Blind Minimax and Maximum Set Estimators: Improving on Least-Squares Estimation

Research Thesis

Submitted in Partial Fulfillment of the Requirements for
the Degree of Master of Science in Electrical Engineering

Zvika Ben-Haim

Submitted to the Senate of the Technion—Israel Institute of Technology

Tevet 5766 Haifa January 2006

Revised and Corrected

Nissan 5766 Haifa May 2006

Acknowledgements

I feel extremely fortunate for having had the opportunity to work under the supervision of Prof. Yonina Eldar. Her advice, wit and intuition were invariably insightful, both in advancing our research and in assisting my first steps in the academic world. Looking back, I feel that our joint work was exactly what I would have wanted to do; and for this I am grateful.

I would also like to thank all my friends from Yonina's research group. They heard my "LS-domination" mantra far too many times, and their comments were both constructive and encouraging. I am particularly indebted to Tsvika Dvorkind and Ami Wiesel, who provided several original contributions which proved crucial in the course of my work.

My wife Yael and the rest of my family deserve my gratitude for their patience and support. Special thanks go to my father Yakov, with whom I held many fruitful discussions in the initial stages of the thesis.

Contents

Abstract **1**

Notation **3**

1 Introduction **5**

1.1 Problem Statement 5

1.2 Previous Work 5

1.3 Main Contributions 7

1.4 Thesis Outline 8

2 Background **9**

2.1 Parameter Estimation 9

2.1.1 Bayesian vs. Deterministic Estimation 9

2.1.2 The System Model 10

2.2 The Least-Squares Estimator 12

2.2.1 The Maximum Likelihood Criterion 12

2.2.2 The Gauss-Markov Theorem 13

2.2.3 Measurement Error Minimization 14

2.3 Stein's Phenomenon 15

2.3.1 Dominance and Admissibility 15

2.3.2 The James-Stein Estimator 17

2.3.3 Other LS-Dominating Estimators 20

2.4 Tikhonov Regularization 23

3 Minimax Estimation **27**

3.1 Problem Statement 27

3.2 Examples of Minimax Estimators 29

3.2.1 Minimax MSE Estimators . . . 29
3.2.2 Minimax Regret Estimators . . . 31
3.2.3 Noncentral Estimation . . . 32
3.3 Conditional Dominance . . . 33

4 Blind Minimax Estimation **35**
4.1 The Blind Minimax Approach . . . 35
4.2 The Spherical Blind Minimax Estimator . . . 36
4.3 The Ellipsoidal Blind Minimax Estimator . . . 39
4.4 Relation to Stein-type Estimation . . . 45
4.5 Numerical Results . . . 46
4.5.1 Comparison with the LS Approach . . . 47
4.5.2 Comparison with Bock's Estimator . . . 48
4.5.3 Comparison with Tikhonov Regularization . . . 50
4.6 Discussion . . . 51

5 Maximum Set Estimation **55**
5.1 Maximum Parameter Set Estimation . . . 56
5.1.1 A Useful Special Case . . . 56
5.1.2 General Form of MPS Estimators . . . 59
5.1.3 Relation to Minimax Estimation . . . 61
5.1.4 Linear MSE Estimators . . . 62
5.1.5 Linear Regret Estimators . . . 66
5.2 Maximum Noise Level Estimation . . . 67
5.3 Application: Channel Estimation . . . 70
5.4 Discussion . . . 74

Bibliography **77**

List of Figures

4.1 MSE vs. SNR for a typical operating condition: effective dimension 5.1, $m = n = 15$. 47
4.2 Estimator(s) achieving lowest MSE, among the five estimators tested ($m = n = 10$) 48
4.3 Estimator MSE vs. condition number; $m = n = 10$, SNR 0 dB 49
4.4 Tikhonov regularization does not dominate the LS estimator 51

5.1 The worst-case error of various minimax MSE channel estimators 72
5.2 Channel estimation error of MPS and LS estimators for various channels 73
5.3 BER for various channels with the LS and MPS channel estimators 74

Abstract

We consider the linear regression problem of estimating an unknown, deterministic parameter vector, observed through colored Gaussian noise. This classical problem is generally solved using the least-squares (LS) estimator. We explore alternatives to this approach, and demonstrate analytically that our techniques outperform the LS method in terms of mean-squared error (MSE).

We begin by presenting blind minimax estimators (BMEs), which consist of a minimax estimator on a parameter set which is itself estimated from measurements. We demonstrate analytically that the BMEs dominate the least-squares technique, i.e., they always achieve lower MSE. Furthermore, we explore the relation of this approach to the James-Stein estimator, and demonstrate its advantage over various Stein-type methods.

We next consider the problem of finding a linear estimator whose MSE does not exceed a given maximum. We develop estimators guaranteeing the required error for as large a parameter set as possible and for as large a noise level as possible. We then discuss methods for finding these estimators and demonstrate that in many cases, the proposed estimators outperform the LS approach.

Notation

Throughout the document, scalars are denoted by italicized lowercase letters, as in m; vectors are denoted by boldface lowercase letters, as in $\mathbf{x}$; and matrices are denoted by boldface uppercase letters, as in $\mathbf{A}$. The ith component of a vector $\mathbf{x}$ is denoted x_i.

