Control and Communication with Signal-to-Noise Ratio Constraints

Johannesson, Erik

2011

Document Version:
Publisher's PDF, also known as Version of record

Control and Communication with Signal-to-Noise Ratio Constraints

Control and Communication with Signal-to-Noise Ratio Constraints

Erik Johannesson

Department of Automatic Control
Lund University
Lund, October 2011

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

ISSN 0280–5316
ISRN LUTFD2/TFRT--1087--SE

Printed in Sweden by Media-Tryck.
Lund 2011

"If I have seen a little further it is by standing on the shoulders of Giants."

Isaac Newton

Abstract

This thesis is about two problems in the intersection of communication and control theory. Their common feature is that they involve communication over an additive white noise channel with a signal-to-noise ratio (SNR) constraint.

The first problem concerns the transmission of a real-valued signal from a partially observed Markov source. The distortion criterion is the mean squared error and the transmission is subject to a delay constraint, which introduces the need for real-time coding. The problem is first considered for scalar-valued signals when the channel has no feedback and then, in turn, generalized to each of the cases with non-white channel noise, vector-valued signals or channel feedback.

It is shown that jointly optimal encoders and decoders within the linear time-invariant (LTI) class can be obtained by solving a convex optimization problem and performing a spectral factorization. The functional to minimize is the sum of the well-known cost in a corresponding Wiener filtering problem and a new term that is induced by the channel noise.

The second problem, which can be viewed as a generalization of the first problem, concerns a networked control system where an LTI plant, subject to a stochastic disturbance, is to be controlled over the channel. The controller is based on output feedback and consists of an encoder/observer that measures the plant output and transmits over the channel, and a decoder/controller that receives the channel output and issues the control signal. The objective is to stabilize the plant, satisfy the SNR constraint and minimize the variance of the disturbance response. The problem is studied for channels without and with feedback.

In both cases, it is shown that optimal controllers within the LTI class can be obtained by solving a convex optimization problem and performing a spectral factorization. Previously known conditions on the SNR for stabilizability follow directly from the constraints of these optimization problems.

Acknowledgements

It seems that being grateful makes you happy [15]. If it is so then working with this thesis has provided me with many reasons for happiness. My gratitude extends to a number of people who have supported and helped me along the way. I could not have done it alone.

First of all, I want to thank Anders Rantzer and Bo Bernhardsson. It has been a privilege to work with two world-class researchers such as Anders and Bo. I have deep respect and admiration for them both, because of their technical knowledge, remarkable intuition and teaching skills. It is a challenge to keep up with them when standing in front of a blackboard, trying to solve a problem. But I have learned a great deal from doing it, and it was fun too!

Anders and Bo deserve credit for their generosity, both with ideas and their time. As my main supervisor, Anders gave me the freedom to find a topic that interested me, helped me to gain momentum, and encouraged me to be proud of my achievements when the results started to come. Bo has done more than anyone could expect from a co-supervisor. Somehow he always found the time for me, even when he was out in the industry, tuning a billion control loops.

I got to know Andrey Ghulchak at a critical stage in my research. I want to thank him for helping me with some of the most technical and difficult parts of this thesis. Andrey is a mathematical wizard who could have grown tired of my questions about functional analysis a long time ago, but he never did.

I want to thank all of my helpful colleagues for giving me valuable suggestions, and for taking the time to explain things that I didn't understand. In particular, Toivo Henningsson, Karl Mårtensson, Aivar Sootla and Giacomo Como have given me important comments that helped me improve this thesis. It has been a pleasure to share office first with Peter Alriksson and then with Daria Madjidian, who I hope enjoyed our discussions as much as I did. It has also been very nice to be able to share the

worries of thesis-writing together with Per-Ola Larsson.

I want to thank Ather Gattami for inspiring me to look into the interesting field of control with communication constraints. Henrik Sandberg left the department before I joined, but he has always been supportive when we have met.

I want to thank Anton Cervin and Toivo Henningsson for the nice work that we did together on event-based control. It taught me that fine research doesn't have to be so difficult if you have a good problem formulation. It also helps to only consider first-order systems...

I have really enjoyed being at the Department of Automatic Control at Lund University. It is pervaded by a great social and friendly atmosphere and a can-do attitude that makes it easy to do your best. I am honored to have been a member of this talented group of people. I believe that the helpful, supportive and efficient administrative and technical staff is an important factor behind the department's success, and they deserve recognition for this.

I have had the opportunity to travel a lot during my Ph.D. studies: I want to thank Nuno Martins for hosting me during an inspiring visit at the University of Maryland, USA. I was also very happy to get the chance to teach a control class at Zhejiang University, China, together with Anton Cervin and Daria Madjidian. I am grateful for the hospitality that our hosts showed there.

My friends and my family may not always understand what it is that I do, but they have constantly encouraged me and helped me to keep my mood up. I hope they realize how important they are to me.

Finally, Emma has given me more love and support than I could ever deserve. Thank you for always believing in me, even when I didn't.

Erik

Financial Support

The author gratefully acknowledges funding received for this research from the Swedish Research Council through the Linnaeus Center LCCC; from the European Union's Seventh Framework Programme under grant agreement number 224428, project acronym CHAT; and from the ELLIIT Strategic Research Center.

Contents

Preface

Motivation

This thesis is about communication and control problems, and aims to be of interest for researchers in both disciplines. The main difference between communication theory and control theory lies in the importance of delays: In communication theory, it is often assumed that delays are of little or no importance. The most famous results are asymptotical in nature, and practical communication systems typically employ block coding with large blocks in order to achieve high performance. On the other hand, it is well-known that time delays can have a detrimental effect on the stability and performance of control systems.

This contrast naturally elicits two questions: How to design a communication system when there is a bound on the accepted delay? And how to design a control system when there are communication limitations? Two problems, each related to one of these questions, are studied in this thesis. It will be seen that the questions are closely related to each other. In fact, under the circumstances studied here, the answer to the second can also give an answer to the first.

In the first type of problem a real-valued source signal is to be communicated over a noisy communication channel under a real-time constraint, using noisy measurements of the source. The objective is to design an encoder-decoder pair that minimizes the mean squared error distortion. A practical example of this is the transmission of a voice signal from a mobile phone while simultaneously filtering out the background noise. The receiver should then be able to reproduce the sound of the voice as well as possible, with some upper bound on the latency.

The setting differs from the traditional in communication theory, both because of the noise at the source and because of the real-time constraint.

In this thesis, the problem is considered for different communication channel models with signal-to-noise ratio (SNR) constraints. A procedure is developed for finding optimal linear encoders and decoders. Even though this problem is mainly seen as a communication problem, it may also be interpreted as an estimation problem or as a feed-forward control problem.

A current trend in control engineering, both in theory and practice, is for control systems to become more distributed and dependent on communication over different types of networks. Traditional control theory assumes perfect communication, so it has become important to study the impact of communication limitations on the control performance. These limitations may take many forms, such as rate limitations, variable time delays, packet drops or SNR constraints. Here, the focus is on the latter. The reason is that while the analysis and design becomes relatively simple in an SNR framework, the results can still be useful in a broader context.

The second type of problem studied in this thesis concerns a scenario where a plant is to be controlled over a noisy communication channel. The controller is divided into two subsystems: one sensor and encoder part that measures the plant output and encodes information for transmission over the channel, and one decoder and controller part that receives the transmission and determines the control signal. The objective is to design the two subsystems so that the plant is stabilized and the disturbance response is minimized. This problem is considered for two different channel models with SNR constraints. A procedure is developed for finding optimal linear encoders and decoders. Conditions on the SNR for stabilizability of the system are also derived.

The optimization of encoders, decoders and controllers is restricted in this thesis to the linear time-invariant class. This may obviously result in the obtained solutions being suboptimal in a larger class of filters. There are three reasons why this restriction is made: First, the problems become tractable. Optimizing over general mappings is a much more difficult problem than optimizing over linear time-invariant ones. Second, using linear solutions means that all of the control-theoretic tools available for analysis of linear systems can be applied, for example to analyze system robustness. The third reason is that linear filters are generally easier to implement.

As is always the case with theoretical studies, all models in this thesis reflect an idealized version of reality. This is because the aim is not to directly develop solutions for practical problems. However, relatively simple models allow us to study the fundamental aspects of the problems. It is hoped that the results presented here will contribute to a general understanding of these and related problems, for example by providing theoretical limitations, as well as ideas for further development.

Outline and Contributions

This section contains an outline of the thesis and a summary of the contributions. For the most part, the thesis follows the chronological order in which the results were obtained. As a consequence, the problems will be considered in order of increasing generality, in the sense that solutions to problems presented later can be used to solve earlier problems as well. The relationships between the considered problems are illustrated in Figure 0.1.

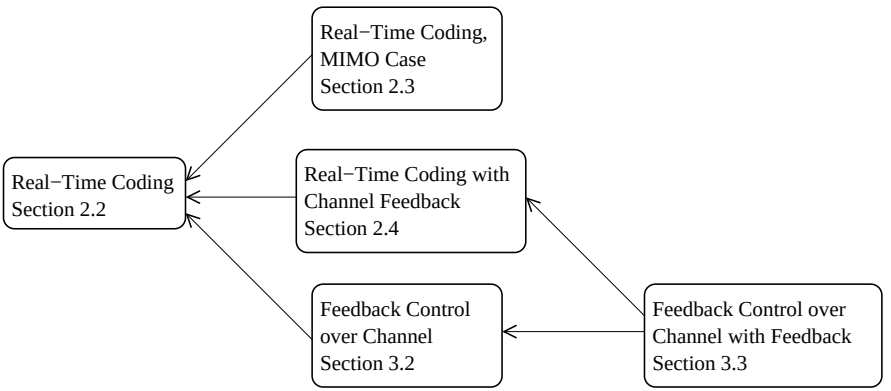

Figure 0.1 Relationships between the problems and sections in this thesis. An arrow pointing from A to B indicates that the solution of A can be used to solve B.

Obviously, a more compact thesis could have been written, in which the most general problems were first solved and the other problems simply presented as special cases. But that thesis would have been more difficult to read. The solution and the main ideas behind it could have become obscured by the additional difficulties posed by the more general problems. Furthermore, the author believes that the special cases are important enough to merit separate presentations and solutions.

Chapter 1: Background

The first chapter gives a brief introduction to communication theory and networked control systems. Since these are vast subjects, the exposition is limited to the parts that are relevant for the forthcoming discussions. Some mathematical preliminaries are also presented.

Chapter 2: Real-Time Coding for a Noisy Channel

The problem of designing jointly optimal encoder-decoder pairs for real-time coding of a partially observed markov source under a mean squared

error distortion measure is considered. The encoder and decoder are constrained to be causal and time-invariant linear filters. The problem is considered for an additive white noise channel with SNR constraint, with and without channel feedback. The case without channel feedback is generalized to cover non-white channel noise and a setting with vector-valued signals and parallel channels.

It is shown that optimal linear[1] encoders and decoders can be obtained by solving a convex optimization problem and performing a spectral factorization. The convex optimization problem is expressed in the product of the encoder and decoder transfer functions and, when channel feedback is available, the transfer function of an additional feedback filter. The functional to be minimized is a weighted sum of a 2-norm, which is the cost in a corresponding Wiener filter problem, and an additional term that is induced by the channel noise. In the case with channel feedback, it is shown how to pose the optimization problem as a semidefinite program. It is also demonstrated by example that channel feedback may improve the performance of linear coding.

Related Publications

E. Johannesson, A. Rantzer, B. Bernhardsson and A. Ghulchak, "*Encoder and Decoder Design for Signal Estimation*," in *Proc. American Control Conference*, Baltimore, USA, June 2010.

E. Johannesson, A. Ghulchak, A. Rantzer and B. Bernhardsson, "*MIMO Encoder and Decoder Design for Signal Estimation*," in *Proc. 19th International Symposium on Mathematical Theory of Networks and Systems*, Budapest, Hungary, July 2010.

E. Johannesson, "*Signal Estimation over Channels with SNR Constraints and Feedback*," in *Proc. 18th IFAC World Congress*, Milano, Italy, August 2011.

Chapter 3: Feedback Control over a Noisy Channel

The problem of designing an optimal linear output feedback controller for a linear plant controlled over an additive noise channel with SNR constraint is considered. The plant has a stochastic disturbance and the controller is divided into two subsystems that are separated by the communication channel. The controller should stabilize the system and minimize the variance of the plant output while satisfying the SNR constraint. The problem is considered with and without channel feedback.

[1]Note that it is not claimed that linear solutions are optimal for any of the problems considered in this thesis, except when so is explicitly stated.

It is shown that optimal linear controllers can be obtained by solving a convex optimization problem and performing a spectral factorization. The convex optimization problems are similar to the one obtained in the channel feedback case in Chapter 2, although it is now expressed in the Youla parameter. The functional to be minimized has been found previously for the case with channel feedback [53] but the problem considered here is slightly more general and the solution has a simpler structure.

Necessary and sufficient conditions for stabilizability of a plant with stochastic disturbance under the SNR constraint follow from the derivations of the convex optimization problems. The obtained conditions are shown to correspond to previously known ones.

It is shown how to pose the optimization problems as semidefinite programs. Finally, it is demonstrated that the solutions to the coding problems with scalar-valued signals in Chapter 2 can be obtained as special cases of the solutions to the feedback control problems.

Related Publications

E. Johannesson, A. Rantzer and B. Bernhardsson, "*Optimal linear control for channels with signal-to-noise ratio constraints*," in *Proc. American Control Conference*, San Francisco, CA, USA., June 2011.

E. Johannesson, A. Rantzer and B. Bernhardsson, "*A Framework for Linear Control over Channels with Signal-to-Noise Ratio Constraints*," accepted for presentation at *The 9th IEEE International Conference on Control & Automation*, Santiago, Chile, December 2011.

Chapter 4: Conclusions

Conclusions are made and a number of areas for further research is suggested.

Appendix A: Some Technical Proofs

The proofs of some of the more technical lemmas have been put in the appendix.

Relation to Research by Silva, Derpich, et al.

Some of the results presented in this thesis are closely related to those recently presented by Eduardo Silva, Milan Derpich and co-authors in the publications [53, 12, 10]. The author would therefore like to comment on the relation between our respective results and their development.

Detailed comparisons of the results are given in Chapters 2 and 3. For now, it is noted that, despite the similarities, there are several technical differences between the results presented here and those in [53, 12, 10]. The fact that this thesis presents original work should be obvious from

these technical differences as well as the differences in the approach used to solve the problems in this thesis and in the work of Silva, Derpich and co-authors.

Regarding the historical development, the author was first made aware of the existence and relevance of [53, 12, 10] in January 2011 by one of the reviewers of [27]. The results in Chapter 2 and Section 3.2 were accordingly obtained without prior knowledge of these publications, with two exceptions:

- Convexity of the functionals in the equivalent minimization problems had not yet been established. It had, however, been shown that they are quasiconvex. The convexity proofs provided in this thesis were independently developed.
- The control problem in Section 3.2 had only been considered for a single-input, single-output (SISO) plant. The extension to the more general structure has, to the author's best knowledge, not been done elsewhere.

Furthermore, the insight, used in [53], that the orthogonality between strictly proper transfer functions and constants can be used to rewrite norm expressions have been helpful in obtaining cleaner solutions, though it was not critical for actually solving the problems.

The results in Section 3.3 were, in contrast, developed using the results in [53] with the purpose of generalizing and improving some of the results presented in that paper.

Other Publications

The following is a list of other publications co-authored by the author of this thesis. These publications are about event-based control and are not directly related to this thesis.

E. Johannesson, T. Henningsson and A. Cervin: "*Sporadic Control of First-Order Linear Stochastic Systems*". In *Proc. 10th International Conference on Hybrid Systems: Computation and Control*, Springer-Verlag, Pisa, Italy, April 2007.

A. Cervin and E. Johannesson: "*Sporadic Control of Scalar Systems with Delay, Jitter and Measurement Noise*". In Proc. 17th IFAC World Congress, Seoul, Korea, July 2008.

T. Henningsson, E. Johannesson and A. Cervin: "*Sporadic Event-Based Control of First-Order Linear Stochastic Systems*". *Automatica*, 44:11, pp. 2890-2895, November 2008.

1

Background

The purpose of this chapter is to provide some context and background that is necessary to understand the problems and results presented in this thesis. The first and second sections give a brief introduction to communication theory and networked control systems, respectively. The third and final section presents the mathematical notation along with some definitions and results that are used in the thesis.

1.1 Introduction to Communication Theory

The fundamental problem of communication, as defined by Claude Shannon in his landmark paper, is to reproduce at one point either exactly or approximately a message selected at another point [48]. See Figure 1.1 for an illustration.

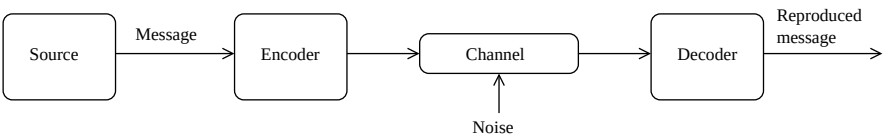

Figure 1.1 The fundamental problem of communication: The source generates a message that is to be reproduced at another point. The encoder determines, based on the message, a signal to transmit over a communication channel. The decoder receives the channel output, which may be affected by noise, and tries to reproduce the original message.

Channel Models

A multitude of different models of communication channels have been proposed and studied in the literature. These can be deterministic or stochastic, have discrete or continuous input and output alphabets, and be dis-

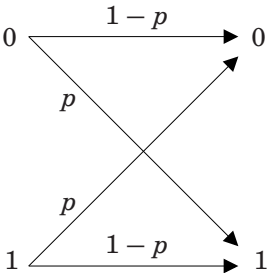

Figure 1.2 Binary Symmetric Channel (BSC)

crete or continuous in time. A channel may have memory, in which case the channel output can depend on its previous input and output. Some simple channel models will now be presented.

Digital Error-Free Channel The digital error-free channel allows error-free transmission of $\mathcal{R}$ bits per time unit. The parameter $\mathcal{R}$ is called the rate of the channel.

Binary Symmetric Channel (BSC) A common simple model of a memoryless noisy channel is the BSC, illustrated in Figure 1.2. As the name suggests, the BSC has binary input and output. For each channel use, the output is equal to the input with probability $1 - p$ and not equal with probability p.

The AWN and AWGN Channels The Additive White Noise (AWN) channel, illustrated in Figure 1.3, takes a real number t as input and the output r is given by

$$r = t + n, \tag{1.1}$$

where the channel noise n is a random number with some specified distribution. The noise is further assumed to be independent between different channel uses. The channel input must satisfy a power constraint. That is,

$$\mathrm{E}(t^2) \leq \sigma_t^2,$$

where $\mathrm{E}(x)$ denotes the expected value of x.

If the variance of n is denoted by σ_n^2, the Signal-to-Noise Ratio (SNR) of the channel is given by $\sigma^2 = \sigma_t^2/\sigma_n^2$. By scaling the input and output properly, any channel with a given SNR can be made equivalent to a channel with $\sigma_t^2 = \sigma^2$ and $\sigma_n^2 = 1$. For this reason, it will from now on be assumed that the transmission power is constrained by σ^2 and that the

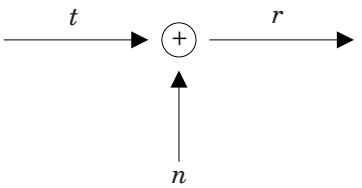

Figure 1.3 Additive White Noise (AWN) channel

channel noise has variance 1. The SNR constraint is thus equivalent to a power constraint.

If the noise n is Gaussian this channel is known as the Additive White Gaussian Noise (AWGN) channel. This is the most common channel model with a continuous alphabet and is used to model many practical channels including radio and satellite links [9].

Parallel Channels A simple example of a channel with multiple inputs and outputs is a channel with n independent parallel AWN or AWGN channels with a common power constraint. This could be used to model a non-white AWN channel, where each of the components represent a different frequency [9]. For this channel, (1.1) holds with r, t and n vector-valued. The noise vector n is assumed to have a diagonal covariance matrix and the power constraint is

$$\sum_i \mathrm{E}(t_i^2) = \mathrm{E}(t^T t) \leq \sigma^2.$$

Channels with Feedback If a channel has feedback it means that the received symbols are sent back to the encoder so that it can use them to decide on the next transmission. This is illustrated in Figure 1.4. This channel can model a physical situation where the communication link is much stronger in one direction, for example as in ground-to-satellite communication [46]. It can also be used as a model of quantization error, since those are known exactly by the encoder [10]. In general, the feedback may be subject to noise, but it will be assumed here that it is error-free.

Information Theory

The fundamental problem of communication has been studied extensively in the field of Information theory, which was established as a result of Shannon's paper. The process of manipulating information for transmission over a communication channel is known as coding. Coding is often divided into source coding and channel coding.

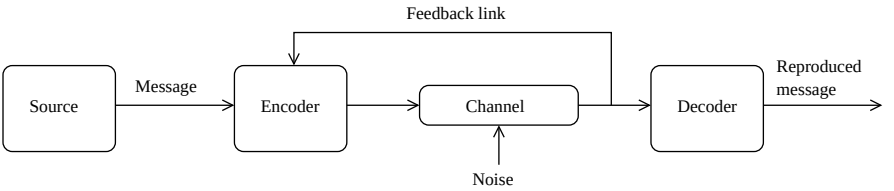

Figure 1.4 Communication channel with feedback.

In general, the message originating from the source contains some redundancy due to statistical dependence. Source coding is a process that exploits this in order to represent the message at a lower rate. The smallest rate at which it is possible to represent a source and perform error-free reconstruction (perhaps after transmission over a digital error-free channel) is called the entropy $\mathcal{H}$ of the source. As an example, a binary source, which has output 0 with probability p and output 1 with probability $1-p$, has entropy

$$\mathcal{H} = -p\log_2 p - (1-p)\log_2(1-p) \text{ bits/time unit.}$$

The maximum value, which is 1, is attained for $p = 0.5$ [48]. Source coding is sufficient for communication over a digital error-free channel. But when the channel is noisy, it becomes necessary to add some kind of redundancy to decrease the sensitivity to the channel noise.

The objective of channel coding is to transmit information over a noisy channel with a minimum of error in the receiving end. A channel code has a rate, which is the amount of information about the message that is being transmitted per time unit. It would be suspected that there is a strict trade-off between this rate and the amount of errors. Shannon showed, however, that for every communication channel there is a rate below which communication is possible with arbitrarily low probability of error. This rate, which can be used to characterize the channel, is called the capacity $\mathcal{C}$ of that channel. In contrast, any code that has a rate above $\mathcal{C}$ will result in an error probability that cannot be made arbitrarily small. Note that the theory does not say how to find practically useful channel codes that achieve capacity rates for general channels [48, 9].

To provide some examples, the digital error-free channel obviously has capacity $\mathcal{R}$ bits/time unit. The BSC has capacity

$$\mathcal{C}_{BSC} = 1 + p\log_2 p + (1-p)\log_2(1-p) \text{ bits/time unit}$$

and the AWGN channel has capacity

$$\mathcal{C}_{AWGN} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_t^2}{\sigma_n^2}\right) \text{ bits/time unit.} \tag{1.2}$$

Surprisingly, the addition of feedback to a memoryless channel does not change its capacity [47]. It may, however, reduce the complexity required of a coding scheme to achieve a given level of performance [46].

A separation of a coding system into source and channel coding operations provides a nice structure for designing practical communication systems. It is one of the main results of information theory that such a separation is in fact optimal. The well-known coding theorem says, accordingly, that it is possible to communicate a source, with arbitrarily low error probability, if and only if $\mathcal{H} \leq \mathcal{C}$ [48, 20].

Rate Distortion Theory

When the source entropy is higher than the channel capacity, it is not possible to communicate without error. This is, for example, the case when the source has a continuous probability distribution and the capacity is finite. The source coder must then make an approximation of the source that satisfies the rate constraint imposed by the capacity. The quality of the approximation is determined by a distortion measure that quantifies the difference between the original message and the approximation. A common distortion measure is the mean squared difference between each source symbol and the corresponding estimate.

In rate distortion theory, the relationship between the rate $\mathcal{R}$ and the distortion $\mathcal{D}$ is studied. Specifically, a rate distortion function $\mathcal{R}(\mathcal{D})$ provides the trade-off between these two quantities for each source and distortion measure. As an example, the rate-distortion function for a white Gaussian source with variance σ^2 is [3]

$$\mathcal{R}(\mathcal{D}) = \begin{cases} \frac{1}{2}\log_2 \frac{\sigma^2}{\mathcal{D}}, & 0 \leq \mathcal{D} \leq \sigma^2 \\ 0, & \mathcal{D} > \sigma^2. \end{cases}$$

Just as for error-free communication, there is a theorem that sets the boundary for approximate communication: Given a distortion $\mathcal{D}$ and the corresponding rate $\mathcal{R}(\mathcal{D})$, there exists a coding system that can communicate the source with a distortion arbitrarily close to $\mathcal{D}$ if and only if $\mathcal{R}(\mathcal{D}) < \mathcal{C}$ [48].

Real-Time Coding

All of the results mentioned so far are asymptotic. The proofs rely on block coding schemes with high complexity, where the block length is allowed to approach infinity. Since this would require an infinite delay, practical communication systems most often operate at a performance below the theoretical optimum. Nevertheless, it should be noted that great progress

in approaching the so called Shannon limit has recently been made due to the development of new coding techniques [31].

If there is a fixed bound on the tolerated delay in a communication problem, then it is called a real-time coding problem. These problems are also referred to as causal coding problems, although often with weaker delay constraints. These problems are very different from, and more difficult than, the classical formulation of the communication problem, which ignores the delay aspect. Consequently, a lot less is known about the solutions. For example, the separation between source and channel coding is not optimal for real-time problems. Instead it is necessary to perform joint source-channel coding. A nice overview of the literature on real-time coding is given in [33].

An interesting observation is the fact that channel feedback, even though it does not change the capacity, is generally useful for real-time coding problems [58]. It is also interesting to note that there are special circumstances when optimal performance can be achieved without any coding, which also implies that communication can occur without delay. This is for example the case when a white Gaussian source is to be transmitted over an AWGN channel, with a quadratic distortion measure [22].

Remote or Partially Observed Sources

Sometimes it can be assumed that the message is distorted before it reaches the encoder. This is referred to as a remote source or, alternatively, a partially observed source and is illustrated in Figure 1.5. These problems are also sometimes called indirect problems. This model is applicable to the problem of transmitting data that is affected by measurement noise, or to the case when a digital communication system must interface with a given analog-to-digital converter [3]. An application example is the presence of background noise in mobile speech communication, which should ideally be filtered out before transmission.

An optimal coding system for a remote source problem with mean squared error criterion and additive noise consists of an optimal estimator followed by optimal encoding and decoding of the estimate. This result, however, holds only asymptotically in the block length and is thus

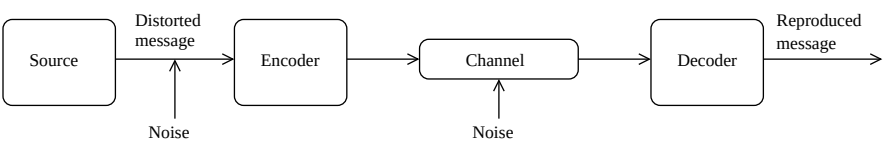

Figure 1.5 If the encoder's observation of the message is affected by noise the source is said to be remote or partially observed.

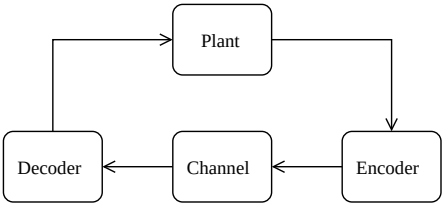

Figure 1.6 Feedback control over a communication channel. The control system consists of an encoder, which also does measurement filtering, and a decoder, which also determines the control signal.

not necessarily valid when there are real-time constraints [63].

1.2 Introduction to Networked Control Systems

Networked control systems are control systems that operate using some kind of communication network. There has been a significant research interest in this field in recent years, driven by a trend of constructing decentralized and large scale control systems. The focus of the research on networked control systems has often been on the interactions between the control and the communication aspects such as random time delays, packet loss and rate limitations.

One way to study the fundamental aspects of networked control systems is to consider a model where a plant is controlled over a communication channel, as depicted in Figure 1.6. The problem of controlling a plant over a communication channel is closely connected to the communication problem. In fact, since control in this case requires some kind of communication, it can be argued that a solution to the control problem also solves a communication problem. This is done for example in [14], where control theory was used to design coding schemes and prove some capacity bounds.

Time delays are critical to the stability and performance of control systems. Therefore, the classical results of information theory can not, due to their asymptotic nature, be directly used to solve control problems with communication constraints. Nevertheless, information theory can for example be useful for obtaining performance bounds. It has been successfully used in combination with control theory to analyze optimal control strategies [2] and to find fundamental limitations of performance both for feedback control [34] and for disturbance attenuation using side information [35].

Stabilization over Channel

A lot of attention has been directed at finding necessary and sufficient conditions for when an unstable linear plant can be stabilized over a communication channel. By now, it is well-known that the intrinsic entropy rate of the plant is crucial in this aspect. In discrete time, this quantity is defined as

$$\mathcal{H}_G = \sum_i \max\left\{0, \log_2 |\lambda_i(G)|\right\}, \tag{1.3}$$

where $\lambda_i(G)$ is the ith pole of the plant G. The intrinsic entropy rate depends only on the unstable poles and can be thought of as the amount of information generated by the plant.

A necessary and sufficient condition for stabilization to be possible over a digital error-free channel is that $\mathcal{R} > \mathcal{H}_G$. This statement, called the data-rate theorem, has been shown both in deterministic [37, 57] and stochastic [38] settings. A thorough exposition of control with rate constraints is given in the review paper [39].

For noisy channels, the situation is a bit more complicated. For discrete channels, $\mathcal{C} > \mathcal{H}_G$ is a necessary and sufficient condition for almost sure asymptotic stabilizability [36]. This is, however, not generally true for other stability notions such as mean square stability. For this reason, the concept of any-time capacity has been proposed as an alternative to the Shannon capacity, in order to characterize moment stabilizability for control over noisy channels [45].

The SNR Framework

Mean square stability is, however, easier to characterize in the special case of control of a linear plant with Gaussian noise over an AWGN channel. In this context, the capacity is often expressed in terms of the Signal-to-Noise Ratio (SNR) of the channel. The associated problems are generally tractable using stochastic control theory, which makes the so-called SNR framework attractive. This was for example demonstrated in [7] and [51] where stabilizability was expressed in terms of the SNR, under different assumptions on plant noise and the controller structure. Specifically, it was shown that in some circumstances, the condition $\mathcal{C} > \mathcal{H}_G$ is actually necessary and sufficient for mean square stabilizability.

Using this framework also makes it relatively simple to design linear time-invariant controllers. Because of this, the SNR framework can be useful for applications such as power control in mobile communication systems [52]. Despite the relative simplicity of the SNR approach, the results obtained using this framework can sometimes be used to draw conclusions about and design controllers for other communication limitations such as rate limitations [49, 53] or packet drops [54].

1.3 Mathematical Preliminaries

This section provides the mathematical notation and conventions that will be used. A number of mathematical definitions and theorems, which are necessary for the development of the results presented in this thesis, are also presented. Some of this material may appear non-standard for many readers, but it is not necessary to understand all of the mathematical details in order to follow the main points of the thesis. The main reason for the necessity of this theory stems from the fact that the solutions to some of the optimization problems presented later on will in general have non-rational transfer functions.

Basic Notation

The logarithm with base b is denoted by $\log_b$. The natural logarithm is simply denoted log.

The real numbers are denoted by $\mathbb{R}$ and the complex numbers by $\mathbb{C}$. The open unit disk, $\{z \in \mathbb{C} : |z| < 1\}$, is denoted by $\mathbb{D}$. Its closure is denoted by $\overline{\mathbb{D}}$ and its boundary, the unit circle, by $\mathbb{T}$.

For a matrix $A \in \mathbb{C}^{m\times n}$, the rank, trace, determinant, transpose, conjugate and conjugate transpose are denoted by $\operatorname{rank} A$, $\operatorname{tr} A$, $\det A$, A^T, $\overline{A}$ and A^*, respectively. The matrix A is Hermitian if and only if $A = A^*$. The singular value decomposition of A is given by $A = U\Sigma V^*$, where $U \in \mathbb{C}^{m\times r}$, $\Sigma \in \mathbb{C}^{r\times r}$, $V \in \mathbb{C}^{n\times r}$ and $r = \min\{m,n\}$. Moreover, $U^*U = V^*V = I$ and Σ is diagonal with diagonal elements $\sigma_k \geq 0$, $k = 1 \dots r$, called singular values, satisfying $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$. The largest singular value of A is accordingly denoted $\sigma_1(A)$.

