Quantum signal processing

Article *in* IEEE Signal Processing Magazine · December 2002

DOI: 10.1109/MSP.2002.1043298 · Source: IEEE Xplore

2 authors:

Yonina Eldar
Weizmann Institute of Science
857 PUBLICATIONS **26,867** CITATIONS
SEE PROFILE

A.V. Oppenheim
Massachusetts Institute of Technology
228 PUBLICATIONS **36,128** CITATIONS
SEE PROFILE

Quantum Signal Processing

by

Yonina Chana Eldar

B.Sc., Physics (1995)
Tel-Aviv University

B.Sc., Electrical Engineering (1996)
Tel-Aviv University

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 2001

Author .....
Department of Electrical Engineering and Computer Science
December 4, 2001

Certified by .....
Alan V. Oppenheim
Ford Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Quantum Signal Processing

by

Yonina Chana Eldar

Submitted to the Department of Electrical Engineering and Computer Science
on December 4, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Quantum signal processing (QSP) as formulated in this thesis, borrows from the formalism and principles of quantum mechanics and some of its interesting axioms and constraints, leading to a novel paradigm for signal processing with applications in areas ranging from frame theory, quantization and sampling methods to detection, parameter estimation, covariance shaping and multiuser wireless communication systems. The QSP framework is aimed at developing new or modifying existing signal processing algorithms by drawing a parallel between quantum mechanical measurements and signal processing algorithms, and by exploiting the rich mathematical structure of quantum mechanics, but not requiring a physical implementation based on quantum mechanics. This framework provides a unifying conceptual structure for a variety of traditional processing techniques, and a precise mathematical setting for developing generalizations and extensions of algorithms.

Emulating the probabilistic nature of quantum mechanics in the QSP framework gives rise to probabilistic and randomized algorithms. As an example we introduce a probabilistic quantizer and derive its statistical properties. Exploiting the concept of generalized quantum measurements we develop frame-theoretical analogues of various quantum-mechanical concepts and results, as well as new classes of frames including oblique frame expansions, that are then applied to the development of a general framework for sampling in arbitrary spaces. Building upon the problem of optimal quantum measurement design, we develop and discuss applications of optimal methods that construct a set of vectors with a given inner product structure that are closest in a least-squares sense to a given set of vectors. We demonstrate that, even for problems without inherent inner product constraints, imposing such constraints in combination with least-squares inner product shaping leads to interesting processing techniques that often exhibit improved performance over traditional methods. In particular, we formulate a new viewpoint toward matched filter detection that leads to the notion of minimum mean-squared error covariance shaping. Using this concept we develop an effective linear estimator for the unknown parameters in a linear model, referred to as the covariance shaping least-squares estimator. Applying this estimator to a multiuser wireless setting, we derive an efficient covariance shaping multiuser receiver for suppressing interference in multiuser communication systems.

Thesis Supervisor: Alan V. Oppenheim
Title: Ford Professor of Electrical Engineering

Acknowledgments

My sincerest thanks to my advisor Prof. Alan Oppenheim, for his mentoring, guidance, encouragement and close collaboration throughout this research. His insistence on creativity and out-of-the-box thinking have had a tremendous impact both on this work and on my professional development. He has provided me with a unique experience—way beyond my expectations. The privilege of working closely with him means a great deal to me and I am looking forward to tugging on many more tails together in the future.

I am extremely grateful for the thoughtful comments and valuable suggestions of my readers Prof. David Forney and Prof. George Verghese. I am sincerely grateful to Prof. Forney for his enthusiasm back at my area exam which gave me the most valuable opportunity to work closely with him. I am very thankful for the many stimulating and thought provoking discussions we had over the past several years that greatly influenced my thinking about research in general and this thesis in particular. It would be a privilege for me to work with him in the future.

I might never have entered the world of DSP would it not be for my close friend and colleague Dr. Arie Yeredor and for Prof. Ehud Weinstein of Tel-Aviv University. I am extremely fortunate to have met Dr. Yeredor and subsequently Prof. Weinstein while an undergraduate at Tel-Aviv. After participating in the course "Parameter Estimation" taught by Prof. Weinstein and assisted by Dr. Yeredor, I knew that DSP is what I would want to do! My sincerest thanks to Dr. Yeredor for exposing me to DSP, for helping me develop my research skills, for the persistent encouragement, and above all for his friendship. Many thanks to Prof. Weinstein for believing in me when I was still green and for encouraging me to come to MIT. I am looking forward to a lifetime of collaboration back in Israel.

I also want to thank Prof. Helmut Bölcskei from University of Illinois, Dianne Egnor from BAE Systems, and Albert Chan from DSPG for collaborations on various parts of this work. Thanks also to Prof. Gilbert Strang for fruitful discussions.

I am indebted to Prof. Akram Aldroubi of Vanderbilt University and Prof. Shlomo Shamai of the Technion—Israel Institute of Technology, for stimulating suggestions for future research related to this work.

Thanks to all the members of DSPG for providing an enjoyable and stimulating environment. Special thanks to Andrew Russell for being a good friend, and for many invaluable research discussions. Thanks also to Petros Boufounos, Stark Draper, Nick Laneman, Maya Said, Matt Secor and Wade Toress, for many discussions on life, religion, family, research and more, and to Giovanni Aliberti for his friendship and sound advice. Thanks to Darla Secor, Dianne Wheeler and Wendy Russell for fun conversations and administrative help.

None of this would have been possible if it where not for my parents who have provided a lifetime of encouragement, support and love. Thank you for teaching me values through a constant living example. Ema, thank you for always being at the other end of the line, and for supporting and guiding me through any obstacle. Abba, thank you for always encouraging me to pursue my interests, for your constant sound advice, and for being a true inspiration.

I also want to thank my brothers and sisters and my extended family for their continuing support during this hectic period. A special thanks to my aunt Hildy and my uncle Dave from Brookline for always being there for us.

Finally, I want to thank my husband, Shalomi, for his boundless love, emotional support, unending patience, and for his many personal sacrifices over the past three years. I also want to thank our son, Yonatan, for filling our lives with happiness which made it all worthwhile.

To my parents

To Shalomi and Yonatan

For God gives wisdom,
out of His mouth come knowledge and understanding.
— Proverbs 2:6

Contents

List of Figures

Chapter 1

Introduction

Quantum signal processing (QSP) as formulated in this thesis, borrows from the principles of quantum mechanics and some of its interesting axioms and constraints. However, in contrast to such fields as quantum computing and quantum information theory, it does not depend on the physics associated with quantum mechanics. Consequently, in developing the QSP framework we are free to impose quantum mechanical constraints that we find useful and to avoid those that are not. In essence, the QSP framework is aimed at developing new or modifying existing signal processing algorithms by drawing a parallel between quantum mechanical measurements and signal processing algorithms, and by exploiting the rich mathematical structure of quantum mechanics, but not requiring a physical implementation based on quantum mechanics. This framework provides a unifying conceptual structure for a variety of traditional processing techniques, and a precise mathematical setting for developing generalizations and extensions of algorithms, leading to a novel paradigm for signal processing with applications in areas ranging from frame theory, quantization and sampling methods to detection, parameter estimation, covariance shaping and multiuser wireless communication systems.

There are many examples in the signal processing literature in which new classes of algorithms have been developed by artificially imposing physical constraints on implementations that are not inherently subject to those constraints. One class of well known examples is wave digital filters [1], which exploit consequences of energy conservation inherent to analog implementations. A direct result of the principle of conservation of energy is that, in contrast to digital filters, analog filters implemented with passive elements have the desir-

able property that they are guaranteed to be stable, even in the presence of element drift or inaccuracies. The general framework of wave digital filters is based on paralleling the energy conservation constraint in the form of a set of concepts referred to as pseudo-energy and pseudo-passivity.

Besides imposing constraints, nature exhibits a variety of behaviors that are potentially interesting to emulate in a wide range of contexts. Using nature as a metaphor we may synthesize systems capitalizing on particular aspects of nature. For example, a diverse collection of natural phenomena exhibit fractal behavior, perhaps suggesting that fractal geometry is somehow optimal or efficient. Whether or not this is truly the case, the fractal-like aspects of nature and related modeling have inspired interesting signal processing paradigms that are not constrained by the physics. For example, fractal modulation [2] emulates the fractal characteristic of nature, resulting in a potentially interesting method for communicating over a particular class of unreliable channels. Likewise, the chaotic behavior of certain features of nature have inspired new classes of signals for secure communications, remote sensing, and a variety of other signal processing applications [3, 4, 5]. Other examples of algorithms using physical systems as a simile are solitons [6], genetic algorithms [7], simulated annealing [8], and neural networks [9].

These examples underscore the fact that even in signal processing contexts that are not constrained by the laws of physics, exploiting laws of nature can inspire new methods for algorithm design and may lead to interesting, efficient and effective processing techniques.

Three fundamental inter-related underlying principles of quantum mechanics are the concept of a measurement, the principle of measurement consistency, and the principle of quantization of the measurement output. In a broad sense, the terms measurement, measurement consistency and output quantization are well known in signal processing, although not with the same precise mathematical interpretation and constraints as in quantum mechanics.

In signal processing, the term measurement can be given a variety of precise or imprecise interpretations. However, as discussed in Section 1.1 of the introduction and in Chapter 3, in quantum mechanics measurement has a very specific definition and meaning, much of which is carried over to the QSP framework. Similarly, in signal processing, quantization is thought of in fairly limited terms. In quantum mechanics, quantization of the measurement output is a fundamental underlying principle and applying this principle along with

the quantum mechanical notions of measurement and consistency, leads to some intriguing generalizations of quantization as typically viewed in signal processing. Measurement consistency also has a precise meaning in quantum mechanics, specifically that repeated applications of a measurement must yield the same outcome. A similar consistency concept is the basis for a variety of signal processing techniques including signal estimation, interpolation, and quantization methods. Some early examples of consistency as it typically arises in signal processing are the interpolation condition in filter design [10], and the condition for no intersymbol interference in waveforms for pulse amplitude modulation [11]. More recent examples include perfect reconstruction filter banks [12, 13], multiresolution and wavelet approximations [14, 15], and sampling methods in which the traditional perfect reconstruction requirement is replaced by the less stringent consistency requirement [16, 17, 18, 19]. Here again, as developed in this thesis, viewing measurement consistency in a broader framework motivated by quantum mechanics leads to some new and interesting signal processing algorithms.

Each of the consistent signal processing algorithms cited above can be described by a linear operator operating on an input signal. Relaxing the requirement for linear processing allows for a broader class of consistent algorithms that can be viewed as generalized quantizers. As part of the thesis we develop a general framework for signal processing algorithms based on the quantum mechanical consistency axiom, which encompasses linear algorithms and nonlinear quantizers as special cases. The algorithms we develop follow from imposing the quantum mechanical interpretation of measurement, quantization and consistency, and by exploiting the formalism and some of the interesting constraints of quantum mechanics to the development of signal processing algorithms, leading to a new framework which we call *Quantum Signal Processing (QSP).*

QSP imposes the quantum mechanical interpretation of the concepts of measurement, quantization and consistency on signal processing algorithms, and borrows further from the formalism and principles of quantum mechanics and some of its interesting constraints. For example, when using quantum systems in a communication context a fundamental problem that arises is the *quantum detection* problem which is subject to the constraints of quantum physics. The constraints imposed in the quantum detection problem suggest some intriguing signal processing algorithms that we explore as part of our framework. As outlined further in this introduction, this approach leads to a new paradigm for signal processing

with applications in a wide range of areas including frame theory, quantization, sampling, parameter estimation, covariance shaping, detection and multiuser wireless communication systems. This thesis is about development and application of this new framework.

In the next section, we summarize the basic principles of measurement, consistency and quantization as they relate to quantum mechanics, and outline the key elements and constraints in the quantum detection problem. In Section 1.2 we indicate how these principles and constraints will be applied in the framework of QSP.

1.1 Quantum Systems

In both signal processing and quantum mechanics, the setting we consider is an arbitrary Hilbert space $\mathcal{H}$. The elements of $\mathcal{H}$ are referred to as vectors or signals interchangeably.

A quantum system in a pure state is characterized by a normalized vector in $\mathcal{H}$. Information about a quantum system is extracted by subjecting the system to a quantum measurement.

1.1.1 Quantum Measurement

A *quantum measurement* is a nonlinear (probabilistic) mapping, that in the simplest case can be described in terms of a set of measurement vectors $\{\mu_i, i \in \mathcal{I}\}$ that span measurement subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$, where $\mathcal{I}$ denotes an index set. The laws of quantum mechanics impose the constraint that the vectors μ_i must be orthonormal. In the more general case, the quantum measurement is described in terms of a set of projection operators $\{P_i, i \in \mathcal{I}\}$ onto subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$, where from the laws of quantum mechanics these projections must form a complete set of orthogonal projections. In quantum mechanics, the outcome of a measurement is inherently probabilistic, with the probabilities of the outcomes of any conceivable measurement determined by the state vector. The measurement collapses (projects) the state of the quantum system onto a state that is compatible with the measurement outcome so that the final state of the system is in general different than the original state.

Measurement consistency is a fundamental postulate of quantum mechanics, *i.e.,* repeated measurements on a system must yield the same outcomes; otherwise we would not be able to confirm the output of a measurement. Therefore the state of the system after a

measurement must be such that if we re-measure the system in this state, then the final state after this second measurement will be identical to the state after the first measurement.

Quantization of the measurement outcome is a direct consequence of the consistency requirement. Specifically, the consistency requirement leads to a class of states referred to as *determinate states* of the measurement [20]. These are states of the quantum system for which the measurement yields a known outcome with probability one, and are the states that lie completely in one of the measurement subspaces $\mathcal{S}_i$. Furthermore, even when the state of the system is not one of the determinate states, after performing the measurement the system is quantized to one of these states, *i.e.,* is certain to be in one of these states, where the probability of being in a particular determinate state is a function of the inner products between the state of the system and the determinate states.

1.1.2 Quantum Detection

The constraints imposed by the physics on a quantum measurement lead to some interesting problems within the framework of quantum mechanics. In particular, an interesting problem that arises when using quantum states for communication is the *quantum detection* problem. As outlined further in this introduction, this problem suggests some intriguing signal processing applications within the framework of QSP.

In a quantum detection problem a sender conveys classical information to a receiver using a quantum-mechanical channel. The sender represents messages by preparing the quantum channel in a pure quantum state drawn from a collection of known states ϕ_i. The receiver detects the information by subjecting the channel to a quantum measurement with measurement vectors μ_i that are constrained by the physics to be orthogonal. If the states are not orthogonal, then no measurement can distinguish perfectly between them. Therefore, a fundamental problem in quantum mechanics is to construct measurements optimized to distinguish between a set of non-orthogonal pure quantum states.

We may formulate this problem as a quantum detection problem, so that the measurement vectors are chosen to minimize the probability of detection error. Necessary and sufficient conditions for an optimum measurement minimizing the probability of detection error have been derived [21, 22, 23]. However, except in some particular cases [23, 24, 25], obtaining a closed-form analytical expression for the optimal measurement directly from these conditions is a difficult and unsolved problem.

In [26] we take an alternative approach of choosing a different optimality criterion, namely a squared-error criterion, and seeking a measurement that minimizes this criterion. Specifically, the measurement vectors μ_i are chosen to be orthogonal, and closest in a least-squares (LS) sense to the given set of state vectors ϕ_i so that the vectors μ_i are chosen to minimize the sum of the squared norms of the error vectors $e_i = \mu_i - \phi_i$, as illustrated in Fig. 1-1. The optimal measurement is referred to as the LS measurement (LSM).

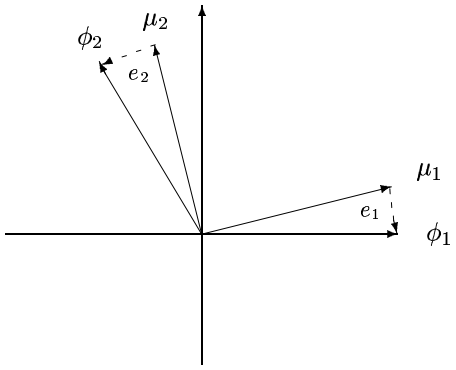


Figure 1-1: 2-dimensional example of the least-squares measurement.

As discussed further in Chapter 3, it turns out that the LSM problem has a simple closed-form solution, with many desirable properties. In particular, this measurement minimizes the probability of a detection error in many cases of practical interest and is nearly optimal in many other cases.

Thus, in the context of quantum detection the constraints of the physics lead to the interesting problem of choosing an optimal set of orthogonal vectors. Borrowing from quantum detection, a central idea in QSP applications is to impose orthogonality or more general inner product constraints on algorithms, and then use the LSM and the results derived in the context of quantum detection to design optimal algorithms subject to these constraints.

1.2 Quantum Signal Processing

The QSP framework draws heavily on the notions of measurement, consistency and quantization as they relate to quantum systems and borrows further from the interesting constraints imposed by quantum physics. However, the QSP framework is broader and less

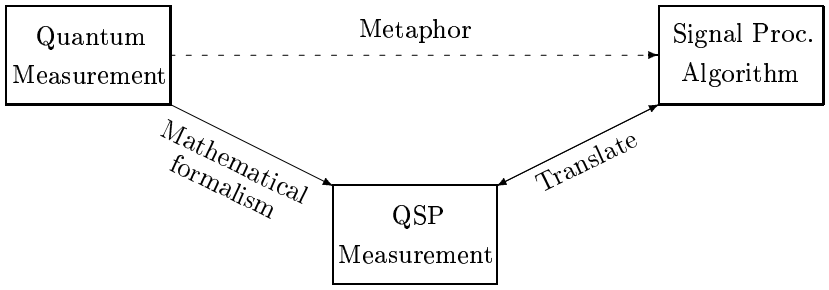


Figure 1-2: Illustration of the QSP framework. In this framework quantum mechanics is used as a metaphor to design new signal processing algorithms by drawing a parallel between a signal processing algorithm and a quantum mechanical measurement. An algorithm is designed by constructing a QSP measurement borrowing from the principles of a quantum measurement, which is then translated into a signal processing algorithm.

restrictive than the quantum measurement framework since in designing algorithms we are not constrained by the physical limitations of quantum mechanics.

In quantum mechanics, systems are "processed" by performing measurements on them. In signal processing signals are processed by putting them through an algorithm. Therefore, to exploit the formalism and rich mathematical structure of quantum mechanics in the design of algorithms we first draw a parallel between a quantum mechanical measurement and a signal processing algorithm by associating a *QSP measurement* with a signal processing algorithm. We then apply the formalism and fundamental principles of quantum measurement to the definition of the QSP measurement. The QSP framework is primarily concerned with the design of the QSP measurement, borrowing from the principles, axioms and constraints of quantum physics and a quantum measurement. As we will show in Chapter 4, the QSP measurement depends on a specific set of measurement parameters, so that this framework provides a convenient and useful setting for deriving new algorithms by choosing different measurement parameters, borrowing from the ideas of quantum mechanics. Furthermore, since the QSP measurement is defined to have a mathematical structure similar to a quantum measurement, the mathematical constraints imposed by the physics on the quantum measurement can also be imposed on the QSP measurement leading to some intriguing new signal processing algorithms. This conceptual framework is illustrated schematically in Fig. 1-2.

1.2.1 The QSP Measurement

We now outline how the quantum-mechanical principles of measurement, consistency and quantization are applied to the definition of the QSP measurement.

Measurement of a signal in the QSP framework corresponds to applying an algorithm to a signal. In the QSP measurement, the signal to be measured may be equal to the signal we wish to process, or may represent this signal in a possibly different signal space. The measurement outcome is a signal in the same signal space as the measured signal, which represents the output of the algorithm, which in turn may be a signal or any other element. As in quantum mechanics, we require that if we re-measure the outcome signal, then the new outcome will be equal to the original outcome.

In analogy with the measurement in quantum mechanics, a *rank-one QSP measurement (ROM)* M on $\mathcal{H}$ is defined by a set of measurement vectors $\{q_i, i \in \mathcal{I}\}$ that span subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$. Since we are not constrained by the physics of quantum mechanics, these vectors are not constrained to be orthonormal. Nonetheless, in some applications we will find it useful to impose such a constraint. A *subspace QSP measurement (SM)* on $\mathcal{H}$ is defined by a set of projection operators $\{E_i, i \in \mathcal{I}\}$ onto subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$. Here again, since we are not constrained by the physics, the projection operators and the subspaces $\mathcal{S}_i$ are not constrained to be orthogonal. The measurement of a signal x is denoted by $M(x)$.

Measurement consistency in our framework is formulated mathematically as

$$M(M(x)) = M(x). \tag{1.1}$$

Note that by our definition of measurement, if x is a signal in a signal space $\mathcal{H}$ then $M(x)$ is also a signal in $\mathcal{H}$, and can therefore be re-measured.

Quantization of the measurement outcome is imposed by requiring that the outcome signal $M(x)$ is one of a set of signals determined by the measurement M. Specifically, in analogy with the quantum mechanical determinate states we define the set of *determinate signals*, which are the signals that lie completely in one of the measurement subspaces $\{\mathcal{S}_i, i \in \mathcal{I}\}$.

The measurement M is then defined to preserve the two fundamental properties of a quantum measurement:

1. The measurement outcome is always equal to one of the determinate signals;

2. For every input signal x, (1.1) is satisfied.

1.2.2 Rank-One Measurements

A ROM M defined by a set of measurement vectors $\{q_i, i \in \mathcal{I}\}$ that span measurement subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$ is in general a nonlinear mapping between $\mathcal{H}$ and the set of determinate signals of M. With E_i denoting a projection onto $\mathcal{S}_i$, the measurement is defined such that if x is a determinate signal then $M(x) = E_i x = x$, and otherwise $M(x) = E_i x$ where

$$i = f_M(\{\langle x, q_k \rangle, k \in \mathcal{I}\}). \tag{1.2}$$

Here f_M is a (possibly probabilistic) mapping between the input signal x and the set of indices $\mathcal{I}$, that depends on the input x only through the inner products between x and the measurement vectors q_i, which are a subset of the determinate signals. For example, we may choose $f_M(x) = \arg\max\langle x, q_k \rangle$.

Note, that since $E_i x \in \mathcal{S}_i$ for any x, the outcome $M(x)$ is always a determinate signal of M, and since for any determinate signal x, $M(x) = x$, this definition of a measurement satisfies the required properties.

As an example of a ROM, suppose that the measurement input is $x = (1/2)q_1 + (\sqrt{3}/2)q_2$ where q_1 and q_2 are two orthonormal measurement vectors. Then the measurement output will be either a vector in the direction of q_1 or a vector in the direction of q_2. The particular output chosen is determined by the mapping f_M which depends on the input x only through the inner products $\langle x, q_1 \rangle = 1/2$ and $\langle x, q_2 \rangle = \sqrt{3}/2$. The measurement process is illustrated in Fig. 1-3.

As developed in detail in Chapter 4, our definition of a rank-one QSP measurement is very similar to the definition of a rank-one quantum measurement, with two main differences: First, we allow for an arbitrary mapping f_M in (1.2); in quantum mechanics f_M is unique, and is the probabilistic mapping in which q_i is chosen with probability $|\langle x, q_i \rangle|^2$. Second, the measurement vectors are not constrained to be orthonormal as in quantum mechanics.

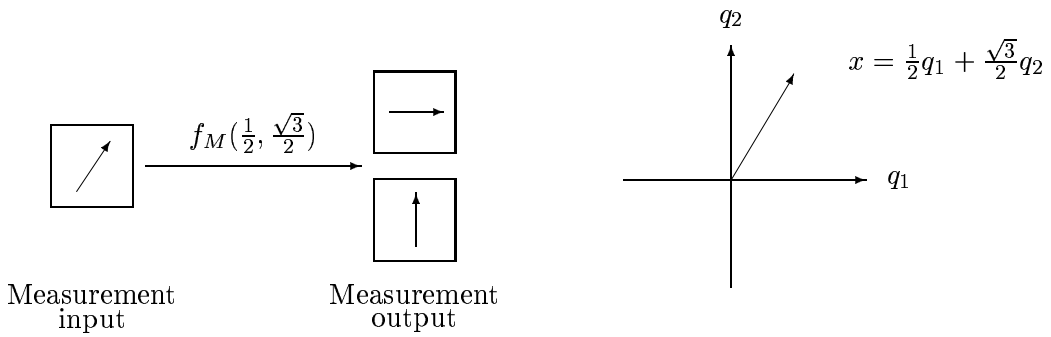


Figure 1-3: Illustration of a rank-one QSP measurement.

1.2.3 Subspace Measurements

The definition of a SM parallels that of a ROM, and borrows from the definition of a higher-rank quantum measurement.

Therefore, a SM M defined by a set of measurement projections $\{E_i, i \in \mathcal{I}\}$ that span measurement subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$ is a nonlinear mapping between $\mathcal{H}$ and the set of determinate signals of M where if x is a determinate signal then $M(x) = E_i x = x$, and otherwise $M(x) = E_i x$ where

$$i = f_M(\{\langle E_k x, E_k x\rangle, k \in \mathcal{I}\}). \tag{1.3}$$

Here f_M is a (possibly probabilistic) mapping between the input signal x and the set of indices $\mathcal{I}$, that depends on the input x only through the inner products $\{\langle E_k x, E_k x\rangle, k \in \mathcal{I}\}$.

A special case of a SM is the case in which the measurement is defined by a single projection. Then $M(x) = Ex$ for all x and the SM reduces to a linear projection operator. We refer to such a measurement as a *simple subspace measurement (SSM)*.

The subspace QSP measurement is very similar to a higher-rank quantum measurement, with three main differences: We allow for an arbitrary mapping f_M, the measurement projections are not constrained to be orthogonal, and the measurement subspaces are not constrained to be orthogonal.

1.3 Algorithm Design in the QSP Framework

Within the QSP framework, the QSP measurement plays a central role in the design of signal processing algorithms. In this framework, signals are processed by either subjecting them to

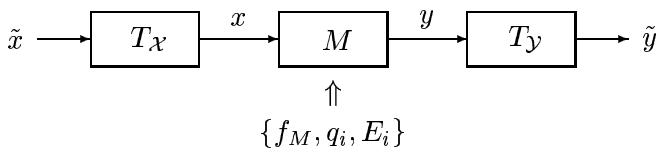


Figure 1-4: Designing algorithms using a QSP measurement.

a QSP measurement as outlined in Section 1.3.1, or by using some of the QSP measurement parameters but not directly applying the measurement, as described in Section 1.3.2.

1.3.1 Algorithm Design Based on QSP Measurements

Algorithm design

To design an algorithm using a QSP measurement we first identify the measurement vectors q_i in a ROM, or the measurement operators E_i in a SM, which specify the possible measurement outcomes. For example, in a detection scenario the measurement vectors may be equal to the transmitted signals, or may represent these signals in a possibly different space. As another example, in a scalar quantizer the measurement vectors may be chosen as a set of vectors that represent the scalar quantization levels. In a SM the measurement operators may be projections onto a set of subspaces used for signalling. We then embed the measurement vectors (projections) in a Hilbert space $\mathcal{H}$. If the signal $\tilde{x}$ to be processed does not lie in $\mathcal{H}$, then we first map it into a signal x in $\mathcal{H}$ using a mapping $T_{\mathcal{X}}$. To obtain the algorithm output we measure the representation x of the signal to be processed. If x is a determinate signal of M, then the measurement outcome is $y = M(x) = x$. Otherwise we approximate x by a determinate signal y using a mapping f_M. If necessary, the measurement outcome y may be mapped to the algorithm output $\tilde{y}$ using a mapping $T_{\mathcal{Y}}$. These basic steps are illustrated in Fig. 1-4.

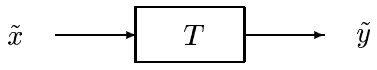
As we explore in the thesis, by choosing different input and output mappings $T_{\mathcal{X}}$ and $T_{\mathcal{Y}}$, and different measurement parameters f_M, q_i and E_i, and using the QSP measurement framework of Fig. 1-4, we can arrive at a variety of new and interesting processing techniques.

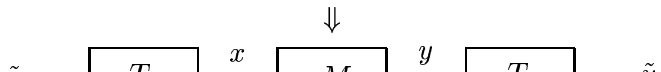
Modifying known algorithms

As we demonstrate throughout the thesis, many traditional detection and processing techniques fit naturally into the framework of Fig. 1-4. Examples include traditional and dithered quantization, sampling methods, matched-filter detection, and multiuser detection. Once an algorithm is described as a QSP measurement, modifications and extensions of the algorithm can be derived by simply changing the measurement parameters. Thus, the QSP framework provides a unified conceptual structure for a variety of traditional processing techniques, and a precise mathematical setting for generating new, potentially effective and efficient processing methods by modifying the measurement parameters.

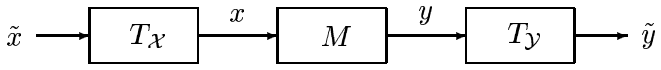
To modify an existing algorithm represented by a mapping T using the QSP framework, we first cast the algorithm as a QSP measurement M, *i.e.,* we choose an input mapping $T_{\mathcal{X}}$ and an output mapping $T_{\mathcal{Y}}$ if necessary, and the measurement parameters f_M, q_i and E_i. We then systematically change some of these parameters resulting in a modified measurement M', which can then be translated into a new signal processing algorithm represented by a mapping T'. The modifications we consider result from either imposing some of the additional constraints of quantum mechanics on the measurement parameters of M, or from relaxing some of these constraints which we do not have to impose in signal processing. These basic steps are summarized as follows:

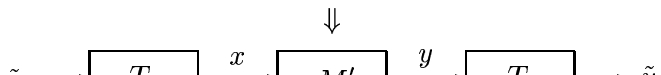
1. Original algorithm



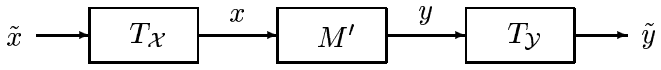


2. QSP representation

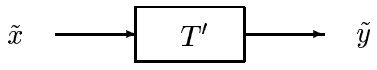




3. Modify parameters



4. Obtain new algorithm



Typical modifications of the parameters that we consider include

1. Using a probabilistic mapping f_M.

2. Imposing inner product constraints on the measurement vectors q_i.

3. Using oblique projections E_i in place of orthogonal projections.

1.3.2 Algorithm Design Using the Measurement Parameters

Another class of algorithms we develop result from processing a signal with some of the measurement parameters, and then imposing quantum mechanical constraints directly on these parameters. For example, we may view any linear processing of a signal as processing with a set of measurement vectors, and then impose inner product constraints on these vectors. Using the ideas of quantum detection we may then design linear algorithms that are optimal subject to these inner product constraints.

To generate new algorithms or modify existing algorithms we describe the algorithm as processing by one of the measurement parameters, and then modify these parameters using one of the three modifications outlined above.

In the remainder of this section we discuss each of these modifications. In Sections 1.4 and 1.5 we indicate how they will be applied to the development of new processing methods.

1.3.3 Probabilistic Mappings

The QSP framework naturally gives rise to probabilistic and randomized algorithms by letting f_M be a probabilistic mapping, emulating the quantum measurement. We expand on this idea in the context of quantization in Chapter 7 and in the context of combined measurements in Chapter 5. However, the full potential benefits of probabilistic algorithms in general resulting from the QSP framework remain an interesting area of future study.

1.3.4 Least-Squares Inner Product Shaping

One of the interesting elements of quantum mechanics is that the measurement vectors are constrained to be orthonormal. This constraint leads to some interesting problems such as the quantum detection problem described in Section 1.1.2. A fundamental problem in quantum mechanics is to construct optimal measurements subject to this constraint, that best represent a given set of state vectors. In analogy to quantum mechanics, an important feature of QSP is the idea of imposing constraints on algorithms. The QSP framework provides a systematic method for imposing such constraints: The measurement vectors are restricted to have a certain inner product structure, as in quantum mechanics. However, since we are not limited by physical laws, we are not confined to an orthogonality constraint. As part of our work, we develop methods for choosing a set of measurement vectors that

"best" represent the signals of interest, and have a specified inner product structure [27]; these methods rely on ideas and results we obtained in the context of quantum detection [26], which unlike QSP are subject to the constraints of quantum physics. Specifically, we construct measurement vectors q_i with a given inner product structure that are closest in a LS sense to a given set of vectors s_i, so that the vectors q_i are chosen to minimize the sum of the squared norms of the error vectors $e_i = q_i - s_i$. These techniques are referred to as LS inner product shaping.

The concept of LS inner product shaping is used in Chapters 8–12 to develop effective solutions to a variety of problems that result from imposing a deterministic or stochastic inner product constraint on the algorithm, and then designing optimal algorithms subject to this constraint. In each of these problems we either describe the algorithm as a QSP measurement and impose an inner product constraint on the corresponding measurement vectors, or we consider linear algorithms on which the inner product constraints can be imposed directly. We demonstrate that, even for problems without inherent inner product constraints, imposing such constraints in combination with least-squares inner product shaping leads to new processing techniques in diverse areas including frame theory, detection, covariance shaping, linear estimation and multiuser wireless communication, that often exhibit improved performance over traditional methods.

1.3.5 Oblique Projections

In a quantum measurement defined by a set of projection operators, the rules of quantum mechanics impose the constraint that the projections must be orthogonal. In QSP we may explore more general types of measurements defined by projection operators that are not restricted to be orthogonal, *i.e.,* oblique projections.

An oblique projection is a projection operator E satisfying $E^2 = E$ that is not necessarily Hermitian. The notation $E_{\mathcal{US}}$ denotes an oblique projection with range space $\mathcal{U}$ and null space $\mathcal{S}$. If $\mathcal{S} = \mathcal{U}^{\perp}$, then $E_{\mathcal{US}}$ is an orthogonal projection onto $\mathcal{U}$. An oblique projection $E_{\mathcal{US}}$ can be used to decompose x into its components in two disjoint spaces $\mathcal{U}$ and $\mathcal{S}$ that are not constrained to be orthogonal, as illustrated in Fig. 1-5.

Oblique projections are used in Chapter 5 to develop new classes of frames and effective subspace detectors, and in Chapter 6 to develop a general sampling framework for sampling and reconstruction in arbitrary spaces.

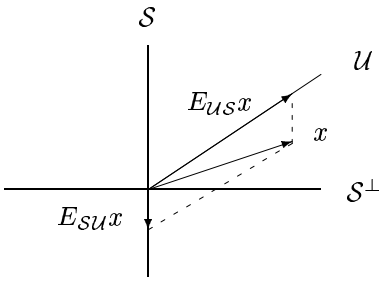


Figure 1-5: Decomposition of x into its components in $\mathcal{U}$ and in $\mathcal{S}$ given by $E_{\mathcal{U}\mathcal{S}}x$ and $E_{\mathcal{S}\mathcal{U}}x$, respectively.

1.4 Applications of Rank-One Measurements

1.4.1 QSP Quantization

In Chapter 7 we explore quantization methods suggested by the QSP framework. In particular, by emulating the quantum measurement we develop a probabilistic quantizer and show that it can be used to efficiently implement a dithered quantizer.

In dithered quantization a random signal called a dither signal is added to the input signal prior to quantization [28, 29, 30, 31]. Dithering techniques have become commonplace in applications in which data is quantized prior to storage or transmission. However, the utility of dithering techniques is limited by the computational complexity associated with generating a random process with an arbitrary joint probability distribution.

As we show in Chapter 7, a probabilistic quantizer can be used to effectively realize a dither signal with an arbitrary joint probability distribution, while requiring only the generation of one uniform random variable per input. By introducing memory into the probabilistic selection rule we derive a probabilistic quantizer that shapes the quantization noise.

1.4.2 Covariance Shaping Matched Filter Detection

As an example of the type of procedure my may follow in using the concept of LS inner product shaping and optimal QSP measurements to derive new processing methods, in Chapter 9 we consider a generic detection problem where one of a set of signals is transmitted over a noisy channel. When the additive noise is white and Gaussian, it is well known (see

e.g., [11, 32]) that the receiver which maximizes the probability of correct detection is the matched filter (MF) receiver. If the noise is not Gaussian, then the MF receiver does not necessarily maximize the probability of correct detection. However, it is still used as the receiver of choice in many applications since the optimal detector for non-Gaussian noise is typically nonlinear (see *e.g.,* [33] and references therein), and depends on the noise distribution which may not be known.

By describing the MF detector as a QSP measurement, and imposing an inner product constraint on the measurement vectors, we derive a new class of receivers consisting of a bank of correlators with correlating signals that are matched to a set of signals with a specified inner product structure, and are closest in a LS sense to the transmitted signals. These receivers depend only on the transmitted signals, so that they do not require knowledge of the noise distribution or the channel signal-to-noise ratio (SNR).

Alternatively, we show that the modified receivers can be implemented as an MF demodulator followed by an optimal covariance shaping transformation, that optimally shapes the correlation of the outputs of the MF prior to detection. This equivalent representation leads to the concept of minimum mean-squared error (MMSE) covariance shaping, which we consider in its most general form in Chapter 10.

As we demonstrate through simulation, when the additive noise is non-Gaussian these receivers can significantly increase the probability of correct detection over the MF receiver, with only a minor impact in performance when the noise is Gaussian.

1.4.3 MMSE Covariance Shaping

Drawing from the quantum detection problem, we can develop new classes of linear algorithms that result from imposing a deterministic or stochastic inner product constraint on the algorithm *i.e.,* a covariance constraint, and then using the results we obtained in the context of quantum detection to derive optimal algorithms subject to this constraint. In particular, we may extend the concept of LS inner product shaping suggested by the quantum detection framework to develop optimal algorithms that minimize a stochastic mean-squared error (MSE) criterion subject to a covariance constraint.

As an example of this approach, in Chapter 10 we exploit the concept of LS inner product shaping to the development of a new viewpoint towards whitening and other covariance shaping problems.

Data shaping arises in a variety of contexts in which it is useful to shape the covariance of a data vector either prior to subsequent processing, or to control the spectral shape after processing [34, 35]. As is well known, the linear transformation that shapes the covariance of a data vector is not unique. While in some applications certain conditions might be imposed on the transformation such as causality or symmetry, with the exception of the work in [36, 37, 38, 39, 40, 41] which explicitly relies on the optimality properties developed in this thesis, there have been no general assertions of optimality for various choices of a linear shaping transformation. In particular, the shaped vector may not be "close" to the original data vector. If this vector undergoes some noninvertible processing, or is used as an estimator of some unknown parameters represented by the data, then we may wish to choose the covariance shaping transformation so that the shaped output is close to the original data in some sense.

Building upon the concept of LS inner product shaping, we propose choosing an optimal shaping transformation that results in a shaped vector that is as close as possible to the original vector in an MSE sense, which we refer to as *MMSE covariance shaping*. The MMSE covariance shaping problem can be interpreted as a stochastic analogue of the LS inner product shaping, in which the covariance shaping transformation is designed to minimize the MSE between its input and output.

1.4.4 Covariance Shaping Least-Squares Estimation

As another example of an algorithm suggested by the quantum detection framework, where we use the ideas of least-squares inner product shaping to design an optimal linear algorithm subject to a stochastic inner product constraint, in Chapter 11 we derive a new linear estimator for the unknown deterministic parameters in a linear model. The estimator is chosen to minimize an MSE criterion, subject to a constraint on the covariance of the estimator. This new estimator is defined as the *covariance shaping least-squares* (CSLS) estimator.

Many signal processing estimation problems can be represented by the linear model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, where $\mathbf{H}$ is a known matrix, $\mathbf{x}$ is a vector of unknown deterministic parameters to be estimated, and $\mathbf{w}$ is a random vector. A common approach to estimating the parameters $\mathbf{x}$ is to restrict the estimator to be linear in the data $\mathbf{y}$, and then find the linear estimate of $\mathbf{x}$ that results in an estimated data vector that is as close as possible to the given data vector

$\mathbf{y}$ in a (weighted) LS sense, so that it minimizes the total squared error in the observations [42, 43, 44, 45]. A difficulty often encountered when using the LS estimator to estimate the parameters $\mathbf{x}$, is that the error in estimating $\mathbf{x}$ can have a large variance and a covariance structure with a very high dynamic range. This is due to the fact that in many cases the data vector $\mathbf{y}$ is not very sensitive to changes in $\mathbf{x}$, so that a large error in estimating $\mathbf{x}$ may translate into a small error in estimating the data vector $\mathbf{y}$, in which case the LS estimate may result in a poor estimate of $\mathbf{x}$. This effect is especially predominant at low to moderate SNR, where the data vector $\mathbf{y}$ is typically affected more by the noise than by changes in $\mathbf{x}$; the exact SNR range will depend on the properties of the model matrix $\mathbf{H}$.

The CSLS estimator is a biased estimator directed at improving the performance of the traditional LS estimator at low to moderate SNR by choosing the estimate to minimize the (weighted) total error variance in the observations subject to a constraint on the covariance of the estimation error, so that we control the dynamic range and spectral shape of the covariance of the estimation error.

Various modifications of the LS estimator under the linear model assumption have been previously proposed in the literature. Among the more prominent alternatives are the ridge estimator [46] (also known as Tikhonov regularization [47]) and the shrunken estimator [48]. We show that both the ridge estimator and the shrunken estimator can be formulated as CSLS estimators, which allows us to interpret these estimators as the estimators that minimize the total error variance in the observations, from all linear estimators with the same covariance.

We develop two equivalent representations of the CSLS estimator. In the first, the CSLS estimator is expressed as a LS estimator followed by a weighted MMSE (WMMSE) covariance shaping transformation, and in the second, the CSLS estimator is expressed as an MF estimator followed by an MMSE covariance shaping transformation.

Analysis of the MSE of the CSLS estimator demonstrates that over a wide range of SNR, the CSLS estimator results in a lower MSE than the traditional LS estimator, for all values of the unknown parameters. The simulations presented in Chapter 11 strongly suggest that the CSLS estimator can significantly decrease the MSE of the estimation error over the LS estimator for a wide range of SNR values.

1.4.5 Covariance Shaping Multiuser Detection

In Chapter 12 we consider an application of the CSLS estimator to the problem of suppressing interference in multiuser wireless communication systems. Specifically, we develop a new linear multiuser receiver for synchronous code-division multiple-access (CDMA) systems, in which different users transmit information over a joint channel by modulating distinct signature vectors. The receiver is referred to as the *covariance shaping multiuser (CSMU) receiver*.

Multiuser receivers for detection of CDMA signals try to mitigate the effect of multiple-access interference (MAI) and background noise. These include the optimal multiuser receiver, the linear MMSE receiver, the decorrelator, and the MF receiver [49]. Both the optimal receiver and the linear MMSE receiver require knowledge of the channel parameters, namely the noise level and the received amplitudes of the users' signals. On the other hand, the MF and the decorrelator receivers are linear receivers that only require knowledge of the signature vectors. The MF optimally compensates for the white noise, but does not exploit the structure of the MAI; the decorrelator optimally rejects the MAI, but does not consider the white noise. Like the MF and the decorrelator, the CSMU receiver does not require knowledge of the channel parameters and relies only on knowledge of the signature vectors. However, by contrast to the MF and the decorrelator, this receiver takes both the background noise and the MAI into account.

Building upon the properties of the CSLS estimator, we develop three equivalent representations of the CSMU receiver. In the first, the receiver consists of a bank of correlators with correlating vectors that have a specified inner product structure, and are closest in a LS sense to the users' signature vectors. In the second, the receiver consists of a decorrelator demodulator [50] followed by a WMMSE covariance shaping transformation. In the third, the receiver consists of an MF demodulator followed by an MMSE covariance shaping transformation.

To evaluate the performance of the receiver, we derive exact and approximate expressions for the probability of bit error. We also develop methods to analyze the output signal-to-interference+noise ratio (SINR) in the large system limit. We show that the SINR converges to a deterministic limit, and compare this limit to the known SINR limits for the decorrelator, MF and linear MMSE receivers [51, 52, 53]. The analysis suggests that this

modified receiver can lead to improved performance over the decorrelator and MF receiver, and can approach the performance of the linear MMSE receiver over a wide range of channel parameters without requiring knowledge of these parameters.

1.5 Applications of Subspace Measurements

1.5.1 Simple Subspace Measurements

A SSM is equivalent to a linear projection operator. Numerous signal processing and detection algorithms based on orthogonal projections have been developed. Algorithms based on oblique projections have received much less attention in the signal processing literature. Recently, oblique projections have been applied to various detection problems [54, 55], to signal and parameter estimation [56], to computation of wavelet transforms [57], and to the formulation of consistent interpolation and sampling methods [16, 58].

In [16] the authors develop consistent reconstruction algorithms, in which the reconstructed signal is in general not equal to the original signal, but nonetheless yields the same samples. Using a SSM corresponding to an oblique projection operator, in Chapter 6 we extend the results of [16] to a broader framework that can be applied to arbitrary subspaces of an arbitrary Hilbert space. The algorithms we develop yield perfect reconstruction for signals in a subspace of $\mathcal{H}$, and consistent reconstruction for arbitrary signals. This framework leads to some new sampling theorems, and can also be used to construct signals from a certain class with prescribed properties. For example, we can use this framework to construct a finite-length signal with specified low-pass coefficients, or an odd signal with specified local averages.

1.5.2 Subspace Coding and Decoding

Subspace measurements also lead to interesting and potentially useful coding and decoding methods for communication-based applications over a variety of channel models. In particular, in Chapter 4 we develop a subspace approach for transmitting information over a noisy channel in which the information is encoded in disjoint subspaces. To detect the information, we design a receiver based on a SM and show that for a certain class of channel models this receiver implements a generalized likelihood ratio test. Although the discussion constitutes a rather preliminary exploration of such coding techniques, it represents an in-

teresting and potentially useful model for communication in many contexts. In particular, decoding methods suggested by the QSP framework may prove useful in the context of recent advances in multiple-antenna coding techniques [59, 60].

1.6 Combined Measurements

An interesting class of measurements in quantum mechanics results from restricting measurements to a subspace in which the quantum system is known a priori to lie. This leads to the notion of generalized measurements, or positive operator-valued measures (POVMs) [61, 62]. It can be shown that a generalized measurement on a quantum system can be implemented by performing a standard measurement on a larger system. Alternatively, we can view a generalized quantum measurement as a combination of a standard measurement followed by an orthogonal projection onto a lower space.

Drawing from the quantum mechanical POVM, in Chapter 5 we consider combined QSP measurements. The QSP analogue of a quantum POVM is a ROM followed by an SSM corresponding to an orthogonal projection operator. Since the QSP framework does not depend on the physics associated with quantum mechanics, we may extend the notion of a (physically realizable) POVM to include other forms of combined QSP measurements, where we perform any two measurements successively. We show that such measurements lead to a variety of extensions and rich insights into frames, to new classes of frames, and to the concept of oblique frame expansions. This framework also leads to subspace MF detectors and randomized algorithms for improving worst-case performance.

1.6.1 Combined Measurements and Tight Frames

Emulating the quantum POVM leads to combined measurements where a ROM is followed by an orthogonal projection onto a subspace $\mathcal{U}$. Such measurements are characterized by an effective set of measurement vectors. We show that the family of possible effective measurement vectors in $\mathcal{U}$ is equal to the family of rank-one POVMs on $\mathcal{U}$, and is precisely the family of (normalized) tight frames for $\mathcal{U}$.

Frames are generalizations of bases which lead to redundant signal expansions [63, 64]. A *frame* for a Hilbert space $\mathcal{U}$ is a set of not necessarily linearly independent vectors that spans $\mathcal{U}$ and has some additional properties. Frames were first introduced by Duffin and

Schaeffer [63] in the context of nonharmonic Fourier series, and play an important role in the theory of nonuniform sampling [63, 64, 65]. Recent interest in frames has been motivated in part by their utility in analyzing wavelet expansions [66, 67]. A *tight frame* is a special case of a frame for which the reconstruction formula is particularly simple, and is reminiscent of an orthogonal basis expansion, even though the frame vectors in the expansion are linearly dependent.

Exploiting the equivalence between tight frames and quantum POVMs, we develop frame-theoretic analogues of various quantum-mechanical concepts and results [68]. In particular, motivated by the construction of optimal LS quantum measurements [26], we consider the problem of constructing optimal LS tight frames for a subspace $\mathcal{U}$ from a given set of vectors that span $\mathcal{U}$.

The problem of frame design has received relatively little attention in the frame literature. A popular frame construction from a given set of vectors is the canonical frame [69, 70, 71, 72], first proposed in the context of wavelets in [73]. The canonical frame is relatively simple to construct, can be determined directly from the given vectors, and plays an important role in wavelet theory [74, 14, 75]. In Chapter 8 we show that the canonical frame vectors are proportional to the LS frame vectors.

This relationship between combined measurements and frames suggests an alternative definition of frames in terms of projections of a set of linearly independent signals in a larger space. This perspective provides additional insights into frames, and suggests a systematic approach for generating new classes of frames by changing the properties of the signals or changing the properties of the projection.

1.6.2 Geometrically Uniform Frames

In the context of a single QSP measurement, we have seen that imposing inner product constraints on the measurement vectors of a ROM leads to interesting new processing techniques. Similarly, in the context of combined measurements imposing such constraints leads to the definition of *geometrically uniform frames* [76]. This class of frames is highly structured resulting in nice computational properties, and possesses strong symmetries that may be advantageous in a variety of applications such as channel coding [77, 78, 79] and multiple description source coding [80]. In Chapter 5 we present some results regarding these frames which also appear in [76].

1.6.3 Consistent Sampling and Oblique Dual Frame Vectors

We also explore extensions of frames that result from choosing an oblique projection operator onto $\mathcal{U}$. In this case the the measurement is described in terms of two sets of effective measurement vectors, where the first set forms a frame for $\mathcal{U}$ and the second set forms what we define as an *oblique dual frame*. The frame operator corresponding to these vectors is referred to as an *oblique dual frame operator*, and is a generalization of the well known dual frame operator [69]. As we show in Chapter 5, these frame vectors have properties that are very similar to those of the conventional dual frame vectors. However, in contrast with the dual frame vectors, they are not constrained to lie in the same space as the original frame vectors. Thus, using oblique dual frame vectors we can extend the notion of a frame expansion to include redundant expansions in which the analysis frame vectors and the synthesis frame vectors lie in different spaces.

Based on the concept of oblique dual frame vectors, in Chapter 6 we develop *redundant* consistent sampling procedures with (almost) arbitrary sampling and reconstruction spaces. By allowing for arbitrary spaces, the sampling and reconstruction algorithms can be greatly simplified in many cases with only a minor increase in approximation error [16, 81, 17, 82, 83, 84]. Using oblique dual frame vectors we can further simplify the sampling and reconstruction processes while still retaining the flexibility of choosing the spaces almost arbitrarily, due to the extra degrees of freedom offered by the use of frames that allow us to construct frames with prescribed properties [66, 85]. Furthermore, if the measurements are quantized prior to reconstruction, then as we show the average power of the reconstruction error using this redundant procedure can be reduced by as much as the redundancy of the frame in comparison with the nonredundant procedure.

1.7 Thesis Outline

This thesis can roughly be divided into two parts: Chapters 2–5 provide the necessary background, and develop the QSP framework. Chapters 6–12 discuss applications of QSP to sampling procedures, quantization, detection, covariance shaping and estimation problems.

Chapters 2–3 summarize relevant background material: In Chapter 2 we review elements of linear algebra that are used in the development of QSP, and derive new results that are used in applications throughout the thesis. In Chapter 3 we provide a brief introduction

to quantum states and measurements, and recapitulate some results on optimal quantum measurements in the context of the quantum detection problem.

In Chapters 4–5 we develop the QSP measurement framework: Chapter 4 considers ROMs and SMs. Throughout the chapter we provide examples of signal processing techniques that can be cast in terms of QSP measurements, as well as generalizations of these algorithms and new algorithms that stem from changing the measurement parameters. Chapter 5 considers various forms of combined measurements. In particular, we develop the oblique dual frame vectors and discuss their key properties.

Chapters 6-12 consider applications of QSP, focusing on applications of SSMs in Chapter 6 and on applications of ROMs in Chapters 7–12.

In Chapter 6 we provide a general framework for redundant and nonredundant consistent sampling procedures with arbitrary sampling and reconstruction spaces.

Chapter 7 discusses quantization methods resulting from ROMs.

In Chapter 8 we systematically construct optimal ROMs from a given set of vectors, with measurement vectors that have a specified inner product structure and are closest in a LS sense to a given set of vectors.

Chapters 9–12 focus specifically on applications of LS inner product shaping. Chapter 9 develops a new viewpoint toward MF detection based on optimal ROMs, that leads to a stochastic analogue of the LS inner product shaping problem, taking on the form of an MMSE covariance shaping problem, considered in Chapter 10. In Chapter 11 we derive the covariance shaping LS estimator, based on which, in Chapter 12, we develop efficient techniques for suppressing interference in multiuser wireless settings.

Chapter 2

Signal Spaces

Underlying the development of QSP is the signal space viewpoint toward signal processing we take on in this thesis, in which signals are regarded as vectors in an abstract Hilbert space referred to as the *signal space*.

This chapter is intended to summarize the key results that we exploit in the development of QSP, and establish the signal space notation used throughout the thesis. We also derive new results that will be used in applications in subsequent chapters. In particular, we provide an explicit construction of oblique projections on arbitrary Hilbert spaces, develop an alternative characterization of oblique pseudoinverses, and establish the key properties of transjectors[1] (partial isometries) [68, 86].

Background material on Hilbert spaces helpful to understanding the material presented in this chapter can be found in [87, 88, 64].

Throughout the thesis $\mathbb{Z}$ denotes the set of integers and $\mathcal{I} \subseteq \mathbb{Z}$ denotes a countable index set.

2.1 Hilbert Spaces

2.1.1 Vector Spaces

A complex vector space $\mathcal{V}$ over the complex numbers $\mathbb{C}$ is a set of elements called vectors, together with vector addition and scalar multiplication by elements of $\mathbb{C}$ such that $\mathcal{V}$ is closed under both operations. We will assume throughout that all vector spaces are complex.

[1] This nomenclature was suggested by G. D. Forney, Jr..

A set $\mathcal{W}$ of vectors in $\mathcal{V}$ is a *subspace* of $\mathcal{V}$ if it is closed under addition and scalar multiplication.

Two subspaces $\mathcal{W}$ and $\mathcal{S}$ are said to be *disjoint* if they intersect only at the zero vector, *i.e.*, $\mathcal{V} \cap \mathcal{W} = \{0\}$.

The *sum* of two closed subspaces $\mathcal{V}$ and $\mathcal{W}$, denoted $\mathcal{V} + \mathcal{W}$, is the set of all vectors of the form $x = v + w$ where $v \in \mathcal{V}$ and $w \in \mathcal{W}$. The *direct sum* of $\mathcal{V}$ and $\mathcal{W}$, denoted $\mathcal{V} \oplus \mathcal{W}$, is the sum of two *disjoint* subspaces. If $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$, then $\mathcal{V} \cap \mathcal{W} = \{0\}$ and any $x \in \mathcal{H}$ can be decomposed *uniquely* as $x = v + w$, with $v \in \mathcal{V}$ and $w \in \mathcal{W}$.

A vector x is a *limit point* of a subspace $\mathcal{W}$ if there exists a sequence of vectors $x_i \in \mathcal{W}$ such that $x_i \to x$. The *closure* $\mathcal{W}^c$ of a subspace $\mathcal{W}$ is the set of all points x that are limit points of $\mathcal{W}$. A subspace $\mathcal{W}$ is *closed* if $\mathcal{W} = \mathcal{W}^c$.

2.1.2 Hilbert Spaces

We now add geometric structure to a vector space in the form of an *inner product* relation between pairs of vectors, which also induces a distance measure or metric on the space.

Definition 2.1. *The inner product on the vector space $\mathcal{V}$, denoted $\langle x, y \rangle$, is a mapping from $\mathcal{V}$ to $\mathbb{C}$ that satisfies:*

1. $\langle x, y \rangle = \langle y, x \rangle^*$;
2. $\langle x, ay + bz \rangle = a\langle x, y \rangle + b\langle x, z \rangle$;
3. $\langle x, x \rangle \geq 0$ *and* $\langle x, x \rangle = 0$ *if and only if* $x = 0$,

where $(\cdot)^*$ denotes the conjugate. The norm *of a vector x is defined by $\|x\| = \sqrt{\langle x, x \rangle}$, and the* distance *between x and y is defined by $\|x - y\|$.*

Any mapping satisfying properties (1)–(3) above is a valid inner product. The choice of inner product will depend on the properties of the underlying vector space $\mathcal{V}$, as well as the particular application. For example, let $\mathcal{V}$ be the vector space of all finite-energy signals. Then a vector $x \in \mathcal{V}$ represents a signal $x(t)$ with $\int_{t=-\infty}^{\infty} |x(t)|^2 dt < \infty$. We can immediately verify that $\langle x, y \rangle = \int_{t=-\infty}^{\infty} x^*(t) y(t) dt$ is a valid inner product on this space.

Two vectors x, y are said to be orthogonal in $\mathcal{V}$ if $\langle x, y \rangle = 0$. If $\mathcal{W}$ is a subspace of $\mathcal{V}$,

then the *orthogonal complement* of $\mathcal{W}$ in $\mathcal{V}$, denoted by $\mathcal{W}^\perp$, is defined as

$$\mathcal{W}^\perp = \{x \in \mathcal{V} | \langle x, y \rangle = 0 \text{ for all } y \in \mathcal{W}\}. \tag{2.1}$$

We are now ready to define a *Hilbert space.*

Definition 2.2. *A* Hilbert space *is a complete*[2] *vector space together with an inner product.*

Some examples of Hilbert spaces that are used throughout the thesis are considered below.

Example 2.1 (The Hilbert space l_2). $\mathcal{H} = l_2$ denotes the set of all sequences $x = \{x_i\}$ with $x_i \in \mathbb{C}$ that are absolutely square summable, *i.e.,* $\sum_{i=1}^{\infty} |x_i|^2 < \infty$. The inner product on l_2 is defined by $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i^* y_i$.

Example 2.2 (The Hilbert space L_2). $\mathcal{H} = L_2$ denotes the set of all functions $x = x(t)$ that are absolutely square integrable, *i.e.,* $\int_{t=-\infty}^{\infty} |x(t)|^2 dt < \infty$. The inner product on L_2 is defined by $\langle x, y \rangle = \int_{t=-\infty}^{\infty} x^*(t) y(t) dt$.

Example 2.3 (The Hilbert space $\mathbb{C}^m$). $\mathcal{H} = \mathbb{C}^m$ denotes the set of all m-dimensional vectors $x = \mathbf{x}$ with components in $\mathbb{C}$. The inner product on $\mathbb{C}^m$ is defined by $\langle x, y \rangle = \mathbf{x}^* \mathbf{y} = \sum_{i=1}^{m} \mathbf{x}_i^* \mathbf{y}_i$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the ith component of $\mathbf{x}$ and $\mathbf{y}$, respectively.

A *signal space* is a Hilbert space whose elements are signals; we refer to these elements as vectors or signals interchangeably. Vectors in $\mathbb{C}^m$ (m arbitrary) are denoted by boldface lowercase letters, *e.g.,* $\mathbf{x}$. The ith component of $\mathbf{x}$ will be denoted by $\mathbf{x}_i$. A sequence in l_2 is denoted by a lowercase letter, *e.g.,* x, and the ith element of x is denoted by x_i.

2.2 Bases

One of the useful features of a signal space $\mathcal{H}$ is that every signal $x \in \mathcal{H}$ can be represented by a unique sequence of scalars using a set of vectors that form a *basis* for $\mathcal{H}$.

Definition 2.3. *A set of vectors $\{x_i \in \mathcal{H}, i \in \mathcal{I}\}$ is a* Schauder basis[3] *for $\mathcal{H}$ if to each vector $x \in \mathcal{H}$ there corresponds a unique sequence of scalars $a_i \in \mathbb{C}$ such that $x = \sum_{i \in \mathcal{I}} a_i x_i$. The* dimension *of $\mathcal{H}$ is equal to the cardinality of $\mathcal{I}$.*

[2] A subspace $\mathcal{V}$ is *complete* if any Cauchy sequence $x_i \in \mathcal{V}$ converges to a vector $x \in \mathcal{V}$, where a *Cauchy sequence* is a sequence of vectors x_i that satisfies $\|x_k - x_i\| \to 0$ as $k, i \to \infty$.

[3] We will use the term basis to denote a Schauder basis.

If the vectors $\{x_i, i \in \mathcal{I}\}$ form a basis for $\mathcal{H}$, then any $x \in \mathcal{H}$ has a unique decomposition of the form $x = \sum_{i\in\mathcal{I}} a_i x_i$, where $a_i \in \mathbb{C}$. However, the coefficients a_i are in general not guaranteed to be in l_2, leading to expansions that are not necessarily stable. To ensure stability, we introduce the definition of a Riesz basis.

Definition 2.4. *A sequence $\{x_i \in \mathcal{H}, i \in \mathcal{I}\}$ is a Riesz basis for $\mathcal{H}$ if it is complete*[4]*, and there exists constants $\alpha > 0$ and $\beta < \infty$ such that*

$$\alpha \sum_{i\in\mathcal{I}} |a_i|^2 \leq \left\| \sum_{i\in\mathcal{I}} a_i x_i \right\|^2 \leq \beta \sum_{i\in\mathcal{I}} |a_i|^2, \tag{2.2}$$

for all $a \in l_2$. Then for any $x \in \mathcal{H}$

$$\alpha \|x\|^2 \leq \sum_{i\in\mathcal{I}} |\langle x, x_i\rangle|^2 \leq \beta \|x\|^2. \tag{2.3}$$

After introducing set transformations in Section 2.6, in Section 2.6.1 we will use these transformations to determine the coefficients a_i in a Riesz basis expansion of an arbitrary signal x, and we will show that (2.3) implies that the sequence of coefficients a is in l_2.

Note that any basis for a finite-dimensional space is a Riesz basis. We also note that (2.2) implies that the span of a Riesz basis is closed [89].

Basis expansions are pervasive in signal processing applications. In particular, they form the foundation of Fourier analysis and modern sampling techniques, in which the coefficients a_i are interpreted as "samples" of a signal we wish to represent [16, 14, 89, 81, 17, 18, 19].

2.3 Linear Transformations

Linear transformations play an important role in the development of QSP. Of particular importance are the projection operator discussed in Section 2.4, and the set transformation defined in Section 2.6.

Definition 2.5. *Let $\mathcal{H}$ and $\mathcal{S}$ be Hilbert spaces. T is a linear transformation from $\mathcal{H}$ to $\mathcal{S}$, denoted T: $\mathcal{H} \to \mathcal{S}$, if every $x \in \mathcal{H}$ is mapped to one and only one $y \in \mathcal{S}$, and $T(c_1x_1 + c_2x_2) = c_1T(x_1) + c_2T(x_2)$ for all $x_1, x_2 \in \mathcal{H}$ and $c_1, c_2 \in \mathbb{C}$.*

[4] A sequence $\{x_i\}$ is complete in $\mathcal{H}$ if the closure of the span of $\{x_i\}$ equals $\mathcal{H}$.

We denote general linear transformations by uppercase letters. $I_{\mathcal{H}}$ denotes the identity transformation on the space $\mathcal{H}$, and 0 denotes the zero transformation. Matrices are denoted by boldface uppercase letters. In particular, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix.

A transformation $T\colon\ \mathcal{H} \to \mathcal{S}$ is *continuous* if $x_i \to x$ implies that $Tx_i \to Tx$ for every $x \in \mathcal{H}$. The *adjoint* of a continuous linear transformation is defined as follows.

Definition 2.6. *The adjoint of a continuous linear transformation $T\colon\ \mathcal{H} \to \mathcal{S}$ is the unique continuous linear transformation $T^*\colon\ \mathcal{S} \to \mathcal{H}$ such that $\langle Tx, y\rangle = \langle x, T^*y\rangle$ for all $x \in \mathcal{H}$, $y \in \mathcal{S}$.*

A *linear operator* is a continuous linear transformation of a Hilbert space onto itself: $T\colon\ \mathcal{H} \to \mathcal{H}$. An operator $T\colon\ \mathcal{H} \to \mathcal{H}$ is *unitary* if $T^*T = TT^* = I_{\mathcal{H}}$. An operator $T\colon\ \mathcal{H} \to \mathcal{H}$ is *Hermitian (self-adjoint)* if $T^* = T$.

2.3.1 Subspaces Associated with a Linear Transformation

With every linear transformation $T\colon\ \mathcal{H} \to \mathcal{S}$ we associate 4 subspaces: The null space (kernel) $\mathcal{N}(T)$ of T, the orthogonal complement $\mathcal{N}(T)^\perp$ of $\mathcal{N}(T)$ in $\mathcal{H}$, the range space (image) $\mathcal{R}(T)$ of T, and the orthogonal complement $\mathcal{R}(T)^\perp$ of $\mathcal{R}(T)$ in $\mathcal{S}$.

The null space of T is the set of vectors $x \in \mathcal{H}$ for which $Tx = 0$. The range of T is the set of vectors $y \in \mathcal{S}$ for which there exists an $x \in \mathcal{H}$ such that $y = Tx$. The definition of $\mathcal{N}(T)^\perp$ and $\mathcal{R}(T)^\perp$ follow immediately from (2.1).

A transformation $T\colon\ \mathcal{H} \to \mathcal{S}$ is *injective* if for any $x \neq y$ we have that $Tx \neq Ty$, which implies that $\mathcal{N}(T) = \{0\}$. $T\colon\ \mathcal{H} \to \mathcal{S}$ is *surjective* if $\mathcal{R}(T) = \mathcal{S}$. A transformation is *bijective* if it is both injective and surjective.

A linear operator $T\colon\ \mathcal{H} \to \mathcal{H}$ is invertible if and only if it is bijective.

The subspaces associated with a linear transformation T are related to the subspaces associated with T^*, as incorporated in the following proposition [87].

Proposition 2.1. *Let $T\colon\ \mathcal{H} \to \mathcal{S}$ be a continuous linear transformation. Then*

1. $\mathcal{N}(T) = \mathcal{R}(T^*)^\perp$;
2. $\mathcal{N}(T)^\perp = \mathcal{R}(T^*)^c$;
3. $\mathcal{N}(T^*) = \mathcal{R}(T)^\perp$;

4. $\mathcal{N}(T^*)^\perp = \mathcal{R}(T)^c$,

where $\mathcal{R}(\cdot)^c$ *denotes the closure of* $\mathcal{R}(\cdot)$*. If* T *is Hermitian, then* $\mathcal{N}(T) = \mathcal{R}(T)^\perp$.

The spaces associated with a transformation $T\colon\ \mathcal{H} \to \mathcal{S}$ can be used to decompose $\mathcal{H}$ and $\mathcal{S}$ into direct sums of smaller subspaces. To this end we rely on the following pair of propositions [87].

Proposition 2.2. *If* $\mathcal{V}$ *is a closed linear subspace of a Hilbert space* $\mathcal{H}$*, then* $\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp$.

Proposition 2.3. *The null space* $\mathcal{N}(T)$ *of a linear transformation* $T\colon\ \mathcal{H} \to \mathcal{S}$ *is a closed subspace of* $\mathcal{H}$*. The orthogonal complement* $\mathcal{V}^\perp$ *of an arbitrary subspace* $\mathcal{V} \subseteq \mathcal{H}$ *is also a closed subspace of* $\mathcal{H}$.

Combing Propositions 2.3 and 2.2, we can decompose $\mathcal{H}$ as

$$\mathcal{H} = \mathcal{N}(T) \oplus \mathcal{N}(T)^\perp. \tag{2.4}$$

Then any $x \in \mathcal{H}$ can be expressed uniquely as $x = x_{\mathcal{N}} + x_{\mathcal{N}^\perp}$ where $x_{\mathcal{N}} \in \mathcal{N}(T)$, $x_{\mathcal{N}^\perp} \in \mathcal{N}(T)^\perp$, and $\langle x_{\mathcal{N}}, x_{\mathcal{N}^\perp} \rangle = 0$.

We would like to obtain a similar decomposition of $\mathcal{S}$ in terms of $\mathcal{R}(T)$ and $\mathcal{R}(T)^\perp$. However, since $\mathcal{R}(T)$ is not necessarily closed we cannot apply Proposition 2.2 directly. Instead, we substitute $\mathcal{V} = \mathcal{R}(T)^c$ in Proposition 2.2, which leads to the decomposition

$$\mathcal{S} = \mathcal{R}(T)^c \oplus \mathcal{R}(T)^\perp. \tag{2.5}$$

Then any $y \in \mathcal{S}$ can be expressed uniquely as $y = y_{\mathcal{R}} + y_{\mathcal{R}^\perp}$ where $y_{\mathcal{R}} \in \mathcal{R}(T)^c$, $y_{\mathcal{R}^\perp} \in \mathcal{R}(T)^\perp$, and $\langle y_{\mathcal{R}}, y_{\mathcal{R}^\perp} \rangle = 0$.

Decomposing $\mathcal{H}$ and $\mathcal{S}$ as in (2.4) and (2.5) respectively, we may describe the action of T and T^* on each of these subspaces. Specifically, from Proposition 2.1 we have that T maps $\mathcal{N}(T)^\perp \subseteq \mathcal{H}$ to $\mathcal{R}(T) \subseteq \mathcal{S}$, and $\mathcal{N}(T) \subseteq \mathcal{H}$ to 0. T^* maps $\mathcal{R}(T)^c \subseteq \mathcal{S}$ to $\mathcal{N}(T)^\perp \subseteq \mathcal{H}$, and $\mathcal{R}(T)^\perp \subseteq \mathcal{S}$ to 0. The action of T and T^* is illustrated in Fig. 2-1.

The direct sum decompositions of (2.4) and (2.5) are not unique. Specifically, there are many choices of subspaces $\mathcal{V} \subseteq \mathcal{H}$ with $\mathcal{V} \neq \mathcal{N}(T)^\perp$ such that $\mathcal{H} = \mathcal{N}(T) \oplus \mathcal{V}$. Then any $x \in \mathcal{H}$ can be decomposed uniquely into its components in $\mathcal{N}(T)$ and $\mathcal{V}$; however, these components are not necessarily orthogonal. Similarly, there are many choices of subspaces

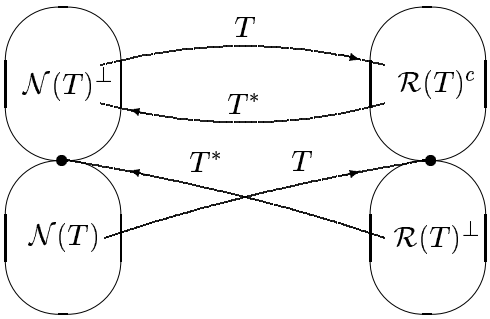


Figure 2-1: The action of T and T^* on the subspaces $\mathcal{N}(T), \mathcal{N}(T)^{\perp}, \mathcal{R}(T)^{c}$ and $\mathcal{R}(T)^{\perp}$.

$\mathcal{W} \subseteq \mathcal{S}$ with $\mathcal{W} \neq \mathcal{R}(T)^{c}$ such that $\mathcal{S} = \mathcal{R}(T)^{c} \oplus \mathcal{W}$, so that any $y \in \mathcal{S}$ can be decomposed uniquely into its components in $\mathcal{R}(T)^{c}$ and $\mathcal{W}$, where these components are not necessarily orthogonal. To decompose an arbitrary $x \in \mathcal{H}$ into its possibly non-orthogonal components in the appropriate subspaces we now discuss projection operators.

2.4 Projection Operators

A linear operator T: $\mathcal{H} \rightarrow \mathcal{H}$ is a *projection* if $T = T^2$. We distinguish between two different types of projections: Hermitian projections (orthogonal projections) for which $T = T^*$ and non-Hermitian projections (oblique projections). Orthogonal projections have enjoyed widespread use in the signal processing literature; oblique projections have received much less attention.

A projection with range equal to $\mathcal{V}$ and null space equal to $\mathcal{W}$ is denoted by $E_{\mathcal{V}\mathcal{W}}$, and is called a projection onto $\mathcal{V}$ along $\mathcal{W}$. Since $\mathcal{W}$ is not necessarily equal to $\mathcal{V}^{\perp}$, this projection in general is not constrained to be an orthogonal projection, *i.e.,* it is an oblique projection [90, 91, 92]. If $\mathcal{W} = \mathcal{V}^{\perp}$, then $E_{\mathcal{V}\mathcal{W}}$ is an orthogonal projection onto $\mathcal{V}$, denoted by $P_{\mathcal{V}}$.

2.4.1 Orthogonal Projection Operators

An orthogonal projection with range equal to $\mathcal{V} \subseteq \mathcal{H}$ is denoted by $P_{\mathcal{V}}$. Since $P_{\mathcal{V}}$ is Hermitian, we have immediately from Proposition 2.1 that $\mathcal{N}(P_{\mathcal{V}}) = \mathcal{V}^{\perp}$.

An orthogonal projection can be used to decompose a signal space into *orthogonal* subspaces. Specifically, given an orthogonal projection $P_{\mathcal{V}}$ on $\mathcal{H}$, we can decompose $\mathcal{H}$ as

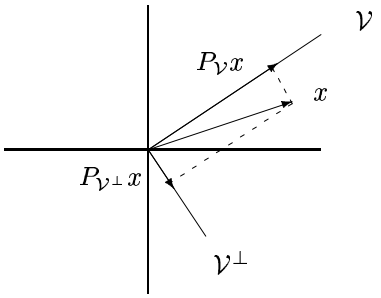


Figure 2-2: Decomposition of x into its orthogonal components in $\mathcal{V}$ and $\mathcal{V}^\perp$ given by $P_{\mathcal{V}}x$ and $P_{\mathcal{V}^\perp}x$, respectively.

$\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp$ so that any $x \in \mathcal{H}$ can be expressed uniquely as $x = x_{\mathcal{V}} + x_{\mathcal{V}^\perp}$, where $x_{\mathcal{V}} = P_{\mathcal{V}}x \in \mathcal{V}$ and $x_{\mathcal{V}^\perp} = (I_{\mathcal{H}} - P_{\mathcal{V}})x = P_{\mathcal{V}^\perp}x \in \mathcal{V}^\perp$, as illustrated in Fig. 2-2. The proof of this well-known result is given in Theorem 2.1 in Section 2.4.2, in the context of more general oblique projections.

The vectors $x_{\mathcal{V}}$ and $x_{\mathcal{V}^\perp}$ are called the projections onto $\mathcal{V}$ and $\mathcal{V}^\perp$ respectively, and have the additional property that $\langle x_{\mathcal{V}}, x_{\mathcal{V}^\perp}\rangle = 0$. Then $\|x\|^2 = \|x_{\mathcal{V}}\|^2 + \|x_{\mathcal{V}^\perp}\|^2$, from which it follows that the norm of the projection is never greater then the norm of the vector:

$$\|P_{\mathcal{V}}x\|^2 = \|x_{\mathcal{V}}\|^2 \leq \|x\|^2. \tag{2.6}$$

This property does not necessarily hold true for an oblique projection onto $\mathcal{V}$ [54].

The orthogonal projection $x_{\mathcal{V}} = P_{\mathcal{V}}x$ has another well-known characterization; it is the closest vector to x in $\mathcal{V}$.

Proposition 2.4. *Let $\mathcal{V} \subseteq \mathcal{H}$ be a closed subspace of $\mathcal{H}$, let $P_{\mathcal{V}}$ denote the orthogonal projection onto $\mathcal{V}$, and let x be an arbitrary vector in $\mathcal{H}$. Then*

$$x_{\mathcal{V}} = \arg\min_{v \in \mathcal{V}} \|x - v\|^2. \tag{2.7}$$

Proof: Expressing x as $x = x_{\mathcal{V}} + x_{\mathcal{V}^\perp}$ where $x_{\mathcal{V}} = P_{\mathcal{V}}x \in \mathcal{V}$ and $x_{\mathcal{V}^\perp} \in \mathcal{V}^\perp$, we have that $x_{\mathcal{V}} - v \in \mathcal{V}$ for any $v \in \mathcal{V}$, so that $\langle x_{\mathcal{V}} - v, x_{\mathcal{V}^\perp}\rangle = 0$ and

$$\|x - v\|^2 = \|x_{\mathcal{V}^\perp} + x_{\mathcal{V}} - v\|^2 = \|x_{\mathcal{V}^\perp}\|^2 + \|x_{\mathcal{V}} - v\|^2 \geq \|x_{\mathcal{V}^\perp}\|^2, \tag{2.8}$$

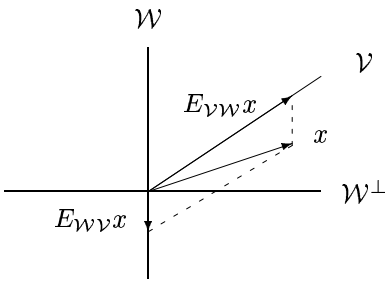


Figure 2-3: Decomposition of x into its components in $\mathcal{V}$ and in $\mathcal{W}$ given by $E_{\mathcal{V}\mathcal{W}}x$ and $E_{\mathcal{W}\mathcal{V}}x$, respectively.

with equality if and only if $x_{\mathcal{V}} = v$. □

An orthogonal projection is a special case of a transjector, defined in Section 2.5. Explicit constructions of orthogonal projections are given in Sections 2.5 and 2.6.2.

2.4.2 Oblique Projection Operators

As with orthogonal projections, oblique projection operators can also be used to decompose a signal space into smaller subspaces; however, when using oblique projections these spaces are no longer constrained to be orthogonal.

Specifically, given a projection $E_{\mathcal{V}\mathcal{W}}$ on $\mathcal{H}$, we can decompose $\mathcal{H}$ as $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$ so that any $x \in \mathcal{H}$ can be expressed uniquely as $x = x_{\mathcal{V}} + x_{\mathcal{W}}$, where $x_{\mathcal{V}} = E_{\mathcal{V}\mathcal{W}}x \in \mathcal{V}$ and $x_{\mathcal{W}} = (I_{\mathcal{H}} - E_{\mathcal{V}\mathcal{W}})x = E_{\mathcal{W}\mathcal{V}}x \in \mathcal{W}$, as illustrated in Fig. 2-3. Note, however, that $x_{\mathcal{V}}$ and $x_{\mathcal{W}}$ are not necessarily orthogonal. This basic property of an oblique projection is incorporated in the following theorem.

Theorem 2.1. *Let $E_{\mathcal{V}\mathcal{W}}$ be a projection on $\mathcal{H}$, with $\mathcal{R}(E) = \mathcal{V}$ and $\mathcal{N}(E) = \mathcal{W}$. Then*

1. *$E_{\mathcal{V}\mathcal{W}}v = v$ for any $v \in \mathcal{V}$;*

2. *$\mathcal{V}$ and $\mathcal{W}$ are both closed subspaces of $\mathcal{H}$;*

3. *$\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$, and any $x \in \mathcal{H}$ can be decomposed uniquely as $x = x_{\mathcal{V}} + x_{\mathcal{W}}$ where $x_{\mathcal{V}} = E_{\mathcal{V}\mathcal{W}}x \in \mathcal{V}$ and is called the projection onto $\mathcal{V}$ along $\mathcal{W}$, and $x_{\mathcal{W}} = (I_{\mathcal{H}} - E_{\mathcal{V}\mathcal{W}})x = E_{\mathcal{W}\mathcal{V}}x \in \mathcal{W}$ and is called the projection onto $\mathcal{W}$ along $\mathcal{V}$.*

Proof: 1. Since $\mathcal{R}(E_{\mathcal{VW}}) = \mathcal{V}$, for any $v \in \mathcal{V}$ there exists an $x \in \mathcal{H}$ such that $v = E_{\mathcal{VW}}x$. Then $E_{\mathcal{VW}}v = E_{\mathcal{VW}}E_{\mathcal{VW}}x = E_{\mathcal{VW}}x = v$.

2. Since $\mathcal{W} = \mathcal{N}(E_{\mathcal{VW}})$ from Proposition 2.3 it is a closed subspace of $\mathcal{H}$. From part (1), $\mathcal{V} = \mathcal{R}(E_{\mathcal{VW}})$ is the space of vectors v satisfying $v = E_{\mathcal{VW}}v$. So, $\mathcal{V} = \mathcal{R}(E_{\mathcal{VW}}) = \mathcal{N}(E_{\mathcal{VW}} - I_{\mathcal{H}})$ and is therefore also a closed subspace of $\mathcal{H}$.

3. If $v \in \mathcal{V}$ then from part (1) $E_{\mathcal{VW}}v = v$, and if $v \in \mathcal{W}$ then $E_{\mathcal{VW}}v = 0$. Thus, if $v \in \mathcal{V} \cap \mathcal{W}$ then $v = 0$ so $\mathcal{V} \cap \mathcal{W} = \{0\}$. Now, any $x \in \mathcal{H}$ can be expressed as $x = E_{\mathcal{VW}}x + (I_{\mathcal{H}} - E_{\mathcal{VW}})x = x_{\mathcal{V}} + x_{\mathcal{W}}$, where $x_{\mathcal{V}} = E_{\mathcal{VW}}x \in \mathcal{V}$ and $x_{\mathcal{W}} = (I_{\mathcal{H}} - E_{\mathcal{VW}})x$. But $E_{\mathcal{VW}}x_{\mathcal{W}} = E_{\mathcal{VW}}(I_{\mathcal{H}} - E_{\mathcal{VW}})x = 0$, so $x_{\mathcal{W}} \in \mathcal{W}$. Since any $x \in \mathcal{H}$ can be expressed as $x = x_{\mathcal{V}} + x_{\mathcal{W}}$ with $x_{\mathcal{V}} \in \mathcal{V}$, $x_{\mathcal{W}} \in \mathcal{W}$, and $\mathcal{V} \cap \mathcal{W} = \{0\}$, $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$.

□

Proposition 2.5 below is a converse to part (3) of Theorem 2.1, and complements Proposition 2.2 by specifying how to determine the components of a signal in a direct sum decomposition.

Proposition 2.5. *Suppose $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$ where $\mathcal{V}$ and $\mathcal{W}$ are closed subspaces so that any $x \in \mathcal{H}$ can be written uniquely as $x = x_{\mathcal{V}} + x_{\mathcal{W}}$ with $x_{\mathcal{V}} \in \mathcal{V}$ and $x_{\mathcal{W}} \in \mathcal{W}$. Then there exists a projection operator $E_{\mathcal{VW}}$ such that $x_{\mathcal{V}} = E_{\mathcal{VW}}x$, and $x_{\mathcal{W}} = (I_{\mathcal{H}} - E_{\mathcal{VW}})x = E_{\mathcal{WV}}x$.*

Proof: Let $x = x_{\mathcal{V}} + x_{\mathcal{W}}$ where $x_{\mathcal{V}} \in \mathcal{V}$ and $x_{\mathcal{W}} \in \mathcal{W}$, and let E be the operator defined by $Ex = x_{\mathcal{V}}$ for all $x \in \mathcal{H}$. Then $EEx = Ex_{\mathcal{V}} = x_{\mathcal{V}} = Ex$ for all $x \in \mathcal{H}$ and E is a projection operator. Since $\mathcal{R}(E) = \mathcal{V}$ and $\mathcal{N}(E) = \mathcal{W}$, $E = E_{\mathcal{VW}}$. Then, $x_{\mathcal{W}} = x - x_{\mathcal{V}} = (I_{\mathcal{H}} - E)x = (I_{\mathcal{H}} - E_{\mathcal{VW}})x = E_{\mathcal{WV}}x$. □

In Section 2.6.2 we develop an explicit construction of an oblique projection operator using set transformations, which are defined in Section 2.6.

In summary, oblique projections and orthogonal projections can both be used to decompose a signal into components in disjoint subspaces as illustrated in Figs. 2-2 and 2-3. However, contrary to decompositions using orthogonal projections, when using oblique projections the components are not necessarily orthogonal. Furthermore, while the norm of the components in an orthogonal projection are no larger than the norm of the original vector, the norm of the components when using an oblique projection can be larger than the norm of the original vector, as can be seen in Fig. 2-3.

2.5 Transjectors

An orthogonal projection is a special case of a transjector (partial isometry) [68, 86], developed in Section 2.5.2. As we will see in Chapter 5, the transjector is useful for studying quantum measurements and combined QSP measurements. To characterize transjectors we rely on the singular value decomposition (SVD), described in Section 2.5.1.

2.5.1 Singular Value Decomposition

It is often useful to decompose a transformation T into elementary transformations that reveal its properties. Such a decomposition is the *singular value decomposition* (SVD) [93]. For simplicity we focus on the case in which $T = \mathbf{T}$ is a matrix; however the results extend to the case in which T is an arbitrary bounded transformation[5] [94].

Proposition 2.6 (Singular Value Decomposition (SVD)). *Let* $\mathbf{T}$ *be an arbitrary* $n \times m$ *matrix with rank* r*. Then*

$$\mathbf{T} = \mathbf{U}\Sigma\mathbf{V}^* = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*,$$

where

1. $\mathbf{T}^*\mathbf{T} = \mathbf{V}(\Sigma^*\Sigma)\mathbf{V}^* = \sum_{i=1}^{r} \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^*$ *is an eigendecomposition of the rank-*r $m \times m$ *matrix* $\mathbf{G} = \mathbf{T}^*\mathbf{T}$*, in which*

   (a) $\{\sigma_i^2, 1 \leq i \leq r\}$ *are the nonzero eigenvalues of* $\mathbf{G}$*, and* $\sigma_i > 0$*;*

   (b) $\{\mathbf{v}_i \in \mathbb{C}^m, 1 \leq i \leq r\}$ *are the corresponding orthonormal eigenvectors;*

   (c) Σ *is a diagonal* $n \times m$ *matrix whose first* r *diagonal elements are* σ_i*, and whose remaining diagonal elements are* 0*;*

   (d) $\mathbf{V}$ *is an* $m \times m$ *unitary matrix whose first* r *columns are the eigenvectors* $\mathbf{v}_i$*, which span the subspace* $\mathcal{V} = \mathcal{N}(\mathbf{T})^\perp \subseteq \mathbb{C}^m$*, and whose remaining* $m - r$ *columns* $\mathbf{v}_i$ *span the orthogonal complement* $\mathcal{V}^\perp \subseteq \mathbb{C}^m$*;*

   and

[5] A transformation T on $\mathcal{H}$ is bounded if $\|Tx\| \leq \alpha\|x\|$ for some $\alpha > 0$ and all $x \in \mathcal{H}$.

2. $\mathbf{T}\mathbf{T}^* = \mathbf{U}(\Sigma\Sigma^*)\mathbf{U}^* = \sum_{i=1}^{r} \sigma_i^2 \mathbf{u}_i\mathbf{u}_i$ *is an eigendecomposition of the rank-r* $n\times n$ *matrix* $\mathbf{S} = \mathbf{T}\mathbf{T}^*$, *in which*

   (a) $\{\sigma_i^2, 1 \le i \le r\}$ *are now identified as the nonzero eigenvalues of* $\mathbf{S}$;

   (b) $\{\mathbf{u}_i \in \mathcal{H}, 1 \le i \le r\}$ *are the corresponding orthonormal eigenvectors;*

   (c) $\mathbf{U}$ *is an* $n \times n$ *unitary matrix whose first* r *columns are the eigenvectors* $\mathbf{u}_i$, *which span the subspace* $\mathcal{U} = \mathcal{R}(\mathbf{T}) \subseteq \mathcal{H}$, *and whose remaining* $n - r$ *columns* $\mathbf{u}_i$ *span the orthogonal complement* $\mathcal{U}^\perp \subseteq \mathcal{H}$.

The matrix $\mathbf{T}$ may be viewed as defining a linear transformation $\mathbf{T} : \mathbb{C}^m \to \mathcal{H}$ according to $\mathbf{v} \mapsto \mathbf{T}\mathbf{v}$. The SVD allows us to interpret this map as follows. A vector $\mathbf{v} \in \mathbb{C}^m$ is first decomposed as $\mathbf{v} = \sum_i \mathbf{v}_i\langle\mathbf{v}_i, \mathbf{v}\rangle$. Since $\mathbf{T}$ maps $\mathbf{v}_i$ to $\sigma_i\mathbf{u}_i$, $\mathbf{T}$ maps the ith component $\mathbf{v}_i\langle\mathbf{v}_i, \mathbf{v}\rangle$ to $\sigma_i\mathbf{u}_i\langle\mathbf{v}_i, \mathbf{v}\rangle$. Therefore, by superposition, $\mathbf{T}$ maps $\mathbf{v}$ to $\sum_i \sigma_i\mathbf{u}_i\langle\mathbf{v}_i, \mathbf{v}\rangle$. Similarly, the conjugate Hermitian matrix $\mathbf{T}^*$ defines the adjoint linear transformation $\mathbf{T}^* : \mathcal{H} \to \mathbb{C}^m$ where $\mathbf{T}^*$ maps $\mathbf{u} \in \mathcal{H}$ to $\sum_i \sigma_i\mathbf{v}_i\langle\mathbf{u}_i, \mathbf{u}\rangle \in \mathbb{C}^m$.

The key element in these maps is the one-dimensional "transjector" (partial isometry) $\mathbf{u}_i\mathbf{v}_i^*$, which "transjects" a basis vector $\mathbf{v}_i \in \mathbb{C}^m$ to the corresponding basis vector $\mathbf{u}_i \in \mathcal{H}$, and the adjoint transjector $\mathbf{v}_i\mathbf{u}_i^*$, which performs the inverse map.

2.5.2 Transjectors (Partial Isometries)

A rank-r $n \times m$ matrix $\mathbf{T}$ is called an *r-dimensional transjector* if its r nonzero singular values are all equal to 1. The special case of a Hermitian operator whose nonzero eigenvalues are all equal to 1 is an orthogonal projector. In other words, $\mathbf{T} = \mathbf{U}\mathbf{Z}_r\mathbf{V}^*$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary and

$$\mathbf{Z}_r = \overbrace{\left[\begin{array}{c|c} \mathbf{I}_r & 0 \\ \hline 0 & 0 \end{array}\right]}^{m}. \tag{2.9}$$

Equivalently, $\mathbf{T}\mathbf{T}^* = \mathbf{U}(\mathbf{Z}_r\mathbf{Z}_r^*)\mathbf{U}^* = P_{\mathcal{U}}$ is an r-dimensional orthogonal projector onto an r-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ with an orthonormal basis $\{\mathbf{u}_i \in \mathcal{H}, 1 \le i \le r\}$ consisting of the first r columns of $\mathbf{U}$, and $\mathbf{T}^*\mathbf{T} = \mathbf{V}(\mathbf{Z}_r^*\mathbf{Z}_r)\mathbf{V}^* = P_{\mathcal{V}}$ is an r-dimensional orthogonal projector onto an r-dimensional subspace $\mathcal{V} \subseteq \mathbb{C}^m$ with an orthonormal basis $\{\mathbf{v}_i \in \mathbb{C}^m, 1 \le i \le r\}$ consisting of the first r columns of V.

An r-dimensional transjector $\mathbf{T}$ is also called a *partial isometry*, because it is an isometry (distance-preserving transformation) between the subspaces $\mathcal{U} \subseteq \mathcal{H}$ and $\mathcal{V} \subseteq \mathbb{C}^m$. Indeed, if $\mathbf{v}, \mathbf{v}' \in \mathcal{V}$ and $\mathbf{u} = \mathbf{T}\mathbf{v}, \mathbf{u}' = \mathbf{T}\mathbf{v}'$, then

$$\langle \mathbf{u}, \mathbf{u}' \rangle = \mathbf{u}^* \mathbf{u}' = \mathbf{v}^* \mathbf{T}^* \mathbf{T} \mathbf{v}' = \mathbf{v}^* P_{\mathcal{V}} \mathbf{v}' = \mathbf{v}^* \mathbf{v}' = \langle \mathbf{v}, \mathbf{v}' \rangle, \tag{2.10}$$

so inner products and *a fortiori* squared norms and distances are preserved. Similarly, if $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, then $\langle \mathbf{T}^* \mathbf{u}, \mathbf{T}^* \mathbf{u}' \rangle = \langle \mathbf{u}, \mathbf{u}' \rangle$. However, inner products are not preserved if $\mathbf{u}, \mathbf{u}' \notin \mathcal{U}$ or $\mathbf{v}, \mathbf{v}' \notin \mathcal{V}$.

The properties of the transjector are summarized in the following theorem [68]:

Theorem 2.2 (Transjectors (partial isometries)). *The following statements are equivalent for a matrix $\mathbf{T}$ whose columns are m vectors in a complex Hilbert space $\mathcal{H}$:*

1. *$\mathbf{T}$ is a transjector between r-dimensional subspaces $\mathcal{U} \subseteq \mathcal{H}$ and $\mathcal{V} \subseteq \mathbb{C}^m$;*
2. *$\mathbf{T}\mathbf{T}^* = P_{\mathcal{U}}$ for an r-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$;*
3. *$\mathbf{T}^*\mathbf{T} = P_{\mathcal{V}}$ for an r-dimensional subspace $\mathcal{V} \subseteq \mathbb{C}^m$.*

A transjector $\mathbf{T}$ between r-dimensional subspaces $\mathcal{U} \subseteq \mathcal{H}$ and $\mathcal{V} \subseteq \mathbb{C}^m$ may be expressed as $\mathbf{T} = \mathbf{U}\mathbf{Z}_r\mathbf{V}^*$, where $\mathbf{U}$ is a unitary matrix whose first r columns $\{\mathbf{u}_i, 1 \leq i \leq r\}$ are an orthonormal basis for $\mathcal{U}$, $\mathbf{V}$ is an $m \times m$ unitary matrix whose first r columns $\{\mathbf{v}_i, 1 \leq i \leq r\}$ are an orthonormal basis for $\mathcal{V}$, and $\mathbf{Z}_r$ is given by (2.9). Equivalently, $\mathbf{T} = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{v}_i^*$.
A transjector $\mathbf{T} : \mathbb{C}^m \to \mathcal{U}$ (resp. $\mathbf{T}^* : \mathcal{H} \to \mathcal{V}$) is an isometry if restricted to $\mathcal{V}$ (resp. $\mathcal{U}$).

2.6 Set Transformations

A useful mathematical tool for describing linear combinations of vectors is the *set transformation*[6], which we now define.

Definition 2.7. *Let $\{x_i, i \in \mathcal{I}\}$ be a set of vectors in a Hilbert space $\mathcal{H}$. The* set transformation *X: $l_2 \to \mathcal{H}$ corresponding to these vectors is defined by $Xa = \sum_{i \in \mathcal{I}} a_i x_i$ for any $a \in l_2$.*

[6] In [95] the set transformation corresponding to the vectors $\{x_i\}$ is referred to as a hypervector and denoted by $(x_1, \ldots, x_i, \ldots)$. We prefer the more compact transformation notation.

From the definition of the adjoint X^*: $\mathcal{H} \to l_2$, if $a = X^* y$, then

$$a_i = \langle x_i, y \rangle. \tag{2.11}$$

A set transformation can be regarded as a possibly infinite-dimensional matrix.

A set transformation corresponding to a set of orthonormal vectors is referred to as an *orthonormal set transformation.* We summarize its key properties in the following theorem.

Theorem 2.3 (Orthonormal set transformation). *Let $\{x_i, i \in \mathcal{I}\}$ denote a set of orthonormal vectors in a Hilbert space $\mathcal{H}$. The set transformation X corresponding to these vectors satisfies $X^*X = I_{\mathcal{H}}$. If in addition the vectors $\{x_i, i \in \mathcal{I}\}$ span $\mathcal{H}$, then $XX^* = I_{\mathcal{H}}$. Conversely, if a set transformation X corresponding to a set of vectors $\{x_i \in \mathcal{H}, i \in \mathcal{I}\}$ satisfies $X^*X = I_{\mathcal{H}}$, then the vectors $\{x_i, i \in \mathcal{I}\}$ are orthonormal. If $XX^* = I_{\mathcal{H}}$, then the vectors $\{x_i, i \in \mathcal{I}\}$ span $\mathcal{H}$.*

Proof: Suppose the vectors $\{x_i, i \in \mathcal{I}\}$ are orthonormal, and let $b = X^*Xa$ with $a \in l_2$. Then,

$$b_k = \langle x_k, \sum_{i \in \mathcal{I}} a_i x_i \rangle = \sum_{i \in \mathcal{I}} a_i \langle x_k, x_i \rangle = \sum_{i \in \mathcal{I}} a_i \delta_{ki} = a_k, \tag{2.12}$$

for all k so that $X^*Xa = a$ for any $a \in l_2$ and $X^*X = I_{\mathcal{H}}$. If in addition the vectors $\{x_i\}$ span $\mathcal{H}$, then any $x \in \mathcal{H}$ can be expressed as $x = \sum_{i \in \mathcal{I}} a_i x_i = Xa$ for some a so that

$$XX^*x = X(X^*X)a = Xa = x, \tag{2.13}$$

and $XX^* = I_{\mathcal{H}}$. Next, suppose $X^*X = I_{\mathcal{H}}$ and let $e^k \in l_2$ denote the sequence with ith element $e_i^k = \delta_{ki}$. Then, $e^k = X^*Xe^k = X^*x_k$, so that $\langle x_i, x_k \rangle = e_i^k = \delta_{ki}$, and the vectors $\{x_i\}$ are orthonormal. Finally if $XX^* = I_{\mathcal{H}}$, then for any $x \in \mathcal{H}$

$$x = XX^*x = \sum_{i \in \mathcal{I}} \langle x_i, x \rangle x_i, \tag{2.14}$$

and the vectors $\{x_i\}$ span $\mathcal{H}$. □

In the next section we use set transformations to determine the coefficients in a basis expansion of a signal. In Section 2.6.2 we use set transformations to develop an explicit

construction of oblique projections.

2.6.1 Basis Expansions

Let $\{x_i, i \in \mathcal{I}\}$ denote a Riesz basis for $\mathcal{H}$. Then we have seen in Section 2.6.1 that any $x \in \mathcal{H}$ can be expressed uniquely as

$$x = \sum_{i \in \mathcal{I}} a_i x_i, \tag{2.15}$$

where $a_i \in \mathbb{C}$. With X denoting the set transformation corresponding to the vectors $\{x_i\}$ we may write (2.15) as $x = Xa$.

If the vectors $\{x_i\}$ are orthonormal then from (2.14), $a_i = \langle x_i, x \rangle$. To determine the coefficients a_i when the basis vectors are not orthonormal we introduce the dual basis of $\{x_i\}$, denoted $\{\tilde{x}_i\}$, defined by

$$\langle \tilde{x}_i, x_k \rangle = \delta_{ik} \tag{2.16}$$

for all i and k. The motivation behind this definition is that if we find a set of vectors satisfying (2.16), then we can take the inner product of x given by (2.15) with $\tilde{x}_k$ to obtain

$$\langle \tilde{x}_k, x \rangle = \sum_{i \in \mathcal{I}} a_i \langle \tilde{x}_k, x_i \rangle = a_k, \tag{2.17}$$

so that any $x \in \mathcal{H}$ can be expressed as

$$x = \sum_{i \in \mathcal{I}} \langle \tilde{x}_i, x \rangle x_i. \tag{2.18}$$

It is well known that every Riesz basis $\{x_i, i \in \mathcal{I}\}$ for a Hilbert space possesses a unique dual basis $\{\tilde{x}_i, i \in \mathcal{I}\}$, which is also a Riesz basis [64]. From (2.3) it then follows that the expansion coefficients $a_i = \langle \tilde{x}_i, x \rangle$ are in l_2.

To explicitly compute these coefficients we need to determine the dual basis $\{\tilde{x}_i\}$ of $\{x_i\}$. Let X and $\widetilde{X}$ denote the set transformations corresponding to the Riesz bases $\{x_i\}$ and $\{\tilde{x}_i\}$, respectively. Then in a similar manner to the proof of Theorem 2.3 it can be

shown that (2.16) is equivalent to

$$\widetilde{X}^* X = I_{\mathcal{H}}. \tag{2.19}$$

To solve for $\widetilde{X}$ we first show that X^*X is invertible, and then express $\widetilde{X}$ in terms of this inverse.

Since the vectors $\{x_i\}$ form a Riesz basis, from (2.2), $\alpha\|a\|^2 \leq \|Xa\|^2 \leq \beta\|a\|^2$, for any $a \in l_2$. But, $\|Xa\|^2 = \langle Xa, Xa\rangle = \langle X^*Xa, a\rangle$ so that

$$\alpha I_{\mathcal{H}} \leq X^*X \leq \beta I_{\mathcal{H}}, \tag{2.20}$$

which implies that X^*X is invertible [69, Lemma 3.2.2]. We may then readily verify that

$$\widetilde{X} = X(X^*X)^{-1} \tag{2.21}$$

is a solution to (2.19): $\widetilde{X}^*X = (X^*X)^{-1}X^*X = I_{\mathcal{H}}$.

Note that if the vectors x_i are orthonormal, then from Theorem 2.3 and (2.21) it follows that $\tilde{x}_i = x_i$ for all i, so that (2.18) reduces to $x = \sum_{i\in\mathcal{I}}\langle x_i, x\rangle x_i$, which agrees with (2.14).

This discussion is summarized in the following theorem:

Theorem 2.4 (Basis expansion). *Let the vectors $\{x_i, i \in \mathcal{I}\}$ form a Riesz basis for a Hilbert space $\mathcal{H}$ and let X be the set transformation corresponding to the vectors $\{x_i, i \in \mathcal{I}\}$. Let the vectors $\{\tilde{x}_i, i \in \mathcal{I}\}$ be the unique dual Riesz basis of $\{x_i, i \in \mathcal{I}\}$ such that $\langle \tilde{x}_k, x_i\rangle = \delta_{ki}$ for all k and i, and let $\widetilde{X}$ be the set transformation corresponding to the vectors $\{\tilde{x}_i, i \in \mathcal{I}\}$. Then*

$$\widetilde{X} = X(X^*X)^{-1}.$$

Any $x \in \mathcal{H}$ can then be expressed uniquely as $x = \sum_{i\in\mathcal{I}}\langle \tilde{x}_i, x\rangle x_i$ where the sequence of coefficients $\langle \tilde{x}_i, x\rangle$ is in l_2.

2.6.2 Construction of Projection Operators

Using set transformations we now explicitly construct a projection $E_{\mathcal{V}\mathcal{W}}$ in terms of arbitrary Riesz bases for $\mathcal{V}$ and $\mathcal{W}^\perp$. This new construction will be used in Chapter 6 to develop

consistent reconstruction algorithms with arbitrary sampling and reconstruction spaces. Explicit formulas for finite-dimensional oblique projections have been given in [54, 91], but are different then the formula presented in Theorem 2.5. We suspect the formula in Theorem 2.5 to be known in finite-dimensions; however, our construction includes infinite-dimensional projections as well. Formulas for oblique projections in atomic spaces [92] and in shift invariant spaces [16] have also been previously considered.

To construct an oblique projection $E_{\mathcal{V}\mathcal{W}}$, we first prove the following lemma[7].

Lemma 2.1. *Let the vectors $\{v_i, i \in \mathcal{I}\}$ form a Riesz basis for a subspace $\mathcal{V}$ of a Hilbert space $\mathcal{H}$, and let the vectors $\{w_i, i \in \mathcal{I}\}$ form a Riesz basis for a subspace $\mathcal{W}^\perp$ of $\mathcal{H}$ such that $\mathcal{V} + \mathcal{W} = \mathcal{H}$. Let $V\colon\ l_2 \to \mathcal{H}$ and $W\colon\ l_2 \to \mathcal{H}$ denote the set transformations corresponding to the vectors $\{v_i, i \in \mathcal{I}\}$ and $\{w_i, i \in \mathcal{I}\}$, respectively. Then W^*V is invertible if and only if $\mathcal{V} \cap \mathcal{W} = \{0\}$.*

Proof: Suppose that x is a nonzero vector in $\mathcal{V} \cap \mathcal{W}$. Then since the vectors v_i form a basis for $\mathcal{V}$, $x = Va$ for some nonzero $a \in l_2$. But since $x \in \mathcal{W}$ we have that $W^*x = 0$. Thus $W^*x = W^*Va = 0$ for a nonzero $a \in l_2$, so that W^*V is not invertible.