$\mathbf{b}(\hat{\mathbf{x}}, \mathbf{x})$	Bias of an estimator, defined in (2.7)
$\mathbf{C_w}$	Measurement noise covariance matrix
d	Effective dimension, defined in (2.22)
$\text{diag}(\mathbf{v})$	Diagonal matrix whose diagonal elements are the elements of vector $\mathbf{v}$
$E\{\mathbf{v}\}$	Expectation of a random vector $\mathbf{v}$
$\mathbf{G}$	Linear estimator matrix, $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$
$\mathbf{H}$	System transformation matrix (size $n \times m$)
$\mathbf{I}$	Identity matrix
$\hat{L}$	Parameter robustness, defined in (5.2)
$\text{MSE}(\hat{\mathbf{x}}, \mathbf{x})$	Mean-squared error, defined in (2.2)
$\Pr\{A\}$	Probability of an event A
$\mathbf{Q}$	Defined as $\mathbf{H}^* \mathbf{C}_\mathbf{w}^{-1} \mathbf{H}$
$\mathcal{S}$	Parameter set (a set known to contain the parameter vector $\mathbf{x}$)
$\text{sgn}(a)$	Sign of a number a (1 if positive, -1 if negative, 0 if zero)
$\text{Tr}(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$\text{Var}(\hat{\mathbf{x}})$	Variance of a random vector, defined in (2.8)
$\mathbf{w}$	Measurement noise with zero mean and known covariance $\mathbf{C_w}$
$\mathbf{x}$	Unknown parameter vector (length m)
$\hat{\mathbf{x}}$	Estimate of $\mathbf{x}$ from measurements $\mathbf{y}$
$\hat{\mathbf{x}}_{\text{EBM}}$	Ellipsoidal blind minimax estimator, defined in Section 4.3
$\hat{\mathbf{x}}_{\text{JS}}$	James-Stein estimator, defined in (2.17)
$\hat{\mathbf{x}}_{\text{LS}}$	Least-squares estimator, defined in (2.4)

$\hat{\mathbf{x}}_{\mathrm{M}}$	Minimax estimator, defined in (3.1)
$\hat{\mathbf{x}}_{\mathrm{NL}}$	Maximum noise level estimator, defined in (5.44)
$\hat{\mathbf{x}}_{\mathrm{PS}}$	Maximum parameter set estimator, defined in (5.3)
$\hat{\mathbf{x}}_{\mathrm{SBM}}$	Spherical blind minimax estimator, defined in Section 4.2
$\mathbf{y}$	Measurement vector (length n)
$\epsilon(\hat{\mathbf{x}}, \mathbf{x})$	Risk function, measuring the discrepancy between $\hat{\mathbf{x}}$ and $\mathbf{x}$
ϵ_0	Mean-squared error of the least-squares estimator, defined in (2.5)
ϵ_m	Maximum allowed estimation risk, defined in Section 5.1
$\lambda_{\max}(\mathbf{A})$	Largest eigenvalue of a matrix $\mathbf{A}$
$\lambda_{\min}(\mathbf{A})$	Smallest eigenvalue of a matrix $\mathbf{A}$
$\hat{\sigma}^2$	Noise robustness, defined in (5.43)
$\mathbf{A}^*$	Hermitian conjugate, i.e., transpose and complex conjugate of $\mathbf{A}$
$\mathbf{A} \succeq 0$	Matrix $\mathbf{A}$ is positive semidefinite
$\mathbf{A} \succeq \mathbf{B}$	Matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite
$\mathbf{A}^{1/2}$	For $\mathbf{A} \succeq 0$, indicates the unique matrix satisfying $(\mathbf{A}^{1/2})^2 = \mathbf{A}$ and $\mathbf{A}^{1/2} \succeq 0$
$\|\mathbf{x}\|$	ℓ_2-norm of a vector $\mathbf{x}$, i.e., $(\mathbf{x}^*\mathbf{x})^{1/2}$
$\|\mathbf{x}\|_{\mathbf{T}}$	$\mathbf{T}$-norm of a vector $\mathbf{x}$, i.e., $(\mathbf{x}^*\mathbf{T}\mathbf{x})^{1/2}$, for some positive-definite matrix $\mathbf{T}$.
$\mathbf{0}_k$	k-vector containing only zeroes
$\mathbf{1}_k$	k-vector containing only ones

Chapter 1

Introduction

1.1 Problem Statement

The problem of estimating parameters from noisy measurements has countless applications in science and engineering. Such estimation problems are typically modelled either in a Bayesian setting, in which a prior distribution on the parameters is assumed, or in a deterministic setting, in which no prior exists [1, 2]. Our work focuses on the deterministic estimation problem. We further assume a linear regression model, in which the measurements $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ are linear combinations of the parameter vector $\mathbf{x}$ with additive noise $\mathbf{w}$. Here the transformation matrix $\mathbf{H}$ and the noise covariance $E\{\mathbf{w}\mathbf{w}^*\}$ are assumed to be known. The success of an estimate $\hat{\mathbf{x}}$ is quantified by its risk, which measures the distance between $\hat{\mathbf{x}}$ and the true value $\mathbf{x}$; the most common risk function is the mean-squared error (MSE).