For a matrix $A \in \mathbb{C}^{m\times n}$ with $r = \min\{m,n\}$, define the Nuclear norm

$$\|A\|_* = \operatorname{tr}\sqrt{A^*A} = \sum_{i=1}^{r} \sigma_i$$

and the Frobenius norm

$$\|A\|_F = \sqrt{\operatorname{tr}\,(A^*A)} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}.$$

Transfer Matrices and Function Spaces

Given a sequence $\{x(k)\}_{k=0}^{\infty}$, the z-transform $F(z)$ is defined as

$$F(z) = \sum_{k=0}^{\infty} f(k) z^{-k}.$$

An LTI system can be represented by its transfer matrix, which is the z-transform of its impulse response. For this reason, complex functions will often be referred to as transfer matrices (in the general matrix-valued case) or as transfer functions (to emphasize that they are scalar).

A rational transfer matrix $X(z)$ is said to be proper if $\lim_{z\to\infty} X(z)$ exists and is bounded. If $\lim_{z\to\infty} X(z) = 0$, then $X(z)$ is said to be strictly proper. The space of all rational and proper transfer matrices with real coefficients is denoted by $\mathcal{R}$.

A non-rational transfer matrix $X(z)$ is said to be proper if the mapping $z \mapsto X(1/z)$ is analytic at 0. It is strictly proper if it is proper and $\lim_{z\to\infty} X(z) = 0$.

The arguments of transfer matrices will often be omitted when they are clear from the context. Equalities and inequalities involving functions evaluated on $\mathbb{T}$ are to be interpreted as holding almost everywhere on $\mathbb{T}$. That is, the subset of $\mathbb{T}$ in which the (in)equality does not hold is of measure zero.

For matrix-valued functions $X(z), Y(z)$ defined on $\mathbb{T}$, define

$$\langle X, Y\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}\left(X(e^{i\omega})^* Y(e^{i\omega})\right) d\omega$$

and the norms

$$\begin{aligned}
\|X\|_1 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left\|X(e^{i\omega})\right\|_* \, d\omega \\
\|X\|_2 &= \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \|X(e^{i\omega})\|_F^2 \, d\omega} \\
\|X\|_\infty &= \operatorname*{ess\,sup}_{\omega} \sigma_1\left(X(e^{i\omega})\right).
\end{aligned}$$

In the case when $X(z)$ is scalar, these definitions correspond to

$$\begin{aligned}
\|X\|_p &= \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|X(e^{i\omega})\right|^p d\omega\right)^{1/p}, \quad 1 \le p < \infty \\
\|X\|_\infty &= \operatorname*{ess\,sup}_{\omega} \left|X(e^{i\omega})\right|.
\end{aligned}$$

Definition 1.1—Lebesgue space
For $p = 1, 2, \infty$, the Lebesgue space $\mathcal{L}_p$ is defined as the space of matrix-valued functions $X(z)$, defined on $\mathbb{T}$, that satisfy $\|X\|_p < \infty$. The subspace $\mathcal{RL}_p$ consists of all real, rational and proper transfer matrices with no poles on $\mathbb{T}$. □

Note that $\mathcal{L}_\infty \subset \mathcal{L}_2 \subset \mathcal{L}_1$.

EXAMPLE 1.1
It holds that

$$e^z \in \mathcal{L}_\infty, \quad \frac{z^3}{(z-2)(2z-1)} \in \mathcal{L}_\infty, \quad \frac{z^2}{(z-2)(2z-1)} \in \mathcal{RL}_\infty$$

because these functions are bounded on $\mathbb{T}$. Note that transfer functions in $\mathcal{L}_p$ may be non-proper. □

For X such that $z \mapsto X(1/z)$ is defined on $\mathbb{D}$ (that is, functions defined outside the closed unit disk) define

$$X_r(z) = X(rz), \quad r > 1.$$

DEFINITION 1.2—HARDY SPACE
For $p = 1, 2, \infty$, the Hardy space $\mathcal{H}_p$ is defined as the space of matrix-valued functions $X(z)$ such that $z \mapsto X(1/z)$ is analytic on $\mathbb{D}$ and

$$\sup_{r>1} \|X_r\|_p < \infty.$$

The subspace $\mathcal{RH}_p$ consists of all real, rational, stable and proper transfer matrices. □

Note that $\mathcal{H}_\infty \subset \mathcal{H}_2 \subset \mathcal{H}_1$. Note also that $\mathcal{H}_p$ can be viewed as a closed subspace of $\mathcal{L}_p$ due to Fatou's Theorem, which says that if $X \in \mathcal{H}_p$, then $\hat{X} = \lim_{r \to 1^+} X_r$ exists almost everywhere on $\mathbb{T}$ and $\hat{X} \in \mathcal{L}_p$. Moreover, if $X \in \mathcal{H}_p$ and $X \in \mathcal{L}_q$, where $0 < p \leq \infty$ and $0 < q \leq \infty$, then it holds that $X \in \mathcal{H}_q$ [21]. In the following, when a function in $\mathcal{H}_p$ is evaluated on $\mathbb{T}$, it is to be understood as the limit $\hat{X}$.

REMARK 1.1
The convention in mathematics is to define $\mathcal{H}_p$ functions to be analytic on $\mathbb{D}$ [44]. Definition 1.2 (and other definitions that will follow) follows the convention in the control community and is related the definition of the z-transform with negative exponents of z [1]. Consequently, a rational transfer matrix is said to be stable if it has all of its poles in $\mathbb{D}$. □

EXAMPLE 1.2

$$X \in \mathcal{H}_2 \iff X(z) = \sum_{k=0}^{\infty} x_k z^{-k} \text{ and } \sum_{k=0}^{\infty} |x_k|^2 < \infty.$$

□

The following function will be used to define the Nevanlinna and Smirnov function classes. For $x \geq 0$, define

$$\log^+(x) = \max(\log(x), 0).$$

DEFINITION 1.3—NEVANLINNA CLASS
For scalar functions, the Nevanlinna class $\mathcal{N}$ is defined as all functions X such that $z \mapsto X(1/z)$ is analytic on $\mathbb{D}$ and

$$\sup_{r>1} \frac{1}{2\pi} \int_{-\pi}^{\pi} \log^+ |X_r(e^{i\omega})| d\omega < \infty.$$

A matrix-valued $X \in \mathcal{N}$ if and only if all its elements are in $\mathcal{N}$. □

DEFINITION 1.4—SMIRNOV CLASS
For scalar functions, the Smirnov class $\mathcal{N}^+$ is defined as all functions $X \in \mathcal{N}$ that satisfy

$$\lim_{r\to 1} \int_{-\pi}^{\pi} \log^+ |X_r(e^{i\omega})| d\omega = \int_{-\pi}^{\pi} \log^+ |X(e^{i\omega})| d\omega.$$

A matrix-valued $X \in \mathcal{N}^+$ if and only if all of its elements are in $\mathcal{N}^+$. □

EXAMPLE 1.3
The transfer function of a PI controller is not in $\mathcal{L}_p$ or $\mathcal{H}_p$ since it has a pole on the unit circle. However, it is of class $\mathcal{N}^+$. For example,

$$\frac{1}{z-1} \in \mathcal{N}^+.$$

As an example of a function that is in $\mathcal{N}$ but not $\mathcal{N}^+$, consider

$$\exp\left(\frac{z+1}{z-1}\right) \in \mathcal{N}.$$

□

The following lemma from [21] establishes some relationships between the introduced function classes.

LEMMA 1.1
It holds that

$$\mathcal{H}_p \subset \mathcal{N}^+ \subset \mathcal{N}.$$

Furthermore, $X \in \mathcal{H}_p$ if and only if $X \in \mathcal{N}^+$ and $X \in \mathcal{L}_p$. That is,

$$\mathcal{H}_p = \mathcal{N}^+ \cap \mathcal{L}_p.$$

□

For more details about the $\mathcal{L}_p, \mathcal{H}_p, \mathcal{N}$ and $\mathcal{N}^+$ function classes, the reader may consult standard textbooks such as [44, 21, 65].

Inner and Outer functions

Inner and outer functions are generalizations of all-pass and minimum phase functions, respectively.

Definition 1.5—Inner function
An inner function is a function $X \in \mathcal{H}_\infty$ such that

$$X(e^{i\omega})^* X(e^{i\omega}) = I.$$

$X \in \mathcal{H}_\infty$ is said to be co-inner if $X(z)^T$ is inner. □

It follows that a scalar function $X(z)$ is inner if and only if $|X(e^{i\omega})| = 1$.

Example 1.4
The functions

$$X(z) = \frac{z+2}{2z+1}, \quad Y(z) = \exp\left(\frac{z+1}{z-1}\right),$$

are inner. Note that $Y(z)$ is not well-defined at $z = 1$, but that $|Y(z)| = 1$ for almost all $z \in \mathbb{T}$. □

Definition 1.6—Outer function
The square matrix-valued function X is said to be outer if and only if $X \in \mathcal{N}^+$ and

$$\det(X(z)) = c \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i\omega}+z}{e^{i\omega}-z} \log\varphi(e^{i\omega}) d\omega\right\},$$

where c is a constant with $|c| = 1$ and φ is a non-negative function on $\mathbb{T}$ with $\log\varphi \in \mathcal{L}_1$. □

Note that outer functions are proper since they belong to $\mathcal{N}^+$. An equivalent definition of outer functions is that X is outer if and only if $X \in \mathcal{N}^+$, $\det X$ is not identically zero and $X^{-1} \in \mathcal{N}^+$ [28]. From this, a simpler condition can be deduced in the scalar, rational case: A scalar and rational function X is outer if and only if it satisfies the following conditions:

- $X(z) = p(z)/q(z)$, where p and q are polynomials of the same degree.
- All zeros and poles of X are in $\overline{\mathbb{D}}$.

In the literature, outer functions are commonly required to be in $\mathcal{H}_p$. If that definition is used then rational outer functions must have all their poles in $\mathbb{D}$.

EXAMPLE 1.5
The function

$$\frac{z+a}{z+b}$$

is outer if and only if $|a| \leq 1$ and $|b| \leq 1$. □

For non-square matrix-valued functions, the concept of row outer functions will be used. The definition is slightly more complicated:

DEFINITION 1.7—ROW OUTER FUNCTION
An $m \times n$ matrix-valued function $X \in \mathcal{N}^+$ is said to be row outer if it has full row rank almost everywhere on $\mathbb{T}$, that is,

$$\operatorname{rank} X(e^{i\omega}) = m$$

and

$$X(0)^* X(0) \geq Y(0)^* Y(0)$$

for any $m \times n$ matrix-valued function $Y \in \mathcal{N}^+$ with

$$Y(e^{i\omega})^* Y(e^{i\omega}) = X(e^{i\omega})^* X(e^{i\omega}).$$

X is said to be co-outer if X^T is row outer. □

The definition of row outer functions can be found in [26], which also contains the following facts:

LEMMA 1.2
Suppose X is a square function. Then X is outer if and only if it is row outer. Moreover, if X is outer, then $X^{-1} \in \mathcal{N}^+$ is outer. □

The first statement says that the definition of row outer is a generalization of the definition of outer to the case of non-square functions. The second statement can be used to prove the following lemma.

LEMMA 1.3
Suppose $Y \in \mathcal{N}^+$ is square and outer, $X \in \mathcal{N}^+$, and that $Y^{-1}X \in \mathcal{L}_p$. Then $Y^{-1}X \in \mathcal{H}_p$. □

PROOF
$Y^{-1} \in \mathcal{N}^+$ by Lemma 1.2. It is easy to verify that the product of two $\mathcal{N}^+$ functions is $\mathcal{N}^+$. The statement then follows from Lemma 1.1. □

The following lemma appears as Corollary 4.7 in [21].

LEMMA 1.4
Suppose $X \in \mathcal{H}_p$ and $X^{-1} \in \mathcal{H}_q$ for some $0 < p \leq \infty$, $0 < q \leq \infty$. Then X is outer. □

Factorizations

A number of different factorization techniques will be used in this thesis. The definition of the singular value decomposition can be extended to transfer matrices: A singular value decomposition of $X \in \mathcal{L}_p$ is defined pointwise on $\mathbb{T}$ as

$$X(e^{i\omega}) = U(e^{i\omega})\Sigma(e^{i\omega})V(e^{i\omega}),$$

where $U, V \in \mathcal{L}_\infty$ and $\Sigma \in \mathcal{L}_p$.

The standard coprime factorization [65] will also prove useful.

THEOREM 1.1—COPRIME FACTORIZATION
Suppose that the scalar transfer function $G(z)$ has a state-space realization

$$G = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right],$$

where (A, B) is stabilizable and (C, A) is detectable. Then there exists a coprime factorization $G = NM^{-1}$, where $N, M \in \mathcal{RH}_\infty$ satisfies the Bezout identity $VM + UN = 1$ for some $V, U \in \mathcal{RH}_\infty$. □

The following theorem from [26] shows that every function in $\mathcal{N}^+$ can be written as the product of an inner and an outer function.

THEOREM 1.2—INNER-OUTER FACTORIZATION
For every $X \in \mathcal{N}^+$ there exists an inner-outer factorization $X = X_i X_o$ where $X_i \in \mathcal{H}_\infty$ is inner and $X_o \in \mathcal{N}^+$ is row outer. Similarly there exists a co-inner-outer factorization $X = X_{co} X_{ci}$ where $X_{co} \in \mathcal{N}^+$ is co-outer and $X_{ci} \in \mathcal{H}_\infty$ is co-inner. □

EXAMPLE 1.6
An inner-outer factorization of $1/(z+2)$ is given by

$$\frac{1}{z+2} = \frac{2z+1}{z(z+2)} \cdot \frac{z}{2z+1},$$

where the first factor is inner and the second is outer. □

REMARK 1.2
If $X \in \mathcal{H}_p$, then the outer factor $X_o \in \mathcal{H}_p$. □

Many of the results in this thesis depend on spectral factorization of non-rational transfer matrices. The following theorem provides conditions for when this is possible. It is the matrix version of a classical theorem by Szegő, which is in turn a generalization of the Fejér-Riesz Factorization Theorem [59, 56].

THEOREM 1.3—SPECTRAL FACTORIZATION
Suppose that $Y \in \mathcal{L}_1$ is Hermitian, positive definite on $\mathbb{T}$ and satisfies

$$\int_{-\pi}^{\pi} \log\det Y(e^{i\omega})\, d\omega > -\infty. \tag{1.4}$$

Then there exists a square outer $X \in \mathcal{H}_2$ such that

$$Y(e^{i\omega}) = X(e^{i\omega})X(e^{i\omega})^*.$$

□

REMARK 1.3
If Y satisfies $Y(e^{-i\omega}) = \overline{Y(e^{i\omega})}$ then it is always possible to choose $X(z)$ such that

$$X(z) = \sum_{k=0}^{\infty} x_k z^{-k}, \quad x_k \in \mathbb{R}.$$

Then X also satisfies $X(e^{-i\omega}) = \overline{X(e^{i\omega})}$. In the rational case, this condition corresponds to X having numerator and denominator polynomials with real coefficients. □

The condition (1.4) is known as the Paley-Wiener condition. The following result, stated in Theorem 17.17 and under 17.19 in [44] will be useful for showing that this condition is satisfied.

LEMMA 1.5
Suppose $X \in \mathcal{N}$ is scalar and not identically zero. Then $\log|X| \in \mathcal{L}_1$. □

In the matrix case, the following lemma will instead be used for the same purpose.

LEMMA 1.6
Suppose that $m \leq n$ and that the $m \times n$ transfer matrix $X \in \mathcal{H}_p$, $p \in \{1, 2, \infty\}$, is row outer. Then the singular values of X satisfy

$$\log\sigma_k \in \mathcal{L}_1, \quad k = 1 \ldots m.$$

□

PROOF

By Theorem 1.2 there exists a co-inner-outer factorization $X = X_{co}X_{ci}$. Since X_{co} has full column rank on $\mathbb{T}$ it cannot have more columns than rows, and since X is row outer X_{co} cannot have fewer rows than columns. Thus X_{co} is $m \times m$.

By Lemma 1.2, X_{co}^T is outer since it is row outer and square. The determinant of an outer function is by definition also outer, so $\det X_{co}$ is outer and hence $\det X_{co} \in \mathcal{N}^+$. Applying Lemma 1.5 gives that $\log|\det X_{co}| \in \mathcal{L}_1$.

For the singular values of X, it holds that

$$\begin{aligned}\sum_{k=1}^{m} \log \sigma_k &= \frac{1}{2} \log \prod_{k=1}^{m} \sigma_k^2 = \frac{1}{2} \log \det X X^* \\ &= \frac{1}{2} \log \det X_{co} X_{ci} X_{ci}^* X_{co}^* = \frac{1}{2} \log \det X_{co} X_{co}^* \\ &= \log|\det X_{co}| \in \mathcal{L}_1.\end{aligned}$$

Furthermore, $\sigma_k \in \mathcal{L}_1$ since $X \in \mathcal{H}_p$. Because $\log \sigma_k < \sigma_k$ it holds that

$$\int_{-\pi}^{\pi} \log \sigma_k \, d\omega < \int_{-\pi}^{\pi} \sigma_k \, d\omega < \infty, \quad k = 1 \ldots m$$

Since the sum of the logarithms is $\mathcal{L}_1$ and every term has an integral bounded from above, it follows that the integral of every term also must be bounded from below. That is,

$$\int_{-\pi}^{\pi} \log \sigma_k \, d\omega > -\infty, \quad k = 1 \ldots m$$

and hence $\log \sigma_k \in \mathcal{L}_1$, $k = 1 \ldots m$ □

It is fairly easy to show that the product of two $\mathcal{H}_2$ functions is a $\mathcal{H}_1$ function. The converse, from Theorem 17.10 in [44], will also be useful.

LEMMA 1.7

Every $X \in \mathcal{H}_1$ is a product $X = YZ$, where $Y, Z \in \mathcal{H}_2$. □

2

Real-Time Coding for a Noisy Channel

2.1 Introduction

The problem studied in this chapter lies in the intersection of communication, estimation and control theory. Its main interpretation is as a real-time coding problem, but it is also related to Wiener filtering and feed-forward control.

Figure 2.1 gives a schematic representation of the problem. A source signal is measured with additive noise by an encoder. The spectra of both signals are known. The encoder filters and encodes information about its measurements and transmits over an AWN channel to a decoder that forms an estimate of the source signal after it has gone through the filter P. The objective is to find a causal encoder-decoder pair that together minimize the estimation error variance.

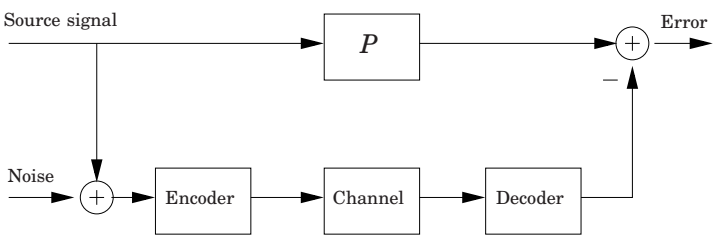

Figure 2.1 Illustration of the problem in this chapter. The encoder and the decoder should be designed to minimize the error. Nominally, P is a fixed time delay but it could be any linear time-invariant (LTI) filter. The channel is an AWN channel.

This problem is a real-time coding problem with partially observed source and a quadratic distortion measure. Other interpretations can also be made (see below). In most cases, P is a fixed time delay, but it could be any stable LTI filter.

In the literature, the problem of coding with a partially observed source often includes the possibility of noise at the receiver as well. The main motivation for excluding that possibility here is the fact, noted in [63], that the optimality of an encoder-decoder design is independent of additive and independent zero-mean noise at the receiver.

A possible application where this problem is relevant is the transmission of speech in mobile communication. The source signal to be estimated at the receiver is the speech signal. The delay constraint is based on the acceptable latency and the noise is any background sound present at the microphone. Speech coders are typically designed using source- and sink-specific models based on the assumption that the speaker's voice is the only input. The effect of other sounds can therefore be substantial, affecting the user experience negatively. Incorporating filtering in the encoder could perhaps help in this aspect [4, 24].

Outline and Main Results

The rest of this section will present the relevant previous research and alternative interpretations of the problem.

The optimal factorization idea, which provides the basis for all the results in this thesis, is introduced in Section 2.2. Thereafter, Theorem 2.1 shows that the jointly optimal linear encoder and decoder can be found in the scalar-valued case by first minimizing a functional of the form

$$\|R - X\|_2^2 + \frac{1}{\sigma^2}\|X\|_1^2 \tag{2.1}$$

over $X \in \mathcal{H}_2$, for a given $R \in \mathcal{L}_\infty$, and then performing a spectral factorization. The solution is also generalized to handle frequency weighted error criteria and non-white channel noise.

In Section 2.3, the case with vector-valued signals and parallel AWN channels is treated. Theorem 2.2 shows that the jointly optimal linear encoder and decoder can be obtained by minimizing the matrix version of (2.1) and performing a matrix spectral factorization.

In Section 2.4, the scalar-valued case is considered for an AWN channel with noiseless feedback. It is shown by Theorem 2.3 that the jointly optimal linear encoder and decoder can be obtained by minimizing another convex functional and performing a spectral factorization. By example, it is shown that channel feedback may improve the performance of linear coding systems.

The restriction to linear encoder and decoder may result in suboptimal solutions. Nevertheless, the linear solution to any problem instance will provide an upper bound to the minimum distortion possible given the SNR and the signal spectras. Moreover, the proposed design methods are relatively simple and computationally feasible.

Previous Research

A family of similar problems was considered in [12, 10] as a means to design optimal scalar feedback quantization schemes. In particular, it is possible to solve some instances of the problems considered in sections 2.2 and 2.4 by by using the solutions in [10] (in the former case using a fixed feedback filter that is set to zero). There are, however, differences in that [12, 10] only consider the scalar-valued case with zero delay tolerance and no noise at the source. That is, $P = 1$ and $G = 0$. It is not obvious to the author if the solutions can be easily modified to accommodate for other delays and for the addition of noise at the source.

Partially Observed Source The problem of coding with a partially observed source was first considered for the Gaussian case with additive noise and mean squared error distortion in [13]. It was shown that the problem is asymptotically equivalent to, and can thus be reduced to, the fully observed case and that an optimal encoder generally has a structure consisting of an optimal estimator followed by optimal encoding for a noise-free source. This structural result was generalized to the non-gaussian and finite time horizon cases in [63]. The problem was further studied in [3], where it was noted that in the case of white source noise, the criterion in the reduced problem is given by the conditional expectation of the original criterion given the encoder input. It was pointed out in [61] that the equivalence in [13] actually was proved for the one-shot problem as well. Moreover, it was shown that the reduction to the non-remote problem follows from a general "disconnection principle". It would thus be possible to formulate a finite horizon version of the problem studied in this chapter as a vectorized one-shot rate distortion problem. The causality requirement would give structural constraints on the solution in form of lower-triangular encoder and decoder mappings and the delay constraint would determine the structure of the performance criterion. However it is not clear how this insight translates into a practical method for actually solving the problem.

Real-Time Coding Even without consideration of a partially observed source, there is not yet a satisfactory solution to the real-time coding problem, although a number of structural results have been obtained. The optimal causal source coder for a white source has been found to be

memoryless [40]. For a Markov source of order k and delay constraint d, an optimal real-time source coder only needs to use the last $\max\{k, d+1\}$ source symbols plus the current state of the decoder. No such memory bound is given, however, when the encoder does not have access to the decoder state [62]. Joint source-channel coding with noiseless feedback was considered for finite alphabet sources in [58] where it was demonstrated that feedback is useful in general, but that coding is useless for a class of channels with a certain symmetry property. Conditions have also been found for when optimal performance can be achieved without coding (even when allowing coding systems with arbitrary delay) [22].

Real-time source coding for a partially observed source has been considered in [5]. The structural results of [62, 58] were extended to the partially observed case in [64], which also presented a separation result for the linear-quadratic Gaussian case similar to the one in [13]. A method for design of optimal real-time coding systems for noisy channels was presented in [32] using noisy feedback and in [33] without feedback. However, there seems to be no method for efficient numerical application of the solution.

Alternative Interpretations

As a Feed-Forward Problem It is possible to make the following interpretation of the problem in Figure 2.1: The source signal is a disturbance that will affect some system where a controller (the decoder) can compensate. The controller has a remote sensor that measures the disturbance and transmits information to the controller over the channel. In this interpretation P may also model any dynamics that the disturbance passes through on the way. A similar interpretation was presented in [61].

A similar problem setup was studied in [35], where information theory was used to find a lower bound on the reduction of entropy rate made possible by side information communicated through a channel with some given capacity. Under stationarity assumptions, this was used to derive a lower bound, which is a generalization of Bode's integral equation, on a sensitivity-like function. The problem architectures is quite similar to the one here, but there are some important differences: The main one is that [35] gives performance bounds for a general communication channel, while a constructive method is studied here for a specific channel model. Comparisons of the results are difficult due to differing performance metrics: Here, the variance of the error is minimized. In [35], a lower bound is achieved on the integral of the logarithm of a sensitivity-like function. Further differences in [35] include the placement of a feedback controller at the receiving end, error-free observations of the source, and the restriction of P into fixed time delays.

Connection to Wiener Filter The problem of estimating a signal that is measured with additive noise, under a mean squared error criterion, is solved by the Wiener filter [60]. The filter is usually obtained by solving the Wiener-Hopf equations, but can also be expressed in the frequency domain as the stable filter K that minimizes

$$\left\|(z^{-k} - K)F\right\|_2^2 + \|KG\|_2^2, \tag{2.2}$$

where k is the allowed time delay and F and G are transfer functions that represent the frequency characteristics of the signal of interest and the measurement noise, respectively.

It is possible to interpret the problem in Figure 2.1 as a distributed Wiener filtering problem, where the estimation is separated into two different locations. The communication channel is used to model the communication constraint between the two locations. This interpretation is strengthened by the result in Section 2.2 that the solution to the problem (without channel feedback) is based on minimizing

$$\|(P - K)F\|_2^2 + \|KG\|_2^2 + \frac{1}{\sigma^2}\|K\begin{bmatrix} F & G \end{bmatrix}\|_1^2.$$

where σ^2 is the SNR and P a transfer function (possibly a pure time delay). This shows that the cost is similar to the one in (2.2), but that the communication channel between the two parts of the filter induces an additional term, which is a weighted 1-norm of K. The size of the additional term scales inversely with the SNR of the channel. For infinite SNR, the cost is equal to that in the Wiener filtering problem. When the channel has feedback, the additional term takes another, slightly more complex, form.

2.2 Optimal Linear Encoder and Decoder

In this section, the problem will be described in detail and solved for the single-input, single-output (SISO) case. The solutions to the multi-input, multi-output (MIMO) case, considered in Section 2.3, and the case when channel feedback is available, considered in Section 2.4, are more general and can be used to solve the problem in this section. This simpler case is, however, presented first in order to make the ideas behind the solution appear more clearly, as the solutions in the next two sections are more complex.

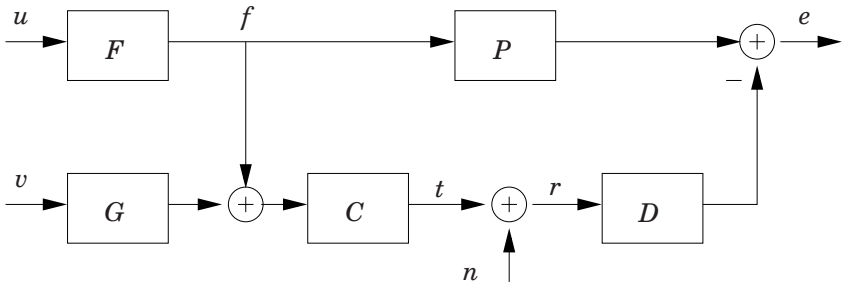

Figure 2.2 Structure of the system. With F, G and P given, the objective is to design C and D such that the stationary variance of e is minimized.

Problem Formulation and Assumptions

The detailed structure of the problem is shown in Figure 2.2. The input signals u, v, w are assumed to be mutually independent scalar white noise sequences with zero mean and variance 1. Every block in the figure represents a linear, time-invariant, single-input, single-output system described by a corresponding transfer function.

The transfer functions $F \in \mathcal{H}_\infty$ and $G \in \mathcal{H}_\infty$ are given shaping filters for the source signal and the measurement noise, respectively. It is assumed that

$$\exists \varepsilon > 0 \text{ such that } FF^* + GG^* \geq \varepsilon \text{ on } \mathbb{T}, \tag{2.3}$$

which implies that F and G have no common zeros on the unit circle (if F and G are rational then the two statements are equivalent). The given transfer function $P \in \mathcal{H}_\infty$ represents the dynamics that the source signal undergoes between the points where it is measured and where it is to be estimated. The encoder $C \in \mathcal{H}_2$ and the decoder $D \in \mathcal{H}_2$ are the design variables. The causality is imposed by their restriction to $\mathcal{H}_2$. Note that it is not assumed that any of these transfer functions are rational.

The problem is studied in a stationary setting. Since all transfer functions are assumed to be stable, the initial values can, without loss of generality, be assumed to be zero. The objective is to minimize the stationary variance of the estimation error e,

$$\lim_{k\to\infty} \mathrm{E}(e(k)^2).$$

The communication channel is an AWN[1] channel with SNR $\sigma^2 > 0$.

[1]Since only linear solutions are considered, it does not matter if the channel noise or the other inputs are Gaussian or not. Linear solutions may, of course, be more or less suboptimal depending on the distributions.

The SNR constraint is assumed to hold in stationarity, that is

$$\lim_{k\to\infty} \mathrm{E}(t(k)^2) \le \sigma^2.$$

By expressing e and t in terms of the transfer functions in Figure 2.2, the objective function and the SNR constraint can be written as

$$J(C,D) = \|(P - DC)F\|_2^2 + \|DCG\|_2^2 + \|D\|_2^2 = \lim_{k\to\infty} \mathrm{E}(e(k)^2) \qquad (2.4)$$

and

$$\sigma^2 \ge \|CF\|_2^2 + \|CG\|_2^2 = \lim_{k\to\infty} \mathrm{E}(t(k)^2), \qquad (2.5)$$

respectively. The problem can thus be formulated as follows.

Problem 2.1

$$\underset{C,D\in\mathcal{H}_2}{\text{minimize}}\, J(C,D)$$

subject to (2.5). □

Initial Observations

The objective function $J(C,D)$ is clearly not convex in the pair (C,D) due to the appearance of the product DC. A simple example for which the minimum can be calculated analytically shows that there are several solutions to the problem:

Example 2.1—Static Case
Suppose that $P = 1$, $F = F_0 > 0$, and $G = G_0 \ge 0$. Expressing C and D as

$$C(z) = \sum_{k=0}^{\infty} C_k z^{-k}, \qquad D(z) = \sum_{k=0}^{\infty} D_k z^{-k},$$

it is easy to see that all coefficients except the first should be zero. For such C and D, the objective reduces to

$$J(C,D) = F_0^2(1 - D_0C_0)^2 + G_0^2D_0^2C_0^2 + D_0^2$$

and the SNR constraint becomes

$$(F_0^2 + G_0^2)C_0^2 \le \sigma^2.$$

This problem can be solved by standard methods. The solution is given by

$$C(z) = (-1)^k\sqrt{\frac{\sigma^2}{F_0^2 + G_0^2}}, \qquad D(z) = (-1)^k\frac{F_0^2}{\sigma^2 + 1}\sqrt{\frac{\sigma^2}{F_0^2 + G_0^2}}$$

where $k \in \{0, 1\}$. The minimum value is

$$J(C, D) = \frac{F_0^2}{\sigma^2 + 1} + \frac{\sigma^2 F_0^2 G_0^2}{(\sigma^2 + 1)(F_0^2 + G_0^2)}.$$

If $G_0 = 0$, the solution in this case coincides with the one in the example given for the AWGN channel in [35].

Note that the SNR constraint is active at optimality. It will be seen that this is always true, except for the trivial and non-interesting cases when either $F = 0$ or $P = 0$ (when the minimum clearly is obtained by $C = D = 0$). □

The general problem, when F, G and P are dynamic, is significantly more difficult. In this section, optimal LTI encoders and decoders will be characterized and it will be shown how they can be obtained by solving a convex optimization problem and performing a spectral factorization.

First, it will be shown that the SNR constraint (2.5) can be written as

$$\|CH\|_2^2 \leq \sigma^2, \tag{2.6}$$

where the function H has some nice properties.