Now, suppose that $\mathcal{V} \cap \mathcal{W} = \{0\}$. Since the vectors $\{v_i\}$ from a Riesz basis, V is a bijection from l_2 to $\mathcal{V}$. We now show that W^* is bijective on $\mathcal{V}$.

Suppose that $W^*v = 0$ for some $v \in \mathcal{V}$. With $v_1 = P_{\mathcal{W}}v$ and $v_2 = P_{\mathcal{W}^\perp}v$, we have that $0 = W^*v = W^*v_1 + W^*v_2 = W^*v_2$. Thus $\sum_i |\langle w_i, v_2\rangle|^2 = 0$ where $v_2 \in \mathcal{W}^\perp$ which implies from (2.3) that $v_2 = 0$. But then $v = v_1$ where $v \in \mathcal{V}$ and $v_1 \in \mathcal{W}$ which implies that $v = 0$ since $\mathcal{V} \cap \mathcal{W} = \{0\}$, so W^* is injective on $\mathcal{V}$.

Now, let $\{\tilde{w}_i\}$ denote the dual Riesz basis of $\{w_i\}$, and let $w = \sum_i a_i \tilde{w}_i$ for some $a \in l_2$. Decompose w as $w = w_1 + w_2$ with $w_1 \in \mathcal{V}$ and $w_2 \in \mathcal{W}$. Then we have that $\langle w_k, w\rangle = a_k$. But $\langle w_k, w\rangle = \langle w_k, w_1\rangle + \langle w_k, w_2\rangle = \langle w_k, w_1\rangle$. We therefore conclude that $a = W^*w_1$ where $w_1 \in \mathcal{V}$, so that W^* is surjective on $\mathcal{V}$.

Since V is bijective, W^*V is a bijection from l_2 to l_2 and consequently invertible. □

Using Lemma 2.1 we immediately deduce our main result:

[7] This proof, due to A. Aldroubi [96], is a more elegant version of my proof of this result as originally constructed.

Theorem 2.5. *Let the vectors $\{v_i, i \in \mathcal{I}\}$ form a Riesz basis for a subspace $\mathcal{V}$ of a Hilbert space $\mathcal{H}$, and let the vectors $\{w_i, i \in \mathcal{I}\}$ form a Riesz basis for a subspace $\mathcal{W}^\perp$ of $\mathcal{H}$ such that $\mathcal{V} \oplus \mathcal{W} = \mathcal{H}$. Let $V\colon\ l_2 \to \mathcal{H}$ and $W\colon\ l_2 \to \mathcal{H}$ denote the set transformations corresponding to the vectors $\{v_i, i \in \mathcal{I}\}$ and $\{w_i, i \in \mathcal{I}\}$, respectively. Then $E_{\mathcal{V}\mathcal{W}} = V(W^*V)^{-1}W^*$.*

Proof: Let $T = V(W^*V)^{-1}W^*$. From Lemma 2.1, W^*V is invertible so that T is well defined. Since $w_i \in \mathcal{W}^\perp$, $W^*w = 0$ for any $w \in \mathcal{W}$, and $Tw = 0$. Furthermore, any $v \in \mathcal{V}$ can be expressed as $v = Va$ for some $a \in l_2$, so that $Tv = TVa = Va = v$. Thus, $T = E_{\mathcal{V}\mathcal{W}}$. □

Using Theorem 2.5 we can construct an orthogonal projection $P_\mathcal{V}$ onto $\mathcal{V}$ using any Riesz basis $\{v_i, i \in \mathcal{I}\}$ for $\mathcal{V}$. With V denoting the set transformation corresponding to the vectors v_i, $P_\mathcal{V} = V(V^*V)^{-1}V^*$. If the vectors $\{v_i\}$ are orthonormal, then from Theorem 2.3 $V^*V = I_\mathcal{H}$ and

$$P_\mathcal{V} = VV^* = \sum_{i \in \mathcal{I}} v_i v_i^*. \tag{2.22}$$

2.7 Pseudoinverse of a Transformation

A linear transformation between spaces of different dimension (*e.g.,* a rectangular matrix), or an operator that is not bijective (*e.g.,* a singular matrix) does not have an inverse in the usual sense. Nevertheless in Section 2.7.1 we define the pseudoinverse which has properties closely related to those of an inverse. Using oblique projections, this concept is extended in Section 2.7.2 to a class of pseudoinverses called oblique pseudoinverses.

In this section we obtain alternative characterizations of the pseudoinverse and the oblique pseudoinverse that are important in applications in this thesis.

2.7.1 Pseudoinverse

Let T be a continuous linear mapping $T\colon\ \mathcal{H} \to \mathcal{S}$. The *pseudoinverse* of T, denoted $T^\dagger$, is defined as the unique mapping that satisfies the *Moore-Penrose conditions* [93]:

$$TT^\dagger T = T; \tag{2.23}$$

$$T^\dagger TT^\dagger = T^\dagger; \tag{2.24}$$

$$(TT^\dagger)^* = TT^\dagger; \tag{2.25}$$

$$(T^\dagger T)^* = T^\dagger T. \tag{2.26}$$

These conditions are equivalent to a more intuitive set of conditions given in Theorem 2.6 below. Essentially this theorem states that if we restrict T to the subspace $\mathcal{N}(T)^\perp \subseteq \mathcal{H}$ and $T^\dagger$ to the subspace $\mathcal{R}(T)^c \subseteq \mathcal{S}$, then T is invertible and its inverse is $T^\dagger$, as illustrated in Fig. 2-4.

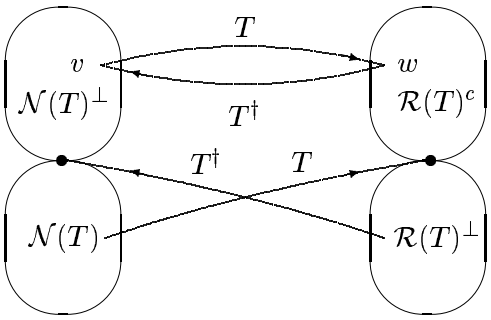


Figure 2-4: The action of T and $T^\dagger$ on the subspaces $\mathcal{N}(T)^\perp, \mathcal{N}(T), \mathcal{R}(T)^c$ and $\mathcal{R}(T)^\perp$.

Theorem 2.6 (Pseudoinverse). *Let T be a continuous linear mapping $T\colon\ \mathcal{H} \to \mathcal{S}$. The* pseudoinverse *of T, denoted $T^\dagger$, is the unique mapping that satisfies the conditions:*

$$TT^\dagger = P_{\mathcal{R}(T)^c}; \tag{2.27}$$

$$T^\dagger T = P_{\mathcal{R}(T^*)^c}; \tag{2.28}$$

$$\mathcal{R}(T^\dagger) = \mathcal{R}(T^*). \tag{2.29}$$

We remark that while it is known that the Moore-Penrose conditions imply (2.27)–(2.29),

the converse does not appear to be established.

Proof: We first prove that any $T^\dagger$ satisfying (2.27)–(2.29) satisfies the Moore-Penrose conditions. Since an orthogonal projection operator is Hermitian, (2.25) and (2.26) are immediately satisfied. Next, for any $y \in \mathcal{H}$, $Ty \in \mathcal{R}(T)$ and $TT^\dagger Ty = P_{\mathcal{R}(T)^c} Ty = Ty$ for all $y \in \mathcal{H}$, establishing (2.23). Finally from (2.29), $T^\dagger x \in \mathcal{R}(T^*)^c$ for any $x \in \mathcal{S}$, and $T^\dagger TT^\dagger x = P_{\mathcal{R}(T^*)^c} T^\dagger x = T^\dagger x$ for all $x \in \mathcal{S}$, establishing (2.24).

We now show that any $T^\dagger$ satisfying the Moore-Penrose conditions satisfies (2.27)–(2.29). From (2.25) and (2.26), $TT^\dagger$ and $T^\dagger T$ are Hermitian. Furthermore, $TT^\dagger TT^\dagger = T(T^\dagger TT^\dagger) = TT^\dagger$ and $T^\dagger TT^\dagger T = T^\dagger(TT^\dagger T) = T^\dagger T$, so that $TT^\dagger$ and $T^\dagger T$ are both orthogonal projections. It remains to determine $\mathcal{R}(TT^\dagger)$ and $\mathcal{R}(T^\dagger T)$.

First we show that $\mathcal{R}(TT^\dagger) = \mathcal{R}(T)^c$. If $x \in \mathcal{R}(T)$, then $x = Ty = TT^\dagger(Ty)$ for some $y \in \mathcal{H}$ so that $x \in \mathcal{R}(TT^\dagger)$. If x is a limit point of $\mathcal{R}(T)$, then there exists a sequence $x_n \in \mathcal{R}(T)$ such that $x = \lim x_n$. Then $x = \lim x_n = \lim Ty_n = \lim TT^\dagger(Ty_n)$ for some $y_n \in \mathcal{H}$, and $x \in \mathcal{R}(TT^\dagger)^c = \mathcal{R}(TT^\dagger)$, because the range space of a projection operator is closed[8]. If $x \in \mathcal{R}(TT^\dagger)$, then $x = T(T^\dagger y)$ for some $y \in \mathcal{S}$ and $x \in \mathcal{R}(T)$.

We now show that $\mathcal{R}(T^\dagger T) = \mathcal{R}(T^*)^c$. If $x \in \mathcal{R}(T^*)$, then $x = T^*y = T^*(T^\dagger)^* T^* y = (T^\dagger T)^* T^* y = T^\dagger T(T^* y)$ for some $y \in \mathcal{S}$, and $x \in \mathcal{R}(T^\dagger T)$. If x is a limit point of $\mathcal{R}(T^*)$, then $x = \lim T^* y_n = \lim T^\dagger T(T^* y_n)$ for some $y_n \in \mathcal{S}$, so that $x \in \mathcal{R}(T^\dagger T)^c = \mathcal{R}(T^\dagger T)$. If $x \in \mathcal{R}(T^\dagger T)$ and $x \neq 0$, then $x = T^\dagger Ty$ for some $y \in \mathcal{H}$ and $Tx = TT^\dagger Ty = Ty \neq 0$, since if $Ty = 0$ then $x = 0$. Thus $x \in \mathcal{N}(T)^\perp = \mathcal{R}(T^*)^c$.

Finally, we show that $\mathcal{R}(T^\dagger) = \mathcal{R}(T^*)$. If $y \in \mathcal{R}(T^\dagger)$, then $y = T^\dagger x = T^\dagger TT^\dagger x = (T^\dagger T)^* T^\dagger x = T^*(T^\dagger)^* T^\dagger x = T^* u$ where $u = (T^\dagger)^* T^\dagger x$, so that $y \in \mathcal{R}(T^*)$. If $y \in \mathcal{R}(T^*)$, then $y = T^* x = (TT^\dagger T)^* x = (T^\dagger T) T^* x = T^\dagger u$ where $u = TT^* x$, and $y \in \mathcal{R}(T^\dagger)$. □

In summary, if we apply T to a $y \in \mathcal{N}(T)^\perp$, then we can invert this mapping by applying $T^\dagger$ to the result: $T^\dagger Ty = y$. Similarly, if we apply $T^\dagger$ to an $x \in \mathcal{R}(T)^c$, then we can invert this mapping by applying T to the result: $TT^\dagger x = x$.

Since $TT^\dagger = (TT^\dagger)^* = (T^\dagger)^* T^*$ and $T^\dagger T = (T^\dagger T)^* = T^*(T^\dagger)^*$, for any $x \in \mathcal{R}(T)^c = \mathcal{N}(T^*)^\perp$, $(T^\dagger)^* T^* x = x$ and for any $y \in \mathcal{R}(T^*)^c = \mathcal{N}(T)^\perp$, $T^*(T^\dagger)^* y = y$. If we apply T^* to a $x \in \mathcal{N}(T^*)^\perp$, then we can invert this mapping by applying $(T^\dagger)^*$ to the result:

[8]For any projection E on $\mathcal{H}$, $\mathcal{R}(E) = \mathcal{N}(I_\mathcal{H} - E)$ which from Proposition 2.3 is a closed subspace of $\mathcal{H}$.

$(T^\dagger)^* T^* x = x$. Similarly, if we apply $(T^\dagger)^*$ to a $y \in \mathcal{N}(T)^\perp$, then we can invert this mapping by applying T^* to the result: $T^*(T^\dagger)^* y = y$.

We note that if T^*T is invertible, then $T^\dagger = (T^*T)^{-1}T^*$. Similarly if TT^* is invertible, then $T^\dagger = T^*(TT^*)^{-1}$.

2.7.2 Oblique Pseudoinverse

The oblique pseudoinverse [97] of a matrix is not very well known in the signal processing literature. In [98] the oblique pseudoinverse is used in a solution to a constrained least-squares problem, introduced in [54]. In Chapter 5 we show that oblique pseudoinverse can be used to generalize frame expansions, and in Chapter 6 we derive a redundant sampling scheme with arbitrary sampling and reconstruction spaces, based on oblique pseudoinverses.

Let $T\colon\ \mathcal{H} \to \mathcal{S}$ be an arbitrary linear transformation, and let $\mathcal{H} = \mathcal{G} \oplus \mathcal{N}(T)$ and $\mathcal{S} = \mathcal{R}(T)^c \oplus \mathcal{Z}$. The *oblique pseudoinverse* of T on $\mathcal{G}$ along $\mathcal{Z}$, denoted $T^\#_{\mathcal{G}\mathcal{Z}}$, is the unique transformation satisfying the Milne conditions [97]:

$$T^\#_{\mathcal{G}\mathcal{Z}} T v = v \text{ for all } v \in \mathcal{G}; \tag{2.30}$$

$$T^\#_{\mathcal{G}\mathcal{Z}} w = 0 \text{ for all } w \in \mathcal{Z}. \tag{2.31}$$

These conditions are equivalent to a more intuitive set of conditions given in Theorem 2.7 below. Essentially this theorem states that the oblique pseudoinverse of T on $\mathcal{G}$ along $\mathcal{Z}$ inverts T between $\mathcal{G}$ and $\mathcal{R}(T)$, while nulling out any vector in $\mathcal{Z}$, as illustrated in Fig. 2-5. From Fig. 2-4 we see that the pseudoinverse $T^\dagger$ of T is a special case of the oblique pseudoinverse $T^\#_{\mathcal{G}\mathcal{Z}}$ for which $\mathcal{G} = \mathcal{N}(T)^\perp$ and $\mathcal{Z} = \mathcal{R}(T)^\perp$.

Theorem 2.7 (Oblique pseudoinverse). *Let T be a continuous linear mapping $T\colon\ \mathcal{H} \to \mathcal{S}$. The* oblique pseudoinverse *of T on $\mathcal{G}$ along $\mathcal{Z}$, denoted $T^\#_{\mathcal{G}\mathcal{Z}}$, is the unique mapping that satisfies the conditions:*

$$T T^\#_{\mathcal{G}\mathcal{Z}} = E_{\mathcal{R}(T)^c \mathcal{Z}}; \tag{2.32}$$

$$T^\#_{\mathcal{G}\mathcal{Z}} T = E_{\mathcal{G}\mathcal{N}(T)}; \tag{2.33}$$

$$\mathcal{R}(T^\#_{\mathcal{G}\mathcal{Z}}) = \mathcal{G}. \tag{2.34}$$

While it is known that the oblique pseudoinverse satisfies (2.32)–(2.34), the converse does

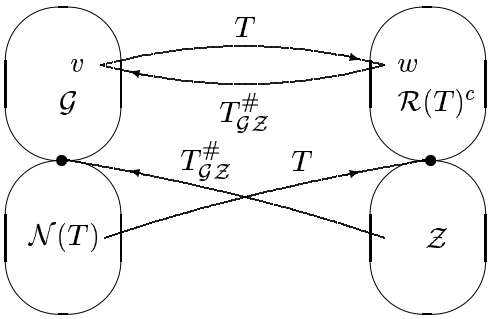


Figure 2-5: The action of T and $T^{\#}_{\mathcal{GZ}}$ on the subspaces $\mathcal{G}, \mathcal{N}(T), \mathcal{R}(T)^c$ and $\mathcal{Z}$. In the special case in which $\mathcal{G} = \mathcal{N}(T)^{\perp}$ and $\mathcal{Z} = \mathcal{R}(T)^{\perp}$, $T^{\#}_{\mathcal{GZ}}$ reduces to the pseudoinverse $T^{\dagger}$.

not appear to be established.

Proof: We first prove that any $T^{\#}_{\mathcal{GZ}}$ satisfying (2.32)–(2.34) satisfies the Milne conditions. From (2.32) $TT^{\#}_{\mathcal{GZ}}w = 0$ for any $w \in \mathcal{Z}$, so that $T^{\#}_{\mathcal{GZ}}w \in \mathcal{N}(T)$. But from (2.34), $\mathcal{R}(T^{\#}_{\mathcal{GZ}}) = \mathcal{G}$ so $T^{\#}_{\mathcal{GZ}}w \in \mathcal{G}$. Since $\mathcal{N}(T)$ and $\mathcal{G}$ are disjoint, $T^{\#}_{\mathcal{GZ}}w = 0$, establishing (2.31). From (2.33), $T^{\#}_{\mathcal{GZ}}Tv = v$ for any $v \in \mathcal{G}$, establishing (2.30).

We now show that any $T^{\#}_{\mathcal{GZ}}$ satisfying the Milne conditions satisfies (2.32)–(2.34). From (2.31), $TT^{\#}_{\mathcal{GZ}}w = 0$ for all $w \in \mathcal{Z}$. If $w \in \mathcal{R}(T)$, then $w = Tv$ for some $v \in \mathcal{G}$ and using (2.30), $TT^{\#}_{\mathcal{GZ}}w = TT^{\#}_{\mathcal{GZ}}Tv = Tv = w$. A similar argument can be used to show that if $w \in \mathcal{R}(T)^c$ then $TT^{\#}_{\mathcal{GZ}}w = w$, establishing (2.32). Combining (2.30) with $T^{\#}_{\mathcal{GZ}}Tv = 0$ for any $v \in \mathcal{N}(T)$ establishes (2.33). Finally, suppose that $v \in \mathcal{R}(T^{\#}_{\mathcal{GZ}})$. Then $v = T^{\#}_{\mathcal{GZ}}w$ for some $w \in \mathcal{R}(T)^c$. We may therefore express w as $w = \lim Ty_n$ for some sequence $y_n \in \mathcal{G}$. Thus $v = \lim T^{\#}_{\mathcal{GZ}}Ty_n$, and from (2.30) we conclude that $v = \lim y_n$ so $v \in \mathcal{G}^c = \mathcal{G}$. Also from (2.30), any $v \in \mathcal{G}$ is in $\mathcal{R}(T^{\#}_{\mathcal{GZ}})$ so that $\mathcal{R}(T^{\#}_{\mathcal{GZ}}) = \mathcal{G}$. □

In summary, if we apply T to a $v \in \mathcal{G}$, then we can invert this mapping by applying $T^{\#}_{\mathcal{GZ}}$ to the result: $T^{\#}_{\mathcal{GZ}}Tv = v$. Similarly, if we apply $T^{\#}_{\mathcal{GZ}}$ to a $w \in \mathcal{R}(T)^c$, then we can invert this mapping by applying T to the result: $TT^{\#}_{\mathcal{GZ}}w = w$. Thus T and $T^{\#}_{\mathcal{GZ}}$ are mutual inverses on $\mathcal{G}$ and $\mathcal{R}(T)^c$, respectively.

Chapter 3

Quantum States and Measurement

In this chapter we present some elements of the theory of quantum mechanics. Rather than attempting to provide a comprehensive survey, we concentrate on portions of the theory that we will draw upon in the development of the QSP framework. In addition to establishing notation and summarizing the important results, this overview provides a particular perspective on quantum mechanics that is central to the development of the QSP measurement. This presentation is based on work with G. D. Forney, Jr. [68, 86], and on the book by D. G. Griffiths [20]. Excellent introductions to quantum mechanics can be found in [99, 100, 101].

In both quantum mechanics and in QSP, the setting we consider is an arbitrary Hilbert space $\mathcal{H}$, whose elements are referred to as vectors (or signals).

The state of a closed quantum system is characterized by a normalized (unit-norm) vector $\phi \in \mathcal{H}$, and is referred to as a pure state. Information about a quantum system is extracted by subjecting the system to a measurement. In quantum theory, the outcome of a measurement is inherently probabilistic, with the probabilities of the outcomes of any conceivable measurement determined by the state vector $\phi \in \mathcal{H}$.

3.1 Standard Measurements

A standard (von Neumann) measurement in quantum mechanics is defined by a collection of projection operators $\{P_i, i \in \mathcal{I}\}$ onto subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$, where $\mathcal{I}$ denotes an index set and the index $i \in \mathcal{I}$ corresponds to a possible measurement outcome. The operators P_i and the subspaces $\mathcal{S}_i$ are referred to as the measurement operators and the measurement

subspaces, respectively. The laws of quantum mechanics impose the constraint that the operators $\{P_i, i \in \mathcal{I}\}$ form a complete set of orthogonal projections so that for any $i, k \in \mathcal{I}$,

$$\begin{aligned} P_i &= P_i^*; && (3.1)\\ P_i^2 &= P_i; && (3.2)\\ P_i P_k &= 0, \quad \text{if } i \neq k; && (3.3)\\ \sum_{i \in \mathcal{I}} P_i &= I_{\mathcal{H}}. && (3.4) \end{aligned}$$

Conditions (3.3) and (3.4) imply that the measurement subspaces $\mathcal{S}_i$ are orthogonal, and that their direct sum is equal to $\mathcal{H}$.

If the state vector is ϕ, then from the rules of quantum mechanics, the probability of observing the ith outcome is

$$p(i) = \langle P_i \phi, \phi \rangle. \tag{3.5}$$

Since the state is normalized, $\sum_i p(i) = \langle \sum_i P_i \phi, \phi \rangle = \langle \phi, \phi \rangle = 1$.

We note that from (3.5) it follows that any state of the form $e^{j\theta}\phi$ where θ is an arbitrary angle, leads to the same probabilities on the output as the state ϕ. Therefore these states are considered to be equivalent[1].

In the simplest case the projection operators are rank-one operators and have the outer-product form $P_i = \mu_i \mu_i^*$ for some nonzero vectors $\{\mu_i \in \mathcal{H}, i \in \mathcal{I}\}$. We refer to such measurements as *rank-one quantum measurements*. Then (3.3) implies that $\langle \mu_i, \mu_k \rangle = \delta_{ik}$, while (3.4) implies that

$$x = I_{\mathcal{H}} x = \sum_{i \in \mathcal{I}} \langle \mu_i, x \rangle \mu_i, \quad \forall x \in \mathcal{H}, \tag{3.6}$$

so the measurement vectors $\{\mu_i, i \in \mathcal{I}\}$ form an orthonormal basis for $\mathcal{H}$. If the state vector is ϕ, then the probability of observing the ith outcome is

$$p(i) = |\langle \mu_i, \phi \rangle|^2. \tag{3.7}$$

[1] A state is therefore defined as a *ray* in a Hilbert space, where a ray is an equivalence class of vectors that differ by multiplication by a nonzero complex scalar. We may then always choose a unit-norm representative of this class.

Due to the probabilistic nature of a quantum measurement, in general identical measurements on identically prepared systems do not produce the same outcome. However, associated with every measurement are particular preparation states that are determinate so that if the system is prepared in one of these states then the measurement will yield the same outcome with probability one (w.p. 1). For a rank-one quantum measurement defined by orthonormal measurement vectors μ_i, it follows from (3.7) that if $\phi = \mu_i$ for some i, then $p(i) = 1$ and output i is obtained w.p. 1. The states $\{\phi = \mu_i\}$ are therefore called the *determinate states* of the measurement. More generally, the determinate states are the states that lie completely in one of the measurement spaces $\mathcal{S}_i$. Indeed, if $\phi \in \mathcal{S}_i$, then $P_i\phi = \phi$ and from (3.5), $p(i) = 1$.

A fundamental postulate of quantum mechanics is that repeated measurements on a quantum system yield the same outcome; if the outcome cannot be confirmed by immediate repetition of the measurement, then we cannot prove that the output was actually observed. Evidently, the measurement alters the state of the system. In quantum terminology, the state is said to collapse onto a state consistent with the measurement outcome so that if we re-measure the system in this state, then the final state after this second measurement will be identical to the state after the first measurement. Suppose that we measure a system in a state ϕ, using a rank-one measurement with orthonormal measurement vectors μ_i, and let ϕ' denote the state of the system after the measurement. If the kth output is observed and a second measurement is performed, then we must have that $p(k) = |\langle \mu_k, \phi' \rangle|^2 = 1$, which implies that (up to a possible phase factor) $\phi' = \mu_k$. In the more general case in which the measurement corresponds to orthogonal projections P_i onto subspaces $\mathcal{S}_i$, if the kth output is observed, then ϕ' is a normalized vector in the direction of $P_k\phi$.

Thus, the state after a measurement is quantized to one of the determinate states. In a rank-one measurement, the probability of being in any particular determinate state is a function of the inner product between the state of the system and the determinate state. More generally, the probability is a function of the norm of the projection of the state onto the corresponding measurement space.

The discussion above leads to a particular viewpoint towards quantum measurement which can be formulated in terms of a probabilistic mapping between $\mathcal{H}$ and the determinate states. Given a state space $\mathcal{X}$ and an observation space $\mathcal{Y}$, a *probabilistic mapping* from $\mathcal{X}$ to $\mathcal{Y}$ is a function $f\colon\ \mathcal{X} \times \mathcal{W} \to \mathcal{Y}$, where $\mathcal{W}$ is the sample space of an auxiliary chance

variable $W(x) = \{\mathcal{W}, p_{W|X}(w|x)\}$ with a probability distribution $p_{W|X}(w|x)$ on $\mathcal{W}$ that in general depends on $x \in \mathcal{X}$. Note that a deterministic mapping $f\colon\ \mathcal{X} \to \mathcal{Y}$ is a special case of a probabilistic mapping in which the auxiliary chance variable is deterministic; *i.e.,* has one outcome w w.p. 1. In this case the the function f is independent of W.

A rank-one quantum measurement corresponding to orthonormal measurement vectors $\{\mu_i \in \mathcal{H}, i \in \mathcal{I}\}$ that span subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$ can be viewed as a probabilistic mapping between $\mathcal{H}$ and the determinate states that is

1. a deterministic identity mapping for $\phi \in \mathcal{S}_i$;

2. a probabilistic mapping for nondeterminate states that maps ϕ to a normalized vector in the direction of the orthogonal projection $P_i\phi$ for some value $i \in \mathcal{I}$, where $i = f(\{\langle \mu_k, \phi\rangle, k \in \mathcal{I}\}, w_i)$.

Here $f\colon\ \mathcal{H} \times \mathcal{W} \to \mathcal{I}$ is a probabilistic mapping between elements ϕ of $\mathcal{H}$ and indices $i \in \mathcal{I}$, that depends on a chance variable W with a discrete alphabet $\mathcal{W} = \mathcal{I}$ such that the probability of outcome $w_i \in \mathcal{W}$ depends on the input ϕ only through the inner products $\{\langle \mu_k, \phi\rangle, k \in \mathcal{I}\}$. Specifically, the probability of outcome w_i is $|\langle \mu_i, \phi\rangle|^2$. If w_i is observed, then $f(\{\langle \mu_k, x\rangle, k \in \mathcal{I}\}, w_i) = i$.

A general quantum measurement corresponding to a complete set of orthogonal projection operators $\{P_i, i \in \mathcal{I}\}$ onto subspaces $\{\mathcal{S}_i \subseteq \mathcal{H}, i \in \mathcal{I}\}$ can be viewed as a probabilistic mapping between $\mathcal{H}$ and the determinate states that is

1. a deterministic identity mapping for $\phi \in \mathcal{S}_i$;

2. a probabilistic mapping for nondeterminate states that maps ϕ to a normalized vector in the direction of the orthogonal projection $P_i\phi$ for some value $i \in \mathcal{I}$, where $i = f(\{\langle P_k\phi, P_k\phi\rangle, k \in \mathcal{I}\}, w_i)$.

Here $f\colon\ \mathcal{H} \times \mathcal{W} \to \mathcal{I}$ is a probabilistic mapping between elements ϕ of $\mathcal{H}$ and indices $i \in \mathcal{I}$, that depends on a chance variable W with a discrete alphabet $\mathcal{W} = \mathcal{I}$ such that the probability of outcome $w_i \in \mathcal{W}$ depends on the input ϕ only through the inner products $\{\langle P_k\phi, P_k\phi\rangle, k \in \mathcal{I}\}$. Specifically, the probability of outcome w_i is $\langle P_i\phi, P_i\phi\rangle$. If w_i is observed, then $f(\{\langle P_k\phi, P_k\phi\rangle, k \in \mathcal{I}\}, w_i) = i$.

This perspective on the quantum measurement is the underpinning of the QSP measurement, defined in Chapter 4.

3.2 Generalized Measurements

By adding an auxiliary system and performing orthogonal measurements on the combined system, we can implement *generalized measurements* [61, 62], which is sometimes a more efficient way of obtaining information about the state of a quantum system than a standard measurement. Alternatively, we can view a generalized measurement as a standard measurement followed by an orthogonal projection onto a lower space [62, 68]. We expand on this interpretation in Section 3.3.1 in the context of Neumark's theorem.

A generalized measurement on a subspace $\mathcal{U} \subseteq \mathcal{H}$ in which the system to be measured is known *a priori* to lie is defined by a set $\{Q_i, i \in \mathcal{I}\}$ of nonnegative Hermitian operators, not necessarily projectors, that satisfy $\sum_{i\in\mathcal{I}} Q_i = I_{\mathcal{U}}$. Such a set of operators is termed a positive operator-valued measure (POVM).

A rank-one POVM acting on a subspace $\mathcal{U} \subseteq \mathcal{H}$ is defined by a set of measurement vectors $\{\mu_i, i \in \mathcal{I}\}$ that satisfy

$$\sum_{i\in\mathcal{I}} \mu_i \mu_i^* = P_{\mathcal{U}}, \tag{3.8}$$

i.e., the operators $Q_i = \mu_i \mu_i^*$ must be a resolution of the identity[2] on $\mathcal{U}$. A POVM is more general than a standard measurement in that the measurement vectors μ_i are not required to be either normalized or orthogonal.

3.3 Measurement Matrices

The *measurement matrix* $\mathbf{M}$ corresponding to a set of m measurement vectors $\{\mu_i \in \mathcal{U}, 1 \leq i \leq m\}$ is defined as the matrix of columns μ_i [68]. We have immediately from (3.8) that

$$\mathbf{M}\mathbf{M}^* = P_{\mathcal{U}}. \tag{3.9}$$

Thus a matrix $\mathbf{M}$ with m columns in $\mathcal{H}$ is a measurement matrix for states in the subspace $\mathcal{U} \subseteq \mathcal{H}$ if and only if $\mathbf{M}$ satisfies (3.9). Note that in the special case in which $\mathbf{M}$ has full column rank, (3.9) implies that the vectors μ_i are orthonormal; however if $\mathbf{M}$ does not have

[2]Often these operators are supplemented by a projection $Q_0 = P_{\mathcal{U}^\perp} = I_{\mathcal{H}} - P_{\mathcal{U}}$ onto the orthogonal subspace $\mathcal{U}^\perp \subseteq \mathcal{H}$, so that the augmented POVM is a resolution of the identity on $\mathcal{H}$.

full column rank, then the measurement vectors μ_i are not orthonormal.

It follows from Theorem 2.2 that a measurement matrix $\mathbf{M}$ with m columns in $\mathcal{H}$ corresponds to a rank-one POVM acting on an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ if and only if $\mathbf{M}$ is a transjector (partial isometry) between $\mathcal{U}$ and an n-dimensional subspace $\mathcal{V} \subseteq \mathbb{C}^m$. We summarize the properties of measurement matrices in the following theorem [68].

Theorem 3.1 (Measurement matrices). *The following statements are equivalent for a matrix $\mathbf{M}$ whose columns are m vectors in a Hilbert space $\mathcal{H}$:*

1. *$\mathbf{M}$ is a measurement matrix corresponding to a rank-one POVM acting on an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$;*
2. *$\mathbf{M}$ is a transjector between n-dimensional subspaces $\mathcal{U} \subseteq \mathcal{H}$ and $\mathcal{V} \subseteq \mathbb{C}^m$;*
3. *$\mathbf{M}\mathbf{M}^* = P_{\mathcal{U}}$ for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$;*
4. *$\mathbf{M}^*\mathbf{M} = P_{\mathcal{V}}$ for an n-dimensional subspace $\mathcal{V} \subseteq \mathbb{C}^m$.*

A measurement matrix $\mathbf{M}$ corresponding to a rank-one POVM on an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ may be expressed as $\mathbf{M} = \mathbf{U}\mathbf{Z}_n\mathbf{V}^*$, where $\mathbf{U}$ is a unitary matrix whose first n columns $\{\mathbf{u}_i, 1 \leq i \leq n\}$ are an orthonormal basis for $\mathcal{U}$, $\mathbf{V}$ is an $m \times m$ unitary matrix whose first n columns $\{\mathbf{v}_i, 1 \leq i \leq n\}$ are an orthonormal basis for $\mathcal{V}$, and $\mathbf{Z}_i, 1 \leq i \leq m$ is given by

$$\mathbf{Z}_i = \overbrace{\left[\begin{array}{c|c} \mathbf{I}_i & 0 \\ \hline 0 & 0 \end{array}\right]}^{m}. \tag{3.10}$$

A measurement matrix $\mathbf{M}$ is an isometry if restricted to $\mathcal{V}$.
A measurement matrix $\mathbf{M}$ whose columns are m vectors in $\mathcal{H}$ represents a standard measurement if and only if its rank is m. Then $\mathbf{M} = \mathbf{U}\mathbf{Z}_m\mathbf{V}^*$, and $\mathbf{M}^*\mathbf{M} = \mathbf{I}_m$.

3.3.1 Neumark's Theorem

Neumark's theorem [62, 68] guarantees that any POVM with measurement vectors $\mu_i \in \mathcal{U}$ can be realized by a set of orthonormal vectors $\tilde{\mu}_i$ in an extended space $\widetilde{\mathcal{U}}$ such that $\mathcal{U} \subseteq \widetilde{\mathcal{U}}$, so that $\mu_i = P_{\mathcal{U}}\tilde{\mu}_i$.

Using the measurement matrix characterization of a POVM and the SVD, we now obtain a simple statement and proof of Neumark's theorem. Moreover, our proof is constructive;

we explicitly construct a set of orthogonal measurement vectors such that their projections onto $\mathcal{U}$ are the original measurement vectors.

In Chapter 5 we develop a relationship between POVMs and tight frames. We then apply ideas and results derived in the context of quantum measurement to the construction and characterization of combined QSP measurements and tight frames. In particular, we use the construction given in this proof to extend a tight frame into an orthogonal basis for a larger space.

Theorem 3.2 (Neumark's theorem). *Let* $\mathbf{M}$ *be a rank-n measurement matrix of an arbitrary POVM, with m columns in a Hilbert space* $\mathcal{H}$*. In other words,* $\mathbf{M}$ *is a transjector between an n-dimensional subspace* $\mathcal{U} \subseteq \mathcal{H}$ *and an n-dimensional subspace* $\mathcal{V} \subseteq \mathbb{C}^m$*. Then there exists a standard measurement with measurement matrix* $\widetilde{\mathbf{M}}$ *which is a transjector between an expanded m-dimensional subspace* $\widetilde{\mathcal{U}} \supseteq \mathcal{U}$ *in a possibly expanded Hilbert space* $\widetilde{\mathcal{H}} \supseteq \mathcal{H}$ *and* $\mathbb{C}^m$*, and whose projection onto* $\mathcal{U}$ *is* $\mathbf{M} = P_{\mathcal{U}}\widetilde{\mathbf{M}}$*.*

Proof: Using Theorem 3.1 we may express $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\mathbf{Z}_n\mathbf{V}^*$. Let $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the columns of $\mathbf{U}$ and $\mathbf{V}$ respectively, and let $k = \dim\ \mathcal{H}$.

We distinguish between the case $k \geq m$ (*i.e.,* $\mathbf{M}$ has at least as many rows as columns), and the case $k < m$ (*i.e.,* $\mathbf{M}$ has more columns than rows).

In the case $k \geq m$, define $\widetilde{\mathbf{M}} = \sum_{i=1}^{m} \mathbf{u}_i\mathbf{v}_i^*$; then $\widetilde{\mathcal{U}} \subseteq \mathcal{H}$ is the m-dimensional subspace spanned by $\{\mathbf{u}_i, 1 \leq i \leq m\}$. The projection of $\widetilde{\mathbf{M}}$ onto $\mathcal{U}$ is

$$P_{\mathcal{U}}\widetilde{\mathbf{M}} = \sum_{j=1}^{n} \mathbf{u}_j\mathbf{u}_j^* \sum_{i=1}^{m} \mathbf{u}_i\mathbf{v}_i^* = \sum_{i=1}^{n} \mathbf{u}_i\mathbf{v}_i^* = \mathbf{M}. \tag{3.11}$$

Moreover, the columns of $\widetilde{\mathbf{M}}$ are orthonormal, since its Gram matrix is

$$\widetilde{\mathbf{M}}^*\widetilde{\mathbf{M}} = \sum_{j=1}^{m} \mathbf{v}_j\mathbf{u}_j^* \sum_{i=1}^{m} \mathbf{u}_i\mathbf{v}_i^* = \sum_{i=1}^{m} \mathbf{v}_i\mathbf{v}_i^* = \mathbf{I}_m. \tag{3.12}$$

In the case $k < m$, first embed $\mathcal{U}$ in an m-dimensional space $\widetilde{\mathcal{U}}$ in an expanded Hilbert space $\widetilde{\mathcal{H}} \supseteq \mathcal{H}$, and let $\{\tilde{\mathbf{u}}_i, 1 \leq i \leq m\}$ be an orthonormal basis for $\widetilde{\mathcal{U}}$ of which the first n vectors are the $\mathcal{U}$-basis. Then proceed as before, using $\tilde{\mathbf{u}}_i$ in place of $\mathbf{u}_i$. □

It is instructive to consider the matrix representation of $\widetilde{\mathbf{M}}$ in both cases. Recall that $\mathbf{M} = \mathbf{U}\mathbf{Z}_n\mathbf{V}^*$, where $\mathbf{Z}_n$ is given by (3.10).

In the case $k \geq m$, we construct $\widetilde{\mathbf{M}}$ simply by extending the identity matrix along the diagonal so that $\widetilde{\mathbf{M}} = \mathbf{U}\mathbf{Z}_m\mathbf{V}^*$. (If $k = m$, then $\mathbf{Z}_m = \mathbf{I}_m$ and $\widetilde{\mathbf{M}} = \mathbf{U}\mathbf{V}^*$).

In the case $k < m$, we first replace the left unitary matrix $\mathbf{U}$ by $\widetilde{\mathbf{U}}$, and thus replace k by $\tilde{k} = m$; then $\widetilde{\mathbf{U}}$ is an $m \times m$ unitary matrix whose first n columns are the $\mathcal{U}$-basis (where we append $m - k$ zeros to each basis vector $\mathbf{u}_i$). We then define $\widetilde{\mathbf{M}} = \widetilde{\mathbf{U}}\mathbf{V}^*$.

3.4 Quantum Detection and Optimal Quantum Measurements

The constraints imposed by the physics on a quantum measurement lead to some interesting problems within the framework of quantum mechanics. In particular, when using quantum systems in a communication context a problem that arises is the *quantum detection* problem. The constraints imposed in this problem suggest some intriguing signal processing algorithms that we explore in Chapters 8–12.

We now describe the quantum detection problem and recapitulate some results on optimal quantum measurements according to various criteria, which will be relevant to the construction of optimal QSP measurements (Chapter 8), to the design of optimal detectors (Chapters 9 and 12), to the development of a new viewpoint towards whitening and other covariance shaping problems (Chapter 10), and to the derivation of a new linear estimator (Chapter 11).

In a quantum detection problem a sender, Alice, conveys classical information to a receiver, Bob, using a quantum-mechanical channel. Alice represents messages by preparing the quantum channel in a pure quantum state drawn from a collection of known states. Bob detects the information by subjecting the channel to a measurement. If the states are mutually orthogonal, then Bob can perform an orthogonal measurement that will determine the state correctly w.p. 1 [99]. The optimal measurement consists of projections onto the given states. However, if the states are not orthogonal, then no measurement will allow Bob to distinguish perfectly between them. Bob's problem is therefore to construct a measurement optimized to distinguish between non-orthogonal quantum states.

Therefore, let $\{\phi_i, 1 \leq i \leq m\}$ be a collection of $m \leq k$ normalized vectors ϕ_i in a k-dimensional Hilbert space $\mathcal{H}$, representing different preparations of a quantum system. In general these vectors are non-orthogonal and span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$.

To distinguish between the different preparations, we subject the system to a measure-

ment. For our measurement, we restrict our attention to rank-one POVMs consisting of m operators of the form $Q_i = \mu_i \mu_i^*$ with measurement vectors $\mu_i \in \mathcal{U}$. We do not require the vectors μ_i to be orthogonal or normalized. However, to constitute a POVM on $\mathcal{U}$ the measurement vectors must satisfy (3.8).

If the states are prepared with equal prior probabilities, then the probability of correct detection using the measurement vectors μ_i is given from (3.5) by

$$P_d = \frac{1}{m} \sum_{i=1}^{m} |\langle \mu_i, \phi_i \rangle|^2. \tag{3.13}$$

If the vectors μ_i are orthonormal, then choosing $\mu_i = \phi_i$ results in $P_d = 1$. However, if the given vectors are not orthonormal, then no measurement can distinguish perfectly between them. Therefore, a fundamental problem in quantum mechanics is to construct measurements optimized to distinguish between a set of non-orthogonal pure quantum states.

This problem may be formulated as a quantum detection problem, so that the measurement vectors are chosen to minimize the probability of detection error, or more generally, minimize the Bayes cost. Necessary and sufficient conditions for an optimum measurement minimizing the Bayes cost have been derived [21, 22, 23]. However, except in some particular cases [23, 24, 25], obtaining a closed-form analytical expression for the optimal measurement directly from these conditions is a difficult and unsolved problem.

In [26] we take an alternative approach[3] of choosing a different optimality criterion, namely a squared-error criterion, and seeking a measurement that minimizes this criterion. Specifically, the measurement vectors μ_i are chosen to minimize the sum of the squared norms of the error vectors $e_i = \mu_i - \phi_i$, as illustrated in Figure 3-1. The optimizing measurement is referred to as the *least-squares measurement* (LSM).

It turns out that the LSM problem has a simple closed-form solution, which we discuss in the general context of least-squares (LS) inner-product shaping in Chapter 8, with many desirable properties. Its construction is relatively simple; it can be determined directly from the given collection of states; it minimizes the probability of detection error when the states exhibit certain symmetries [26]; it is "pretty good" when the states to be distinguished are equally likely and almost orthogonal [102]; it achieves a probability of error within a factor of two of the optimal probability of error [103]; and it is asymptotically optimal [104].

[3]This work was done in collaboration with G. D. Forney, Jr..

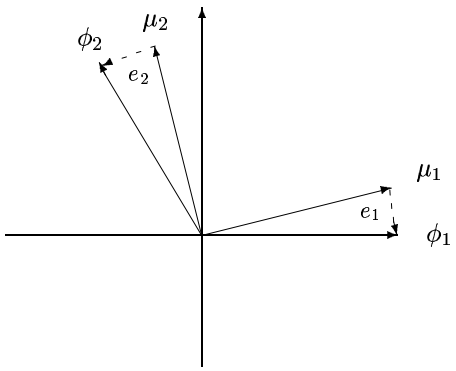


Figure 3-1: 2-dimensional example of the least-squares measurement.

The LSM can be used to motivate a new method of orthogonalization that constructs a set of orthogonal vectors that is closest in a LS sense to a given set of vectors. This orthogonalization method is the basis for constructing optimal QSP measurements in Chapter 8, in which we construct a set of vectors to have any desired inner product structure, and to be closest in a LS sense to a given set of vectors, which we refer to as LS inner product shaping. These new methods are then used in Chapters 8–12 to develop effective solutions to a variety of problems in areas ranging from frame theory to multiuser wireless communication. We demonstrate that, even for problems without inherent orthogonality or other inner product constraints, imposing such constraints in combination with LS inner product shaping can lead to new processing techniques that often exhibit improved performance over traditional methods.

Another interesting aspect of the quantum detection problem is that, as we will see for example in Chapters 8, 9 and 12, the mathematical form of the quantum detection problem appears in a variety of different signal processing problems, so that the ideas and results we developed in the context of quantum detection are also useful in many signal processing applications.

A special class of vector sets that plays an important role in the context of quantum detection is the class of *geometrically uniform (GU)* vectors [77]. Exploiting the strong symmetry properties of these sets, we show in [26] that for such vector sets the LSM minimizes the probability of detection error, so that in this case the LSM constitutes a solution to the unsolved quantum detection problem. The proof of this result reveals some nice properties of these vector sets which we develop further in [27, 76].

As we will demonstrate in the ensuing chapters, GU vector sets are also central to a variety of QSP applications. In particular, in Chapter 5 and in [76] we introduce the class of GU frames which are highly structured frames that posses nice computational properties. In Chapters 9–12 we show that GU vector sets play an important role in several classical detection and estimation problems such as matched filter detection and multiuser detection.

In the next chapter we exploit the basic principles of measurement, consistency and quantization as formulated in this chapter to the development of the QSP measurement, and we indicate how the constraints and ideas of quantum detection will be applied in the framework of QSP.

Chapter 4

QSP Measurement

The QSP measurement framework derives from the formalism of quantum mechanics by drawing a parallel between a quantum mechanical measurement and a signal processing algorithm. This framework is aimed at developing new or modifying existing signal processing algorithms by drawing heavily on the notions of measurement, consistency and quantization as they apply to quantum systems and by borrowing further from the interesting constraints imposed by quantum physics. However, it is broader and less restrictive than the quantum measurement framework since in designing algorithms we are not constrained by the physical limitations of quantum mechanics.

To exploit the formalism and rich mathematical structure of quantum mechanics in the design of algorithms we associate a *QSP measurement* with a signal processing algorithm. We then apply the formalism and fundamental principles of quantum measurement to the definition of the QSP measurement. In the QSP framework, a signal is processed by either subjecting it (or its representation in a possibly different signal space) to a QSP measurement, or by processing it using some of the measurement parameters. This framework leads to a variety of interesting processing techniques which we explore in the thesis. In particular, since the QSP measurement is defined to have a similar mathematical structure as a quantum measurement, the mathematical constraints imposed by the physics on the quantum measurement can also be imposed on the QSP measurement leading to some intriguing new signal processing algorithms. Furthermore, as will become apparent in examples throughout the thesis, many known signal processing algorithms and techniques can be described in terms of a QSP measurement by choosing the appropriate measurement

parameters. The power of the QSP measurement framework and its pivotal contribution is in providing a common umbrella for a multitude of different processing techniques and a systematic framework for generating new, potentially effective and efficient processing methods by modifying the measurement parameters. The modifications we consider result from imposing some of the additional constraints of quantum mechanics on the parameters or relaxing some of these constraints. Throughout the thesis we demonstrate that this new framework leads to a variety of interesting new algorithms in areas ranging from frame theory, sampling and quantization to detection, estimation and covariance shaping.

In this chapter we introduce and develop the QSP measurement. We begin by defining rank-one QSP measurements (ROM) which, in analogy to rank-one quantum measurements, are described by a set of measurement vectors. We then define subspace measurements (SMs) which are described by a set of projection operators, in analogy to higher-rank quantum measurements. A special case of a SM is a simple SM (SSM) which is a SM defined by a single projection operator, so that it is equal to a linear projection.

Throughout the chapter we provide examples demonstrating the utility of the QSP framework in deriving new processing methods by either casting an existing algorithm in terms of a QSP measurement, and then changing some of the parameters of the measurement describing the algorithm, or by directly processing a signal using some of the measurement parameters and then imposing constraints borrowed from quantum mechanics on these parameters. As we show, this framework provides a unified conceptual structure for a variety of traditional processing techniques, and a precise mathematical setting for developing generalizations and extensions of algorithms. While some of these examples are highly preliminary and require further evaluation, they illustrate the potential of the framework and the type of procedure that might be followed in using our framework to generate new processing techniques, as well as highlight some possible directions for future research.

In this chapter we focus primarily on developing the measurement framework. Detailed applications of ROMs and SSMs to frame theory, sampling, quantization, matched filter detection, covariance shaping, parameter estimation and multiuser detection are explored in Chapters 5–12. As an example of a direction for application of SMs, in Section 4.4 we develop a subspace approach for transmitting information over a channel with a particular structure. Although the discussion constitutes a rather preliminary exploration of such coding techniques, it represents an interesting and potentially useful model for communication

in many contexts.

4.1 The QSP Measurement

To derive the concept of QSP measurement we exploit the principles of measurement, consistency and quantization as they relate to quantum mechanics, as discussed in Chapter 3.

Measurement of a signal in the QSP framework corresponds to applying an algorithm to a signal. A QSP measurement M on a Hilbert space $\mathcal{H}$ operates on an input signal $x \in \mathcal{H}$, and returns an output signal $y = M(x) \in \mathcal{H}$. As illustrated in Fig. 4-1, the input signal x represents the signal $\tilde{x} \in \mathcal{X}$ we wish to process, so that $x = T_{\mathcal{X}}(\tilde{x}) \in \mathcal{H}$ for a mapping $T_{\mathcal{X}}\colon\ \mathcal{X} \to \mathcal{H}$, where $T_{\mathcal{X}}$ may be equal to the identity in which case $x = \tilde{x}$. Similarly, the measurement outcome $y \in \mathcal{H}$ represents the algorithm output $\tilde{y} \in \mathcal{Y}$ so that $\tilde{y} = T_{\mathcal{Y}}(y)$ for a mapping $T_{\mathcal{Y}}\colon\ \mathcal{H} \to \mathcal{Y}$, where $T_{\mathcal{Y}}$ may be equal to the identity in which case $y = \tilde{y}$. In developing the QSP measurement framework we explicitly assume that the measurement input x and the measurement output y lie in $\mathcal{H}$. Note, however, that after designing the QSP measurement M, we may always combine $T_{\mathcal{X}}$, M, and $T_{\mathcal{Y}}$ into a single mapping $T_M\colon\ \mathcal{X} \to \mathcal{Y}$, as illustrated in Fig. 4-1.

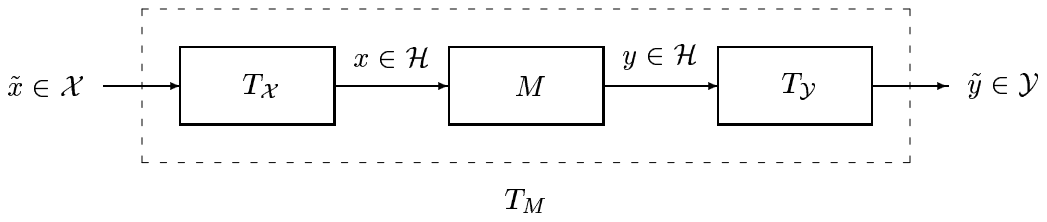


Figure 4-1: Processing a signal $\tilde{x} \in \mathcal{X}$ using a QSP measurement M on $\mathcal{H}$. If necessary, then the algorithm input $\tilde{x} \in \mathcal{X}$ is first mapped to $x = T_{\mathcal{X}}(\tilde{x}) \in \mathcal{H}$. Similarly, the measurement outcome $y = M(x) \in \mathcal{H}$ may be mapped to the algorithm output $\tilde{y} = T_{\mathcal{Y}}(y) \in \mathcal{Y}$ if necessary.

In applications involving QSP measurements that are not equal to linear projection operators, we will explicitly assume that $x \neq 0$ in Fig. 4-1.

4.1.1 Rank-One QSP Measurement

In analogy with the rank-one quantum mechanical measurement, a rank-one QSP measurement (ROM) M on $\mathcal{H}$ is a nonlinear mapping that is described by a set of measurement

vectors $\{q_i, i \in \mathcal{I}\}$, where $\mathcal{I}$ denotes an index set, and $q_i \in \mathcal{H}$ for all $i \in \mathcal{I}$. We assume that no two vectors q_i are multiples of each other, so that the one-dimensional spaces $\mathcal{S}_i$ spanned by the measurement vectors q_i intersect only at 0, *i.e.,* are disjoint. Since we are not constrained by the physics of quantum mechanics, in QSP, in contrast to quantum mechanics, these vectors are not constrained to be orthonormal. Nonetheless, in some applications we will find it useful to impose such a requirement.

The primary constraint on the measurement $M(x)$ is the quantum mechanical notion of consistency: successive measurements must yield identical outcomes. Mathematically,

$$M(M(x)) = M(x). \tag{4.1}$$

Quantization of the measurement outcome is imposed by requiring that the outcome signal is one of a set of signals determined by the measurement vectors. Specifically, in analogy with the quantum mechanical determinate states we define the *determinate signals*, which are the signals $x \in \mathcal{H}$ such that $x \in \mathcal{S}_i$ for some $i \in \mathcal{I}$. Equivalently, x is determinate if $x = cq_i$ for some $c \in \mathbb{C}$ and $i \in \mathcal{I}$. We denote the set of determinate signals of M by X_M.

The definition of $M(x)$ derives from the definition of a quantum measurement, and preserves the two fundamental principles of a quantum measurement, *i.e.,* consistency and quantization of the measurement output. Recall from Chapter 3 that a rank-one quantum measurement on $\mathcal{H}$ is a probabilistic mapping between $\mathcal{H}$ and the set of determinate states that is

1. a deterministic identity mapping for determinate states;

2. a probabilistic mapping for nondeterminate states that maps the input state to a determinate state where the probability of mapping to a particular determinate state is a function of the inner product between the nondeterminate state and the determinate state.

In defining the QSP measurement of x we emulate these properties of a quantum measurement. With E_i denoting a projection onto the one-dimensional space $\mathcal{S}_i$ spanned by the vector q_i, $E_i x \in \mathcal{S}_i$ for any $x \in \mathcal{H}$ so that $E_i x$ is a determinate signal. Hence, we define $M(x)$ as a probabilistic mapping[1] between $\mathcal{H}$ and the determinate signals X_M that is

[1] We use the terminology probabilistic mapping whenever the mapping may be probabilistic.

1. a deterministic identity mapping for $x \in \mathcal{S}_i$ defined by $M(x) = E_i x = x$;

2. a probabilistic mapping for nondeterminate signals x defined by $M(x) = E_i x$ for some value $i \in \mathcal{I}$, where $i = f_M(\{\langle q_k, x\rangle, k \in \mathcal{I}\})$.

Here f_M: $\mathcal{H} \to \mathcal{I}$ is a probabilistic mapping between elements x of $\mathcal{H}$ and indices $i \in \mathcal{I}$, that depends on the input x only through the inner products $\langle q_k, x\rangle$, where k ranges over $\mathcal{I}$. The role of the probabilistic mapping f_M is to quantize a nondeterminate signal x to one of the determinate signals $\{E_i x, i \in \mathcal{I}\}$. Then, like a quantum measurement, the output of a QSP measurement M is always a determinate signal of M and is a function of the inner products between the input signal and the measurement vectors which are a subset of the determinate signals.

Note, that we can always define the mapping f_M so that for $x \in \mathcal{S}_i$, $f_M(x) = i$. Then for all x, $M(x) = E_i x$ where $i = f_M(\{\langle q_k, x\rangle, k \in \mathcal{I}\})$. However we prefer to distinguish between the two cases so as not to impose additional constraints on f_M.

As an example of a (deterministic) mapping f_M, we may choose

$$f_M(x) = \arg\max_{k \in \mathcal{I}} \langle q_k, x\rangle. \tag{4.2}$$

As another example, we may choose

$$f_M(x) = \arg\max_{k \in \mathcal{I}} (\langle q_k, x\rangle - \langle q_k, q_k\rangle). \tag{4.3}$$

The mapping f_M of (4.3) chooses the vector q_k that minimizes the distance $\|x - q_k\|$ so that it maps x to the closest vector q_k.

As an example of a probabilistic mapping, emulating the quantum mechanical rule we may choose f_M: $\mathcal{H} \times \mathcal{W} \to \mathcal{I}$ as a probabilistic mapping from $\mathcal{H}$ to $\mathcal{I}$, where $\mathcal{W} = \mathcal{I}$ is the sample space of an auxiliary chance variable w, such that w can take on a value $w_i \in \mathcal{I}$ with probability $c|\langle q_i, x\rangle|^2$, where c is an appropriate normalization constant. Then let $f_M(x, w_i) = i$. The outcome $M(x)$ when x is not determinate is then given by

$$M(x) = E_i x \text{ with probability } c|\langle q_i, x\rangle|^2. \tag{4.4}$$

We note that the output of the algorithm resulting from a measurement M of a non-

determinate x may depend in principle on the choice of projections E_i onto $\mathcal{S}_i$. In some applications, rather than choosing arbitrary projections, it may be of interest to explore "optimal" methods for choosing these projections. In subsequent chapters we will see some examples of how the choice of projection can affect the algorithm output in the context of SMs. In applications such as quantization and detection, a common choice of output mapping $T_{\mathcal{Y}}$ in Fig. 4-1 maps any vector $y \in \mathcal{S}_i$ to some $\tilde{y}_i \in \mathcal{Y}$. Since for any projection E_i onto $\mathcal{S}_i$, $y = E_i x \in \mathcal{S}_i$, in this case the choice of projection will not effect the algorithm output $\tilde{y} = T_{\mathcal{Y}}(y)$.

The discussion above is summarized in the following definition:

Definition 4.1 (Rank-one QSP measurement). *Let M be a rank-one QSP measurement on $\mathcal{H}$ with measurement vectors $\{q_i, i \in \mathcal{I}\}$ that lie in $\mathcal{H}$. Let $\mathcal{S}_i$ denote the one-dimensional space spanned by the vector q_i, and let E_i denote a projection onto $\mathcal{S}_i$. Let X_M denote the determinate signals of M that are the vectors $x \in \mathcal{S}_i$ for some $i \in \mathcal{I}$. Then the outcome of the measurement $M(x)$ of $x \in \mathcal{H}$ is given by*

$$M(x) = \begin{cases} x, & x \in X_M; \\ E_i x \text{ where } i = f_M(\{\langle q_k, x\rangle, k \in \mathcal{I}\}), i \in \mathcal{I}, & x \notin X_M, \end{cases} \tag{4.5}$$

where f_M: $\mathcal{H} \to \mathcal{I}$ is a probabilistic mapping between elements of $\mathcal{H}$ and the index set $\mathcal{I}$, and depends on the input x only through the the inner products $\{\langle q_k, x\rangle, k \in \mathcal{I}\}$.

The QSP measurement leads in general to nonlinear algorithms that are characterized by a set of linear projection operators. These algorithms have a simple structure taking on the form of a linear projection operator E_i operating on an input signal, where the projection is chosen out of a set of possible projections $\{E_i, i \in \mathcal{I}\}$, and the particular choice of projection is determined by the probabilistic mapping f_M. Drawing from the definition of a quantum measurement, the QSP measurement is then formulated in a way that ensures consistency for all inputs. A direct consequence is that the output of the algorithm is always 'quantized' to one of the determinate signals. Note, however, that f_M does not necessarily quantize the signal in the strict sense since typically a quantizer has a countable set of outputs, whereas the images in the output space $\mathcal{Y}$ of the determinate signals may in principle contain an uncountable number of signals, so that the final output in $\mathcal{Y}$ may be chosen from an uncountable set.

Since a ROM depends directly on the choice of probabilistic mapping f_M, the choice of measurement vectors q_i, and possibly the choice of projections E_i, this framework provides a systematic method for deriving new, potentially interesting processing algorithms by, for example, varying the probabilistic mapping f_M or imposing constraints on the measurement vectors q_i.

In summary, our definition of a ROM M is very reminiscent of the quantum measurement as defined in Section 3.1, and preserves the fundamental concepts of consistency and quantization. However, whereas in quantum mechanics f_M is uniquely specified as the probabilistic mapping in which the measurement vector q_i is chosen with probability $|\langle q_i, x\rangle|^2$, in our formulation we allow for more general probabilistic and deterministic mappings f_M. Furthermore, the measurement vectors are not constrained to be orthonormal as in quantum mechanics, and the projections E_i are not constrained to be orthogonal projections.

In later chapters of the thesis we demonstrate the utility of the QSP measurement framework in deriving effective solutions to problems in a wide range of areas. All the processing methods we develop result from either formulating an algorithm as a QSP measurement and then systematically changing some of the parameters on which the measurement depends, or by processing a signal using one of the measurement parameters and then imposing constraints borrowed from quantum mechanics on these parameters. Typical modifications we consider include choosing a probabilistic mapping f_M emulating the quantum mechanical measurement, and imposing inner product constraints on the measurement vectors. There are potentially a host of additional applications of the QSP framework beyond those explored in the subsequent chapters. In particular, we may consider imposing other constraints on the measurement vectors and choosing different probabilistic mappings f_M. Perhaps the most rewarding direction for future research is in discovering and developing further applications of this framework.

4.2 Algorithm Design Using Rank-One Measurements

In the QSP framework signals are processed by either subjecting them to a QSP measurement, or by using some of the QSP measurement parameters but not directly applying the measurement. In Section 4.2.1 we consider designing algorithms using the QSP measurement, and in Section 4.2.2 we consider algorithms that result from using some of the

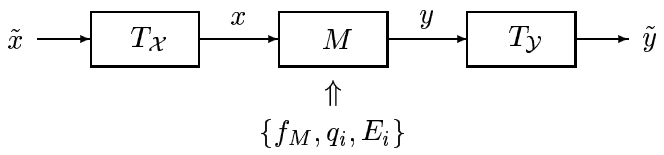


Figure 4-2: Designing algorithms using a rank-one measurement.

measurement parameters.

4.2.1 Algorithm Design by Applying Rank-One Measurements

Algorithm design

To design an algorithm using a ROM we first identify the measurement vectors which specify the possible measurement outcomes. For example, in a detection scenario the measurement vectors may be equal to the transmitted signals, or may represent these signals in a possibly different space. As another example, in a scalar quantizer the measurement vectors may be chosen as a set of vectors that represent the scalar quantization levels. We then embed the measurement vectors in a Hilbert space $\mathcal{H}$ and, if necessary, map the signal $\tilde{x}$ to be processed into a signal x in $\mathcal{H}$ using a mapping $T_{\mathcal{X}}$. To obtain the algorithm output we first measure the representation x of the signal to be processed. If x is a determinate signal of M, then the measurement outcome is $y = M(x) = x$. Otherwise we approximate x by a determinate signal y using the probabilistic mapping f_M. If necessary, the final algorithm output $\tilde{y}$ may be obtained from the measurement outcome y using the mapping $T_{\mathcal{Y}}$. These basic steps are illustrated in Fig. 4-2.

By choosing different input and output mappings $T_{\mathcal{X}}$ and $T_{\mathcal{Y}}$, and different measurement parameters f_M, q_i and E_i in Fig. 4-2, we can arrive at a variety of new and interesting processing techniques.

Modifying known algorithms

As we now demonstrate, many traditional detection and processing techniques fit naturally into the framework of Fig. 4-2. Once an algorithm is described as a QSP measurement M, a myriad of modifications and extensions of the algorithm can then be derived by, for example, varying f_M, or by imposing constraints on the measurement vectors of M. The

modifications we consider result from borrowing from some of the additional constraints of quantum mechanics.

In particular, QSP measurements naturally give rise to probabilistic algorithms by letting f_M be a probabilistic mapping, emulating the quantum measurement. We expand on this idea in the context of quantization in Chapter 7, and in the examples below, and in the context of combined measurements in Chapter 5. However, the full potential benefits of probabilistic algorithms in general resulting from the QSP framework remain an interesting area of future study.

Another possibility for extensions is by imposing constraints on the measurement vectors q_i. One of the interesting constraints of quantum mechanics is that measurement vectors must be orthonormal, which leads to some interesting problems within the framework of quantum mechanics such as the quantum detection problem, described in Section 3.4. A fundamental problem in quantum mechanics is to construct optimal measurements subject to this constraint, that best represent a given set of state vectors. In analogy to quantum mechanics, we may impose inner product constraints on the measurement vectors of a QSP measurement M. However, since we are not limited by physical laws, we are not confined to an orthogonality constraint. Borrowing from the ideas we developed in the context of quantum detection (see Section 3.4 and [26]), in Chapter 8 we consider in detail methods for choosing a set of measurement vectors that best represent the signals of interest and have a specified inner product structure, which we refer to as least-squares (LS) inner product shaping. Applications of this concept to matched filter (MF) detection, MMSE covariance shaping, estimation, and multiuser wireless communication systems are developed in Chapters 9–12. These applications demonstrate that LS inner product shaping, inspired by optimal quantum measurement design, is a very versatile methods with applications spanning many different areas.

We now consider applications of ROMs to MF detection and quantization. We demonstrate both how to formulate these algorithms in terms of a QSP measurement, and also how to use the framework to develop various modifications. Further details on and extensions of these applications are considered in Chapters 9 and 7, respectively.

Example 4.1 (Matched Filter). In this example we demonstrate how to cast an MF detector in terms of a QSP measurement. Suppose that one of m signals $\{s_i(t), 1 \leq i \leq m\}$ is received over an additive noise channel with equal probability, where the signals lie in a

real Hilbert space $\mathcal{H}$ with inner product $\langle x(t), y(t)\rangle = \int_{t=-\infty}^{\infty} x(t)y(t)dt$, and are normalized so that $\langle s_i(t), s_i(t)\rangle = 1$ for all i. The received signal $r(t)$ is also assumed to be in $\mathcal{H}$, and is modeled as $r(t) = s_i(t) + n(t)$ for one value i, where $n(t)$ is a stationary white noise process with zero mean and spectral density σ^2, and with otherwise unknown distribution.

A classical receiver for detecting the received signal is the well known MF detector [11], depicted in Fig. 4-3. The received signal $r(t)$ is cross-correlated with each of the m signals $s_i(t)$ so that $a_i = \langle s_i(t), r(t)\rangle$. The declared detected signal is $s_i(t)$ where $i = \arg\max a_k$. The mapping T in Fig. 4-3 maps the index i into the algorithm output $s_i(t)$.

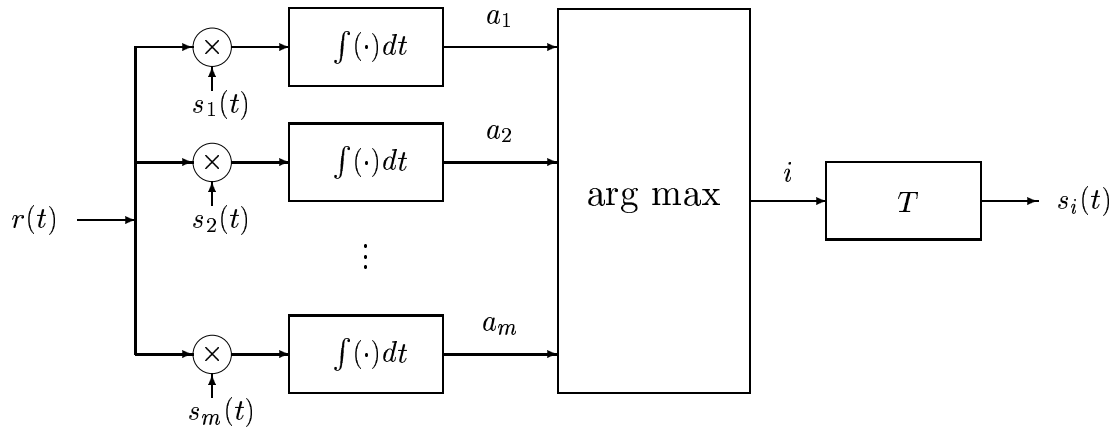


Figure 4-3: Matched filter detector.

To implement the MF detector using a QSP measurement M, we define measurement vectors $\{q_i(t) = s_i(t), 1 \leq i \leq m\}$, and define the mapping f_M by

$$f_M(r(t)) = i, \text{ where } i = \arg \max_{1 \leq k \leq m} \langle s_k(t), r(t)\rangle. \tag{4.6}$$

If $r(t) = cs_i(t)$ for some $c \in \mathbb{C}$ and one value i, then $r(t)$ is a determinate signal of M and $M(r(t)) = r(t) = cs_i(t)$. Otherwise, $M(r(t)) = E_i r(t) = cs_i(t)$ for some $c \in \mathbb{C}$, where $i = f_M(r(t)) = \arg\max\langle s_k(t), r(t)\rangle$, and E_i is a projection onto the space $\mathcal{S}_i$ spanned by $s_i(t)$. Thus the MF output can be obtained by performing the measurement M on the observed signal $r(t)$, followed by a mapping $T_{\mathcal{Y}}$ from $\mathcal{H}$ to $\mathcal{Y} = \mathcal{H}$ defined by $T_{\mathcal{Y}}(y) = s_i(t)$ if $y \in \mathcal{S}_i$. The measurement description of the MF detector is illustrated in Fig. 4-4.

We note that if the signals $s_i(t)$ do not have equal norm, then the MF receiver is

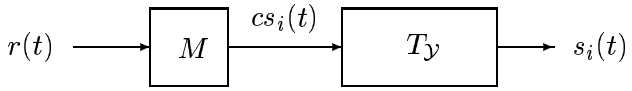


Figure 4-4: Measurement description of the matched filter detector.

modified so that the declared detected signal is $s_i(t)$ where $i = \arg\max(a_k - \|s_k(t)\|^2)$. In this case we can implement the modified MF using a ROM with mapping $f_M(x(t)) = i$ where $i = \arg\max(\langle s_k(t), x(t)\rangle - \|s_k(t)\|^2)$. A similar modification also occurs if the signals are transmitted with unequal prior probabilities. □

A significant advantage to describing a known algorithm in the QSP measurement language, is that it suggests modifications of the algorithm by changing the measurement parameters. In the next two examples we consider modifications of the MF detector that result from changing the parameters of the basic measurement describing the MF detector.

Example 4.2 (Orthogonal Matched Filter). An interesting modification of the MF detector suggested by the QSP framework results from constraining the measurement vectors of the MF measurement M to be orthonormal, as in a quantum mechanical measurement. In this case the measurement vectors can no longer in general be equal to the transmitted signals $s_i(t)$. Instead, borrowing from the construction of optimal quantum measurements in Section 3.4, we construct a measurement with orthonormal measurement vectors $h_i(t)$ that are closest in a LS sense to the transmitted signals. We discuss this construction in the context of LS inner product shaping in Chapter 8. The resulting detector consists of correlating the received signal with each of the m signals $h_i(t)$, and choosing as the declared detected signal the one for which $\langle h_i(t), r(t)\rangle$ is maximum. We refer to this detector as the orthogonal matched filter (OMF) detector [36, 37].

In Chapter 9 we discuss the OMF detector in considerable detail, and develop further modifications of the MF detector by imposing additional inner product constraints on the measurement vectors of M. We provide simulation results that suggest that in certain cases of non-Gaussian noise the OMF detector can significantly increase the probability of correct detection over the MF receiver, and may have only minor impact on performance in the Gaussian case. By exploiting results derived in the context of quantum detection [26], we show that in many practical cases the OMF detector has a property analogous to the MF

detector, namely that it maximizes the total output SNR subject to a whitening constraint on the outputs. □

In the previous example we considered a modification of the MF detector that results from imposing an inner product constraint on the measurement vectors. In the next example we consider a modification of the MF detector that results from choosing a probabilistic mapping f_M in place of the conventional MF mapping (4.6).

Example 4.3 (Probabilistic Matched Filter). Employing the measurement description of the MF detector (Example 4.1), we can derive a probabilistic MF by choosing a probabilistic mapping f_M. Specifically, emulating the quantum measurement we choose $f_M\colon\ \mathcal{H} \times \mathcal{W} \to \mathcal{I}$ as a probabilistic mapping from $\mathcal{H}$ to $\mathcal{I}$, where $\mathcal{W} = \mathcal{I}$ is the sample space of an auxiliary chance variable w with discrete alphabet $\mathcal{W}$, such that w can take on a value $w_i = i$ with probability $c|\langle s_i(t), x(t)\rangle|^2$, where for normalization, $c = 1/\sum_{k=1}^{m} |\langle s_k(t), x(t)\rangle|^2$. Then let $f_M(x(t), i) = i$. The declared detected signal is then $s_i(t)$ with probability $c|\langle s_i(t), r(t)\rangle|^2 = ca_i^2$. Note, that the output of this probabilistic MF is a random variable even when the input to the detector is known.

To allow for more freedom in the design of the probabilistic MF we can first map the transmitted signals $s_i(t)$ and the received signal $r(t)$ onto a different set of vectors in a possibly different Hilbert space. We may then design a new measurement with measurement vectors equal to the representations of the transmitted signals. Using this method we can generate any desired probability distribution on the outputs. □

In Chapter 7 we consider probabilistic algorithms in more detail, in the specific context of quantization. A brief description of a probabilistic quantizer is given in the next example.

Example 4.4 (Probabilistic Quantizer). A scalar quantizer depicted in Fig. 4-5, with m quantization levels $a_1, \ldots, a_m$, quantizes an input value z to the level $a_i = Q(z)$, where $i = \arg\min |z - a_k|$. In this example we formulate the scalar quantizer as a QSP measurement, and develop a probabilistic quantizer by modifying the measurement parameters.

To describe the scalar quantizer as a QSP measurement, we first define an input mapping $T_{\mathcal{X}}\colon\ \mathcal{R} \to \mathcal{R}^m$ that maps the quantization levels a_i and the input signal z onto vectors in $\mathcal{R}^m$. With $\{\mathbf{q}_i, 1 \le i \le m\}$ denoting m orthonormal vectors in $\mathcal{R}^m$, the mapping $T_{\mathcal{X}}$ is defined such that $T_{\mathcal{X}}(a_i) = \mathbf{q}_i$, and $T_{\mathcal{X}}(z) = \mathbf{x}$ for $z \neq a_i$ for all i, where $\mathbf{x} = \sum_{i=1}^{m} (z - a_i)^{-1}\mathbf{q}_i$. We then construct a QSP measurement with measurement vectors equal

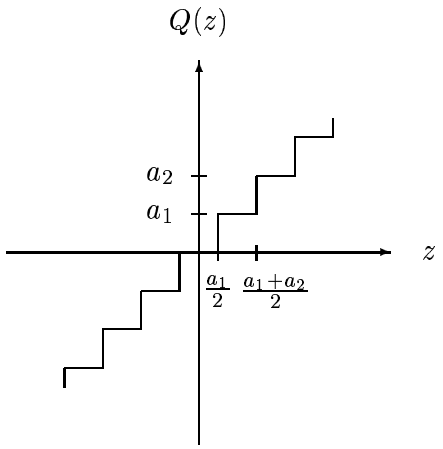


Figure 4-5: Quantizer transfer characteristic.

to the representations $\mathbf{q}_i \in \mathcal{R}^m$ of the quantization levels a_i. To quantize z using the QSP measurement M, we first map z to $\mathbf{x} = T_{\mathcal{X}}(z)$, and then perform the measurement M on $\mathbf{x}$. The outcome of the measurement is mapped to the final quantized level using the mapping $T_{\mathcal{Y}}$: $\mathcal{R}^m \to \mathcal{R}$ defined by $T_{\mathcal{Y}}(y) = a_i$ if y is a multiple of $\mathbf{q}_i$. The resulting measurement description of the quantizer is depicted in Fig. 4-6.

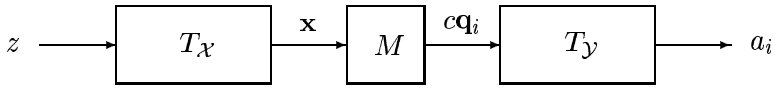


Figure 4-6: Measurement description of quantizer.

If $z = a_i$ for some i, then $\mathbf{x} = \mathbf{q}_i$, and $M(\mathbf{x}) = \mathbf{x}$. The quantized output is then $T_{\mathcal{Y}}(\mathbf{q}_i) = a_i$. If z is not equal to one of the quantization levels a_i, then $M(\mathbf{x}) = c\mathbf{q}_i$ for some $c \in \mathbb{C}$, where $i = f_M(\mathbf{x})$ and f_M is a function of the inner products $\langle \mathbf{q}_i, \mathbf{x} \rangle$. The quantized output is then $T_{\mathcal{Y}}(\mathbf{q}_i) = a_i$. The choice of the probabilistic mapping f_M determines the overall function of the quantizer.

Since $\langle \mathbf{q}_i, \mathbf{x} \rangle = (z - a_i)^{-1}$, the quantizer resulting from the measurement description of Fig. 4-6 can be described directly in $\mathcal{R}$ as $Q_M(z) = a_i$, where $i = f(z)$ and f is a probabilistic mapping that depends on z through the numbers $\{z - a_i, 1 \leq i \leq m\}$. We

refer to the resulting quantizer as a QSP quantizer with mapping f. If we choose

$$f(z) = i, \text{ where } i = \arg\min |z - a_k|, \tag{4.7}$$

then the output of the QSP quantizer is a_i where $i = \arg\min |z - a_k|$ and is equal to the output of the quantizer of Fig. 4-5 with input z.

As an example of a generalization suggested by the QSP measurement framework, suppose instead of the mapping (4.7) we choose the probabilistic mapping f: $\mathcal{R} \times \mathcal{W} \to \mathcal{I}$ where $\mathcal{W} = \mathcal{I}$ is the sample space of an auxiliary chance variable w with discrete alphabet $\mathcal{W}$, such that w can take on a value $w_i = i, 1 \leq i \leq m$ with probability $p_i = g(z - a_i)$, for some function g, that is chosen so that the values p_i represent probabilities for all possible input values z, and such that if z is close to a quantization level a_i, then the corresponding p_i is relatively large. Then let $f(z, w_i) = i$. The output of the QSP quantizer with this choice of mapping is then equal to a_i with probability p_i.

In Chapter 7 we consider the probabilistic quantizer in more detail, and show that it can be viewed as a dithered quantizer [31, 28], in which continuous-time random noise is added to the input signal prior to quantization. The advantage of this implementation of a dithered quantizer is that it can effectively realize a dither signal with an arbitrary joint probability distribution, while requiring only the generation of one uniform random variable per input. By introducing memory into the probabilistic selection rule we also derive a probabilistic quantizer that shapes the quantization noise. □

4.2.2 Algorithm Design Using the Measurement Parameters

Another class of algorithms we develop results from processing a signal with some of the measurement parameters, and then imposing quantum mechanical constraints on these parameters. In particular, we may view any linear processing of a signal as processing with a set of measurement vectors, and then impose inner product constraints on these vectors. Using the ideas of quantum detection we may then design linear algorithms that are optimal subject to these inner product constraints.

Specifically, suppose we are given a linear algorithm described by some linear transformation T: $\mathcal{X} \to \mathcal{Y}$. In an appropriate basis for $\mathcal{X}$ and $\mathcal{Y}$ we can always represent T as a (possibly infinite) matrix. Let the columns of T in this representation be denoted by t_i.

Then for an input x, the algorithm output $y = Tx$ is a linear combination of the vectors t_i. We may therefore interpret the vectors t_i as a set of measurement vectors, and then impose inner product constraints on these vectors. Thus, we may seek the vectors h_i that have a specific inner product structure and are closest in a LS sense to the vectors t_i, and then process x using the transformation H whose columns are the optimal vectors h_i. In this way we are replacing the given linear algorithm T with a modified linear algorithm H that is the 'closest' transformation to T from all linear transformations whose columns satisfy certain inner product constraints.

Drawing from the quantum detection problem, we can also develop new classes of linear algorithms that result from imposing a stochastic inner product constraint on the algorithm *i.e.,* a covariance constraint, and then deriving optimal algorithms subject to this constraint. In particular, we may extend the concept of LS inner product shaping suggested by the quantum detection framework to develop optimal algorithms that minimize a stochastic MSE criterion subject to a covariance constraint. In the next example, we use this approach to develop an interesting new viewpoint towards whitening.

Example 4.5 (MMSE Whitening). Suppose we have a random vector $\mathbf{a} \in \mathbb{C}^m$ with covariance $\mathbf{C}_a$, and we want to whiten the vector $\mathbf{a}$ using a whitening transformation $\mathbf{W}$ to obtain the random vector $\mathbf{b} = \mathbf{W}\mathbf{a}$, where the covariance matrix of $\mathbf{b}$ is given by $\mathbf{C}_b = c^2\mathbf{I}_m$ for some $c > 0$. Thus we seek a transformation $\mathbf{W}$ such that

$$\mathbf{C}_b = \mathbf{W}\mathbf{C}_a\mathbf{W}^* = c^2\mathbf{I}_m, \tag{4.8}$$

for some $c > 0$.

Given a covariance matrix $\mathbf{C}_a$, there are many ways to choose a whitening transformation $\mathbf{W}$ satisfying (4.8). However, no general assertion of optimality is known for the output $\mathbf{b} = \mathbf{W}\mathbf{a}$ of these different transformations. In particular, the white random vector $\mathbf{b} = \mathbf{W}\mathbf{a}$ may not be "close" to the input vector $\mathbf{a}$. If the vector $\mathbf{b}$ undergoes some noninvertible processing, or is used as an estimator of some unknown parameters represented by the data $\mathbf{a}$, then we may wish to choose the whitening transformation in such a way that $\mathbf{b}$ is close to $\mathbf{a}$ in some sense. This can be particularly important in applications in which $\mathbf{b}$ is the input to a detector, so that we may wish to whiten $\mathbf{a}$ prior to detection, but at the same time minimize the distortion to $\mathbf{a}$ by choosing $\mathbf{W}$ so that $\mathbf{b}$ is close to $\mathbf{a}$.

Drawing from the ideas of quantum detection, we propose a whitening transformation that is optimal in the sense that it results in a random vector $\mathbf{b}$ that is as close as possible to $\mathbf{a}$ in MSE. Specifically, among all possible whitening transformations we seek the one that minimizes the total MSE given by

$$\varepsilon_{\text{MSE}} = \sum_{i=1}^{m} E\left((a_i - b_i)^2\right) = E\left((\mathbf{a}-\mathbf{b})^*(\mathbf{a}-\mathbf{b})\right), \tag{4.9}$$

subject to (4.8), where a_i and b_i are the ith components of $\mathbf{a}$ and $\mathbf{b}$ respectively. We refer to such a whitening transformation as an MMSE whitening transformation.

In Chapter 10 we show that the MMSE whitening problem can be interpreted as a LS inner product shaping problem, so that the MMSE whitening transformation can be found by applying results derived in that context. We also consider more general forms of MMSE covariance shaping so that we seek the vector $\mathbf{b}$ with covariance proportional to an arbitrary covariance matrix $\mathbf{R}$, that is closest to $\mathbf{a}$ in an MSE sense.

This new concept of MMSE shaping, inspired by the ideas we derived in the context of the quantum detection problem, can be useful in a variety of signal processing methods that incorporate shaping transformations in which we can imagine using an optimal procedure that shapes the data but at the same time minimizes the distortion to the original data. □

As another example of an algorithm suggested by the quantum detection framework, where we use the ideas of least-squares inner product shaping to design an optimal linear algorithm subject to a stochastic inner product constraint, in the next example we consider a new linear estimator for the unknown deterministic parameters in a linear model.

Example 4.6. A generic estimation problem that has been studied extensively in the literature is that of estimating the unknown deterministic parameters $\mathbf{x}$ observed through a known linear transformation $\mathbf{H}$ and corrupted by zero-mean additive noise $\mathbf{w}$ with covariance $\mathbf{C}_w$. A common approach to estimating the parameters $\mathbf{x}$ is to restrict the estimator to be linear in the data $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, and then find the linear estimate $\hat{\mathbf{x}}_{\text{LS}} = \mathbf{G}\mathbf{y}$ of $\mathbf{x}$ that results in an estimated data vector $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}_{\text{LS}}$ that is as close as possible to the given data vector $\mathbf{y}$ in a (weighted) LS sense [42, 43, 44, 45], so that $\hat{\mathbf{x}}_{\text{LS}} = \mathbf{G}\mathbf{y}$ is chosen to minimize the total squared error

$$\varepsilon_{\text{LS}} = (\mathbf{y} - \mathbf{H}\mathbf{G}\mathbf{y})^* \mathbf{C}_w^{-1} (\mathbf{y} - \mathbf{H}\mathbf{G}\mathbf{y}). \tag{4.10}$$

However, in many cases the data vector $\mathbf{y}$ is not very sensitive to changes in $\mathbf{x}$, so that a large error in estimating $\mathbf{x}$ may translate into a small error in estimating the data vector $\mathbf{y}$, in which case the LS estimate may result in a poor estimate of $\mathbf{x}$. A difficulty often encountered in this estimation problem is that the error in the estimation can have a covariance structure with a very high dynamic range.

To improve the performance of the LS estimator we propose a modification of the LS estimate based on the ideas developed in the context of quantum detection, in which we control the dynamic range and spectral shape of the covariance of the estimation error in estimating $\mathbf{x}$ by choosing the estimator of $\mathbf{x}$ to minimize the total error variance in the observations $\mathbf{y}$, subject to a constraint on the covariance of the estimation error. The resulting estimator of $\mathbf{x}$ is referred to as the covariance shaping LS (CSLS) estimator, and is denoted by $\hat{\mathbf{x}}_{\text{CSLS}}$. Thus, $\hat{\mathbf{x}}_{\text{CSLS}} = \mathbf{G}\mathbf{y}$ is chosen to minimize

$$\varepsilon_{\text{CSLS}} = E\left((\mathbf{y}' - \mathbf{H}\mathbf{G}\mathbf{y}')^* \mathbf{C}_w^{-1} (\mathbf{y}' - \mathbf{H}\mathbf{G}\mathbf{y}')\right), \tag{4.11}$$

where $\mathbf{y}' = \mathbf{y} - E(\mathbf{y})$, subject to the constraint that the covariance of the error in the estimate $\hat{\mathbf{x}}_{\text{CSLS}}$, which is equal to the covariance of the estimate $\hat{\mathbf{x}}_{\text{CSLS}}$, is proportional to a given covariance matrix $\mathbf{R}$. Thus $\mathbf{G}$ must satisfy

$$\mathbf{G}\mathbf{C}_w\mathbf{G}^* = c^2\mathbf{R}, \tag{4.12}$$

where $c > 0$ is a constant that is either specified, or chosen to minimize the error (4.11).

In Chapter 11 we develop the CSLS estimator, *i.e.,* we determine the matrix $\mathbf{G}$ that minimizes (4.11) subject to (4.12). We also analyze the MSE of this estimator from which we conclude that over a wide range of SNR, the CSLS estimator results in a lower MSE than the traditional LS estimator, for all values of the unknown parameters. The simulations presented in Chapter 11 strongly suggest that the CSLS estimator can significantly decrease the MSE of the estimation error over the LS estimator for a wide range of SNR values. □

4.3 Subspace Measurements

In the previous section we developed the properties of a ROM M with measurement vectors $\{q_i\}$. In particular, we showed that a ROM selects an outcome from the determinate signals,

that is "best matched" to the measured signal in some deterministic or probabilistic sense. In this section we consider measurements that select a *subspace* that is "best matched" to the measured signal. We refer to these measurements as *subspace measurements (SMs).*

In analogy with the general quantum mechanical measurement, a SM M^S on $\mathcal{H}$ is a nonlinear mapping that is described by a set of projection operators $\{E_i, i \in \mathcal{I}\}$, where E_i is a projection onto a subspace $\mathcal{S}_i \subseteq \mathcal{H}$. We assume that the subspaces $\mathcal{S}_i$ intersect only at 0. Since we are not constrained by the physics of quantum mechanics, these projections are not constrained to be orthogonal projections and the subspaces $\mathcal{S}_i$ are not constrained to be orthogonal subspaces. Nonetheless, in some applications we may find it useful to impose such requirements. The outcome of a SM is always a determinate signal of M^S where, as in a ROM, a signal $x \in \mathcal{H}$ is a determinate signal of M^S if $x \in \mathcal{S}_i$ for some $i \in \mathcal{I}$. The set of determinate signals of M^S is denoted by X_M^S.

We now define the SM to preserve the fundamental principles of consistency and quantization, drawing from Definition 4.1 of a ROM and the definition of a quantum measurement. Specifically, the SM $M^S(x)$ is a probabilistic mapping between $\mathcal{H}$ and X_M^S that is

1. a deterministic identity mapping for $x \in \mathcal{S}_i$ defined by $M^S(x) = E_i x = x$;

2. a probabilistic mapping for nondeterminate signals x defined by $M(x) = E_i x$ for some value $i \in \mathcal{I}$, where $i = f_M^S(\{\langle E_k x, E_k x\rangle, k \in \mathcal{I}\})$.

Here f_M^S: $\mathcal{H} \to \mathcal{I}$ is a probabilistic mapping between elements x of $\mathcal{H}$ and indices $i \in \mathcal{I}$, that depends on the input x only through the inner products $\langle E_k x, E_k x\rangle$, where k ranges over $\mathcal{I}$. Note that if E_k is an orthogonal projection operator, then $\langle E_k x, E_k x\rangle = \langle x, E_k x\rangle$. For example, we may choose the mapping

$$f_M(x) = \arg\max_{k \in \mathcal{I}} \langle E_k x, E_k x\rangle. \tag{4.13}$$

As another example, emulating the quantum mechanical rule we may choose f_M^S: $\mathcal{H} \times \mathcal{W} \to \mathcal{I}$ as a probabilistic mapping from $\mathcal{H}$ to $\mathcal{I}$, where $\mathcal{W} = \mathcal{I}$ is the sample space of an auxiliary chance variable w, such that w can take on a value $w_i \in \mathcal{I}$ with probability $c\langle E_i x, E_i x\rangle$, where c is an appropriate normalization constant. Then let $f_M^S(x, w_i) = i$. The outcome $M^S(x)$ when x is not determinate is then given by

$$M^S(x) = E_i x \text{ with probability } c\langle E_i x, E_i x\rangle. \tag{4.14}$$

In the special case in which the projections E_i are orthogonal projections onto orthogonal subspaces $\mathcal{S}_i$, this choice of probabilistic mapping is equivalent to the probabilistic mapping in the quantum measurement.

Thus, the SM projects the input signal x onto the subspace $\mathcal{S}_i$ that is "best matched" to x in some deterministic or probabilistic sense. The particular choice of subspace depends on the probabilistic mapping f_M^S which is a function of the "similarity" between x and the subspaces $\mathcal{S}_i$ as measured by the inner products $\langle E_i x, E_i x \rangle$. Since $\langle E_i x, E_i x \rangle$ is the norm of the projection of x onto $\mathcal{S}_i$, we expect this norm to be relatively large if x is "close" to $\mathcal{S}_i$.

The discussion above is summarized in the following definition:

Definition 4.2 (Subspace measurement). *Let M^S be a subspace measurement on $\mathcal{H}$ with measurement projections $\{E_i, i \in \mathcal{I}\}$, where E_i is a projection onto a subspace $\mathcal{S}_i \subseteq \mathcal{H}$, and the subspaces $\{\mathcal{S}_i, i \in \mathcal{I}\}$ intersect only at 0. Let X_M^S denote the determinate signals of M^S that are the vectors $x \in \mathcal{S}_i$ for some $i \in \mathcal{I}$. Then the outcome of the measurement $M^S(x)$ of $x \in \mathcal{H}$ is given by*

$$M^S(x) = \begin{cases} x, & x \in X_M^S; \\ E_i x \text{ where } i = f_M^S(\{\langle E_k x, E_k x \rangle, k \in \mathcal{I}\}), i \in \mathcal{I}, & x \notin X_M^S, \end{cases} \tag{4.15}$$

where f_M^S: $\mathcal{H} \to \mathcal{I}$ is a probabilistic mapping between elements of $\mathcal{H}$ and the index set $\mathcal{I}$, and depends on the input x only through the inner products $\{\langle E_k x, E_k x \rangle, k \in \mathcal{I}\}$.

Note that in the special case in which the SM M^S is defined by a single projection E, $M^S(x) = Ex$ for all x so that the SM is equal to the projection E. In this case we refer to the SM as a *simple SM (SSM)*.

To summarize, in our development of the QSP measurement we distinguished between 3 classes of measurements: ROMs defined by a set of measurement vectors; SMs defined by a set of projections; and SSMs which are linear projections. In contrast to SSMs, ROMs and SMs are typically nonlinear and have a lot of flexibility in their design.

Several applications of ROMs where outlined in Section 4.2. Further applications are considered in detail in Chapter 5, and in Chapters 7–12. Applications of SMs are considered in the next section and in Chapters 5 and 6.

4.4 Applications of Subspace Measurements

In this section we consider some applications of SMs.

4.4.1 Simple Subspace Measurements

Since a SSM is a projection, the only flexibility in the design of a SSM is in choosing the type of projection (namely choosing the range space and null space). As we have seen in Section 2.4, there are two different types of projections: orthogonal projections and oblique projections. In contrast to orthogonal projections that abound in signal processing, oblique projections have received limited attention in the signal processing literature. In Chapter 6 we derive a general framework for consistent sampling and reconstruction procedures using a SSM equal to an oblique projection operator. In Chapter 5 we use a SSM to develop the new concept of oblique dual frame vectors that lead to frame expansions in which the analysis and synthesis frame vectors are not constrained to lie in the same space as with conventional frame expansions. This expansion is then used in Chapter 6 to develop redundant consistent sampling procedures.

Also in Chapter 5, based on oblique projections we construct subspace MF detectors for detecting a signal contaminated by both structured noise and white noise, and illustrate that these detectors can lead to improved performance over conventional GLRT based detectors.

4.4.2 Subspace Coding and Detection

Recall that a SM selects a *subspace* that is best matched to a measured signal. Therefore, SMs are useful in problems where we want to distinguish between subspaces and not between individual signals. This suggests coding strategies in which rather than encoding the desired information in a particular signal, the information is encoded in a disjoint set of not necessarily orthogonal subspaces that intersect only at 0. Detection is performed using a SM.

The concept of transmitting information in disjoint subspaces is not new in signal processing. In fact, many well known communication systems are based on this principle. For example, in a frequency division multiplexing system the desired information is sent over disjoint frequency bands. Similarly, in a time division multiplexing system the information is sent over disjoint time intervals. In both these case the subspaces used for signalling are

orthogonal. A more recent application of subspace coding is to multi-antenna communications [59, 60], in which the subspaces used for signaling are not necessarily orthogonal. The QSP subspace measurement framework provides a unified description of these known techniques, and opens up a new realm of subspace methods by allowing for different probabilistic mappings f_M^S than the typical mapping given by (4.13), by imposing different constraints on the subspaces $\mathcal{S}_i$ used for detection which are not constrained in this framework to be equal to the subspaces used for transmission, and by choosing different projection operators onto $\mathcal{S}_i$. For example we may choose a probabilistic mapping f_M^S, resulting in a probabilistic subspace method. As another example we may constrain the detection subspaces $\mathcal{S}_i$ to be orthogonal, or we may constrain the angle between these subspaces. If the subspaces used for transmission do not satisfy this angle constraint, then we may choose detection subspaces that satisfy this constraint and are as close as possible to the transmission subspaces in some sense, thus extending the idea of LS inner product shaping to higher dimensions. These modifications parallel those we discussed in the case of ROMs. Although we do not pursue these ideas further in the thesis, they constitute an interesting area for further research.

We now consider several coding and detection techniques based on SMs, indicating the direction that some applications may take. We study a general channel model and proceed to demonstrate that the use of SMs in combination with coding techniques in which data is encoded in subspaces, is quite natural. We also undertake a very preliminary investigation of such coding techniques under certain channel models, that establishes the basic viability of these methods.

We begin with two examples that are intuitive outside the QSP framework but shown here to illustrate the main idea.

Example 4.7. Suppose that a transmitter transmits one of m signals $\{s_i(t), 1 \leq i \leq m\}$ over an unknown channel, which can be modeled as an LTI filter with unknown impulse response $h(t)$ and frequency response $H(\omega)$, followed by an additive noise source, as depicted in Fig. 4-7. The received signal $r(t)$ is given by $r(t) = s_i(t) * h(t) + n(t)$ for one value i.

If the Fourier transform $S_i(\omega)$ of $s_i(t)$ is supported on a frequency band Δ_i, then regardless of the choice of $H(\omega)$, when $s_i(t)$ is transmitted the filter output $x(t) = s_i(t) * h(t)$ is also supported on Δ_i. This suggests choosing the signals $s_i(t)$ so that $S_i(\omega) = 0$ for $\omega \notin \Delta_i$ where the frequency bands Δ_i do not overlap. The information regarding the transmitted

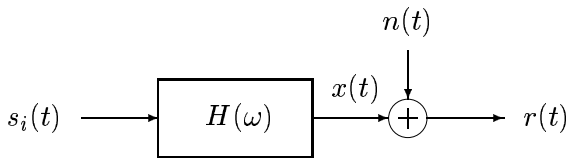


Figure 4-7: Channel model.

signal is therefore encoded in the orthogonal subspaces $\{\mathcal{S}_i, 1 \leq i \leq m\}$, where $\mathcal{S}_i$ is the subspace of signals with frequency support on Δ_i.

In the absence of noise the transmitted signal can be perfectly detected by determining the frequency support of the received signal. To detect the transmitted signal in the presence of noise, we need to determine which of the subspaces $\mathcal{S}_i$ is best matched to the received signal $r(t)$. We therefore propose detecting the transmitting signal by performing a SM corresponding to the projections $\{E_i, 1 \leq i \leq m\}$, where E_i is an orthogonal projection onto $\mathcal{S}_i$ defined by $E_i y(t) = b_i(t) * y(t)$ where the Fourier transform $B_i(\omega)$ of $b_i(t)$ is equal to 1 for $\omega \in \Delta_i$ and 0 otherwise. To complete the description of the SM we need to specify the mapping f_M^S. In analogy to the MF detector of Example 4.1, we choose

$$f_M^S(x(t)) = i \text{ where } i = \arg\max \langle x(t), E_k x(t) \rangle. \tag{4.16}$$

Then $y = M^S(r(t)) = E_i r(t)$ where $i = \arg\max \langle r(t), E_k r(t) \rangle$, and

$$\langle r(t), E_k r(t) \rangle = \frac{1}{2\pi} \int_{\Delta_k} |R(\omega)|^2 d\omega \triangleq F_k, \tag{4.17}$$

where $R(\omega)$ denotes the Fourier transform of $r(t)$. The measurement outcome y is then mapped to one of the signals $s_i(t)$ using the mapping $T_{\mathcal{Y}}$, where $T_{\mathcal{Y}}(y) = s_i(t)$ if $y \in \mathcal{S}_i$. The measurement description of the detector is depicted in Fig. 4-8. This detector is equivalent

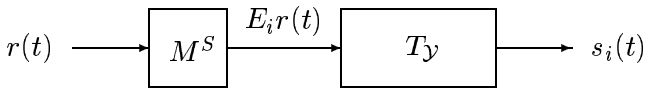


Figure 4-8: Measurement description of detector.

to the detector depicted in Fig. 4-9, in which $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, and T maps the index i to the signal $s_i(t)$.

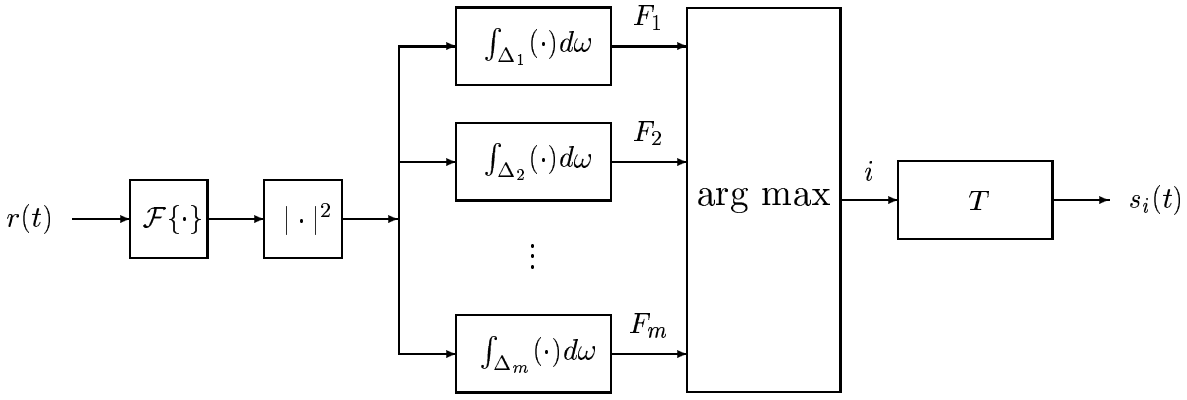


Figure 4-9: Subspace signal detector for the channel of Fig. 4-7.

If the transmitted signal is $s_i(t)$, then for $k \neq i$,

$$F_k = \frac{1}{2\pi}\int_{\Delta_k} |S_i(\omega) + N(\omega)|^2 d\omega = \int_{\Delta_k} |N(\omega)|^2 d\omega, \tag{4.18}$$

where $N(\omega)$ denotes the Fourier transform of $n(t)$, and

$$F_i = \frac{1}{2\pi}\left(\int_{\Delta_i} |S_i(\omega)|^2 d\omega + \int_{\Delta_i} |N(\omega)|^2 d\omega + 2\Re\left\{\int_{\Delta_i} S_i(\omega)N(\omega)d\omega\right\}\right). \tag{4.19}$$

Consequently if the noise has zero mean and is spread out evenly in the different frequency bands, then F_i will tend to be larger than F_k for $k \neq i$, and the subspace detector will correctly identify the transmitted signal. Therefore, this detection scheme resulting from a SM seems intuitively reasonable.

An additional justification for the SM outlined above, or equivalently, for the subspace detector of Fig. 4-9, is that if the noise $n(t)$ is a stationary white Gaussian process, then this detector implements the generalized likelihood ratio test (GLRT) [105] for detecting the signal $s_i(t)$. We prove this result in a more general setting below that does not require the signaling spaces to be orthogonal.

A possible extension of the basic subspace detector results from choosing different map-

pings f_M^S than that considered in (4.16). For example, we may choose a probabilistic mapping which leads to a probabilistic subspace detector.

Another interesting extension suggested by the QSP framework is to the case in which the frequency bands Δ_i overlap, so that the signaling subspaces are no longer orthogonal. Although we may still use the subspace detector of Fig. 4-9 to detect the transmitted signal, we may be able to improve the detection performance by projecting the received signals onto a set of orthogonal subspaces that are closest to the transmission subspaces, using the ideas of LS inner product shaping. □

In the next example we consider an application of the coding and decoding technique proposed in Example 4.7 to the problem of detecting which one of a set of known signals has been received over an additive noise channel, where the covariance of the noise is unknown.

Example 4.8. Suppose that a transmitter transmits one of m signals $\{s_i(t), 1 \leq i \leq m\}$ over an additive noise channel, so that the received signal is modeled as $r(t) = s_i(t) + n(t)$ for some index i. The noise $n(t)$ is stationary, zero mean, with unknown covariance function $R(t)$. This problem is similar to the MF problem discussed in Example 4.1, however now the noise is not assumed to be white.

If the covariance function of the noise is known, then we may first filter $r(t)$ with a whitening filter that whitens the noise component in $r(t)$. The problem then reduces to a conventional MF problem where the signals $s_i(t)$ are now replaced by the signals filtered by the whitening filter. However, if the correlation function of the noise is unknown, then the whitening filter and consequently the filtered signals are unknown, and an MF detector matched to the filtered signals cannot be designed.