The primary difficulty in obtaining low risk is that the risk function typically depends on the unknown value of the parameter $\mathbf{x}$. Thus, an estimator may obtain low risk for some values of $\mathbf{x}$, and high risk for other values. Furthermore, no single estimator achieves optimal risk for all values of $\mathbf{x}$. Nevertheless, a particular estimator can be said to dominate, or improve upon, a different estimator, if its risk is lower for *all* values of $\mathbf{x}$. This work is aimed at obtaining novel estimators which dominate standard solutions to the estimation problem.

1.2 Previous Work

The deterministic estimation problem was first addressed by Gauss and Legendre, who independently proposed the least-squares (LS) estimator in the beginning of the 19th century [3, 4]. Several lines of reasoning can be used to support the LS approach. One argument shows

that the LS estimator minimizes the squared error between the measurements $\mathbf{y}$ and the transformed estimate $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}$. It is also well-known that the LS estimator is the maximum likelihood estimator for Gaussian noise. However, neither of these criteria are directly related to the MSE, or to any other measure of the distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$. Another property of the LS estimator is that it is the linear, unbiased estimator achieving minimal MSE. However, although linearity and unbiasedness may be intuitively appealing properties, they have no relation to the primary goal at hand, namely, achieving low estimation error.

Our work stems from the technique of linear minimax estimation for bounded parameter sets [5–7]. These *minimax estimators* are designed for the situation in which the parameter vector $\mathbf{x}$ is known to lie within a bounded parameter set $\mathcal{S}$; specifically, they minimize the worst-case risk among all possible values of $\mathbf{x}$ within $\mathcal{S}$. Such estimators have been studied extensively in the past, and closed forms are known for many types of parameter sets and risk functions. We formalize the usefulness of minimax estimators by providing a *conditional dominance* theorem, which states that, for any bounded set $\mathcal{S}$, a minimax MSE estimator achieves lower estimation error than the LS estimator, for all values of $\mathbf{x}$ in $\mathcal{S}$.

Although minimax estimators outperform the LS approach, this may be attributed to the fact that they are designed for a particular parameter set; this information is not available to the LS estimator. Similarly, several techniques have been designed for estimation under other special cases. For example, Tikhonov regularization [8, 9] was developed for ill-posed problems, whence the LS estimator is numerically unstable. These techniques are generally inappropriate as general-purpose replacements of the LS estimator. Indeed, it was long believed that the LS estimator was *admissible*, i.e., no estimator could achieve lower MSE for *all* parameter values. Surprisingly, the inadmissibility of the LS approach for Gaussian noise was demonstrated by Stein in 1956, some 150 years after the technique was first introduced [2, 10].

Both the LS approach and Gaussian noise are extremely common in many scientific fields. Therefore, a technique which improves upon the LS estimator for all values of $\mathbf{x}$, under the assumption of Gaussian noise, is of great practical importance. During the 1960s and 1970s, several such estimators were constructed. Early work on so-called LS-dominating estimators considered the independent, identical-distribution (i.i.d.) case, for which $\mathbf{H} = \mathbf{I}$ and the noise is Gaussian and white. Among these, the James-Stein estimator [2, 11] is the best-known example; another is the Thompson estimator [12, 13]. Various extensions of the James-Stein estimator were later constructed for the non-i.i.d. case of colored Gaussian noise [14–16]. Of these, Bock's estimator [15] is quoted most often [17, 18]. However, none of the approaches has become a

standard alternative to the LS estimator, and they are rarely used in signal processing [18]. One reason for this is that the estimators are poorly justified and seem counterintuitive, and as such they are sometimes regarded with skepticism (see discussion following [19]). Another reason is that many of these approaches (including the James-Stein estimator and Bock's extension) result in shrinkage estimators, consisting of a gain factor multiplying the LS estimate. While this can certainly be used to reduce MSE, such estimators are inappropriate for some applications, in which a gain factor has no effect on final estimation quality.

1.3 Main Contributions

The thesis originated from the aforementioned conditional dominance theorem, which demonstrates the advantage of minimax estimation when $\mathbf{x}$ is known to belong to a bounded parameter set. Our proof of the conditional dominance theorem led us to search for extensions of the minimax principle, which would be applicable in alternative settings. This resulted in two extensions of the minimax technique, which form the main body of this work. We refer to these novel techniques as blind minimax estimation and maximum set estimation.

Blind Minimax Estimation. In our work, we use minimax estimators to obtain novel LS-dominating estimators, using a simple, intuitive principle called the blind minimax approach [20–22]. Many blind minimax estimators (BMEs) reduce to Stein-type estimators in the i.i.d. case, and they continue to dominate the LS estimator in the non-i.i.d. case as well. Unlike Bock's estimator, BMEs may be constructed so that they are non-shrinkage, if this is required. Furthermore, extensive simulations show that BMEs usually outperform Bock's estimator by a considerable margin.

Unlike minimax estimation, BMEs do not require the assumption of a parameter set $\mathcal{S}$. Instead, the blind minimax approach consists of a two-stage estimation process. In the first stage, the set $\mathcal{S}$ is estimated from the measurements. In the second stage, a minimax estimator for $\mathcal{S}$ is used to estimate the parameter itself. The result may be viewed as a simple estimator, independent of this two-stage construction process. Indeed, our LS-dominance proofs are not related to the method by which the estimators are generated. In particular, the dominance results do not depend on the parameter actually lying within the estimated parameter set. However, the blind minimax technique is useful in that it provides a framework whereby many different estimators can be generated, and provides insight into the mechanism by which these

estimators outperform the LS estimator.

