equation 3.2.6 then becomes

\[ \begin{bmatrix} u_N \\ u_{(N-1)} \\ \vdots \\ u_N \end{bmatrix} = \Gamma^{-1} \begin{bmatrix} < z_1, b_N > \\ < z_1, b_{(N-1)} > \\ \vdots \\ < z_1, b_N > \end{bmatrix} \]

For this solution to exist, \( \Gamma \) must be non-singular. Note that for real functions

\[ < b_i (x), b_j (x) > = < b_j (x), b_i (x) > \]

so the \( \Gamma \) matrix will be symmetric.

The matrix \( \Gamma \) is known as the Gram matrix of \( b_N, b_{(N-1)}, \ldots, b_N \). Looking at expression 3.2.7 we can see that the \( i,j \)th element of \( \Gamma \) is given by

\[ [\Gamma]_{ij} = < b_i, b_j > \quad i,j \in \{ -N, -(N-1), \ldots, N \} \]

This shows that \( i,j \)th element of \( G \) gives a measure of the "overlap" of the response of the \( i \)th and \( j \)th actuators. We have assumed that the plant has \( 2N + 1 \) identical actuators spaced equally across the sheet. Also, the response of each actuator is assumed to be symmetric.
and to spread over part, but not all, of the sheet so the response of an actuator will overlap with the responses of its neighbours. This implies that $\Gamma$ will be a banded matrix.

The width of the band of $\Gamma$ depends on the width of the actuator response and the separation between actuators. If response of the $i$th actuator overlaps with the response of the $(i+n)$th actuator, but does not overlap with the response of the $(i+n+1)$th actuator then the width of the band of $\Gamma$ is $2n+1$. Note that if the actuators are placed far apart or that the responses of the actuators are very narrow (or both) so that none of the responses overlap, $\Gamma$ reduces to a diagonal matrix.

As pointed out above, in order to solve the minimisation problem and to obtain the optimal set points, we require $\Gamma$ to be invertible. It can be shown that $\Gamma$ will be singular if the $b_k$'s $(k = -N, ..., N)$ are linearly dependent. If there is linear dependence between the $b_k$'s, then this means that at least one of the responses of the actuators can be expressed exactly as a linear combination of the responses of the other actuators. This implies that at least one of the actuators is redundant because its response can be generated from a combination of responses from the other actuators.

Intuitively, we would expect $\Gamma$ to be more ill-conditioned when the actuators are placed close together and their responses cover a wide part of the strip, resulting in a high degree of overlap between the responses. This implies that if the matrix $\Gamma$ has a very wide diagonal band, we would expect $\Gamma$ to be singular or near singular so we cannot position the actuators arbitrarily close together without encountering this singularity problem. The effect of the $\Gamma$ approaching singularity is that it tends to make the elements of $\Gamma^{-1}$ very large which, in turn, makes the size of the actuator set points very large. This point will be returned to in section 4.5.

Returning to equation 3.2.9, we see that the optimal set points are calculated by creating a vector containing the inner product of the desired response, $z_1$, with each of the actuator
responses, b_k, k={-N,...,N}. This vector is then premultiplied by the inverse of \( \Gamma \) to obtain the vector of optimal set points. From above, we know that \( \Gamma \) is a banded, symmetric matrix. \( \Gamma^{-1} \) will also be banded and symmetric, but it is likely that the width of the band in the inverse will be wider than the band of \( \Gamma \) [Duff, 1977] [Allgower, 1973]. This means that the optimal \( u_i \) is determined by \( <z_1, u_{i-n'}>, <z_1, u_{i-n'+1}>, \ldots, <z_1, u_{i+n'}> \) where \( n' \) is the width of the band in \( \Gamma^{-1} \).

We can see that the optimal set point, \( u_i \), depends not only upon the inner product of \( z_1 \) with the responses of the ith actuator, but also on the inner product of \( z_1 \) with the responses of the actuators around the ith actuator. Thus the set point for the ith actuator is determined by the value of \( z_1 \) over a wide range of the web, rather than its value in the region local to the response of the ith actuator. Intuitively, the reason for this slightly surprising result can be seen by noting that the response of the ith actuator overlaps with the response of its neighbours, so applying a change to \( u_i \) in order to match \( z_1 \) in the region of ith actuator affects the profile of the web in the region of the adjacent actuators. In order to compensate for these profile changes, the set points of these adjacent actuators have to be changed which in turn affects the profile in the regions of other actuators, including the ith actuator to which the change was initially applied. In the light of this argument, we can view equation 3.2.9 as a vector of set points being determined by a vector of interactions between the profile \( z_1 \) and the response of the corresponding actuator, with the matrix \( \Gamma^{-1} \) "undoing" the interaction between the actuators. The effect of this interaction was not appreciated in most of the early cross directional control schemes, so they were not very successful [Chen et al, 1986].
3.3 The Spatial Frequency Response of an Array of Actuators

The previous section developed a technique for obtaining the set points which will produce a response from the actuator array which matches the desired response, as closely as possible. In this section, we are interested in the size of the residual between the desired response and the achieved response. The residual between the achieved response of the actuator array and the desired response is

\[ e(x; u_k) = z_1(x) - \sum_{k=-N}^{N} b_k(x) u_k \]  

3.3.1

where we have written \( e(x; u_k) \) to indicate that the residual is a function of \( x \), but is also dependent upon the value of the \( u_k \)'s. When the optimal \( u_k \)'s are used (as obtained from equation 3.2.9) the value of \( \| e(x; u_k) \|_2 \) will be a minimum.

As we will see in chapter 4, when the actuator array is used as part of a feedback control system, \( z_1(x) \) is usually the actuator response that is required to cancel exactly a profile error. In this case, \( e(x; u_k) \) will be profile error that is left at the next sample point, after the control action has been applied (assuming that the profile error is due to a steady disturbance). Analysing the size of this residual is important, because it gives us a measure of the uncontrollable part of the profile error.

One way of calculating \( e(x; u_k) \) for a given \( z_1(x) \), would be to generate the optimal \( u_k \)'s (using equation 3.2.9) and then to substitute these values into equation 3.3.1 to give the residual. This approach requires a lot of calculation and the residual cannot be obtained without first calculating the optimal set points. In this section we develop an alternative approach by investigating the size of the residuals that are left when \( z_1(x) \) is a sinusoid. By calculating the size of the residual as the frequency of the sinusoid is increased, we can build up a spatial frequency response of the actuator array. Since any general function can be expressed as a sum of sinusoids, we can use this spatial frequency response to predict the extent to which the various frequency components of \( z_1(x) \) can be matched. For any
\(z_1(x)\), we can use this approach to give us the size of the residual, without having to calculate the optimal set points. The analysis that is required to set up the spatial frequency response of the actuator array also leads to a criterion for choosing the optimal actuator response and the optimal actuator spacing.

In keeping with the previous analysis, we will use the 2 norm to describe the size of the residual. Sometimes, it will be convenient to normalise this value with respect to the size of the desired actuator response, in which case we will use

\[
\frac{\| e(x; u_k) \|_2}{\| z_1(x) \|_2} \tag{3.3.2}
\]

as the measure of the size of the residual.

We have assumed that the responses of the actuators are identical, symmetrical about their centre line and equally spaced. For notational convenience, we will assume that the centre line of the actuator in the middle of the array is placed at \(x = 0\). If this actuator in the middle of the array has a response

\[
b_0 = b(x) \quad \text{where } b(x) \in L^2[-L, L] \tag{3.3.3}
\]

then, because the responses of each of the actuators are identical, the response of a finite array (which will be denoted by \(p(x; u_k)\)) can be written as

\[
p(x; u_k) = \sum_{k=-N}^{N} b(x-kd) u_k \tag{3.3.4}
\]

where \(d\) is the distance between the actuators. Again we have written \(p(x; u_k)\) to indicate that the response of the actuator array is a function of \(x\), but is also determined by the \(u_k\)'s.

We will assume that the response of each of the actuators, \(b(x-kd)\), is spatially limited, so that

\[
b(x - kd) = 0 \quad \text{for all } |x - kd| > \delta \tag{3.3.5}
\]

where \(\delta \in \mathbb{R}\) is the half width of the response. Since the response of a single actuator is
likely to spread over only part of the sheet, we will take $\delta \ll L$.

In practice, the response of the actuators is smooth, in the sense that $b(x-kd)$ is infinitely differentiable on $x \in [-L,L]$ [Rudin, 1974]. This means that $b(x-kd)$ decays smoothly to zero as $|x-kd| \to \infty$ (provided that the function includes a bounded support) so $b(x-kd)$ never actually attains the value zero for finite $x$. This means that, strictly speaking, $b(x-kd)$ will not be spatially limited. However, for $b(x-kd) \in C^\infty([-L,L])$, we can write

$$|b(x-kd)| < \varepsilon \quad \text{for all} \quad |x-kd| > \delta$$

for some $\varepsilon, \delta \in \mathbb{R}$. For small $\varepsilon$, this means that the response will be negligible beyond a distance $\delta$ from the centre line and we will assume that the effective spatial half width is $\delta$. Defining the half width, $\delta$, in this way allows us to quantify the spread of the effect of each of the actuators in the array.

These ideas can be illustrated by considering the Gaussian actuator response used in example 2.2.1. This has the form

$$b_k = b(x-kd) = \exp\left[-\frac{(x-kd)^2}{2\sigma^2}\right]$$

where $\sigma$ is the standard deviation of the Gaussian distribution. This function is continuously differentiable on $[-L,L]$ and decays to zero as $|x-kd|$ becomes large. For the Gaussian distribution it is well known that $|b(x-kd)| < 0.012$ when $|x-kd| > 3\sigma$, so if $\varepsilon$ is taken as 0.012 then the effective half width of the actuator response can be considered to be $3\sigma$ (see figure 3.3.1).

It should be noted, that, so far, we have described arrays with an odd number of identical actuators $(2N+1)$ arranged symmetrically about the centre of the sheet. While this restriction is not necessary, it does simplify the exposition of the subsequent analysis. The analysis could easily be extended to cases which do not have such simple symmetry. However, the analysis cannot handle arrays which do not have equally spaced actuators or actuators with non-identical responses. In these instances, the residual error has to be
Figure 3.3.1 The shape of the Gaussian actuator response, $b(x)$, with standard deviation, $\sigma$, used in example 2.2.1, showing the effective spatial half width, $\delta$, which is equal to $3\sigma$. 
calculated by generating the optimal set points from equation 3.2.9 and then using equation 3.3.1 to generate the residual.

Having refined our model of the spatial response of the actuator array, we are in a position to set up a spatial frequency response for the array which will predict the size of the residual component when the response of the array is required to match sinusoids of different frequencies. We will start by considering the artificial case of a plant producing a sheet of infinite width with an infinite array of actuators. The analysis is most easily applied to this case and the results that are obtained can then be adapted firstly to a plant of infinite width with a finite array of actuators and finally to the practical case of a finite width plant and a finite array of actuators.

For a plant and actuator array of infinite width, both \( L \) (the half width of the sheet) and \( N \) (the number of actuators on one side of the sheet) become infinite. The response of the array is changed to

\[
p'(x;u_k') = \sum_{k=-\infty}^{\infty} b(x-kd) u_k' \]  

3.3.8

where \( x \in (-\infty, \infty) \) and \( b(x-kd) \in L^2(-\infty, \infty) \) for \( k = \{..., -1, 0, 1, ... \} \). The primes, ' , on the total response \( p'(x;u_k') \) and the set points \( u_k' \), indicate that these variables refer to the infinite array.

Note that although each of the \( b(x-kd) \) are in \( L^2(-\infty, \infty) \), the infinite sum of their responses is only square integrable on \( x \in (-\infty, \infty) \) if all of the \( u_k' \) are finite and if \( u_k' \to 0 \) as \( |k| \to \infty \) (ie the set points of those actuators away from the centre of the array, tend to zero). It will turn out that this will be true if the required response of the array, \( z_1(x) \), also tends to zero as \( |x| \to \infty \). Unfortunately, we are considering the case where \( z_1(x) \) is a sinusoid, which does not go to zero as \( |x| \to \infty \), so neither \( z_1(x) \) nor the corresponding \( p'(x; u_k') \) will strictly be in \( L^2(-\infty, \infty) \). However, it is possible to multiply both \( z_1(x) \) and \( p'(x; u_k') \) by a slowly decaying function, to bring them into \( L^2(-\infty, \infty) \) and then to consider their behaviour

72
as the rate of decay \( \to 0 \). This problem will be discussed more fully below.

We will proceed by converting the minimisation problem of obtaining the optimal set points, into the spatial frequency domain. This will be done using the Fourier transform \( \mathcal{F}(.) \) [Rudin, 1974].

\[
\mathcal{F}(\omega) = \mathcal{F}[f(x)] = \int \limits_{-\infty}^{\infty} f(x) \, e^{-i\omega x} \, dx \tag{3.3.9}
\]

We will denote transformed functions in the spatial frequency domain by capital letters. The inverse Fourier transform, \( \mathcal{F}^{-1}(.) \) is given by

\[
f(x) = \mathcal{F}^{-1}[F(\omega)] = \frac{1}{2\pi} \int \limits_{-\infty}^{\infty} F(\omega) \, e^{i\omega x} \, d\omega \tag{3.3.10}
\]

It can be shown [Dym and MacKean, 1972, chapter 2] that if \( f(x) \in L^2(-\infty, \infty) \), then the Fourier transform exists and \( F(\omega) \in L^2(-\infty, \infty) \). Essentially, the Fourier transform is a Hilbert space isomorphism between the \( L^2 \) space defined on \( x \in (-\infty, \infty) \) and the \( L^2 \) space defined on \( \omega \in (-\infty, \infty) \) [Rudin, 1974]. One of the consequences of this is the Plancherel identity [Rudin, 1974]

\[
\| f(x) \|_2 = \| F(\omega) \|_2 \tag{3.3.11}
\]

Applying this to the residual of an infinite array, \( e'(x; u_k) \) gives

\[
\| e'(x; u_k') \|_2 = \| E'(\omega; u_k') \|_2 \tag{3.3.12}
\]

\[
= \| Z_1(\omega) - P'(\omega; u_k') \|_2 \tag{3.3.13}
\]

where \( E'(\omega; u_k'), Z_1(\omega) \) and \( P'(\omega; u_k') \) are the Fourier transforms of \( e'(x; u_k') \), \( z_1(x) \) and \( p'(x; u_k') \) respectively.

Let us consider the Fourier transform of the response of an infinite array of actuators, \( P'(\omega; u_k') \).
\begin{align*}
P'(\omega; u_k') &= \mathcal{F} \left[ p'(x; u_k') \right] \quad 3.3.14 \\
&= \mathcal{F} \left[ \sum_{k=-\infty}^{\infty} b(x-kd) u_k' \right] \quad 3.3.15 \\
&= B(\omega) \sum_{k=-\infty}^{\infty} u_k' e^{-ik\omega} \quad 3.3.16
\end{align*}

where \( B(\omega) = \mathcal{F}[b(x)] \) and we have used the identity [Riley, 1974].

\[ \mathcal{F}[b(x-kd)] = e^{-ik\omega} \mathcal{F}[b(x)] = e^{-ik\omega} B(\omega) \quad 3.3.17 \]

We can see that \( P'(\omega; u_k') \) consists of two parts:

(i) \( B(\omega) \), the Fourier transform of the response of a single actuator. This can be considered as the spatial frequency response of an actuator;

(ii) the term

\[ \sum_{k=-\infty}^{\infty} u_k' e^{-ik\omega} \]

We will now examine the second term in more detail by developing the following theorem.

**Theorem 3.3.1** There exists a function, \( u'(x) \in L^2(-\infty, \infty) \) with a Fourier transform, \( U'(\omega) \in L^2(-\infty, \infty) \). We will assume that the Fourier transform, \( U'(\omega) \), is band limited so that \( U'(\omega) = 0 \) for all \( |\omega| \geq \frac{1}{2}\omega_s \), where \( \omega_s = \frac{2\pi}{d} \), is the repeat frequency of the actuators. The following two statements hold:

(i) the actuator set points, \( u_k' \), in equation 3.3.16 are determined by taking samples of this function, \( u'(x) \), at \( x=kd \), for \( k=\{..,-1,0,1,...\} \) (ie the samples are taken at the positions of the central lines of the actuators);

(ii) the \( u_k' \) are the coefficients of a Fourier series expansion of a periodic extension of period \( \omega_s \), of \( U'(\omega) \) (to within a scaling factor).
Proof Given \( u'(x) \in L^2(-\infty, \infty) \), then \( U'(\omega) \) is given by

\[
U'(\omega) = \mathcal{F} [u'(x)] = \int_{-\infty}^{\infty} u'(x) e^{-i\omega x} \, dx \quad 3.3.18
\]

Similarly, \( u'(x) \) is given by

\[
u'(x) = \mathcal{F}^{-1} [U'(\omega)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} U'(\omega) e^{i\omega x} \, d\omega \quad 3.3.19
\]

This means that the samples of \( u'(x) \) at the points \( x=kd, \, k=\{..., -1, 0, 1, ...\} \), are given by

\[
u'(x=kd) = \frac{1}{2\pi} \int_{-\infty}^{\infty} U'(\omega) e^{ik\omega_0} \, d\omega \quad 3.3.20
\]

where we have simply evaluated equation 3.3.19 at the sample points.

We have taken \( U'(\omega) = 0 \) for all \( \omega \) such that \( |\omega| \geq \frac{1}{2}\omega_s \), so \( U'(\omega) \) is zero outside the frequency range \( \omega \in [-1/2\omega_s, 1/2\omega_s] \). Given the periodic extension of \( U'(\omega) \), with \( \frac{1}{\omega_s} \) periodic, which we will write as

\[
u_s'(\omega) = \frac{1}{d} \sum_{k=-\infty}^{\infty} U(\omega + k\omega_s) \quad 3.3.21
\]

where the \( \frac{1}{d} \) term is a scaling factor. This function is illustrated in figure 3.3.2. \( U_s'(\omega) \), being periodic, has the Fourier series

\[
u_s'(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{-ik\omega_0} \quad 3.3.22
\]

where \( c_k, \, k=\{..., -1, 0, 1, ...\} \) are the Fourier coefficients given by

\[
c_k = \frac{1}{\omega_s} \int_{-\frac{1}{2}\omega_s}^{\frac{1}{2}\omega_s} e^{ik\omega_0} U_s'(\omega) \, d\omega \quad 3.3.23
\]

Over the range of this integration, \( \omega \in [-1/2\omega_s, 1/2\omega_s] \), \( U_s'(\omega) = \frac{1}{d} U'(\omega) \), (from the definition of \( U_s'(\omega) \) in equation 3.3.21) so the expression for the Fourier coefficients can be written as
Figure 3.3.2  A bandlimited function, $U'(\omega)$, and its periodic extension, $U_s'(\omega)$. Note that $U_s'(\omega)$ has been scaled by a factor $1/d$, where $d$ is the separation between the actuators.
\[ c_k = \frac{1}{\omega d} \int_{-\frac{\omega}{2}}^{\frac{\omega}{2}} e^{ikd\omega} U'(\omega) \, d\omega \]

3.3.24

\[ = \frac{1}{2\pi} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{ikd\omega} U'(\omega) \, d\omega \]

3.3.25

since \( \omega_s = \frac{2\pi}{d} \).

We can compare equation 3.3.25 with the expression for the samples of \( u'(x) \) (equation 3.3.20)

\[ u'(x=kd) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikd\omega} U'(\omega) \, d\omega \]

3.3.26

\[ = \frac{1}{2\pi} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} e^{ikd\omega} U'(\omega) \, d\omega \]

3.3.27

where the limits of the integration can be reduced because \( U'(\omega) \) is non zero only in the range \( |\omega| \leq \frac{1}{2}\omega_s \). From equations 3.3.25 and 3.3.27, we can see that

\[ c_k = u'(x=kd) \]

3.3.28

so the Fourier series of \( U_s'(\omega) \) can be written as

\[ U_s'(\omega) = \sum_{k=-\infty}^{\infty} u'(x=kd) \, e^{-ikd\omega} \]

3.3.29

by substituting for \( c_k \) in equation 3.3.22. This links the two statements in the theorem.

This expression is very similar to the second part of \( P'(\omega; u_k') \) in equation 3.3.16. If we write \( u'(x=kd) = u_k' \), then

\[ U_s'(\omega) = \sum_{k=-\infty}^{\infty} u_k' \, e^{-ikd\omega} \]

3.3.30

So, if we identify the actuator set points, \( u_k' \), with the samples of the function \( u'(x) \) at
x = kd, we can replace the second part of the expression for \( P'(\omega; u_k') \) by \( U_s'(\omega) \).

This ends the proof.

The above proof is based on the Generalised Sampling Theorem [Jerri, 1977]. \( U_s'(\omega) \) is the Fourier transform of the sampled signal \( u'(x) \). The factor, \( \frac{1}{d} \), in the periodic extension of \( U'(\omega) \) (equation 3.3.21) is important and we will return to it later. Astrom and Wittenmark [Astrom and Wittenmark, 1984, section 2.3] remark that, because of this scaling factor, the sampling process has a gain of \( \frac{1}{d} \).

We can now simplify the expression for the frequency response of an infinite array of actuators. From equation 3.3.16

\[
P'(\omega; u_k') = B(\omega) \sum_{k=-\infty}^{\infty} u_k' e^{-ik\omega}
\]

where we have substituted in equation 3.3.30. This shows that the frequency response of an array of actuators is the product of the frequency responses of a single actuator and the periodic function, \( U_s'(\omega) \).

Up to this point we have not made any assumptions about \( B(\omega) \). We have taken \( b(x) \) to be smooth (in the sense that it is continuously differentiable) and to be spatially limited (in the sense that \( |b(x)| \) is negligible beyond a certain distance from the centreline of the actuator). Since \( b(x) \) is assumed to be continuously differentiable, it can be shown that \( B(\omega) \) is "rapidly decreasing" in the sense that \( \omega^n B(\omega) \in L^2(-\infty, \infty) \) for every \( n \geq 0 \). For \( \omega^n \) \( B(\omega) \) to be square integrable on \( \omega \in (-\infty, \infty) \), this means that \( |B(\omega)| \to 0 \) as \( |\omega| \to \infty \) [Dym and MacKean, 1972, section 2.3]. This implies that \( B(\omega) \) is effectively bandlimited, because there will exist an \( \epsilon \) and an \( \omega_c \) such that

\[
|B(\omega)| < \epsilon \quad \text{for all } |\omega| > \omega_c
\]

We will choose the limit \( \omega_c \) so that the size of \( \epsilon \) is negligible in a similar manner to the
choice of the effective spatial band limit. We can then assume that $|B(\omega)| \equiv 0$ for all $|\omega| > \omega_c$.

This concept can be illustrated by returning to the actuator with a Gaussian response in Example 2.2.1, given by

$$b(x) = \exp\left[ -\frac{x^2}{2\sigma^2} \right]$$  \hfill (3.3.34)

As pointed out above this response has an effective spatial "half-width" of $3\sigma$. The corresponding $B(\omega)$ for this response is

$$B(\omega) = \sqrt{2\pi} \sigma \exp\left[ -\frac{\omega^2 \sigma^2}{2} \right]$$  \hfill (3.3.35)

This is another Gaussian curve with standard deviation $1/\sigma$. Arguing as before, since $B(\omega)$ is negligible outside three standard deviations from $\omega=0$, we can say that the effective band limit, $\omega_c$, is

$$\omega_c = \frac{3}{\sigma}$$  \hfill (3.3.36)

In fact, it is not possible to have a response which is strictly band limited in both the spatial domain and the frequency domain. If the response is strictly band limited in one domain, then the transformed response will have infinite bandwidth. This is a consequence of the Generalised Uncertainty Principle [Papoulis, 1962], [Bracewell, 1978]. However, as an approximation for practical systems, we will use the definition of bandwidth (in both the spatial domain and the spatial frequency domain) as being the values of $x$ and $\omega$ for beyond which $|b(x)|$ and $|B(\omega)|$ is negligible.

At this stage, our discussion can be divided between two separate cases, which depend upon the relationship between the frequency bandwidth of an actuator, $\omega_c$ and the actuator separation frequency, $\omega_s = 2\pi/d$:

(i) $B(\omega)$ has a bandwidth, $\omega_c$, which satisfies $\omega_c < 1/2\omega_s$;

(ii) $\omega_c \geq 1/2\omega_s$ (This includes the case where $B(\omega)$ has infinite bandwidth).
It will become apparent that the difference between these two cases is determined by whether spatial aliasing occurs in the system. For this reason, case (i) will be referred to as the non-aliased case and case (ii) as the aliased case. We will consider case (i) first and will return to case (ii) later.

**Case (i): The Non-Aliased Case**

From above (equation 3.3.32) we know that the response of the array of actuators can be written as

\[
P'(\omega; u_k) = B(\omega) \ U'_s(\omega)
\]

which is the product of the bandlimited response of a single actuator and a periodic function, \(U'_s(\omega)\). For case (i), then \(|B(\omega)| = 0\) over the range \(|\omega| > \omega_c\) with \(\omega_c < \frac{1}{2}\omega_s\). This makes the response of the whole array

\[
P'(\omega; u_k) = B'(\omega) \ U'(\omega) \quad \text{for} \quad |\omega| \leq \omega_c
\]

\[
= 0 \quad \text{for} \quad |\omega| > \omega_c
\]

(assuming that \(U'_s(\omega)\) is finite). This is illustrated in figure 3.3.3. Because we have assumed that \(\omega_c < \frac{1}{2}\omega_s\) and from the definition of \(U'_s(\omega)\) (equation 3.3.21) we know that

\[
U'_s(\omega) = \frac{1}{d} U'(\omega) \quad \text{for} \quad \omega \in \left[ -\frac{1}{2}\omega_s, \frac{1}{2}\omega_s \right]
\]

so \(P'(\omega; u_k)\) becomes

\[
P'(\omega; u_k) = \frac{1}{d} B(\omega) \ U'(\omega) \quad \text{for} \quad |\omega| < \omega_c
\]

\[
= 0 \quad \text{for} \quad |\omega| \geq \omega_c
\]

This equation can now be substituted into the expression for the size of the residual. From equation 3.3.13, we can write

\[
\| e'(x; u_k) \|_2 = \| Z_1(\omega) - P'(\omega; u_k) \|_2
\]

\[
= \| Z_1(\omega) - \frac{1}{d} B(\omega) \ U'(\omega) \|_2
\]

This expression shows us how to obtain the size of the residual error from the frequency
Figure 3.3.3 The frequency response of an infinite array of actuators, $P'(\omega;u_k')$, is the product of the frequency response of a single actuator, $B(\omega)$, and a periodic function, $U_s'(\omega)$. 
domain rather than the spatial domain.

From Theorem 3.3.1 we know that the $u_k'$ are the samples of the function $u'(x)$ at $x = kd$ for $k=\{..., -1, 0, 1, ...\}$, so there is an alternative to the approach described in Section 3.2 for choosing the set points, $u_k'$, which will minimise $||e'(x; u_k')||_2$

(i) for a given $Z_1(\omega)$, choose the $U'(\omega)$ which minimises expression for the size of the residual (note that the minimisation can be only be carried out in the region where $B(\omega)$ is non-zero, ie $|\omega| \leq \omega_c$);

(ii) inverse Fourier transform to obtain the optimal $u'(x)$;

(iii) sample $u'(x)$ at $x=kd$, $k=\{..., -1, 0, 1, ...\}$, to obtain $u_k$.

For general $z_1(x)$, there is no advantage in calculating the set points, $u_k'$, in this manner, because the approach developed in section 3.2, using the spatial domain, is simpler. However because the two approaches are equivalent, the size of the minimal residual will be the same in both cases. For certain $z_1(x)$ which have simple Fourier transforms, this allows us to predict the size of this minimal residual, using equation 3.3.32, without having to calculate the set points first. In particular we can use this equation to predict $||e'(x; u_k')||_2$ when $z_1(x)$ is a sinusoid, because the Fourier transform of a sinusoid has a simple form. If $z_1(x)$ is a cosine wave at frequency, $\omega_1$, such that

$$z_1(x) = \cos \omega_1 x$$

then

$$Z_1(\omega) = \frac{1}{2} \delta (\omega - \omega_1) + \frac{1}{2} \delta (\omega + \omega_1)$$

where $\delta(.)$ is the Dirac delta function [Bracewell, 1978]. Similarly, if $z_1(x)$ is a sine wave at frequency, $\omega_1$, then

$$z_1(\omega) = \sin \omega_1 x$$

and

$$Z_1(\omega) = i \left[ \frac{1}{2} \delta (\omega - \omega_1) - \frac{1}{2} \delta (\omega + \omega_1) \right]$$

The Fourier transform of any sinusoid of frequency, $\omega_1$, with arbitrary amplitude and
phase, can be obtained by taking the appropriate linear combination of the expressions in
equations 3.3.44 and 3.3.46.

Before proceeding, it is necessary to return to a technical difficulty which was mentioned
earlier. Strictly, neither a sinusoid nor its Fourier transform have finite square integrals on
the region \( x \in (-\infty, \infty) \). Since our expression for the size of the residual is given by
\[
\| e'(x; u_k) \|_2 = \left\| z_1(\omega) - \frac{1}{d} B(\omega) U'(\omega) \right\|_2
\]
we cannot guarantee that \( \| e'(x; u_k) \|_2 \) can be evaluated when \( z_1(x) \) is a sinusoid.

However, multiplying a sinusoid by a slowly decaying function, such as a Gaussian
distribution of unit square integral and a very large standard deviation \( \sigma \), brings the product
into \( L^2(-\infty, \infty) \) [Lighthill, 1962, chapter 5]. The corresponding Fourier transform of this
product is the convolution of the two delta functions given in equations 3.3.44 and 3.3.46
and an extremely tall and narrow Gaussian distribution, with both the amplitude and
standard deviation being proportional to \( 1/\sigma \). This is illustrated in figure 3.3.4. As the
width of the modulating Gaussian distribution is increased (ie \( \sigma \to \infty \)) the Fourier
transform approaches the two delta functions which were the transform of the original
sinusoid. In future, we will consider a sinusoid as being modulated by a Gaussian
distribution which is so wide enough so that \( 1/\sigma \) is arbitrarily small. This allows us to
make both the sinusoid and its transform square integrable.

It should be noted in figure 3.3.4 that we have represented a delta function be an arrow
whose height corresponds to the coefficient in from of the delta function. In the figure, the
two delta functions have a height of \( 1/2 \), representing the terms in equation 3.3.44. This
form of illustration will be used throughout this study.

We are now in a position to investigate the size of the residual \( \| e'(x; u_k) \|_2 \) when \( z_1(x) \)
is a sinusoid. For the moment, we will concentrate on the case where \( z_1(x) \) is a cosine
Figure 3.3.4 The top line of graphs shows the product of a cosine wave, \( \cos(\omega_1 x) \), and a Gaussian function with a large standard deviation, \( \sigma \). The lower line of graphs shows the frequency domain version of this product, which is the convolution of two delta functions of magnitude, \( 1/2 \), at \( \omega = \pm \omega_1 \) and a narrow Gaussian function with standard deviation, \( 1/\sigma \). Note that the delta functions are represented by arrows whose height corresponds to their magnitudes.
wave of frequency $\omega_1$ and unit amplitude.

$$z_1(x) = \cos \omega_1 x$$

$$Z_1(\omega) = \frac{1}{2} \delta(\omega-\omega_1) + \frac{1}{2} \delta(\omega+\omega_1)$$

From equation 3.3.32, we know that $\| e'(x; u) \|_2$ is given by

$$\| e'(x; u_k) \|_2 = \| Z_1(\omega) - \frac{1}{d} B(\omega) U'(\omega) \|_2$$

For $z_1(x)$ being a cosine wave, the size of the residual is determined by how well the term

$$P'(\omega; u_k) = \frac{1}{d} B(\omega) U'(\omega)$$

matches the two delta functions at $\omega = \pm \omega_1$, in equation 3.3.50.

The first thing to be noticed is that, since $|B(\omega)| = 0$ for $|\omega| > \omega_c$, then $P'(\omega; u_k) = 0$ for $|\omega| > \omega_c$ (assuming that $U'(\omega_1)$ is finite). So any cosine wave with a frequency greater than $\omega_c$ cannot be matched at all and

$$\| e'(x; u_k) \|_2 = \| Z_1(\omega) \|_2 = \| z_1(x) \|_2$$

for $z_1(x) = \cos \omega_1 x$ with $\omega_1 > \omega_c$

If $\omega_1 \leq \omega_c$, then the two delta functions of $Z_1(\omega)$ can be matched by making $U_1(\omega)$ two delta functions of magnitude $1/2$, scaled by the factor

$$\frac{d}{B(\omega-\omega_1)}$$

This will make

$$P'(\omega; u_k) = \frac{1}{d} B(\omega) \left\{ \frac{d}{B(\omega-\omega_1)} \left[ \frac{1}{2} \delta(\omega-\omega_1) + \frac{1}{2} \delta(\omega+\omega_1) \right] \right\}$$

$$= \frac{1}{2} \delta(\omega-\omega_1) + \frac{1}{2} \delta(\omega+\omega_1)$$

$$= Z_1(\omega)$$

where we have used the fact that that $P'(\omega; u_k)$ is non zero only when $\omega = \omega_1$ because of the properties of the $\delta$ functions. Equation 3.3.57 implies that a cosine wave of frequency $\omega < \omega_c$ can be matched exactly by the response of an infinite array of actuators.
The actuator set points for achieving this match can be obtained by noting that

\[ U'(\omega) = \frac{d}{B(\omega=\omega_1)} \left[ \frac{1}{2} \delta(\omega-\omega_1) + \frac{1}{2} \delta(\omega+\omega_1) \right] \] 3.3.58

so that

\[ u'(x) = \frac{d}{B(\omega=\omega_1)} \cos \omega_1 x \] 3.3.59

\[ u'_k = u'(x=kd) = \frac{d}{B(\omega=\omega_1)} \cos \omega_1 kd \quad \text{for } k = \{\ldots,-1,0,1,\ldots\} \] 3.3.60

This shows that the pattern of the set points follows the shape of the profile to be matched, but their amplitude is proportional to the separation between actuators, \(d\), and is dependent upon the size of the component at \(\omega_1\) in the frequency response of a single actuator.

Intuitively, this answer makes sense. We would expect the actuator set points to be positive in regions where the response was required to be positive and negative in regions where the response needed to be negative. Also, if the required component of the frequency response of each of the actuators, is small, then it is likely that the amplitudes of the set points will need to be large to compensate for this lack of power in response.

The dependence of the set points' amplitude upon the actuator separation, \(d\), can be seen by considering the case where \(d\) is halved, so that there are now twice as many actuators per unit length in the array. Looking at equation 3.3.60, we can see that this means that the value of the set points will be halved. This can be understood because there are now two actuators are covering the area of web where there was previously one actuator, so each of the setpoints need to be halved to produce the same effect. Note that although the size of the individual set points are halved, the total size of the set points (ie the sum of the absolute values of the set points) does not change because there are now twice as many actuators in the array.

The price that is paid for reducing the actuator separation (and hence reducing the size of
the set points) is that the response of the array increases because
\[
P'(\omega; u_k) = \frac{1}{d} B(\omega) U'(\omega) \quad \text{for all } |\omega| < \omega_c \tag{3.3.61}
\]
If \(d \to 0\) then \(P'(\omega; u_k) \to \infty\) and the frequency response of the array "blows up". This is the frequency domain analogue of the ill-conditioning of the Gram matrix as the actuator separation is reduced, which was discussed in section 3.2.

These ideas that have been developed for making the response of the actuator array match a cosine wave are illustrated in figure 3.3.5. In this illustration we assume that we have an infinite array of actuators, each with a Gaussian frequency response as in figure 3.3.5a. Figure 3.3.5b shows part of the desired response of the array, which is a cosine wave with a low frequency \(\omega_1\), and the actuator set points that are required to generate this response exactly, (it is assumed that \(\omega_c < 1/2 \omega_0\)). We can see that the pattern of the actuator set points follows the desired response. If the frequency of the desired response is increased to \(\omega_2\) (while keeping the amplitude the same) then the optimal set points are shown in figure 3.3.5c. Again the pattern of the set points follows the desired response, but the amplitudes of the set points are increased because
\[
B(\omega_2) < B(\omega_1) \tag{3.3.62}
\]
In figure 3.3.5d, the desired response is the same as in figure 3.3.5b, but the actuator separation has been halved. Consequently the amplitudes of the individual set points are also halved.

The above analysis has been applied to cosine waves, but it can easily be adapted to sine waves by replacing the frequency response of the cosine wave
\[
\frac{1}{2} \delta (\omega-\omega_1) + \frac{1}{2} \delta (\omega+\omega_1) \tag{3.3.63}
\]
in equation 3.3.55, with the response of a sine wave
\[
i \left[ \frac{1}{2} \delta (\omega-\omega_1) - \frac{1}{2} \delta (\omega+\omega_1) \right] \tag{3.3.64}
\]
Following the analysis through, we can see that any response which is a sine wave with a frequency \(\omega_1 < \omega_c\) can also be matched. By taking a linear combination of a sine and a
Figure 3.3.5 Assume that we have an infinite array of actuators, each of which has the Gaussian response shown in figure (a). The left hand side of figure (b) shows the desired response which is a cosine wave of frequency, $\omega_1$. The right hand side shows the actuator set points which are required to achieve this response. Figure (c) shows the set points that are required to match a cosine response of the same amplitude but a higher frequency, $\omega_2$. Figure (d) shows the set points that are required to match the same response as in figure (b) when the separation between the actuators is halved.
cosine the argument can be extended to a sinusoid of arbitrary phase. Again, the pattern of the set points will match the shape of the desired response.