To derive a coding and decoding technique in this case we exploit the ideas of Example 4.7 by modeling the received signal $r(t)$ as the output of the channel depicted in Fig. 4-7, where now $H(\omega)$ is the unknown whitening filter and $n(t)$ is a white noise process. This suggests designing the transmitted signals $s_i(t)$ to lie in different frequency bands, and then detecting the transmitted signal using the subspace detector of Fig. 4-9. □

Examples 4.7 and 4.8 demonstrate that SMs can lead to effective detectors for problems where the information regarding a signal is conveyed by a *subspace* in which the signal is known a priori to lie. This motivates coding techniques for transmitting information over an unknown channel, that is not necessarily linear or time-invariant, in which the information

is embedded in subspaces that intersect only at 0. The subspaces are chosen so that the channel can be decomposed into two components, where the subspaces are invariant to the first channel component, and are perturbed only by the second channel component. This idea is illustrated in Fig. 4-10, where the input to the channel lies in one of the subspaces $\{\mathcal{S}_i, 1 \leq i \leq m\}$. The output of the first channel component lies in the same subspace as the input, while the second component perturbs the signal out of the subspace and the output is no longer guaranteed to be in the input space. A detector based on a SM then chooses as the detected signal the one that lies in the subspace "closest" to the received signal r.

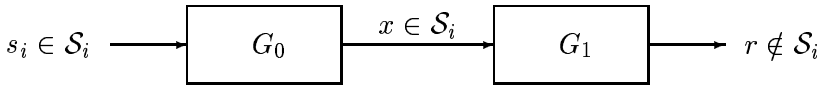


Figure 4-10: Decomposition of a channel into two components. The first channel G_0 operates within the subspace associated with the input signal. The second channel G_1 perturbs the signal out of the subspace.

In the special case depicted in Fig. 4-11, where G_1 is an additive white Gaussian noise source n and G_0 is an arbitrary not necessarily linear or time-invariant channel, we can construct a SM that implements the GLRT for detecting the transmitted signal. This provides further justification for the use of SMs.

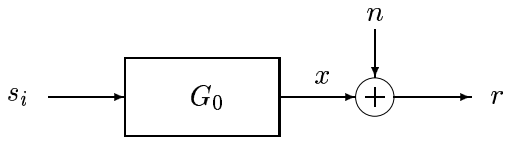


Figure 4-11: Special case of Fig. 4-10 in which G_1 is an additive white Gaussian noise source.

4.4.3 Generalized Likelihood Ratio Test for Subspace Detection

Suppose that a transmitter transmits one of m signals $\{s_i \in \mathcal{S}_i, 1 \leq i \leq m\}$ where the subspaces $\{\mathcal{S}_i \subset \mathcal{H}, 1 \leq i \leq m\}$ intersect only at 0 but are not assumed to be orthogonal, and $\mathcal{H}$ is an arbitrary Hilbert space with inner product $\langle x, y \rangle$ for any $x, y \in \mathcal{H}$. The received

signal is modeled as $r = G_0(s_i) + n$ for one value of i, where n is a white Gaussian noise vector whose components have variance σ^2, and $G_0(\cdot)$ is a channel invariant to the subspaces $\{\mathcal{S}_i\}$ so that if $s_i \in \mathcal{S}_i$ then $x = G_0(s_i) \in \mathcal{S}_i$, and otherwise unknown.

Based on the received signal r we detect the transmitted signal using a SM M^S with projections $\{E_i, 1 \leq i \leq m\}$, where E_i is an orthogonal projection onto $\mathcal{S}_i$, and

$$f_M^S(r) = i \text{ where } i = \arg\max\langle E_k r, E_k r\rangle. \tag{4.20}$$

The declared detected signal using this SM is s_i where $i = \arg\max\langle E_k r, E_k r\rangle$.

We now show that this SM implements the GLRT for detecting the transmitted signal. The GLRT [49, 32] chooses as the detected signal the signal s_i where $i = \arg\max f(r|s_k, \hat{G}_0^k)$. Here $f(r|s_k, \hat{G}_0^k)$ is the probability density function of r given s_k and $\hat{G}_0^k$, and $\hat{G}_0^k$ is the maximum likelihood (ML) estimate of G_0 given that the transmitted signal is s_k, and is chosen to maximize $f(r|s_k, G_0)$.

Since n is white and Gaussian,

$$\log f(r|s_k, G_0) = K - \frac{1}{2\sigma^2}\langle r - G_0(s_k), r - G_0(s_k)\rangle, \tag{4.21}$$

where K is a constant independent of G_0. The ML estimate $\hat{G}_0^k$ is thus chosen to minimize

$$e = \langle r - G_0(s_k), r - G_0(s_k)\rangle. \tag{4.22}$$

Expressing r as $r = r_k + r_k^\perp$ where $r_k \in \mathcal{S}_k$ and $r_k^\perp \in \mathcal{S}_k^\perp$, we may rewrite (4.22) as

$$e = \langle r_k + r_k^\perp - G_0(s_k), r_k + r_k^\perp - G_0(s_k)\rangle = \langle r_k - G_0(s_k), r_k - G_0(s_k)\rangle + \langle r_k^\perp, r_k^\perp\rangle, \tag{4.23}$$

where we used the fact that $\langle r_k^\perp, G_0(s_k)\rangle = 0$ since $G_0(s_k) \in \mathcal{S}_k$ for any choice of G_0. Thus,

$$\hat{G}_0^k = \arg\min\langle r_k - G_0(s_k), r_k - G_0(s_k)\rangle. \tag{4.24}$$

Since $r_k \in \mathcal{S}_k$ and the only restriction on G_0 is that $G_0(x) \in \mathcal{S}_k$ for any $x \in \mathcal{S}_k$, we can always choose G_0 such that $G_0(s_k) = r_k$. Therefore the ML estimate satisfies

$$\hat{G}_0^k(s_k) = r_k. \tag{4.25}$$

The detected signal is then chosen to maximize

$$\log f(r|s_k, \hat{G}_0^k) = K - \frac{1}{2\sigma^2}\langle r - \hat{G}_0^k(s_k), r - \hat{G}_0^k(s_k)\rangle = K - \frac{1}{2\sigma^2}\langle r - r_k, r - r_k\rangle, \qquad (4.26)$$

so that with E_k denoting the orthogonal projection operator onto $\mathcal{S}_k$, the detected signal is equal to s_i where $i = \arg\min\langle r - r_k, r - r_k\rangle = \arg\max\langle r_k, r_k\rangle = \arg\max\langle E_k r, E_k r\rangle$. Evidently, the detected signal is equal to the detected signal using the SM described previously.

We note that no specific model was assumed for the channel G_0 in our analysis above. In particular, G_0 is not constrained to be an LTI system as in Examples 4.7 and 4.8. Regardless of the properties of G_0, under the assumptions of the analysis, a subspace detector corresponding to the proposed SM will always implement the GLRT for detecting the transmitted signal. However, the structure of G_0 will determine the invariant subspaces and subsequently the coding strategy.

Although using orthogonal projection operators onto the invariant subspaces implements a GLRT, in many applications we may be able to improve the detection performance by using oblique projections. The use of oblique projections for detection is considered in more detail in Section 5.6.

If G_0 is an LTI system, then the invariant subspaces can be chosen as subspaces of signals with support on different frequency bands. We now consider two other choices of channel models that are of practical importance: linear memoryless (LM) channels and nonlinear memoryless (NLM) channels.

4.4.4 Subspace Detection for Linear Memoryless Channels

Suppose that G_0 in Fig. 4-11 is a linear memoryless (LM) channel. To employ the coding and detection strategy outlined in the previous section, we need to find subspaces $\{\mathcal{S}_i, 1 \leq i \leq m\}$ that are invariant to G_0. Since G_0 is a LM channel, if the input to the channel is $s_i(t)$, then $x(t)$ must have the form $x(t) = c(t)s_i(t)$ for some function $c(t)$. Consequently, if $s_i(t) = 0$ for $t \notin \Delta_i$ where Δ_i is an arbitrary time interval, then $x(t) = 0$ for $t \notin \Delta_i$. We thus conclude that the subspace of signals time-limited to a time interval Δ_i is an invariant subspace of a LM channel.

To transmit information over a LM channel we therefore design a transmitter that transmits one of m signals $\{s_i(t) \in \mathcal{S}_i, 1 \leq i \leq m\}$ where $\mathcal{S}_i$ is the subspace of signals $y(t)$

such that $y(t) = 0, t \notin \Delta_i$, and the time intervals $\{\Delta_i, 1 \leq i \leq m\}$ do not overlap. To detect the transmitted signal we perform a SM corresponding to the projections $\{E_i, 1 \leq i \leq m\}$, where E_i is the orthogonal projection onto $\mathcal{S}_i$ and is defined by

$$E_i y(t) = \begin{cases} y(t), & t \in \Delta_i; \\ 0, & \text{otherwise.} \end{cases} \tag{4.27}$$

The mapping f_M^S is chosen as in (4.20). Then $M^S(r(t)) = E_i r(t)$ where $i = \arg\max\langle r(t), E_k r(t)\rangle$, and from (4.27),

$$\langle r(t), E_k r(t)\rangle = \int_{\Delta_k} |r(t)|^2 dt \triangleq T_k. \tag{4.28}$$

The outcome y of the measurement is then mapped to one of the signals $s_i(t)$ using the mapping $T_{\mathcal{Y}}$ defined by $T_{\mathcal{Y}}(y) = s_i(t)$ in $y \in \mathcal{S}_i$. The resulting subspace detector is depicted in Fig. 4-12, where in the figure $T(i) = s_i(t)$.

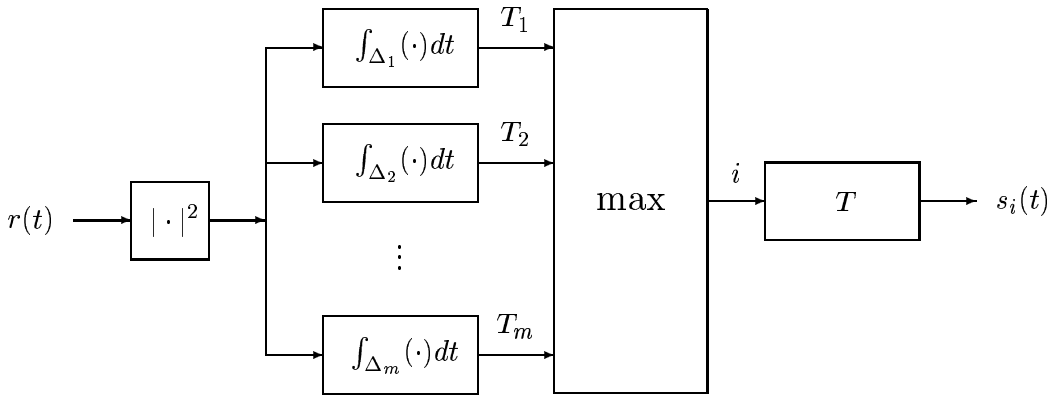


Figure 4-12: Subspace signal detector for a linear memoryless channel.

4.4.5 Subspace Detection for Nonlinear Memoryless Channels

We now consider the case in which G_0 in Fig. 4-11 is a nonlinear memoryless (NLM) channel where the inputs and outputs of the channel lie in the space $\mathcal{H}$ of signals that are zero outside the interval $[0, T]$, with inner product $\langle x(t), y(t)\rangle = \int_{t=0}^{T} x(t)y(t)dt$.

We can immediately verify that in this case the subspaces $\mathcal{S}_i$ of all signals in $\mathcal{H}$ that

are arbitrary on a time interval Δ_i and constant outside this interval, are invariant to G_0. Based on these invariant subspaces, we can design a coding and decoding scheme using the strategy outlined in the beginning of the section. Note, that in this case the invariant subspaces are not orthogonal.

To implement a subspace detector corresponding to a SM we need to determine the orthogonal projection E_i onto $\mathcal{S}_i$. To this end we note that each subspace $\mathcal{S}_i$ can be expressed as a direct sum $\mathcal{S}_i = \mathcal{V}_i \oplus \mathcal{G}_i$ where $\mathcal{V}_i$ is the subspace of signals in $\mathcal{H}$ that are zero outside the interval Δ_i and $\mathcal{G}_i$ is the space of signals in $\mathcal{H}$ that are zero on the interval Δ_i and constant outside this interval. Since $\mathcal{V}_i$ and $\mathcal{G}_i$ are orthogonal, $\langle r(t), E_k r(t)\rangle = \langle r(t), E_k^{\mathcal{V}} r(t)\rangle + \langle r(t), E_k^{\mathcal{G}} r(t)\rangle$ where $E_k^{\mathcal{V}}$ is the orthogonal projection onto $\mathcal{V}_k$ and is given by (4.27), and $E_k^{\mathcal{G}}$ is the orthogonal projection onto $\mathcal{G}_k$, and can be readily derived as

$$E_k^{\mathcal{G}} r(t) = \begin{cases} \frac{1}{T-t_k} \int_{t=0, t\notin\Delta_k}^{T} r(t)dt & t \notin \Delta_k; \\ 0, & t \in \Delta_k, \end{cases} \tag{4.29}$$

where $t_k = \int_{t\in\Delta_k} dt$. The declared detected signal using the resulting subspace detector is $s_i(t)$ where $i = \arg\max R_k$ and

$$R_k = \left| \frac{1}{T-t_k} \int_{t=0, t\notin\Delta_k}^{T} r(t)dt \right|^2 + \int_{t\in\Delta_k} |r(t)|^2 dt. \tag{4.30}$$

In summary, subspace measurements lead to interesting and potentially useful coding and detection methods for communication-based applications over a variety of channel models. However, this section represents a rather preliminary exploration of such techniques and there are several aspects that require further study and evaluation, providing some interesting directions for further research.

4.4.6 Successive Subspace Measurements

Subspace measurements can be performed in sequence to identify smaller and smaller subspaces that are best matched to a signal. For example, we may start by performing a SM corresponding to subspaces $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$. If the outcome lies in $\mathcal{S}_i$, then we measure the projected signal using a SM corresponding to subspaces $\{\mathcal{S}_{i1}, \ldots, \mathcal{S}_{in}\}$ of $\mathcal{S}_i$, and continue recursively. A similar idea is considered in the next example.

Example 4.9 (Adaptive wavelet tree). In this example we use repeated SMs to create a deterministic or probabilistic adaptive wavelet tree, which is "best" fitted to a given signal.

A wavelet tree is used to decompose a signal into a wavelet basis. A standard wavelet tree has the form of Fig. 4-13, where each branch in the tree is depicted in Fig. 4-14. Here $H_0(\omega)$ denotes the the frequency response of a lowpass filter with impulse response $h_0[n]$,

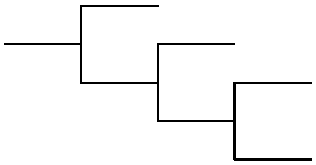


Figure 4-13: Wavelet tree.

and $H_1(\omega)$ denotes the frequency response of a highpass filter with impulse response $h_1[n]$. The coefficients of the wavelet decomposition are given by the nodes of the wavelet tree

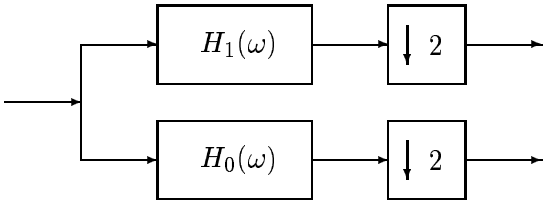


Figure 4-14: Branch in a wavelet tree.

shown in Fig. 4-13.

For some signals it may be more efficient to choose coefficients adaptively, depending on the signal properties. Instead of using a fixed tree, we may use an adaptive tree [106, 107] in which at each node we decide whether to choose the highpass or lowpass branch, based on some criterion. For example, the criterion may be the energy of the signal at each node in the highpass and lowpass regions. A possible tree resulting from such an algorithm is depicted in Fig. 4-15.

We now implement an adaptive decomposition using successive SMs. Let $x_k[n]$ denote the signal at the kth node, where $x_0[n]$ is the input to the tree, and let $X_k(\omega)$ denote the Fourier transform of $x_k[n]$. We define E_0 and E_1 as projections onto $\mathcal{S}_0$ and $\mathcal{S}_1$, where $\mathcal{S}_i$

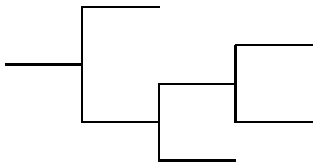


Figure 4-15: Possible wavelet decomposition resulting from an adaptive algorithm.

is the subspace of signals with frequency support on Δ_i. Here Δ_0 and Δ_1 are the effective frequency bands corresponding to the lowpass and highpass filters respectively. We then construct a SM with projections E_0 and E_1, where the properties of the tree will depend on the choice of projections and the choice of mapping $f_M^S(x_k[n])$. For example, we may choose a deterministic mapping

$$f_M^S(x_k[n]) = E_i \text{ where } i = \arg\max\langle x_k[n], E_l x_k[n]\rangle, \tag{4.31}$$

with

$$\langle x_k[n], E_l x_k[n]\rangle = \frac{1}{2\pi}\int_{\Delta_l} |X_k(\omega)|^2 d\omega \triangleq F_k^l, \quad l = 1, 2. \tag{4.32}$$

The resulting SM can be implemented by computing F_k^0 and F_k^1 at each node in the tree, and choosing the branch corresponding to the larger value. The signal is then filtered with the appropriate filter, and downsampled by 2, and the measurement is repeated on the filtered signal.

We may also choose a probabilistic mapping such as

$$f_M^S(x_k[n]) = E_i \text{ with probability } c\langle x_k[n], E_i x_k[n]\rangle, \tag{4.33}$$

where c is a normalization constant. In this case at each node we choose a branch probabilistically, with probabilities proportional to $F_k^i, i = 1, 2$. □

In Example 4.9 we repeated the same SM at each stage. We may also perform different SMs at each stage, that could depend on the output of the previous stage. In the example above this corresponds to using different filters at each node.

In this section we introduced the idea of combining measurements. Here, we focused on

successive SMs. In the next chapter we turn our attention to combined measurements in which two measurements are performed in sequence, where each measurement is either a ROM or a SSM.

Chapter 5

Combined Measurements

An interesting and important class of measurements in quantum mechanics results from restricting measurements to a subspace in which a quantum system is known a priori to lie. This leads to the notion of positive operator-valued measures (POVMs) [62, 61] which, as we discussed in Chapter 3, can be realized by a combination of a standard measurement followed by an orthogonal projection onto a lower space. Therefore, central to the concept of a POVM is the notion of applying measurements successively. Since a combined measurement is sometimes a more efficient way of obtaining information about the state of a quantum system than a standard measurement, such measurements have enjoyed widespread use.

The QSP analogue of a rank-one POVM is a rank-one measurement (ROM) followed by a simple subspace measurement (SSM) *i.e.,* a projection. We begin this chapter by exploring applications of combined QSP measurements of this form. We show that such measurements lead to a variety of extensions and insights into frames, which are generalizations of bases that result in redundant signal expansions [63, 64]. In particular, the combined measurement framework offers an alternative perspective on frames in terms of projections of vectors in a larger space. This viewpoint provides a convenient setting for developing generalizations and extensions of frames, for example, by changing the properties of the vectors in the larger space or by exploring the effect of oblique projections, leading to new classes of frames and to the concept of oblique frame expansions. This framework also offers rich insights into conventional frame expansions that result from exploiting the connection between combined QSP measurements and quantum POVMs. Based on this relationship, we develop frame-theoretic analogues of various quantum-mechanical concepts

and results. In particular, motivated by the construction of optimal quantum measurements [26], we consider the problem of optimal frame design. These applications are considered in more detail in subsequent chapters of the thesis, and in the references [68, 76, 19].

Since the QSP framework does not depend on the physics associated with quantum mechanics, we may extend the notion of a POVM to include other forms of combined QSP measurements, where we perform any two measurements successively. There are a number of interesting potential applications of these more general forms of combined QSP measurements. In the later part of the chapter we consider some specific examples as an indication of the direction that some applications may take. In particular we show that this framework leads naturally to subspace MF detectors and to certain classes of randomized algorithms. We remark at the outset that these examples are highly preliminary and require further study and evaluation. The primary purpose is to identify some concrete applications and discuss some of their merits, which may then suggest future directions to explore.

5.1 Classes of Combined Measurements

In our development, we restrict our attention to combinations of ROMs and SSMs. Therefore let M_1 and M_2 denote QSP measurements that can each be a ROM or a SSM, and suppose that we perform the measurements successively. Then the combined measurement M_{21} is defined by $M_{21}(x) = M_2(M_1(x))$ for any $x \in \mathcal{H}$. We distinguish between 4 classes of combined measurements:

1. M_1 is a ROM and M_2 is a SSM;
2. M_1 is a SSM and M_2 is a ROM;
3. M_1 and M_2 are both ROMs;
4. M_1 and M_2 are both SSMs.

We now briefly consider each of the cases above pointing our their primary applications, which are then explored in more detail in the ensuing sections.

1. If M_2 is a SSM then $M_2 = E$ for some projection E, and the effect of M_2 is to project the outcome of M_1 onto the range $\mathcal{R}(E)$ of E. In this case, the possible outputs of the combined measurement, denoted by M_{E1}, are proportional to the projections of

the measurement vectors of M_1, which we call the *effective measurement vectors*. As we show, these vectors have the property that they form a *frame* for $\mathcal{R}(E)$. Imposing inner product constraints on the measurement vectors of M_1 leads to a variety of interesting insights into existing frame expansions as well as new classes of frames that are explored in Section 5.3 and in [68, 76]. If E is an orthogonal projection and we only measure vectors in $\mathcal{R}(E)$, then the combined measurement is fully characterized by the effective measurement vectors. If, on the other hand, E is an oblique projection, then an additional set of vectors is needed to describe the combined measurement. This new set of vectors forms a frame for $\mathcal{N}(E)^{\perp}$, that is intimately related to the frame for $\mathcal{R}(E)$, and leads to the definition of the *oblique dual frame vectors*. Oblique frame expansions are studied in Section 5.4 and in [19], and are subsequently used in Chapter 6 to develop redundant consistent sampling algorithms.

2. If M_1 is a SSM and M_2 is a ROM, then the combined measurement of a signal is equal to the ROM of the projected signal. As we discuss in Section 5.6, combined measurements of this form are useful in applications in which we may benefit from processing only certain components of a signal, *e.g.,* when a signal is contaminated by subspace noise. Based on these combined measurements we develop subspace MF detectors for detecting a signal contaminated by both structured noise and white noise, and illustrate that these detectors can lead to improved performance over conventional GLRT based detectors.

3. If M_1 and M_2 are both ROMs, then the effect of M_2 is to postprocess the outcome of M_1 in a deterministic or probabilistic fashion. Thus, combined measurements of this form can be used to generate new algorithms by processing the outputs of existing algorithms. In particular, these measurements lead to randomized algorithms in which an algorithm is modified so that the original output is used to generate a probability distribution on the final output. As we illustrate in Section 5.7, randomized algorithms of this type have the effect of improving worst case performance.

4. If M_1 and M_2 are both SSMs, then the combined measurement is equivalent to a linear operator which is equal to the concatenation of the two projections.

In the remainder of this chapter we study the first three classes of combined measurements in more detail.

5.2 Frames and Combined Measurements

The first class of combined measurements we consider is the QSP analogue of the quantum POVM, which consists of a ROM M_1 with measurement vectors $\{q_i \in \mathcal{H}, 1 \leq i \leq m\}$, followed by a SSM $M_2 = E$, where E is a projection onto a subspace $\mathcal{U} \subseteq \mathcal{H}$. Since in the QSP framework we are not constrained by physical laws, the measurement vectors q_i are not constrained to be orthogonal and the projection E is not constrained to be an orthogonal projection, as in a quantum POVM. We assume that the vectors $\{q_i\}$ span a subspace $\mathcal{W} \subseteq \mathcal{H}$ such that $\mathcal{U} \subseteq \mathcal{W}$, and that we only measure signals in $\mathcal{U}$. Then, like its quantum analogue, the combined measurement can be viewed as a restriction onto $\mathcal{U}$ of a measurement on $\mathcal{H}$.

Combined measurements of this form lead to a variety of extensions and insights into frame expansions, which are developed in this section and in Sections 5.3 and 5.4. In particular, these measurements lead to a new viewpoint towards frames, which we now discuss.

5.2.1 Alternative Perspective on Frames

To develop the relationship between combined measurements and frames, we note that since $M_1(x)$ is proportional to q_i for some i, the outcome of the combined measurement $M_{E1}(x)$ where $M_E = E$ is proportional to one of the vectors $\{Eq_i \in \mathcal{U}, 1 \leq i \leq m\}$, which we call the *effective measurement vectors* of M_{E1}. In Section 5.2.3 we show that these vectors form a *frame* for $\mathcal{U}$, which will be defined in Section 5.2.2. The properties of the frame depend on the properties of the vectors q_i as well as on the properties of the projection E. We also show that any frame for $\mathcal{U}$ can be viewed as the effective measurement vectors of a combined measurement or, equivalently, that any frame for $\mathcal{U}$ can be viewed as projections of a set of vectors in a larger space containing $\mathcal{U}$.

Typically in the literature frames are defined in terms of their properties in $\mathcal{U}$. The combined measurement framework offers a different perspective on frames in terms of projections of vectors in a larger space. This viewpoint leads to some rich insights into frames as well as a convenient framework for developing extensions of frame expansions. Specifically, by choosing vectors q_i with different inner product constraints, and choosing different projection operators E, a variety of new classes of frames can be developed.

In the context of a single QSP measurement, we have seen that imposing inner product constraints on the measurement vectors of a ROM leads to new, effective processing techniques. Similarly, as we show in Section 5.3, in the context of combined measurements imposing such constraints leads to interesting classes of frames as well as new insights into existing frames. In Section 5.3.1 we consider the case in which M_1 has orthonormal measurement vectors and $E = P_{\mathcal{U}}$ is an orthogonal projection onto $\mathcal{U}$. The corresponding effective measurement vectors are shown to form a *tight frame* for $\mathcal{U}$. Building upon the connection between combined QSP measurements and quantum POVMs, we show that tight frames and rank-one quantum POVMs are intimately related. Using this relationship, we develop frame-theoretical analogues of various quantum-mechanical concepts and results. In Section 5.3.2 we explore the class of *geometrically uniform (GU) frames* that results from a combined measurement M_{E1}, where $M_E = P_{\mathcal{U}}$ and where we impose certain inner product constraints on the measurement vectors of M_1.

Extensions of frames that result from choosing $E = E_{\mathcal{US}}$ as an oblique projection onto $\mathcal{U}$ along $\mathcal{S}$, where $\mathcal{S} \subseteq \mathcal{H}$ is disjoint from $\mathcal{U}$ so that $\mathcal{U} \cap \mathcal{S} = \{0\}$, are explored in Section 5.4. These measurements lead to the new concept of *oblique frame expansions* in which, contrary to conventional frame expansions, the analysis and synthesis vectors do not lie in the same space.

Before proceeding to the detailed development, in the next section we provide a brief introduction to frame expansions.

5.2.2 Frames

Frames are generalizations of bases which lead to redundant signal expansions [63, 64]. A frame for a Hilbert space $\mathcal{U}$ is a set of not necessarily linearly independent vectors that spans $\mathcal{U}$ and has some additional properties. Frames were first introduced by Duffin and Schaeffer [63] in the context of nonharmonic Fourier series, and play an important role in the theory of nonuniform sampling [63, 64, 65]. Recent interest in frames has been motivated in part by their utility in analyzing wavelet expansions [66, 67].

Many efforts have been made to construct bases with specified properties. Since the conditions on bases are quite stringent, in many applications it is hard to find "good" bases. The conditions on frame vectors are usually not as stringent, allowing for increased flexibility in their design [66, 85]. For example, frame expansions admit signal representations that

are localized in both time and frequency [67], as well as sparse representations [108].

Frame expansions have many desirable properties. The coefficients may be computed with less precision than the coefficients in a basis expansion for a given desired reconstruction precision [67]; the effect of additive noise on the coefficients on the reconstructed signal is reduced in comparison with a basis expansion [67, 69, 109, 19]; and the coefficients are more robust to quantization degradations [110, 111]. Recently, frames have been applied to the development of modern uniform and nonuniform sampling techniques [112], to various detection problems [37, 39], and to multiple description source coding [80].

Let $\{\varphi_i, 1 \leq i \leq m\}$ denote a set of m vectors in $\mathcal{H}$. The vectors φ_i form a *frame* for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ if there exist constants $\alpha > 0$ and $\beta < \infty$ such that

$$\alpha^2 \|x\|^2 \leq \sum_{i=1}^{m} |\langle x, \varphi_i \rangle|^2 \leq \beta^2 \|x\|^2, \tag{5.1}$$

for all $x \in \mathcal{U}$ [69]. In this chapter, we restrict our attention to the case where m and n are finite. The lower bound ensures that the vectors φ_i span $\mathcal{U}$; thus we must have $m \geq n$. If $m < \infty$, then the right hand inequality is always satisfied with $\beta^2 = \sum_{i=1}^{m} \langle \varphi_i, \varphi_i \rangle$. Thus, any finite set of vectors that spans $\mathcal{U}$ is a frame for $\mathcal{U}$. In particular, any basis for $\mathcal{U}$ is a frame for $\mathcal{U}$. However in contrast to basis vectors, which are linearly independent, frame vectors with $m > n$ are linearly dependent. If the bounds $\alpha = \beta$ in (5.1), then the frame is called a tight frame. If in addition $\alpha = \beta = 1$, then we call the frame a normalized tight frame; otherwise it is said to be *β-scaled* [68]. The redundancy of the frame is defined as $r = m/n$, *i.e.,* m vectors in a n-dimensional space.

The *frame operator* corresponding to frame vectors $\{\varphi_i, 1 \leq i \leq m\}$ is defined as [69]

$$S = \sum_{i=1}^{m} \varphi_i \varphi_i^*. \tag{5.2}$$

Using the frame operator, (5.1) can be rewritten as

$$\alpha^2 \|x\|^2 \leq \langle Sx, x \rangle \leq \beta^2 \|x\|^2. \tag{5.3}$$

From (5.3) it follows that the tightest possible frame bounds α^2 and β^2 are given by $\alpha^2 = \min_i \lambda_i(S)$ and $\beta^2 = \max_i \lambda_i(S)$, where $\{\lambda_i(S) > 0, 1 \leq i \leq n\}$ are the n positive eigenvalues of the frame operator S.

If the vectors $\{\varphi_i, 1 \leq i \leq m\}$ form a frame for $\mathcal{U}$, then any $x \in \mathcal{U}$ can be expressed as a linear combination of these vectors: $x = \sum_i a_i \varphi_i$. When $m > n$, the coefficients in this expansion are not unique. A possible choice is $a_i = \langle \bar{\varphi}_i, x \rangle$ where $\bar{\varphi}_i$ are the *dual frame vectors* [69] of the frame vectors φ_i, and are given by

$$\bar{\varphi}_i = S^\dagger \varphi_i, \tag{5.4}$$

where S is the frame operator defined by (5.2), and $(\cdot)^\dagger$ denotes the pseudoinverse (see Chapter 2). This choice of coefficients has the property that among all possible coefficients it has the minimal l_2-norm [69, 113]. With $\overline{F}$ and F denoting the set transformations corresponding to the vectors $\bar{\varphi}_i$ and φ_i respectively, we have that $\overline{F} = (F^\dagger)^*$. The transformation $\overline{F}$ is referred to as the dual frame operator [69].

Since for any $x \in \mathcal{U}$,

$$\sum_{i=1}^{m} |\langle x, \varphi_i \rangle|^2 = \sum_{i=1}^{m} x^* \varphi_i \varphi_i^* x = \langle x, \left(\sum_i \varphi_i \varphi_i^* \right) x \rangle, \tag{5.5}$$

for β-scaled tight frames,

$$\sum_{i=1}^{m} \varphi_i \varphi_i^* = \beta^2 P_{\mathcal{U}}. \tag{5.6}$$

Conversely, if the vectors $\varphi_i \in \mathcal{H}$ satisfy (5.6), then (5.5) implies that (5.1) is satisfied with $\alpha = \beta$ for all $x \in \mathcal{U}$. We conclude that a set of m vectors $\varphi_i \in \mathcal{H}$ forms a tight frame for a subspace $\mathcal{U} \subseteq \mathcal{H}$ if and only if the vectors satisfy (5.6) for some $\beta > 0$. In this case the dual frame vectors are $\bar{\varphi}_i = (1/\beta^2)\varphi_i$. Tight frames are very convenient analytically and possess very nice numerical properties [69], and are therefore particularly popular.

5.2.3 Effective Measurement Vectors and Frames

We now establish the relationship between combined measurements and frames, which is the basis for the developments in subsequent sections. The following proposition follows immediately from [85, Theorem 2].

Proposition 5.1. *Let $\{q_i \in \mathcal{H}, 1 \leq i \leq m\}$ be the measurement vectors of a ROM M_1 and let $\mathcal{W} \subseteq \mathcal{H}$ denote the space spanned by the vectors q_i. Let $\{v_i = E_{\mathcal{US}} q_i, 1 \leq i \leq m\}$ be*

the effective measurement vectors of a combined measurement M_{E1} where $M_E = E_{\mathcal{US}}$ is a projection onto $\mathcal{U}$ along $\mathcal{S}$. Here $\mathcal{U}$ is a subset of $\mathcal{W}$ and $\mathcal{S}$ is an arbitrary subspace of $\mathcal{H}$ such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$. Then the vectors $\{v_i, 1 \leq i \leq m\}$ form a frame for $\mathcal{U}$.

Thus, a set of vectors in $\mathcal{U}$ can be the effective measurement vectors of a combined measurement with measurement vectors spanning a subspace $\mathcal{W} \supseteq \mathcal{U}$ only if they form a frame for $\mathcal{U}$. Theorem 5.1 below asserts that this condition is also sufficient.

Theorem 5.1. *Let $\{v_i, 1 \leq i \leq m\}$ be a frame for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$. Then there exists a set of linearly independent vectors $\{q_i, 1 \leq i \leq m\}$ that span an expanded m-dimensional subspace $\mathcal{W} \supseteq \mathcal{U}$ in a possibly expanded Hilbert space $\widetilde{\mathcal{H}} \supseteq \mathcal{H}$ such that $\{v_i = E_{\mathcal{US}} q_i, 1 \leq i \leq m\}$, where $E_{\mathcal{US}}$ is a projection onto $\mathcal{U}$ along $\mathcal{S}$, and $\mathcal{S}$ is an arbitrary subspace of $\mathcal{H}$ such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$. Consequently, the vectors $\{v_i\}$ may be viewed as the effective measurement vectors of a combined measurement M_{E1}, where $M_E = E_{\mathcal{US}}$ and M_1 is a ROM with measurement vectors $\{q_i\}$.*

Proof: Let V be the set transformation corresponding to the vectors v_i. Since the vectors v_i form a frame for the n-dimensional subspace $\mathcal{U}$, V has rank n and can be expressed using the SVD as $V = \sum_{i=1}^{n} \sigma_i u_i \mathbf{z}_i^*$ where the vectors $\{u_i \in \mathcal{H}, 1 \leq i \leq n\}$ form an orthonormal basis for $\mathcal{U}$, the vectors $\{\mathbf{z}_i \in \mathbb{C}^m, 1 \leq i \leq n\}$ are orthonormal, and $\{\sigma_i > 0, 1 \leq i \leq n\}$.

We distinguish between the case $k = \dim\ \mathcal{H} \geq m$, and the case $k < m$.

In the case $k \geq m$, let $\{s_i, 1 \leq i \leq m - n\}$ denote $m - n$ linearly independent vectors in $\mathcal{S}$ and define the vectors x_i such that $x_i = u_i, 1 \leq i \leq n$ and $x_{i+n} = s_i, 1 \leq i \leq m - n$. Then the vectors x_i are linearly independent and satisfy

$$E_{\mathcal{US}} x_i = \begin{cases} u_i, & 1 \leq i \leq n; \\ 0, & n+1 \leq i \leq m, \end{cases} \tag{5.7}$$

since $E_{\mathcal{US}} u = u$ for any $u \in \mathcal{U}$ and $E_{\mathcal{US}} s = 0$ for any $s \in \mathcal{S}$. Next, extend the vectors $\{\mathbf{z}_i, 1 \leq i \leq n\}$ to an orthonormal set $\{\mathbf{z}_i, 1 \leq i \leq m\}$ and define $Q = \sum_{i=1}^{m} \sigma_i x_i \mathbf{z}_i^*$, where $\sigma_i, n+1 \leq i \leq m$ are arbitrary positive numbers. Let $\{q_i, 1 \leq i \leq m\}$ be the vectors corresponding to Q. Then $\mathcal{W} \subseteq \mathcal{H}$ is the m-dimensional subspace spanned by

$\{x_i, 1 \leq i \leq m\}$, and using (5.7), the projection of Q onto $\mathcal{U}$ is

$$E_{\mathcal{US}} Q = \sum_{i=1}^{m} \sigma_i E_{\mathcal{US}} x_i \mathbf{z}_i^* = \sum_{i=1}^{n} \sigma_i u_i \mathbf{z}_i^* = V, \tag{5.8}$$

so that $v_i = E_{\mathcal{US}} q_i$. Moreover, the vectors q_i are linearly independent since Q has rank m.

In the case $k < m$, first embed $\mathcal{U}$ and $\mathcal{S}$ in an expanded Hilbert space $\widetilde{\mathcal{H}} \supseteq \mathcal{H}$, and then proceed as before. □

Combining Proposition 5.1 with Theorem 5.1 we can conclude that a set of vectors $\{v_i, 1 \leq i \leq m\}$ in $\mathcal{U}$ may be the effective measurement vectors of a combined measurement M_{E1}, where $M_E = E$ is an arbitrary projection onto $\mathcal{U}$ and M_1 is a ROM with measurement vectors that span a subspace $\mathcal{W} \supseteq \mathcal{U}$, if and only if they form a frame for $\mathcal{U}$. This relationship between combined measurements and frames suggests an alternative definition of frames in terms of projections of a set of vectors in a larger space. Specifically, from Proposition 5.1 and Theorem 5.1 we conclude that:

Theorem 5.2 (Frames). *A set of vectors $\{\varphi_i \in \mathcal{H}, 1 \leq i \leq m\}$ forms a frame for $\mathcal{U}$ if and only if there exists a set of linearly independent vectors $\{\tilde{\varphi}_i, 1 \leq i \leq m\}$ that span an expanded m-dimensional subspace $\mathcal{W} \supseteq \mathcal{U}$ in a possibly expanded Hilbert space $\widetilde{\mathcal{H}} \supseteq \mathcal{H}$ such that $\{\varphi_i = E_{\mathcal{US}} \tilde{\varphi}_i, 1 \leq i \leq m\}$ where $E_{\mathcal{US}}$ is a projection onto $\mathcal{U}$ along $\mathcal{S}$, and $\mathcal{S}$ is an arbitrary subspace of $\mathcal{H}$ such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$. Moreover, the vectors $\tilde{\varphi}_i$ can always be chosen to be linearly independent.*

This perspective on frames provides additional insights into frames, and suggests a systematic approach for generating new classes of frames by changing the properties of the vectors in the larger space or changing the properties of the projection.

In Section 5.3 we consider frames that result from choosing $E = P_{\mathcal{U}}$ as an orthogonal projection operator onto $\mathcal{U}$ and imposing various inner product constraints on the measurement vectors of M_1. Extensions of frames resulting from using oblique projections are considered in Section 5.4.

5.3 ROM Followed by an Orthogonal Projection Operator

In this section we consider combined measurements M_{E1} where M_1 is a ROM with measurement vectors q_i that span a subspace $\mathcal{W}$, and $E = P_{\mathcal{U}}$ where $\mathcal{U} \subseteq \mathcal{W}$. If we assume

that we only measure signals $x \in \mathcal{U}$, then $M_{E1}(x)$ depends only on the effective measurement vectors $v_i = P_{\mathcal{U}} q_i$. Indeed, for any $x \in \mathcal{U}$, $M_1(x) = cq_i$ for some $c \in \mathbb{C}$ and some value i, where i depends on the input x only through the inner products $\langle q_i, x \rangle$, and $M_{E1}(x) = P_{\mathcal{U}} M_1(x) = cP_{\mathcal{U}} q_i = cv_i$. Since $x \in \mathcal{U}$, $x = P_{\mathcal{U}} x$ and

$$\langle q_i, x \rangle = \langle q_i, P_{\mathcal{U}} x \rangle = \langle P_{\mathcal{U}} q_i, x \rangle = \langle v_i, x \rangle. \tag{5.9}$$

Thus for any $x \in \mathcal{U}$, $M_{E1}(x)$ is proportional to one of the vectors v_i, where the particular outcome depends on the input x only through the inner products $\langle v_i, x \rangle$.

We have seen in Theorem 5.1 that the vectors v_i form a frame for $\mathcal{U}$. In the remainder of this section we consider the frame properties when different inner product constraints are imposed on the measurement vectors q_i.

5.3.1 Tight Frames

Suppose now that the measurement vectors $\{q_i, 1 \leq i \leq m\}$ are orthonormal and span a subspace $\mathcal{W}$ where $\mathcal{U} \subseteq \mathcal{W}$, and let $\{v_i = P_{\mathcal{U}} q_i, 1 \leq i \leq m\}$ denote the effective measurement vectors of M_{E1} with $M_E = P_{\mathcal{U}}$. Then for any $x \in \mathcal{U}$,

$$\|x\|^2 = \sum_{i=1}^{m} |\langle q_i, x \rangle|^2 = \sum_{i=1}^{m} |\langle q_i, P_{\mathcal{U}} x \rangle|^2 = \sum_{i=1}^{m} |\langle v_i, x \rangle|^2, \tag{5.10}$$

so that the vectors $\{v_i\}$ form a normalized tight frame for $\mathcal{U}$.

Since a ROM with orthonormal measurement vectors followed by an orthogonal projection corresponds to a quantum POVM, this connection between the effective measurement vectors and tight frames suggests that there is a relationship between quantum measurements and frames. Indeed, Theorem 5.3 below asserts that the family of normalized tight frames for a subspace $\mathcal{U}$ in which a quantum mechanical system is known to lie is precisely the family of POVMs on $\mathcal{U}$. Exploiting this equivalence, we can apply ideas and results derived in the context of quantum measurement to the theory of frames and *vice versa*.

Specifically, comparing (5.6) with the definition of a POVM (3.8), we conclude that:

Theorem 5.3 (Tight frames). *A set of vectors $\varphi_i \in \mathcal{H}$ forms a β-scaled tight frame for $\mathcal{U}$ if and only if the scaled vectors $\beta^{-1}\varphi_i$ are the measurement vectors of a rank-one POVM on $\mathcal{U}$. In particular, the vectors φ_i form a normalized tight frame for $\mathcal{U}$ if and only if they*

are the measurement vectors of a rank-one POVM on $\mathcal{U}$.

This fundamental relationship between rank-one quantum measurements and tight frames can be used to develop frame analogues of various results in quantum measurement [68]. In particular, we define frame transformations in analogy to the measurement matrices of quantum mechanics. We then use Neumark's theorem to extend tight frames to orthogonal bases. Finally, motivated by the least-squares measurement of quantum mechanics [26], we address the problem of constructing optimal tight frames.

Frame transformations

In analogy to the measurement matrix described in Section 3.3, we define the *frame transformation* F as the set transformation corresponding to the vectors φ_i, where the vectors φ_i form a tight frame for $\mathcal{U}$. From (5.6) it then follows that

$$FF^* = \beta^2 P_{\mathcal{U}}. \tag{5.11}$$

The properties of F follow immediately from Theorem 5.3 and Theorem 3.1:

Theorem 5.4 (Frame Transformations). *For a set transformation F corresponding to m vectors in a Hilbert space $\mathcal{H}$ and for $\beta > 0$, the following statements are equivalent:*

1. *F is the frame transformation of a β-scaled tight frame for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$;*
2. *$\beta^{-1}F$ is a transjector between n-dimensional subspaces $\mathcal{U} \subseteq \mathcal{H}$ and $\mathcal{V} \subseteq \mathbb{C}^m$;*
3. *$FF^* = \beta^2 P_{\mathcal{U}}$ for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$;*
4. *$F^*F = \beta^2 P_{\mathcal{V}}$ for an n-dimensional subspace $\mathcal{V} \subseteq \mathbb{C}^m$.*

A frame transformation F of a β-scaled tight frame for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ may be expressed as $F = \beta U \mathbf{Z}^* = \beta \sum_{i=1}^{n} u_i \mathbf{z}_i$, where U is a set transformation corresponding to n vectors $\{u_i, 1 \leq i \leq n\}$ that form an orthonormal basis for $\mathcal{U}$, and $\mathbf{Z}$ is an $m \times n$ matrix whose columns $\{\mathbf{z}_i, 1 \leq i \leq n\}$ form an orthonormal basis for $\mathcal{V}$.
A frame transformation F of a β-scaled tight frame is an isometry if restricted to $\mathcal{V}$ and scaled by β^{-1}.

A frame transformation F of a β-scaled tight frame represents an orthogonal basis for $\mathcal{U}$ (i.e., is an orthogonal frame transformation*) if and only if its rank is m. Then $F^*F = \beta^2 I_m$; i.e., all frame vectors have squared norm β^2.*

Neumark's theorem and construction of tight frames

Neumark's theorem (Theorem 3.2) was derived based on the properties of measurement matrices. Since by Theorem 5.3 frame transformations of tight frames have essentially the same properties as measurement matrices of rank-one POVMs, we can now obtain an equivalent of Neumark's theorem for tight frames. The proof is essentially the same as the proof of Theorem 3.2, so we omit it.

Theorem 5.5 (Neumark's theorem for tight frames). *Let F be a rank-n frame transformation, corresponding to m vectors in a Hilbert space $\mathcal{H}$ that span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$. Then there exists an orthogonal frame transformation $\widetilde{F}$ corresponding to equal-norm orthogonal vectors that span an expanded m-dimensional subspace $\widetilde{\mathcal{U}} \supseteq \mathcal{U}$ in a possibly expanded Hilbert space $\widetilde{\mathcal{H}} \supseteq \mathcal{H}$ such that the projection $P_{\mathcal{U}}\widetilde{F}$ of $\widetilde{F}$ onto $\mathcal{U}$ is F.*

We remark that given a set of equal-norm orthogonal vectors in $\widetilde{\mathcal{U}} \supseteq \mathcal{U}$, their projections onto $\mathcal{U}$ will always form a tight frame for $\mathcal{U}$ [85]. Combining this result with Theorem 5.5, we can conclude that a set of vectors forms a tight frame for $\mathcal{U}$ if and only if the vectors can be expressed as a projection onto $\mathcal{U}$ of a set of orthogonal vectors with equal norm in a larger space $\widetilde{\mathcal{U}}$ containing $\mathcal{U}$. This then implies that a set of vectors $\{v_i \in \mathcal{U}\}$ can be the effective measurement vectors of a combined measurement M_{E1} where $M_E = P_{\mathcal{U}}$ and M_1 is a ROM with measurement vectors $\{q_i\}$ that are orthogonal and have equal norm, if and only if they form a tight frame for $\mathcal{U}$.

Starting with a given frame transformation F in $\mathcal{U}$, the proof of Theorem 3.2 gives a concrete construction of an orthogonal frame transformation $\widetilde{F}$ in $\widetilde{\mathcal{U}} \supseteq \mathcal{U}$ such that $P_{\mathcal{U}}\widetilde{F} = F$. We now give two examples of this construction in which $\mathcal{H} = \mathbb{C}^k$ for some k. (These examples were also given in [68].) We consider first an example in which $k < m$, and then one in which $k > m$.

Example 5.1. Consider the four frame vectors $\varphi_1 = [0.35 \;\; -0.61]^*$, $\varphi_2 = [0.61 \;\; 0.35]^*$,

$\varphi_3 = [0.5 \ -0.5]^*$, and $\varphi_4 = [0.5 \ 0.5]^*$. The frame matrix associated with this frame is

$$\mathbf{F} = \begin{bmatrix} 0.35 & 0.61 & 0.5 & 0.5 \\ -0.61 & 0.35 & -0.5 & 0.5 \end{bmatrix}; \tag{5.12}$$

we may check that $\mathbf{F}$ is indeed the frame matrix of a tight frame since $\mathbf{F}\mathbf{F}^* = \mathbf{I}_2$.

We wish to construct an orthogonal frame matrix $\widetilde{\mathbf{F}}$ such that $\mathbf{F} = P_{\mathcal{U}}\widetilde{\mathbf{F}}$. In the proof of Theorem 3.2 for the case $k < m$, we constructed an $m \times m$ unitary matrix $\widetilde{\mathbf{F}}$ using the SVD $\mathbf{F} = \mathbf{U}\Sigma\mathbf{V}^*$. Using this construction here, we obtain:

$$\mathbf{U} = \begin{bmatrix} 0.5 & -0.87 \\ -0.87 & -0.5 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \ \mathbf{V} = \begin{bmatrix} 0.70 & 0 & 0.70 & 0 \\ 0 & -0.70 & 0 & -0.70 \\ 0.68 & -0.18 & -0.68 & 0.18 \\ -0.18 & -0.68 & 0.18 & 0.68 \end{bmatrix}. \tag{5.13}$$

We now define the extended frame matrix $\widetilde{\mathbf{U}}$ in accordance with the proof of Theorem 3.2. The first two columns of $\widetilde{\mathbf{U}}$ are uniquely defined as the first two columns of $\mathbf{U}$ with zeroes appended. The remaining two columns are arbitrary, as long as the resulting $\widetilde{\mathbf{U}}$ is unitary. A possible choice is:

$$\widetilde{\mathbf{U}} = \begin{bmatrix} 0.5 & -0.87 & 0 & 0 \\ -0.87 & -0.5 & 0 & 0 \\ 0 & 0 & 0.5 & -0.87 \\ 0 & 0 & -0.87 & -0.5 \end{bmatrix}. \tag{5.14}$$

Then

$$\widetilde{\mathbf{F}} = \widetilde{\mathbf{U}}\mathbf{V}^* = \begin{bmatrix} 0.35 & 0.61 & 0.5 & 0.5 \\ -0.61 & 0.35 & -0.5 & 0.5 \\ 0.35 & 0.61 & -0.5 & -0.5 \\ -0.61 & 0.35 & 0.5 & -0.5 \end{bmatrix}. \tag{5.15}$$

We may immediately verify that $\widetilde{\mathbf{F}}^*\widetilde{\mathbf{F}} = \mathbf{I}_4$; *i.e.,* $\widetilde{\mathbf{F}}$ represents an orthonormal set of vectors.

Since the columns of $\mathbf{F}$ span a 2-dimensional Hilbert space $\mathcal{U} = \mathcal{H}$,

$$P_{\mathcal{U}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{5.16}$$

and indeed $\mathbf{F} = P_{\mathcal{U}}\widetilde{\mathbf{F}}$. □

Example 5.2. We now consider an example in which $k > m$. Constructing $\widetilde{\mathbf{F}}$ is simpler than in the previous case because we do not have to extend $\mathcal{H}$. Consider the frame vectors $\varphi_1 = \frac{1}{2}[1\ 1\ 1]^*$, $\varphi_2 = \frac{1}{2}[-1\ 1\ 1]^*$, and $\varphi_3 = \frac{1}{2}[\sqrt{2}\ 0\ 0]^*$. The corresponding frame matrix is

$$\mathbf{F} = \frac{1}{2}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}. \tag{5.17}$$

To verify that $\mathbf{F}$ is a frame matrix of a tight frame, we compute the SVD $\mathbf{F} = \mathbf{U}\Sigma\mathbf{V}^*$, where

$$\mathbf{U} = \begin{bmatrix} 0.58 & 0.82 & 0 \\ 0.58 & -0.4 & 0.7 \\ 0.58 & -0.4 & -0.7 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \ \mathbf{V} = \begin{bmatrix} 0.87 & 0 & 0.5 \\ 0.29 & -0.82 & -0.5 \\ 0.4 & 0.58 & -0.7 \end{bmatrix}. \tag{5.18}$$

From Theorem 5.4 we conclude that $\mathbf{F}$ is indeed the frame matrix of a tight frame since its nonzero singular values are all equal to 1; *i.e.,* $\mathbf{F}$ is a transjector. A basis for the subspace $\mathcal{U}$ spanned by the columns of $\mathbf{F}$ is the two vectors

$$\mathbf{u}_1 = \begin{bmatrix} 0.58 & 0.58 & 0.58 \end{bmatrix}^*, \ \mathbf{u}_2 = \begin{bmatrix} 0.82 & -0.4 & -0.4 \end{bmatrix}^*. \tag{5.19}$$

Thus, $P_{\mathcal{U}}$ is given by

$$P_{\mathcal{U}} = \sum_{i=1}^{2} \mathbf{u}_i\mathbf{u}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}; \tag{5.20}$$

and indeed $\mathbf{F}\mathbf{F}^* = P_{\mathcal{U}}$.

We now define an extended frame matrix $\widetilde{\mathbf{F}}$ such that $\mathbf{F} = P_{\mathcal{U}}\widetilde{\mathbf{F}}$ and $\widetilde{\mathbf{F}}^*\widetilde{\mathbf{F}} = \mathbf{I}_3$. From the proof of Theorem 3.2, we have

$$\widetilde{\mathbf{F}} = \mathbf{U}\mathbf{V}^* = \mathbf{F} + \mathbf{u}_3\mathbf{v}_3^* = \begin{bmatrix} 0.5 & -0.5 & 0.7 \\ 0.85 & 0.15 & -0.5 \\ 0.15 & 0.85 & 0.5 \end{bmatrix}, \tag{5.21}$$

where

$$\mathbf{u}_3 = \begin{bmatrix} 0 & 0.7 & -0.7 \end{bmatrix}^*, \quad \mathbf{v}_3 = \begin{bmatrix} 0.5 & -0.5 & 0.7 \end{bmatrix}^*. \tag{5.22}$$

Since $P_{\mathcal{U}}\mathbf{u}_3\mathbf{v}_3^* = 0$, we have immediately that $\mathbf{F} = P_{\mathcal{U}}\widetilde{\mathbf{F}}$. □

Optimal tight frames

In the special case of a tight frame the dual frame vectors are proportional to the original frame vectors so that the reconstruction formula is particularly simple. In many applications it is therefore desirable to construct a tight frame from an arbitrary set of frame vectors.

The problem of frame design has received relatively little attention in the frame literature. Typically in applications the frame vectors are chosen, rather than optimized. Iterative algorithms for constructing frames that are optimal in some sense are given in [114]. Methods for generating frames starting from a given frame are described in [85].

A popular frame construction from a given set of vectors is the canonical frame [69, 70, 71, 72, 115, 68], first proposed in the context of wavelets in [73]. The *canonical tight frame vectors* $\{\mu_i, 1 \leq i \leq m\}$ associated with the vectors $\{\varphi_i, 1 \leq i \leq m\}$ are given by

$$\mu_i = \left(S^{1/2}\right)^{\dagger} \varphi_i, \tag{5.23}$$

where S is the frame operator. The canonical frame is relatively simple to construct, can be determined directly from the given vectors, and plays an important role in wavelet theory [74, 14, 75]. Some optimality properties of canonical frames have been discussed in [115].

Exploiting the equivalence between normalized tight frames and POVMs, in [68] we use the least-squares measurement developed in the context of quantum detection [26], to

systematically construct optimal tight frames from a given set of vectors. Specifically, we seek a tight frame consisting of frame vectors that minimize the sum of the squared norms of the error vectors, where the ith error vector is defined as the difference between the ith given vector and the ith frame vector. The optimal tight frame is derived both for the case in which the scaling of the frame is specified, and for the the case in which the scaling is such that the error is minimized. It turns out that the canonical frame vectors are proportional to the optimal frame vectors. We discuss these results in more detail in Chapter 8, in the context of general least-squares inner product shaping.

5.3.2 Geometrically Uniform Frames

We have seen in the previous section that imposing an orthogonality constraint on the measurement vectors $\{q_i\}$ of M_1 leads to effective measurement vectors that form a tight frame. In this section we explore a class of frames that results from imposing more general inner product constraints on the vectors $\{q_i\}$. Specifically, we consider the case in which the measurement vectors $\{q_i\}$ have a strong symmetry property called *geometric uniformity*, which as we show is equivalent to a particular inner product constraint. We have seen in Chapter 3 that such vector sets play an important role in quantum detection theory, since in contrast with general vector sets, the optimal measurement for distinguishing between geometrically uniform sets is known and is equal to the least-squares measurement [26]. In the context of combined measurements, the effective measurement vectors corresponding to these sets form what we define as a *geometrically uniform frame*[1] [76]. This class of frames is highly structured resulting in nice computational properties, and possess strong symmetries that may be advantageous in a variety of applications.

Geometrically uniform vector sets

A set of vectors $\{\varphi_i \in \mathcal{H}, 1 \leq i \leq m\}$ is geometrically uniform (GU) [77, 116, 26] if every vector in the set has the form $\varphi_i = U_i\varphi$, where φ is an arbitrary *generating vector* and the transformations $\{U_i, 1 \leq i \leq n\}$ are unitary and form an abelian group[2] $\mathcal{Q}$. For concreteness we assume that $U_1 = I$. The group $\mathcal{Q}$ will be called the *generating group* of $\mathcal{S}$.

[1] The work on geometrically uniform frames was done in collaboration with H. Bölcskei.

[2] That is, $\mathcal{Q}$ contains the identity transformation I; if $\mathcal{Q}$ contains U_i, then it also contains its inverse U_i^{-1}; the product U_iU_j of any two elements of $\mathcal{Q}$ is again in $\mathcal{Q}$; and $U_iU_j = U_jU_i$ for any two elements in $\mathcal{Q}$ [117].

Alternatively, a vector set is GU if given any two vectors φ_i and φ_j in the set, there is an isometry Z_{ij} that transforms φ_i into φ_j while leaving the set invariant [77]. Intuitively, a vector set is GU if it "looks the same" geometrically from any of the points in the set. Some examples of GU vector sets are considered in [77].

We remark that in our development we focus our attention on GU vector sets that are generated by finite abelian groups, although GU vector sets can also be defined over possibly infinite, non-abelian groups.

The Gram matrix of inner products $\{\mathbf{G} = \langle\varphi_i, \varphi_j\rangle\}$ of a GU signal set has the property that every row (column) is a permutation of the first row (column) [26]. Such a matrix is called a *permuted matrix*[3]. Furthermore, if the Gram matrix $\mathbf{G} = \{\langle\varphi_i, \varphi_j\rangle\}$ is a permuted matrix and in addition $\mathbf{G} = \mathbf{G}^T$, then we show in [27] that the vectors $\{\varphi_i\}$ are GU. For example, a set of vectors with the property that the inner products between distinct vectors in the set are all equal is GU.

To further characterize the properties of a GU vector set, it will be convenient to replace the multiplicative group $\mathcal{Q}$ by an additive group Q to which $\mathcal{Q}$ is isomorphic[4]. Specifically, it is well known (see *e.g.,* [117]) that every finite abelian group $\mathcal{Q}$ is isomorphic to a direct product Q of a finite number of cyclic groups: $\mathcal{Q} \cong Q = \mathcal{Z}_{n_1} \times \cdots \times \mathcal{Z}_{n_p}$, where $\mathcal{Z}_{n_t}$ is the cyclic additive group of integers modulo n_t, and $m = \prod_t n_t$. Thus every element $U_i \in \mathcal{Q}$ can be associated with an element $q \in Q$ of the form $q = (q_1, q_2, \ldots, q_p)$, where $q_t \in \mathcal{Z}_{n_t}$; this correspondence is denoted by $U_i \leftrightarrow q$.

Each vector $\varphi_i = U_i\varphi$ is then denoted as $\varphi(q)$, where $U_i \leftrightarrow q$. The zero element $0 = (0, 0, \ldots, 0) \in Q$ corresponds to the identity $I \in \mathcal{Q}$, and an additive inverse $-q \in Q$ corresponds to a multiplicative inverse $U_i^{-1} = U_i^* \in \mathcal{Q}$. The Gram matrix is then the matrix

$$\mathbf{G} = \{\langle\varphi(q'), \varphi(q)\rangle, q', q \in Q\} = \{s(q - q'), q', q \in Q\}, \tag{5.24}$$

[3]An example of a permuted matrix is $\begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_2 & a_1 & a_4 & a_3 \\ a_3 & a_4 & a_1 & a_2 \\ a_4 & a_3 & a_2 & a_1 \end{bmatrix}$.

[4]Two groups $\mathcal{Q}$ and $\mathcal{Q}'$ are *isomorphic*, denoted by $\mathcal{Q} \cong \mathcal{Q}'$, if there is a bijection (one-to-one and onto map) $\varphi : \mathcal{Q} \to \mathcal{Q}'$ which satisfies $\varphi(xy) = \varphi(x)\varphi(y)$ for all $x, y \in \mathcal{Q}$ [117].

with row and column indices $q', q \in Q$, where s is now the function on Q defined by

$$s(q) = \langle \varphi(0), \varphi(q) \rangle. \tag{5.25}$$

The Fourier transform (FT) of a complex-valued function $\varphi : Q \to \mathbb{C}$ defined on $Q = \mathcal{Z}_{n_1} \times \cdots \times \mathcal{Z}_{n_p}$ is the complex-valued function $\hat{\varphi} : Q \to \mathbb{C}$ defined by [26, 118]

$$\hat{\varphi}(h) = \frac{1}{\sqrt{m}} \sum_{q \in Q} \langle h, q \rangle \varphi(q), \tag{5.26}$$

where the Fourier kernel $\langle h, q \rangle$ is

$$\langle h, q \rangle = \prod_{t=1}^{p} e^{-2\pi i h_t q_t / n_t}. \tag{5.27}$$

Here h_t and q_t are the tth components of h and q respectively, and the product $h_t q_t$ is taken as an ordinary integer modulo n_t.

The FT matrix over Q is defined as the $m \times m$ matrix $\mathcal{F} = \{\frac{1}{\sqrt{m}} \langle h, q \rangle, h, q \in Q\}$. The FT of a column vector $\varphi = \{\varphi(q), q \in Q\}$ is then the column vector $\hat{\varphi} = \{\hat{\varphi}(h), h \in Q\}$ given by $\hat{\varphi} = \mathcal{F}\varphi$. Since $\mathcal{F}$ is unitary, we obtain the inverse FT formula

$$\varphi = \mathcal{F}^* \hat{\varphi} = \left\{ \frac{1}{\sqrt{m}} \sum_{h \in Q} \langle h, q \rangle^* \hat{\varphi}(h), q \in Q \right\}. \tag{5.28}$$

The FT matrix plays an important role in defining GU vector sets, as incorporated in the theorem below. We provide a proof of this theorem in [76].

Theorem 5.6. *A set of vectors $\{\varphi_i \in \mathcal{H}, 1 \leq i \leq n\}$ is geometrically uniform if and only if the Gram matrix $\mathbf{G} = \{\langle \varphi_i, \varphi_j \rangle\}$ is diagonalized by a Fourier transform matrix $\mathcal{F}$ over a finite product of cyclic groups Q.*

As a consequence of Theorem 5.6 and its proof, we have the following corollary.

Corollary 5.1. *A set of vectors is geometrically uniform if and only if the corresponding set transformation F has an SVD of the form $F = U\Sigma\mathcal{F}^*$, where U is a set transformation corresponding to m orthonormal vectors u_i, Σ is an $m \times m$ diagonal matrix with diagonal elements σ_i, and $\mathcal{F}$ is an $m \times m$ Fourier transform matrix over a direct product Q of cyclic groups. Then the vectors corresponding to F may be expressed as $\{\varphi(q) = U(q)\varphi, q \in Q\}$*

where $U(q) = UB(q)U^*$ *with* $B(q)$ *denoting the diagonal matrix with diagonal elements* $\{\langle h, q\rangle, h \in Q\}$ *where* $\langle h, q\rangle$ *is defined by (5.27), and* $\varphi = (1/\sqrt{n})\sum_i \sigma_i u_i$.

Geometrically uniform frames

Suppose now that M_1 is a ROM with measurement vectors $\{q_i\}$ that are GU and span a subspace $\mathcal{W} \subseteq \mathcal{H}$. Then the effective measurement vectors $\{v_i = P_{\mathcal{U}} q_i\}$ of the combined measurement M_{E1} with $M_E = P_{\mathcal{U}}$, are also GU. Indeed, let Q and V denote the set transformations corresponding to the vectors q_i and v_i, respectively. Then From Corollary 5.1 Q has the form $Q = U\Sigma\mathcal{F}^*$, where the vectors $\{u_i, 1 \leq i \leq m\}$ corresponding to U form an orthonormal basis for $\mathcal{W}$ in which the first n vectors form an orthonormal basis for $\mathcal{U}$. Then $V = P_{\mathcal{U}} Q = U\Sigma'\mathcal{F}^*$ where $P_{\mathcal{U}} = \sum_{i=1}^{n} u_i u_i^*$, and Σ' is a diagonal matrix with the first n diagonal elements equal to the diagonal elements of Σ and the remaining diagonal elements all equal zero, and from Corollary 5.1 the vectors v_i are also GU. In addition, from Proposition 5.1 the vectors $v_i = P_{\mathcal{U}} q_i$ form a frame for $\mathcal{U}$. Consequently, the effective measurement vectors form what we define as a *geometrically uniform frame* for $\mathcal{U}$.

Definition 5.1. *A finite set of vectors* $\{\varphi_i \in \mathcal{H}, 1 \leq i \leq m\}$ *form a* geometrically uniform frame *for a subspace* $\mathcal{U} \subseteq \mathcal{H}$, *if the vectors* $\{\varphi_i\}$ *are GU and span* $\mathcal{U}$.

If a set of vectors forms a GU frame for $\mathcal{U}$, then they can always be the effective measurement vectors of a combined measurement M_{E1}, where $M_E = P_{\mathcal{U}}$ and M_1 is a ROM with measurement vectors that are GU. Equivalently, a GU frame for $\mathcal{U}$ can always be viewed as the orthogonal projection onto $\mathcal{U}$ of a set of GU vectors in a larger space containing $\mathcal{U}$. The proof of this result follows immediately from the proof of Theorem 5.1 and Corollary 5.1, and is therefore omitted.

GU frames, that arise from a ROM with an inner product constraint on the measurement vectors followed by a SSM, have many desirable properties, which we explore in [76]. A fundamental characteristic of these frames is that they are highly structured and posses strong symmetry properties that may be desirable in a variety of applications such as channel coding [77, 78, 79] and multiple description source coding [80].

Two important classes of highly structured frames are Gabor (Weyl-Heisenberg (WH)) frames [119, 120] and wavelet frames [66, 67, 69]. Both classes of frames are generated by a single generating function. WH frames are obtained by translations and modulations of the

generating function (referred to as the window function), and wavelet frames are obtained by shifts and dilations of the generating function (referred to as the mother wavelet). Like WH and wavelet frames, GU frames are also generated from a single generating vector. Furthermore, we show in [76] that the dual frame vectors and canonical tight frame vectors associated with GU frames are also GU, and are therefore generated by a single generating vector which can be computed very efficiently using a FT matrix defined over the generating group $\mathcal{Q}$ of the frame. These properties are summarized in the following theorem [76]:

Theorem 5.7 (GU frames). *Let $\{\varphi_i = U_i\varphi, 1 \leq i \leq m, U_i \in \mathcal{Q},\}$ be a geometrically uniform frame generated by a finite abelian group $\mathcal{Q}$ of unitary transformations, where φ is an arbitrary generating vector, and let F be the corresponding frame transformation. Let Q be an additive abelian group isomorphic to $\mathcal{Q}$, let $\{\varphi(q), q \in Q\}$ be the elements of the geometrically uniform set under this isomorphism, and let $\mathcal{F}$ be the Fourier transform matrix over Q. Then*

1. *the dual frame vectors $\{\bar{\varphi}_i, 1 \leq i \leq m\}$ are geometrically uniform with generating group $\mathcal{Q}$ and generating vector $\bar{\varphi} = (1/\sqrt{m}) \sum_{h\in\mathcal{I}} (1/\sigma(h))u(h)$, where*
   - *(a) $\{\sigma(h) = m^{1/4}\sqrt{\hat{s}(h)}, h \in Q\}$ are the singular values of F,*
   - *(b) $\{\hat{s}(h), h \in Q\}$ is the Fourier transform of the inner-product sequence $\{\langle\varphi(0), \varphi(q)\rangle, q \in Q\}$,*
   - *(c) $\mathcal{I}$ is the set of indices $h \in Q$ for which $\sigma(h) \neq 0$,*
   - *(d) $u(h) = \hat{\varphi}(h)/\sigma(h)$ for $h \in \mathcal{I}$,*
   - *(e) $\{\hat{\varphi}(h), h \in Q\}$ is the Fourier transform of $\{\varphi(q), q \in Q\}$,*
2. *the canonical tight frame vectors $\{\mu_i, 1 \leq i \leq m\}$ are geometrically uniform with generating group $\mathcal{Q}$ and generating vector $\mu = (1/\sqrt{m}) \sum_{h\in\mathcal{I}} u(h)$.*

An important special case of Theorem 5.7 is the case in which the generating group $\mathcal{Q}$ is *cyclic* so that $U_i = Z^{i-1}, 1 \leq i \leq m$, where Z is a unitary transformation with $Z^m = I$. A cyclic group generates a cyclic vector set $\{\varphi_i = Z^{i-1}\varphi,\ 1 \leq i \leq m\}$, where φ is arbitrary. If $\mathcal{Q}$ is cyclic, then $\mathbf{G}$ is a circulant matrix[5], and Q is the cyclic group $\mathcal{Z}_m$. The FT kernel

[5] A circulant matrix is a matrix where every row (or column) is obtained by a right circular shift (by one position) of the previous row (or column). An example is: $\begin{bmatrix} a_0 & a_2 & a_1 \\ a_1 & a_0 & a_2 \\ a_2 & a_1 & a_0 \end{bmatrix}$.

is then $\langle h, g \rangle = e^{-2\pi i hg/m}$ for $h, g \in \mathcal{Z}_m$, and the FT matrix $\mathcal{F}$ reduces to the $m \times m$ DFT matrix. The singular values of F are then $m^{1/4}$ times the square roots of the DFT values of the inner products $\{\langle \varphi_1, \varphi_j \rangle, 1 \leq j \leq m\}$.

Pruning GU frames

In applications it is often desirable to know how a frame behaves when one or more frame elements are removed. In particular, it is important to know or to be able to estimate the frame bounds of the reduced frame. In general, if no structural constraints are imposed on a frame this behavior will depend critically on the particular frame elements removed.

One of the prime applications of frames is signal analysis and synthesis, where a signal is expanded by computing the inner products of the signal with the frame elements. The resulting coefficients are subsequently stored, transmitted, quantized or manipulated in some way. In particular, a coefficient may be lost (*e.g.,* due to a transmission error) which results in a reconstructed signal that is equivalent to an expansion using a pruned frame obtained by removing the corresponding frame vector.

Recently, there has been increased interest in using frames for multiple description source coding where a signal is expanded into a redundant set of functions and the resulting coefficients are transmitted over a lossy packet network, where one or more of the coefficients can be lost because a packet is dropped [80]. The goal of multiple description source coding is to ensure a gradually behaving reconstruction quality as a function of the number of dropped packets. When using frames in this context, the reconstruction quality is often governed by the frame bound ratio of the pruned frame. If the packets are dropped with equal probability, then it is desirable that the frame bound ratio should deteriorate uniformly irrespective of the particular frame element that is removed. In [76] we show that GU frames have this property. We also demonstrate that if the original frame is a tight GU frame, then the frame bound ratio of the pruned frame obtained by removing one frame element can be computed exactly.

Finally, we consider distance properties of GU frames, which may be of interest when using GU frames for code design (group codes) [116, 77]. In particular, in [76] we introduce a class of GU frames with strictly positive distance spectra for all choices of generating vectors. Such GU frames are shown to be generated by fixed-point-free groups [121].

An interesting direction for further research is to characterize the more general class

of infinite-dimensional GU frames over possibly nonabelian groups using continuous-time Fourier transforms defined over nonabelian groups (see *e.g.*, [118]).

5.4 ROM Followed by an Oblique Projection Operator

In the previous section we considered generalizations of frames that result from viewing frames as orthogonal projections of a set of vectors in a larger space, and imposing various inner product constraints on these vectors. In this section we consider extensions of frames that result from viewing frames as *oblique projections* of vectors in a larger space.

To this end, we now characterize the outcomes of the combined measurement M_{E1} on $\mathcal{U}$, where M_1 is a ROM with measurement vectors $\{q_i, 1 \leq i \leq m\}$ that span a subspace $\mathcal{W} \supseteq \mathcal{U}$, and $M_E = E_{\mathcal{US}}$ is an oblique projection onto $\mathcal{U}$ along $\mathcal{S}$, where $\mathcal{S}$ is an arbitrary subspace of $\mathcal{H}$ such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$. For any $x \in \mathcal{U}$, $M_1(x) = cq_i$ for some $c \in \mathbb{C}$ and one value of i, where i depends on the input x only through the inner products $\langle q_i, x\rangle$, and $M_{E1}(x) = cE_{\mathcal{US}}q_i = cv_i$ where $v_i = E_{\mathcal{US}}q_i$. Since $x \in \mathcal{U}$, $x = E_{\mathcal{US}}x$ and

$$\langle q_i, x\rangle = \langle q_i, E_{\mathcal{US}}x\rangle = \langle E^*_{\mathcal{US}}q_i, x\rangle = \langle \tilde{v}_i, x\rangle, \tag{5.29}$$

where $\tilde{v}_i = E^*_{\mathcal{US}}q_i$. We therefore conclude that the outcome of the combined measurement M_{E1} on $\mathcal{U}$ is proportional to one of the vectors v_i, where the value of i depends on the input x only through the inner products $\langle \tilde{v}_i, x\rangle$. Note that in contrary to the case $M_E = P_{\mathcal{U}}$, here we need two sets of vectors, $\{v_i\}$ and $\{\tilde{v}_i\}$, to fully characterize the outcome of the combined measurement M_{E1}.

From Proposition 5.1 the vectors $\{v_i \in \mathcal{U}, 1 \leq i \leq m\}$ form a frame for $\mathcal{U}$. Since $E^*_{\mathcal{US}}$ is an oblique projection onto $\mathcal{S}^\perp$ along the direction of $\mathcal{U}^\perp$, the vectors $\{\tilde{v}_i \in \mathcal{S}^\perp, 1 \leq i \leq m\}$ form a frame for $\mathcal{S}^\perp$. In the special case in which the vectors $\{q_i, 1 \leq i \leq m\}$ are orthonormal we have that $P_{\mathcal{W}} = \sum_i q_i q_i^*$ so that for any $x \in \mathcal{U}$,

$$x = E_{\mathcal{US}}P_{\mathcal{W}}E_{\mathcal{US}}x = E_{\mathcal{US}}\left(\sum_{i=1}^{m} q_i q_i^*\right)E_{\mathcal{US}}x = \sum_{i=1}^{m}\langle \tilde{v}_i, x\rangle v_i. \tag{5.30}$$

Any vector in $\mathcal{U}$ can then be expressed as a linear combination of the vectors v_i where the coefficients may be chosen as $\langle \tilde{v}_i, x\rangle$. The expansion (5.30) is reminiscent of a frame expansion in $\mathcal{U}$ in terms of a set of frame vectors and dual frame vectors, however here the

analysis vectors $\tilde{v}_i$ do not lie in $\mathcal{U}$ as in a conventional frame expansion and are therefore not equal to the dual frame vectors of v_i. Since the vectors $\tilde{v}_i$ lie in $\mathcal{S}^\perp$, we call these vectors the *oblique dual frame vectors* of the vectors $\{v_i\}$ over $\mathcal{S}^\perp$.

Thus, combined measurements with oblique projection operators lead to interesting new frame expansions that result from viewing frames as oblique projections of vectors in a larger space. Here the oblique dual frame vectors are defined in terms of a set of orthonormal vectors in a larger space; in the next section we suggest a more general and direct definition motivated by the properties of the effective measurement vectors of the combined measurement.

5.4.1 Oblique Dual Frame Vectors

The effective measurement vectors considered in the previous section lead to the notion of signal expansions in which the analysis and synthesis vectors do not lie in the same space. We now would like to obtain a general definition of oblique dual frame vectors to preserve these properties. Therefore, suppose that the vectors $\{\varphi_i, 1 \leq i \leq m\}$ form a frame for $\mathcal{U}$ with frame transformation F. We define the oblique dual frame vectors $\tilde{\varphi}_i$ on $\mathcal{S}^\perp$ of the vectors φ_i such that $\tilde{\varphi}_i \in \mathcal{S}^\perp$ and such that any $x \in \mathcal{U}$ can be expressed as $x = \sum_{i=1}^m \langle \tilde{\varphi}_i, x \rangle \varphi_i$.

First we note that any $x \in \mathcal{U}$ can be expressed as $x = \sum_{i=1}^m \langle \bar{\varphi}_i, x \rangle \varphi_i$, where the vectors $\bar{\varphi}_i \in \mathcal{U}$ are the dual frame vectors of the frame vectors $\varphi_i \in \mathcal{U}$, and are the vectors corresponding to the set transformation $(F^\dagger)^*$. The definition of the dual frame vectors suggests a useful approach for defining the oblique dual frame vectors. Specifically, we suggest defining them as the vectors corresponding to the set transformation $(F^\#_{\mathcal{VS}})^*$, where $F^\#_{\mathcal{VS}}$ is the oblique pseudoinverse of F on $\mathcal{V}$ along $\mathcal{S}$, described in Chapter 2, and $\mathcal{V} = \mathcal{N}(F)^\perp$. We now show that this definition is compatible with our basic requirements.

Let $\{\varphi_i, 1 \leq i \leq m\}$ be a frame for $\mathcal{U}$ with corresponding frame transformation F, and let $F^\#_{\mathcal{VS}}$ be the oblique pseudoinverse of F on $\mathcal{V} = \mathcal{N}(F)^\perp$ along $\mathcal{S}$. Let $\tilde{\varphi}_i$ denote the vectors corresponding to $(F^\#_{\mathcal{VS}})^*$. Then since $\mathcal{R}((F^\#_{\mathcal{VS}})^*) = \mathcal{N}(F^\#_{\mathcal{VS}})^\perp = \mathcal{S}^\perp$, the vectors $\tilde{\varphi}_i$ lie in $\mathcal{S}^\perp$. Furthermore, since $\mathcal{R}(F) = \mathcal{U}$ it follows from the properties of the oblique pseudoinverse (2.27) that $FF^\#_{\mathcal{VS}} = E_{\mathcal{US}}$. Thus,

$$(F^\#_{\mathcal{VS}})^* F^* = (FF^\#_{\mathcal{VS}})^* = E^*_{\mathcal{US}} = E_{\mathcal{S}^\perp \mathcal{U}^\perp}. \tag{5.31}$$

Using (5.31), any $x \in \mathcal{S}^\perp$ can be expressed as

$$x = E_{\mathcal{S}^\perp \mathcal{U}^\perp} x = (F^{\#}_{\mathcal{V}\mathcal{S}})^* F^* x = \sum_{i=1}^{m} a_i \tilde{\varphi}_i, \tag{5.32}$$

where a_i are the elements of $a = F^*x$ so that the vectors $\tilde{\varphi}_i$ span $\mathcal{S}^\perp$ and consequently form a frame for $\mathcal{S}^\perp$. In addition, since $F F^{\#}_{\mathcal{V}\mathcal{S}} = E_{\mathcal{U}\mathcal{S}}$, any $x \in \mathcal{U}$ can be expressed as

$$x = E_{\mathcal{U}\mathcal{S}} x = \sum_{i=1}^{m} \langle \tilde{\varphi}_i, x \rangle \varphi_i. \tag{5.33}$$

Eq. (5.33) is just a frame expansion of a signal $x \in \mathcal{U}$. However, in contrast with conventional frame expansions, here the synthesis frame vectors lie in $\mathcal{U}$, while the analysis frame vectors lie in an arbitrary space $\mathcal{S}^\perp$, such that $\mathcal{U}$ and $\mathcal{S}$ are disjoint.

We therefore propose the following definition.

Definition 5.2. *Let $\{\varphi_i \in \mathcal{U}, 1 \leq i \leq m\}$ denote a frame for a subspace $\mathcal{U}$ of $\mathcal{H}$. The oblique dual frame vectors of φ_i on $\mathcal{S}^\perp$, where $\mathcal{S}$ is an arbitrary subspace of $\mathcal{H}$ such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$, are the frame vectors $\{\tilde{\varphi}_i \in \mathcal{S}^\perp, 1 \leq i \leq m\}$ corresponding to the oblique dual frame operator $(F^{\#}_{\mathcal{V}\mathcal{S}})^*$.*

We may immediately verify that the effective measurement vectors $\tilde{v}_i = E^*_{\mathcal{U}\mathcal{S}} q_i$ resulting from a combined measurement M_{E1} where M_1 is a ROM with orthonormal measurement vectors q_i and $M_E = E_{\mathcal{U}\mathcal{S}}$, satisfy the requirements of Definition 5.2 to be the oblique dual frame vectors on $\mathcal{S}^\perp$ of the effective measurement vectors $v_i = E_{\mathcal{U}\mathcal{S}} q_i$. To this end it is sufficient to show that $E^*_{\mathcal{U}\mathcal{S}} Q = F^{\#}_{\mathcal{V}\mathcal{S}}$, where Q is a set transformation corresponding to the orthonormal vectors q_i and $F = E_{\mathcal{U}\mathcal{S}} Q$, which may be readily verified.

In [19] we derived explicit constructions of $F^{\#}_{\mathcal{V}\mathcal{S}}$. In particular, we have the following proposition.

Proposition 5.2 ([19]). *Let the vectors $\{w_i, 1 \leq i \leq n\}$ denote a basis for an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$, let W denote the set transformation corresponding to the vectors w_i, and let the vectors $\{\varphi_i, 1 \leq i \leq m\}$ denote a frame for $\mathcal{U}$ expressible as $F = WZ$ for some $Z\colon\ \mathbb{C}^n \to \mathbb{C}^m$. Let the vectors $\{s_i, 1 \leq i \leq n\}$ denote a basis for an n-dimensional subspace $\mathcal{S}^\perp \subseteq \mathcal{H}$, such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$, let the vectors $\{x_i, 1 \leq i \leq m\}$ denote a frame for*

$\mathcal{S}^\perp$, *and let* S *and* X *denote the corresponding set transformations. Then*

$$F_{\mathcal{VS}}^{\#} = (X^*F)^\dagger X^* = Z^\dagger (S^*W)^{-1} S^*.$$

If in addition the vectors w_i *are orthonormal and the vectors* φ_i *form a* β*-scaled tight frame, then*

$$F_{\mathcal{VS}}^{\#} = \frac{1}{\beta^2} Z^* (S^*W)^{-1} S^*.$$

Note that since $\mathcal{S} \cap \mathcal{U} = \{0\}$, from Lemma 2.1 it follows that S^*W is invertible.

5.4.2 Properties of the Oblique Dual Frame Vectors

Given frame vectors $\{\varphi_i, 1 \leq i \leq m\}$ for $\mathcal{U}$, there are many ways of choosing coefficients a_i such that for any $x \in \mathcal{U}$, $x = \sum_i a_i \varphi_i$. The particular choice $\tilde{a}_i = \langle \tilde{\varphi}_i, x \rangle$ given by the oblique dual frame vectors has some desirable properties that we now discuss, which are analogous to the properties of the conventional dual frame vectors (see, *e.g.,* [113, pp. 88–89, Theorems 4.7–4.8]), and therefore justify our choice of terminology.

Proposition 5.3. *Let* $\{\varphi_i, 1 \leq i \leq m\}$ *denote a frame for an* n*-dimensional subspace* $\mathcal{U} \subseteq \mathcal{H}$*, let* F *be the matrix of vectors* φ_i*, and let* $\mathcal{V} = \mathcal{N}(F)^\perp$*. Let* $\mathcal{S} \subseteq \mathcal{H}$ *be an arbitrary subspace of* $\mathcal{H}$ *such that* $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$*. Then from all possible coefficients* a_i *that satisfy*

$$x = \sum_{i=1}^{m} a_i \varphi_i \tag{5.34}$$

for all $x \in \mathcal{U}$*, the coefficients* $\tilde{a}_i$ *corresponding to* $\tilde{a} = F_{\mathcal{VS}}^{\#} x$ *have minimal norm.*

Proof: From (5.33) it follows that the coefficients $\tilde{a}_i$ indeed satisfy (5.34). Now, let a_i denote the elements of an arbitrary sequence a such that (5.34) is satisfied. Then

$$\sum_{i=1}^{m} (a_i - \tilde{a}_i) \varphi_i = 0, \tag{5.35}$$

which implies that $a - \tilde{a} \in \mathcal{N}(F)$. Since $\tilde{a} = F_{\mathcal{VS}}^{\#} x$, $\tilde{a} \in \mathcal{R}(F_{\mathcal{VS}}^{\#})$ which from (2.29) is equal

to $\mathcal{V} = \mathcal{N}(F)^\perp$. Thus $a = \tilde{a} + y$ where $y \in \mathcal{N}(F)$ so that $\langle \tilde{a}, y \rangle = 0$. Thus,

$$\|a\|^2 = \|\tilde{a}\|^2 + \|a\|^2 \geq \|\tilde{a}\|^2, \tag{5.36}$$

with equality if and only if $a = \tilde{a}$. □

We can consider the property stated in Proposition 5.3 from a slightly different point of view. Since the vectors $\{\varphi_i, 1 \leq i \leq m\}$ form a frame for $\mathcal{U}$, any $x \in \mathcal{U}$ can be expressed as

$$x = Fa \tag{5.37}$$

for some coefficients a_i. However, these coefficients are not unique because the vectors φ_i are linearly dependent. Suppose we are interested in finding the coefficients with minimal norm. Then a is the unique solution to (5.37) that lies in $\mathcal{N}(F)^\perp = \mathcal{V}$. We may express this solution as $a = F^\dagger x$; indeed $Fa = FF^\dagger x = P_{\mathcal{U}} x = x$. Alternatively, we have that $a = F^\#_{\mathcal{VS}} x$ where $\mathcal{S}$ is an arbitrary subspace of $\mathcal{H}$ such that $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$; $Fa = FF^\#_{\mathcal{VS}} x = E_{\mathcal{US}} x = x$. Thus, although the minimal norm coefficients a_i are unique, the resulting analysis vectors t_i such that $a_i = \langle t_i, f \rangle$ are not unique. If in addition we impose the constraint that $t_i \in \mathcal{S}^\perp$, then the unique vectors that result in coefficients with minimal norm correspond to $(F^\#_{\mathcal{VS}})^*$. This interpretation is useful in applications in which a signal $x \in \mathcal{U}$ is corrupted by noise that is known to lie in a subspace $\mathcal{S}$. By using appropriate analysis vectors in $\mathcal{S}^\perp$, we can totally eliminate this noise and at the same time recover the minimal norm coefficients.

Next, suppose we want to reconstruct a signal in $\mathcal{U}$ from some given coefficients b_i. Among all possible reconstruction vectors we seek the vectors that result in a reconstructed signal whose coefficients using a given set of analysis vectors are as close as possible to b_i. Then the optimal synthesis vectors are given by the oblique dual frame operator.

Proposition 5.4. *Let $\hat{x} = \sum_{i=1}^m b_i w_i$ for some vectors $\{w_i, 1 \leq i \leq m\}$ that form a frame for $\mathcal{U}$, and are to be determined. Let $\{t_i, 1 \leq i \leq m\}$ denote a set of analysis vectors corresponding to T. Then the vectors w_i corresponding to the set transformation $(T^\#_{\mathcal{N}(T)^\perp \mathcal{U}^\perp})^*$ result in a reconstruction $\hat{x}$ with coefficients $\langle t_i, \hat{f} \rangle$ that are as close as possible to b_i in l_2-norm sense.*

Proof: Let z denote the coefficients of $\hat{x}$ with analysis vectors t_i, so that $z = T^* W b$. Then $z \in \mathcal{R}(T^*) = \mathcal{N}(T)^\perp$. To minimize $\|z - b\|$ we need to choose a W such that

$z = P_{\mathcal{N}(T)^\perp} b$, *i.e.*, such that $T^*W = P_{\mathcal{N}(T)^\perp}$. In addition we must have that $\mathcal{R}(W) = \mathcal{U}$. Let $W = (T^{\#}_{\mathcal{N}(T)^\perp \mathcal{U}^\perp})^*$. Then from (2.28), $T^*W = (T^{\#}_{\mathcal{N}(T)^\perp \mathcal{U}^\perp} T)^* = P_{\mathcal{N}(T)^\perp}$, and $\mathcal{R}(W) = \mathcal{N}(T^{\#}_{\mathcal{N}(T)^\perp \mathcal{U}^\perp})^\perp = \mathcal{U}$. □

In summary, the oblique dual frame vectors are very similar to the conventional dual frame vectors: Given a set of vectors $\{\varphi_i\}$ that form a frame for $\mathcal{U}$, the dual frame vectors $\{\bar{\varphi}_i\}$ are the unique vectors in $\mathcal{U}$ such that any $x \in \mathcal{U}$ can be expressed as $x = \sum_i \langle \bar{\varphi}_i, x \rangle \varphi_i$, and the coefficients $\langle \bar{\varphi}_i, x \rangle$ have minimal norm from all possible coefficients. Similarly, the oblique dual frame vectors of φ_i on $\mathcal{S}^\perp$, with $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$, are the unique vectors in $\mathcal{S}^\perp$ such that any $x \in \mathcal{U}$ can be expressed as $x = \sum_i \langle \tilde{\varphi}_i, x \rangle \varphi_i$, and the coefficients $\langle \tilde{\varphi}_i, x \rangle$ have minimal norm from all possible coefficients. Thus, using the concept of oblique dual frame vectors we can extend the notion of a frame expansion to the case in which the analysis frame vectors do not lie in the same space as the synthesis frame vectors, but rather lie in an arbitrary subspace $\mathcal{S}^\perp \subseteq \mathcal{H}$, with $\mathcal{H} = \mathcal{U} \oplus \mathcal{S}$.

It is interesting to note that the oblique dual frame vectors of $\tilde{\varphi}_i$ on $\mathcal{U}$ are the frame vectors φ_i. Thus not only do we have $f = \sum_{i=1}^m \langle \tilde{\varphi}_i, f \rangle \varphi_i$ for any $f \in \mathcal{U}$ but we also have $f = \sum_{i=1}^m \langle \varphi_i, f \rangle \tilde{\varphi}_i$ for any $f \in \mathcal{S}^\perp$ [19].

5.5 Summary of Combined Measurements and Frames

To conclude, we presented various generalizations and extensions of frames that result from QSP analogues of the quantum POVM, taking on the form of a combined measurement where a ROM is followed by a SSM. We demonstrated that viewing tight frames in the context of the quantum measurement framework provides additional insight and perspective and leads to frame-theoretical analogues of various results in quantum measurement. Furthermore, imposing inner product constraints on the measurement vectors of the ROM leads to the new class of GU frames that constitute an interesting and potentially important class of frames for various signal processing and communication applications due to their inherent symmetry properties. Choosing the SSM as an oblique projection leads to the definition of oblique dual frame vectors, which can be useful in applications in which it is desirable to work in two different spaces. As one possible application, in Chapter 6 we consider a very general sampling problem with almost no restrictions on it, and we use the oblique dual frame vectors to derive very general reconstruction algorithms. Specifically,

we develop redundant sampling and reconstruction procedures with (almost) arbitrary sampling and reconstruction spaces, that can be used to reduce the quantization error when quantizing the samples prior to reconstruction by as much as the redundancy of the frame in comparison with a nonredundant procedure.

In our closing remarks, there appear to be other important connections to be explored between frame theory and quantum POVMs. There are also a host of applications in which various properties of GU frames can be exploited *e.g.,* multiple description coding, and multiple-antenna code design. These applications require further study of the properties of GU frames, in particular their robustness to erasures and their distance properties.

Another interesting direction for further research is applications of oblique dual frame vectors to sampling and reconstruction algorithms, beyond those explored in Chapter 6. Since the oblique dual frames can be defined over an almost arbitrary space, they can be used to develop sampling procedures with arbitrary sampling and reconstruction spaces. It is well known that by allowing for arbitrary sampling and reconstruction spaces the sampling and reconstruction algorithms can be greatly simplified in many cases with only a minor increase in approximation error [16, 81, 17, 82, 83, 84]. Using oblique dual frame vectors we can further simplify the sampling and reconstruction processes while still retaining the flexibility of choosing the spaces almost arbitrarily, due to the extra degrees of freedom offered by the use of frames that allow us to construct frames with prescribed properties [66, 85].

5.6 SSM Followed by a ROM

We now consider the second class of combined measurements in which a SSM $M_1 = E$ is followed by a ROM M_2. Combined measurements of this form amount to applying a projection as a preprocessor to an existing algorithm, so that the input to the algorithm represented by the ROM M_2 is a projection of the original signal we wish to process.

There are a variety of applications in which orthogonal projection operators have been used as preprocessors, *e.g.,* in various detection scenarios. One justification for the widespread use of orthogonal projections in detection applications is that they arise naturally as part of a GLRT in many problems (see *e.g.,* [122]). By contrast, oblique projections have received comparatively less attention. Although GLRT based detectors are popular,

they are not necessarily optimal. In fact, as we show in the context of a concrete example below, using an oblique projection rather than an orthogonal projection may be advantageous in a variety of contexts and may lead to improved detection performance. Furthermore, in many cases the oblique projection can also be derived as part of a GLRT, albeit under non-standard assumptions.

5.6.1 Subspace Matched Filter Detection

There are a multitude of applications in which oblique projections may be used as a preprocessor as suggested by the combined measurement framework, even though they may not arise as naturally as solutions to popular processing criteria. In this section we consider one such example in a rather preliminary manner to highlight some of the key merits and issues with this approach, as well as identify some directions for further research.

Suppose we have a transmitter that transmits one of m known signals $\{s_i(t), 1 \leq i \leq m\}$ that span a subspace $\mathcal{U} \subseteq \mathcal{H}$ with equal probability, where the signals lie in a real Hilbert space $\mathcal{H}$ with inner product inner product $\langle x(t), y(t) \rangle = \int_{t=-\infty}^{\infty} x(t)y(t)dt$, and are normalized so that $\langle s_i(t), s_i(t) \rangle = 1$ for all i. The channel is assumed to corrupt the transmitted signal by both additive white noise and structured (or low-rank) noise, *i.e.,* noise that lies in a linear subspace [54]. The structured noise component lies in a known subspace $\mathcal{S}$ of $\mathcal{H}$, where we assume that $\mathcal{U}$ and $\mathcal{S}$ are disjoint, but not necessarily orthogonal. Thus, the received signal $r(t)$ is modeled as

$$r(t) = s_i(t) + n_s(t) + n_w(t), \tag{5.38}$$

for one value i where $n_s(t) \in \mathcal{S}$ is a structured noise component, and $n_w(t)$ is a stationary white noise process with zero mean and spectral density σ^2.

Based on the observation $r(t)$ we wish to detect the transmitted signal. In Example 4.1 of Chapter 4 we considered the case where $n_s(t) = 0$, and constructed a ROM, which we denote here by M_2, that implements the MF detector. Since the structured noise lies in a known subspace it seems intuitively that we may be able to improve upon simple MF detection by entirely or partially eliminating this noise. Therefore we propose detecting the transmitted signal using a SSM $M_1 = E$ where E is a projection operator, followed by the ROM M_2. Equivalently, we first project the received signal onto an appropriate subspace,

and then process the projected signal with an MF detector matched to the transmitted signals. The properties of the resulting detector will depend on the projection E. To eliminate the structured noise entirely we choose the null space $\mathcal{N}(E) = \mathcal{S}$. Then, if E is chosen as an orthogonal projection, $E = P_{\mathcal{S}^\perp}$. If, on the the other hand, we choose E as an oblique projection, then we can choose the range of E arbitrarily as long as it is disjoint from $\mathcal{S}$. Since the transmitted signal lies in $\mathcal{U}$, a reasonable choice is $E = E_{\mathcal{US}}$. We now consider the detectors corresponding to the combined measurements resulting from these two choices.

With $E = E_{\mathcal{US}}$, the combined measurement is equivalent to the detector depicted in Fig. 5-1, which we refer to as the oblique subspace MF (OBSMF) detector. The projected signal $r_{\mathcal{US}}(t) = E_{\mathcal{US}} r(t)$ is cross-correlated with the m signals $s_i(t)$ so that $a_i = \langle s_i(t), r_{\mathcal{US}}(t) \rangle$, and the declared detected signal is $s_i(t)$ where $i = \arg\max a_k$. We note that we can construct $E_{\mathcal{US}}$ explicitly using Theorem 2.5 and Proposition 5.2. Specifically, $E_{\mathcal{US}} = S(V^*S)^\dagger V^*$ where S is the set transformation corresponding to the m signals $s_i(t)$ and V is a set transformation corresponding to a set of m signals that span $\mathcal{S}^\perp$.

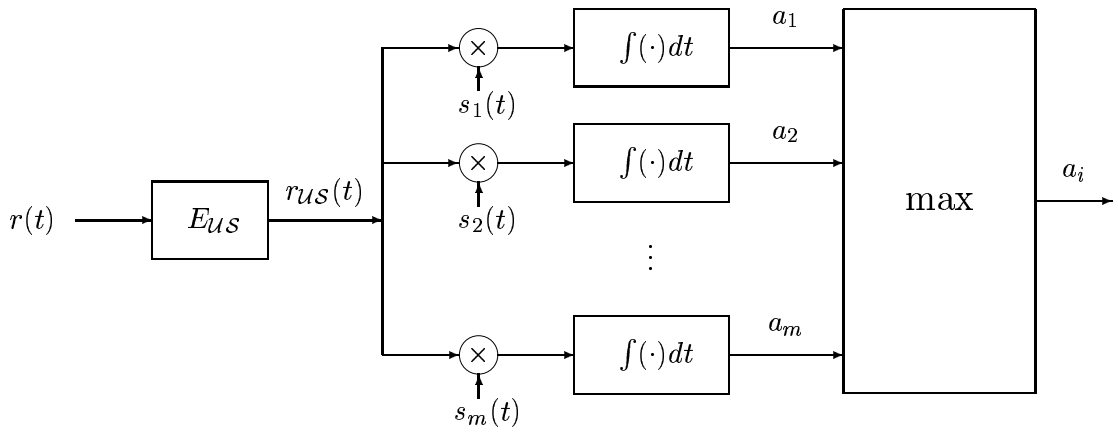


Figure 5-1: Oblique subspace matched filter detector.

If the transmitted signal is $s_i(t)$, then

$$E_{\mathcal{US}} r(t) = E_{\mathcal{US}}(s_i(t) + n_s(t) + n_w(t)) = s_i(t) + E_{\mathcal{US}} n_w(t), \tag{5.39}$$

where we used the fact that $E_{\mathcal{US}} s_i(t) = s_i(t)$ since $s_i(t) \in \mathcal{U}$, and $E_{\mathcal{US}} n_s(t) = 0$ since

$n_s(t) \in \mathcal{S}$. Thus, projecting the received signal using the oblique projection $E_{\mathcal{US}}$ totally eliminates the structured noise without modifying the transmitted signal. However, we note that the noise power of the projected white noise $\|E_{\mathcal{US}} n_w(t)\|^2$ may be larger than the noise power of the original noise $\|n_w(t)\|^2$, due to the oblique projection [54]. A desirable property that the OBSMF detector has is that when the variance of the background noise $\sigma \to 0$, $E_{\mathcal{US}} r(t) \to s_i(t)$ and the detector will correctly detect the transmitted signal with probability 1, irrespective of the transmitted signals and the noise subspace $\mathcal{S}$.

If we choose $E = P_{\mathcal{S}^\perp}$ as an orthogonal projection operator onto $\mathcal{S}^\perp$, then the detector resulting from the combined measurement is equivalent to the detector depicted in Fig. 5-2, which we refer to as the orthogonal subspace MF (OTSMF) detector. The projected signal $r_{\mathcal{S}^\perp}(t) = P_{\mathcal{S}^\perp} r(t)$ is cross-correlated with the m signals $s_i(t)$ so that $c_i = \langle s_i(t), r_{\mathcal{S}^\perp}(t) \rangle$, and the declared detected signal is $s_i(t)$ where $i = \arg\max c_k$.

In this case,

$$P_{\mathcal{S}^\perp} r(t) = P_{\mathcal{S}^\perp}(s_i(t) + n_s(t) + n_w(t)) = P_{\mathcal{S}^\perp} s_i(t) + P_{\mathcal{S}^\perp} n_w(t), \tag{5.40}$$

where we used the fact that $P_{\mathcal{S}^\perp} n_s(t) = 0$ since $n_s(t) \in \mathcal{S}$. Thus, the orthogonal projection operator eliminates the structured noise, but at the same time alters the transmitted signal. Since $P_{\mathcal{S}^\perp}$ is an orthogonal projection operator, the power of the projected noise is no greater than the power of the original noise: $\|P_{\mathcal{S}^\perp} n_w(t)\| \leq \|n_w(t)\|$. However, when $\sigma \to 0$, $P_{\mathcal{S}^\perp} r(t) \to P_{\mathcal{S}^\perp} s_i(t)$ so that the OTSMF will not necessarily correctly detect the transmitted signal. Thus, clearly there are situations in which the OBSMF detector leads to improved performance over the OTSMF detector, particularly in the high SNR regime.

In summary, using a combined measurement we arrived at two subspace MF detectors corresponding to different choices of projection operators. Both projections eliminate the structured noise entirely but have different effects on the background noise and on the transmitted signal. The oblique projection does not alter the signal, but tends to enhance the background noise. We therefore expect the OBSMF detector to be particularly useful in the high SNR regime. By contrast, the orthogonal projection modifies the signal but does not enhance the background noise, and will therefore be advantageous in the low SNR regime. In general, we expect the relative performance of the detectors to depend on the power of the white noise as well as on the structure of the projections onto $\mathcal{S}^\perp$ of the

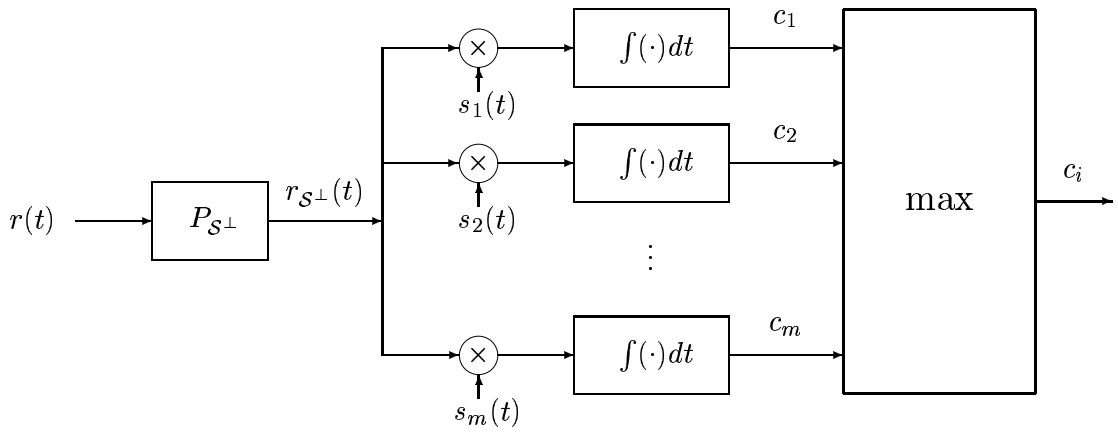


Figure 5-2: Orthogonal subspace matched filter detector.

transmitted signals.