Maximum Set Estimation. As a second application of minimax estimation theory, we seek a linear estimator satisfying given maximum error requirements. The maximum error ϵ_m is a design choice, based on known properties of the system at hand. For example, ϵ_m may be chosen to guarantee a required signal-to-noise ratio (SNR) at the estimator output. In particular, ϵ_m may be chosen to be smaller than the error obtained by the LS estimator. Motivated by information-gap decision theory [23, 24], we seek a *maximum set estimator*, namely, a linear estimator guaranteeing an error not exceeding ϵ_m, for as large a range of conditions as possible [25,26]. Thus, we may seek an estimator achieving the required error for as large a parameter set as possible; alternatively, we may seek to maximize the noise level for which error requirements are satisfied. In this way, it is possible to outperform the LS estimator by any desired amount — although choosing excessive error requirements reduces the set of supported operating conditions.

We demonstrate the relation of this problem to minimax estimation. In particular, we show that under suitable conditions, any maximum set estimator is also a minimax estimator, and vice versa. Thus, one can find a maximum set estimator for a given problem by obtaining a minimax estimator whose worst-case error is ϵ_m. This allows us to use the substantial body of knowledge on minimax estimation in studying maximum set estimators, resulting in closed-form estimators for many types of maximum set estimation problems.

1.4 Thesis Outline

The remainder of this thesis is organized as follows. In Chapter 2, we review classical results in estimation theory, including the LS estimator, various Stein-type estimators, and Tikhonov regularization. Chapter 3 discusses minimax estimation, presenting several known examples of minimax estimators as well as the aforementioned conditional dominance theorem. Chapter 4 is devoted to blind minimax estimation; it develops several types of BMEs, proves their dominance over the LS technique, and presents simulations comparing them with other estimators. Finally, Chapter 5 presents the maximum set estimation approach, derives methods for obtaining closed forms of such estimators, and compares them with the LS estimator.

Chapter 2

Background

In this chapter, we review some preliminary estimation concepts which will be used throughout the remainder of the work. We begin with a general overview of the parameter estimation problem, defining the setting and objective (Section 2.1). We next discuss the advantages and limitations of several well-known estimators, including the least-squares estimator (Section 2.2), Stein-type estimators (Section 2.3), and Tikhonov regularization (Section 2.4).

2.1 Parameter Estimation

In a parameter estimation problem, one is given an observation vector, from which an unknown parameter vector must be estimated. To this end, a model describing the relation between observations and parameters is required. The model typically includes some uncertainties, e.g., random noise added to the measurements. Furthermore, some formal estimation goal must be defined, such as obtaining low mean-squared error. In this section, we provide definitions of the setting and model in which our work is to take place.

2.1.1 Bayesian vs. Deterministic Estimation

It is helpful to distinguish between two estimation settings: the Bayesian scheme and the deterministic (or frequentist) scheme [1, 2]. Although sometimes similar in terminology, these approaches differ substantively. In the Bayesian model, the parameter vector is random, and its distribution (called the prior) is usually known. Estimators are judged based on average performance over different realizations of the parameter. As a result, the quality of an estimator is a number representing its average performance over all possible parameter values; any two estimators can be compared in this way.

The situation is more complex in the frequentist setting, which is the framework we will adopt in this work. Here, the parameter vector is modelled as a deterministic value. Nothing is assumed about the parameter in advance; the estimate is based solely on the measurements. Thus, this is an information-sparse approach. One consequence of this setting is that estimator performance is difficult to judge: a particular estimator may be suitable for some parameter values and inaccurate for others. With no way of deciding which condition is more probable, we cannot combine the differing performance levels to a single value representing estimator quality. We will return to this point in Subsection 2.3.1, in which a partial ordering among estimators is defined.

The differences between the Bayesian and deterministic viewpoints are deep, and each is appropriate in a different scenario. Let us demonstrate this by examining two classical applications of estimation theory. Consider first the problem of symbol detection in a communication system. Here, the problem is to estimate the value of a transmitted symbol in a noisy channel. In this case, possible symbol values and their respective probabilities are completely known, implying that the Bayesian approach is a suitable model for such problems. On the other hand, consider the problem of estimating an unknown physical constant, such as the mass of an electron or the noise figure of a given receiver. There is only one correct value for such quantities, rather than a spectrum of possible values. It is therefore undesirable to assign a probability distribution to this parameter, and deterministic estimation techniques should be applied.

Some devout believers of either Bayesian or frequentist philosophies claim that any estimator can be expressed as a solution to a properly stated problem stemming from their approach. However, an estimator's mathematical formula is a consequence of the problem setting, and not vice versa. As we have seen, this setting can inherently be either deterministic or Bayesian. In the remainder of this work, we consider the deterministic estimation problem. This is done with the understanding that such a setting is appropriate to many real-world problems, though certainly not to all problems.

2.1.2 The System Model

Thus far, we have spoken vaguely about the relation between parameters and measurements. We now turn to explicitly defining this relation in terms of a system model and an estimation goal.