From equation 3.3.60, we know that the set points for matching a cosine wave of frequency, \( \omega_1 \), are given by

\[
u'_k = \frac{d}{B(\omega=\omega_1)} \cos \omega_1 kd \quad k = \ldots, -1, 0, 1, \ldots \tag{3.3.65}
\]

If there is some \( \omega < \omega_c \) for which \( B(\omega) = 0 \), then these responses at this frequency cannot be matched without using infinite actuator set points. An example of an actuator frequency response with this property is shown in figure 3.3.6. In addition, in regions of \( \omega \) where \( |B(\omega)| \) is very small, the actuator set points will be very large. This is likely to be the case for frequencies approaching \( \omega_c \) and any other regions around where \( B(\omega) = 0 \). In practice, the constraints of the system will limit the size of the actuator set points, so sinusoids whose frequencies are in this region will not be matched completely by the response of the array.

We can summarise the response of the array of actuators for the case where the bandwidth of each actuator, \( \omega_c \), is less than \( \frac{1}{2} \omega_s \). If the desired response is a sinusoid of frequency, \( \omega_1 \leq \omega_c \), then it can be matched exactly by the response of the array, irrespective of its phase. Thus

\[
\| e'(x;u'_k) \|_2 = 0 \quad \text{for} \quad \omega_1 \leq \omega_c
\]

3.3.66

If the frequency of the desired response is above \( \omega_c \), it cannot be matched at all by the response of the array and

\[
\| e'(x;u'_k) \|_2 = 1 \quad \text{for} \quad \omega_1 > \omega_c
\]

3.3.67

In general, the desired response will be some combination of sinusoidal components, in which case the response of the actuator array will be able to match the components with frequencies below \( \omega_c \), but it will not be able to match higher frequency components. In the language of section 3.1, the low frequency components lie in the controllable subspace,
Figure 3.3.6  An example of an actuator frequency response, $B(\omega)$, for which $B(\omega) = 0$ for some $\omega$ in the region $\omega < \omega_c$. In this case, $B(\omega_1) = 0$. 
while the higher frequency components are in the uncontrollable subspace.

Case (ii): The Aliased Case

The discussion above has concentrated on the non-aliased case, where the band width of the response of the actuators, $\omega_a$, is less than half of the frequency of the actuator spacing, $\omega_s$. As pointed out previously, Theorem 3.3.1 is a version of the Generalised Sampling Theorem, so we can identify $1/2\omega_s$ as the Nyquist frequency of the actuator separation. We now wish to investigate the situation where the bandwidth, $\omega_c$, is greater than this Nyquist frequency. This corresponds to case (ii) above.

As before, the frequency response of the actuator array is

$$P'(\omega u_k) = B(\omega) \quad U_s'(\omega) \quad |\omega| \leq \omega_c$$

$$= 0 \quad |\omega| > \omega_c$$

which is the product of the frequency response of a single actuator and the frequency response of a sampled function. Although $B(\omega)$ now has components at frequencies greater than $1/2\omega_s$, the frequency response of the array is still limited by $\omega_c$. This means that any desired response which includes frequency components greater than $\omega_c$ still cannot be matched exactly. However, from the arguments above, we would expect that components in the range $1/2\omega_s \leq \omega < \omega_c$ could be matched by the response of the array. To illustrate this, consider a cosine wave with a frequency $\omega_1$, in this range. From above, the pattern of the optimal set points, $u_k'$, will follow the shape of the desired response and their size will be scaled by a factor

$$\frac{d}{B(\omega=\omega_1)}$$

We know that the set points can be considered as the samples of the function $u'(x)$, which will be a scaled cosine wave of frequency, $\omega_1$. $U_s'(\omega)$ is the frequency response of the sampled version of this function. Unfortunately, if $\omega_1 > 1/2\omega_s$ this sampled function will suffer from spatial aliasing. This means that the actuator set points for matching a cosine
response of frequency, \( \omega_1 \), are identical to the set points for matching a response at the aliasing frequency, \( \omega_s - \omega_1 \) (see figure 3.3.7).

Figure 3.3.8 illustrates this effect in the frequency domain. The desired response is shown on the left of the figure. The achievable response (which is the product of the frequency response of a single actuator and the frequency response of the sampled function) is shown on the right. We can see that it is impossible to match the desired response at frequency \( \omega_1 \) without introducing a component at \( \omega_s - \omega_1 \). This means that the effectiveness of the actuator array is reduced, because, unlike the case where \( \omega_c < \frac{1}{2} \omega_s \), the array cannot match exactly all responses with a frequency \( \omega_1 < \omega_c \).

It is illuminating to consider the spatial frequency response of the array when it is required to match firstly a cosine wave of increasing spatial frequency and later a sine wave. We start by attempting to match a response with a low frequency, with \( \omega_c < (\omega_s - \omega_1) \). This is illustrated in figure 3.3.9a. Because the "aliased" component at frequency \( (\omega_s - \omega_1) \) is outside the bandwidth of \( B(\omega) \), it does not contribute the response of the array, so the desired response can be matched exactly. If the bandwidth of \( B(\omega) \), \( \omega_c \), is so large that \( \omega_c > \omega_s \) then there will be no responses which can be matched exactly, however low the frequency, because there will always be a component in the response at \( (\omega_s - \omega_1) \) (See figure 3.3.9b).

Returning to the case where the bandwidth of \( B(\omega) \) is not quite as large (ie where \( \frac{1}{2} \omega_s < \omega_c < \omega_s \)) we consider a desired response of frequency \( \omega_1 \) where \( \omega_1 < \frac{1}{2} \omega_s \), but \( \omega_s - \omega_1 < \omega_c \) so there is aliasing. (See figure 3.3.9c). To minimise the size of the residual, we have to balance matching the component at \( \omega_1 \) against the introduction of the unwanted component at \( (\omega_s - \omega_1) \). In this example, with a Gaussian actuator distribution, \( B(\omega=\omega_1) \) is much larger than \( B(\omega=\omega_s - \omega_1) \) so the match is relatively good. This should be contrasted with the case where the desired response has a higher frequency, \( \frac{1}{2} \omega_s < \omega_1 < \omega_c \), as in
Figure 3.3.7  The effect of aliasing. The upper graph shows the actuator set points (represented by •) that are required to generate a response which matches a cosine wave of frequency, $(\omega_s - \omega_1)$ (indicated by the solid curve). The lower graph shows that the same actuator set points are required to match a cosine response at the lower frequency of $\omega_1$, assuming that $B(\omega_s - \omega_1)$ is equal to $B(\omega_1)$ (if this is not the case then the graphs can be scaled to compensate for the difference between $B(\omega_s - \omega_1)$ and $B(\omega_1)$). The cosine response at the lower frequency is shown as a dashed line on the upper graph.
Figure 3.3.8 The effect of aliasing in the spatial frequency domain when generating a response which is a cosine wave of frequency, $\omega_1$. The Fourier transform of the desired response is two delta functions at $\omega = \pm \omega_1$, as shown in the graph on the left. The achieved response is shown on the right as the product of the response of a single actuator, $B(\omega)$, and the periodic function, $U_4'(\omega; u_k)$. Because $\omega_c$ for the actuator response is greater than $1/2\omega_9$, it can be seen that it is impossible to match the response at $\omega = \omega_1$, without introducing a component at $\omega = \omega_9 - \omega_1$. 
Figure 3.3.9 The effect of aliasing when matching cosine responses of different frequencies, $\omega_1$. In each case, the desired frequency response is shown on the left and the achieved response, which is the product of the frequency response of a single actuator, $B(\omega)$, and a periodic function, $U_\alpha(\omega;u_k)$, is shown on the right. In figure (a), $\omega_1$ is low so that $\omega_c < (\omega_s - \omega_1)$ (where $\omega_c$ is the bandwidth of $B(\omega)$). In figure (b), the frequency of the desired response is the same as in (a), but the bandwidth of the response, $\omega_c$, has increased so that $\omega_c > \omega_s$. 
Figure 3.3.9 (cont.) In figure (c), the bandwidth, \( \omega_c \), is the same as in (a), but the frequency of the desired response is increased so that \( \omega_c > (\omega_s - \omega_1) \), but \( \omega_1 < \frac{1}{2} \omega_s \). In figure (d), \( \omega_1 \) is increased still further so that \( \omega_c > (\omega_s - \omega_1) \) and \( \omega_1 > \frac{1}{2} \omega_s \).
Figure 3.3.9 (cont.) In figure (e), the frequency of the desired response, $\omega_1$, equals $1/2\omega_s$. 
Here, \( B(\omega=\omega_1) \) is much smaller than \( B(\omega=\omega_s-\omega_1) \) so only a small fraction of the desired response can be achieved because matching the desired response exactly would introduce a large component at \( \omega = \omega_s - \omega_1 \). For an actuator frequency response, \( B(\omega) \), which decreases monotonically with increasing \( \omega \), as in this example, the size of minimal residual will increase monotonically with increasing \( \omega \).

There is, however, an anomaly when trying to match a cosine response at a frequency \( \omega_1 = \frac{1}{2}\omega_s \). At this frequency the delta function due to the component at \( \omega_1 \) coincides with the delta function due to the aliasing component at \( \omega_s - \omega_1 \) (see figure 3.3.9e). This means that the desired response can be achieved without introducing any unwanted frequency components, so \( ||e'(x;u_k')||_2 \) is zero and the desired response can be matched exactly.

The discussion above has concentrated on matching response which are cosine waves. We need to consider sine wave responses as well. From equation 3.3.46, we know that if the desired response is a sine wave a frequency, \( \omega_1 \), then the frequency domain equivalent of this response is

\[
Z_1(\omega) = i \left[ \frac{1}{2} \delta(\omega-\omega_1) - \frac{1}{2} \delta(\omega+\omega_1) \right]
\]

which is a positive imaginary delta function at \( \omega=\omega_1 \) and a negative imaginary delta function at \( \omega=-\omega_1 \). For general sinewave frequencies, the response of the array is similar to that of a cosine response. For example, in figure 3.3.10a, the desired response at \( \omega=\omega_1 \) cannot be matched without introducing a (negative) component at \( \omega = \omega_s - \omega_1 \). However, at \( \omega_1 = \frac{1}{2}\omega_s \), the component of the actuator array's response at \( \omega=\omega_1 \) cancels with the aliasing component at \( \omega = \omega_s - \omega_1 \) (unlike the cosine case where the two components add together). This means that there is no sine wave component of the response of the array at \( \omega_1 = \frac{1}{2}\omega_s \) and the required response at this frequency will not be achieved at all (see figure 3.3.10b).

Some insight into this anomalous behaviour can be obtained by noting that \( \omega = \frac{1}{2}\omega_s \) is the Nyquist frequency of the actuator spacing. Looking at figure 3.3.11, we can see that a
Figure 3.3.10 The effect of aliasing when matching sine responses of different frequencies, $\omega_1$. In both cases, the desired frequency response is shown on the left and the achieved response, which is the product of the frequency response of a single actuator, $B(\omega)$, and a periodic function, $U_s(\omega; u_k)$, is shown on the right. In figure (a), the frequency of the desired response, $\omega_1$, satisfies $\omega_1 > (\omega_0 - \omega_1)$, but $\omega_1 < \frac{1}{2}\omega_0$. In figure (b), $\omega_1$ equals $\frac{1}{2}\omega_0$. 
Figure 3.3.11 The top graph shows the actuator set points (represented by •) for matching the desired response (shown as a solid curve) of a cosine wave at frequency $\omega_1 = 1/2 \omega_1$ ($B(\omega=1/2 \omega_1)$ is taken to be unity). The lower graph shows the corresponding set points for matching a sinewave at the same frequency.
cosine wave at the Nyquist frequency can be matched well by the actuator set points, but a sine wave at the same frequency goes through zero at \( x = kd, k = \{-1, 0, 1, \ldots\} \) so the set points will be zero.

We can quantify the effect of the aliasing component by evaluating the relative size of the error between the desired response and the achieved response. Returning to the case of a cosine wave of frequency \( \omega_1 \), we know that

\[
Z_1(\omega) = \frac{1}{2} \delta(\omega - \omega_1) + \frac{1}{2} \delta(\omega + \omega_1)
\]

If there is no aliasing component then we can choose a \( U'(\omega) \) which minimises

\[
\| Z_1(\omega) - P'(\omega; u_k) \|_2 = \| Z_1(\omega) - \frac{1}{d} B(\omega) U'(\omega) \|_2
\]

From equation 3.3.58 we see that the size of the error can be made equal to zero by making

\[
U'(\omega) = \rho \left[ \frac{1}{2} \delta(\omega - \omega_1) + \frac{1}{2} \delta(\omega + \omega_1) \right]
\]

which is the Fourier transform of another cosine wave, at the same frequency, scaled by a factor \( \rho \) where

\[
\rho = \frac{d}{B(\omega = \omega_1)}
\]

As mentioned above, this factor compensates for the size of the actuator’s frequency response at the frequency \( \omega_1 \) and the actuator separation.

In the case where aliasing occurs, the optimal choice of the scaling factor \( \rho \) is a trade-off between matching the desired component at \( \omega = \omega_1 \) and avoiding introducing a component at \( \omega = \omega_k - \omega_1 \) (assuming that there is only one significant aliasing component). Thus we wish to choose a value of \( \rho \) which minimises
\[ \| \mathbf{Z}_1(\omega) - P'(\omega; \mathbf{u}_k^*) \|_2 = \| \left[ \frac{1}{2} \delta(\omega - \omega_1) + \frac{1}{2} \delta(\omega + \omega_1) \right] \]
\[ - \rho \frac{1}{d} B(\omega_1) \left[ \frac{1}{2} \delta(\omega - \omega_1) + \frac{1}{2} \delta(\omega + \omega_1) \right] \]
\[ - \rho \frac{1}{d} B(\omega_s - \omega_1) \left[ \frac{1}{2} \delta(\omega - (\omega_s - \omega_1)) + \frac{1}{2} \delta(\omega + (\omega_s - \omega_1)) \right] \|_2 \] 3.3.75

Because the delta functions at \( \omega = \omega_1 \) and \( \omega = \omega_s - \omega_1 \) do not overlap, the cross product terms in the evaluation of the norm are zero and using the fact that
\[ \| \frac{1}{2} \delta(\omega - \omega_1) + \frac{1}{2} \delta(\omega + \omega_1) \|_2 = \| \frac{1}{2} \delta(\omega - (\omega_s - \omega_1)) + \frac{1}{2} \delta(\omega + (\omega_s - \omega_1)) \|_2 \]
3.3.76

it is readily shown that the minimisation problem reduces to choosing the value of \( \rho \) which minimises the expression
\[ \| 1 - \rho \frac{1}{d} B(\omega_1) \|^2 + \| \rho \frac{1}{d} B(\omega_s - \omega_1) \|^2 \] 3.3.77

Expanding this gives
\[ 1 - \rho \frac{1}{d} B(\omega_1) + \rho^2 \frac{1}{d^2} \left[ B(\omega_1)^2 + B(\omega_s - \omega_1)^2 \right] \] 3.3.78

which is quadratic in \( \rho \). The optimal value is
\[ \rho = \frac{\frac{1}{d} B(\omega_1)}{\frac{1}{d^2} \left[ B(\omega_1)^2 + B(\omega_s - \omega_1)^2 \right]} \] 3.3.79

which means that the set points are given by
\[ u_k^* = u'(x=kd) = \frac{\frac{1}{d} B(\omega_1)}{\frac{1}{d^2} \left[ B(\omega_1)^2 + B(\omega_s - \omega_1)^2 \right]} \cos \omega_1 kd \] 3.3.80

The size of the residual error between the desired response and the achieved response, relative to the size of \( \| \mathbf{Z}_1(\omega) \|_2 \) is given by
\[
\frac{\| E'(\omega; \psi_k) \|_2^2}{\| Z_1(\omega) \|_2^2} = \| 1 - \rho \frac{1}{d} B(\omega_1) \|^2 + \| \rho \frac{1}{d} B(\omega_s - \omega_1) \|^2
\]

\[
= \frac{B(\omega_s - \omega_1)^4 + B(\omega_1)^2 B(\omega_s - \omega_1)^2}{[B(\omega_1)^2 + B(\omega_s - \omega_1)^2]^2}
\]

3.3.81

Note that when there is no aliasing \(B(\omega_s - \omega_1)\) is zero and the value of \(\rho\) reduces to that given in equation 3.3.74. Also the relative error reduces to zero (provided that \(B(\omega_1)\) remains finite) which agrees with the result obtained from the non-aliased case.

The description of the effect of the aliasing components which are generated when \(\omega_c > \frac{1}{2}\omega_s\), has been presented in the frequency domain. The corresponding effects in the spatial domain can be seen by noting that if \(\omega_s\) is decreased (so that \(\frac{1}{2}\omega_s < \omega_c\)) then this is equivalent to increasing the distance between the actuators, because

\[
\omega_s = \frac{2\pi}{d}
\]

3.3.82

Intuitively, we can see that if the actuators are placed too far apart, their responses do not overlap sufficiently so that there are "gaps" in the spatial response of the array. If \(\omega_c\) is only just larger than \(\frac{1}{2}\omega_s\), only the frequency components near \(\frac{1}{2}\omega_s\) are affected by the aliasing components. If the separation between the actuators is increased so that \(\omega_c > \omega_s\), then aliasing components will be introduced even when matching response which have low frequency components. This is illustrated in figure 3.3.12. Here an array with widely spaced actuators is being used to generate a low frequency response. Because the actuators are too far apart, the response cannot be matched without introducing a higher frequency component. It should be stressed that this aliasing effect will only arise if the actuator spacing is large enough so that \(\omega_s > \frac{1}{2}\omega_s\).

As a summary of the aliased case, we can bring together the results to characterise the frequency response of an infinite array of actuators, where the response of a single actuator, \(B(\omega)\), is a monotonically decreasing, band limited function of bandwidth, \(\omega_c\) (as
Figure 3.3.12  The effect of aliasing in the spatial domain. The dashed curve (---) shows the desired response which is a low frequency cosine wave. The achieved response (the solid line) is the sum of a set of actuator responses (each shown by a dotted line (...) ) where the spacing between the actuators is too large.
in the case of a Gaussian actuator response). The size of the residual that remains when trying to achieve responses which are sinusoidal waves of increasing frequency, \( \omega_1 \), divides into three regions:

(i) \[
\frac{\| e'(x;u_k') \|_2}{\| z_1(x) \|_2} = 0 \quad \text{for} \quad \omega_1 < (\omega_s - \omega_c) \quad \text{assuming that} \quad (\omega_s - \omega_c) > 0.
\]

(ii) \[
\frac{\| e'(x;u_k') \|_2}{\| z_1(x) \|_2} \quad \text{increases monotonically over the region} \quad \omega_s - \omega_c < \omega_1 < \omega_c,
\]

(assuming that \( B(\omega) \) decreases monotonically as in the illustrations).

(iii) \[
\frac{\| e'(x;u_k') \|_2}{\| z_1(x) \|_2} = 1 \quad \text{for} \quad \omega_1 > \omega_c.
\]

There is also an anomalous behaviour at \( \omega_1 = \frac{1}{2} \omega_s \) where for cosine waves

\[
\frac{\| e'(x;u_k') \|_2}{\| z_1(x) \|_2} = 0 \quad \text{3.3.83}
\]

and for sinewaves

\[
\frac{\| e'(x;u_k') \|_2}{\| z_1(x) \|_2} = 1 \quad \text{3.3.84}
\]

**Summary**

We can combine the results of the non-aliased and aliased cases by plotting the frequency response of the array of actuators under various conditions. Figure 3.3.13 shows the response to cosine waves, where each curve represents an array with a fixed actuator separation (and hence a fixed \( \omega_s \)) but the bandwidth of the response, \( \omega_c \), increases. The shape of the response is taken to be Gaussian as in example 2.2.1. We have limited the frequency response to its nominal bandwidth, \( \omega_c \) (which is taken to be \( \omega_c = \frac{3}{\sigma} \)) so that
Figure 3.3.13 The frequency response on an infinite array of actuators. In each of the graphs, the relative error between the desired actuator response, which has the form of a cosine wave, and the achieved response (as defined in equation 3.3.64) is plotted against the frequency of the desired response. Each graph shows the response for a different actuator bandwidth, $\omega_c$. In figure (a), $\omega_c$ is equal to $0.3\omega_s$. In figure (b), $\omega_c$ is equal to $0.5\omega_s$. In figure (c), $\omega_c$ is equal to $0.75\omega_s$. In figure (d), $\omega_c$ is equal to $1.2\omega_s$. 
B(ω) is zero for all $|ω| > ω_c$. This prevents the actuator array attempting to match frequencies in the region where $B(ω)$ is small so that very large set points are required.

On the x axis is plotted the frequency of the cosine wave and on the y axis, the size of the residual is shown as the fraction of the size of the desired response, given by

$$\frac{||e'(x; u'_k)||_2}{||z_1(x)||_2} = \frac{||E'(ω; u'_k)||_2}{||Z_1(ω)||_2}$$ \hspace{1cm} (3.3.85)

For figure 3.3.13a, $ω_c(a)$ is low so only the low frequencies where $ω_1 < ω_c$ can be matched exactly. For figure 3.3.13b, $ω_c(b) = 1/2ω_s$, so all frequencies up to $ω_1 = 1/2ω_s$ can be matched exactly. For figure 3.3.13c, $1/2ω_s < ω_c(c) < ω_s$ so the low frequencies can be matched exactly, but for frequencies in the range $(ω_c(c) - ω_1) < ω_1 < ω_c(c)$ there will be a residual component due to the aliasing effect. The shape of this curve is determined by the relative sizes of the frequency component of the actuator response at $ω_1$ and the aliasing component at $ω_s - ω_1$. The effect of increasing the bandwidth of the responses of the actuators still further is illustrated in figure 3.3.13d. Here the bandwidth is such that $ω_c(d) > ω_s$, so that aliasing occurs when matching all frequencies. In the two cases where $ω_c > 1/2ω_s$ (figures 3.3.13a and 3.3.13b) there is the anomalous behaviour at $ω_1 = 1/2ω_s$ where there is no residual.

The size of the set points (defined as the amplitude of the $u'(x)$) corresponding to the responses shown in figure 3.3.13, are shown in figure 3.3.14. The values are generated from equations 3.3.80 which describes the amplitude of the set points for the aliased case. Note that equation 3.3.80 collapses to equation 3.3.60 (which describes the amplitude of the set points for the non-aliased case) so the same equation can be used for both aliased and non-aliased cases. For the two cases with no aliasing (figures 3.3.14a and 3.3.14b), the size of the actuator set points increases as $ω_c$ increases. This is because $σ$ decreases (recall that $σ = σ_0/ω_s$) so the size of $B(ω)$ also decreases (see equation 3.3.35) making the amplitude of the set points to match a sinusoid of frequency $ω_1$ increase because $u'_k$ is inversely proportional to $B(ω=ω_1)$ (see equation 3.3.60). For the cases where aliasing
Figure 3.3.14 These graphs show the size of the actuator set points required to generate the frequency responses shown in figure 3.3.13. The size of the set points are plotted against the frequency of the desired response. (Note the difference in scales on the y axes).
occurs (figures 3.3.14c and 3.3.14d) the size of the set points are much smaller, because
the array cannot generate a response to match the sinusoid at $\omega_1$ without introducing a
component at $(\omega_2 - \omega_1)$. Therefore the array does not generate a response which matches the
sinusoid at $\omega_1$ completely which means that the set points are smaller.

Figures 3.3.15 and 3.3.16 show the corresponding responses for sinewaves. The only
differences occur at $\omega_1 = 1/2\omega_s$ where the fractional size of the residual goes to unity for
actuator arrays where $\omega_c > 1/2\omega_s$ and the size of the set points goes to zero.

The above figures have plotted the curves for different values of $\omega_c$ with a fixed actuator
separation, $d$ (and hence a fixed value of $1/2\omega_s$). It is instructive to reverse this and to plot
the residuals for an actuator response with a fixed bandwidth, $\omega_c$, but with a varying $\omega_s$.
This arrangement is shown in figure 3.3.17 for the residuals that are left when matching
responses which are cosine waves of increasing frequency. Figure 3.3.17a shows the case
where the actuator separation is large, so $\omega_s(a)$ is small, with $\omega_c > \omega_s(a)$. In this case,
there are no frequency components which can be matched without introducing an aliasing
component. There is also the anomalous behaviour at $\omega_1 = 1/2\omega_s(a)$. For figure 3.3.17b,
the actuator separation has been reduced, but $\omega_s(b)$ is such that $1/2\omega_s(b) < \omega_c < \omega_s(b)$.
Here, the low frequencies, $\omega_1 < (\omega_s(b) - \omega_1)$, can be matched exactly, but for higher
frequencies, aliasing occurs. In figure 3.3.17c, the actuator separation has been decreased
further, so that $\omega_c = 1/2\omega_s(c)$. For this value of the actuator separation, there are no
aliasing effects and all responses up to a frequency of $\omega_1 = \omega_c$ can be matched exactly. The
important point to note is that decreasing the actuator separation still further as in figure
3.3.17d where $1/2\omega_s(d) > \omega_c$, does not change the response of the actuator array, because
the frequency bandwidth of the whole array is still limited by the frequency bandwidth of a
single actuator, $\omega_c$. Figure 3.3.18 shows the amplitudes of the set points corresponding to
the response illustrated in figure 3.3.17.
Figure 3.3.15 These graphs show the size of the residual error when generating responses to match the sinewaves of increasing frequencies. In each case the size of the error is plotted against frequency for the different bandwidths of actuator response, as detailed in figure 3.3.13.
Figure 3.3.16 These graphs show the size of the actuator set points required to generate the frequency responses shown in figure 3.3.15. The size of the set points are plotted against the frequency of the desired response. (Note the difference in scales on the y axes).
Figure 3.3.17 As in figure 3.3.13, these graphs show the size of the error when matching responses which are cosine waves of increasing frequency. The size of the error (as defined in equation 3.3.64) is plotted against the frequency of the response. In figure 3.3.13, each graph represented the response of an array where the actuator separation was fixed and the actuator bandwidth changed. In this figure, the bandwidth, $\omega_c$, remains fixed, but the actuator separation (and hence the actuator repeat frequency, $\omega_s$) varies. In graph (a), the actuator separation is such that $1/2\omega_s$ is equal to $0.35\omega_c$. In graph (b), $1/2\omega_s$ is equal to $0.75\omega_c$. In graph (c), $1/2\omega_s$ is equal to $\omega_c$. In graph (d), $1/2\omega_s$ is equal to $1.25\omega_c$. 

112
Figure 3.3.18 These graphs show the size of the actuator set points required to generate the frequency responses shown in figure 3.3.17. The size of the set points are plotted against the frequency of the desired response. (Note the difference in scales on the y axes).
3.4 The Design of an Array of Actuators

The analysis of the previous section gives us a method for designing an array of actuators. The most important result from the analysis is that the spatial frequency response of an actuator array is limited by the spatial frequency response of a single actuator. In particular, the response of the whole array cannot contain any frequency components above the spatial bandwidth of an actuator, $\omega_c$. This means that the components of profile errors with frequencies above $\omega_c$ are uncontrollable. This discussion can be turned around to provide a criterion for the design of the response of the actuators. If the quality specifications of the web require that all of the components of the cross directional profile errors with frequencies up to $\omega_c$ must be removed from the final web, then we can immediately specify that the spatial bandwidth of each actuator must also be $\omega_c$.

Given the bandwidth of the response of the actuators, we can determine the spacing between the actuators. From the previous section, we know that if the actuators are too far apart so that $\omega_c > \frac{1}{2} \omega_s$ ($\omega_s$ being the repeat frequency of the actuators) then the response of the whole array will be degraded by the effects of aliasing. If the actuator separation is reduced so that $\omega_c = \frac{1}{2} \omega_s$, then the response of the whole array matches the response of a single actuator in the sense that the array can generate any response containing frequency components up to $\omega_c$. Reducing the actuator separation still further does not increase the bandwidth of the response of the whole array (it is still limited by the bandwidth of the response of a single actuator, $\omega_c$) but it does have the effect of making the system more difficult to control because reducing the distance between the actuators makes the actuators less linearly independent, which makes the system more ill-conditioned. The optimal spacing makes $\omega_c = \frac{1}{2} \omega_s$ because this maximises the bandwidth of the response of the actuator array while minimising the ill-conditioning. Using the fact that

$$\omega_s = \frac{2\pi}{d} \quad (3.4.1)$$

with $d$ being the actuator separation, leads to the expression for the optimal separation
So far, we have specified the required bandwidth of the response of each actuator and the optimal actuator spacing, but we have not said anything about the shape of the actuator's response. We know from equation 3.3.60 in the previous section that the actuator set points that are required for generating a response which is a cosine of frequency, \( \omega_1 \), are

\[
\mathbf{u}_k = \frac{d}{B(\omega=\omega_1)} \cos \omega_1 kd \quad \text{for } k=\{..., -1, 0, 1, ...\}
\]

The set points that are required to generate a more general response are the sum of the set points which generate the individual components of this response. We can see from this equation that in regions where the frequency response of the actuators, \( B(\omega) \), is small, the amplitude of the set points will be large. In order to keep the amplitude of the set points uniform over the whole of the controllable frequency range, we need to make \( B(\omega) \) as flat as possible over the range \( |\omega| < \omega_c \). Using this argument allows us to design an actuator response that is optimal in the sense of minimising the size of the actuator set points. If \( B(\omega) \) has the form

\[
B(\omega) = \begin{cases} 
  k & \text{for all } |\omega| < \omega_c \\
  0 & \text{for all } |\omega| \geq \omega_c
\end{cases}
\]

where \( k \) is a constant (see figure 3.4.1a) then the amplitude of the pattern of the set points is the same for generating a response which is any sinusoid with a frequency up to \( \omega_c \). The corresponding optimal spatial response is shown in figure 3.4.1b and is given by

\[
b(x) = \frac{k \sin \omega_c x}{\pi x}
\]

Obviously, there may be physical constraints which mean that this response cannot be achieved in practice, but, in general, the frequency response of the actuators should be designed to be as flat as possible in the frequency range \([-\omega_c, \omega_c]\). This form of frequency response also has the advantage that it avoids the problem of \( B(\omega) \) going to zero for values of \( \omega \) less than \( \omega_c \).
Figure 3.4.1 The shape of the optimal frequency response of an actuator is shown in figure (a). The corresponding optimal shape of the spatial response is shown in figure (b).
The analysis of section 3.3 has allowed us to define the optimal bandwidth, shape of response and separation of the actuators in the array. It is important to note that these results for the design of the actuator array rely on the analysis of the spatial frequency response of the system. Lindeborg [Lindeborg, 1986] has commented that these results cannot be obtained from a simple analysis performed in the spatial domain (rather than the spatial frequency domain) and he was forced to resort to computer simulations in an attempt to generate the same results.

3.5 The Response of a Finite Array of Actuators

The results of the previous two sections have been based on the artificial case of an infinite array of actuators producing a response on a web of infinite width. The reason for this artificial case being used is that it allows us to make the limits of the integration in in the spatial domain equal to ±∞ (see equation 3.2.5) which means that the problem can be transferred to the frequency domain by using Fourier transforms and the Plancherel identity (see equation 3.3.11). We now wish to modify the arguments of the previous section to accommodate the more practical case of a finite array of actuators and a finite width of web. We start by considering a finite actuator array on a web of infinite width and then develop the analysis for the case where both the actuator array and the web width are finite.

Case (i): Finite Actuator Array, Infinite Web Width.

The spatial response of a finite array is given by

\[ p(x;u_k) = \sum_{k=-N}^{N} u_k b(x-kd) \]  3.5.1

Note that we have dropped the primes from \( p(x;u_k) \) and \( u_k \) because we are now dealing with a finite array. The Fourier transform of \( p(x;u_k) \) is given by
\[ P(\omega;u_k) = \mathcal{F} [p(x;u_k)] = B(\omega) \sum_{k=-N}^{N} u_k e^{-ik\omega} \quad 3.5.2 \]

However, from Theorem 3.3.1, we know that the \( u_k \)'s are the coefficients of a Fourier series of \( U_s'(\omega) \). Thus the second half of the term on the right hand side of this equation can be considered as the partial Fourier series (up to \( N \) terms) of \( U_s'(\omega) \). We will write this partial sum as

\[ U_s(\omega;N) = \sum_{k=-N}^{N} u_k e^{-ik\omega} \quad 3.5.3 \]

which means that

\[ P(\omega;u_k) = B(\omega) U_s(\omega;N) \quad 3.5.4 \]

The partial sum \( U_s(\omega;N) \) is illustrated in figure 3.5.1. This figure should be compared with figure 3.3.2 which shows the full Fourier series expansion, \( U_s'(\omega) \). It can be seen that \( U_s(\omega;N) \) is still a periodic function (stretching to \( \omega=\pm\infty \)) but that the shape of the repeated section is different because it is only a partial sum.

Since \( B(\omega) \) has a bandlimit, \( \omega_c \), we can write (as illustrated in figure 3.5.2)

\[ P(\omega;u_k) = B(\omega) U_s(\omega;N) \quad \text{for } |\omega| \leq \omega_c \]
\[ = 0 \quad \text{for } |\omega| > \omega_c \quad 3.5.5 \]

The important point to note about this equation is that the frequency response of the finite array is still limited to the bandwidth of an individual actuator, ie \( \omega_c \).

If we restrict our attention to the case where there is no aliasing (ie \( \omega_c < 1/2\omega_s \)) then we can extend this analysis in a similar fashion to the case of an infinite array. We can write \( U_s(\omega;N) \) as the scaled periodic extension of a function \( U^*(\omega;N) \) which is the partial Fourier sum of \( U'(\omega) \) and is defined on the region \( |\omega| < 1/2\omega_s \) (see equation 3.3.21 in the derivation of Theorem 3.3.1). This gives

\[ U_s(\omega;N) = \frac{1}{d} \sum_{p=-\infty}^{\infty} U^*(\omega + p\omega_s;N) \quad 3.5.6 \]

Because we have assumed that we are dealing with the non aliased case.
Figure 3.5.1 The partial Fourier series, $U_N(\omega;N)$, of the periodic extension of the function, $U'(\omega)$. Note that the shape of the repeated pattern is not the same as the shape of $U'(\omega)$. Compare with figure 3.3.2 which shows the full Fourier series.
Figure 3.5.2 The frequency response of a finite array containing 2N+1 actuators, $P(\omega; u_k)$, which is the product of the response of a single actuator, $B(\omega)$ and the partial Fourier series, $U_s(\omega;N)$. 
\[ U_s(\omega; N) = \frac{1}{d} U^*(\omega; N) \quad \text{for} \quad |\omega| \leq \omega_c \]  \hspace{1cm} 3.5.7

so the frequency response of the finite array becomes
\[ P(\omega; u_k) = \frac{1}{d} B(\omega) U^*(\omega; N) \quad \text{for} \quad |\omega| \leq \omega_c \]
\[ = 0 \quad \text{for} \quad |\omega| > \omega_c \]  \hspace{1cm} 3.5.8

With the expression for the frequency response of the finite array in this form, we can define a frequency domain approach to choosing the optimal set points, \( u_k \), to match a desired response, \( z_1(x) \).

i) choose the optimal \( U'(\omega) \) whose partial Fourier series (to N terms) \( U^*(\omega; N) \) minimises
\[ \|Z_1(\omega) - \frac{1}{d} B(\omega) U^*(\omega; N)\|_2 \]  \hspace{1cm} 3.5.9
(note that this minimisation can only be carried out in the region where \( B(\omega) \) is non-zero, ie \( |\omega| \leq \omega_c \));

ii) inverse Fourier transform the optimal \( U'(\omega) \) to obtain the optimal \( u'(x) \);

iii) sample the optimal \( u'(x) \) at the points \( x=kd \) for \( k=(-N,...,-1,0,1,...,N) \) to obtain the set points for the 2N+1 actuators.

At this stage we are using a finite array of actuators to control a profile on a web of infinite width. The case of a web of finite width will be developed later.

As it stands, the procedure above is only of limited use unless we can relate the partial Fourier series, \( U^*(\omega; N) \), to the full function \( U'(\omega) \). To do this we need to return to the case of the infinite array of actuators to create a modified approach to the analysis. From equation 3.3.21, the infinite periodic extension of \( U'(\omega) \) is written as
\[ U'_*(\omega) = \frac{1}{d} \sum_{p=-\infty}^{\infty} U'(\omega + p\omega_s) \]  \hspace{1cm} 3.5.10

An alternative way of looking at this is to write [Bracewell, 1978], [Brigham, 1974]
\[ U'_*(\omega) = U'(\omega) * \frac{1}{d} \text{III}(\omega; \omega_s) \]  \hspace{1cm} 3.5.11
where $*$ signifies convolution and $\Pi(\omega;\omega_s)$ is defined as a set of Dirac delta functions defined on $\omega \in (-\infty, \infty)$, each spaced at $\omega_s$.

$$\Pi(\omega;\omega_s) = \sum_{p=-\infty}^{\infty} \delta(\omega - p\omega_s)$$  \hspace{1cm} 3.5.12

The inverse Fourier transform of $\frac{1}{d}\Pi(\omega;\omega_s)$ is another set of delta functions in the spatial domain, each separated by $d$ [Lighthill, 1962, chapter 5]

$$F^{-1}\left[ \frac{1}{d}\Pi(\omega;\omega_s) \right] = \Pi(x;d) = \sum_{p=-\infty}^{\infty} \delta(x-pd)$$  \hspace{1cm} 3.5.13

Thus the sampled spatial signal $u_s(x)$ (which is equal to the $u_k$ for $k=\ldots,-1,0,1,\ldots$) is given by

$$u_k = u_s(x) = u'(x) \Pi(x;d)$$  \hspace{1cm} 3.5.14

where we have inverse Fourier transformed equation 3.5.11 and used the fact that convolution in the frequency domain corresponds to multiplication in the spatial domain. This implies that the process of sampling a continuous function is equivalent to multiplying the function by a set of delta functions. This is illustrated in figure 3.5.3. Considering sampling in this manner provides a useful, alternative analysis of sampling (see [Bracewell, 1978] for more details).