LEMMA 2.1
Suppose that $F, G \in \mathcal{H}_\infty$ and that (2.3) holds. Then there exists $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$ such that

$$HH^* = FF^* + GG^* \text{ on } \mathbb{T}. \tag{2.7}$$

□

PROOF
By (2.3) and Theorem 1.3 there exists an outer function $H \in \mathcal{H}_2$ such that (2.7) holds. Since $F, G \in \mathcal{H}_\infty$ it follows that $H \in \mathcal{H}_\infty$. Moreover, it follows from (2.3) that $\left\|H^{-1}\right\|_\infty \leq 1/\sqrt{\varepsilon}$ and since H is outer it then follows from Lemma 1.3 that $H^{-1} \in \mathcal{H}_\infty$. □

Optimal Factorization

The optimal factorization approach presented here will provide the basis for the solution of all the problems solved in this thesis. The idea is to consider the product DC as given and then to find an optimal factorization of this product. The factorization gives an analytical expression for the cost in terms of the product, which means that optimization of the objective

may then be performed over the product. When an optimal product is found, the optimality conditions from the optimal factorization can then be applied to find optimal C and D.

Introduce $K = DC$. The objective can then be written as

$$\|(P-K)F\|_2^2 + \|KG\|_2^2 + \|D\|_2^2. \tag{2.8}$$

Note that the first two terms are constant for any given K. The minimum over C and D, given K, is thus obtained by minimizing the third term in (2.8) subject to (2.6) and $K = DC$. This minimization problem is the *optimal factorization problem*.

The interpretation is that for any given product of the encoder and decoder, the contribution to the objective of the signals that pass through both the encoder and the decoder is not affected by the choice of the factors C and D — only their product matters. The channel noise, however, only passes through the decoder, which means that D (and implicitly C since $C = D^{-1}K$) should be chosen to minimize the impact of the channel noise on the objective.

The optimal factorization problem and its solution will appear many times in this thesis, in slightly different versions depending on the problem. The solution to the present version is given by the following lemma.

LEMMA 2.2—OPTIMAL FACTORIZATION, SISO CASE
Suppose that $\sigma^2 > 0$, $K \in \mathcal{H}_1$ and $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$. Then the optimization problem

$$\underset{C,D \in \mathcal{H}_2}{\text{minimize}} \|D\|_2^2 \tag{2.9}$$

subject to

$$K = DC, \quad \|CH\|_2^2 \leq \sigma^2 \tag{2.10}$$

attains the minimum value

$$\frac{1}{\sigma^2}\|KH\|_1^2. \tag{2.11}$$

Moreover, if K is not identically zero then $C, D \in \mathcal{H}_2$ are optimal if and only if $DC = K$ and

$$|C|^2 = \frac{\sigma^2}{\|KH\|_1}\left|\frac{K}{H}\right| \text{ on } \mathbb{T}. \tag{2.12}$$

If $K = 0$, then the minimum is achieved by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies (2.10). □

PROOF
If $K = 0$ the proof is trivial, so assume that K is not identically zero. Then C is not identically zero and $D = KC^{-1}$. Then (2.10) and Cauchy-Schwarz's inequality gives

$$\begin{aligned}\|D\|_2^2 &= \left\|KC^{-1}\right\|_2^2 \\ &\geq \frac{\|CH\|_2^2}{\sigma^2}\left\|KC^{-1}\right\|_2^2 \\ &\geq \frac{1}{\sigma^2}\left\langle |CH|, \left|KC^{-1}\right|\right\rangle^2 \\ &= \frac{1}{\sigma^2}\|KH\|_1^2\end{aligned}$$

This shows that (2.11) is a lower bound on (2.9). Equality holds if and only if $|KC^{-1}|$ and $|CH|$ are proportional on the unit circle and $\|CH\|_2^2 = \sigma^2$. It is easily verified that this is equivalent to (2.12). Thus, C and D achieve the lower bound if and only if $D = KC^{-1}$ and (2.12) holds.

It remains to show existence of such $C, D \in \mathcal{H}_2$. Note that $KH^{-1} \in \mathcal{H}_1$ is not identically zero. Hence, by Lemma 1.5, $\log|KH^{-1}| \in \mathcal{L}_1$. It follows by Theorem 1.3 that there exists an outer $C \in \mathcal{H}_2$ that satisfies (2.12). Thus

$$\left\|KC^{-1}\right\|_2^2 = \frac{1}{\sigma^2}\|KH\|_1^2 < \infty,$$

so $KC^{-1} \in \mathcal{L}_2$. Since $K \in \mathcal{H}_1$ and $C \in \mathcal{H}_2$ is outer it follows from Lemma 1.3 that $D = KC^{-1} \in \mathcal{H}_2$. □

REMARK 2.1
Optimal D will satisfy

$$|D|^2 = \frac{\|KH\|_1}{\sigma^2}|KH| \text{ on } \mathbb{T}. \tag{2.13}$$

Apparently, the magnitudes of C and D are both proportional to the square root of the magnitude of K. This provides some intuition to why the minimum value depends on the 1-norm of K. □

REMARK 2.2
The existence part of Lemma 2.2 shows that one specific solution, where C is outer, can be obtained. By using the freedom available in spectral factorization, it is possible to obtain other solutions, for example by changing

the sign of both C and D, or by instead choosing D to be outer. More generally, in the rational case, any non-minimum phase zeros or time delays could be located in C or D. □

For any given product of the encoder and decoder, an optimal encoder and decoder are specified by (2.12) and (2.13), respectively. Their transfer functions can be obtained by a spectral factorization of $|KH^{-1}|$.

Equivalent Convex Problem

A heuristic solution to the main problem could now, for example, be to use the Wiener filter, which minimizes (2.2), for the factorization, but this is not optimal. An optimal solution is obtained by inserting the minimum value of $\|D\|_2^2$ into (2.8) and minimizing over K. That is, minimizing

$$\varphi(K) = \|(P-K)F\|_2^2 + \|KG\|_2^2 + \frac{1}{\sigma^2}\|K\begin{bmatrix} F & G \end{bmatrix}\|_1^2, \tag{2.14}$$

which is a convex problem. That this procedure in fact solves the main problem is shown by the following theorem, which is the main result of this section.

THEOREM 2.1
Suppose that $\sigma^2 > 0$, $F, G, P \in \mathcal{H}_\infty$ and that (2.3) holds. Then the optimization problem

$$\underset{C,D\in\mathcal{H}_2}{\text{minimize}}\, J(C,D) \tag{2.15}$$

subject to

$$\|CF\|_2^2 + \|CG\|_2^2 \le \sigma^2 \tag{2.16}$$

attains a minimum value that is equal to the minimum of the convex optimization problem

$$\underset{K\in\mathcal{H}_2}{\text{minimize}}\, \varphi(K), \tag{2.17}$$

which is attained by a unique minimizer.

Moreover, suppose $K \in \mathcal{H}_2$ is a solution to (2.17). If K is not identically zero, then C and D solve (2.15) subject to (2.16) if and only if $C \in \mathcal{H}_2$, $D = KC^{-1} \in \mathcal{H}_2$ and

$$|C|^2 = \frac{\sigma^2}{\|K\begin{bmatrix} F & G \end{bmatrix}\|_1} \frac{|K|}{\sqrt{|F|^2+|G|^2}} \text{ on } \mathbb{T}. \tag{2.18}$$

If $K = 0$, then the solution to (2.15) and (2.16) is given by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies (2.16). □

PROOF

By Lemma 2.1, there exists $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$ such that

$$HH^* = FF^* + GG^* \text{ on } \mathbb{T}. \tag{2.19}$$

and (2.16) is equivalent to

$$\|CH\|_2^2 \leq \sigma^2. \tag{2.20}$$

Define the sets

$$\begin{aligned}\Theta &= \{(C,D) : C, D \in \mathcal{H}_2,\ (2.20)\}\\ \Theta(K) &= \{(C,D) : C, D \in \mathcal{H}_2,\ (2.20),\ K = DC\}.\end{aligned}$$

Then the infimum of $J(C,D)$ subject to (2.16) can be written

$$\begin{aligned}\inf_{C,D\in\Theta} J(C,D) &= \inf_{K\in\mathcal{H}_1} \inf_{C,D\in\Theta(K)} J(C,D)\\ &= \inf_{K\in\mathcal{H}_1} \left(\|(P-K)F\|_2^2 + \|KG\|_2^2 + \inf_{C,D\in\Theta(K)} \|D\|_2^2 \right)\\ &= \inf_{K\in\mathcal{H}_1} \|(P-K)F\|_2^2 + \|KG\|_2^2 + \frac{1}{\sigma^2}\|KH\|_1^2\\ &= \inf_{K\in\mathcal{H}_1} \varphi(K) \end{aligned} \tag{2.21}$$

The first equality is true by Lemma 1.7. The second equality follows because the first two terms in $\inf_{C,D\in\Theta(K)} J(C,D)$ are constant. The third equality follows from application of Lemma 2.2 to perform the inner minimization. The final equality follows from (2.19).

It will now be shown that the minimum in (2.21) is attained by a unique $K \in \mathcal{H}_2$. Completion of squares gives that

$$\begin{aligned}\varphi(K) &= \|(P-K)F\|_2^2 + \|KG\|_2^2 + \frac{1}{\sigma^2}\|KH\|_1^2\\ &= \|PF\|_2^2 + \|KH\|_2^2 - 2\,\mathrm{Re}\langle PFF^*, KHH^{-1}\rangle + \frac{1}{\sigma^2}\|KH\|_1^2\\ &= \|PFF^*H^{-*} - KH\|_2^2 + \frac{1}{\sigma^2}\|KH\|_1^2 + \text{const.}\end{aligned}$$

Let $X = KH \in \mathcal{H}_1$ and $R = PFF^*H^{-*} \in \mathcal{L}_\infty$. Minimizing $\varphi(K)$ over $K \in \mathcal{H}_1$ is then equivalent to minimizing

$$\psi(X) = \|R - X\|_2^2 + \frac{1}{\sigma^2}\|X\|_1^2 \tag{2.22}$$

over $X \in \mathcal{H}_1$. In the latter problem, it is sufficient to consider X such that $\psi(X) \leq \psi(0) = \|R\|_2^2$. That is, only X satisfying

$$\begin{aligned}\|X\|_2 = \|R - X - R\|_2 &\leq \|R - X\|_2 + \|R\|_2 \\ &\leq \sqrt{\psi(X)} + \|R\|_2 \leq 2\|R\|_2 \stackrel{\text{def}}{=} r.\end{aligned}$$

Now, in the weak topology, $\psi(X)$ is lower semicontinuous on $\mathcal{L}_2$ and the set $\{X : \|X\|_2 \leq r\}$ is compact. This proves the existence of a minimum. The minimum is unique since $\psi(X)$ is strictly convex. Moreover, since $\|X\|_2 \leq r$, it is sufficient to minimize over $X \in \mathcal{H}_2$ instead of $\mathcal{H}_1$.

Suppose now that $X \in \mathcal{H}_2$ minimizes $\psi(X)$. From $H^{-1} \in \mathcal{H}_\infty$ it follows that $K = H^{-1}X \in \mathcal{H}_2$ attains the infimum value in (2.21) and that this value is equal to the minimum of (2.17).

Since the minimum is attained in (2.21) and, by Lemma 2.2, there exists $(C, D) \in \Theta$ such that $J(C, D) = \varphi(K)$, it follows that the minimum of (2.15) subject to (2.16) is attained.

Finally, the optimality condition (2.18) follows from the application of Lemma 2.2, using that $|H| = \sqrt{|F|^2 + |G|^2}$. □

Remark 2.3
Preliminary results suggest that the optimal K will have a non-rational transfer function [23]. This implies that optimal C and D are non-rational as well. A system with non-rational transfer function cannot be implemented with finite memory, meaning that some approximation has to be performed.

This may explain the assumption that was made in [62], that the encoder has access to the decoder state. Without this type of feedback, it would be impossible to bound the encoder's memory requirement. □

Remark 2.4
$\varphi(K)$ is convex, and $\varphi(\overline{K}) = \varphi(K)$. Thus,

$$\varphi\left(\frac{K + \overline{K}}{2}\right) \leq \frac{1}{2}\left(\varphi(K) + \varphi(\overline{K})\right) = \varphi(K).$$

Since the optimal K is unique, this shows that the minimizing K satisfies $K(e^{-i\omega}) = \overline{K(e^{i\omega})}$. Thus, C can be chosen to have this property as well, meaning that C can be approximated by a rational function with real coefficients. The same holds for D.

Similar remarks can be made regarding the corresponding optimization problems presented in sections 2.3, 2.4, 3.2 and 3.3. □

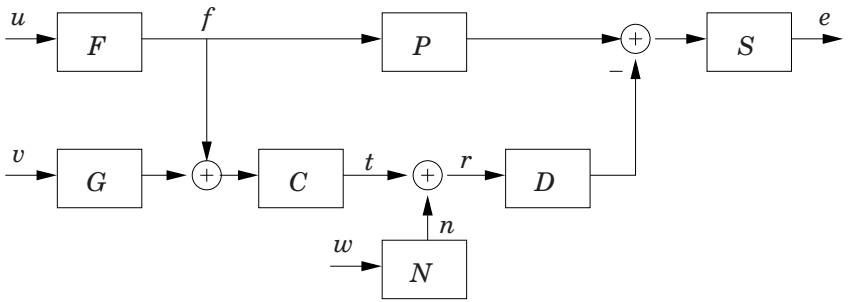

Figure 2.3 Extended version of the problem in Figure 2.2. Here, the error is frequency weighted by S and the channel noise is shaped by N.

REMARK 2.5
It was noted in Remark 2.2 that the optimal factorization problem can have multiple solutions. To clarify, the optimal K is unique but there are multiple factorizations of K into C and D that achieve the minimum value of $J(C,D)$. □

It is noted that the solution of the problem essentially amounts to minimizing the sum of a 2-norm and a 1-norm of the decision variable. The 2-norm represents the cost in the Wiener filter problem, and the 1-norm represents the contribution of the channel noise to the error variance. The SNR σ^2 determines the relative importance of the two terms. For small SNR, the optimal K will have small magnitude since the channel noise dominates the transmitted signal. As the SNR becomes larger, the magnitude of K will become larger, and it will approach the Wiener filter in the limit when the SNR goes to infinity.

Frequency Weighting and Non-White Channel Noise

Consider the extended problem structure illustrated in Figure 2.3. The difference from the original problem is that the error signal is frequency weighted by $S \in \mathcal{H}_\infty$ and that the channel noise is colored by $N \in \mathcal{H}_\infty$. This means that the channel now has memory. In a feed-forward context, S could represent the sensitivity function of a given closed-loop system that is affected by the disturbance.

The objective function for the extended problem is

$$J_{ext}(C,D) = \|S(P - DC)F\|_2^2 + \|SDCG\|_2^2 + \|SDN\|_2^2 .$$

The SNR constraint is not changed. Assuming that $S^{-1}, N^{-1} \in \mathcal{H}_\infty$, it is straightforward to modify Lemma 2.2 and Theorem 2.1 to cover the

extended problem, so the proofs are omitted. Solving the optimal factorization problem results in the equivalent convex optimization problem

$$\underset{K \in \mathcal{H}_2}{\text{minimize}} \; \|S(P-K)F\|_2^2 + \|SKG\|_2^2 + \frac{1}{\sigma^2} \|SKN \begin{bmatrix} F & G \end{bmatrix}\|_1^2 \tag{2.23}$$

and the optimality condition

$$|C|^2 = \frac{\sigma^2}{\|SKN \begin{bmatrix} F & G \end{bmatrix}\|_1} \left| \frac{SKN}{\sqrt{|F|^2 + |G|^2}} \right| \text{ on } \mathbb{T}. \tag{2.24}$$

Numerical Solution

A procedure for numerical solution of the (extended) problem will now be outlined.

1. The first step is to solve the optimization problem (2.23). Since this problem is infinite-dimensional and an analytical solution seems to be out of reach, this is done approximately using a finite basis representation of K and sum approximations of the integrals. The approximated problem can then be cast as a quadratic program with second-order cone constraints.
2. Use a finite basis approximation $A(\omega)$ of CC^*, for example using the parametrization

$$A(\omega) = A_0 + \sum_{k=1}^{N_c} A_k \left(e^{ki\omega} + e^{-ki\omega} \right) \tag{2.25}$$

   and fit $A(\omega)$ to the right hand side of (2.24), for example by minimizing the mean squared deviation.
3. Perform a spectral factorization of $A(\omega)$, choosing C as the stable and minimum phase spectral factor.
4. Let $D = KC^{-1}$.

The numerical solution will be illustrated by an example.

Example 2.2
Consider the problem with $P = z^{-2} + 0.5z^{-7}$, $F = 1/(z - 0.5)$, $G = 1$, $S = N = 1$ and $\sigma = 1$. This P can be given the interpretation that the coding system has two opportunities to estimate each sample of the source signal. First after 2 samples delay, and then again after 7 samples delay. The second estimate does not count as much as the first because of the smaller coefficient.

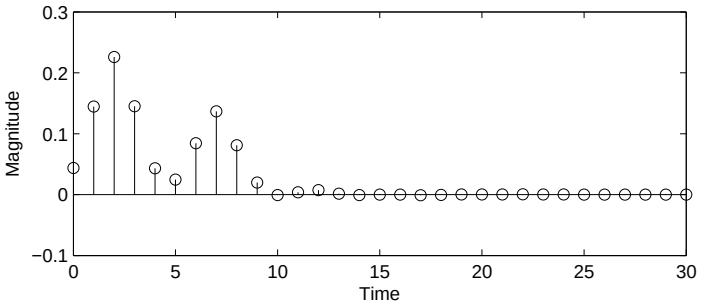

Figure 2.4 Impulse response of K, the product of the encoder and the decoder, in Example 2.2.

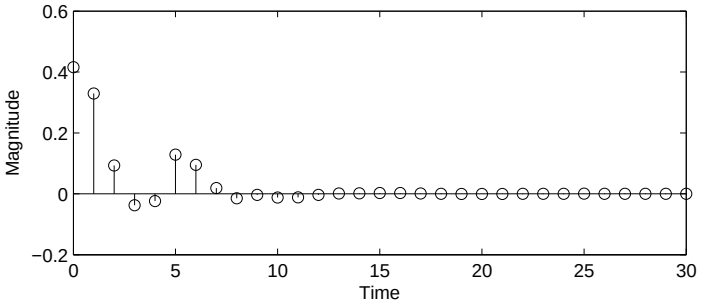

Figure 2.5 Impulse response of the encoder C in Example 2.2.

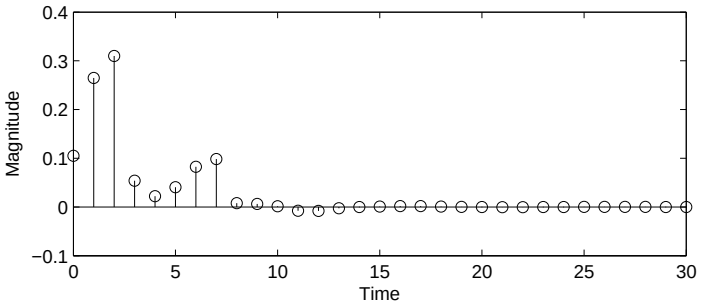

Figure 2.6 Impulse response of the decoder D in Example 2.2.

The functional (2.22) was minimized with $X = KH$ parametrized as an FIR filter:

$$X(z) = \sum_{k=0}^{N_x} x_k z^{-k},$$

with $N_x = 30$. The minimization was implemented in Matlab using Yalmip [30] and SeDuMi [55]. The grid distance 0.001 was used for numerical computation of the integrals. Figure 2.4 shows impulse response of the obtained K. Note that the two peaks in the impulse response correspond to the non-zero coefficients of P.

The spectrum of C was parametrized as in (2.25) with $N_c = 30$. The spectrum coefficients were obtained by solving a least squares problem. Finally, C was obtained through spectral factorization and $D = KC^{-1}$. The impulse responses of C and D are shown in figures 2.5 and 2.6, respectively. The obtained value for this problem is 1.00. □

2.3 The MIMO Case

In this section, the results in the previous section will be generalized to the case of vector-valued signals.

Problem Formulation and Assumptions

Consider again the system in Figure 2.2, but with the modification that all signals are vector-valued and all systems are MIMO with corresponding transfer matrices. The number of elements in signal f is denoted n_f and so forth. Matrix dimensions are not explicitly stated in this section except when necessary. It is generally assumed that all matrices are of appropriate size. In addition to all the assumptions made previously, it is now also assumed that:

1. The number of elements in the signals satisfy

$$n_t \geq \min\{n_f, n_e\}, \text{ where } C \text{ is } n_t \times n_f \text{ and } D \text{ is } n_e \times n_t. \qquad (2.26)$$

   If the number of channels n_t would be smaller than n_f and n_e, then the product DC could not have full rank. This means that optimization over $K = DC$ would have to include a rank constraint, which is very difficult to handle.

2. The inequality (2.3) is replaced by the matrix version

$$\exists \varepsilon > 0 \text{ such that } FF^* + GG^* \geq \varepsilon I \text{ on } \mathbb{T}. \qquad (2.27)$$

3. The communication channel consists of n_t parallel AWN channels. The SNR constraint (2.5) is replaced by the power constraint

$$\lim_{k\to\infty} \mathrm{E}(t(k)^T t(k)) \leq \sigma^2.$$

4. The input signals u, v, n have identity covariance matrices. Note that this is non-restrictive for f or the measurement noise, since these are scaled by F and G, respectively. But it does mean that only the case when all the channel components have noise with the same variance is considered.

The objective is to minimize

$$J(C,D) = \|(P - DC)F\|_2^2 + \|DCG\|_2^2 + \|D\|_2^2 = \lim_{k\to\infty} \mathrm{E}(e(k)^T e(k))$$

subject to

$$\sigma^2 \geq \|CF\|_2^2 + \|CG\|_2^2 = \lim_{k\to\infty} \mathrm{E}(t(k)^T t(k)). \tag{2.28}$$

Accordingly, the problem in this section is defined as follows.

Problem 2.2

$$\underset{C,D\in\mathcal{H}_2}{\text{minimize}}\, J(C,D)$$

subject to (2.28). □

In the norm notation, the objective and the constraint are written in the same way as in the SISO case. It will be seen that the equivalent convex problem also looks the same as in the SISO case. The optimality condition will, however, be more complicated.

Optimal Factorization

The solution to the optimal factorization problem is given by the following lemma. This problem is much more difficult to solve in the MIMO case.

Lemma 2.3—Optimal Factorization, MIMO Case
Suppose that $\sigma^2 > 0$, $K \in \mathcal{H}_1$, $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$ and that (2.26) and (2.27) hold. Then the optimization problem

$$\underset{C,D\in\mathcal{H}_2}{\text{minimize}}\, \|D\|_2^2$$

subject to

$$K = DC, \quad \|CH\|_2^2 \leq \sigma^2$$

attains the minimum value $\frac{1}{\sigma^2}\|KH\|_1^2$.

Moreover, suppose K is not identically zero and let $K = K_i K_o$ be an inner-outer factorization and $K_o H = U_o \Sigma V^*$ be a singular value decomposition. Then $C, D \in \mathcal{H}_2$ are optimal if and only if

$$K = DC, \quad \|CH\|_2^2 = \sigma^2, \quad DD^* = \frac{\|KH\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*.$$

If $K = 0$ then the minimum is achieved by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies $\|CH\|_2^2 \leq \sigma^2$. □

PROOF

If $K = 0$ the proof is trivial, so assume that K is not identically zero. Then neither C nor D are identically zero and $\alpha = \|CH\|_2 > 0$. Now, suppose that C, D are feasible and that $\alpha < \sigma$. Then

$$C_\alpha = \frac{\sigma}{\alpha} C, \quad D_\alpha = \frac{\alpha}{\sigma} D$$

are feasible and $\|D_\alpha\|_2 < \|D\|_2$. Hence, a necessary condition for optimality is that $\|CH\|_2^2 = \sigma^2$.

The remainder of this proof is divided into three parts. First, the dual problem is considered. Then, it is shown that there is a saddle point and the optimality criteria are derived. Finally, existence of the solution is proven by construction.

DUAL PROBLEM: In order to avoid dealing with analyticity constraints associated with $\mathcal{H}_2$, the search will temporarily be relaxed to $C, D \in \mathcal{L}_2$. Later, it will be shown that there are $C, D \in \mathcal{H}_2$ that satisfy the derived optimality criteria. For $\lambda \geq 0$ and $\Phi \in \mathcal{L}_\infty$, where Φ is matrix-valued, introduce the Lagrangian

$$\begin{aligned}
&L(C, D, \lambda, \Phi) = \\
&= \|D\|_2^2 + \lambda \left(\|CH\|_2^2 - \sigma^2 \right) - \langle \operatorname{Re} \Phi, \operatorname{Re} DC - K \rangle - \langle \operatorname{Im} \Phi, \operatorname{Im} DC - K \rangle \\
&= \|D\|_2^2 + \lambda \left(\|CH\|_2^2 - \sigma^2 \right) - \operatorname{Re} \langle \Phi, DC - K \rangle \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \|D\|_F^2 + \lambda \|CH\|_F^2 - \operatorname{Re} \operatorname{tr} \left(\Phi^* (DC - K) \right) d\omega - \lambda \sigma^2 \qquad (2.29)
\end{aligned}$$

The integrand in (2.29) can be rewritten, by a completion of squares, as

$$\begin{aligned}
&\|D\|_F^2 + \lambda \|CH\|_F^2 - \operatorname{Re} \operatorname{tr} \left(C\Phi^* D - \Phi^* K \right) \\
&= \left\| D - \frac{1}{2} \Phi C^* \right\|_F^2 + \lambda \|CH\|_F^2 - \frac{1}{4} \|C\Phi^*\|_F^2 + \operatorname{Re} \operatorname{tr} \left(\Phi^* K \right) \\
&= \left\| D - \frac{1}{2} \Phi C^* \right\|_F^2 + \operatorname{tr} \left[C \left(\lambda H H^* - \frac{1}{4} \Phi^* \Phi \right) C^* \right] + \operatorname{Re} \operatorname{tr} \left(\Phi^* K \right) \qquad (2.30)
\end{aligned}$$

Only the first term depends on D. The contribution of this term is minimized by

$$D = \frac{1}{2} \Phi C^*. \qquad (2.31)$$

If (2.31) holds, then L only depends on C through the second term of (2.30). Pointwise minimization of that term gives

$$\inf_{C\in\mathcal{L}_2} \operatorname{tr}\left[C\left(\lambda HH^* - \frac{1}{4}\Phi^*\Phi\right)C^*\right] = \begin{cases} 0, & 4\lambda HH^* \geq \Phi^*\Phi \text{ on } \mathbb{T} \\ -\infty, & \text{otherwise.} \end{cases}$$

Moreover, the third term in (2.30) can be written

$$\operatorname{tr}(\Phi^*K) = \operatorname{tr}(\Phi^*DC) = \frac{1}{2}\operatorname{tr}(C\Phi^*\Phi C^*) = \frac{1}{2}\|\Phi C^*\|_F^2.$$

Thus, $\operatorname{tr}(\Phi^*K)$ is real and non-negative, and

$$\inf_{C,D\in\mathcal{L}_2} L = \begin{cases} \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} \operatorname{tr}(\Phi^*K)\,d\omega - \lambda\sigma^2, & 4\lambda HH^* \geq \Phi^*\Phi \text{ on } \mathbb{T} \\ -\infty, & \text{otherwise.} \end{cases}$$

Introduce

$$\Psi = \frac{1}{2\sqrt{\lambda}}\Phi H^{-*}.$$

Then the dual problem can be written as

$$\underset{\lambda\geq 0,\Psi\in\mathcal{L}_\infty}{\text{maximize}} \ \frac{2\sqrt{\lambda}}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}(\Psi^*KH)\,d\omega - \lambda\sigma^2$$

subject to

$$\Psi^*\Psi \leq I \text{ on } \mathbb{T}. \tag{2.32}$$

The dual function is concave in λ. Letting $\lambda = 0$ gives the value 0. Since $\operatorname{tr}(\Psi^*KH) \geq 0$ there exists $\lambda > 0$ that gives a positive value, so the optimal λ is given by the first-order condition

$$\left(\frac{1}{\sigma^2 2\pi}\int_{-\pi}^{\pi} \operatorname{tr}(\Psi^*KH)\,d\omega\right)^2 = \lambda,$$

obtained by differentiation with respect to λ. With this λ the dual problem simplifies to

$$\underset{\Psi\in\mathcal{L}_\infty}{\text{maximize}} \ \frac{1}{\sigma^2}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}(\Psi^*KH)\,d\omega\right)^2 \tag{2.33}$$

subject to (2.32).

The integrand in (2.33) will now be maximized pointwise. Recall that $KH = K_i K_o H = K_i U_o \Sigma V^*$ and denote the number of rows of K_o by m. Then Σ is diagonal with diagonal elements σ_k, $k = 1 \dots m$. Since K is $n_e \times n_f$ the rank of K is not greater than $\min\{n_e, n_f\}$ and thus

$$m \leq \min\{n_e, n_f\}. \tag{2.34}$$

K_o is row outer by definition and H is outer by Lemma 1.4. It follows that $K_o H$ is row outer and thus has full row rank. It follows that the singular values are positive: $\sigma_k > 0$, $k = 1 \dots m$. Since $K_o H$ is wide (it has $n_f \geq m$ columns) it follows that U_o is square and thus unitary.

Define $U = K_i U_o$ and $\widetilde{\Psi} = U^* \Psi V$. Then it follows from (2.32) and $UU^* \leq I$ that

$$\widetilde{\Psi}^* \widetilde{\Psi} = V^* \Psi^* U U^* \Psi V \leq V^* \Psi^* \Psi V \leq V^* V = I.$$

Using $\widetilde{\Psi}$, an upper bound can be obtained for the integrand in (2.33):

$$\begin{aligned} \sup_{\Psi^*\Psi \leq I} \operatorname{tr}(\Psi^* K H) &= \sup_{\Psi^*\Psi \leq I} \operatorname{tr}(\Psi^* U \Sigma V^*) = \sup_{\Psi^*\Psi \leq I} \operatorname{tr}(V^* \Psi^* U \Sigma) \\ &\leq \sup_{\widetilde{\Psi}^*\widetilde{\Psi} \leq I} \operatorname{tr}\left(\widetilde{\Psi}^* \Sigma\right) = \sum_{k=1}^{m} \sup_{|\widetilde{\Psi}_{kk}| \leq 1} \sigma_k \widetilde{\Psi}_{kk} = \sum_{k=1}^{m} \sigma_k \end{aligned}$$

The supremum is achieved if and only if $\widetilde{\Psi} = I$. Therefore, the upper bound is achieved by Ψ if and only if $U^* \Psi V = I$ and $\Psi^* \Psi \leq I$. The set of Ψ satisfying these conditions can be parametrized as:

$$\Psi = UV^* + \Psi_0 = K_i U_o V^* + \Psi_0 \tag{2.35}$$

$$I \geq \Psi^* \Psi, \tag{2.36}$$

where Ψ_0 satisfies

$$0 = U^* \Psi_0 V = U_o^* K_i^* \Psi_0 V. \tag{2.37}$$

Pre-multiplying (2.37) with U_o gives the equivalent condition

$$K_i^* \Psi_0 V = 0. \tag{2.38}$$

Choosing, for example, $\Psi_0 = 0$ gives $\Psi = UV^*$, which attains the upper bound. Hence, the value of the dual problem is

$$\begin{aligned} \max_{\Psi^*\Psi \leq I} \frac{1}{\sigma^2} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr}(\Psi^* K H)\, d\omega \right)^2 &= \frac{1}{\sigma^2} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr}(V U^* U \Sigma V^*)\, d\omega \right)^2 \\ &= \frac{1}{\sigma^2} \|KH\|_1^2. \end{aligned}$$

The maximizing dual variables are given by

$$\Phi = 2\sqrt{\lambda}\Psi H^* = 2\sqrt{\lambda}(K_i U_o V^* + \Psi_0)H^* \tag{2.39}$$

where Ψ_0 is such that (2.35), (2.36) and (2.38) hold, and

$$\lambda = \left(\frac{1}{\sigma^2}\|KH\|_1\right)^2. \tag{2.40}$$

SADDLE POINT: It will now be shown that there is a saddle point, which implies that the duality gap is zero.

In the following, assume that (2.35), (2.36), (2.38), (2.39) and (2.40) hold. Then λ and Φ are dual feasible. The point (C, D, λ, Φ) is a saddle point if and only if $C, D \in \mathcal{H}_2$ are primal feasible,

$$\lambda\left(\|CH\|_2^2 - \sigma^2\right) = 0 \tag{2.41}$$

and

$$L(C, D, \lambda, \Phi) = \inf_{\widehat{C},\widehat{D}\in\mathcal{H}_2} L(\widehat{C}, \widehat{D}, \lambda, \Phi). \tag{2.42}$$

The saddle point conditions imply that $\|CH\|_2 = \sigma$ since $\lambda > 0$ and that $D = \frac{1}{2}\Phi C^*$ as it was seen earlier that this follows from minimization of the Lagrangian.