We restrict our attention to the case in which the measurements $\mathbf{y} \in \mathbb{C}^n$ are described as a

linear function of the parameters $\mathbf{x} \in \mathbb{C}^m$ with additive noise,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}. \tag{2.1}$$

Here, $\mathbf{H} \in \mathbb{C}^{n \times m}$ is a known matrix of full rank m, and $\mathbf{w}$ is zero-mean additive noise, whose covariance $\mathbf{C_w}$ is positive-definite and known. Unless otherwise specified, the noise distribution is unknown; in many cases, however, we will assume that the noise is Gaussian.

The definition of an estimation objective, or goal, is a crucial step of the problem formulation: it often happens that an estimator exhibits excellent performance under one criterion but achieves poor results when the criterion is changed. An estimation goal is often defined in terms of a *risk function* $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$, which measures the discrepancy or error[1] between the estimate $\hat{\mathbf{x}}$ and the true parameter value $\mathbf{x}$. In this section, we describe two different risk functions: the mean-squared error (MSE) and the regret. The choice of a risk function is necessarily a combination of reasonable requirements and historical prejudice. However, we will attempt to justify, insofar as possible, the choice of risk functions in this section.

The most commonly used risk function is the MSE, also called squared-error risk. This is defined as

$$\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = E\left\{ \|\hat{\mathbf{x}} - \mathbf{x}\|^2 \right\}, \tag{2.2}$$

and will be used throughout the majority of this work. Since the parameter vector $\mathbf{x}$ is deterministic, the expectation is taken only over $\hat{\mathbf{x}}$, and the result is a function not only of the estimator used, but also of the unknown value of $\mathbf{x}$.

It is also possible to quantify the estimation error using other measures, such as the *regret* [7], which is a useful measure of the quality of linear estimators. The regret of an estimator $\hat{\mathbf{x}}$ is defined as the difference between the MSE of $\hat{\mathbf{x}}$ and the best MSE obtainable using a linear estimator $\hat{\mathbf{x}}_o = \mathbf{G}(\mathbf{x})\mathbf{y}$ which is a function of $\mathbf{x}$. Because we are limiting the discussion to linear estimators, even an estimator with knowledge of the value of $\mathbf{x}$ cannot achieve zero MSE. By calculating the MSE of $\hat{\mathbf{x}}_o$, it can be shown [7] that the regret is given by

$$\mathrm{Reg}(\hat{\mathbf{x}}, \mathbf{x}) = \mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) - \frac{\mathbf{x}^*\mathbf{x}}{1 + \mathbf{x}^*\mathbf{H}^*\mathbf{C}_\mathbf{w}^{-1}\mathbf{H}\mathbf{x}}. \tag{2.3}$$

Minimizing the regret is an intuitively appealing method for measuring the quality of linear estimators, as it attempts to disregard errors resulting from the limitation of linear estimators. Thus, we shall return to the regret when discussing linear estimators in Chapters 3 and 5.

[1] An alternative approach, which quantifies the error as expressed in the measurements $\mathbf{y}$, is briefly presented in Subsection 2.2.3.

2.2 The Least-Squares Estimator

The problem of estimating a parameter vector from noisy measurements was first addressed in a modern framework by Gauss and Legendre, who worked independently in the beginning of the 19th century. Their solution became known as the least-squares (LS) method [1,3,4], and is given, in our notation, as

$$\hat{\mathbf{x}}_{\mathrm{LS}} = (\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H})^{-1} \mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{y}. \tag{2.4}$$

The LS estimator is, without a doubt, the most common parameter estimation technique used to this day. It has several convenient properties. One such property is linearity, which allows efficient computation of the estimate. Another important property is that the MSE achieved by the LS estimator is constant for all $\mathbf{x}$, and can be calculated from the known matrices $\mathbf{H}$ and $\mathbf{C}_{\mathbf{w}}$; it is given by

$$\epsilon_0 = \mathrm{Tr}((\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H})^{-1}). \tag{2.5}$$

Several lines of reasoning are often cited to justify the use of the LS estimator; these include the maximum likelihood criterion, the Gauss-Markov theorem, and the measurement error minimization criterion. Since these arguments are often used to claim "optimality" of the LS approach, we shall take some time to explore their meaning in depth. In doing so, we note that estimator choice depends on the estimation goal, and recall our goal of minimizing the MSE.

2.2.1 The Maximum Likelihood Criterion

Given a measurement vector and a proposed estimator, one can calculate the probability of obtaining the observed measurements under the assumption that the proposed estimator is correct. The result is referred to as the *likelihood* of the measurements, and the maximum likelihood (ML) criterion seeks estimators optimizing this value. It is straightforward to show that for zero-mean Gaussian noise, the ML approach leads to the LS estimator.

The use of likelihood as a figure of merit is convenient, since it quantifies the performance of an estimator, for a given set of measurements, without requiring knowledge of the true parameter value. However, for the same reason, the ML criterion is not directly related to the estimator's ability to predict parameter values. For example, consider a scalar measurement y which simply equals $x + w$, with the noise w having exponential distribution. In this case, the ML estimate is given by $\hat{x} = y$, despite the fact that the true value x is smaller than y with probability 1. Clearly, the average noise behavior is more relevant than its maximum.

For symmetrical noise distributions, such as the Gaussian distribution, the ML criterion can be viewed as a technique for selecting an unbiased estimate, i.e., an estimate whose expectation equals the true parameter value. Furthermore, in the Gaussian case, the likelihood depends quadratically on the term $\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}$. Thus, finding the ML estimator by differentiating the likelihood function is equivalent to finding a linear unbiased estimator. We thus proceed to discuss the Gauss-Markov approach, which specifically seeks linear unbiased estimators.