An outstanding issue that would be interesting to investigate is under what conditions on the background noise the OBSMF detector leads to improved performance over the (more common) OTSMF detector. We expect that there is a threshold SNR, above which the OBSMF detector will lead to a higher probability of correct detection than the OTSMF detector. We expect this threshold to depend on the transmitted signals and on the norm of the oblique projection, which in turn depends on the angle $\theta_{\mathcal{US}^\perp}$ between the spaces $\mathcal{U}$ and $\mathcal{S}^\perp$, defined as [16]

$$\cos(\theta_{\mathcal{US}^\perp}) = \inf_{x\in\mathcal{U},\|x\|=1} \|P_{\mathcal{S}}^{\perp} x\|. \tag{5.41}$$

Specifically, it is shown in [82] that for all $x \in \mathcal{H}$,

$$\|E_{\mathcal{US}} x\| \leq \frac{1}{\cos(\theta_{\mathcal{US}^\perp})} \|x\|. \tag{5.42}$$

Thus, if the angle is large then the oblique projection will tend to enhance the background noise while if the angle is small, the enhancement will be marginal.

To conclude, even though the orthogonal projection may result as part of a standard GLRT, the oblique projection may lead to improved performance in many cases. In fact, as we show in the next section, the OBSMF detector can also be derived as a GLRT detector

for detecting the transmitted signal, under certain (nonstandard) assumptions.

5.6.2 Oblique Projections and the Generalized Likelihood Ratio Test

We now assume that the background noise $n_w(t)$ has a Gaussian distribution. Then under a variety of assumptions on the background and structured noise it has been shown in [122] that a (standard) GLRT detector consists of a preprocessor equal to $P_{\mathcal{S}^\perp}$. As we discussed in the previous section in many scenarios we can improve the performance by using an oblique projection operator rather than an orthogonal projection. In this section we establish the viability of the OBSMF detector by showing that it too can be derived as a GLRT detector.

The received signal is now modeled as

$$r(t) = s_i(t) + n_s(t) + n_w(t), \tag{5.43}$$

where $s_i(t)$ and $n_s(t) \in \mathcal{S}$ are as before, and $n_w(t)$ is a stationary white noise Gaussian process with zero mean and spectral density σ^2. Based on the observation $r(t)$ we wish to detect the transmitted signal. Since the structured noise $n_s(t)$ is unknown we cannot derive a detector that minimizes the probability of a detection error, or maximizes the likelihood of the received signal. Instead, we use a GLRT detector in which we first find the maximum likelihood (ML) estimate of the unknown structured noise, and then detect the transmitted signal as the signal that maximizes the likelihood of the received signal given the transmitted signal and the estimate of the structured noise. The ML estimate of the structured noise is derived based on the assumption that the transmitted signal is an unknown signal in the signal subspace[6] $\mathcal{U}$. As we now show, the resulting GRLT detector is equivalent to the OBSMF detector of Fig. 5-1.

To derive the ML estimate of $n_s(t)$ we seek $n_s(t) \in \mathcal{S}$ and $s(t) \in \mathcal{U}$ that maximize the likelihood function $\log f(r(t)|n_s(t), s(t))$ where $f(x(t)|y(t))$ is the probability density function of $x(t)$ given $y(t)$. Since $n_w(t)$ is a white Gaussian process,

$$\log f(r(t)|n_s(t), s(t)) = K - \frac{1}{2\sigma^2}\langle r(t) - s(t) - n_s(t), r(t) - s(t) - n_s(t)\rangle, \tag{5.44}$$

where K is a constant. The maximizing $n_s(t)$ is [54] $\hat{n}_s(t) = E_{\mathcal{SU}} r(t)$, where $E_{\mathcal{SU}}$ is

[6] In a conventional GRLT the structured noise is estimated under each of the hypotheses $s_i(t)$. However, as we have seen, the resulting detector does not always yield satisfactory performance.

the oblique projection onto $\mathcal{S}$ along $\mathcal{U}$. The declared detected signal is then $s_i(t)$ where $i = \arg\max \log f(r(t)|s_i(t), \hat{n}_s(t))$, and

$$\begin{aligned}\log f(r(t)|s_i(t), \hat{n}_s(t)) &= K - \frac{1}{2\sigma^2}\langle r(t) - s_i(t) - E_{\mathcal{SU}} r(t), r(t) - s_i(t) - E_{\mathcal{SU}} r(t)\rangle \\ &= K - \frac{1}{2\sigma^2}\langle (I - E_{\mathcal{SU}})r(t) - s_i(t), (I - E_{\mathcal{SU}})r(t) - s_i(t)\rangle. \quad (5.45)\end{aligned}$$

Finally, $\arg\max\langle (I - E_{\mathcal{SU}})r(t), s_k(t)\rangle = \arg\max\langle E_{\mathcal{US}} r(t), s_k(t)\rangle = \arg\max \tilde{a}_k$, where $\tilde{a}_k$ is the kth output of the OBSMF demodulator depicted in Fig. 5-1.

In our closing remarks, we note that we may view both projections $P_{\mathcal{S}^\perp}$ and $E_{\mathcal{US}}$ as special cases of an oblique projection $E_{\mathcal{VS}}$ in which $\mathcal{V} \subset \mathcal{H}$ is disjoint from $\mathcal{S}$. The oblique projection $E_{\mathcal{US}}$ in the OBSMF detector corresponds to the choice $\mathcal{V} = \mathcal{U}$, while the orthogonal projection $P_{\mathcal{S}^\perp}$ in the OTSMF detector corresponds to the choice $\mathcal{V} = \mathcal{S}^\perp$. We may be able to improve the performance over these detectors by choosing a preprocessor $E_{\mathcal{VS}}$ where $\mathcal{V}$ is an "optimal" subspace in some sense. If $\mathcal{V} = \mathcal{U}$, then the resulting projection does not modify the signal but from (5.41) the norm of the noise can be enhanced by as much as $1/\cos(\theta_{\mathcal{US}^\perp})$. On the other hand if $\mathcal{V} = \mathcal{S}^\perp$, then the corresponding projection does not increase the norm of the noise, but the norm of the desired signal is reduced by as much as $\arg\min(\|P_{\mathcal{S}^\perp} s_i(t)\|/\|s_i(t)\|)$. We may therefore consider choosing an optimal $\mathcal{V}$ to maximize a possibly weighted combination of $\cos(\theta_{\mathcal{VS}^\perp})$ and $\arg\min \|E_{\mathcal{VS}} s_i(t)\|$. Alternatively, we can consider the average effect on the transmitted signal $\sum_i \|E_{\mathcal{VS}} s_i(t)\|$.

5.7 Combined ROMs

The last class of combined measurements we consider are combinations of two ROMs M_1 and M_2. The effect of the ROM M_2 is to map the possible outputs of the ROM M_1 to a new set of outputs, in some deterministic or probabilistic fashion. Thus, combined measurements of this form are useful for developing algorithms based on existing algorithms, where we postprocess the output of the original algorithm. For example, by choosing the mapping f_2 of the ROM M_2 as a probabilistic mapping we can generate randomized algorithms in which the original outputs induce a probability distribution on the final outputs. This can be useful, *e.g.,* in detection algorithms where we implement a randomized decision rule, by first obtaining a deterministic decision which then begets a distribution on the final outputs

(*e.g.,* a randomized likelihood ratio test).

More specifically, suppose that M_1 is a ROM with measurement vectors $\{q_i, 1 \leq i \leq m\}$, and that M_2 is a ROM with measurement vectors $\{h_i, 1 \leq i \leq n\}$. We assume that $q_i \neq h_k$ for all i and k. Then $M_1(x) = c_i q_i$ for some value i and some $c_i \in \mathbb{C}$, and $M_{21}(x) = M_2(c_i q_i) = d_l h_l$ where $l = f_2(\{\langle h_k, c_i q_i \rangle, 1 \leq k \leq n\})$, for some $d_l \in \mathbb{C}$. Evidently, f_2 depends only on the nm values $\{\langle h_k, c_i q_i \rangle, 1 \leq k \leq n, 1 \leq i \leq m\}$. The mapping f_2 maps the ith output of M_1 to the lth output of M_2, where the value of l may be selected deterministically by the value of i, or may be chosen randomly with probabilities determined by the value of i.

To deterministically map the ith output of M_1 to the lth output of M_2 we may choose the vectors $\{h_k, 1 \leq k \leq n\}$ such that $\langle h_k, c_i q_i \rangle = f_{ki}$ where $\max_k f_{ki} = f_{li}$, and then choose $f_2(x) = h_l$ where $l = \arg\max \langle h_k, x \rangle$.

5.7.1 Randomized Algorithms

A more interesting class of algorithms results from choosing f_2 to be a probabilistic mapping. Then each output of M_1 generates a (possibly) different probability distribution on the outputs of M_2, as illustrated in Fig. 5-3. Specifically, if the ith outcome is obtained from the measurement M_1, then the lth outcome of M_{21} is obtained with probability $p(l|i)$. To realize a particular distribution we may choose the vectors h_l such that $\langle h_l, c_i q_i \rangle = p(l|i)$, and choose the mapping $f_2 \colon \mathcal{H} \times \mathcal{W} \to \mathcal{I}$ as a probabilistic mapping from $\mathcal{H}$ to $\mathcal{I} = \{1, 2, \ldots, n\}$, where $\mathcal{W} = \mathcal{I}$ is the sample space of an auxiliary chance variable w, such that w can take on a value $w_l \in \mathcal{I}$ with probability $\langle h_l, x \rangle$. Then let $f_2(x, w_l) = l$. If the ith output of M_1 is obtained, then the output of the combined measurement will be proportional to h_l with probability $p(l|i)$.

We now consider some examples of randomized algorithms resulting from combined ROMs, that highlight some of their merits.

Example 5.3 (Randomized MF). Suppose that one of m signals $\{s_i(t), 1 \leq i \leq m\}$ is received over an additive noise channel with equal probability, where the signals lie in a real Hilbert space $\mathcal{H}$ with inner product $\langle x(t), y(t) \rangle = \int_{t=-\infty}^{\infty} x(t) y(t) dt$, and are assumed to be normalized. The received signal $r(t)$ is also assumed to be in $\mathcal{H}$, and is modeled as $r(t) = s_i(t) + n(t)$ for one value i, where $n(t)$ is a stationary white noise process with zero mean and spectral density σ^2, and with otherwise unknown distribution.

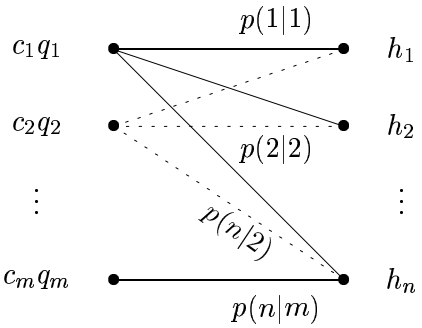


Figure 5-3: Action of a probabilistic mapping f_2.

In Example 4.3 we introduced a probabilistic MF for detecting the transmitted signal, based on a ROM with a probabilistic mapping. In that example the observed signal $r(t)$ induced a probability distribution on the declared detected signal so that $s_i(t)$ is declared with probability $c|\langle r(t), s_i(t)\rangle|^2$, where c is a normalization constant. Since this probability distribution depends explicitly on the value of $r(t)$, in principle there are an unlimited number of possible distributions. To reduce the complexity of the probabilistic MF we can instead use a combined measurement to obtain a randomized MF, where like in the probabilistic MF the declared detected signal is chosen probabilistically, however the probability distribution on the output is now chosen from one of a set of m distributions, where the particular choice depends on the signal $r(t)$.

To describe the combined measurement let M_1 be the ROM that implements the MF detector as in Example 4.1, and let M_2 be a ROM with measurement vectors $\{h_k, 1 \leq k \leq m\}$ and probabilistic mapping f_2: $\mathcal{H} \times \mathcal{W} \to \mathcal{I}$ where $\mathcal{I} = \{1, 2, \ldots, m\}$, and $\mathcal{W} = \mathcal{I}$ is the sample space of an auxiliary chance variable w, such that w can take on a value $w_l \in \mathcal{I}$ with probability $\langle h_l(t), x(t)\rangle$. Then let $f_2(x(t), w_l) = l$. The signals $\{h_k(t), 1 \leq k \leq m\}$ are chosen to satisfy $\langle h_k(t), s_i(t)\rangle = p(k|i)$ where the probabilities $p(k|i), 1 \leq i, k \leq m$ are prespecified. The output of the combined measurement is then mapped to one of the signals $s_l(t)$ by the mapping $T_{\mathcal{Y}}$: $\mathcal{H} \to \mathcal{H}$ which maps any multiple of $h_i(t)$ to $s_i(t)$, as depicted in Fig. 5-4. If the outcome of the measurement M_1 is $c_i s_i(t)$, then the declared detected signal using the combined measurement is $s_l(t)$ with probability $p(l|i)$. Thus, the MF output generates a probability distribution on the final declared detected signal where the probability distribution is chosen from one of m prespecified distributions depending on

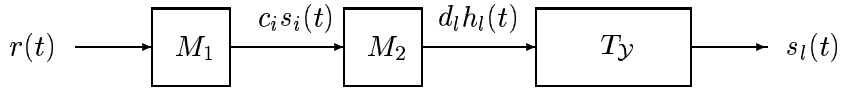


Figure 5-4: Measurement description of randomized matched filter detector.

the output of the MF demodulator. The detector resulting from the combined measurement can equivalently be represented as in Fig. 5-5, and is referred to as a randomized MF (RMF) detector. In this figure, T: $\mathcal{I} \to \mathcal{H}$ is a mapping that maps the index l to the signal $s_l(t)$.

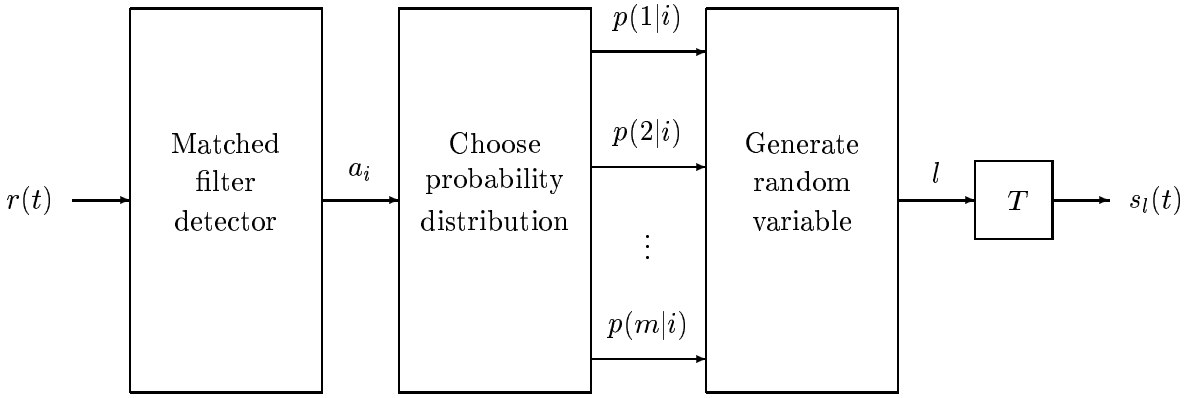


Figure 5-5: Randomized matched filter detector.

To gain some insight into the effect of the randomized MF, suppose that $m = 2$ and that $\langle s_1(t), s_2(t) \rangle$ is close to 1. Then we expect the probability of detection error using an MF detector to be pretty large. If the probability of detection error is greater than $1/2$, then we can reduce the probability of detection error by reversing the detector outputs, so that we declare $s_2(t)$ when the outcome of the detector is $s_1(t)$, and *vice versa*. However, if we do not know the probability of detection error, then by deterministically reversing the outputs we may actually increase the probability of detection error. Instead we use a combined measurement to randomly reverse the outputs. Specifically, if the output of the MF detector is $s_1(t)$, then we declare that $s_1(t)$ was transmitted with probability p_0 and that $s_2(t)$ was transmitted with probability $1 - p_0$, and *vice versa*. As we now show, by using such a randomized decision rule we can improve the worst case behavior of the

detector.

With P_e^{MF} and P_e^{RMF} denoting the probability of detection error using the MF detector and the RMF detector respectively, we have that

$$P_e^{\mathrm{RMF}} = p_0 P_e^{\mathrm{MF}} + (1-p_0)(1-P_e^{\mathrm{MF}}) = (2p_0-1)P_e^{\mathrm{MF}} + 1 - p_0. \tag{5.46}$$

From (5.46) it follows that, as we expect intuitively, $P_e^{\mathrm{RMF}} < P_e^{\mathrm{MF}}$ only if $P_e^{\mathrm{MF}} > 1/2$. If we know that $P_e^{\mathrm{MF}} < 1/2$, then we do not gain using the RMF. On the other hand, if we know that $P_e^{\mathrm{MF}} > 1/2$, then we can always reverse the outputs of the MF and again we do not gain using the RMF. However, if we do not know P_e^{MF}, then using the RMF we can reduce the worst case probability of detection error. Indeed, regardless of the value of P_e^{MF}, using the randomized MF we always have that $P_e^{\mathrm{RMF}} \leq p_0$. In the extreme case in which $p_0 = 1/2$, the output of the RMF is $s_1(t)$ or $s_2(t)$ with probability $1/2$, independent of the received signal $r(t)$, resulting in a constant probability of detection error. □

The previous example suggests that in a detection scenario combined ROMs can be used to improve worst case behavior. The next example demonstrates the same characteristic of the combined measurement in a different context.

Example 5.4 (Modular redundancy). Suppose we have a system S that performs some computation which we want to protect using modular redundancy [123]. Thus, we have 3 copies of the system operating in parallel using the same data. The outputs are then compared with voter circuitry. For simplicity we assume that the output of the computation is binary; thus the possible outputs of each system are 0 and 1, and the possible outputs of the three systems are all possible triplets of 0 and 1, *e.g.,* $(0,0,0), (0,0,1)$ *etc.,* where the first entry corresponds to the output of the first system, and so forth. The outputs of the systems could either all agree, or two outputs will be the same and different than the third. The voter circuitry declares as the final output the output that is common to at least two of the systems. If each system has a probability of failure p, then without redundancy the probability of an output error is p. Using modular redundancy there will be an error if all 3 systems fail, or if 2 of the systems fail, and the probability of error is $P_e^{\mathrm{MR}} = p^3 + 3p^2(1-p)$.

We now propose using randomized modular redundancy (RMR) based on a combined measurement to improve the worst case performance of the overall system. Specifically, we first construct a ROM M_1 to implement the voter circuitry by mapping the 8 possible input

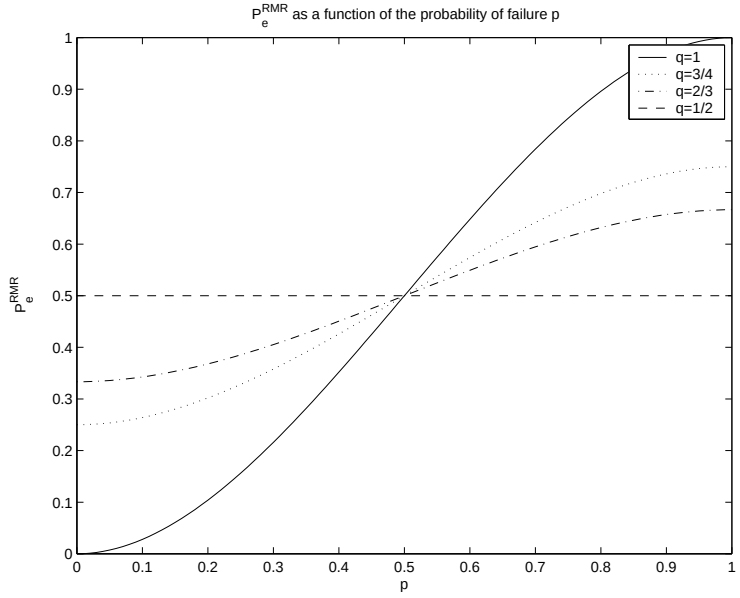


Figure 5-6: Probability of error P_e^{RMR} using randomized MR, as a function of the probability of failure p, for different values of q, where q is the probability of reversing the output of the voter circuitry.

triplets to vectors q_i and the two possible outputs 0 and 1 to vectors h_i. The mapping f_1 is chosen to realize the required mapping between the q_is and the h_is, using the techniques described at the beginning of this section. The second measurement M_2 is constructed such that the outcome is the same as the outcome of the first measurement with probability q, and reversed with probability $1-q$. This construction is similar to the construction of the measurement M_2 in the combined measurement implementing the RMF, described in Example 5.3, and is therefore omitted. We refer to this approach as randomized modular redundancy (RMR). The probability of error using RMR is given by

$$P_e^{\mathrm{RMR}} = qP_e^{\mathrm{MR}} + (1-q)(1-P_e^{\mathrm{MR}}) = (2q-1)P_e^{\mathrm{MR}} + 1 - q. \tag{5.47}$$

In Fig. 5.4 we plot the probability of error using RMR, as a function of the probability of failure p. The case $q = 1$ corresponds to the case of conventional modular redundancy. As we see from the figure, by choosing values of q such that $0 < q < 1$, we can improve the worst case performance of the system. □

In this chapter we suggested some new processing techniques that result from combined

QSP measurements. We have seen that combined measurements lead to various generalizations and insights into frame expansions, to the notion of subspace MF detectors, and to certain types of randomized algorithms. There are potentially a host of additional applications of the combined measurement framework beyond those considered in this chapter.

As indicated at the outset, this chapter represents a preliminary exploration of the combined measurement framework and many topics and issues within these examples remain to be explored. While some of the applications are not sufficiently developed, they potentially represent interesting novel methods for signal processing.

Chapter 6

Sampling With Arbitrary Sampling and Reconstruction Spaces

In this chapter we exploit our results regarding oblique projections and oblique dual frame vectors, derived in Chapters 2 and 5 respectively, to develop a general framework for sampling and reconstruction procedures. The procedures we develop allow for almost arbitrary sampling and reconstruction spaces, as well as arbitrary input signals. We first derive a nonredundant sampling procedure. Based on the concept of oblique dual frame vectors, we then develop a redundant sampling procedure that can be used to reduce the quantization error when quantizing the measurements prior to reconstruction. The algorithms we develop satisfy the consistency requirement, introduced in the context of sampling by Unser and Aldroubi in [16]. Building upon this property of our algorithms, we develop a general procedure for constructing signals with prescribed properties.

6.1 Sampling in Signal Spaces

Many methods exist for representing a signal by a sequence of numbers, which can be interpreted as measurements of the signal we wish to represent. The classical approach is to choose the measurements as samples of the signal. A more recent approach [16, 14, 89, 81, 17, 18, 19, 124, 125] is to consider measurements that may be expressed as the inner products of the signal with a set of vectors that span some subspace $\mathcal{S}$, which is referred to as the sampling space. Examples include multiresolution decompositions [14], and spline decompositions [81]. The problem then is to reconstruct the signal from these

measurements, using a set of vectors that span a subspace $\mathcal{W}$, which we refer to as the reconstruction space. If the signal we wish to reconstruct does not lie in $\mathcal{W}$, then it can not be perfectly reconstructed using only reconstruction vectors that span $\mathcal{W}$. Therefore, if we allow for signals out of $\mathcal{W}$, then we must relax the requirement for perfect reconstruction.

Given a reconstruction method, we can always choose a sampling method so that the reconstructed signal is equal to the orthogonal projection of the original signal onto the reconstruction space, which is the minimal-error approximation to the original signal. However, this requires the sampling space $\mathcal{S}$ to be equal to the reconstruction space $\mathcal{W}$. If the sampling scheme is specified such that $\mathcal{S}$ is not equal to $\mathcal{W}$, then the minimal-error approximation can not be obtained. Our problem therefore is to construct a 'good' approximation of the signal given both a sampling method and a reconstruction method.

The rudimentary constraint we impose on the reconstruction is that if the original signal lies in $\mathcal{W}$, then the reconstruction will be equal to the original signal. We will show that this requirement uniquely determines the reconstructed signal. Furthermore this reconstructed signal is a *consistent reconstruction* of the original signal, namely it has the property that although if the original signal does not lie in $\mathcal{W}$ then it is not equal to the original signal, it nonetheless yields the same measurements.

In [16], Unser and Aldroubi introduce the concept of consistent reconstruction, based on which they develop a new sampling procedure for the special case in which the signals lie in L_2, and where the sampling and reconstruction spaces are not necessarily equal but are both generated by integer translates of appropriately chosen functions.

In this chapter we extend the results of [16] in several ways. First, we expand their results to a broader framework that does not require the sampling and reconstruction spaces to be generated by integer translates, and does not require the signals to lie in L_2, but rather can be applied to arbitrary subspaces of an arbitrary Hilbert space. This framework leads to some new sampling theorems, as well as further insight into the results of [16].

Second, we exploit the new concept of oblique dual frame vectors introduced in Section 5.4 of the previous chapter to develop *redundant* sampling procedures in which the measurements constitute an overcomplete representation of the signal. These measurements correspond to inner products of the signal with a set of linearly dependent vectors that form a frame for $\mathcal{S}$, and reconstruction is obtained using a set of linearly dependent vectors which form a frame for $\mathcal{W}$. Using oblique dual frame vectors we can simplify the

sampling and reconstruction processes while still retaining the flexibility of choosing the spaces almost arbitrarily, due to the extra degrees of freedom offered by the use of frames that allow us to construct frames with prescribed properties [66, 85]. Furthermore, if the measurements are quantized prior to reconstruction, then as we show the average power of the reconstruction error using this redundant procedure can be reduced by as much as the redundancy of the frame in comparison with the nonredundant procedure. This generalizes a similar result of Goyal *et al.* [111] for the case in which the sampling and reconstruction spaces are equal.

Third, building upon a geometric interpretation of the consistent sampling procedures we develop a general framework for constructing signals with prescribed properties. For example, using this framework we can construct a signal with specified odd part and specified local averages, or a signal with specified lowpass coefficients and specified values over a time interval.

In Section 6.2 we consider the sampling framework in detail, and develop a geometric interpretation of the sampling and reconstruction scheme that provides further insight into the problem. In Section 6.3 we consider nonredundant sampling schemes, and derive explicit consistent reconstruction methods. Section 6.4 illustrates the reconstruction in the context of a concrete example. The aliasing error and reconstruction error resulting from our general scheme are analyzed in Section 6.5. In Section 6.6 we use the oblique dual frame vectors to develop a redundant sampling procedure, and show that it can be used to reduce the quantization error when quantizing the measurements prior to reconstruction by as much as the redundancy of the frame in comparison with a nonredundant procedure. Based on our consistent reconstruction algorithms, in Section 6.7 we develop a general framework for constructing signals with prescribed properties.

6.2 Consistent Reconstruction

6.2.1 Consistency Condition

Suppose we are given measurements[1] $c[i]$ of a signal f that lies in an arbitrary Hilbert space $\mathcal{H}$. The measurements are obtained by taking the inner products of the signal with

[1] To be consistent with the notation typically used in the sampling literature, throughout this section we use the notation $c[i]$ to denote the elements of a sequence $c \in l_2$.

a set of sampling vectors $\{s_i\}$ that span a subspace $\mathcal{S} \subseteq \mathcal{H}$, which is referred to as the sampling space, so that $c[i] = \langle s_i, f \rangle$. In the case of nonredundant sampling the vectors form a Riesz basis for $\mathcal{S}$; in the case of redundant sampling the vectors form a frame for $\mathcal{S}$. We construct an approximation $\hat{f}$ of f using a given set of reconstruction vectors $\{w_i\}$ that span a subspace $\mathcal{W} \subseteq \mathcal{H}$, which we refer to as the reconstruction space. In the case of nonredundant sampling the reconstruction vectors form a Riesz basis for $\mathcal{W}$, and in the redundant case the vectors form a frame for $\mathcal{W}$. We do not require the sampling space $\mathcal{S}$ and the reconstruction space $\mathcal{W}$ to be equal. However, we assume that they have equal dimension.

The reconstruction $\hat{f}$ has the form $\hat{f} = \sum_i d[i] w_i$ for for some coefficients $\{d[i]\}$ that are a linear transformation of the measurements $\{c[i]\}$, so that $d = Hc$ for some H. Then, with W and S denoting the set transformations corresponding to the vectors w_i and s_i respectively,

$$\hat{f} = \sum_i d[i] w_i = W d = W H c = W H S^* f. \tag{6.1}$$

The sampling and reconstruction scheme is illustrated in Fig. 6-1.

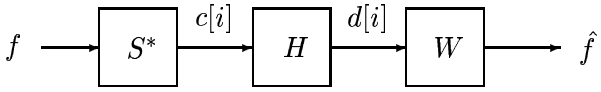


Figure 6-1: General sampling and reconstruction scheme.

If f does not lie in $\mathcal{W}$, then it cannot be perfectly reconstructed using only vectors in $\mathcal{W}$ since $\hat{f}$ given by (6.1) will always lie in $\mathcal{W}$. Therefore, our problem is to choose the transformation H in Fig. 6-1 so that $\hat{f}$ is a good approximation of f. In addition we require that $\hat{f}$ reduces to a perfect reconstruction of f when f lies in $\mathcal{W}$. If $\mathcal{W}$ and $\mathcal{S}^\perp$ are not disjoint, *i.e.*, $\mathcal{W} \cap \mathcal{S}^\perp \neq \{0\}$, then perfect reconstruction for all $f \in \mathcal{W}$ is not possible. For suppose that x is a nonzero signal in $\mathcal{W} \cap \mathcal{S}^\perp$. Then $c[i] = \langle s_i, x \rangle = 0$ for all i, and clearly x can not be reconstructed from the measurements $c[i]$. Consequently, we assume throughout the chapter that $\mathcal{W}$ and $\mathcal{S}^\perp$ are disjoint, and that $\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp$. Subject to this condition, we will show that any $f \in \mathcal{W}$ can be perfectly reconstructed from the measurements $\{c[i]\}$ using reconstruction vectors $\{w_i\}$.

Since we are requiring that $\hat{f} = WHS^*f = f$ for all $f \in \mathcal{W}$ it follows immediately that with $G = WHS^*$, $GGf = Gf$ for any $f \in \mathcal{W}$. Furthermore, since any $x \in \mathcal{H}$ can be expressed as $x = w + v$ with $w \in \mathcal{W}$ and $v \in \mathcal{S}^\perp$ and $S^*v = 0$, $Gx = Gw = w$, so that for any $f \in \mathcal{H}$, $GGf = Gf$. We conclude that G must be a projection operator. To specify G, we need to determine its null space $\mathcal{N}(G)$ and its range space $\mathcal{R}(G)$. Since $G = WHS^*$, $\mathcal{N}(G) \supseteq \mathcal{N}(S^*) = \mathcal{S}^\perp$ and $\mathcal{R}(G) \subseteq \mathcal{R}(W) = \mathcal{W}$. But since $Gf = f$ for all $f \in \mathcal{W}$ we have that $\mathcal{R}(G) = \mathcal{W}$ which immediately implies that $\mathcal{N}(G) = \mathcal{S}^\perp$, so that $G = E_{\mathcal{W}\mathcal{S}^\perp}$. We therefore have the following theorem:

Theorem 6.1. *Let $\{c[i] = \langle s_i, f\rangle\}$ denote measurements of $f \in \mathcal{H}$ with sampling vectors $\{s_i\}$ that span a subspace $\mathcal{S} \subseteq \mathcal{H}$, and let the reconstruction vectors $\{w_i\}$ span a subspace $\mathcal{W} \subseteq \mathcal{H}$ such that $\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp$. Then $\hat{f}$ is a linear reconstruction of f that reduces to a perfect reconstruction for all $f \in \mathcal{W}$ if and only if*

$$\hat{f} = E_{\mathcal{W}\mathcal{S}^\perp} f. \tag{6.2}$$

The reconstruction (6.2) has the additional property that it satisfies the consistency requirement as formulated by Unser and Aldroubi in [16]. A consistent reconstruction $\hat{f}$ of f has the property that if we measure it using the measurement vectors s_i, then the measurements will be equal to the measurements $c[i]$ of f. Since $\hat{f} = E_{\mathcal{W}\mathcal{S}^\perp} f$ it follows immediately that $S^*\hat{f} = S^*E_{\mathcal{W}\mathcal{S}^\perp} f = S^*f$, so that $\hat{f}$ is a consistent reconstruction of f. Furthermore, any consistent reconstruction $\hat{f}$ of f reduces to a perfect reconstruction for $f \in \mathcal{W}$. Indeed, if $f \in \mathcal{W}$ and $\hat{f}$ is a consistent reconstruction of f, then $\langle s_i, \hat{f}\rangle = \langle s_i, f\rangle$ for all i, so that $\langle s_i, f - \hat{f}\rangle = 0$, which implies that $f - \hat{f} \in \mathcal{S}^\perp$. But $f - \hat{f}$ also lies in $\mathcal{W}$, and since $\mathcal{W}$ and $\mathcal{S}^\perp$ are disjoint we conclude that $f = \hat{f}$. We therefore have the following corollary to Theorem 6.1:

Corollary 6.1. *Let $\{c[i] = \langle s_i, f\rangle\}$ denote measurements of $f \in \mathcal{H}$ with sampling vectors $\{s_i\}$ that span a subspace $\mathcal{S} \subseteq \mathcal{H}$, and let the reconstruction vectors $\{w_i\}$ span a subspace $\mathcal{W} \subseteq \mathcal{H}$ such that $\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp$. Then $\hat{f}$ is a consistent linear reconstruction of f if and only if*

$$\hat{f} = E_{\mathcal{W}\mathcal{S}^\perp} f. \tag{6.3}$$

Theorem 6.1 describes the form of the unique consistent reconstruction if it exists, however it does not establish the existence of such a reconstruction. In Sections 6.3 and 6.6 we use our results regarding oblique projections (Section 2.4) and oblique dual frame vectors (Section 5.4) to show that a consistent reconstruction can always be obtained, and we derive explicit reconstruction procedures. This then implies that if $f \in \mathcal{W}$, then f can be perfectly reconstructed from the measurements $c[i]$. Therefore, our results can also be used to generate new sampling theorems that yield perfect reconstruction. We will illustrate these ideas in the context of a concrete example in Section 6.4. In that example $\mathcal{H}$ is the space of length n discrete-time sequences $x[k]$, the reconstruction space $\mathcal{W}$ is the space of length $m = 2m' + 1 < n$ sequences, and the sampling space $\mathcal{S}$ is the space of "bandlimited" sequences in $\mathcal{H}$ so that $x[k] \in \mathcal{S}$ if and only if $X[k] = 0$ for $m' < k < n - m'$, where $X[k]$ is the n point DFT of $x[k]$. Using our framework we obtain a consistent "time-limited" reconstruction of any signal in $\mathcal{H}$, so that the lowpass DFT coefficients of the time-limited sequence and the original sequence are equal. Furthermore, we show that since any signal in $\mathcal{W}$ can be perfectly reconstructed from its samples in $\mathcal{S}$, a time-limited sequence can be reconstructed from a lowpass segment of its DFT transform.

Before proceeding to the detailed methods, in the next section we present a geometric interpretation of the sampling and reconstruction that provide further insight into the problem.

6.2.2 Geometric Interpretation of Sampling and Reconstruction

Let us first consider the case of perfect reconstruction for signals in $\mathcal{W}$. Thus, we would like to determine conditions under which any $f \in \mathcal{W}$ can be reconstructed from the measurements $c[i] = \langle f, s_i \rangle$. We first note that sampling f with measurement vectors in $\mathcal{S}$, is equivalent to sampling the orthogonal projection of f onto $\mathcal{S}$, denoted by $f_{\mathcal{S}} = P_{\mathcal{S}} f$. This follows from the relation

$$\langle s_i, f \rangle = \langle P_{\mathcal{S}} s_i, f \rangle = \langle s_i, P_{\mathcal{S}} f \rangle. \tag{6.4}$$

We may therefore decompose the sampling process into two stages, as illustrated in Fig. 6-2. In the first stage the signal f is orthogonally projected onto the sampling space $\mathcal{S}$, and in the second stage the projected signal $f_{\mathcal{S}}$ is measured. Since $f_{\mathcal{S}} \in \mathcal{S}$ and the vectors s_i span

$\mathcal{S}$, $f_{\mathcal{S}}$ is uniquely determined by the measurements $c[i]$. Therefore, knowing $c[i]$ is equivalent to knowing $f_{\mathcal{S}}$.

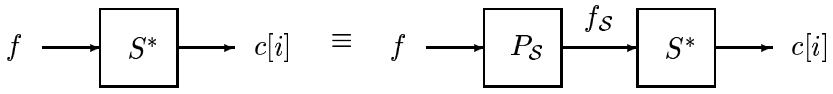


Figure 6-2: Decomposition of the sampling process into two stages.

In view of the interpretation of Fig. 6-2, our problem can be rephrased as follows. Can we reconstruct a signal in $\mathcal{W}$, given the orthogonal projection of the signal onto $\mathcal{S}$, with $\mathcal{W}$ and $\mathcal{S}^\perp$ disjoint? Fig. 6-3(a) depicts the orthogonal projection $f_{\mathcal{S}}$ of an unknown signal $f \in \mathcal{W}$ onto $\mathcal{S}$. The problem is to determine f from this projection. Since the direction of $\mathcal{W}$ is known, there is only one vector in $\mathcal{W}$ whose projection onto $\mathcal{S}$ is $f_{\mathcal{S}}$; this vector is illustrated in Fig. 6-3(b). From this geometrical interpretation we conclude that if $\mathcal{W}$ and $\mathcal{S}^\perp$ are disjoint, then perfect reconstruction of any $f \in \mathcal{W}$ from the measurements $c[i]$ is always possible.

We now discuss consistent reconstruction for signals $f \in \mathcal{H}$. If $\hat{f}$ is a consistent reconstruction of f, then f and $\hat{f}$ have the same measurements: $c[i] = \langle s_i, f\rangle = \langle s_i, \hat{f}\rangle$. From our previous discussion it follows that $f_{\mathcal{S}} = \hat{f}_{\mathcal{S}}$ where $\hat{f}_{\mathcal{S}} = P_{\mathcal{S}}\hat{f}$. Thus, geometrically a consistent reconstruction $\hat{f}$ of f is a signal in $\mathcal{W}$ whose orthogonal projection onto $\mathcal{S}$ is equal to the orthogonal projection of f onto $\mathcal{S}$, as illustrated in Fig. 6-4. Evidently, the consistent reconstruction is unique and always exists. We have seen in Theorem 6.1 that this reconstruction has a nice geometrical interpretation: It is the oblique projection of f onto $\mathcal{W}$ along $\mathcal{S}^\perp$. This interpretation is illustrated in Fig. 6-5, from which it is apparent that $E_{\mathcal{W}\mathcal{S}^\perp}f$ and f have the same orthogonal projection onto $\mathcal{S}$ and consequently yield the same measurements.

In summary, by considering a geometric interpretation of the sampling process and the consistency requirement we have demonstrated that perfect reconstruction for signals in $\mathcal{W}$ is always possible as long as $\mathcal{W}$ and $\mathcal{S}^\perp$ are disjoint, and we illustrated the reconstruction geometrically. We also showed that consistent reconstruction is always possible, and illustrated the reconstruction. It is important to note that the geometric interpretation (and Theorem 6.1) hold irrespective of whether the sampling process is nonredundant or

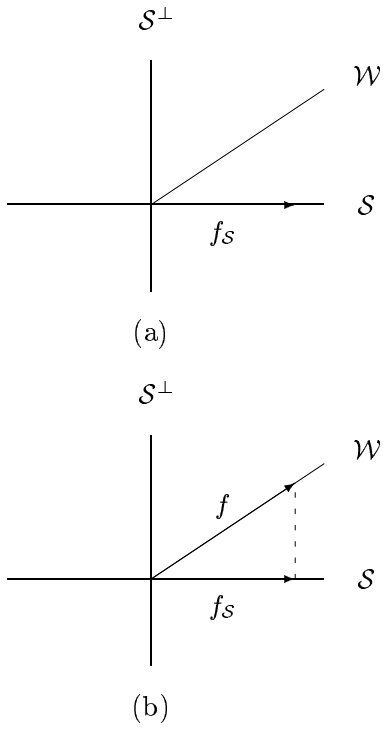


Figure 6-3: Illustration of perfect reconstruction of $f \in \mathcal{W}$ from $f_\mathcal{S} = P_\mathcal{S} f$, with $\mathcal{W}$ and $\mathcal{S}^\perp$ disjoint (a) projection of unknown signal in $\mathcal{W}$ onto $\mathcal{S}$ (b) unique signal in $\mathcal{W}$ with the given orthogonal projection.

redundant. However, the specific reconstruction algorithms will be different in both cases. In the next section we provide mathematical proof of these results for the case of nonredundant sampling, and derive an explicit reconstruction scheme; redundant procedures are considered in Section 6.6.

6.3 Reconstruction From Nonredundant Measurements

Suppose that the sampling vectors $\{s_i\}$ form a Riesz basis for $\mathcal{S}$ and the reconstruction vectors $\{w_i\}$ form a Riesz basis for $\mathcal{W}$. Then we can always find an invertible transformation H such that $G = WHS^* = E_{\mathcal{W}\mathcal{S}^\perp}$, which from Theorem 6.1 implies consistent reconstruction for all $f \in \mathcal{H}$ and perfect reconstruction for all $f \in \mathcal{W}$. Specifically, from Theorem 2.5 it follows that with $H = (S^*W)^{-1}$, $G = WHS = E_{\mathcal{W}\mathcal{S}^\perp}$. Thus, reconstruction is obtained by first transforming the measurements $c[i]$ into "corrected" measurements $d[i]$ corresponding

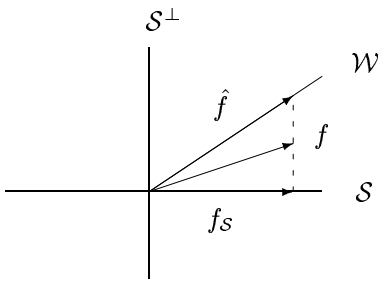


Figure 6-4: Illustration of consistent reconstruction of an arbitrary f from $f_{\mathcal{S}}$, with $\mathcal{W}$ and $\mathcal{S}^\perp$ disjoint.

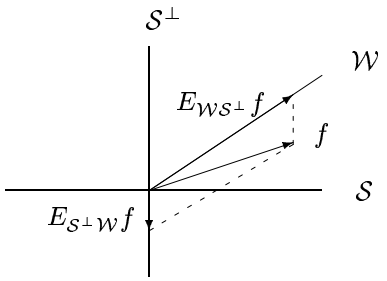


Figure 6-5: Decomposition of f into its components in $\mathcal{W}$ and in $\mathcal{S}^\perp$ given by $E_{\mathcal{W}\mathcal{S}^\perp} f$ and $E_{\mathcal{S}^\perp\mathcal{W}} f$, respectively.

to the sequence $d = Hc = (S^*W)^{-1}c$, which by Lemma 2.1 is well defined. Then

$$\hat{f} = \sum_i d[i] w_i = Wd = W(S^*W)^{-1}S^*f = E_{\mathcal{W}\mathcal{S}^\perp} f. \tag{6.5}$$

The resulting measurement and reconstruction scheme is depicted in Fig. 6-6.

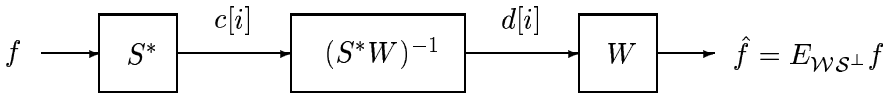


Figure 6-6: Consistent reconstruction of f using sampling vectors s_i and reconstruction vectors w_i, with $\mathcal{W}$ and $\mathcal{S}^\perp$ disjoint.

If $f \in \mathcal{W}$ then $\hat{f} = E_{\mathcal{W}\mathcal{S}^\perp} f = f$, and f can be perfectly reconstructed from the measurements $c[i]$ using the scheme depicted in Fig. 6-6. By choosing different spaces $\mathcal{H}$, $\mathcal{W}$

and $\mathcal{S}$ and using the measurement and reconstruction scheme of Fig. 6-6, we can arrive at a variety of new and interesting perfect reconstruction sampling theorems.

We can interpret the sampling and reconstruction scheme of Fig. 6-6 in terms of a basis expansion for signals in $\mathcal{W}$. Specifically, for any $f \in \mathcal{W}$ we have that $\hat{f} = f$ so that f can be represented as $f = \sum_i d[i]w_i$. The coefficients $d[i]$ can be expressed as $d[i] = \langle v_i, f \rangle$ where the vectors $v_i \in \mathcal{S}$ correspond to the set transformation $V = S(W^*S)^{-1}$, and are biorthogonal to w_i: $\langle v_i, w_k \rangle = \delta_{ik}$. This follows immediately from $V^*W = (S^*W)^{-1}S^*W = I$. Therefore Fig. 6-6 provides an explicit method for constructing basis vectors for an arbitrary space $\mathcal{S}$ with $\mathcal{W}$ and $\mathcal{S}^\perp$ disjoint, that are biorthogonal to the basis vectors w_i.

We summarize our results regarding nonredundant sampling in the following theorem:

Theorem 6.2 (Nonredundant sampling and reconstruction). *Let $\{c_i = \langle s_i, f \rangle\}$ denote measurements of a signal $f \in \mathcal{H}$ with sampling vectors $\{s_i\}$ that form a Riesz basis for a subspace $\mathcal{S} \subseteq \mathcal{H}$. Let $\{w_i\}$ denote a set of reconstruction vectors that form a Riesz basis for a subspace $\mathcal{W} \subseteq \mathcal{H}$, with $\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp$. Then*

1. *Any $f \in \mathcal{W}$ can be perfectly reconstructed from the measurements $c[i]$ using the reconstruction vectors w_i as $f = \sum_i d[i]w_i$ with $d = (S^*W)^{-1}c$. In addition,*
   (a) *$d[i] = \langle v_i, f \rangle$ where the vectors $\{v_i\}$ are the unique vectors in $\mathcal{S}$ biorthogonal to the vectors $\{w_i\}$;*
   (b) *the coefficients $d[i]$ are unique.*
2. *Any $f \in \mathcal{H}$ can be consistently reconstructed from the measurements $c[i]$ using the reconstruction vectors w_i as $\hat{f} = \sum_i d[i]w_i$ with $d = (S^*W)^{-1}c$. In addition,*
   (a) *the consistent reconstruction $\hat{f}$ is unique;*
   (b) *the coefficients $d[i]$ are unique.*

6.4 Bandlimited Sampling of Time-Limited Sequences

To illustrate the details of the sampling and reconstruction scheme of Fig. 6-6, we now consider in detail the example outlined in Section 6.2.1. $\mathcal{H}$ is the space of sequences $x[k]$ such that $x[k] = 0$ for $k < 0, k \geq n$, $\mathcal{W}$ is the space of sequences $x[k]$ such that $x[k] = 0$ for $k < 0, k \geq m$ where $m = 2m' + 1 < n$, and $\mathcal{S}$ is the space of "bandlimited" sequences $x[k]$

such that $X[k] = 0$ for $m' < k < n - m'$, where $X[k], 0 \leq k \leq n-1$ denotes the n point DFT of $x[k]$. The bases for $\mathcal{S}$ and $\mathcal{W}$ are chosen as the sequences $s_i[k], 0 \leq i \leq m-1$ and $w_i[k], 0 \leq i \leq m-1$ respectively, given by $s_i[k] = e^{j2\pi(i-m')k/n}$ for $0 \leq k \leq n-1$ and 0 otherwise, and $w_i[k] = \delta[i-k]$.

The measurements $c[i], 0 \leq i \leq m-1$ of an arbitrary sequence $f \in \mathcal{H}$ are equal to

$$c[i] = \langle s_i, f \rangle = \sum_{k=0}^{n-1} s_i^*[k] f[k] = \sum_{k=0}^{n-1} f[k] e^{-j2\pi(i-m')k/n} = F[((i-m'))_n], \tag{6.6}$$

where $F[k], 0 \leq k \leq n-1$ is the n point DFT of $f[k]$, and $((p))_n = p \bmod n$. Thus, the measurements $c[i]$ are the m lowpass DFT coefficients of the n point DFT of f. To obtain a consistent reconstruction of f from $c[i]$ we need to determine $(S^*W)^{-1}$. The jlth element of the $m \times m$ matrix S^*W is

$$\langle s_j, w_l \rangle = \sum_{k=0}^{n-1} s_j^*[k] w_l[k] = s_j^*[l] = Z^{jl} B^l, \tag{6.7}$$

where $Z = e^{-j2\pi/n}$ and $B = e^{j2\pi m'/n}$. We can therefore express S^*W in the form

$$S^*W = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & Z & Z^2 & \cdots & Z^{m-1} \\ & & \vdots & & \\ 1 & Z^{m-1} & Z^{2(m-1)} & \cdots & Z^{(m-1)^2} \end{bmatrix} D. \tag{6.8}$$

Eq. (6.8) is the product of a Vandermonde matrix and a diagonal matrix D with nonzero diagonal elements $B^l, 0 \leq l \leq m-1$. Therefore, S^*W is always invertible which implies by Lemma 2.1 that $\mathcal{W}$ and $\mathcal{S}^\perp$ are disjoint. From Theorem 6.2 it then follows that consistent reconstruction is possible for all f. We can compute the inverse of S^*W using any of the formulas for the inverse of a Vandermonde matrix (see *e.g.,* [126, 127]). The corrected measurements $d[i]$ are then given by the elements of $d = (S^*W)^{-1}c$ where c is the vector with elements $c[i]$ given by (6.6), and $\hat{f}[k] = \sum_{i=0}^{n-1} w_i[k] d[i] = d_k$ for $0 \leq k \leq m-1$ and 0 otherwise. The consistency requirement implies that $\hat{F}[((k-m'))_n] = F[((k-m'))_n]$ for $0 \leq k \leq m-1$, where $\hat{F}[k]$ is the n point DFT of $\hat{f}[k]$. Thus $\hat{f}$ is a "time-limited" sequence that has the same lowpass DFT coefficients as f.

In Section 6.7 we develop a systematic method for constructing signals in a subspace

$\mathcal{W}$ with specified properties in a subspace $\mathcal{S}$. We also consider the more general problem of constructing a signal in $\mathcal{H}$ with specified properties in both $\mathcal{W}$ and $\mathcal{S}$. Using these methods we can generalize our construction here to produce a signal with specified lowpass coefficients *and* specified values on a given time interval.

Now, suppose that f is a length m sequence in $\mathcal{W}$, and we are given m lowpass DFT coefficients $F[((k-m'))_n]$, $0 \leq k \leq m-1$. We can then perfectly reconstruct f from these coefficients using the method described above. This implies the intuitive result that a time-limited discrete-time sequence can be reconstructed from a lowpass segment of its DFT transform. This result is the analogue for the finite length discrete-time case of Papoulis' theorem [128], which implies that a time-limited function can be recovered from a lowpass segment of its Fourier transform. The reconstruction based on Papoulis' theorem is typically obtained using iterative algorithms such as those discussed in [128, 129]. By choosing appropriate sampling and reconstruction vectors in the general scheme of Fig. 6-6, we obtained a finite length discrete-time version of this theorem together with a simple non-iterative reconstruction method. This example illustrates the type of procedure that might be followed in using our framework to generate new sampling theorems.

6.5 Aliasing and Error Bounds

Since in general f does not lie in $\mathcal{W}$, the reconstruction scheme of Fig. 6-6 may result in aliasing at the output. Intuitively, aliasing will occur when components of f that lie out of $\mathcal{W}$ are aliased into $\hat{f}$. A very nice and intuitive way to think about aliasing was proposed in [130] in the context of multiresolution spaces in terms of the norm of the "out-of-space" component. Let Γ denote the sampling operator defined by $\hat{f} = \Gamma f$ where f is the original signal and $\hat{f}$ is the reconstructed signal. Then the aliasing norm is defined as [130, 131]

$$A_\Gamma = \sup_{f \in \mathcal{W}^\perp} \frac{\|\Gamma f\|}{\|f\|}. \tag{6.9}$$

As we expect intuitively, $A_\Gamma = 0$ if $\Gamma f = 0$ for all $f \in \mathcal{W}^\perp$.

In our case $\Gamma = E_{\mathcal{W}\mathcal{S}^\perp}$, and

$$A_\Gamma = \sup_{f \in \mathcal{W}^\perp} \frac{\|E_{\mathcal{W}\mathcal{S}^\perp} f\|}{\|f\|}. \tag{6.10}$$

From (6.10) we conclude that $A_\Gamma = 0$ only if $E_{\mathcal{WS}^\perp} = 0$ for all $f \in \mathcal{W}^\perp$ which implies that $\mathcal{S} = \mathcal{W}$. To avoid aliasing when $\mathcal{S} \neq \mathcal{W}$, we can first orthogonally project f onto $\mathcal{W}$, and then measure the projection, so that the part of the signal that lies in $\mathcal{W}^\perp$ will not contribute to the reconstruction $\hat{f}$. The measurements are then given by $c = S^* P_\mathcal{W} f$, or $c[i] = \langle t_i, f \rangle$ where $t_i = P_\mathcal{W} s_i$ and consequently $t_i \in \mathcal{W}$; as we expect the effective sampling space is equal to the reconstruction space. When the spaces are not equal, we can obtain a bound on the aliasing norm A_Γ in terms of the angle $\theta_{\mathcal{WS}}$ between the spaces $\mathcal{W}$ and $\mathcal{S}$, defined in (5.41) [16]. Specifically, using (5.42) we have that

$$A_\Gamma \leq \frac{1}{\cos(\theta_{\mathcal{WS}})}, \tag{6.11}$$

where $\cos(\theta_{\mathcal{WS}}) = \inf_{f \in \mathcal{W}, \|f\|=1} \|P_\mathcal{S} f\|$. As we expect, the bound decreases as the angle between the spaces $\mathcal{W}$ and $\mathcal{S}$ decreases, in which case $\mathcal{S}^\perp$ is "close" to $\mathcal{W}^\perp$.

To decrease the aliasing norm when $\mathcal{W}$ and $\mathcal{S}$ are not equal, we may oversample the signal using a larger set of sampling and reconstruction vectors. Intuitively, we can reduce the aliasing norm by effectively decreasing the angle between $\mathcal{W}$ and $\mathcal{S}$. This can be done by expanding $\mathcal{W}$ and $\mathcal{S}$ to larger spaces $\mathcal{W}'$ and $\mathcal{S}'$ such that $\mathcal{W} \subset \mathcal{W}'$ and $\mathcal{S} \subset \mathcal{S}'$, and such that $\cos(\theta_{\mathcal{W}'\mathcal{S}'}) > \cos(\theta_{\mathcal{WS}})$. Mathematically, by enlarging the reconstruction space the supremum in (6.10) can not increase. We then oversample the signal using sampling vectors $\{s_i'\}$ that span the larger space $\mathcal{S}'$ and include the sampling vectors $\{s_i\}$. To reconstruct the signal, we use reconstruction vectors $\{w_i'\}$ that span $\mathcal{W}'$ and include the reconstruction vectors $\{w_i\}$. As long as $\mathcal{W}'$ and $(\mathcal{S}')^\perp$ are disjoint, we can still perfectly reconstruct any signal $f \in \mathcal{W}$ from the new measurements, and at the same time decrease the aliasing norm when measuring signals that do not lie in $\mathcal{W}$.

The reconstruction error using the general scheme of Fig. 6-6 can be bounded based on results derived in [16],

$$\|f - P_\mathcal{W} f\| \leq \|f - E_{\mathcal{WS}^\perp} f\| \leq \frac{1}{\cos(\theta_{\mathcal{WS}})} \|f - P_\mathcal{W} f\|, \tag{6.12}$$

where $\|f - P_\mathcal{W} f\|$ is the minimal norm of the reconstruction error corresponding to the case in which $\mathcal{W} = \mathcal{S}$. From (6.12) we see that there is a price to pay for the flexibility offered by choosing the sampling space (almost) arbitrarily: The norm of the reconstruction error for input signals that do not lie in the reconstruction space is increased. However, in many

practical applications this increase in error is very small [17, 82, 83, 84].

Since the error bounds presented in this section depend only on the reconstructed signal, they also hold in the case of redundant sampling and reconstruction, which we discuss in the next section.

6.6 Reconstruction From Redundant Measurements

6.6.1 Reconstruction Scheme

In Section 6.3 we considered consistent reconstruction of $f \in \mathcal{H}$ from a nonredundant set of measurements, given by inner products of f with a set of basis vectors. We now consider the problem of consistent reconstruction from *redundant* measurements, given by inner products of f with a set of frame vectors. Throughout this section we assume that $\mathcal{S}$ and $\mathcal{W}$ both have finite dimension m; the results extend to the infinite-dimensional case as well.

Suppose we are given a set of $n > m$ measurements $\tilde{c}[i] = \langle x_i, f \rangle$ of a signal $f \in \mathcal{H}$, where the sampling vectors $\{x_i, 1 \leq i \leq n\}$ form a frame for $\mathcal{S}$, and reconstruction is obtained using reconstruction vectors $\{y_i, 1 \leq i \leq n\}$ that form a frame for $\mathcal{W}$. From Theorem 6.1 it follows that to obtain a consistent reconstruction for all $f \in \mathcal{H}$ and perfect reconstruction for all $f \in \mathcal{W}$ we need to find a transformation H such that $G = YHX^* = E_{\mathcal{W}\mathcal{S}^\perp}$, where X and Y are the set transformations corresponding to the vectors x_i and y_i, respectively. Using our construction of oblique dual frame vectors presented in Section 5.4.1, we now show that such a transformation always exists.

Specifically, from Proposition 5.2 and the properties of the oblique pseudoinverse developed in Section 2.7.2, it follows that with $H = (X^*Y)^\dagger$, $G = YHX^* = YY^{\#}_{\mathcal{V}\mathcal{S}^\perp} = E_{\mathcal{W}\mathcal{S}^\perp}$, where $\mathcal{V} = \mathcal{N}(Y)^\perp$ and $Y^{\#}_{\mathcal{V}\mathcal{S}^\perp}$ is the *oblique pseudoinverse* of Y on $\mathcal{V}$ along $\mathcal{S}^\perp$. Thus, reconstruction is obtained by first transforming the measurements $\tilde{c}[i]$ into "corrected" measurements $\tilde{d}[i]$ corresponding to the sequence $\tilde{d} = (X^*Y)^\dagger \tilde{c} = Y^{\#}_{\mathcal{V}\mathcal{S}^\perp} f$ so that $\tilde{d}[i] = \langle \tilde{y}_i, f \rangle$ where the vectors $\tilde{y}_i$ are the oblique dual frame vectors of y_i on $\mathcal{S}$. Then

$$\hat{f} = Y\tilde{d} = YY^{\#}_{\mathcal{V}\mathcal{S}^\perp} f = E_{\mathcal{W}\mathcal{S}^\perp} f. \tag{6.13}$$

The resulting measurement and reconstruction scheme is depicted in Fig. 6-6.

From the properties of the oblique dual frame vectors, developed in Section 5.4.2, it

follows that although the coefficients $\tilde{d}[i]$ are not unique, they have the property that from all possible coefficients $d[i]$ such that $\hat{f} = \sum_{i=1}^{n} d[i]y_i$, they have the minimal l_2-norm.

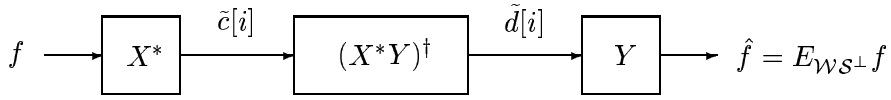


Figure 6-7: Consistent reconstruction of f using redundant sampling vectors x_i and redundant reconstruction vectors y_i.

Now, any frame $\{y_i, 1 \leq i \leq n\}$ for $\mathcal{W}$ can be expressed as $Y = WZ$ where W is a set transformation corresponding to an arbitrary basis for $\mathcal{W}$, and Z: $\mathbb{C}^m \to \mathbb{C}^n$ satisfies $ZZ^\dagger = I_m$. Then, from Proposition 5.2

$$Y^{\#}_{\mathcal{V}\mathcal{S}^\perp} = Z^\dagger (S^*W)^{-1} S^*, \tag{6.14}$$

where S is a set transformation corresponding to an arbitrary basis for $\mathcal{S}$. From (6.14) it follows that we can obtain the redundant corrected measurements $\tilde{d}[i]$ directly from the nonredundant corrected measurements $d = (S^*W)^{-1}S^*f = (S^*W)^{-1}c$, via $\tilde{d} = Z^\dagger d$, where $c[i] = \langle s_i, f\rangle$ are the nonredundant measurements obtained using the vectors s_i. This interpretation is illustrated in Fig. 6-8.

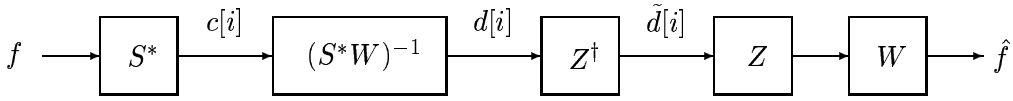


Figure 6-8: Equivalent form of Fig. 6-7.

We summarize our results regarding redundant sampling in the following theorem:

Theorem 6.3 (Redundant sampling and reconstruction). *Let $\{\tilde{c}_i = \langle x_i, f\rangle, 1 \leq i \leq n\}$ denote redundant measurements of a signal $f \in \mathcal{H}$ with sampling vectors $\{x_i, 1 \leq i \leq n\}$ that form a frame for an m-dimensional subspace $\mathcal{S} \subseteq \mathcal{H}$. Let $\{y_i, 1 \leq i \leq n\}$ denote a set of reconstruction vectors that form a frame for an m-dimensional subspace $\mathcal{W} \subseteq \mathcal{H}$, with $\mathcal{W} \oplus \mathcal{S}^\perp$. Then*

1. *Any $f \in \mathcal{W}$ can be perfectly reconstructed from the measurements $\tilde{c}[i]$ using the reconstruction vectors y_i as $f = \sum_{i=1}^{n} \tilde{d}[i] y_i$ with $\tilde{d} = (X^*Y)^\dagger c$. In addition,*

   (a) *the coefficients $\tilde{d}[i]$ are not unique;*

   (b) *the coefficients $\tilde{d}[i]$ have minimal norm among all possible coefficients $d[i]$ such that $f = \sum_{i=1}^{n} d[i] y_i$;*

   (c) *$\tilde{d}[i] = \langle \tilde{y}_i, f \rangle$ where the vectors $\{\tilde{y}_i, 1 \leq i \leq n\}$ are the oblique dual frame vectors of y_i on $\mathcal{S}$.*

2. *Any $f \in \mathcal{H}$ can be consistently reconstructed from the measurements $\tilde{c}[i]$ using the reconstruction vectors y_i as $\hat{f} = \sum_{i=1}^{n} \tilde{d}[i] y_i$ with $\tilde{d} = (X^*Y)^\dagger c$. In addition,*

   (a) *the consistent reconstruction $\hat{f}$ is unique;*

   (b) *the coefficients $\tilde{d}[i]$ are not unique;*

   (c) *the coefficients $\tilde{d}[i]$ have minimal norm among all possible coefficients such that $\hat{f} = \sum_{i=1}^{n} d[i] y_i$.*

6.6.2 Reducing Quantization Error

One of the reasons for using redundant measurements is to reduce the average power of the quantization error, when quantizing the corrected measurements prior to reconstruction. If the sampling and reconstruction spaces are equal, then $\hat{f} = P_{\mathcal{W}} f$ is the unique consistent reconstruction of $f \in \mathcal{H}$. Since we can express this reconstruction as $P_{\mathcal{W}} f = \sum_i \langle y_i, f \rangle y_i$ where the vectors y_i form a normalized tight frame for $\mathcal{W}$, we can consistently reconstruct f using the vectors y_i, where the corrected measurements are $\tilde{d}[i] = \langle y_i, f \rangle$. Alternatively, we can use a nonredundant scheme where the reconstruction vectors w_i form an orthonormal basis for $\mathcal{W}$, and the corrected measurements are $d[i] = \langle w_i, f \rangle$. Suppose now we quantize the measurements $\tilde{d}[i]$ and $d[i]$ prior to reconstruction. Then using the redundant procedure, *i.e.,* quantizing the measurements $\tilde{d}[i]$, it is well known that we can reduce the quantization error by the redundancy $r = n/m$ of the frame [69, 111] in comparison with quantizing the measurements $d[i]$. We now extend this result to the case where the sampling and reconstruction spaces are not constrained to be equal. In particular, we show that we can choose a tight frame y_i for $\mathcal{W}$ such that when using the redundant sampling procedure

of Figs. 6-7 and 6-8 we can reduce the average power of the reconstruction error by r, in comparison with the nonredundant scheme of Fig. 6-6.

Let $\{w_i, 1 \leq i \leq m\}$ denote a set of reconstruction vectors that form an orthonormal basis for $\mathcal{W}$, and let $\{s_i, 1 \leq i \leq m\}$ denote a set of sampling vectors that form a basis for $\mathcal{S}$. Let $c[i] = \langle s_i, f \rangle$ denote nonredundant measurements of a signal f. From Theorem 6.2, the consistent reconstruction $\hat{f}$ of f is obtained using the corrected measurements $d[i]$ corresponding to $d = (S^*W)^{-1}c$, which can be expressed as $d[i] = \langle v_i, f \rangle$, where $\{v_i, 1 \leq i \leq m\}$ are the vectors corresponding to the set transformation $V = S(W^*S)^{-1}$, and are biorthogonal to the vectors w_i. Thus,

$$\hat{f} = \sum_{i=1}^{m} \langle v_i, f \rangle w_i = \sum_{i=1}^{m} q[i] \langle \overline{v}_i, f \rangle \overline{w}_i, \tag{6.15}$$

where $q[i] = \sqrt{a[i]b[i]}$, $a[i] = \langle w_i, w_i \rangle = 1$, $b[i] = \langle v_i, v_i \rangle$, $\overline{w}_i = w_i/\sqrt{a[i]}$, and $\overline{v}_i = v_i/\sqrt{b[i]}$.

Assume we quantize the normalized measurements $\overline{d}[i] = \langle \overline{v}_i, f \rangle$ prior to reconstruction, and model the quantization error as an additive zero-mean white noise source, so that the quantized measurements are $\overline{d}[i] + e[i]$ where $E(e[i]e[j]) = \sigma^2 \delta_{ij}$. Then the reconstruction error is $\epsilon = \sum_{i=1}^{m} q[i]e[i]\overline{w}_i$ and the average power of the reconstruction error, denoted by D, is

$$D = E(\|\epsilon\|^2) = \sigma^2 \sum_{i=1}^{m} q^2[i] = \sigma^2 \sum_{i=1}^{m} b[i]. \tag{6.16}$$

Note, that D does not depend on the particular choice of orthonormal basis vectors w_i. For suppose that the vectors w_i' corresponding to the set transformation W' form a different orthonormal basis for $\mathcal{W}$. Then $W' = W\mathbf{U}$ for some unitary matrix $\mathbf{U}$. The biorthogonal vectors are then the vectors v_i' corresponding to the set transformation $V' = V\mathbf{U} = S(W^*S)^{-1}\mathbf{U}$. But then $\sum_i \langle v_i', v_i' \rangle = \mathrm{Tr}(V\mathbf{U}\mathbf{U}^*V^*) = \mathrm{Tr}(VV^*)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of the corresponding matrix, and the average power of the reconstruction error is again equal to D.

Suppose now we use a redundant procedure so that we reconstruct the signal using a β-scaled tight frame $\{y_i, 1 \leq i \leq n\}$ for $\mathcal{W}$, with redundancy $r = n/m$. Then $Y = WZ$ for some $Z\colon\ \mathbb{C}^m \to \mathbb{C}^n$ such that $ZZ^* = \beta^2 I_m$. From Theorem 6.3 and Proposition 5.2 it follows that the sampling vectors leading to consistent reconstruction correspond to the set

transformation $X = (Y^{\#}_{\mathcal{V}\mathcal{S}^{\perp}})^* = (1/\beta^2)VZ$, so that in this case

$$\hat{f} = \sum_{i=1}^{n} \langle x_i, f \rangle y_i = \sum_{i=1}^{n} \tilde{q}[i] \langle \overline{x}_i, f \rangle \overline{y}_i, \tag{6.17}$$

where $\tilde{q}[i] = \sqrt{\tilde{a}[i]\tilde{b}[i]}$, $\tilde{a}[i] = \langle y_i, y_i \rangle$, $\tilde{b}[i] = \langle x_i, x_i \rangle$, $\overline{y}_i = y_i/\sqrt{\tilde{a}[i]}$, and $\overline{x}_i = x_i/\sqrt{\tilde{b}[i]}$. If we quantize the normalized redundant measurements $\langle \overline{x}_i, f \rangle$ and model the quantization error as before, then the average power of the reconstruction error using the redundant procedure, denoted by $\tilde{D}$, is

$$\tilde{D} = \sigma^2 \sum_{i=1}^{n} \tilde{q}^2[i] = \sigma^2 \sum_{i=1}^{n} \tilde{a}[i]\tilde{b}[i]. \tag{6.18}$$

We now show that we can choose a tight frame y_i such that $\tilde{D} = (m/n)D = (1/r)D$.

Let $Y = W\tilde{\mathcal{F}}$, where $\tilde{\mathcal{F}}$ is an $m \times n$ matrix whose rows are equal to the first m rows of the $n \times n$ Fourier matrix $\mathcal{F}$ with elements $1/\sqrt{n}e^{-j2\pi kl/n}$. Since $YY^* = P_{\mathcal{W}}$, the vectors y_i corresponding to Y form a normalized tight frame for $\mathcal{W}$. The oblique dual frame vectors x_i of y_i on $\mathcal{S}$ are the vectors corresponding to $X = V\tilde{\mathcal{F}}$. We now show that for this choice of sampling and reconstruction vectors $\tilde{D} = (1/r)D$. Let f_i denote the ith column of $\tilde{\mathcal{F}}$. From the definition of $\tilde{\mathcal{F}}$, $\langle f_i, f_i \rangle = m/n$ for all i so that,

$$a_i = \langle y_i, y_i \rangle = \langle Wf_i, Wf_i \rangle = \langle f_i, f_i \rangle = \frac{m}{n}, \tag{6.19}$$

since $W^*W = I_m$, and (6.18) reduces to

$$\tilde{D} = \sigma^2 \frac{m}{n} \sum_{i=1}^{n} \tilde{b}[i]. \tag{6.20}$$

Now,

$$\sum_{i=1}^{n} \tilde{b}[i] = \mathrm{Tr}(X^*X) = \mathrm{Tr}(V^*V) = \sum_{i=1}^{m} b[i]. \tag{6.21}$$

Substituting (6.21) into (6.20), and comparing with (6.16) we conclude that $\tilde{D} = (m/n)D$.

We note that the average power of the reconstruction error is the same if we choose $Y = \beta W\tilde{\mathcal{F}}$ for any $\beta > 0$, in which case the vectors y_i form a β-scaled tight frame for $\mathcal{W}$.

Therefore, to reduce the quantization error in the sampling and reconstruction scheme of Fig. 6-6, we propose the following. Instead of directly quantizing the measurements $d[i]$ in Fig. 6-6, we first take the n point DFT of the length m sequence of measurements $d[i]$, and then quantize the DFT coefficients. The reconstructed signal is then a linear combination of the reconstruction vectors w_i, where the coefficients are the first m values of the inverse DFT transform of the quantized DFT coefficients, as depicted in Fig. 6-9. If we take out the quantizer in Fig. 6-9, then $\hat{f} = E_{\mathcal{W}\mathcal{S}^\perp} f$ as in Fig. 6-6. However, using the redundant sampling scheme of Fig. 6-9 the average power of the quantization error is reduced by the redundancy $r = n/m$ in comparison with a nonredundant scheme.

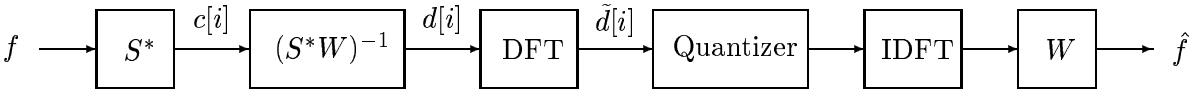


Figure 6-9: Reconstruction of f from quantized measurements using a redundant sampling scheme.

There are many other choices of frame vectors $\{y_i\}$ for $\mathcal{W}$ and corresponding oblique dual frame vectors $\{x_i\}$ on $\mathcal{S}$, that lead to reduction by a factor of r in the average power of the reconstruction error. In particular we have the following theorem.

Theorem 6.4. *Let $\{w_i, 1 \le i \le m\}$ denote an orthonormal basis for $\mathcal{W} \subseteq \mathcal{H}$, and let $\{v_i, 1 \le i \le m\}$ denote the biorthogonal basis for $\mathcal{S} \subseteq \mathcal{H}$, with $\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp$. Let $\{y_i, 1 \le i \le n\}$ denote a frame for $\mathcal{W}$, and let $\{x_i, 1 \le i \le n\}$ denote the oblique dual frame vectors of y_i on $\mathcal{S}$. Suppose we quantize the coefficients in the frame and basis expansions and model the quantization error as additive zero mean white noise. We consider a 'good' oblique frame expansion $\{y_i, x_i\}$ as one for which the average power of the reconstruction error is reduced by the redundancy $r = n/m$ in comparison with the basis expansion $\{w_i, v_i\}$. Let $\mathcal{F}$ denote the $n \times n$ Fourier matrix, and let $\tilde{\mathcal{F}}$ denote the first m rows of $\mathcal{F}$. Then*

1. *The frame vectors corresponding to $Y = \beta W \tilde{\mathcal{F}}$, $X = (1/\beta) V \tilde{\mathcal{F}}$ where $\beta > 0$, form a good oblique frame expansion;*

2. *The frame vectors corresponding to $Y = \beta W \tilde{\mathcal{F}} T$, $X = (1/\beta) V \tilde{\mathcal{F}} T$ where T is a unitary circulant matrix and $\beta > 0$, form a 'good' oblique frame expansion.*

Proof: We already proved the first part of the theorem; it remains to prove the second part. Since the choice of β does not affect the derivation we assume for simplicity that $\beta = 1$. Using Proposition 5.2 we can immediately verify that $X = V\tilde{\mathcal{F}}T$ is in fact the oblique dual frame operator on $\mathcal{S}$ of $Y = W\tilde{\mathcal{F}}T$. Next, since T is a circulant matrix it is diagonalized by $\mathcal{F}^*$ [27], so we can express T as $T = \mathcal{F}^*\Lambda\mathcal{F}$ where Λ is a diagonal matrix with diagonal elements λ_i. Since T is also unitary, $|\lambda_i| = 1$ for all i. Then,

$$Y = W\tilde{\mathcal{F}}T = W\tilde{\mathcal{F}}\mathcal{F}^*\Lambda\mathcal{F} = W\tilde{I}\Lambda\mathcal{F}, \tag{6.22}$$

where $\tilde{I} = [I_m \ 0]$, and

$$Y^*Y = \mathcal{F}^*\Lambda^*\tilde{I}^*\tilde{I}\Lambda\mathcal{F} = \mathcal{F}^*\tilde{I}^*\tilde{I}\mathcal{F} = \tilde{\mathcal{F}}^*\tilde{\mathcal{F}}. \tag{6.23}$$

Combining (6.23) and (6.19), we conclude that $\langle y_i, y_i\rangle = m/n$ for all i. From (6.18) we have that the average power of the reconstruction error is given by $\tilde{D} = \sigma^2(m/n)\sum_{i=1}^n \langle x_i, x_i\rangle$. Now, $X = V\tilde{\mathcal{F}}T = V\tilde{I}\Lambda\mathcal{F}$ so that

$$\sum_{i=1}^{n}\langle x_i, x_i\rangle = \mathrm{Tr}(X^*X) = \mathrm{Tr}(\mathcal{F}^*\Lambda^*\tilde{I}^*V^*V\tilde{I}\Lambda\mathcal{F}) = \mathrm{Tr}(V^*V), \tag{6.24}$$

and $\tilde{D} = (m/n)D$ where D is the average power of the reconstruction error using the basis expansion consisting of the orthonormal vectors w_i and the biorthogonal vectors v_i. □

Based on results derived in [27, 26] we can show that Theorem 6.4 still holds when we replace $\mathcal{F}$ by a generalized Fourier transform matrix defined on a direct product of cyclic groups, and replace T by a real unitary permuted matrix. This is because such a permuted matrix is diagonalized by a generalized Fourier transform matrix [27], and the magnitude of the elements of an $n \times n$ generalized Fourier transform matrix are all equal $1/\sqrt{n}$.

In the special case in which $\mathcal{W} = \mathcal{S}$ Goyal *et al.* [111] proved that it is possible to choose a tight frame for $\mathcal{W}$ such that the average power of the reconstruction error is reduced by r in comparison with an orthonormal basis expansion. Theorem 6.4 extends this result to the case in which $\mathcal{W}$ is not necessarily equal to $\mathcal{S}$.

6.7 Constructing Signals With Prescribed Properties

A potential class of interesting applications of the consistent sampling procedures we developed in the previous sections is to the problem of constructing signals with prescribed properties that can be described in terms of inner products of the signal with a set of vectors. For example, we may consider constructing an odd signal with specified local averages, or constructing a signal with specified odd part *and* specified local averages. Exploiting the results we derived in the context of consistent reconstruction, in this section we develop a general framework for constructing signals of this form.

We first consider the simpler case, in which we wish to construct a signal f to lie in a subspace $\mathcal{W}$, and to have some additional properties in a subspace $\mathcal{S}$ that can be described in terms of a set of mathematical constraints of the form $\langle s_i, f\rangle$ for a set of vectors s_i that span $\mathcal{S}$. We then consider the problem of constructing a signal f with properties in two subspaces $\mathcal{W}$ and $\mathcal{S}$ that can be described in terms of mathematical constraints of the form $\langle s_i, f\rangle$ for a set of vectors s_i that span $\mathcal{S}$, and $\langle w_i, f\rangle$ for a set of vectors w_i that span $\mathcal{W}$.

Throughout this section we assume for simplicity that the constraints are nonredundant, so that each set of vectors $\{s_i\}$ and $\{w_i\}$ is a linearly independent set. We further assume that the vectors s_i form a Riesz basis for $\mathcal{S}$ and the vectors w_i from a Riesz basis for $\mathcal{W}$. Using the construction of oblique dual frame vectors the results in this section extend in a straightforward manner to the redundant case as well.

Our first problem can be solved immediately by noting that it is equivalent to a consistent reconstruction problem. Specifically, let $c[i] = \langle s_i, f\rangle$ denote the constraints on the signal f. Then the problem is to construct a signal $f \in \mathcal{W}$ so that its measurements taken with respect to the sampling vectors s_i are equal to $c[i]$. If $\mathcal{S}$ and $\mathcal{W}^\perp$ are disjoint, then the unique signal f follows immediately from Theorem 6.2,

$$f = W(S^*W)^{-1}c, \tag{6.25}$$

where W is a set transformation corresponding to a Riesz basis for $\mathcal{W}$, and S is the set transformation corresponding to the vectors s_i.

Next, suppose that we want to construct a signal f with specific properties in two disjoint spaces $\mathcal{W}$ and $\mathcal{S}$, *i.e.,* we want to construct f such that $\langle s_i, f\rangle = c[i]$ and $\langle w_i, f\rangle = d[i]$. In view of the geometric interpretation of Fig. 6-2 it follows that constructing f such that

$\langle s_i, f\rangle = c[i]$ and $\langle w_i, f\rangle = d[i]$ is equivalent to constructing f to have a specified orthogonal projection $f_{\mathcal{S}}$ onto $\mathcal{S}$ and a specified orthogonal projection $f_{\mathcal{W}}$ onto $\mathcal{W}$. Fig. 6-3(a) depicts the orthogonal projections of an unknown signal f onto $\mathcal{S}$ and $\mathcal{W}$. The problem then is to construct a signal f with these orthogonal projections. With $\mathcal{U} = \mathcal{W} \oplus \mathcal{S}$, it is obvious that f can be arbitrary on $\mathcal{U}^{\perp}$. However, there is a unique vector $f \in \mathcal{U}$ compatible with the given projections; this vector is illustrated in Fig. 6-3(b). From this geometrical interpretation we conclude that for $\mathcal{W}$ and $\mathcal{S}$ disjoint, we can always construct a signal with the desired properties. Furthermore, the orthogonal projection of this signal onto $\mathcal{U}$ is unique.

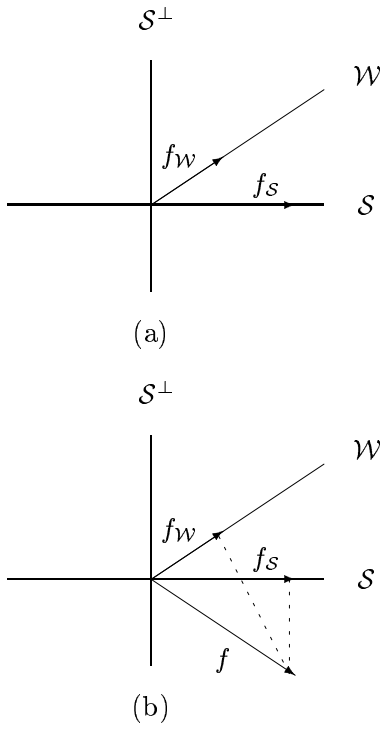


Figure 6-10: Illustration of a construction of a signal f with specified orthogonal projections $f_{\mathcal{S}} = P_{\mathcal{S}} f$ and $f_{\mathcal{W}} = P_{\mathcal{W}} f$ with $\mathcal{W}$ and $\mathcal{S}$ disjoint (a) orthogonal projection of unknown signal onto $\mathcal{S}$ and $\mathcal{W}$ (b) unique signal in $\mathcal{U} = \mathcal{W} \oplus \mathcal{S}$ with the given projections.

We now use Theorem 6.2 to explicitly construct the unique vector $f \in \mathcal{U}$ satisfying the required constraints. First we note that any signal $f \in \mathcal{U}$ can be written as $f = s + v$ where $s \in \mathcal{S}$ and $v \in \widetilde{\mathcal{S}}$ with $\widetilde{\mathcal{S}} = \mathcal{S}^{\perp} \cap \mathcal{U}$. Then, since $\langle s_i, f\rangle = \langle s_i, s\rangle$ for all i, constructing a signal f such that $\langle s_i, f\rangle = c[i]$ is equivalent to constructing a signal $s \in \mathcal{S}$ such that $\langle s_i, s\rangle = c[i]$. Since the vectors s_i form a Riesz basis for $\mathcal{S}$, S^*S is invertible and the unique vector $s \in \mathcal{S}$ such that $S^*s = c$ is given by $s = S(S^*S)^{-1}c$. Once we determined s, the

problem reduces to finding $v \in \widetilde{\mathcal{S}}$ such that $\langle w_i, v\rangle = d[i] - \langle w_i, s\rangle \stackrel{\triangle}{=} d'[i]$, which is again equivalent to a consistent reconstruction problem: We need to construct a signal $v \in \widetilde{\mathcal{S}}$ so that its measurements using the sampling vectors w_i are equal to $d'[i]$. Since the orthogonal complement $\widetilde{\mathcal{S}}^{\perp}$ of $\widetilde{\mathcal{S}}$ in $\mathcal{U}$ is equal to $\mathcal{S}$, $\widetilde{\mathcal{S}}^{\perp}$ and $\mathcal{W}$ are disjoint, and we can apply (6.25) to obtain $v = V(W^*V)^{-1}d' = V(W^*V)^{-1}(d - W^*s)$, where V is a set transformation corresponding to a basis for $\widetilde{\mathcal{S}}$. Finally, the unique $f \in \mathcal{U}$ satisfying the desired constraints is

$$f = S(S^*S)^{-1}c + V(W^*V)^{-1}(d - W^*S(S^*S)^{-1}c). \tag{6.26}$$

We can immediately verify that indeed $S^*f = c$ and $W^*f = d$.

Note that there are many alternative methods of constructing f. Specifically, instead of utilizing the decomposition $f = s + v$ we can decompose f as $f = x + v$ where $v \in \widetilde{\mathcal{S}}$ and x is a subspace $\mathcal{X}$ such that $\mathcal{X} \oplus \widetilde{\mathcal{S}} = \mathcal{U}$. For example, if $\mathcal{W}$ and $\mathcal{S}^{\perp}$ are disjoint then we may choose $\mathcal{X} = \mathcal{W}$. We then construct f by first finding the unique vector $x \in \mathcal{X}$ such that $\langle s_i, x\rangle = c[i]$, and then finding the unique $v \in \widetilde{\mathcal{S}}$ such that $\langle w_i, v\rangle = d[i] - \langle w_i, x\rangle$. With X denoting a set transformation corresponding to a Riesz basis for $\mathcal{X}$,

$$f = X(S^*X)^{-1}c + V(W^*V)^{-1}(d - W^*X(S^*X)^{-1}c). \tag{6.27}$$

In Section 6.4 we considered an application of consistent sampling to the construction of a time-limited signal with specified lowpass coefficients. Using (6.26) we can now extend this construction to produce a signal with specified lowpass coefficients and specified values on a time interval. By choosing different spaces $\mathcal{W}$ and $\mathcal{S}$ and using (6.26), we can construct signals with a variety of different properties. We consider some specific examples in the next section.

6.7.1 Examples of Signal Construction

To illustrate the details of the framework for constructing signals with prescribed properties, in this section we consider the problem of constructing a signal with prescribed local averages and prescribed odd part, the problem of constructing a signal with prescribed recurrent nonuniform samples, and the problem of constructing a signal with prescribed samples

using a given reconstruction filter.

Constructing a signal with prescribed local averages and prescribed odd part

As an illustration of the framework, we consider an example in which we wish to construct a signal with prescribed local averages and prescribed odd part. Specifically, we want to construct a sequence $f \in l_2$ such that $f[2k]+f[2k+1] = c[k]$ for all k, and $f[k]-f[-k] = d[k]$ for $k \geq 1$, where the sequences c and d are given and are assumed to be absolutely summable.

To construct such a signal f using our framework we first determine a set of vectors s_i and a set of vectors w_i such that the desired properties can be expressed in the form $\langle s_i, f\rangle = \tilde{c}[i], i \geq 1$ and $\langle w_i, f\rangle = d[i], i \geq 1$ where $\tilde{c}[i]$ is a reordering[2] of $c[i]$:

$$\tilde{c}[i] = \begin{cases} c[(i-1)/2], & i \geq 1, i \text{ odd}; \\ c[-i/2], & i \geq 2, i \text{ even}. \end{cases} \tag{6.28}$$

Let

$$s_i[k] = \begin{cases} \delta[k-i+1] + \delta[k-i], & i \geq 1, i \text{ odd}; \\ \delta[k+i-1] + \delta[k+i], & i \geq 2, i \text{ even}, \end{cases} \tag{6.29}$$

and $w_i[k] = \delta[k-i] - \delta[k+i]$. Then, $\tilde{c}[i] = \langle s_i, f\rangle$ and $d[i] = \langle w_i, f\rangle$ for $i \geq 1$.

In this example, $\mathcal{S}$ is the subspace of signals x that satisfy $x[2k] = x[2k+1]$ for all k, and $\mathcal{W}$ is the subspace of odd signals. It is immediate that $\mathcal{S}$ and $\mathcal{W}$ are disjoint. To apply (6.26) we need to select a basis v_i for $\mathcal{S}^\perp$, which is the subspace of signals x that satisfy $x[2k] = -x[2k+1]$ for all k. A possible basis is

$$v_i[k] = \begin{cases} \delta[k-i+1] - \delta[k-i], & i \geq 1, i \text{ odd}; \\ \delta[k+i-1] - \delta[k+i], & i \geq 2, i \text{ even}. \end{cases} \tag{6.30}$$

Since $\mathcal{U} = \mathcal{S} \oplus \mathcal{W} = l_2$, there is a unique signal $f \in l_2$ with the desired properties.

To determine f we need to calculate the semi-infinite matrices $(S^*S)^{-1}$, $(W^*V)^{-1}$, and W^*S, where S, W and V are the set transformations corresponding to the vectors s_i, w_i

[2]The purpose of the reordering is to ensure that the index set of the vectors s_i and the vectors w_i is the same.

and v_i, respectively. Since $S^*S = 2I$, $(S^*S)^{-1} = (1/2)I$. Now,

$$\langle w_i, v_j \rangle = (-1)^{j+1}(\delta_{i,j-1} - \delta_{i,j}), \quad i, j \geq 1, \tag{6.31}$$

so that

$$W^*V = \begin{bmatrix} -1 & -1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & -1 & -1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 1 & \cdots \\ & & \vdots & & & \end{bmatrix}. \tag{6.32}$$

We can immediately verify that

$$(W^*V)^{-1} = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & \cdots \\ 0 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 0 & -1 & -1 & -1 & \cdots \\ 0 & 0 & 0 & 1 & 1 & \cdots \\ & & \vdots & & & \end{bmatrix}, \tag{6.33}$$

and if $g = (W^*V)^{-1}h$ for some sequence h, then

$$g[i] = (-1)^i \sum_{k=i}^{\infty} h[k]. \tag{6.34}$$

Finally,

$$\langle w_i, s_j \rangle = (-1)^{j+1}(\delta_{i,j-1} + \delta_{i,j}), \quad i, j \geq 1, \tag{6.35}$$

so that

$$W^*S = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -1 & 0 & \cdots \\ 0 & 0 & 0 & -1 & 1 & \cdots \\ & & \vdots & & & \end{bmatrix}, \tag{6.36}$$

and with $e = (W^*S)\tilde{c}$,

$$e[i] = (-1)^{i+1}(\tilde{c}[i] - \tilde{c}[i+1]). \tag{6.37}$$

Applying (6.26) results in

$$f = \frac{1}{2}S\tilde{c} + Vg, \tag{6.38}$$

where $g = (W^*V)^{-1}h$ with $h = d - (1/2)e$ and $e = W^*S\tilde{c}$. Thus, $f[k] = f_1[k] + f_2[k]$ where $f_1[k] = (1/2)\sum_{i=1}^{\infty} \tilde{c}[i]s_i[k]$ and $f_2[k] = \sum_{i=1}^{\infty} g[i]v_i[k]$ with $\tilde{c}[i]$, $s_i[k]$ and $v_i[k]$ given by (6.28), (6.29) and (6.30) respectively, and from (6.34)

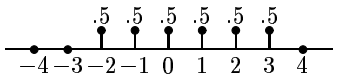
$$g[i] = (-1)^i \sum_{k=i}^{\infty} (d[k] - \frac{1}{2}e[k]), \tag{6.39}$$

where $e[i]$ is given by (6.37). The sequence f_1 lies in $\mathcal{S}$ and has the desired local averages: $f_1[2k] + f_1[2k+1] = c[k]$ for all k. The sequence f_2 lies in $\mathcal{S}^\perp$, and completes the odd part of f_1 to the desired odd part. Finally, we can write $f[k]$ explicitly as

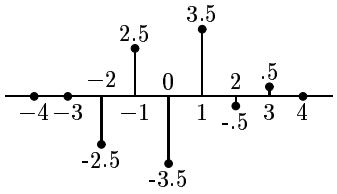
$$f[k] = \begin{cases} \frac{1}{2}c[k/2] + g[k+1], & k \geq 0, k \text{ even}; \\ \frac{1}{2}c[(k-1)/2] - g[k], & k > 0, k \text{ odd}; \\ \frac{1}{2}c[k/2] - g[-k], & k < 0, k \text{ even}; \\ \frac{1}{2}c[(k-1)/2] + g[-k+1], & k < 0, k \text{ odd}. \end{cases} \tag{6.40}$$

We now consider a concrete example of this construction.

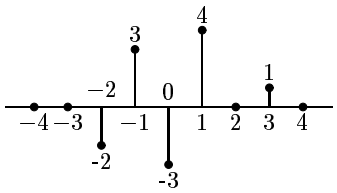
Example 6.1. Suppose we want to construct a signal f such that $f[2k]+f[2k+1] = c[k]$ for all k, where $c[-1] = c[0] = c[1] = 1$ and $c[k] = 0$ otherwise, and such that $f[k]-f[-k] = d[k]$



(a)



(b)



(c)

Figure 6-11: Constructing a sequence f with specified local averages and specified odd part (a) unique signal $f_1 \in \mathcal{S}$ with required local averages (b) unique signal $f_2 \in \mathcal{S}^\perp$ with odd part equal to the difference between the required odd part and the odd part of f_1 (c) unique signal $f = f_1 + f_2$ with both the required local averages and the required odd part.

for $k \geq 1$, where $d[1] = 1, d[2] = 2, d[3] = 1$. We determine f using the construction of (6.40). From (6.28), $\tilde{c}[1] = \tilde{c}[2] = \tilde{c}[3] = 1$ and $\tilde{c}[k] = 0$ for $k \geq 4$. Then from (6.37), $e[1] = e[2] = 0$, $e[3] = 1$, and $e[k] = 0$ for $k \geq 4$. Finally, $g[1] = -\sum_{k=1}^{3}(d[k] - 1/2e[k]) = -3.5$, $g[2] = \sum_{k=2}^{3}(d[k] - 1/2e[k]) = 2.5$, $g[3] = -d[3] + 1/2e[3] = -0.5$.

Thus, f is the sum of the two sequences depicted in Fig. 6-11. Fig. 6-11(a) depicts the unique signal $f_1 \in \mathcal{S}$ with the desired local averages, so that $f_1[2k] + f_1[2k+1] = c[k]$. Fig. 6-11(b) depicts the unique signal $f_2 \in \mathcal{S}^\perp$ with odd part satisfying $f_2[k] - f_2[-k] = d[k] - x[k]$, where $x = W^* f_1$ is the odd part of f_1. Note that, as we expect, the local averages of f_2 are all equal 0. Fig. 6-11(c) depicts $f = f_1 + f_2$ which is the unique sequence with the desired local averages and the desired odd part. □

Constructing a signal with prescribed recurrent nonuniform samples

As a second illustration of the framework, suppose we want to construct a continuous-time signal $f(t)$ bandlimited to $\omega_0 = \pi/T_Q$ with specified samples, where the sampling points are divided into groups of N points each, and the group has a recurrent period $T = NT_Q$. Each period consists of N nonuniform sampling points. Denoting the points in one period by $t_k, k = 1, 2, \ldots N$, the complete set of sampling points is

$$t_k + lT, \quad k = 1, 2, \ldots N, \; l \in \mathbb{Z}. \tag{6.41}$$

Thus our problem is to find an $f \in \mathcal{W}$ where $\mathcal{W}$ is the space of all signals bandlimited to $\omega_0 = \pi/T_Q$, such that $\langle s_i, f\rangle = c[i]$ where $s_i(t) = \delta(t - lT - t_k)$ with $i = lN + k, 0 \leq k \leq N-1$ and $\langle s_i, f\rangle = \int s_i^*(t) f(t) dt$. The unique f with these samples is the signal given by $f = (S^*W)^{-1}c$ where S is the set transformation corresponding to the signals $s_i(t)$, and W is a set transformation corresponding to a basis $w_i(t)$ for $\mathcal{W}$. A possible choice is $w_i(t) = \sin(\omega_0(t - iT_Q))/(\omega_0(t - iT_Q))$. With this choice, if $y = S^*Wc$, then y can be obtained as the output of the filter bank depicted in Fig. 6-12, where the filters $H_k(\omega)$ have impulse response $h_k[i] = (-1)^i \sin(\omega_o t_k)/(\omega_o t_k - i\pi)$.

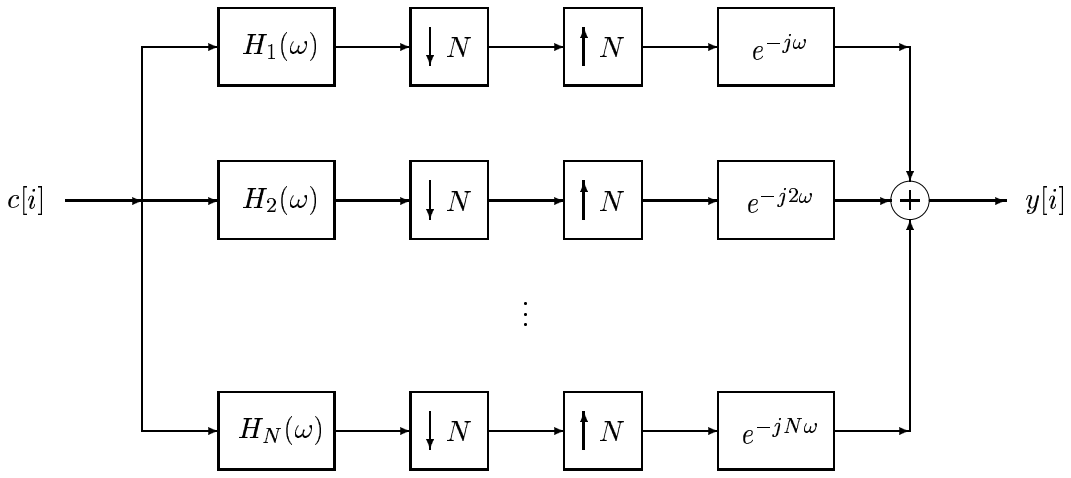


Figure 6-12: Filter bank implementation of $y = S^*Wc$.

To determine $(S^*W)^{-1}$ we need to invert the filter bank of Fig. 6-12. The inverse filter bank has the form depicted in Fig. 6-13, where the filters $G_k(\omega)$ have been determined in [132] and are equal to the filters in [132, Fig. 9] given by $G_k(\omega) = (1/T_Q)R_k(\omega/T_Q)e^{-jt_k\omega/T_Q}$,

for $|\omega| \leq \pi$, where $R_k(\omega)$ is the frequency response of the filter with impulse response

$$r_k(t) = a_k T \frac{\sin(\pi t/T)}{\pi t} \prod_{\substack{q=0 \\ q \neq k}}^{N-1} \sin(\pi(t + t_k - t_q)/T), \tag{6.42}$$

and

$$a_k = \frac{1}{\prod_{q=0, q\neq k}^{N-1} \sin(\pi(t_k - t_q)/T)}. \tag{6.43}$$

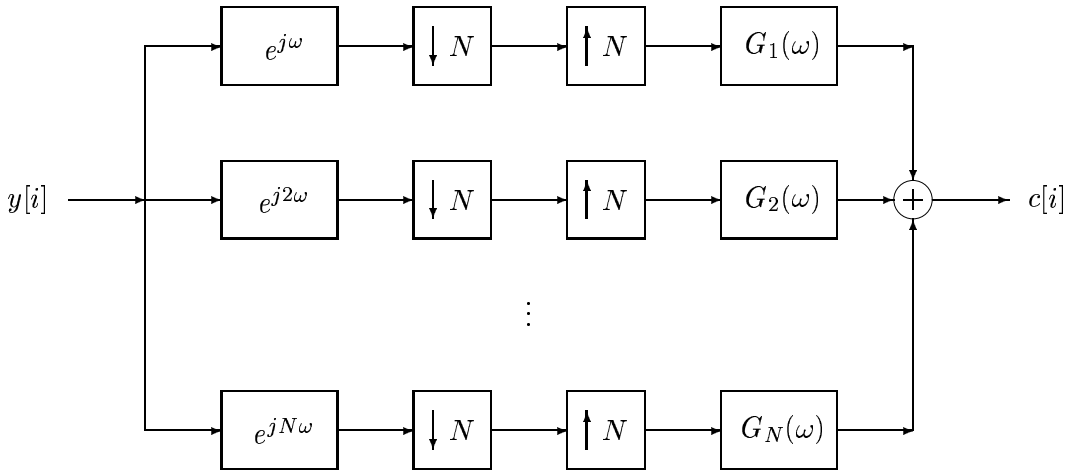


Figure 6-13: Filter bank implementation of $c = (S^*W)^{-1}y$.

Therefore to construct $f(t)$, we first obtain y using the filter bank of Fig. 6-13. Then $f(t) = \sum_i y[i]w_i(t)$ which can be implemented by modulating the samples $y[i]$ onto a uniformly spaced impulse train with period T_Q, and then filtering the modulated impulse train with a continuous-time lowpass filter with cutoff frequency π/ω_0.

Constructing signals with prescribed samples

As a third illustration of our framework, suppose we wish to construct a continuous-time signal $f(t)$ to have prescribed samples so that $f(i) = c[i], i \in \mathbb{Z}$, where $f(i)$ denotes the value of $f(t)$ at $t = i$. The signal $f(t)$ is constrained to lie in the subspace $\mathcal{W}$ generated by the integer translates $\{w(t-i), i \in \mathbb{Z}\}$ of a given function $w(t)$, so that $f(t) = \sum_i x[i]w(t-i)$ for some coefficients $x[i]$. We assume that $\alpha \leq \sum_i |W(\omega - 2\pi i)|^2 \leq \beta$ where $0 < \alpha < \beta < \infty$ and $W(\omega)$ is the continuous-time Fourier transform of $w(t)$, which ensures that $\{w(t-i)\}$

forms a Riesz basis for $\mathcal{W}$ [89]. We can equivalently obtain the signal $f(t)$ as the the output of the block diagram of Fig. 6-14.

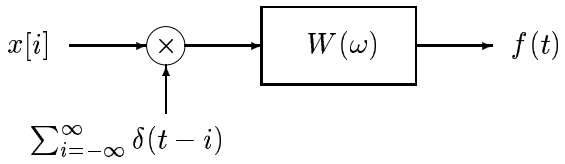


Figure 6-14: Constructing a signal $f(t)$ from the sequence $x[i]$ using a given filter with frequency response $W(\omega)$ and impulse response $w(t)$.

The problem then is to find the coefficients $x[i]$ so that $f(i) = c[i]$. We can express $f(i)$ as $f(i) = \langle s_i(t), f(t) \rangle$ where $s_i(t) = \delta(t-i)$ and $\langle y(t), r(t) \rangle = \int y^*(t) r(t) dt$. From (6.25) it then follows that $x = (S^*W)^{-1}c$ where S and W are the set transformations corresponding to the vectors $s_i(t)$ and $w_i(t)$ respectively. Since $\langle s_i(t), w_k(t) \rangle = w(k-i)$, S^*W is an infinite Toeplitz matrix, and is therefore equivalent to a filtering operation with a filter whose impulse response is given by $\langle s_0(t), w_k(t) \rangle = w(k)$. The frequency response of the filter is

$$\sum_{k=-\infty}^{\infty} w(k) e^{-j\omega k} = 2\pi \sum_{k=-\infty}^{\infty} W(\omega + 2\pi k), \tag{6.44}$$

where we used the Poisson sum formula [133]. It follows that if $x = (S^*W)^{-1}c$, then x is obtained by filtering the sequence c with a discrete-time filter with frequency response

$$G(\omega) = \frac{1}{2\pi \sum_k W(\omega + 2\pi k)}, \tag{6.45}$$

as depicted in Fig. 6-15.

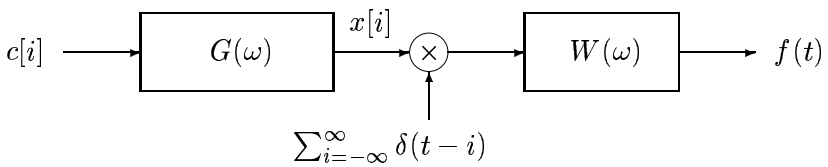


Figure 6-15: Constructing a signal $f(t)$ with samples $f(i) = c[i]$ using a given filter with frequency response $W(\omega)$, where $G(\omega)$ is given by (6.45).

In the special case in which $W(\omega)$ is the frequency response of an ideal lowpass filter with cutoff frequency $\omega_0 = \pi$, $G(\omega) = 1$ so that $x[i] = c[i]$.

To conclude this chapter, based on consistency, oblique projections and oblique dual frame vectors we developed a general framework for sampling and reconstruction in arbitrary spaces as well as a general framework for constructing signals with prescribed properties.

In the remainder of the thesis we consider applications of ROMs. Specifically, we develop new algorithms that result from either processing a signal using a ROM or processing a signal using the measurement vectors of a ROM and imposing inner product constraints directly on these vectors.

Chapter 7

QSP Quantization

In this chapter we consider some quantization methods that result from the QSP measurement framework. In particular, we develop an efficient practical implementation of a dithered quantizer based on a QSP measurement, that can be used to control the statistical properties of the quantization error, and to efficiently shape the quantization noise. This new implementation requires only the generation of one uniform random variable per input regardless of the distribution of the dither signal, thus alleviating the computational complexity associated with a general dithered quantizer.