Equation 3.5.14 gives the sampled spatial signal for an infinite array of actuators. We require $u_k$ to be defined for a finite array, i.e. for $k=\{-N,\ldots,-1,0,1,\ldots,N\}$. This can be done by multiplying the expression in 3.5.14 by a window which is zero outside the extent of the finite array. Thus the finite sampled signal is given by

$$u_s(x;N) = u'(x) \Pi(x;b) w[x;(N+\frac{1}{2})d]$$  \hspace{1cm} 3.5.15

where $w[x;(N+\frac{1}{2})d]$ is a window function given by

$$w[x;(N+\frac{1}{2})d] = 1 \text{ \ for } |x| \leq (N+\frac{1}{2})d$$
$$= 0 \text{ \ for } |x| > (N+\frac{1}{2})d$$  \hspace{1cm} 3.5.16

We can now convert 3.5.15 back into the frequency domain by taking Fourier transforms.
Figure 3.5.3

The effect of sampling a bandlimited signal, \( u'(x) \), with a distance, \( d \), between the samples. The top set of graphs in the spatial domain show \( u'(x) \) being multiplied by a function, \( \text{III}(x;d) \), consisting of an array of delta functions at a separation, \( d \), to generate the sampled signal, \( u_s'(x) \). In the frequency domain, this corresponds to convoluting the Fourier transform of \( u'(x) \), \( U'(\omega) \), by \( \frac{1}{d} \text{III}(\omega;\omega_s) \) to give \( U_s'(\omega) \).
\[ U_s(\omega; N) = \mathcal{F} \left[ u'(x) \Pi(x;d) w(x;(N+\frac{1}{2})d) \right] = U'(\omega) \ast \frac{1}{d} \Pi(\omega;\omega') \ast 2(N+\frac{1}{2})d \text{sinc}[(N+\frac{1}{2})d\omega] \]

where

\[ \mathcal{F} \left[ w(x;(N+\frac{1}{2})d) \right] = 2(N+\frac{1}{2})d \frac{\text{sinc}[(N+\frac{1}{2})d\omega]}{(N+\frac{1}{2})d\omega} \]

\[ = 2(N+\frac{1}{2})d \text{sinc}[(N+\frac{1}{2})d\omega] \]

\[ u_s(x; N) \text{ and } U_s(\omega; N) \text{ are illustrated in figure 3.5.4.} \]

We are now in a position to compare the frequency response of an infinite array and a finite array. For an infinite array

\[ P'(\omega; u_k) = B(\omega) \sum_{k=-\infty}^{\infty} u_k e^{-ikd\omega} \]

\[ = B(\omega) U'_s(\omega) \]

\[ = B(\omega) \left[ U'(\omega) \ast \frac{1}{d} \Pi(\omega;\omega') \right] \]

where we have used equations 3.3.30 and 3.5.11. The corresponding response for the finite array is

\[ P(\omega; u_k) = B(\omega) \sum_{k=-N}^{N} u_k e^{-ikd\omega} \]

\[ = B(\omega) U_s(\omega; N) \]

\[ = B(\omega) \left[ U'(\omega) \ast 2(N+\frac{1}{2})d \text{sinc}[(N+\frac{1}{2})d\omega] \ast \frac{1}{d} \Pi(\omega;\omega') \right] \]

Here we have used equations 3.5.3 and 3.5.17 and we have used the fact that the convolution operation is associative [Dym and McKeans, 1972] to reorder the expression in the square brackets in 3.5.26. By comparing equations 3.5.23 and 3.5.26 we see that the difference between the two responses is that in the response of the finite array, \( U'(\omega) \) is convoluted by a sinc function. However from equations 3.5.21 and 3.5.24, we can also
Figure 3.5.4  

The effect of sampling a bandlimited signal, $u'(x)$, of finite extent, $(2N+1)d$, where $d$ is the distance between the samples, to give a sampled signal, $u_s'(x;N)$. The top graphs show the effect in the spatial domain and are the same as in figure 3.5.3, except that the signal, $u'(x)$, is multiplied by a window function, $w[x;(N+1/2)d]$. In the frequency domain, this is equivalent to convoluting $U'(\omega)$ with the function $2(N+1/2)d \cdot \text{sinc}[(N+1/2)d\omega]$ as shown in the lower line of graphs.
see that the difference between the two responses is that the response of the infinite array contains an infinite Fourier expansion of the Fourier transform of the sampled actuator signal \( U_\sigma'(\omega) \) while the response of the finite array contains a partial Fourier series of the same function. This allows us to relate the fundamental functions of the two series by

\[
U^*(\omega; N) = U'(\omega) \ast 2(\text{N}+\frac{1}{2})d \text{sinc}[(\text{N}+\frac{1}{2})d\omega] \tag{3.5.27}
\]

The finite Fourier series has a fundamental function defined on \( \omega \in [-\frac{1}{2}\omega_s, \frac{1}{2}\omega_s] \) which is the fundamental function of the infinite series convoluted with a sinc function. We can recast the minimisation problem given above, in terms \( U'(\omega) \) without having to involve the partial Fourier series at all. By substituting this expression for \( U^*(\omega; N) \) into equation 3.5.8 we obtain

\[
P(\omega; u_k) = \frac{1}{d} B(\omega) \left[ U'(\omega) \ast 2(\text{N}+\frac{1}{2})d \text{sinc}[(\text{N}+\frac{1}{2})d\omega] \right] \quad \text{for } |\omega| \leq \omega_c
\]

\[
= 0 \quad \text{for } |\omega| > \omega_c \tag{3.5.28}
\]

The procedure for the choice of the optimal set points to match a desired response \( z_1(x) \) as closely as possible, now becomes:

1. choose the optimal \( U'(\omega) \) to minimise
2. inverse Fourier transform the optimal \( U'(\omega) \) to obtain the optimal \( u'(x) \);
3. sample the optimal \( u'(x) \) at the points \( x = kd \) for \( k = \{-N, ..., 0, 1, ..., N\} \) to obtain the set points for the \( 2N+1 \) actuators

This gives us a procedure for obtaining the optimal \( u_k \)'s which does not involve the partial Fourier series.

The only problem with this analysis is that although \( U'(\omega) \) may be band limited in the frequency domain, so that \( U'(\omega) = 0 \) for \( |\omega| > \frac{1}{2}\omega_s \), this does not mean that the partial sum \( U^*(\omega; N) \) will be necessarily band limited to the same region. From 3.5.27, \( U^*(\omega; N) \) is the convolution of a bandlimited function \( U'(\omega) \) with a function of infinite width (the
sinc function). This means that $U^*(\omega;N)$ has infinite width which causes aliasing in the partial Fourier series. However, if the window function $w[x;(N+1/2)d]$ is wide relative to the actuator spacing, $d$, the corresponding sinc function will be very narrow. When this narrow function is convoluted with $U'(\omega)$ the magnitude of the resultant function is negligible beyond a certain bandwidth. This means that $U^*(\omega;N)$ is effectively band limited and, if necessary, the actuator spacing, $d$, can always be reduced so that the Nyquist frequency, $1/2\omega_s$ is greater than this effective bandwidth and the analysis will hold.

Case (ii): Finite Actuator Array, Finite Web Width.

So far in this discussion of the frequency response of the actuator array, we have been assuming that we have a web of infinite width. We now modify the analysis to accommodate finite width sheets. If the web is defined on the region $x \in [-L, L]$, the minimisation problem of choosing the optimal set points can be carried out over this region, while keeping the limits of integration at $\pm L$ by incorporating another window function, $w(x; L)$, which is defined as

$$
w(x; L) = 1 \text{ for } |x| \leq L$$

$$= 0 \text{ for } |x| > L$$

3.5.30

Using this window we can write the 2 norm defined on $x \in [-L, L]$ in terms of the 2 norm defined on $x \in [-\infty, \infty]$

$$\| e(x; u_k) \|_2 = \left[ \int_{-L}^{L} |e(x; u_k)|^2 \, dx \right]^{1/2}$$

$$= \left[ \int_{-\infty}^{\infty} w(x; L) e(x; u_k) |^2 \, dx \right]^{1/2}$$

3.5.31

We can now transfer this expression into the frequency domain so that

$$\| e(x; u_k) \|_2 = \| 2 \omega \tilde{sinc}(\omega) * \tilde{E}(\omega; u_k) \|_2$$

3.5.32
where the spatial domain norm is defined on $x \in [-L, L]$ and the frequency domain norm is defined on $\omega \in [-\infty, \infty]$. We have also used the fact that

$$\mathcal{F}[w(x;1)] = 2L \text{sinc}(\omega)$$

3.5.33

By writing

$$E(\omega;u_k) = Z_1(\omega) - P(\omega;u_k)$$

3.5.34

and using the expression for $P(\omega;u_k)$ from equation 3.5.28, we can write the 2 norm of the residual error between the desired response and the achieved response from a finite array on a web of finite width as

$$\| e(x;u_k) \|_2 = \| 2L \text{sinc}(\omega) \ast Z_1(\omega) -$$

$$2L \text{sinc}(\omega) \ast \frac{1}{d} B(\omega) \left[ U'(\omega) \ast 2(N+\frac{1}{2})d \text{sinc}(N+\frac{1}{2})d\omega \right] \|_2$$

3.5.35d

In principle, we could choose a $U'(\omega)$ which minimises this expression and then inverse Fourier transform and sample this function at the points $x=kd$ for $k=-N,-...,0,1,...,N$ to obtain the optimal actuator set points $u_k$.

As it stands, this expression cannot be simplified any further and it is impossible to obtain an explicit expression for the optimal set points even in cases where the desired response is simple (e.g., if $z_1(x)$ is a cosine wave). It is much simpler to use the spatial domain approach (as described in section 3.2) for calculating set points. Part of the reason why the finite width case does not produce results which are as clear as the results in the infinite width case is that we are continuing to express the actuator responses in terms of sinusoids. The problem is that sinusoids of different frequencies are not orthogonal on the finite width $x \in [-L, L]$, so sinusoids do not form an orthogonal basis on this region. There are a number of sets of functions which are mutually orthogonal on the finite region, $x \in [-L, L]$. In particular, prolate spherical waves [Slepian and Pollak, 1961] [Landau and Pollak, 1961] have this property with the additional advantage that they are also mutually orthogonal on the region $x \in (-\infty, \infty)$. A future topic of research is to consider the possibility of using these functions as a basis for describing the response of the actuators and to apply the analysis of section 3.3 using this basis rather than a basis of sinusoids. Hopefully the analysis will
then be applicable to webs of both finite and infinite widths. The major disadvantage of this approach is that decomposing the response of an actuator in terms of generalised prolate spherical functions is not as intuitive as decomposing the response in terms of sinusoids, so the practical consequences of the analysis (for example, the criteria for the design of the actuator array) are not as readily available.

Up to now we have restricted our attention to arrays of actuators which are arranged in a straight line. There is one common type actuator array that does not fit into this category and that is the annular die used in the bubble process for producing plastic sheet (as described in section 1.3). Because the arrangement of the actuators is circular, there is not a problem of finite length as in the case of a linear array. Instead, if we assume that the response of the array is continuous, then the response can only contain sinusoidal components whose frequencies are an integral multiple of the fundamental frequency, $\omega_f$, given by

$$\omega_f = \frac{1}{\chi}$$

where $\chi$ is the radius of the die. This means that any response can be expressed in terms of a Fourier series of sinusoids rather than a Fourier transform. Another future topic of research is to investigate the response of annular dies using this approach.

Although equation 3.5.35 for the response of a finite width die acting on a finite width web does not immediately reveal as much information as the response of an infinite width array, if the width of the sheet and the array are much larger than the spatial width of the response of an individual actuator (ie $L$ and $(N+1/2)d$ are much larger than $\delta$) then a good approximation is to replace the sinc functions by delta functions and equation 3.5.35 reduces to the same expression as for an infinite array. This allows us to use the frequency response of the infinite array as an approximation to the response of the finite array. The validity of the approximation can be seen by calculating the true frequency response curves of a finite array and comparing it with the corresponding curves described in section 3.3 (figures 3.3.13 and 3.3.14) For a finite array, the size of the residuals produced when
matching responses which are cosine waves of increasing frequencies, is calculated by:

i) generating the optimal set points using the approach of section 3.2;

ii) calculating the response of the actuator array when these set points are applied;

iii) subtracting the response of the actuator array from the desired response to obtain

the residual.

The solid lines in figure 3.5.5 shows the size of the residual calculated using this approach for a finite array of actuators each of which has a shape of response which is Guassian. The dashed lines show the size of the residual obtained from the spatial frequency analysis for the corresponding infinite array, as described in section 3.3. Each plot illustrates the response of an array with a different width of actuator response, but a fixed actuator separation (corresponding to figure 3.3.13 for the infinite array) the width of the finite array being between 10 and 35 times the width of the spatial response of a single actuator.

Comparing figure 3.3.13b, c and with figures 3.5.5b, c and d, it can be seen that the response of the finite array is well approximated by the response of the infinite array, particularly in figure 3.5.5c and d where \( \omega_c \) is large so that the width of the response is much less than the width of the array. In general, the response of the finite array is wider and less smooth than the corresponding response of the infinite array which can be attributed to the convolution by the sinc functions in equation 3.5.35. Figure 3.5.6 shows the amplitude of the set points of the actuators which generate these responses. Here again, there is good correspondence between the values from the finite array and those from the infinite array.

The approximation falls down for the array where \( \omega_c \) is less than \( \frac{1}{2} \omega_s \) (eg when \( \omega_c=0.3 \omega_s \) as in figure 3.5.5a) There are two reasons for this. Firstly, when \( \omega_c \) is small the spatial width of the response, \( \delta \), is large. In the case illustrated in figure 3.5.5a, \( L=10 \delta \) so the approximation of the sinc function by the delta function is less valid than for cases where the width of the array is much larger than the width of a single response. Secondly and more importantly, the wide spatial response means that there is a large degree of overlap between the responses of adjacent actuators, indicating that the array is highly
Figure 3.5.5 The solid lines show the frequency responses of a finite array consisting of 45 actuators. In each of the graphs, the relative error between the desired actuator response, which has the form of a cosine wave, and the achieved response is plotted against the frequency of the desired response. Each graph shows the response for a different actuator bandwidth, \( \omega_c \). In figure (a), \( \omega_c \) is equal to 0.3\( \omega_s \). In figure (b), \( \omega_c \) is equal to 0.5\( \omega_s \). In figure (c), \( \omega_c \) is equal to 0.75\( \omega_s \). In figure (d), \( \omega_c \) is equal to 1.2\( \omega_s \). The dashed lines show the corresponding results for an infinite array of actuators (as in figure 3.3.13).
**Figure 3.5.6** The solid lines in these graphs show the size of the actuator set points required to generate the frequency responses shown in figure 3.5.5 for a finite array of 45 actuators. The dashed lines show the corresponding actuator set points for an infinite array (as shown in figure 3.3.14). The size of the set points are plotted against the frequency of the desired response. (Note the difference in scales on the y axes).
ill-conditioned. This is borne out by the very large set point amplitudes shown in figure 3.5.6a. In practice, the ill-conditioning would mean that an array with this width of response could not be used because, as will be seen in the next chapter, it would be almost impossible to design a stable control system for it. Instead, as indicated in the previous section, it would make more sense to either decrease the width of the response or to increase the spacing between the actuators.

Figure 3.5.7 and 3.5.8 show the corresponding results when matching responses which are sinewaves. The shapes of the responses and the size of the set points are similar to the case when matching cosines, except for the anomalous response at the Nyquist frequency where the fractional error approaches unity and the size of the set points goes to zero for both the finite and infinite arrays.

For most practical systems, the results from the infinite width arrays are good approximations to the responses of finite arrays which allows us to use the criteria that were developed in section 3.4 for the design of the actuator arrays of finite width.
Figure 3.5.7 The solid lines show the frequency responses of the finite array used in figure 3.5.5, when matching sinewaves. In each of the graphs, the relative error between the desired actuator response and the achieved response is plotted against the frequency of the desired response. Each graph shows the response for a different actuator bandwidth, $\omega_c$. In figure (a), $\omega_c$ is equal to $0.3\omega_s$. In figure (b), $\omega_c$ is equal to $0.5\omega_s$. In figure (c), $\omega_c$ is equal to $0.75\omega_s$. In figure (d), $\omega_c$ is equal to $1.2\omega_s$. The dashed lines show the corresponding results for an infinite array of actuators (as in figure 3.3.15).
Figure 3.5.8  The solid lines in these graphs show the size of the actuator set points required to generate the frequency responses shown in figure 3.5.7 for a finite array of 45 actuators. The dashed lines show the corresponding actuator set points for an infinite array (as shown in figure 3.3.16). The size of the set points are plotted against the frequency of the desired response. (Note the difference in scales on the y axes).
CHAPTER 4: DESIGNING THE CONTROL SCHEME

The analysis in the previous chapters has been based on a distributed parameter model of the array of actuators, where it is assumed that we can measure the profile as a continuous function across the width of the web. In practice, this form of measurement can not be achieved and profile readings are taken at a finite number of points across the web. In section 4.1, we use the ideas developed in the previous chapter to determine the optimal positioning of these measurement points and the pre-processing of the signal that needs to be performed before the measurements can be made. Sampling the profile signal results in a system which consists of a finite number of actuators and a finite number of measuring points, so that the response of the actuator array can be described by a multivariable, lumped parameter system. Section 4.2 sets up such a model of the system. In section 4.3 this model is used as the basis for the design of a control system. Usually the design of a controller for a multivariable plant is a complex procedure, but by assuming a simple structure for the system and using specific decompositions, it is possible to simplify the design to the choice of a single input, single output dynamic term. Section 4.4 describes an efficient algorithm for implementing the controller on a computer. Section 4.5 discusses the robustness of the controlled system, concentrating on the case where there is uncertainty in the shape and position of the response of the actuators in the array. Finally, in section 4.6, we consider the behaviour of the system when there is a failure in one of the actuators in the array. This also provides an insight into the response of the system when the set point of an actuator is limited by the physical constraints of the plant.

All of the cross directional control designs reported in the literature are based on multivariable, lumped parameter models of the array of actuators. Unfortunately, none of these designs consider the positioning of the samples of the web profile nor the pre-processing of the signal that is required. The methods of taking the profile readings appear to be designed on an ad hoc basis or in a form to comply with the restriction of the plant layout, rather than from the analysis of sections 4.1 and 4.2. Some of the control
designs, particularly [Chen et al, 1986], [Wilhelm and Fjeld, 1983], [Karlsson et al, 1981], [Karlsson et al, 1985] and [Ringwood and Grimble, 1986], derive a similar form of controller to that developed in section 4.3, but the link between the controller design and the controllability concepts of chapter 3, together with the procedures for being able to handle general actuator dynamics that are developed in this chapter, appear to be unique. There are numerous reports of cross directional control schemes not behaving as expected and some of these problems can be attributed to a failure to consider the robustness of the control design. The robustness issue is starting to be considered in the literature (see, for example, [Lindeborg, 1986], [Ringwood and Grimble, 1986], [Karlsson et al, 1985]), but, as yet, there has not been a treatment as full as the discussion in section 4.5 which draws on the concepts developed in chapter 3. Finally, the behaviour of the plant when an actuator fails or reaches its physical limits has been considered, particularly in the paper industry, where actuators such as the headbox slice have a limited set point range [Wahren, 1986] [Chen et al, 1986], but unlike the treatment in section 4.6, a full analysis of the response of the plant under these conditions does not appear to have been attempted.

4.1 Sampling the Profile Measurement

So far in this study, we have assumed that the profile measurement at time, \( t \), \( z(x,t) \) is a continuous function on \( x \in [-L,L] \) (ie across the whole web). In practice, the measurements are usually made at a number of discrete points across the web, either by sampling a continuous, cross directional measurement or by using a finite set of fixed measuring devices arranged across the web. This raises the question: how far apart should these measurements be made in order to obtain the maximum information from a sampled profile? This question can be answered by using the fact that an actuator array cannot generate a response which contains any spatial frequency components above \( \omega_c \), the frequency bandwidth of a single actuator. Therefore there is no advantage in including these higher frequency components in the profile measurement, because the actuator set
points required to cancel a given profile error will be the same whether the these uncontrollable components are present or not. This means that the measurement signal can be low pass filtered to remove all frequency components above $\omega_c$ without affecting the response of the array of actuators.

It is important that the filter does not distort either the magnitude or the phase of the measured profile, otherwise the response of the actuator will attempt to cancel the distorted error profile rather than the true error profile. The ideal magnitude response is shown in figure 4.1.1. Clearly, this ideal response cannot be achieved in reality [Rabiner and Gold, 1975], so a more practical response is shown in figure 4.1.2 where the response is virtually flat up to $\omega_c$ and then falls off rapidly so that it is negligible beyond $\omega_c$. The errors introduced by phase shifts are potentially more serious. As an example, consider a filter which applies a 180° phase shift to a controllable frequency component. For this component the filtered signal is in anti-phase to the true signal so the applied actuator response will be increasing the profile error rather than decreasing it. This can make the system unstable.

For this reason it is important to use a filter with a linear phase response over the pass band (eg a finite impulse response filter [Rabiner and Gold, 1975]) If the filter is being employed online, the delay introduced by the linear phase shift can be compensated. One method of doing this is to reverse the filtered signal and to run it back through the filter [Hamming, 1977]. This has the effect of squaring the magnitude response while cancelling the phase shift. Alternatively, if the data is processed offline, a finite impulse response filter can be used as a smoother [Rabiner and Gold, 1975] to give a smoothed signal which has no phase shift. These two techniques are best suited to discrete digital signals rather than the continuous profile measurements that we have assumed are available. However, it is possible to sample an analogue signal extremely rapidly (relative to $\omega_c$, the ideal cutoff frequency of the filter) so that the digital signal is virtually continuous. Handling the signal in this fashion allows us to use digital filtering techniques.
Figure 4.1.1 The frequency response of an ideal low pass filter with bandwidth, \( \omega_c \).
Figure 4.1.2  The frequency response of a more practical low pass filter with a nominal bandwidth of $\omega_c$ and an cut-off frequency at $\omega_c'$. 

140
After filtering, the measurement profile is band limited, so it does not contain any frequency components above \( \omega_c \) (we will assume, for this argument, that an ideal filter with a sharp cutoff at \( \omega_c \) as in figure 4.1.1, has been used). By Shannon's theorem [Jerri, 1977], [Brigham, 1974] this means that the signal can be sampled at a frequency, \( \omega_r \), where

\[
\omega_r \geq 2\omega_c \quad 4.1.1
\]

with

\[
\omega_r = \frac{2\pi}{r} \quad 4.1.2
\]

\( r \) being the separation between samples. Sampling at this separation will ensure that there is no information loss due to aliasing. Simple rearrangement of 4.1.1 and 4.1.2 gives

\[
r \leq \frac{\pi}{\omega_c} \quad 4.1.3
\]

Notice that this expression relates the sample interval, \( r \), to the cut off frequency, \( \omega_c \), which is determined by the bandwidth of the frequency response of a single actuator. This means that the sampling strategy is tied to the response of an actuator.

It was shown in the previous chapter that for arrays containing a large number of actuators (strictly speaking an infinite number of actuators) the optimal design is to choose an actuator spacing, \( d \), such that

\[
\frac{1}{2}\omega_s = \omega_c \quad 4.1.4
\]

where

\[
\omega_s = \frac{2\pi}{d} \quad 4.1.5
\]

By substituting in for \( \omega_c \) in 4.1.3, we can see that the largest possible value for the distance between samples is
This implies that, for an ideal system, the separation of the profile samples is the same as the separation of the actuators. In practice, it is impossible to design an ideal low pass filter with the impulse response shown in figure 4.1.1 and a realistic filter is likely to have a response similar to that shown in figure 4.1.2, with an effective band width of $\omega_{c}' (> \omega_c)$. In this case the sampling frequency, $\omega_r$, must be increased so that

$$\omega_r \geq 2\omega_{c}'$$

in order to prevent aliasing. This means that the separation between the samples will be less than the separation between the actuators.

An important point to note is that it is essential to low pass filter the profile signal before sampling. If the signal is not filtered, aliasing can occur, which can cause the actuator array to respond to profile components which are uncontrollable. This is illustrated in figure 4.1.3. If the samples are taken at a separation $r$, then it is impossible to distinguish between a sinusoidal component at frequency $\omega_1$ and a component at $\omega_r - \omega_1$. If the actuator and sample separations are properly designed then it is likely that the component at $\omega_1$ is controllable, while the component at $\omega_r - \omega_1$ is uncontrollable. This means that a control system will attempt to control the uncontrollable component at $\omega_r - \omega_1$, which will lead to the performance of the controller being degraded. It is therefore essential to remove the uncontrollable components before sampling.

It is also worth pointing out the dangers of using an incorrectly designed filter. A common practice is to sample the profile by dividing the web up into sections of length $r$ (where $r$ is the distance between samples) with one section for each sampling point as in figure 4.1.4. The profile is averaged over each of these sections and the average value is used as the sample. This is equivalent to passing the data through a filter of the form shown in figure 4.1.5a and then sampling the filtered signal at points separated by $r$. The magnitude
Figure 4.1.3 The aliasing problem. If samples of a signal are taken at a separation, $r$, it is impossible to distinguish between the samples of a sinusoid of frequency, $\omega_r - \omega_1$ (as shown in the top graph) and the samples of a sinusoid of frequency, $\omega_1$ (as shown in the lower graph). In both graphs, the sinusoids are represented by dashed lines and the samples by •. The dotted line in the upper graph represents the sinusoid at frequency, $\omega_1$. 
Figure 4.1.4  Block averaging. The web is divided up into regions of length, $r$, and the profile is averaged over each region to generate a sample.
Figure 4.1.5  Block averaging is equivalent to passing the profile signal through a filter with a spatial response as in figure (a) and then sampling at an interval of \( r \). The frequency response of this filter is shown in figure (b). The frequency response of the sampled signal is shown in figure (c), where the response of the unsampled signal is repeated along the frequency axis at a separation of \( \omega_r \).
The frequency response of this filter is shown in figure 4.1.5b and if the raw profile signal is considered as white noise then the spectrum of the sampled signal is shown in figure 4.1.5c, where the spectrum of the unsampled signal has been repeated along the frequency axis at a separation of $\omega_r$. It can be seen that there is a large degree of aliasing in the low frequency region (which corresponds to the region of controllable profile components) so the sampled controllable components will be indistinguishable from the uncontrollable components, leading to a degradation in the performance of the actuator array.

The aliasing problem is illustrated in the spatial domain in figure 4.1.6. The web has been divided up into regions as before and the spatial averages over these regions for two sinusoids are shown. It can be seen that the spatial averages in each block are the same for both sinusoids (note that for the higher frequency wave, the speckled part of the wave cancels out when averaged, leaving only the hatched region) so the sampled versions of the two waves are indistinguishable.

So far the discussion of the sampling has considered an ideal measuring device which takes an instantaneous, continuous reading across the whole of the web. We now need to apply the ideas of filtering and sampling to the two common forms of measuring devices:

i) a fixed array of individual measuring devices which is typical for measuring the shape variations of metals (see section 1.3);

ii) a single measuring device which produces an analogue profile reading as it scans across the web, which is used in most other cross directional control schemes.

In case (i), the individual measuring devices can be regarded as providing an inherent form of sampling. The analysis of the previous section suggests that these measuring devices should be arranged so that they have a separation $r$. The difficulty is that it is impossible to filter the signal in the conventional manner for this form of device, so it is essential to design the devices so that they do not respond to any spatial frequency components above $\omega_r$. If this can be achieved, then there will be no aliasing from the measuring array.
Figure 4.1.6  The block averaging of two sinusoidal profiles of different frequencies in the spatial domain. Because the average over each of hatched the regions is the same for both profiles, the sampled profiles will be identical. (Note that the average of the two dotted areas in the higher frequency profile are equal in size, but opposite in sign, so they cancel out).
For the second type of measuring device, the analogue signal can be filtered and sampled as described above. However, because the gauge is scanning across the web, the measured profile will contain machine directional components as well as the cross directional and residual components. As illustrated in figure 4.1.7, it is impossible to distinguish between a machine directional component and a cross directional component from a single scanned measurement, so it is feasible for the cross directional control system to respond to machine directional variations. Separating the two components is a current topic of research, but a promising approach is to correlate the scanned signal with a profile signal from a fixed gauge, which measures the machine direction plus residual components at a fixed point across the web. It seems likely that subtracting the correlated component from the scanned signal will remove the machine direction variations from the scanned signal, leaving behind only the cross directional and residual components.

4.2 A Lumped Parameter Model of the Array of Actuators

The previous section has shown us how to sample the cross directional profile signal so that there are now a finite number of measuring points. We have also based our analysis on a system which uses profile measurements taken at discrete times, \( t=nT \) with \( n=\{1,2,\ldots\} \). As described in section 2.3, this is a common method of operation for practical systems. This means that we have now sampled the continuous variations in the web profile in two directions, so we have reduced the distributed parameter model of the system to a model which is discrete in both the cross directional spatial domain and the time (or machine directional) domain. We can represent the system by a sampled data lumped parameter system which describes the response between an array containing a finite number of actuators and a finite number of measuring points. If we assume that there are \( m \) measuring points and \( 2N+1 \) actuators, then the response of the actuator array is given by

\[
y(nT) = g(q^{-1})G u(nT) \tag{4.2.1}
\]

where
Both of these diagrams show the variations of profile, $P(x,t)$, over a web. The upper diagram shows a profile which consists of a pure sinusoid in the cross direction and in the lower diagram the profile is a pure sinusoid in the machine direction. If the profile is measured by a gauge scanning across the web, as indicated, it is impossible to distinguish the machine direction variation from the cross direction variation.
y(nT) \in \mathbb{R}^m \text{ is the vector of measurements at time } t=nT; \\
u(nT) \in \mathbb{R}^{(2N+1)} \text{ is the vector of actuator set points at time } t=nT; \\
G \in \mathbb{R}^{mx(2N+1)} \text{ is the impulse response matrix whose } ij \text{th element contains the} \\
response of the } i \text{th measuring point to a unit change of the set point for the } j \text{th} \\
actuator (-NS \leq i \leq N, j \leq m); \\
g(q^{-1}) \text{ is the (scalar) function describing the dynamics of each of the actuators.} \\

Although this equation describes a multivariable system, it should be noted that the 
assumption about each of the actuators having identical dynamics (see Assumption 2.1.1a) 
allows us to make } g(q^{-1}) \text{ scalar, so that the } G \text{ matrix contains only real values. In practice,} 
it is likely that this assumption will not necessarily be true because the dynamics are likely 
to be different for actuators at edge of the array compared to those in the centre, or because 
the dynamics change as individual actuators age at different rates. If non-identical dynamics 
are to be modelled, the response of the array should be written as 
y(nT) = G(q^{-1}) u(nT)  
where } G(q^{-1}) \text{ is a } mx(2N+1) \text{ matrix, each of whose elements is a rational function of } q^{-1}. 

For even quite modest numbers of actuators this can result in a multivariable control 
problem of very large dimensions which can be extremely complex. However, if a control 
system is designed on the simpler model using identical actuator dynamics (as in equation 
4.2.1) and the design incorporates a degree of robustness, it is likely that the controller will 
still perform well in the presence of non-identical dynamics. This will be discussed further 
in the section 4.5. 

The impulse response matrix, } G, \text{ in equation 4.2.1 gives the steady state, profile response} 
at each of the measuring points in response to unit changes in each of the actuators. Thus if 
the } m \text{ measuring points are positioned at } x_1, x_2, ..., x_m, \text{ across the web, then the } ij \text{th} 
element of } G \text{ is given by 

$$[G]_{ij} = b_j(x_i) \text{ for } i = \{1, 2, ..., m\}  
$$

where } b_j(x) \text{ is the continuous response of the } j \text{th actuator as defined in section 2.2. The } j\text{th}
column of G will contain a sampled version of $b_j(x)$. Since the shape of the responses of each of the actuators has been assumed to be the same the shape of each of the columns is the same, except that the centre of each response is shifted. This is illustrated in figure 4.2.1 for the case where $b_j(x)$ is a Gaussian response (as in example 2.2.1). The special structure of the G matrix will be used in the procedure described in the next chapter for estimating G.

Equation 4.2.1 describes the effect of the actuators on the profile. The actual measured profile will be the sum of the effect of the actuators and any external disturbances. The external disturbances can be divided into two parts (see section 1.2):

i) a fixed cross directional disturbance which does not change with time (or at least varies slowly relative to the time between samples, T);

ii) a residual component which is not fixed in either the cross direction or the machine direction.

Note that although there may be machine direction disturbances present, we have assumed that the cross directional profile measurements are taken relative to the instantaneous mean level (ie the mean profile across the web at the time of the sample, as described in section 1.2). Machine direction variations simply shift this mean level so this implies that the machine direction variations will not be seen by the cross directional measuring device, provided that all of the cross directional samples are taken at the same time. This will not be true of a scanning cross directional measuring device where the samples are taken at different times as the measuring device scans across the web. In this case steps must be taken to separate the machine direction and cross direction components of the profile as described at the end of the last section.

We can now express the observed profile measurements, $z(nT) \in \mathbb{R}^m$ as

$$z(nT) = g(q^{-1}) G u(nT) + d + e(nT) \quad 4.2.4$$

where

$d \in \mathbb{R}^m$ is the steady state cross directional disturbance;
Figure 4.2.1  A representation of the G matrix when the response of each of the actuators has the shape of a Gaussian. Each column contains a Gaussian response whose centre is shifted to align with the centreline of the corresponding actuator. Any actuator response which overlaps the edge of the sheet is truncated.
\(e(nT) \in \mathbb{R}^m\) is the resididual component of the disturbance.

This gives a complete lumped parameter model of the effects of the cross directional actuator and the disturbances.

4.3 Controlling the Lumped Parameter System

Having developed a lumped parameter model of the cross directional web variations, we now wish to design a control system which will keep the web profile as flat as possible (in a sense to be defined below). This means that we want to make \(x(nT)\) as close as possible to the zero vector for all times \(t=nT, n=\{1, 2, ...\}\). In addition we also require that

i) the controlled system is stable;

ii) the controlled system is robust enough to remain stable and to perform adequately when the actual plant does not match the true plant.

This section concentrates on the design of a stable control system. A discussion of the robustness of the controlled system will be deferred to the next section.

From equation 4.2.4, we know that the observed profile at the measuring points is given by

\[
x(nT) = g(q^{-1}) G u(nT) + d + e(nT) \tag{4.3.1}
\]

However, from section 4.1, we know that the profile has to be passed through a spatial, low pass filter before it is sampled. This means that \(x(nT)\) does not contain any high frequency, spatial components. These components will not be "seen" by the control system, so they will remain as variations in the true profile of the web. Provided that the filter has been designed correctly, these high frequency components are uncontrollable, so it does not matter that they are ignored by the controller. It should be noted that this does not mean that \(x(nT)\) contains only controllable components. Because it is impossible to implement the ideal filter (with a response as in figure 4.1.1) there will be some components at frequencies above \(\omega_c\) which are passed by the filter.
Before proceeding with the controller design, it is necessary to consider the controllability of the actuator array. This was investigated for the distributed parameter system in section 3.1, but we now wish to transfer the concept to the lumped parameter system. We will consider controllability in terms of the following definition.

**Definition 4.3.1** Consider a system at time $t=0$ with a measured profile $z(0) \in \mathbb{R}^m$. The profile is controllable if the profile of the web can be brought to the state $z(nT) = 0 \in \mathbb{R}^m$ in the absence of any external influences (e.g., noise) for some finite $n$.

While we are considering spatial controllability, we will assume that the actuator dynamics are simply a delay, so that

$$g(q^{-1}) = q^{-1}$$

4.3.2

and the response of the actuator array is

$$y(nT) = q^{-1} G u(nT)$$

4.3.3

$$= G u((n-1)T)$$

4.3.4

These simple dynamics mean that if a profile is controllable, it can be brought from $z(0)$ to zero in one step.

From the discussion of sampling in section 4.1, we know that there will be at least as many measuring points as actuators (i.e., $m \geq 2N + 1$). Equality will be achieved if an ideal low pass filter can be implemented so that the distance between samples can be made equal to the actuator separation without any danger of aliasing.

Looking at equation 4.3.4, we can see that the response of the actuator array is a linear combination of the columns of $G$, the coefficients of the linear combination being the actuator set points, $u((n-1)T)$. As we have identified the columns of $G$ as the sampled spatial impulse responses of each of the actuators, this implies that the spatial response of the array is a linear combination of the spatial responses of the individual
actuators. Since there are only $2N+1$ actuators, this means that the response of the array can span a space of only (at most) $2N+1$ dimensions. If the number of measuring points, $m$, is greater than the number of actuators, then the combined response of the whole array cannot span the whole space of possible (sampled) web profiles. This means that it will not be possible to match exactly all of the profiles in $\mathbb{R}^m$, so the system will not be completely controllable. We can identify these profiles which are controllable as belonging to the subspace which is spanned by the columns of $G$. This will be referred to as the controllable subspace.

We will assume in the subsequent analysis, that all of the spatial responses of the actuators are linearly independent even after they have been sampled. This means that the $2N+1$ actuators span a space of dimension $2N+1$, so the controllable subspace also has dimension $2N+1$.