2.2.2 The Gauss-Markov Theorem

The Gauss-Markov theorem states that the LS estimator achieves minimum MSE within the class of linear, unbiased estimators. The theorem holds for non-Gaussian noise as well, as long as the noise covariance matrix $\mathbf{C_w}$ is finite and known.

To prove this result, we first note that the MSE of any estimator $\hat{\mathbf{x}}$ is given by

$$\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = \|\mathbf{b}(\hat{\mathbf{x}}, \mathbf{x})\|^2 + \mathrm{Var}(\hat{\mathbf{x}}), \tag{2.6}$$

where

$$\mathbf{b}(\hat{\mathbf{x}}, \mathbf{x}) = E\{\hat{\mathbf{x}}\} - \mathbf{x} \tag{2.7}$$

is the bias and

$$\mathrm{Var}(\hat{\mathbf{x}}) = E\big\{\|\hat{\mathbf{x}} - E\{\hat{\mathbf{x}}\}\|^2\big\} \tag{2.8}$$

is the variance. For unbiased estimators, $\mathbf{b}(\hat{\mathbf{x}}, \mathbf{x}) = \mathbf{0}$, so that minimizing the MSE is equivalent to minimizing the variance. If we further restrict attention to linear estimators by writing $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$, the requirement of zero bias for all $\mathbf{x}$ becomes

$$\mathbf{I} - \mathbf{G}\mathbf{H} = \mathbf{0}. \tag{2.9}$$

Minimizing the variance of $\hat{\mathbf{x}}$, subject to (2.9), results in the LS estimator. The LS estimator is therefore referred to as the best linear unbiased estimator, or BLUE [1]. As mentioned earlier, in the Gaussian case, the linearity restriction can be removed, so that in this case the LS estimator is also the uniformly minimum variance unbiased (UMVU) estimator [2].

The restriction to zero bias is a crucial step in this construction of the LS estimator. We must therefore ask whether the unbiasedness restriction is in order. If an estimator is unbiased, then using the estimator repeatedly, with the same parameter $\mathbf{x}$ but with different noise realizations, results in a set of estimates whose average tends to the true value $\mathbf{x}$. This would have been a useful property if we were to confront a series of estimation problems, at the end of which

the estimates are all averaged into a final, combined result. However, if this were truly the scenario, one would do better by constructing a single estimation problem which combines all measurements. Thus, the zero bias restriction, although intuitively appealing, is not directly related to achieving low risk. Indeed, in many cases, it is possible to drastically reduce the variance by introducing a small bias; by (2.6), this reduces the total MSE. It is still possible to define one's problem as the search for the unbiased estimator achieving minimum MSE; but in light of the above discussion, such a problem statement is somewhat arbitrary.

A different objection to biased estimation is the breaking of symmetry. If we are to add bias to an estimator, how are we to decide on the desired bias direction? This claim implicitly assumes that the unbiased solution is also the most "symmetric" in terms of risk. In fact, as we shall see in Subsection 2.3.2, the unbiased estimator consistently overestimates the true parameter value, just as the ML estimate in the example of Subsection 2.2.1 was an overestimate in the case of exponential noise. The answer to this objection, then, is that bias should be used to decrease, or shrink, the unbiased estimate.

2.2.3 Measurement Error Minimization

Measurement error is the discrepancy between the actual observations and the expected value of the observations assuming the estimate is accurate. For example, the squared measurement error is defined as

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2, \tag{2.10}$$

where $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}$ is the expected observation vector when the parameter equals $\hat{\mathbf{x}}$. This differs substantively from the risk minimization approach discussed previously, which is defined as the mismatch between the true (unknown) parameter vector and its estimate.

The LS estimator can be derived from the measurement error criterion by first whitening the noise. This is done by multiplying (2.1) by $\mathbf{C}_{\mathbf{w}}^{-1/2}$, obtaining

$$\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{w}}, \tag{2.11}$$

where $\tilde{\mathbf{y}} = \mathbf{C}_{\mathbf{w}}^{-1/2}\mathbf{y}$, $\tilde{\mathbf{H}} = \mathbf{C}_{\mathbf{w}}^{-1/2}\mathbf{H}$, and $\tilde{\mathbf{w}} = \mathbf{C}_{\mathbf{w}}^{-1/2}\mathbf{w}$. The result (2.11) is referred to as the *whitened* estimation problem, since $\tilde{\mathbf{w}}$ is white noise with covariance $\mathbf{I}$, and $\tilde{\mathbf{H}}$ remains full-rank. Thus, $\tilde{\mathbf{y}}$ is an alternative measurement vector, which is equivalent to $\mathbf{y}$: any estimator based on $\tilde{\mathbf{y}}$ can be converted to an estimator based on $\mathbf{y}$ by first multiplying its measurements by $\mathbf{C}_{\mathbf{w}}^{-1/2}$. By whitening the noise as in (2.11) and then differentiating (2.10) with respect to $\hat{\mathbf{x}}$, it

can be shown that the LS estimator is the estimator achieving minimum squared measurement error.