Suppose that C, D satisfy $K = DC$ and $D = \frac{1}{2}\Phi C^*$. Then

$$\begin{aligned} DD^* = \frac{1}{2}DC\Phi^* = \frac{1}{2}K\Phi^* &= \sqrt{\lambda}K_i K_o H(V U_o^* K_i^* + \Psi_0^*) \\ &= \sqrt{\lambda}(K_i U_o \Sigma U_o^* K_i^* + K_i U_o \Sigma V^* \Psi_0^*). \end{aligned}$$

Clearly, DD^* and $K_i U_o \Sigma U_o^* K_i^*$ are Hermitian. Accordingly,

$$A = K_i U_o \Sigma V^* \Psi_0^*$$

must be Hermitian. Now, by (2.38),

$$\begin{aligned} AK_i = K_i U_o \Sigma V^* \Psi_0^* K_i = 0 \\ \Rightarrow 0 = AK_i = A^* K_i = \Psi_0 V \Sigma U_o^* K_i^* K_i = \Psi_0 V \Sigma U_o^*. \end{aligned}$$

Hence, $A = 0$ and

$$DD^* = \sqrt{\lambda}K_i U_o \Sigma U_o^* K_i^* = \frac{\|KH\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*. \tag{2.43}$$

Now, suppose instead that $C, D \in \mathcal{H}_2$ satisfy $K = DC$, $\|CH\|_2 = \sigma$ and (2.43). Then C, D are primal feasible and (2.41) is satisfied. Moreover,

$$\begin{aligned} L(C, D, \lambda, \Phi) = \|D\|_2^2 &= \frac{\sqrt{\lambda}}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr}\,(K_i U_o \Sigma U_o^* K_i^*)\, d\omega \\ &= \frac{\sqrt{\lambda}}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr}\,(\Sigma)\, d\omega = \frac{1}{\sigma^2} \|KH\|_1^2, \end{aligned}$$

so (2.42) holds and thus the saddle point conditions are satisfied. Since these assumptions and the saddle point conditions imply each other, they are equivalent.

To conclude, it has been shown that (C, D, λ, Φ) is a saddle point, which implies that $C, D \in \mathcal{H}_2$ achieve the claimed optimal value, if and only if $K = DC$, $\|CH\|_2^2 = \sigma^2$ and (2.43) holds.

Existence of Solution: Define $M = \sqrt{\lambda} U_o \Sigma U_o^* \in \mathcal{L}_1$, which is Hermitian with real diagonal. Recall that $K_o H$ is row outer with singular values $\sigma_k > 0$, $k = 1 \ldots m$. From this and Lemma 1.6 it follows that $\log \sigma_k \in \mathcal{L}_1$. Since U_o is unitary it also follows that M is positive definite. Moreover,

$$\log \det M = \frac{m}{2} \log \lambda + \sum_{k=1}^{m} \log \sigma_k \in \mathcal{L}_1$$

Therefore, according to Theorem 1.3, there is an outer transfer matrix $D_o \in \mathcal{H}_2$ such that $M = D_o D_o^*$. Let $\widetilde{D} = K_i D_o \in \mathcal{H}_2$ and $\widetilde{C} = D_o^{-1} K_o$. Then

$$\begin{aligned} \widetilde{C} &= D_o^{-1} K_o H H^{-1} = D_o^{-1} U_o \Sigma V^* H^{-1} \\ &= D_o^{-1} U_o \Sigma U_o^* U_o V^* H^{-1} = \frac{1}{\sqrt{\lambda}} D_o^* U_o V^* H^{-1} \in \mathcal{L}_2 \end{aligned}$$

Since D_o is outer it follows from Lemma 1.3 that $\widetilde{C} \in \mathcal{H}_2$.

It can now be verified that $\widetilde{C}$ and $\widetilde{D}$ satisfy the optimality conditions:

$$\widetilde{D}\widetilde{C} = K_i D_o D_o^{-1} K_o = K_i K_o = K,$$

$$\begin{aligned} \left\| \widetilde{C} H \right\|_2^2 = \left\| D_o^{-1} K_o H \right\|_2^2 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr}\,\left(H^* K_o^* D_o^{-*} D_o^{-1} K_o H \right) d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{tr}\,\left(V \Sigma U_o^* M^{-1} U_o \Sigma V^* \right) d\omega \\ &= \frac{1}{\sqrt{\lambda} 2\pi} \int_{-\pi}^{\pi} \operatorname{tr}\,(\Sigma)\, d\omega = \sigma^2 \end{aligned}$$

and

$$\widetilde{D}\widetilde{D}^* = K_i D_o D_o^* K_i^* = \sqrt{\lambda} K_i U_o \Sigma U_o^* K_i^*.$$

If the rank of K does not equal n_t, then $\widetilde{C}$ and $\widetilde{D}$ are not of the required dimensions. $\widetilde{C}$ is $m \times n_f$ and $\widetilde{D}$ is $n_e \times m$, where, by (2.26) and (2.34), $m \leq \min\{n_e, n_f\} \leq n_t$. It is required that C is $n_t \times n_f$ and that D is $n_e \times n_t$. To solve this problem, let

$$D = \begin{bmatrix} \widetilde{D} & 0_{n_e \times n_t - m} \end{bmatrix} \in \mathcal{H}_2, \quad C = \begin{bmatrix} \widetilde{C} \\ 0_{n_t - m \times n_f} \end{bmatrix} \in \mathcal{H}_2.$$

Noting that $DC = \widetilde{D}\widetilde{C} = K$, that $\|CH\|_2 = \left\|\widetilde{C}H\right\|_2$ and that $DD^* = \widetilde{D}\widetilde{D}^*$ it is *finally* concluded that C, D are optimal. □

Remark 2.6
In the scalar case $K_i = U_o = 1$ and $\Sigma = |KH|$, so it is easily verified that the optimality conditions in that case are equivalent to those in Section 2.2. □

Equivalent Convex Problem

Just as in the SISO case, the solution to the optimal factorization problem can be used to find an equivalent convex problem. This problem looks exactly the same in the MIMO case as in the SISO case.

Theorem 2.2
Suppose that $\sigma^2 > 0$, $F, G, P \in \mathcal{H}_\infty$ and that (2.26) and (2.27) hold. Then the optimization problem

$$\underset{C,D \in \mathcal{H}_2}{\text{minimize}}\, J(C, D) \tag{2.44}$$

subject to

$$\|CF\|_2^2 + \|CG\|_2^2 \leq \sigma^2 \tag{2.45}$$

attains a minimum value that is equal to the minimum of the convex optimization problem

$$\underset{K \in \mathcal{H}_2}{\text{minimize}}\, \|(P - K)F\|_2^2 + \|KG\|_2^2 + \frac{1}{\sigma^2} \|K \begin{bmatrix} F & G \end{bmatrix}\|_1^2, \tag{2.46}$$

which is attained by a unique minimizer.

Moreover, suppose $K \in \mathcal{H}_2$ is a solution to (2.46). If K is not identically zero, then $C, D \in \mathcal{H}_2$ solve (2.44) subject to (2.45) if and only if

$$K = DC, \quad \|C \begin{bmatrix} F & G \end{bmatrix}\|_2^2 = \sigma^2, \quad DD^* = \frac{\|K \begin{bmatrix} F & G \end{bmatrix}\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*,$$

where K_i is defined by an inner-outer factorization $K = K_i K_o$ and U_o and Σ are given by a singular value decomposition $K_o H = U_o \Sigma V^*$, where $H \in \mathcal{H}_\infty$ satisfies $H^{-1} \in \mathcal{H}_\infty$ and $HH^* = FF^* + GG^*$.

If $K = 0$, then the solution to (2.44) and (2.45) is given by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies (2.45). □

PROOF

With the assumption (2.27), Lemma 2.1 holds in the matrix case as well. The rest of the proof is identical to the proof of Theorem 2.1, except that Lemma 2.3 is used instead of Lemma 2.2, with the obvious implications for the optimality conditions. □

REMARK 2.7

The assumption (2.26) may deserve some explanation. If there are too few communication channels relative to the dimensionality of f and e, the maximum rank of the product DC may be smaller than the smallest dimension of K. Then not all K would be realizable as a product of D and C, and a rank condition would have to be imposed on K in Theorem 2.2. In principle, this changes nothing, but the assumption is included in order to avoid formulating the solution in terms of an optimization problems that cannot be reliably solved. □

Numerical Solution

A procedure for numerical solution of the MIMO version of the problem will now be outlined.

1. The first step is to solve the optimization problem (2.46). An approximate solution can be obtained using a finite basis representation of K and sum approximations of the integrals. The approximated problem can then be cast as a quadratic program with second-order cone constraints.
2. Perform a matrix spectral factorization to obtain $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$ that satisfies $HH^* = FF^* + GG^*$ on $\mathbb{T}$.
3. Perform an inner-outer factorization to obtain $K_i K_o = K$.
4. Perform a singular value decomposition to obtain $U_o \Sigma V^* = K_o H$.

5. Use a finite basis approximation $A(\omega)$ of DD^*, for example using the parametrization

$$A(\omega) = A_0 + \sum_{k=1}^{N_c} A_k \left(e^{ki\omega} + e^{-ki\omega}\right)$$

and fit $A(\omega)$ to

$$\frac{\|K\,[F \quad G\,]\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*,$$

for example by minimizing the mean squared deviation.

6. Perform a spectral factorization of $A(\omega)$, choosing D_o as the stable and outer spectral factor.

7. Let $D = K_i D_o$ and $C = D_o^{-1} K_o$.

8. If C and D are of incorrect size, add rows of zeros to C and columns of zeros to D until they are of correct size.

2.4 Using Channel Feedback

In this section, the same problem as in sections 2.2 and 2.3 will be considered. A SISO setting is again considered, but this time under the assumption of channel feedback, as illustrated in Figure 2.7. The encoder now has two inputs: one is the same as before, and the other is a one-sample delayed version of the decoder input.

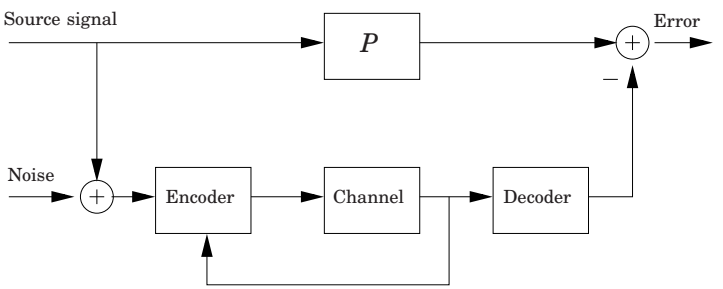

Figure 2.7 Illustration of the problem studied in this section. It is assumed that the channel has feedback, that is, the encoder has access to the signal received by the decoder.

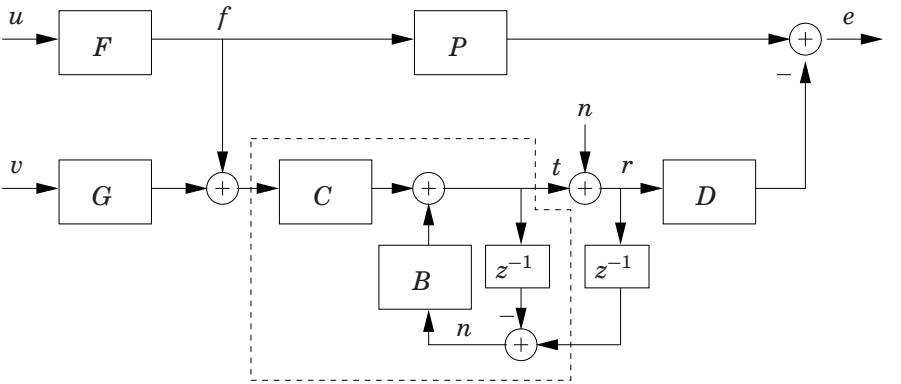

Figure 2.8 Structure of the system with channel feedback. With F, G and P given, the objective is to design B, C and D such that the stationary variance of e is minimized. The dashed box represents the encoder.

Problem Formulation and Assumptions

Make the same assumptions as in Section 2.2. Note that it is again assumed that all signals are scalar-valued and that all systems are SISO. The encoder can be parametrized in a number of ways. The structure and parametrization illustrated in Figure 2.8 will be used. Since the encoder remembers its past output, it can subtract t from r to obtain the channel noise n, delayed by one time step. This is the input to B, which represents the part of the encoder that processes the feedback signal. By linearity, this encoder structure can be assumed without loss of generality.

It will be assumed, for technical reasons, that B is rational. Since it must also be proper and stable, it follows that $B \in \mathcal{RH}_\infty$.

By expressing e and t in terms of the transfer functions in Figure 2.8, the objective function and the SNR constraint can be written as

$$J(B,C,D) = \|(P - DC)F\|_2^2 + \|DCG\|_2^2 + \left\|D(1 + Bz^{-1})\right\|_2^2 = \lim_{k\to\infty} \mathrm{E}(e(k)^2)$$

and

$$\sigma^2 \geq \|CF\|_2^2 + \|CG\|_2^2 + \|B\|_2^2 = \lim_{k\to\infty} \mathrm{E}(t(k)^2), \tag{2.47}$$

respectively. The problem in this section is thus:

Problem 2.3

$$\underset{B\in\mathcal{RH}_\infty, C,D\in\mathcal{H}_2}{\text{minimize}} \quad J(B,C,D)$$

subject to (2.47). □

Note that if $B = 0$, then the objective function and the SNR constraint are the same as those in Problem 2.1.

In contrast with the two previous section, it will be seen here that the minimum is not generally attained. However, an approximate solution, that gives a performance arbitrary close to the infimum, can be constructed. This shortcoming is mainly theoretical, since the numerical solution of Problems 2.1 and 2.2 is performed approximately anyway.

Optimal Factorization

Compared to before, the problem now has an additional variable in B. The optimal factorization approach is consequently modified to assume that, in addition to the product $K = DC$, B is also given. For feasible B, it holds that

$$\|B\|_2^2 \leq \sigma^2.$$

Actually, if $\|B\|_2^2 = \sigma^2$ then $C = 0$ and

$$J(B, 0, D) = \|PF\|_2^2 + \left\|D(1 + Bz^{-1})\right\|_2^2.$$

Then it is optimal to let $D = 0$. But $J(B, 0, 0)$ does not depend on B, so the same cost can be achieved with, for example, $B = 0$. Thus, it can be assumed without loss of generality that $\|B\|_2^2 < \sigma^2$.

By Lemma 2.1, there exists $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$ such that

$$HH^* = FF^* + GG^* \text{ on } \mathbb{T}. \tag{2.48}$$

and $\|C\begin{bmatrix} F & G \end{bmatrix}\|_2^2 = \|CH\|_2^2$. The set of feasible (C, D), parametrized by K and B, is thus defined as

$$\Theta_{C,D}(B, K) = \left\{ (C, D) : C, D \in \mathcal{H}_2, DC = K, \|CH\|_2^2 \leq \sigma^2 - \|B\|_2^2 \right\}.$$

The solution to the optimal factorization problem is given by the following lemma.

Lemma 2.4—Optimal Factorization, Channel Feedback Case
Suppose $K \in \mathcal{H}_1$, $B \in \mathcal{RH}_\infty$, $\|B\|_2^2 < \sigma^2$ and $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$. Then

$$\inf_{(C,D) \in \Theta_{C,D}(B,K)} \left\|D(1 + Bz^{-1})\right\|_2^2 \geq \frac{\left\|KH(1 + Bz^{-1})\right\|_1^2}{\sigma^2 - \|B\|_2^2}. \tag{2.49}$$

Suppose furthermore that $1 + Bz^{-1}$ has no zeros on $\mathbb{T}$. Then there exists $(C, D) \in \Theta_{C,D}(B, K)$ with $C, D \in \mathcal{H}_2$ and C outer, such that the minimum is attained and (2.49) holds with equality.

If K is not identically zero, then $(C, D) \in \mathcal{H}_2$ are optimal if and only if $DC = K$ and

$$|C|^2 = \frac{\sigma^2 - \|B\|_2^2}{\|KH(1 + Bz^{-1})\|_1} \left| \frac{K(1 + Bz^{-1})}{H} \right| \text{ on } \mathbb{T}. \tag{2.50}$$

If $K = 0$, then the minimum is achieved by $D = 0$ and any $C \in \mathcal{H}_2$ that satisfies $\|CH\|_2^2 \leq \sigma^2 - \|B\|_2^2$. □

PROOF

The proof of the lower bound is nearly identical to the first part in the proof of Lemma 2.2 and can be obtained by simply inserting the factor $(1+Bz^{-1})$ in the derivations. The optimality conditions follow in the same way.

The only significant difference lies in the existence part: It is necessary to verify the existence of $C \in \mathcal{H}_2$ and $D = KC^{-1} \in \mathcal{H}_2$ such that (2.50) holds. Note first that $K(1 + Bz^{-1})H^{-1} \in \mathcal{H}_1$. Moreover, $KH^{-1} \in \mathcal{H}_1$ is not identically zero. $1 + Bz^{-1} \in \mathcal{RH}_\infty$ is also not identically zero since B is causal. It then follows from Lemma 1.5 that

$$\log \left| KH^{-1}(1 + Bz^{-1}) \right| = \log \left| KH^{-1} \right| + \log \left| 1 + Bz^{-1} \right| \in \mathcal{L}_1$$

Hence, it follows from Theorem 1.3 that there is an outer function $C \in \mathcal{H}_2$ such that (2.50) holds.

Since $1 + Bz^{-1}$ is rational and has no zeros on $\mathbb{T}$, there exists $\varepsilon > 0$ such that $|1 + Bz^{-1}| \geq \varepsilon$ on $\mathbb{T}$. Thus,

$$\begin{aligned} \|D\|_2^2 &= \left\| KC^{-1} \right\|_2^2 \\ &= \frac{\left\| KH(1 + Bz^{-1}) \right\|_1}{\sigma^2 - \|B\|_2^2} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{KH}{1 + Bz^{-1}} \right| d\omega \\ &\leq \frac{\left\| KH(1 + Bz^{-1}) \right\|_1}{\sigma^2 - \|B\|_2^2} \frac{\|KH\|_1}{\varepsilon} < \infty, \end{aligned}$$

and so $D \in \mathcal{L}_2$. Since $K \in \mathcal{H}_1$ and $C \in \mathcal{H}_2$ is outer, Lemma 1.3 implies that $D \in \mathcal{H}_2$. □

REMARK 2.8

The existence proof fails for $D \in \mathcal{H}_2$ if $1 + Bz^{-1}$ is allowed to have zeros on $\mathbb{T}$. However, in that case there exists $D \in \mathcal{N}^+$ such that the lower bound is attained. The implications of allowing $D \in \mathcal{N}^+$ will be discussed later. □

Equivalent Convex Problem

Using the optimal factorization of Lemma 2.4, it will now be shown that the infimum of $J(B,C,D)$ is equal to the infimum of a convex minimization problem.

The feasible set is

$$\Theta_{B,C,D} = \left\{ (B,C,D) : B \in \mathcal{RH}_\infty, C, D \in \mathcal{H}_2, \| C \begin{bmatrix} F & G \end{bmatrix} \|_2^2 + \|B\|_2^2 \leq \sigma^2 \right\}.$$

It will be seen that minimization of $J(B,C,D)$ over $\Theta_{B,C,D}$ is equivalent to minimization of the convex functional

$$\varphi(B,K) = \|(P-K)F\|_2^2 + \|KG\|_2^2 + \frac{\left\| K \begin{bmatrix} F & G \end{bmatrix} (1+Bz^{-1}) \right\|_1^2}{\sigma^2 + 1 - \|1+Bz^{-1}\|_2^2}, \quad (2.51)$$

over the convex set

$$\Theta_{B,K} = \left\{ (B,K) : B \in \mathcal{RH}_\infty, K \in \mathcal{H}_1, \|B\|_2^2 < \sigma^2 \right\}.$$

The (B,K) obtained from the minimization of $\varphi(B,K)$ will be used to construct (C,D) so that $(B,C,D) \in \Theta_{B,C,D}$. This will, however, only be possible if $1+Bz^{-1}$ has no zeros on the unit circle. If there are such zeros, then a small perturbation will be applied, as detailed by the following lemma.

LEMMA 2.5
Suppose $(B,K) \in \Theta_{B,K}$ and $\varepsilon > 0$. Then there exists $\hat{B}$ such that $1+\hat{B}z^{-1}$ has no zeros on $\mathbb{T}$, $(\hat{B},K) \in \Theta_{B,K}$ and

$$\varphi(\hat{B},K) < \varphi(B,K) + \varepsilon.$$

□

The proof of Lemma 2.5 is based on a perturbation argument and can be found in Appendix A.

THEOREM 2.3
Suppose that $\sigma^2 > 0$, $F, G, P \in \mathcal{H}_\infty$, and that (2.3) holds. Then

$$\inf_{(B,C,D)\in\Theta_{B,C,D}} J(B,C,D) = \inf_{(B,K)\in\Theta_{B,K}} \varphi(B,K). \quad (2.52)$$

Moreover, suppose $(B,K) \in \Theta_{B,K}$ and $\varepsilon > 0$. Let $\hat{B}$ be as given by Lemma 2.5. Then there exists (C,D) such that the following conditions hold:

- If K is not identically zero: $(C,D) \in \mathcal{H}_2$, with C outer and

$$|C|^2 = \frac{\sigma^2 - \left\|\hat{B}\right\|_2^2}{\left\|K\begin{bmatrix} F & G \end{bmatrix}(1+\hat{B}z^{-1})\right\|_1} \frac{\left|K(1+\hat{B}z^{-1})\right|}{\sqrt{|F|^2+|G|^2}} \text{ on } \mathbb{T} \tag{2.53}$$

$$D = KC^{-1}. \tag{2.54}$$

- If $K = 0$: $C = D = 0$.

If (C,D) satisfy these conditions, then $(\hat{B},C,D) \in \Theta_{B,C,D}$ and

$$J(\hat{B},C,D) < \varphi(B,K) + \varepsilon.$$

□

PROOF

Consider $(B,C,D) \in \Theta_{B,C,D}$ and let $K = DC$. Then $(C,D) \in \Theta_{C,D}(B,K)$. Moreover, $K \in \mathcal{H}_1$ and since the SNR constraint is satisfied by (B,C,D) it follows that if $\|B\|_2^2 \neq \sigma^2$ then $(B,K) \in \Theta_{B,K}$.

A lower bound will now be determined for $J(B,C,D)$. This will be accomplished through a series of inequalities and equalities, where each step will be explained afterwards.

$$\begin{aligned}
&\inf_{(B,C,D)\in\Theta_{B,C,D}} J(B,C,D) \\
&\overset{(1)}{\geq} \inf_{(B,K)\in\Theta_{B,K}} \inf_{(C,D)\in\Theta_{C,D}(B,K)} J(B,C,D) \\
&\overset{(2)}{=} \inf_{(B,K)\in\Theta_{B,K}} \|(P-K)F\|_2^2 + \|KG\|_2^2 + \inf_{(C,D)\in\Theta_{C,D}(B,K)} \left\|D(1+Bz^{-1})\right\| \\
&\overset{(3)}{\geq} \inf_{(B,K)\in\Theta_{B,K}} \|(P-K)F\|_2^2 + \|KG\|_2^2 + \frac{\left\|KH(1+Bz^{-1})\right\|_1^2}{\sigma^2 - \|B\|_2^2} \\
&\overset{(4)}{=} \inf_{(B,K)\in\Theta_{B,K}} \varphi(B,K).
\end{aligned}$$

In the first step, the minimization over (C,D) is parametrized by B and the product $K = DC$. The inequality follows from the discussion in the beginning of this proof and from the fact, shown earlier, that if $\|B\|_2^2 = \sigma^2$ then the same value can be achieved with $B = 0$.

In the second step, the first two terms of $J(B,C,D)$ are moved out from the inner minimization since they are constant for fixed (B,K). The

third step follows from application of Lemma 2.4. The fourth step follows from (2.48) and the fact that

$$\|B\|_2^2 = \left\|Bz^{-1}\right\|_2^2 + 1 - 1 = \left\|1 + Bz^{-1}\right\|_2^2 - 1,$$

which follows from orthogonality.

Now, a suboptimal solution will be constructed for Problem 2.3. Suppose that $(B,K) \in \Theta_{B,K}$ and $\varepsilon > 0$ and let $\hat{B}$ be as given by Lemma 2.5. Then $(\hat{B},K) \in \Theta_{B,K}$ and

$$\varphi(\hat{B},K) = \|(P-K)F\|_2^2 + \|KG\|_2^2 + \frac{\left\|KH(1+\hat{B}z^{-1})\right\|_1^2}{\sigma^2 - \left\|\hat{B}\right\|_2^2}$$

If $K = 0$ then

$$J(\hat{B},0,0) = \|PF\|_2^2 = \varphi(\hat{B},K) < \varphi(B,K) + \varepsilon,$$

and the proof is complete.

If, on the other hand, K is not identically zero then according to Lemma 2.4 there then exist $C,D \in \mathcal{H}_2$, with C outer, such that (2.53) and (2.54) hold. The lemma also says that such (C,D) satisfy

$$\left\|D(1+\hat{B}z^{-1})\right\| = \frac{\left\|KH(1+\hat{B}z^{-1})\right\|_1^2}{\sigma^2 - \left\|\hat{B}\right\|_2^2}$$

and

$$\|CH\|_2^2 = \sigma^2 - \left\|\hat{B}\right\|_2^2.$$

Thus, $(B,C,D) \in \Theta_{B,C,D}$ and

$$\begin{aligned} J(\hat{B},C,D) &= \|(P-K)F\|_2^2 + \|KG\|_2^2 + \left\|D(1+\hat{B}z^{-1})\right\| \\ &= \varphi(\hat{B},K) \\ &< \varphi(B,K) + \varepsilon. \end{aligned}$$

Since ε can be made arbitrarily small this shows that (2.52) holds and hence the proof is complete. □

REMARK 2.9
The purpose of the perturbation of B is to make sure that $1+Bz^{-1}$ has no zeros on the unit circle. This is needed by Lemma 2.4 to ensure existence of $D \in \mathcal{H}_2$. It was mentioned earlier that if the search of D is relaxed from $\mathcal{H}_2$ to $\mathcal{N}^+$, then optimal D exists even if $1 + Bz^{-1}$ has zeros on the unit circle.

Having $D \in \mathcal{N}^+$ means that D potentially has poles on the unit circle. This is fine with regards to the error variance that is being minimized, since these poles would be cancelled by zeros in C and $(1 + Bz^{-1})$. (If they were not cancelled then $J(B, C, D)$ would be infinite, which explains why D has no unit circle poles if $(1 + Bz^{-1})$ has no unit circle zeros.) This does, however, mean that the system is not internally stable in the following sense: If a signal with finite variance is added to the input of D, but not to the input of B, then the variance of the output of D would grow unbounded. Since the input of B is constructed by subtracting t from r this means that any numerical error in that operation could cause the output of D to grow indefinitely.

Of course, having a system without internal stability is not acceptable in practice, which is why the perturbation is performed. Still, since the perturbation is small, $(1+Bz^{-1})$ will potentially have zeros that are very close to the unit circle, and thus have a very small magnitude at the corresponding frequencies. Since it is the product of D and $(1+Bz^{-1})$ that is minimized by an optimal factorization, it is possible that D will have large magnitude peaks at those frequencies, resulting in poor robustness. □

REMARK 2.10
Note that $\varphi(0, K)$ is equal to the functional (2.14). This should have been expected, since $B = 0$ corresponds to not utilizing the channel feedback. The solution that is obtained by minimizing $\varphi(0, K)$ is thus a solution to Problem 2.1 in Section 2.2. □

It will now be shown that the minimization of $\varphi(B, K)$ over $\Theta_{B,K}$ is a convex problem. To this end, define the functional

$$\rho(a, e) = \frac{1}{2\pi} \int_{-\pi}^{\pi} a(\omega)^2 d\omega + \frac{\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} a(\omega)e(\omega)d\omega\right)^2}{\sigma^2 + 1 - \frac{1}{2\pi} \int_{-\pi}^{\pi} e(\omega)^2 d\omega}$$

with domain

$$\Theta_\rho = \left\{ (a, e) : a(\omega), e(\omega) \in \mathbb{R}\ \forall \omega, \frac{1}{2\pi} \int_{-\pi}^{\pi} e(\omega)^2 d\omega < \sigma^2 + 1 \right\}.$$

LEMMA 2.6
The functional $\rho(a,e)$ is convex. □

PROOF
Take $n \geq 2$. The function

$$\begin{aligned} f(x,y,v) &= (x+yv)^T(x+yv) - v^2, \\ &= x^T x + 2vx^T y + v^2(y^T y - 1) \end{aligned}$$

with domain $\{(x,y,v) : x,y \in \mathbb{R}^n,\ v \in \mathbb{R},\ y^T y < 1\}$, is convex in (x,y) for any $v \in \mathbb{R}$. Thus,

$$g(x,y) = \max_{v \in \mathbb{R}} f(x,y,v) = x^T x + \frac{(x^T y)^2}{1 - y^T y},$$

with domain $\{(x,y) : x,y \in \mathbb{R}^n,\ y^T y < 1\}$, is convex in (x,y) since it is the pointwise maximum of a set of convex functions [6].

Now, suppose $(a,e) \in \Theta_\rho$. Let

$$\begin{aligned} &\omega_1 = 0, \quad \omega_{k+1} - \omega_k = 2\pi/n, \quad k = 1,\ldots,n-1 \\ &\hat{a} = [\, a(\omega_1) \quad a(\omega_2) \quad \ldots \quad a(\omega_n)\,]^T \\ &\hat{e} = [\, e(\omega_1) \quad e(\omega_2) \quad \ldots \quad e(\omega_n)\,]^T . \end{aligned}$$

By definition of the integral, it holds that

$$\lim_{n\to\infty} \frac{\hat{e}^T \hat{e}}{(\sigma^2+1)n} = \frac{1}{(\sigma^2+1)} \frac{1}{2\pi} \int_{-\pi}^{\pi} e(\omega)^2 d\omega < 1.$$

So for large n, $\left(\hat{a}, (\sigma^2+1)^{-1/2}\hat{e}\right)/\sqrt{n}$ belongs to the domain of g and

$$\rho(a,e) = \lim_{n\to\infty} g\left(\frac{\hat{a}}{\sqrt{n}}, \frac{\hat{e}}{\sqrt{(\sigma^2+1)n}}\right).$$

Since the right hand side is convex in $(\hat{a},\hat{e})$, and thus in (a,e), it follows that $\rho(a,e)$ is convex. □

REMARK 2.11
Convexity of $\rho(a,e)$ has previously been shown in Lemma 4 in [11]. The proof given here is, however, substantially shorter. □

The convex functional ρ will be used in a relaxation of the minimization of φ. To compare the two functionals the constant and linear parts of φ will be removed. They do not matter for the convexity properties and will be added again later. Define the functional

$$\begin{aligned}\varphi_0(B,K) &= \varphi(B,K) - \left(\|PF\|_2^2 - 2\,\mathrm{Re}\langle PF, KF\rangle\right) \\ &= \|K\,[\,F \quad G\,]\|_2^2 + \frac{\left\|K\,[\,F \quad G\,]\,(1+Bz^{-1})\right\|_1^2}{\sigma^2 + 1 - \|1+Bz^{-1}\|_2^2},\end{aligned}$$

LEMMA 2.7
Suppose $(B,K) \in \Theta_{B,K}$. Then $\varphi_0(B,K) \le \gamma$ if and only if there exists $(a,e) \in \Theta_\rho$ such that $\rho(a,e) \le \gamma$ and

$$a(\omega) \ge |K|\sqrt{FF^* + GG^*}, \qquad e(\omega) \ge |1 + Bz^{-1}| \quad \forall\omega. \tag{2.55}$$

□

PROOF
Suppose $(B,K) \in \Theta_{B,K}$ and $\varphi_0(B,K) \le \gamma$. Let

$$a(\omega) = |K|\sqrt{FF^* + GG^*}, \qquad e(\omega) = |1 + Bz^{-1}|$$

and it follows that $\rho(a,e) = \varphi_0(B,K)$. Conversely, suppose that $(a,e) \in \Theta_\rho$ satisfy (2.55) and that $\rho(a,e) \le \gamma$. Then it follows from inspection of $\varphi_0(B,K)$ and $\rho(a,e)$ that $\varphi_0(B,K) \le \rho(a,e) \le \gamma$. □

Convexity can now be proved.