In dithered quantization a random signal called a dither signal is added to the input signal prior to quantization [28, 29, 30, 31]. Dithering techniques have become commonplace in applications in which it is necessary to quantize data prior to storage or transmission. However, the utility of dithering techniques is limited by the computational complexity associated with generating a random process with an arbitrary joint probability distribution. Therefore, in practice, dithering signals are typically restricted to a weighted combination of uniform independent, identically distributed (iid) random variables. In this chapter we show that a probabilistic quantizer can be used to effectively realize a dither signal with an arbitrary joint probability distribution, while requiring only the generation of one uniform random variable per input.

7.1 Classical Model of Quantization

Quantization is the operation of mapping an input into a set of discrete outputs. The input can be, for example, a continuous-time signal or a discrete-time signal that takes on

continuous values. For concreteness, we assume that the input signal to the quantizer is a sequence in l_2 with elements x_n. The quantized output at time n is denoted by $y_n = Q(x_n)$. When considering memoryless quantizers the time index n is not important, so we omit it. A memoryless quantizer has a "staircase" transfer characteristic, illustrated in Fig. 7-1.

We assume that no overflow occurs in the quantizer, so that the signal is never clipped by saturation of the quantizer. A convenient way of incorporating this premise is by assuming that the quantizer has infinitely many quantization levels. For simplicity, we assume that the quantizers are uniform, *i.e.,* the separation between adjacent quantization levels is fixed and is equal to Δ, as depicted in Fig. 7-1. The quantizer in Fig. 7-1 is called a uniform mid-tread quantizer with quantizer step size Δ. We will assume this quantizer model throughout this chapter. The input-output relation of this quantizer is given by

$$Q(x) = \Delta \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor . \tag{7.1}$$

Alternatively,

$$Q(x) = i\Delta, \text{ where } i = \arg\min |x - k\Delta|. \tag{7.2}$$

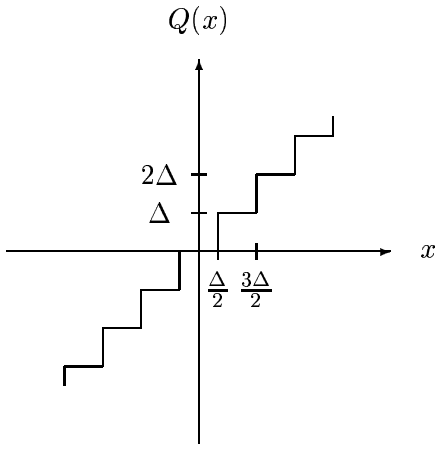


Figure 7-1: Quantizer transfer characteristic.

7.2 Measurement Description of Quantizer

We now formulate the uniform quantizer as a QSP measurement, and then use the measurement framework to develop modifications of the uniform quantizer.

Suppose we want to quantize x to obtain the output $y = Q(x)$, where Q is the uniform quantizer of Fig. 7-1. Example 4.4 in Chapter 4 gives a concrete construction of a ROM which can be used to implement the quantizer. Using this construction here, we map the quantization levels $i\Delta$ and the input x into vectors in l_2 using an input mapping $T_{\mathcal{X}}\colon\ \mathcal{R} \to l_2$. We then construct a ROM on l_2 with measurement vectors equal to $q_i = T_{\mathcal{X}}(i\Delta)$ and a mapping f_M that depends on $z = T_{\mathcal{X}}(x)$ only through the inner products $\langle z, q_i \rangle$, where the mapping $T_{\mathcal{X}}$ is chosen so that $\langle z, q_i \rangle = 1/(x - i\Delta)$. The measurement outcome is then mapped to the final quantized level using the mapping $T_{\mathcal{Y}}\colon\ l_2 \to \mathcal{R}$ defined by $T_{\mathcal{Y}}(y) = i\Delta$ if y is a multiple of q_i. Since this construction is analogous to the construction in Example 4.4 we omit the details. The resulting measurement description of the quantizer is depicted in Fig. 7-2.

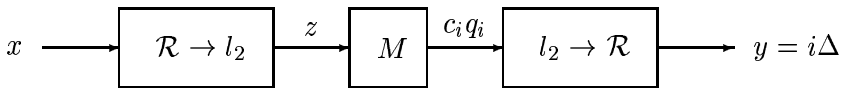


Figure 7-2: Measurement description of quantizer.

Since $\langle z, q_i \rangle = 1/(x - i\Delta)$, the input-output relation of the quantizer resulting from the measurement description of Fig. 7-2 can be described directly in $\mathcal{R}$ as $y = i\Delta$, where $i = \tilde{f}(x)$ and $\tilde{f}$ is a mapping that depends on the input x only through the numbers $\{x - i\Delta\}$. We refer to this quantizer as a QSP quantizer with mapping $\tilde{f}$. In the uniform quantizer, $\tilde{f}(x) = \arg\min |x - k\Delta|$. By choosing different mappings $\tilde{f}$, a variety of new potentially interesting quantizers can be obtained.

In the remainder of this chapter we focus on QSP quantizers that result from choosing $\tilde{f}$ as a probabilistic mapping, which we refer to as probabilistic quantizers. In our discussion we consider both the memoryless case, in which $\tilde{f}$ is a probabilistic mapping that depends only on the input x, and the more general case in which the mapping $\tilde{f}$ may depend on previous values of x.

7.3 Memoryless Probabilistic Quantizer

The main result we establish in this section is that a memoryless probabilistic quantizer is equivalent to a dithered quantizer [28, 31] with an iid dither signal. However, these equivalent representations have different implications in terms of implementation. We will show that the probabilistic quantizer generally results in a more practical and efficient implementation, which allows for much more flexibility in the design of dithered quantizers.

7.3.1 The Memoryless Probabilistic Quantizer

We now consider the case in which we choose $\tilde{f}\colon\ \mathcal{R} \times \mathcal{W} \to \mathbb{Z}$ as a probabilistic mapping from $\mathcal{R}$ to $\mathbb{Z}$, where $\mathcal{W} = \mathbb{Z}$ is the sample space of an auxiliary chance variable w with discrete alphabet $\mathbb{Z}$, such that w can take on a value $i \in \mathbb{Z}$ with probability $p_i = g(x - i\Delta)$, for some function $g(x)$. Then let $\tilde{f}(x, i) = i$. The output of the resulting QSP quantizer with mapping $\tilde{f}$ is then given by

$$y = i\Delta, \text{ with probability } p_i = g(x - i\Delta). \tag{7.3}$$

We refer to this quantizer as a probabilistic quantizer with mapping $g(x)$.

To ensure that the values p_i represent probabilities, the function $g(x)$ must satisfy

$$g(x) \geq 0; \tag{7.4}$$

$$\sum_{k=-\infty}^{\infty} g(x - k\Delta) = 1. \tag{7.5}$$

Since,

$$\sum_{k=-\infty}^{\infty} g(x - k\Delta) = g(x) * \sum_{k=-\infty}^{\infty} \delta(x - k\Delta), \tag{7.6}$$

where $*$ denotes convolution, we can express the condition (7.5) in the frequency domain as

$$\frac{2\pi}{\Delta} \sum_{k=-\infty}^{\infty} G(\omega)\delta\left(\omega - \frac{2\pi}{\Delta}k\right) = 2\pi\delta(\omega), \tag{7.7}$$

where $G(\omega)$ denotes the Fourier transform of $g(x)$. Thus, $G(\omega)$ must satisfy

$$G\left(\frac{2\pi}{\Delta}k\right) = \Delta\delta[k], \quad k \in \mathbb{Z}. \tag{7.8}$$

We conclude that $g(x)$ in (7.3) must be a nonnegative function with Fourier transform $G(\omega)$ satisfying (7.8).

To implement the probabilistic quantizer with mapping g, we first compute the probabilities p_i according to (7.3), and then generate a chance variable w with alphabet $\mathbb{Z}$ such that $w = i$ with probability p_i. The quantized output is given by $w\Delta$. The random variable w can be generated using a uniform random number generator. Specifically, let a be a uniform random variable on $[0, 1]$, let $s_i = \sum_{k=1}^{i} p_k$, and let Pr $(\cdot)$ denote the probability of the corresponding event. Then Pr $(s_i \leq a \leq s_{i+1}) = s_{i+1} - s_i = p_i$. We can therefore implement the probabilistic quantizer by generating a realization of a, and then quantizing the output to $i\Delta$ where $s_i \leq a \leq s_{i+1}$.

7.3.2 Probabilistic Quantizer and Nonsubtractive Dithered Quantizer

We now show that the output of the memoryless probabilistic quantizer, is equivalent[1] to the output of a nonsubtractive dithered (NSD) quantizer [28, 31], depicted in Fig. 7-3.

In NSD quantization, the input to the quantizer is the system input x_n plus an additive random signal v_n called the dither signal, which is assumed to be stationary and independent of x_n. In this section we assume that v_n is an iid signal so that we may omit the index n. The output of the NSD quantizer is then given by $y' = Q(x + v)$, where v is a random variable with probability density function (pdf) $f_v(v)$.

Although in a uniform (undithered) quantizer the error is a deterministic function of the input, the classical model of quantization treats this error as an additive random process, that is independent of the input signal, iid, and uniformly distributed [134, 135]. This model of quantization has been shown to be reasonable for quasi-random input signals with large magnitude relative to the step size Δ [134]. However, it fails for small relatively simple signals in which case the error signal retains many of the characteristics of the original signal, which causes undesirable audio and visual distortions [136, 31].

[1] By equivalent we mean that the statistics of the outputs of both quantizers are identical. Thus, no experiment can be performed on the two systems that will allow us to distinguish between them.

The main objective of dithering is to control the statistical properties of the error $\epsilon = y' - x$ and its relationship to the input x. By using dither, some of the statistical properties of the error can be adjusted so that *e.g.,* the error can be rendered white and uncorrelated with the input. The drawback of this approach, alleviated by the use of the probabilistic quantizer, is that it requires generating a random process with a pdf that may be difficult to implement in practice. We will show that using a probabilistic quantizer we can realize any desired pdf on the dither signal, in a practical efficient manner.

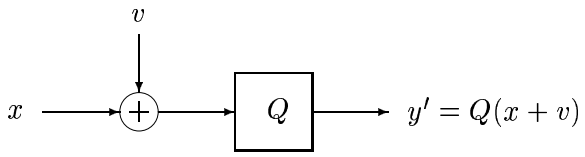


Figure 7-3: Nonsubtractive dithered quantization.

Using an NSD quantizer, the probability that the output y' is equal to $i\Delta$ is given by

$$\begin{aligned}
\Pr\,(y' = i\Delta) &= \Pr\left(|x + v - i\Delta| \leq \frac{\Delta}{2}\right) \\
&= \Pr\left(-\frac{\Delta}{2} + i\Delta - x \leq v \leq \frac{\Delta}{2} + i\Delta - x\right) \\
&= \int_{-\frac{\Delta}{2}+i\Delta-x}^{\frac{\Delta}{2}+i\Delta-x} f_v(v)dv \\
&= \Pi_\Delta(x - i\Delta) * f_v(-x), \qquad (7.9)
\end{aligned}$$

where

$$\Pi_\Delta(x) = \begin{cases} 1, & |x| \leq \frac{\Delta}{2}; \\ 0, & \text{otherwise.} \end{cases} \qquad (7.10)$$

When using a probabilistic quantizer with mapping $g(x)$,

$$\Pr\,(y = i\Delta) = g(x - i\Delta). \qquad (7.11)$$

Comparing (7.11) with (7.9), we conclude that the probability distribution of the output y of the probabilistic quantizer is equivalent to that of the output y' of the NSD quantizer

if we choose

$$g(x) = \Pi_\Delta(x) * f_v(-x). \tag{7.12}$$

We can express this relation in the frequency domain as

$$G(\omega) = \Delta \ \mathrm{sinc}(\omega\Delta/2) F_v(-\omega), \tag{7.13}$$

where $F_v(\omega)$ is the Fourier transform of $f_v(v)$ and $\mathrm{sinc}(x) = \sin(x)/x$.

Since $F_v(0) = \int f_v(v)dv = 1$ and $\mathrm{sinc}(\pi i) = \delta[i]$, for any choice of noise pdf $f_v(v)$, $G(\omega)$ generated according to (7.13) always satisfies (7.8). In addition, since $f_v(x)$ and $\Pi_\Delta(x)$ are nonnegative, from (7.12) it follows that $g(x)$ is always nonnegative. We therefore conclude that any $g(x)$ generated according to (7.12) is guaranteed to satisfy the required conditions (7.8) and (7.4).

Thus, any NSD quantizer can be realized as a probabilistic quantizer where $g(x)$ is chosen in accordance with (7.12) or (7.13). We can also implement any probabilistic quantizer with function $g(x)$ as an NSD quantizer, where the dither signal v is chosen to have a pdf $f_v(v)$ satisfying (7.12) or (7.13). Specifically,

$$F_v(\omega) = \frac{G(-\omega)}{\Delta \ \mathrm{sinc}(\omega\Delta/2)}, \quad \omega \neq \frac{2\pi}{\Delta} i, \ i = \pm 1, \pm 2, \ldots, \tag{7.14}$$

and the values of $F_v(\omega)$ on $\omega = 2\pi i/\Delta$ must be chosen such that $f_v(v) \geq 0$ for all v.

In summary, an NSD quantizer with dither signal with pdf $f_v(v)$ and a probabilistic quantizer with mapping $g(x)$ are equivalent, where $g(x)$ and $f_v(v)$ are related through (7.12). However, they have different implications in terms of implementation. The implementation of the probabilistic quantizer requires only a uniform random number generator, regardless of the choice of $g(x)$. By contrast, to implement an NSD quantizer we need to generate a random variable v with distribution $f_v(v)$ which may be hard to do in practice. In fact, typically in applications [28, 29, 30, 31] $f_v(v)$ is constrained to be a sum of iid random variables distributed uniformly on $[-\Delta/2, \Delta/2]$, so that the random variable v can be easily generated. By using a probabilistic quantizer we can alleviate this constraint and effectively implement any desired distribution on v by simply choosing a different function g.

Since the probabilistic quantizer is equivalent to an NSD quantizer, we can apply the

many known results regarding dither quantization to the probabilistic quantizer. One of the most important considerations in using an NSD quantizer is to render the moments of the error at the output of the quantizer, independent of the input. We can achieve the same effect using a probabilistic quantizer, with an appropriate choice of $g(x)$. Specifically, let $\epsilon = y - x$ denote the error at the output of the probabilistic quantizer, which we refer to in the sequel as the quantization error. Then the following result follows immediately from [28, Theorem 6].

Theorem 7.1. *When using a probabilistic quantizer with mapping $g(x)$, $E(\epsilon^k|x)$ is functionally independent of the input x for $k \geq 1$ if and only if*

$$\left.\frac{d^k G(\omega)}{d\omega^k}\right|_{\omega=\frac{2\pi}{\Delta}i} = 0, \quad i = \pm 1, \pm 2, \dots , \tag{7.15}$$

where $G(\omega)$ is the Fourier transform of $g(x)$. In particular, if (7.15) is satisfied for $k = 1$, then ϵ and x are uncorrelated.
When (7.15) is satisfied,

$$E(\epsilon^k|x) = E(\epsilon^k) = \frac{(-j)^k}{\Delta}\left.\frac{d^k G(\omega)}{d\omega^k}\right|_{\omega=0} = \frac{1}{\Delta}\int_x x^k g(x)dx. \tag{7.16}$$

Thus, by choosing $g(x)$ appropriately we can control the moments of the quantization error.

In applications, we can always construct $g(x)$ by first choosing a distribution $f_v(v)$ and then using (7.12). In the next section we mention a few of the techniques that can be used to directly construct $g(x)$, by exploiting the connection with other well known problems.

7.3.3 Constructing the Mapping in the Probabilistic Quantizer

There are a variety of contexts in which it is desired to construct a function $g(x)$ satisfying (7.5). In many of these problems $g(x)$ is also constructed to be nonnegative. We can therefore exploit these known constructions to the design of $g(x)$.

One context in which the criterion (7.5) arises is in communication over a bandlimited channel. In this context, a linearly modulated signal $\sum_i a_i p(t - iT)$ is received in additive noise over a bandlimited channel, where $\{a_i\}$ represents the digital information-bearing sequence of symbols, and $p(t)$ is the bandlimited received pulse shape. The received signal

is sampled at times iT, and the sequence of samples is then processed by an appropriate slicer. It is well known that to ensure that there is no intersymbol interference (ISI) at the output of the channel, the received pulse shape $p(t)$ must satisfy the *Nyquist criterion* $\sum_{i=-\infty}^{\infty} P(\omega - 2\pi i/T) = T$, where $P(\omega)$ is the Fourier transform of $p(t)$ [137, 11]. There are a variety of well known methods for constructing Nyquist pulses that satisfy the Nyquist criterion. Using any one of these methods we can construct $g(x) = (1/T)P(x)$ where $P(\omega)$ is a nonnegative Nyquist pulse. Then $g(x)$ satisfies (7.4) and (7.5) with $\Delta = 2\pi/T$.

Another context in which the criterion (7.5) arises is in the construction of tight frames. Specifically, in [138] it is shown that the family of functions $h_{mn}(x) = e^{2j\pi nx/L}h(x + m\Delta)$, $0 < \Delta < L$, $n, m \in \mathbb{Z}$, is a tight frame if and only if $\sum_{k=-\infty}^{\infty} |h(x + k\Delta)|^2 = C$ for some constant $C \neq 0$. Thus if $h(x)$ is a function that results in a tight frame, then choosing $g(x)$ proportional to $|h(x)|^2$ will satisfy (7.4) and (7.5).

Based on results in [138] we now provide an explicit construction of symmetric nonnegative functions $g(x)$ satisfying (7.5). Let $r(x)$ be a function such that $r(0) = 0, r(1) = 1$, $0 \leq r(x) \leq 1$ for $0 \leq x \leq 1$, and $r(x) = 1 - r(1 - x)$ for $0 \leq x \leq 1$. Let $\Delta < L \leq 2\Delta$. Then it can be readily verified that the following $g(x)$ satisfies (7.4) and (7.5):

$$g(x) = \begin{cases} 1, & |x| \leq \Delta - \frac{L}{2}; \\ r\left(\frac{L-2|x|}{2(L-\Delta)}\right), & \Delta - \frac{L}{2} \leq |x| \leq \frac{L}{2}; \\ 0, & |x| \geq \frac{L}{2}. \end{cases} \tag{7.17}$$

The resulting probabilistic quantizer is in general hard to implement as an NSD quantizer, since the equivalent noise distribution $f_v(v)$ is typically quite involved.

Example 7.1. As a special case of (7.17), suppose we choose $r(x) = 1/2 + (1/2)\sin(\pi(x - 1/2))$, and let $\alpha = L/\Delta - 1$ so that $0 \leq \alpha \leq 1$. Then

$$g(x) = \begin{cases} 1, & |x| \leq (1-\alpha)\Delta/2; \\ \frac{1}{2}\left(1 - \sin(\frac{\pi}{\alpha\Delta}(|x| - \Delta/2)\right), & (1-\alpha)\Delta/2 \leq |x| \leq (1+\alpha)\Delta/2; \\ 0, & |x| \geq (1+\alpha)\Delta/2. \end{cases} \tag{7.18}$$

The function $g(x)$ is the well-known raised-cosine function [137, 11] which is widely used in practice in the context of communication. □

Example 7.2. As another example of (7.17), suppose that $L = 2\Delta$ and $r(x) = \sin^2(\pi x/2)$,

$0 \leq x \leq 1$. The corresponding $g(x)$ is given by

$$g(x) = \begin{cases} \cos^2(2\pi x/\Delta), & 0 \leq |x| \leq \Delta; \\ 0, & |x| \geq \Delta. \end{cases} \tag{7.19}$$

With $G(\omega)$ denoting the Fourier transform of $g(x)$,

$$\begin{aligned} G(\omega) &= \operatorname{sinc}(\omega\Delta)\left(\Delta + \frac{2\omega^2\Delta^3}{\pi^2 - \Delta^2\omega^2}\right) \\ &= \operatorname{sinc}(\omega\Delta/2)\cos(\omega\Delta/2)\left(\Delta + \frac{2\omega^2\Delta^3}{\pi^2 - \Delta^2\omega^2}\right) \\ &= \operatorname{sinc}(\omega\Delta/2)F_v(\omega), \end{aligned} \tag{7.20}$$

where

$$F_v(\omega) = \cos(\Delta\omega/2)\left(\Delta + \frac{2\omega^2\Delta^3}{\pi^2 - \Delta^2\omega^2}\right). \tag{7.21}$$

We see that although $g(x)$ given by (7.19) is a simple function to realize, the distribution of the corresponding dither noise is quite involved and hard to practically implement. Evidently in this case the probabilistic quantizer offers a more practical, efficient method for implementing the NSD with dither pdf $f_v(v)$ whose Fourier transform is given by (7.21). □

7.4 Probabilistic Quantizer With Memory

We have seen in the previous section that by choosing a memoryless function $g(x)$ in the probabilistic quantizer we can control the moments of the quantization error. It is also important in a variety of applications to control second order statistics of the quantization error. For example, in an audio system it may be preferable to shape the quantization error such that most of its power resides in high frequency bands where the human ear is relatively insensitive. Noise shaping techniques together with oversampling of the continuous-time signal also allow for high resolution conversion of low bandwidth signals, *e.g.,* using sigma-delta modulation [139, 140]. In this section we show that using a probabilistic quantizer with memory, we can control the second order statistics of the quantization error to allow for various forms of noise shaping.

Specifically, we now consider the case in which the probabilistic mapping has memory, so

that the mapping g depends on the current value x_n, and on previous values $x_{n-1}, \ldots, x_{n-m}$ for some $m \geq 1$. We assume throughout that the function g does not depend on n, so we denote $x_{n-k} = x_k$ for $0 \leq k \leq m$. We then show that the probabilistic quantizer with memory is equivalent to an NSD quantizer with a dither signal that is not iid. Based on results derived in the context of NSD quantization, we show that we can choose the mapping g to control the correlation of the error sequence. The advantage of implementing an NSD quantizer as a probabilistic quantizer with memory is even more pronounced then in the memoryless case, since generating a random process with arbitrary joint pdf can be computationally very demanding, while the probabilistic quantizer with memory still only requires the generation of one uniform random variable per input.

7.4.1 The Probabilistic Quantizer With Memory

To describe the probabilistic quantizer with memory, suppose that we now choose the mapping $\tilde{f}$ in the QSP quantizer as a probabilistic mapping $\tilde{f}$: $\mathcal{V} \times \mathcal{W} \to \mathbb{Z}$ from $\mathcal{V} = \mathcal{R} \times \cdots \times \mathcal{R}$ to $\mathbb{Z}$, where $\mathcal{W} = \mathbb{Z} \times \cdots \times \mathbb{Z}$ is the sample space of $m+1$ auxiliary chance variables $w_0, \ldots, w_m$ with discrete alphabets $\mathbb{Z}$ such that each chance variable w_k can take on a value $i_k \in \mathbb{Z}$, where the joint probability of outcomes $i_0, \ldots, i_m \in \mathcal{W}$ is given by $p(i_0, \ldots, i_m) = g(x_0 - i_0\Delta, \ldots, x_m - i_m\Delta)$, for some function $g(x_0, \ldots, x_m)$. Given the previous outcomes $i_1, \ldots, i_m$, the conditional probability of outcome i_0 is

$$\begin{aligned} p(i_0|i_1, \ldots, i_m) &= \frac{p(i_0, i_1, \ldots, i_m)}{p(i_1, \ldots, i_m)} \\ &= \frac{g(x_0 - i_0\Delta, \ldots, x_m - i_m\Delta)}{\sum_{i_0} g(x_0 - i_0\Delta, x_1 - i_1\Delta, \ldots, x_m - i_m\Delta)}. \end{aligned} \tag{7.22}$$

Then let $\tilde{f}(x_0, \ldots, x_m, i_0, \ldots, i_m) = i_0$. The output of the QSP quantizer with mapping $\tilde{f}$ is then given by

$$y_0 = i_0\Delta, \text{ with probability } p(i_0|i_1, \ldots, i_m). \tag{7.23}$$

We refer to this quantizer as a probabilistic quantizer with mapping $g(x_0, \ldots, x_m)$.

To ensure that the values $p(i_0, \ldots, i_m)$ represent probabilities, the function $g(x_0, \ldots, x_m)$

must satisfy

$$g(x_0, \ldots, x_m) \geq 0; \tag{7.24}$$

$$\sum_{i_0, \ldots, i_m = -\infty}^{\infty} g(x_0 - i_0\Delta, \ldots, x_m - i_m\Delta) = 1. \tag{7.25}$$

In addition, we also require that the joint probability of any subset of variables $i_{t_1}, \ldots, i_{t_n}, 0 \leq t_1, \ldots, t_n \leq m$ depends only on $x_{t_1}, \ldots, x_{t_n}$ and is independent of x_t for $t \neq t_i, 1 \leq i \leq n$.

We now translate the last two constraints on $g(x_0, \ldots, x_m)$ to constraints on the $(m+1)$-dimensional Fourier transform $G(\omega_0, \ldots, \omega_m)$ of $g(x_0, \ldots, x_m)$. First, taking the $(m+1)$-dimensional Fourier transform of (7.25), we have that

$$G\left(\frac{2\pi}{\Delta}k_0, \ldots, \frac{2\pi}{\Delta}k_m\right) = \Delta^{m+1}\delta[k_0, \ldots, k_m], \quad k_0, \ldots, k_m \in \mathbb{Z}, \tag{7.26}$$

where $\delta[k_0, \ldots, k_m] = 1$ if $k_0 = \ldots = k_m = 0$, and 0 otherwise.

Next, suppose that we require $p(i_1, \ldots, i_m)$ to be independent of x_0, so that

$$\begin{aligned} p(i_1, \ldots, i_m) &= \sum_{i_0} g(x_0 - i_0\Delta, x_1 - i_1\Delta, \ldots, x_m - i_m\Delta) \\ &= u(x_1 - i_1\Delta, \ldots, x_m - i_m\Delta), \end{aligned} \tag{7.27}$$

for some function $u(x_1, \ldots, x_m)$, where

$$\begin{aligned} u(x_1, \ldots, x_m) &= \sum_{i_0} g(x_0 - i_0\Delta, x_1, \ldots, x_m) \\ &= g(x_0, x_1, \ldots, x_m) * \sum_{i_0} \delta(x_0 - i_0\Delta, x_1, \ldots, x_m). \end{aligned} \tag{7.28}$$

We refer to the function $u(x_1, \ldots, x_m)$ as the joint probability function of $x_1, \ldots, x_m$. With $U(\omega_1, \ldots, \omega_m)$ denoting the m-dimensional Fourier transform of $u(x_1, \ldots, x_m)$, and taking the $(m+1)$-dimensional Fourier transform of (7.28),

$$2\pi U(\omega_1, \ldots, \omega_m)\delta(\omega_0) = \frac{2\pi}{\Delta} \sum_{k_0 = -\infty}^{\infty} G\left(\frac{2\pi}{\Delta}k_0, \omega_1, \ldots, \omega_m\right) \delta\left(\omega_0 - \frac{2\pi}{\Delta}k\right). \tag{7.29}$$

Thus we must have that for all $\omega_1, \ldots, \omega_m$,

$$G\left(\frac{2\pi}{\Delta}k_0, \omega_1, \ldots, \omega_m\right) = 0, \quad k_0 = \pm 1, \pm 2, \ldots . \tag{7.30}$$

The joint probability function is then given by

$$u(x_1, \ldots, x_m) = \frac{1}{\Delta}\mathcal{F}^{-m}\{G(0, \omega_1, \ldots, \omega_m)\} = \frac{1}{\Delta}\int_{x_0} g(x_0, \ldots, x_m), \tag{7.31}$$

where $\mathcal{F}^{-m}\{\cdot\}$ denotes the m-dimensional inverse Fourier transform.

Similarly, we can show that for any subset of variables $t_1, \ldots, t_n$, $p(i_{t_1}, \ldots, i_{t_n})$ is independent of x_t for $t \neq t_i, 1 \leq i \leq n$, if and only if for any $\omega_1, \ldots, \omega_m$,

$$\begin{aligned} G\left(\frac{2\pi}{\Delta}k_0, \omega_1, \ldots, \omega_m\right) &= G\left(\omega_0, \frac{2\pi}{\Delta}k_1, \omega_2, \ldots, \omega_m\right) = \cdots \\ &= G\left(\omega_0, \ldots, \omega_{m-1}, \frac{2\pi}{\Delta}k_m\right) = 0, \quad k_0, \ldots, k_m = \pm 1, \pm 2, \ldots . \end{aligned} \tag{7.32}$$

Then $p(i_{t_1}, \ldots, i_{t_n}) = u(x_{t_1} - i_{t_1}\Delta, \ldots, x_{t_n} - i_{t_n}\Delta)$ where $u(x_{t_1}, \ldots, x_{t_n})$ is the joint probability function of $x_{t_1}, \ldots, x_{t_n}$ and is given by

$$u(x_{t_1}, \ldots, x_{t_n}) = \frac{1}{\Delta^{m+1-n}}\mathcal{F}^{-n}\{G(0, \omega_{t_1}, 0, \ldots, 0, \omega_{t_n}, 0, \ldots, 0)\}. \tag{7.33}$$

We conclude that $g(x_0, \ldots, x_m)$ in (7.22) must be a nonnegative function with $(m+1)$-dimensional Fourier transform $G(\omega_1, \ldots, \omega_m)$ satisfying (7.26) and (7.32).

To implement the probabilistic quantizer with memory, given $i_1, \ldots, i_m$ we compute the probabilities $p(i_0, \ldots, i_m)$, which for notational brevity we denote here by p_{i_0}. Then $p(i_0|i_1, \ldots, i_m) = Cp_{i_0}$, where $C = 1/\sum_{i_0} p_{i_0}$. As in the memoryless case, we next generate a chance variable w with alphabet $\mathbb{Z}$ such that $w = i$ with probability Cp_{i_0}. The quantized output is given by $w\Delta$.

7.4.2 Probabilistic Quantizer With Memory and Nonsubtractive Dithered Quantizer

We now show that the resulting probabilistic quantizer is equivalent to an NSD with a stationary dither signal $\{v_n\}$, where v_n is statistically dependent on $v_{n-1}, \ldots, v_{n-m}$ and v_n, v_{n-k} for $k > m$ are statistically independent. Let $f_v(v_0, \ldots, v_m)$ denote the joint pdf of

$v_0 = v_n, \ldots, v_m = v_{n-m}$, and let $y'_0, \ldots, y'_m$ denote the outputs of the NSD quantizer at times $n, \ldots, n-m$, respectively. Then,

$$\begin{aligned}
&\Pr\,(y'_0 = i_0\Delta, \ldots, y'_m = i_m\Delta) = \\
&\quad = \Pr\left(|x_0 + v_0 - i_0\Delta| \le \frac{\Delta}{2}, \ldots, |x_m + v_m - i_m\Delta| \le \frac{\Delta}{2}\right) \\
&\quad = \int_{-\frac{\Delta}{2}+i_0\Delta-x_0}^{\frac{\Delta}{2}+i_0\Delta-x_0} \cdots \int_{-\frac{\Delta}{2}+i_m\Delta-x_m}^{\frac{\Delta}{2}+i_m\Delta-x_m} f_v(v_0, \ldots, v_m)dv_0 \cdots dv_m \\
&\quad = \Pi_\Delta(x_0 - i_0\Delta, \ldots, x_m - m\Delta) * f_v(-x_0, \ldots, -x_m), \qquad (7.34)
\end{aligned}$$

where

$$\Pi_\Delta(x_0, \ldots, x_m) = \begin{cases} 1, & |x_0| \le \frac{\Delta}{2}, \ldots, |x_m| \le \frac{\Delta}{2}; \\ 0, & \text{otherwise.} \end{cases} \qquad (7.35)$$

When using a probabilistic quantizer with mapping $g(x_0, \ldots, x_m)$,

$$\Pr\,(y_0 = i_0\Delta, \ldots, y_m = i_m\Delta) = g(x_0 - i_0\Delta, \ldots, x_m - i_m\Delta). \qquad (7.36)$$

As in the memoryless case, comparing (7.36) with (7.34) we conclude that the probability distribution of the output using a probabilistic quantizer is equivalent to that of the NSD quantizer if

$$g(x_0, \ldots, x_m) = \Pi_\Delta(x_0 \ldots, x_m) * f_v(-x_0, \ldots, -x_m). \qquad (7.37)$$

We can express this relation in the frequency domain as

$$G(\omega_0, \ldots, \omega_m) = \Delta^{m+1}\,\mathrm{sinc}(\omega_0\Delta/2) \cdots\, \mathrm{sinc}(\omega_m\Delta/2)F_v(-\omega_0, \ldots, -\omega_m), \qquad (7.38)$$

where $F_v(\omega_0, \ldots, \omega_m)$ is the $(m+1)$-dimensional Fourier transform of $f_v(v_0, \ldots, v_m)$.

We can immediately verify that if (7.37) is satisfied, then the joint probability distribution of any set of outputs using a probabilistic quantizer is equal to the joint probability distribution of the corresponding set of outputs using an NSD quantizer. For example,

using an NSD quantizer, for any $1 \leq k \leq m$,

$$\Pr\ (y'_0 = i_0\Delta, y'_k = i_k\Delta) = \Pi_\Delta(x_0 - i_0\Delta, x_k - k\Delta) * f_v(-x_0, -x_k), \tag{7.39}$$

where $f_v(x_0, x_k)$ is the joint distribution of x_0 and x_k and is given by integrating $f_v(x_0, \ldots, x_m)$ over all variables different than x_0 and x_k. Thus, denoting by $F_v(\omega_0, \omega_k)$ the 2-dimensional Fourier transform of $f_v(x_0, x_k)$ we have that

$$F_v(\omega_0, \omega_k) = F_v(\omega_0, 0, \ldots, 0, \omega_k, 0, \ldots, 0). \tag{7.40}$$

When using the probabilistic quantizer,

$$\Pr\ (y_0 = i_0\Delta, y_k = i_k\Delta) = \sum_{i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_m} g(x_0 - i_0\Delta, \ldots, x_m - i_m\Delta). \tag{7.41}$$

These joint probabilities are equal if

$$\sum_{i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_m} g(x_0 - i_0\Delta, \ldots, x_m - i_m\Delta) = \Pi_\Delta(x_0 - i_0\Delta, x_k - k\Delta) * f_v(-x_0, -x_m), \tag{7.42}$$

or, equivalently, if

$$\frac{1}{\Delta^{m-1}} G(\omega_0, 0, \ldots, 0, \omega_k, 0, \ldots, 0) = \Delta^2\ \mathrm{sinc}(\omega_0\Delta/2)\ \mathrm{sinc}(\omega_k\Delta/2) F_v(-\omega_0, -\omega_k). \tag{7.43}$$

From (7.40) and (7.38) we can immediately verify that (7.43) is satisfied. In a similar fashion we can verify that all other joint probabilities are equal.

Using a probabilistic quantizer with memory we can control the shape of the correlation function of the quantization error. Specifically, let $\epsilon_n = y_n - x_n$ denote the quantization error at time n. In Theorem 7.2 below we show that with ϵ_0 and ϵ_k denoting two error values separated in time by $k > 0$, $E(\epsilon_0\epsilon_k)$ can be determined directly from the 2-dimensional Fourier transform $U(\omega_0, \omega_k)$ of the joint probability function $u(x_0, x_k)$, where

$$U(\omega_0, \omega_k) = \begin{cases} \frac{1}{\Delta^{m-1}} G(\omega_0, 0, \ldots, 0, \omega_k, 0, \ldots, 0), & 1 \leq k \leq m; \\ U(\omega_0)U(\omega_k), & k > m, \end{cases} \tag{7.44}$$

and,

$$U(\omega_0) = \frac{1}{\Delta^m} G(\omega_0, 0, \dots, 0). \tag{7.45}$$

With this notation, the next result follows from [29, Theorem 3].

Theorem 7.2. *When using a probabilistic quantizer with mapping $g(x_0, \dots, x_m)$, $E(\epsilon_0^{k_1} \epsilon_k^{k_2})$ of two error values ϵ_0 and ϵ_k separated in time by $k \neq 0$ is independent of the system input if and only if*

$$\left. \frac{d^{k_1+k_2} U(\omega_0, \omega_k)}{d\omega_0^{k_1} d\omega_k^{k_2}} \right|_{\omega_0 = \frac{2\pi}{\Delta} i_0,\, \omega_k = \frac{2\pi}{\Delta} i_k} = 0, \quad i_0, i_k = \pm 1 \pm 2, \dots, \tag{7.46}$$

where $U(\omega_0, \omega_k)$ is given by (7.44) and (7.45).
When (7.46) is satisfied,

$$E(\epsilon_0^{k_1} \epsilon_k^{k_2}) = \frac{(-j)^{k_1+k_2}}{\Delta^2} \left. \frac{d^{k_1+k_2} U(\omega_0, \omega_k)}{d\omega_0^{k_1} d\omega_k^{k_2}} \right|_{\omega_0 = \omega_k = 0}. \tag{7.47}$$

Furthermore, $E(\epsilon_0^{k_1})$ is independent of the system input if and only if

$$\left. \frac{d^{k_1} U(\omega_0)}{d\omega_0^{k_1}} \right|_{\omega_0 = \frac{2\pi}{\Delta} i_0} = 0, \quad i_0 = \pm 1 \pm 2, \dots, \tag{7.48}$$

where $U(\omega_0)$ is given by (7.45).
When (7.48) is satisfied,

$$E(\epsilon_0^{k_1}) = \frac{(-j)^{k_1}}{\Delta} \left. \frac{d^{k_1} U(\omega_0)}{d\omega_0^{k_1}} \right|_{\omega_0 = 0}. \tag{7.49}$$

As a consequence of Theorems 7.1 and 7.2 we have the following Corollary.

Corollary 7.1. *Suppose we use a memoryless probabilistic quantizer with a function $g(x)$. Then the error sequence ϵ_n is uncorrelated with the input and is white if and only if*

$$\left. \frac{dG(\omega)}{d\omega} \right|_{\omega = \frac{2\pi}{\Delta} i} = 0, \quad i = \pm 1, \pm 2, \dots. \tag{7.50}$$

Using Theorem 7.2 we can choose the mapping g to control the correlation function of the quantizer error, as we show in the following example.

Example 7.3. Suppose we want to design a probabilistic quantizer with a mapping g so that the correlation function of the resulting error ϵ_n is

$$E(\epsilon_n \epsilon_{n+m}) = (\Delta^2/12)(3\delta[m] - \delta[m-1] - \delta[m+1]). \tag{7.51}$$

As we now show, this can be done by choosing $g = g(x_0, x_1)$ to be a function of $x_0 = x_n$ and $x_1 = x_{n-1}$ with Fourier transform

$$G(\omega_0, \omega_1) = \Delta^2 \operatorname{sinc}((\omega_1 - \omega_0)\Delta/2) \operatorname{sinc}^2(\omega_0\Delta/2) \operatorname{sinc}^2(\omega_1\Delta/2). \tag{7.52}$$

Note that with this choice of $G(\omega_0, \omega_1)$, (7.26) and (7.32) are satisfied.

With $k_1 = k_2 = 1$, we can immediately verify that (7.46) is satisfied, so that from (7.47)

$$E(\epsilon_0\epsilon_1) = -\frac{1}{\Delta^2} \frac{d^2 G(\omega_0, \omega_1)}{d\omega_0 d\omega_1} \bigg|_{\omega_0=\omega_1=0} = -\frac{\Delta^2}{12}, \tag{7.53}$$

where we used the fact that with $r(x) = \operatorname{sinc}(x)$, $r'(0) = 0$ and $r''(0) = -1/3$, where $r'(x)$ and $r''(x)$ denote the first and second derivatives of $r(x)$, respectively. From (7.45),

$$U(\omega_0) = \Delta \operatorname{sinc}^3(\omega_0\Delta/2). \tag{7.54}$$

For $k > 1$,

$$E(\epsilon_0\epsilon_k) = -\frac{1}{\Delta^2} \left(\frac{dU(\omega_0)}{d\omega_0} \bigg|_{\omega_0=0} \right)^2 = 0. \tag{7.55}$$

Finally, from (7.49),

$$E(\epsilon_0^2) = -\frac{1}{\Delta} \frac{d^2 U(\omega_0)}{d\omega_0^2} \bigg|_{\omega_0=0} = \frac{\Delta^2}{4}. \tag{7.56}$$

□

There are variety of applications in which shaping the noise at the output of the quantizer

is desirable. The probabilistic quantizer with memory offers a general method for shaping the correlation function of the output error, with a simple and practical implementation. All that is required is a uniform random number generator, regardless of the desired shape we wish to achieve at the output. This is particularly important in high-speed applications in which it may be prohibitively time-consuming to generate a random process with an arbitrary joint pdf.

In our closing remarks we note that the ideas presented in this chapter can also be applied to requantization. Requantization is a similar operation to quantization in which the input to the quantizer is discrete in amplitude, and the quantizer maps these discrete values to a different smaller set of discrete values. Requantization is typically used to reduce the wordlength of digital data after processing. Using a similar analysis to that presented in this chapter we can show that probabilistic requantization corresponds to NSD quantization with a discrete dither signal [29].

Chapter 8

Optimal QSP Measurements

As we discussed in Chapter 3, one of the interesting elements of quantum mechanics is that measurement vectors are constrained to be orthonormal. A fundamental problem in quantum mechanics is to construct optimal measurements subject to this constraint, that best represent a given set of state vectors. In analogy to quantum mechanics, central to the concept of QSP is the idea of imposing constraints on algorithms. The QSP framework provides a systematic method for imposing such constraints: The measurement vectors of the QSP measurement are constrained to have a certain inner product structure, as in quantum mechanics. However, since we are not limited by physical laws, we are not confined to an orthogonality constraint. Therefore, a fundamental problem in QSP is to construct optimal QSP measurements subject to a general inner product constraint on the measurement vectors.

As outlined in Chapter 4, in applications involving ROMs, we typically first identify a set of signals of interest, and then construct a measurement with measurement vectors equal to these signals. If we constrain these vectors to have a specified inner product structure, then in general they cannot be chosen to be equal to the desired signals. Instead, we choose the measurement vectors to have the required inner products, and to "best" represent the desired signals in some sense.

There are many ways to construct vectors with specified inner products. In this chapter we consider new methods that construct vectors that are closest in a least-squares (LS) sense to a given set of vectors. Specifically, the constructed vectors are chosen to minimize the sum of the squared norms of the error vectors between the constructed vectors and the

given vectors. These techniques are referred to as LS inner product shaping [27], and rely on ideas and results obtained in the context of quantum detection, which unlike QSP are subject to the constraints of quantum physics.

LS inner product shaping has potential applications to a variety of problems. One application, outlined in Chapter 3 and developed further in [26], is to a detection problem in quantum mechanics. In this context, the concept of LS orthogonalization leads to a quantum measurement with many desirable properties. In Chapters 9–12 we consider applications of LS inner product shaping to matched filter detection, minimum mean-squared error (MMSE) covariance shaping, linear estimation, and multiuser detection. These applications demonstrate that, even for problems without inherent inner product constraints, imposing such constraints in combination with optimal inner product shaping can lead to new processing techniques that in many cases exhibit improved performance over traditional methods.

Finally, we note that most signals used in digital communications are geometrically uniform (GU) [77, 78], so that the Gram matrix of inner products is a permuted matrix diagonalized by a Fourier transform matrix. Such signal sets have strong symmetry properties that are desirable in various applications such as channel coding [77, 78, 79], and multiple description source coding [76, 80]. It may therefore be useful to have a method for constructing optimal signal sets of this form, which is equivalent to constructing optimal signal sets with a specific inner product structure.

8.1 Problem Formulation

Suppose we are given a set of m vectors $\{s_i, 1 \leq i \leq m\}$ in a complex Hilbert space $\mathcal{H}$, with inner product $\langle x, y \rangle$ for any $x, y \in \mathcal{H}$. The vectors $\{s_i\}$ span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$. If the vectors are linearly independent, then $n = m$; otherwise $n < m$. Our objective is to construct a set of vectors $\{h_i, 1 \leq i \leq m\}$ with a specified inner product structure, from the given vectors $\{s_i, 1 \leq i \leq m\}$.

Based on the LS measurement we developed in the context of quantum detection (see Chapter 3), we now propose a systematic method for constructing optimal vectors with a specified inner product structure. Specifically, we seek the vectors h_i that are "closest" to

the vectors s_i in the LS sense. Thus, the vectors are chosen to minimize

$$\varepsilon_{\text{LS}} = \sum_{i=1}^{m} \langle s_i - h_i, s_i - h_i \rangle, \tag{8.1}$$

subject to the constraint

$$\langle h_i, h_k \rangle = c^2 r_{ik}, \tag{8.2}$$

for some $c > 0$ and constants r_{ik}. With S and H denoting the set transformations corresponding to s_i and h_i respectively, we may rewrite the error ε_{LS} as

$$\varepsilon_{\text{LS}} = \text{Tr}\left((S - H)^*(S - H)\right), \tag{8.3}$$

where $\text{Tr}(\cdot)$ denotes the trace, and the constraint (8.2) as

$$H^*H = c^2\mathbf{R}, \tag{8.4}$$

where $[\mathbf{R}]_{ik} = r_{ik}$ and $[\cdot]_{ik}$ denotes the ikth element of the matrix.

We may wish to constrain the constant c in (8.2), or may choose c such that the LS error ε_{LS} is minimized. Similarly, we may wish to constrain the elements r_{ik} of the matrix $\mathbf{R}$, or we may choose $\mathbf{R}$ to have a specified structure, so that the eigenvectors of $\mathbf{R}$ are fixed, but choose the eigenvalues to minimize the LS error.

For example, we may wish to construct a set of orthogonal vectors, so that $\mathbf{R}$ is a diagonal matrix with eigenvector matrix equal to $\mathbf{I}_m$, but choose the norms of the vectors, *i.e.,* the eigenvalues of $\mathbf{R}$, to minimize the LS error. As another example, we may wish to construct a cyclic set h_i so that $\langle h_i, h_k \rangle$ depends only on $k - i \bmod m$. In this case $\{\langle h_i, h_k \rangle, 1 \leq i \leq m\}$ is a cyclic permutation of $\{\langle h_1, h_k \rangle, 1 \leq k \leq m\}$ for all k, and the Gram matrix H^*H is a circulant matrix diagonalized by a DFT matrix[1], so that the eigenvectors of $\mathbf{R}$ are fixed. We may then wish to specify the values $\{\langle h_1, h_k \rangle, 1 \leq k \leq m\}$ (possibly up to a scale factor), which corresponds to specifying the eigenvalues of $\mathbf{R}$, or we may choose these values, equivalently the eigenvalues of $\mathbf{R}$, so that the LS error is minimized.

In our development we consider both the case in which $\mathbf{R}$ is fixed, and the case in which

[1] In [27] we show that a vector set has a circulant Gram matrix if and only if the set is cyclic.

the eigenvalues of $\mathbf{R}$ are chosen to minimize the LS error ε_{LS}. We will see that for fixed $\mathbf{R}$ the LS inner product shaping problem has a simple closed form solution; by contrast, if the eigenvalues of $\mathbf{R}$ are not specified, then there is no known analytical solution to the LS inner product shaping problem for arbitrary vectors s_i.

To develop the solution to the general LS inner product shaping problem, in Section 8.2 we first consider the special case in which $\mathbf{R} = \mathbf{I}_m$, so that the vectors h_i are constrained to be orthogonal with equal norm c, which we refer to as c-scaled orthonormal vectors. The optimal vectors are defined as the scaled-orthonormal LS vectors (SOLSV). In the special case in which $c = 1$, the vectors h_i are orthonormal and the optimal vectors are referred to as the orthonormal LS vectors (OLSV). As an outgrowth of the development of the SOLSV, we derive the LS tight frame which is the closest frame in a LS sense to the given vectors. Section 8.3 generalizes these results to allow for a weighted LS error. The resulting vectors are referred to as the weighted SOLSV (WSOLSV). Section 8.5 considers the case in which $\mathbf{R}$ is diagonal, so that the vectors h_i are orthogonal, but are not constrained to have equal norm. The optimal orthogonal vectors are referred to as the orthogonal LS vectors (OGLSV). We first derive the OGLSV in the case in which the norms of the vectors are specified. We then show that obtaining a closed form analytical expression for the OGLSV when the norms are chosen to minimize the LS error is in general a difficult problem. We consider a special case for which an analytical solution is derived, and then propose an iterative algorithm to construct the OGLSV in the general case. In Section 8.6 we derive the solution to the LS inner product shaping problem with arbitrary $\mathbf{R}$, by showing that it can be reduced to a LS orthogonalization problem.

8.2 Least-Squares Scaled Orthonormalization

In this section we consider the problem of constructing a set of c-scaled orthonormal vectors $\{h_i, 1 \leq i \leq m\}$ that minimize (8.1) subject to

$$\langle h_k, h_i \rangle = c^2 \delta_{ki}, \tag{8.5}$$

for some $c > 0$, or

$$H^* H = c^2 \mathbf{I}_m. \tag{8.6}$$

The minimizing vectors are referred to as the scaled-orthonormal LS vectors (SOLSV). In the special case in which $c = 1$, the vectors h_i are orthonormal and the minimizing vectors are referred to as the orthonormal LS vectors (OLSV).

Our approach to determining the SOLSV is to perform a unitary change of coordinates $\mathbf{U}$ so that in the new coordinate system, S is mapped to $\overline{S} = S\mathbf{U}$ and H is mapped to $\overline{H} = H\mathbf{U}$. With $\bar{s}_i$ denoting the vectors corresponding to $\overline{S}$, the transformation $\mathbf{U}$ is chosen such that the vectors $\{\bar{s}_i, 1 \leq i \leq n\}$ are orthogonal and $\{\bar{s}_i = 0, n+1 \leq i \leq m\}$. Since $\mathbf{U}$ is unitary and $H^*H = c^2\mathbf{I}_m$, $\overline{H}^*\overline{H} = c^2\mathbf{I}_m$ so that the vectors $\{\bar{h}_i, 1 \leq i \leq m\}$ corresponding to $\overline{H}$ are orthogonal with equal norm c^2, and the LS error defined by (8.3) between S and H is equal to the LS error between $\overline{S}$ and $\overline{H}$:

$$\mathrm{Tr}\left((S-H)^*(S-H)\right) = \mathrm{Tr}\left(\mathbf{U}^*(S-H)^*(S-H)\mathbf{U}\right) = \mathrm{Tr}\left((\overline{S}-\overline{H})^*(\overline{S}-\overline{H})\right). \tag{8.7}$$

Thus, we may first solve the LS scaled orthonormalization problem in the new coordinate system. Then, with $\widehat{\overline{H}}$ and $\widehat{H}$ denoting the optimal set transformations in the new and original coordinate systems respectively,

$$\widehat{H} = \widehat{\overline{H}}\mathbf{U}^*. \tag{8.8}$$

Such a unitary transformation is provided by the singular value decomposition (SVD) of S. Since S is a finite-dimensional transformation, it has an SVD of the form

$$S = U\Sigma\mathbf{V}^*,$$

where $U\colon\ \mathbb{C}^m \to \mathcal{H}$ is a set transformation corresponding to a set of orthonormal vectors $\{u_i, 1 \leq i \leq m\}$, so that $U^*U = \mathbf{I}_m$, Σ is an $m \times m$ matrix with the first n diagonal elements equal to $\sigma_i > 0$, and the remaining diagonal elements all equal to 0, and $\mathbf{V}$ is an $m \times m$ unitary matrix. If we choose $\overline{S} = S\mathbf{V} = U\Sigma$, then the vectors $\{\bar{s}_i, 1 \leq i \leq n\}$ are orthogonal with $\langle \bar{s}_i, \bar{s}_i \rangle = \sigma_i^2$, and $\bar{s}_i = 0$ for $n+1 \leq i \leq m$.

To determine $\widehat{\overline{H}}$, we express $\varepsilon_{\mathrm{LS}}$ of (8.1) as

$$\varepsilon_{\mathrm{LS}} = \sum_{i=1}^{m} \langle \bar{h}_i - \bar{s}_i, \bar{h}_i - \bar{s}_i \rangle = mc^2 + \sum_{i=1}^{n} \sigma_i^2 - 2\sum_{i=1}^{n} \Re\{\langle \bar{h}_i, \bar{s}_i \rangle\}. \tag{8.9}$$

We first minimize (8.9) with respect to the vectors $\{\bar{h}_i, 1 \leq i \leq n\}$. From the Cauchy-Schwarz inequality,

$$\Re\{\langle \bar{h}_i, \bar{s}_i \rangle\} \leq |\langle \bar{h}_i, \bar{s}_i \rangle| \leq \langle \bar{h}_i, \bar{h}_i \rangle^{1/2} \langle \bar{s}_i, \bar{s}_i \rangle^{1/2} = c\sigma_i, \tag{8.10}$$

with equality if and only if $\bar{h}_i = x_i \bar{s}_i$ for some $x_i > 0$, in which case we also have $\langle \bar{h}_i, \bar{h}_i \rangle = x_i^2 \langle \bar{s}_i, \bar{s}_i \rangle = c^2$, so $x_i = c/\langle \bar{s}_i, \bar{s}_i \rangle^{1/2} = c/\sigma_i$, and $\bar{h}_i = (c/\sigma_i)\bar{s}_i = cu_i$. Note, that $\bar{h}_i$ can always be chosen proportional to $\bar{s}_i$ since the vectors $\{\bar{s}_i, 1 \leq i \leq n\}$ are orthogonal.

If $n < m$, then we may choose the remaining vectors $\bar{h}_i,\ n+1 \leq i \leq m$, arbitrarily, as long as the resulting n vectors $\bar{h}_i$ are mutually orthogonal with equal norm c. This choice will not affect the residual squared error. A convenient choice which leads to the SOLSV is $\bar{h}_i = u_i, n+1 \leq i \leq m$.

If the constant c in (8.5) is specified, then $\widehat{\overline{H}} = cU$, and from (8.8),

$$\widehat{H} = c\widehat{\overline{H}}\mathbf{V}^* = cU\mathbf{V}^*. \tag{8.11}$$

The SOLSV are then defined as $\hat{h}_i = c\widehat{H}\mathbf{i}_i$, where $\mathbf{i}_i \in \mathbb{C}^m$ is the vector with kth component δ_{ik}. In particular, the OLSV corresponding to $c = 1$ are the vectors $U\mathbf{V}^*\mathbf{i}_i$.

Alternatively, we may choose to further minimize (8.9) with respect to c. Substituting the optimal vectors $\hat{\bar{h}}_i = cu_i$ back into (8.9), we choose c to minimize

$$mc^2 - 2c\sum_{i=1}^{n} \sigma_i. \tag{8.12}$$

The optimal value of c, denoted by $\hat{c}$, is therefore given by

$$\hat{c} = \frac{1}{m}\sum_{i=1}^{n} \sigma_i = \frac{1}{m}\mathrm{Tr}\left((S^*S)^{1/2}\right), \tag{8.13}$$

and the SOLSV are the vectors corresponding to the set transformation

$$\widehat{H} = \hat{c}U\mathbf{V}^*. \tag{8.14}$$

The SOLSV in both cases can be described in a unified way as the vectors corresponding

to the set transformation

$$\widehat{H} = \alpha U \mathbf{V}^*, \tag{8.15}$$

where if c in (8.5) is specified then $\alpha = c$, and if c is chosen to minimize the LS error then $\alpha = \hat{c}$ given by (8.13). Evidently, the SOLSV $\{\hat{h}_i\}$ have the property that they do not depend on the order of the vectors $\{s_i\}$.

If the singular values σ_i of S are distinct, then the vectors $u_i, 1 \leq i \leq n$ are unique (up to a phase factor $e^{j\theta_i}$). Given the vectors u_i, the columns $\{\mathbf{v}_i, 1 \leq i \leq n\}$ of $\mathbf{V}$ are uniquely determined, so $P_{\mathcal{U}}\widehat{H} = \alpha \sum_{i=1}^{n} u_i \mathbf{v}_i^*$ is unique. If, on the other hand, there are repeated singular values, then the corresponding eigenvectors are not unique. Nonetheless, the choice of singular vectors does not affect $P_{\mathcal{U}}\widehat{H}$. Indeed, if the vectors corresponding to a repeated singular value are $\{u_j\}$, then $\sum_j u_j u_j^*$ is a projection onto the corresponding eigenspace, and therefore is the same regardless of the choice of the vectors $\{u_j\}$. Thus

$$\sum_j u_j \mathbf{v}_j^* = \frac{1}{\sigma} \sum_j u_j u_j^* S, \tag{8.16}$$

independent of the choice of $\{u_j\}$, and $P_{\mathcal{U}}\widehat{H}$ is unique.

Therefore, if the vectors s_i are linearly independent so that $n = m$, then $\{\hat{h}_i = \widehat{H}\mathbf{i}_i, 1 \leq i \leq m\}$ are the unique vectors that minimize (8.1) subject to (8.5), where $\widehat{H}$ is given by (8.15). Furthermore, in this case we may express $\widehat{H}$ directly in terms of S as

$$\widehat{H} = \alpha S (S^* S)^{-1/2}. \tag{8.17}$$

Indeed, $(S^*S)^{-1/2} = \mathbf{V}(\Sigma^*\Sigma)^{-1/2}\mathbf{V}^* = \mathbf{V}\Sigma^{-1}\mathbf{V}^*$, and $S(S^*S)^{-1/2} = U\mathbf{V}^*$. From (8.17) we see that, as we expect, the optimal signals $\hat{h}_i$ lie in the space $\mathcal{U}$ spanned by the vectors s_i.

Alternatively, as can be readily verified, $\widehat{H}$ may be expressed as

$$\widehat{H} = \alpha ((SS^*)^{1/2})^{\dagger} S. \tag{8.18}$$

If the vectors s_i are linearly dependent, so that $n < m$, then the optimal $\widehat{H}$ that minimizes (8.3) subject to (8.6) is not strictly unique. However, its action in the subspace $\mathcal{U}$ spanned by the vectors s_i and the resulting residual squared error are unique. Specifically,

for any minimizing H,

$$P_{\mathcal{U}} H = \alpha U \tilde{\mathbf{I}}_n \mathbf{V}^* = \alpha S((S^*S)^{1/2})^\dagger = \alpha((SS^*)^{1/2})^\dagger S, \tag{8.19}$$

where $\tilde{\mathbf{I}}_r, 1 \leq r \leq m$ is the $m \times m$ matrix

$$\tilde{\mathbf{I}}_r = \left[\begin{array}{c|c} \mathbf{I}_r & 0 \\ \hline 0 & 0 \end{array}\right]. \tag{8.20}$$

We note that since the right-hand partial isometries and the left-hand unitary matrices in the SVD of $\widehat{H}$ and S are equal, it follows from Corollary 5.1 that if the vectors s_i are GU so that $\mathbf{V}$ is a Fourier transform matrix, then the vectors h_i are also GU with the same generating group, and generating vector $(c/\sqrt{m}) \sum_{i=1}^m u_i$.

Finally, the optimal vectors $\hat{h}_i$ satisfy

$$\langle \hat{h}_i, s_i \rangle = \langle P_{\mathcal{U}} \hat{h}_i, s_i \rangle = [\widehat{H}^* P_{\mathcal{U}} S]_{ii} = \alpha [S^*S]_{ii}^{1/2}. \tag{8.21}$$

This relation may be used to derive bounds on the inner products $\langle \hat{h}_i, s_i \rangle$ in terms of the inner products $\langle s_i, s_j \rangle$; see [104].

8.2.1 Optimal Tight Frames

We now try to gain further insight into the SOLSV in the linearly dependent case. The perspective which we now develop also leads to the construction of optimal LS tight frames.

Our problem is to find a set of scaled-orthonormal vectors that are as close as possible to the vectors s_i, where the vectors lie in an n-dimensional subspace $\mathcal{U}$. When $n < m$, there are at most n orthonormal vectors in $\mathcal{U}$. Therefore, the optimal scaled-orthonormal vectors h_i must lie partly in the orthogonal complement $\mathcal{U}^\perp$. Thus the vectors h_i span an m-dimensional subspace $\mathcal{V}$, where $\mathcal{U} \subset \mathcal{V}$. Each vector has a component in $\mathcal{U}$, $h_i^{\mathcal{U}} = P_{\mathcal{U}} h_i$, and a component in $\mathcal{U}^\perp$, $h_i^{\mathcal{U}^\perp} = P_{\mathcal{U}^\perp} h_i$. Now, we may rewrite the error ε_{LS} of (8.1) as

$$\begin{aligned} \varepsilon_{\text{LS}} &= \sum_{i=1}^m \langle s_i - h_i^{\mathcal{U}} - h_i^{\mathcal{U}^\perp}, s_i - h_i^{\mathcal{U}} - h_i^{\mathcal{U}^\perp} \rangle \\ &= \sum_{i=1}^m \left(\langle s_i - h_i^{\mathcal{U}}, s_i - h_i^{\mathcal{U}} \rangle + \langle h_i^{\mathcal{U}^\perp}, h_i^{\mathcal{U}^\perp} \rangle \right), \end{aligned} \tag{8.22}$$

since $\langle s_i - h_i^{\mathcal{U}}, h_i^{\mathcal{U}^\perp} \rangle = 0$. For any choice of scaled-orthonormal vectors h_i,

$$\sum_{i=1}^{m} \langle h_i^{\mathcal{U}}, h_i^{\mathcal{U}} \rangle = \mathrm{Tr}((H^{\mathcal{U}})^* H^{\mathcal{U}}) = \mathrm{Tr}(P_{\mathcal{U}} H H^*) = c^2 \mathrm{Tr}(P_{\mathcal{U}} P_{\mathcal{V}}) = c^2 \mathrm{Tr}(P_{\mathcal{U}}) = c^2 n, \qquad (8.23)$$

where $H^{\mathcal{U}} = P_{\mathcal{U}} H$ is the set transformation corresponding to $h_i^{\mathcal{U}} = P_{\mathcal{U}} h_i$, so that

$$\sum_{i=1}^{m} \langle h_i^{\mathcal{U}^\perp}, h_i^{\mathcal{U}^\perp} \rangle = \sum_{i=1}^{m} \left(\langle h_i, h_i \rangle - \langle h_i^{\mathcal{U}}, h_i^{\mathcal{U}} \rangle \right) = c^2 (m - n). \qquad (8.24)$$

Thus minimization of $\varepsilon_{\mathrm{LS}}$ is equivalent to minimization of

$$\varepsilon'_{\mathrm{LS}} = \sum_{i=1}^{m} \langle s_i - h_i^{\mathcal{U}}, s_i - h_i^{\mathcal{U}} \rangle + c^2 (m - n). \qquad (8.25)$$

Since $H^{\mathcal{U}} = P_{\mathcal{U}} H$, the vectors $P_{\mathcal{U}} h_i$ form a c-scaled tight frame for $\mathcal{U}$. Furthermore, Theorem 5.5 shows that any c-scaled tight frame for $\mathcal{U}$ can be expressed as $P_{\mathcal{U}} h_i$ for a set of c-scaled orthonormal vectors h_i. Therefore, finding a set of c-scaled orthonormal vectors that minimize the LS error is equivalent to finding a c-scaled tight frame for $\mathcal{U}$ that minimizes (8.25).

If c is fixed, then minimizing $\varepsilon'_{\mathrm{LS}}$ is equivalent to minimizing

$$\sum_{i=1}^{m} \langle s_i - h_i^{\mathcal{U}}, s_i - h_i^{\mathcal{U}} \rangle, \qquad (8.26)$$

so that choosing a set of c-scaled orthonormal vectors with fixed scaling c that minimize $\varepsilon_{\mathrm{LS}}$, is equivalent to choosing a c-scaled tight frame for $\mathcal{U}$ that is closest in a LS sense to the vectors s_i. The unique optimal frame is defined as the LS frame, and follows directly from (8.19) as the frame corresponding to $cS((S^*S)^\dagger)^{1/2}$.

If the scaling c is chosen to minimize the LS error, then the LS frame vectors minimizing (8.26) can be determined from the solution to the LS scaled orthonormalization problem. Specifically, as in the case in which c is fixed, minimizing (8.26) with respect to $h_i^{\mathcal{U}}$, the optimal frame vectors are the vectors corresponding to $cS((S^*S)^\dagger)^{1/2}$. Substituting these vectors back into (8.26) and minimizing with respect to c, the optimal value of c, denoted

by $\tilde{c}$, is given by

$$\tilde{c} = \frac{1}{n}\sum_{i=1}^{n}\sigma_i = \frac{1}{n}\mathrm{Tr}\left((S^*S)^{1/2}\right) = \frac{m}{n}\hat{c}, \tag{8.27}$$

where $\hat{c}$ is given by (8.13). Thus, the optimal tight frame is proportional to the orthogonal projections onto $\mathcal{U}$ of the optimal scaled-orthonormal vectors.

We conclude that choosing an optimal c-scaled tight frame for $\mathcal{U}$ with fixed c that minimizes the LS error (8.26), is equivalent to choosing a set of c-scaled orthonormal vectors that minimize the LS error $\varepsilon_{\mathrm{LS}}$. The unique orthonormal projections onto $\mathcal{U}$ of the optimal c-scaled orthonormal vectors are the optimal c-scaled frame vectors. Furthermore, choosing an optimal tight frame with unconstrained scaling that minimizes the LS error (8.26), is equivalent to choosing a set of orthogonal vectors with unconstrained equal norm that minimize $\varepsilon_{\mathrm{LS}}$, and scaling these optimal vectors by m/n.

8.2.2 Orthonormalization in $\mathbb{C}^k$

We now consider the particular case $\mathcal{H} = \mathbb{C}^k$ for some $k \geq m$. Thus, the vectors $s_i = \mathbf{s}_i$ and $h_i = \mathbf{h}_i$ are vectors in $\mathbb{C}^k$, and the set transformations S and H reduce to the $k \times m$ matrices $\mathbf{S}$ and $\mathbf{H}$ of columns $\mathbf{s}_i$ and $\mathbf{h}_i$, respectively.

Let $\mathbf{S}$ have an SVD $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$ where $\mathbf{U}$ is a $k \times k$ unitary matrix, Σ is a $k \times m$ diagonal matrix with the first n diagonal elements equal to $\sigma_i > 0$ and the remaining diagonal elements all equal to 0, and $\mathbf{V}$ is an $m \times m$ unitary matrix. Then, the optimal matrix $\widehat{\mathbf{H}}$ whose columns $\hat{\mathbf{h}}_i$ are the SOLSV that minimize the LS error of (8.1), follows from (8.15),

$$\widehat{\mathbf{H}} = \alpha\mathbf{U}\tilde{\mathbf{I}}'_m\mathbf{V}^*, \tag{8.28}$$

where $\tilde{\mathbf{I}}'_r, 1 \leq r \leq m$ is the $k \times m$ matrix

$$\tilde{\mathbf{I}}'_r = \left[\frac{\mathbf{I}_r}{0}\right]. \tag{8.29}$$

If the vectors $\mathbf{s}_i$ are linearly independent, then we may express $\widehat{\mathbf{H}}$ as

$$\widehat{\mathbf{H}} = \alpha\mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}. \tag{8.30}$$

Connection with the polar decomposition

We now show that the SOLSV are related to the polar decomposition (PD) of $\mathbf{S}$.

Let $\mathbf{A}$ denote a $k \times m$ matrix with $k \geq m$. Then $\mathbf{A}$ has a *polar decomposition* [141, 142],

$$\mathbf{A} = \mathbf{Q}\mathbf{P}, \tag{8.31}$$

where $\mathbf{Q}$ is a $k \times m$ partial isometry that satisfies $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_m$, and $\mathbf{P} = (\mathbf{A}^*\mathbf{A})^{1/2}$. The Hermitian factor $\mathbf{P}$ is always unique; the partial isometry $\mathbf{Q}$ is unique if and only if $\mathbf{A}$ has full rank. Let $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$ be the SVD of $\mathbf{A}$. Then a natural choice of $\mathbf{Q}$, which is the unique choice if $n = m$, is $\mathbf{Q} = \mathbf{U}\tilde{\mathbf{I}}'_m\mathbf{V}^*$, where $\tilde{\mathbf{I}}'_m$ is given by (8.29).

We have seen in (8.28) that the SOLSV are the columns of $\widehat{\mathbf{H}} = \alpha\mathbf{U}\tilde{\mathbf{I}}'_m\mathbf{V}^*$, where $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$ is the SVD of $\mathbf{S}$. Therefore, $(1/\alpha)\widehat{\mathbf{H}}$ is just the partial isometry in the PD of $\mathbf{S}$: $\mathbf{S} = (1/\alpha)\widehat{\mathbf{H}}(\mathbf{S}^*\mathbf{S})^{1/2}$. In particular the OLSV, corresponding to $\alpha = 1$, are the columns of the partial isometry in the PD of $\mathbf{S}$. This is consistent with the known result, that the best partial isometry approximation to a matrix $\mathbf{A}$ with PD $\mathbf{A} = \mathbf{Q}\mathbf{P}$ is the matrix $\mathbf{Q}$ [141].

Exploiting the relationship between the SOLSV and the PD, the SOLSV can be computed very efficiently by use of the many known efficient algorithms for computing the PD (see *e.g.*, [93, 143, 141, 144]).

Recently the truncated polar decomposition (TPD), a variation on the PD, has been introduced [145] and has proved to be useful for various estimation and detection problems. As we now show, the columns of the TPD of a matrix $\mathbf{S}$ are just the closest normalized frame vectors to the columns $\mathbf{s}_i$ of $\mathbf{S}$.

Let $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$ denote an arbitrary $k \times m$ matrix with rank n. Then the order-p TPD of $\mathbf{S}$ is the factorization

$$P_{\mathcal{U}_p}\mathbf{S} = [\mathbf{U}\tilde{\mathbf{I}}'_p\mathbf{V}^*][\mathbf{V}\Sigma^*\tilde{\mathbf{I}}'_p\mathbf{V}^*] = \widetilde{\mathbf{Q}}\widetilde{\mathbf{P}}, \tag{8.32}$$

where $P_{\mathcal{U}_p}$ is the orthogonal projection onto the space spanned by the first p singular vectors $\mathbf{u}_i$ of $\mathbf{S}$. From (8.32) it follows that the columns of the left-hand matrix in the order-n TPD of $\mathbf{S}$ are the optimal normalized frame vectors. Similarly, the columns of the left-hand matrix in the order-p TPD of $\mathbf{S}$, with $p < n$, are the optimal normalized tight frame vectors corresponding to the vectors $P_{\mathcal{U}_p}\mathbf{s}_i$.

Since the SOLSV are related to the PD of $\mathbf{S}$, properties of these optimal vectors can be deduced from properties of the PD (see *e.g.*, [141, 142, 143, 146]). For example, the SOLSV corresponding to columns of a symmetric nonnegative definite matrix $\mathbf{S}$ are proportional to the columns of $\mathbf{I}$. This follows immediately from the fact that if $\mathbf{S}$ is symmetric and nonnegative definite, then $\mathbf{P} = \mathbf{S}$. As another example, the SOLSV with fixed c corresponding to two vector sets $\{\mathbf{s}_i\}$ and $\{\mathbf{g}_i\}$ are the same if and only if the corresponding matrices satisfy $\mathbf{S}\mathbf{G}^* = (\mathbf{S}\mathbf{S}^*)^{1/2}(\mathbf{G}\mathbf{G}^*)^{1/2}$ [146].

Connection with the orthogonal Procrustes problem

The LS scaled orthonormalization problem is also related to the *orthogonal Procrustes problem* [147, 148, 149]. In particular, the OLSV can be obtained as a solution to an orthogonal Procrustes problem. In this problem, we are given two $k \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$, and we want to rotate $\mathbf{B}$ into $\mathbf{A}$ by seeking a unitary matrix $\mathbf{Z}$ to minimize

$$\mathrm{Tr}\left((\mathbf{A} - \mathbf{B}\mathbf{Z})^*(\mathbf{A} - \mathbf{B}\mathbf{Z})\right). \tag{8.33}$$

We can pose the minimization problem of (8.3) with $c = 1$ in (8.4) as an orthogonal Procrustes problem by choosing $\mathbf{A} = \mathbf{S}$ and $\mathbf{B}$ as a matrix whose columns form an orthonormal basis for $\mathcal{U}$. The OLSV are then the columns of $\mathbf{B}\widehat{\mathbf{Z}}$, where $\widehat{\mathbf{Z}}$ is the solution to (8.33).

We summarize our results regarding the SOLSV in the following theorem:

Theorem 8.1 (Scaled-orthonormal least-squares vectors (SOLSV)). *Let $\{s_i, 1 \leq i \leq m\}$ be a set of m vectors in a Hilbert space $\mathcal{H}$, that span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$. Let $\{\hat{h}_i, 1 \leq i \leq m\}$ be the scaled-orthonormal least-squares vectors that minimize the least-squares error defined by (8.1) subject to the constraint (8.5). In the special case in which $c = 1$, the vectors $\hat{h}_i$ are orthonormal and are defined as the orthonormal least-squares vectors. Let $S = U\Sigma\mathbf{V}^*$ and $\widehat{H}$ denote the set transformations corresponding to the vectors s_i and $\hat{h}_i$, respectively. Then the optimal $\widehat{H}$ can be chosen as*

$$\widehat{H} = \alpha U\mathbf{V}^*,$$

where

1. *if c in (8.5) is specified then $\alpha = c$;*

2. *if c is chosen to minimize the least-squares error then $\alpha = \hat{c}$ with $\hat{c} = (1/m)\,Tr\left((S^*S)^{1/2}\right)$.*

In addition,

1. *(a) If $n = m$, then*

   i. $\widehat{H} = \alpha S(S^*S)^{-1/2}$ and the scaled-orthonormal least-squares vectors lie in $\mathcal{U}$;

   ii. the scaled-orthonormal least-squares vectors are unique.

   (b) If $n < m$, then

   i. the projection of $\widehat{H}$ onto $\mathcal{U}$ is unique and is given by $P_{\mathcal{U}}\widehat{H} = \alpha U\tilde{\mathbf{I}}_n\mathbf{V}^* = \alpha S((S^*S)^{1/2})^\dagger = \alpha((SS^*)^{1/2})^\dagger S$, where $\tilde{\mathbf{I}}_n$ is given by (8.20);

   ii. if c is fixed, then the vectors $\{P_{\mathcal{U}}\hat{h}_i, 1 \le i \le m\}$ are the closest c-scaled tight frame vectors to the vectors $\{s_i, 1 \le i \le m\}$, in the least-squares sense;

   iii. if c is chosen to minimize the least-squares error, then the vectors $\{(m/n)P_{\mathcal{U}}\hat{h}_i, 1 \le i \le m\}$ are the closest tight frame vectors to the vectors $\{s_i, 1 \le i \le m\}$, in the least-squares sense.

2. *If $\mathcal{H} = \mathbb{C}^k$ for some $k \ge m$, then*

   (a) $\widehat{\mathbf{H}} = \alpha\mathbf{U}\tilde{\mathbf{I}}'_m\mathbf{V}^*$ where $\widehat{\mathbf{H}}$ is the matrix of columns $\hat{\mathbf{h}}_i$, $\mathbf{U}$ and $\mathbf{V}$ are the unitary matrices in the SVD of the matrix $\mathbf{S}$ of columns $\mathbf{s}_i$, and $\tilde{\mathbf{I}}'_m$ is given by (8.29).

   (b) $(1/\alpha)\widehat{\mathbf{H}}$ is the partial isometry in the polar decomposition of $\mathbf{S}$.

3. *If the vectors $\{s_i, 1 \le i \le m\}$ are geometrically uniform with generating group $\mathcal{Q}$, then the vectors $\{\hat{h}_i, 1 \le i \le m\}$ are also geometrically uniform with generating group $\mathcal{Q}$ and generating vector $(\alpha/\sqrt{m})\sum_{i=1}^m u_i$, where u_i are the vectors corresponding to U.*

8.3 Weighted Least-Squares Scaled Orthonormalization

In Section 8.2 we sought the scaled-orthonormal vectors $\{h_i, 1 \le i \le m\}$ to minimize the error (8.1). Essentially, we are assigning equal weights to the different errors. However, in many cases we might choose to weight these errors according to some prior knowledge regarding the vectors s_i. For example, in a detection scenario if the vector s_j is transmitted with high probability, then we might wish to assign a large weight to $\langle s_j - h_j, s_j - h_j\rangle$.

More generally, we consider the problem of minimizing the weighted squared error,

$$\varepsilon_{\mathrm{LS}}^{w} = \sum_{i,k=1}^{m} a_{ik}\langle s_i - h_i, s_k - h_k\rangle = \mathrm{Tr}\left((S-H)^*(S-H)\mathbf{A}\right) \tag{8.34}$$

subject to (8.5), where $\mathbf{A}$ is a symmetric nonnegative definite $m \times m$ weighting matrix with $[\mathbf{A}]_{ik} = a_{ik}$. The minimizing vectors are defined as the weighted SOLSV (WSOLSV). If $c = 1$ in (8.5), then the vectors h_i are orthonormal and the minimizing vectors are defined as the weighted OLSV (WOLSV).

The derivation of the solution to this minimization problem is analogous to the derivation of the SOLSV with a slight modification. With $S_w = S\mathbf{A}$, we can express $\varepsilon_{\mathrm{LS}}^{w}$ as

$$\begin{aligned}
\varepsilon_{\mathrm{LS}}^{w} &= \mathrm{Tr}\left((S-H)^*(S-H)\mathbf{A}\right) \\
&= \mathrm{Tr}\left((S_w-H)^*(S_w-H)\right) + \mathrm{Tr}\left((\mathbf{A}-\mathbf{I}_m)H^*H\right) + \mathrm{Tr}\left(\mathbf{A}(\mathbf{I}_m-\mathbf{A})S^*S\right) \\
&= \mathrm{Tr}\left((S_w-H)^*(S_w-H)\right) + c^2\mathrm{Tr}(\mathbf{A}-\mathbf{I}_m) + K,
\end{aligned} \tag{8.35}$$

where K is independent of the choice of H.

If c is fixed, then minimizing $\varepsilon_{\mathrm{LS}}^{w}$ is equivalent to minimizing $\varepsilon_{\mathrm{LS}}^{\prime w}$ where

$$\varepsilon_{\mathrm{LS}}^{\prime w} = \mathrm{Tr}\left((S_w-H)^*(S_w-H)\right). \tag{8.36}$$

Furthermore, this minimization problem is equivalent to the LS scaled orthonormalization problem, if we substitute S_w for S. Therefore we now employ the SVD of S_w, namely $S_w = U_w\Sigma_w\mathbf{V}_w^*$, and follow the derivation of Section 8.2, where we substitute S_w for S and $U_w, \mathbf{V}_w$ and Σ_w for $U, \mathbf{V}$ and Σ, respectively. The optimal set transformation $\widehat{H}_w$ then follows from Theorem 8.1,

$$\widehat{H}_w = cU_w\mathbf{V}_w^*. \tag{8.37}$$

If we further wish to minimize the weighted LS error with respect to c, then substituting $\widehat{H}_w$ back into (8.35) and minimizing with respect to c, the optimal value of c is

$$\hat{c}_w = \frac{\mathrm{Tr}\left((S_w^*S_w)^{1/2}\right)}{\mathrm{Tr}(\mathbf{A})} = \frac{\mathrm{Tr}\left((\mathbf{A}S^*S\mathbf{A})^{1/2}\right)}{\mathrm{Tr}(\mathbf{A})}. \tag{8.38}$$

The WSOLSV vectors are thus given by $\hat{h}_i^w = \widehat{H}_w \mathbf{i}_i$, where

$$\widehat{H}_w = \alpha_w U_w \mathbf{V}_w^*. \tag{8.39}$$

If c is specified then $\alpha_w = c$, and if c is chosen to minimize $\varepsilon_{\text{LS}}^w$ then $\alpha_w = \hat{c}_w$ given by (8.38). If the vectors s_i are linearly independent and $\mathbf{A}$ is invertible, then

$$\widehat{H}_w = \alpha_w S_w (S_w^* S_w)^{-1/2} = \alpha_w S\mathbf{A}(\mathbf{A}^* S^* S\mathbf{A})^{-1/2}, \tag{8.40}$$

and, as we expect, the WSOLSV lie in the space $\mathcal{U}$ spanned by the vectors s_i.

8.4 Example of the SOLSV and the WSOLSV

We now give an example illustrating the SOLSV and the WSOLSV.

Consider the two linearly independent vectors,

$$\mathbf{s}_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}^*, \ \mathbf{s}_2 = \frac{1}{2}\begin{bmatrix} -1 & \sqrt{3} \end{bmatrix}^*. \tag{8.41}$$

We wish to construct the SOLSV with scaling $c = 1$; thus we seek the orthonormal vectors that are closest in a LS sense to the vectors $\mathbf{s}_i$. We begin by forming the matrix $\mathbf{S}$,

$$\mathbf{S} = \frac{1}{2}\begin{bmatrix} 2 & -1 \\ 0 & \sqrt{3} \end{bmatrix}. \tag{8.42}$$

We then determine the SVD $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$, which yields

$$\mathbf{U} = \frac{1}{2}\begin{bmatrix} \sqrt{3} & -1 \\ -1 & -\sqrt{3} \end{bmatrix}, \ \Sigma = \frac{1}{\sqrt{2}}\begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}, \ \mathbf{V} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}. \tag{8.43}$$

From (8.28) and (8.29), we now have

$$\widehat{\mathbf{H}} = \mathbf{U}\mathbf{V}^* = \begin{bmatrix} 0.97 & -0.26 \\ 0.26 & 0.97 \end{bmatrix}, \tag{8.44}$$

and

$$\hat{\mathbf{h}}_1 = \begin{bmatrix} 0.97 & 0.26 \end{bmatrix}^*, \ \hat{\mathbf{h}}_2 = \begin{bmatrix} -0.26 & 0.97 \end{bmatrix}^*, \tag{8.45}$$

where $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ are the optimal orthonormal vectors that minimize the LS error defined by (8.1). Using (8.30) we may express the optimal vectors directly in terms of the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$,

$$\widehat{\mathbf{H}} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2} = \mathbf{S} \begin{bmatrix} 1.12 & 0.30 \\ 0.30 & 1.12 \end{bmatrix}, \tag{8.46}$$

thus

$$\hat{\mathbf{h}}_1 = 1.12\mathbf{s}_1 + 0.30\mathbf{s}_2, \ \ \hat{\mathbf{h}}_2 = 0.30\mathbf{s}_1 + 1.12\mathbf{s}_2. \tag{8.47}$$

Figure 8-1 depicts the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ together with the optimal orthonormal vectors $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$. As is evident from (8.47) and from Fig. 8-1, the optimal vectors are as close as possible to the vectors $\mathbf{s}_i$, given that they must be orthonormal.

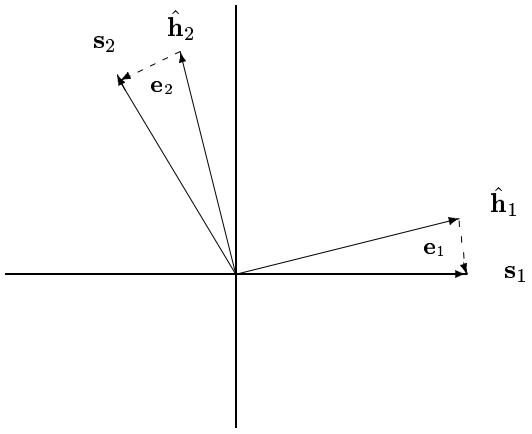


Figure 8-1: 2-dimensional example of the OLSV. $\mathbf{s}_1$ and $\mathbf{s}_2$ are given by (8.41), the optimal OLSV $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$ are given by (8.45) and are orthonormal, and $\mathbf{e}_i = \mathbf{s}_i - \hat{\mathbf{h}}_i, i = 1, 2$.

Suppose now we want to minimize a weighted LS error as in (8.34), where we choose

the weighting matrix $\mathbf{A}$ as the diagonal matrix

$$\mathbf{A} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.8 \end{bmatrix}. \tag{8.48}$$

This choice of weights corresponds to heavily penalizing deviations of $\mathbf{h}_2$ from $\mathbf{s}_2$. Consequently, we expect the WOLSV to be such that $\hat{\mathbf{h}}_2^w$ is closer to $\mathbf{s}_2$ than $\hat{\mathbf{h}}_1^w$ is to $\mathbf{s}_1$.

To determine the WOLSV we compute the SVD of $\mathbf{SA} = \mathbf{U}_w \Sigma_w \mathbf{V}_w^*$. The optimal WOLSV are then given by the columns of $\widehat{\mathbf{H}}_w = \mathbf{U}_w \mathbf{V}_w^*$, which yields

$$\hat{\mathbf{h}}_1^w = \begin{bmatrix} 0.91 & 0.41 \end{bmatrix}^*, \ \hat{\mathbf{h}}_2^w = \begin{bmatrix} -0.41 & 0.91 \end{bmatrix}^*. \tag{8.49}$$

We may express the optimal WOLSV directly in terms of the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$,

$$\widehat{\mathbf{H}}_w = \mathbf{SA}(\mathbf{AS}^*\mathbf{SA})^{-1/2} = \mathbf{S} \begin{bmatrix} 1.15 & 0.12 \\ 0.47 & 1.05 \end{bmatrix}, \tag{8.50}$$

thus

$$\hat{\mathbf{h}}_1^w = 1.15\mathbf{s}_1 + 0.47\mathbf{s}_2, \ \ \hat{\mathbf{h}}_2^w = 0.12\mathbf{s}_1 + 1.05\mathbf{s}_2. \tag{8.51}$$

Figure 8-2 depicts the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ together with the optimal orthonormal vectors $\hat{\mathbf{h}}_1^w$ and $\hat{\mathbf{h}}_2^w$. As is evident from (8.51) and from Fig. 8-2, $\hat{\mathbf{h}}_2^w$ is much closer to $\mathbf{s}_2$ then $\hat{\mathbf{h}}_1^w$ is to $\mathbf{s}_1$. Comparing Fig. 8-1 and 8-2 we see that the weighting results in a rotation of the OLSV $\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2$ towards $\mathbf{s}_2$.

8.5 Least-Squares Orthogonalization

We now consider the problem of constructing an *orthogonal* set of vectors from a given set of vectors $\{s_i, 1 \leq i \leq m\}$, where we do not constrain the norms of the vectors to be equal to a constant c. Instead, we may wish to constrain the vectors $\{h_i, 1 \leq i \leq m\}$ to have some specified norm, *e.g.*, $\langle h_i, h_i \rangle = \langle s_i, s_i \rangle$, or we may choose the vectors $\{h_i\}$ to be orthogonal and to minimize the LS error ε_{LS} of (8.1). In this case $\langle h_i, h_i \rangle$ will be such that ε_{LS} is minimized. The orthogonal vectors minimizing the LS error are defined as the orthogonal LS vectors (OGLSV).

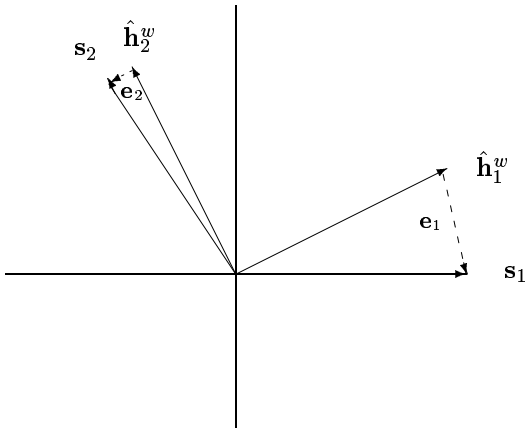


Figure 8-2: 2-dimensional example of the WOLSV. The weights are chosen as $a_{11} = 0.2, a_{22} = 0.8$, and $a_{12} = a_{21} = 0$, the optimal vectors $\hat{\mathbf{h}}_1^w$ and $\hat{\mathbf{h}}_2^w$ are given by (8.49) and are orthonormal, and $\mathbf{e}_i = \mathbf{s}_i - \hat{\mathbf{h}}_i^w, i = 1, 2$.

8.5.1 Orthogonalization With Constrained Norms

We first consider the problem of constructing a set of vectors $\{h_i, 1 \leq i \leq m\}$ that minimize ε_{LS} of (8.1), subject to the constraint

$$\langle h_i, h_k \rangle = c^2 c_i^2 \delta_{ik}, \tag{8.52}$$

where the scalars $c_i \geq 0$ are specified and $c > 0$ may be specified, or may be chosen to minimize the LS error.

We assume without loss of generality that $c_i > 0$ for $1 \leq i \leq k$. Then minimizing ε_{LS} is equivalent to minimizing

$$\varepsilon'_{\text{LS}} = \sum_{i=1}^{m} c_i^2 \langle \tilde{s}_i - \tilde{h}_i, \tilde{s}_i - \tilde{h}_i \rangle, \tag{8.53}$$

where the vectors $\tilde{h}_i$ are c-scaled orthonormal vectors such that $\tilde{h}_i = (1/c_i)h_i, 1 \leq i \leq k$, and the vectors $\tilde{s}_i$ are defined by $\tilde{s}_i = (1/c_i)s_i, 1 \leq i \leq k$ and $\tilde{s}_i = 0, k+1 \leq i \leq m$. Comparing (8.53) with (8.34) we see that the scaled-orthonormal vectors $\tilde{h}_i$ that minimize (8.53) are the WSOLSV corresponding to the vectors $\{\tilde{s}_i, 1 \leq i \leq m\}$ with weighting matrix

$\mathbf{A} = \mathbf{C}^2$, where $\mathbf{C}$ is the diagonal matrix with diagonal elements c_i.

Thus the vectors $\hat{\tilde{h}}_i$ that minimize (8.53) are given by $\hat{\tilde{h}}_i = \widehat{\widetilde{H}}\mathbf{i}_i$, where from (8.39),

$$\widehat{\widetilde{H}} = \tilde{\alpha}\widetilde{U}\widetilde{\mathbf{V}}^*. \tag{8.54}$$

Here $\widetilde{S} = S\mathbf{C}^\dagger$ is the set transformation corresponding to the vectors $\tilde{s}_i$, $\mathbf{C}^\dagger$ is the diagonal matrix with diagonal elements $1/c_i, 1 \le i \le k$ and 0 otherwise, and $\widetilde{S}\mathbf{C}^2 = \widetilde{U}\widetilde{\Sigma}\widetilde{\mathbf{V}}^*$ is the SVD of $\widetilde{S}\mathbf{C}^2 = S\mathbf{C}$. If c is specified then $\tilde{\alpha} = c$, and if c is chosen to minimize the LS error then $\tilde{\alpha} = \tilde{c}$ where from (8.38),

$$\tilde{c} = \frac{\mathrm{Tr}\left((\mathbf{C}S^*S\mathbf{C})^{1/2}\right)}{\mathrm{Tr}(\mathbf{C}^2)}. \tag{8.55}$$

The OGLSV are then given by $\hat{h}_i = c_i\hat{\tilde{h}}_i = \widehat{H}\mathbf{i}_i$, where

$$\widehat{H} = \tilde{\alpha}\widetilde{U}\widetilde{\mathbf{V}}^*\mathbf{C}. \tag{8.56}$$

Note that the vectors $\hat{h}_i$ lie in the space $\mathcal{V}$ spanned by the first k columns of $\widetilde{U}$. This follows from the fact that with $T = S\mathbf{C}$, T^*T is a block diagonal matrix whose lower $(m-k)\times(m-k)$ block is a 0 matrix with all 0 entries, so that $\widetilde{\mathbf{V}}$ is also a block diagonal matrix. Then, the the last $m-k$ elements of each of the first k columns of the matrix $\widetilde{\mathbf{V}}\mathbf{C}$ are all equal 0, and the remaining columns of $\widetilde{\mathbf{V}}\mathbf{C}$ are all 0.

Therefore, as we now show, if $k \le n$, then we can always choose $\widehat{H}$ so that the vectors $\hat{h}_i$ lie in the n-dimensional space $\mathcal{U}$ spanned by the vectors s_i. Specifically, if $l = \mathrm{rank}(S\mathbf{C})$ is equal to k, then $\mathcal{V}$ is equal to the space spanned by the first k columns of $S\mathbf{C}$ so that $\mathcal{V} \subseteq \mathcal{U}$. Since $\hat{h}_i \in \mathcal{V}$, it follows mediately that $\hat{h}_i \in \mathcal{U}$. If $l < k$ then only the first l columns of $\widetilde{\mathcal{U}}$ are specified, and the remaining columns can be chosen arbitrarily. Since these l columns span a subset of $\mathcal{U}$, we can always choose the next $k-l$ columns so that the space $\mathcal{V}$ spanned by the first k columns of $\widetilde{\mathcal{U}}$ is also a subset of $\mathcal{U}$. Then since $\hat{h}_i \in \mathcal{V}$ and $\mathcal{V} \subseteq \mathcal{U}$, $\hat{h}_i \in \mathcal{U}$.

If the vectors s_i are linearly independent and $c_i > 0$ for all i, then

$$\widehat{H} = \tilde{\alpha}S\mathbf{C}(\mathbf{C}S^*S\mathbf{C})^{-1/2}\mathbf{C}. \tag{8.57}$$

8.5.2 Unconstrained Orthogonalization

We now consider the orthogonalization problem with unconstrained norms. Thus, we seek a set of orthogonal vectors h_i that minimize (8.1), subject to

$$\langle h_i, h_k \rangle = 0, \quad i \neq k. \tag{8.58}$$

Expressing ε_{LS} as

$$\varepsilon_{\text{LS}} = \sum_{i=1}^{m} \left(\langle s_i, s_i \rangle + \langle h_i, h_i \rangle - 2\Re\{\langle h_i, s_i \rangle\} \right), \tag{8.59}$$

it follows that minimization of ε_{LS} is equivalent to minimization of

$$\varepsilon'_{\text{LS}} = \sum_{i=1}^{m} \left(\langle h_i, h_i \rangle - 2\Re\{\langle h_i, s_i \rangle\} \right). \tag{8.60}$$

Let $h_i = b_i \tilde{h}_i$, where $b_i^2 = \langle h_i, h_i \rangle$ and $\langle \tilde{h}_i, \tilde{h}_i \rangle = 1$. Then

$$\varepsilon'_{\text{LS}} = \sum_{i=1}^{m} \left(b_i^2 - 2b_i \Re\{\langle \tilde{h}_i, s_i \rangle\} \right). \tag{8.61}$$

To determine the optimal vectors h_i we have to minimize ε'_{LS} with respect to b_i and $\tilde{h}_i$. Fixing $\tilde{h}_i$ and minimizing with respect to b_i, the optimal value of b_i, denoted $\hat{b}_i$, is given by

$$\hat{b}_i = \Re\{\langle \tilde{h}_i, s_i \rangle\}. \tag{8.62}$$

Substituting $\hat{b}_i$ back into (8.61), we get that the vectors $\tilde{h}_i$ are chosen to maximize

$$R_{hs} = \sum_{i=1}^{m} \Re^2\{\langle \tilde{h}_i, s_i \rangle\}, \tag{8.63}$$

subject to the constraint

$$\langle \tilde{h}_i, \tilde{h}_k \rangle = \delta_{ik}. \tag{8.64}$$

Obtaining a closed form analytical expression for the vectors $\tilde{h}_i$ that maximize (8.63) subject to (8.64) is in general a difficult problem. In fact, in the case in which the vectors

h_i and s_i are real, this problem is equivalent to a quantum detection problem discussed in Section 3.4, for which there is no known analytical solution in general [26, 23]. Based on results we obtained in the context of quantum detection [26], we now show that in the special case in which the vectors $\{s_i\}$ are geometrically uniform (GU), an analytical solution can be obtained. An iterative algorithm for constructing the orthonormal vectors that maximize (8.63) for arbitrary vectors $\{s_i\}$ is considered in Section 8.5.4.

8.5.3 Maximizing R_{hs} for Geometrically Uniform Vector Sets

To obtain a more convenient expression for R_{hs}, let $S = U\Sigma\mathbf{V}^*$ and $\widetilde{H}$ denote the set transformations corresponding to s_i and $\tilde{h}_i$, respectively. Since $\{\tilde{h}_i\}$ are proportional to the vectors closest to the vectors $\{s_i\}$, the space spanned by the vectors s_i is a subspace of the space spanned by the vectors $\tilde{h}_i$. In addition, the vectors $\tilde{h}_i$ are orthonormal. Consequently, $\widetilde{H}$ has an SVD of the form $\widetilde{H} = U\mathbf{Q}$ for some unitary $m \times m$ matrix $\mathbf{Q}$. With $\mathbf{v}_i$ and $\mathbf{q}_i$ denoting the columns of $\mathbf{V}^*$ and $\mathbf{Q}^*$ respectively, we can express R_{hs} as

$$R_{hs} = \sum_{i=1}^{m} \Re^2\{\langle \tilde{h}_i, s_i\rangle\} = \sum_{i=1}^{m} \Re^2\{\langle U^*\tilde{h}_i, U^*s_i\rangle\} = \sum_{i=1}^{m} \Re^2\{\langle \mathbf{q}_i, \Sigma\mathbf{v}_i\rangle\}.$$

Our problem thus reduces to finding a set of orthonormal vectors $\mathbf{q}_i$ that maximize $\sum_{i=1}^{m} \Re^2\{\langle \mathbf{q}_i, \Sigma\mathbf{v}_i\rangle\}$, where the vectors $\mathbf{v}_i$ are also orthonormal. Using the Cauchy-Schwarz inequality,

$$\begin{aligned} R_{hs} &= \sum_{i=1}^{m} \Re^2\{\langle \Sigma^{1/2}\mathbf{q}_i, \Sigma^{1/2}\mathbf{v}_i\rangle\} \le \sum_{i=1}^{m} |\langle \Sigma^{1/2}\mathbf{q}_i, \Sigma^{1/2}\mathbf{v}_i\rangle|^2 \\ &\le \sum_{i=1}^{m} \langle \mathbf{q}_i, \Sigma\mathbf{q}_i\rangle\langle \mathbf{v}_i, \Sigma\mathbf{v}_i\rangle, \end{aligned} \tag{8.65}$$

with equality if and only if $\Sigma^{1/2}\mathbf{q}_i = x_i\Sigma^{1/2}\mathbf{v}_i$ for some $x_i > 0$ and all i. In particular, we have equality for $\mathbf{q}_i = \mathbf{v}_i$. In general the bound of (8.65) depends on the unknown vectors $\mathbf{q}_i$. However, in the special case where $\langle \mathbf{v}_i, \Sigma\mathbf{v}_i\rangle = K$ independent of i, (8.65) reduces to

$$R_{hs} \le K\sum_{i=1}^{m} \langle \mathbf{q}_i, \Sigma\mathbf{q}_i\rangle = K\mathrm{Tr}(\mathbf{Q}\Sigma\mathbf{Q}^*) = K\mathrm{Tr}(\Sigma), \tag{8.66}$$

with equality if $\mathbf{v}_i = \mathbf{q}_i$. Since $\langle \mathbf{v}_i, \Sigma\mathbf{v}_i\rangle = [(S^*S)^{1/2}]_{ii}$, we conclude that if the diagonal

elements of $(S^*S)^{1/2}$ are equal, then the optimal orthonormal vectors maximizing R_{hs} correspond to $U\mathbf{V}^*$, and from Theorem 8.1 are just the OLSV with respect to the vectors $\{s_i\}$. We note that as with the OLSV, if the vectors $\{s_i\}$ are linearly dependent, then the vectors maximizing R_{hs} are not unique. However, their projections onto $\mathcal{U}$ are always unique.

If the vectors $\{s_i\}$ are GU, then from Corollary 5.1 $\mathbf{V}$ is a Fourier transform matrix and the components of the vectors $\mathbf{v}_i$ have equal magnitude $1/\sqrt{m}$ so that for all i, $\langle \mathbf{v}_i, \Sigma \mathbf{v}_i \rangle = (1/m)\sum_j \sigma_j$. Therefore, in this case the OLSV maximize R_{hs}. The optimal vectors $\hat{\tilde{h}}_i$ are then the vectors corresponding to $\hat{\tilde{H}} = U\mathbf{V}^*$, and from (8.62), $\hat{b}_i = [(S^*S)^{1/2}]_{ii} = (1/m)\sum_j \sigma_j = \hat{c}$ for all i, where $\hat{c}$ is the optimal scaling in the SOLSV, given by (8.13). Thus for GU vector sets the OGLSV are equal to the SOLSV with scaling c that minimizes the LS error.

8.5.4 Iterative Algorithm Maximizing R_{hs} for Arbitrary Vector Sets

For simplicity of exposition, we assume throughout this section that $\mathcal{H} = \mathcal{R}^k$ for some $k \geq m$, so that $R_{hs} = \sum_{i=1}^m \langle \tilde{\mathbf{h}}_i, \mathbf{s}_i \rangle^2$.

The proposed algorithm proceeds as follows. Starting with an arbitrary matrix with orthonormal columns, at each iteration we construct a new matrix with orthonormal columns by multiplying the current matrix by an orthogonal matrix. The orthogonal matrix is chosen so that R_{hs} does not decrease form iteration to iteration, where at each iteration R_{hs} is computed using the orthonormal columns of the new matrix. Since R_{hs} is bounded above for any choice of orthonormal vectors, the iterations are guaranteed to converge.

The algorithm is initialized by choosing a matrix $\mathbf{H}^{(0)}$ with orthonormal columns $\mathbf{h}_i^{(0)}$. A good choice is the matrix $\mathbf{H}^{(0)} = \mathbf{U}\tilde{\mathbf{I}}'_m\mathbf{V}^*$ where $\mathbf{U}$ and $\mathbf{V}$ are the unitary matrices in the SVD of the matrix $\mathbf{S}$ of columns $\mathbf{s}_i$, and $\tilde{\mathbf{I}}'_m$ is given by (8.29), so that the columns $\mathbf{h}_i^{(0)}$ are the closest orthonormal vectors in a LS sense to the vectors $\mathbf{s}_i$. If the vectors $\mathbf{s}_i$ are linearly independent, then $\mathbf{H}^{(0)} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}$. Then, for $j = 0, 1, 2, \ldots$ we choose an orthogonal matrix $\mathbf{Q}^{(j)}$, and set $\mathbf{H}^{(j+1)} = \mathbf{H}^{(j)}\mathbf{Q}^{(j)}$. If the columns of $\mathbf{H}^{(j)}$ are orthonormal, and $\mathbf{Q}^{(j)}$ is an orthogonal matrix, then the columns of $\mathbf{H}^{(j+1)}$ are also orthonormal. Since the columns of $\mathbf{H}^{(0)}$ are orthonormal, $\mathbf{H}^{(j)}$ will have orthonormal columns for all j.

Suppose we can choose $\mathbf{Q}^{(j)}$ so that for all j,

$$R_{hs}^{(j+1)} \geq R_{hs}^{(j)}. \tag{8.67}$$

Here $R_{hs}^{(j)} = \sum_{i=1}^{m} \left(d_{ii}^{(j)}\right)^2$ where $d_{ii}^{(j)} = \langle \mathbf{h}_i^{(j)}, \mathbf{s}_i \rangle$, and $\mathbf{h}_i^{(j)}$ denotes the ith column of $\mathbf{H}^{(j)}$. Then, since $R_{hs}^{(j+1)} \leq \sum_{i=1}^{m} \langle \mathbf{s}_i, \mathbf{s}_i \rangle$ for all j, the iterations are guaranteed to converge.