Measurement error does not directly depend on the unknown parameter $\mathbf{x}$, which makes it a convenient estimation tool. But the distinction between measurement error and risk is meaningful in the underlying estimation context as well. Measurement error is applicable when the goal is to obtain a reconstruction of the observations. For example, in an image compression scheme, we may seek a low-dimensional parameter vector $\hat{\mathbf{x}}$ which reconstructs the observed image $\mathbf{y}$ as closely as possible; the "true" value $\mathbf{x}$ is irrelevant. In estimation problems, however, we wish to find the underlying parameter vector itself, and treat the measurements merely as chance occurrences from which something about the value of $\mathbf{x}$ may be learned.

In an estimation context, then, it is more appropriate to minimize the risk than the measurement error. Unfortunately, measurement error is not necessarily indicative of risk: large measurement errors may translate to low risk, and vice versa. This is because the system model (2.1) can cause some measurements to be very loosely dependent on the parameters, for example, if a certain row of $\mathbf{H}$ contains small values. Minimizing the measurement error would then result in fitting the observations to the noise vector, rather than fitting them to the parameters. In some cases, the resulting estimate $\hat{\mathbf{x}}$ is completely unrelated to the parameter $\mathbf{x}$ [7]. In effect, this approach replaces the estimation goal with a simpler, but quite different, objective.

2.3 Stein's Phenomenon

We have already mentioned the fact that, for Gaussian noise, some estimators achieve lower MSE than the LS technique, for all values of $\mathbf{x}$. In this section, we adopt the assumption of Gaussian noise; we discuss these so-called LS-dominating estimators, and attempt to explain their surprising properties.

2.3.1 Dominance and Admissibility

Estimators achieving optimal MSE for a particular value of $\mathbf{x}$ may not perform well for other values of $\mathbf{x}$. For example, the trivial estimator $\hat{\mathbf{x}} = \mathbf{0}$ achieves zero MSE when $\mathbf{x} = \mathbf{0}$; lower risk is clearly not possible. However, $\hat{\mathbf{x}} = \mathbf{0}$ is obviously a very poor estimator for other values of $\mathbf{x}$. Since we have no prior information about the likelihood of parameter values, it is entirely possible that $\mathbf{0}$ is the true value of $\mathbf{x}$, in which case $\hat{\mathbf{x}} = \mathbf{0}$ is indeed the optimal estimator. Who is to say that other estimators should be preferred over it? Any such claim implicitly assumes

that $\mathbf{x} \neq \mathbf{0}$ is more likely, in some sense, than $\mathbf{x} = \mathbf{0}$, which is prior information we do not have the luxury of using.

Thus, not all estimators are comparable in terms of MSE performance. However, some estimators are uniformly better than others. For example, the estimator $\hat{\mathbf{x}} = -\hat{\mathbf{x}}_{\mathrm{LS}}$, which takes the LS estimate (2.4) and goes in the opposite direction, obtains higher MSE than the LS estimator, regardless of the value of $\mathbf{x}$. We say that $\hat{\mathbf{x}}_{\mathrm{LS}}$ *dominates* $\hat{\mathbf{x}}$. In general, we have the following definition.

Definition 2.1. An estimator $\hat{\mathbf{x}}_1$ is said to *dominate* an estimator $\hat{\mathbf{x}}_2$ if

$$\begin{aligned} &\mathrm{MSE}(\hat{\mathbf{x}}_1, \mathbf{x}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_2, \mathbf{x}) \quad \text{for all } \mathbf{x}, \\ &\mathrm{MSE}(\hat{\mathbf{x}}_1, \mathbf{x}) < \mathrm{MSE}(\hat{\mathbf{x}}_2, \mathbf{x}) \quad \text{for some } \mathbf{x}. \end{aligned} \tag{2.12}$$

If the stronger condition

$$\mathrm{MSE}(\hat{\mathbf{x}}_1, \mathbf{x}) < \mathrm{MSE}(\hat{\mathbf{x}}_2, \mathbf{x}) \quad \text{for all } \mathbf{x} \tag{2.13}$$

also holds, we say that $\hat{\mathbf{x}}_1$ *strictly dominates* $\hat{\mathbf{x}}_2$.

If $\hat{\mathbf{x}}_1$ dominates $\hat{\mathbf{x}}_2$, then one would prefer the use of $\hat{\mathbf{x}}_1$ over $\hat{\mathbf{x}}_2$. Hence, given a particular estimator, an important question is whether it can be dominated. We thus have the following definition.

Definition 2.2. An estimator $\hat{\mathbf{x}}_1$ is said to be *inadmissible* if there exists some estimator which dominates it. Otherwise, $\hat{\mathbf{x}}_1$ is said to be *admissible*.

The trivial estimator $\hat{\mathbf{x}} = \mathbf{0}$ is admissible, since no substantially different estimator[2] can achieve zero MSE at $\mathbf{x} = \mathbf{0}$. Surprisingly, however, the LS estimator turns out to be inadmissible [10]. Thus, it is of interest to characterize the class of admissible estimators, and to find estimators which dominate the LS estimator.