THEOREM 2.4
The problem of minimizing $\varphi(B,K)$ over $\Theta_{B,K}$ is convex. □

PROOF
Suppose $(B_1,K_1) \in \Theta_{B,K}$ and $(B_2,K_2) \in \Theta_{B,K}$. Then by Lemma 2.7 there exists $(a_1,e_1) \in \Theta_\rho$ and $(a_2,e_2) \in \Theta_\rho$ such that $\rho(a_1,e_1) \le \varphi_0(B_1,K_1)$ and $\rho(a_2,e_2) \le \varphi_0(B_2,K_2)$. For $0 \le \theta \le 1$, it thus holds that

$$\begin{aligned}\theta\varphi_0(B_1,K_1) + (1-\theta)\varphi_0(B_2,K_2) &\ge \theta\rho(a_1,e_1) + (1-\theta)\rho(a_2,e_2) \\ &\ge \rho\left(\theta a_1 + (1-\theta)a_2, \theta e_1 + (1-\theta)e_2\right) \\ &\ge \varphi_0(\theta B_1 + (1-\theta)B_2, \theta K_1 + (1-\theta)K_2).\end{aligned}$$

The second inequality follows from Lemma 2.6. The third inequality follows from Lemma 2.7 and that the constraints (2.55) are convex. It is thus proved that $\varphi_0(B,K)$ is convex in (B,K). Since

$$\varphi(B,K) - \varphi_0(B,K) = \|PF\|_2^2 - 2\,\mathrm{Re}\langle PF, KF\rangle \stackrel{\text{def}}{=} \Delta(K)$$

is also convex it follows that $\varphi(B,K)$ is convex. It is finally noted that $\Theta_{B,K}$ is convex. □

Numerical Solution

Even though convexity of $\varphi(B,K)$ has been established, it is not immediately clear how to formulate the minimization problem in a standard form. Such a formulation will be derived in this subsection.

Minimizing $\varphi(B,K)$ over $\Theta_{B,K}$ is, by Lemma 2.7, equivalent to minimizing $\rho(a,e)+\Delta(K)$ over $\Theta_\rho \times \Theta_{B,K}$ subject to (2.55). A finite dimensional approximation with n grid points gives

$$\rho_n(\hat{a},\hat{e}) = \frac{1}{n}\hat{a}^T\hat{a} + \frac{\left(\frac{1}{n}\hat{a}^T\hat{e}\right)^2}{\sigma^2 + 1 - \frac{1}{n}\hat{e}^T\hat{e}} \approx \rho(a,e)$$

$$\Delta_n(K) = \|PF\|_2^2 - \frac{2}{n}\,\mathrm{Re}\sum_{k=1}^{n} P(e^{i\omega_k})^*|F(e^{i\omega_k})|^2 K(e^{i\omega_k}) \approx \Delta(K).$$

By the definition of the integral it holds that

$$\lim_{n\to\infty} \rho_n(\hat{a},\hat{e}) + \Delta_n(K) = \rho(a,e) + \Delta(K),$$

so the minimum of the approximation can be made to come arbitrarily close to $\inf_{(B,K)\in\Theta_{B,K}} \varphi(B,K)$ if n is chosen sufficiently large. When implementing the minimization program, the transfer functions K and B are parametrized using finite basis representations. The accuracy of the approximated problem obviously depends on this representation as well.

Noting that $\rho_n(\hat{a},\hat{e}) + \Delta_n(K)$ can be written as a Schur complement and that the denominator of $\rho_n(\hat{a},\hat{e})$ is positive for sufficiently large n, it follows that $\rho_n(\hat{a},\hat{e}) + \Delta_n(K) \le \gamma$ if and only if

$$\begin{bmatrix} \frac{1}{n}\hat{e}^T\hat{e} - \sigma^2 - 1 & \frac{1}{n}\hat{a}^T\hat{e} \\ \frac{1}{n}\hat{e}^T\hat{a} & \frac{1}{n}\hat{a}^T\hat{a} + \Delta_n(K) - \gamma \end{bmatrix} \preceq 0,$$

or, equivalently,

$$\begin{bmatrix} n(\sigma^2+1) & 0 \\ 0 & n(\gamma - \Delta_n(K)) \end{bmatrix} - [\,\hat{e} \quad \hat{a}\,]^T I\,[\,\hat{e} \quad \hat{a}\,] \succeq 0.$$

Since the left hand side of this inequality also can be written as a Schur complement, this is equivalent to

$$\begin{bmatrix} I & \hat{e} & \hat{a} \\ \hat{e}^T & n(\sigma^2+1) & 0 \\ \hat{a}^T & 0 & n(\gamma - \Delta_n(K)) \end{bmatrix} \succeq 0. \tag{2.56}$$

The constraints can be approximated by

$$a(\omega_k) \geq |K(e^{i\omega_k})| \sqrt{|F(e^{i\omega_k})|^2 + |G(e^{i\omega_k})|^2}, \; k = 1 \dots n \tag{2.57}$$

$$e(\omega_k) \geq \left|1 + B(e^{i\omega_k}) e^{-i\omega_k}\right|, \; k = 1 \dots n \tag{2.58}$$

$$\sigma^2 + 1 > \frac{1}{n} \sum_{k=1}^{n} e(\omega_k)^2. \tag{2.59}$$

Minimizing γ subject to (2.56)–(2.59) is a semidefinite program.

A procedure for numerical solution of the coding problem with channel feedback will now be outlined.

1. Choose n to be sufficiently large and determine the grid points ω_k, $k = 1 \dots n$. Then solve the optimization problem of minimizing γ subject to (2.56)–(2.59). The transfer functions K and B are parametrized by finite basis representations, for example as FIR filters.
2. Use a finite basis approximation $A(\omega)$ of CC^*, for example the parametrization (2.25), and fit it to the right hand side of (2.53), for example by minimizing the mean squared deviation.
3. Perform a spectral factorization of $A(\omega)$, choosing C as the stable and minimum phase spectral factor.
4. Let $D = KC^{-1}$.

The following example illustrates the numerical solution. The example also shows that channel feedback can enhance the performance. It is thus established that feedback is useful for linear real-time coding with a partially observed source.

EXAMPLE 2.3
Suppose $P = z^{-2} + 0.5z^{-7}$, $F = \frac{1}{z-0.5}$, $G = 1$ and $\sigma = 1$. K and B are parametrized as FIR filters with 20 and 19 coefficients, respectively. The minimization is implemented in Matlab using Yalmip [30] and SeDuMi [55], with a grid distance of 0.0025. The impulse responses of the resulting transfer functions can be seen in figures 2.9–2.12.

The minimum value for this problem is 0.90, to compare with the minimum value of 1.00 for the corresponding problem with no access to feedback from the channel, as seen in Section 2.2. □

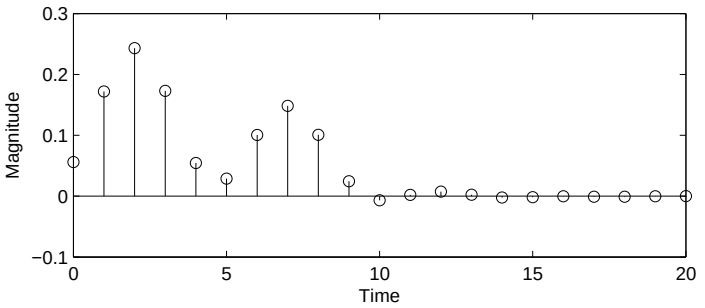

Figure 2.9 Impulse response of K, the product of C and D, in Example 2.3. Note that the timing of the peaks coincide with the exponents of P.

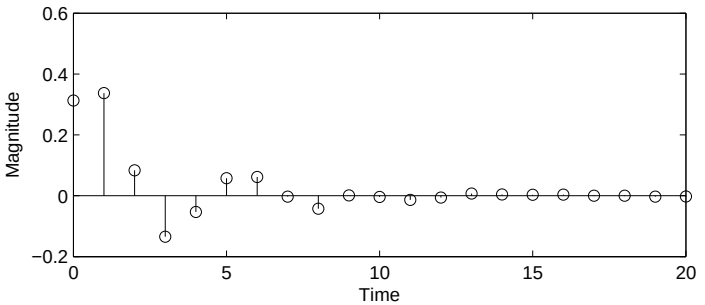

Figure 2.10 Impulse response of the encoder C in Example 2.3.

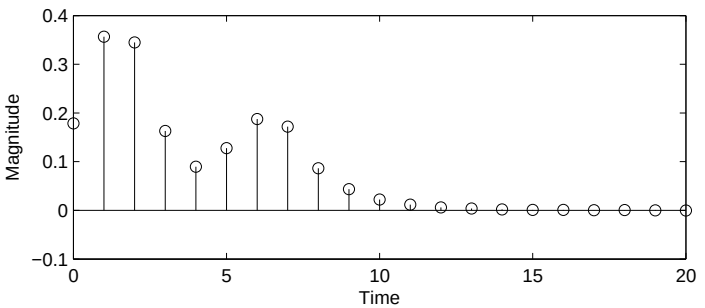

Figure 2.11 Impulse response of the decoder D in Example 2.3.

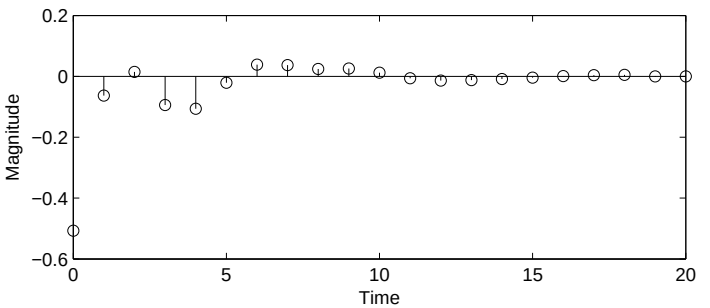

Figure 2.12 Impulse response of B, the feedback part of the encoder, in Example 2.3. Note that the direct term is strongly negative, resulting in a transmission that is negatively correlated with the channel noise in the previous time step.

3

Feedback Control over a Noisy Channel

3.1 Introduction

This chapter is about the problem of feedback control of a plant over a communication channel. The problem is illustrated in Figure 3.1. The controller is based on output feedback and consists of an encoder and a decoder. The encoder measures the plant output, filters the measurements and encodes them for transmission over the communication channel, which is an AWN channel. The decoder decodes the received signal and determines the control signal. The plant is LTI, possibly unstable and subject to a stochastic disturbance. The objective of the controller is to stabilize the system and minimize the plant output, while satisfying the SNR constraint of the channel.

Outline and Main Results

The case when the channel has no feedback and the controller two degrees of freedom is considered in Section 3.2. The case when the channel has noiseless feedback and the controller has three degrees of freedom is considered in Section 3.3. In both sections, it is shown that an optimal LTI output feedback controller is obtained by minimizing a convex functional and performing a spectral factorization. These results are given in Theorem 3.1 and Theorem 3.3. Previously known necessary and sufficient conditions on the SNR for stabilizability follow as by-products of the theorems. In both cases, it is also shown how to pose the minimization problem as a semidefinite program.

Given a nominal LTI controller K, designed for a classical feedback

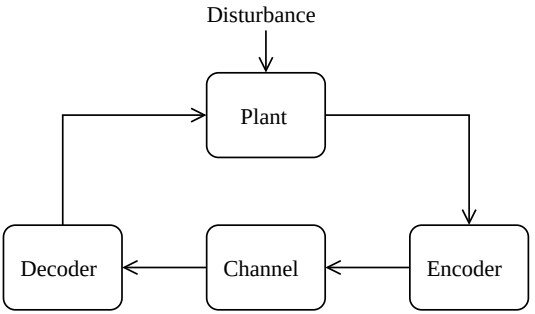

Figure 3.1 Feedback control of a disturbed plant over a noisy communication channel. The controller consists of an encoder and a decoder.

system, Lemma 3.3 and Lemma 3.8 shows, in the case without and with channel feedback, respectively, how to factorize K into an encoder and a decoder such that the impact of the channel noise is minimized, while preserving the original closed loop transfer functions from plant input to output.

In the end of both sections, it is shown that the coding problems with scalar-valued signals, considered earlier in this thesis, can be solved using the solutions to the feedback control problems.

The rest of this section will present the relevant previous research.

Previous Research

The general problem of controlling a process over a communication channel has received a lot of interest recently. As a consequence, many results have been obtained regarding stabilizability, performance bounds, and controller design, both for general channels and for specific channel models. Some of these results were presented in Section 1.2. The exposition here will be focused mainly on the study of linear control over AWN channels, also known as the SNR framework.

It has been argued that the SNR framework, despite its relative simplicity, offers the possibility of studying a variety of control problems with communication constraints and that the usage of linear controllers admits application of established performance and robustness tools [7].

Stability Necessary and sufficient criteria for stabilizability of undisturbed LTI plants under an SNR constraint were developed for state and output feedback in [7]. For static linear state feedback, the condition is

that the SNR σ^2 satisfies

$$\sigma^2 > \left(\prod_i |\max\{1, \lambda_i\}|^2 \right) - 1, \tag{3.1}$$

where λ_i is the ith pole of the plant. For AWGN channels, it can be seen from (1.2) and (1.3) that this condition exactly matches the data-rate theorem. For LTI output feedback, a higher SNR will be required if the plant is not minimum phase. The condition is that

$$\sigma^2 > \left(\prod_i |\max\{1, \lambda_i\}|^2 \right) - 1 + \eta + \delta, \tag{3.2}$$

where $\eta \geq 0$ depends on the non-minimum phase zeros and $\delta \geq 0$ on the relative degree of the plant. It was shown that the condition (3.1) is recovered for output feedback if the controller is allowed to be time-varying. However, this leads to poor robustness and sensitivity [7].

The SNR that is necessary and sufficient for stabilization if there are bandwidth limitations and colored channel noise has been characterized in [43].

For stabilization of a plant driven by Gaussian noise over an AWGN channel, the condition (3.1) has been shown to be necessary for stabilization, even if nonlinear and time varying state feedback controllers are allowed [19]. It has been shown that if the encoder is a unity gain and the decoder is the only design variable, then a plant disturbance may increase the required SNR so that the condition (3.2) is no longer sufficient [42]. However, regardless of the channel and plant noise distributions, it turns out that necessary and sufficient conditions for stabilizability of an LTI plant using LTI output feedback are given by (3.2) when the controller has two degrees of freedom, and by (3.1) if the channel has feedback [51].

Performance An early formulation of a feedback control problem over an AWN channel was made in [2], although the encoder design was there referred to as designing a measurement strategy. In that paper, the plant is modelled as a time-varying ARMA process and the optimal control problem is considered both for hard and soft power constraints (the former being equivalent to an SNR constraint). The channel has feedback and the encoder has access to the complete history of the decoder input. The assumed information structure is, however, a bit unusual since it is assumed that the encoder only remembers the current state of the plant. It was shown how to find the optimal linear solution and it was established, using information theory, that linear mappings are optimal for first-order

plants. A counterexample is provided, showing that non-linear solutions may outperform linear ones for higher order plants [2]. It should, however, be noted that the provided counterexample requires the plant to be time-varying and that it fails if the encoder is has more memory of the process output.

Many authors have since then considered the design of optimal controllers with one degree of freedom: It has been shown how to find the optimal controller for a structure where the encoder is a constant scaling factor and the decoder an LTI filter, and that this structure is optimal for first order plants [17]. Optimal coding systems with unity transfer function (that is, the decoder is the inverse of the encoder) was designed for fixed nominal controllers in different architectures in [25]. A more general approach to the optimization of an output feedback controller in a general LTI architecture, possibly including channel feedback, has been proposed, in which the optimal performance is obtained by solving a convex optimization problem. It was, however, noted that the approach leads to a difficult optimization problem with sparsity constraints when the controller has two degrees of freedom [51]. A state feedback controller, where the encoder consists of a scalar product between a static vector and the states and the decoder is based on an LQG solution was developed in [8]. Optimal control design for MIMO plants and parallel AWN channels was studied in [29], where the encoder was a constant scaling factor and the decoder an LTI filter, and in [41] for static state feedback.

A lower bound on the variance of the plant state was obtained for feedback control over AWGN channels without feedback, using general controllers with two degrees of freedom, in [19]. This bound tends to infinity as the SNR approaches the limit for when stabilization is possible. For first-order plants it was also shown that the performance bound is tight and achievable with linear controllers.

An important contribution was made in [53]. Although the aim of that paper is to establish bounds and coding schemes for control over a rate-limited channel, it accomplishes this by designing an LTI output feedback controller in an SNR framework where the channel has feedback. The optimal performance is shown to be obtained by solving a convex optimization problem with the same structure as the ones found in this chapter. An optimal controller is then acquired by finding rational transfer functions that approximate certain frequency responses.

The case without channel feedback is not mentioned in [53]. The presented convex functional that gives the optimal cost for the case with channel feedback can, however, with some minor work be modified to give the optimal cost for this case as well. Moreover, the expressions for the optimal transfer functions that are given can, with some additional work, also be modified to give solutions to the case without channel feedback.

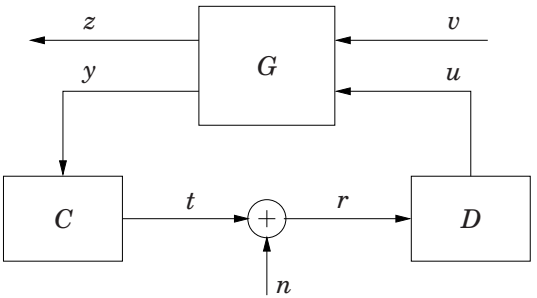

Figure 3.2 Feedback system with noisy communication channel. The objective is to design the controller components C and D so that the system is stabilized and the variance of z is minimized under the SNR constraint (3.3).

Regarding the case with channel feedback, the solution in [53] uses an over-parametrized controller with four degrees of freedom (the encoder has two filters that both handle the channel feedback). The solution in Section 3.3 is related, but requires only three degrees of freedom. Moreover, the plant is assumed to be SISO in [53]. Here, it is allowed to have a slightly more general structure, making it possible to, for example, include any number of noise and reference signals and to penalize the control signal variance. A final contribution of this chapter relative to [53] is that it is shown how to pose the optimization problems as semidefinite programs.

Results related to [53], in the case where the controller is pre-designed and the coding system should have unity transfer function, are given in [49] and [50].

The problem of optimizing the control performance at a given terminal time was considered in [18] and [16]. The solutions may however yield poor transient performance and can therefore be unsuitable for closed-loop control.

3.2 Optimal Linear Controller

In this section, the problem of controlling a plant over an AWN channel will be considered. It will be shown how to design an optimal LTI output feedback controller with two degrees of freedom.

Problem Formulation and Assumptions

A detailed block diagram representation of the system is shown in Fig-

ure 3.2. The plant G is a MIMO LTI system with state space realization

$$G(z) = \begin{bmatrix} G_{zv}(z) & G_{zu}(z) \\ G_{yv}(z) & G_{yu}(z) \end{bmatrix} = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right],$$

where (A, B_2) is stabilizable and (C_2, A) is detectable. The signals v and z are vector-valued with n_v and n_z elements, respectively. All other signals are scalar-valued. Accordingly, G_{zv} is $n_z \times n_v$, G_{yv} is $1 \times n_v$, G_{zu} is $n_z \times 1$ and G_{yu} is scalar and strictly proper. It is assumed that $G_{zu}^* G_{zu}$ and $G_{yv} G_{yv}^*$ have no zeros or poles on the unit circle.

The input v is used to model exogenous signals such as load disturbances, measurement noise and reference signals. It is assumed that v and the channel noise n are mutually independent white noise sequences with zero mean and identity variance. The other signals in the figure include the control signal u, the measurement y and the control error or performance index z.

This type of plant model is quite general. Actually, G is the coefficient matrix of a linear fractional transformation (LFT) [65]. The generality of the structure makes it possible to formulate many different problems as special cases of this problem. This will be illustrated in the end of this and the next section.

The communication channel is an AWN[1] channel with SNR $\sigma^2 > 0$. The SNR constraint is assumed to hold in stationarity, that is

$$\lim_{k\to\infty} \mathrm{E}(t(k)^2) \le \sigma^2. \tag{3.3}$$

The feedback system is said to be internally stable if no additive injection of a stochastic signal with finite variance, at any point in the block diagram, leads to another signal having unbounded variance. This is true if and only if all closed loop transfer functions are in $\mathcal{H}_2$.

The objective is to find causal LTI systems C and D that make the system internally stable, satisfy the SNR constraint (3.3) and minimize the sum of the variances of z in stationarity:

$$\lim_{k\to\infty} \mathrm{E}(z(k)^T z(k)).$$

By expressing z and t in terms of the transfer functions in Figure 3.2, the objective and the SNR constraint can be written as

$$J(C,D) = \left\| G_{zv} + \frac{DCG_{zu}G_{yv}}{1 - DCG_{yu}} \right\|_2^2 + \left\| \frac{DG_{zu}}{1 - DCG_{yu}} \right\|_2^2 = \lim_{k\to\infty} \mathrm{E}(z(k)^T z(k))$$

[1]Since only linear controllers are considered, it does not matter if n or v are Gaussian or not. Linear solutions may, of course, be more or less suboptimal depending on their distributions.

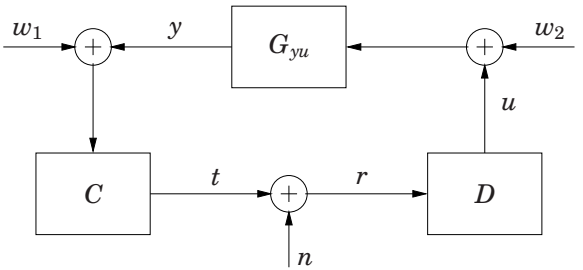

Figure 3.3 Block diagram for internal stability analysis.

and

$$\sigma^2 \geq \left\| \frac{CG_{yv}}{1 - DCG_{yu}} \right\|_2^2 + \left\| \frac{DCG_{yu}}{1 - DCG_{yu}} \right\|_2^2 = \lim_{k \to \infty} \mathrm{E}(t(k)^2), \qquad (3.4)$$

respectively. The problem can thus be stated as follows.

PROBLEM 3.1

$$\underset{C,D}{\text{minimize}}\, J(C, D)$$

subject to (3.4) and internal stability of the feedback system. □

For technical reasons, only solutions where the product DC is a rational transfer function will be considered. This may exclude the possibility of achieving the minimum value, but the infimum can still be arbitrarily well approximated by rational functions.

Since D and C are required to be proper, DC has to be proper has well. That is, $DC \in \mathcal{R}$. Though the latter will be enforced, it is not explicitly required in this problem formulation that C and D are proper. It will, however, be seen that the solution can be constructed so that $C \in \mathcal{H}_2$ is outer. Then C, C^{-1} are proper, and $D = (DC)C^{-1}$ is also proper.

Internal Stability

The solution to Problem 3.1 will be found by using the same optimal factorization approach as in Chapter 2. Therefore, introduce

$$K = DC.$$

Following the same reasoning as in [65], it is concluded that internal stability of the systems in Figure 3.2 and Figure 3.3 are equivalent. The

latter can be represented by the closed loop map T, defined by

$$\begin{bmatrix} y \\ t \\ u \end{bmatrix} = T \begin{bmatrix} w_1 \\ w_2 \\ n \end{bmatrix}.$$

Hence, the system in Figure 3.2 is internally stable if and only if

$$T = \begin{bmatrix} \frac{KG_{yu}}{1-KG_{yu}} & \frac{G_{yu}}{1-KG_{yu}} & \frac{DG_{yu}}{1-KG_{yu}} \\ \frac{C}{1-KG_{yu}} & \frac{CG_{yu}}{1-KG_{yu}} & \frac{KG_{yu}}{1-KG_{yu}} \\ \frac{K}{1-KG_{yu}} & \frac{KG_{yu}}{1-KG_{yu}} & \frac{D}{1-KG_{yu}} \end{bmatrix} \in \mathcal{H}_2. \tag{3.5}$$

The following two lemmas will give necessary and sufficient conditions for internal stability, respectively.

LEMMA 3.1
Suppose that $T \in \mathcal{H}_2$, that $G_{yu} = NM^{-1}$ is a coprime factorization over $\mathcal{RH}_\infty$ and that $U, V \in \mathcal{RH}_\infty$ satisfy the Bezout identity $VM + UN = 1$. Then

$$K = \frac{MQ - U}{NQ + V}, \quad Q \in \mathcal{RH}_\infty. \tag{3.6}$$

□

PROOF
It follows directly from (3.5) that

$$\frac{G_{yu}}{1-KG_{yu}} \in \mathcal{H}_2, \quad \frac{K}{1-KG_{yu}} \in \mathcal{H}_2, \quad \frac{KG_{yu}}{1-KG_{yu}} = \frac{1}{1-KG_{yu}} - 1 \in \mathcal{H}_2.$$

These transfer functions are rational and have no poles on or outside the unit circle, so it follows that

$$\begin{bmatrix} 1 & -K \\ -G_{yu} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{1-KG_{yu}} & \frac{K}{1-KG_{yu}} \\ \frac{G_{yu}}{1-KG_{yu}} & \frac{1}{1-KG_{yu}} \end{bmatrix} \in \mathcal{RH}_\infty, \tag{3.7}$$

It is well-known that the set of K satisfying (3.7) can be parametrized using the Youla parametrization of all stabilizing controllers [65]. That is, K can be written as in (3.6). □

LEMMA 3.2
Suppose that

$$K = DC = \frac{MQ - U}{NQ + V}, \quad Q \in \mathcal{RH}_\infty, \tag{3.8}$$

where $G_{yu} = NM^{-1}$ is a coprime factorization over $\mathcal{RH}_\infty$ and $U, V \in \mathcal{RH}_\infty$ satisfy the Bezout identity $VM + UN = 1$. Suppose also that $C \in \mathcal{H}_2$ is outer and that $D \in \mathcal{L}_2$. Then $T \in \mathcal{H}_2$. □

PROOF
It follows from (3.8) that

$$\frac{G_{yu}}{1 - KG_{yu}} \in \mathcal{RH}_\infty, \quad \frac{K}{1 - KG_{yu}} \in \mathcal{RH}_\infty,$$
$$\frac{KG_{yu}}{1 - KG_{yu}} = \frac{1}{1 - KG_{yu}} - 1 \in \mathcal{RH}_\infty.$$

Moreover,

$$\frac{DG_{yu}}{1 - KG_{yu}} = \frac{KG_{yu}}{1 - KG_{yu}} C^{-1},$$

where the left hand side is in $\mathcal{L}_2$ since it is the product of an $\mathcal{L}_2$ function and a $\mathcal{RH}_\infty$ function. Since C is outer, application of Lemma 1.3 gives that the right hand side is in $\mathcal{H}_2$. A similar argument shows that

$$\frac{D}{1 - KG_{yu}} \in \mathcal{H}_2.$$

Finally,

$$\frac{C}{1 - KG_{yu}} \in \mathcal{H}_2, \quad \frac{CG_{yu}}{1 - KG_{yu}} \in \mathcal{H}_2,$$

since these functions are products of an $\mathcal{H}_2$ function and an $\mathcal{RH}_\infty$ function.

Noting that $\mathcal{RH}_\infty \subseteq \mathcal{H}_2$, it has then been proved that all elements of T are in $\mathcal{H}_2$ and thus $T \in \mathcal{H}_2$. □

Optimal Factorization

Suppose for now that K is a given stabilizing controller for the classical feedback system in Figure 3.4. Thus, K satisfies (3.6). Nothing else is assumed about the design of K, it might for example be the $\mathcal{H}_2$ optimal controller or have some other desirable properties in terms of step responses or closed loop sensitivity.

In either case, it is a natural question to ask what the best way is to implement this controller in the architecture of Figure 3.2. If the nominal

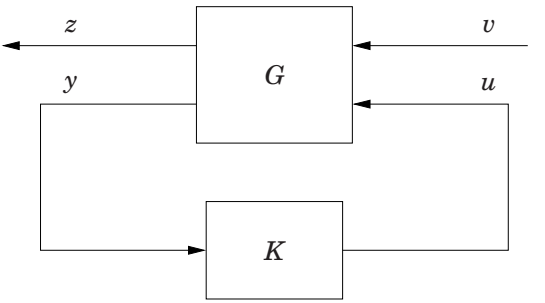

Figure 3.4 Classical feedback system without communication channel.

design is to be preserved then C and D should satisfy $K = DC$ since the transfer matrix from v to z would then be the same. Given this relationship, choosing C and D can be thought of as factorizing K. The factorization should be chosen to minimize the negative effect of the communication channel. That is, they should keep the system stable, satisfy the SNR constraint and minimize the impact of the channel noise. The latter can be thought of as minimizing the contribution of n to the variance of z.

Rewriting $J(C, D)$ and the SNR constraint in terms of K gives

$$\left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} \right\|_2^2 + \left\| \frac{DG_{zu}}{1 - KG_{yu}} \right\|_2^2 \tag{3.9}$$

and

$$\left\| \frac{CG_{yv}}{1 - KG_{yu}} \right\|_2^2 + \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2 \leq \sigma^2. \tag{3.10}$$

The SNR constraint will be impossible to satisfy unless K satisfies

$$\alpha = \sigma^2 - \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2 \geq 0.$$

Actually, if $\alpha = 0$ then

$$\left\| \frac{CG_{yv}}{1 - DCG_{yu}} \right\|_2^2 = 0 \Rightarrow \frac{CG_{yv}}{1 - DCG_{yu}} = 0 \Rightarrow \frac{DCG_{yv}}{1 - DCG_{yu}} = \frac{KG_{yu}}{1 - KG_{yu}} = 0,$$

which is a contradiction. Thus, it will be assumed for the optimal factorization problem that $\alpha > 0$.

The objective of the optimal factorization problem is to find C and D such that (3.9) is minimized subject to (3.10) and $K = DC$. Stability considerations are temporarily disregarded and will be handled later. Introducing

$$S = \frac{1}{1 - KG_{yu}} \in \mathcal{RH}_\infty$$

for notational convenience, the set of feasible (C, D), parametrized by K, is thus defined as

$$\Theta_{C,D}(K) = \left\{ (C, D) : \|CSG_{yv}\|_2^2 \leq \alpha, DC = K \right\}.$$

It is noted that the first term in (3.9) is constant and that the second term is a weighted norm of D. In the left hand side of (3.10), the first term is a weighted norm of C and the second is constant. This means that an optimal factorization problem can be formulated with the same structure as those presented in Chapter 2.

The optimal factorization will then be used to solve Problem 3.1. It could also be used to factorize a nominal controller that was designed for a classical feedback architecture. The solution to the optimal factorization problem is given by the following lemma.

LEMMA 3.3—OPTIMAL FACTORIZATION, FEEDBACK CONTROL CASE
Suppose that $\alpha > 0$, $S \in \mathcal{RH}_\infty$, $K \in \mathcal{R}$ and that $G_{zu}^* G_{zu} \in \mathcal{RL}_\infty$ and $G_{yv} G_{yv}^* \in \mathcal{RL}_\infty$ have no zeros on $\mathbb{T}$. Then

$$\inf_{(C,D) \in \Theta_{C,D}(K)} \|DSG_{zu}\|_2^2 \geq \frac{1}{\alpha} \left\| KS^2 G_{zu} G_{yv} \right\|_1^2. \tag{3.11}$$

Suppose furthermore that $K \in \mathcal{RL}_1$ satisfies (3.6). Then there exists $(C, D) \in \Theta_{C,D}(K)$ with $C \in \mathcal{H}_2$ outer and $D \in \mathcal{L}_2$, such that the minimum is attained and (3.11) holds with equality.

If K is not identically zero, then (C, D) is optimal if and only if $DC = K$ and

$$|C|^2 = \frac{\alpha}{\|KS^2 G_{zu} G_{yv}\|_1} \sqrt{\frac{G_{zu}^* G_{zu}}{G_{yv} G_{yv}^*}} |K| \text{ on } \mathbb{T}. \tag{3.12}$$

If $K = 0$, then the minimum is achieved by $D = 0$ and any C that satisfies $\|CSG_{yv}\|_2^2 \leq \alpha$. □

PROOF

Suppose first that $K = 0$. Then the right hand side of (3.11) is 0. Letting $D = 0$ gives $\|SDG_{zu}\|_2^2 = 0$ and it is clear that $(C, D) \in \Theta_{C,D}$ if C is as stated.

Thus, it can now be assumed that K is not identically zero. Then C is not identically zero and $D = KC^{-1}$.