This discussion of controllability fits in with the design of the optimal filtering and sampling given in section 4.1, which came out of the idea of controllability in the continuous spatial domain (chapter 3). The optimal filter removes all frequencies above $\omega_c$, so that the profile signal can be sampled at intervals equal to the distance between actuators (see equation 4.1.7). Under these circumstances, there will be $2N+1$ actuators and $2N+1$ measuring points. This means that the space of possible web profiles is $\mathbb{R}^{2N+1}$, which can be spanned by the $2N+1$ actuator responses (assuming that they are linearly independent) so the web is completely controllable. A system which is only using sampled profiles can only control the profile at the sample points and cannot control the web in between the sample points. Although a correctly filtered and sampled system is completely controllable, it can only drive the profile to zero at the sample positions. However, the filter allows through only low frequency components, and if the samples of these components are zero, then the continuous filtered profile across the whole of the web must also be zero because it is impossible to have any low frequency variations between the sample points. Thus if the profile of an optimally filtered and sampled system is controlled to zero at the sample
points, all of the continuous controllable components will be removed. The uncontrollable components, which have been removed from the profile measurements by the low pass filter, will not be "seen" by the control system and will remain on the web.

At this stage, we introduce the singular value decomposition (SVD) of the $G$ matrix. If $G \in \mathbb{R}^{m \times (2N+1)}$, then there exists orthogonal matrices $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{(2N+1) \times (2N+1)}$ such that

$$U^T G V = \Sigma$$ \hspace{2cm} 4.3.5

where $\Sigma \in \mathbb{R}^{m \times (2N+1)}$ and can be written as

$$\Sigma = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$ \hspace{2cm} 4.3.6

$\Sigma \in \mathbb{R}^{(2N+1) \times (2N+1)}$ is a diagonal matrix

$$\Sigma = \text{diag} (\sigma_i)$$ \hspace{2cm} 4.3.7

with

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{2N+1} \geq 0$$ \hspace{2cm} 4.3.8

The $\sigma_i$'s are the singular values of $G$. If the responses of the actuators are linearly independent (as we have assumed) then $G$ is of full rank and the $\sigma_i$ for $i=\{1,\ldots,2N+1\}$ are greater than zero.

The columns of the $U$ matrix are the left singular vectors of $G$. We can divide $U$ into two parts

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$$ \hspace{2cm} 4.3.9

with $U_1 \in \mathbb{R}^{m \times (2N+1)}$, $U_2 \in \mathbb{R}^{m \times (m-(2N+1))}$. The columns of $U_1$ (which are the first $2N+1$ columns of $U$) span the range of $G$ [Golub and van Loan], ie they span the controllable subspace of $G$. The columns of $U_2$ span the rest of the $\mathbb{R}^m$ space, ie the uncontrollable subspace.

If we rearrange equation 4.3.5 as
\[ G = U \bar{\Sigma} V^T \quad 4.3.10 \]

(where we have used the fact that \( U \) and \( V \) are orthogonal matrices so that \( UU^T = I_m \) and \( VV^T = I_{2N+1} \)) then we can write

\[ G = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T \quad 4.3.11 \]

\[ = U_1 \Sigma V^T \quad 4.3.12 \]

We will use both of these decompositions for \( G \) below.

The analysis of the spatial controllability of the actuator array has shown that (for \( m>2N+1 \)) there are some uncontrollable web profiles. From definition 4.3.1, this means that it is impossible to generate a set of actuator set points which can bring uncontrollable web profiles to zero. The best that can be done is to bring the uncontrollable profile to a profile which is as close as possible (in some sense) to zero. In keeping with the rest of this study, we will use the 2 norm as a measure of "closeness". This means that we wish to chose the set points \( u^* \in \mathbb{R}^{2N+1} \) which will optimise

\[ \min_{u} || z - Gu ||_2 \quad 4.3.13 \]

where \( z \in \mathbb{R}^m \) is a general web profile and \( || . ||_2 \) signifies the 2 norm on \( \mathbb{R}^m \) which is defined as

\[ || a ||_2 = a^T a \quad \text{for any } a \in \mathbb{R}^m \quad 4.3.14 \]

\( || z - Gu ||_2 \) gives the size of the residual that is left after applying the set points \( u \) to control a profile, \( z \) (once the transient response has died away). The optimal set points, \( u^* \), will make this residual as small as possible. If \( z \) is in the controllable subspace, the size of the residual will be zero.

Obtaining the optimal \( u^* \) which minimises expression 4.3.13 is a least squares problem which has a solution [Davis, 1975]

\[ u^* = (G^T G)^{-1} G^T z \quad 4.3.15 \]
We will denote the matrix \((G^T G)^{-1} G^T E_{R(2N+1)x m}\) by \(K\). As might be expected, there is a striking similarity between this expression for the optimal set points in the discrete spatial domain and the corresponding expression in chapter 3 (equation 3.2.9) for the optimal set points in the continuous spatial domain. We can see this by noting that the columns of \(G\) correspond to the sampled response of the actuators. If we write \(g_i\) for the \(i\)th column of \(G\) (corresponding to the response of the \(i\)th actuator) then we can express \(G^T G\) (which is part of 4.3.15) as
\[
[ G^T G ]_{ij} = < g_i, g_j > \tag{4.3.16}
\]
where \(< ., . >\) indicates the inner product in \(\mathbb{R}^m\). The rest of the right hand side of 4.3.15, \(G^T z\), is a vector in \(\mathbb{R}^{(2N+1)}\) whose \(i\)th element is given by
\[
[ G^T z ]_i = < g_i, z > \tag{4.3.17}
\]
As we have identified \(g_i\) as the discrete actuator response and \(b_i(x)\) as the continuous actuator response, the correspondence between equations 4.3.15 and 3.2.9 is readily seen.

The matrix \(K\) can be expressed in terms of the singular value decomposition of \(G\) (see equation 4.3.12)
\[
K = (G^T G)^{-1} G^T = [ V [ \Sigma^2 ] V^T ]\tag{4.3.18}
\]
\[
= [ V [ \Sigma 0 ] U^T U [ \Sigma 0 ] V^T ]^{-1} V [ \Sigma 0 ] U^T \tag{4.3.19}
\]
\[
= V [ \Sigma^{-2} ]^{-1} V^T V [ \Sigma 0 ] U^T \tag{4.3.20}
\]
\[
= V [ \Sigma^{-1} 0 ] U^T \tag{4.3.21}
\]
Because \(\Sigma\) is diagonal, \(\Sigma^{-1}\) can be identified as (using equation 4.3.7)
\[
\Sigma^{-1} = \text{diag} \left[ \frac{1}{\sigma_i} \right]_{i=1,2N+1} \tag{4.3.22}
\]
We can use the expression for \(K\) in equation 4.3.15 for the determination of the optimal set points.
\[ u^* = V \begin{bmatrix} \Sigma^{-1} 0 \\ U_1^T \\ U_2^T \end{bmatrix} z \]

\[ = V \Sigma^{-1} U_1^T z \]

where we have partitioned the columns of the U matrix as in equation 4.3.9.

As remarked earlier, the columns of \( U_1 \) (and hence the rows of \( U_1^T \)) span the range of \( G \), which is the controllable subspace. Because \( U \) is an orthogonal matrix, the columns can be considered as an orthonormal basis for the controllable subspace. The term \( U_1^T z \) in 4.3.24 can be thought of as extracting the controllable component of the profile, \( z \). We can see this by splitting \( z \) into two parts

\[ z = z_c + z_u \]

where \( z_c \) is the component which lies in the controllable subspace (ie the range of \( G \)) and \( z_u \) the component of the profile which is orthogonal to the controllable subspace. It is always possible to express any \( z \) in this form [Golub and van Loan, 1983, section 6.1].

The optimal set points can now be written as

\[ u^* = V \Sigma^{-1} U_1^T z_c + V \Sigma^{-1} U_1^T z_u \]

\[ = V \Sigma^{-1} U_1^T z_c \]

because \( U_1^T z_u = 0 \) since \( z_u \) must be orthogonal to each of the rows of \( U_1 \). This means that the matrix \( K \) can be viewed as extracting the controllable component of any general profile \( z \). A similar result can be obtained by considering \( K \) as the orthogonal projection of \( z \) onto the controllable subspace [Davis and Vinter, 1985].

The discussion of controllability leads to a simple design procedure for choosing a feedback control system for the plant. We wish to design a controller \( K(q^{-1}) \) which will satisfy the criteria given at the start of this section, ie stability, disturbance rejection and robustness. We propose to use a specific structure for the controller which is
\[ K(q^{-1}) = k(q^{-1})K \quad \text{(4.3.28)} \]

where \( K \) is the real matrix defined above, ie
\[ K = (G^T G)^{-1} G^T \in \mathbb{R}^{2N+1 \times m} \quad \text{(4.3.29)} \]

and \( k(q^{-1}) \) is a scalar dynamic term.

The reasoning behind this choice of structure will become clear in the subsequent discussion, but intuitively this is a sensible form for the controller. With reference to figure 4.3.1, we can see that the \( K \) matrix extracts the controllable component from the error signal and generates the optimal set points for removing this component from the error signal. The controller dynamics determine the transient response of the actuators as they achieve these set points. Since we have assumed that we have actuators with identical dynamics, it is reasonable to make the controller dynamics the same, hence the scalar \( k(q^{-1}) \) term.

With a controller of this form, the open loop transfer function between the error signal, \( \eta(nT) \) and the response of the actuators, \( y(nT) \), is given by
\[ y(nT) = g(q^{-1}) k(q^{-1}) G K \eta(nT) \quad \text{(4.3.30)} \]

Using the SVD’s of \( G \) and \( K \) (equations 4.3.5 and 4.3.21) the term \( GK \) can be written as
\[ GK = U \left[ \begin{array}{c} \Sigma \\ 0 \end{array} \right] V^T V \left[ \begin{array}{c} \Sigma^{-1} \\ 0 \end{array} \right] U^T \quad \text{(4.3.31)} \]
\[ = U \left[ \begin{array}{cc} I_{2N+1} & 0 \\ 0 & 0 \end{array} \right] U^T \quad \text{(4.3.32)} \]
\[ = \left[ \begin{array}{c} U_1 \\ U_2 \end{array} \right] \left[ \begin{array}{c} I_{2N+1} \\ 0 \end{array} \right] \left[ \begin{array}{c} U_1^T \\ U_2^T \end{array} \right] \quad \text{(4.3.33)} \]
\[ = U_1 U_1^T \quad \text{(4.3.34)} \]

Note that it is \( UU^T \) which equals the identity matrix and not \( U_1 U_1^T \), so equation 4.3.34 does not collapse to the identity matrix. The open loop transfer function is illustrated in figure 4.3.2.
Figure 4.3.1  Block diagram of the feedback control system. The controller is separated into a constant compensator matrix, K, and a set of identical dynamics on each loop, k(q^-1). u(nT) represents the actuator set points, y(nT) the effect on the profile of actuator array, d the steady state disturbance, e(nT) the noise sequence, z(nT) the observed profile and η(nT) the profile error relative to zero.
Figure 4.3.2 The open loop transfer function between the error signal, $\eta(nT)$ and the response of the actuators, $y(nT)$, can be described in terms of a set of identical SISO dynamics, $g(q^{-1})k(q^{-1})$, between two real matrices $U_1$ and $U_1^T$, where the columns of $U_1$ span the controllable subspace of the actuator response matrix, $G$. 
This expansion can be applied to the closed loop return difference. If the loop is broken at point A in figure 4.3.3 and a signal $\xi_A(nT)$ is injected, the signal that is seen at the other side of the break, $\xi_B(nT)$ is given by [Rosenbrock, 1974]

$$\xi_B(nT) = \left\{ I + g(q^{-1}) k(q^{-1}) Gk \right\} \xi_A(nT)$$  \hspace{1cm} 4.3.35

$$= \left\{ UU^T + U \begin{bmatrix} g(q^{-1}) k(q^{-1}) I_{2N+1} & 0 \\ 0 & 0 \end{bmatrix} U^T \right\} \xi_A(nT)$$  \hspace{1cm} 4.3.36

$$= U \begin{bmatrix} [1+ g(q^{-1}) k(q^{-1}) ] I_{2N+1} & 0 \\ 0 & I_{m-(2N+1)} \end{bmatrix} U^T \xi_A(nT)$$  \hspace{1cm} 4.3.37

If $U$ is split into $[ U_1 U_2 ]$ as in equation 4.3.9, then we can rearrange this expression as

$$\begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \xi_B(nT) = \begin{bmatrix} [1+ g(q^{-1}) k(q^{-1}) ] I_{2N+1} & 0 \\ 0 & I_{m-(2N+1)} \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \xi_A(nT)$$  \hspace{1cm} 4.3.38

From above, we can view premultiplication by $U_1$ and $U_2$ as extracting the controllable and uncontrollable components of a signal, so 4.3.38 can be written as

$$\begin{bmatrix} \xi_Bc(nT) \\ \xi_Bu(nT) \end{bmatrix} = \begin{bmatrix} [1+ g(q^{-1}) k(q^{-1}) ] I_{2N+1} & 0 \\ 0 & I_{m-(2N+1)} \end{bmatrix} \begin{bmatrix} \xi_Ac(nT) \\ \xi_Au(nT) \end{bmatrix}$$  \hspace{1cm} 4.3.39

where $\xi_Ac(nT)$ and $\xi_Bc(nT)$ are the controllable components of $\xi_A(nT)$ and $\xi_B(nT)$ respectively and $\xi_Au(nT)$ and $\xi_Bu(nT)$ are the corresponding uncontrollable components.

This equation can be simplified to

$$\xi_Bc(nT) = [1+ g(q^{-1}) k(q^{-1}) ] \xi_Ac(nT)$$  \hspace{1cm} 4.3.40

$$\xi_Bu(nT) = \xi_Au(nT)$$  \hspace{1cm} 4.3.41

These expressions show us (as might be expected) that the uncontrollable components are
The closed loop return difference of the system illustrated in figure 4.3.1.

The loop is broken at point A in the feedback arm and a signal, \( \xi_A(nT) \), injected. \( \xi_B(nT) \) is the signal that is seen on the other side of the break.
projected around the loop without being affected by the controller, while it is only the controllable components of the signal which respond to the controller dynamics. We can also see that the design of the controller has been reduced to the design of $2N+1$ decoupled SISO loops. Because each of the actuators has the same dynamic response, these loops are all identical, so we have reduced the multivariable control problem to the design of a single SISO closed loop.

We need to design this SISO closed loop so that it is stable, rejects steady state disturbances and has a good transient response. We also require the system to remain stable and to perform reasonably well when the dynamics of the true system do not correspond to the modelled dynamics. This design problem is well understood for SISO systems [Dorf, 1980]. However, because it is necessary to remove the controllable part of any steady state spatial disturbance from the loop it is essential that the controller dynamics contain an integrator. This can be seen by looking at the component of the error signal, $\eta(nT)$, which is due to the steady state, cross directional disturbance, $d$. This is

$$\eta(nT) = \left[ I + g(q^{-1}) k(q^{-1}) Gk \right]^{-1} d$$  \hspace{1cm} (4.3.42)

$$= U \begin{bmatrix} 1 & 0 \\ 1+ g(q^{-1}) k(q^{-1}) & I_{2N+1} \\ 0 & I_{m-(2N+1)} \end{bmatrix} U^T d$$  \hspace{1cm} (4.3.43)

Splitting both $\eta(nT)$ and $d$ into controllable and uncontrollable components as before, we can see that the controllable component of $\eta(nT)$ due to $d$ is given by

$$\eta_c(nT) = \frac{1}{1+ g(q^{-1}) k(q^{-1})} d_e$$  \hspace{1cm} (4.3.44)

The integrator in $k(q^{-1})$ will ensure that, once the transient response of the system has passed, there is no controllable component of the disturbance left in the error signal. However, from 4.3.43, we can also see that

$$\eta_u(nT) = d_u$$  \hspace{1cm} (4.3.45)

so the uncontrollable component of the steady state disturbance remains in the web profile.
This approach to the design of a controller for the system, where we have decomposed the plant into the form

\[ g(q^{-1})G = U \begin{bmatrix} A(q^{-1}) \\ 0 \end{bmatrix} V^T \]  
4.3.46

(where \( A(q^{-1}) \) is a diagonal matrix) and used a compensator

\[ K = (V^T)^{-1} \begin{bmatrix} r \\ 0 \end{bmatrix} (U)^{-1} \]  
4.3.47

\[ = V \begin{bmatrix} \Sigma^{-1} \\ 0 \end{bmatrix} U^T \]  
4.3.48

to decouple the interaction between the controllable loops, can be viewed as a special case of the dyadic design procedure suggested by Owens [Owens, 1973] where the "decoupling matrices", \( U \) and \( V^T \) are orthogonal. In addition, if the SISO dynamics are chosen so that \( k(q^{-1}) \) minimises the error signal over some horizon, then it can be shown that the controller, \( k(q^{-1})K \), is equivalent to a linear quadratic regulator [Goowin and Sin, 1984] [Kwakernaak and Sivan, 1972].

It is interesting to compare the design procedure developed in this study with the results suggested by Chen et al [Chen et al, 1986] for the control of basis weight through the adjustment of an array of control bolts on the headbox slice of a paper machine. Essentially, the model of the actuator in this system is given by

\[ y(nT) = q^{-1}G u(nT) \]  
4.3.49

(i.e. the actuator dynamics consist of a simple delay between applying a control and observing its effect). Chen et al approach the control design by creating a "one step ahead" minimum prediction error controller, giving

\[ K(q^{-1}) = \frac{1}{1-q^{-1}} (G^T G)^{-1} G^T \]  
4.3.50

which is simply the controller matrix, \( K \) and integral controller dynamics. This matches the approach outlined in this study.

In paper making it is important to limit the bending of the headbox slice, because it is not
possible to apply an arbitrarily large deflection to the slice. Chen et al suggest a number of
techniques for avoiding this problem (this will be discussed in more detail in section 4.5).
One of these techniques is a factor to minimise the size of the actuator set point changes. In
the terms of this study, this feature can be incorporated by choosing the set points as a
linear combination of the columns of G which match a desired response, \( z \), as closely as
possible, while simultaneously minimising the size of the coefficients of the linear
combination. This can be expressed as

\[
\begin{align*}
\min_u & \quad \| z - Gu \|_2 + \mu \| u \|_2 \\
\end{align*}
\]

where \( \mu \) determines the relative importance between minimising \( \| z - Gu \|_2 \) and \( \| u \|_2 \). The optimal answer, \( u^* \), to this minimisation problem is readily shown to be [Chen et
al, 1986]

\[
u^* = (G^T G + \mu I_{2N+1})^{-1} G^T z
\]

which we will write as

\[
u^* = K_\mu z
\]

with

\[
K_\mu = (G^T G + \mu I_{2N+1})^{-1} G^T
\]

Using the singular value decomposition it is easily shown that

\[
K_\mu = V \begin{bmatrix} \Sigma & 0 \\ \mu & 0 \end{bmatrix} U^T
\]

where

\[
\Sigma = \text{diag} \left\{ \sigma_i \right\}_{i=1,2N+1} \quad \text{and} \quad \sigma_i = \sigma_i^2 + \mu
\]

If \( K_\mu \) is used instead of \( K \) as the compensator matrix, then the open loop transfer function
of the controlled system becomes

\[
g(q^{-1}) k(q^{-1}) G K_\mu = U \begin{bmatrix} \Gamma_\mu & 0 \\ 0 & 0 \end{bmatrix} U^T
\]

where
\[ \Gamma_{\mu}(q^{-1}) = \text{diag} \sum_{i=1,2N+1}^{N} \gamma_i g(q^{-1}) k(q^{-1}) \]  

with

\[ \gamma_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu} \]

The values of \( \gamma \) depend on the relative sizes of the singular values, \( \sigma_i \) and the factor, \( \mu \), but \( \gamma \) will take a range of values which we will define as

\[ \gamma_{\text{min}} \leq \gamma_i \leq \gamma_{\text{max}} \]

The range of \( \gamma_i \) means that the SISO loops are now no longer identical. However, we can still use a single set of actuator dynamics, \( k(q^{-1}) \), if we view the \( \gamma_i \)'s as loop gains. We can then design the controller dynamics with enough gain margin so that each of the loops performs satisfactorily (in terms of stability and transient response) for the range of gains in expression 4.3.56.

Because of the range of gains, it will be impossible for all of the closed loops to have the optimal transient response. The transient response of the whole system is dependent upon the response of the "worst" loop (ie the loop which has a \( \gamma_i \) which is furthest away from the nominal gain for which the controller is designed). This means that the transient response will be sub-optimal. However, this does not mean that the steady state response will be degraded. Because we have assumed that the controller dynamics contain integrators (in order to remove the steady state disturbances) in the absence of noise, the actuator set points will only obtain their steady values when the controllable component has been removed from the error signal. This can be seen by writing

\[ u(nT) = k(q^{-1}) K_{\mu} \eta(nT) \]  

\[ = V \begin{bmatrix} k(q^{-1}) \Sigma_{\mu} & 0 \end{bmatrix} U^T \eta(nT) \]  

\[ = V \begin{bmatrix} k(q^{-1}) \Sigma_{\mu} & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \eta(nT) \]
where $\eta_c(nT)$ and $\eta_u(nT)$ are controllable and uncontrollable components of the error signal. From equation 4.3.65, we can see that if $k(q^{-1})$ contains an integral term, $\eta_c(nT)$ must be zero before $u(nT)$ reaches steady state.

This shows us that including a factor for costing the set point changes does not affect the steady state behaviour of the controlled system, but it does degrade the transient response. Essentially, limiting the set point changes means that the controller takes a number of small steps before reaching steady state, rather than taking one large step. We will see in the next section that limiting the set point changes can improve the robustness of the system.

4.4 Controller Implementation

We now consider the calculation of the optimal set points for matching a specific response, $z_1$, with particular reference to implementing the controller on a computer. The major computational effort lies in solving

$$\min_u \| z_1 - Gu \|_2$$

where (for the moment) we have not weighed the size of $u$. From above, we know that the solution of this minimisation problem is given by

$$u^* = \left( G^T G \right)^{-1} G^T z_1 \quad 4.4.2$$

$$= Kz_1 \quad 4.4.3$$

One approach to calculating the $u^*$ is to precalculate the $mx(2N+1)$ matrix $(G^T G)^{-1} G^T$ directly. By careful use of the symmetry of the $G^T G$, it is possible to complete this in
Having obtained this matrix, it can be used to generate a new vector of set points at each time interval by multiplying by the vector \( z_1 \in \mathbb{R}^{(2N+1)} \). This requires a further \( [mx(2N+1)] \) flops or 10k flops for this example. In addition, \( mx(2N+1) \) elements of storage are required to hold the matrix, \((G^T G)^{-1}G^T\).

Noble [Noble, 1973] has indicated that provided \( G \) is of full rank (as we have assumed) this procedure is perfectly adequate. However, in general, it pays to avoid performing the inversion of \((G^T G)\) explicitly, because this can introduce numerical errors, particularly if \( G \) is nearly rank deficient [Gill et al., 1974]. There are a number of numerical techniques for solving the least squares problem efficiently and accurately, including LU decompositions, Cholesky factorisation and QR factorisations. Noble [Noble, 1973] provides a complete description of these methods, but for the purposes of this study we will concentrate on the approach based on the Cholesky factorisation. This technique has three major advantages:

i) the inclusion of a term for weighting the size of the set points (as in equation 4.3.51) is straightforward;

ii) the computer implementation of the algorithm is more efficient and numerically more stable than the direct approach described above;

iii) the special structure of the Cholesky factorisation simplifies the adjustments that need to be made to the algorithm when one (or more) of the actuator set points reaches its physical limit. This will be discussed in more detail in section 4.5.

To investigate the technique, we start by including a weighting factor, \( \mu \), in equation 4.4.2 and rearranging to give

\[
\left[ G^T G + \mu I_{2N+1} \right] u^* = G^T z_1
\]  

4.4.4

If the \( \mu \) term is not required, it can be made equal to zero without affecting the subsequent analysis. We have assumed that \( G \) is not rank deficient, so \((G^T G + \mu I_{2N+1})\) must be
positive definite and we can obtain a Cholesky factorisation

\[ G^T G + \mu I_{2N+1} = R^T R \]  \hspace{1cm} (4.4.5)

where \(R \in \mathbb{R}^{(2N+1) \times (2N+1)}\) is an upper triangular matrix with strictly positive diagonal elements. Performing this factorisation requires \((2N+1)^3/6\) flops. For our example with 50 actuators, this is equivalent to 20.8 kflops, which is significantly less than the effort required for the direct calculation of \((G^T G)^{-1} G^T\).

We can now recast the least squares problems as choosing the \(u^*\) which satisfies

\[ R^T R u^* = G^T z_1 \]  \hspace{1cm} (4.4.6)

If we write

\[ f = G^T z_1 \]  \hspace{1cm} (4.4.7)

and

\[ R u^* = g \]  \hspace{1cm} (4.4.8)

then the problem can be split into two parts. Firstly, find the \(g\) that solves

\[ R^T g = f \]  \hspace{1cm} (4.4.9)

and then find the \(u^*\) that satisfies

\[ R u^* = g \]  \hspace{1cm} (4.4.10)

Because \(R^T\) is lower triangular, equation 4.4.9 can be solved efficiently by using forward substitution, without having to invert \(R^T\) explicitly. Similarly, equation 4.4.10 can be solved by back substitution because \(R\) is upper triangular (see algorithms 4.1-1 and 4.1-2 in [Golub and van Loan, 1983]).

The number of computational operations required for each step are:

i) \(m(2N+1)\) flops for generating the vector, \(f\) (equation 4.4.7);

ii) \((2N+1)^2/2\) flops for the forward substitution in equation 4.4.9;

iii) \((2N+1)^2/2\) flops for the backward substitution in equation 4.4.10.

This gives a total of \((2N+1)[m+(2N+1)]\) flops for each calculation of a new vector of set points. In addition it is necessary to store the \(G\) matrix (\(m \times (2N+1)\) elements) for step (i)
and the R matrix \((2N+1)^2/2\) elements) for steps (ii) and (iii).

At first sight, there does not appear to be any saving in either storage or the number of operations required for this technique relative to the direct multiplication by the K matrix as in equation 4.4.3. However, we can take advantage of the special structure of the G matrix in step (i). Because the columns of G (and hence the rows of \(G^T\)) contain sampled versions of each actuator response, calculating

\[
f = G^T z_1
\]

is equivalent to taking the inner product of \(z_1\) with each of the actuator responses in turn. We have assumed that the shape of each actuator response is the same, so it is necessary to store only one response in a vector. The centre of the response can then be shifted to the appropriate point in the vector corresponding to the centre of the response of a given actuator, in order to calculate the inner product between the response and \(z_1\). Also, we have assumed that the response is spatially limited so that only a fraction of the elements in the vector containing the response are non-zero. If we take the width of the response to be \(p\) elements, then the calculation of \(f\) in equation 4.4.7 above can be performed in \([px(2N+1)]\) flops. The storage of the G matrix is reduced to holding a vector of length \(p\).

Using this simplification, the number of operations required to calculate a new vector of set points now becomes \((2N+1)[p+(2N+1)]\) flops and the total storage requirements are \(p+(2N+1)^2/2\) elements. For a system of 50 actuators and 200 measuring points with a width of actuator response of 30, this corresponds to 4kflops per set point calculation and 1280 elements of storage. This is a considerable saving over the corresponding figures of 10 kflops and 10000 storage elements that were quoted above for the direct calculation of \(u^*\). In addition, using the Cholesky factorisation reduces the amount of computational effort required to accommodate the adjustments that are necessary when the set point of an actuator reaches its physical limit (see section 4.6). Once the vector of set points has been obtained, the controller dynamics can be implemented by applying each element of the vector to the recurrence relation defined by \(k(q^{-1})\).
4.5 A Discussion of Robustness

The design of the control system in the previous section was based on the (implicit) assumption that the model of the response of the actuator array

\[ y(nT) = g(q^{-1}) G u(nT) \]  

was known accurately. In practice it is almost impossible to know \( g(q^{-1}) \) and \( G \) exactly for the actual plant, so it is necessary to consider the controller design when there is an uncertainty about this model. We need to have some method of characterising this uncertainty, so we will describe the true plant as being an additive perturbation from a nominal plant. We will divide the possible model uncertainties into three areas:

i) uncertainty in the dynamics of the actuators, so that the true actuator response is described by

\[ y(nT) = \left[ g(q^{-1}) + \delta(q^{-1}) \right] G u(nT) \]  

where \( \delta(q^{-1}) \) belongs to a set of possible additive perturbations;

ii) uncertainty in the impulse response of the actuators, so that the true actuator response is described by

\[ y(nT) = g(q^{-1}) \left[ G + \Delta \right] u(nT) \]  

where \( \Delta \in \mathbb{R}^{m \times (2N+1)} \) belongs to a set, \( S \), of possible perturbations;

iii) uncertainty in the model, so that it is not possible to divide the actuator's response into a constant impulse response matrix and scalar dynamics, so the response has to be described as

\[ y(nT) = \left[ G(q^{-1}) + \Delta(q^{-1}) \right] u(nT) \]  

where \( G(q^{-1}) \) and \( \Delta(q^{-1}) \) are matrices of dimension \( m \times (2N+1) \), each of whose elements is a rational function of the delay operator, \( q^{-1} \).

Through the use of the compensator matrix, \( K \), the previous section has shown us that the choice of the actuator dynamics is a SISO controller design problem. This means that if
there are uncertainties in the dynamics $g(q^{-1})$, as described in case (i), these can be handled through the use of a robust SISO design. The techniques of robust SISO design are well established (see, for example, [Astrom and Wittenmark, 1984]) so they will not be considered here.

Case (iii) describes the actuator response as a full MIMO system, which cannot be reduced to the simple form of equation 4.5.1. The design of robust control systems for this type of problem is a major topic of current research (see [Ljung, 1988] for an overview of the current research effort in this field) and a full discussion of this area is beyond the scope of this study. We will, therefore, concentrate on the control of systems which include uncertainties of the type described in case (ii). It will be seen below that a number of common plant uncertainties fit into this category.

As mentioned earlier, we will view the matrix $G$ as a nominal actuator response matrix that is a "good representation" of the true actuator response matrix. We have chosen to represent the actual actuator response matrix of the physical plant by $G + \Delta$ (see equation 4.5.3) where $\Delta$ is a small perturbation around the nominal matrix, $G$. We will assume that the true plant can be described in this form for some perturbation in the set, $S$.

Initially, we will investigate the response of a true plant of this form when it is controlled by the controller described in the previous section. The controller was designed for the nominal plant, $G$, and is given by

$$K(q^{-1}) = k(q^{-1}) K$$

with

$$K = (G^T G)^{-1} G^T$$

and $k(q^{-1})$ chosen to compensate for the actuator dynamics, $g(q^{-1})$. (Because we are interested in uncertainties of the type described in case (ii) above, we will assume that $g(q^{-1})$ is known exactly for the plant). We will start by considering the steady state response of the true system. For this reason, we will assume that there is no random
component of the noise, but that there is a fixed disturbance, $d$, which contains both
controllable and uncontrollable components. From the previous section, we know that the
controller dynamics will have been designed to include an integral term so that this fixed
disturbance can be removed. We will define the steady state of the system as the point
where the actuator set points stop changing following the transient response. If we examine
the steady state of the true plant with an actuator response $G+\Delta$, when it is controlled by a
controller designed for the nominal plant, then there are two possible outcomes:
either  i) the controlled true plant removes the same disturbance components from the
web profile as the controlled nominal plant;
or ii) the controlled true plant is unstable (ie the actuator set points never settle, but
either oscillate or increase monotonically with time).

This result can be seen by expressing the actuator set points in terms of the profile error,
$\eta(nT)$ as in the previous section
$$u(nT) = k(q^{-1}) K \eta(nT)$$ 4.5.7
Because of the integrator in $k(q^{-1})$, $u(nT)$ can only attain steady state when
$$K \eta(nT) = 0$$ 4.5.8
From the previous section we know that multiplying a vector by $K = (G^T G)^{-1} G^T$
evaluates the component of the vector which lies in the controllable subspace of the nominal
plant response, $G$. This means that $K \eta(nT)$ will only become zero when all of the
components of the profile error that are controllable by the nominal plant have been
removed. This implies that the nominally controllable component of the fixed disturbance,
$d$, must be removed from the web profile by the effect of the true actuator responses,
before the system can achieve steady state. If it is impossible to make $K \eta(nT) = 0$ because
the true actuator response cannot remove all of the nominally controllable components, then
through the effects of the integrator, the actuator set points will ramp up and the system will
be unstable.

This argument gives a slightly surprising result because it says that, provided that the
system remains stable, the steady state response of the true (perturbed) plant is the same as the response of the nominal plant.

It is worth illustrating this point with four types of perturbation that are extremely common on practical plants:

i) the gain on all of the actuators on the true plant is different from the gain on the nominal plant (although the gain of each individual actuator is the same);
ii) the shape of the true actuator response is different from the nominal response (although the shape of the response of the individual actuators remains the same);
iii) the position of the effect of the true actuator array is shifted from the position of the nominal array;
iv) the separation between the centres of the responses of the actuators on the true plant is not the same as on the nominal plant (although the separation remains the same between each of the responses).

The first two perturbations are a result of uncertainty in the shape and size of the responses of the actuators. The second pair of perturbations are associated with the a problem known as mapping. Usually the position and spacing of the actuators on the plant is well known. However, it is not as easy to determine the position and separation of the effect of the actuators on the web, particularly if the web has been stretched or has wandered. Determining the position of the effect of the actuators is known as the mapping problem [Chen et al, 1986].

We can gain some insight into the effects of these four classes of perturbation by analysing them in terms of the continuous spatial frequency response approach that was developed in section 3.3. It should be emphasised that the techniques of this section have the impractical restriction of requiring a web of infinite width and an infinite array of actuators, so the results described below should be regarded as a qualitative (but valuable) insight into the effects of the perturbations, rather than an exact analysis.
We will assume that we have a feedback system which is subjected to a steady disturbance which may contain both controllable and uncontrollable components, but no noise. We will also assume that an optimal "one step ahead" controller [Goodwin and Sin, 1984] has been designed for the nominal plant, so that the actuator set points are chosen to generate a response which exactly cancels the (nominally) controllable component of the profile error in one step. This means that the response of the actuator array will be equal and opposite to the controllable component of the error. It will be taken that the controller dynamics contain an integrator so that the steady disturbance can be removed.

If we consider the case of a disturbance which is a cosine wave of unit amplitude

\[ d(x) = \cos(\omega_1 x) \quad 4.5.9 \]

where \( \omega_1 < \omega_c \), with \( \omega_c \) being the controllable bandwidth of a nominal actuator, then we would expect the optimal controller to generate a response for the actuator array which is given by

\[ p'(x; u_k') = -\cos(\omega_1 x) \quad 4.5.10 \]

in order to cancel the disturbance. However, if the actuators in the true plant have a different gain from those in the nominal plant, then the response of the array can be written as

\[ p'(x; u_k') = -p \cos(\omega_1 x) \quad 4.5.11 \]

where \( p \) is the ratio of the true actuator gain to the nominal actuator gain. The residual error that remains after applying the control action on the true plant is

\[ e'(x; u_k') = d(x) - p'(x; u_k') \quad 4.5.12 \]

\[ = (1-p) \cos(\omega_1 x) \quad 4.5.13 \]

We can see that unless the true actuator gain is equal to the nominal gain (ie \( p=1 \)) then there will be a residual error. However, provided that the amplitude of the residual error is smaller than the amplitude of the original error, this residual can be removed by subsequent control actions. If the residual error has a larger amplitude than the original error, then the
system will become unstable, as at each subsequent control action the size of the residual profile error will increase. We can see that the system will remain stable if

$$|1 - \rho| < 1$$

4.5.14

which will be satisfied if $\rho$ lies in the range

$$0 < \rho < 2$$

4.5.15

This allows us to see that, provided $\rho$ is positive (i.e., the direction of the effect of a true actuator is in the same direction as the effect of a nominal actuator) the "one step ahead" optimal controller has a gain margin of 2.

We now turn our attention to case (ii) where the shape of the true actuator response is different from the nominal actuator response. This means that there will be a difference between the spatial frequency responses of the true array and the nominal array. This is illustrated in figure 4.5.1. Because the sizes of the actuator set points are scaled to compensate for the size of the spatial response of the array at specific frequencies (see equation 3.3.60) we can use the analysis that was developed above for uncertainty in the gain, to investigate the response of the true array when controlling sinusoidal disturbances of different frequencies. For example, if there is a frequency where the true response is greater than twice the nominal response (e.g., $\omega_1$ in figure 4.5.1) then the controller will generate a response which is too big (by a factor of 2) to control a disturbance at this frequency and the residual will be bigger than the original disturbance. Subsequent control actions will amplify the size of the residual, causing the set points to oscillate with increasing amplitude and making the system unstable.

If there is a region where the frequency response of the true array is zero while the nominal response is non zero (e.g., $\omega_2$ in figure 4.5.1) then the controller will generate the set points which produce a nominal response which cancels a disturbance in this frequency region, but the true array will be unable to create a response at this frequency. This means that the original disturbance will remain on the web. At the next control action, the controller will again attempt to cancel the disturbance, so that, under the effect of the integrator, the set
Figure 4.5.1  An example of a true actuator frequency response (represented by the solid line) and a nominal frequency response (represented by the dashed line). At frequency $\omega_1$, the true response is greater than twice the nominal response. At $\omega_2$, the true response is zero while the nominal response is non-zero and at $\omega_3$, the true response is negative while the nominal response is positive.
points will ramp up. A similar problem arises in region where the frequency response of
the true array has the opposite sign to the response of the nominal array (e.g., \(\omega_3\) in figure
4.5.1). Here the true array generates a response which is in the wrong direction and so
reinforces the disturbance. Again, this will cause the set points to ramp under the influence
of the integrators.