Thus, the crux of the algorithm is choosing $\mathbf{Q}^{(j)}$ so that (8.67) is satisfied for all j. This can be accomplished by choosing $\mathbf{Q}^{(j)}$ as an "optimal" *Givens rotation* [93]. A Givens rotation $\mathbf{J}(r, l, \theta)$ with $1 \leq r, l \leq m, r \neq l$, is an orthogonal matrix that is equal to the identity matrix except for the entries

$$\begin{bmatrix} \mathbf{J}_{rr} & \mathbf{J}_{rl} \\ \mathbf{J}_{lr} & \mathbf{J}_{ll} \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \triangleq \mathbf{J}(\theta), \tag{8.68}$$

where $\mathbf{J}_{ik} = [\mathbf{J}(r, l, \theta)]_{ik}$, $c = \cos(\theta)$, and $s = \sin(\theta)$.

Now, let $\mathbf{Q}^{(j)} = \mathbf{J}(r, l, \hat{\theta})$, where $\hat{\theta}$ is chosen to maximize $R_{hs}^{(j+1)}$. Since for $\theta = 0$, $\mathbf{J}(r, l, 0) = \mathbf{I}_m$ and $R_{hs}^{(j+1)} = R_{hs}^{(j)}$, we are guaranteed that for an optimal choice of θ, (8.67) is satisfied. note, that if $\mathbf{H}^{(j+1)} = \mathbf{H}^{(j)}\mathbf{J}(r, l, \theta)$, then $d_{kk}^{(j+1)} = d_{kk}^{(j)}$ for $k \neq l, r$. Therefore, choosing θ to maximize $R_{hs}^{(j+1)}$ is equivalent to choosing θ to maximize

$$R_{hs}^{\prime(j+1)} = \left(d_{ll}^{(j+1)}\right)^2 + \left(d_{rr}^{(j+1)}\right)^2. \tag{8.69}$$

Let j, r, l be fixed, and let $\mathbf{D}^{(j)} = (\mathbf{H}^{(j)})^* \mathbf{S}$. Denote by $\mathbf{D}^{\prime(j)}$ the 2×2 matrix

$$\mathbf{D}^{\prime(j)} = \begin{bmatrix} [\mathbf{D}^{(j)}]_{rr} & [\mathbf{D}^{(j)}]_{rl} \\ [\mathbf{D}^{(j)}]_{lr} & [\mathbf{D}^{(j)}]_{ll} \end{bmatrix} \triangleq \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}. \tag{8.70}$$

Then $R_{hs}^{\prime(j+1)}$ is equal to the sum of the squares of the diagonal elements of $\mathbf{D}^{\prime(j+1)} = \mathbf{J}^*(\theta)\mathbf{D}^{\prime(j)}$, where $\mathbf{J}(\theta)$ is defined by (8.68). Thus,

$$R_{hs}^{\prime(j+1)} = (cb_{11} + sb_{21})^2 + (cb_{22} - sb_{12})^2. \tag{8.71}$$

In Appendix A we show that $R_{hs}^{\prime(j+1)}$ is maximized when

$$\hat{\theta} = \begin{cases} \frac{1}{2}\tan^{-1}\left(\frac{y}{x}\right) \text{ and } x\cos(2\hat{\theta}) > 0, & x \neq 0; \\ \pi/4, & x = 0, y > 0; \\ -\pi/4, & x = 0, y < 0; \\ 0, & x = y = 0, \end{cases} \tag{8.72}$$

where $x = b_{11}^2 + b_{22}^2 - b_{21}^2 - b_{12}^2$ and $y = b_{11}b_{21} - b_{22}b_{12}$.

The iterations are continued until convergence, where in each iteration we choose different values of r and l.

We now summarize the iterative algorithm:

1. Choose the vectors $\mathbf{h}_i^{(0)}$ as the columns of $\mathbf{H}^{(0)} = \mathbf{U}\tilde{\mathbf{I}}'_m\mathbf{V}^*$ where $\mathbf{U}$ and $\mathbf{V}$ are the unitary matrices in the SVD of the matrix $\mathbf{S}$ of columns $\mathbf{s}_i$, and $\tilde{\mathbf{I}}'_m$ is given by (8.29), and set $j = 0$;

2. choose r and l arbitrarily so that $1 \leq r, l \leq m$ and $r \neq l$;

3. compute $b_{11} = \langle \mathbf{h}_r^{(j)}, \mathbf{s}_r \rangle, b_{12} = \langle \mathbf{h}_r^{(j)}, \mathbf{s}_l \rangle, b_{21} = \langle \mathbf{h}_l^{(j)}, \mathbf{s}_r \rangle$ and $b_{22} = \langle \mathbf{h}_l^{(j)}, \mathbf{s}_l \rangle$;

4. compute $\mathbf{H}^{(j+1)} = \mathbf{H}^{(j)}\mathbf{J}(r, l, \hat{\theta})$ where $\hat{\theta}$ is given by (8.72);

5. set $j = j + 1$ and go to step (2).

Our iterative algorithm can be shown to be equivalent to the algorithm developed in [150] in the context of quantum detection, which is derived using quantum-mechanical ideas and concepts. However, since our algorithm does not invoke such considerations, its derivation is more straightforward.

Based on results derived in that context it can be shown that the vectors $\mathbf{h}_i$ maximizing R_{hs} are unique (up to multiplication by -1) [150, 21, 23, 22]. Furthermore, the optimal vectors $\mathbf{h}_i$ must be such that the matrix Υ defined by $[\Upsilon]_{ik} = \langle \mathbf{h}_i, \mathbf{s}_k \rangle \langle \mathbf{h}_k, \mathbf{s}_k \rangle$, is nonnegative definite [150]. Therefore, upon convergence of the algorithm we can test whether or not the vectors $\mathbf{h}_i$ are the optimal vectors maximizing R_{hs} or whether the algorithm converged to a local maximum, by checking if Υ is nonnegative definite. If the algorithm converged to a local minimum, then we may slightly rotate the matrix $\mathbf{H}$ of columns $\mathbf{h}_i$, *i.e.,* multiply $\mathbf{H}$ by a unitary matrix $\mathbf{U}$, such that the rotated vectors $\mathbf{h}'_i = \mathbf{U}\mathbf{h}_i$ result in a higher R_{hs}; these vectors form the initial conditions $\mathbf{h}_i^{(0)}$ for resumption of the main algorithm. A formula for determining $\mathbf{U}$ can be found in [150, Appendix I]. We note that when the initial conditions for the algorithm are chosen as the OLSV, $R_{hs}^{(0)}$ will be pretty close to the maximal value, and the algorithm is unlikely to converge to a local maximum.

We summarize our results regarding the OGLSV in the following theorem:

Theorem 8.2 (Orthogonal least-squares vectors (OGLSV)). *Let $\{s_i, 1 \leq i \leq m\}$ be a set of m vectors in a Hilbert space $\mathcal{H}$, that span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$.*

Let $\{\hat{h}_i, 1 \le i \le m\}$ denote the orthogonal least-squares vectors that minimize the least-squares error (8.1), subject to the constraint (8.52), and let $\widehat{H}$ denote the set transformation corresponding to the vectors $\hat{h}_i$. Then,

1. *If $\langle h_i, h_i \rangle = c^2 c_i^2$ where the constants $c_i \ge 0$ are given, then let $\mathbf{C}$ denote the diagonal matrix with diagonal elements c_i, and let $S\mathbf{C} = \widetilde{U}\widetilde{\Sigma}\widetilde{\mathbf{V}}^*$. Then $\widehat{H}$ can be chosen as*

$$\widehat{H} = \tilde{\alpha}\widetilde{U}\widetilde{\mathbf{V}}^*\mathbf{C},$$

where

   (a) *if c is specified then $\tilde{\alpha} = c$;*

   (b) *if c is chosen to minimize the least-squares error then $\tilde{\alpha} = \tilde{c}$ where $\tilde{c} = Tr\left((\mathbf{C}S^*S\mathbf{C})^{1/2}\right) / Tr(\mathbf{C}^2)$.*

In addition,

   (a) *If $rank(\mathbf{C}) \le n$, then the vectors $\hat{h}_i$ can be chosen to lie in $\mathcal{U}$;*

   (b) *If $c_i > 0$ for all i, then*

      i. *the projection of $\widehat{H}$ onto $\mathcal{U}$ is unique and is given by $P_{\mathcal{U}}\widehat{H} = \tilde{\alpha}S\mathbf{C}((\mathbf{C}S^*S\mathbf{C})^{1/2})^\dagger\mathbf{C} = \tilde{\alpha}((\mathbf{C}SS^*\mathbf{C})^{1/2})^\dagger S\mathbf{C}^2$;*

      ii. *if in addition $n = m$, then*

         A. *$\widehat{H} = \tilde{\alpha}S\mathbf{C}(\mathbf{C}S^*S\mathbf{C})^{-1/2}\mathbf{C}$;*

         B. *the orthogonal least-squares vectors are unique.*

2. *If the squared norms $\langle h_i, h_i \rangle$ are chosen to minimize the least-squares error, then $\widehat{H}$ can be computed using the iterative algorithm of Section 8.5.4. If the vectors $\{s_i, 1 \le i \le m\}$ are geometrically uniform, then $\widehat{H} = \alpha U\mathbf{V}^*$ where $S = U\Sigma\mathbf{V}^*$ and $\alpha = (1/m)\,Tr\left((S^*S)^{1/2}\right)$, and the vectors $\hat{h}_i$ are equal to the scaled-orthonormal least-squares vectors with optimal scaling, with respect to the vectors s_i.*

8.6 Least-Squares Inner Product Shaping

We now consider the general LS inner product shaping problem in which we seek the vectors $\{h_i, 1 \le i \le m\}$ with inner products $\langle h_i, h_k \rangle = c^2[\mathbf{R}]_{ik}$ for some $\mathbf{R}$ and $c > 0$, that are closest

in a LS sense to the vectors $\{s_i, 1 \leq i \leq m\}$. We assume that $\mathbf{R}$ has an eigendecomposition $\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^*$, where $\mathbf{Q}$ is unitary and Λ is diagonal. In our development we consider both the case in which $\mathbf{R}$ is specified, and the case in which the eigenvectors of $\mathbf{R}$, or the matrix $\mathbf{Q}$, are specified and the eigenvalues, or the matrix Λ, are chosen to minimize the LS error.

8.6.1 Constrained Least-Squares Inner Product Shaping

Suppose that $\mathbf{R}$ is a specified rank-r matrix, so that $[\Lambda]_{ii} > 0, 1 \leq i \leq r$, and $[\Lambda]_{ii} > 0, r+1 \leq i \leq m$. We then solve the LS inner product shaping problem by performing a unitary change of coordinates $\mathbf{Q}$, so that in the new coordinate system, S is mapped to $\overline{S} = S\mathbf{Q}$ and H is mapped to $\overline{H} = H\mathbf{Q}$. With $\bar{h}_i$ denoting the vectors corresponding to $\overline{H}$, $\langle \bar{h}_i, \bar{h}_k \rangle = c^2[\Lambda]_{ik}$, so that the vectors $\{\bar{h}_i, 1 \leq i \leq r\}$ are orthogonal with $\langle \bar{h}_i, \bar{h}_i \rangle = c^2[\Lambda]_{ii}$, and $\bar{h}_i = 0, r+1 \leq i \leq m$. Since $\mathbf{Q}$ is unitary,

$$\varepsilon_{\text{LS}} = \text{Tr}\left((S-H)^*(S-H)\right) = \text{Tr}\left(\mathbf{Q}^*(S-H)^*(S-H)\mathbf{Q}\right) = \text{Tr}\left((\overline{S}-\overline{H})^*(\overline{S}-\overline{H})\right). \tag{8.73}$$

The constraint (8.4) can be restated as

$$\overline{H}^*\overline{H} = c^2\Lambda. \tag{8.74}$$

Thus we now seek a set of orthogonal vectors $\bar{h}_i$ with $\langle \bar{h}_i, \bar{h}_i \rangle = c^2[\Lambda]_{ii}$ that are closest in a LS sense to the vectors $\bar{s}_i$. The optimal $\overline{H}$ follows from Theorem 8.2,

$$\widehat{\overline{H}} = \tilde{\alpha}\widetilde{U}\widetilde{\mathbf{V}}^*\Lambda^{1/2}, \tag{8.75}$$

where $\widetilde{U}$ and $\widetilde{\mathbf{V}}^*$ are the partial isometry and the unitary matrix respectively in the SVD of $\overline{S}\Lambda^{1/2}$. If c in (8.4) is fixed then $\tilde{\alpha} = c$, and if c is chosen to minimize the LS error then $\tilde{\alpha} = \tilde{c}$ where

$$\tilde{c} = \frac{\text{Tr}\left((\Lambda^{1/2}\overline{S}^*\overline{S}\Lambda^{1/2})^{1/2}\right)}{\text{Tr}(\Lambda)} = \frac{\text{Tr}\left((\Lambda^{1/2}\mathbf{Q}^*S^*S\mathbf{Q}\Lambda^{1/2})^{1/2}\right)}{\text{Tr}(\mathbf{R})}. \tag{8.76}$$

We can simplify the expression for $\tilde{c}$ using the equality (B.1) derived in Appendix B,

$\mathbf{Y}\mathbf{X}^{1/2}\mathbf{Y}^\dagger = (\mathbf{Y}\mathbf{X}\mathbf{Y}^\dagger)^{1/2}$ for any $\mathbf{X}$ and $\mathbf{Y}$ such that $\mathcal{R}(\mathbf{X}^{1/2}) \subseteq \mathcal{N}(\mathbf{Y})^\perp$. Then,

$$\begin{aligned}
\mathrm{Tr}\left((\Lambda^{1/2}\mathbf{Q}^*S^*S\mathbf{Q}\Lambda^{1/2})^{1/2}\right) &= \mathrm{Tr}\left(\mathbf{Q}(\Lambda^{1/2})^\dagger(\Lambda^{1/2}\mathbf{Q}^*S^*S\mathbf{Q}\Lambda^{1/2})^{1/2}\Lambda^{1/2}\mathbf{Q}^*\right) \\
&= \mathrm{Tr}\left((P_{\mathcal{N}(\mathbf{R})^\perp}S^*S\mathbf{R}P_{\mathcal{N}(\mathbf{R})^\perp})^{1/2}\right) \\
&= \mathrm{Tr}\left((S^*S\mathbf{R})^{1/2}P_{\mathcal{N}(\mathbf{R})^\perp}\right) \\
&= \mathrm{Tr}\left((S^*S\mathbf{R})^{1/2}\right), \qquad (8.77)
\end{aligned}$$

where $P_{\mathcal{N}(\mathbf{R})^\perp} = \mathbf{Q}\tilde{\mathbf{I}}_r\mathbf{Q}^*$ is an orthogonal projection onto $\mathcal{N}(\mathbf{R})^\perp = \mathcal{R}(\mathbf{R})$, and we used the fact that $P_{\mathcal{N}(\mathbf{R})^\perp} = P^\dagger_{\mathcal{N}(\mathbf{R})^\perp}$. Thus,

$$\tilde{c} = \frac{\mathrm{Tr}\left((S^*S\mathbf{R})^{1/2}\right)}{\mathrm{Tr}(\mathbf{R})}. \qquad (8.78)$$

The optimal $\widehat{H}$ that minimizes (8.3) subject to (8.4) is then given by

$$\widehat{H} = \widehat{\overline{H}}\mathbf{Q}^* = \tilde{\alpha}\widetilde{U}\widetilde{\mathbf{V}}^*\Lambda^{1/2}\mathbf{Q}^*. \qquad (8.79)$$

If $r \leq n$, then from Theorem 8.2 it follows that $\widetilde{U}$ can always be chosen so that the optimal vectors $\hat{h}_i$ lie in the space $\mathcal{U}$ spanned by the vectors s_i, and $\widehat{H} = P_{\mathcal{U}}\widehat{H}$.

Suppose in addition that $\mathbf{R}$ is such that $\mathcal{N}(S) = \mathcal{N}(\mathbf{R}) = \mathcal{R}(\mathbf{R})^\perp$. Then $\mathcal{R}(S\mathbf{Q}\Lambda^{1/2}) = \mathcal{U}$ so that we can express $P_{\mathcal{U}}$ as $P_{\mathcal{U}} = \widetilde{U}\tilde{\mathbf{I}}_n\widetilde{U}^*$, and (8.79) reduces to

$$\begin{aligned}
\widehat{H} &= \tilde{\alpha}P_{\mathcal{U}}\widetilde{U}\widetilde{\mathbf{V}}^*\Lambda^{1/2}\mathbf{Q}^* \\
&= \tilde{\alpha}\widetilde{U}\tilde{\mathbf{I}}_n\widetilde{\mathbf{V}}^*\Lambda^{1/2}\mathbf{Q}^* \\
&= \tilde{\alpha}S\mathbf{Q}\Lambda^{1/2}((\Lambda^{1/2}\mathbf{Q}^*S^*S\mathbf{Q}\Lambda^{1/2})^{1/2})^\dagger\Lambda^{1/2}\mathbf{Q}^* \\
&= \tilde{\alpha}S((\mathbf{R}S^*S)^{1/2})^\dagger\mathbf{R} \\
&= \tilde{\alpha}S\mathbf{R}((S^*S\mathbf{R})^{1/2})^\dagger. \qquad (8.80)
\end{aligned}$$

In particular, if $r = n = m$ then

$$\widehat{H} = \tilde{\alpha}S(\mathbf{R}S^*S)^{-1/2}\mathbf{R} = \tilde{\alpha}S\mathbf{R}(S^*S\mathbf{R})^{-1/2}. \qquad (8.81)$$

8.6.2 Unconstrained Least-Squares Inner Product Shaping

We now consider the LS inner product shaping problem in which the eigenvectors of $\mathbf{R}$ are specified, and the eigenvalues are chosen to minimize the LS error. We may again solve this minimization problem by performing a change of coordinates $\mathbf{Q}$ and solving an optimal orthogonalization problem in the new coordinate system. However, now the orthogonal vectors are not norm constrained. Thus our problem reduces to seeking the orthogonal vectors $\bar{h}_i$ corresponding to $\overline{H} = H\mathbf{Q}$ that minimize the LS error (8.73) subject to $\langle \bar{h}_i, \bar{h}_k \rangle = 0, i \neq k$.

This minimization problem is equivalent to the unconstrained LS orthogonalization problem discussed in Section 8.5.2. We may solve this problem using the iterative algorithm described in Section 8.5.4. The optimal LS vectors are then the vectors corresponding to $\widehat{H} = \widehat{\overline{H}}\mathbf{Q}^*$, where $\widehat{\overline{H}}$ is the set transformation corresponding to the optimal LS orthogonal vectors that minimize (8.73).

From Theorem 8.2 it follows that if the vectors $\bar{s}_i$ are GU, then $\widehat{\overline{H}}$ is equal to the optimal set transformation corresponding to the SOLSV with respect to $\bar{s}_i$. With $S = U\Sigma\mathbf{V}^*$ denoting the SVD of S, $\overline{S} = U\Sigma\mathbf{V}^*\mathbf{Q}$ so that From Corollary 5.1, the vectors $\bar{s}_i$ are GU if and only if $\mathbf{Q}^*\mathbf{V}$ is equal to a Fourier transform matrix. Then, $\widehat{\overline{H}} = \alpha U\mathbf{V}^*\mathbf{Q}$ with

$$\alpha = \frac{1}{m}\mathrm{Tr}\left((\mathbf{Q}^*S^*S\mathbf{Q})^{1/2}\right) = \frac{1}{m}\mathrm{Tr}\left((S^*S)^{1/2}\right), \tag{8.82}$$

where we used (B.3), and

$$\widehat{H} = \widehat{\overline{H}}\mathbf{Q}^* = \alpha U\mathbf{V}^*. \tag{8.83}$$

We conclude that if the vectors $\bar{s}_i$ are GU, then the closest vectors to the vectors s_i in a LS sense with Gram matrix that is diagonalized by $\mathbf{Q}$ are scaled-orthonormal, and are equal to the SOLSV with respect to the vectors s_i.

We summarize our results regarding LS inner product shaping in the following theorem:

Theorem 8.3 (Least-squares inner product shaping). *Let $\{s_i, 1 \leq i \leq m\}$ be a set of m vectors in a Hilbert space $\mathcal{H}$, that span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$. Let $\{\hat{h}_i, 1 \leq i \leq m\}$ denote the least-squares vectors that minimize the least-squares error (8.1), subject to the constraint (8.2). Let $\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^*$ be the matrix of inner products $\langle \hat{h}_i, \hat{h}_k \rangle = r_{ik}$,*

and let $\widehat{H}$ denote the set transformation corresponding to the vectors $\hat{h}_i$. Then

1. *If $\mathbf{R}$ is a specified rank-r matrix, then let $S\mathbf{Q}\Lambda^{1/2} = \widetilde{U}\Sigma\widetilde{\mathbf{V}}^*$ where S is the set transformation corresponding to the vectors s_i. Then $\widehat{H}$ can be chosen as*

$$\widehat{H} = \tilde{\alpha}\widetilde{U}\widetilde{\mathbf{V}}^*\Lambda^{1/2}\mathbf{Q}^*,$$

where

   (a) *if c in (8.2) is specified then $\tilde{\alpha} = c$;*

   (b) *if c is chosen to minimize the least-squares error then $\tilde{\alpha} = \tilde{c}$ where $\tilde{c} = Tr\left((S^*S\mathbf{R})^{1/2}\right) / Tr(\mathbf{R})$.*

   In addition,

   (a) *If $r \leq n$, then the vectors $\hat{h}_i$ can always be chosen to lie in $\mathcal{U}$;*

   (b) *If $r = n$ and $\mathcal{N}(S) = \mathcal{N}(\mathbf{R})$, then*

      i. *$\widehat{H} = \tilde{\alpha}S((\mathbf{R}S^*S)^{1/2})^{\dagger}\mathbf{R} = \tilde{\alpha}S\mathbf{R}((S^*S\mathbf{R})^{1/2})^{\dagger}$;*

      ii. *the least-squares vectors are unique.*

   (c) *If $r = n = m$, then $\widehat{H} = \tilde{\alpha}S(\mathbf{R}S^*S)^{-1/2}\mathbf{R} = \tilde{\alpha}S\mathbf{R}(S^*S\mathbf{R})^{-1/2}$.*

2. *If the eigenvalues of $\mathbf{R}$ are chosen to minimize the least-squares error, then $\widehat{H}$ can be chosen as $\widehat{H} = \widehat{\overline{H}}\mathbf{Q}^*$, where $\widehat{\overline{H}}$ is the set transformation corresponding to an orthogonal set of vectors that minimize the least-squares error defined by (8.73) with $\overline{S} = S\mathbf{Q}$, and can be computed using the iterative algorithm of Section 8.5.4.*
   Let $S = U\Sigma\mathbf{V}^*$. If $\mathbf{Q}^*\mathbf{V}$ is a Fourier transform matrix, then $\widehat{H} = \alpha U\mathbf{V}^*$ where $\alpha = (1/m)\,Tr\left((S^*S)^{1/2}\right)$, and the vectors $\hat{h}_i$ are equal to the scaled-orthonormal least-squares vectors with optimal scaling, with respect to the vectors s_i.

8.6.3 Weighted Least-Squares Inner Product Shaping

We may also consider a weighted LS inner product shaping problem in which we seek the vectors h_i corresponding to H to minimize the weighted squared error $\varepsilon_{\mathrm{LS}}^{w}$ given by (8.34),

subject to (8.4) where $\mathbf{R}$ is fixed. With $S_w = S\mathbf{A}$, we can express $\varepsilon_{\text{LS}}^w$ as

$$\begin{aligned}\varepsilon_{\text{LS}}^w &= \text{Tr}\left((S-H)^*(S-H)\mathbf{A}\right) \\ &= \text{Tr}\left((S_w-H)^*(S_w-H)\right) + \text{Tr}\left((\mathbf{A}-\mathbf{I}_m)H^*H\right) + \text{Tr}\left(\mathbf{A}(\mathbf{I}_m-\mathbf{A})S^*S\right) \\ &= \text{Tr}\left((S_w-H)^*(S_w-H)\right) + c^2\text{Tr}\left((\mathbf{A}-\mathbf{I}_m)\mathbf{R}\right) + K, \end{aligned} \tag{8.84}$$

where K is independent of the choice of H.

If c is fixed, then minimizing $\varepsilon_{\text{LS}}^w$ is equivalent to minimizing $\varepsilon'^w_{\text{LS}}$ where

$$\varepsilon'^w_{\text{LS}} = \text{Tr}\left((S_w-H)^*(S_w-H)\right). \tag{8.85}$$

Furthermore, this minimization problem is equivalent to the general LS inner product shaping problem, if we substitute S_w for S. The optimal set transformation $\widehat{H}_w$ then follows from Theorem 8.3,

$$\widehat{H}_w = c\widetilde{U}_w\widetilde{\mathbf{V}}_w^*\Lambda^{1/2}\mathbf{Q}^*, \tag{8.86}$$

where $\widetilde{U}_w$ and $\widetilde{\mathbf{V}}_w^*$ are the partial isometry and the unitary matrix respectively in the SVD of $S\mathbf{A}\mathbf{Q}\Lambda^{1/2}$. If the vectors s_i are linearly independent and $\mathbf{A}$ and $\mathbf{R}$ are invertible, then

$$\widehat{H}_w = cS\mathbf{A}(\mathbf{R}\mathbf{A}S^*S\mathbf{A})^{-1/2}\mathbf{R} = cS\mathbf{A}\mathbf{R}(\mathbf{A}S^*S\mathbf{A}\mathbf{R})^{-1/2}. \tag{8.87}$$

If we further wish to minimize the weighted LS error $\varepsilon_{\text{LS}}^w$ with respect to c, then substituting $\widehat{H}_w$ back into (8.84) and minimizing with respect to c, the optimal value of c is given by

$$\tilde{c}_w = \frac{\Re\{\text{Tr}(\widehat{H}_w^*S\mathbf{A})\}}{\text{Tr}(\mathbf{A}\mathbf{R})} = \frac{\text{Tr}\left((\Lambda^{1/2}\mathbf{Q}^*\mathbf{A}S^*S\mathbf{A}\mathbf{Q}\Lambda^{1/2})^{1/2}\right)}{\text{Tr}(\mathbf{A}\mathbf{R})} = \frac{\text{Tr}\left((\mathbf{A}S^*S\mathbf{A}\mathbf{R})^{1/2}\right)}{\text{Tr}(\mathbf{A}\mathbf{R})}. \tag{8.88}$$

8.7 Summary

In this chapter we developed new methods that construct a set of vectors with specified inner product structure, that are closest in a LS sense to a given set of vectors. These methods are based on the LS measurement, which we developed in the context of quantum detection. Using these methods we can now construct optimal QSP measurements with

measurement vectors that have a specified inner product structure, and are closest in a LS sense to some desired set of vectors. We can also use these methods to construct optimal linear algorithms subject to inner product constraints, by viewing linear processing of a signal as processing with a set of measurement vectors, and then imposing inner product constraints on these vectors.

In the remainder of the thesis we demonstrate that LS inner product shaping, inspired by the quantum detection problem, is a very versatile method with applications spanning many different areas. We have already seen in Chapter 3 that by applying LS orthogonalization to a detection problem in quantum mechanics, we can derive a solution to a previously unsolved problem in this field, for a very important special case that often arises in practice. In the ensuing chapters we focus on more subtle applications of LS inner product shaping to problems with no inherent inner product constraints. By casting various signal processing algorithms as a QSP measurement or as processing with a set of measurement vectors, and then imposing an inner product constraint on the corresponding measurement vectors, we can derive a variety of effective new processing techniques that often exhibit improved performance over existing methods.

As an example of the type of procedure we may follow in using the concept of optimal QSP measurements to derive new processing methods, in Chapter 9 we consider a generic detection problem where one of a set of signals is transmitted over a noisy channel. By describing the conventional matched filter (MF) detector as a QSP measurement, and imposing inner product constraints on the MF measurement vectors, we derive a new class of detectors. We then demonstrate through simulation that when the additive noise is non-Gaussian these detectors can significantly increase the probability of correct detection over the MF receiver, with only a minor impact in performance when the noise is Gaussian.

In Chapter 10 we show that the concept of LS inner product shaping leads to a new viewpoint towards whitening and other covariance shaping problems which arise frequently in signal processing applications. Specifically, we derive a stochastic analogue of the LS inner product shaping problem that takes on the form of an MMSE covariance shaping problem, in which the covariance shaping transformation is designed to minimize the MSE between its input and output.

Drawing from LS inner product shaping, we can develop new classes of linear algorithms that result from imposing a deterministic or stochastic inner product constraint on the algo-

rithm *i.e.,* a covariance constraint, and then using the concept of LS inner product shaping to derive optimal linear algorithms subject to this constraint. In Chapter 11 we use this basic idea to derive a new linear estimator for estimating a set of unknown deterministic parameters observed through a known linear transformation and corrupted by additive noise. This new estimator is defined as the covariance shaping LS (CSLS) estimator. Analysis of the MSE of the CSLS estimator demonstrates that over a wide range of SNR, the CSLS estimator results in a lower MSE than the traditional LS estimator, for all values of the unknown parameters.

Finally, in Chapter 12 we consider an application of the CSLS estimator to the problem of suppressing interference in multiuser wireless communication systems. Specifically, we develop a new linear multiuser receiver for synchronous code-division multiple-access (CDMA) systems, in which different users transmit information over a joint channel by modulating distinct signature vectors. This receiver can be viewed as an MF receiver matched to a set of vectors with a specified inner product structure, that are closest in a LS sense to the users' signature vectors. Alternatively, we may view the receiver as a decorrelator receiver [50] followed by an optimal covariance shaping transformation, that optimally shapes the covariance of the decorrelator output prior to detection. We demonstrate that this modified receiver can lead to improved performance over the decorrelator and MF receiver, and can approach the performance of the linear MMSE receiver, which is the optimal linear receiver that assumes knowledge of the channel parameters and maximizes the output signal-to-interference+noise ratio, over a wide range of channel parameters without requiring knowledge of these parameters.

These applications demonstrate that drawing from the ideas and constraints of the quantum detection problem outlined in Chapter 3, and imposing inner product and covariance constraints in combination with optimal inner product and covariance shaping methods, can be advantageous in a variety of problems and can lead to a multitude of new, potentially effective processing techniques.

In our closing remarks, we mention that there would appear to be many other potential applications of LS inner product shaping beyond those explored in the thesis. An interesting and potentially fruitful direction for future research is to identify and explore such new applications.

Chapter 9

Inner Product Shaping Matched Filter Detection

In this chapter we consider an application of optimal QSP measurements and LS inner product shaping to the generic problem of detecting in the presence of additive noise, which one from a set of known signals has been received.

In place of the classical MF receiver we propose a class of modified receivers that are derived by formulating the MF receiver as a QSP measurement, and then imposing an inner product constraint on the corresponding measurement vectors. Specifically, as outlined in Chapter 4, the MF receiver for detecting the transmitted signal can be implemented as a ROM with measurement vectors that are equal to the transmitted signals. Building upon the notion of optimal QSP measurements, we develop a class of modified receivers by imposing an inner product constraint on the measurement vectors of the ROM describing the MF receiver, and then designing an optimal QSP measurement subject to this constraint. The resulting receiver consists of a bank of correlators with correlating signals that are matched to a set of signals with a specified inner product structure, and are closest in a LS sense to the transmitted signals.

Alternatively, we show that the modified receiver can be implemented as an MF demodulator followed by an optimal covariance shaping transformation, that optimally shapes the correlation of the outputs of the MF prior to detection. This equivalent representation leads to the concept of minimum MSE (MMSE) covariance shaping, which we consider in its most general form in Chapter 10.

In our development we focus primarily on the case in which the correlating signals are chosen to be orthonormal or to form a normalized tight frame, so that the outputs of the receiver are uncorrelated on a space formed by the transmitted signals. However, the ideas we develop in this chapter can be readily applied to other forms of inner product constraints on the correlating signals.

9.1 Detection Problem

9.1.1 Problem Formulation

A generic problem which has been studied extensively is that of detecting which one of a set of known signals is received over an additive noise channel. This problem arises in a wide variety of contexts including target classification, signature analysis, and other multi-signature problems.

When the additive noise is white and Gaussian, it is well known (see *e.g.,* [11, 32]) that the receiver which maximizes the probability of correct detection consists of an MF demodulator comprised of a bank of correlators with correlating signals equal to the transmitted set, followed by a detector which chooses as the detected signal the one for which the output of the correlator is maximum.

If the noise is not Gaussian, then the MF receiver does not necessarily maximize the probability of correct detection. However, it is still used as the receiver of choice in many applications since the optimal detector for non-Gaussian noise is typically nonlinear (see *e.g.,* [33] and references therein), and depends on the noise distribution which may not be known. One justification often given for its use is that if a signal is corrupted by Gaussian or non-Gaussian additive white noise, then the filter matched to that signal maximizes the output signal-to-noise ratio (SNR) [11].

In this chapter we develop modifications of the MF receiver by imposing inner product constraints on the measurement vectors of the ROM describing the MF receiver. If we constrain these vectors to have a specified inner product structure, then in general they can no longer be chosen to be equal to the transmitted signals. We therefore choose the measurement vectors to have the required inner products, and to "best" represent the transmitted signals in some sense. The resulting receiver again consists of a bank of correlators followed by the same detector used in the MF receiver, where now the correlating signals are matched

to a set of optimal signals with the desired inner products. This receiver depends only on the transmitted signals, so that it does not require knowledge of the noise distribution or the channel SNR. The simulations presented in Section 9.6 strongly suggest that when the additive noise is non-Gaussian this modified receiver can significantly increase the probability of correct detection over the MF receiver particularly when the probability of correct detection with the MF is marginal. When the additive noise is Gaussian, the reduction in performance over the MF receiver is minor.

9.1.2 Receiver Design

Suppose that one of m signals $\{s_i(t), 1 \leq i \leq m\}$ is received over an additive noise channel with equal probability, where the signals lie in a real Hilbert space $\mathcal{H}$ with inner product $\langle x(t), y(t) \rangle = \int_{t=-\infty}^{\infty} x(t)y(t)dt$, and span an n-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$. We assume that the signals are normalized[1] so that $\int_{t=-\infty}^{\infty} s_i^2(t)dt = 1$ for all i. The received signal $r(t)$ is also assumed to be in $\mathcal{H}$, and is modeled as $r(t) = s_i(t) + n(t)$ for one value i, where $n(t)$ is a stationary white noise process with zero mean and spectral density σ^2, and with otherwise unknown distribution.

The receiver we design consists of the correlation demodulator depicted in Fig. 9-1, that cross-correlates the received signal $r(t)$ with each of the m signals $\{q_i(t) \in \mathcal{U}, 1 \leq i \leq m\}$ so that $a_i = \langle q_i(t), r(t) \rangle$, where the signals $\{q_i(t)\}$ are to be determined. The declared detected signal is $s_i(t)$ where $i = \arg\max a_i$. (We can equivalently obtain a_i by filtering $r(t)$ using a filter with impulse response given by $q_i(-t)$, and sampling the output at $t = 0$.) The difference between the modified receivers and the MF receiver lies in the choice of the signals $q_i(t)$.

If we choose the signals $q_i(t) = s_i(t)$ in Fig. 9-1, then the resulting demodulator is equivalent to the MF demodulator [11]. If the noise is not Gaussian, then the MF receiver does not necessarily minimize the probability of detection error. However, it is still used as the receiver of choice in many applications since the optimal receiver for non-Gaussian noise is typically nonlinear, and requires knowledge of the noise distribution.

For a correlation demodulator in the form of Fig. 9-1, we would like to choose the

[1]The normalization assumption as well as the assumption that the signals are transmitted with equal probability is for notational convenience only. As we discuss in Section 9.8, the results readily extend to the more general case of unequal norms and unequal probabilities

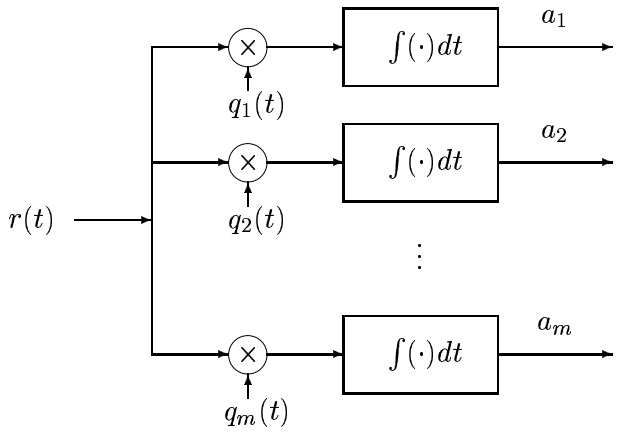


Figure 9-1: Correlation demodulator.

signals $q_i(t)$ so that when the noise is non-Gaussian the resulting detector leads to improved performance over MF detection. Drawing from the quantum detection problem, we propose imposing an inner product constraint on the signals $q_i(t)$, which as we show is equivalent to imposing a constraint on the correlation between the demodulator outputs. Building upon our results regarding optimal QSP measurements (Chapter 8), we then develop a new class of correlation receivers that, like the MF, depend only on the transmitted signals, and that can lead to improved performance over the MF for some classes of non-Gaussian noise, with essentially negligible loss of performance for Gaussian noise.

In the system of Fig. 9-1, the correlation between the outputs a_i of the correlation demodulator is proportional to the inner products between the signals $q_i(t)$:

$$\mathrm{cov}(a_i, a_k) \quad = \quad E\left(\langle q_i(t), n(t)\rangle\langle n(t), q_k(t)\rangle\right) = \sigma^2\langle q_i(t), q_k(t)\rangle. \tag{9.1}$$

In our modification of the MF demodulator we propose shaping the correlation of the outputs prior to detection. Thus, we propose choosing the signals $q_i(t)$ to have a specified inner product structure, so that the outputs a_i have the desired correlation. In the ensuing sections we consider the special case in which we choose the signals $q_i(t)$ so that the outputs a_i are uncorrelated. In Section 9.7 we briefly consider the more general case in which the signals $q_i(t)$ are chosen to have an arbitrary inner product structure.

The subsequent development considers separately the case in which the transmitted signals are linearly independent and the case in which they are linearly dependent.

If the signals $s_i(t)$ are linearly independent, then we may decorrelate the outputs a_i by choosing the signals $q_i(t)$ to be orthonormal. The resulting demodulator is referred to as the orthogonal MF (OMF) demodulator, and the overall detector is referred to as the OMF detector.

If the signals $s_i(t)$ are linearly dependent, so that they span an n-dimensional subspace $\mathcal{U}$, then there are at most n orthonormal signals in $\mathcal{U}$, and we cannot choose the correlating signals to whiten the outputs a_i in the conventional sense. Instead we choose the correlating signals as projections of a set of orthonormal signals in a larger space containing $\mathcal{U}$, *i.e.,* we chose the correlating signals to form a normalized tight frame for $\mathcal{U}$. As we show, the outputs a_i are then uncorrelated on a space formed by the transmitted signals. The resulting demodulator is referred to as the projected orthogonal MF (POMF) demodulator, and the overall detector is referred to as the POMF detector.

Alternative derivations of the OMF and POMF receivers are considered in [36, 37].

9.2 The Orthogonal Matched Filter Demodulator

9.2.1 Design Criterion

We first consider the case in which the signals $\{s_i(t), 1 \leq i \leq m\}$ are linearly independent.

From (9.1) it follows that if the correlating signals $q_i(t)$ in Fig. 9-1 are not orthogonal, then the outputs of the demodulator are correlated. To improve the performance of the detector, we propose eliminating this common (linear) information prior to detection, so that the detection is based on uncorrelated outputs. We therefore consider choosing the correlating signals denoted by $q_i(t) = g_i(t)$ to be orthonormal. Then the outputs of the correlation demodulator are uncorrelated and have equal variance.

There are many ways of choosing a set of orthonormal signals $g_i(t)$. In our modification of the MF receiver we would like to choose these signals so that when the noise is non-Gaussian the resulting detector leads to improved performance over MF detection. In addition we want our design criterion to depend only on the transmitted signals so that the modified receiver does not depend on the noise distribution or the channel SNR.

The MF demodulator has the well known property that if the transmitted signal is $s_i(t)$,

then choosing $q_i(t) = s_i(t)$ in Fig. 9-1 maximizes the SNR of a_i, denoted SNR_i. Indeed,

$$\mathrm{SNR}_i = \frac{\langle q_i(t), s_i(t)\rangle^2}{E\left(\langle q_i(t), n(t)\rangle^2\right)} = \frac{1}{\sigma^2}\langle q_i(t), s_i(t)\rangle^2 \leq \frac{1}{\sigma^2}\langle q_i(t), q_i(t)\rangle^2\langle s_i(t), s_i(t)\rangle^2 = \frac{1}{\sigma^2}, \quad (9.2)$$

with equality in (9.2) if and only if $q_i(t)$ is proportional to $s_i(t)$. The choice $q_i(t) = s_i(t)$ also of course maximizes the total SNR defined by $\mathrm{SNR}_T = \sum_{i=1}^m \mathrm{SNR}_i = \frac{1}{\sigma^2}\sum_{i=1}^m |\langle q_i(t), s_i(t)\rangle|^2$, since the individual terms are maximized by this choice.

To derive a modification of the MF receiver we may consider choosing a set of orthonormal signals $g_i(t)$ to maximize the output SNR. If we constrain the signals $g_i(t)$ to be orthonormal, then in general we cannot maximize SNR_i individually subject to this constraint. However, we may seek a set of orthonormal signals $g_i(t)$ to maximize the total output SNR,

$$\mathrm{SNR}_T = \frac{1}{\sigma^2}\sum_{i=1}^m |\langle g_i(t), s_i(t)\rangle|^2. \quad (9.3)$$

This problem is equivalent to the quantum detection problem discussed in Section 3.4. Specifically, comparing (9.3) with (3.13) we see that choosing a set of orthonormal quantum measurement vectors to maximize the probability of correct detection in a quantum detection problem, is equivalent to choosing a set of orthonormal correlating signals to maximize SNR_T. We may therefore interpret the design problem of (9.3) as a quantum detection problem, and then apply results derived in that context. In particular, from our discussion in Section 3.4 it follows that for arbitrary signals $s_i(t)$, there is no known closed-form analytical expression for the orthonormal signals $g_i(t)$ maximizing SNR_T.

Therefore, in analogy to the quantum detection problem, we propose taking an alternative approach of choosing a different optimality criterion, namely a squared-error criterion, and seeking the signals $g_i(t)$ that minimize this criterion. Thus, in the OMF demodulator the signals $g_i(t)$ are chosen to be orthonormal, and to minimize the LS error

$$\varepsilon_{\mathrm{LS}} = \sum_{i=1}^m \langle s_i(t) - g_i(t), s_i(t) - g_i(t)\rangle. \quad (9.4)$$

This problem is equivalent to the LS orthonormalization problem discussed in Section 8.2, so that the minimizing signals $\hat{g}_i(t)$, which we refer to as the OMF signals, follow immediately from Theorem 8.1. With S and $\widehat{G}$ denoting the set transformations corre-

sponding to the signals $s_i(t)$ and $\hat{g}_i(t)$ respectively,

$$\widehat{G} = S(S^*S)^{-1/2}. \tag{9.5}$$

Thus, the OMF demodulator consists of a correlation demodulator with orthonormal signals $\hat{g}_i(t)$ defined by (9.5), that are closest in the LS sense to the signals $s_i(t)$.

9.2.2 OMF Signals

To implement the OMF demodulator, we may find it more convenient to reformulate the OMF signals in terms of their coefficients in a basis expansion for the m-dimensional space $\mathcal{U}$ spanned by the signals $s_i(t)$. These coefficients can be viewed as vectors in $\mathbb{C}^m$.

Let X denote a set transformation corresponding to a set of m signals that form an orthonormal basis for $\mathcal{U}$. Then $s_i(t) = X\mathbf{s}_i$ for some $\mathbf{s}_i \in \mathbb{C}^m$, and $S = X\mathbf{S}$ where $\mathbf{S}$ is the $m \times m$ matrix of columns $\mathbf{s}_i$. We may then express $\widehat{G}$ of (9.5) in terms of X and $\mathbf{S}$ as

$$\widehat{G} = S(S^*S)^{-1/2} = X\mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}. \tag{9.6}$$

Thus, $\hat{g}_i(t) = X\hat{\mathbf{g}}_i$ where $\hat{\mathbf{g}}_i$ is the ith column of the $m \times m$ matrix $\widehat{\mathbf{G}}$, and

$$\widehat{\mathbf{G}} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}. \tag{9.7}$$

Since (9.7) has the same form as (9.6), we conclude that the vectors $\{\hat{\mathbf{g}}_i, 1 \leq i \leq m\}$ are the closest orthonormal vectors to the vectors $\{\mathbf{s}_i, 1 \leq i \leq m\}$, in the LS sense. From the discussion in Section 8.2.2 it then follows that $\widehat{\mathbf{G}}$ is just the partial isometry in the polar decomposition (PD) of $\mathbf{S}$, and can be expressed in terms of the elements of the SVD $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$ as $\widehat{\mathbf{G}} = \mathbf{U}\mathbf{V}^*$.

Exploiting the relationship between the OMF signals and the PD, these signals can be computed very efficiently by use of the many known efficient algorithms for computing the PD (see *e.g.,* [93, 143, 141, 144]).

9.3 The Projected Orthogonal Matched Filter Demodulator

Suppose now that the transmitted signals $\{s_i(t), 1 \leq i \leq m\}$ are linearly dependent, and span an n-dimensional subspace $\mathcal{U} \subset \mathcal{H}$, where $n < m$. As in the case of linearly independent signals, we can choose the signals $q_i(t) = g_i(t)$ in Fig. 9-1 to be orthonormal, and to minimize the LS error (9.4). Since $\langle g_i(t), g_i(t) \rangle = 1$ for any choice of signals $g_i(t)$, minimizing the LS error is equivalent to maximizing

$$\sum_{i=1}^{m} \langle g_i(t), s_i(t) \rangle = \sum_{i=1}^{m} \langle g_i^{\mathcal{U}}(t), s_i(t) \rangle, \tag{9.8}$$

where the signals $g_i^{\mathcal{U}}(t) = P_{\mathcal{U}} g_i(t)$ form a normalized tight frame for $\mathcal{U}$. Since, as we showed in (8.23), for any normalized tight frame for $\mathcal{U}$, $\sum_{i=1}^{m} \langle g_i^{\mathcal{U}}(t), g_i^{\mathcal{U}}(t) \rangle = n$, maximizing (9.8) is equivalent to minimizing

$$\sum_{i=1}^{m} \langle g_i^{\mathcal{U}}(t) - s_i(t), g_i^{\mathcal{U}}(t) - s_i(t) \rangle. \tag{9.9}$$

Thus, when the signals $s_i(t)$ are linearly dependent, choosing a set of orthonormal signals to maximize (9.8) is equivalent to choosing a normalized tight frame for $\mathcal{U}$ to minimize the LS error (9.9). Furthermore, if the transmitted signal is $s_i(t)$, then the ith output of the correlation demodulator with signals $g_i(t)$ is

$$a_i = \langle g_i(t), r(t) \rangle = \langle g_i^{\mathcal{U}}(t), s_i(t) + n(t) \rangle + \langle g_i^{\mathcal{U}^\perp}(t), n(t) \rangle = r_i + n_i, \tag{9.10}$$

where $g_i^{\mathcal{U}^\perp}(t) = P_{\mathcal{U}^\perp} g_i(t)$, $n_i = \langle g_i^{\mathcal{U}^\perp}(t), n(t) \rangle$, and $r_i = \langle g_i^{\mathcal{U}}(t), s_i(t) + n(t) \rangle$. Since r_i and n_i are uncorrelated, n_i does not contain any linear information that is relevant to the detection of $s_i(t)$. Therefore in the case of linearly dependent signals $s_i(t)$, we propose choosing the signals $q_i(t)$ in Fig. 9-1 to be a normalized tight frame for $\mathcal{U}$, which we denote by $q_i(t) = f_i(t)$, that minimizes the LS error.

Thus we seek the signals $\{f_i(t), 1 \leq i \leq m\}$ corresponding to F to minimize

$$\varepsilon_{\text{LS}} = \sum_{i=1}^{m} \langle s_i(t) - f_i(t), s_i(t) - f_i(t) \rangle, \tag{9.11}$$

subject to the constraint

$$FF^* = P_{\mathcal{U}}. \tag{9.12}$$

This problem is equivalent to the LS tight frame problem discussed in Section 8.2.1, so that the minimizing signals $\hat{f}_i(t)$, which we refer to as the POMF signals, follow immediately from Theorem 8.1. With $\widehat{F}$ denoting the set transformations corresponding to the signals $\hat{f}_i(t)$,

$$\widehat{F} = S((S^*S)^{1/2})^\dagger. \tag{9.13}$$

Thus, the POMF demodulator consists of a correlation demodulator with signals $\hat{f}_i(t)$ defined by (9.5) that form a tight frame for $\mathcal{U}$, and are closest in the LS sense to the signals $s_i(t)$.

9.3.1 POMF Signals

As with the OMF demodulator, to implement the POMF demodulator we may find it more convenient to reformulate the POMF signals in terms of their coefficients in a basis expansion for the n-dimensional space $\mathcal{U}$, which can be viewed as vectors in $\mathbb{C}^n$.

Let X denote a set transformation corresponding to a set of n signals that form an orthonormal basis for $\mathcal{U}$. Then $s_i(t) = X\mathbf{s}_i$ for some $\mathbf{s}_i \in \mathbb{C}^n$, and $S = X\mathbf{S}$ where $\mathbf{S}$ is the $n \times m$ matrix of columns $\mathbf{s}_i$. We can now express $\widehat{F}$ in terms of X and $\mathbf{S}$ as

$$\widehat{F} = S((S^*S)^{1/2})^\dagger = X\mathbf{S}((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger. \tag{9.14}$$

Thus, $\hat{f}_i(t) = X\hat{\mathbf{f}}_i$ where $\hat{\mathbf{f}}_i$ is the ith column of the $n \times m$ matrix $\widehat{\mathbf{F}}$, and

$$\widehat{\mathbf{F}} = \mathbf{S}((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger. \tag{9.15}$$

Since (9.15) has the same form as (9.14), we conclude that the vectors $\{\hat{\mathbf{f}}_i, 1 \le i \le m\}$ form the closest normalized tight frame to the vectors $\{\mathbf{s}_i, 1 \le i \le m\}$, in the LS sense. From the discussion in Section 8.2.2, $\widehat{\mathbf{F}}$ is just the projection onto the space spanned by the vectors $\mathbf{s}_i$ of the partial isometry in a PD of $\mathbf{S}$, and can be expressed in terms of the elements of

the SVD $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$ as $\widehat{\mathbf{F}} = \mathbf{U}\mathbf{I}_{nm}\mathbf{V}^*$, where

$$\mathbf{I}_{nm} = \left[\begin{array}{cc} \mathbf{I}_n & \mathbf{0}_{n\times(m-n)} \end{array} \right]. \tag{9.16}$$

9.4 Matched Filter Representation of the OMF and POMF Demodulators

In many practical receivers the MF demodulator serves as a front-end whose objective is to obtain a vector representation of the received signal. Thus, in many applications we do not have control over the correlating signals of the correlation demodulator, but rather we are given the MF outputs. In this section we show that the OMF and POMF demodulators can be implemented by processing the MF outputs. Specifically, we derive an equivalent implementation of the OMF and POMF demodulators that consists of an MF demodulator followed by an optimal whitening transformation on a space formed by the transmitted signals, that optimally decorrelates the outputs of the MF prior to detection.

9.4.1 Matched Filter Representation of a Correlation Demodulator

We first show that any correlation demodulator of the from of Fig. 9-1 with correlating signals $q_i(t) \in \mathcal{U}$, is equivalent to an MF demodulator followed by a linear transformation $\mathbf{T}$ on the MF outputs, as depicted in Fig. 9-2.

Since the signals $s_i(t)$ span $\mathcal{U}$, any signal $q_i(t) \in \mathcal{U}$ can be expressed as a linear combination of the signals $s_i(t)$, so that the set transformation Q corresponding to the signals $q_i(t)$ may be expressed as $Q = S\mathbf{T}^*$, where $\mathbf{T}$ is an $m \times m$ coefficient matrix. The vector output $\mathbf{a}$ of Fig. 9-1, with components a_i, can then be written as $\mathbf{a} = Q^* r(t) = \mathbf{T}S^* r(t) = \mathbf{T}\tilde{\mathbf{a}}$, where $\tilde{\mathbf{a}} = S^* r(t)$ is the vector output of the MF demodulator. Thus, the correlation demodulator with correlating signals $q_i(t) \in \mathcal{U}$ is equivalent to an MF demodulator followed by a linear transformation $\mathbf{T}$ defined by $Q = S\mathbf{T}^*$.

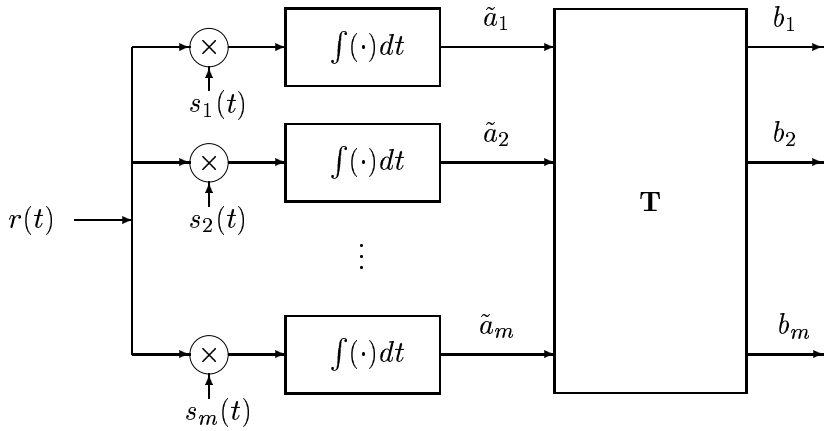


Figure 9-2: Equivalent representation of a correlation demodulator. The linear transformation $\mathbf{T}$ is a function of the transmitted signals $s_i(t)$ and the correlating signals $q_i(t)$ of Fig. 9-1.

9.4.2 Matched Filter Representation of the OMF Demodulator

If the signals $q_i(t) = g_i(t)$ in Fig. 9-1 are orthonormal, then $G^*G = \mathbf{I}_m$ and the corresponding transformation $\mathbf{T}$ in Fig. 9-2 defined by $G = S\mathbf{T}^*$ must satisfy

$$\mathbf{T}S^*S\mathbf{T}^* = \mathbf{I}_m. \tag{9.17}$$

We now try to gain insight into the condition (9.17). From (9.1), the covariance between the outputs $\tilde{a}_i$ of the MF demodulator is given by $\text{cov}(\tilde{a}_i, \tilde{a}_k) = \sigma^2\langle s_i(t), s_k(t)\rangle$, so that the covariance matrix of $\tilde{\mathbf{a}}$, denoted $\mathbf{C}_a$, is $\mathbf{C}_a = \sigma^2 S^*S$. The covariance of $\mathbf{b} = \mathbf{T}\tilde{\mathbf{a}}$, denoted $\mathbf{C}_b$, is then equal to $\mathbf{C}_b = \sigma^2\mathbf{T}S^*S\mathbf{T}^*$, which using (9.17) reduces to $\mathbf{C}_b = \sigma^2\mathbf{I}_m$. Thus, $\mathbf{T}$ is a whitening transformation[2] that whitens the output $\tilde{\mathbf{a}}$ of the MF demodulator.

In summary, a correlation demodulator with orthonormal signals $g_i(t) \in \mathcal{U}$ is equivalent to an MF demodulator followed by a whitening transformation, which we denote by $\mathbf{W}$, that is defined by $G = S\mathbf{W}^*$.

Since every set of orthonormal correlating signals defines a whitening transformation, and the OMF signals are optimal in some sense, we expect the corresponding whitening transformation to also have some form of optimality. Indeed, we now show that the whitening transformation in the OMF demodulator minimizes the MSE between the input $\tilde{\mathbf{a}}$ and

[2] We define a random vector $\mathbf{a} \in \mathbb{C}^m$ to be white if the covariance of $\mathbf{a}$ is proportional to $\mathbf{I}_m$.

the output $\mathbf{b}$.

Let $g_i(t)$ denote an arbitrary set of orthonormal signals with set transformation G expressible as $G = S\mathbf{W}^*$ for some whitening transformation $\mathbf{W}$. Then

$$\begin{aligned}\sigma^2\langle s_i(t) - g_i(t), s_i(t) - g_i(t)\rangle &= E\left(\langle s_i(t) - g_i(t), n(t)\rangle\langle n(t), s_i(t) - g_i(t)\rangle\right) \\ &= E(\tilde{a}_i' - b_i')^2, \end{aligned} \tag{9.18}$$

where $\tilde{a}_i' = \tilde{a}_i - E(\tilde{a}_i), \tilde{a}_i = \langle s_i(t), r(t)\rangle$ is the ith output of the MF demodulator, and $b_i' = b_i - E(b_i)$ where $b_i = \langle g_i(t), r(t)\rangle$ is the ith component of $\mathbf{b} = \mathbf{W}\tilde{\mathbf{a}}$. Thus, seeking a set of orthonormal signals $g_i(t)$ that minimize the LS error (9.4) is equivalent to seeking the whitening transformation $\widehat{\mathbf{W}}$ that minimizes the total MSE given by

$$\varepsilon_{\text{MSE}} = \sum_{i=1}^{m} \text{var}(\tilde{a}_i - b_i) = \sum_{i=1}^{m} E\left((\tilde{a}_i' - b_i')^2\right), \tag{9.19}$$

where b_i is the ith component of the output $\mathbf{b} = \mathbf{W}\tilde{\mathbf{a}}$ of the whitening transformation. That is, from all possible whitening transformations we seek the one that results in a white vector $\mathbf{b}$ that is as close as possible in MSE to the output $\tilde{\mathbf{a}}$ of the MF demodulator. We refer to such whitening as MMSE whitening.

We may therefore interpret the OMF demodulator as an MF demodulator followed by an MMSE whitening transformation that optimally whitens the outputs of the MF prior to detection. Since the OMF signals are the signals corresponding to $\widehat{G} = S\widehat{\mathbf{W}}^*$, from (9.5) it follows that $\widehat{\mathbf{W}} = (S^*S)^{-1/2}$. The general MMSE whitening problem is considered in the next chapter. As we expect, applying the general form of the solution (see Section 10.2.1) to our problem results in the same optimal transformation $\widehat{\mathbf{W}}$.

9.4.3 Matched Filter Representation of the POMF Demodulator

If the signals $q_i(t) = f_i(t)$ form a normalized tight frame for $\mathcal{U}$, then $F^*F = P_{\mathcal{V}}$ where $\mathcal{V} = \mathcal{R}(F^*) = \mathcal{N}(F)^{\perp}$. Thus, in this case $\mathbf{T}$ must satisfy

$$\mathbf{T}S^*S\mathbf{T}^* = P_{\mathcal{V}}, \tag{9.20}$$

and the covariance of $\mathbf{b} = \mathbf{T}\tilde{\mathbf{a}}$ is

$$\mathbf{C}_b = \sigma^2 \mathbf{T} S^* S \mathbf{T}^* = \sigma^2 P_{\mathcal{V}}. \tag{9.21}$$

We therefore conclude that seeking a normalized tight frame for $\mathcal{U}$ that minimizes the LS error defined by (9.4) is equivalent to seeking the optimal transformation $\mathbf{T}$ that satisfies (9.20) and minimizes the MSE given by (9.19).

In the next chapter we show that if a random vector $\mathbf{b}$ has covariance given by (9.21), then $\mathbf{b}$ lies in $\mathcal{V}$ with probability one (w.p. 1), and its representation in terms of any orthonormal basis for $\mathcal{V}$ is white. We then say that $\mathbf{b}$ is white on $\mathcal{V}$. Thus, $\mathbf{T} = \mathbf{W}_s$ is a *subspace whitening transformation* that whitens the vector $\tilde{\mathbf{a}}$ with covariance $\mathbf{C}_a$ on $\mathcal{V} = \mathcal{R}(\mathbf{C}_a)$ (see Section 10.2.2 and [40] for a detailed discussion on subspace whitening). So, a correlation demodulator with signals $f_i(t) \in \mathcal{U}$ that from a normalized tight frame for $\mathcal{U}$ is equivalent to an MF demodulator followed by a subspace whitening transformation, defined by $F = S\mathbf{W}_s^*$.

In a manner analogous to the OMF demodulator, we may show that the POMF demodulator is equivalent to a MF demodulator followed by an MMSE subspace whitening transformation $\widehat{\mathbf{W}}_s$ that minimizes the MSE between $\tilde{\mathbf{a}}$ and $\mathbf{b}$. The POMF signals then correspond to $\widehat{F} = S\widehat{\mathbf{W}}_s$, so that from (9.13) it follows that $\widehat{\mathbf{W}}_s = ((S^*S)^{1/2})^\dagger$. The general MMSE subspace whitening problem is considered in the next chapter. As we expect, applying the general form of the solution (see Section 10.2.2) to our problem results in the same optimal transformation $\widehat{\mathbf{W}}_s$.

The alternative representations of the OMF and POMF demodulators developed in this section, lead to the new broadly applicable concept of MMSE whitening and subspace whitening, which can be viewed as a stochastic analogue of the LS orthonormalization problem. We explore the MMSE whitening problem, together with the more general MMSE covariance shaping problem, in the next chapter.

9.5 Summary of the OMF and POMF Demodulators

We summarize our results regarding the OMF and POMF demodulators in the following theorems:

Theorem 9.1 (Orthogonal matched filter (OMF) demodulator). *Let* $\{s_i(t), 1 \leq i \leq m\}$ *denote a set of* m *transmitted signals in a Hilbert space* $\mathcal{H}$ *that span an* m*-dimensional subspace* $\mathcal{U} \subseteq \mathcal{H}$*. Let* $\{\hat{g}_i(t), 1 \leq i \leq m\}$ *denote the OMF signals that are the correlating signals of the OMF demodulator. Let* S *and* $\widehat{G}$ *denote the set transformations corresponding to the signals* $s_i(t)$ *and* $\hat{g}_i(t)$*, respectively. Then*

$$\widehat{G} = S(S^*S)^{-1/2}.$$

Let X *denote a set transformation corresponding to an orthonormal basis for* $\mathcal{U}$*, let* $s_i(t) = X\mathbf{s}_i$*, and let* $\mathbf{S}$ *be the matrix of columns* $\mathbf{s}_i$ *with SVD* $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$*. Then* $\hat{g}_i(t) = X\hat{\mathbf{g}}_i$ *where the vectors* $\hat{\mathbf{g}}_i$ *are the columns of* $\widehat{\mathbf{G}}$*, and*

$$\widehat{\mathbf{G}} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2} = \mathbf{U}\mathbf{V}^*.$$

In addition,

1. *The OMF demodulator can be realized by an MF demodulator followed by a minimum mean-squared error whitening transformation* $\widehat{\mathbf{W}} = (S^*S)^{-1/2} = (\mathbf{S}^*\mathbf{S})^{-1/2}$*;*

2. *The signals* $\hat{g}_i(t)$ *minimize the least-squares error given by (9.4), i.e., they are the closest orthonormal signals to the signals* $s_i(t)$*;*

3. *The vectors* $\hat{\mathbf{g}}_i$ *are the closest orthonormal vectors to the vectors* $\mathbf{s}_i$*, and are the columns of the partial isometry in the polar decomposition of* $\mathbf{S}$*.*

Theorem 9.2 (Projected orthogonal matched filter (POMF) demodulator). *Let* $\{s_i(t), 1 \leq i \leq m\}$ *denote a set of* m *transmitted signals in a Hilbert space* $\mathcal{H}$ *that span an* n*-dimensional subspace* $\mathcal{U} \subset \mathcal{H}$*. Let* $\{\hat{f}_i(t), 1 \leq i \leq m\}$ *denote the POMF signals that are the correlating signals of the POMF demodulator. Let* S *and* $\widehat{F}$ *denote the set transformations corresponding to the signals* $s_i(t)$ *and* $\hat{f}_i(t)$*, respectively. Then*

$$\widehat{F} = S((S^*S)^{1/2})^{\dagger}.$$

Let X *denote a set transformation corresponding to an orthonormal basis for* $\mathcal{U}$*, let* $s_i(t) = X\mathbf{s}_i$*, and let* $\mathbf{S}$ *be the matrix of columns* $\mathbf{s}_i$ *with SVD* $\mathbf{S} = \mathbf{U}\Sigma\mathbf{V}^*$*. Then* $\hat{f}_i(t) = X\hat{\mathbf{f}}_i$ *where*

the vectors $\hat{\mathbf{f}}_i$ *are the columns of* $\widehat{\mathbf{F}}$*, and*

$$\widehat{\mathbf{F}} = \mathbf{S}((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger = \mathbf{U}\mathbf{I}_{nm}\mathbf{V}^*,$$

where $\mathbf{I}_{nm}$ *is given by (9.16). In addition,*

1. *The POMF demodulator can be realized by an MF demodulator followed by a minimum mean-squared error subspace whitening transformation* $\widehat{\mathbf{W}}_s = ((S^*S)^{1/2})^\dagger = ((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger$*;*

2. *The signals* $\hat{f}_i(t)$ *minimize the least-squares error given by (9.11), i.e., they form the closest normalized tight frame to the signals* $s_i(t)$*;*

3. *The vectors* $\hat{\mathbf{f}}_i$ *form the closest normalized tight frame to the vectors* $\mathbf{s}_i$*, and are the projections onto* $\mathcal{U}$ *of the columns of the partial isometry in a polar decomposition of* $\mathbf{S}$*.*

Finally, we note that based on results derived in the context of quantum detection [26] it can be shown that in many cases the OMF and POMF demodulators have an additional property, analogous to the SNR property of the MF demodulator.

Specifically, from the equivalence between the maximum SNR_T problem and the quantum detection problem established in Section 9.2.1, and the discussion in Section 3.4 regarding the LS quantum measurement, it follows that when the signals $s_i(t)$ are geometrically uniform, the OMF and POMF signals maximize SNR_T subject to the constraint that the outputs of the demodulator are uncorrelated on the space in which they lie. In [77] it is claimed that most practical signal sets used in digital communication are indeed geometrically uniform. Thus, in a communications context the POMF and OMF demodulators have a property analogous to the MF demodulator, namely that they typically maximize the total SNR_T subject to the decorrelation constraint. This provides some additional justification for this class of receivers.

Further results regarding the orthogonal or tight frame signals that maximize SNR_T that follow from results pertaining to quantum detection are that if the signals are nearly orthogonal, then the OMF and POMF signals maximize SNR_T [150]. Iterative algorithms for maximizing SNR_T for arbitrary signal sets are given in Section 8.5.4 and in [150].

9.6 Simulation Results

In this section we provide simulation results suggesting the behavior and performance of the OMF detector, in comparison to the MF detector[3].

The behaviors of the detectors where simulated in non-Gaussian and Gaussian noise using random signal constellations. The signals in the constellation have dimension m equal to the number of the signals in the constellation, and the samples of the signals are mutually independent zero-mean Gaussian random variables with variance $1/\sqrt{m}$, scaled to have norm 1.

We considered two different distributions for the non-Gaussian noise. The first is a Gaussian mixture of two components with equal weights. This choice of distribution is motivated by the fact that Gaussian mixtures have been used extensively to model non-Gaussian noise [151, 152, 153], and in part because the Gaussian mixture model is capable of closely approximating many non-Gaussian distributions. The second distribution is the Beta distribution, which is chosen since it is very flexible and capable of attaining a wide variety of shapes by varying its two parameter values a and b. Depending on the values of these parameters the Beta distribution will have the "U", the "J", the triangle or the general bell shape. In addition, the Beta distribution can model the effect of several noise components since the sum of N Gamma-distributed random variables is Beta-distributed, if N is not too large [154].

We generated 500 realizations of signals. For each signal realization, we determined the probability of correct detection for the detectors in both types of noise by recording the number of successful detections over 500 noise realizations. We then plotted histograms of the probability of correct detection P_d for the different detectors, which indicated that P_d has a unimodal distribution with a bell-shaped appearance. Therefore, it is reasonable to compactly present the results in terms of the mean and standard deviation of P_d for the various detectors.

[3]These simulations where performed in collaboration with D. Egnor and appear in [37]. Additional simulation results appear in [36].

9.6.1 Gaussian Mixture Noise

We first considered a Gaussian mixture of two components each with standard deviation 0.25 centered at ± 1, corresponding to an SNR close to 0 dB.

In Fig. 9-3 we plot the mean of P_d for the OMF detector and the MF detector as a function of the number of signals in the transmitted constellation. The vertical lines indicate the standard deviation of P_d. From the figure it is evident that at this SNR the OMF detector outperforms the MF detector, where the relative improvement in performance of the OMF detector over the MF detector increases for increasing constellation size.

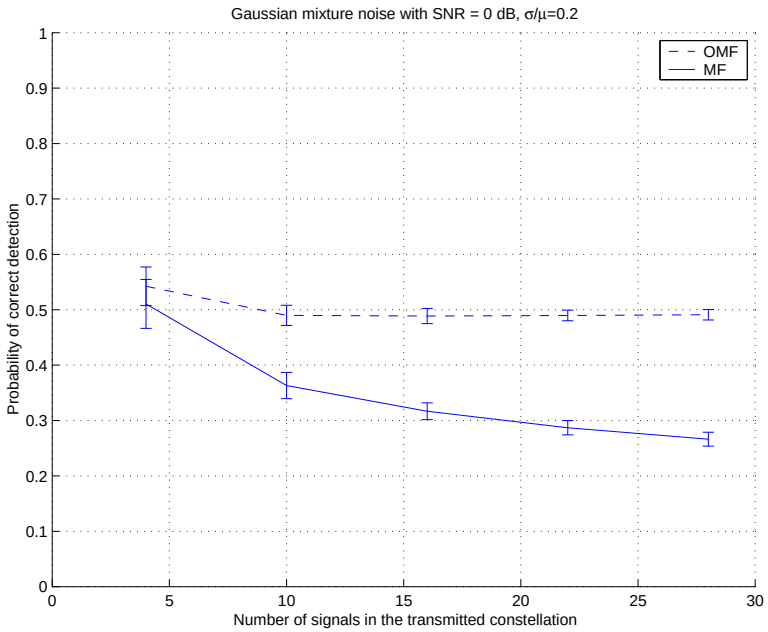


Figure 9-3: Comparison between the OMF and MF in Gaussian mixture noise, as a function of the number of signals in the transmitted constellation. The mixture components have standard deviation of 0.25 and are centered at ± 1. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

We repeated the simulations for different parameters of the Gaussian mixture components, again at an SNR of 0 dB. In general we found that the relative improvement of the OMF detector over the MF detector increased as the separation between the mixtures increased. When the separation is decreased relative to the mixture standard deviation the relative improvement in performance using the OMF detector decreases, consistent with the fact that the Gaussian mixture distribution approaches a Gaussian distribution. The

same behavior is evident when varying the standard deviation of the mixture components for fixed mean separation. In Fig. 9-4 we plot the mean of P_d for the OMF and MF detectors for constellations of 13 signals in Gaussian mixture noise of two components each with standard deviation σ centered around $\pm\mu$, as a function of σ/μ. The vertical lines indicate the standard deviation of P_d. As the standard deviation of the mixture components increases relative to the mixture mean, the Gaussian mixture distribution approaches a Gaussian distribution, in which case the relative improvement in performance using the OMF detector decreases.

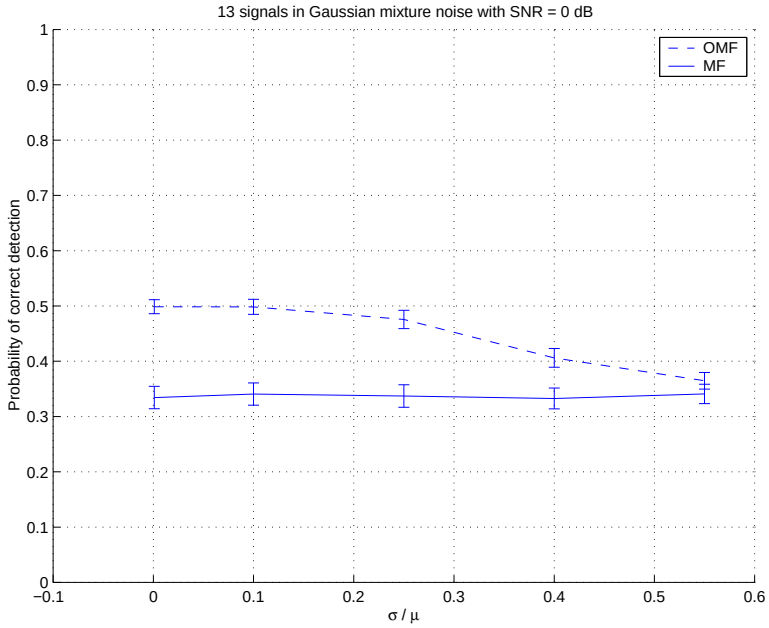


Figure 9-4: Comparison between the OMF and MF detectors for transmitted constellations of 13 signals in Gaussian mixture noise with mixture components with standard deviation σ centered at $\pm\mu$, as a function of σ/μ. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

In general we observed that the relative improvement in performance of the OMF over the MF detector increased with decreasing SNR, and is predominant for large signal constellation size. For increasing values of SNR the relative improvement in performance using the OMF detector decreases.

The qualitative behavior of the POMF detector in comparison to the MF detector when varying the Gaussian mixture parameters and the SNR is similar to that of the OMF

detector.

9.6.2 Beta Distributed Noise

We next consider Beta-distributed noise with a variety of parameter values.

In Fig. 9-5 we plot the mean of P_d for the OMF detector and the MF detector in Beta-distributed noise with $a = b = 1$, as a function of the number of signals in the transmitted constellation. The vertical lines indicate the standard deviation of P_d. From the figure it is evident that the OMF detector outperforms the MF detector, where the relative improvement in performance of the OMF detector over the MF detector increases for increasing constellation size.

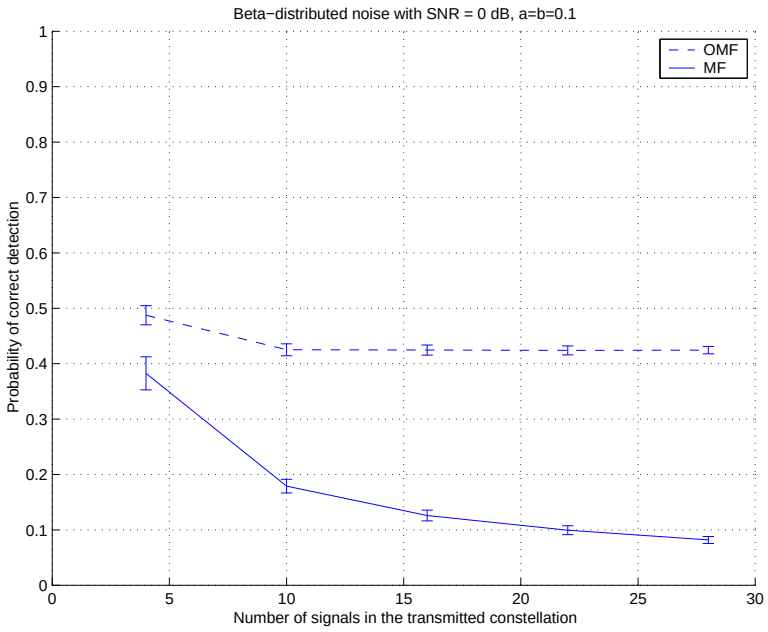


Figure 9-5: Comparison between the OMF and MF detectors in Beta-distributed noise, as a function of the number of signals in the transmitted constellation. The parameters of the distribution are $a = b = 0.1$. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

We repeated the simulations for different parameter values. In general we found that the relative improvement of the OMF detector over the MF detector increased as the distribution became more bimodal. In Figs. 9-6–9-8 we plot the mean of P_d for the OMF and MF detectors for constellations of 13 signals in Beta-distributed noise with varying

parameters. The vertical lines indicate the standard deviation of P_d. As the b parameter increases, the Beta distribution approaches a unimodal distribution, in which case the relative improvement in performance using the OMF detector decreases.

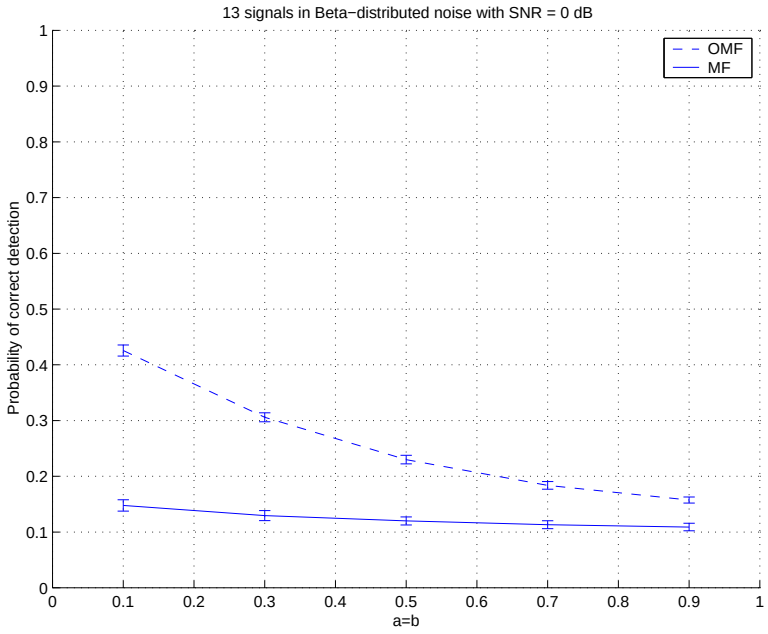


Figure 9-6: Comparison between the OMF and MF detectors for transmitted constellations of 13 signals in Beta-distributed noise, as a function of the parameters with $a = b$. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

In Fig. 9-9 we plot the mean and standard deviation of P_d for the OMF and MF detectors as a function of SNR for transmitted constellations of 13 signals, in Beta-distributed noise with $a = b = 0.1$. The SNR is given by $10 \log P_s/\sigma^2$, where P_s is the signal power and the variance of the Beta distribution is given in terms of the parameters a and b as

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}. \tag{9.22}$$

The improvement in performance of the OMF over the MF detector is predominant for low to intermediate values of SNR.

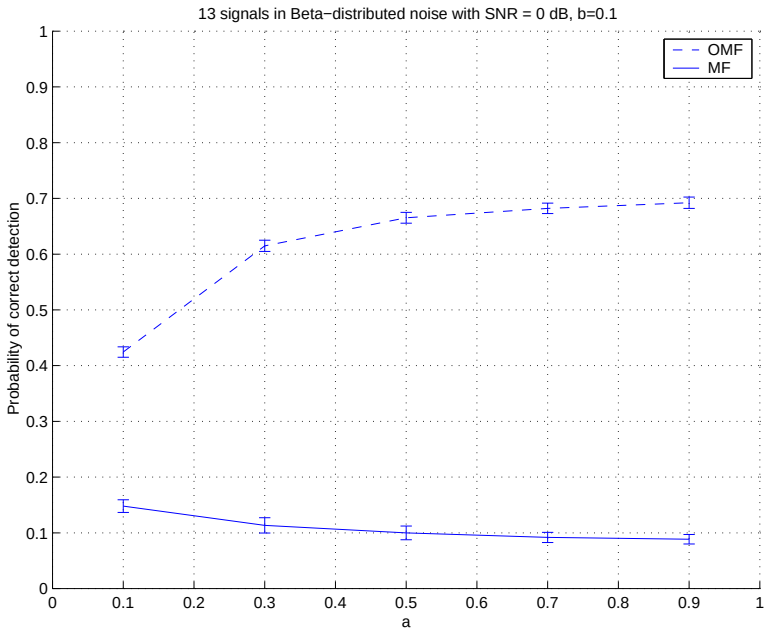


Figure 9-7: Comparison between the OMF and MF detectors for transmitted constellations of 13 signals in Beta-distributed noise with $b = 0.1$, as a function of the parameter a. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

9.6.3 Gaussian Noise

We repeated the simulations leading to the results previously presented with zero-mean Gaussian noise. In Fig. 9-10 we plot the mean of P_d for the OMF detector and the MF detector in Gaussian noise, as a function of the number of signals in the transmitted constellation. The vertical lines indicate the standard deviation of P_d. In Fig. 9-11 we plot the mean of P_d using the OMF and MF detectors for transmitted constellations of 7 signals in Gaussian noise, as a function of SNR.

As expected, for Gaussian noise the MF detector outperforms the OMF detector. This is consistent with the fact that the MF detector maximizes the probability of correct detection for Gaussian noise. However, it is evident from Figs. 9-10–9-11 that the relative improvement in performance using the MF detector over the OMF is not very significant. Specifically, in Fig. 9-11 note that the maximum (mean) difference in probability of correct detection is less than 0.08. These results are encouraging since they suggest that if the receiver is designed to operate in different noise environments, or in an unknown noise

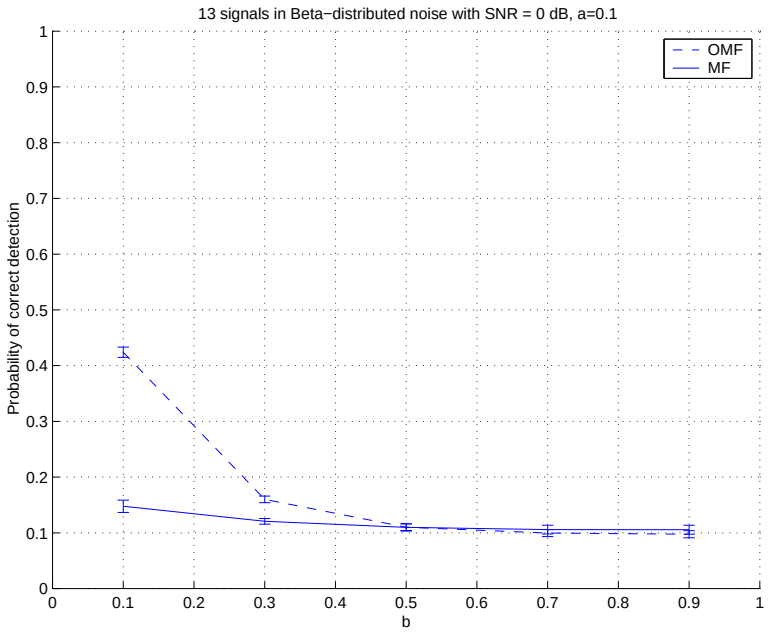


Figure 9-8: Comparison between the OMF and MF detectors for transmitted constellations of 13 signals in Beta-distributed noise with $a = 0.1$, as a function of the parameter b. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

environment, than we may prefer using an OMF or POMF detector since for certain non-Gaussian noise distributions these detectors may result in a substantial improvement in performance over an MF detector, without significantly degrading the performance if the noise is Gaussian.

9.7 Inner Product Shaping Matched Filter Detection

In this chapter we focused our attention primarily on modified receivers that result from constraining the outputs of the correlation demodulator to be uncorrelated on an appropriate subspace. Equivalently, the correlating signals are constrained to be orthonormal or to be projections of orthonormal signals. We may also consider modified receivers in which the correlating signals are chosen to shape the correlation of the demodulator outputs. In this case the modified receiver consists of signals $q_i(t)$ with the desired inner product structure, that are closest in a LS sense to the transmitted signals $s_i(t)$. The optimal signals $q_i(t)$ can be readily determined from the results of Chapter 8.

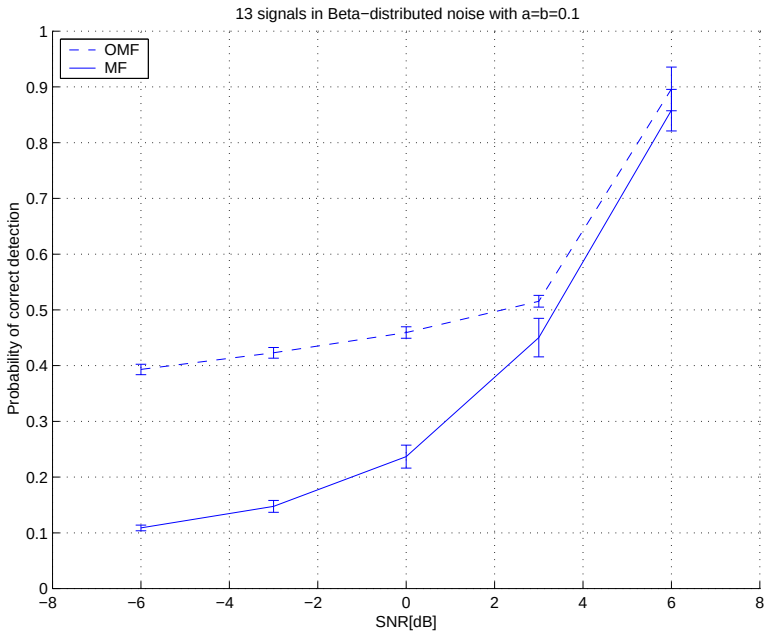


Figure 9-9: Comparison between the OMF and MF detectors for transmitted constellations of 13 signals in Beta-distributed noise with $a = b = 0.1$, as a function of the SNR. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

As in Section 9.4, we can show that a correlation demodulator with signals $q_i(t)$ with inner products $Q^*Q = \mathbf{R}$, that are closest in a LS sense to the signals $s_i(t)$, can be equivalently implemented as an MF demodulator followed be an optimal covariance shaping transformation, that optimally shapes the covariance of the MF output prior to detection. The optimal shaping transformation has the property that it minimizes the MSE between the MF output and the shaped output. In the next chapter we consider the MMSE covariance shaping problem in its most general form.

Preliminary simulations demonstrate that in a variety of cases choosing a set of non-orthogonal correlating signals with a specified inner product structure, can further improve the performance of the modified receiver over the MF receiver. An interesting and potentially promising direction for future research is to investigate the relationship between the inner product structure of the correlating signals, the inner products of the transmitted signals, and the receiver performance. In particular, it would appear useful to design an optimality criterion for choosing the desired inner product structure, based on knowledge

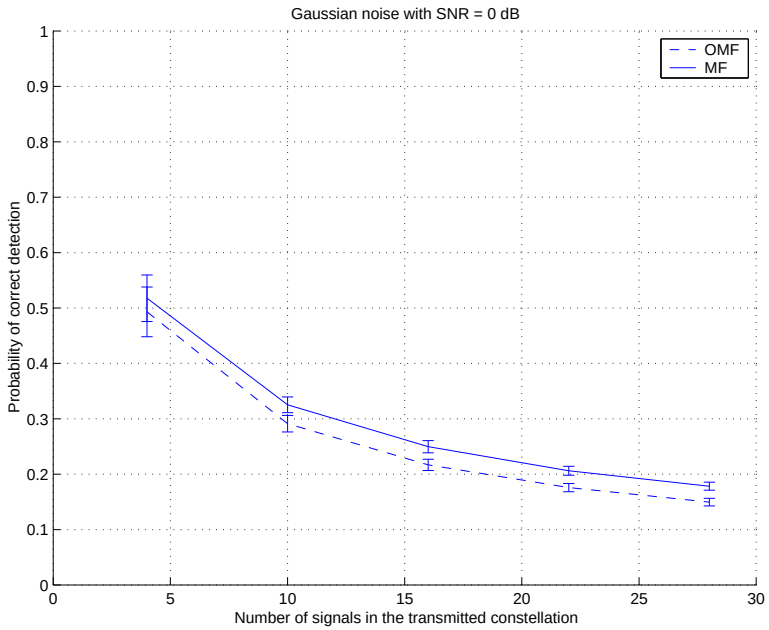


Figure 9-10: Comparison between the OMF and MF detectors in zero mean, unit variance Gaussian noise, as a function of the number of signals in the transmitted constellation. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

of the transmitted signals.

It is also interesting to evaluate the performance of the modified receivers when the signals $q_i(t)$ are designed to have a specified inner product structure, so that the eigenvectors of $\mathbf{R} = Q^*Q$ are specified, and the eigenvalues are chosen to minimize the LS error. The optimal signals $q_i(t)$ of this form can again be determined using the results of Chapter 8.

9.8 Summary and Remarks

In this chapter we provided a preliminary development of inner product shaping matched filter detection. However, there are various aspects of this new class of receivers that warrant further study and evaluation.

As noted in the previous section, an interesting and important direction to explore is the use of other forms of inner product shaping in the design of the modified receiver. Of particular importance would be to determine methods for optimal design of the correlating signals inner product structure, or equivalently, the output covariance shape.

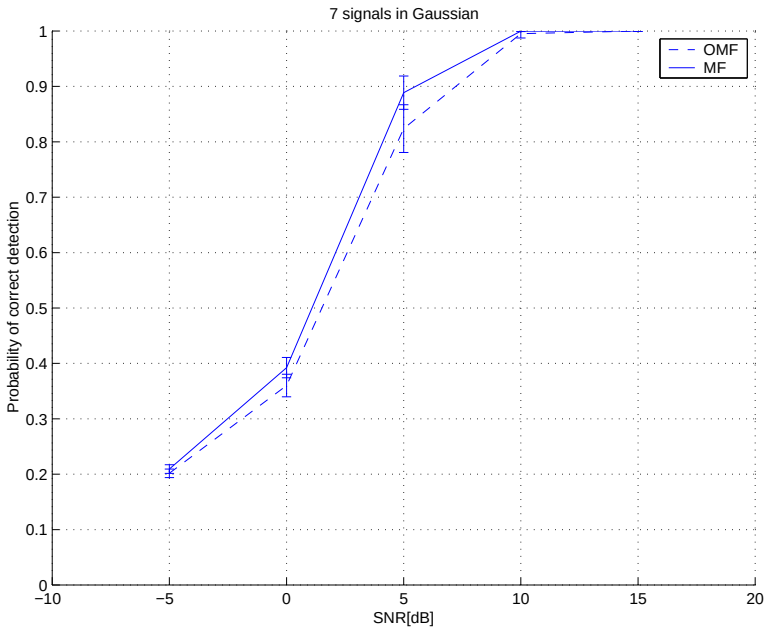


Figure 9-11: Comparison between the OMF and MF detectors for transmitted constellations of 7 signals in Gaussian noise, as a function of SNR. The dashed line is the mean P_d using the OMF detector, and the solid line is the mean P_d using the MF detector. The vertical lines indicate the standard deviation of the corresponding P_d.

The performance improvement using the modified receiver was demonstrated through simulation only. It would be extremely valuable to analyze the behavior of the proposed receivers analytically. Of related interest is the development of analytical methods for determining under which non-Gaussian distributions the modified receivers lead to improved performance over the MF receiver, for example using large deviation theory and the Chernoff bound.

Another issue that deserves attention is whether inner product shaping matched filter detection is optimal in some sense, for example under certain channel assumptions. Although we demonstrated through simulation that in some cases of non-Gaussian noise the OMF receiver can lead to a significant improvement in detection performance, we have not established that it is optimal under any specific criterion.

Of significant practical interest is the robustness of the modified receivers with respect to modeling errors, for example in the case in which the additive noise is not white or does not have zero mean, or in the case in which the transmitted signals are not known exactly. Preliminary simulations demonstrate that the modified receivers have improved

resilience to modeling errors over the MF receiver, however this requires further evaluation and investigation. It is also interesting to evaluate the performance of the receivers for coded systems.

In our closing remarks, we note that we can readily extend the results developed in this chapter to the case in which the signals $s_i(t)$ have unequal norm, or the case in which the signals are transmitted with unequal probability. In such cases, we may choose the correlating signals $q_i(t)$ to minimize a weighted LS error as in Section 8.3, where the weights may be chosen to reflect the signal priors, or the signal norms.

Chapter 10

MMSE Covariance Shaping

In this chapter we demonstrate how the ideas and results derived in the context of quantum detection lead to an interesting new perspective on covariance shaping problems. Specifically, we develop linear covariance shaping transformations that minimize the MSE between the original and shaped data, *i.e.,* that result in an output with the desired covariance that is as close as possible to the input, in an MSE sense. As we have seen in the previous chapter, the concept of MMSE covariance shaping is closely related to the concept of LS inner product shaping developed in Chapter 8. Thus, we may view MMSE covariance shaping as a stochastic analog of the optimal QSP measurement design problem.

Data shaping arises in a variety of contexts in which it is useful to decorrelate or otherwise shape a data sequence either prior to subsequent processing, or to control the spectral shape after processing. For example, in a multi-signature system the received signal is typically processed by a bank of filters matched to the signatures, with the declared signature being that associated with the maximum value in the MF vector output. In general there may be correlation between the elements of the vector output of the MF bank. As shown in the previous chapter and in [36, 37], the probability of correct detection can be improved in many cases by first shaping the MF output prior to applying a simple detection algorithm. Other contexts in which data shaping has been used to advantage include enhancing direction of arrival algorithms by pre-whitening [34, 35].

As is well known, the transformation that shapes a data vector is not unique. While in some applications of covariance shaping certain conditions might be imposed on the transformation such as causality or symmetry, with the exception of the work in [36, 37, 38,

39, 41] which explicitly relies on the optimality properties developed here, there have been no general assertions of optimality for various choices of a covariance shaping transformation.

Shaping the covariance of a data vector or signal introduces distortion to the values of the data relative to the unshaped data. In certain applications of shaping, it is desirable to shape the data while minimizing this distortion, *i.e.,* to choose the shaping transformation in an optimal sense. Drawing from the notion of optimal QSP measurements, in this chapter we develop a linear shaping transformation that minimizes the MSE between the original and shaped data, *i.e.,* that results in an output that is as close as possible to the input, in an MSE sense. We refer to such a covariance shaping transformation as an MMSE covariance shaping transformation.

10.1 Optimal Covariance Shaping Transformation

Let $\mathbf{a} \in \mathbb{C}^m$ denote a zero-mean[1] random vector with rank-n covariance matrix $\mathbf{C}_a$. We wish to shape the covariance of $\mathbf{a}$ using a shaping transformation $\mathbf{T}$ to obtain the random vector $\mathbf{b} = \mathbf{T}\mathbf{a}$, where the covariance matrix of $\mathbf{b}$ is given by $\mathbf{C}_b = c^2\mathbf{R}$ for some $c > 0$ and rank-r matrix $\mathbf{R}$ with $r \leq n$. Thus we seek a transformation $\mathbf{T}$ such that

$$\mathbf{C}_b = \mathbf{T}\mathbf{C}_a\mathbf{T}^* = c^2\mathbf{R}, \tag{10.1}$$

for some $c > 0$. We refer to any $\mathbf{T}$ satisfying (10.1) as a covariance shaping transformation.

Given a covariance matrix $\mathbf{C}_a$, there are many ways to choose a covariance shaping transformation $\mathbf{T}$ satisfying (10.1). Although there are an unlimited number of covariance shaping transformations satisfying (10.1), no general assertion of optimality is known for the output $\mathbf{b} = \mathbf{T}\mathbf{a}$ of these different transformations. In particular, the random vector $\mathbf{b} = \mathbf{T}\mathbf{a}$ may not be "close" to the input vector $\mathbf{a}$. If the vector $\mathbf{b}$ undergoes some noninvertible processing, or is used as an estimator of some unknown parameters represented by the data $\mathbf{a}$, then we may wish to choose the shaping transformation in a way that $\mathbf{b}$ is close to $\mathbf{a}$ in some sense. This can be particularly important in applications in which $\mathbf{b}$ is the input to a detector, so that we may wish to shape the covariance of $\mathbf{a}$ prior to detection, but at the same time minimize the distortion to $\mathbf{a}$ by choosing $\mathbf{T}$ so that $\mathbf{b}$ is close to $\mathbf{a}$.

[1] If the mean $E(\mathbf{a})$ is not zero, then we can always define $\mathbf{a}' = \mathbf{a} - E(\mathbf{a})$ so that the results hold for $\mathbf{a}'$.

We therefore propose a shaping transformation that is optimal in the sense that it results in a random vector $\mathbf{b}$ that is as close as possible to $\mathbf{a}$ in MSE. Specifically, among all possible covariance shaping transformations we seek the one that minimizes the total MSE given by

$$\varepsilon_{\text{MSE}} = \sum_{i=1}^{m} E\left((a_i - b_i)^2\right) = E\left((\mathbf{a}-\mathbf{b})^*(\mathbf{a}-\mathbf{b})\right), \tag{10.2}$$

subject to (10.1), where a_i and b_i are the ith components of $\mathbf{a}$ and $\mathbf{b}$ respectively. We may wish to constraint the constant c in (10.1), or may choose c to minimize the total MSE.

Our approach to determining the shaping transformation $\widehat{\mathbf{T}}$ that minimizes (10.2) is to interpret the MMSE covariance shaping problem as a LS inner product shaping problem, discussed in Chapter 8, and then apply results derived in that context.

To this end, let $\mathbf{C}_a$ have an eigendecomposition $\mathbf{C}_a = \mathbf{V}\mathbf{D}\mathbf{V}^*$ where $\mathbf{V}$ is a unitary matrix and $\mathbf{D}$ is a diagonal matrix whose first n diagonal elements are positive, and whose remaining diagonal elements are all equal 0, and let $\mathbf{R}$ have an eigendecomposition $\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^*$ where $\mathbf{Q}$ is a unitary matrix and Λ is a diagonal matrix whose first $r \leq n$ diagonal elements are positive, and whose remaining diagonal elements are all equal 0. Define $\mathbf{S} = \mathbf{D}^{1/2}\mathbf{V}^*$ and $\mathbf{H} = c\Lambda^{1/2}\mathbf{Q}^*$ so that $\mathbf{S}^*\mathbf{S} = \mathbf{C}_a$ and $\mathbf{H}^*\mathbf{H} = c^2\mathbf{R}$, and let $\mathbf{n}$ be a zero-mean random vector with covariance $\mathbf{I}_m$. Then $\mathbf{a}$ and $\mathbf{b}$ have the same mean and covariance as $\mathbf{S}^*\mathbf{n}$ and $\mathbf{H}^*\mathbf{n}$ respectively, so that as long as we are concerned with first and second order statistics only, we may write $\mathbf{a} = \mathbf{S}^*\mathbf{n}$ and $\mathbf{b} = \mathbf{H}^*\mathbf{n}$. We can then express ε_{MSE} as

$$\begin{aligned}
\varepsilon_{\text{MSE}} &= E\left((\mathbf{a}-\mathbf{b})^*(\mathbf{a}-\mathbf{b})\right) \\
&= \text{Tr}\left(E\left((\mathbf{a}-\mathbf{b})(\mathbf{a}-\mathbf{b})^*\right)\right) \\
&= \text{Tr}\left((\mathbf{S}-\mathbf{H})^* E(\mathbf{n}\mathbf{n}^*)(\mathbf{S}-\mathbf{H})\right) \\
&= \text{Tr}\left((\mathbf{S}-\mathbf{H})^*(\mathbf{S}-\mathbf{H})\right).
\end{aligned} \tag{10.3}$$

Thus, seeking a covariance shaping transformation $\mathbf{T}$ to minimize the MSE is equivalent to seeking the matrix $\mathbf{H}$ with columns $\mathbf{h}_i$ that are closest to the columns $\mathbf{s}_i$ of $\mathbf{S}$ in a LS error sense, subject to $\mathbf{H}^*\mathbf{H} = c^2\mathbf{R}$. This problem is just the LS inner product shaping problem discussed in Chapter 8. From Theorem 8.3 it follows that the optimal vectors $\hat{\mathbf{h}}_i$

are the columns of

$$\widehat{\mathbf{H}} = \alpha \mathbf{U}\mathbf{Z}^* \Lambda^{1/2} \mathbf{Q}^*, \tag{10.4}$$

where $\mathbf{S}\mathbf{Q}\Lambda^{1/2} = \mathbf{U}\Sigma\mathbf{Z}^*$ is the SVD of $\mathbf{S}\mathbf{Q}\Lambda^{1/2}$. If c in (10.1) is fixed then $\alpha = c$, and if c is chosen to minimize the MSE then $\alpha = \hat{c}$ with

$$\hat{c} = \frac{\mathrm{Tr}\left((\mathbf{C}_a\mathbf{R})^{1/2}\right)}{\mathrm{Tr}(\mathbf{R})}. \tag{10.5}$$

It also follows from the derivation of the LS vectors that $\widehat{\mathbf{H}}$ can always be chosen so that the columns $\hat{\mathbf{h}}_i$ of $\widehat{\mathbf{H}}$ lie in the space $\mathcal{U} = \mathcal{R}(\mathbf{D})$ spanned by the vectors $\mathbf{s}_i$. Specifically, with $\mathbf{Y} = \mathbf{S}\mathbf{Q}\Lambda^{1/2}$ and $l = \mathrm{rank}(\mathbf{Y})$, only the first l columns of $\mathbf{U}$ are determined by the SVD of $\mathbf{Y}$. These columns span a subspace of $\mathcal{U}$. We can therefore always chose the next $r - l$ columns of $\mathbf{U}$ so that together, the first r columns span a subspace of $\mathcal{U}$. Since, as showed in Chapter 8, the range space of $\widehat{\mathbf{H}}$ is spanned by the first r columns of $\mathbf{U}$, this then implies that the vectors $\hat{\mathbf{h}}_i$ lie in $\mathcal{U}$.

Thus with an appropriate choice of $\mathbf{U}$, $\hat{\mathbf{h}}_i \in \mathcal{U}$, and we may write $\widehat{\mathbf{H}}$ as

$$\widehat{\mathbf{H}} = P_{\mathcal{U}}\widehat{\mathbf{H}} = \mathbf{S}\mathbf{S}^\dagger\widehat{\mathbf{H}} = \mathbf{S}\widehat{\mathbf{X}}, \tag{10.6}$$

where $\widehat{\mathbf{X}} = \mathbf{S}^\dagger\widehat{\mathbf{H}}$, and $\mathbf{S}^\dagger = \mathbf{V}(\mathbf{D}^{1/2})^\dagger$. The optimal $\mathbf{b}$ is then given by $\mathbf{b} = \widehat{\mathbf{H}}^*\mathbf{n} = \widehat{\mathbf{X}}^*\mathbf{S}^*\mathbf{n} = \widehat{\mathbf{X}}^*\mathbf{a}$, so that the MMSE covariance shaping transformation $\widehat{\mathbf{T}}$ is related to $\widehat{\mathbf{H}}$ through

$$\widehat{\mathbf{T}} = \widehat{\mathbf{X}}^* = \widehat{\mathbf{H}}^*(\mathbf{S}^\dagger)^*. \tag{10.7}$$

Finally, from (10.4)

$$\widehat{\mathbf{T}} = \alpha\mathbf{Q}\Lambda^{1/2}\mathbf{Z}\mathbf{U}^*(\mathbf{S}^\dagger)^* = \alpha\mathbf{Q}\Lambda^{1/2}\mathbf{Z}\mathbf{U}^*(\mathbf{D}^{1/2})^\dagger\mathbf{V}^*. \tag{10.8}$$

In the special case in which $r = n = m$ so that $\mathbf{C}_a$ and $\mathbf{R}$ are both positive definite,

$$\begin{aligned} \mathbf{U}\mathbf{Z}^* &= \mathbf{S}\mathbf{Q}\Lambda^{1/2}(\Lambda^{1/2}\mathbf{Q}^*\mathbf{S}^*\mathbf{S}\mathbf{Q}\Lambda^{1/2})^{-1/2} \\ &= \mathbf{D}^{1/2}\mathbf{V}^*(\mathbf{R}\mathbf{C}_a)^{-1/2}\mathbf{Q}\Lambda^{1/2}, \end{aligned} \tag{10.9}$$

where we used (B.3), and $\widehat{\mathbf{T}}$ reduces to

$$\widehat{\mathbf{T}} = \alpha \mathbf{Q}\Lambda^{1/2}\mathbf{Z}\mathbf{U}^*\mathbf{D}^{-1/2}\mathbf{V}^* = \alpha \mathbf{R}(\mathbf{C}_a\mathbf{R})^{-1/2} = (\mathbf{R}\mathbf{C}_a)^{-1/2}\mathbf{R}. \tag{10.10}$$

We summarize our results regarding MMSE covariance shaping in the following theorem:

Theorem 10.1 (MMSE covariance shaping). *Let* $\mathbf{a} \in \mathbb{C}^m$ *be a random vector with rank-n covariance matrix* $\mathbf{C}_a = \mathbf{V}\mathbf{D}\mathbf{V}^*$. *Let* $\widehat{\mathbf{T}}$ *be the optimal covariance shaping transformation that minimizes the MSE defined by (10.2), between the input* $\mathbf{a}$ *and the output* $\mathbf{b} = \mathbf{T}\mathbf{a}$ *with rank-r covariance matrix* $\mathbf{C}_b = c^2\mathbf{R}$ *where* $\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^*$ *and* $c > 0$. *Let* $\mathbf{D}^{1/2}\mathbf{V}^*\mathbf{Q}\Lambda^{1/2} = \mathbf{U}\Sigma\mathbf{Z}^*$ *where the first r columns of* $\mathbf{U}$ *are chosen so that they span a subspace of* $\mathcal{R}(\mathbf{D})$. *Then*

$$\widehat{\mathbf{T}} = \alpha \mathbf{Q}\Lambda^{1/2}\mathbf{Z}\mathbf{U}^*(\mathbf{D}^{1/2})^{\dagger}\mathbf{V}^*,$$

where

1. *if c is specified then* $\alpha = c$;

2. *if c is chosen to minimize the MSE then* $\alpha = \hat{c}$ *where* $\hat{c} = Tr\left((\mathbf{C}_a\mathbf{R})^{1/2}\right) / Tr(\mathbf{R})$.