By assumption both $G_{zu}^* G_{zu}$ and $G_{yv} G_{yv}^*$ are positive on the unit circle. Since these functions are rational this implies that

$$\exists \varepsilon > 0 \text{ such that } G_{zu}^* G_{zu} \geq \varepsilon \text{ and } G_{yv} G_{yv}^* \geq \varepsilon, \text{ on } \mathbb{T}. \tag{3.13}$$

Thus by Theorem 1.3 there exist scalar minimum phase transfer functions $\hat{G}_{zu}, \hat{G}_{yv} \in \mathcal{H}_2$ such that

$$G_{zu}^* G_{zu} = \hat{G}_{zu}^* \hat{G}_{zu}, \qquad G_{yv} G_{yv}^* = \hat{G}_{yv} \hat{G}_{yv}^*.$$

Now, $\|CSG_{yv}\|_2^2 \leq \alpha$ and Cauchy-Schwarz's inequality gives

$$\begin{aligned}
\|DSG_{zu}\|_2^2 &= \left\| KC^{-1} S \hat{G}_{zu} \right\|_2^2 \\
&\geq \frac{\left\| CS\hat{G}_{yv} \right\|_2^2}{\alpha} \left\| KC^{-1} S \hat{G}_{zu} \right\|_2^2 \\
&\geq \frac{1}{\alpha} \left\langle \left| CS\hat{G}_{yv} \right|, \left| KC^{-1} S \hat{G}_{zu} \right| \right\rangle^2 \\
&= \frac{1}{\alpha} \left\| KS^2 \hat{G}_{zu} \hat{G}_{yv} \right\|_1^2 \\
&= \frac{1}{\alpha} \left\| KS^2 G_{zu} G_{yv} \right\|_1^2.
\end{aligned}$$

This proves the lower bound (3.11).

Equality holds if and only if $|KC^{-1}S\hat{G}_{zu}|$ and $|CS\hat{G}_{yv}|$ are proportional on the unit circle and $\|CSG_{yv}\|_2^2 = \alpha$. It is easily verified that this is equivalent to (3.12). Thus, (C, D) achieves the lower bound if and only if $D = KC^{-1}$ and (3.12) holds, since these conditions imply that $(C, D) \in \Theta_{C,D}(K)$.

Under the additional assumptions that $K \in \mathcal{RL}_1$ satisfies (3.6), it will now be shown that there exists such $(C, D) \in \mathcal{H}_2 \times \mathcal{L}_2$ with C outer. Since K satisfies (3.6) with $M, N, Q, U, V \in \mathcal{RH}_\infty$ it holds that

$$\log|K| = \log|MQ - U| - \log|NQ + V|$$

By Lemma 1.5, $\log|MQ - U| \in \mathcal{L}_1$ and $\log|NQ + V| \in \mathcal{L}_1$ and thus $\log|K| \in \mathcal{L}_1$. It follows from (3.13) and the boundedness of $\hat{G}_{yv}$ and $\hat{G}_{zu}$ on $\mathbb{T}$ that

$$\int_{-\pi}^{\pi} \log \left| \frac{\hat{G}_{zu}}{\hat{G}_{yv}} K \right| d\omega > -\infty$$

and

$$\left| \frac{\hat{G}_{zu}}{\hat{G}_{yv}} K \right| \in \mathcal{L}_1.$$

Then by Theorem 1.3 there exists an outer function $C \in \mathcal{H}_2$ such that (3.12) holds. Also, $D = KC^{-1} \in \mathcal{L}_2$ since

$$\left\| KC^{-1} \right\|_2^2 = \frac{1}{\alpha} \left\| KS^2 G_{zu} G_{yv} \right\|_1 \left\| \frac{K\hat{G}_{yv}}{\hat{G}_{zu}} \right\|_1 < \infty.$$

□

Remark 3.1
The spectral factorization gives some freedom in the choice of (C, D) that attain the bound. For example, D instead of C could be chosen to be $\mathcal{H}_2$ and outer. That would result in having $C \in \mathcal{L}_2$. Considering more solutions than the one selected would require more a slightly more complicated stability characterization, so this is not done. □

Remark 3.2
Optimal D will satisfy

$$|D|^2 = \frac{\left\| KS^2 G_{zu} G_{yv} \right\|_1}{\alpha} \sqrt{\frac{G_{yv} G_{yv}^*}{G_{zu}^* G_{zu}}} |K| \text{ on } \mathbb{T}.$$

It is interesting that the magnitudes of C and D both are directly proportional, on the unit circle, to the square root of the magnitude of K. In other words, the dynamics of a nominal controller K is "evenly" distributed on both sides of the communication channel. The static gain of C (and D) is tuned so that the SNR constraint is active. In the case when $G_{yv} = G_{zu}$ then finding an optimal factorization amounts to performing a spectral factorization of $|K|$ and tuning the static gain. The magnitudes of the frequency responses of C and D will then be proportional. □

Equivalent Convex Problem

It will now be shown that an approximate solution to Problem 3.1 can be obtained, with arbitrary accuracy, by solving a convex minimization problem in the Youla parameter.

As discussed in the problem formulation, (C, D) should satisfy the SNR constraint (3.4) and stabilize the system. The latter corresponds to

$T \in \mathcal{H}_2$ or (3.5). Also, it was assumed that $CD \in \mathcal{R}$. Thus, the set of feasible (C, D) is defined as

$$\Theta_{C,D} = \left\{ (C, D) : DC \in \mathcal{R}, (3.4), T \in \mathcal{H}_2 \right\}.$$

It will be shown that minimization of $J(C, D)$ over $\Theta_{C,D}$ can be performed by minimizing the convex functional

$$\varphi(Q) = \|G_{zv} + G_{zu}G_{yv}(AQ + B)\|_2^2 + \frac{\|G_{zu}G_{yv}\,(AQ + B)\,(EQ + F)\|_1^2}{\sigma^2 + 1 - \|EQ + F\|_2^2},$$

where $A = M^2$, $B = -MU$, $E = MN$ and $F = MV$, with M, N, U, V determined by a coprime factorization of G_{yu}, over the convex set

$$\Theta_Q = \left\{ Q : Q \in \mathcal{RH}_\infty, \|EQ + F\|_2^2 < \sigma^2 + 1 \right\}.$$

The $Q \in \Theta_Q$ obtained from minimizing $\varphi(Q)$ will be used to construct $(C, D) \in \Theta_{C,D}$. However, this will not be possible for Q for which the corresponding K has poles on $\mathbb{T}$. For such Q a small perturbation can then be applied first. This will result in an increased cost, but this increase can be made arbitrarily small. That this is possible is established by the following lemma.

LEMMA 3.4
Suppose $Q \in \Theta_Q$ and $\varepsilon > 0$. Then there exists $\hat{Q} \in \Theta_Q$ such that

$$K = \frac{M\hat{Q} - U}{N\hat{Q} + V} \in \mathcal{RL}_1, \tag{3.14}$$

and

$$\varphi(\hat{Q}) < \varphi(Q) + \varepsilon.$$

□

The proof of Lemma 3.4 is based on a perturbation argument and can be found in Appendix A.

The main theorem of this section can now be formulated.

THEOREM 3.1
Suppose that $\sigma^2 > 0$, $G_{zu}^* G_{zu} \in \mathcal{RL}_\infty$ and $G_{yv}G_{yv}^* \in \mathcal{RL}_\infty$ have no zeros on $\mathbb{T}$, that $G_{yu} = NM^{-1}$ is a coprime factorization over $\mathcal{RH}_\infty$, and that $U, V \in \mathcal{RH}_\infty$ satisfy the Bezout identity $VM + UN = 1$. Then

$$\inf_{(C,D)\in\Theta_{C,D}} J(C, D) = \inf_{Q\in\Theta_Q} \varphi(Q). \tag{3.15}$$

Furthermore, suppose $Q \in \Theta_Q$, $\varepsilon > 0$ and let $\hat{Q} \in \Theta_Q$ be as in Lemma 3.4. Then there exists (C, D) such that the following conditions hold:

- If $M\hat{Q} - U$ is not identically zero: $(C, D) \in \mathcal{H}_2 \times \mathcal{L}_2$, where C is outer and

$$K = \frac{M\hat{Q} - U}{N\hat{Q} + V} \tag{3.16}$$

$$|C|^2 = \frac{\sigma^2 + 1 - \left\| \dfrac{1}{1 - KG_{yu}} \right\|_2^2}{\left\| \dfrac{KG_{zu}G_{yv}}{(1 - KG_{yu})^2} \right\|_1} \sqrt{\frac{G_{zu}^* G_{zu}}{G_{yv} G_{yv}^*}}\, |K| \text{ on } \mathbb{T} \tag{3.17}$$

$$D = KC^{-1} \tag{3.18}$$

- If $M\hat{Q} - U = 0$: $C = D = 0$.

If (C, D) satisfy these conditions, then $(C, D) \in \Theta_{C,D}$ and

$$J(C, D) < \varphi(Q) + \varepsilon.$$

□

Proof

Recall that

$$\Theta_{C,D}(K) = \left\{ (C, D) : DC = K, \left\| \frac{CG_{yv}}{1 - KG_{yu}} \right\|_2^2 \leq \sigma^2 - \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2 \right\}.$$

Consider $(C, D) \in \Theta_{C,D}$ and define $K = DC$. Then $(C, D) \in \Theta_{C,D}(K)$ for this choice of K. Moreover, because $T \in \mathcal{H}_2$ it follows from Lemma 3.1 that K can be written using the Youla parametrization (3.6). Since the SNR constraint (3.4) is satisfied by (C, D) it follows that $K \in \Theta_K$, where Θ_K is defined by

$$\Theta_K = \left\{ K : (3.6), \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2 < \sigma^2 \right\}.$$

The inequality in this definition is strict because it was shown earlier that equality cannot hold. It has thus been proved that

$$(C, D) \in \Theta_{C,D} \Rightarrow (C, D) \in \Theta_{C,D}(K) \text{ for some } K \in \Theta_K. \tag{3.19}$$

A lower bound will now be determined for $J(C,D)$. This will be accomplished through a series of inequalities and equalities, where each step will be explained afterwards.

$$
\begin{aligned}
&\inf_{(C,D)\in\Theta_{C,D}} J(C,D) \\
&\overset{(1)}{\geq} \inf_{K\in\Theta_K} \inf_{(C,D)\in\Theta_{C,D}(K)} \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1-KG_{yu}} \right\|_2^2 + \left\| \frac{DG_{zu}}{1-KG_{zu}} \right\|_2^2 \\
&\overset{(2)}{=} \inf_{K\in\Theta_K} \left(\left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1-KG_{yu}} \right\|_2^2 + \inf_{(C,D)\in\Theta_{C,D}(K)} \left\| \frac{DG_{zu}}{1-KG_{zu}} \right\|_2^2 \right) \\
&\overset{(3)}{\geq} \inf_{K\in\Theta_K} \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1-KG_{yu}} \right\|_2^2 + \frac{\left\| \frac{KG_{zu}G_{yv}}{(1-KG_{yu})^2} \right\|_1^2}{\sigma^2 - \left\| \frac{KG_{yu}}{1-KG_{yu}} \right\|_2^2} \\
&\overset{(4)}{=} \inf_{Q\in\Theta_Q} \| G_{zv} + G_{zu}G_{yv}(AQ+B) \|_2^2 + \frac{\| G_{zu}G_{yv}(AQ+B)(EQ+F) \|_1^2}{\sigma^2 + 1 - \| EQ+F \|_2^2} \\
&\overset{(5)}{=} \inf_{Q\in\Theta_Q} \varphi(Q)
\end{aligned}
$$

The first step follows from (3.19) and a rewriting of the functional in terms of K. In the second step, the first term of the functional has been moved out since it is constant in the inner minimization. The third step follows from application of Lemma 3.3 with

$$
\alpha = \sigma^2 - \left\| \frac{KG_{yu}}{1-KG_{yu}} \right\|_2^2 > 0, \quad S = \frac{1}{1-KG_{yu}} \in \mathcal{RH}_\infty.
$$

The fourth step follows from

$$
\left\| \frac{KG_{yu}}{1-KG_{yu}} \right\|_2^2 = \left\| \frac{1}{1-KG_{yu}} - 1 \right\|_2^2 + 1 - 1 = \left\| \frac{1}{1-KG_{yu}} \right\|_2^2 - 1,
$$

where the second equality is due to orthogonality, since G_{yu} is strictly proper, and application of the Youla parametrization, which gives

$$
\frac{K}{1-KG_{yu}} = AQ + B, \quad \frac{1}{1-KG_{yu}} = EQ + F.
$$

The fifth step follows from the definition of $\varphi(Q)$.

Now a suboptimal solution will be constructed for Problem 3.1. Suppose that $Q \in \Theta_Q$ and $\varepsilon > 0$ and let $\hat{Q} \in \Theta_Q$ be as given by Lemma 3.4 and define $K \in \mathcal{RL}_1$ by (3.16). Then $K \in \Theta_K$ and

$$\varphi(\hat{Q}) = \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} \right\|_2^2 + \frac{\left\| \frac{KG_{zu}G_{yv}}{(1 - KG_{yu})^2} \right\|_1^2}{\sigma^2 - \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2}$$

If $M\hat{Q} - U = 0$ then $K = 0$,

$$J(0,0) = \|G_{zv}\|_2^2 = \varphi(\hat{Q}) < \varphi(Q) + \varepsilon,$$

and the proof is complete.

If, on the other hand, $M\hat{Q} - U$ is not identically zero then K is not identically zero. By Lemma 3.8 there then exists an outer $C \in \mathcal{H}_2$ and $D \in \mathcal{L}_2$ such that (3.17) and (3.18) are satisfied. The lemma also says that such (C, D) satisfy

$$\left\| \frac{DG_{zu}}{1 - KG_{yu}} \right\|_2^2 = \frac{\left\| \frac{KG_{zu}G_{yv}}{(1 - KG_{yu})^2} \right\|_1^2}{\sigma^2 - \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2}$$

and

$$\left\| \frac{CG_{yv}}{1 - KG_{yu}} \right\|_2^2 \leq \sigma^2 - \left\| \frac{KG_{yu}}{1 - KG_{yu}} \right\|_2^2.$$

D, C and K satisfy the conditions of Lemma 3.2, so $T \in \mathcal{H}_2$, which implies that $(C, D) \in \Theta_{C,D}$. Moreover,

$$J(C,D) = \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} \right\|_2^2 + \left\| \frac{DG_{zu}}{1 - KG_{yu}} \right\|_2^2 = \varphi(\hat{Q}) = \varphi(Q) + \varepsilon.$$

Since ε can be made arbitrarily small this shows that (3.15) holds and hence the proof is complete. □

Remark 3.3

Theorem 3.1 shows that an ε-suboptimal solution of Problem 3.1 can be found by minimizing $\varphi(Q)$ over Θ_Q. The obtained Q may have to be perturbed so that the resulting K has no poles on the unit circle. Then C is given by a spectral factorization and D is then obtained from C. □

A by-product of Theorem 3.1 is a necessary and sufficient criterion for the existence of a stabilizing controller that satisfies the SNR constraint.

COROLLARY 3.1
There exists (C, D) that stabilize the closed loop system of Figure 3.2 subject to the SNR constraint (3.4) if and only if there exists $Q \in \mathcal{RH}_\infty$ such that

$$\|MNQ + MV\|_2^2 < \sigma^2 + 1. \tag{3.20}$$

□

REMARK 3.4
Corollary 3.1 implies that the minimum SNR compatible with stabilization of a stochastically disturbed plant by an output feedback LTI controller with two degrees of freedom can be found by minimizing the left hand side of (3.20) over $Q \in \mathcal{RH}_\infty$. The analytical condition (3.2), presented in [7], is derived from a minimization of the left hand side of (3.20). This means that the condition (3.2) is also necessary and sufficient in the present problem setting. This has been noted previously in [51].

There is no plant disturbance in the setup of [7]. In that case, the critical SNR for stabilizability will be the same regardless if the controller has one or two degrees of freedom. However, [42] considered the case when there is a plant disturbance and showed that the SNR required for stabilizability may then be larger than prescribed by (3.2) (the case when $\eta = \delta = 0$ was considered). However, the controller in [42] was assumed to only have one degree of freedom (the encoder was fixed to be a unity gain). This corollary, and Theorem 17 in [51], shows that if the controller has two degrees of freedom, then (3.2) is again a necessary and sufficient criterion for stabilizability. □

It will now be shown that the minimization of $\varphi(Q)$ over Θ_Q is a convex problem. This will be done in the same way as in Section 2.4. Recall that the functional

$$\rho(a, e) = \frac{1}{2\pi}\int_{-\pi}^{\pi} a(\omega)^2 d\omega + \frac{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} a(\omega)e(\omega)d\omega\right)^2}{\sigma^2 + 1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} e(\omega)^2 d\omega}$$

with domain

$$\Theta_\rho = \left\{(a, e) : a(\omega), e(\omega) \in \mathbb{R}\ \forall \omega, \frac{1}{2\pi}\int_{-\pi}^{\pi} e(\omega)^2 d\omega < \sigma^2 + 1\right\}$$

is convex by Lemma 2.6. Define the functional

$$\begin{aligned}\varphi_0(Q) &= \varphi(Q) - \left(\|G_{zv}\|_2^2 + 2\,\mathrm{Re}\,\langle G_{zv}, G_{zu}G_{yv}(AQ + B)\rangle\right) \\ &= \|G_{zu}G_{yv}(AQ + B)\|_2^2 + \frac{\|G_{zu}G_{yv}\,(AQ + B)\,(EQ + F)\|_1^2}{\sigma^2 + 1 - \|EQ + F\|_2^2}.\end{aligned}$$

LEMMA 3.5
Suppose $Q \in \Theta_Q$. Then $\varphi_0(Q) \leq \gamma$ if and only if there exists $(a, e) \in \Theta_\rho$ such that $\rho(a, e) \leq \gamma$ and

$$a(\omega) \geq \sqrt{G_{zu}^* G_{zu} G_{yv} G_{yv}^*}\,|AQ + B|\,, \qquad e(\omega) \geq |EQ + F| \quad \forall \omega. \tag{3.21}$$

□

PROOF
The proof is a simple modification of the proof of Lemma 2.7. □

THEOREM 3.2
The problem of minimizing $\varphi(Q)$ over Θ_Q is convex. □

PROOF
The proof is a simple modification of the proof of Theorem 2.4. □

Numerical Solution

Denote

$$\Delta(Q) = \varphi(Q) - \varphi_0(Q) = \|G_{zv}\|_2^2 + 2\,\mathrm{Re}\,\langle G_{zv}, G_{zu} G_{yv}(AQ + B)\rangle\,.$$

By Lemma 3.5, minimizing $\varphi(Q)$ over Θ_Q is equivalent to minimizing $\rho(a, e) + \Delta(Q)$ over $\Theta_\rho \times \Theta_Q$ subject to (3.21). This problem is infinite-dimensional, so the integrals are discretized for numerical solution. It will now be shown how the discretized problem can be posed as a semidefinite program.

Let $n \geq 2$ and introduce

$$\begin{aligned} &\omega_1 = 0, \quad \omega_{k+1} - \omega_k = 2\pi/n, \quad k = 1, \ldots, n-1 \\ &\hat{a} = [\, a(\omega_1) \quad a(\omega_2) \quad \ldots \quad a(\omega_n)\,]^T \\ &\hat{e} = [\, e(\omega_1) \quad e(\omega_2) \quad \ldots \quad e(\omega_n)\,]^T. \end{aligned}$$

An approximation with n grid points is then given by

$$\rho_n(\hat{a}, \hat{e}) = \frac{1}{n}\hat{a}^T\hat{a} + \frac{\left(\frac{1}{n}\hat{a}^T\hat{e}\right)^2}{\sigma^2 + 1 - \frac{1}{n}\hat{e}^T\hat{e}} \approx \rho(a, e)$$

$$\Delta_n(Q) = \|G_{zv}\|_2^2 + \frac{2}{n}\,\mathrm{Re}\sum_{k=1}^{n} \mathrm{tr}\;\left(G_{zv}^* G_{zu} G_{yv}(AQ + B)\right)\big|_{z=e^{i\omega_k}} \approx \Delta(Q).$$

By the definition of the integral it holds that

$$\lim_{n\to\infty} \rho_n(\hat{a},\hat{e}) + \Delta_n(Q) = \rho(a,e) + \Delta(Q),$$

so the minimum of the approximation can be made to come arbitrarily close to $\inf_{Q\in\Theta_Q} \varphi(Q)$ if n is chosen sufficiently large. When implementing the minimization program, Q is parametrized using a finite basis representation. The accuracy of the approximated problem obviously depends on this representation as well.

Noting that $\rho_n(\hat{a},\hat{e}) + \Delta_n(Q)$ can be written as a Schur complement and that the denominator of $\rho_n(\hat{a},\hat{e})$ is positive for sufficiently large n, it follows that $\rho_n(\hat{a},\hat{e}) + \Delta_n(Q) \le \gamma$ if and only if

$$\begin{bmatrix} \frac{1}{n}\hat{e}^T\hat{e} - \sigma^2 - 1 & \frac{1}{n}\hat{a}^T\hat{e} \\ \frac{1}{n}\hat{e}^T\hat{a} & \frac{1}{n}\hat{a}^T\hat{a} + \Delta_n(Q) - \gamma \end{bmatrix} \preceq 0,$$

or, equivalently,

$$\begin{bmatrix} n(\sigma^2+1) & 0 \\ 0 & n\gamma - n\Delta_n(Q) \end{bmatrix} - \begin{bmatrix} \hat{e} & \hat{a} \end{bmatrix}^T I \begin{bmatrix} \hat{e} & \hat{a} \end{bmatrix} \succeq 0.$$

Using Schur complement again, this is equivalent to

$$\begin{bmatrix} I & \hat{e} & \hat{a} \\ \hat{e}^T & n(\sigma^2+1) & 0 \\ \hat{a}^T & 0 & n\gamma - n\Delta_n(Q) \end{bmatrix} \succeq 0. \tag{3.22}$$

Let $g_k = \sqrt{G_{zu}(e^{i\omega_k})^* G_{zu}(e^{i\omega_k}) G_{yv}(e^{i\omega_k}) G_{yv}(e^{i\omega_k})^*}$. The constraints can then be approximated by

$$a(\omega_k) \ge g_k \left| A(e^{i\omega_k})Q(e^{i\omega_k}) + B(e^{i\omega_k}) \right|, \quad k = 1 \dots n \tag{3.23}$$

$$e(\omega_k) \ge \left| E(e^{i\omega_k})Q(e^{i\omega_k}) + F(e^{i\omega_k}) \right|, \quad k = 1 \dots n \tag{3.24}$$

$$\sigma^2 + 1 > \frac{1}{n}\sum_{k=1}^{n} e(\omega_k)^2. \tag{3.25}$$

Minimizing γ subject to (3.22)–(3.25) is a semidefinite program. The value of the approximated problem is arbitrarily close to $\inf_{Q\in\Theta_Q} \varphi(Q)$ for sufficiently large n.

A procedure for numerical solution of Problem 3.1 will now be outlined.

1. Determine a $N, M, U, V \in \mathcal{RH}_\infty$ by a coprime factorization of G_{yu} and calculate $A, B, E, F \in \mathcal{RH}_\infty$.

2. Choose n sufficiently large, determine the grid points ω_k, $k = 1 \dots n$ and solve the optimization problem of minimizing γ subject to (3.22)–(3.25). The transfer function Q is parametrized with a finite basis representation, for example as an FIR filter. If the problem is infeasible it could mean that a larger σ^2 is needed to stabilize the plant. This can be checked analytically using the condition in [7]. If σ^2 is sufficiently large according to this condition, the problem could still become infeasible if n is too small or Q is too coarsely parametrized.

3. If $NQ + V$ has zeros on the unit circle, determine a small perturbation $\hat{Q}$ of Q as outlined by Lemma 3.4.

4. Determine K from (3.14).

5. Use a finite basis approximation $A(\omega)$ of CC^*, for example the parametrization (2.25), and fit $A(\omega)$ to the right hand side of (3.17), for example by minimizing the mean squared deviation.

6. Perform a spectral factorization of $A(\omega)$, choosing C as the stable and minimum phase spectral factor.

7. Let $D = KC^{-1}$.

Special Cases and Examples

Feedback Control of SISO Plant with SNR Constraint Consider the system in Figure 3.5. The SISO plant represents a special case where

$$G(z) = \begin{bmatrix} G_{zv}(z) & G_{zu}(z) \\ G_{yv}(z) & G_{yu}(z) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} P(z).$$

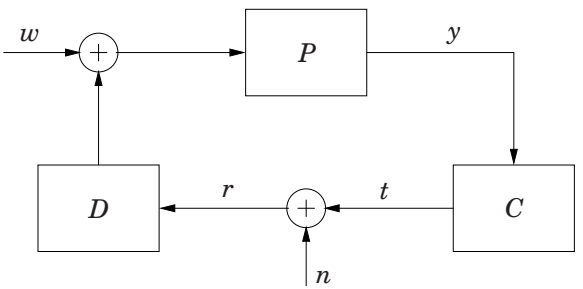

Figure 3.5 Control of a SISO plant over an AWN channel.

The functional to be minimized can in this case be written

$$\begin{aligned}\varphi(Q) &= \|P + P(AQ + B)\|_2^2 + \frac{\|P\,(AQ + B)\,(EQ + F)\|_1^2}{\sigma^2 + 1 - \|EQ + F\|_2^2} \\ &= \left\|N^2Q + NV\right\|_2^2 + \frac{\left\|\left(N^2Q + NV\right)(MNQ - NU)\right\|_1^2}{\sigma^2 - \|MNQ - NU\|_2^2}\end{aligned}$$

EXAMPLE 3.1
Consider the plant $G = 1/(z(z-2))$. It has one unstable pole at $z = 2$ and a one-sample time delay. Using the condition (3.2) from [7], it is determined that stabilization is possible for $\sigma^2 > 12$. ($\eta = 0$, since there are no non-minimum phase zeros, and $\delta = 9$, because of the relative degree, which is 2, and the location of the unstable pole. For details, see [7]).

A controller was determined for various values of σ^2, using the previously outlined algorithm. The optimization was performed in Matlab, using the toolboxes Yalmip [30] and SeDuMi [55]. In the optimization program, $n = 629$ grid points were used and Q was parametrized as an FIR filter with length 20. The plant output variance is plotted in Figure 3.6 for a number of different σ^2. It can be seen that the variance grows unbounded as σ^2 approaches 12 and the feedback system comes closer to instability. □

Incorporating the Control Signal Variance In the previous example, only the plant output variance was minimized. Frequently, it is desirable to include the control signal variance in the minimization, using a criterion of the form

$$\lim_{k\to\infty} \mathrm{E}(z(k)^T z(k)) + \rho \mathrm{E}(u(k)^2), \tag{3.26}$$

where the parameter $\rho \geq 0$ determines the relative weight of the variances in the minimization.

In general, given a plant

$$G(z) = \begin{bmatrix} G_{zv}(z) & G_{zu}(z) \\ G_{yv}(z) & G_{yu}(z) \end{bmatrix},$$

it is possible to minimize (3.26) by instead performing the control design for the plant

$$\hat{G}(z) = \begin{bmatrix} G_{zv}(z) & G_{zu}(z) \\ 0 & \sqrt{\rho} \\ G_{yv}(z) & G_{yu}(z) \end{bmatrix}.$$

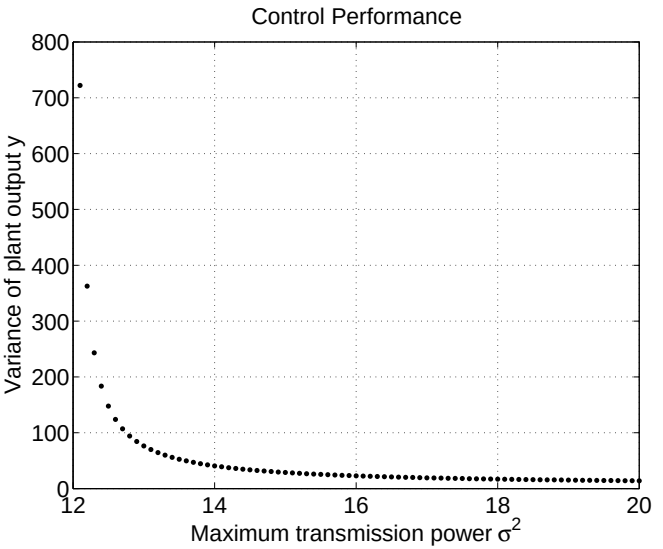

Figure 3.6 Variance of the plant output y as a function of the SNR/maximum allowed transmission power σ^2, for the plant $G = 1/(z(z-2))$. The variance grows unbounded as σ^2 approaches the lower limit for stabilization.

Real-Time Coding for Noisy Channel Consider again Problem 2.1 in Section 2.2. Comparing the block diagram in Figure 2.2 with the one in Figure 3.2, Problem 2.1 is seen to be a special case of the problem studied in this section, provided that

$$G(z) = \begin{bmatrix} G_{zv}(z) & G_{zu}(z) \\ G_{yv}(z) & G_{yu}(z) \end{bmatrix} = \begin{bmatrix} PF & 0 & -1 \\ F & G & 0 \end{bmatrix}.$$

Since $G_{yu} = 0$, a coprime factorization is given by $N = 0$, $M = 1$, $U = 0$ and $V = 1$. Then $Q = K$ and the corresponding functional to minimize is

$$\varphi(K) = \| [\, PF - FK \quad -GK \,] \|_2^2 + \frac{1}{\sigma^2} \| [\, F \quad G \,] \, K \|_1^2,$$

which is equal to the one defined by (2.14) in Section 2.2. Moreover, the SNR constraint reduces to $\sigma^2 + 1 > 1$, which trivially holds for any K. It is easily verified that the optimality criteria for C and D are also equivalent.

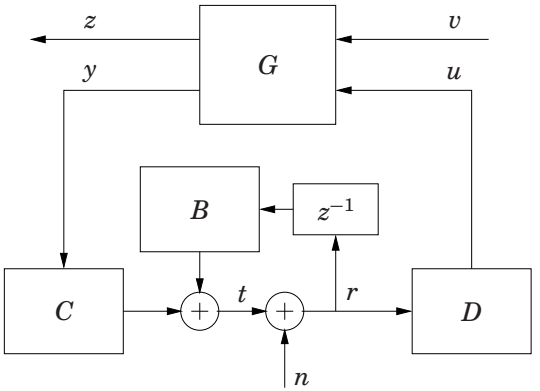

Figure 3.7 Feedback system with noisy communication channel with feedback. The objective is to design the controller components B, C and D so that the system is stabilized and the variance of z is minimized under an SNR constraint.

3.3 Using Channel Feedback

It will now be assumed that the channel has noise-free feedback. This means that the encoder has access to the channel output and may use it to modify the transmitted signal. Hence, the controller now has an additional degree of freedom. Note that there are now two feedback loops in the system. The problem may be referred to as feedback control over a channel with feedback.

Problem Formulation and Assumptions

Make the same assumptions as in Section 3.2 and assume additionally that the encoder has access to the channel output, delayed by one sample. Assume also that G_{yu} has no poles on the unit circle. That is, $G_{yu} \in \mathcal{RL}_\infty$.

Since the controller is linear, it can be assumed without additional loss of generality that it has the structure illustrated in Figure 3.7. The encoder now consists of C and the feedback part B. Note that the feedback part of the encoder is not structured as in Section 2.4. There, the transmitted signal was subtracted from the input to B, in order to avoid forming a loop. The present structure is essentially equivalent.

The objective is to find causal LTI systems B, C and D that make the system internally stable, satisfy the SNR constraint (3.3) and minimize

$$\lim_{k\to\infty} \mathrm{E}(z(k)^T z(k)).$$

Modifying the expressions in Problem 3.1 to take the channel feedback

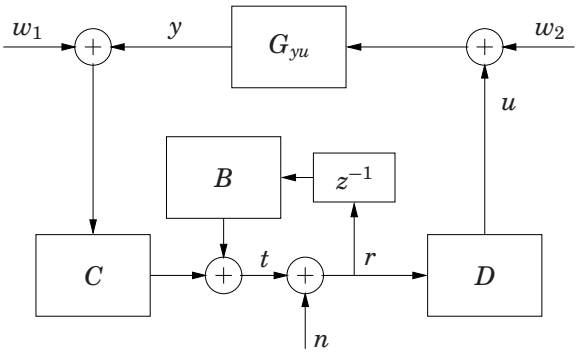

Figure 3.8 Block diagram for internal stability analysis in the case with channel feedback.

path into account, the following problem formulation is obtained.