We can express the requirement on the frequency response of the true array which ensure
stability by recasting equation 4.5.15 and defining

\[
\rho(\omega) = \frac{P'_{(G+\Delta)}(\omega; u_k')}{P_G'(\omega; u_k')}
\]

where \(P'_{(G+\Delta)}(\omega; u_k')\) is the frequency response of the true array and \(P_G'(\omega; u_k')\) is the
response of the nominal array. The true system will be stable provided that

\[
0 < \rho(\omega) < 2 \quad \text{for all} \quad |\omega| < \omega_c(G)
\]

where \(\omega_c(G)\) is the bandwidth of the response of the nominal array. This shows us that it is
the errors in the spatial frequency response of the actuator array rather than the errors in the
shape of the responses themselves which directly determines the robustness of the system.

If figure 4.5.1 is representative of the difference between a true response and a nominal
response, then it would appear to be very difficult to satisfy the criterion in equation
4.5.17. In fact, practical experience with finite arrays (rather than the infinite arrays of this
analysis) suggest that errors in the shape of the actuator response can be tolerated,
particularly if the gain of the controller is reduced or if the size of the set point changes is
limited (see end of last section and below). This suggests that the changes due to small
errors in the shape of the actuator's response, do not greatly affect the frequency response
of the array.

The third type of perturbation described above is a result of the position of the response of
the true array being shifted relative to the response of the nominal array. Thus if the
controller sees a profile error of
\[ d(x) = \cos(\omega_1 x) \] 4.5.18

it will attempt to generate a response which cancels this error. Because the response of the true array is shifted, the generated response will effectively be phase shifted, so that

\[ p'_{(G+\Delta)}(x;u^*_k) = -\cos(\omega_1 x + \phi) \] 4.5.19

where \( \phi = \omega_1 \delta \) with \( \delta \) being the distance between the true and nominal responses. The error that remains after the control action has been applied is

\[ e'(x;u^*_k) = \cos(\omega_1 x) - \cos(\omega_1 x + \phi) \] 4.5.20

\[ = -2\sin(\frac{1}{2}\phi) \cos(\omega_1 x + \frac{1}{2}\phi) \] 4.5.21

The residual error is a sinusoid of the same frequency as the original error, but it has been shifted in phase by \( \frac{1}{2}\phi \). This residual will be controlled again on the next pass. In order for the system to remain stable, we require that the amplitude of the residual is less than the amplitude of the original disturbance, i.e.

\[ |2\sin(\frac{1}{2}\phi)| < 1 \] 4.5.22

which is equivalent to requiring that

\[ |\phi| < \frac{\pi}{3} \] 4.5.23

or

\[ |\delta| < \frac{\pi}{3\omega_1} \] 4.5.24

The maximum shift which can be tolerated while the system remains stable, will be determined by the highest nominally controllable frequency, \( \omega_c \) because a fixed error in distance will produce the largest phase error at this frequency. This gives

\[ |\delta| < \frac{\pi}{3\omega_c} \] 4.5.25

If the actuator spacing is designed correctly, then from section 3.4

\[ \omega_c = \frac{1}{2}\omega_s \] 4.5.26

where \( \omega_s \) is the repeat frequency of the actuator in the array, given by

180
with \( d \) being the separation between the actuators. Combining 4.5.26 and 4.5.27 and substituting into 4.5.25 gives
\[
| \delta | < \frac{d}{3} \tag{4.5.28}
\]
This means that the positioning of the true actuator response must be correct to within \( \pm \frac{1}{3} \) of the actuator spacing in order to ensure that the "one step ahead" controller remains stable.

Finally, we consider the case where the spacing between the responses of the true actuators is different from the nominal spacing. We have assumed that the nominal spacing between the actuators is \( d \). If the true spacing is \( d' \) and we are attempting to control a cosine profile error
\[
d(x) = \cos(\omega_1 x) \tag{4.5.29}
\]
then the actual actuator response that is generated to cancel this error, will be
\[
p'_{(G+\Delta)}(x;u_k') = -\cos\left(\frac{\omega_1 d'}{d'} x\right) \tag{4.5.30}
\]
The residual error is
\[
e'(x;u_k') = \cos(\omega_1 x) - \cos\left(\frac{\omega_1 d'}{d'} x\right) \tag{4.5.31}
= -2 \sin\left[\frac{1}{2}(1-d')\omega_1 x\right] \cos\left[\frac{1}{2}(1+d')\omega_1 x\right] \tag{4.5.32}
\]
which is a cosine wave of nearly the same frequency (assuming that \( d' \approx d \)) modulated by the term
\[
-2 \sin\left[\frac{1}{2}(1-d')\omega_1 x\right] \tag{4.5.33}
\]
For the system to remain stable, we require that
\[
| 2 \sin\left[\frac{1}{2}(1-d')\omega_1 x\right] | < 1 \tag{4.5.34}
\]
It is impossible to satisfy this relation for all \( x \) on an infinite web. However, if we transfer
this idea to a finite web, we only require that this inequality is satisfied on \( x \in [-L,L] \). Also, the expression on the left hand side of 4.5.34 will be largest when the spatial frequency of the disturbance equals the largest controllable frequency (ie when \( \omega_1 = \omega_c \)). This means that the criterion for stability can be written as

\[
|2 \sin \left[ \frac{1}{2} (1 - \frac{d}{d'}) \omega_c L \right]| < 1
\]

which reduces to

\[
1 - \frac{\pi}{3 \omega_c L} < \frac{d}{d'} < 1 + \frac{\pi}{3 \omega_c L}
\]

As before, substituting \( \omega_c = \frac{1}{2} \omega_a \) for an optimally designed actuator array spacing, gives

\[
1 - \frac{d}{3L} < \frac{d}{d'} < 1 + \frac{d}{3L}
\]

This gives us an allowable range on the relative separations between the responses of the true actuators and the nominal actuators, for the controlled system to remain stable.

Although the analysis for the four types of common robustness problems has been applied to the artificial case of the responses of an infinite array of actuators, the results do give us a valuable indication of the sizes of the differences between the nominal plant and the true plant that can be accommodated while the system remains stable. It should be pointed out that the analysis has been applied to a "one-step ahead" controller which attempts to cancel the error at the next sample interval. If the dynamics of the system are simple, then it is generally true that if the gain of the controller is reduced or the control is applied over a longer horizon, then the ranges of model error that can be tolerated without the system becoming unstable are likely to be increased. This comment only applies to those perturbations which create a response which is too large to cancel the profile error (ie where \( \rho \) is greater than 2). For cases where the true plant does not create any response or where the response is in the wrong direction (\( \rho \leq 0 \)), reducing the controller gain will not improve the stability of the true system.

This raises the question of how far the does the dynamic response of the system have to be
degraded before we can guarantee that a true plant described by
\[ y(nT) = (G + \Delta) u(nT) \quad \text{for some } \Delta \in \mathcal{S} \]
remains stable? To solve this problem we need some method of characterising the set, \( \mathcal{S} \), of possible perturbations of the true plant about the nominal plant. There are a number of approaches to this problem, but we will choose to define the set, \( \mathcal{S} \), as all perturbations, \( \Delta \), which have a 2 norm which does not exceed a given value, \( \Omega \). This allows us to define a perturbation as belonging to \( \mathcal{S} \) as
\[ \Delta \in \mathcal{S} \text{ if } \| \Delta \|_2 \leq \Omega \]
The suitability and consequences of this definition of \( \mathcal{S} \) will be discussed below.

The design of the compensator, \( K \) (as described in equation 4.3.29) is based on the nominal impulse response matrix, \( G \) and is chosen so that
\[ GK = U \begin{bmatrix} I_{2N+1} & 0 \\ 0 & 0 \end{bmatrix} U^T \]
If the actual impulse response matrix is described by \((G+\Delta)\) for some \( \Delta \in \mathcal{S} \), then it is unlikely that \((G+\Delta)K\) will have as simple a structure as 4.5.40. However, it will be possible to find a singular value decomposition for \((G+\Delta)K\) given by
\[ (G+\Delta)K = \tilde{U} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \tilde{V}^T \]
where \( \tilde{U}, \tilde{V} \in \mathbb{R}^{m \times m} \) are orthogonal matrices (different from \( U \) and \( V \)) and
\[ \Sigma = \text{diag} \sigma_i[(G+\Delta)K] \]
with \( \sigma_i[(G+\Delta)K] \) being the singular values of \((G+\Delta)K\) (some of which may be zero). This decomposition of \((G+\Delta)K\) allows us to redraw the closed loops for the perturbed plant as a set of \( 2N+1 \) SISO closed loops, each of which has open loop dynamics \( g(q^{-1}) \) \( k(q^{-1}) \) and an open loop gain of \( \sigma_i[(G+\Delta)K] \).

We can now relate the singular values of \((G+\Delta)K\) to the singular values of \( GK \). It can be shown [Golub and van Loan, 1983, section 8.3] that
\[ \sigma_i[(G+A)K] - \sigma_i[GK] \leq \|\Delta K\|_2 \text{ for } i=1, \ldots, 2N+1 \]

for any \( \Delta \in S \) provided that \( \|\Delta K\|_2 < \|GK\|_2 \)

\[ \leq \|\Delta\|_2 \|K\|_2 \]

\[ \leq \Omega \|K\|_2 \]

From equation 4.5.40, we can see that if \( K \) has been chosen so that

\[ \sigma_i(GK) = 1 \text{ for } i=1, \ldots, 2N+1 \]

then equation 4.5.45 implies that the singular values of \((G+\Delta)K\) (for any \( \Delta \) which satisfies the condition in equation 4.5.43) must lie in the range

\[ 1 - \Omega \|K\|_2 \leq \sigma_i[(G+\Delta)K] \leq 1 + \Omega \|K\|_2 \text{ for any } i=1, \ldots, 2N+1 \]

As indicated in section 4.3, the singular values of \((G+\Delta)K\) can be viewed as the loop gains on the individual SISO loops. If we can design the controller dynamics, \( k(q^{-1}) \) for a nominal loop gain of unity, but with sufficient gain margin so that the loops remain stable if the loop gain is changed to \( 1 \pm \Omega \|K\|_2 \), then the whole system will remain stable for any true actuator response matrix \( G+\Delta \), where \( \Delta \in S \), subject to \( \|\Delta K\|_2 \) being less than 1.

From the arguments above, we know that, provided the system remains stable, the steady state response of the true system will be the same as the steady state response of the nominal system, so designing the controller with sufficient gain margin will ensure that steady state response of the system is not degraded by any uncertainties in the nominal model of the plant. The price that is paid for this robustness is that the dynamic response of the system is likely to be degraded by including a large gain margin in the design of the controller.

The condition in equation 4.5.43 that requires that \( \|\Delta K\|_2 < \|GK\|_2 = 1 \) ensures that the singular values of \((G+\Delta)K\) must be positive (see equation 4.5.47). Using the interpretation of the singular values being the SISO loop gains, this is equivalent to ensuring that there is negative feedback in the loops. This can be viewed as an alternative way of requiring that the true plant generates a true response which is in the correct direction (ie \( \rho \) must not be \( \leq 0 \)).
It should be stressed that the inequalities involved in equations 4.5.43 to 4.5.45 introduce a high degree of conservativeness into the analysis, because the calculation of the gain margin that is required to ensure stability must accommodate the worst possible perturbation. This has two consequences:

i) the definition of the set, $S$, as those perturbations which satisfy $\|A\|_2 < \Omega$ may be too wide in the sense that it includes perturbations which are unlikely to occur from a physical point of view. For example, it is likely that the response of the true plant will be concentrated around the nominal position of each actuator, so there is no point in allowing for a perturbation which includes a response at the opposite side of the web from the position of an actuator. This implies that an ideal analysis would redefine the set, $S$, so that it contained only realistic perturbations. This is a future topic of research.

ii) it is possible that a true plant which is described by a perturbation, $A$, from the nominal plant, $G$, will remain stable for a controller designed with the specified gain margin, even though the size of $A$ is such that $\|A\|_2$ is greater than $\Omega$ (ie even if it does not strictly belong to $S$).

This analysis has shown us how to accommodate perturbations into a system which has a compensator matrix given by

$$K = (G^T G)^{-1} G^T \quad 4.5.48$$

through the adjustment of the controller dynamics. An alternative approach is to build the robustness into the design of the compensator matrix, $K$. The $K$ matrix in 4.5.48 was obtained from the minimisation problem

$$\min_u \| z - Gu \|_2 \quad 4.5.49$$

If we know that the impulse response matrix of the true system can be described by $(G+A)$ for some $A \in S$, where $S$ is defined in 4.5.49, then we would like the solution to the problem
\[
\min_{u} \max_{\Delta \in \mathcal{S}} \| z - (G+\Delta)u \|_2^n
\]

As it stands, this problem has no explicit solution, but we can generate an associated problem for which a closed solution does exist. For any \( \Delta \in \mathcal{S} \), we can write
\[
\| z - (G+\Delta)u \|_2 \leq \| z - Gu \|_2 + \| \Delta u \|_2
\]
\[
\leq \| z - Gu \|_2 + \Omega \| u \|_2
\]
because \( \| \Delta \|_2 \leq \Omega \) for all \( \Delta \in \mathcal{S} \). If we replace the minimisation problem in 4.5.50 by
\[
\min_{u} \| z - Gu \|_2 + \Omega \| u \|_2
\]
then this problem has an explicit solution given by
\[
u^* = \left[ G^T G + \Omega I_{2N+1} \right]^{-1} G^T z
\]
which can be written as
\[
u^* = K_{\Omega} z
\]
where
\[
K_{\Omega} = \left[ G^T G + \Omega I_{2N+1} \right]^{-1} G^T
\]
This compensator, \( K_{\Omega} \), has the same form as the compensator in equation 4.3.54 which was obtained by minimising the size of the actuator set point changes. This provides a link with the intuitive concept that reducing the size of the actuator set point changes (and hence degrading the dynamic response) improves the robustness of the system. The maximum size, \( \Omega \) of the allowable perturbations in \( \mathcal{S} \), determines the relative importance of minimising the error between the desired response and the achieved response against minimising the size of the actuator set point changes.

This analysis also suffers from the two problems mentioned above. The conversion of the minimisation problem in equation 4.5.50 to the simpler problem in 4.5.53 requires the use of a number of inequalities which introduce conservativeness into the solution. Also the definition of the set of possible perturbations, \( \mathcal{S} \), can include perturbations which are unrealistic. The derivation of a compensator directly from the minimisation problem in
4.5.50 using a more realistic definition for the set, \( S \), is another future topic of research.

Finally, it is worth investigating the linear independence of the actuator responses. The smallest singular value of \( G \) gives a measure of how far away \( G \) is from being rank deficient. This can be seen by considering a general additive perturbation, \( \Delta' \) (which is not necessarily in \( S \)). We use the fact that [Golub and van Loan, 1983, section 2.3]

\[
\min_{\Delta' \text{ s.t. } \det(G+\Delta') = 0} ||\Delta'||_2 = \sigma(G)
\]

4.5.57

This says that the size of the perturbation that can be added to \( G \) before \( G \) loses rank, is equal to the smallest singular value of \( G \) which means that the smaller the minimum singular value of \( G \), the closer \( G \) is to having columns (or actuators) which are linearly dependent. This fact has a bearing on the size of the actuator set points that are required to remove a profile disturbance. We can see this by recalling that

\[
K = V \left[ \Sigma^{-1} \ 0 \right] U^T
\]

4.5.58

where \( \Sigma \) is a diagonal matrix containing the singular values of \( G \). This allows us to see that the maximum singular value of \( K \) (and hence the 2 norm of \( K \)) is related to the smallest singular value of \( G \) by

\[
||K||_2 = \max_{i=1,2N+1} \sigma_i(K) = \min_{i=1,2N+1} \frac{1}{\sigma_i(G)}
\]

4.5.59

Thus the closer \( G \) is to having linearly dependent columns (so that the lowest singular value of \( G \) is small) the larger the 2 norm of \( K \). Because

\[
u^* = Kz
\]

4.5.60

we can see that

\[
||u||_2 = ||Kz||_2 \leq ||K||_2 ||z||_2
\]

4.5.61

For profile errors of unit norm, the 2 norm of \( K \) gives a measure of the size of the actuator set points, so we can see that plants with nearly linear dependent columns are likely to use large actuator set points when controlling profile errors. This agrees with the comments on ill-conditioning and linear dependence in chapter 3.
4.6 The Integrity of the Closed Loop System

In this section we wish to investigate the integrity of the closed loop control system that was described in section 4.3, initially concentrating on the case where there is a failure in one of the actuators. Although the behaviour of the system when there is an actuator failure is important in its own right, the analysis gives an insight into the response of the system under the more common situation when the set point of an actuator reaches its physical limits. Under these circumstances, the actuator is effectively redundant. This analysis is particularly relevant to actuator arrays such as the lip slice on the headbox of a paper machine (as described in section 1.3) which has a restriction on the amount of bending that can be applied by the lip screws.

We will take the response of the actuator array to be
\[ y(nT) = g(q^{-1}) G u(nT) \]  \hspace{1cm} 4.6.1
and we will assume that the controller has the form
\[ k(q^{-1}) K \]  \hspace{1cm} 4.6.2
Further, as indicated in section 4.3, the dynamics of the controller will be taken to include an integral term so that steady state disturbances can be removed. This allows us to write
\[ k(q^{-1}) = \frac{1}{1 - q^{-1}} k'(q^{-1}) \]  \hspace{1cm} 4.6.3

If there is a failure in the \( r \)th actuator, then this means that the \( r \)th set point is always zero. This can be represented by placing the matrix
\[ I_{2N+1} - e_r e_r^T \]  \hspace{1cm} 4.6.4
(where \( e_r \) is a vector of zeros except for a unit entry in the \( r \)th element) in between the controller and the plant as in figure 4.6.1. This will ensure that the set point applied to the \( r \)th actuator is always zero, irrespective of the value generated by the controller. If the closed loop plant has been described so that it is stable when all of the actuators are
Figure 4.6.1  The failure of the $r$th actuator can be represented by placing the matrix, $I_{2N+1} - e_r e_r^T$ (where $e_r$ is a zero vector except for a unity in the $r$th element) between the controller and the plant. This ensures that the set point given to the $r$th actuator is always zero.
functioning normally, then it is readily shown [Owens, 1978] that the stability of the closed loop system with a failure in the rth actuator is determined by the zeros of the rth diagonal element of the matrix

\[
\left[ I_{2N+1} + k(q^{-1}) g(q^{-1}) KG \right]^{-1}
\]

Because we have assumed that we have scalar dynamics, the zeros of the rth diagonal element of this matrix are determined by the poles of \( k(q^{-1}) \) and \( g(q^{-1}) \). Even if the dynamics of the open loop plant are stable, the controller dynamics include an integral term (see equation 4.6.3) so the closed loop system in the presence of actuator failure will be unstable.

One possible solution to this problem is to replace the integrator in \( k(q^{-1}) \) by the term

\[
\frac{1}{1 - \alpha q^{-1}}
\]

with \( \alpha \) being a real constant which is less than, but close to, unity. Provided that there are no unstable poles in \( g(q^{-1}) \) or in the rest of \( k(q^{-1}) \), then this will ensure that the system remains stable when one of the actuators fails. However, using this "near-integrator" is not really a sensible solution because the system is only just stable and the web profile will be degraded quite dramatically following the actuator failure. A better approach is to identify that the actuator has failed and to correct the problem. One technique for monitoring the operation of actuators while the system is running in closed loop will be described in the next chapter. In practice, however, the malfunction of an actuator, in an otherwise stable system, can usually be readily inferred from the degradation of the web profile and the ramping of the set points.

The failure of an actuator is an abnormal occurrence during the running of the plant. However, there is a similar situation when an actuator set point reaches its physical limit, which can be thought of as part of the normal behaviour of the plant. Typically, this situation arises when there is a large cross directional steady disturbance, usually containing a big high frequency component (recall from chapter 3 that the actuator set
points tend to be large when controlling disturbances in the high frequency regions where
the response of an actuator tends to be small). Under these conditions, if the allowable set
point range for any actuator is
\[ u_{\min} \leq u_k \leq u_{\max} \quad \text{for } k=\{-N, \ldots, N\} \]
and the set point demanded for, say, the rth actuator is greater than \( u_{\max} \), then the actual
applied set point is limited to \( u_{\max} \). This is a slightly different situation from complete
actuator failure, where the set point applied to the actuator is zero, because for the case of a
limited set point, the actual value applied to the actuator is a constant, \( u_{\max} \). We can model
this behaviour by placing the matrix
\[ L_{2N+1}^{-1} e_r e_r^T \]
between the integrators and the rest of the controllers, as in figure 4.6.2. Here the input to
the rth integrator is zero, so the set point applied to the rth actuator is the constant, \( u_{\max} \).
However, because all of the dynamic terms are scalar, it is readily seen that the criterion for
the stability of the system with a limited actuator is the same as that for an actuator failure.
This criterion is described in equation 4.6.5, above. We can see that the presence of the
integrator will cause a closed loop system with a limited actuator to be unstable.

This situation is unsatisfactory because the limiting of actuator set points is a common
occurrence on some plants, so the instability cannot be tolerated. There are two basic
approaches to overcoming the problem:

i) adjust the control strategy to prevent the actuator set points exceeding their limits;
ii) allow set points to exceed their limits where appropriate and recalculate the set
points of the other actuators to compensate for the limited actuators.

[Chen et al, 1986] advocate the first approach. Recall from section 4.3, that the inclusion of
the \( \mu \) factor weighted the size of the set points (see equation 4.3.51). By the time that these
values have been applied to the integrator in the controller dynamics, including the \( \mu \) factor
is equivalent to weighting the changes in the set points, rather than the set points
themselves. Chen et al suggest that the actuator set points are chosen to match the.
Figure 4.6.2 The case where the set point of the rth actuator reaches its physical limits can be represented by separating out the integral term from the controller dynamics and placing the matrix, $I_{2N+1}^{-1}e_r e_r^T$ before it (where $e_r$ is a zero vector except for a unity in the rth element). This ensures that the set point given to the rth actuator is always constant (ie equal to the physical limit of the set point).
desired response as closely as possible, while weighting both the size of the set point changes and the absolute values of the set points. Costing large set point values, tends to keep the set points away from their limits.

There are two problems with this approach. Firstly, weighting the the absolute value of the set points may mean that the desired response is not matched as closely as possible. For example, a better response may be achieved by letting the actuator set points reach their limits, rather than holding them to a lower value. Secondly, unless non-linear weightings are used, it is still possible that certain profile errors will cause the actuators to saturate, even if the set point weightings are included. Chen et al acknowledge these problem and suggest that the optimal answer is to choose the set points using quadratic minimisation combined with linear constraints for incorporating the limits of the set points. As they point out, this technique requires a major computational effort at each control action.

In this study, we will concentrate on the second strategy listed above, where we allow the actuator set points to reach their physical limits and then adjust the set points of other actuators to compensate for this situation. To illustrate this approach, we will assume that a new vector of set points has been generated and that the value of the set point of the rth actuator, $u_r$, is greater than the physical limit, $u_{\text{max}}$. This means that the achieved response of the actuator array is the optimal response, less the component due to the limited actuator.

We will denote this component as $z_r$ and it can be seen that

$$z_r = (u_r - u_{\text{max}}) g_r$$

4.6.9

where $g_r$ is the rth column of the $G$ matrix which describes the response of the rth actuator.

We can now adjust the set points of the other actuators in an attempt to match $z_r$ and thus compensate for the limiting of the rth actuator.

If we use the same set point weighting factor, $\mu$, as in the original calculation of the set points, the adjustments to the set points are given by

$$u_r = \left[ G_r^T G_r + \mu I_{2N} \right]^{-1} G_r^T z_r$$

4.6.10
where $G_r \in \mathbb{R}^{(mx2N)}$ is the $G$ matrix with its $r$th column removed. The vector of set points, $u_r \in \mathbb{R}^{2N}$, (note that there will not be an entry for the $r$th actuator) can be added to the set points that were obtained from the original calculation and applied to the plant.

At first sight it would appear that calculating $u_r$ would require a large amount of computational effort because it is necessary to invert the $2N x 2N$ matrix $(G_r^T G_r + \mu I_{2N})$. However if we decompose the problem using the Cholesky factorisation as described in section 4.4, then we see that the calculation of $u_r$ in equation 4.6.10 is closely related to the calculation of the set points using all actuators (i.e. the standard case where the $r$th actuator has not reached its limit) which results in a considerable simplification.

To show this, we can rearrange 4.6.10 to give

$$\left[ G_r^T G_r + \mu I_{2N} \right] u_r = G_r z_r \quad 4.6.11$$

As in section 4.4, we will decompose $(G_r^T G_r + \mu I_{2N})$ using a Cholesky factorisation, to give

$$R_r^T R_r u_r = G_r z_r \quad 4.6.12$$

where

$$\left[ G_r^T G_r + \mu I_{2N} \right] = R_r^T R_r \quad 4.6.13$$

Defining

$$f_r = G_r z_r \quad 4.6.14$$

and

$$g_r = R_r u_r \quad 4.6.15$$

the vector of set points can be obtained by solving for $g_r$ in

$$R_r^T g_r = f_r \quad 4.6.16$$

and then solving for $u_r$ in

$$R_r u_r = g_r \quad 4.6.17$$

As described in section 4.4, this procedure is efficient in terms of both the number of
operations used and the storage requirements, as well as being numerically stable. However, the major computational effort appears to be in calculating the Cholesky factor, \( R_r \). Fortunately, \( R_r \) can be obtained directly from the Cholesky factor, \( R \), of the full matrix which includes all of the columns, i.e.

\[
R^T R = \left[ G^T G + \mu I_{2N+1} \right]
\]

This factorisation will always be available because it is required for calculating the unlimited set points.

In order to see the connection between \( R \) and \( R_r \), we note that \( G \) is related to \( G_r \) by

\[
G_r = G E_r
\]

with \( E_r \in \mathbb{R}^{(2N+1)\times 2N} \) is the identity matrix with the \( r \)th column eliminated.

\[
E_r = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\]

and post multiplication of \( G \) by \( E_r \) has the effect of removing the \( r \)th column of \( G \). This allows us to write

\[
\left[ G_r^T G_r + \mu I_{2N} \right] = E_r^T \left[ G^T G + \mu I_{2N+1} \right] E_r
\]

where we have used the fact that

\[
E_r^T I_{2N+1} E_r = I_{2N}
\]

Substituting the Cholesky factorisations into 4.6.21 we obtain

\[
R_r^T R_r = E_r^T R^T R E_r
\]

\[
= \tilde{R}_r^T \tilde{R}_r
\]

where \( \tilde{R}_r \) is the Cholesky factor, \( R \), with the \( r \)th column removed. It should be stressed that \( \tilde{R}_r \) is not equal to the Cholesky factor \( R_r \). This can be seen by noting that \( \tilde{R}_r \) is an upper
Hessenberg matrix rather than upper triangular. This means that there will be extra elements in the positions \([\tilde{R}_r]_{(r+1),r}, [\tilde{R}_r]_{(r+2),r+1}, \ldots, [\tilde{R}_r]_{(2N+1),2N}\). These elements need to be zeroed before the matrix can be made upper triangular.

We can illustrate this with a G matrix which is in \(R^{6 \times 5}\), where we remove the second column. \(\tilde{R}_r\) will have the form

\[
\tilde{R}_r = \begin{bmatrix}
x & x & x & x \\
x & x & & & \\
x & x & & & \\
x & & & & \\
x & & & & \\
x & & & & \\
\end{bmatrix}
\]

The elements \([\tilde{R}_r]_{3,2}, [\tilde{R}_r]_{4,3}\) and \([\tilde{R}_r]_{5,4}\) must be made zero before the matrix becomes upper triangular. The selective zeroing of elements of a matrix can be accomplished using Givens transformations [Golub and van Loan, 1983, section 3.4]. These have the form

\[
J(p,q) = \begin{bmatrix}
1 & & & & \\
& & & & \\
& c & s & & \\
& -s & c & & \\
& & & & 1
\end{bmatrix} \quad \leftarrow \text{pth row}
\]

\[
J(p,q) = \begin{bmatrix}
& & & & \\
& & & & \\
c & s & & & \\
-s & c & & & \\
& & & & 1
\end{bmatrix} \quad \leftarrow \text{qth row}
\]

where \(c\) and \(s\) are chosen so that the pth element of the qth column of a matrix is zeroed by adjusting the qth row. Thus applying this transformation to the matrix in equation 4.6.25, results in

\[
J(2,3) \tilde{R}_r = \begin{bmatrix}
x & x & x & x \\
x & x & & & \\
x & & & & \\
x & & & & \\
x & & & & \\
x & & & & \\
\end{bmatrix}
\]

Here the(3,2) element of the matrix has been zeroed and the second row has been amended.
Applying a number of these transformations successfully reduces all of the elements in the subdiagonal to zero. Thus

\[
\begin{bmatrix}
\times & \times & \times \\
\times & \times & & \\
\times & & & \\
& & & \\
& & & \\
& & & \\
\end{bmatrix}
\]

4.6.28

Note that this is an upper triangular \((2N+1)x2N\) matrix. Although we have reduced \(\mathbf{R}_c\) to an upper triangular matrix, it does not immediately follow that the result is in fact the Cholesky decomposition, \(\mathbf{R}_c\). However, it can be shown that the first \(2N\) rows of \(\mathbf{R}_c\) (ie ignoring the last row of zeros in 4.6.28) is equal to \(\mathbf{R}_c\) [Stewart, 1979].

This procedure is best illustrated by a simple example. Suppose \(G\) is a 6x5 matrix

\[
G = \begin{bmatrix}
8 & 5 & 3 & 2 & 1 \\
4 & 7 & 4 & 3 & 2 \\
3 & 6 & 8 & 5 & 2 \\
2 & 5 & 7 & 5 & \\
2 & 3 & 6 & 8 & \\
1 & 2 & 4 & 6 & 8 \\
\end{bmatrix}
\]

4.6.29

and we calculate the Cholesky factorisation of \((G^T G + \mu I_{2N+1})\) where \(\mu=1\). This gives

\[
\begin{bmatrix}
9.95 & 10.45 & 9.45 & 7.94 & 5.03 \\
0 & 6.30 & 8.45 & 7.77 & 5.47 \\
0 & 0 & 5.51 & 6.95 & 5.50 \\
0 & 0 & 0 & 4.03 & 5.05 \\
0 & 0 & 0 & 0 & 3.61 \\
\end{bmatrix}
\]

4.6.30

We now remove the second column from \(G\) to give
then the Cholesky factorisation of \((G_2^T G_2 + \mu I_{2N})\) is
\[
R_2 = \begin{bmatrix}
9.95 & 9.45 & 7.94 & 5.03 \\
0 & 10.09 & 10.31 & 7.59 \\
0 & 0 & 4.32 & 5.29 \\
0 & 0 & 0 & 3.63
\end{bmatrix}
\]

Returning to the original factorisation, \(R\), we can remove the second column to give
\[
\tilde{R}_2 = \begin{bmatrix}
9.95 & 9.45 & 7.94 & 5.03 \\
0 & 8.45 & 7.77 & 5.47 \\
0 & 5.51 & 6.95 & 5.50 \\
0 & 0 & 4.03 & 5.05 \\
0 & 0 & 0 & 3.61
\end{bmatrix}
\]

This is an upper Hessenberg matrix. Applying a Givens transformation to zero the (3,2) element gives
\[
\begin{bmatrix}
9.95 & 9.45 & 7.94 & 5.03 \\
0 & 10.09 & 10.31 & 7.59 \\
0 & 0 & 1.57 & 1.62 \\
0 & 0 & 4.03 & 5.05 \\
0 & 0 & 0 & 3.61
\end{bmatrix}
\]

Repeating this for the (4,3) and (5,4) elements leads to
which agrees with the Cholesky factorisation in equation 4.6.32.

We can now see how to use this procedure to provide a technique for modifying the other actuator set points when the rth actuator set point is limited. We take the Cholesky factorisation for the full array, delete the rth column and then apply a series of Givens transformations to eliminate the non-zero elements of the first sub-diagonal, resulting in a new Cholesky factorisation, \( R_r \). This factorisation can then be used in equations 4.6.16 and 4.6.17 to solve for \( u_r \). The values in this vector can then be used to correct the set points of the unlimited actuators. The number of Givens transformations required depends upon which actuator has reached its physical limit (because 2N transformations are required if the first column of \( R \) is deleted while no transformations are required if the last column of \( R \) is removed). At most there will be 2N transformations. Since the maximum number of calculations required for each transformation is 4(2N+1) flops and one square root, the total number of operations performed does not exceed 8N(2N+1) flops and 2N square roots. This is a significant saving in computational operations relative to performing a complete Cholesky decomposition for \( (G_r^T G_r + \mu I_{2N}) \) which requires \( 8(2N+1)^{3/6} \) flops [Golub and van Loan, 1983].
The procedure for adjusting the Cholesky factorisation when a column is deleted could also be used for adjusting the control algorithm in the presence of an actuator failure. Once the failure has been noticed, then the effect of the actuator can be removed by calculating all future set points with a reduced Cholesky factorisation which does not include the corresponding column. Once the actuator has been repaired or replaced, there are efficient techniques for adding a column back into the factorisation (ie for doing the reverse of the above). These techniques are described in detail in [Gill et al, 1974].
CHAPTER 5: ESTIMATING THE RESPONSE OF THE SYSTEM

From the previous chapter, we know that an important component of the control design is a knowledge of the G matrix. Although section 4.5 has shown us how to design a system which is robust to errors in the response of the actuators, it is advantageous to estimate the actuator responses in the G matrix because:

i) an estimation procedure allows us to learn the responses of a system when commissioning a new plant or when changing between different grades of web on an existing system;

ii) we can use the estimation procedure to monitor plant for any changes in the response of the system, so that the control design can be adapted to accommodate the changes (recall that although the control design is robust, the performance of the system is degraded if the true G matrix does not match the matrix that was used for the design).

An estimation procedure is particularly useful for monitoring the mapping of the position of an actuator in the array to its effect on the web (this was described in section 4.5). The mapping of the system can vary quite markedly on some web forming processes and mapping errors have been identified as the major sources of instability and poor performance of cross directional control schemes [Chen et al, 1986] [Dolphin, 1988] [Karlsson et al, 1985] [Lindeborg, 1986]. Most current control schemes measure the position of the effect of an actuator using manual methods (for example, marking the web at the centre line of an actuator and measuring the position of the mark at the point of gauging). This technique is susceptible to measurement error and can not track any variations in the mapping, but if we can estimate the G matrix online, then we have an immediate measure of the mapping because we know the position of the effect of every actuator.
The columns of the matrix $G$ hold the impulse responses of the $2N+1$ actuators and estimating the effect of every actuator at every measuring point would require a massive amount of computation. We have assumed (see Assumption 2.1.1c) that the shape of each of the responses is the same and that the responses are equally spaced. However we are interested in only the shape and position of the responses so we do not need to estimate all of the elements of the $G$ matrix. The elements of a single column of a $G$ describe the shape of the response of one of the actuators and given estimates of the elements of this column, we can estimate the shape of the response of the other actuators. Also, because the response of a single actuator does not cover all of the measuring points across the web, it is necessary to perform the estimation only at the elements of the column corresponding to the positions of the likely effect of the response.

If we are given estimates of two of the columns of $G$, we can obtain the centre lines of their responses. Because we have assumed that the actuators are equally spaced, from the positions of the two estimated responses, we can deduce the positions of the responses of all of the actuators by interpolation. In order to minimise the errors in deducing the positions of the other responses, it is sensible to estimate the responses of actuators which are well separated, for example on opposite sides of the array of actuators.

This argument shows us that, under the assumptions that we have made about the model, it is possible to deduce an estimate of the whole of the $G$ matrix if we are given estimates of the non zero terms in just two of its columns, which leads to a major reduction in the size of the estimation problem. Although the analysis of this chapter concentrates on estimating the $G$ matrix by estimating the response of two of the actuators, it should be apparent that the techniques described can be used to monitor the responses of individual actuators. By "polling" each actuator in turn, any aging or failed actuators can be identified and replaced or the control algorithm can be adjusted to accommodate the malfunction, as described in section 4.6.
Section 5.1 develops an online estimation procedure which operates while the system is running in closed loop. This is done by adding a small perturbation signal onto the set points of two of the actuators and estimating the closed loop response between these perturbation signals and the web profile. An estimate of the G matrix is obtained from the closed loop response. Although the injection of a perturbation signal means that the estimator is not strictly "non-invasive", the amplitude of the perturbation can be made small enough so that its effect on the web profile is not seen above the level of the process noise. Section 5.2 describes the implementation of a recursive version of this estimator and section 5.3 considers the design of the perturbation signal in order to optimise the estimate of G and to reduce the dependence of the technique on the choice of model order.

Online estimation of actuator responses does not appear to have been used extensively on web forming processes. The only reported implementation is described in Chen et al [Chen et al, 1986] where a large step change is applied to one actuator and its effect on the web profile is observed. Because the perturbation needs to be large and the system has to operate in open loop mode while the estimate is made, this procedure is less practical than the approach described in this chapter which uses a small perturbation and estimates the response while the system runs in closed loop.

Adding a small perturbation signal into a feedback loop is a well known technique for estimating a transfer function while a system is running in closed loop [Gustavsson et al, 1977] [Gustavsson et al, 1981] but the simplification of the estimation algorithm provided by the special structure of the G matrix in cross directional control systems has not been exploited. Similarly, the recursive implementation of the estimator that is described in section 5.2 is based on standard theory [Ljung and Soderstrom, 1983] but some of the computational issues, such as the use of one covariance matrix for a number of estimators (described in more detail below) have not been previously reported. The choice of the optimal perturbation signal has received extensive research [Mehra, 1981] [Goodwin and Payne, 1977] and section 5.3 uses some of these results. However, for the particular
model structure used for describing the plant, there is a link between the choice of input signal and reducing the sensitivity of the estimator to the choice of model order, which has not been made before.