In addition, if $r = n = m$ *then* $\widehat{\mathbf{T}} = \alpha \mathbf{R}(\mathbf{C}_a\mathbf{R})^{-1/2} = \alpha(\mathbf{R}\mathbf{C}_a)^{-1/2}\mathbf{R}$.

10.2 Examples of MMSE Covariance Shaping

In this section we consider some special cases of MMSE covariance shaping. We first consider MMSE whitening and subspace whitening, in which the whitening transformation is designed to optimally whiten the vector on a subspace in which it is contained. These concepts are developed in more detail in [40]. Applications of MMSE whitening and subspace whitening to MF detection where previously considered in Chapter 9. We then consider the case of MMSE unwhitening and subspace unwhitening, in which the random vector $\mathbf{a}$ is white on the subspace in which it is contained, and the problem is to optimally shape its covariance.

10.2.1 MMSE Whitening

Suppose that $\mathbf{a} \in \mathbb{C}^m$ is a random vector with positive definite covariance matrix $\mathbf{C}_a$. We seek a whitening transformation $\mathbf{T} = \mathbf{W}$ such that the vector $\mathbf{b} = \mathbf{W}\mathbf{a}$ has covariance $\mathbf{C}_b = c^2\mathbf{I}_m$ for some $c > 0$, and is as close as possible to $\mathbf{a}$ in the MSE sense.

The MMSE whitening transformation follows from Theorem 10.1 with $\mathbf{R} = \mathbf{I}_m$,

$$\widehat{\mathbf{W}} = \alpha \mathbf{C}_a{}^{-1/2}, \tag{10.11}$$

where $\alpha = c$ if c is fixed and $\alpha = \hat{c}$ if c is chosen to minimize the MSE where

$$\hat{c} = \frac{1}{m}\mathrm{Tr}\left(\mathbf{C}_a^{1/2}\right). \tag{10.12}$$

It is interesting to note that the MMSE whitening transformation (10.11) has the additional property that it is the unique *symmetric* whitening transformation [155]. It is also proportional to the Mahalanobis transformation, that is frequently used in signal processing applications incorporating whitening (see *e.g.*, [44, 34, 35]).

10.2.2 MMSE Subspace Whitening

Suppose now that $\mathbf{C}_a$ has rank $n < m$. In this case there is no whitening transformation $\mathbf{W}$ such that $\mathbf{W}\mathbf{C}_a\mathbf{W}^* = c^2\mathbf{I}_m$. Instead, we propose whitening $\mathbf{a}$ on the space in which it is contained, which we refer to as *subspace whitening*.

Subspace whitening

Let $\mathbf{a}$ be a zero-mean random vector in $\mathbb{C}^m$ with rank-n covariance matrix $\mathbf{C}_a$, and let $\mathcal{V} \subset \mathbb{C}^m$ denote the range space $\mathcal{R}(\mathbf{C}_a)$. If $\mathbf{C}_a$ is not invertible, then the elements of $\mathbf{a}$ are deterministically linearly dependent with probability one (w.p. 1)[2], and consequently any realization of the random vector $\mathbf{a}$ lies in $\mathcal{V}$ (see Appendix C.1). We may therefore consider whitening $\mathbf{a}$ on $\mathcal{V}$, which we refer to as subspace whitening.

First consider a zero mean random vector $\mathbf{q} \in \mathbb{C}^m$ with full-rank covariance matrix, and let $\mathbf{y} = \mathbf{W}\mathbf{q}$ where $\mathbf{W}$ is a whitening transformation, so that $\mathbf{y}$ is white. Then $\mathbf{y}$ and $\mathbf{q}$ lie in

[2]Throughout this section when we say that the elements of a random vector are linearly dependent we mean w.p. 1; similarly, when we say that a random vector lies in a subspace we mean w.p. 1.

the same space $\mathbb{C}^m$. Furthermore, if $\mathbf{y}$ is white then the representation of $\mathbf{y}$ in terms of any orthonormal basis for $\mathbb{C}^m$ is also white. This follows from the fact that any two orthonormal bases for $\mathbb{C}^m$ are related through a unitary transformation. We define subspace whitening to preserve these two properties.

Let $\mathbf{a}$ be a random vector with covariance $\mathbf{C}_a$ with range space $\mathcal{V}$, and let $\mathbf{b}$ denote the output of a subspace whitening transformation of $\mathbf{a}$. Since $\mathbf{a} \in \mathcal{V}$ we require that $\mathbf{b} \in \mathcal{V}$. In addition, we require that the representation of $\mathbf{b}$ in terms of some orthonormal basis for $\mathcal{V}$ is white, which then implies that the representation in terms of any orthonormal basis for $\mathcal{V}$ is white.

In Appendix C.2 we translate the conditions such that a random vector $\mathbf{b}$ is white on $\mathcal{V}$, to conditions on the covariance matrix $\mathbf{C}_b$. Specifically, we show that $\mathbf{C}_b$ must satisfy,

$$\mathbf{C}_b = c^2 P_{\mathcal{V}} = c^2 \mathbf{V} \tilde{\mathbf{I}}_n \mathbf{V}^*, \tag{10.13}$$

where the first n columns of $\mathbf{V}$ form an orthonormal basis for $\mathcal{V}$, and

$$\tilde{\mathbf{I}}_n = \begin{bmatrix} \mathbf{I}_n & 0 \\ 0 & 0 \end{bmatrix}. \tag{10.14}$$

MMSE subspace whitening transformation

To restate the MMSE subspace whitening problem, let $\mathbf{a} \in \mathbb{C}^m$ be a random vector with rank-n covariance matrix $\mathbf{C}_a = \mathbf{V}\mathbf{D}\mathbf{V}^*$ where $n < m$, and let $\mathcal{V} = \mathcal{R}(\mathbf{C}_a)$. We seek a subspace whitening transformation $\mathbf{T} = \mathbf{W}_s$ such that the vector $\mathbf{b} = \mathbf{W}_s \mathbf{a}$ is white on $\mathcal{V}$, namely such that $\mathbf{b}$ has a covariance matrix $\mathbf{C}_b = c^2 P_{\mathcal{V}} = c^2 \mathbf{V} \tilde{\mathbf{I}}_n \mathbf{V}^*$, where $\tilde{\mathbf{I}}_n$ is given by (10.14), $c > 0$, and is as close as possible to $\mathbf{a}$ in the MSE sense.

The MMSE subspace whitening transformation, denoted by $\widehat{\mathbf{W}}_s$, follows from Theorem 10.1 where we substitute $\mathbf{V}$ and $\tilde{\mathbf{I}}_n$ for $\mathbf{Q}$ and Λ respectively. With these substitutions, $\mathbf{U}$ and $\mathbf{Z}$ are determined from the SVD of $\mathbf{D}^{1/2}\mathbf{V}^*\mathbf{Q}\Lambda^{1/2} = \mathbf{D}^{1/2}$, so $\mathbf{U} = \mathbf{Z} = \mathbf{I}_m$. Then

$$\widehat{\mathbf{W}}_s = \alpha \mathbf{V} (\mathbf{D}^{1/2})^\dagger \mathbf{V}^* = \alpha (\mathbf{C}_a{}^{1/2})^\dagger, \tag{10.15}$$

where if c is fixed then $\alpha = c$, and if c is chosen to minimize the MSE then $\alpha = \hat{c}$ with

$$\hat{c} = \frac{1}{n}\mathrm{Tr}(\mathbf{C}_a{}^{1/2}). \tag{10.16}$$

It is intuitively reasonable and follows from the derivation (see [40]) that $\widehat{\mathbf{W}}_s$ is uniquely specified on $\mathcal{V}$, but can be arbitrary on $\mathcal{V}^\perp$. However, since the input $\mathbf{a}$ to the whitening transformation lies in $\mathcal{V}$ w.p. 1, the choice of $\widehat{\mathbf{W}}_s$ on $\mathcal{V}^\perp$ does not affect the output $\mathbf{b}$ (w.p. 1).

The results above are summarized in the following theorem:

Theorem 10.2 (MMSE subspace whitening). *Let $\mathbf{a} \in \mathbb{C}^m$ be a random vector with rank-n covariance matrix $\mathbf{C}_a = \mathbf{V}\mathbf{D}\mathbf{V}^*$, and let $\mathcal{V}$ denote the range space of $\mathbf{C}_a$. Let $\widehat{\mathbf{W}}_s$ be any subspace whitening transformation that minimizes the MSE defined by (10.2), between the input $\mathbf{a}$ and the output $\mathbf{b}$ with covariance matrix $\mathbf{C}_b = c^2 P_{\mathcal{V}} = c^2\mathbf{V}\tilde{\mathbf{I}}_n\mathbf{V}^*$, where $\tilde{\mathbf{I}}_n$ is given by (10.14) and $c > 0$. Then*

1. *$\widehat{\mathbf{W}}_s$ is not unique;*
2. *$\widehat{\mathbf{W}}_s = \alpha(\mathbf{C}_a{}^{1/2})^\dagger$ is an optimal subspace whitening transformation where*
   (a) *if c is specified then $\alpha = c$;*
   (b) *if c is chosen to minimize the MSE then $\alpha = (1/n)\,Tr(\mathbf{C}_a{}^{1/2})$;*
3. *Define $\mathbf{W}_s^{\mathcal{V}} = \widehat{\mathbf{W}}_s P_{\mathcal{V}}$ where $\widehat{\mathbf{W}}_s$ is any optimal subspace whitening transformation; then*
   (a) *$\mathbf{W}_s^{\mathcal{V}}$ is unique, and is given by $\mathbf{W}_s^{\mathcal{V}} = \alpha(\mathbf{C}_a{}^{1/2})^\dagger$;*
   (b) *$\widehat{\mathbf{W}}_s\mathbf{a} = \mathbf{W}_s^{\mathcal{V}}\mathbf{a}$ w.p. 1;*
   (c) *$\mathbf{b} = \widehat{\mathbf{W}}_s\mathbf{a}$ is unique w.p. 1.*

10.2.3 MMSE Unwhitening

Suppose now that $\mathbf{a} \in \mathbb{C}^m$ is a white random vector with covariance matrix $\mathbf{C}_a = \mathbf{I}_m$, and we want to "unwhiten" $\mathbf{a}$ to obtain the vector $\mathbf{b} = \mathbf{T}\mathbf{a}$ where the covariance of $\mathbf{b}$ is $\mathbf{C}_b = c^2\mathbf{R}$ for some $c > 0$, and is as close as possible to $\mathbf{a}$ in the MSE sense.

The MMSE transformation in this case can be determined from Theorem 10.1, where now $\mathbf{D}^{1/2}\mathbf{V}^* = \mathbf{I}_m$ so that $\mathbf{U} = \mathbf{Q}$ and $\mathbf{Z} = \mathbf{I}_m$. Then

$$\widehat{\mathbf{T}} = \alpha \mathbf{Q}\Lambda^{1/2}\mathbf{Q}^* = \alpha \mathbf{R}^{1/2}, \tag{10.17}$$

where if c is fixed then $\alpha = c$, and if c is chosen to minimize the MSE then $\alpha = \hat{c}$ with

$$\hat{c} = \frac{\mathrm{Tr}(\mathbf{R}^{1/2})}{\mathrm{Tr}(\mathbf{R})}. \tag{10.18}$$

The MMSE unwhitening transformation will be used in the next chapter to derive the covariance shaping LS estimator.

10.2.4 MMSE Subspace Unwhitening

We may also consider subspace unwhitening in which $\mathbf{a} \in \mathbb{C}^m$ is a random vector that is white on an n-dimensional subspace $\mathcal{V} \subseteq \mathbb{C}^m$ so that $\mathbf{C}_a = P_{\mathcal{V}} = \mathbf{V}\tilde{\mathbf{I}}_n\mathbf{V}^*$, where $\mathbf{V}$ is a unitary matrix whose first n columns span $\mathcal{V}$. We want to "unwhiten" $\mathbf{a}$ on $\mathcal{V}$ to obtain the vector $\mathbf{b} = \mathbf{T}\mathbf{a}$ where the covariance of $\mathbf{b}$ is $\mathbf{C}_b = c^2\mathbf{R}$ for some $c > 0$, and covariance matrix $\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^*$ with $\mathcal{R}(\mathbf{R}) = \mathcal{N}(\mathbf{R})^{\perp} = \mathcal{V}$, and is as close as possible to $\mathbf{a}$ in the MSE sense.

The MMSE transformation in this case can be determined from Theorem 10.1 as follows. Since $\mathcal{R}(\mathbf{R}) = \mathcal{V}$, the first n columns of $\mathbf{Q}$ span $\mathcal{V}$, the remaining columns span $\mathcal{V}^{\perp}$, and the last $m-n$ diagonal elements of Λ are all equal 0. Then $\mathbf{V}^*\mathbf{Q}$ is a block diagonal matrix so that $\tilde{\mathbf{I}}_n\mathbf{V}^*\mathbf{Q}\Lambda^{1/2} = \mathbf{V}^*\mathbf{Q}\Lambda^{1/2}$. Thus $\mathbf{U} = \mathbf{V}^*\mathbf{Q}, \mathbf{Z} = \mathbf{I}_m$ and

$$\widehat{\mathbf{T}} = \alpha \mathbf{Q}\Lambda^{1/2}\mathbf{Q}^*\mathbf{V}\tilde{\mathbf{I}}_n\mathbf{V}^* = \alpha \mathbf{R}^{1/2}P_{\mathcal{V}} = \alpha \mathbf{R}^{1/2}, \tag{10.19}$$

where if c is fixed then $\alpha = c$, and if c is chosen to minimize the MSE then $\alpha = \hat{c}$ with

$$\hat{c} = \frac{\mathrm{Tr}(P_{\mathcal{V}}\mathbf{R}^{1/2})}{\mathrm{Tr}(\mathbf{R})} = \frac{\mathrm{Tr}(\mathbf{R}^{1/2})}{\mathrm{Tr}(\mathbf{R})}. \tag{10.20}$$

Comparing (10.19) and (10.20) with (10.17) and (10.18) respectively, we see that the MMSE subspace unwhitening transformation is equal to the MMSE unwhitening transformation.

10.3 Weighted Covariance Shaping Transformation

We may also consider a weighted MMSE covariance shaping problem in which the shaping $\mathbf{T}$ is chosen to minimize a weighted MSE. Thus we seek a transformation $\mathbf{T}$ such that $\mathbf{b} = \mathbf{T}\mathbf{a}$ has covariance $\mathbf{C}_b = c^2\mathbf{R}$ for some $c > 0$ and rank-r matrix $\mathbf{R}$ with $r \leq n$, and such that

$$\varepsilon^w_{\text{MSE}} = E\left((\mathbf{a}-\mathbf{b})^*\mathbf{A}(\mathbf{a}-\mathbf{b})\right), \tag{10.21}$$

is minimized, where $\mathbf{A}$ is some nonnegative definite Hermitian weighting matrix.

To determine the weighted MMSE covariance shaping transformation we note that

$$E\left((\mathbf{a}-\mathbf{b})^*\mathbf{A}(\mathbf{a}-\mathbf{b})\right) = E\left((\bar{\mathbf{a}}-\bar{\mathbf{b}})^*(\bar{\mathbf{a}}-\bar{\mathbf{b}})\right), \tag{10.22}$$

where $\bar{\mathbf{a}} = \mathbf{A}^{1/2}\mathbf{a}$ and $\bar{\mathbf{b}} = \mathbf{A}^{1/2}\mathbf{b}$. Thus we may first seek the transformation $\widehat{\overline{\mathbf{T}}}$ that minimizes the MSE between the random vector $\bar{\mathbf{a}}$ with covariance $\mathbf{C}_{\bar{a}} = \mathbf{A}^{1/2}\mathbf{C}_a\mathbf{A}^{1/2}$, and the random vector $\bar{\mathbf{b}} = \overline{\mathbf{T}}\bar{\mathbf{a}}$ with covariance $\mathbf{C}_{\bar{b}} = \overline{\mathbf{T}}\mathbf{C}_{\bar{a}}\overline{\mathbf{T}} = c^2\overline{\mathbf{R}}$, where $\overline{\mathbf{R}} = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$. The optimal $\widehat{\overline{\mathbf{T}}}$ follows immediately from Theorem 10.1. Then, the weighted MMSE covariance shaping transformation satisfies

$$\mathbf{A}^{1/2}\widehat{\mathbf{T}} = \widehat{\overline{\mathbf{T}}}\mathbf{A}^{1/2}. \tag{10.23}$$

If $\mathbf{A}$ is positive definite, then

$$\widehat{\mathbf{T}} = \mathbf{A}^{-1/2}\widehat{\overline{\mathbf{T}}}\mathbf{A}^{1/2}. \tag{10.24}$$

We now consider two special cases of (10.23). First, if $\mathbf{C}_a, \mathbf{R}$ and $\mathbf{A}$ are all positive definite, then from Theorem 10.1,

$$\begin{aligned}
\widehat{\overline{\mathbf{T}}} &= \alpha(\overline{\mathbf{R}}\mathbf{C}_{\bar{a}})^{-1/2}\overline{\mathbf{R}} \\
&= \alpha(\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}\mathbf{C}_a\mathbf{A}^{1/2})^{-1/2}\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2} \\
&= \alpha\mathbf{A}^{1/2}(\mathbf{R}\mathbf{A}\mathbf{C}_a\mathbf{A})^{-1/2}\mathbf{R}\mathbf{A}^{1/2},
\end{aligned} \tag{10.25}$$

where $\alpha = c$ if c is fixed and $\alpha = \hat{c}$ if c is chosen to minimize the MSE with

$$\hat{c} = \frac{\mathrm{Tr}\left((\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}\mathbf{C}_a\mathbf{A}^{1/2})^{1/2}\right)}{\mathrm{Tr}(\mathbf{R}\mathbf{A})} = \frac{\mathrm{Tr}\left((\mathbf{R}\mathbf{A}\mathbf{C}_a\mathbf{A})^{1/2}\right)}{\mathrm{Tr}(\mathbf{R}\mathbf{A})}. \tag{10.26}$$

The weighted MMSE covariance shaping transformation is then given by

$$\widehat{\mathbf{T}} = \alpha\mathbf{A}^{-1/2}\widehat{\overline{\mathbf{T}}}\mathbf{A}^{1/2} = \alpha(\mathbf{R}\mathbf{A}\mathbf{C}_a\mathbf{A})^{-1/2}\mathbf{R}\mathbf{A}. \tag{10.27}$$

The second case we consider is when $\mathbf{A} = {\mathbf{C}_a}^{\dagger}$ and $\mathbf{R}$ is a covariance matrix with $\mathcal{R}(\mathbf{R}) = \mathcal{R}(\mathbf{C}_a) = \mathcal{V}$. Note that this choice of weighting matrix is reminiscent of the Gauss-Markov weighting in LS estimation [43]. In this case $\mathbf{C}_{\bar{a}} = ({\mathbf{C}_a}^{1/2})^{\dagger}\mathbf{C}_a({\mathbf{C}_a}^{1/2})^{\dagger} = P_{\mathcal{V}}$, so that $\widehat{\overline{\mathbf{T}}}$ is equal to the MMSE subspace unwhitening transformation with $\overline{\mathbf{R}} = ({\mathbf{C}_a}^{1/2})^{\dagger}\mathbf{R}({\mathbf{C}_a}^{1/2})^{\dagger}$. From (10.19), $\widehat{\overline{\mathbf{T}}} = \alpha\overline{\mathbf{R}}^{1/2}$ where if c is fixed then $\alpha = c$, and if c is chosen to minimize the MSE then $\alpha = \hat{c}$ where from (10.20) and (B.1),

$$\hat{c} = \frac{\mathrm{Tr}\left((\mathbf{R}{\mathbf{C}_a}^{\dagger})^{1/2}\right)}{\mathrm{Tr}(\mathbf{R}{\mathbf{C}_a}^{\dagger})}. \tag{10.28}$$

The weighted MMSE covariance shaping transformation then satisfies

$$({\mathbf{C}_a}^{1/2})^{\dagger}\widehat{\mathbf{T}} = \alpha\widehat{\overline{\mathbf{T}}}({\mathbf{C}_a}^{1/2})^{\dagger}. \tag{10.29}$$

Thus,

$$P_{\mathcal{V}}\widehat{\mathbf{T}} = \alpha{\mathbf{C}_a}^{1/2}\widehat{\overline{\mathbf{T}}}({\mathbf{C}_a}^{1/2})^{\dagger} = \alpha(P_{\mathcal{V}}\mathbf{R}{\mathbf{C}_a}^{\dagger})^{1/2} = \alpha(\mathbf{R}{\mathbf{C}_a}^{\dagger})^{1/2}. \tag{10.30}$$

Evidently if $\mathbf{C}_a$ is not invertible, then the optimal transformation $\widehat{\mathbf{T}}$ is not unique. A possible choice is

$$\widehat{\mathbf{T}} = \alpha(\mathbf{R}{\mathbf{C}_a}^{\dagger})^{1/2}. \tag{10.31}$$

If $\mathbf{C}_a$ is invertible, then $P_{\mathcal{V}} = \mathbf{I}_m$ and

$$\widehat{\mathbf{T}} = \alpha(\mathbf{R}{\mathbf{C}_a}^{-1})^{1/2}. \tag{10.32}$$

If the scaling c is fixed, $\mathbf{C}_a$ is invertible, and $\mathbf{R} = \mathbf{I}_m$, then the WMMSE whitening transformation with $\mathbf{A} = {\mathbf{C}_a}^{-1}$ is equal to the MMSE whitening transformation; however the optimal scaling values are different in both cases. Similarly, if the scaling c is fixed, $\mathcal{R}(\mathbf{C}_a) = \mathcal{V}$, and $\mathbf{R} = P_{\mathcal{V}}$, then the WMMSE subspace whitening transformation with $\mathbf{A} = {\mathbf{C}_a}^{\dagger}$ is equal to the MMSE subspace whitening transformation.

In analogy to the LS inner product shaping problem, we may also consider the MMSE covariance shaping problem in which the eigenvectors of the desired covariance matrix $\mathbf{R}$ are specified, and the eigenvalues are chosen to minimize the MSE. The MMSE covariance shaping transformation in this case can then be determined by exploiting the equivalence between the MMSE covariance shaping problem and the LS inner product shaping problem and relying on results obtained in that context.

In the next chapter we consider an application of MMSE covariance shaping to the problem of estimating the unknown deterministic parameters in a linear model. Based on the concept of MMSE covariance shaping we propose a new linear estimator and show that it many cases the MSE of the proposed estimator is lower than the MSE of the conventional LS estimator.

Chapter 11

Covariance Shaping Least-Squares Estimation

The essential idea underlying optimal QSP measurements is to construct optimal measurements or algorithms subject to an inner product constraint, borrowing from the ideas of quantum detection. A stochastic analogue of this idea is to construct optimal algorithms subject to a covariance constraint. In particular, in MMSE covariance shaping developed in the previous chapter, a random vector is constructed to minimize an MSE criterion, subject to a constraint on the covariance of the constructed vector. In this chapter we further exploit this fundamental concept inspired by the quantum detection problem, to develop a new linear estimator for the unknown parameters in a linear model, where we choose the estimator to minimize an MSE criterion, subject to a constraint on the covariance of the estimator. We refer to this estimator as the covariance shaping LS (CSLS) estimator.

The CSLS estimator is a biased estimator directed at improving the performance of the traditional LS estimator at low to moderate SNR by choosing the estimate to minimize the (weighted) total error variance in the observations subject to a constraint on the covariance of the estimation error, so that we control the dynamic range and spectral shape of the covariance of the estimation error.

We develop two equivalent representations of the CSLS estimator. In the first, the CSLS estimator is expressed as a LS estimator followed by a weighted MMSE (WMMSE) covariance shaping transformation, that optimally shapes the covariance of the LS estimate. In the second, the CSLS estimator is expressed as an MF estimator followed by an MMSE

covariance shaping transformation, that optimally shapes the covariance of the MF estimate.

Analysis of the MSE of both the CSLS estimator and the LS estimator demonstrates that there is a threshold SNR, such that for SNR values lower than this threshold the MSE of the CSLS estimator is lower than the MSE of the LS estimator, for all values of the unknown parameters.

As we show, some of the well-known modifications of the LS estimator can be formulated as CSLS estimators. This allows us to interpret these estimators as the estimators that minimize the total error variance in the observations, from all linear estimators with the same covariance.

11.1 Least-Squares Estimation

A generic estimation problem that has been studied extensively in the literature is that of estimating the unknown deterministic parameters $\mathbf{x}$ in the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{11.1}$$

where $\mathbf{H}$ is a known $n \times m$ matrix, and $\mathbf{w}$ is a zero-mean random vector with covariance $\mathbf{C}_w$. For simplicity of exposition we assume that $\text{rank}(\mathbf{H}) = m$; the results extend in a straightforward way to the case in which $\text{rank}(\mathbf{H}) < m$. An important special case of a non full-rank model is considered in Section 11.8.

A common approach to estimating the parameters $\mathbf{x}$ is to restrict the estimator to be linear in the data $\mathbf{y}$, and then find the linear estimate of $\mathbf{x}$ that results in an estimated data vector $\hat{\mathbf{y}}$ that is as close as possible to the given data vector $\mathbf{y}$ in a (weighted) LS sense, so that $\hat{\mathbf{y}}$ is chosen to minimize the total squared error in the observations. The Gauss-Markov theorem [43] states that the weighting matrix that leads to an unbiased estimator of $\mathbf{x}$ with minimum variance is $\mathbf{C}_w^{-1}$. Thus, the LS estimate $\hat{\mathbf{x}}_{\text{LS}} = \mathbf{G}\mathbf{y}$ is chosen to minimize the total squared error

$$\varepsilon_{\text{LS}} = (\mathbf{y} - \mathbf{H}\mathbf{G}\mathbf{y})^* \mathbf{C}_w^{-1} (\mathbf{y} - \mathbf{H}\mathbf{G}\mathbf{y}), \tag{11.2}$$

and is given by

$$\hat{\mathbf{x}}_{\text{LS}} = (\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{y}. \tag{11.3}$$

The LS method is widely employed in diverse fields, both as an estimation criterion and as a method for parametric modeling of data (see *e.g.,* [42, 43, 44, 45]). Numerous extensions of the LS method have been previously proposed in the literature. The Total LS method, first proposed by Golub and Van Loan in [156] (see also [157]), assumes that the model matrix $\mathbf{H}$ may not be known exactly and seeks the parameters $\mathbf{x}$ and the minimum perturbation to the model matrix that minimize the LS error. The Extended LS method proposed by Yeredor in [158] seeks the parameters and some presumed underlying data that together minimize a weighted combination of model errors and measurement errors. In both of these extensions it is assumed that the data model (11.1) does not hold perfectly, either due to errors in $\mathbf{H}$ or errors in the data $\mathbf{y}$.

In our method we assume that the data model holds *i.e.,* $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ with $\mathbf{H}$ and $\mathbf{y}$ known exactly, and our objective is to minimize the error between $\mathbf{x}$ and the estimate of $\mathbf{x}$. In many cases the data vector $\mathbf{y}$ is not very sensitive to changes in $\mathbf{x}$, so that a large error in estimating $\mathbf{x}$ may translate into a small error in estimating the data vector $\mathbf{y}$, in which case the LS estimate may result in a poor estimate of $\mathbf{x}$. This effect is especially predominant at low to moderate SNR, where the data vector $\mathbf{y}$ is typically affected more by the noise than by changes in $\mathbf{x}$; the exact SNR range will depend on the properties of the model matrix $\mathbf{H}$. A difficulty often encountered in this estimation problem is that the error in the estimation can have a covariance structure with a very high dynamic range.

Various modifications of the LS estimator for the case in which the model (11.1) is assumed to hold perfectly have been proposed [159]. In [160], Stein showed that the LS estimator for the mean vector in a multivariate Gaussian distribution with dimension greater than 2 is inadmissible, *i.e.,* for certain parameter values, other estimators exist with lower MSE. An explicit (nonlinear) estimator with this property, referred to as the James-Stein estimator, was later proposed and analyzed in [161]. This work appears to have been the starting point for the study of alternatives to LS estimators. Among the more prominent alternatives are the ridge estimator [46] (also known as Tikhonov regularization [47]) and the shrunken estimator [48].

To improve the performance over the LS estimator at low to moderate SNR we choose the estimator of $\mathbf{x}$ to minimize the total error variance in the observations $\mathbf{y}$, subject to a constraint on the covariance of the error in the estimate of $\mathbf{x}$, so that we control the dynamic range and spectral shape of the covariance of the estimation error. As we will show, in many cases the CSLS estimator can reduce the MSE of the estimator by allowing for a bias, where the MSE of an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ is defined by

$$\mathrm{MSE}(\hat{\mathbf{x}}) = E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) = \mathrm{Tr}\left(E((\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^*)\right). \tag{11.4}$$

In Section 11.7 we show that both the ridge estimator and the shrunken estimator can be formulated as CSLS estimators.

11.2 The Covariance Shaping Least-Squares Estimator

The CSLS estimate of $\mathbf{x}$, denoted $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, is chosen to minimize the total variance of the weighted error between $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}_{\mathrm{CSLS}} = \mathbf{HGy}$ and $\mathbf{y}$, subject to the constraint that the covariance of the error in the estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ is proportional to a given covariance matrix $\mathbf{R}$. From (11.1) it follows that the covariance of $\mathbf{y}$ is equal to $\mathbf{C}_w$, so that the covariance of $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, which is equal to the covariance of the error in the estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, is given by $\mathbf{GC}_w\mathbf{G}^*$. Thus, $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \mathbf{Gy}$ is chosen to minimize

$$\varepsilon_{\mathrm{CSLS}} = E\left((\mathbf{y}' - \mathbf{HGy}')^*\mathbf{C}_w^{-1}(\mathbf{y}' - \mathbf{HGy}')\right), \tag{11.5}$$

subject to

$$\mathbf{GC}_w\mathbf{G}^* = c^2\mathbf{R}, \tag{11.6}$$

where $\mathbf{y}' = \mathbf{y} - E(\mathbf{y})$, $\mathbf{R}$ is a given covariance matrix, and $c > 0$ is a constant that is either specified, or chosen to minimize the error (11.5).

The minimization problem of (11.5) and (11.6) is a special case of the general WMMSE shaping problem, considered in Section 10.3. Specifically, with $\mathbf{a} = \mathbf{y}'$, $\mathbf{C}_a = \mathbf{C}_w$ and $\mathbf{T} = \mathbf{HG}$, the problem of (11.5) and (11.6) is equivalent to the problem of finding $\mathbf{T}$ to

minimize

$$E\left((\mathbf{a}-\mathbf{b})^*\mathbf{C}_a^{-1}(\mathbf{a}-\mathbf{b})\right), \tag{11.7}$$

where $\mathbf{b}=\mathbf{T}\mathbf{a}$, subject to

$$\mathbf{C}_b=\mathbf{T}\mathbf{C}_a\mathbf{T}^*=c^2\mathbf{Q}, \tag{11.8}$$

with $\mathbf{Q}=\mathbf{H}\mathbf{R}\mathbf{H}^*$. This problem is equivalent to the WMMSE shaping problem of (10.21) with weighting $\mathbf{A}=\mathbf{C}_a^{-1}$.

Denoting by $\widetilde{\mathbf{G}}=(1/c)\mathbf{G}$ we then have from (10.32) that the optimal value of $\widetilde{\mathbf{G}}$ satisfies

$$\mathbf{H}\widehat{\widetilde{\mathbf{G}}}=(\mathbf{Q}\mathbf{C}_w^{-1})^{1/2}=(\mathbf{H}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1})^{1/2}. \tag{11.9}$$

Multiplying both sides by $(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}$,

$$\begin{aligned}\widehat{\widetilde{\mathbf{G}}} &= (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}(\mathbf{H}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1})^{1/2}\\ &= (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\\ &= \mathbf{R}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R})^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1},\end{aligned} \tag{11.10}$$

where we used (B.3).

Note that from Theorem 8.3, the columns of $\mathbf{C}_w^{1/2}\widehat{\widetilde{\mathbf{G}}}^*=\mathbf{C}_w^{-1/2}\mathbf{H}\mathbf{R}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R})^{-1/2}$ are the closest vectors with Gram matrix $\mathbf{R}$ to the columns of $\mathbf{C}_w^{-1/2}\mathbf{H}$, in a LS sense.

If the scaling c in (11.6) is specified, then the CSLS estimator is given by $\hat{\mathbf{x}}_{\mathrm{CSLS}}=c\widehat{\widetilde{\mathbf{G}}}\mathbf{y}$. If c is chosen to minimize $\varepsilon_{\mathrm{CSLS}}$, then $\hat{\mathbf{x}}_{\mathrm{CSLS}}=\hat{c}\widehat{\widetilde{\mathbf{G}}}\mathbf{y}$, where from (10.28),

$$\hat{c}=\frac{\mathrm{Tr}((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})}{\mathrm{Tr}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})}. \tag{11.11}$$

In general $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ is a biased estimator of $\mathbf{x}$, so that when $\sigma^2\to 0$, $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ does not converge to $\mathbf{x}$. At high SNR we therefore expect the LS estimator to perform better than the CSLS estimator. The advantage of the CSLS estimator is at low to moderate SNR, where we reduce the MSE of the estimator by allowing for a biased estimator. Indeed, as we show in Section 11.3, in the case in which $\mathbf{R}=\mathbf{I}_m$, there is always a threshold SNR, below which

the CSLS estimator yields a lower MSE than the LS estimator, for all values of $\mathbf{x}$. Some examples demonstrating the values of these thresholds are presented in Section 11.4. As we show in Sections 11.4 and 11.9, in applications this threshold value can be pretty large.

Since the covariance of the LS estimate is given from (11.3) by $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$ and the covariance of the CSLS estimate is proportional to $c^2\mathbf{R}$, it is immediate that $\hat{\mathbf{x}}_{\text{CSLS}}$ can be equal to $\hat{\mathbf{x}}_{\text{LS}}$ only if $\mathbf{R}$ is chosen to be proportional to $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$. In fact, using the CSLS estimator with optimal scaling we have that $\hat{\mathbf{x}}_{\text{CSLS}} = \hat{\mathbf{x}}_{\text{LS}}$ if and only if $\mathbf{R} = d^2(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$ for some $d > 0$. Then $\hat{\mathbf{x}}_{\text{LS}} = (1/d^2)\mathbf{R}\mathbf{H}^*\mathbf{C}_w\mathbf{y}$, and $\hat{\mathbf{x}}_{\text{CSLS}} = (\hat{c}/d)\mathbf{R}\mathbf{H}^*\mathbf{C}_w\mathbf{y}$. From (11.11), $\hat{c} = d/d^2 = 1/d$, so that for any choice of d, $\hat{\mathbf{x}}_{\text{CSLS}} = \hat{\mathbf{x}}_{\text{LS}}$.

Thus, the CSLS estimator offers a new interpretation of the LS estimator as the estimator that minimizes the error variance in the observations, from all estimators with covariance proportional to $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$.

Finally, we note that the CSLS estimator with optimal scaling is invariant to an overall gain in $\mathbf{C}_w$. Thus if $\mathbf{C}_w = \sigma^2\mathbf{C}$ for some covariance matrix $\mathbf{C}$, then the CSLS estimator does not depend on σ. This property does not hold in the case in which c is chosen as a constant, independent of σ. In this case the CSLS estimator depends explicitly on σ which therefore must be known.

The CSLS estimator is summarized in the following theorem:

Theorem 11.1 (CSLS estimator). *Let $\mathbf{x}$ denote the deterministic unknown parameters in the model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, where $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{w}$ is a zero-mean random vector with covariance $\mathbf{C}_w$. Let $\hat{\mathbf{x}}_{\text{CSLS}}$ denote the covariance shaping least-squares estimator of $\mathbf{x}$ that minimizes the error (11.5) subject to (11.6), for some $c > 0$. Then*

$$\hat{\mathbf{x}}_{\text{CSLS}} = \beta\mathbf{R}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R})^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y} = \beta(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y},$$

where

1. *if c is specified then $\beta = c$;*

2. *if c minimizes the error then $\beta = \hat{c}$ where $\hat{c} = Tr((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})/\,Tr(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})$.*

In addition,

1. *The columns of* $\mathbf{C}_w^{1/2}\widehat{\widetilde{\mathbf{G}}}^*$ *are the closest vectors with Gram matrix* $\mathbf{R}$ *to the columns of* $\mathbf{C}_w^{-1/2}\mathbf{H}$*, in a least-squares sense;*

2. *With* $\beta = \hat{c}$*,* $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ *is equal to the least-squares estimate* $\hat{\mathbf{x}}_{\mathrm{LS}}$ *if and only if* $\mathbf{R} = d^2\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$ *for some* $d > 0$*.*

11.3 Performance Analysis of the CSLS Estimator

To compare the performance of the LS and CSLS estimators, we evaluate the MSE of the estimators, where the MSE of an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ is given by (11.4). In the analysis we assume that the scaling of the CSLS estimator is chosen as $\beta = \hat{c}$ given by (11.11), unless explicitly stated otherwise.

From (11.3) and (11.1),

$$\hat{\mathbf{x}}_{\mathrm{LS}} - \mathbf{x} = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{w}, \tag{11.12}$$

so that

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}}) = \mathrm{Tr}((\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}). \tag{11.13}$$

From Theorem 11.1 with $\beta = \hat{c}$,

$$\hat{\mathbf{x}}_{\mathrm{CSLS}} - \mathbf{x} = (\hat{c}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2} - \mathbf{I}_m)\mathbf{x} + \hat{c}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{w}, \tag{11.14}$$

where $\hat{c}$ is given by (11.11). So,

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) &= \|(\hat{c}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + \hat{c}^2\mathrm{Tr}\left((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2}\right) \\ &= \|(\hat{c}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + \hat{c}^2\mathrm{Tr}(\mathbf{R}). \end{aligned} \tag{11.15}$$

We now compare the performance of the LS and CSLS estimators in the special case in which $\mathbf{R} = \mathbf{I}_m$ and $\mathbf{C}_w = \sigma^2\mathbf{C}$, where the diagonal elements of $\mathbf{C}$ are all equal to 1, so that the variance of each of the noise components of $\mathbf{C}_w$ is σ^2. With $\mathbf{B} = \mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$ and

$\lambda_i, 1 \leq i \leq m$ denoting the eigenvalues of $\mathbf{B}$, from (11.13)

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}}) = \sigma^2 \mathrm{Tr}(\mathbf{B}^{-1}) = \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1}, \tag{11.16}$$

and from (11.15),

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) = \|(\alpha \mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + m\alpha^2\sigma^2, \tag{11.17}$$

where,

$$\alpha = \frac{\hat{c}}{\sigma} = \frac{\mathrm{Tr}(\mathbf{B}^{1/2})}{\mathrm{Tr}(\mathbf{B})} = \frac{\sum_{i=1}^{m} \lambda_i^{1/2}}{\sum_{i=1}^{m} \lambda_i}. \tag{11.18}$$

The first term in (11.17) is the squared norm of the bias of the estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, and the second term in (11.17) is the total variance of $\hat{\mathbf{x}}_{\mathrm{CSLS}}$.

For large values of σ^2, the first term in (11.17) is negligible and $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \approx m\alpha^2\sigma^2$. Thus, at sufficiently low SNR both $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ and $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}})$ are proportional to σ^2 and, as we show in Appendix D, the proportionality constant $m\alpha^2$ of the CSLS estimator is smaller then the proportionality constant $\sum_i \lambda_i^{-1}$ of the LS estimator. At sufficiently high SNR, the second term in (11.17) can be considered negligible and as $\sigma \to 0$, $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}})$ converges to the constant $\|(\alpha \mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2$. These trends in the behavior of the MSE can be seen in the simulations in Section 11.9. From this qualitative analysis it is clear that there is a threshold SNR that will depend in general on $\mathbf{x}$, below which the CSLS estimator outperforms the LS estimator.

From (11.17) we have that,

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \|\alpha \mathbf{B}^{1/2} - \mathbf{I}_m\|^2 \|\mathbf{x}\|^2 + m\alpha^2\sigma^2 = |\alpha\lambda_\gamma^{1/2} - 1|^2 \|\mathbf{x}\|^2 + m\alpha^2\sigma^2, \tag{11.19}$$

where $\|\mathbf{A}\|$ denotes the spectral norm [142] of the matrix $\mathbf{A}$ defined by $\|\mathbf{A}\| = \max \sigma_i^{1/2}$ with $\{\sigma_i\}$ denoting the eigenvalues of $\mathbf{A}^*\mathbf{A}$, and

$$\gamma = \arg\max |\alpha\lambda_i^{1/2} - 1|^2, \tag{11.20}$$

with α given by (11.18). We have equality in (11.19) only in the event in which $\mathbf{x}$ is in the

direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_γ.

We note that if σ is known, then we can choose the scaling of the CSLS estimator as $\beta = c$ where c is a predetermined constant independent of σ. In this case (11.19) becomes

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq |(c/\sigma)\lambda_\gamma^{1/2} - 1|^2 \|\mathbf{x}\|^2 + mc^2. \tag{11.21}$$

In the low SNR limit in which σ is large, this estimator has the desirable property that its MSE is bounded. By contrast, the MSE of both the LS estimator and the CSLS estimator with optimal scaling is proportional to σ^2 and will therefore increase without bound with decreasing SNR.

Let $\zeta = \|\mathbf{x}\|^2/(\sigma^2 m)$ denote the SNR per dimension. Then combining (11.16) and (11.19) we have that $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$|\alpha\lambda_\gamma^{1/2} - 1|^2 \zeta + \alpha^2 \leq \frac{1}{m}\sum_{i=1}^{m} \lambda_i^{-1}. \tag{11.22}$$

The expression $|\alpha\lambda_\gamma^{1/2} - 1|$ is equal to zero only in the case in which $\lambda_i^{1/2} = 1/\alpha$ for all i, so that $\mathbf{B} = (1/\alpha^2)\mathbf{I}_m$. From Theorem 11.1 it then follows that $|\alpha\lambda_\gamma^{1/2} - 1| = 0$ if and only if $\hat{\mathbf{x}}_{\mathrm{LS}} = \hat{\mathbf{x}}_{\mathrm{CSLS}}$. If $\hat{\mathbf{x}}_{\mathrm{LS}} \neq \hat{\mathbf{x}}_{\mathrm{CSLS}}$, then we have that $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$\zeta \leq \frac{(1/m)\sum_{i=1}^{m} \lambda_i^{-1} - \alpha^2}{|\alpha\lambda_\gamma^{1/2} - 1|^2} \triangleq \zeta_{\mathrm{WC}}. \tag{11.23}$$

Note that ζ_{WC} given by (11.23) is a worst case bound, since it corresponds to the worst possible choice of parameters, namely when the unknown vector $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_γ. In practice the CSLS estimator with $\mathbf{R} = \mathbf{I}_m$ will outperform the LS estimator for higher values of SNR than ζ_{WC}.

In Appendix D we show that when $\hat{\mathbf{x}}_{\mathrm{CSLS}} \neq \hat{\mathbf{x}}_{\mathrm{LS}}$, $\zeta_{\mathrm{WC}} > 0$ so that there is always a range of SNR values for which $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$.

Starting from (11.17) in a similar manner we can show that

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \geq |\alpha\lambda_\kappa^{1/2} - 1|^2 \|\mathbf{x}\|^2 + m\alpha^2\sigma^2, \tag{11.24}$$

where

$$\kappa = \arg\min |\alpha\lambda_i^{1/2} - 1|^2, \tag{11.25}$$

with equality in (11.24) only if $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_κ. Thus, $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \geq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$\zeta \geq \frac{(1/m)\sum_{i=1}^{m} \lambda_i^{-1} - \alpha^2}{|\alpha\lambda_\kappa^{1/2} - 1|^2} \triangleq \zeta_{\mathrm{BC}}. \tag{11.26}$$

The performance analysis of the CSLS estimator $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ in the case in which $\mathbf{R} = \mathbf{I}_m$ and $\mathbf{C}_w = \sigma^2\mathbf{C}$ can be summarized as follows: Let $\zeta = \|\mathbf{x}\|^2/(\sigma^2 m)$ denote the SNR per dimension. Then with $\{\lambda_i, 1 \leq i \leq m\}$ denoting the eigenvalues of $\mathbf{B} = \mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$, and α given by (11.18),

1. $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta \leq \zeta_{\mathrm{WC}}$, where ζ_{WC} is the worst case bound given by (11.23);

2. $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \geq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta \geq \zeta_{\mathrm{BC}}$, where ζ_{BC} is the best case bound given by (11.26);

3. $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}})$ may be smaller or larger than $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta_{\mathrm{WC}} \leq \zeta \leq \zeta_{\mathrm{BC}}$, depending on the value of $\mathbf{x}$. Thus, the true threshold value in a particular application will be between ζ_{WC} and ζ_{BC}.

In addition,

1. if $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_γ given by (11.20), then $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta \leq \zeta_{\mathrm{WC}}$, and $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \geq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta \geq \zeta_{\mathrm{WC}}$;

2. if $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_κ given by (11.25), then $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta \leq \zeta_{\mathrm{BC}}$, and $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \geq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for $\zeta \geq \zeta_{\mathrm{BC}}$.

A possibly more realistic threshold can be derived by considering an average performance rather than a worst case or best case performance. Specifically, in deriving ζ_{WC} we assumed that $\mathbf{x}$ was in the direction of the eigenvector corresponding to the eigenvalue λ_γ. Similarly,

in deriving ζ_{BC} we assumed that $\mathbf{x}$ was in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_κ. To obtain an average threshold, we now assume that $\mathbf{x}$ has equal energy in each of the eigenvector directions. Specifically, let $\mathbf{v}_i, 1 \leq i \leq m$ denote the orthonormal eigenvectors of $\mathbf{B}$. The thresholds ζ_{WC} and ζ_{BC} were obtained by assuming that $\mathbf{x} = \|\mathbf{x}\|\mathbf{v}_\gamma$ and $\mathbf{x} = \|\mathbf{x}\|\mathbf{v}_\kappa$ respectively, where $\mathbf{v}_\gamma$ is the eigenvector corresponding to λ_γ, and $\mathbf{v}_\kappa$ is the eigenvector corresponding to λ_κ.

Now we assume that

$$\mathbf{x} = \frac{\|\mathbf{x}\|}{\sqrt{m}} \sum_{i=1}^{m} \mathbf{v}_i. \tag{11.27}$$

Then,

$$(\alpha \mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x} = \frac{\|\mathbf{x}\|}{\sqrt{m}} \sum_{i=1}^{m} (\alpha \lambda_i^{1/2} - 1)\mathbf{v}_i, \tag{11.28}$$

and

$$\|(\alpha \mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 = \frac{\|\mathbf{x}\|^2}{m} \sum_{i=1}^{m} (\alpha \lambda_i^{1/2} - 1)^2. \tag{11.29}$$

Thus, the "average" MSE (averaged over $\mathbf{x}$) using the CSLS estimator is

$$\frac{\|\mathbf{x}\|^2}{m} \sum_{i=1}^{m} (\alpha \lambda_i^{1/2} - 1)^2 + \alpha^2 \sigma^2 m, \tag{11.30}$$

and the average threshold is

$$\overline{\zeta} = \frac{\sum_{i=1}^{m} \lambda_i^{-1} - m\alpha^2}{\sum_{i=1}^{m} (\alpha \lambda_i^{1/2} - 1)^2}. \tag{11.31}$$

If $\mathbf{x}$ is given by (11.27), then for SNR values lower than $\overline{\zeta}$ the CSLS estimator will yield a lower MSE than the LS estimator.

11.4 Examples of Threshold Values

In this section we consider some examples illustrating the threshold values when $\mathbf{R} = \mathbf{I}_m$ for different matrices $\mathbf{B} = \mathbf{H}^* \mathbf{C}^{-1} \mathbf{H}$, where $\mathbf{C}_w = \sigma^2 \mathbf{C}$. These examples indicate that in a

variety of applications the threshold values are pretty large.

A. Line fitting: A popular application of LS modeling is fitting a line to given data. Specifically, suppose we are given measurements $\{y_i, 1 \leq i \leq n\}$ taken at times $\{t_i = i/n, 1 \leq i \leq n\}$ which we model as $y_i = at_i + b + w_i$ for some parameters a and b, where w_i is additive white noise. In this case $\mathbf{y}$ is the vector of components y_i, $\mathbf{x}$ is the vector of components a and b, $\mathbf{w}$ is the vector of components w_i so that $\mathbf{C} = \mathbf{I}_n$, and

$$\mathbf{H} = \begin{bmatrix} 1 & t_0 \\ 1 & t_1 \\ \vdots & \vdots \\ 1 & t_{n-1} \end{bmatrix}. \tag{11.32}$$

In Fig. 11-1 we plot ζ_{WC}, ζ_{BC} and $\overline{\zeta}$ as a function of n.

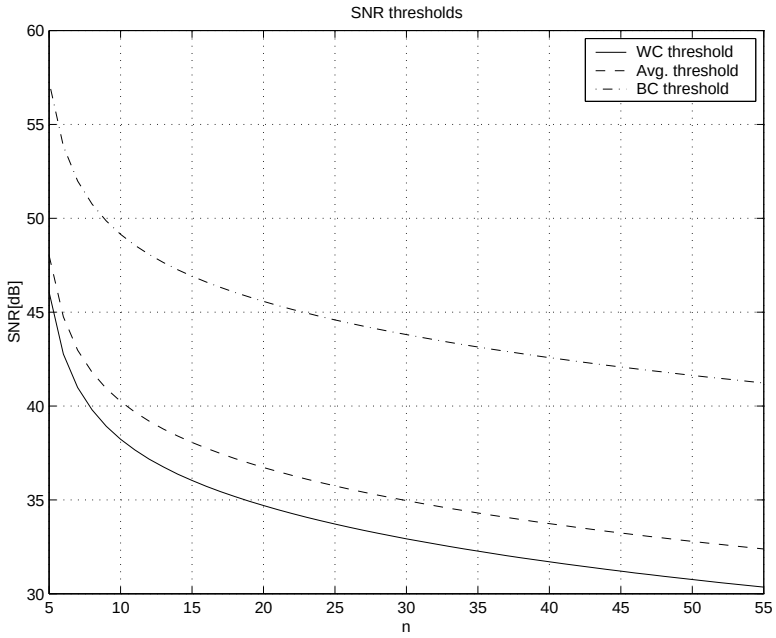


Figure 11-1: SNR worst case threshold ζ_{WC} (11.23), best case threshold ζ_{BC} (11.26), and average threshold $\overline{\zeta}$ (11.31), for line fitting with $t_i = i/n$, where n is the number of sampling points.

B. Two-dimensional case: Consider the case where there are two parameters to estimate, so that $\mathbf{B} = \mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$ is a 2×2 matrix. In this case the thresholds are a function of the eigenvalues λ_1 and λ_2 of $\mathbf{B}$. If we fix the trace of $\mathbf{B}$ to $\text{Tr}(\mathbf{B}) = a$, then $\lambda_1 + \lambda_2 = a$ and

the thresholds are a function of $\lambda = \lambda_1$.

In Fig. 11-2 we plot $\zeta_{\rm WC}$, $\zeta_{\rm BC}$ and $\overline{\zeta}$ as a function of λ with $\mathrm{Tr}(\mathbf{B}) = 1$. When $\lambda = 0.5$, $\mathbf{B} = 0.5\mathbf{I}_2$, and $\hat{\mathbf{x}}_{\rm LS} = \hat{\mathbf{x}}_{\rm CSLS}$.

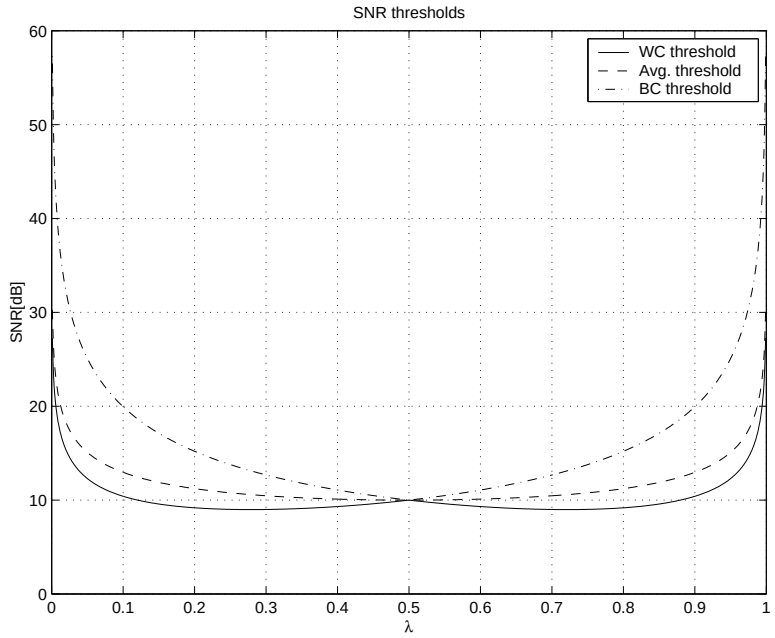


Figure 11-2: SNR worst case threshold $\zeta_{\rm WC}$ (11.23), best case threshold $\zeta_{\rm BC}$ (11.26), and average threshold $\overline{\zeta}$ (11.31) as a function of λ with $\mathrm{Tr}(\mathbf{B}) = 1$.

The behavior of the thresholds as a function of λ for the case in which $\mathrm{Tr}(\mathbf{B}) = a$ with a arbitrary is similar to that plotted in Fig. 11-2, where the thresholds increase as a decreases.

11.5 Least-Squares Estimator Followed by WMMSE Shaping

The CSLS was derived to minimize the total variance in the data error subject to a constraint on the covariance of the estimator of $\mathbf{x}$. In this section we show that the CSLS estimator can alternatively be expressed as a LS estimator $\hat{\mathbf{x}}_{\rm LS}$ followed by a WMMSE covariance shaping transformation that optimally shapes the covariance of $\hat{\mathbf{x}}_{\rm LS}$.

Specifically, suppose we estimate the parameters $\mathbf{x}$ using the LS estimator $\hat{\mathbf{x}}_{\rm LS}$. Since $\hat{\mathbf{x}}_{\rm LS} = \mathbf{x} + \tilde{\mathbf{w}}$ where $\tilde{\mathbf{w}} = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w\mathbf{w}$, the covariance of the noise component $\tilde{\mathbf{w}}$ in $\hat{\mathbf{x}}_{\rm LS}$ is equal to the covariance of $\hat{\mathbf{x}}_{\rm LS}$, $\mathbf{C}_{\hat{\mathbf{x}}_{\rm LS}} = \sigma^2(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$. To improve the performance of the LS estimator, we consider shaping the covariance of the noise component

in the estimator $\hat{\mathbf{x}}_{\rm LS}$. Thus we seek a transformation $\mathbf{T}$ such that the covariance matrix of $\hat{\mathbf{x}} = \mathbf{T}\hat{\mathbf{x}}_{\rm LS}$, denoted by $\mathbf{C}_{\hat{x}}$, satisfies

$$\mathbf{C}_{\hat{x}} = \mathbf{T}\mathbf{C}_{\hat{\mathbf{x}}_{\rm LS}}\mathbf{T}^* = c^2\mathbf{R}, \tag{11.33}$$

for some $c > 0$. To minimize the distortion to the estimator $\hat{\mathbf{x}}_{\rm LS}$, from all possible transformations $\mathbf{T}$ satisfying (11.33) we choose the one that minimizes the weighted MSE

$$E\left((\hat{\mathbf{x}}'_{\rm LS} - \mathbf{T}\hat{\mathbf{x}}'_{\rm LS})^*\mathbf{C}(\hat{\mathbf{x}}'_{\rm LS} - \mathbf{T}\hat{\mathbf{x}}'_{\rm LS})\right), \tag{11.34}$$

where $\hat{\mathbf{x}}'_{\rm LS} = \hat{\mathbf{x}}_{\rm LS} - E(\hat{\mathbf{x}}_{\rm LS})$ and $\mathbf{C}$ is an arbitrary weighting matrix.

We now show that if we choose $\mathbf{C} = \mathbf{C}^{-1}_{\hat{\mathbf{x}}_{\rm LS}}$ so that $\mathbf{T}$ minimizes

$$E\left((\hat{\mathbf{x}}'_{\rm LS} - \mathbf{T}\hat{\mathbf{x}}'_{\rm LS})^*\mathbf{C}^{-1}_{\hat{\mathbf{x}}_{\rm LS}}(\hat{\mathbf{x}}'_{\rm LS} - \mathbf{T}\hat{\mathbf{x}}'_{\rm LS})\right), \tag{11.35}$$

then the resulting estimator $\hat{\mathbf{x}} = \mathbf{T}\hat{\mathbf{x}}_{\rm LS}$ is equal to $\hat{\mathbf{x}}_{\rm CSLS}$. Note that this choice of weighting matrix is reminiscent of the Gauss-Markov weighting in LS estimation [43].

The minimization problem of (11.35) is a special case of the general WMMSE shaping problem discussed in Section 10.3 with $\mathbf{a} = \hat{\mathbf{x}}'_{\rm LS}$, $\mathbf{C}_a = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$, $\mathbf{A} = \mathbf{C}_a^{-1}$, and $\mathbf{b} = \hat{\mathbf{x}}$. Thus from (10.32),

$$\begin{aligned}\hat{\mathbf{x}} &= \hat{c}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\hat{\mathbf{x}}_{\rm LS} \\ &= \hat{c}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y},\end{aligned} \tag{11.36}$$

and from (10.28),

$$\hat{c} = \frac{\mathrm{Tr}((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})}{\mathrm{Tr}(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})}. \tag{11.37}$$

Comparing (11.36) with $\hat{\mathbf{x}}_{\rm CSLS}$ given by Theorem 11.1 we conclude that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\rm CSLS}$, so that the CSLS estimator can be determined by first finding the LS estimator $\hat{\mathbf{x}}_{\rm LS}$, and then optimally shaping its covariance.

11.6 Matched Filter Estimator Followed by MMSE Shaping

We now show that the CSLS estimator with fixed scaling can also be expressed as an MF estimator followed by MMSE shaping. Specifically, suppose we estimate the parameters $\mathbf{x}$ using a simple MF transformation $\hat{\mathbf{x}}_{\text{MF}} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}$. Then the covariance of the noise component in $\hat{\mathbf{x}}_{\text{MF}}$, which is equal to the covariance of $\hat{\mathbf{x}}_{\text{MF}}$, is $\mathbf{C}_{\hat{\mathbf{x}}_{\text{MF}}} = \sigma^2\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$. To improve the performance of $\hat{\mathbf{x}}_{\text{MF}}$ we consider shaping its covariance, so that we seek a transformation $\mathbf{T}$ such that the covariance matrix of $\hat{\mathbf{x}} = \mathbf{T}\hat{\mathbf{x}}_{\text{MF}}$, denoted by $\mathbf{C}_{\hat{x}}$, satisfies

$$\mathbf{C}_{\hat{x}} = \mathbf{T}\mathbf{C}_{\hat{\mathbf{x}}_{\text{MF}}}\mathbf{T}^* = c^2\mathbf{R}, \tag{11.38}$$

where c is given. To minimize the distortion to the estimator $\hat{\mathbf{x}}_{\text{MF}}$, from all possible transformations $\mathbf{T}$ satisfying (11.38) we choose the one that minimizes the MSE

$$E\left((\hat{\mathbf{x}}'_{\text{MF}} - \mathbf{T}\hat{\mathbf{x}}'_{\text{MF}})^*(\hat{\mathbf{x}}'_{\text{MF}} - \mathbf{T}\hat{\mathbf{x}}'_{\text{MF}})\right), \tag{11.39}$$

where $\hat{\mathbf{x}}'_{\text{MF}} = \hat{\mathbf{x}}_{\text{MF}} - E(\hat{\mathbf{x}}_{\text{MF}})$.

This minimization problem is a special case of the general MMSE shaping problem discussed in Section 10.1 with $\mathbf{a} = \hat{\mathbf{x}}'_{\text{MF}}$, $\mathbf{C}_a = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$, and $\mathbf{b} = \hat{\mathbf{x}}$. Thus from Theorem 10.1,

$$\hat{\mathbf{x}} = c(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\hat{\mathbf{x}}_{\text{MF}} = c(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}. \tag{11.40}$$

Comparing (11.40) with $\hat{\mathbf{x}}_{\text{CSLS}}$ given by Theorem 11.1 we conclude that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\text{CSLS}}$, so that the CSLS estimator with fixed scaling can be determined by first finding the MF estimator $\hat{\mathbf{x}}_{\text{MF}}$, and then optimally shaping its covariance. The optimal scaling can be found by choosing c to minimize (11.5) with $\mathbf{G} = c(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}$.

This interpretation will be useful in the next chapter when we consider applications of CSLS estimation to multiuser detection.

11.7 Connection With Other Least-Squares Modifications

We now compare the CSLS estimator with the ridge estimator proposed by Hoerl and Kennard [46], and Tikhonov [47], and with the shrunken estimator proposed by Mayer and Willke [48].

The ridge estimator for the linear model (11.1), denoted by $\hat{\mathbf{x}}_{\mathrm{R}}$, is defined by

$$\hat{\mathbf{x}}_{\mathrm{R}} = (\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H} + \delta \mathbf{T})^{-1} \mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{y}, \tag{11.41}$$

where $\mathbf{T}$ is some positive definite matrix and δ is a regularization parameter. Often $\mathbf{T}$ is chosen to be equal to $\mathbf{I}_m$. It can be shown that $\hat{\mathbf{x}}_{\mathrm{R}}$ minimizes the LS error (11.2) subject to a constraint on the norm of $\hat{\mathbf{x}}_{\mathrm{R}}$. Thus, for all estimators with fixed norm, $\hat{\mathbf{x}}_{\mathrm{R}}$ given by (11.41) minimizes the LS error, where δ is chosen to satisfy the norm constraint.

We now show that $\hat{\mathbf{x}}_{\mathrm{R}}$ is equal to a CSLS estimator with an appropriate choice of $\mathbf{R}$. Specifically, let $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ be the CSLS estimator with covariance $\mathbf{R}_{\mathrm{R}}$, where $\mathbf{R}_{\mathrm{R}}$ is the covariance of the estimate $\hat{\mathbf{x}}_{\mathrm{R}}$ and is given by

$$\mathbf{R}_{\mathrm{R}} = (\mathbf{I} + \delta (\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{T})^{-1} (\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H} + \delta \mathbf{T})^{-1}. \tag{11.42}$$

Then by direct substitution of (11.42) into the expression for $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ from Theorem 11.1, $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_{\mathrm{R}}$. Based on this connection between the ridge estimator and the CSLS estimator, we may interpret the ridge estimator as the estimator that minimizes the error $\varepsilon_{\mathrm{CSLS}}$ given by (11.5) from all estimators with covariance $\mathbf{R}_{\mathrm{R}}$.

The shrunken estimator for the linear model (11.1), denoted by $\hat{\mathbf{x}}_{\mathrm{S}}$, is a scaled version of the LS estimator and is defined by

$$\hat{\mathbf{x}}_{\mathrm{S}} = \kappa \hat{\mathbf{x}}_{\mathrm{LS}} = \kappa (\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{y}, \tag{11.43}$$

where κ is a regularization parameter. A stochastically (nonlinear) shrunken estimator is a shrunken estimator in which κ is a function of the data $\mathbf{y}$, an example of which is the well known James-Stein estimator [161].

The shrunken estimator $\hat{\mathbf{x}}_{\mathrm{S}}$ can be formulated as a CSLS estimator where the covariance of $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ is chosen to be equal to the covariance of $\hat{\mathbf{x}}_{\mathrm{S}}$ given by

$$\mathbf{R}_{\mathrm{S}} = \kappa^2 (\mathbf{H}^* \mathbf{C}_w \mathbf{H})^{-1}. \tag{11.44}$$

Substituting (11.44) into the expression for $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ from Theorem 11.1, we have indeed that $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_{\mathrm{S}}$. Thus, we may interpret $\hat{\mathbf{x}}_{\mathrm{S}}$ as the estimator that minimizes the error $\varepsilon_{\mathrm{CSLS}}$ of

(11.5) from all estimators with covariance $\mathbf{R}_S$.

In summary, some of the more popular alternatives to the LS estimator under the model (11.1) are in fact CSLS estimators. This provides additional insight and further optimality properties of these estimators. However, the CSLS estimator is more general since we are not constrained to a specific choice of covariance $\mathbf{R}$. By choosing $\mathbf{R}$ to "best" shape the estimator covariance in some sense we can improve the performance of the estimator over these LS alternatives.

As a final note, suppose we are given an arbitrary linear estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ with covariance $\mathbf{C}_x$. Then we can compute the CSLS estimate $\hat{\mathbf{x}}_{\text{CSLS}}$ with $\mathbf{R} = \mathbf{C}_x$. If $\hat{\mathbf{x}}_{\text{CSLS}} = \hat{\mathbf{x}}$, then the estimate $\hat{\mathbf{x}}$ has the additional property that from all estimators with covariance $\mathbf{C}_x$ it minimizes the (weighted) total error variance in the observations. If, on the other hand, $\hat{\mathbf{x}}_{\text{CSLS}} \neq \hat{\mathbf{x}}$, then we can always improve the total error variance of the estimate without altering its covariance by using $\hat{\mathbf{x}}_{\text{CSLS}}$. Therefore an estimate with covariance $\mathbf{C}_x$ is said to be consistent with the total error variance criterion if it minimizes this criterion from all estimators with covariance $\mathbf{C}_x$, in which case it is equal to the CSLS estimate with $\mathbf{R} = \mathbf{C}_x$.

11.8 Example of a Non Full-Rank CSLS Estimator

The CSLS estimator was derived in Section 11.2 for the case in which $\mathbf{H}$ has full rank and $\mathbf{R}$ is positive definite. Using similar techniques we can derive the CSLS estimator for the more general case in which $\mathbf{H}$ and $\mathbf{R}$ are not assumed to have full rank. Specifically, as in the full-rank case the CSLS problem of (11.5) and (11.6) is equivalent to the WMMSE covariance shaping problem of (11.7) and (11.8). Thus, as before, with $\widetilde{\mathbf{G}} = (1/c)\mathbf{G}$, the optimal value of $\widetilde{\mathbf{G}}$ must satisfy

$$\mathbf{H}\widehat{\widetilde{\mathbf{G}}} = (\mathbf{H}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1})^{1/2}. \tag{11.45}$$

Since the error (11.5) depends on $\mathbf{G}$ only through $\mathbf{H}\mathbf{G}$, $P_{\mathcal{N}(\mathbf{H})}\mathbf{G}$ does not figure in the error. Therefore, it is reasonable to choose $\mathbf{R}$ so that $\mathcal{R}(\mathbf{R}) = \mathcal{R}(\mathbf{G}) = \mathcal{N}(\mathbf{H})^\perp$. In the remainder of this section we assume that $\mathbf{R}$ is chosen to satisfy this condition.

To solve (11.45) for $\widehat{\widetilde{\mathbf{G}}}$ in this case we note that since $\mathcal{R}(\widehat{\widetilde{\mathbf{G}}}) = \mathcal{R}(\mathbf{G}) = \mathcal{N}(\mathbf{H})^\perp$, $P_{\mathcal{N}(\mathbf{H})^\perp}\widehat{\widetilde{\mathbf{G}}} = \widehat{\widetilde{\mathbf{G}}}$. Furthermore with $\mathbf{B} = \mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$, $\mathcal{N}(\mathbf{B}) = \mathcal{N}(\mathbf{H})$ so that $P_{\mathcal{N}(\mathbf{H})^\perp} = \mathbf{B}^\dagger\mathbf{B}$.

Thus,

$$\begin{aligned}\widehat{\widetilde{\mathbf{G}}} &= (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\widehat{\widetilde{\mathbf{G}}} \\ &= (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}(\mathbf{H}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1})^{1/2} \\ &= (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^\dagger\mathbf{R}\mathbf{H}^*(\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\mathbf{H}^*)^{1/2}\mathbf{C}_w^{-1}, \qquad (11.46)\end{aligned}$$

where we used (B.3). To simplify (11.46) we use the fact that $\mathcal{N}((\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\mathbf{H}^*)^{1/2}) = \mathcal{N}(\mathbf{H}\mathbf{R}\mathbf{H}^*) = \mathcal{N}(\mathbf{R}\mathbf{H}^*)$, to write

$$(\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\mathbf{H}^*)^{1/2} = (\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\mathbf{H}^*)^{1/2}P_{\mathcal{N}(\mathbf{R}\mathbf{H}^*)^\perp} = (\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\mathbf{H}^*)^{1/2}(\mathbf{R}\mathbf{H}^*)^\dagger\mathbf{R}\mathbf{H}^*. \qquad (11.47)$$

Substituting (11.47) into (11.46) and using (B.1),

$$\widehat{\widetilde{\mathbf{G}}} = (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^\dagger(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1} = ((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}. \qquad (11.48)$$

Thus

$$\hat{\mathbf{x}}_{\mathrm{CSLS}} = \beta((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}, \qquad (11.49)$$

where if c is specified then $\beta = c$ and if c is chosen to minimize the error then $\beta = \hat{c}$ given by (11.11).

Note that from Theorem 8.3, the columns of $\mathbf{C}_w^{1/2}\widehat{\widetilde{\mathbf{G}}}^*$ are the closest vectors with Gram matrix $\mathbf{R}$ to the columns of $\mathbf{C}_w^{-1/2}\mathbf{H}$, in a LS sense.

As in the full rank case, using the results of Section 10.3 it is straightforward to show that in this case the CSLS estimator can again be expressed as a LS estimator followed by a WMMSE covariance shaping transformation $\mathbf{T}$ that minimizes

$$E\left((\hat{\mathbf{x}}'_{\mathrm{LS}} - \mathbf{T}\hat{\mathbf{x}}'_{\mathrm{LS}})^*\mathbf{C}^\dagger_{\hat{\mathbf{x}}_{\mathrm{LS}}}(\hat{\mathbf{x}}'_{\mathrm{LS}} - \mathbf{T}\hat{\mathbf{x}}'_{\mathrm{LS}})\right), \qquad (11.50)$$

subject to

$$\mathbf{T}\mathbf{C}_{\hat{\mathbf{x}}_{\mathrm{LS}}}\mathbf{T}^* = c^2\mathbf{R}, \qquad (11.51)$$

for some $c > 0$, where now $\mathbf{C}_{\hat{\mathbf{x}}_{\mathrm{LS}}} = (\mathbf{H}\mathbf{C}_w^{-1}\mathbf{H})^\dagger$. Indeed, this minimization problem is a

special case of the general WMMSE shaping problem discussed in Section 10.3 with $\mathbf{a} = \hat{\mathbf{x}}'_{\text{LS}}$, $\mathbf{C}_a = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^\dagger$, $\mathbf{A} = \mathbf{C}_a^\dagger$, $\mathcal{R}(\mathbf{R}) = \mathcal{R}(\mathbf{C}_a)$, and $\mathbf{b} = \hat{\mathbf{x}}$. Thus from (10.30),

$$\widehat{\mathbf{T}} = (\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2} = ((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}, \tag{11.52}$$

so that

$$\begin{aligned}\hat{\mathbf{x}} &= \hat{c}((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\hat{\mathbf{x}}_{\text{LS}} \\ &= \hat{c}((\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{1/2})^\dagger\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y},\end{aligned} \tag{11.53}$$

where from (10.28), $\hat{c}$ is given by (11.11). Comparing (11.53) with $\hat{\mathbf{x}}_{\text{CSLS}}$ given by (11.49) we conclude that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\text{CSLS}}$.

Similarly, as in the full-rank case, the CSLS estimator can also be expressed as an MF estimator followed by an MMSE covariance shaping transformation $\mathbf{T}$ that minimizes (11.39) subject to (11.38).

11.9 Applications of CSLS Estimation

In this section we consider some applications of CSLS estimation.

11.9.1 System Identification

As a first application of CSLS estimation, we consider the problem of estimating the parameters in an ARMA model, and compare the estimated parameters to those obtained by using the modified Yule-Walker equations in combination with Shanks' method [162, 44].

Suppose we are given noisy measurements $y[l]$ of a signal $x[l]$ with z-transform

$$X(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_q z^{-q}}{1 + a_1 z^{-1} + \ldots + a_p z^{-p}} \triangleq B(z)H(z), \tag{11.54}$$

where $B(z)$ denotes the numerator polynomial, $H(z)$ denotes the inverse of the denominator polynomial, and $q < p$. The MA parameters of $x[l]$ are the coefficients b_i in (11.54), and the coefficients a_i in (11.54) are the AR parameters. The problem then is to estimate the AR and MA parameters from the data $y[0], \ldots, y[n-1]$ where $y[l] = x[l] + w[l]$, and $w[l]$

denotes a zero-mean white Gaussian noise process with variance σ^2.

Various methods exist for estimating these parameters based on different applications of LS estimation [44]. A popular method is to estimate the AR parameters using the modified Yule-Walker equations [44], and then use these estimates in combination with Shanks' method [162] to estimate the MA parameters. We use this method as a basis for comparison.

Specifically, for $l > q$ we have from (11.54) that the data $y[l]$ satisfies

$$y[l] = \sum_{i=1}^{p} a_i y[l-i] + w[l]. \tag{11.55}$$

Now, let $\mathbf{a}$ denote the vector with components $a_i, 1 \leq i \leq p$, let $\mathbf{y}$ denote the data vector with components $y[l], p \leq l \leq n-1$, let $\mathbf{w}$ denote the vector with components $w[l], p \leq l \leq n-1$ and let

$$\mathbf{H}_{\text{AR}} = \begin{bmatrix} y[p-1] & y[p-2] & \cdots & y[0] \\ y[p] & y[p-1] & \cdots & y[1] \\ \vdots & \vdots & & \vdots \\ y[n-2] & y[n-3] & \cdots & y[n-p-1] \end{bmatrix}. \tag{11.56}$$

Then $\mathbf{y} = \mathbf{H}_{\text{AR}}\mathbf{a} + \mathbf{w}$, and the LS estimator of the AR parameters is given by

$$\hat{\mathbf{a}}_{\text{LS}} = (\mathbf{H}_{\text{AR}}^* \mathbf{H}_{\text{AR}})^{-1} \mathbf{H}_{\text{AR}}^* \mathbf{y}. \tag{11.57}$$

The CSLS estimator of the AR parameters is given from Theorem 11.1 by

$$\hat{\mathbf{a}}_{\text{CSLS}} = \hat{c}(\mathbf{R}\mathbf{H}_{\text{AR}}^* \mathbf{H}_{\text{AR}})^{-1/2} \mathbf{R}\mathbf{H}_{\text{AR}}^* \mathbf{y}, \tag{11.58}$$

where $\hat{c}$ is given by (11.11).

We now use these estimates of $\mathbf{a}$ to estimate the MA parameters using Shanks' method. Specifically, let $e[l] = y[l] - h[l] * b[l]$ where $h[l]$ is the impulse response of the filter with z-transform $H(z)$, which is computed using the estimates of the AR parameters, and $b[l]$ is the (unknown) impulse response of the filter with z-transform $B(z)$. Shanks proposed estimating the unknown sequence $b[l]$ by minimizing $\sum_{l=0}^{n-1} e^2[l]$. With $\mathbf{e}$ denoting the error vector with components $e[l], 0 \leq l \leq n-1$, $\mathbf{e} = \mathbf{y} - \mathbf{H}_{M_A}\mathbf{b}$ where $\mathbf{b}$ is the vector with

components $b_i, 1 \leq i \leq q$, $\mathbf{y}$ is the data vector with components $y[l], 0 \leq l \leq n-1$, and

$$\mathbf{H}_{M_A} = \begin{bmatrix} h[0] & 0 & \cdots & 0 \\ h[1] & h[0] & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ h[n-1] & h[n-2] & \cdots & h[n-q] \end{bmatrix}, \tag{11.59}$$

so that Shanks' method reduces to a LS problem. The LS estimator of the MA parameters is then given by

$$\hat{\mathbf{b}}_{\text{LS}} = (\mathbf{H}_{M_A}^* \mathbf{H}_{M_A})^{-1} \mathbf{H}_{M_A}^* \mathbf{y}, \tag{11.60}$$

where $\mathbf{H}_{M_A}$ is computed using the the LS estimate $\hat{\mathbf{a}}_{\text{LS}}$ given by (11.57).

We can modify Shanks' estimator by using the CSLS estimator of the parameters $\mathbf{b}$, which leads to the estimator

$$\hat{\mathbf{b}}_{\text{CSLS}} = \hat{c}(\mathbf{R}\mathbf{H}_{M_A}^* \mathbf{H}_{M_A})^{-1/2} \mathbf{R}\mathbf{H}_{M_A}^* \mathbf{y}, \tag{11.61}$$

where $\hat{c}$ is given by (11.11), and now $\mathbf{H}_{M_A}$ is computed using the the CSLS estimate $\hat{\mathbf{a}}_{\text{CSLS}}$ given by (11.58).

To evaluate the performance of both estimators we consider an example in which the ARMA parameters are given by

$$a_1 = 0.9, a_2 = 0.6, a_3 = 0.4, b_0 = 1, b_2 = 0.5, \tag{11.62}$$

and the matrix $\mathbf{R}$ is chosen as $\mathbf{R} = \mathbf{I}_m$.

In Fig. 11-3 we plot the MSE in estimating the AR parameters using $\hat{\mathbf{a}}_{\text{CSLS}}$ and $\hat{\mathbf{a}}_{\text{LS}}$ for $n = 20$ averaged over 2000 noise realizations, as a function of $-10 \log \sigma^2$ where σ^2 is the noise variance. As we expect, the MSE of the CSLS estimator decreases with σ^2 for low SNR and converges to a constant in the high SNR limit. The MSE of the LS estimator decreases with σ^2 at a much slower rate. The experimental threshold is ≈ 61 dB so that for σ^2 greater than ≈ -61 dB the CSLS estimator yields a lower MSE than the LS estimator.

In Fig. 11-4 we plot the MSE in estimating the MA parameters using $\hat{\mathbf{b}}_{\text{CSLS}}$ and $\hat{\mathbf{b}}_{\text{LS}}$ for $n = 20$ averaged over 2000 noise realizations, as a function of $-10 \log \sigma^2$. The experimental

threshold is ≈ 32 dB.

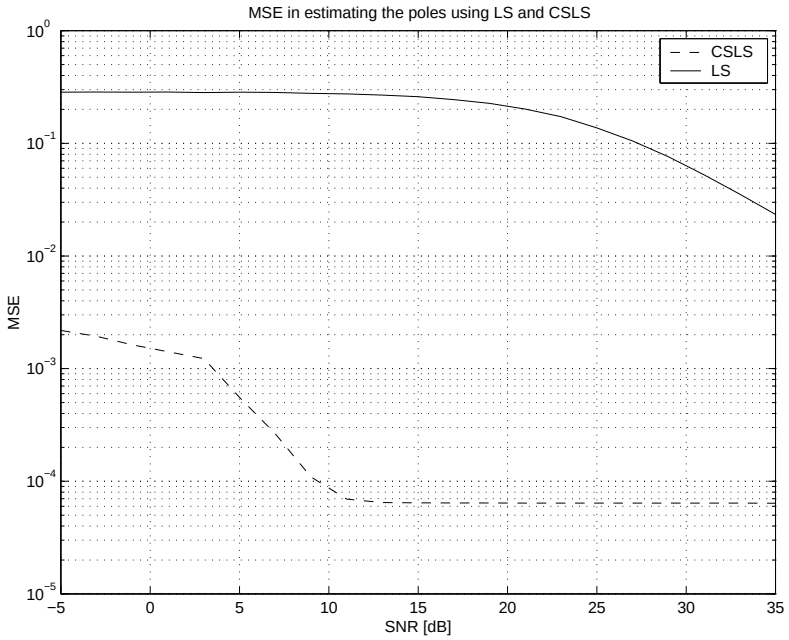


Figure 11-3: Mean-squared error in estimating the AR parameters a_i given by (11.62) using the LS estimator (11.57) and the CSLS estimator (11.58).

11.9.2 Exponential Signal Modeling

As a second application of the CSLS estimator, we consider the problem of estimating the amplitudes of two complex exponentials with known frequencies and damping factor, in complex additive white Gaussian noise. The data is thus given by

$$y[l] = a_1 e^{s_1 l} + a_2 e^{s_2 l} + w[l], \quad l = 0, 1, \ldots, n-1, \tag{11.63}$$

where $w[l]$ is a white complex Gaussian noise process with variance σ^2, and n is the number of data points.

Denoting by $\mathbf{y}$ the vector of components $y[l]$, we have that $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ where $\mathbf{x}$ is the

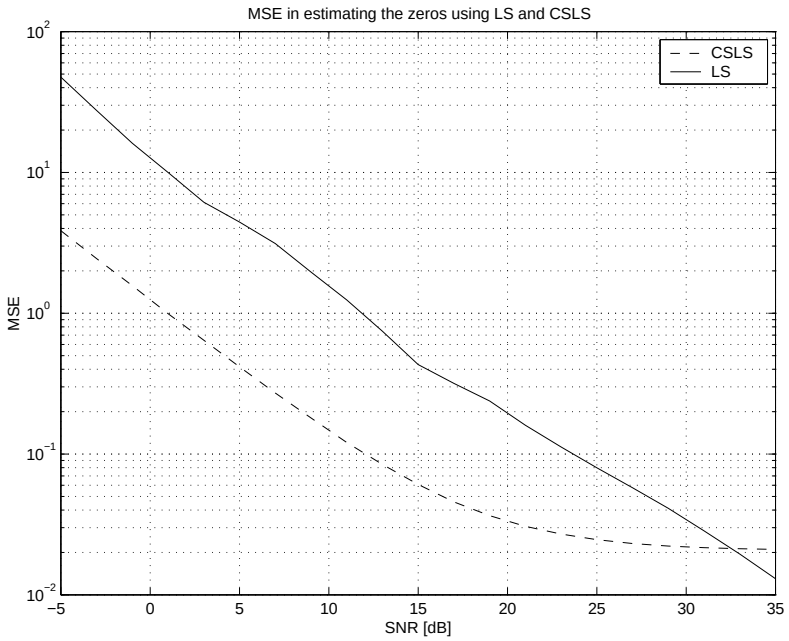


Figure 11-4: Mean-squared error in estimating the MA parameters b_i given by (11.62) based on the estimated values of the AR parameters, using the LS estimator (11.60) and the CSLS estimator (11.61).

vector of components a_1 and a_2, $\mathbf{w}$ is the vector of components $w[l]$, and

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ e^{s_1} & e^{s_2} \\ \vdots & \vdots \\ e^{s_2(n-1)} & e^{s_2(n-1)} \end{bmatrix}. \tag{11.64}$$

In Fig. 11-5 we plot the MSE in estimating the parameters a_1 and a_2 using the CSLS estimator and the LS estimator, for the case in which $s_1 = -0.6 + j2\pi(0.40)$, $s_2 = -0.6 + j2\pi(0.41)$ and $n = 15$. The true parameter values are $a_1 = a_2 = 1$. For the noise variance range shown, the CSLS estimator performs better than the LS estimator. In this example the experimental threshold variance is $\approx$ 56 dB, so that for values of σ^2 greater than ≈ -56 dB the CSLS estimator yields a lower MSE than the LS estimator.

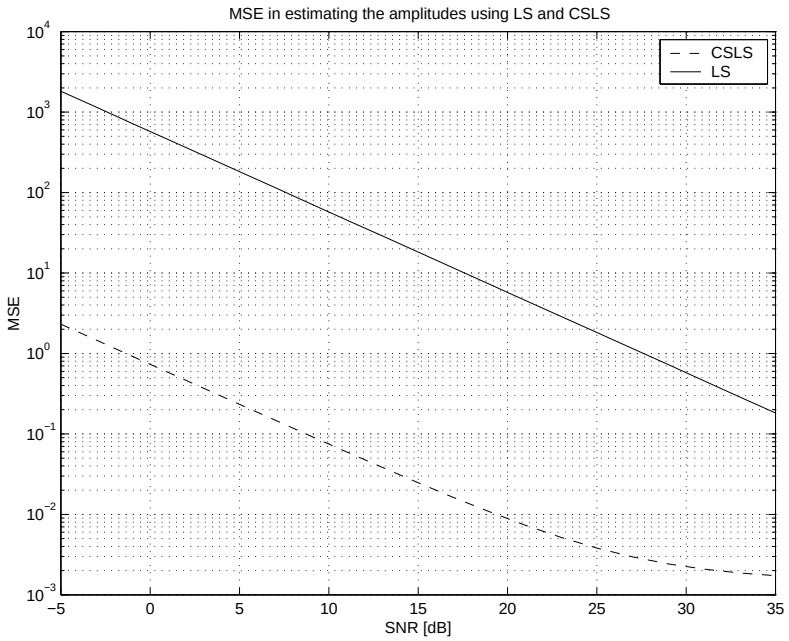


Figure 11-5: Mean-squared error in estimating the amplitudes a_1 and a_2 in the model (11.63) using the LS estimator and the CSLS estimator. The parameter values are given by $s_1 = -0.6 + j2\pi(0.40), s_2 = -0.6 + j2\pi(0.41), n = 15$ and $a_1 = a_2 = 1$.

11.9.3 Multiuser Detection

In the next chapter we use the concept of CSLS estimation to derive a new class of linear receivers for synchronous code-division multiple-access (CDMA) systems. These receivers depend only on the users' signatures and do not require knowledge of the channel parameters. Nonetheless, over a wide range of these parameters the performance of these receivers can approach the performance of the linear MMSE receiver which is the optimal linear receiver that assumes knowledge of the channel parameters and maximizes the output signal-to-interference ratio (SIR).

11.10 Summary

In this chapter we used the constraints of the quantum detection problem and the results we derived in that context to develop a new linear estimator for the unknown parameters in a linear model. We demonstrated both through simulation and analytically that this modification of the LS estimator can significantly outperform the LS estimator, particularly

over practical ranges of SNR.

In our closing remarks we note that the performance analysis of the CSLS estimator presented in Section 11.3 only considered the special case in which $\mathbf{R} = \mathbf{I}_m$. It would also be valuable to analyze the MSE of the CSLS estimator for other choices of $\mathbf{R}$.

Another particularly important direction for future research is to develop methods for optimally choosing the desired covariance shape $\mathbf{R}$ in a specific application, based on knowledge of the model matrix $\mathbf{H}$. We have seen in Sections 11.4 and 11.9 that the choice $\mathbf{R} = \mathbf{I}_m$ leads to good performance in a variety of problems. However, preliminary simulations demonstrate that further improvement in performance can be obtained by tailoring the covariance $\mathbf{R}$ to the specific problem at hand. Some results for the case in which $\mathbf{R} \neq \mathbf{I}_m$ will be presented in the next chapter in the context of multiuser detection.

A possible direction to explore is choosing $\mathbf{R}$ to be equal to the covariance of the noise component in some optimal nonlinear estimate of the parameters $\mathbf{x}$. Then $\hat{\mathbf{x}}_{\text{CSLS}}$ will be an optimal linear estimator with the same covariance as the optimal nonlinear estimator. Another interesting direction to pursue is choosing $\mathbf{R}$ to have some particular structure, *i.e.,* fixing the eigenvectors of $\mathbf{R}$, and then finding the estimate $\hat{\mathbf{x}}_{\text{CSLS}}$ and the eigenvalues of $\mathbf{R}$ that minimize the MSE error $\varepsilon_{\text{CSLS}}$. The CSLS estimator in this case can again be determined by exploiting the equivalence between the CSLS estimation problem and the LS inner product shaping problem and then relying on results obtained in that context.

Chapter 12

Covariance Shaping Multiuser Detection

In this chapter we consider an application of CSLS estimation developed in the previous chapter, to the problem of suppressing interference in multiuser wireless communication systems. Based on the concept of CSLS estimation we propose a new class of linear multiuser receivers for synchronous code-division multiple-access (CDMA) systems. These receivers depend only on the signature vectors and do not require knowledge of the received amplitudes or the channel SNR.

Building on the properties of the CSLS estimator, we develop three equivalent representations of the receiver that are mathematically equivalent but may have different implications in terms of implementation. In the first, the receiver consists of a bank of correlators with correlating vectors that have a specified inner product structure, and are closest in a LS sense to the users' signature vectors. In the second, the receiver consists of a decorrelator demodulator followed by a WMMSE covariance shaping transformation that optimally shapes the noise component in the output of the decorrelator prior to detection. In the third, the receiver consists of an MF demodulator followed by an MMSE covariance shaping transformation that optimally shapes the noise component in the output of the MF.

To evaluate the performance of the receivers, we derive exact and approximate expressions for the probability of bit error. We then derive a result regarding random matrices that is used to show that for the case in which the outputs of the receiver are constrained

to be uncorrelated on a space spanned by the signature vectors, the output signal-to-interference+noise ratio (SINR) converges in the large system limit. This limit is then compared to the known SINR limits for the decorrelator, MF and linear MMSE receivers [51, 52, 53]. The analysis presented in Section 12.5 strongly suggests that in a variety of cases the CSMU receiver can outperform both the MF and the decorrelator receivers, and can approach the performance of the linear MMSE receiver, even though it does not rely on knowledge of the channel parameters.

12.1 Multiuser Detection

Multiuser receivers for detection of CDMA signals try to mitigate the effect of multiple-access interference (MAI) and background noise. These include the optimal multiuser receiver, the linear MMSE receiver, the decorrelator, and the MF receiver [49].

Both the optimal receiver and the linear MMSE receiver require knowledge of the channel parameters, namely the noise level and the received amplitudes of the users' signals. On the other hand, the MF and the decorrelator receivers are linear receivers that only require knowledge of the signature vectors. The MF optimally compensates for the white noise, but does not exploit the structure of the MAI; the decorrelator optimally rejects the MAI, but does not consider the white noise. Like the MF and the decorrelator, the receivers we develop in this chapter do not require knowledge of the channel parameters and rely only on knowledge of the signature vectors. However, in contrast to the MF and the decorrelator, these receivers take both the background noise and the MAI into account.

Consider an m-user white Gaussian synchronous CDMA system where each user transmits information by modulating a signature sequence. The discrete-time model for the received signal $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{S}\mathbf{A}\mathbf{b} + \mathbf{w}, \tag{12.1}$$

where $\mathbf{S}$ is the $n \times m$ matrix of signatures $\mathbf{s}_i$ with $\mathbf{s}_i \in \mathbb{C}^n$ being the signature vector of the ith user, $\mathbf{A} = \text{diag}(A_1, \ldots, A_m)$ is the matrix of received amplitudes with $A_i > 0$ being the amplitude of the ith user's signal, $\mathbf{b}$ is the data vector of components $b_i \in \{1, -1\}$ with b_i being the ith user's transmitted symbol, and $\mathbf{w}$ is a noise vector whose elements are independent $\mathcal{CN}(0, \sigma^2)$. We assume that all data vectors are equally likely with covariance

$\mathbf{I}_m$, and that $\mathbf{s}_i^*\mathbf{s}_i = 1$ for all i.

Based on the observed signal $\mathbf{y}$, we design a receiver to detect the information transmitted by each user. The receiver consists of two parts, the signal demodulator and the detector. We restrict our attention to linear demodulators that do not require knowledge of the received amplitudes A_i or the noise level σ^2. The demodulator estimates the vector $\mathbf{x} = \mathbf{A}\mathbf{b}$ as $\hat{\mathbf{x}} = \mathbf{Q}^*\mathbf{y}$ for some matrix $\mathbf{Q}$. The ith user's symbol is then detected as $\hat{b}_i = \text{sgn}\,(\hat{x}_i)$ where $\hat{x}_i = \mathbf{q}_i^*\mathbf{y}$ is the ith component of $\hat{\mathbf{x}}$, and $\mathbf{q}_i$ are the columns of $\mathbf{Q}$. A receiver of this form can be implemented using a bank of correlators with correlating vectors $\mathbf{q}_i$, as depicted in Fig. 12-1.

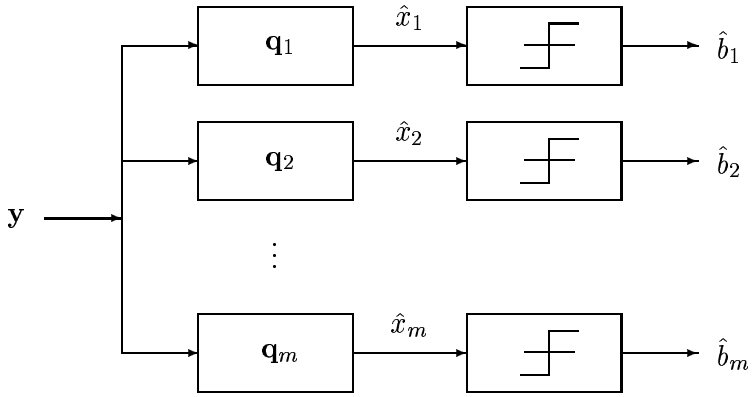


Figure 12-1: General linear receiver comprised of a bank of correlators with correlating vectors $\mathbf{q}_i$ followed by a bank of detectors.

The observed signal $\mathbf{y}$ is related to the unknown vector of parameters $\mathbf{x}$ through

$$\mathbf{y} = \mathbf{S}\mathbf{x} + \mathbf{w}, \tag{12.2}$$

which is equivalent to the linear model (11.1) considered in the previous chapter, with $\mathbf{H} = \mathbf{S}$. Therefore the design problem associated with Fig. 12-1 is equivalent to the problem of estimating $\mathbf{x}$ in the linear model (12.2), where we treat $\mathbf{x}$ as a deterministic unknown vector of parameters.

If we estimate $\mathbf{x}$ using the LS estimator, then $\hat{\mathbf{x}}_{\text{LS}} = (\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*\mathbf{y} = \mathbf{Z}^*\mathbf{y}$. The resulting receiver can be implemented using the bank of correlators of Fig. 12-1 with $\mathbf{q}_i = \mathbf{z}_i$, where the vectors $\mathbf{z}_i$ are the columns of $\mathbf{Z} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger$. This receiver is equal to the well

known decorrelator receiver, introduced by Lupas and Verdu [50]. The decorrelator optimally rejects the MAI when the signature vectors are linearly independent, but does not compensate for the white noise. Indeed, denoting by $\mathbf{a} = \hat{\mathbf{x}}_{\text{LS}}$ the output of the decorrelator demodulator, it follows from (12.1) that for linearly independent signature vectors

$$\mathbf{a} = \mathbf{Z}^*\mathbf{y} = \mathbf{A}\mathbf{b} + \mathbf{Z}^*\mathbf{w}, \tag{12.3}$$

and in the absence of noise $\hat{b}_i = \text{sgn}\,(A_i b_i) = b_i$ for all i. However, when noise is present the inverse operation of the decorrelator may enhance the white noise, resulting in degraded performance.

Alternatively, we may estimate $\mathbf{x}$ using the MF estimator, $\hat{\mathbf{x}}_{\text{MF}} = \mathbf{S}^*\mathbf{y}$. The resulting receiver can be implemented using the bank of correlators of Fig. 12-1 with $\mathbf{q}_i = \mathbf{s}_i$, which is equivalent to the single-user MF receiver. The MF receiver optimally compensates for the white noise on the channel, but it does not take the structure of the MAI into account.

12.2 The Covariance Shaping Multiuser Detector

To improve the performance of the decorrelator and MF receivers without assuming knowledge of the channel parameters, we propose estimating $\mathbf{x}$ using a CSLS estimator, which leads to a class of receivers that we define as the *covariance shaping multiuser (CSMU) receivers*. The CSMU demodulator minimizes the total error variance in the received signal subject to the constraint that the covariance of the noise component in the output of the demodulator of Fig. 12-1 is equal to $\sigma^2\mathbf{R}$, where $\mathbf{R}$ is a given covariance matrix and σ^2 is the noise variance, so that we control the dynamic range and spectral shape of the noise at the output of the demodulator. The particular shaping $\mathbf{R}$ can be tailored to the specific set of signatures. Like the MF and the decorrelator, this receiver requires knowledge of the signature vectors only.

To ensure that the corresponding correlating vectors in Fig. 12-1 lie in the space $\mathcal{U}$ spanned by the signature vectors $\mathbf{s}_i$, we choose the shaping $\mathbf{R}$ at the output of the demodulator to be such that $\mathcal{R}(\mathbf{R}) = \mathcal{N}(\mathbf{S})^{\perp} = \mathcal{V}$. In particular, if the signature vectors are linearly independent then $\mathbf{R}$ is chosen to be positive definite. With this choice of $\mathbf{R}$ the

CSLS estimator of $\mathbf{x}$ follows from Theorem 11.1 and (11.49) as

$$\hat{\mathbf{x}}_{\text{CSLS}} = c((\mathbf{RS}^*\mathbf{S})^{1/2})^\dagger \mathbf{RS}^*\mathbf{y} = c\mathbf{C}^*\mathbf{y}, \tag{12.4}$$

where $\mathbf{C} = \mathbf{SR}((\mathbf{S}^*\mathbf{SR})^{1/2})^\dagger$. Note, that the scaling of $\hat{\mathbf{x}}_{\text{CSLS}}$ will not effect the detector output and therefore can be chosen arbitrarily. In our derivation we assume that $c = 1$. Henceforth we denote $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\text{CSLS}} = \mathbf{C}^*\mathbf{y}$.

Thus, the CSMU receiver can be implemented using the bank of correlators of Fig. 12-1 with $\mathbf{q}_i = \mathbf{c}_i$, where $\mathbf{c}_i$ are the columns of $\mathbf{C}$. From Theorem 11.1 and 8.3, it follows that the vectors $\mathbf{c}_i$ are the closest vectors with Gram matrix $\mathbf{C}^*\mathbf{C} = \mathbf{R}$ to the signature vectors $\mathbf{s}_i$, in a LS sense. Therefore we may interpret the CSMU demodulator as a correlation demodulator with correlating vectors $\mathbf{c}_i$ with Gram matrix $\mathbf{R}$, that are closest in a LS sense to the signature vectors.

Since the output $\mathbf{a}$ of the decorrelator demodulator is equal to $\hat{\mathbf{x}}_{\text{LS}}$, from the discussion in Section 11.5 it follows that the CSMU receiver can equivalently be implemented as a decorrelator receiver followed by a WMMSE covariance shaping transformation $\widehat{\mathbf{T}}^w$ with weighting $\mathbf{C}_a^\dagger$, as depicted in Fig. 12-2. Here $\mathbf{C}_a$ is the covariance of the noise component in $\mathbf{a}$, and the shaping transformation $\widehat{\mathbf{T}}^w$ is designed to optimally shape[1] this covariance prior to detection.

The covariance of the noise component $\mathbf{Z}^*\mathbf{w}$ in $\mathbf{a}$ is given by

$$\mathbf{C}_a = \sigma^2\mathbf{Z}^*\mathbf{Z} = \sigma^2(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*\mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger = \sigma^2(\mathbf{S}^*\mathbf{S})^\dagger. \tag{12.5}$$

The WMMSE shaping transformation with weighting ${\mathbf{C}_a}^\dagger$, shaping $\mathbf{R}$, and constant $c = \sigma$ then follows from (10.31) as

$$\widehat{\mathbf{T}}^w = \sigma(\mathbf{R}{\mathbf{C}_a}^\dagger)^{1/2} = (\mathbf{RS}^*\mathbf{S})^{1/2}, \tag{12.6}$$

and indeed, $\widehat{\mathbf{T}}^w\mathbf{a} = ((\mathbf{RS}^*\mathbf{S})^{1/2})^\dagger\mathbf{RS}^*\mathbf{Sa} = \hat{\mathbf{x}}$.

From the representation of the CSLS estimator of Section 11.6 it follows that the CSMU

[1] In this chapter when we refer to shaping a random vector $\mathbf{a}$, we explicitly mean shaping the noise component in $\mathbf{a}$. Equivalently, this corresponds to shaping $\mathbf{a} - E(\mathbf{a}|\mathbf{b})$. Similarly, when we say that a random vector $\mathbf{a}$ has covariance $\mathbf{C}_a$ we explicitly mean that the noise component $\mathbf{a} - E(\mathbf{a}|\mathbf{b})$ in $\mathbf{a}$ has covariance $\mathbf{C}_a$.

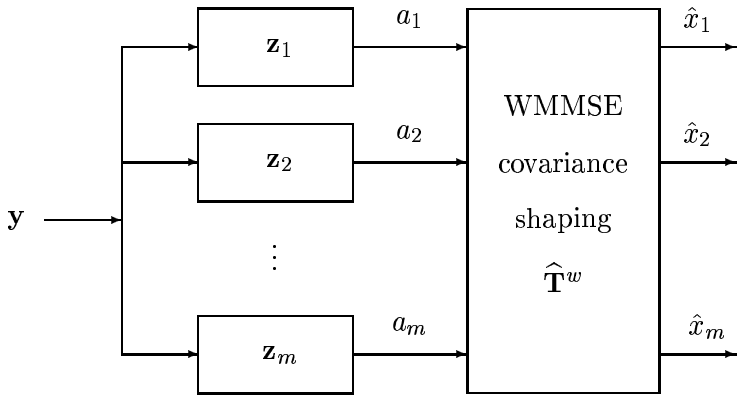


Figure 12-2: Representation of the CSMU demodulator in terms of a decorrelator demodulator followed by WMMSE covariance shaping.

receiver can also be implemented as an MF demodulator followed by an MMSE covariance shaping transformation $\widehat{\mathbf{T}}$, as depicted in Fig. 12-3. The transformation $\widehat{\mathbf{T}}$ is designed to optimally shape the covariance $\mathbf{C}_{\tilde{a}}$ of the noise component in the MF output $\tilde{\mathbf{a}} = \mathbf{S}^*\mathbf{y}$, prior to detection. Since $\mathbf{C}_{\tilde{a}} = \mathbf{S}^*\mathbf{S}$, the MMSE shaping transformation with shaping $\mathbf{R}$ and constant $c = \sigma$ is

$$\widehat{\mathbf{T}} = \sigma((\mathbf{R}\mathbf{C}_{\tilde{a}})^{1/2})^\dagger\mathbf{R} = ((\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2})^\dagger\mathbf{R}, \tag{12.7}$$

and indeed, $\widehat{\mathbf{T}}\tilde{\mathbf{a}} = \hat{\mathbf{x}}$.