PROBLEM 3.2

$$\underset{B,C,D}{\text{minimize}} \left\| G_{zv} + \frac{DCG_{zu}G_{yv}}{1 - Bz^{-1} - DCG_{yu}} \right\|_2^2 + \left\| \frac{DG_{zu}}{1 - Bz^{-1} - DCG_{yu}} \right\|_2^2$$

subject to

$$\left\| \frac{CG_{yv}}{1 - Bz^{-1} - DCG_{yu}} \right\|_2^2 + \left\| \frac{Bz^{-1} + DCG_{yu}}{1 - Bz^{-1} - DCG_{yu}} \right\|_2^2 \leq \sigma^2 \tag{3.27}$$

while achieving internal stability of the feedback system. □

Due to technical reasons, only solutions where $B \in \mathcal{R}$ and $DC \in \mathcal{R}$ are considered. This may exclude the possibility of achieving the minimum value, but the infimum can still be arbitrarily well approximated by rational functions. It is not explicitly required that C and D are proper but it will be seen also here that the solution can be constructed so that $C \in \mathcal{H}_2$ is outer. Then C, C^{-1} are proper, and $D = (DC)C^{-1}$ is also proper.

Internal Stability

The factorization approach used previously has to be modified to handle the channel feedback and the additional design variable B. As before, K is defined as the open loop transfer function from y to u. In this structure, this means that

$$K = D(1 - Bz^{-1})^{-1}C. \tag{3.28}$$

Introduce also the transfer function

$$S = \frac{1}{1 - Bz^{-1} - DCG_{yu}} = \frac{(1 - Bz^{-1})^{-1}}{1 - D(1 - Bz^{-1})^{-1}CG_{yu}} = \frac{(1 - Bz^{-1})^{-1}}{1 - KG_{yu}}. \tag{3.29}$$

It is noted that if B is proper then

$$\lim_{z\to\infty} S(z) = \lim_{z\to\infty} \frac{1}{(1 - B(z)z^{-1})(1 - K(z)G_{yu}(z))} = 1, \tag{3.30}$$

since G_{yu} is strictly proper. It will thus be assumed that S satisfies (3.30). This condition guarantees that the feedback system is well-posed, since both loops then contain a strictly proper transfer function.

Internal stability means that all closed loop transfer functions are in $\mathcal{H}_2$. Following the same reasoning as in [65], it is concluded that internal stability of the systems in Figure 3.7 and Figure 3.8 are equivalent. The latter can be represented by the closed loop map T, defined by

$$\begin{bmatrix} y \\ t \\ u \end{bmatrix} = T \begin{bmatrix} w_1 \\ w_2 \\ n \end{bmatrix}.$$

It follows that the system in Figure 3.7 is internally stable if and only if

$$T = \begin{bmatrix} SDCG_{yu} & S(1 - Bz^{-1})G_{yu} & SDG_{yu} \\ SC & SCG_{yu} & S(Bz^{-1} + DCG_{yu}) \\ SDC & SDCG_{yu} & SD \end{bmatrix} \in \mathcal{H}_2. \tag{3.31}$$

The following two lemmas will give necessary and sufficient conditions for internal stability, respectively.

Lemma 3.6
Suppose that $T \in \mathcal{H}_2$, that $G_{yu} = NM^{-1}$ is a coprime factorization over $\mathcal{RH}_\infty$ and that $U, V \in \mathcal{RH}_\infty$ satisfy the Bezout identity $VM + UN = 1$. Then $S \in \mathcal{RH}_\infty$ and

$$K = \frac{MQ - U}{NQ + V}, \quad Q \in \mathcal{RH}_\infty. \tag{3.32}$$

Moreover, if G_{yu} has a pole of multiplicity n at z, where $|z| \geq 1$, then S has a zero of multiplicity greater than or equal to n at z. □

PROOF

It is seen in (3.31) that $S(Bz^{-1} + DCG_{yu}) \in \mathcal{H}_2$ and since

$$S(Bz^{-1} + DCG_{yu}) = S(Bz^{-1} - 1 + 1) + \frac{KG_{yu}}{1 - KG_{yu}} = S - 1, \qquad (3.33)$$

it follows that $S \in \mathcal{H}_2$. Since S is rational it has no poles on or outside the unit circle and thus $S \in \mathcal{RH}_\infty$.

It also follows directly from (3.31) that

$$\begin{aligned}
\frac{G_{yu}}{1 - KG_{yu}} &= S(1 - Bz^{-1})G_{yu} \in \mathcal{H}_2, \\
\frac{K}{1 - KG_{yu}} &= SDC \in \mathcal{H}_2, \\
\frac{KG_{yu}}{1 - KG_{yu}} &= \frac{1}{1 - KG_{yu}} - 1 = SDCG_{yu} \in \mathcal{H}_2.
\end{aligned}$$

Since these transfer functions are rational and have no poles on or outside the unit circle it follows that

$$\begin{bmatrix} 1 & -K \\ -G_{yu} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{1}{1 - KG_{yu}} & \dfrac{K}{1 - KG_{yu}} \\ \dfrac{G_{yu}}{1 - KG_{yu}} & \dfrac{1}{1 - KG_{yu}} \end{bmatrix} \in \mathcal{RH}_\infty, \qquad (3.34)$$

It is well-known that the set of K satisfying (3.34) can be parametrized using the Youla parametrization of all stabilizing controllers [65]. That is, K can be written as in (3.32).

To prove the final statement, introduce the function

$$\Upsilon(X, z) = \begin{cases} n, & X \text{ has a zero of multiplicity } n \text{ at } z \\ 0, & X \text{ has no pole or zero at } z \\ -n, & X \text{ has a pole of multiplicity } n \text{ at } z. \end{cases}$$

Suppose that (3.31) holds and $\Upsilon(G_{yu}, z) < 0$ for some $|z| \geq 1$. Then

$$\frac{G_{yu}}{1 - KG_{yu}} \in \mathcal{RH}_\infty \Rightarrow \Upsilon(1 - KG_{yu}, z) \leq \Upsilon(G_{yu}, z).$$

The definition of S in (3.29) implies that

$$\Upsilon((1 - Bz^{-1})^{-1}, z) = \Upsilon(S, z) + \Upsilon(1 - KG_{yu})$$

KG_{yu} and $1 - KG_{yu}$ have the same poles, so

$$\Upsilon(KG_{yu}, z) = \Upsilon(1 - KG_{yu}, z).$$

From $SCG_{yu} \in \mathcal{H}_2$ and $SDG_{yu} \in \mathcal{H}_2$ it follows that $S^2DCG_{yu}^2 \in \mathcal{H}_1$. Moreover, $S^2DCG_{yu}^2 \in \mathcal{RH}_\infty$ since it is rational and has no poles on or outside the unit circle. Thus

$$\Upsilon(DC, z) \geq -2\Upsilon(S, z) - 2\Upsilon(G_{yu}, z)$$

Putting these relationships together gives that

$$\begin{aligned}\Upsilon(KG_{yu}, z) &= \Upsilon(DC, z) + \Upsilon((1 - Bz^{-1})^{-1}, z) + \Upsilon(G_{yu}, z) \\ &\geq -2\Upsilon(S, z) - 2\Upsilon(G_{yu}, z) + \Upsilon(S, z) + \Upsilon(KG_{yu}, z) + \Upsilon(G_{yu}, z),\end{aligned}$$

which gives

$$\Upsilon(S, z) \geq -\Upsilon(G_{yu}, z). \tag{3.35}$$

This means that if G_{yu} has an unstable pole of multiplicity n, then S will have a zero with at least the same multiplicity in the same location. □

LEMMA 3.7
Suppose that

$$K = D(1 - Bz^{-1})^{-1}C = \frac{MQ - U}{NQ + V}, \quad Q \in \mathcal{RH}_\infty, \tag{3.36}$$

where $G_{yu} = NM^{-1}$ is a coprime factorization over $\mathcal{RH}_\infty$ and $U, V \in \mathcal{RH}_\infty$ satisfy the Bezout identity $VM + UN = 1$. Suppose also that $C \in \mathcal{H}_2$ is outer, $D \in \mathcal{L}_2$,

$$S = \frac{(1 - Bz^{-1})^{-1}}{1 - KG_{yu}} \in \mathcal{RH}_\infty,$$

and that S satisfies (3.35). Then $T \in \mathcal{H}_2$. □

PROOF
It follows from (3.36) that

$$\begin{aligned}SDCG_{yu} &= \frac{KG_{yu}}{1 - KGyu} \in \mathcal{RH}_\infty \\ S(1 - Bz^{-1})G_{yu} &= \frac{G_{yu}}{1 - KGyu} \in \mathcal{RH}_\infty \\ SDC &= \frac{K}{1 - KGyu} \in \mathcal{RH}_\infty.\end{aligned}$$

Moreover,

$$SDG_{yu} = \frac{KG_{yu}}{1 - KG_{yu}} C^{-1}.$$

Since $S \in \mathcal{RH}_\infty$ it follows from (3.35) that $SG_{yu} \in \mathcal{RH}_\infty$. This gives $SDG_{yu} \in \mathcal{L}_2$. Since C is outer, application of Lemma 1.3 gives that $SDG_{yu} \in \mathcal{H}_2$. $SC \in \mathcal{H}_2$ since $S \in \mathcal{RH}_\infty$ and $C \in \mathcal{H}_2$. It follows from $SG_{yu} \in \mathcal{RH}_\infty$ that $SCG_{yu} \in \mathcal{H}_2$. From (3.33) it is seen that

$$S(Bz^{-1} + DCG_{yu}) = S - 1 \in \mathcal{RH}_\infty.$$

Finally, $SD \in \mathcal{L}_2$ and thus

$$SD = \frac{K}{1 - KG_{yu}} C^{-1} \in \mathcal{H}_2$$

due to Lemma 1.3.

Noting that $\mathcal{RH}_\infty \subseteq \mathcal{H}_2$, it has then been proved that all elements of T are in $\mathcal{H}_2$ and thus $T \in \mathcal{H}_2$. □

Optimal Factorization

Suppose for now that K and S are given and that they satisfy (3.30) and the conditions necessary for stability derived in Lemma 3.6. Then B is proper and can be obtained from (3.29). C and D are still, however, left to determine.

Rewriting the objective and the SNR constraint in terms of K and S gives

$$\left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} \right\|_2^2 + \|SDG_{zu}\|_2^2 \tag{3.37}$$

and

$$\|SCG_{yv}\|_2^2 + \|S - 1\|_2^2 \leq \sigma^2. \tag{3.38}$$

By (3.30), $S - 1$ is strictly proper and thus orthogonal to 1. Hence

$$\|S - 1\|_2^2 + 1 - 1 = \|S\|_2^2 - 1. \tag{3.39}$$

Thus, the SNR constraint will be impossible to satisfy unless S satisfies

$$\alpha = \sigma^2 + 1 - \|S\|_2^2 \geq 0.$$

If $\alpha = 0$ then S is non-zero and $\|SCG_{yv}\|_2^2 = 0$. Since G_{yv} is non-zero by assumption, this implies that $C = 0$ and $K = 0$. This is not possible if G_{yu}

is unstable, since (3.34) would be violated. In the case that G_{yu} is stable, the objective becomes

$$\|G_{zv}\|_2^2 + \|SDG_{zu}\|_2^2$$

and it is clear that it is optimal to let $D = 0$. In that case, the objective does not depend on S and it gives no loss to set $S = 1$, giving $\alpha = \sigma^2 > 0$. Therefore, it can be assumed without loss of generality that $\alpha > 0$.

The objective of the optimal factorization problem is to find C and D such that (3.37) is minimized subject to (3.38) and $K = D(1 - Bz^{-1})^{-1}C$. Stability considerations are temporarily disregarded and will be handled later. Thus, the set of feasible (C, D), parametrized by K and S, is defined as

$$\Theta_{C,D}(K,S) = \{(C,D) : \|SCG_{yv}\|_2^2 \le \alpha, K(1 - Bz^{-1}) = DC, (3.29)\}.$$

(The final condition in $\Theta_{C,D}$ is included to define B by K and S.)

It is noted that the first term in (3.37) is constant and that the second term is a weighted norm of D. In the left hand side of (3.38), the first term is a weighted norm of C and the second term is constant. This is similar to the observations in Section 3.2 and means that the optimal factorization problem can be formulated and solved in the same manner as was done there. The solution to the optimal factorization problem is given by the following lemma.

Lemma 3.8—Optimal Factorization, Feedback Control using Channel Feedback Case
Suppose that $\alpha > 0$, $S \in \mathcal{RH}_\infty$, $K \in \mathcal{R}$, $B \in \mathcal{R}$ satisfies (3.29) and that $G_{zu}^* G_{zu} \in \mathcal{RL}_\infty$ and $G_{yv} G_{yv}^* \in \mathcal{RL}_\infty$ have no zeros on $\mathbb{T}$. Then

$$\inf_{(C,D)\in\Theta_{C,D}(K,S)} \|SDG_{zu}\|_2^2 \ge \frac{1}{\alpha} \left\|K(1 - Bz^{-1})S^2 G_{zu} G_{yv}\right\|_1^2. \tag{3.40}$$

Suppose furthermore that $K \in \mathcal{RL}_1$ satisfies (3.32) and $B \in \mathcal{RH}_\infty$. Then there exists $(C, D) \in \Theta_{C,D}(K, S)$ with $C \in \mathcal{H}_2$ outer and $D \in \mathcal{L}_2$ such that the minimum is attained and (3.40) holds with equality. If K is not identically zero, then (C, D) is optimal if and only if $DC = K(1 - Bz^{-1})$ and

$$|C|^2 = \frac{\alpha}{\|K(1 - Bz^{-1})S^2 G_{zu} G_{yv}\|_1} \sqrt{\frac{G_{zu}^* G_{zu}}{G_{yv} G_{yv}^*}} \left|K(1 - Bz^{-1})\right| \text{ on } \mathbb{T}. \tag{3.41}$$

If $K = 0$, then the minimum is achieved by $D = 0$ and any C that satisfies $\|SCG_{yv}\|_2^2 \le \alpha$. □

PROOF

Suppose first that $K = 0$. Then the right hand side of (3.40) is 0. Letting $D = 0$ gives $\|SDG_{zu}\|_2^2 = 0$ and it is clear that $(C, D) \in \Theta_{C,D}$ if C is as stated.

Thus, it can now be assumed that K is not identically zero. Then $K(1 - Bz^{-1})$ is not identically zero, since B is proper, and hence C is not identically zero and $D = K(1 - Bz^{-1})C^{-1}$.

By assumption both $G_{zu}^* G_{zu}$ and $G_{yv} G_{yv}^*$ are positive on the unit circle. Since these functions are rational this implies that

$$\exists \varepsilon > 0 \text{ such that } G_{zu}^* G_{zu} \geq \varepsilon \text{ and } G_{yv} G_{yv}^* \geq \varepsilon, \text{ on } \mathbb{T}. \tag{3.42}$$

Thus by Theorem 1.3 there exist scalar minimum phase transfer functions $\hat{G}_{zu}, \hat{G}_{yv} \in \mathcal{H}_2$ such that

$$G_{zu}^* G_{zu} = \hat{G}_{zu}^* \hat{G}_{zu}, \qquad G_{yv} G_{yv}^* = \hat{G}_{yv} \hat{G}_{yv}^*.$$

Now, $\|SCG_{yv}\|_2^2 \leq \alpha$ and Cauchy-Schwarz's inequality gives

$$\begin{aligned}
\|DSG_{zu}\|_2^2 &= \left\| K(1 - Bz^{-1})C^{-1} S \hat{G}_{zu} \right\|_2^2 \\
&\geq \frac{\left\| SC\hat{G}_{yv} \right\|_2^2}{\alpha} \left\| K(1 - Bz^{-1})C^{-1} S \hat{G}_{zu} \right\|_2^2 \\
&\geq \frac{1}{\alpha} \left\langle \left| SC\hat{G}_{yv} \right|, \left| K(1 - Bz^{-1})C^{-1} S \hat{G}_{zu} \right| \right\rangle^2 \\
&= \frac{1}{\alpha} \left\| K(1 - Bz^{-1}) S^2 \hat{G}_{zu} \hat{G}_{yv} \right\|_1^2 \\
&= \frac{1}{\alpha} \left\| K(1 - Bz^{-1}) S^2 G_{zu} G_{yv} \right\|_1^2.
\end{aligned}$$

This proves the lower bound (3.40).

Equality holds if and only if $|K(1 - Bz^{-1})C^{-1} S \hat{G}_{zu}|$ and $|SC\hat{G}_{yv}|$ are proportional on the unit circle and $\|SCG_{yv}\|_2^2 = \alpha$. It is easily verified that this is equivalent to (3.41). Thus, (C, D) achieves the lower bound if and only if $D = K(1 - Bz^{-1})C^{-1}$ and (3.41) holds, since these conditions imply that $(C, D) \in \Theta_{C,D}(K, S)$.

Under the additional assumptions that $K \in \mathcal{RL}_1$ satisfies (3.32) and $B \in \mathcal{RH}_\infty$, it will now be shown that there exists such $(C, D) \in \mathcal{H}_2 \times \mathcal{L}_2$ with C outer. Since K satisfies (3.32) with $M, N, Q, U, V \in \mathcal{RH}_\infty$ it holds that

$$\log \left| K(1 - Bz^{-1}) \right| = \log |MQ - U| - \log |NQ + V| + \log \left| 1 - Bz^{-1} \right|$$

By Lemma 1.5, all three terms in the right hand side are $\mathcal{L}_1$ functions and thus $\log\left|K(1-Bz^{-1})\right| \in \mathcal{L}_1$. It follows from (3.42) and the boundedness of $\hat{G}_{yv}$ and $\hat{G}_{zu}$ on $\mathbb{T}$ that

$$\int_{-\pi}^{\pi} \log\left|\frac{\hat{G}_{zu}}{\hat{G}_{yv}}K(1-Bz^{-1})\right| d\omega > -\infty$$

and

$$\left|\frac{\hat{G}_{zu}}{\hat{G}_{yv}}K(1-Bz^{-1})\right| \in \mathcal{L}_1.$$

Then by Theorem 1.3 there exists an outer function $C \in \mathcal{H}_2$ such that (3.41) holds. Also, $D = K(1-Bz^{-1})C^{-1} \in \mathcal{L}_2$ since

$$\begin{aligned}\|D\|_2^2 &= \left\|K(1-Bz^{-1})C^{-1}\right\|_2^2 \\ &= \frac{1}{\alpha}\left\|K(1-Bz^{-1})S^2G_{zu}G_{yv}\right\|_1 \left\|\frac{K(1-Bz^{-1})\hat{G}_{yv}}{\hat{G}_{zu}}\right\|_1 < \infty.\end{aligned}$$

□

Remark 3.5
Optimal D will satisfy

$$|D|^2 = \frac{\left\|K(1-Bz^{-1})S^2G_{zu}G_{yv}\right\|_1}{\alpha}\sqrt{\frac{G_{yv}G_{yv}^*}{G_{zu}^*G_{zu}}}\left|K(1-Bz^{-1})\right| \text{ on } \mathbb{T}.$$

□

Equivalent Convex Problem

Define the objective functional

$$J(B,C,D) = \left\|G_{zv} + \frac{DCG_{zu}G_{yv}}{1-Bz^{-1}-DCG_{yu}}\right\|_2^2 + \left\|\frac{DG_{zu}}{1-Bz^{-1}-DCG_{yu}}\right\|_2^2$$

and the feasible set

$$\Theta_{B,C,D} = \{(B,C,D) : B, DC \in \mathcal{R}, (3.27), T \in \mathcal{H}_2\},$$

consisting of all controllers that stabilize the plant under the SNR constraint. Just as in the previous section, the minimization will be performed

over Q, which parametrizes all K that satisfy the necessary conditions for stability. Stability also requires S to have zeros where the plant has unstable poles. In order to deal with this interpolation constraint, the transfer function X is introduced, from which S can be obtained by multiplication with all-pass transfer functions.

It will be shown that minimization of $J(B,C,D)$ over $\Theta_{B,C,D}$ can be performed by minimizing the convex functional

$$\varphi(Q,X) = \left\| G_{zv} + G_{zu}G_{yv}(M^2Q - MU) \right\|_2^2 + \frac{\left\| G_{zu}G_{yv}(M^2Q - MU)X \right\|_1^2}{\sigma^2 + 1 - \|X\|_2^2}$$

over the convex set

$$\Theta_{Q,X} = \{(Q,X) : Q, X \in \mathcal{RH}_\infty, \|X\|_2^2 < \sigma^2 + 1, \lim_{z\to\infty} X(z) = \prod_{\lambda\in\Lambda(G_{yu})} |\lambda|\},$$

where $\Lambda(G_{yu})$ is the set of unstable poles of G_{yu},

$$\Lambda(G_{yu}) = \{z : |z| > 1,\ z \text{ is a pole of } G_{yu}\}.$$

The $(Q,X) \in \Theta_{Q,X}$ obtained from minimizing $\varphi(Q,X)$ will be used to construct $(B,C,D) \in \Theta_{B,C,D}$. However, this will not be possible for (Q,X) such that the corresponding K has poles on $\mathbb{T}$ or X has zeros on $\mathbb{T}$. For such (Q,X) a small perturbation can then be applied first. This will result in an increased cost, but this increase can be made arbitrarily small. That this is possible is established by the following lemma.

Lemma 3.9
Suppose $(Q,X) \in \Theta_{Q,X}$ and $\varepsilon > 0$. Then there exists $(\hat{Q},\hat{X}) \in \Theta_{Q,X}$ such that $\hat{X}$ has no zeros for $|z| \geq 1$,

$$K = \frac{M\hat{Q} - U}{N\hat{Q} + V} \in \mathcal{RL}_1,$$

and

$$\varphi(\hat{Q},\hat{X}) < \varphi(Q,X) + \varepsilon.$$

□

Proof
$\hat{Q}$ is obtained by a perturbation using the same technique as in the proof of Lemma 3.4 (that proof is found in Appendix A). If X has zeros on the unit circle they can also be moved in the same way. X will then be

perturbed into $X + x_0 + x_1 z^{-1}$, where $x_0, x_1 \in \mathbb{R}$ are small. In general, this perturbation does not satisfy the condition on the limit of $X(z)$ as $z \to \infty$. Therefore, introduce

$$\widetilde{X} = \frac{X(\infty)}{X(\infty) + x_0} \left(X + x_0 + x_1 z^{-1} \right),$$

where

$$X(\infty) = \lim_{z \to \infty} X(z) = \prod_{\lambda \in \Lambda(G_{yu})} |\lambda|.$$

Then $\widetilde{X}$ has the same zeros as $X + x_0 + x_1 z^{-1}$ and $\lim_{z \to \infty} \widetilde{X}(z) = X(\infty)$.

Since both the perturbations of Q and X can be made small, and since $|x_0|$ is small, it holds that $\varphi(\hat{Q}, \widetilde{X}) < \varphi(Q, X) + \varepsilon$ and $\left\| \widetilde{X} \right\|_2^2 < \sigma^2 + 1$.

The non-minimum phase zeros of $\widetilde{X}$ will now be mirrored into the unit disk by multiplication with all-pass factors. This will decrease $|\widetilde{X}|$ on $\mathbb{T}$ and thus decrease the value of φ, which shows that it is suboptimal to have non-minimum phase zeros in X when minimizing φ. Consider the set of non-minimum phase zeros of $\widetilde{X}$,

$$\Omega = \{ z : |z| > 1, \widetilde{X}(z) = 0 \}$$

and let

$$\hat{X} = \widetilde{X} \prod_{\zeta \in \Omega} \frac{z\zeta - 1}{(z - \zeta)\zeta}.$$

Then

$$\lim_{z \to \infty} \hat{X}(z) = \lim_{z \to \infty} \widetilde{X}(z) = X(\infty)$$

and

$$|\hat{X}| = |\widetilde{X}| \prod_{\zeta \in \Omega} \frac{1}{\zeta} \leq |\widetilde{X}|.$$

Hence $\varphi(\hat{Q}, \hat{X}) \leq \varphi(\hat{Q}, \widetilde{X}) < \varphi(Q, X) + \varepsilon$ and $(\hat{Q}, \hat{X}) \in \Theta_{Q,X}$. □

THEOREM 3.3

Suppose that $\sigma^2 > 0$, $G_{yu} \in \mathcal{RL}_\infty$, $G_{yu} = NM^{-1}$ is a coprime factorization over $\mathcal{RH}_\infty$, $U, V \in \mathcal{RH}_\infty$ satisfy the Bezout identity $VM + UN = 1$ and that $G_{zu}^* G_{zu} \in \mathcal{RL}_\infty$ and $G_{yv} G_{yv}^* \in \mathcal{RL}_\infty$ have no zeros on $\mathbb{T}$. Then

$$\inf_{(B,C,D) \in \Theta_{B,C,D}} J(B, C, D) = \inf_{Q, X \in \Theta_{Q,X}} \varphi(Q, X). \tag{3.43}$$

Furthermore, suppose $(Q, X) \in \Theta_{Q,X}$ and $\varepsilon > 0$. Let $(\hat{Q}, \hat{X}) \in \Theta_{Q,X}$ be as in Lemma 3.9. Then there exists (B, C, D) such that the following conditions hold:

- If $M\hat{Q} - U$ is not identically zero: $(B, C, D) \in \mathcal{R} \times \mathcal{H}_2 \times \mathcal{L}_2$, where C is outer and

$$K = \frac{M\hat{Q} - U}{N\hat{Q} + V} \tag{3.44}$$

$$S = \hat{X} \prod_{\lambda \in \Lambda(G_{yu})} \frac{z - \lambda}{z\lambda^* - 1} \tag{3.45}$$

$$B = z\left(1 - \frac{1}{S(1 - KG_{yu})}\right) \tag{3.46}$$

$$|C|^2 = \frac{\sigma^2 + 1 - \|S\|_2^2}{\|K(1 - Bz^{-1})S^2 G_{zu} G_{yv}\|_1} \sqrt{\frac{G_{zu}^* G_{zu}}{G_{yv} G_{yv}^*}} \left|K(1 - Bz^{-1})\right| \text{ on } \mathbb{T} \tag{3.47}$$

$$D = K(1 - Bz^{-1})C^{-1}. \tag{3.48}$$

- If $M\hat{Q} - U = 0$: $B = C = D = 0$.

If (B, C, D) satisfy these conditions, then $(B, C, D) \in \Theta_{B,C,D}$ and

$$J(B, C, D) < \varphi(Q, X) + \varepsilon.$$

□

Proof

Recall that

$$\Theta_{C,D}(K, S) = \{(C, D) : \|SCG_{yv}\|_2^2 \leq \alpha, K(1 - Bz^{-1}) = DC, (3.29)\},$$

where $\alpha = \sigma^2 - \|S - 1\|_2^2$.

Consider $(B, C, D) \in \Theta_{B,C,D}$ and define K and S according to (3.28) and (3.29). Rewriting the SNR constraint (3.27) in terms of S gives that $\|SCG_{yv}\|_2^2 \leq \alpha$. Thus $(C, D) \in \Theta_{C,D}(K, S)$ for this choice of (K, S).

Moreover, with this choice of (K, S), it follows that $\lim_{z\to\infty} S(z) = 1$ since B is proper. Because $T \in \mathcal{H}_2$ it follows from Lemma 3.6 that $S \in \mathcal{RH}_\infty$ will have zeros according to (3.35) and that K can be written using the Youla parametrization (3.32). Since the SNR constraint (3.27) is satisfied by (B, C, D) it follows that $\|S - 1\|_2^2 \leq \sigma^2$. Thus $(K, S) \in \Theta_{K,S}$, where $\Theta_{K,S}$ is defined by

$$\Theta_{K,S} = \{(K, S) : S \in \mathcal{RH}_\infty, \|S - 1\|_2^2 \leq \sigma^2, \lim_{z\to\infty} S(z) = 1, (3.35), (3.32)\}.$$

It has thus been proved that

$$(B,C,D) \in \Theta_{B,C,D} \Rightarrow (C,D) \in \Theta_{C,D}(K,S) \text{ for some } (K,S) \in \Theta_{K,S}. \tag{3.49}$$

As discussed previously, the case $\|S-1\|_2^2 = \sigma^2$ can be disregarded without loss of generality. Removing such S from $\Theta_{K,S}$ gives the set

$$\widetilde{\Theta}_{K,S} = \{(K,S) : (K,S) \in \Theta_{K,S}, \|S-1\|_2^2 < \sigma^2\}.$$

Finally, parametrizing over the Youla parameter $Q \in \mathcal{RH}_\infty$ instead of K gives the set

$$\Theta_{Q,S} = \{(Q,S) : Q,S \in \mathcal{RH}_\infty, \|S\|_2^2 < \sigma^2 + 1, \lim_{z\to\infty} S(z) = 1, (3.35)\}.$$

A lower bound will now be determined for $J(B,C,D)$. This will be accomplished through a series of inequalities and equalities, where each step will be explained afterwards.

$$\begin{aligned}
&\inf_{(B,C,D)\in\Theta_{B,C,D}} J(B,C,D) \\
&\overset{(1)}{\geq} \inf_{(K,S)\in\Theta_{K,S}} \inf_{(C,D)\in\Theta_{C,D}(K,S)} \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1-KG_{yu}} \right\|_2^2 + \|SDG_{zu}\|_2^2 \\
&\overset{(2)}{=} \inf_{(K,S)\in\widetilde{\Theta}_{K,S}} \left(\left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1-KG_{yu}} \right\|_2^2 + \inf_{(C,D)\in\Theta_{C,D}(K,S)} \|SDG_{zu}\|_2^2 \right) \\
&\overset{(3)}{\geq} \inf_{(K,S)\in\widetilde{\Theta}_{K,S}} \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1-KG_{yu}} \right\|_2^2 + \frac{\left\| \frac{KG_{zu}G_{yv}}{1-KG_{yu}} S \right\|_1^2}{\sigma^2 - \|S-1\|_2^2} \\
&\overset{(4)}{=} \inf_{(Q,S)\in\Theta_{Q,S}} \left\| G_{zv} + G_{zu}G_{yv}(M^2Q - MU) \right\|_2^2 + \frac{\left\| G_{zu}G_{yv}(M^2Q - MU)S \right\|_1^2}{\sigma^2 + 1 - \|S\|_2^2} \\
&\overset{(5)}{=} \inf_{(Q,X)\in\Theta_{Q,X}} \varphi(Q,X)
\end{aligned}$$

The first step follows from (3.49) and a rewriting of the functional in terms of K and S. The second step follows from the fact that if

$$\|S-1\| = \sigma^2$$

then the minimum of the functional is attained by $K = C = D = 0$, in which case $S = 1$ gives the same value. The first term of the functional has also been moved out since it is constant in the inner minimization.

The third step follows from Lemma 3.8, using that

$$S(1 - Bz^{-1}) = \frac{1}{1 - KG_{yu}}.$$

The fourth step follows from (3.39) and application of the Youla parametrization.

Given $(Q, S) \in \Theta_{Q,S}$, define

$$X = S \prod_{\lambda \in \Lambda(G_{yu})} \frac{z\lambda^* - 1}{z - \lambda}. \tag{3.50}$$

Then $|X| = |S|$ on $\mathbb{T}$, so the value of the functional does not change if S is swapped for X. Furthermore, $X \in \mathcal{RH}_\infty$, $\|X\|_2^2 = \|S\|_2^2$ and

$$\lim_{z \to \infty} X(z) = \prod_{\lambda \in \Lambda(G_{yu})} |\lambda|.$$

Hence $(Q, X) \in \Theta_{Q,X}$. Conversely if $(Q, X) \in \Theta_{Q,X}$ then S can be defined by (3.50) and then $(Q, S) \in \Theta_{Q,S}$. From this, step 5 follows and the lower bound is obtained.

Now a suboptimal solution will be constructed for Problem 3.2. Suppose that $(Q, X) \in \Theta_{Q,X}$ and $\varepsilon > 0$ and let $(\hat{Q}, \hat{X}) \in \Theta_{Q,X}$ be as given by Lemma 3.9 and define $K \in \mathcal{RL}_1$ by (3.44) and S by (3.45). Then it holds that $(K, S) \in \widetilde{\Theta}_{K,S}$ and

$$\varphi(\hat{Q}, \hat{X}) = \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} \right\|_2^2 + \frac{\left\| \dfrac{KG_{zu}G_{yv}}{1 - KG_{yu}} S \right\|_1^2}{\sigma^2 - \|S - 1\|_2^2}.$$

If $M\hat{Q} - U = 0$ then $K = 0$,

$$J(0, 0, 0) = \|G_{zv}\|_2^2 = \varphi(\hat{Q}, \hat{X}) < \varphi(Q, X) + \varepsilon,$$

and the proof is complete.