5.1 Estimating the Impulse Response of an Actuator

In this section we describe a technique for obtaining an estimate of the impulse response of one of the actuators. As discussed above, a practical constraint on this procedure is that it must be performed in closed loop. We will consider the discrete time, discrete space model of the plant which was developed in chapter 4. From equation 4.2.1 we know that the response of the array of actuators is given by

\[ y(nT) = g(q^{-1}) \cdot G \cdot u(nT) \]  

5.1.1

where \( y(nT) \in \mathbb{R}^m \) is the vector of observed profile changes at the \( m \) measuring points across the web, \( u(nT) \in \mathbb{R}^{2N+1} \) is the vector of actuator set points, \( G \in \mathbb{R}^{mx(2N+1)} \) is the matrix containing the impulse responses of the \( 2N+1 \) actuators at the \( m \) measuring points, \( g(q^{-1}) \) is the scalar dynamic term describing the dynamics of all of the actuators (note that each of the actuators is taken to have identical dynamics, so that all of the dynamics can be described by a scalar function). \( T \) is the time between samples and \( n \) is the number of the sample.

Equation 5.1.1 describes the profile response due to the actuators. The observed profile can be written as

\[ z(nT) = y(nT) + e(nT) + d \]  

5.1.2

where \( e(nT) \in \mathbb{R}^m \) is a vector of noise terms and \( d \in \mathbb{R}^m \) is the steady state disturbance in the system (as discussed in chapter 4). The steady disturbance, \( d \), can be split into two parts

\[ d = d_c + d_u \]  

5.1.3

with \( d_c \) being the controllable part of \( d \) (ie the components of \( d \) which have frequencies less than \( \omega_c \)) and \( d_u \) being the uncontrollable part.
We will assume that the system is operating under a closed loop control law given by
\[ u(nT) = -k(q^{-1}) K z(nT) \] 5.1.4
where \( K \in \mathbb{R}^{(2N+1) \times m} \) is the controller matrix and \( k(q^{-1}) \) is the scalar dynamic term (see section 4.3).

Initially we will assume that the actuator dynamics, \( g(q^{-1}) \), are known, either from a separate experiment or from a priori knowledge of the system. It will also be taken that the controller matrix, \( K \) and the controller dynamics, \( k(q^{-1}) \) are known. For the moment we will assume that there is no steady state disturbance (ie \( d=0 \)). This means that we can write the system equations as
\[ z(nT) = g(q^{-1}) G u(nT) + e(nT) \] 5.1.5
\[ u(nT) = -k(q^{-1}) K z(nT) \] 5.1.6
The most obvious approach to estimating the impulse response matrix \( G \) for a given set of measurements, \( z(nT) \) and set points \( u(nT) \), for \( n = 1,2,...,n_{\text{max}} \), is to take the estimate of \( G \) (which will be denoted \( \hat{G} \)) as the value which minimises
\[ \min_{\hat{G}} \sum_{n=1}^{n_{\text{max}}} \| z(nT) - g(q^{-1}) \hat{G} u(nT) \|_2 \] 5.1.7
We will simplify the notation by writing
\[ e(nT;\hat{G}) = z(nT) - g(q^{-1}) \hat{G} u(nT) \] 5.1.8
so equation 5.1.5 becomes
\[ \min_{\hat{G}} \sum_{n=1}^{n_{\text{max}}} \| e(nT;\hat{G}) \|_2 \] 5.1.9
This minimisation problem can be viewed as choosing the value of \( \hat{G} \) which minimises (in a 2 norm sense) the prediction error, \( e(nT;\hat{G}) \) between the observed profile, \( z(nT) \) and the expected profile due to the actuator response, \( g(q^{-1}) \hat{G} u(nT) \).

From equations 5.1.5 and 5.1.6 we can see that
\[ z(nT) = \left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} e(nT) \] 5.1.10

\[ u(nT) = -k(q^{-1}) K \left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} e(nT) \] 5.1.11

Substituting these values into expression 5.1.8 for \( e(nT;\hat{G}) \) gives

\[
e(nT;\hat{G}) = \left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} e(nT) \]

\[ + g(q^{-1}) k(q^{-1}) \hat{G} K \left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} e(nT) \] 5.1.12

\[ = \left[ I + g(q^{-1}) k(q^{-1}) \hat{G} K \right] \left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} e(nT) \] 5.1.13

It can be shown [Soderstrom et al, 1976] that the optimal value of \( \hat{G} \) makes

\[ e(nT;\hat{G}) = e(nT) \] 5.1.14

for \( n = 1,2,...,n_{\text{max}} \) provided that \( e(nT) \) is independent of the input sequence \( u[(n-1)T], u[(n-2)T] \) etc. Intuitively, this is a sensible result because it means that the optimal estimate is the one which makes the estimation error equal to the noise alone and there is no error due to an incorrect estimate of \( G \).

Applying this result to equation 5.1.13 we can see that the optimal value of \( \hat{G} \) is the one which satisfies

\[ I = \left[ I + g(q^{-1}) k(q^{-1}) \hat{G} K \right] \left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} \] 5.1.15

for all values of \( q \) in the complex plane. Provided that both \( G \) and \( K \) are of full rank, then the optimal solution is given by [Soderstrom et al, 1976]

\[ \hat{G} = G \] 5.1.16

as required. Thus if we were to perform the minimisation given in equation 5.1.7, using standard least squares techniques, we would be able to obtain a true estimate of the impulse response matrix, \( G \).

There are a number of disadvantages with this approach:

(i) The minimisation procedure requires readings from all \( m \) measuring points and all of the \( 2N+1 \) set points, in order to estimate the whole of the \( G \) matrix. For practical systems, this results in a massive estimation procedure and no use is made of the fact that the whole of the \( G \) matrix can be generated from estimates.
of two of its columns.

(ii) If there is a steady state disturbance on the system (as will be likely in practice) the set points, \(u(nT)\) will be offset to compensate for this disturbance. Unless this offset is accounted for in the estimation procedure, there will be bias in the elements of \(\hat{G}\).

(iii) Because we require the noise \(e(nT)\) to be independent of the input sequence so that the optimal value of \(\hat{G}\) can be obtained (see equation 5.1.14) this means that the noise sequence must be white (in the sense that \(E[e^T(n_1 T) e(n_2 T)] = \delta_{n_1,n_2}\sigma^2\) where \(\sigma^2\) is the variance of the noise vector). The noise is not white in practice as a result of the feedback control law which causes the input sequence to be correlated with the past noise, because \(u(nT)\) is determined from \(z(nT)\) via equation 5.1.6.

The procedure that has just been described is a direct approach to the estimation problem where \(\hat{G}\) is obtained from the open loop response given the actuator set points \(u(nT)\) and the profile measurements \(z(nT)\). We now develop an indirect approach to estimating \(G\), where an estimate of the closed loop is generated and then \(G\) is obtained from this estimate. This turns out to be much simpler than the direct approach. The price that is paid for this simplification is that it is necessary to introduce a small perturbation into the system, so the method is not strictly non-invasive.

Soderstrom et al [Soderstrom et al, 1976] suggest that one approach to closed loop estimation is to introduce a perturbation, \(v(nT)\in \mathbb{R}^{(2N+1)}\) which is added to the actuator set points. This makes the system equations

\[
\begin{align*}
z(nT) &= g(q^{-1}) G u(nT) + e(nT) + d \\
u(nT) &= -k(q^{-1}) K z(nT) + v(nT)
\end{align*}
\]

The system is illustrated in figure 5.1.1. By substituting for \(u(nT)\) in equation 5.1.17 we
Figure 5.1.1 The addition of a perturbation, \( v(nT) \), to the actuator set points for the purpose of actuator response estimation.
obtain the closed loop response between \( v(nT) \) and \( z(nT) \)

\[
z(nT) = \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} \left( g(q^{-1}) G v(nT) + \left[ I + g(q^{-1}) k(q^{-1}) G \right]^{-1} e(nT) \right) + \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} d(nT) \tag{5.1.19}
\]

As mentioned above, the steady disturbance can be split into a controllable part and an uncontrollable part, so the last term becomes

\[
\left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} d = \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} d_e + \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} d_u \tag{5.1.20}
\]

From chapter 4 (equations 4.3.48 and 4.3.49), we know that if the controller dynamics include an integrator, then the effect of the controllable part of the disturbance will be removed from the closed loop response (so the first term on the right hand side of 5.1.20 is zero) and the uncontrollable part is unaffected by the controller. This means that when the system is working in steady state

\[
\left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} d = d_u \tag{5.1.21}
\]

We will simplify the notation in 5.1.19 by writing

\[
\bar{e}(nT) = \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} e(nT) \tag{5.1.22}
\]

Note that if \( e(nT) \) is white, \( \bar{e}(nT) \) is coloured by the term \( \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} \).

Assume that the system is operating in steady state, so that the controllable part of the disturbance has been removed, then the closed response of the system becomes

\[
z(nT) = \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} g(q^{-1}) G v(nT) + \bar{e}(nT) + d_u \tag{5.1.23}
\]

In order to convert this expression into the form required for the subsequent analysis, it is necessary to expand the return difference term, \( \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} \), using the following theorem.

**Theorem 5.1.1** Given an open loop transfer function, \( Q(q^{-1}) \in \mathbb{C}^{m \times m} \), with a state space realisation
\[ x[(n+1)T] = A_Q x(nT) + B_Q u(nT) \]  \hspace{1cm} 5.1.24
\[ y(nT) = C_Q x(nT) + D_Q u(nT) \]  \hspace{1cm} 5.1.25

then provided that the matrix \((I+D_Q)^{-1}\in\mathbb{R}^{m\times m}\) is invertible, the closed loop return difference, \(T(q^{-1})\in\mathbb{C}^{m\times m}\), given by
\[ T(q^{-1}) = \left[ I_m + Q(q^{-1}) \right]^{-1} \]  \hspace{1cm} 5.1.26

can be written as an infinite series of the form
\[ T(q^{-1}) = \left[ I_m + D_Q \right] + T_1 q^{-1} + T_2 q^{-2} + ... \]  \hspace{1cm} 5.1.27

where \(T_1, T_2, ... \in \mathbb{R}^{m\times m}\). Further, if the closed loop is stable (ie all the poles of \(T(q^{-1})\) are inside the unit disc) then successive terms of the series become smaller in the sense that
\[ \| T_j \|_2 \rightarrow 0 \text{ as } j \rightarrow \infty \]  \hspace{1cm} 5.1.28

**Proof** We will use the notation \([A, B, C, D]\) to represent a state space realisation, so that the realisation of \(Q(q^{-1})\) is denoted by \([A_Q, B_Q, C_Q, D_Q]\). It is readily seen that the \(I_m\) term in the expression for the return difference corresponds to the realisation \([0, 0, 0, I_m]\). This allows us to write the state space realisation of \([I_m + Q(q^{-1})]\) as \([A_Q, B_Q, C_Q, I_m+D_Q]\).

Using the fact that the realisation of the inverse transfer function is given by [Francis, 1987]
\[ [A, B, C, D]^{-1} = [A-BD^{-1}C, BD^{-1}C, -D, D^{-1}] \]  \hspace{1cm} 5.1.29
(provided that \(D^{-1}\) exists) a state space realisation of the return difference, \(T(q^{-1})\), is
\[ [A_T, B_T, C_T, D_T] = [A_Q^{-1}B_Q[I_m+D_Q^{-1}C_Q, B_Q[I_m+D_Q^{-1}C_Q, -I_m+D_Q^{-1}]C_Q, [I_m+D_Q^{-1}]C_Q]^T] \]  \hspace{1cm} 5.1.30

It is well known that \(T(q^{-1})\) can be expressed in terms of the Markov parameters of the state space realisation (see for example [Kailath, 1980, section 2.3])
\[ T(q^{-1}) = D_T + \sum_{j=1}^{\infty} C_T A_T^{j-1} B_T q^j \]  \hspace{1cm} 5.1.31

By noting that \(D_T\) is \(I_m+D_Q\) and by identifying the \(T_j\)'s as
\[ T_j = C_T A_T^{j-1} B_T \]  \hspace{1cm} 5.1.32
we obtain the expression in 5.1.27.
In order to prove that the size of the terms in the $5.1.27$ decrease as $j$ becomes large, we use the fact that

\[
\|T_j\|_2 = \|C_T A_T^{j-1} B_T\|_2 \\
\leq \|C_T\|_2 \|A_T^{j-1}\|_2 \|B_T\|_2
\]

If the closed loop is stable, then this means that all of the eigenvalues of $A_T$ must lie inside the unit circle and in this case it is readily shown that [Davis and Vinter, 1986, appendix D]

\[
\|A_T^{j-1}\|_2 \rightarrow 0 \text{ as } j \rightarrow \infty
\]

Because $\|C_T\|_2$ and $\|B_T\|_2$ are fixed, this means that the closed loop system is stable, $\|T_j\|_2$ must decrease to zero as $j$ increases.

This ends the proof.

From this theorem we can see that if a unit pulse is injected into the closed loop, then the size of the signal due to this pulse $j$ steps later is given by $\|T_j\|_2$ and if the closed loop is stable, the magnitude of the signal decays with time.

The theorem does involve two conditions: that the closed loop is stable and that $I_m + D_Q$ is invertible. The stability condition will be discussed in more detail below, but the requirement for $I_m + D_Q$ is readily satisfied for practical systems. This is a result of the transport delay which is inherent in web forming processes and means that if a change is applied to the set point of an actuator at time sample $t=nT$, then its effect is not seen until at least one sample later, at $t=(n+1)T$. This means that a state space realisation of the plant must have a zero $D$ matrix. Even if the controller does have a non-zero $D$ matrix in its realisation, the effect of the transport delay in the plant will ensure that the $D_Q$ matrix in the realisation of the open loop transfer function, $Q(q^{-1})$, is also zero. If $D_Q$ is zero then this will ensure that $I_m + D_Q$ is invertible and in all of the subsequent discussion we will assume this to be the case.
We can now use the expansion of $T(q^{-1})$ in equation 5.1.27 to obtain an expression for the closed loop response between $z(nT)$ and $v(nT)$ given in equation 5.1.23. For convenience, we will denote the term

$$\left[ I + g(q^{-1}) k(q^{-1}) GK \right]^{-1} g(q^{-1}) G$$  \hspace{1cm} 5.1.36

by $H(q^{-1})$. This means that

$$H(q^{-1}) = T(q^{-1}) g(q^{-1}) G$$  \hspace{1cm} 5.1.37

Given a state space realisation for the open loop transfer function

$$Q(q^{-1}) = g(q^{-1}) k(q^{-1}) GK$$  \hspace{1cm} 5.1.38

and provided that the system is closed loop stable, then $H(q^{-1})$ can be written as

$$H(q^{-1}) = \left[ I_m + T_1 q^{-1} + T_2 q^{-2} + \ldots \right] g(q^{-1}) G$$  \hspace{1cm} 5.1.39

where we have taken $D_Q$ to be zero. If the dynamics of the plant are stable (as is likely to be the case for practical plants as discussed below) then it is possible to expand $g(q^{-1})$ as

$$g(q^{-1}) = g_0 + g_1 q^{-1} + g_2 q^{-2} + \ldots$$  \hspace{1cm} 5.1.40

and using the same argument as in the second half of the proof of theorem 5.1.1, for stable dynamics, $|g_j|$ will decrease to zero as $j$ increases. In addition, because we have assumed that the system has a transport delay (and hence the realisation of the plant has no $D$ matrix) $g_0$ at least will be zero. If the transport delay is longer than one sample then the terms $g_1$, $g_2$, etc may also be zero. For this analysis we will assume that the transport delay is less than one sample so that $g_0$ is zero and $g_1$ (at least) is non-zero, although the arguments are readily adapted to the case of a longer delay.

We can now write

$$H(q^{-1}) = \left[ I_m + T_1 q^{-1} + T_2 q^{-2} + \ldots \right] \left[ g_1 G q^{-1} + g_2 G q^{-2} + \ldots \right]$$  \hspace{1cm} 5.1.41

which can be simplified by writing

$$H(q^{-1}) = H_1 q^{-1} + H_2 q^{-2} + H_3 q^{-3} + \ldots$$  \hspace{1cm} 5.1.42

where
\[ H_1 = g_1G \] 5.1.43
\[ H_2 = g_1G T_1 + g_2G \] 5.1.44
\[ H_3 = g_1G T_2 + g_2G T_1 + g_3G \] 5.1.45

From equation 5.1.41, it is evident that \( H(q^{-1}) \) is the product of two series where the size of the coefficients decrease with increasing powers of \( q^{-1} \) and it is readily seen that the size of the matrix coefficients in the series for \( H(q^{-1}) \) must also decrease with increasing powers of \( q^{-1} \). The value of \( j \) for which \( \| H_j \|_2 \) becomes negligible can be identified as the settling time for the closed loop response, because if a unit pulse is applied to the input of the loop, after \( j \) samples, the size of its effect will have virtually died to zero. In the subsequent analysis, we will truncate \( H(q^{-1}) \) so that it becomes

\[ H(q^{-1}) = H_1 q^{-1} + H_2 q^{-2} + H_3 q^{-3} + \ldots + H_p q^{-p} \] 5.1.47

where \( p \) is the number of samples required for the closed loop to settle. In addition, we will be mainly interested in the first term of the series so we will usually write \( H_1 \) explicitly as \( g_1G \).

We can now use this expansion in developing a procedure for estimating the closed loop response. For the moment, we will assume that there is no uncontrollable disturbance, so equation 5.1.23 becomes

\[ z(nT) = H(q^{-1}) v(nT) + \epsilon(nT) \] 5.1.48

Given a set of exogenous perturbations, \( v(nT) \), and observations, \( z(nT) \), we can estimate the closed loop response as the value of \( H(q^{-1}) \) which minimises the expression

\[ \min_{\hat{H}(q^{-1})} \sum_{n=1}^{n_{\text{max}}} \| z(nT) - \hat{H}(q^{-1}) v(nT) \|_2 \] 5.1.49

Writing the prediction error as

\[ \epsilon(nT;\hat{H}(q^{-1})) = z(nT) - \hat{H}(q^{-1}) v(nT) \] 5.1.50

the minimisation procedure becomes

\[ \min_{\hat{H}(q^{-1})} \sum_{n=1}^{n_{\text{max}}} \| \epsilon(nT;\hat{H}(q^{-1})) \|_2 \] 5.1.51

213
As before, provided that:

(i) the noise sequence, $\varepsilon(nT)$, is independent of the exogenous input sequences, $v(nT)$;

(ii) the order of the $\hat{H}(q^{-1})$ polynomial is chosen so that $\hat{H}(q^{-1})$ can model the true closed response, $H(q^{-1})$;

then the value of $\hat{H}(q^{-1})$ which minimises this problem is the one which makes $\varepsilon(nT;\hat{H}(q^{-1}))$ equal to $\varepsilon(nT)$ for $n=\{1,\ldots,n_{\text{max}}\}$ [Soderstrom et al, 1976]. From equation 5.1.50, we know that

\[\varepsilon(nT;\hat{H}(q^{-1})) = H(q^{-1}) v(nT) + \varepsilon(nT) - 11(q^{-1}) v(nT) \]

\[= [H(q^{-1}) - \hat{H}(q^{-1})] v(nT) + \varepsilon(nT)\]

(where we have used equation 5.1.48). For $\varepsilon(nT;\hat{H}(q^{-1}))$ to equal $\varepsilon(nT)$, this implies that

\[H(q^{-1}) = \hat{H}(q^{-1})\]

(provided that $H(q^{-1})$ and $\hat{H}(q^{-1})$ are not rank deficient) so the estimate obtained matches the true closed response of the system.

The requirement that the noise sequence, $\varepsilon(nT)$, being independent of $v(nT)$ is not severe, because we have the freedom to choose an exogenous input which will satisfy this condition. A typical example of such an input is a pseudo random binary sequence which will be independent of any general noise sequence. This means that the estimation procedure will obtain the correct $\hat{H}(q^{-1})$ even if $\varepsilon(nT)$ or the underlying sequence $\varepsilon(nT)$ is coloured noise. The second convergence criterion concerning the model order is more restrictive, but, as will be seen in section 5.3, the sensitivity to errors in the model order can be reduced by careful selection of the input sequence, $v(nT)$.

So far, we have ignored the effect of the uncontrollable disturbance, $d_u$, on the system. If such a disturbance exists on the plant, then it is important to include it in the model and to estimate its value, otherwise the elements of $\hat{H}(q^{-1})$ will be biased. This bias can be seen by writing the estimation error as
\[ \varepsilon(nT;H(q^{-1})) = z(nT) - \hat{H}(q^{-1}) v(nT) \] 5.1.55

when

\[ z(nT) = H(q^{-1}) v(nT) + \bar{e}(nT) + d_u \] 5.1.56

This means that \( \varepsilon(nT;H(q^{-1})) \) becomes

\[ \varepsilon(nT;\hat{H}(q^{-1})) = \left[ H(q^{-1}) - \hat{H}(q^{-1}) \right] v(nT) + d_u + \bar{e}(nT) \] 5.1.57

If the optimal \( \hat{H}(q^{-1}) \) is obtained when \( \varepsilon(nT;\hat{H}(q^{-1})) \) is equal to \( \bar{e}(nT) \) for \( n=\{1,\ldots,n_{\text{max}}\} \), then it can be seen that \( \hat{H}(q^{-1}) \) will not converge to its true value, \( H(q^{-1}) \).

The uncontrollable disturbance, \( d_u \), can be considered as a steady state offset to the data, \( z(nT) \). One approach to removing an offset is to difference the data, but as pointed out by Ljung [Ljung, 1987, section 14.6] differencing has the effect of amplifying the high frequency components of the noise and so should be avoided. However, if our model of the system is amended, so that

\[ \varepsilon(nT;\hat{H}(q^{-1})) = z(nT) - H(q^{-1}) v(nT) - \hat{d}_u \] 5.1.58

\[ = \left[ H(q^{-1}) - \hat{H}(q^{-1}) \right] v(nT) + \left[ d_u - \hat{d}_u \right] + \bar{e}(nT) \] 5.1.59

then \( \hat{H}(q^{-1}) \) will converge to \( H(q^{-1}) \). In fact, estimating \( d_u \) is equivalent to calculating the mean level of the web profile at each measuring point and then subtracting these mean values from the measurements, \( z(nT) \) [Ljung, 1987, section 14.6]. As a by-product, the uncontrollable component of the steady error is also estimated. It will be seen below, that it is easy to incorporate the estimation of \( d_u \) into a practical estimation scheme.

The important point that comes out of this analysis is that, given an estimate of \( H(q^{-1}) \), then we can obtain an estimate of the impulse response matrix of the actuator array, \( G \), from the first term of \( \hat{H}(q^{-1}) \) because \( h_1 \) is equal to \( G \) scaled by \( g_1 \) (see equation 5.1.43). Even if the dynamics are unknown, \( g_1 \) is only a scaling factor, so the shape and the positions of the impulse responses in the \( G \) matrix will be unaffected by any uncertainties in \( g_1 \).

\( H(q^{-1}) \) describes the closed loop response between the exogenous input \( v(nT) \) and the
observed output $z(nT)$. If an exogenous input is applied at $t=nT$, then its effect will first be seen at the measuring position at time $t=(n+1)T$ (because the plant inherently contains a delay) but because of the closed loop feedback law, there will also be components in the measured output at future time samples $(i.e. t=(n+2)T, (n+3)T,...)$ which are a result of the exogenous input at $t=nT$. The size and shape of these effects at the output are determined by $H(q^{-1})$ and the order of this polynomial is determined by the length of time that the effect of $v(nT)$ is projected around the feedback loop which is the settling time of the closed loop system. Although we are estimating the closed loop response, we require an estimate of the open loop response. However, if an exogenous input is applied at $t=nT$, the component of the measured profile at $t=(n+1)T$ which is a result of this input is determined solely by the open loop response of the system. This is because the feedback controller has not yet "seen" the effect of the input. This allows us to extract the open loop response from the first term of the closed loop response.

So far we have not made use of the observation that it is not necessary to estimate every element in the $G$ matrix, because $\hat{G}$ can be generated from estimates of two of its columns (by using the assumption that the impulse responses of all of the actuators are identical and equally spaced). This simplification can be accommodated into the analysis by applying the exogenous input to only two of the actuators. As an illustration of this idea, consider the case where the exogenous input is applied to just the $j$th actuator, so that

$$v^T(nT) = [0, 0,...v_j(nT),...,0] \quad 5.1.60$$

where $v_j(nT) \in \mathbb{R}$. Equation 5.1.56 which describes the response of the closed loop system, can now be written as

$$z(nT) = [H(q^{-1})]_j v_j(nT) + \bar{e}_j(nT) + d_u \quad 5.1.61$$

where $[H(q^{-1})]_j$ is the $j$th column of $H(q^{-1})$. Because $v_j(nT)$ is a scalar, we can write down an individual expression for each of the components of $z(nT)$.

$$z_i(nT) = [H(q^{-1})]_{ij} v_j(nT) + \bar{e}_i(nT) + d_{ui} \quad 5.1.62$$

where $[H(q^{-1})]_{ij}$ is the $ij$th element of $[H(q^{-1})]$ and $z_i(nT)$, $\bar{e}_i(nT)$ and $d_{ui}$ are the $i$th components of $z(nT)$, $\bar{e}(nT)$ and $d_u$ respectively.
From equations 5.1.42 to 5.1.45, we know that

\[ H(q^{-1}) = g_1 G q^{-1} + H_2 q^{-2} + H_3 q^{-3} + \ldots + H_p q^{-p} \]  

5.1.63

where \( g_1 \in \mathbb{R} \) and \( G, H_2, H_3, \ldots \in \mathbb{R}^{m \times (2N+1)} \) and \( p \) is the number of samples required for the closed loop to settle. This allows us to write equation 5.1.62 as

\[
\begin{align*}
   z_i(nT) &= \left[ g_1 [G]_{ij} q^{-1} + [H_2]_{ij} q^{-2} + [H_3]_{ij} q^{-3} + \ldots + [H_p]_{ij} q^{-p} \right] v_j(nT) + \\
   &+ \varepsilon_i(nT) + d_{ui}
\end{align*}
\]  

5.1.64

If can estimate the scalar transfer function between \( v_j(nT) \) and \( z_i(nT) \), then the first coefficient of this transfer function is an estimate of \( g_1[G]_{ij} \). Provided that \( g_1 \) is known (either from a priori knowledge of the actuator dynamics or from a separate identification experiment) this gives us an estimate of the \( ij \)th element of the impulse response matrix, \( G \). If we repeat the estimation procedure for each of the measuring points across the web (ie at \( z_i(nT) \) for \( i=\{1,\ldots,m\} \)) then we can build up an estimate of the \( j \)th column of \( G \) from the individual estimates of \( [G]_{ij} \) for \( i=\{1,\ldots,m\} \).

We have now reduced the problem of estimating a column of \( G \) to a set of \( m \) SISO estimators which run concurrently. In fact, because it is unlikely that the open loop response of a single actuator will cover the whole of the web, it is necessary to look only at the measuring points where \( [G]_{ij} \) is expected to be non-zero. This reduces the size of the estimation problem even further. Note that even though the open loop response of a single actuator does not affect the whole of the sheet, the closed loop response will probably spread the response over a wider portion of the web. This means that the columns of the higher order terms in the \( H(q^{-1}) \) polynomial (ie \( H_2, H_3, \) etc.) will be non-zero over a wider range of measurement points than the columns of \( G \). It does not matter that we do not estimate all of the non-zero points of \( H_2, H_3, \) etc. because we do not need them if we only require an estimate of a single column of the open loop response, \( G \).

Given an estimate of the \( j \)th column of \( G \), this gives us the shape of the impulse response of the \( j \)th actuator. Because we have assumed that shape of the responses of all of the
actuators are the same, we can now deduce the "shape" of all of the other columns of $G$. However, we do not know the separation between the actuators so we cannot deduce the position of the centres of the responses. As pointed out above, this problem can be solved if we have estimates of two columns of $G$.

We can generate estimates of two columns of $G$ by applying the same perturbation to two widely separated actuators. So if $v(nT)$ has the form

$$v^T = [0, \ldots, 0, v_{j1}(nT), 0, \ldots, 0, v_{j2}(nT), 0, \ldots, 0]$$  \hspace{1cm} 5.1.65

where $v_{j1}(nT), v_{j2}(nT) \in \mathbb{R}$ and $v_{j1}(nT) = v_{j2}(nT)$, then the response of the actuator array to this perturbation can be split into two parts

$$z(nT) = [H(q^{-1})]_{j1} v_{j1}(nT) + [H(q^{-1})]_{j2} v_{j2}(nT) + e(nT) + d_u$$  \hspace{1cm} 5.1.66

$$= [g_1 [G]_{j1} q^{-1} + [H_2]_{j1} q^{-2} + \ldots + [H_p]_{j1} q^{-p}] v_{j1}(nT) + [g_1 [G]_{j2} q^{-1} + [H_2]_{j2} q^{-2} + \ldots + [H_p]_{j2} q^{-p}] v_{j2}(nT) + e(nT) + d_u$$  \hspace{1cm} 5.1.67

Provided that the open loop responses of the $j1$th and $j2$th actuators do not overlap, then the response can be split into two. The response at the measuring points in the region of the $j1$th actuator is given by

$$z_{1}(nT) = [g_1 [G]_{ij1} q^{-1} + [H_2]_{ij1} q^{-2} + \ldots + [H_p]_{ij1} q^{-p}] v_{j1}(nT) + e_{1}(nT) + d_{ui}$$  \hspace{1cm} 5.1.68

and in the region of the $j2$th actuator, by

$$z_{2}(nT) = [g_2 [G]_{ij2} q^{-1} + [H_2]_{ij2} q^{-2} + \ldots + [H_p]_{ij2} q^{-p}] v_{j2}(nT) + e_{2}(nT) + d_{ui}$$  \hspace{1cm} 5.1.69

This means that if we can estimate the closed loop response between $v_{j1}(nT)$ and $z_{1}(nT)$ around the position of the $j1$th actuator, then the first term of this response gives us an estimate of the non-zero elements of the $j1$th column of $G$. Similarly, estimating the closed loop response in the region around the $j2$th actuator leads to an estimate of the non-zero elements of the $j2$th column of $G$. Now that we have the shape and position of the impulse responses of two actuators, we can deduce all of the elements of the $G$ matrix (because we have assumed that the shape of the responses of all of the actuators are identical and that
they are equally spaced).

This analysis for estimating the response of two actuators simultaneously requires that there is no overlap between the open loop responses of the two actuators (this criterion can also be thought of as requiring that there is no overlap between the non-zero elements of the corresponding columns of G). If the responses do overlap, for example at the measuring point $z_i(nT)$, then it is impossible to separate the response at $z_i(nT)$ which is due to the j1th actuator from the response due to the j2th actuator because they are both driven with the same perturbation. For most practical systems, the actuators have responses with finite widths which are small relative to the width of the web, so it is always possible to choose two widely separated actuators with non-overlapping responses. Also, choosing widely separated actuators reduces the relative size of any errors in the estimated value of the distance between actuators. If the overlap is unavoidable then the two actuators can be driven with two independent, mutually uncorrelated perturbation sequences, so that the estimators for each of the actuators do not interact. However, for the rest of this study we will assume this approach is unnecessary so that the same perturbation will be applied to each actuator.

Although it is possible to choose two actuators whose open loop responses do not overlap, it is likely that their closed loop responses will overlap, because the controller causes the effect of a change in one actuator to be spread across the whole web. However, because the effect of the controller first appears in the $q^{-2}$ term of the closed loop response via the return difference matrix (see equation 5.1.44) this closed loop interaction between actuators will not affect the open loop response in the $q^{-1}$ term. The effect of the overlap is to introduce bias in the higher order terms of $H(q^{-1})$ polynomial.

We have now shown that it is possible to reduce the problem of estimating all of the elements of the $mx(2N+1)$ real matrix, G, to a set of SISO problems which estimate the elements of two (non-overlapping) columns of G. Using our assumptions about the
actuators, it is possible to deduce estimates of the other elements of $G$ from the estimates of these two columns. All that remains to be done is to show that the SISO estimators will converge to their true values.

For ease of notation, we will consider the convergence of a single estimator which looks at the closed loop response between an exogenous input sequence applied to the $j$th actuator and the $i$th measuring point. The analysis is easily extended to cover all of the SISO estimators. From equation 5.1.64, we know that the response at $z_i(nT)$ is given by

$$z_i(nT) = [ g_1 [G]_{ij} q^{-1} + [H_2]_{ij} q^{-2} + \ldots + [H_p]_{ij} q^{-p} ] v_j(nT) + \bar{e}_i(nT) + d_{ui}$$  

$$= h_{ij}(q^{-1}) v_j(nT) + \bar{e}_i(nT) + d_{ui}$$  

where $p$ is the order of the closed loop polynomial and $h_{ij}(q^{-1})$ is a scalar polynomial given by

$$h_{ij}(q^{-1}) = g_1 [G]_{ij} q^{-1} + [H_2]_{ij} q^{-2} + \ldots + [H_p]_{ij} q^{-p}$$  

As in the multivariable case, we can estimate $h_{ij}(q^{-1})$ and $d_{ui}$ by choosing the values which minimise, in a least squares sense, the (scalar) estimation error

$$\varepsilon_i(nT; h_{ij}(q^{-1}), d_{ui}) = z_i(nT) - h_{ij}(q^{-1}) v_j(nT) - d_{ui}$$  

Substituting for $z_i(nT)$ from equation 5.1.71 gives

$$\varepsilon_i(nT; \hat{h}_{ij}(q^{-1}), \hat{d}_{ui}) = [ h_{ij}(q^{-1}) - \hat{h}_{ij}(q^{-1}) ] v_j(nT) + [ d_{ui} - \hat{d}_{ui} ] + \bar{e}_i(nT)$$  

Provided that [Soderstrom et al, 1976]

i) the order of the $\hat{h}_{ij}(q^{-1})$ polynomial is greater than or equal to the the true order of the closed loop response, $h_{ij}(q^{-1})$;

ii) the exogenous input sequence, $v_j(nT)$, is independent of $\bar{e}_i(nT)$;

then the values of $\hat{h}_{ij}(q^{-1})$ and $\hat{d}_{ui}$ which minimise $\varepsilon(nT; \hat{h}_{ij}(q^{-1}), \hat{d}_{ui})$ for all $n=\{1,2,\ldots\}$ are those which make

220
\[
e_i(nT; \hat{h}_{ij}(q^{-1}), \hat{d}_{ui}) = \bar{e}_i(nT) \quad \text{for } n=\{1,2,...\}
\]

To satisfy this criterion at all time steps requires that \( \hat{h}_{ij}(q^{-1}) = h_{ij}(q^{-1}) \) and \( \hat{d}_{ui} = d_{ui} \), which means that the estimate of the closed loop response will converge to its correct value so we can obtain a valid estimate of \([G]_{ij}\) from the first term of \(\hat{h}_{ij}(q^{-1})\).

The two conditions for convergence are easily satisfied. The noise sequence, \(\bar{e}_i(nT)\), is unlikely to be white, even if the underlying process noise, \(e_i(nT)\), is white, because of the colouring introduced by the feedback loop (recall that \(e(nT)\) is given by

\[
\bar{e}(nT) = \left[ I + g(q^{-1}) k(q^{-1}) G K \right] e(nT)
\]

However, as long as the exogenous input sequence is chosen independently of the plant noise (for example if \(v_j(nT)\) is a pseudo random binary sequence [Ljung, 1987, section 14.3]) then there will be no correlation between \(v_j(nT)\) and \(e_i(nT)\) irrespective of whether \(\bar{e}(nT)\) is coloured or not.

Strictly speaking, the estimation procedure does require that the noise sequence has a mean of zero. If \(e(nT)\) (and hence \(\bar{e}(nT)\)) does not have zero mean, then the finite value of the mean can be considered as part of the uncontrollable disturbance, \(d_u\). This means that the estimate of \(d_u\) will be biased (by the value of the mean of the noise) but the estimate of \(h_{ij}(q^{-1})\) will not be affected. This implies that a true estimate of \(G\) can be obtained, even if the underlying noise process is coloured and has a non-zero mean.

The restriction on the order of the polynomial \(\hat{h}_{ij}(q^{-1})\) can be satisfied by making the order arbitrarily large. However, it is well known that increasing the order of the model above the order of the true plant polynomial, \(h_{ij}(q^{-1})\), increases the variance of the estimates of the coefficients in \(\hat{h}_{ij}(q^{-1})\) [Ljung and Soderstrom, 1983]. So it is sensible to obtain an idea of the length of the true \(h_{ij}(q^{-1})\) polynomial from the settling time of the closed loop system.

The number of samples that are required before the closed loop settles following a transient, determines the order of \(h_{ij}(q^{-1})\). Measuring this settling time on the plant gives a rough estimate of the order of \(\hat{h}_{ij}(q^{-1})\). The increase in the variance of the estimates as a
result of using a model order which is too large can be reduced by careful choice of the input sequence. This will be discussed in more detail in section 5.3

We now have an indirect approach for estimating the elements of the G matrix. The advantages of this method over the direct method which was described earlier, are

i) The estimation of the mx(2N+1) real matrix, G, has been reduced to a set of SISO estimators, with one estimator for each of the non-zero elements of two columns of G.

ii) The estimators give the correct value in the presence of both controllable and uncontrollable steady state disturbances.

iii) The estimators give correct values in the presence of non-zero mean, coloured noise.