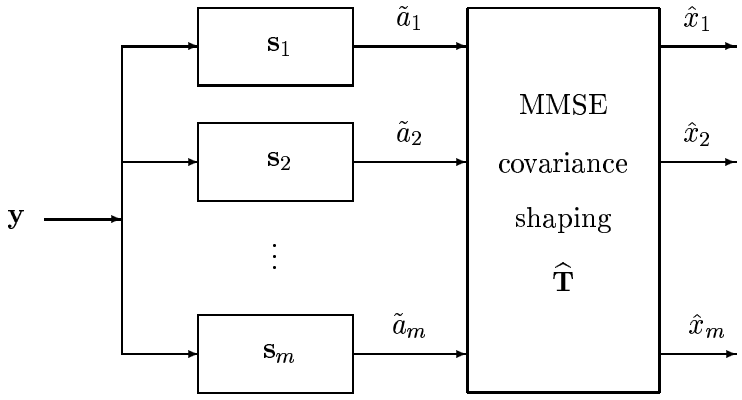


Figure 12-3: Alternative representation of the CSMU receiver in terms of an MF demodulator followed by MMSE covariance shaping.

We summarize our results regarding the CSMU demodulator in the following theorem.

Theorem 12.1 (CSMU demodulator). *Let* $\{\mathbf{s}_i, 1 \leq i \leq m\}$ *denote* m *signature vectors, and let* $\{\mathbf{c}_i, 1 \leq i \leq m\}$ *denote the correlating vectors of the CSMU demodulator. Let* $\mathbf{S}$ *and* $\mathbf{C}$ *denote the matrices of columns* $\mathbf{s}_i$ *and* $\mathbf{c}_i$*, respectively. Then*

$$\mathbf{C} = \mathbf{SR}((\mathbf{S}^*\mathbf{SR})^{1/2})^\dagger,$$

where $\mathbf{R}$ *is a nonnegative definite Hermitian matrix with* $\mathcal{R}(\mathbf{R}) = \mathcal{R}(\mathbf{S}^*)$*. In addition,*

1. *the vectors* $\mathbf{c}_i$ *are the closest vectors with Gram matrix* $\mathbf{R}$ *to the signature vectors* $\mathbf{s}_i$*, in the least-squares sense.*

2. *the CSMU demodulator can be realized by a decorrelator demodulator followed by a WMMSE covariance shaping transformation* $\widehat{\mathbf{T}}^w = (\mathbf{RS}^*\mathbf{S})^{1/2}$*;*

3. *the CSMU demodulator can be realized by an MF demodulator followed by an MMSE covariance shaping transformation* $\widehat{\mathbf{T}} = ((\mathbf{RS}^*\mathbf{S})^{1/2})^\dagger\mathbf{R}$*.*

The analysis presented in Section 12.5 strongly suggests that in a variety of cases the CSMU receiver outperforms the MF and the decorrelator receivers and approaches the performance of the linear MMSE receiver, even though the received amplitudes of the users are unknown.

12.3 The OMU and POMU Demodulators

12.3.1 The OMU Demodulator

The *orthogonal multiuser (OMU) demodulator* is a special case of the CSMU demodulator in which the signature vectors $\mathbf{s}_i$ are linearly independent and the covariance of the noise at the output of the CSMU demodulator is proportional to $\mathbf{R} = \mathbf{I}_m$. The OMU demodulator can be implemented as a correlation demodulator with correlating vectors $\mathbf{g}_i$, that are the columns of

$$\mathbf{G} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}. \tag{12.8}$$

Since $\mathbf{G}^*\mathbf{G} = \mathbf{R} = \mathbf{I}_m$, these correlating vectors are orthonormal.

From the general properties of the CSMU demodulator, it follows that the vectors $\mathbf{g}_i$ are the closest orthonormal vectors to the vectors $\mathbf{s}_i$, in a LS sense. We may therefore interpret the OMU demodulator as a correlation demodulator matched to a set of orthonormal signals that are closest in a LS sense to the signature vectors.

When implementing the OMU demodulator in the form of Fig. 12-2, the covariance shaping transformation $\widehat{\mathbf{T}}^w = \widehat{\mathbf{W}}^w$ is a whitening transformation that optimally whitens the noise component in the output of the decorrelator, prior to detection. In Section 10.3 we showed that the WMMSE whitening transformation for a random vector $\mathbf{a}$ with weighting $\mathbf{C}_a^\dagger$, is equal to the MMSE whitening transformation of $\mathbf{a}$. Therefore, the OMU demodulator can be interpreted as a decorrelator demodulator followed by an MMSE whitening transformation $\widehat{\mathbf{W}}^w = (\mathbf{S}^*\mathbf{S})^{1/2}$. Equivalently, the OMU demodulator can be interpreted as an MF demodulator followed by an MMSE whitening transformation $\widehat{\mathbf{T}} = \widehat{\mathbf{W}} = (\mathbf{S}^*\mathbf{S})^{-1/2}$.

We now show that the vectors $\mathbf{g}_i$ have the additional property that they are the closest orthonormal vectors to the decorrelator vectors $\mathbf{z}_i$, in a LS sense. From Theorem 8.1, the orthonormal vectors $\mathbf{d}_i$ closest to the vectors $\mathbf{z}_i$ are the columns of $\mathbf{D}$ where

$$\mathbf{D} = \mathbf{Z}(\mathbf{Z}^*\mathbf{Z})^{-1/2} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1}((\mathbf{S}^*\mathbf{S})^{-1})^{-1/2} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}. \tag{12.9}$$

Comparing (12.9) and (12.8), $\mathbf{D} = \mathbf{G}$, so that $\mathbf{g}_i = \mathbf{d}_i$ as claimed.

We summarize our results regarding the OMU demodulator in the following theorem.

Theorem 12.2 (OMU demodulator). *Let $\{\mathbf{s}_i, 1 \le i \le m\}$ denote m linearly independent signature vectors, and let $\{\mathbf{g}_i, 1 \le i \le m\}$ denote the correlating vectors of the OMU demodulator. Let $\mathbf{S}$ and $\mathbf{G}$ denote the matrices of columns $\mathbf{s}_i$ and $\mathbf{g}_i$, respectively. Then*

$$\mathbf{G} = \mathbf{S}(\mathbf{S}^*\mathbf{S})^{-1/2}.$$

In addition,

1. *the OMU demodulator can be realized by a decorrelator demodulator followed by an MMSE whitening transformation $\widehat{\mathbf{W}}^w = (\mathbf{S}^*\mathbf{S})^{1/2}$;*
2. *the OMU demodulator can be realized by an MF demodulator followed by an MMSE whitening transformation $\widehat{\mathbf{W}} = (\mathbf{S}^*\mathbf{S})^{-1/2}$;*
3. *the vectors $\mathbf{g}_i$ are the closest orthonormal vectors to the decorrelator vectors $\mathbf{z}_i$;*

4. the vectors $\mathbf{g}_i$ are the closest orthonormal vectors to the signature vectors $\mathbf{s}_i$.

12.3.2 The POMU Demodulator

The *projected orthogonal multiuser (POMU) demodulator* is also a special case of the CSMU demodulator in which the signature vectors $\mathbf{s}_i$ are linearly dependent and the covariance of the noise at the output of the CSMU demodulator is proportional to $\mathbf{R} = P_{\mathcal{V}}$, where $\mathcal{V} = \mathcal{N}(\mathbf{S})^{\perp}$. The POMU demodulator can be implemented as a correlation demodulator with correlating vectors $\mathbf{f}_i$, that are the columns of

$$\mathbf{F} = \mathbf{S}((\mathbf{S}^*\mathbf{S})^{1/2})^{\dagger}. \tag{12.10}$$

Since $\mathbf{F}^*\mathbf{F} = \mathbf{R} = P_{\mathcal{V}}$, these correlating vectors form a normalized tight frame for the space $\mathcal{U} = \mathcal{R}(\mathbf{S})$ spanned by the signature vectors $\mathbf{s}_i$.

From the general properties of the CSMU demodulator, it follows that the vectors $\mathbf{f}_i$ are the closest normalized tight frame vectors for $\mathcal{U}$, to the vectors $\mathbf{s}_i$ in a LS sense. We may therefore interpret the POMU demodulator as a correlation demodulator matched to a set of frame vectors for $\mathcal{U}$, that are closest in a LS sense to the signature vectors.

When implementing the POMU demodulator in the form of Fig. 12-2, the covariance shaping transformation $\widehat{\mathbf{T}}^w = \widehat{\mathbf{W}}_s^w$ is a subspace whitening transformation that optimally whitens the noise component in the output of the decorrelator on the space $\mathcal{V}$ in which it is contained, prior to detection. In analogy to the case in which $\mathbf{R} = \mathbf{I}_m$, the WMMSE subspace whitening transformation of $\mathbf{a}$ with weighting $\mathbf{C}_a^{\dagger}$ is equal to the MMSE subspace whitening transformation of $\mathbf{a}$. Therefore, the POMU demodulator can be interpreted as a decorrelator demodulator followed by an MMSE subspace whitening transformation $\widehat{\mathbf{W}}_s^w = (\mathbf{S}^*\mathbf{S})^{1/2}$.

Note, that the MMSE subspace whitening transformation $\widehat{\mathbf{W}}_s^w$ is equal to the MMSE whitening transformation $\widehat{\mathbf{W}}^w$. However, when the signature vectors are linearly independent, the vector output $\hat{\mathbf{x}} = \widehat{\mathbf{W}}^w\mathbf{a}$ is white, while for linearly dependent signature vectors the output $\hat{\mathbf{x}} = \widehat{\mathbf{W}}_s^w\mathbf{a}$ is white only on a subspace. Furthermore, for linearly dependent signature vectors the whitening transformation is not invertible, while for linearly independent signature vectors the transformation is invertible.

Finally, the POMU demodulator can also be interpreted as an MF demodulator followed

by an MMSE subspace whitening transformation $\widehat{\mathbf{T}} = \widehat{\mathbf{W}}_s = ((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger$.

As in the OMU demodulator, the vectors $\mathbf{f}_i$ have the additional property that they are the closest normalized tight frame vectors to the decorrelator vectors $\mathbf{z}_i$, in a LS sense. From Theorem 8.1, the closest normalized tight frame vectors $\mathbf{d}_i$ to the vectors $\mathbf{z}_i$ are the columns of $\mathbf{D}$ where now

$$\mathbf{D} = \mathbf{Z}((\mathbf{Z}^*\mathbf{Z})^{1/2})^\dagger = \mathbf{S}((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger. \tag{12.11}$$

Comparing (12.11) and (12.10), $\mathbf{D} = \mathbf{F}$, and indeed $\mathbf{f}_i = \mathbf{d}_i$.

We summarize our results regarding the POMU demodulator in the following theorem.

Theorem 12.3 (POMU demodulator). *Let $\{\mathbf{s}_i, 1 \le i \le m\}$ denote m linearly dependent signature vectors that span a subspace $\mathcal{U}$, and let $\{\mathbf{f}_i, 1 \le i \le m\}$ denote the correlating vectors of the POMU demodulator. Let $\mathbf{S}$ and $\mathbf{F}$ denote the matrices of columns $\mathbf{s}_i$ and $\mathbf{g}_i$, respectively. Then*

$$\mathbf{F} = \mathbf{S}((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger.$$

In addition,

1. *the POMU demodulator can be realized by a decorrelator demodulator followed by an MMSE subspace whitening transformation $\widehat{\mathbf{W}}_s^w = (\mathbf{S}^*\mathbf{S})^{1/2}$;*

2. *the POMU demodulator can be realized by an MF demodulator followed by an MMSE subspace whitening transformation $\widehat{\mathbf{W}}_s = ((\mathbf{S}^*\mathbf{S})^{1/2})^\dagger$;*

3. *the vectors $\mathbf{f}_i$ are the closest normalized tight frame vectors for $\mathcal{U}$ to the decorrelator vectors $\mathbf{z}_i$;*

4. *the vectors $\mathbf{f}_i$ are the closest normalized tight frame vectors for $\mathcal{U}$ to the signature vectors $\mathbf{s}_i$.*

Alternative derivations of the OMU and POMU receivers are developed in [38, 39].

Exploiting results we derived in the context of quantum detection [26], in the next section we show that if the signature vectors are GU, which is the case, for example, for pseudo noise (PN) sequences corresponding to maximal-length shift-register sequences [49,

163], then the OMU and POMU receivers minimize the total MAI and maximize the total output SNR, subject to the constraint that the outputs of the demodulator are uncorrelated on the appropriate space. These properties of the OMU and POMU demodulators also holds approximately for nearly orthogonal signature vectors. This provides some additional justification for this class of receivers.

12.4 OMU, POMU Demodulators and Minimizing MAI

In the implementation of Fig. 12-2, the OMU and POMU demodulators are expressed as a decorrelator demodulator followed by an MMSE whitening transformation that optimally whitens the noise component in the output of the decorrelator on the space in which it is contained. In this section we show that this whitening transformation has the additional property that among all possible whitening transformations, it minimizes the total MAI in the output of the transformation for GU signature vectors. Furthermore, for nearly orthogonal signature vectors the MMSE whitening transformation approximately minimizes the total MAI.

For simplicity of exposition, we assume throughout this section that $E(A_i^2) = 1$ for all i; the results extend in a straightforward way to the general case in which the powers $E(A_i^2)$ are not equal.

We have seen that (in the linearly independent case) the decorrelator eliminates the MAI by inverting the multiuser channel, but in the process may enhance the white noise. The OMU and POMU demodulators try and compensate for this possible noise enhancement by whitening the noise component in the output $\mathbf{a}$ of the decorrelator prior to detection. However, the whitening transformation introduces some MAI into the outputs $\hat{x}_i$. Indeed, the data component in the output $\hat{\mathbf{x}}$ of the whitening transformation is the vector $\mathbf{WAb}$, whose ith component is

$$\sum_{k=1}^{m} [\mathbf{W}]_{ik} A_k b_k = \hat{x}_i^S + \hat{x}_i^I, \tag{12.12}$$

where $\hat{x}_i^S = [\mathbf{W}]_{ii} A_i b_i$ is the signal component in $\hat{x}_i$, and $\hat{x}_i^I = \sum_{k=1, k\neq i}^{m} [\mathbf{W}]_{ik} A_k b_k$, is the MAI component in $\hat{x}_i$. We may therefore choose $\mathbf{W}$ to minimize the total MAI in the output $\hat{\mathbf{x}}$, or equivalently to maximize the the total signal-to-interference ratio (SIR) in $\hat{\mathbf{x}}$.

Thus, we seek the $\mathbf{W}$ that maximizes the total SIR given by

$$\mathrm{SIR}_T = \frac{\sum_{i=1}^m E((\hat{x}_i^S)^2)}{\sum_{i=1}^m E((\hat{x}_i^I)^2)} = \frac{\sum_{i=1}^m |[\mathbf{W}]_{ii}|^2}{\sum_{i=1}^m \sum_{k=1,k\neq i}^m |[\mathbf{W}]_{ik}|^2}, \tag{12.13}$$

subject to the whitening constraint,

$$\mathbf{W}\mathbf{C}_a\mathbf{W}^* = \sigma^2\mathbf{I}_m, \tag{12.14}$$

or the subspace whitening constraint,

$$\mathbf{W}\mathbf{C}_a\mathbf{W}^* = \sigma^2 P_{\mathcal{V}}, \tag{12.15}$$

where $\mathbf{C}_a$ is the covariance of the noise component in $\mathbf{a}$ given by (12.5), and $\mathcal{V} = \mathcal{N}(\mathbf{C}_a)^\perp$.

We first consider the case of linearly independent signature vectors. In this case, we can simplify the expression for SIR_T given by (12.13) as follows. Since $\mathbf{W}$ must be invertible, (12.14) reduces to

$$\mathbf{W}^*\mathbf{W} = \sigma^2 {\mathbf{C}_a}^{-1}. \tag{12.16}$$

From (12.16) we conclude that $\mathrm{Tr}(\mathbf{W}^*\mathbf{W}) = \sum_{i,k=1}^m |[\mathbf{W}]_{ik}|^2$ is constant, independent of the choice of $\mathbf{W}$. With $\alpha = \sum_{i,k=1}^m |[\mathbf{W}]_{ik}|^2$, we can write (12.13) as

$$\mathrm{SIR}_T = \frac{\sum_{i=1}^m |[\mathbf{W}]_{ii}|^2}{\alpha - \sum_{i=1}^m |[\mathbf{W}]_{ii}|^2}, \tag{12.17}$$

so that maximizing SIR_T subject to (12.14) is equivalent to maximizing

$$\Gamma = \sum_{i=1}^m |[\mathbf{W}]_{ii}|^2 \tag{12.18}$$

subject to this constraint.

Let $\mathbf{C}_a$ have an eigendecomposition $\mathbf{C}_a = \mathbf{V}\Lambda\mathbf{V}^*$, where Λ is a diagonal matrix with diagonal elements $\lambda_i > 0$. Then from (12.16), and using the properties of the SVD,

$$\mathbf{W} = \sigma\mathbf{U}^*\Lambda^{-1/2}\mathbf{V}^*, \tag{12.19}$$

for some unitary matrix $\mathbf{U}^*$. Let $\mathbf{u}_i$ denote the columns of $\mathbf{U}$, and let $\mathbf{t}_i$ denote the columns of $\mathbf{T} = \sigma\Lambda^{-1/2}\mathbf{V}^*$. Then $[\mathbf{W}]_{ii} = \mathbf{u}_i^*\mathbf{t}_i$, and

$$\Gamma = \sum_{i=1}^{m} |[\mathbf{W}]_{ii}|^2 = \sum_{i=1}^{m} |\mathbf{u}_i^*\mathbf{t}_i|^2. \tag{12.20}$$

Thus, the problem of maximizing (12.13) subject to (12.14) reduces to seeking a set of orthonormal vectors $\mathbf{u}_i$ that maximize (12.20).

When the signature vectors are linearly dependent, we can show that the design problem of (12.13) and (12.15) reduces to seeking a set of vectors $\mathbf{u}_i$ that form a normalized tight frame and maximize (12.20).

This problem is equivalent to the quantum detection problem discussed in Section 3.4. Specifically, comparing (12.20) with (3.13) we see that choosing a set of quantum measurement vectors to maximize the probability of correct detection in a quantum detection problem subject to an orthogonality or tight frame constraint, is equivalent to choosing a set of correlating vectors to maximize Γ subject to the corresponding constraint. From our discussion in Section 3.4 it follows that for arbitrary vectors $\mathbf{t}_i$, or equivalently arbitrary vectors $\mathbf{s}_i$, there is no known closed-form analytical expression for the vectors maximizing Γ.

Based on results we derived in a quantum detection context (see Section 3.4 and [26]), it can be shown that when the signature vectors are GU, the vectors $\mathbf{u}_i$ maximizing (12.20) are equal to the columns of the unitary matrix $\mathbf{V}^*$ in the eigendecomposition of $\mathbf{C}_a$. From (12.19) it then follows that the whitening transformation that maximizes SIR_T is given by $\mathbf{W} = \sigma\mathbf{V}\Lambda^{-1/2}\mathbf{V}^* = \sigma{\mathbf{C}_a}^{-1/2} = (\mathbf{S}^*\mathbf{S})^{1/2}$, which is equal to the MMSE whitening transformation $\widehat{\mathbf{W}}^w$ given by Theorem 12.2. Similarly, the subspace whitening transformation that maximizes SIR_T is equal to the MMSE subspace whitening $\sigma({\mathbf{C}_a}^{1/2})^\dagger = (\mathbf{S}^*\mathbf{S})^{1/2} = \widehat{\mathbf{W}}_s^w$, given by Theorem 12.3.

A common choice for signature vectors in a direct-sequence CDMA system are PN sequences corresponding to maximal-length shift-register sequences [49, 163]. These sequences have the property that the inner product between any two distinct sequences is equal to a constant. Thus, for this choice of signature vectors the OMU and POMU demodulators maximize SIR_T subject to the constraint that the outputs of the demodulator are uncorrelated on the space in which they lie.

Further results regarding the whitening or subspace whitening transformation maximizing SIR_T that follow from results pertaining to quantum detection are that if the signature vectors are nearly orthogonal, then the MMSE whitening and subspace whitening transformations also approximately maximize SIR_T [150].

Based on the results developed in the context of MF detection in Chapter 9, these whitening transformations have the additional property that for GU signature vectors they maximize the total output SNR defined by $\mathrm{SNR}_T = \sum_{i=1}^{m} \mathrm{SNR}_i$, where SNR_i is the SNR at the ith output of the whitening transformation.

12.5 Performance Analysis of the CSMU Receiver

In this section, we discuss the theoretical performance of the CSMU receiver. We first derive exact and approximate expressions for the probability of detection error for any choice of shaping $\mathbf{R}$. We then derive the asymptotic SINR of the output of the OMU and POMU receivers, corresponding to the choice $\mathbf{R} = P_{\mathcal{V}}$ where $\mathcal{V} = \mathcal{N}(\mathbf{S})^{\perp}$, in the large system limit.

12.5.1 Exact Probability of Detection Error

The detector input of the CSMU receiver defined via (12.4) is

$$\hat{\mathbf{x}} = ((\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2})^{\dagger}\mathbf{R}\mathbf{S}^*\mathbf{y} = (\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}\mathbf{A}\mathbf{b} + ((\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2})^{\dagger}\mathbf{R}\mathbf{S}^*\mathbf{w}. \tag{12.21}$$

Each component of the detector input vector can be decomposed into

$$\hat{x}_i = \hat{x}_i^S + \hat{x}_i^I + \hat{x}_i^N, \tag{12.22}$$

where the terms

$$\hat{x}_i^S = [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii} A_i b_i \tag{12.23}$$

$$\hat{x}_i^I = \sum_{k \neq i} [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ik} A_k b_k \tag{12.24}$$

$$\hat{x}_i^N = [((\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2})^{\dagger}]_i^* \mathbf{R}\mathbf{S}^*\mathbf{w} \tag{12.25}$$

represent the desired signal, the MAI, and the noise respectively. Here $[\cdot]_i$ denotes the ith column of the corresponding matrix. Conditioned on $\mathbf{b}$, the decision statistic $\hat{x}_i$ is Gaussian

with mean $\hat{x}_i^S + \hat{x}_i^I$ and variance $\sigma^2[\mathbf{R}]_{ii}$. Taking into consideration all possibilities of $\mathbf{b}$, the resulting probability of detection error for the ith user is

$$P_i(\sigma) = \frac{1}{2^{m-1}} \sum_{e_1 \in \{-1,1\}} \cdots \sum_{\substack{e_k \in \{-1,1\} \\ k \neq i}} \cdots \sum_{e_m \in \{-1,1\}} \mathcal{Q}\left(\frac{[(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii} A_i}{\sigma \sqrt{[\mathbf{R}]_{ii}}} + \sum_{k \neq i} \frac{[(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ik} A_k}{\sigma \sqrt{[\mathbf{R}]_{ii}}} e_k \right), \tag{12.26}$$

where

$$\mathcal{Q}(v) = \frac{1}{\sqrt{2\pi}} \int_v^\infty e^{-t^2/2} dt. \tag{12.27}$$

From (12.26), we see that the probability of detection error of the CSMU detector for the ith user goes to zero as $\sigma \to 0$ if and only if the argument of each of the $\mathcal{Q}$-functions is positive.

For example, in the special case in which $\mathbf{R} = P_{\mathcal{V}}$ and all cross-correlations of the signature vectors are identically equal to ρ, it can be shown that

$$[(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ik} = [(\mathbf{S}^*\mathbf{S})^{1/2}]_{ik} = \begin{cases} \alpha_1 + (m-1)\alpha_2, & \text{if } i = k, \\ \alpha_1 - \alpha_2, & \text{if } i \neq k, \end{cases} \tag{12.28}$$

where $\alpha_1 = (1/m)\sqrt{1 + (m-1)\rho}$ and $\alpha_2 = (1/m)\sqrt{1-\rho}$. In this case the probability of detection error of the CSMU detector for the ith user goes to zero as $\sigma \to 0$ when

$$(\alpha_1 + (m-1)\alpha_2) A_i > (\alpha_1 - \alpha_2) \sum_{k \neq i} A_k. \tag{12.29}$$

In the special case of m equal-energy users the condition in (12.29) simplifies to

$$\rho < \frac{3m-4}{m(m-1)}. \tag{12.30}$$

In the analysis below we choose the shaping $\mathbf{R}$ as a circulant matrix with parameter δ:

$$\mathbf{R} = \begin{bmatrix} 1 & \delta & \delta & \dots & \delta \\ \delta & 1 & \delta & \dots & \delta \\ \vdots & & \ddots & & \vdots \\ \delta & \delta & \dots & \delta & 1 \end{bmatrix}. \tag{12.31}$$

When $\delta = 0$, $\mathbf{R} = \mathbf{I}_m$ and the CSMU receiver reduces to the OMU receiver.

Figure 12-4 evaluates (12.26) in the special case of two users with cross-correlation $\rho = 0.8$ and with $\mathbf{R}$ given by (12.31) with $\delta = 0$ (*i.e.*, $\mathbf{R} = \mathbf{I}_2$), where the desired user has an SNR of 8 dB. The probability of bit error of the CSMU receiver is plotted as a function of the near-far ratio A_2/A_1, where A_1 is the amplitude of the desired user. The corresponding curves for the single-user MF, decorrelator, and linear MMSE receiver are plotted for comparison. We see that for all values of the near-far ratio shown, the CSMU receiver performs better than the decorrelator. When the power of the interferer is negligible, the MF performs better than the CSMU receiver which is expected since the MF is optimal in the absence of MAI. Thus, the CSMU receiver performs better than both the decorrelator and the MF when A_2/A_1 is roughly between 0.35 and 1. In this regime, the CSMU receiver performs similarly to the linear MMSE receiver.

In Fig. 12-5 we plot the probability of bit error for two users with cross-correlation $\rho = 0.8$ with $\mathbf{R}$ given by (12.31) as a function of the near-far ratio A_2/A_1, for several values of δ. The desired user has an SNR of 10 dB. In general we see that increasing δ results in improved performance at low near-far ratios A_2/A_1, but degraded performance at high near-far ratios. When $\rho = -0.8$ the opposite behavior is observed. The probability of error plots in this case are identical to those shown in Fig. 12-5 in reversed order, so that *e.g.*, the $\delta = 0.2$ plot in Fig. 12-5 is equivalent to the performance for $\delta = -0.2$ when $\rho = -0.8$.

In Fig. 12-6 we evaluate (12.26) for two users with cross-correlation $\rho = 0.8$ and with $\mathbf{R} = \mathbf{I}_2$, where the desired user has an SNR of 15 dB. The trends are similar to those seen in Fig. 12-4. The CSMU receiver performs better than both the decorrelator and the MF when A_2/A_1 is roughly between 0.3 and 0.75, and in most of this regime the CSMU receiver actually performs better than the linear MMSE receiver. The reason this is possible is that while the linear MMSE receiver is the linear receiver that maximizes the SINR at the slicer input, the linear MMSE receiver does not necessarily minimize the probability of detection

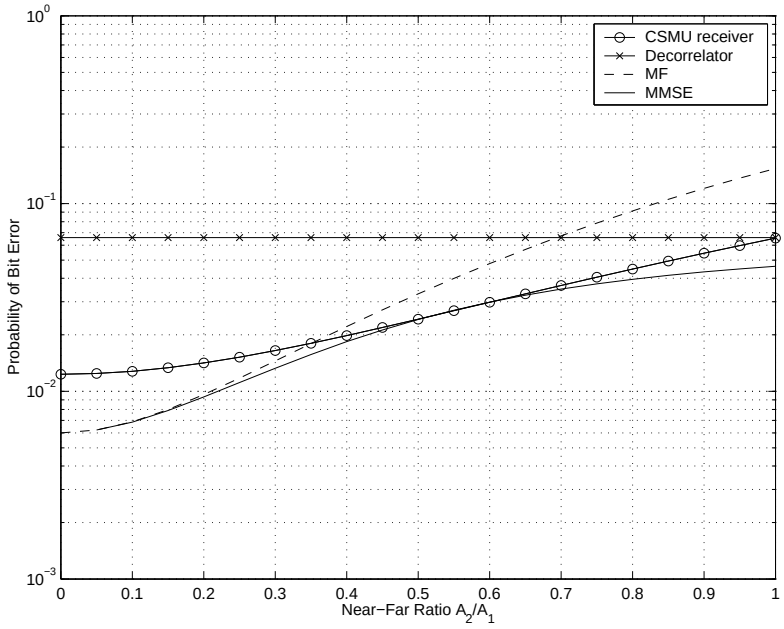


Figure 12-4: Probability of bit error with two users and cross-correlation $\rho = 0.8$, as a function of the near-far ratio A_2/A_1. In the CSMU receiver, $\mathbf{R} = \mathbf{I}_2$. The SNR of the first user, the desired user, is 8 dB.

error since the noise at the slicer is not strictly Gaussian due to the MAI.

We now look at the case of 3 users with equal cross-correlation $\rho = 0.8$ and $\mathbf{R} = \mathbf{I}_3$. Figure 12-7 depicts the bit-error rate when the first user, the desired user, has two interferers such that $A_2/A_1 = 0.5$ and $A_3/A_1 = 0.25$. The CSMU receiver performs similarly to the linear MMSE receiver at all SNR, and better than the decorrelator and the single-user MF.

In Fig. 12-8 we examine the scenario in which the desired user has 4 interferers such that $A_i/A_1 = 0.2$ for $i = 2, 3, 4, 5$, where all 5 users have equal cross-correlation $\rho = 0.8$, and $\mathbf{R} = \mathbf{I}_5$. The CSMU receiver performs significantly better than the decorrelator and the MF. Moreover, the CSMU receiver performs slightly better than the linear MMSE receiver at high SNR. Fig. 12-9 evaluates the probability of bit error with $\mathbf{R}$ equal to a circulant matrix with parameter $\delta = 0.2$. Now $A_i/A_1 = 0.5$ for $i = 2, 3, 4, 5$, and all 5 users have equal cross-correlation $\rho = -0.2$. Even though the magnitude of the cross-correlations are low, the CSMU receiver still performs better than the decorrelator and the MF and performs similarly to the linear MMSE receiver.

Finally, in Fig. 12-10 we evaluate the probability of bit error of the CSMU receiver

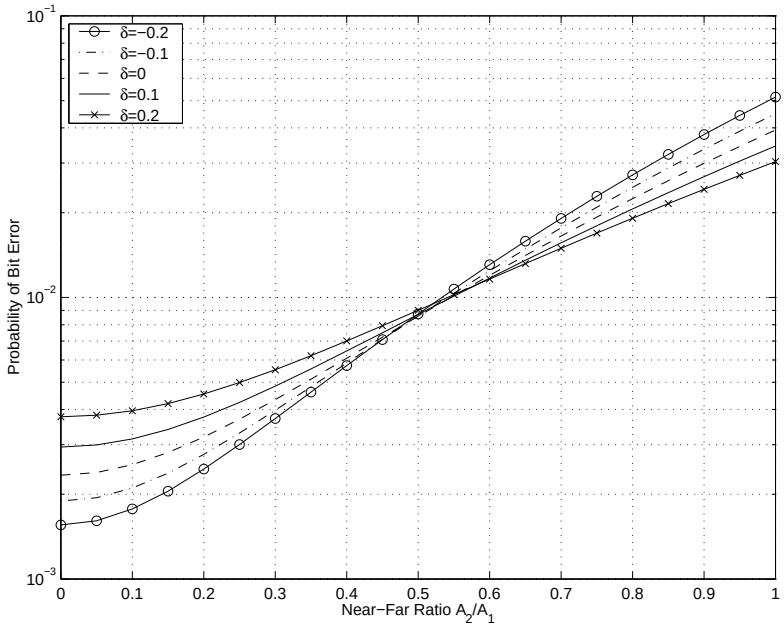


Figure 12-5: Probability of bit error with two users and cross-correlation $\rho = 0.8$ as a function of the near-far ratio A_2/A_1, where $\mathbf{R}$ is a circulant matrix with parameter δ. The SNR of the first user, the desired user, is 10 dB.

with $\mathbf{R}$ equal to a circulant matrix with parameter $\delta = 0.35$, in the case of 10 users with $\rho = -0.1$, and with accurate power control so that $A_i = 1$ for all i. Here again, the CSMU receiver performs better than the decorrelator and the MF and performs similarly to the linear MMSE receiver.

The analysis results presented in this section demonstrate that over a wide variety of the channel parameters, the CSMU receiver can outperform both the decorrelator and the MF receivers and can perform similarly to the linear MMSE receiver, even though it does not rely on knowledge of the channel parameters.

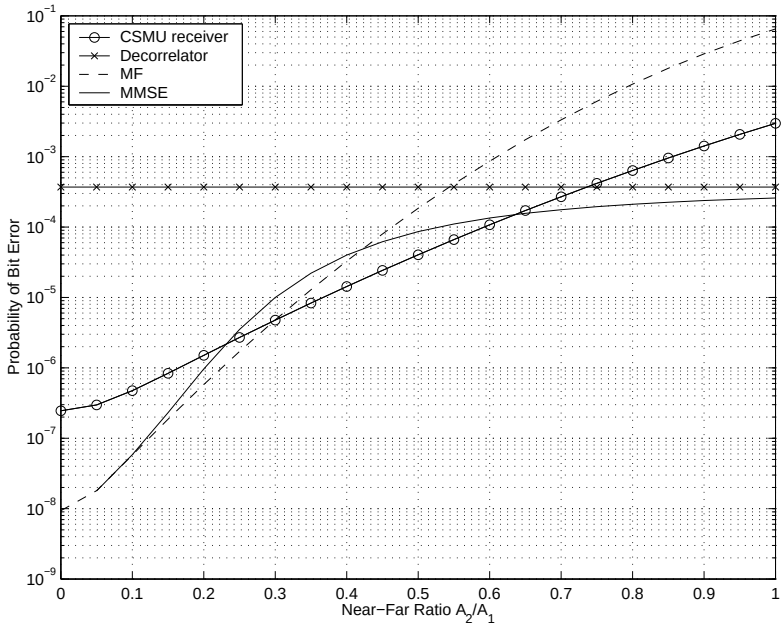


Figure 12-6: Probability of bit error with two users and cross-correlation $\rho = 0.8$, as a function of the near-far ratio A_2/A_1. In the CSMU receiver, $\mathbf{R} = \mathbf{I}_2$. The SNR of the first user, the desired user, is 15 dB.

12.5.2 SINR and Approximating the Probability of Detection Error

From (12.23)–(12.25), the terms $\hat{x}_i^S$, $\hat{x}_i^I$, and $\hat{x}_i^N$ are mutually independent and zero-mean, and have variances

$$\begin{aligned} \mathrm{var}(\hat{x}_i^S) &= [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2 A_i^2 && (12.32) \\ \mathrm{var}(\hat{x}_i^I) &= [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_i^* \mathbf{A}^2 [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_i - [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2 A_i^2 && (12.33) \\ \mathrm{var}(\hat{x}_i^N) &= \sigma^2 [\mathbf{R}]_{ii}. && (12.34) \end{aligned}$$

The SINR at the detector for the ith user is therefore

$$\gamma_i = \frac{[(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2 A_i^2}{\sigma^2 [\mathbf{R}]_{ii} + [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_i^* \mathbf{A}^2 [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_i - [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2 A_i^2}. \quad (12.35)$$

In the case of accurate power control, *i.e.,* $\mathbf{A} = A\mathbf{I}_m$, we can simplify (12.35) to

$$\gamma_i = \frac{[(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2}{\zeta [\mathbf{R}]_{ii} + [\mathbf{R}\mathbf{S}^*\mathbf{S}]_{ii} - [(\mathbf{R}\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2}, \quad (12.36)$$

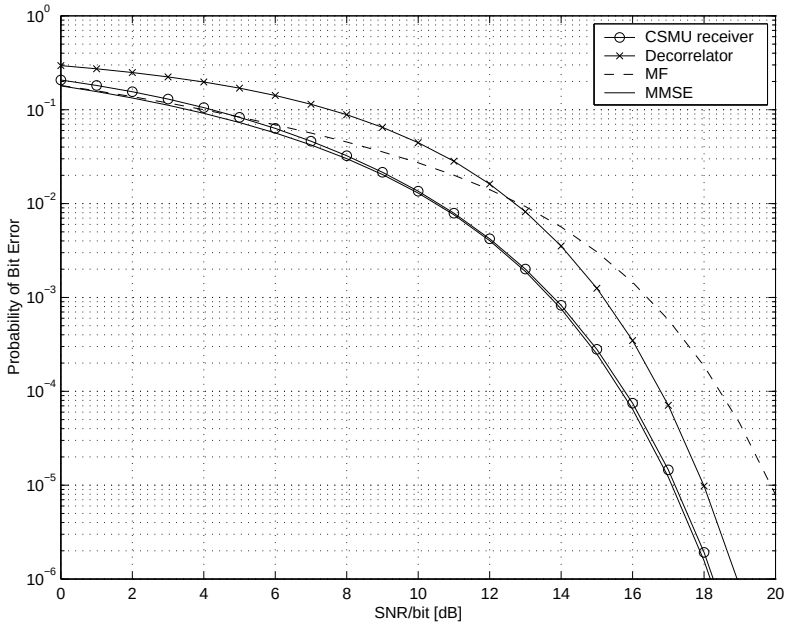


Figure 12-7: Probability of bit error with three users and cross-correlation $\rho = 0.8$, as a function of SNR. In the CSMU receiver, $\mathbf{R} = \mathbf{I}_3$. The amplitude A_1 of the desired user is 2 times greater than the amplitude A_2 of the second user and 4 times greater than the amplitude A_3 of the third user.

where

$$\frac{1}{\zeta} = \frac{A^2}{\sigma^2} \tag{12.37}$$

is the received SNR. An alternate form for (12.36), which will be more convenient for the analysis in Section 12.5.3, is

$$\gamma_i = \frac{1}{1 - \frac{[(\mathbf{RS}^*\mathbf{S})^{1/2}]^2_{ii}}{\zeta[\mathbf{R}]_{ii} + [\mathbf{RS}^*\mathbf{S}]_{ii}}} - 1. \tag{12.38}$$

Assuming $\hat{x}_i^I + \hat{x}_i^N$ is Gaussian, the probability of detection error can then be approximated as

$$P_i(\sigma) \approx \mathcal{Q}(\sqrt{\gamma_i}). \tag{12.39}$$

At low SNR, the Gaussian approximation is acceptable because Gaussian noise is the dominant impairment. However, at high SNR, the discrete distribution of the MAI is poorly

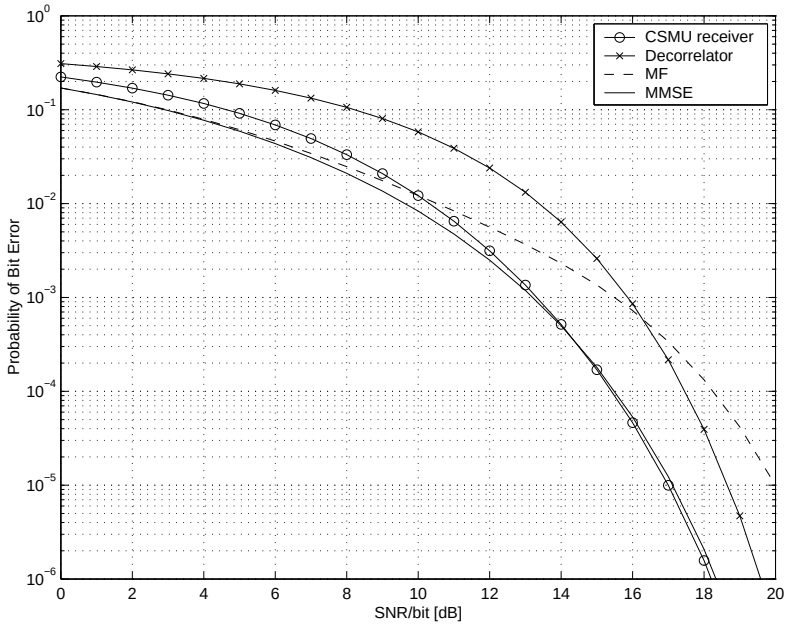


Figure 12-8: Probability of bit error with five users and cross-correlation $\rho = 0.8$, as a function of SNR. In the CSMU receiver, $\mathbf{R} = \mathbf{I}_5$. The amplitude A_1 of the desired user is 5 times greater than the amplitude A_i of any of the other interferers.

approximated by a Gaussian distribution, especially at the tails of the distribution where the bit-error rate is determined. Thus we do not expect (12.39) to be particularly accurate at high SNR.

12.5.3 Asymptotic Large System Performance

We now evaluate the performance of the CSMU receiver in the large system limit[2] for the special case in which $\mathbf{R} = P_{\mathcal{V}}$ where $\mathcal{V} = \mathcal{N}(\mathbf{S})^{\perp}$. If the signature vectors are linearly independent then $P_{\mathcal{V}} = \mathbf{I}_m$, and the CSMU receiver reduces to the OMU receiver. If they are linearly dependent, then the CSMU receiver reduces to the POMU receiver.

When $\mathbf{R} = P_{\mathcal{V}}$, the SINR at the detector for the ith user follows from (12.38) as

$$\gamma_i = \frac{1}{1 - \frac{[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}^2}{\zeta[P_{\mathcal{V}}]_{ii} + [\mathbf{S}^*\mathbf{S}]_{ii}}} - 1. \tag{12.40}$$

The following theorem characterizes the performance of the OMU and POMU receivers in

[2]The analysis presented in this section is the product of a joint collaboration with A. Chan, and appears in [39].

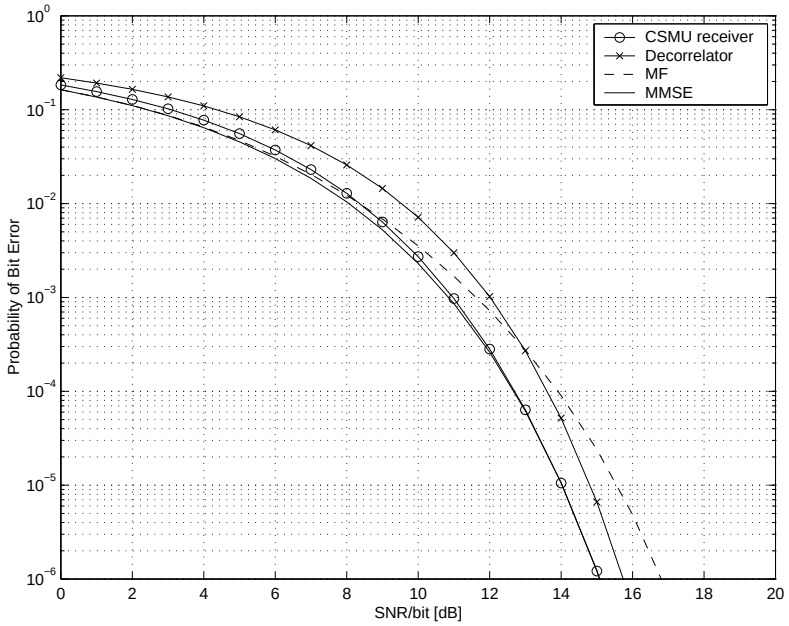


Figure 12-9: Probability of bit error with five users and cross-correlation $\rho = -0.2$, as a function of SNR. In the CSMU receiver, $\mathbf{R}$ is a circulant matrix with parameter $\delta = 0.2$. The amplitude A_1 of the desired user is 2 times greater than the amplitude A_i of any of the other interferers.

the large system limit when random Gaussian signatures and accurate power control are used. The method we use in its proof can be easily modified to characterize the performance of other multiuser detectors in the large system limit as well. For example, the method can be used to derive the asymptotic SINR for the matched filter detector, and we have recently used it to derive the asymptotic SINR for the decorrelator [53].

Theorem 12.4 (Asymptotic SINR). *Let the elements of the $n \times m$ signature matrix $\mathbf{S}$ be independent $\mathcal{CN}(0, 1/n)$, and let the matrix of amplitudes $\mathbf{A}$ be expressible as $A\mathbf{I}_m$. Then in the limit as $m \to \infty$ with $\beta \triangleq m/n$ held constant, the SINR for each user at the OMU and POMU demodulator output satisfy*[3]

$$\gamma_i \xrightarrow{\text{m.s.}} \frac{1}{1 - \frac{\eta_2[(\eta_1+\eta_2)E(\sqrt{1-\eta_1/\eta_2}) - 2\eta_1 K(\sqrt{1-\eta_1/\eta_2})]^2}{9\pi^2\beta^2(\zeta+1)}} - 1 \tag{12.41}$$

[3]We use the notation $\xrightarrow{\text{m.s.}}$ to denote convergence in the mean-squared (L^2) sense [164].

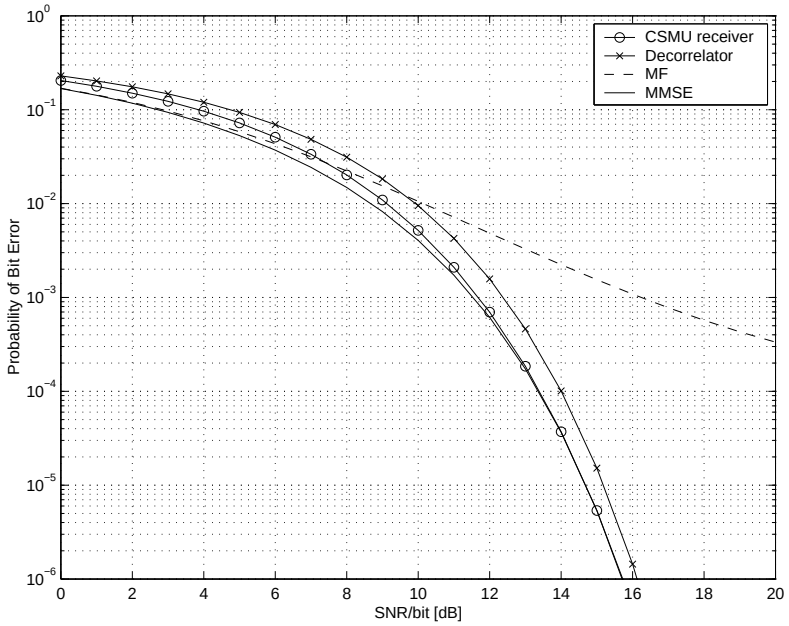


Figure 12-10: Probability of bit error with 10 users, cross-correlation $\rho = -0.1$, and accurate power control, as a function of SNR. In the CSMU receiver, $\mathbf{R}$ is a circulant matrix with parameter $\delta = 0.35$.

and

$$\gamma_i \xrightarrow{\text{m.s.}} \frac{1}{1 - \frac{\eta_2[(\eta_1+\eta_2)E(\sqrt{1-\eta_1/\eta_2})-2\eta_1 K(\sqrt{1-\eta_1/\eta_2})]^2}{9\pi^2\beta^2(\zeta/\beta+1)}} - 1 \tag{12.42}$$

respectively, where [165]

$$K(k) = \int_0^{\pi/2} \frac{dt}{\sqrt{1-k^2\sin^2 t}} = \int_0^1 \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} \tag{12.43}$$

$$E(k) = \int_0^{\pi/2} \sqrt{1-k^2\sin^2 t}\,dt = \int_0^1 \sqrt{\frac{1-k^2x^2}{1-x^2}}\,dx \tag{12.44}$$

are the complete elliptic integrals of the first and second kinds respectively, and

$$\eta_1 = (1-\sqrt{\beta})^2 \tag{12.45}$$

$$\eta_2 = (1+\sqrt{\beta})^2. \tag{12.46}$$

Proof: We begin by presenting the following lemma on Wishart matrices[4], which have the form $\mathbf{S}^*\mathbf{S}$ with the elements of $\mathbf{S}$ being independent $\mathcal{CN}(0, \sigma^2)$. The lemma and its proof rely on the concepts of isotropically distributed vectors and matrices, which are reviewed in Appendix E.

Lemma 12.1. *Let the elements of an $n \times m$ matrix $\mathbf{S}$ be independent $\mathcal{CN}(0, \sigma^2)$. Then the eigenvector matrix of $\mathbf{S}^*\mathbf{S}$ is isotropically distributed unitary and independent of the eigenvalues.*

We emphasize that this lemma is true for Wishart matrices of *any* size.

Proof: Let $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be the SVD of $\mathbf{S}$, where $\mathbf{U}$ is an $n \times n$ unitary matrix, $\mathbf{V}$ is an $m \times m$ unitary matrix, and Σ is a diagonal $n \times m$ matrix with diagonal elements $\sigma_i \geq 0$. Then

$$\mathbf{S}^*\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}^*, \tag{12.47}$$

where $\Lambda = \mathbf{\Sigma}^*\mathbf{\Sigma}$ is a diagonal matrix of eigenvalues of $\mathbf{S}^*\mathbf{S}$, and $\mathbf{V}$ is a matrix of eigenvectors of $\mathbf{S}^*\mathbf{S}$.

Let $\mathbf{\Upsilon}$ denote an independent, isotropically distributed unitary matrix. By premultiplying and postmultiplying (12.47) by $\mathbf{\Upsilon}^*$ and $\mathbf{\Upsilon}$ respectively, we have that

$$\mathbf{\Upsilon}^*\mathbf{S}^*\mathbf{S}\mathbf{\Upsilon} = \mathbf{\Upsilon}^*\mathbf{V}\Lambda\mathbf{V}^*\mathbf{\Upsilon}, \tag{12.48}$$

or, equivalently,

$$(\mathbf{S}\mathbf{\Upsilon})^*(\mathbf{S}\mathbf{\Upsilon}) = (\mathbf{V}^*\mathbf{\Upsilon})^*\Lambda(\mathbf{V}^*\mathbf{\Upsilon}). \tag{12.49}$$

Let us examine the left-hand side of (12.49). Since the elements of $\mathbf{S}$ are $\mathcal{CN}(0, \sigma^2)$, $\mathbf{S}$ is an isotropically distributed matrix. With $\mathbf{S}$ being isotropically distributed and $\mathbf{\Upsilon}$ being unitary, $\mathbf{S}\mathbf{\Upsilon}$ has the same distribution as $\mathbf{S}$, and consequently $(\mathbf{S}\mathbf{\Upsilon})^*(\mathbf{S}\mathbf{\Upsilon})$ has the same distribution as $\mathbf{S}^*\mathbf{S}$.

[4]A similar lemma exists for matrices $\mathbf{S}^*\mathbf{S}$, where the n-dimensional columns of $\mathbf{S}$ are independent and drawn uniformly from the surface of the unit n-sphere.

We now focus on the right-hand side of (12.49). Note that $\mathbf{V}^*\boldsymbol{\Upsilon}$ is unitary and Λ is diagonal, so that the right-hand side (12.49) is an eigendecomposition. Now, since $\boldsymbol{\Upsilon}$ is an isotropically distributed unitary matrix and $\mathbf{V}^*$ is a unitary matrix, the eigenvector matrix $\mathbf{V}^*\boldsymbol{\Upsilon}$ is an isotropically distributed unitary matrix. Furthermore, the eigenvector matrix $\mathbf{V}^*\boldsymbol{\Upsilon}$ is independent of the eigenvalue matrix Λ because $\boldsymbol{\Upsilon}$ is independent of Λ. □

To prove Theorem 12.4, we need to determine the limits of $[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}$, $[P_{\mathcal{V}}]_{ii}$, and $[\mathbf{S}^*\mathbf{S}]_{ii}$ as $m \to \infty$ with β held constant.

From (12.47), the quantity $[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}$ can be written as

$$[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii} = [\mathbf{V}\Lambda^{1/2}\mathbf{V}^*]_{ii} = \mathbf{v}_i^*\Lambda^{1/2}\mathbf{v}_i, \tag{12.50}$$

where $\mathbf{v}_i$ is the ith column of $\mathbf{V}^*$. Now, $\mathbf{V}$ and Λ are the eigenvector matrix and the eigenvalue matrix respectively in the eigendecomposition of the Wishart matrix $\mathbf{S}^*\mathbf{S}$. Thus using Lemma 12.1, we conclude that $\mathbf{V}^*$ is an isotropically distributed unitary matrix independent of Λ. Since $\mathbf{v}_i$ is a column of an isotropically distributed unitary matrix, from Appendix E it then follows that $\mathbf{v}_i$ is an isotropically distributed unit vector. Consequently, $\mathbf{v}_i$ has the same distribution as $\mathbf{z}/\sqrt{\mathbf{z}^*\mathbf{z}}$, where $\mathbf{z}$ is an m-dimensional vector of independent $\mathcal{CN}(0,1)$ random variables. Thus $[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii}$ has the same distribution as

$$\frac{\mathbf{z}^*\Lambda\mathbf{z}}{\mathbf{z}^*\mathbf{z}} = \frac{\sum_{j=1}^m \sqrt{\lambda_j}|z_j|^2/m}{\sum_{j=1}^m |z_j|^2/m}, \tag{12.51}$$

with the λ_j's denoting the eigenvalues of $\mathbf{S}^*\mathbf{S}$ and the z_j's denoting the components of $\mathbf{z}$. To evaluate the limit of (12.51) when $m \to \infty$, we rely on the following pair of lemmas.

Lemma 12.2 ([166]). *If the ratio of the number of users to the signature length is, or converges to a constant:*

$$\lim_{i\to\infty} \frac{m}{n} = \beta \in (0, \infty), \tag{12.52}$$

then the percentage of the m eigenvalues of $\mathbf{S}^*\mathbf{S}$ that lie below x converges to the cumulative distribution function of the probability density function

$$f_\beta(x) = [1-\beta^{-1}]^+\delta(x) + \frac{\sqrt{[x-\eta_1]^+[\eta_2-x]^+}}{2\pi\beta x} \tag{12.53}$$

here η_1 and η_2 are defined according to (12.45) and (12.46), and the operator $[\cdot]^+$ is defined according to

$$[u]^+ \stackrel{\triangle}{=} \max\{0, u\}. \tag{12.54}$$

Lemma 12.3 ([53]). *Let $\{c_j\}$ denote a set of independent identically distributed (iid) random variables independent of $\{\lambda_j\}$, where $\{\lambda_j\}$ denote the eigenvalues of a Wishart matrix under the conditions of Lemma 12.2. Furthermore, let $g(\cdot)$ be a function such that $E((g(\lambda_1))^2) < \infty$ when evaluated according to the probability density function $f_\beta(x)$ of (12.53). Then as $m \to \infty$,*

$$\frac{1}{m}\sum_{j=1}^{m} g(\lambda_j)c_j \xrightarrow{\text{m.s.}} E(g(\lambda_1))E(c_1), \tag{12.55}$$

where $E(g(\lambda_1))$ is evaluated according to $f_\beta(x)$.

Applying Lemma 12.3 and the strong law of large numbers to the numerator and denominator of (12.51) respectively, we have

$$[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii} \xrightarrow{\text{m.s.}} \frac{E(\sqrt{\lambda_1})E(|z_1|^2)}{E(|z_1|^2)} = E(\sqrt{\lambda_1}) \tag{12.56}$$

as $m \to \infty$, where $E(\sqrt{\lambda_1})$ is evaluated according to the probability density function $f_\beta(x)$ of (12.53). Thus,

$$\begin{aligned}
[(\mathbf{S}^*\mathbf{S})^{1/2}]_{ii} \xrightarrow{\text{m.s.}} E(\sqrt{\lambda_1}) &= \int_0^\infty \sqrt{x} f_\beta(x) dx \\
&= \int_0^\infty \sqrt{x}\left\{[1-\beta^{-1}]^+\delta(x) + \frac{\sqrt{[x-\eta_1]^+[\eta_2-x]^+}}{2\pi\beta x}\right\} dx \\
&= \int_{\eta_1}^{\eta_2} \frac{\sqrt{(x-\eta_1)(\eta_2-x)}}{2\pi\beta\sqrt{x}} dx \\
&= \frac{\sqrt{\eta_2}}{3\pi\beta}[(\eta_1+\eta_2)E(\sqrt{1-\eta_1/\eta_2}) - 2\eta_1 K(\sqrt{1-\eta_1/\eta_2})],
\end{aligned} \tag{12.57}$$

where the last equality is from [165], and where $K(\cdot)$, $E(\cdot)$, η_1, and η_2 are defined by (12.43), (12.44), (12.45), and (12.46) respectively.

Similarly, $[P_{\mathcal{V}}]_{ii}$ can be written as

$$[P_{\mathcal{V}}]_{ii} = [\mathbf{V}\tilde{\mathbf{I}}\mathbf{V}^*]_{ii} = \mathbf{v}_i^*\tilde{\mathbf{I}}\mathbf{v}_i, \tag{12.58}$$

where $\tilde{\mathbf{I}}$ is a diagonal matrix with ith diagonal element equal to μ_i given by

$$\mu_i = \begin{cases} 1, & \lambda_i \neq 0, \\ 0, & \lambda_i = 0, \end{cases} \tag{12.59}$$

so that $[P_{\mathcal{V}}]_{ii}$ has the same distribution as

$$\frac{\mathbf{z}^*\tilde{\mathbf{I}}\mathbf{z}}{\mathbf{z}^*\mathbf{z}} = \frac{\sum_{j=1}^{m} \mu_j |z_j|^2/m}{\sum_{j=1}^{m} |z_j|^2/m}. \tag{12.60}$$

Applying Lemma 12.3 and the strong law of large numbers to the numerator and denominator of (12.60) respectively, we have

$$[P_{\mathcal{V}}]_{ii} \xrightarrow{\text{m.s.}} \frac{E(\mu_1)E(|z_1|^2)}{E(|z_1|^2)} = E(\mu_1) \tag{12.61}$$

as $m \to \infty$, where $E(\mu_1)$ is evaluated according to the probability density function $f_\beta(x)$ of (12.53). Thus,

$$[P_{\mathcal{V}}]_{ii} \xrightarrow{\text{m.s.}} E(\mu_1) = \lim_{i\to\infty} P(\lambda_1 \neq 0) = \begin{cases} 1, & \beta \leq 1, \\ \frac{1}{\beta}, & \beta > 1. \end{cases} \tag{12.62}$$

Lastly,

$$[\mathbf{S}^*\mathbf{S}]_{ii} = \mathbf{s}_i^*\mathbf{s}_i = \sum_{j=1}^{n} s_{ji}^2 \xrightarrow{\text{m.s.}} 1 \tag{12.63}$$

by the strong law of large numbers, with the s_{ji}'s denoting the components of $\mathbf{s}_i$.

It is well known that if $x_n \xrightarrow{\text{m.s.}} \bar{x}$ and $y_n \xrightarrow{\text{m.s.}} \bar{y}$, then $x_n \pm y_n \xrightarrow{\text{m.s.}} \bar{x} + \bar{y}$ and $x_n y_n \xrightarrow{\text{m.s.}} \bar{x}\bar{y}$ [164]. The following lemma which involves the convergence of $1/x_n$ is now required to complete the proof of Theorem 12.4.

Lemma 12.4 ([53]). *Let $x_n \xrightarrow{\text{m.s.}} \bar{x}$, where $\{x_n\}$ is a sequence of random variables such*

that $|1/x_n| \leq B$ *for all* n, *and* $\bar{x} \neq 0$. *Then*

$$\frac{1}{x_n} \xrightarrow{\text{m.s.}} \frac{1}{\bar{x}}. \tag{12.64}$$

Substituting (12.57), (12.62), and (12.63) into (12.40), and using the fact that $\gamma_i \leq 1/\zeta$ with Lemma 12.4 completes the proof of Theorem 12.4. □

Since the MAI is asymptotically Gaussian in the infinite-user limit, we expect (12.39) to be an accurate approximation to the bit-error rate at all SNR, where γ_i is given by Theorem 12.4. We will use this approximation to compute the bit-error rate for the remainder of this section.

From (12.41) and (12.42), we note that as $\sigma \to 0$, i.e. $\zeta \to 0$, the SINR γ_i converges to a finite constant. This behavior is expected, since the condition (12.30) cannot be satisfied as $m \to \infty$. Thus the probability of bit error does not go to zero in this scenario.

In Fig. 12-11, the bit-error rate in the infinite-user limit for the OMU receiver is compared to the single-user MF, the decorrelator, and the linear MMSE receiver, for $\beta = 0.95$. For the SNR range shown, the OMU receiver performs better than the decorrelator and the MF. At low SNR, the performance of the OMU receiver is close to that of the linear MMSE receiver. Note that at high SNR, the bit-error rate of the OMU receiver begins to converge to its high-SNR, infinite-user limit. In Fig. 12-12, we plot the probability of bit error in the infinite-user limit[5] as a function of β, with an SNR of 8 dB. For β roughly greater than 0.55 but less than 1.45, the OMU/POMU receiver performs significantly better than both the decorrelator and the MF.

In summary, the performance analysis shows that over a wide range of channel parameters the CSMU receiver performs similarly to the linear MMSE receiver and outperforms both the decorrelator and the single-user MF.

The asymptotic SINR in the infinite-user limit was derived assuming that the outputs of the CSMU receiver are uncorrelated on the space in which they are contained. By allowing for other choices of the output covariance shape $\mathbf{R}$ we can further improve the performance of the CSMU receiver in many cases. An important direction for future work is to compute the asymptotic SINR for other choices of $\mathbf{R}$. Based on knowledge of these asymptotic limits

[5] We derive the asymptotic large system performance of the decorrelator for the case $\beta > 1$ in [53].

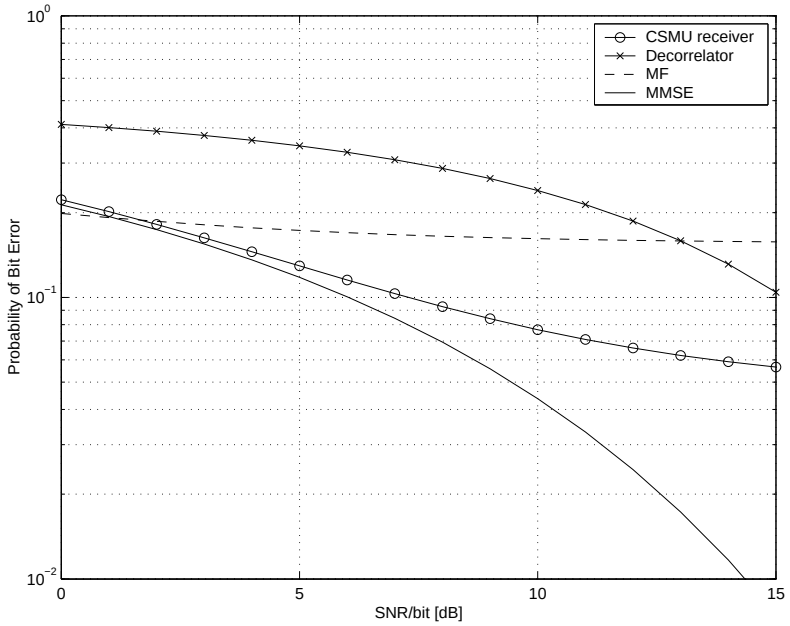


Figure 12-11: Probability of bit error in the large-system limit, with equal-power users, random signatures, and $\beta = 0.95$. In the CSMU receiver, $\mathbf{R} = \mathbf{I}$.

we can then develop optimal methods for choosing $\mathbf{R}$.

Another important direction to investigate is the performance of the CSMU receiver for coded systems. In particular, it would be useful to derive the spectral efficiency in the infinite-user limit of the CDMA channel when using a CSMU receiver, and then compare it to the known spectral efficiencies when using other multiuser receivers [52].

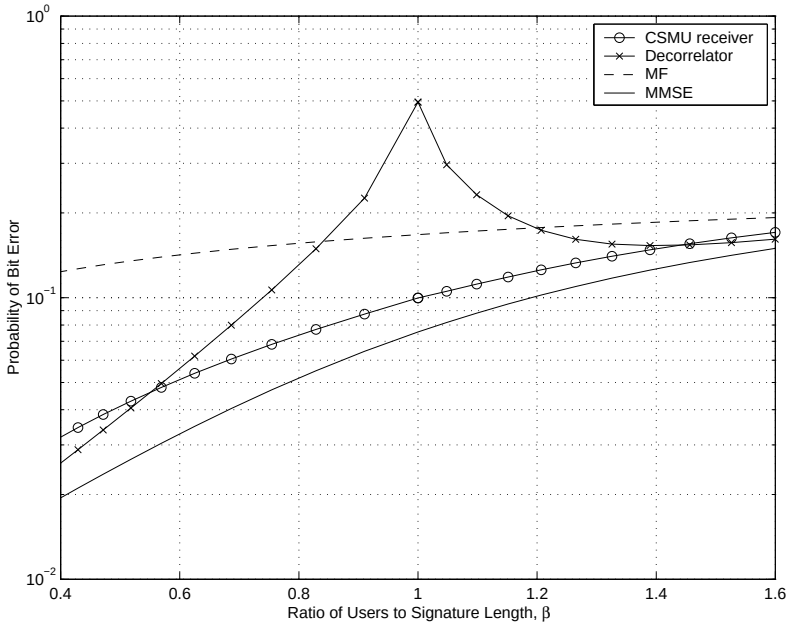


Figure 12-12: Probability of bit error as a function of β in the large-system limit, with equal-power users, random signatures, and SNR of 8 dB. In the CSMU receiver, $\mathbf{R} = P_{\mathcal{V}}$.

Appendix A

Iterative Algorithm Computing the Least-Squares Orthogonal Vectors

In this appendix we derive the angle θ for the optimal Givens rotation that maximizes $R'^{(j+1)}_{hs}$ of (8.71).

Recall that,

$$\begin{aligned} R'^{(j+1)}_{hs} &= (cb_{11} + sb_{21})^2 + (cb_{22} - sb_{12})^2 \\ &= c^2(b_{11}^2 + b_{22}^2) + s^2(b_{21}^2 + b_{12}^2) + 2cs(b_{11}b_{21} - b_{22}b_{12}), \end{aligned} \tag{A.1}$$

where $c = \cos(\theta)$ and $s = \sin(\theta)$. Differentiating (A.1) with respect to θ,

$$\begin{aligned} \frac{dR'^{(j+1)}_{hs}}{d\theta} &= -2cs(b_{11}^2 + b_{22}^2) + 2cs(b_{21}^2 + b_{12}^2) + 2(c^2 - s^2)(b_{11}b_{21} - b_{22}b_{12}) \\ &= -x\sin(2\theta) + 2y\cos(2\theta), \end{aligned} \tag{A.2}$$

where $x = b_{11}^2 + b_{22}^2 - b_{21}^2 - b_{12}^2$, $y = b_{11}b_{21} - b_{22}b_{12}$, and we used the relations $c^2 - s^2 = \cos(2\theta)$, $2cs = \sin(2\theta)$. Equating (A.2) to 0 yields,

$$x\sin(2\theta) = 2y\cos(2\theta). \tag{A.3}$$

Note, that $\mathbf{J}(r, l, \theta) = \mathbf{J}(r, l, \theta + 2\pi k)$ for any integer k, thus it is sufficient to consider

solutions $\theta \in (-\pi, \pi]$. If $x \neq 0$, then the solutions to (A.3) are,

$$\theta = \frac{1}{2} \tan^{-1} \left(\frac{2y}{x} \right). \tag{A.4}$$

If $x = 0$ and $y \neq 0$, then the solutions to (A.3) are $\theta = \pm\pi/4$. If $x = y = 0$, then from (A.1) we see that $R'^{(j+1)}_{hs}$ does not depend on θ, and we choose $\theta = 0$.

Taking the second derivative of $R'^{(j+1)}_{hs}$ with respect to θ yields,

$$\frac{d^2 R'^{(j+1)}_{hs}}{d\theta^2} = -2x \cos(2\theta) - 4y \sin(2\theta). \tag{A.5}$$

Thus, a solution θ of (A.3) maximizes $R'^{(j+1)}_{hs}$ if

$$x \cos(2\theta) + 2y \sin(2\theta) > 0, \tag{A.6}$$

which for $x \neq 0$ reduces to

$$x \cos(2\theta)(1 + 4y^2/x^2) > 0. \tag{A.7}$$

Thus if $x \neq 0$, then $\hat{\theta}$ is a maximum of $R'^{(j+1)}_{hs}$ if $\hat{\theta}$ has the form (A.4) and $\cos(2\hat{\theta})$ and x have the same sign. Similarly, for $x = 0$, $\sin(2\hat{\theta})$ and y must have the same sign. So, for $y > 0$, $\hat{\theta} = \pi/4$, and for $y < 0$, $\hat{\theta} = -\pi/4$.

Appendix B

Matrix Equalities

In this appendix we prove some useful matrix equalities which we use throughout the thesis.

Let $\mathbf{S}$ be an arbitrary $m \times n$ matrix, and let $\mathbf{T}$ be an arbitrary $n \times n$ matrix with $\mathcal{R}(\mathbf{T}) \subseteq \mathcal{N}(\mathbf{S})^{\perp}$. Then with $\mathbf{S}^{\dagger}$ denoting the Moore-Penrose pseudo inverse of $\mathbf{S}$,

$$\mathbf{S}\mathbf{T}^{1/2}\mathbf{S}^{\dagger} = (\mathbf{S}\mathbf{T}\mathbf{S}^{\dagger})^{1/2}. \tag{B.1}$$

We establish (B.1) by showing that $\mathbf{A}^2 = \mathbf{S}\mathbf{T}\mathbf{S}^{\dagger}$ where $\mathbf{A} = \mathbf{S}\mathbf{T}^{1/2}\mathbf{S}^{\dagger}$. Indeed,

$$\mathbf{A}^2 = \mathbf{S}\mathbf{T}^{1/2}\mathbf{S}^{\dagger}\mathbf{S}\mathbf{T}^{1/2}\mathbf{S}^{\dagger} = \mathbf{S}\mathbf{T}^{1/2}P_{\mathcal{N}(\mathbf{S})^{\perp}}\mathbf{T}^{1/2}\mathbf{S}^{\dagger} = \mathbf{S}\mathbf{T}\mathbf{S}^{\dagger}, \tag{B.2}$$

where we used the fact that $\mathcal{R}(\mathbf{T}^{1/2}) \subseteq \mathcal{N}(\mathbf{S})^{\perp}$ so that $P_{\mathcal{N}(\mathbf{S})^{\perp}}\mathbf{T}^{1/2} = \mathbf{T}^{1/2}$.

An important special case of (B.1) is the case in which $n = m$ and $\mathbf{S}$ is an invertible $m \times m$ matrix. Then $\mathcal{N}(\mathbf{S})^{\perp} = \mathbb{C}^m$ so that for any $m \times m$ matrix $\mathbf{T}$, $\mathcal{R}(\mathbf{T}) \subseteq \mathcal{N}(\mathbf{S})^{\perp}$, and

$$\mathbf{S}\mathbf{T}^{1/2}\mathbf{S}^{-1} = (\mathbf{S}\mathbf{T}\mathbf{S}^{-1})^{1/2}. \tag{B.3}$$

If in addition $\mathbf{T}$ is invertible, then substituting $\mathbf{T}^{-1}$ for $\mathbf{T}$ in (B.3),

$$\mathbf{S}\mathbf{T}^{-1/2}\mathbf{S}^{-1} = (\mathbf{S}\mathbf{T}\mathbf{S}^{-1})^{-1/2}. \tag{B.4}$$

Appendix C

Subspace Whitening

C.1 Implication of Noninvertible Covariance Matrix

Let $\mathbf{a}$ be a zero-mean random vector in $\mathbb{C}^m$ with rank-n covariance matrix $\mathbf{C}_a$ were $n < m$, and let $\mathcal{V}$ denote the range space $\mathcal{R}(\mathbf{C}_a)$. Then since $\mathbf{C}_a$ is Hermitian, the null space $\mathcal{N}(\mathbf{C}_a) = \mathcal{V}^\perp$. Let $\{\mathbf{v}_i, 1 \leq i \leq n\}$ denote an orthonormal basis for $\mathcal{V}$, and let $\{\mathbf{v}_i, n+1 \leq i \leq m\}$ denote an orthonormal basis for $\mathcal{V}^\perp$. Since $\mathbf{v}_i \in \mathcal{N}(\mathbf{C}_a)$ for $n+1 \leq i \leq m$,

$$\mathbf{C}_a \mathbf{v}_i = \mathbf{0}, \ n+1 \leq i \leq m. \tag{C.1}$$

We now try to gain some insight into (C.1). Let $\lambda_i = \mathbf{v}_i^* \mathbf{a}$, $n+1 \leq i \leq m$. Then the variance of λ_i, denoted σ_i, is $\sigma_i = E(\mathbf{v}_i^* \mathbf{a} \mathbf{a}^* \mathbf{v}_i) = \mathbf{v}_i^* \mathbf{C}_a \mathbf{v}_i = 0$, $n+1 \leq i \leq m$, where we used (C.1). Thus $\lambda_i = E(\lambda_i) = 0$ w.p. 1, or

$$\mathbf{v}_i^* \mathbf{a} = 0, \ n+1 \leq i \leq m, \tag{C.2}$$

for any realization of $\mathbf{a}$, w.p. 1. From (C.2) we conclude that the elements of $\mathbf{a}$ are deterministically linearly dependent and any realization of the random vector $\mathbf{a}$ lies in a subspace of $\mathbb{C}^m$. Specifically, $\mathbf{a}$ lies in the orthogonal complement in $\mathbb{C}^m$ of the space spanned by the vectors $\{\mathbf{v}_i, n+1 \leq i \leq m\}$. Thus, $\mathbf{a} \in \mathcal{V}$.

C.2 Subspace Whitening

We now translate the conditions on a random vector $\mathbf{b}$ to be white on $\mathcal{V}$, to conditions on the covariance $\mathbf{C}_b$ of $\mathbf{b}$. The first condition on the vector $\mathbf{b}$ is that $\mathbf{b} \in \mathcal{V}$. Suppose that $\mathbf{b} \in \mathcal{V}$. Then $\mathbf{v}_i^* \mathbf{b} = 0, m+1 \leq i \leq n$ (w.p. 1), since the vectors $\{\mathbf{v}_i, m+1 \leq i \leq n\}$ span $\mathcal{V}^\perp$. This in turn implies that

$$\mathbf{C}_b \mathbf{v}_i = \mathbf{0}, \ n+1 \leq i \leq m, \tag{C.3}$$

so that the null space of $\mathbf{C}_b$ contains $\mathcal{V}^\perp$. Conversely, suppose that the null space of $\mathbf{C}_b$ contains $\mathcal{V}^\perp$. Then (C.3) holds, and we have already shown that this implies that $\mathbf{b} \in \mathcal{V}$. We conclude that $\mathbf{b} \in \mathcal{V}$ if and only if the null space of $\mathbf{C}_b$ contains $\mathcal{V}^\perp$, so that $\mathbf{C}_b$ satisfies (C.3).

We now discuss the requirement that the representation of $\mathbf{b}$ in terms of any orthonormal basis for $\mathcal{V}$ is white. Let $\mathbf{V}_1$ denote the matrix of columns $\{\mathbf{v}_i, 1 \leq i \leq n\}$, that form a basis for $\mathcal{V}$. The representation of $\mathbf{b}$ in this basis for $\mathcal{V}$ is $\mathbf{b}_v = \mathbf{V}_1^* \mathbf{b}$, $\mathbf{b}_v \in \mathbb{C}^n$. We require that $\mathbf{b}_v$ is white, namely that the covariance matrix of $\mathbf{b}_v$ is equal to $c^2 \mathbf{I}_n$. Since the covariance of $\mathbf{b}_v$ is given by $\mathbf{V}_1^* \mathbf{C}_b \mathbf{V}_1$, our requirement on $\mathbf{C}_b$ is

$$\mathbf{V}_1^* \mathbf{C}_b \mathbf{V}_1 = c^2 \mathbf{I}_n, \tag{C.4}$$

for some $c > 0$. Thus the matrix $\mathbf{C}_b$ has to satisfy (C.3) and (C.4), which can be combined into the single condition

$$\mathbf{C}_b = c^2 P_{\mathcal{V}} = c^2 \mathbf{V} \tilde{\mathbf{I}}_n \mathbf{V}^*, \tag{C.5}$$

where $\mathbf{V}$ is the matrix of columns $\{\mathbf{v}_i, 1 \leq i \leq m\}$, and $\tilde{\mathbf{I}}_n$ is given by

$$\tilde{\mathbf{I}}_n = \begin{bmatrix} \mathbf{I}_n & 0 \\ 0 & 0 \end{bmatrix}. \tag{C.6}$$

Appendix D

Positive Worst-Case Threshold in CSLS Estimation

In this appendix we show that for $\hat{\mathbf{x}}_{\text{CSLS}} \neq \hat{\mathbf{x}}_{\text{LS}}$, $\zeta_{\text{WC}} \geq 0$, where ζ_{WC} is the worst case bound in CSLS estimation, and is given by (11.23). To this end we need to prove that $(1/m)\sum_{i=1}^{m} \lambda_i^{-1} \geq \alpha^2$ or, equivalently,

$$\frac{1}{m}\sum_{i=1}^{m} \lambda_i^{-1} \left(\sum_{i=1}^{m} \lambda_i\right)^2 \geq \left(\sum_{i=1}^{m} \lambda_i^{1/2}\right)^2, \tag{D.1}$$

with equality if and only if $\hat{\mathbf{x}}_{\text{CSLS}} = \hat{\mathbf{x}}_{\text{LS}}$.

From the Cauchy-Schwarz inequality,

$$\frac{1}{m}\sum_{i=1}^{m} \lambda_i^{-1} \sum_{i=1}^{m} \lambda_i \geq \frac{1}{m}\left(\sum_{i=1}^{m} \lambda_i^{-1/2}\lambda_i^{1/2}\right)^2 = m, \tag{D.2}$$

and

$$\sum_{i=1}^{m} \lambda_i = \frac{1}{m}\sum_{i=1}^{m} 1 \sum_{i=1}^{m} \lambda_i \geq \frac{1}{m}\left(\sum_{i=1}^{m} \lambda_i^{1/2}\right)^2. \tag{D.3}$$

Combining (D.2) with (D.3) proves the inequality (D.1).

We have equality in (D.2) if and only if $\lambda_i^{-1/2} = a\lambda_i^{1/2}$ for a constant $a \neq 0$, which implies that all the eigenvalues λ_i are equal, so that $\mathbf{B}$ is proportional to $\mathbf{I}_m$ and from Theorem 11.1, $\hat{\mathbf{x}}_{\text{CSLS}} = \hat{\mathbf{x}}_{\text{LS}}$. Under the same condition we have equality in (D.3). We

therefore conclude that when $\hat{\mathbf{x}}_{\text{CSLS}} \neq \hat{\mathbf{x}}_{\text{LS}}$, there is always a range of SNR values for which $\text{MSE}(\hat{\mathbf{x}}_{\text{CSLS}}) \leq \text{MSE}(\hat{\mathbf{x}}_{\text{LS}})$.

Appendix E

Isotropically Distributed Vectors and Matrices

In this appendix, we define the concept of isotropically distributed vectors and matrices and highlight the key properties that are used to prove Theorem 12.4. A more detailed discussion can be found in [167].

Definition E.1. *An m-dimensional complex random vector $\boldsymbol{\phi}$ is isotropically distributed if its probability density is invariant to all unitary transformations; i.e., $f(\boldsymbol{\phi}) = f(\boldsymbol{\Theta}^*\boldsymbol{\phi})$ for all $\boldsymbol{\Theta}$ such that $\boldsymbol{\Theta}^*\boldsymbol{\Theta} = \mathbf{I}_m$.*

Intuitively, an isotropically distributed complex vector is equally likely to point in any direction in complex space. Thus, the probability density of $\boldsymbol{\phi}$ is a function of its magnitude but not its direction. If, in addition, $\boldsymbol{\phi}$ is constrained to be a unit vector, then the probability density is

$$f(\boldsymbol{\phi}) = \frac{\Gamma(m)}{\pi^m}\delta(\boldsymbol{\phi}^*\boldsymbol{\phi} - 1), \tag{E.1}$$

and $\boldsymbol{\phi}$ is conveniently generated by $\boldsymbol{\phi} = \mathbf{z}/\sqrt{\mathbf{z}^*\mathbf{z}}$, where $\mathbf{z}$ is an m-dimensional vector of independent $\mathcal{CN}(0,1)$ random variables.

Definition E.2. *An $n \times m$ complex random matrix $\boldsymbol{\Phi}$ is isotropically distributed if its probability density is unchanged when premultiplied by an $n \times n$ unitary matrix; i.e., $f(\boldsymbol{\Phi}) = f(\boldsymbol{\Theta}^*\boldsymbol{\Phi})$ for all $\boldsymbol{\Theta}$ such that $\boldsymbol{\Theta}^*\boldsymbol{\Theta} = \mathbf{I}_n$.*

From the definition of an isotropically distributed matrix, it can be shown that the probability density is also unchanged when the matrix is postmultiplied by an $m \times m$ unitary matrix; i.e., $f(\mathbf{\Phi}) = f(\mathbf{\Phi\Theta})$ for all $\mathbf{\Theta}$ such that $\mathbf{\Theta}^*\mathbf{\Theta} = \mathbf{I}_m$. Furthermore, by combining Definitions E.1 and E.2, we can readily see that the column vectors of $\mathbf{\Phi}$ are themselves isotropically distributed vectors.

Bibliography

[1] A. Fettweis, "Wave digital filters: Theory and practice," *IEEE Proc.*, vol. 74, pp. 270–316, Feb. 1986.

[2] G. W. Wornell and A. V. Oppenheim, "Wavelet-based representations for a class of self similar signals with application to fractal modulation," *IEEE Trans. Inform. Theory*, vol. 38, pp. 785–800, Mar. 1992.

[3] T. L. Carroll and L. M. Pecora, "Synchronizing chaotic circuits," *IEEE Trans. Circuit Syst.*, vol. 38, pp. 453–456, Apr. 1991.

[4] L. M. Pecora and T. L. Carroll, "Driving systems with chaotic signals," *Phys. Rev. A*, vol. 44, pp. 2374–2383, Aug. 1991.

[5] K. M. Cuomo, A. V. Oppenheim, and S. H. Strogatz, "Synchronization of Lorenz-based chaotic circuits with applications to communications," *IEEE Trans. Circuit Syst. II*, vol. 40, pp. 626–639, Oct. 1993.

[6] A. C. Singer, A. V. Oppenheim, and G. W. Wornell, "Detection and estimation of multiplexed soliton signals," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2768–2782, Oct. 1999.

[7] K. S. Tang, K. F. Man, S. Kwong, and Q. He, "Genetic algorithms and their applications," *IEEE Signal Processing Mag.*, pp. 22–37, Nov. 1996.

[8] M. Pirlot and R. V. V. Vidal, "Simulated annealing: A tutorial," *Contr. Cybernetics*, vol. 25, no. 1, pp. 9–31, 1996.

[9] S. S. Haykin, *Neural Networks: A Comprehensive Foundation*, Upper Saddle River, NJ: Prentice Hall, Inc., second edition, 1999.

[10] H. W. Schüssler and P. Steffen, "Some advanced topics in filter design," in *Advanced Topics in Signal Processing*, J. S. Lim and A. V. Oppenheim, Eds. Englewood Cliffs, NJ: Prentice Hall, Inc., 1988.

[11] J. G. Proakis, *Digital Communications*, McGraw-Hill, Inc., third edition, 1995.

[12] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice Hall, Inc., 1992.

[13] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley, MA: Wellesley-Cambridge Press, 1996.

[14] S. G. Mallat, "A theory of multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, 1989.

[15] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Englewood Cliffs, NJ: Prentice Hall, Inc., 1995.

[16] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 2915–2925, Nov. 1994.

[17] M. Unser, "Sampling — 50 years after Shannon," *IEEE Proc.*, vol. 88, pp. 569–587, Apr. 2000.

[18] Y. C. Eldar and A. V. Oppenheim, "Nonredundant and redundant sampling with arbitrary sampling and reconstruction spaces," *Proceedings of the 2001 Workshop on Sampling Theory and Applications, SampTA'01*, pp. 229–234, May 2001.

[19] Y. C. Eldar, "Sampling and reconstruction in arbitrary spaces and oblique dual frame vectors," submitted to *J. Fourier Analys. Appl.*, May 2001.

[20] D. G. Griffiths, *Introduction to Quantum Mechanics*, Upper Saddle River, NJ: Prentice Hall, Inc., 1995.

[21] A. S. Holevo, "Statistical decisions in quantum theory," *J. Multivar. Anal.*, vol. 3, pp. 337–394, Dec. 1973.

[22] H. P. Yuen, R. S. Kennedy, and M. Lax, "Optimum testing of multiple hypotheses in quantum detection theory," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 125–134, Mar. 1975.

[23] C. W. Helstrom, *Quantum Detection and Estimation Theory*, New York: Academic Press, 1976.

[24] M. Charbit, C. Bendjaballah, and C. W. Helstrom, "Cutoff rate for the m-ary PSK modulation channel with optimal quantum detection," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1131–1133, Sep. 1989.

[25] M. Ban, K. Kurokawa, R. Momose, and O. Hirota, "Optimum measurements for discrimination among symmetric quantum states and parameter estimation," *Int. J. Theor. Phys.*, vol. 36, pp. 1269–1288, 1997.

[26] Y. C. Eldar and G. D. Forney, Jr., "On quantum detection and the square-root measurement," *IEEE Trans. Inform. Theory*, vol. 47, pp. 858–872, Mar. 2001.

[27] Y. C. Eldar, "Least-squares inner product shaping," *Linear Algebra Appl.*, to appear.

[28] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.

[29] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of nonsubtractive dither," *IEEE Trans. Signal Processing*, vol. 48, no. 2, pp. 499–516, Feb. 2000.

[30] R. A. Wannamaker, *The Theory of Dithered Quantization*, Ph.D. thesis, University of Waterloo, Waterloo, Canada, 1997.

[31] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[32] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*, Addison-Wesley, Inc., 1991.

[33] T. Kailath and V. Poor, "Detection of stochastic processes," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2230–2259, Oct. 1998.

[34] E. M. Friel and K. M. Pasala, "Direction finding with compensation for a near field scatterer," in *International Symposium Antennas and Propagation Society*, 1995, pp. 106–109.

[35] R. J. Piechocki, N. Canagarajah, and J. P. McGeehan, "Improving the direction-of-arrival resolution via double code filtering in WCDMA," in *First International Conference on 3G Mobile Communication Technologies*, Mar. 2000, pp. 204–207.

[36] Y. C. Eldar and A. V. Oppenheim, "Orthogonal matched filter detection," *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP-2001), (Salt Lake, UT)*, 2001.

[37] Y. C. Eldar, A. V. Oppenheim, and D. Egnor, "Orthogonal and projected orthogonal matched filter detection," submitted to *IEEE Trans. Signal Processing*, Jan. 2001.

[38] Y. C. Eldar and A. V. Oppenheim, "Orthogonal multiuser detection," *Signal Processing*, to appear.

[39] Y. C. Eldar and A. M. Chan, "Orthogonal and projected orthogonal multiuser detection," submitted to *IEEE Trans. Inform. Theory*, May 2001.

[40] Y. C. Eldar and A. V. Oppenheim, "MMSE whitening and subspace whitening," submitted to *IEEE Trans. Signal Processing*, July 2001.

[41] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," submitted to *IEEE Trans. Signal Processing*, Sep. 2001.

[42] T. Kailath, *Lectures On Linear Least-Squares Estimation*, Wein, NY: Springer, 1976.

[43] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Upper Saddle River, NJ: Prentice Hall, Inc., 1993.

[44] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, Inc., 1992.

[45] J. A. Cadzow, "Signal processing via least squares modeling," *IEEE ASSP Mag.*, pp. 12–31, Oct. 1990.

[46] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, Feb. 1970.

[47] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, Washington, DC: V.H. Winston, 1977.

[48] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497–508, Aug. 1973.

[49] S. Verdu, *Multiuser Detection*, Cambridge, UK: Cambridge Univ. Press, 1998.

[50] R. Lupas and S. Verdu, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 123–136, Jan. 1989.

[51] D. N. C. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inform. Theory*, vol. 45, pp. 641–657, Mar. 1999.

[52] S. Verdu and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 622–640, Mar. 1999.

[53] Y. C. Eldar and A. M. Chan, "On Wishart matrix eigenvalues and eigenvectors and the asymptotic performance of the decorrelator," submitted to *IEEE Trans. Inform. Theory*, May 2001.

[54] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. Signal Processing*, vol. 42, no. 6, pp. 1413–1424, June 1994.

[55] Y. C. Eldar, "On geometrical properties of the decorrelator," *IEEE Comm. Lett.*, to appear.

[56] R. T. Behrens and L. L. Scharf, "Parameter estimation in the presence of low-rank noise," in *Proc. Twenty-Second Asilomar Conf. Signals Syst. Comput.*, (Pacific Grove, CA), Nov. 1988.

[57] M. J. Vrhel, C. Lee, and M. Unser, "Rapid computation of the continuous wavelet transform by oblique projections," *IEEE Trans. Signal Processing*, vol. 45, no. 4, pp. 891–900, Apr. 1997.

[58] C. Lee, M. Eden, and M. Unser, "High-quality image resizing using oblique projection operators," *IEEE Trans. Image Processing*, vol. 7, no. 5, pp. 670–692, May 1998.

[59] B. M. Hochwald and W. Sweldens, "Space-time modulation for unknown fading," in *Digital Wireless Communication*, R. M. Rao, S. A. Dianat, and M. D. Zoltowski, Eds., June 1999, vol. 3708 of *SPIE Proc.*

[60] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 543–564, 2000.

[61] M. A. Neumark, "On a representation of additive operator set functions," *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, vol. 41, pp. 359–361, 1943.

[62] A. Peres, "Neumark's theorem and quantum inseparability," *Found. Phys.*, vol. 20, no. 12, pp. 1441–1453, 1990.

[63] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 314–366, 1952.

[64] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, New York: Academic Press, 1980.

[65] J. J. Benedetto, "Irregular sampling and frames," in *Wavelets — A Tutorial in Theory and Applications*, C. K. Chui, Ed., pp. 445–507. Boca Raton, FL: CRC Press, 1992.

[66] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, no. 4, pp. 628–666, Dec. 1989.

[67] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, Sep. 1990.

[68] Y. C. Eldar and G. D. Forney, Jr., "Optimal tight frames and quantum measurement," *IEEE Trans. Inform. Theory*, to appear; also available at `http://xxx.lanl.gov/abs/quant-ph/0106070`.

[69] I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia, PA: SIAM, 1992.

[70] H. Bölcskei, *Oversampled Filter Banks and Predictive Subband Coders*, Ph.D. thesis, Vienna University of Technology, Nov. 1997.

[71] H. Bölcskei and A. J. E. M. Janssen, "Gabor frames, unimodularity, and window decay," *J. Fourier Analys. Appl.*, vol. 6, no. 3, pp. 255–276, 2000.

[72] A. J. E. M. Janssen and H. Bölcskei, "Equivalence of two methods for constructing tight Gabor frames," *IEEE Signal Processing Lett.*, vol. 7, pp. 79–82, Apr. 2000.

[73] Y. Meyer, "Ondelettes et functions splines," Dec. 1986, Seminaire EDP, Ecole Polytechnique, Paris, France.

[74] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, pp. 909–996, 1988.

[75] M. Unser and A. Aldroubi, "Families of multiresolution and wavelet spaces with optimal properties," *Numer. Funct. Anal. Optimiz.*, vol. 14, pp. 417–446, 1993.

[76] Y. C. Eldar and H. Bölcskei, "Geometrically uniform frames," submitted to *IEEE Trans. Inform. Theory*, August 2001; also available at `http://arXiv.org/abs/math.FA/0108096`.

[77] G. D. Forney, Jr., "Geometrically uniform codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1241–1260, Sep. 1991.

[78] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. 28, pp. 55–67, 1982.

[79] D. Raphaeli, "On multidimensional coded modulations having uniform error property for generalized decoding and flat-fading channels," *IEEE Trans. Commun.*, vol. 46, pp. 34–40, 1998.

[80] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures," *Appl. Comp. Harmonic Analys.*, vol. 10, pp. 203–233, May 2001.

[81] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Processing Mag.*, pp. 22–38, Nov. 1999.

[82] M. Unser and J. Zerubia, "Generalized sampling: Stability and performance analysis," *IEEE Trans. Signal Processing*, vol. 45, no. 12, pp. 2941–2950, Dec. 1997.

[83] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques: Part I — interpolators and projectors," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2783–2795, Oct. 1999.

[84] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques: Part II — wavelets," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2796–2806, Oct. 1999.

[85] A. Aldroubi, "Portraits of frames," *Proc. Amer. Math. Soc.*, vol. 123, pp. 1661–1668, 1995.

[86] Y. C. Eldar and G. D. Forney, Jr., "On measurements and density operators," in preparation.

[87] S. K. Berberian, *Introduction to Hilbert Space*, New York, NY: Oxford Univ. Press, 1961.

[88] P. R. Halmos, *Introduction to Hilbert Space*, New York, NY: Chelsea Publishing Company, second edition, 1957.

[89] A. Aldroubi and M. Unser, "Sampling procedures in function spaces and asymptotic equivalence with Shannon's sampling theory," *Numer. Funct. Anal. Optimiz.*, vol. 15, pp. 1–21, Feb. 1994.

[90] K. Hoffman and R. Kunze, *Linear Algebra*, Englewood Cliffs, NJ: Prentice Hall, Inc., second edition, 1971.

[91] S. Kayalar and H. L. Weinert, "Oblique projections: Formulas, algorithms, and error bounds," *Math. Contr. Signals Syst.*, vol. 2, no. 1, pp. 33–45, 1989.

[92] A. Aldroubi, "Oblique projections in atomic spaces," *Proc. Amer. Math. Soc.*, vol. 124, no. 7, pp. 2051–2060, 1996.

[93] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Baltimore MD: Johns Hopkins Univ. Press, third edition, 1996.

[94] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Providence, RI: AMS, 1969.

[95] M. A. Artin, *Algebra*, Upper Saddle River, NJ: Prentice Hall, Inc., 1991.

[96] A. Aldroubi, Private communication.

[97] R. D. Milne, "An oblique matrix pseudoinverse," *SIAM J. Appl. Math.*, vol. 16, no. 5, pp. 931–944, Sep. 1968.

[98] P. L. Ainsleigh, "Observations on oblique projectors and pseudoinverses," *IEEE Trans. Signal Processing*, vol. 45, no. 7, pp. 1886–1889, July 1997.

[99] A. Peres, *Quantum Theory: Concepts and Methods*, Boston: Kluwer, 1995.

[100] J. Preskill, *Notes on Quantum Computation*, available at `http://www.theory.caltech.edu/people/preskill/ph229`, 1997.

[101] J. J. Sakurai, *Modern Quantum Mechanics*, Addison-Wesley, Inc., 1985.

[102] P. Hausladen and W. K. Wootters, "A 'pretty good' measurement for distinguishing quantum states," *J. Mod. Opt.*, vol. 41, pp. 2385–2390, 1994.

[103] H. Barnum and E. Knill, "Reversing quantum dynamics with near-optimal quantum and classical fidelity," `http://xxx.lanl.gov/abs/quant-ph/0004088`.

[104] P. Hausladen, R. Josza, B. Schumacher, M. Westmoreland, and W. K. Wootters, "Classical information capacity of a quantum channel," *Phys. Rev. A*, vol. 54, pp. 1869–1876, Sep. 1996.

[105] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, New York, NY: John Willey and Sons, 1968.

[106] R. R. Coifman, Y. Meyer, S. R. Quake, and M. V. Wickerhauser, "Signal processing and compression with wavelet packets," in *Proc. Conf. Wavelets and Applications (Toulouse, France)*, pp. 77–93. June 1992.

[107] R. R. Coifman and M. V. Wickerhauser, "Wavelets and adapted waveform analysis," in *Wavelets: Mathematics and Applications*, J. J. Benedetto and M. Frazier, Eds., pp. 399–423. Boca Raton, FL: CRC Press, 1992.

[108] S. O. Aase, J. H. Husøy, K. Skretting, and K. Engan, "Optimized signal expansions for sparse representations," *IEEE Trans. Signal Processing*, vol. 49, no. 5, pp. 1087–1096, May 2001.

[109] N. J. Munch, "Noise reduction in tight Weyl-Heisenberg frames," *IEEE Trans. Inform. Theory*, vol. 38, pp. 608–616, Mar. 1992.

[110] Z. Cvetković and M. Vetterli, "Overcomplete expansions and robustness," in *Proc. IEEE TFTS-96*, Paris, France, June 1996, pp. 325–328.

[111] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathcal{R}^N$: Analysis, synthesis, and algorithms," *IEEE Trans. Inform. Theory*, vol. 44, pp. 16–31, Jan. 1998.

[112] A. Aldroubi and K. Gröchenig, "Non-uniform sampling and reconstruction in shift-invariant spaces," *Siam Rev.*, to appear.

[113] G. Kaiser, *A Friendly Guide to Wavelets*, Boston: Birkhauser, 1994.

[114] K. Engan, S. O. Aase, and J. H. Husøy, "Designing frames for matching pursuit algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP-1998), (Seattle, WA)*, 1998, pp. 1817–1820.

[115] A. J. E. M. Janssen and T. Strohmer, "Characterization and computation of canonical tight windows for Gabor frames," `http://xxx.lanl.gov/abs/math.FA/0010245`.

[116] D. Slepian, "Group codes for the Gaussian channel," *Bell Syst. Tech. J.*, pp. 575–602, Apr. 1968.

[117] M. A. Armstrong, *Groups and Symmetry*, New York: Springer-Verlag, 1988.

[118] A. Terras, *Fourier Analysis on Finite Groups and Applications*, Cambridge, UK: Cambridge Univ. Press, 1999.

[119] D. Gabor, "Theory of communication," *J. Inst. Elec. Eng.*, vol. 93, pp. 429–439, Nov. 1946.

[120] H. G. Feichtinger and T. Strohmer, Eds., *Gabor Analysis: Theory, Algorithms, and Applications*, Birkhäuser, 1998.

[121] A. Shokrollahi, B. Hassibi, B. M. Hochwald, and W. Sweldens, "Representation theory for high-rate multiple-antenna code design," submitted to *IEEE Trans. Inform. Theory*, 2000.

[122] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 2146–2157, Aug. 1994.

[123] D. A. Rennels, "Fault-tolerant computing — concepts and examples," *IEEE Trans. Computers*, vol. C-33, pp. 1116–1129, Dec. 1984.

[124] P. P. Vaidyanathan and B. Vrcelj, "Biorthogonal partners and applications," *IEEE Trans. Signal Processing*, vol. 49, no. 5, pp. 1013–1027, May 2001.

[125] P. P. Vaidyanathan, "Generalizations of the sampling theorem: Seven decades after Nyquist," *IEEE Trans. Circuit Syst.*, to appear.

[126] N. Macon and A. Spitzbart, "Inverses of Vandermonde matrices," *American Mathematical Monthly*, vol. 65, no. 2, pp. 95–100, Feb. 1958.

[127] F. D. Parker, "Inverses of Vandermonde matrices," *American Mathematical Monthly*, vol. 71, no. 4, pp. 410–411, Apr. 1964.

[128] A. Papoulis, "A new algorithm in spectral analysis and bandlimited extrapolation," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 735–742, 1975.

[129] H. Stark and Y. Yang, *Vector Space Projections*, John Wiley and Sons, Inc., 1998.

[130] A. J. E. M. Janssen, "The Zak transform and sampling theorems for wavelet subspaces," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3360–3364, Dec. 1993.

[131] J. A. Hogan and J. Lakey, "Sampling and aliasing without translation-invariance," *Proceedings of the 2001 Workshop on Sampling Theory and Applications, SampTA'01*, pp. 61–66, May 2001.

[132] Y. C. Eldar and A. V. Oppenheim, "Filter bank reconstruction of bandlimited signals from nonuniform and generalized samples," *IEEE Trans. Signal Processing*, vol. 48, pp. 2864–2875, Oct. 2000.

[133] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw Hill, Inc., third edition, 1991.

[134] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.

[135] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, Oct. 1998.

[136] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.

[137] E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Norwell, MA: Kluwer Academic Publishers, second edition, 1994.

[138] I. Daubechies and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 5, no. 27, pp. 1271–1283, May 1986.

[139] J. Candy and G. Temes, "Oversampling methods for A/D and D/A conversion," in *Oversampling Delta-Sigma Data Converters*, pp. 1–25. New York, NY: IEEE Press, 1992.

[140] P. M. Aziz, H. V. Sorenson, and J. Van der Spiegel, "An overview of sigma-delta converters," *IEEE Signal Processing Mag.*, pp. 61–84, Jan. 1996.

[141] N. J. Higham, "Computing the polar decomposition — with applications," *SIAM J. Sci. Stat. Comput.*, vol. 7, pp. 1160–1174, 1986.

[142] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge, UK: Cambridge Univ. Press, 1985.

[143] P. Zielinski and K. Zietak, "The polar decomposition — properties, applications and algorithms," *Matematyka Stosowana*, vol. 38, pp. 23–40, 1995.

[144] A. A. Dubrulle, "An optimum iteration for the matrix polar decomposition," *Electron. Trans. Numer. Anal.*, vol. 8, pp. 21–25, 1999.

[145] O. Besson and P. Stoica, "Exponential signals with time-varying amplitude: Parameter estimation via polar decomposition," *Signal Processing*, vol. 66, pp. 27–43, 1998.

[146] A. Markiewicz, "Simultaneous polar decomposition of rectangular complex matrices," *Linear Algebra Appl.*, vol. 289, pp. 279–284, 1999.

[147] J. C. Gower, "Multivariate analysis: Ordination, multidimensional scaling and allied topics," in *Handbook of Applicable Mathematics*, vol. 6, pp. 727–781. Chichester, UK: Wiley, 1984.

[148] B. F. Green, "The orthogonal approximation of an oblique structure in factor analysis," *Psychometrika*, vol. 17, pp. 429–440, 1952.

[149] P. H. Schonemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, pp. 1–10, 1966.

[150] C. W. Helstrom, "Bayes-cost reduction algorithm in quantum hypothesis testing," *IEEE Trans. Inform. Theory*, vol. 28, pp. 359–366, Mar. 1982.

[151] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Chichester, U.K.: Wiley, 1985.

[152] G. J. McLachlan and K. E. Basford, *Mixture Models*, New York: Marcel Dekker, 1988.

[153] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*, London, U.K.: Chapman and Hall, 1981.

[154] D. D'Addio, S. Giannatempo, and G. Galati, "Generation of K-distributed random variables," *Trans. Society Comp. Simulation*, vol. 5, pp. 159–174, 1988.

[155] J. J. Atick and A. N. Redlich, "Convergent algorithm for sensory receptive filed development," *Neurak Comp.*, vol. 5, 1993.

[156] G. H. Golub and C. F. Van Loan, "An analysis of the total least-squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 4, pp. 883–893, 1979.

[157] S. Van Huffel and J. Vandewalle, *The Total Least-Squares Problem: Computational Aspects and Analysis*, vol. 9 of *Frontier in Applied Mathematics*, Philadelphia, PA: SIAM, 1991.

[158] A. Yeredor, "The extended least-squares criterion: Minimization algorithms and applications," *IEEE Trans. Signal Processing*, vol. 49, no. 1, pp. 74–86, Jan. 2001.

[159] M. H. J. Gruber, *Regression Estimators: A Comparative Study*, San Diego, CA: Academic Press, Inc., 1990.

[160] C. M. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proc. 3rd Berkeley Symposium Math. Stat. Prob.*, vol. 1, pp. 197–206.

[161] W. James and C. M. Stein, "Estimation with quadratic loss," *Proc. 4th Berkeley Symposium Math. Stat. Prob.*, vol. 1, pp. 361–379.

[162] J. L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. 32, no. 1, pp. 33–51, Feb. 1967.

[163] S. Ulukus and R. D. Yates, "Optimum multiuser detection is tractable for synchronous CDMA systems using m-sequences," *IEEE Commun. Letters*, vol. 2, pp. 89–91, Apr. 1998.

[164] K. L. Chung, *A Course in Probability Theory*, San Diego, CA: Academic Press, 2001.

[165] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, New York: Academic Press, 1980.

[166] Z. D. Bai and Y. Q. Yin, "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix," *Annals of Probability*, vol. 21, pp. 1275–1294, 1993.

[167] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999.