If, on the other hand, $M\hat{Q} - U$ is not identically zero then K is not identically zero. Define B by (3.46) and note that B is proper since $\lim_{z \to \infty} S(z) = 1$ and G_{yu} is strictly proper. Moreover, since G_{yu} has no poles on $\mathbb{T}$ and $\hat{X}$ has no zeros for $|z| \geq 1$, the zeros of S for $|z| \geq 1$ correspond, with multiplicity, to poles of G_{yu}, which are also the zeros of $(1 - KG_{yu})^{-1} \in \mathcal{RH}_\infty$. Thus $B \in \mathcal{RH}_\infty$.

According to Lemma 3.8 there then exists an outer $C \in \mathcal{H}_2$ and $D \in \mathcal{L}_2$ such that

$$K = D(1 - Bz^{-1})^{-1}C, \quad S = \frac{1}{1 - Bz^{-1} - DCG_{yu}},$$

and (3.47) and (3.48) are satisfied. The lemma also says that such (C, D) satisfy

$$\|SDG_{zu}\|_2^2 = \frac{\left\| \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} S \right\|_1^2}{\sigma^2 - \|S - 1\|_2^2}, \quad \|SCG_{yv}\|_2^2 \le \sigma^2 - \|S - 1\|_2^2 .$$

K, C, D and S satisfy the conditions of Lemma 3.7, so $T \in \mathcal{H}_2$, which implies that $(B, C, D) \in \Theta_{B,C,D}$. Moreover,

$$J(B, C, D) = \left\| G_{zv} + \frac{KG_{zu}G_{yv}}{1 - KG_{yu}} \right\|_2^2 + \|SDG_{zu}\|_2^2 = \varphi(\hat{Q}, \hat{X}) < \varphi(Q, X) + \varepsilon.$$

Since ε can be made arbitrarily small this shows that (3.43) holds and hence the proof is complete. □

Remark 3.6
Theorem 3.3 shows that an ε-suboptimal solution of Problem 3.2 can be found by minimizing $\varphi(Q, X)$ over $\Theta_{Q,X}$. The obtained (Q, X) may have to be perturbed so that the resulting K has no poles on the unit circle and S has no zeros on the unit circle. Then B is directly calculated and C given by a spectral factorization. D is then obtained from C. □

A by-product of Theorem 3.3 is a necessary and sufficient criterion for the existence of a stabilizing controller that satisfies the SNR constraint.

Corollary 3.2
There exists (B, C, D) that stabilize the closed loop system of Figure 3.7 subject to the SNR constraint (3.27) if and only if

$$\prod_{\lambda \in \Lambda(G_{yu})} |\lambda|^2 < \sigma^2 + 1. \tag{3.51}$$

□

PROOF
The theorem showed that there exists $(B,C,D) \in \Theta_{B,C,D}$ if and only if there exists $(Q,X) \in \Theta_{Q,X}$. In the definition of $\Theta_{Q,X}$, it is seen that Q can be chosen freely in $\mathcal{RH}_\infty$, but that $X \in \mathcal{RH}_\infty$ must satisfy

$$\|X\|_2^2 < \sigma^2 + 1, \quad \lim_{z\to\infty} X(z) = \prod_{\lambda\in\Lambda(G_{yu})} |\lambda|$$

The second of these conditions states that

$$X(z) = \prod_{\lambda\in\Lambda(G_{yu})} |\lambda| + \sum_{k=1}^{\infty} x_k z^{-k}.$$

Clearly, the 2-norm of X is then bounded from below by

$$\|X\|_2^2 = \prod_{\lambda\in\Lambda(G_{yu})} |\lambda|^2 + \sum_{k=1}^{\infty} |x_k|^2 \geq \prod_{\lambda\in\Lambda(G_{yu})} |\lambda|^2 . \tag{3.52}$$

This bound is obviously tight, which shows that there can exist such an X if and only if (3.51) holds. □

REMARK 3.7
The condition (3.51) is the same as (3.1). This result was previously shown in [51].

An interesting consequence of (3.52) is that if

$$\sigma^2 \to \prod_{\lambda\in\Lambda(G_{yu})} |\lambda|^2 - 1, \text{ then } |S| \to \prod_{\lambda\in\Lambda(G_{yu})} \left| \frac{z-\lambda}{z-1/\lambda^*} \right| = 1 \text{ on } \mathbb{T}.$$

That is, S approaches an all-pass filter. □

It will now be shown that the minimization of $\varphi(Q,X)$ over $\Theta_{Q,X}$ is a convex problem. This will be done in the same way as in sections 2.4 and 3.2.

Recall that the functional

$$\rho(a,e) = \frac{1}{2\pi}\int_{-\pi}^{\pi} a(\omega)^2 d\omega + \frac{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} a(\omega)e(\omega)d\omega\right)^2}{\sigma^2 + 1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} e(\omega)^2 d\omega}$$

with domain

$$\Theta_\rho = \left\{ (a,e) : a(\omega), e(\omega) \in \mathbb{R}\ \forall\omega, \frac{1}{2\pi}\int_{-\pi}^{\pi} e(\omega)^2 d\omega < \sigma^2 + 1 \right\}$$

is convex by Lemma 2.6. Define the functional

$$\begin{aligned}\varphi_0(Q,X) &= \varphi(Q,X) - \left(\|G_{zv}\|_2^2 + 2\,\mathrm{Re}\,\langle G_{zv}, G_{zu}G_{yv}(M^2Q - MU)\rangle\right) \\ &= \left\|G_{zu}G_{yv}(M^2Q - MU)\right\|_2^2 + \frac{\left\|G_{zu}G_{yv}\left(M^2Q - MU\right)X\right\|_1^2}{\sigma^2 + 1 - \|X\|_2^2}.\end{aligned}$$

LEMMA 3.10
Suppose $(Q,X) \in \Theta_{Q,X}$. Then $\varphi_0(Q,X) \le \gamma$ if and only if there exists $(a,e) \in \Theta_\rho$ such that $\rho(a,e) \le \gamma$ and

$$a(\omega) \ge \sqrt{G_{zu}^* G_{zu} G_{yv} G_{yv}^*}\,\left|M^2Q - MU\right|, \qquad e(\omega) \ge |X| \quad \forall \omega. \tag{3.53}$$

□

PROOF
The proof is a simple modification of the proof of Lemma 2.7. □

THEOREM 3.4
The problem of minimizing $\varphi(Q,X)$ over $\Theta_{Q,X}$ is convex. □

PROOF
The proof is a simple modification of the proof of Theorem 2.4. □

Numerical Solution

Denote

$$\Delta(Q) = \varphi(Q,X) - \varphi_0(Q) = \|G_{zv}\|_2^2 + 2\,\mathrm{Re}\,\langle G_{zv}, G_{zu}G_{yv}(M^2Q - MU)\rangle\,.$$

By Lemma 3.10, minimizing $\varphi(Q,X)$ over $\Theta_{Q,X}$ is equivalent to minimizing $\rho(a,e) + \Delta(Q)$ over $\Theta_\rho \times \Theta_{Q,X}$ subject to (3.53). This problem is infinite-dimensional, so the integrals are discretized for numerical solution. It will now be shown how the discretized problem can be posed as a semidefinite program.

Let $n \ge 2$ and introduce

$$\begin{aligned}&\omega_1 = 0, \quad \omega_{k+1} - \omega_k = 2\pi/n, \quad k = 1,\ldots,n-1 \\ &\hat{a} = [\,a(\omega_1) \quad a(\omega_2) \quad \ldots \quad a(\omega_n)\,]^T \\ &\hat{e} = [\,e(\omega_1) \quad e(\omega_2) \quad \ldots \quad e(\omega_n)\,]^T\,.\end{aligned}$$

An approximation with n grid points is then given by

$$\rho_n(\hat{a},\hat{e}) = \frac{1}{n}\hat{a}^T\hat{a} + \frac{\left(\frac{1}{n}\hat{a}^T\hat{e}\right)^2}{\sigma^2 + 1 - \frac{1}{n}\hat{e}^T\hat{e}} \approx \rho(a,e)$$

$$\Delta_n(Q) = \|G_{zv}\|_2^2 + \frac{2}{n}\operatorname{Re}\sum_{k=1}^{n}\operatorname{tr}\ \left(G_{zv}^* G_{zu} G_{yv}(M^2 Q - MU)\right)\Big|_{z=e^{i\omega_k}} \approx \Delta(Q).$$

By the definition of the integral it holds that

$$\lim_{n\to\infty} \rho_n(\hat{a},\hat{e}) + \Delta_n(Q) = \rho(a,e) + \Delta(Q),$$

so the minimum of the approximation can be made to come arbitrarily close to $\inf_{Q\in\Theta_{Q,X}} \varphi(Q,X)$ if n is chosen sufficiently large. When implementing the minimization program, Q and X are parametrized using finite basis representations. The accuracy of the approximated problem obviously depends on this representation as well.

Noting that $\rho_n(\hat{a},\hat{e}) + \Delta_n(Q)$ can be written as a Schur complement and that the denominator of $\rho_n(\hat{a},\hat{e})$ is positive for sufficiently large n, it follows that $\rho_n(\hat{a},\hat{e}) + \Delta_n(Q) \le \gamma$ if and only if

$$\begin{bmatrix} \frac{1}{n}\hat{e}^T\hat{e} - \sigma^2 - 1 & \frac{1}{n}\hat{a}^T\hat{e} \\ \frac{1}{n}\hat{e}^T\hat{a} & \frac{1}{n}\hat{a}^T\hat{a} + \Delta_n(Q) - \gamma \end{bmatrix} \preceq 0,$$

or, equivalently,

$$\begin{bmatrix} n(\sigma^2+1) & 0 \\ 0 & n\gamma - n\Delta_n(Q) \end{bmatrix} - \begin{bmatrix} \hat{e} & \hat{a} \end{bmatrix}^T I \begin{bmatrix} \hat{e} & \hat{a} \end{bmatrix} \succeq 0.$$

Using Schur complement again, this is equivalent to

$$\begin{bmatrix} I & \hat{e} & \hat{a} \\ \hat{e}^T & n(\sigma^2+1) & 0 \\ \hat{a}^T & 0 & n\gamma - n\Delta_n(Q) \end{bmatrix} \succeq 0. \tag{3.54}$$

Let $g_k = \sqrt{G_{zu}(e^{i\omega_k})^* G_{zu}(e^{i\omega_k}) G_{yv}(e^{i\omega_k}) G_{yv}(e^{i\omega_k})^*}$. The constraints can then be approximated by

$$a(\omega_k) \ge g_k \left| M(e^{i\omega_k})^2 Q(e^{i\omega_k}) - M(e^{i\omega_k}) U(e^{i\omega_k}) \right|, \quad k = 1 \dots n \tag{3.55}$$

$$e(\omega_k) \ge \left| X(e^{i\omega_k}) \right|, \quad k = 1 \dots n \tag{3.56}$$

$$\sigma^2 + 1 > \frac{1}{n}\sum_{k=1}^{n} e(\omega_k)^2 \tag{3.57}$$

$$X(z) = \prod_{\lambda\in\Lambda(G_{yu})} |\lambda| + Y(z), \text{ where } Y(z) \text{ is strictly proper.} \tag{3.58}$$

Minimizing γ subject to (3.54)–(3.58) is a semidefinite program. The value of the approximated problem is arbitrarily close to $\inf_{Q\in\Theta_{Q,X}} \varphi(Q,X)$ for sufficiently large n.

A procedure for numerical solution of Problem 3.2 will now be outlined.

1. Check if the plant is stabilizable under the SNR constraint by using condition (3.51).
2. Determine a $N,M,U,V \in \mathcal{RH}_\infty$ by a coprime factorization of G_{yu}.
3. Choose n sufficiently large, determine the grid points ω_k, $k = 1\ldots n$ and solve the optimization problem of minimizing γ subject to (3.54)–(3.58). The transfer functions Q,X are parametrized using finite basis representations, for example as FIR filters.
4. If $NQ+V$ has zeros on the unit circle, or if X has zeros on or outside the unit circle, determine $\hat{Q},\hat{X}$ as outlined by Lemma 3.9.
5. Determine S from (3.45) and B from (3.46).
6. Use a finite basis approximation $A(\omega)$ of CC^*, for example the parametrization (2.25), and fit $A(\omega)$ to the right hand side of (3.47), for example by minimizing the mean squared deviation.
7. Perform a spectral factorization of $A(\omega)$, choosing C as the stable and minimum phase spectral factor.
8. Determine D from (3.48).

Special Cases

It will now be shown that the problems considered in sections 2.4 and 3.2 can be solved using the equivalent optimizaton problem that was obtained in this section.

Control with SNR constraint without channel feedback A solution to Problem 3.1 can be obtained by letting $B = 0$. Then

$$S = \frac{1}{1 - KG_{yu}} = MNQ + MV,$$

and the functional to minimize becomes

$$\varphi(Q) = \|G_{zv} + G_{zu}G_{yv}(AQ + B)\|_2^2 + \frac{\|G_{zu}G_{yv}(AQ + B)(EQ + F)\|_1^2}{\sigma^2 + 1 - \|(EQ + F)\|_2^2},$$

where $A = M^2$, $B = -MU$, $E = MN$ and $F = MV$. The SNR constraint becomes $\|EQ + F\|_2^2 < \sigma^2+1$. This is the same problem as the one obtained in Section 3.2.

Real-Time Coding for a Noisy Channel with Feedback Just as the coding problem without channel feedback (Problem 2.1) was seen to be a special case of the control problem without channel feedback (Problem 3.1), it can be seen that the corresponding problems with channel feedback (Problem 2.3 and Problem 3.2 in this section) have the same relationship.

The feedback part of the encoder is parametrized differently in Section 2.4 compared to here (compare the block diagram in Figure 2.8 with the one in Figure 3.7). These parametrizations are, however, equivalent, and it is possible to obtain one of them given the other. Denote the filters that are called B and C in Section 2.4 instead by $\widetilde{B}$ and $\widetilde{C}$, respectively. Then it is easy to show that the encoders are equivalent if and only if

$$\begin{aligned} \widetilde{C} &= (1 - Bz^{-1})^{-1}C \\ 1 + \widetilde{B}z^{-1} &= (1 - Bz^{-1})^{-1}. \end{aligned}$$

Solving Problem 2.3 corresponds to solving Problem 3.2 for the plant

$$G(z) = \begin{bmatrix} G_{zv}(z) & G_{zu}(z) \\ G_{yv}(z) & G_{yu}(z) \end{bmatrix} = \begin{bmatrix} PF & 0 & -1 \\ F & G & 0 \end{bmatrix}.$$

Since $G_{yu} = 0$, a coprime factorization is given by $N = 0$, $M = 1$, $U = 0$ and $V = 1$. Then $Q = K$, $S = (1 - Bz^{-1})^{-1}$ and the corresponding functional to minimize is

$$\begin{aligned} \varphi(B,K) &= \left\| \begin{bmatrix} PF - FK & -GK \end{bmatrix} \right\|_2^2 + \frac{\left\| \begin{bmatrix} F & G \end{bmatrix} K(1 - Bz^{-1})^{-1} \right\|_1^2}{\sigma^2 + 1 - \left\| (1 - Bz^{-1})^{-1} \right\|} \\ &= \left\| \begin{bmatrix} PF - FK & -GK \end{bmatrix} \right\|_2^2 + \frac{\left\| \begin{bmatrix} F & G \end{bmatrix} K(1 + \widetilde{B}z^{-1}) \right\|_1^2}{\sigma^2 + 1 - \left\| 1 + \widetilde{B}z^{-1} \right\|}, \end{aligned}$$

which is equal to the one defined by (2.51) in Section 2.4. Moreover, the SNR constraint becomes $\sigma^2 + 1 > \left\| (1 + Bz^{-1})^{-1} \right\|_2^2$, which is equivalent to $\sigma^2 > \left\| \widetilde{B} \right\|_2^2$. Finally, the optimality conditions are easily verified to be equivalent.

4

Conclusions

This thesis has introduced a new method, based on a technique called optimal factorization, for solving a number of communication and control problems. It is a fairly simple idea that there for every product of the encoder and decoder transfer functions should exist an optimal factorization that minimizes the negative impact of the channel noise. This approach can also be viewed as a variable change followed by a sequential minimization over different variables. Luckily, it turns out that the optimal factorization problem has a closed-form expression for the value and that the minimization over the encoder-decoder product becomes a convex problem.

Lately, information theory has been successfully applied to control problems with communication constraints, which has resulted in stabilizability conditions and performance bounds. Results on the design of optimal controllers have, however, been more scarce. This thesis represents an attempt to instead apply control theoretic tools to a problem of this kind. The advantage of the approach developed here is, as has been shown, that optimal solutions may be obtained using well-known techniques, such as convex optimization. The disadvantage is that the optimization can only be performed over the relatively simple class of LTI systems. Still, linear solutions have nice features in that they are amenable to analysis using well-known tools and are easy to implement.

Two questions were asked in the preface: How should a communication system be designed when there is a bound on the accepted delay? And, how should a control system be designed when there are communication limitations? The results presented in this thesis give a partial answer to these two questions, while at the same time demonstrating how intertwined they are. In order to control there has to be communication, and the stability and performance of the system depends critically on any delay induced by coding. Thus, it is no wonder that the solution of a communication-constrained control problem also can be used to solve

a communication problem. Interestingly, this does not reflect the chronological order in which the results in this thesis were obtained.

A number of topics remain for further research. The most obvious is the investigation whether linear solutions are in fact optimal or not, for these problems. And if they are not, what is then an optimal solution? For the linear solutions, it should be investigated if the memory requirements are infinite for the problems in this where the channel has feedback. When infinite memory is required, some approximation is necessary for implementation. It would be good to find suboptimality bounds, for example when using truncated FIR filters to approximate the optimal filters.

It does not seem to be possible to directly apply the techniques used in this thesis to solve the feedback control problem in the case of a general MIMO plant, where also the control loop has a vector-valued signal. But perhaps the techniques can be modified to handle this case as well? The case with noisy channel feedback is also of interest.

Finally, perhaps there are other types of problems that could be solved using an approach similar to the optimal factorization idea.

5

Bibliography

[1] K. J. Åström and B. Wittenmark, *Computer-Controlled Systems*. Prentice Hall, Jan. 1997.

[2] R. Bansal and T. Basar, "Simultaneous design of measurement and control strategies for stochastic systems with feedback," *Automatica*, vol. 25, no. 5, pp. 679–694, 1989.

[3] T. Berger, *Rate distortion theory: a mathematical basis for data compression*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[4] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2693–2723, 1998.

[5] V. Borkar, S. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of markov sources," in *Proc. IEEE Conference on Decision and Control*, vol. 1, 2001, pp. 205–210.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[7] J. Braslavsky, R. Middleton, and J. Freudenberg, "Feedback stabilization over signal-to-noise ratio constrained channels," *IEEE Transactions on Automatic Control*, vol. 52, no. 8, pp. 1391–1403, Aug. 2007.

[8] P. Breun and W. Utschick, "On transmitter design in power constrained LQG control," in *Proc. American Control Conference*, June 2008, pp. 4979–4984.

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[10] M. Derpich, "Optimal source coding with signal transfer function constraints," Ph.D. dissertation, University of Newcastle, 2009.

[11] M. S. Derpich and J. Østergaard, "Improved upper bounds to the causal quadratic rate-distortion function for gaussian stationary sources," *IEEE Transactions on Information Theory*, 2010, submitted. [Online]. Available: http://arxiv.org/abs/1001.4181v2

[12] M. Derpich, E. Silva, D. Quevedo, and G. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3871–3890, Aug. 2008.

[13] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, Sept. 1962.

[14] N. Elia, "When Bode meets Shannon: Control-oriented feedback communication schemes," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1477–1488, 2004.

[15] R. A. Emmons and M. E. McCullough, "Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life," *Journal of Personality and Social Psychology*, vol. 84, no. 2, pp. 377–389, 2003.

[16] J. S. Freudenberg and R. H. Middleton, "Stabilization and performance over a gaussian communication channel for a plant with time delay," in *Proc. American Control Conference*, 2009, pp. 2148–2153.

[17] J. Freudenberg, R. Middleton, and J. Braslavsky, "Stabilization with disturbance attenuation over a gaussian channel," in *Proc. IEEE Conference on Decision and Control*, dec. 2007, pp. 3958–3963.

[18] ——, "Minimum variance control over a gaussian communication channel," in *Proc. American Control Conference*, 2008, pp. 2625–2630.

[19] J. Freudenberg, R. Middleton, and V. Solo, "Stabilization and disturbance attenuation over a gaussian communication channel," *IEEE Transactions on Automatic Control*, vol. 55, no. 3, pp. 795–799, Mar. 2010.

[20] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[21] J. Garnett, *Bounded analytic functions*, revised 1st ed. New York, NY, USA: Springer, 2007.

[22] M. Gastpar, "To code or not to code," Ph.D. dissertation, EPFL, Lausanne, 2002.

[23] A. Ghulchak, personal communication, 2010.

[24] J. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[25] G. C. Goodwin, D. E. Quevedo, and E. I. Silva, "Architectures and coder design for networked control systems," *Automatica*, vol. 44, no. 1, pp. 248–257, 2008.

[26] Y. Inouye, "Linear systems with transfer functions of bounded type: Canonical factorization," *IEEE Transactions on Circuits and Systems*, vol. 33, no. 6, pp. 581–589, June 1986.

[27] E. Johannesson, A. Rantzer, and B. Bernhardsson, "Optimal linear control for channels with signal-to-noise ratio constraints," in *Proc. American Control Conference*, San Francisco, CA, USA., June 2011.

[28] V. Katsnelson and B. Kirstein, "On the Theory of Matrix Valued Functions Belonging to the Smirnov Class," in *Topics in interpolation theory*, ser. Operator theory, advances and applications, H. Dym, B. Fritzshe, V. Katsnelson, and B. Kirstein, Eds. Birkhäuser, 1997. [Online]. Available: http://arxiv.org/abs/0706.1901

[29] Y. Li, E. Tuncel, J. Chen, and W. Su, "Optimal tracking performance of discrete-time systems over an additive white noise channel," in *Proc. IEEE Conference on Decision and Control, held jointly with the Chinese Control Conference.*, Dec. 2009, pp. 2070–2075.

[30] J. Löfberg, "Yalmip : A toolbox for modeling and optimization in MATLAB," in *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[31] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003, available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

[32] A. Mahajan and D. Teneketzis, "On the design of globally optimal communication strategies for real-time noisy communication systems with noisy feedback," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 580–595, May 2008.

[33] ——, "Optimal design of sequential real-time communication systems," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5317–5338, Nov. 2009.

[34] N. Martins and M. Dahleh, "Feedback control in the presence of noisy channels: "Bode-like" fundamental limitations of performance," *IEEE Transactions on Automatic Control*, vol. 53, no. 7, pp. 1604–1615, Aug. 2008.

[35] N. Martins, M. Dahleh, and J. Doyle, "Fundamental limitations of disturbance attenuation in the presence of side information," *IEEE Transactions on Automatic Control*, vol. 52, no. 1, pp. 56–66, Jan. 2007.

[36] A. Matveev and A. Savkin, "An analogue of Shannon information theory for networked control systems. stabilization via a noisy discrete channel," in *Proc. IEEE Conference on Decision and Control*, vol. 4, 2004, pp. 4491–4496 Vol.4.

[37] G. N. Nair and R. J. Evans, "Exponential stabilisability of finite-dimensional linear systems with limited data rates," *Automatica*, vol. 39, no. 4, pp. 585–593, 2003.

[38] ——, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM J. Control Optim.*, vol. 43, pp. 413–436, Feb. 2004.

[39] G. Nair, F. Fagnani, S. Zampieri, and R. Evans, "Feedback control under data rate constraints: An overview," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 108–137, Jan. 2007.

[40] D. Neuhoff and R. Gilbert, "Causal source codes," *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 701–713, Sept. 1982.

[41] S. Pulgar, E. Silva, and M. Salgado, "Optimal state-feedback design for MIMO systems subject to multiple SNR constraints," in *Proc. 18th IFAC World Congress*, Milano, Italy, Aug. 2011.

[42] A. J. Rojas, "Signal-to-noise ratio performance limitations for input disturbance rejection in output feedback control," *Systems & Control Letters*, vol. 58, no. 5, pp. 353–358, 2009.

[43] A. J. Rojas, J. H. Braslavsky, and R. H. Middleton, "Fundamental limitations in control over a communication channel," *Automatica*, vol. 44, no. 12, pp. 3147–3151, 2008.

[44] W. Rudin, *Real and Complex Analysis*, 3rd ed. McGraw-Hill Science/Engineering/Math, May 1986.

[45] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link — Part I: Scalar systems," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[46] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback — I: No bandwidth constraint," *IEEE Transactions on Information Theory*, vol. 12, no. 2, pp. 172–182, Apr. 1966.

[47] C. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, Sept. 1956.

[48] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, 1948.

[49] E. I. Silva, M. S. Derpich, and J. Ostergaard, "A framework for control system design subject to average data-rate constraints," *IEEE Transactions on Automatic Control*, vol. PP, no. 99, p. 1, 2010.

[50] ——, "An achievable data-rate region subject to a stationary performance constraint for LTI plants," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1968–1973, Aug. 2011.

[51] E. I. Silva, G. C. Goodwin, and D. E. Quevedo, "Control system design subject to SNR constraints," *Automatica*, vol. 46, no. 2, pp. 428–436, 2010.

[52] E. Silva, J. Agüero, G. Goodwin, K. Lau, and M. Wang, *The SNR approach to Networked Control*, 2nd ed. CRC Press, 2011, ch. 25.

[53] E. Silva, M. Derpich, and J. Østergaard, "On the minimal average data-rate that guarantees a given closed loop performance level," in *Proceedings of the 2nd IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys)*, Annecy, France, July 2010.

[54] E. Silva, G. Goodwin, and D. Quevedo, "On the design of control systems over unreliable channels," in *Proc. European Control Conference*, Budapest, Hungary, 2009.

[55] J. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11–12, pp. 625–653, 1999, version 1.3 available from `http://sedumi.ie.lehigh.edu/`.

[56] G. Szegő, *Orthogonal polynomials*, 4th ed. American Mathematical Society, Providence, RI, 1975.

[57] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, July 2004.

[58] J. Walrand and P. Varaiya, "Optimal causal coding–decoding problems," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 814–820, Nov. 1983.

[59] N. Wiener and E. Akutowicz, "A factorization of positive Hermitian matrices," *Indiana Univ. Math. J.*, vol. 8, pp. 111–120, 1959.

[60] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

[61] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 518–521, Sept. 1980.

[62] ——, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, no. 6, pp. 1437–1451, July-Aug 1979.

[63] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 406–411, July 1970.

[64] S. Yüksel, "On optimal causal coding of partially observed markov sources in single and multi-terminal settings," 2010. [Online]. Available: http://arxiv.org/abs/1010.4824v2

[65] K. Zhou, J. C. Doyle, and K. Glover, *Robust and optimal control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

A

Some Technical Proofs

This appendix contains the proofs of Lemma 2.5 and Lemma 3.4. The proof of Lemma 2.5 follows easily from the proof of Lemma 3.4, which is therefore presented first.

Proof of 3.4

The proof is based on construction of $\hat{Q}$ through a perturbation of Q. Take $Q \in \Theta_Q$ and let

$$K = \frac{MQ - U}{NQ + V}.$$

If $K \in \mathcal{RL}_1$ then let $\hat{Q} = Q$ and the construction is complete. Suppose instead that K has at least one pole on $\mathbb{T}$. Since $MQ - U \in \mathcal{RH}_\infty$, z is a pole of K if and only if

$$N(z)Q(z) + V(z) = 0. \tag{A.1}$$

Moreover, suppose that (A.1) holds and that $N(z) = 0$. Then it follows from the Bezout identity that $V(z) \neq 0$, which is a contradiction. Thus if $NQ + V$ has a zero at z then $N(z) \neq 0$.

Suppose now that $NQ + V$ has a zero at $z_0 \in \mathbb{T}$ and that $z_0 \notin \mathbb{R}$ (the case when $z_0 \in \mathbb{R}$ is discussed later). Let

$$\hat{Q} = Q + \lambda_0 + \lambda_1 z^{-1}, \quad \lambda_0, \lambda_1 \in \mathbb{R}.$$

Then $\left\| E\hat{Q} + F \right\|_2 < \sigma^2 + 1$ if $|\lambda_0| + |\lambda_1| < \delta_\lambda$ for small enough δ_λ.

The coefficients λ_0, λ_1 will be chosen so that the zero at z_0 is perturbed away from $\mathbb{T}$. It must also be made sure that none of the other zeros can reach $\mathbb{T}$ under the same perturbation. For this reason, define the set of zeros not on the unit circle,

$$\Omega = \{z : z \notin \mathbb{T}, N(z)Q(z) + V(z) = 0\},$$

and the smallest distance from that set to the unit circle,

$$r = \inf_{z_1 \in \Omega, z_2 \in \mathbb{T}} |z_1 - z_2|,$$

where $r > 0$ since Ω has a finite number of elements. The location of the zeros of $N\hat{Q} + V$ depend continuously on (λ_0, λ_1). Thus, there exists $\delta_r > 0$ such that if $|\lambda_0| + |\lambda_1| < \delta_r$ then all zeros are displaced strictly less than r.

Introduce the function

$$X(z, \lambda_0, \lambda_1) = N\hat{Q} + V = NQ + V + N(\lambda_0 + \lambda_1 z^{-1}).$$

Then

$$\det \begin{bmatrix} \operatorname{Re} \dfrac{\partial X}{\partial \lambda_0} & \operatorname{Re} \dfrac{\partial X}{\partial \lambda_1} \\ \operatorname{Im} \dfrac{\partial X}{\partial \lambda_0} & \operatorname{Im} \dfrac{\partial X}{\partial \lambda_1} \end{bmatrix} = \det \begin{bmatrix} \operatorname{Re} N & \operatorname{Re} Nz^{-1} \\ \operatorname{Im} N & \operatorname{Im} Nz^{-1} \end{bmatrix} \neq 0 \text{ at } z = z_0$$

since $N(z_0) \neq 0$ and $z_0 \in \mathbb{T} \setminus \mathbb{R}$. Then, by the implicit function theorem, there is a differentiable mapping $z \mapsto (\lambda_0, \lambda_1)$ defined in a neighborhood of z_0, such that

$$N(z)\hat{Q}(z) + V(z) = N(z)Q(z) + V(z) + N(z)(\lambda_0(z) + \lambda_1(z)z^{-1}) = 0.$$

This means that a new location z can be determined for the zero, and the mapping gives the corresponding λ_0, λ_1.

Take $\varepsilon > 0$. Since $\varphi(Q)$ is continuous there exists $\delta_Q > 0$ such that

$$\left\| \hat{Q} - Q \right\|_\infty < \delta_Q \Rightarrow \left| \varphi(\hat{Q}) - \varphi(Q) \right| < \varepsilon.$$

Continuity of the mapping from z to (λ_0, λ_1) implies that there exists $\delta_z > 0$ such that

$$|z - z_0| < \delta_z \Rightarrow |\lambda_0(z)| + |\lambda_1(z)| < \min\{\delta_Q, \delta_\lambda, \delta_r\}.$$

Now pick $z \notin \mathbb{T}$ such that $|z - z_0| < \delta_z$ and the mapping to λ_0, λ_1 is defined. Then

$$\left\| \hat{Q} - Q \right\|_\infty \leq |\lambda_0(z)| + |\lambda_1(z)| < \min\{\delta_Q, \delta_\lambda, \delta_r\},$$

which implies that $|\varphi(\hat{Q}) - \varphi(Q)| < \varepsilon$, $\left\| E\hat{Q} + F \right\|_2 < \sigma^2 + 1$ and that there are no new zeros on $\mathbb{T}$. Since $z \notin \mathbb{T}$ it follows that $N\hat{Q} + V$ has at least one zero less than $NQ + V$ on $\mathbb{T}$.

If z_0 is real, then define instead

$$\hat{Q} = Q + \lambda_0, \quad \lambda_0 \in \mathbb{R}$$

and determine λ_0 analogously. Note, however, that the zero must be kept on the real axis.

If $\hat{Q}$ is such that $N\hat{Q}+V$ has zeros on $\mathbb{T}$, the procedure may be repeated again, with ε appropriately chosen, until there are no such zeros. Thus, for every $Q \in \Theta_Q$ and $\varepsilon > 0$ it is possible to construct $\hat{Q}$ such that $N\hat{Q}+V$ has no zeros on $\mathbb{T}$, $|\varphi(\hat{Q}) - \varphi(Q)| < \varepsilon$ and $\left\| E\hat{Q} + F \right\|_2 < \sigma^2 + 1$. □

Proof of 2.5

Noting that $\varphi(B, K)$ is continuous in B, the proof follows the same reasoning as the proof of Lemma 3.4 with $Q = B$, $\hat{Q} = \hat{B}$, $N = z^{-1}$, $V = 1$, $E = z^{-1}$ and $F = 1$. □