The major price that is paid for these simplifications is that the estimation procedure is no longer non-invasive, because it is necessary to introduce an exogenous input to two actuators. In practice, it is always possible to reduce the amplitude of these inputs so that the size of their effects on the web are less than the amplitude of the process noise. However, as will be seen in the next section, reducing the amplitude of the perturbation, increases the variance of the estimates of the elements of G, so choosing the amplitude of the exogenous inputs has to be a compromise between obtaining accurate estimates while not disrupting the profile of the web.

There are also some severe restrictions imposed upon the estimation procedure by the requirements that were introduced during the analysis. The three major requirements are:

(i) the closed loop response must be stable (so that the in the expansion of the return difference matrix polynomial in equation 5.1.27, the size of the coefficients decrease with increasing order of q⁻¹);

(ii) the open loop response of the plant must be stable (so that the plant dynamics,
(i) \( g(q^{-1}) \) can also be expanded as a polynomial with decreasing coefficients as the order of \( q^{-1} \) increases;

(ii) the system should be operating in steady state (so that the controllable part of the steady state disturbance is removed from the profile signal).

In practical systems, the dynamics of the actuators are rarely so complicated that the uncontrolled plant is open loop unstable, so requirement (ii) is met. The requirement on the steady state operation of the plant is to ensure that the disturbance seen in the profile signal remains fixed. During a startup of the machine or a change in the operating conditions which produces a change in the cross directional disturbance, the controller will take a number of steps to remove the controllable part of the disturbance and in this transient period, the disturbance will consist of a controllable component, \( d_c \), as well as the uncontrollable component, \( d_u \). Because \( d_c \) is changing with time as it is controlled, the estimator, which is expecting a constant disturbance component, will introduce bias into the other terms, including the estimate of the \( G \) matrix, until the controllable component is removed. Although section 4.5 showed that, provided the system remained stable, the controllable component of the disturbance would eventually be removed, if the controller is badly tuned or if there is an error between the response of the true plant and the nominal response which was used as the basis of the control design, then the transient period is likely to be long, so it will take some time before unbiased estimates of \( G \) are available.

If requirement (i) is violated, then closed loop polynomial \( H(q^{-1}) \) cannot be truncated because the settling time is infinite and all of the terms in the estimator will be biased. This means that the plant has to be started up with some initial, a priori estimate of the \( G \) matrix which is sufficiently accurate to allow a controller to be designed which will maintain stability. Once the transient response has died away (so that the controllable component of the disturbance has been removed) and an estimate of the \( G \) matrix has been obtained, then this can be used to refine the design of the controller. The need for an a priori estimate and the effects of the transients mean that the estimator does not perform well during startups and changes to the process and it is at these times that accurate estimates are most valuable.
Current research is aimed at improving this area of the performance of the estimator.

Finally, there are some practical points that need to be considered before the estimation procedure can be implemented. Firstly, as mentioned earlier, it is necessary to know the value of the first term of the actuator dynamics (the $g_1$ term in equation 5.1.43) in order to extract an estimate of $G$ from an estimate of the closed loop response. This can be obtained from a priori knowledge of the actuator dynamics or from a separate estimation procedure. As pointed out above, even if $g_1$ is not known exactly, we can see from equation 5.1.43 that any uncertainty in $g_1$ simply scales the estimate of $G$, so it will not affect the shape or position of the estimates of the impulse responses of the actuators. Effectively, this can be considered as uncertainty in the gain of the matrix $G$. If the estimated $G$ is used to design a controller for the system (as in section 4.3) then we know from the discussion of robustness in section 4.5, that errors in the gain need not degrade the performance of the controller too severely. Provided that the gain margin of the controller dynamics, $k(q^{-1})$ is large enough so that the system remains stable in the presence of uncertainty in the gain, then the analysis of section 4.5 showed that the controllable component of the steady state profile disturbance will still be removed from the web by the action of the control system, although the transient response may be degraded.

We also require some idea of the width of the response of an actuator. This allows us to define the number of SISO estimators that need to be set up in order to estimate the impulse response. In addition, we need to know the width of the response to ensure that the effects of the actuators, to which the exogenous inputs are being applied, do not overlap. In practice, a rough estimate of the width of the response can usually be obtained from an understanding of the physical behaviour of the actuators.

As mentioned in section 4.1, the measured profile will have been put through a low pass filter to remove any cross directional spatial frequencies above $\omega_c$ (which is the limit of the frequency response of a nominal actuator). The filtering has the advantage of improving the
signal to noise ratio of the measured profile values, \( z_i(nT) \). Unfortunately, without previously estimating the response of an actuator, it is impossible to obtain an accurate value of \( \omega_c \). This means that the filter may be removing frequency components from the profile measurements which are part of the true response. The estimated response will therefore be distorted because the estimate does not include all of its frequency components. In practice, an upper bound on the true value of \( \omega_c \) can usually be obtained (as before, either from an understanding of the physical system or from offline measurements) which can be used as the cut off frequency of the low pass filter. In addition, practical low pass filters do not have a sharp frequency cut off, so frequency components above \( \omega_c \) will be attenuated rather than removed completely. This reduces the distortion of the estimate of the actuator response.

5.2 The Implementation of a Recursive Estimator

The previous section showed how the estimation of the impulse response matrix, \( G \), can be reduced to a set of SISO estimators, each of which minimises (in a least squares sense) the prediction error given by

\[
\varepsilon_i(nT; \hat{h}_{ij}(q^{-1}) \hat{d}_{ui}) = z_i(nT) - \hat{h}_{ij}(nT) v_j(nT) - \hat{d}_{ui}
\]

5.2.1

where \( \hat{h}_{ij}(q^{-1}) \) is an estimate of \( h_{ij}(q^{-1}) \) with

\[
h_{ij}(q^{-1}) = g_1 [G]_{ij} q^{-1} + [H_2]_{ij} q^{-2} + \ldots + [H_p]_{ij} q^{-p}
\]

5.2.2

\( p \) being the order of the closed loop polynomial. The notation can be reduced by defining a parameter vector, \( \theta_{ij} \), as

\[
\theta_{ij}^T = \left[ g_1 [G]_{ij}, [H_2]_{ij}, \ldots, [H_p]_{ij}, d_{ui} \right]
\]

5.2.3

and a state vector, \( \phi(nT) \) as

\[
\phi^T(nT) = \left[ v_j([(n-1)T], v_j([(n-2)T], \ldots, v_j([(n-p)T], 1 \right]
\]

5.2.4

Note that \( \phi(nT) \) contains the past perturbations and the constant, 1. If the same perturbation is applied to both of the actuators whose responses are being estimated (as was assumed in
the previous section) then the state vector is the same for all of the estimators.

Equation 5.2.1 can now be simplified to

\[ \epsilon_i(nT;\hat{\theta}_{ij}(nT)) = z_i(nT) - \hat{\theta}_{ij}^T(nT) \phi(nT) \]  

5.2.5

Given a set profile measurements and perturbations for the time periods \( t=1T, \ldots, n \), we require the value of \( \theta_{ij}(nT) \) which minimises the least squares criterion

\[ \hat{\theta}_{ij}(nT) = \arg \min_{\theta_{ij}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2(lT;\theta_{ij}) \]  

5.2.6

Because this is a quadratic criterion, the optimal \( \hat{\theta}_{ij}(nT) \) can be evaluated analytically as [Ljung and Soderstrom, 1983]

\[ \hat{\theta}_{ij}(nT) = R(nT)^{-1} \sum_{l=1}^{n} \phi(lT) z_i(lT) \]  

5.2.7

where

\[ R(nT) = \sum_{l=1}^{n} \phi(lT) \phi(lT)^T \]  

5.2.8

Provided that the inverse of \( R(nT) \) exists and that \( \phi(nT) \) is independent of the noise in the process, \( \epsilon_i(lT) \) for \( l=1,2,\ldots, n \), then \( \hat{\theta}_{ij}(nT) \) gives an unbiased estimate of the true parameter vector, \( \theta_{ij} \). Also, if the order of the estimated model matches the true order of the closed loop system, then the expected value of \( \hat{\theta}_{ij}(nT) \) coincides with the true parameters in \( \theta_{ij} \) [Davis and Vinter, 1985] [Ljung, 1987].

This estimation procedure is an example of a least squares estimator. It is easily shown [Ljung and Soderstrom, 1983] that \( \hat{\theta}_{ij}(nT) \) can be computed recursively as

\[ \hat{\theta}_{ij}(nT) = \hat{\theta}_{ij}[(n-1)T] + L(nT) \left[ z_i(nT) - \hat{\theta}_{ij}^T[(n-1)T] \phi(nT) \right] \]  

5.2.9

\[ L(nT) = \frac{P[(n-1)T] \phi(nT)}{1 + \phi(nT)^T P[(n-1)T] \phi(nT)} \]  

5.2.10

\[ P(nT) = P[(n-1)T] - \frac{P[(n-1)T] \phi(nT) \phi(nT)^T P[(n-1)T]}{1 + \phi(nT)^T P[(n-1)T] \phi(nT)} \]  

5.2.11
The matrix \( P(nT) \in \mathbb{R}^{(p+1) \times (p+1)} \) is given by
\[
P(nT) = R(nT)^{-1} \tag{5.2.12}
\]
\[
= \left[ \sum_{i=1}^{n} \phi(IT) \phi(IT)^T \right]^{-1} \tag{5.2.13}
\]
and equation 5.2.11 simply provides a recursive update for \( P(nT) \) which uses the matrix inversion lemma to avoid inverting the \((p+1)\times(p+1)\) matrix in 5.2.13 at every step.

Usually, we wish to obtain estimates of \( \hat{\theta}_{ij}(nT) \) at every time step, i.e., at \( t=nT \) for \( n=\{1,2,\ldots\} \). In order to start the recursive estimator, we require values for \( \hat{\theta}_{ij}(0) \) and \( P(0) \). Given these values, then it is possible to show [Ljung and Soderstrom, 1983, section 2.2] that
\[
\hat{\theta}_{ij}(nT) = \left[ P(0)^{-1} + \sum_{i=1}^{n} \phi(IT) \phi(IT)^T \right]^{-1} \left[ P(0)^{-1} \hat{\theta}_{ij}(0) + \sum_{i=1}^{n} \phi(IT) z_{i}(IT) \right] \tag{5.2.14}
\]
If \( P(0) \) is chosen so that \( P(0)^{-1} \) is small relative to
\[
\sum_{i=1}^{n} \phi(IT) \phi(IT)^T \tag{5.2.15}
\]
and \( \hat{\theta}_{ij}(0) \) is chosen as the zero vector, then by comparing equations 5.2.7 and 5.2.14 it can be seen that the effect of the initial conditions is negligible and the recursive estimate approaches the offline estimate in 5.2.7. It is usual to start the recursion with \( \hat{\theta}_{ij}(0) = 0 \) and \( P(0) = cI_{(p+1)} \) where \( c \) is a constant which is much larger than the elements of \( R(T) \).

In practice, it is not usual to implement the \( P(nT) \) update as in equation 5.2.10, because the expression can be very sensitive to any numerical errors (e.g., rounding). The problem can be seen by noting that
\[
[ P(nT) ]^{-1} = \sum_{i=1}^{n} \phi(IT) \phi(IT)^T \tag{5.2.16}
\]
is a sum of symmetric outer products and therefore must be positive definite. As more samples are read and \( n \) becomes large, the expression tends to become large. This means
that the inverse, \( P(nT) \) becomes small while remaining positive definite, so \( P(nT) \) is very susceptible to any numerical errors which make \( P(nT) \) indefinite. This can lead to the estimation procedure becoming unstable [Bierman, 1977]. One way around this problem is to factorise \( P(nT) \) as

\[
P(nT) = U(nT) D(nT) U(nT)^T
\]

where \( U(nT) \) is an upper triangular matrix with unit diagonals and \( D(nT) \) is a diagonal matrix. If the recursion is applied to \( U(nT) \) and \( D(nT) \), rather than to \( P(nT) \), the numerical problems are reduced [Bierman, 1977] [Ljung and Soderstrom, 1983].

Using this factorisation, equation 5.2.11 becomes

\[
P(nT) = U[(n-1)T] D[(n-1)T] U[(n-1)T]^T - \frac{U[(n-1)T] \psi(nT) \psi(nT)^T U[(n-1)T]}{\beta(nT)}
\]

\[
= U[(n-1)T] \left[ D[(n-1)T] - \frac{\psi(nT) \psi(nT)^T}{\beta(nT)} \right] U[(n-1)T]^T
\]

where

\[
\phi(nT) = U[(n-1)T]^T \phi(nT)
\]

\[
\psi(nT) = D[(n-1)T] \phi(nT)
\]

\[
\beta(nT) = 1 + \phi(nT)^T \psi(nT)
\]

If the middle term in equation 5.2.17 can be factorised as

\[
\bar{U}(nT) \bar{D}(nT) \bar{U}(nT)^T = D[(n-1)T] - \frac{\psi(nT) \psi(nT)^T}{\beta(nT)}
\]

Combining equations 5.2.21 and 5.2.17, \( P(nT) \) can be expressed as

\[
P(nT) = U[(n-1)T] \bar{U}(nT) \bar{D}(nT) \bar{U}(nT)^T U[(n-1)T]^T
\]

so the update can be reduced to

\[
U(nT) = U[(n-1)T] \bar{U}(nT)
\]

\[
D(nT) = \bar{D}(nT)
\]

An extremely elegant algorithm for implementing this procedure is described in [Bierman,
Another potential difficulty with the estimator is that it is possible for the matrix \( R(nT) = [P(nT)]^{-1} \) (see equation 5.2.8) to become singular, so that its inverse becomes singular and the estimate of \( \hat{\Theta}_{ij}(nT) \) cannot be evaluated (see equation 5.2.7). For the least squares problem being considered here this is unlikely to be a problem. This is because the vector \( \phi(nT) \) contains only past perturbations \( v_j(nT), v_j[(l-1)T], \ldots, v_j[(l-p)T] \) and the constant 1, so the \( R(nT) \) matrix (which depends solely on \( \phi(nT) \) for \( l=\{1,2,\ldots,n\} \)) can only become singular when the perturbations are not "persistently exciting" (see [Ljung and Soderstrom, 1983] and section 5.3 of this report). In principle, it is always possible to choose a set of perturbations which avoids this problem, but for general inputs, it is worth applying a simple check to \( P(nT) \) to prevent its occurrence. This check is known as regularisation [Ljung and Soderstrom, 1983, section 6.5].

We are trying to prevent \( R(nT) \) from becoming singular. If \( P(nT) = R(nT)^{-1} \) is factorised as \( U(nT)D(nT)U(nT)^T \), then \( R(nT) \) can be written as

\[
R(nT) = P(nT)^{-1} = U(nT)^T D(nT)^{-1} U(nT)^{-1}
\]

Because \( U(nT) \) is upper triangular with unit diagonals, \( \det[U(nT)] = 1 \), so \( U(nT)^{-1} \) and \( U(nT)^T \) are both non-singular. This means that \( R(nT) \) can only be singular if \( D(nT)^{-1} \) has zero elements on its diagonal (recall that \( D(nT) \) is defined to be diagonal). We can prevent singularity by ensuring that the diagonal elements of \( D(nT) \) do not become too large (so the corresponding elements of \( D(nT)^{-1} \) do not become too small). This can be done by comparing each of the elements of \( D(nT) \) with a prechosen large value and replacing any that exceed this value with the value itself.

Although it is possible to run recursive estimators without taking precautions against numerical errors and matrix singularities, if an estimator has to be left running on the plant...
for long periods, it is advisable to take steps to avoid these problems, so it is sensible to use the techniques outlined above.

So far, we have assumed that the parameter vector remains constant. In practice, it is likely that the plant will change (e.g., as the actuators age) which produce a change in the true \( \theta_{ij} \) vector. If these changes are occurring slowly, then one method of tracking them is to introduce a forgetting factor, \( \lambda \) [Ljung and Soderstom, 1983, section 2.6]. The minimisation criterion in equation 5.2.6 can be amened to

\[
\hat{\theta}_{ij}(nT) = \arg\min_{\theta_{ij}} \sum_{i=1}^{n} \lambda^{(n-i)} e_i(\lambda; \theta_{ij})^2
\]

where \( \lambda < 1 \). Essentially, the forgetting factor applies more weight to the most recent observations and reduces the influence of earlier observations.

The forgetting factor can be incorporated into the recursive update by making

\[
L(nT) = \frac{P[(n-1)T] \phi(nT)}{\lambda + \phi(nT)^{T} P[(n-1)T] \phi(nT)}
\]

\[
P(nT) = \frac{1}{\lambda} \left[ P[(n-1)T] - \frac{P[(n-1)T] \phi(nT) \phi(nT)^{T} P[(n-1)T]}{\lambda + \phi(nT)^{T} P[(n-1)T] \phi(nT)} \right]
\]

(The \( U(nT) D(nT) U(nT)^{T} \) factorised form can also be readily amended to accommodate a forgetting factor, [Ljung and Soderstom, 1983, section 6.2]). A consequence of the forgetting factor is to increase the size of \( P(nT) \) and to prevent it from approaching zero. This means that \( L(nT) \) is also kept larger so that any changes in the parameter vector are transmitted to the estimates via the \( \hat{\theta}_{ij}(nT) \) update expression (equation 5.2.9) which depends on \( L(nT) \). The price that is paid for preventing \( P(nT) \) from approaching zero is that the variance of the parameter estimates is increased, as will be seen below.

An interesting secondary advantage of keeping \( P(nT) \) large through the use of a forgetting factor, is a reduction in the sensitivity of the estimator to numerical errors, because \( P(nT) \) does not approach zero.
We have now shown how to implement the SISO estimator developed in the previous section for estimating the closed loop response between an exogenous input on one actuator and a single measuring point, with an estimate of the open loop response being obtained from the first term in the estimated parameter vector, $\hat{\theta}_{ij}(nT)$. It has also been shown how to obtain an estimate of the variance of the estimated response. An estimate of the complete G matrix can be built up by repeating the SISO estimation for the measuring points around the position of two actuators. For practical systems, this can result in a large number of SISO estimators, each of which needs updating when a new set of measurements is received. The computational load required to perform these updates can be quite severe, particularly when calculating $P(nT)$ and $L(nT)$ (see equations 5.2.9 and 5.2.10). However, a major simplification can be made by noting that both $P(nT)$ and $L(nT)$ depend upon only $\phi(nT)$ for their updates. From equation 5.2.4, we know that $\phi(nT)$ is a function of only the past exogenous inputs and the constant, 1. Because we have assumed that the same exogenous input is applied to both actuators, this means that the $\phi(nT)$ (and hence $P(nT)$ and $L(nT)$) will be the same for all of the SISO estimators. This allows us to update $P(nT)$ and $L(nT)$ just once and then we can use these values for each of the $\hat{\theta}_{ij}(nT)$ updates. This simplification results in a major reduction in the computational effort required to generate the estimate of the matrix, G.

5.3 The Choice of the Optimal Input Sequence

So far in this discussion, we have not said anything about choosing the exogenous input, $v_j(nT)$ for $n=\{1,2,...\}$. We have, however, placed some conditions on $v_j(nT)$:

i) the input sequence must be uncorrelated with the noise sequence, $\bar{e}_i(nT)$ at all of the measuring points that are used for SISO estimators, so that the parameter estimates are unbiased;

ii) the amplitude of the input sequence must be small enough so that its effect on
the web is not too large;

iii) the input sequence must ensure that the matrix

\[
R(nT) = \sum_{l=1}^{n} \phi(lT) \phi(lT)^T
\]

is invertible (recall that \( \phi(lT) \) is dependent upon \( v_j[(l-1)T], v_j[(l-2)T], \ldots, v_j[(l-p)T] \) as described in equation 5.2.4).

In addition we would like to choose an input sequence which (in some sense) optimises the estimates of the parameter vector, \( \theta_{ij} \).

The first condition can be satisfied by ensuring that the input sequence, \( v_j(nT) \), is chosen without reference to the process noise. Under these circumstances, \( e_i(nT) \) (and hence \( \bar{e}_i(nT) \)) will be independent of \( v_j(nT) \) for \( n=1,2,\ldots \) and for all of the measuring points used in the estimation procedure.

In this section we consider the problem of choosing an input sequence which optimises the parameter estimates. It will be seen that requirements (ii) and (iii) above will be satisfied in the course of solving the optimisation problem. We start the discussion by investigating some of the properties of the estimator that was developed in the previous two sections. To simplify the notation, we introduce

\[
Z_i(nT) = \begin{bmatrix} z_i(T) \\ z_i(2T) \\ \vdots \\ z_i(nT) \end{bmatrix}, \quad X(nT) = \begin{bmatrix} \phi(T)^T \\ \phi(2T)^T \\ \vdots \\ \phi(nT)^T \end{bmatrix}, \quad N_i(nT) = \begin{bmatrix} \bar{e}_i(T) \\ \bar{e}_i(2T) \\ \vdots \\ \bar{e}_i(nT) \end{bmatrix}
\]

This allows us to write

\[
Z_i(nT) = X(nT) \theta_{ij} + N_i(nT)
\]

where \( \theta_{ij} \) is taken to be a vector containing the true parameter values as defined in equation 5.3.2.
5.2.3. Using this notation, we can recast the minimisation problem given in equation 5.2.7 for choosing the optimal estimate of $\theta_{ij}$, as

$$\hat{\theta}_{ij}(nT) = \arg \min_{\theta_{ij}} \frac{1}{n} \left[ (Z_i(nT) - X(nT) \theta_{ij})^T (Z_i(nT) - X(nT) \theta_{ij}) \right]$$  \hspace{1cm} 5.3.3

It is readily shown that the solution to this problem is given by [Goodwin and Payne, 1977]

$$\hat{\theta}_{ij}(nT) = \left[ X(nT)^T X(nT) \right]^{-1} X(nT)^T Z_i(nT)$$  \hspace{1cm} 5.3.4

From equations 5.2.8 and 5.3.1, we can identify $R(nT)$ as

$$R(nT) = X(nT)^T X(nT)$$  \hspace{1cm} 5.3.5

so we can see that the estimate in equation 5.3.4 is the same as the estimate in equation 5.2.7 and also will coincide with the recursive estimate given in equation 5.2.14, assuming that effect of the initial conditions is negligible.

The expected value of the estimate is

$$E[\hat{\theta}_{ij}(nT)] = E[ \left( X(nT)^T X(nT) \right)^{-1} X(nT)^T Z_i(nT) ]$$  \hspace{1cm} 5.3.6

$$= E[ \left( X(nT)^T X(nT) \right)^{-1} X(nT)^T (X(nT) \theta_{ij} + N_i(nT)) ]$$  \hspace{1cm} 5.3.7

$$= \theta_{ij} + \left( X(nT)^T X(nT) \right)^{-1} E[ X(nT) N_i(nT) ]$$  \hspace{1cm} 5.3.8

$$= \theta_{ij}$$  \hspace{1cm} 5.3.9

where we have used the fact that the noise and the sequence of input perturbations are independent, which means that

$$E[ X(nT) N_i(nT) ] = 0$$  \hspace{1cm} 5.3.10

We can see that the estimator is unbiased because the expected value of the estimate coincides with the true value.

The variance of the estimate is given by
\[
\text{var}[\hat{\theta}_{ij}(nT)] = E[(\hat{\theta}_{ij}(nT) - E[\hat{\theta}_{ij}(nT)]) (\hat{\theta}_{ij}(nT) - E[\hat{\theta}_{ij}(nT)])^T] \quad 5.3.11 \\
= E[(\hat{\theta}_{ij}(nT) - \theta_{ij}) (\hat{\theta}_{ij}(nT) - \theta_{ij})^T] \quad 5.3.12
\]

We start by evaluating
\[
\hat{\theta}_{ij}(nT) - \theta_{ij} = (X(nT)^TX(nT))^{-1}X(nT)^TZ_i(nT) - \theta_{ij} \quad 5.3.13 \\
= (X(nT)^TX(nT))^{-1}X(nT)^T\theta_{ij} + (X(nT)^TX(nT))^{-1}X(nT)^TN_i(nT) - \theta_{ij} \quad 5.3.14 \\
= (X(nT)^TX(nT))^{-1}X(nT)^TN_i(nT) \quad 5.3.15
\]

\text{var}[\theta_{ij}(nT)] now becomes
\[
\text{var}[\theta_{ij}(nT)] = (X(nT)^TX(nT))^{-1}X(nT)^TE[N_i(nT)N_i(nT)^T]X(nT) (X(nT)^TX(nT))^{-T} \quad 5.3.16 \\
= (X(nT)^TX(nT))^{-1}X(nT)^T \Xi_iX(nT) (X(nT)^TX(nT))^{-T} \quad 5.3.17
\]

where \(\Xi_i\) is the covariance matrix for the ith coefficient of the noise and is defined as
\[
\Xi_i = E[N_i(nT)N_i(nT)^T] \quad 5.3.18
\]

\(\Xi_i\) will be a symmetric positive definite matrix. If the noise is white, then \(\Xi_i\) is diagonal and can be written as
\[
\Xi_i = \sigma_i^2 I_n \quad 5.3.19
\]

where \(\sigma_i^2\) is the variance of the noise. In this case the variance of the estimate becomes
\[
\text{var}[\theta_{ij}(nT)] = (X(nT)^TX(nT))^{-1}X(nT)^T\sigma_i^2 I_n X(nT) (X(nT)^TX(nT))^{-T} \quad 5.3.20 \\
= \sigma_i^2(X(nT)^TX(nT))^{-1} \quad 5.3.21
\]

Usually the noise variance is unknown, but it is readily shown that an unbiased estimate of \(\sigma_i^2\) can be obtained at time \(t=nT\), from the estimation errors, \(e(nT;\hat{\theta}_{ij}(nT))\) [Davis and Vinter, 1985].
\[ \hat{\sigma}_i^2(nT) = \frac{1}{n-p} \sum_{i=1}^{n} \varepsilon_i(nT; \hat{\theta}_i(nT))^2 \] 5.3.22

where \( p \) is the order of the model of the closed loop response.

Equation 5.2.35 can be rearranged so that it can be computed recursively.

\[ \hat{\sigma}_i^2(nT) = \hat{\sigma}_i^2((n-1)T)^2 + \frac{1}{n-p} \left[ \varepsilon_i(nT; \hat{\theta}_i(nT))^2 - \hat{\sigma}_i^2((n-1)T)^2 \right] \] 5.3.23

Unfortunately, this expression becomes sensitive to rounding errors as \( n \) becomes large, so if a forgetting factor is being used in the estimation of \( \hat{\theta}_ij(nT) \), then it makes sense to incorporate it into the estimate of \( \sigma_i(nT)^2 \) as [Ljung and Soderstrom, 1983, section 3.4]

\[ \hat{\sigma}_i^2(nT) = \lambda \hat{\sigma}_i^2((n-1)T)^2 + (1-\lambda) \varepsilon_i(nT; \hat{\theta}_i(nT))^2 \] 5.3.24

The expression in equation 5.3.21 for the variance of the estimates of the parameters, only applies when the noise is white. However, for the system in this study, we know that the noise in the estimation process is given by (see equation 5.1.22)

\[ \bar{e}_i(nT) = \left[ I + g(q^{-1}) k(q^{-1}) G K \right]^{-1} e_i(nT) \] 5.3.25

so that even if \( e_i(nT) \) is white, \( \bar{e}_i(nT) \) will be coloured by the term \( \left[ I+g(q^{-1}) k(q^{-1}) G K \right]^{-1} \). This means that \( \Xi_i \) will not be diagonal. Although the estimator described above (equation 5.3.4) will give an unbiased estimate of \( \theta_{ij} \), it is not the best estimator because it does not minimise the variance of the estimator. When the noise is coloured, it can be shown that the estimate of \( \theta_{ij} \) with the minimum variance is given by [Goodwin and Payne, 1977, section 2.3]

\[ \hat{\theta}_{ij}(nT) = \left( X(nT)^T \Xi^{-1} X(nT) \right)^{-1} X(nT)^T \Xi^{-1} Z_i(nT) \] 5.3.26

This estimator can be shown to be the best linear unbiased estimator for the system given in equation 5.3.2. It can be viewed as the weighted least squares estimate of \( \theta_{ij} \), which is given by

235
\[
\hat{\theta}_{ij}(nT) = \arg\min_{\theta_{ij}} \left[ (Z_i(nT) - X(nT) \theta_{ij})^T \Xi^{-1} (Z_i(nT) - X(nT) \theta_{ij}) \right]
\]

5.3.27

where the inverse of the noise covariance is used as the weighting factor. Goodwin and Payne [Goodwin and Payne, 1977] point out that this approach can also be thought of as pre-whitening the data. If we define a matrix, \( T \), as

\[
T^T = \Xi^{-1}
\]

5.3.28

then by transforming the data using

\[
\tilde{Z}_i(nT) = T^{-1} Z_i(nT)
\]

5.3.29

and applying the ordinary least squares procedure on the transformed data, the estimate is identical to that produced the weighted least squares approach in equation 5.3.27. This transformation of the data has the effect of decoupling the whitening the noise.

Unfortunately, to apply the weighted least squares procedure requires that the covariance matrix, \( \Xi \) and in most practical cases, \( \Xi \) is not known a priori. It is possible to estimate the matrix \( [1 + g(q^{-1}) k(q^{-1}) G K]^{-1} \) using prediction error methods [Ljung, 1988] but these estimation procedures are non-linear and cannot accommodate the simple structure developed in the previous section, which reduces the multivariable estimation problem to a set SISO estimators. For this reason, we will continue to consider the unweighted least squares approach. It should be stressed that the unweighted least squares estimator will give unbiased estimates of the parameters, the only difference between the two approaches is that the variance of the estimates from the ordinary least squares will be at best equal to and, in general, greater than the variance of the estimates from the efficient weighted least squares approach.

We now wish to investigate the choice of input perturbations, \( v_j(T) \), \( v_j(2T) \), ... , which will optimise the estimates. The literature on the choice of optimal input signals is extensive (see for example [Mehra, 1981] and [Goodwin and Payne, 1977]). The basic aim is to choose an input sequence which maximises the information about the parameters that can be obtained by the estimator. It is well known that for any estimator, the variance of the
Parameter estimates obtained satisfies [Davies and Vinter, 1985]

\[ \text{var}[\hat{\theta}_i] \geq M^{-1} \]  

5.3.30

where \( M \) is Fisher's information matrix (see [Godwin and Payne, 1977]), which for the estimator in this study, is given by

\[ M = \mathbb{E}_{z|\theta} \left[ \left( \frac{\partial \log p(Z|\theta)}{\partial \theta} \right) \left( \frac{\partial \log p(Z|\theta)}{\partial \theta} \right)^T \right] \]  

5.3.31

with \( Z \) being a random variable of which the measurements, \( z_i(T), z_i(2T), \ldots \), are observations. The matrix, \( M^{-1} \), is referred to as the Cramer-Rao lower bound and the variance matrix of the parameter estimates achieves this lower bound (ie there is equality in equation 5.3.30) when

i) the noise process is white or when the colouring of the noise is accommodated by pre-whitening the data or by using a weighted least squares approach as in equation 5.3.26. This assumes that the covariance of the coloured noise is known;

ii) the noise has specific distribution, typically a Normal distribution.

If these two conditions can be satisfied, then it can be seen that by minimising (in some sense) \( M^{-1} \) will minimise the variance of the estimates, which can be viewed as ensuring that the estimator is maximizing the information about the parameters.

Unfortunately, as mentioned earlier, for the system considered in this study, the noise is coloured by the term \( [I + g(q^{-1}) k(q^{-1}) GK]^{-1} \) and in practical systems, we cannot guarantee that the noise will have a Gaussian distribution, so the variance of the parameter estimates will not attain the Cramer-Rao lower bound. However, this lower bound represents the best that is achievable, so minimising \( M^{-1} \) indicates that the estimator is being as efficient as possible. For this analysis, we will assume that the noise sequence, \( \bar{e}_i(nT) \), is white with a Normal distribution with zero mean and variance, \( \sigma_1^2 \). This is done for two reasons; firstly, the analysis of the white noise case is of interest in its own right and leads to some ideas that are applicable to the coloured noise case (particularly in the area of the robustness of the estimator to model order selection) and secondly because the approach
generated for the white noise case has been found to work well in practice, when the noise is coloured.

For the parameter vector set up in the previous section

\[ \theta_{ij}^T = \left[ g_1, [G]_{ij}, [H_2]_{ij}, \ldots, [H_p]_{ij}, d_{ui} \right] \] 5.3.32

and the associated state vector

\[ \phi^T(nT) = \left[ v_j((n-1)T), v_j((n-2)T), \ldots, v_j((n-p)T), 1 \right] \] 5.3.33

it can be shown [Goodwin and Payne, 1977] that Fisher' information matrix for the ith of the SISO estimators at time t=nT, is given by

\[ M_i(nT) = \frac{1}{\sigma_i^2} \sum_{l=1}^{n} \phi(lT) \phi(lT)^T \] 5.3.34

\[ = \frac{1}{\sigma_i^2} R(nT) \] 5.3.35

where \( \sigma_i^2 \) is the variance of the noise process, \( \bar{e}_i(nT) \). Note that if the Cramer-Rao lower bound is achieved, then

\[ \text{var} [ \theta_{ij}(nT) ] = M_i^{-1}(nT) \] 5.3.36

\[ = \sigma_i^2 R(nT) \] 5.3.37

which agrees with the expression for the variance of the parameters in equation 5.3.21.

Looking at the information matrix more closely, we can substitute for \( \phi(lT) \) to show that

\[ M_i(nT) = \frac{1}{\sigma_i^2} \sum_{l=1}^{n} \begin{bmatrix} v_j((l-1)T)v_j((l-1)T) & \ldots & v_j((l-1)T)v_j((l-p)T) & v_j((l-1)T) \\ v_j((l-p)T)v_j((l-1)T) & \ldots & v_j((l-p)T)v_j((l-p)T) & v_j((l-p)T) \\ \vdots & \vdots & \vdots & \vdots \\ v_j((l-1)T) & \ldots & v_j((l-p)T) & 1 \end{bmatrix} \] 5.3.38

A scaled version of the average information matrix can be defined as
Because scaling $M(nT)$ will not affect the minimisation, for convenience we will work with this average information matrix.

If ergodicity applies (so that we can replace ensemble averages by expectations) then $\bar{M}(nT)$ can be written as

$$
\bar{M}_1(nT) = \frac{\sigma_i^2}{n} M_1(nT)
$$

5.3.39

where $\tau_{ij}(v_j)$ is the correlation coefficient, given by

$$
\tau_{pq}(v_j) = E [ \gamma_j[(1-p)T] \gamma_j[(1-q)T] ]
$$

5.3.41

and $v_j$ is given by

$$
\bar{v}_j = E [ v_j(T) ]
$$

5.3.42

We now require a measure of the size of the $\bar{M}$ matrix, so that we can perform the minimisation. A convenient measure is

$$
J = \log \det [ \bar{M}_1^{-1}(nT) ]
$$

5.3.43

$$
= - \log \det [ \bar{M}_1(nT) ]
$$

5.3.44

It can be shown that $J$ is a measure of the volume, in the parameter space, of the uncertainty in the parameters around the estimated value (see [Mehra, 1981] for more details).

We wish to choose a sequence of perturbations, $v_j(T)$ for $j=1,2,...,n$ which will minimise this measure. From 5.3.44, we can see that $J$ is minimised when $\det[\bar{M}_1(nT)]$ is maximised. From equation 5.3.38, $\det[\bar{M}_1(nT)]$ can be increased simply by increasing the size of the input, $v_j(T)$. However, it is necessary to restrict the amplitude of $v_j(T)$ so that the web profile is not significantly degraded by the exogenous perturbations. If the
amplitude of the input is restricted to $v^*$, then the optimisation problem can be recast as choosing the sequence $v_j(lT)$ for $l=\{1,2,...,n\}$ which minimises

$$ J = -\log \det [\tilde{M}_j(nT)] $$ \hspace{1cm} 5.3.45

subject to

$$ |v_j(nT)| \leq v^* \quad \text{for} \quad l=\{1,2,...,n\} $$ \hspace{1cm} 5.3.46

Goodwin and Payne [Goodwin and Payne, 1977] show that the solution to this problem is to choose an input sequence which makes

$$ M(nT) = \begin{bmatrix} (v^*)^2 & 0 \\ (v^*)^2 & \ddots \\ 0 & \ddots & (v^*)^2 \\ \vdots & \ddots & \ddots & 0 \end{bmatrix} $$ \hspace{1cm} 5.3.47

This is equivalent to making all of the off-diagonal elements equal to zero while maximising the size of the diagonal elements. Comparing the matrices in equations 5.3.40 and 5.3.47, we can see that this requirement are satisfied by making

$$ \overline{v}_j(lT) = 0 $$ \hspace{1cm} 5.3.48

and

$$ \tau_{pq}(v_j) = (v^*)^2 \quad \text{if} \quad p=q $$

$$ = 0 \quad \text{if} \quad p\neq q $$ \hspace{1cm} 5.3.49

For large $n$, this requirements can be achieved by making $v_j(lT)$ a zero mean white noise sequence. For finite $n$, it is usually more convenient to make $v_j(lT)$ a psuedo random binary sequence (PRBS) [Ljung, 1987, section 14.3] where $v_j(lT)$ flips between $+v^*$ and $-v^*$. Under these circumstances, the conditions in 5.3.48 and 5.3.49 will be very nearly satisfied.
Another advantage of choosing an input sequence which minimises \(-\log \det [\overline{M}_f^{-1}(nT)]\) (and hence maximises \(\log \det [\overline{M}_f(nT)]\)) is that it ensures that the matrix \(R(nT)\) is invertible. This means that condition (iii) (given at the beginning of this section) will be satisfied, so that a parameter estimate will exist. The invertibility of this matrix can be seen by combining equations 5.3.34 and 5.3.39 to give
\[
R(nT) = nM_f(nT)
\]
5.3.50

\[
\begin{bmatrix}
  n(v^*)^2 & 0 & \\
  0 & n(v^*)^2 & \\
  & & n(v^*)^2
\end{bmatrix}
\]
5.3.51

Provided that \(|v^*|\) is non zero, this matrix is non-singular.

The analysis above has concluded that the optimal input sequence is a white noise or psuedo random binary sequence. In practice, it is usually easier to implement the latter, because a PRBS can be generated easily on a computer. However, it should be noted that if a forgetting factor is used, we require that the matrix
\[
\sum_{i=1}^{n} \lambda^{(n-i)} \phi(IT) \phi(IT)^T
\]
5.3.52

is minimised by making it diagonal. If \(\lambda\) is not unity, this matrix will not be diagonalised by a white noise sequence or a PRBS. While techniques exist for choosing an input sequence which will make this diagonal over a fixed time interval, if the estimator is to run recursively for a long (but unspecified) time it is better to use a recursive procedure for generating the inputs as described in [Goodwin and Payne, 1977].

The inputs can be derived in conjunction with a recursive least squares estimator. We have identified the information matrix, \(M_f(nT)\), with the \(R(nT)\) matrix for an efficient estimator (see equation 5.3.37) Thus minimising
\[ J = - \log \det [ M_i(nT) ] \] 5.3.53

can be written as
\[
J = \log \det [ R(nT)^{-1} ] \\
= \log \det [ P(nT) ] 
\] 5.3.54 5.3.55

where \( P(nT) \) is as defined in equation 5.2.11. Unlike the previous analysis, the recursive procedure attempts to minimise this criterion in a "one-step ahead" sense (whereas the above analysis minimises \( M_i(nT) \) for \( n=\{1,2,...,n\} \)). At time sample \( t=[(n-1)T] \), given a value of \( P((n-1)T) \), we aim to find the value of the input, \( v_j((n-1)T) \) which will minimise \( \log \det [ P(nT) ] \). From section 5.2, we know that
\[
\phi^T(nT) = [ v_j((n-1)T), v_j((n-2)T), ... , v_j((n-p)T), 1 ]
\] 5.3.56

and
\[
P(nT) = \frac{1}{\lambda} \left[ I - P((n-1)T) \frac{\phi(nT) \phi(nT)^T}{\lambda + \phi(nT)^T P((n-1)T) \phi(nT)} \right] P((n-1)T) 
\] 5.3.57

This means that
\[
\log \det [ P(nT) ] = -\log \lambda + \log \det P((n-1)T) \\
+ \log \det \left[ I - P((n-1)T) \frac{\phi(nT) \phi(nT)^T}{\lambda + \phi(nT)^T P((n-1)T) \phi(nT)} \right] 
\] 5.3.58

At time \( t=[(n-1)T] \) the first two terms on the right hand side of this equation are fixed, so minimising
\[
\log \det \left[ I - P((n-1)T) \frac{\phi(nT) \phi(nT)^T}{\lambda + \phi(nT)^T P((n-1)T) \phi(nT)} \right] \\
= \log \left[ 1 - \frac{\phi(nT)^T P((n-1)T) \phi(nT)}{\lambda + \phi(nT)^T P((n-1)T) \phi(nT)} \right] \\
= \log \left[ \frac{\lambda}{\lambda + \phi(nT)^T P((n-1)T) \phi(nT)} \right] \\
= \log \lambda - \log \left[ \lambda + \phi(nT)^T P((n-1)T) \phi(nT) \right]
\] 5.3.59 5.3.60 5.3.61

where we have twice used the identity
\[
\det [ I + BC ] = \det [ I + CB ] 
\] 5.3.62
(see [Goodwin and Payne, 1977]). Since $\lambda$ is fixed, this means that minimising $\log \det [P(nT)]$ is equivalent to maximising the scalar

$$\phi(nT)^T P[(n-1)T] \phi(nT)$$

through the choice of $v_j[(n-1)T]$.

If we partition $\phi(nT)$ as

$$\phi(nT)^T = \begin{bmatrix} v_j[(n-2)T] & v_j[(n-3)T] & \ldots & v_j[(n-p)T] \end{bmatrix}$$

where

$$\xi(nT)^T = \begin{bmatrix} v_j[(n-2)T] & v_j[(n-3)T] & \ldots & v_j[(n-p)T] \end{bmatrix} \in \mathbb{R}^{(p-1)}$$

and we partition $P[(n-1)T]$ correspondingly as

$$P[(n-1)T] = \begin{bmatrix} P_{11}[(n-1)T] & P_{12}[(n-1)T] \\ P_{12}[(n-1)T]^T & P_{22}[(n-1)T] \end{bmatrix}$$

where $P_{11}[(n-1)T] \in \mathbb{R}$, $P_{12}[(n-1)T] \in \mathbb{R}^{1 \times (p-1)}$, $P_{22}[(n-1)T] \in \mathbb{R}^{(p-1) \times (p-1)}$, then equation 5.3.63 can be written as

$$P_{11}[(n-1)T] v_j[(n-1)T]^2 + 2 P_{12}[(n-1)T] \xi(nT)^T v_j[(n-1)T] + \xi(nT)^T P[(n-1)T] \xi(nT)$$

This expression is quadratic in $v_j[(n-1)T]$ and is therefore maximised by making

$$v_j[(n-1)T] = v^* \text{ sign } \left[ \frac{P_{12}[(n-1)T] \xi(nT)}{P_{11}[(n-1)T]} \right]$$

where $v^*$ is the maximum allowable amplitude of the input and the sign function is defined as

$$\text{sign}(\alpha) = +1 \text{ if } \alpha \geq 0$$

$$-1 \text{ if } \alpha < 0$$

Choosing $v_j[(n-1)T]$ according to this expression will drive $P(nT)$ as close as possible to its optimal value.

Regardless of which technique is used for choosing the input sequence, the aim is to make the matrix
\[ R(nT) = \sum_{i=1}^{n} \phi(iT) \phi(iT)^T \] 

5.3.70

diagonal, with the diagonal elements being as large as possible, subject to the constraint on
the amplitude of \( v_j(nT) \). This has the effect of minimising the variance of the parameter
estimate and also maximising the information matrix (provided that the estimator is
efficient, so that \( \text{var}[\hat{\theta}_{ij}(nT)] \) coincides with the Cramer-Rao lower bound).

Another advantage of making the \( R(nT) \) matrix diagonal is that it has the effect of reducing
the sensitivity of the estimator to the model order. Up to now, we have not considered the
choice of model order explicitly. Because the minimisation criterion for the least squares
estimator (as in equation 5.2.7) is quadratic in \( \theta_{ij} \), it is easy to show that there is only one
minimum for the least squares problem [Ljung and Soderstrom, 1983]. This is a result of
the fact that the quadratic cost function is convex, so the unique minimum must be a global
minimum.

Although the least squares procedure will give an estimate, \( \hat{\theta}_{ij}(nT) \), which is a global
minimum, this is not a guarantee that the estimated parameters will be good estimates of the
true system parameters. To obtain good estimates, the first requirement is that the model
used in the estimator must be able to describe the response of the true system [Ljung,
1988]. Since we are attempting to estimate the closed loop response between the the
exogenous input and the measured output, it is reasonable to use the model developed in
section 5.1 (see equation 5.1.56) because the output will be a function of the past inputs
plus the coloured noise and a fixed disturbance. This means that the model can describe the
ture system accurately, so the problem is to choose the order of the model. This is
equivalent to asking how many of the past inputs affect the output at a specified time. As
mentioned above, this is related to the settling time of the closed loop response.

There are three possible relationships between the order of the true system and the order of
the model:
i) the model order equals the true order;

ii) the model order is less than the true order;

iii) the model order is greater than the true order.

In the first case, because the estimator generates a global minimum and unbiased estimates, the expected value of the parameter estimates must coincide with the true values of the parameters. For the case where the model order is less than the true order, the model does not have enough parameters to describe the true system, so the global minimum is achieved when the prediction error is minimised (i.e., when the best possible estimator is obtained for the given model). Under these circumstances, the model parameters may be biased to accommodate the higher order terms that are missing from the model. We will see below that the exact values of the parameter estimates, $\hat{g}_{ij}(nT)$, depend upon the the input sequence, $v_j(T)$, $v_j(2T)$, ..., $v_j(nT)$.

In the final case, where the model order exceeds the true order, the estimator is able to obtain unbiased estimates of the true parameters, but it also generates estimates of some non-existent parameters. The estimation of these extra parameters can increase the variance of the estimates of the true parameters, the size of this increase depending (for the model in this study) on the exogenous inputs.

This discussion is an illustration of the trade-off between flexibility and parsimony which is always present in model order selection. The model order must be flexible enough to include a description of the true system (i.e., the model order must be greater than or equal to the order of the true system) but parsimonious so that there are no redundant terms in the model.

As indicated above, the size of the effect of an incorrect model is determined by the exogenous input sequence. Strictly speaking, it is a function of $\phi(IT)$, but for the system in this study, $\phi(IT)$ is a function of only the past inputs and the constant unity. For a least squares estimator, it is possible to see this effect, because given an estimate of a model of a
particular order, we can generate an estimate for a model of a lower order from the same input sequence.

Suppose that we have a system that is modelled by

\[ z(nT) = \theta_T^T \phi(nT) + \bar{e}(nT) \]  

(5.3.71)

(where, for notational simplicity, we have dropped the ij subscripts which indicate the particular SISO estimator). In the absence of a forgetting factor, the optimal offline estimate at time \( t=nT \) is given by (see equation 5.2.7)

\[ \hat{\theta}(nT) = R(nT)^{-1} \sum_{l=1}^{n} \phi(lT) z(lT) \]  

(5.3.72)

This offline estimate will coincide with the recursive estimate once the effect of the initial conditions become negligible (see equation 5.2.13). We will take the order of this model to be \( p \) (ie \( \theta \in \mathbb{R}^p \). Suppose that the parameter vector is split into two parts

\[ \theta^T = [ \theta_1^T, \theta_2^T ] \]  

(5.3.73)

where \( \theta_1 \in \mathbb{R}^{p_1} \) and \( \theta_2 \in \mathbb{R}^{p_2} \) with \( p_1 + p_2 = p \). Then the original model in 5.3.41 can be recast as

\[ z(nT) = \begin{bmatrix} \phi_1(nT) \\ \phi_2(nT) \end{bmatrix} + \bar{e}(nT) \]  

(5.3.74)

where \( \phi(nT) \) has been split into two parts, \( \phi_1(nT) \in \mathbb{R}^{p_1} \) and \( \phi_2(nT) \in \mathbb{R}^{p_2} \). The \( R(nT) \) matrix can be partitioned correspondingly as

\[ R(nT) = \begin{bmatrix} R_{11}(nT) & R_{12}(nT) \\ R_{12}(nT)^T & R_{22}(nT) \end{bmatrix} \]  

(5.3.75)

We are now in a position to generate an estimate of a lower order model from an estimate of the existing model given in 5.3.71. A model of the system of order \( p_1 \) is given by

\[ z(nT) = \theta_1^T \phi_1(nT) + \bar{e}(nT) \]  

(5.3.76)

and the least squares estimate of this model is

\[ \hat{\theta}_1^*(nT) = R_{11}(nT)^{-1} \sum_{l=1}^{n} \phi_1(lT) z(lT) \]  

(5.3.77)
Note that this estimate, \( \hat{\theta}_1^*(nT) \) is not the same as \( \hat{\theta}_1(nT) \) which is the first \( p_1 \) terms of the estimate of the model with an order of \( p \). This new estimate can be expressed in terms of the original estimate of the higher order model by substituting for \( z(nT) \) from the equation 5.3.74 into 5.3.76 to give [Ljung and Soderstrom, 1983]

\[
\hat{\theta}_1^*(nT) = \hat{\theta}_1(nT) + R_{11}(nT)^{-1} R_{12}(nT) \hat{\theta}_2(nT)
\]

5.3.78

This expression shows us how to obtain an estimate of the parameters of a lower order model from an existing estimate of the parameters of a higher order model. The important point to note is that if \( R(nT) \) is diagonal, \( R_{12}(nT) \) will be a zero matrix, so the second term on the right hand side of 5.3.48 disappears, giving

\[
\hat{\theta}_1^*(nT) = \hat{\theta}_1(nT)
\]

5.3.79

This means that when \( R(nT) \) is diagonal the estimate of the parameters in a lower order model coincides with the estimate of the corresponding parameters in a model of higher order. A corollary of this observation is that if \( p_1 \) is the true order of the model, then the estimate of the true parameters are not affected by including higher order terms in the model, provided that the input sequence, \( v_j(nT) \), can be chosen so as to keep \( R(nT) \) diagonal.

Essentially choosing an input sequence to keep \( R(nT) \) diagonal has the effect of decoupling the terms in the model. This is because diagonalising \( R(nT) \) is equivalent to making \( v_j(nT) \) a zero mean, uncorrelated sequence (as described above, see equations 5.3.48 and 5.3.49). If \( v_j(nT) \) is not correlated with \( v_j([n-1]T), v_j([n-2]T), ... \) etc, there is no interaction between the response of each of the elements of the model of the closed loop system given by (from equation 5.1.64)

\[
z_1(nT) = g_1 [G]_{ij} v_j([n-1]T) + [H_2]_{ij} v_j([n-2]T) + ... + [H_p]_{ij} v_j([n-p]T) + d_u + \bar{e}_1(nT)
\]

5.3.80

This means that choosing an uncorrelated input sequence has effectively decoupled the estimation problem into a series of single parameter estimators.
To obtain an estimate of the open loop impulse response of an actuator, we are interested in an estimate of only the first parameter of the closed loop response. This analysis has shown us that, provided the \( R(nT) \) matrix for the model with an order corresponding to the true model order, is diagonal, then the estimate of the first parameter is decoupled from the estimates of the other parameters. This means that a valid estimate of the first parameter could be obtained from the same data when using a model of order one. Replacing a full order estimator with an estimator of order one gives a major reduction in the amount of computation required because \( R(nT) \) is now a scalar, so it is not necessary to perform any matrix inversions. Unfortunately, in practice it is not possible to choose an exogenous input sequence which will make the \( R(nT) \) matrix associated with the full order model, exactly diagonal for all times, particularly if we are using the technique described above where we minimises the determinant of \( R(nT) \) on a "one step ahead" basis. Under these circumstances, if we are using a model order of one, we will be choosing the exogenous input on the basis of minimising the determinant of a scalar, instead of minimising the determinant of the \( R(nT) \) matrix associated with the full order model. It is therefore better to use a model which has a model order equal to an upper bound on the true model order (obtained from an estimate of the closed loop settling time, as described above) and to use the estimate of the first element in the parameter vector to obtain an estimate of the impulse response. Making \( R(nT) \) diagonal or near diagonal will ensure that the estimate of this first element is unaffected by the estimates of the higher order terms.

It is important to note that this discussion of the effects of using different model orders has not placed any restriction on the noise sequence, so it is applicable to the case where the noise is coloured. However, the argument for making \( R(nT) \) diagonal with large elements for minimising the variance of the estimate cannot be used when the noise is not white. This can be seen by recalling that the variance of the parameter estimates from an ordinary least squares estimator when the noise is coloured, is given by (from equation 5.3.17)

\[
\text{var}[\hat{\theta}_j(nT)] = R(nT)^{-1} X(nT)^T X(nT) R(nT)^{-T} \tag{5.3.81}
\]

where we have substituted \( R(nT) \) for \( X(nT)^T X(nT) \). If we consider minimising the
determinant of the variance, i.e.

\[
\det[\text{var}(\hat{\theta}_j(nT))] = \det[\begin{bmatrix} R(nT)^{-1}X(nT)^T & \Xi X(nT)^T R(nT)^{-T} \end{bmatrix}] \quad 5.3.82
\]

\[
= \left[\det[R(nT)^{-1}]\right]^2 \det[X(nT)^T \Xi X(nT)] \quad 5.3.83
\]

Although choosing the input sequence to minimise \(\det[R(nT)^{-1}]\) will minimise the first term, this does not necessarily mean that the whole expression will be minimised because the input sequence will not, in general, simultaneously minimise the second term. However, using the approach to choosing the \(v_j(nT)\) described above has been found in practice, to produce estimates with satisfactory variances, with the additional advantage that the sensitivity of the estimator to uncertainty in the model order is reduced.
CHAPTER 6: CONCLUSION

6.1 Summary of Results

In this study we have investigated the design of a scheme which uses an array of identical actuators for controlling the cross directional variations of web properties. This section summarises the major results that have come out of the study.

• By measuring the cross directional profile relative to the average profile across the width of the web and by ensuring that the array of actuators does not affect the mean level of the profile by adjusting the set points so that their weighted average is the same for all control actions, the cross directional controller can be decoupled from the machine directional control scheme (section 1.2).

• We started by creating a distributed parameter model describing the response of the actuators, which was continuous in both the cross and machine directions. We then formed a sample and hold model of the response of the actuator array by measuring the cross directional web profile at fixed time intervals and holding the set points constant between samples (chapter 2).

• This model was used to investigate the spatial controllability of the system. By assuming that the web was of infinite width and by transferring the analysis to the spatial frequency domain, we were able to identify the controllable frequency components of any cross directional disturbance. We showed that the frequency range of controllable components is limited by \( \omega_c \), the frequency bandwidth of the response of a single actuator (section 3.3).

• The analysis also allowed us to examine the effect of changing the spacing between the actuators. If the frequency of the actuator spacing is \( \omega_s \) (where
\( \omega_s = 2\pi / d \), with \( d \) being the distance between actuators) then if \( \omega_c \leq 1/2 \omega_s \), the array can control all cross directional components with frequencies up to \( \omega_c \), irrespective of their phase. For \( \omega_c > 1/2 \omega_s \), the frequency response of the system is degraded because it is impossible to control certain frequency components without introducing an "aliasing" component at a different frequency. In addition there is an anomalous behaviour when controlling frequencies of \( 1/2 \omega_s \), where the cosine component can be controlled exactly, but the sine component is completely uncontrollable (section 3.3).

- These results define the optimal distance between the actuators to be \( d = \pi / \omega_c \) (or \( 1/2 \omega_s = \omega_c \)). There is no advantage in moving the actuators closer together than this, because the range of controllable frequencies is not increased and the only effect of reducing the separation is to make the system ill conditioned. Increasing the distance between the actuators beyond the optimal spacing introduces the effects of aliasing which impairs the frequency response (section 3.4).

- The ideal shape of the response of each actuator is one which has a frequency response that is flat over the range \( \omega \in [-\omega_c, \omega_c] \) and is zero outside this region. This corresponds to a spatial response of the form \( \sin(x)/x \) (or sinc(x)). In practice, the physical restraints of the plant will mean that this shape of response cannot be achieved, but the actuators should be designed so that their frequency response (rather than their spatial response) are as close as possible to the optimal. In particular, it is important that the frequency response does not go to zero within \( \omega \in [-\omega_c, \omega_c] \) because this means that there is an uncontrollable component in this frequency range (section 3.4).

- Although the spatial frequency analysis was applied to webs and actuator arrays of infinite width, we showed that the results were a good approximation to the answers obtained from the more practical case of finite width systems, provided
that the width of the spatial response of each actuator is small relative to the width of the web (section 3.5). It is important to stress that all of the results for the design of the shape of response and the spacing of the actuators require the frequency domain analysis of section 3.3 and are not available simply by looking at the response in the spatial domain.

- Because the plant cannot control any components with frequencies above \( \omega_c \), the web profile can be put through a low pass filter with a cutoff at \( \omega_c \) to remove the uncontrollable components, without affecting the response of the system. Once the profile has been filtered, the signal can be sampled in the cross direction at twice the cutoff frequency without any loss of information due to the effects of aliasing. The spacing between profile samples is directly related to the shape of the actuator response via the frequency bandwidth, \( \omega_c \) (section 4.1).

- It is important to design the profile filter carefully so that the filter removes all frequency components above \( \omega_c \) and does not introduce any phase shifts. Spatial averaging, which is the most common form of filtering and sampling used on current cross directional control systems, does not completely eliminate all high frequency components and causes the actuator array to respond to uncontrollable components in the web profile (section 4.1).

- Once the web profile has been filtered and sampled, there are a finite number of measurement points so the response of the actuators at each of these measurement positions can be represented by a multivariable, lumped parameter model of the system. Because we assumed that all of the actuators have identical dynamics, the model consists of a real matrix, \( G \), whose columns contain the responses of the actuators, together with a scalar dynamic term (section 4.2).

- We described the design of a control scheme for this model which incorporates a
real compensator matrix and a dynamic term. By applying a singular value decomposition to the system, it is possible to reduce the choice of the controller dynamics to a SISO design (section 4.3). This controller can be implemented efficiently on a computer with an algorithm based on the Cholesky factorisation (section 4.4).

- The design of the controller requires the matrix of actuator responses, \( G \). We investigated the behaviour of the system when the matrix of the actual responses of the plant does not match the nominal matrix which is used as the basis of the control design and showed that provided the system remains stable, the steady state response of the true system is the same as the response of the nominal system, in the sense that the same components of the profile disturbance will be removed in both systems. By using the spatial frequency analysis from section 3.3, we were able to give bounds on the size the mismatch between the true and nominal \( G \) matrices which could be accommodated before the system became unstable, for common modelling faults such as incorrect gains on the actuator responses, inaccurate shapes of responses and mapping errors. Given that the true \( G \) matrix could be described by an additive perturbation to the nominal matrix, then we indicated how the stability of the true plant could be guaranteed by degrading the dynamic response of the controller. In addition we showed that the robustness of the system could be improved by including a term in the compensator matrix which costed the size of the set point changes at each control action (section 4.5).

- Because the controller dynamics are likely to include an integral term so that steady state cross directional disturbances can be removed, the system will become unstable if one (or more) of the actuators fails or if its set point is limited by the physical constraints of the system. Under these circumstances, the special structure of the algorithm which implements the controller can be readily
modified to accommodate the faulty actuator (section 4.6)

- Although the design of the control system is robust enough to handle errors in the G matrix, it is useful to have an estimate of the matrix for monitoring any changes in the responses of the actuators or for commissioning the controller on a new plant. We showed that the G matrix could be estimated, while the plant was running in closed loop, by adding a small, random perturbation signal to the set points of two actuators. We also developed an online, recursive algorithm for implementing this estimator together with a technique for choosing a perturbation signal which minimised the variance of the estimates and reduced the sensitivity of the results to errors in model order (chapter 5).

6.2 Future Research

The study has provided a model for analysing the response of cross directional control systems and a general method of designing these systems but there are a number of areas where the research could be extended.

- Although the spatial frequency analysis provided the frequency response for a web and an array of actuators of infinite width, the corresponding results for finite arrays were not as clear. This is partly due to choosing to expand the web profile in terms of sinusoids of increasing frequency. These functions form an orthogonal basis on the region $x \in (-\infty, \infty)$, but are not mutually orthogonal on a web of finite width, unless some boundary conditions are imposed so that the value of the profile is restricted at the edges of the web in which case the profile can be expanded in terms of a Fourier series. For finite width webs with no boundary conditions, it would be worth investigating the use of other basis functions. As mentioned in section 3.5, one possibility is to use the set of prolate
spheroidal waves [Slepian and Pollak, 1961] [Landau and Pollak, 1961] which are the solutions of a Legendre equation and have the useful property that the solutions are mutually orthogonal on $x \in (-L, L)$ as well as $x \in (-\infty, \infty)$.

- An annular die for the extrusion of blown plastic film (see section 1.3) is an example of a practical actuator whose web profiles can be expanded in terms of a basis of sinusoids without any error. This is because a web profile must be continuous around the die, so the profile can only contain components which have wavelengths which are an integral number times twice the circumference of the die. We can therefore express any profile in terms of a Fourier series whose fundamental harmonic has a wavelength equal to twice the die circumference. It would be interesting to use the Fourier series, rather than the Fourier transform, to generate a frequency response for the die, in the manner of the analysis in chapter 3.

- A common method of measuring the web profile is a gauge which scans across the web in a finite time. It is not possible to obtain a cross directional profile directly from such a gauge because the measurement will be corrupted by machine direction components. However, it should be possible to process the profile to remove these machine direction components, particularly if an additional measurement is available from a gauge which is not scanning.

- The analysis of section 4.6 showed how to modify the actuator set points when one actuator reached its physical limit. A better approach would be to modify the minimisation procedure in control algorithm by including constraints which ensure that the actuator set points did not reach their limits. This could be done using quadratic programming techniques, although there is a large computational load involved in implementing such a procedure.
• In section 4.5, we derived a robust compensator matrix by limiting the size of the actuator set point changes (see equation 4.5.56). The analysis used a norm based approach together with a very conservative definition of the allowable deviations between the true G matrix and the nominal G matrix. A more elegant method would be to obtain the compensator matrix directly from the minimisation problem (from equation 4.5.50)
\[
\min_{u} \max_{\Delta \in S} \left\| z - (G+\Delta)u \right\|_2
\]
without using any norm inequalities but including a more realistic definition of the set S, which reflects the special structure of the allowable uncertainties, as described in section 4.5.

• Throughout this study we have used the 2 norm to choose the optimal set points. This was mainly done for mathematical simplicity, but it is possible to imagine cases where other norms would be more appropriate. For example, if the quality requirements for the web were such that the cross directional profile should not deviate from the mean by more than a given level, it would make sense to use the \( \infty \) norm in the choice of the optimal set points because this would minimise the maximum variation in the profile, rather than minimising the square error over the whole web. It would be worthwhile investigating the properties and behaviour of controllers based on other norms.

• The procedure developed in chapter 5 for estimating the response of the actuators was described in the context of commissioning a new plant or monitoring an existing plant. However, it is possible that the estimated response could be used to update the calculation of the compensator matrix, K, which would make the control system adaptive. The design of the adaption law and the stability of such a procedure would need to be investigated before it could be implemented.

• The performance of the estimation procedure during startups and changes to the
process needs to be improved to reduce the effects of the transients involved in removing the controllable component of the disturbance.
REFERENCES

ALLGOWER E. L. (1973)


BANKS S.P. (1983)
   State-Space and Frequency-Domain Methods in the Control of Distributed Parameter
   Systems. Peter Peregrinus.

BIERMAN G.J. (1977)

BRACEWELL R.N. (1978)

   The Fast Fourier Transform. Prentice-Hall.

   Plastics Film (2nd ed.). The Plastics and Rubber Institute, London.

BRYANT G.F. (ed.) (1973)
   Automation of Tandem Mills. The Iron and Steel Institute, London.

BRYANT G.F., EDWARDS W.J. and McCLURE C.H. (1973)
   Cold-rolling-mill control-system design. Part 1: system description and control
   objectives. In: Automation of Tandem Mills (Bryant G.F., ed.), The Iron and Steel
   Institute, London.

BRYANT G.F. and SPOONER P.D. (1973)
   On-line adaption of tandem-mill schedules. In: Automation of Tandem Mills (Bryant
   G.F. ed.). The Iron and Steel Institute, London.

CHEN S-C., SYNDER R.M. and WILHELM Jr. R.G.
   Adaptive profile control for sheetmaking processes. Proc. IFAC PRP Control
   Conference, Akron, Ohio, U.S.A., pp. 77-83.

258
CURTAIN R.F. and PRITCHARD A.J. (1977)

CURTAIN R.F. and PRITCHARD A.J. (1978)

DAVIS P.J. (1975)
Interpolation and Approximation, Dover.

DAVIS M.H.A. and VINTER R.B. (1985)
Stochastic Modelling and Control, Chapman Hall.

DOLPHIN M.P. (1988)
New requirements and opportunities in profile control. Proc Process Control Conference, PIRA Paper and Board Division, Eastbourne, Sussex, U.K.

DORF R.C. (1980)
Modern Control Systems (3rd ed.). Addison-Wesley.

DUFF I.S. (1977)

Application of advanced control methods in the paper and pulp industry - a survey. Automatica, 22, no.2, pp. 143-153

DYM H. and MACKEAN H.P. (1972)
Fourier Series and Integrals. Academic.


FEISTKORN W. and SENSEN K. (1987)
FEISTKORN W. and SENSEN K. (1988)


FRANCIS, B.A. (1987)
A course in $H\infty$ control theory. Springer Verlag.


Dynamic System Identification: Experiment Design and Data Analysis. Academic.


GUSTAVSSON I., LJUNG L. and SODERSTROM T. (1977)
GUSTAVSSON I., LJUNG L. and SODERSTROM T. (1981)

HADRA M.J. (1985)
Selecting actuators for cross-machine control. TAPPI J., 68, no. 6, pp. 60-62.

HAMMING R.W. (1977)
Digital Filters. Prentice-Hall.


HAVERINEN W. (1983)
Results report with a new caliper sensor and CD control on a Canadian newsprint machine. Proc 3rd International Pulp and Paper Control Symposium, Vancouver, BC., Canada, pp. 111-120.


JERRI A.J. (1977)

KAILATH T. (1980)
Linear Systems. Prentice-Hall.

KARLSSON H., HAGLUND L., and HANSON A. (1985)
Optimal cross-direction basis weight and moisture profile control on paper machines. Pulp and Paper Canada, 86, no. 8, pp. 213-218.

KUBRUSLY C.S. and MALEBRANCHE M. (1985)
   Sensors and controllers location in distributed systems - a survey. Automatica, 21, no. 2, pp. 117-128.

KWAKERNAAK H. and SIVAN R. (1972)

LANDAU H.J. and POLLAK H.O. (1961)

LARKE E.C. (1963)
   The Rolling of Strip, Sheet, and Plate (2nd ed.). Chapman and Hall.

LINDEBORG C. (1986)

LIGHTHILL M.J. (1962)
   Introduction to Fourier Analysis and Generalised Functions. Cambridge University Press.

LJUNG L. (1987)
   System Identification: Theory for the User. Prentice-Hall.


LJUNG L. and SODERSTROM T. (1983)
   Theory and Practice of Recursive Identification. MIT Press.

   Measurement and Control in Papermaking. Adam Hilyer Ltd.


NOBLE B. (1976)

PAPOULIS A. (1962)

OWENS D.H. (1973)

OWENS D.H. (1978)
Feedback and Multivariable Systems. Peter Peregrinus.

RABINER L.R. and GOLD B. (1975)


RINGWOOD J.V. and GRIMBLE M.J.

ROBERTS W.L. (1978)
Cold Rolling of Steel. Marcel Decker.
ROBINSON A.C. (1971)


ROTH R.C. (1986)

RUDD N.E. (1978)

Real and Complex Analysis. McGraw Hill.

SILER S.J. (1984)

SLEPIAN D. and POLLAK H.O. (1961)

SMITH K.E. (1985)
Cross direction control is still top process automation trend. Pulp and Paper, 59, no. 2, pp. 72-76.

SODERSTROM T., LJUNG L. and GUSTAVSSON I. (1976)

SPOONER P.D. and BRYANT G.F. (1973)
Analysis of shape and discussion of problems of scheduling, set-up and shape control. Proc. The Metals Society Conference on Shape Control, Chester, U.K.
STEWART G.W. (1979)

The effects of rounding error on an algorithm for downdating a Cholesky factorization.


THOMAS C. (1977)

The Paper Chain. Earth Resources Research Ltd.

WAHREN D. (1986)

Recent highlights in paper technology. TAPPI J., 62, no. 3, pp. 36-45.

WALLER M.H. (1987)

Recent developments in headboxes. TAPPI J., 70, no.1, pp. 33-36.

WANG P.K.C. (1968)

Theory of stability and control for distributed parameter systems (a bibliography). Int.


On the controllability of CD variations in sheet properties. Proc. TAPPI Engineering


Control algorithms for cross-directional control: the state of the art. Proc. 5th

International IFAC/IMEKO PRP Control Conference, Antwerp, Belgium, pp.

139-150.


A new control technique for cross machine control of basis weight. Proc. 5th

International IFAC/IMEKO PRP Control Conference, Antwerp, Belgium, pp.

151-158.
APPENDIX A: MATHEMATICAL DEFINITIONS REQUIRED FOR SECTION 2.2

Definition A1 A real Banach space, $X$, is a complete normed vector space over $\mathbb{R}$. For reference, a vector norm satisfies

(i) $||x||_X = 0$ if and only if $x = 0$
(ii) $||\alpha x||_X = |\alpha||x||_X$ for all $\alpha \in \mathbb{R}, x \in X$
(iii) $||x + y||_X \leq ||x||_X + ||y||_X$ for all $x, y \in X$

(see [Rudin, 1974, chapters 2 and 3] for definitions of vector spaces and completeness).

The subscript, $X$, on the norm indicates the space over which the norm is taken.

In general, because we are dealing with real quantities, we shall restrict our discussion to real Banach spaces (rather than complex ones).

Definition A2 We will define an inner product between two elements of $X$, $x$ and $y$, as $\langle x, y \rangle$ where

(i) $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in X$
(ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ for all $x, y, z \in X$
(iii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in X$ and $\alpha \in \mathbb{R}$
(iv) $\langle x, x \rangle \geq 0$ for all $x \in X$
(v) $\langle x, x \rangle = 0$ only if $x = 0$

Definition A3 A norm on a real Banach space, $X$, defined by an inner product, is a map given by

$$||x||_X = \sqrt{\langle x, x \rangle}$$  \hspace{1cm} (A1)

(see [Rudin, 1974, chapter 4] to show that this definition satisfies the properties of a norm given in definition A2).
Definition A4 A Hilbert space, $H$, is a Banach space whose norm is defined from an inner product as in equation A1.

Most of our discussions in the study will be based upon functions which are members of $L^2[a,b]$ (i.e., those functions which are Lebesgue square integrable on the segment $x \in [a,b]$). See [Rudin, 1974, chapter 1]). This set of functions forms a Hilbert space under the scalar product [Rudin, 1974, chapter 4]

$$< f, g > = \int_a^b f(x) \overline{g(x)} \, dx \quad A2$$

and the norm

$$|| f || = \sqrt{< f, f >} = \left[ \int_a^b |f(x)|^2 \, dx \right]^{1/2} \quad A3$$

We now turn our attention to transformations between Banach spaces.

Definition A5 A linear transformation, $T$, from a real linear space, $X$, to a real linear space, $Y$, is a map $T: X \rightarrow Y$, such that

$$T(\alpha x + \beta y) = \alpha Tx + \beta Ty \quad \text{for all } x, y \in X \text{ and } \alpha, \beta \in \mathbb{R} \quad A4$$

We will be interested in transformations which are bounded.

Definition A6 A transformation, $T: X \rightarrow Y$ between two normal spaces is said to be bounded if

$$||Tx||_Y \leq k ||x||_X \quad A5$$

for some finite, real constant, $k$ and for all $x \in X$. Note that the subscripts, $X$ and $Y$ on the norms indicate which of the spaces that the norm is measured over.
Definition A7 If $T$ is a bounded transformation between two normed, linear spaces, $x$ and $y$, we define the norm of this transformation, $\|T\|$, by

$$\|T\| = \sup_{x \in X, x \neq 0} \left[ \frac{\|Tx\|_Y}{\|x\|_X} \right]$$

It can be shown [Rudin, 1974] [Curtain and Pritchard, 1977] that this definition satisfies the properties of a norm given in Definition A2.

This definition allows us to recast Definition A6 for bounded transformations in a slightly different form. From Definition A7 we can see that

$$\|Tx\|_Y \leq \|T\| \|x\|_X \text{ for all } x \in X$$

For a bounded transformation, a value of the constant, $k$, which satisfies the criterion in Definition A6 is given by $k = \|T\|$. Thus, bounded transformations must have finite norms.

Definition A8 If $X$ and $Y$ are normed linear spaces, then we can define $L(X,Y)$ as the normed linear space of bounded transformations $T:X \rightarrow Y$ with the norm given in Definition A7. It can be shown [Curtain and Pritchard, 1977] that if $Y$ is a Banach space, then so is $L(X,Y)$.

For the special case where $X=Y$ (i.e., the range of the transformation is the same as the domain) we will write $L(X,X)$ as $L(X)$. Since $L(X)$ is a linear space, it is closed under addition and scalar multiplication. It is also closed under transformation composition [Curtain and Pritchard, 1977]

$$T_1 T_2 x \in X \text{ for all } T_1, T_2 \in L(X) \text{ and } x \in X$$

Since $L(X)$ is a space of bounded transformations, this means that the combined transformation is also bounded.
\[
\| T_1 T_2 x \| \leq \| T_1 \| \| T_2 x \| \leq \| T_1 \| \| T_2 \| \| x \|
\]

for all \( x \in X \) and for all \( T_1, T_2 \in \mathcal{L}(X) \) with \( \| T_1 \| \| T_2 \| \) \( A9 \)

This gives us
\[
\| T_1 T_2 \| \leq \| T_1 \| \| T_2 \| \quad A10
\]

We can use this result to investigate an important transformation; the exponential transformation.

**Definition A9**

Let \( T_1 \in \mathcal{L}(X) \), then the exponential transformation is defined by
\[
e^{T_1} = I + T_1 + \frac{T_1^2}{2!} + \frac{T_1^3}{3!} + \ldots \quad A11
\]

(where \( I \) is the identity transformation).

Although \( T_1 \) is bounded (because \( T_1 \in \mathcal{L}(X) \)) it is not immediately obvious that \( e^{T_1} \) is bounded. However, using the result in equation \( A8 \), we can see that
\[
\| e^{T_1} \| \leq 1 + k + \frac{k^2}{2!} + \frac{k^3}{3!} + \ldots = e^k \quad A12
\]

where \( \| T_1 \| = k \). Since \( T_1 \) is bounded, \( k \) is finite, so \( \| e^{T_1} \| \) is finite and \( e^{T_1} \) is bounded. This means that \( e^{T_1} \in \mathcal{L}(X) \). For unbounded \( T_1, e^{T_1} \) is not well defined.