The study of admissibility is sometimes restricted to the set of linear estimators, i.e., estimators of the form $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$. A linear admissible estimator is one which is not dominated by any other linear estimator. A simple rule characterizes the class of such estimators [27], and, given any linear inadmissible estimator, it is possible to construct a linear admissible alternative which dominates it [28]. However, the problem of admissibility is considerably more intricate when the linearity restriction is removed; generally, admissible estimators are either

[2]There are other estimators which achieve zero MSE at $\mathbf{x} = \mathbf{0}$: any estimator which returns $\mathbf{0}$ with probability 1 satisfies this property. However, no such estimator improves upon the trivial estimator, neither at the point $\mathbf{x} = \mathbf{0}$ nor at any other point; thus, the estimator $\hat{\mathbf{x}} = \mathbf{0}$ is admissible.

trivial (e.g., $\hat{\mathbf{x}} = \mathbf{0}$) or exceedingly complex [16, 29, 30]. As a result, much research has focused on finding simple nonlinear techniques which dominate the LS estimator. A review of these results is presented in the following subsections.

Despite the difficulty in obtaining practical admissible estimators, there are several interesting theoretical results concerning admissibility. One important example is the following [2, Theorem 5.2.4].

Theorem 2.1. *Suppose $\hat{\mathbf{x}}$ is the unique Bayes estimator for some given prior distribution. Then, $\hat{\mathbf{x}}$ is an admissible estimator.*

Theorem 2.1 ties together two seemingly unrelated worlds: the Bayesian and deterministic estimation problems. The theorem can be intuitively explained as follows. In a deterministic setting, we have no knowledge of the prior distribution of $\mathbf{x}$. This setting can thus be viewed as one in which all priors are possible, and it is thus impossible to prefer one prior over another. In this view, dominance is obtained only when one estimator is better than another, for all possible priors. If a particular estimator is the unique optimum for a certain prior, then it cannot be dominated, and hence it is admissible.

An immediate consequence of Theorem 2.1 is the admissibility of $\hat{\mathbf{x}} = \mathbf{0}$, since it is the unique Bayes estimator for the prior given by $\Pr\{\mathbf{x} = \mathbf{0}\} = 1$. However, no prior yields the LS estimator as a Bayes estimator; indeed, we will now present an estimator which dominates $\hat{\mathbf{x}}_{\mathrm{LS}}$.

2.3.2 The James-Stein Estimator

Consider a simple deterministic estimation problem, in which

$$\mathbf{y} = \mathbf{x} + \mathbf{w}, \tag{2.14}$$

where $\mathbf{y}$ is an observation vector, $\mathbf{x}$ is unknown, and $\mathbf{w}$ is i.i.d. Gaussian noise with zero mean and known variance σ^2. We refer to this problem as the i.i.d. case of our system model (2.1) (note, however, that in addition to i.i.d. noise, this version assumes $\mathbf{H} = \mathbf{I}$).

The system (2.14) describes an everyday situation in which a set of parameters is measured, and the measurements obtained have independent noise. Since the noise has zero mean, it is very reasonable to use the measurements themselves as an estimate of the parameters. This is the approach of the LS estimator, which becomes simply $\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{y}$ in the i.i.d. case. As a result, there was considerable shock and disbelief when Stein demonstrated in 1956 that, in terms of

MSE, this approach is suboptimal [10]. The result became known as Stein's phenomenon.[3]

Stein's idea stemmed from an odd property of the expectation of $\|\mathbf{y}\|^2$. It is straightforward to show that

$$E\{\|\mathbf{y}\|^2\} = \|\mathbf{x}\|^2 + m\sigma^2, \tag{2.15}$$

where m is the length of the vectors $\mathbf{y}$ and $\mathbf{x}$. Thus, the average squared norm of the vector $\mathbf{y}$ is larger than the squared norm of the vector $\mathbf{x}$; the LS estimate is consistently an overestimate of the true parameter values. Stein therefore proposed to decrease the LS estimator by a factor of $\frac{\|\mathbf{y}\|^2 - m\sigma^2}{\|\mathbf{y}\|^2}$. The resulting *Stein estimator* is given by

$$\hat{\mathbf{x}}_{\mathrm{S}} = \left(1 - \frac{m\sigma^2}{\|\mathbf{y}\|^2}\right)\mathbf{y}. \tag{2.16}$$

Estimators such as $\hat{\mathbf{x}}_{\mathrm{S}}$, which consist of a scalar multiplying $\hat{\mathbf{x}}_{\mathrm{LS}}$, are referred to as *shrinkage estimators*.

Although (2.16) turns out to be a rather good estimator, it should be noted that the original reasoning provided by Stein is somewhat vague. In particular, it is the *squared* norm of $\mathbf{x}$ which is overestimated by a factor of $1 - \frac{m\sigma^2}{\|\mathbf{y}\|^2}$. Thus, Stein's argument would seem to suggest that a shrinkage factor of $\sqrt{1 - \frac{m\sigma^2}{\|\mathbf{y}\|^2}}$ is more appropriate. We will present alternative justifications for Stein's estimator later in this section, and in Chapter 4.

James and Stein [11] later showed that (2.16) dominates the LS estimate when $m \geq 4$. They also presented an improved estimator,

$$\hat{\mathbf{x}}_{\mathrm{JS}} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right)\mathbf{y}, \tag{2.17}$$

and showed that this so-called James-Stein estimator dominates the Stein estimator, and dominates the LS estimator for all $m \geq 3$. Stein [10] had previously shown that the LS estimator is admissible when $m \leq 2$.

This leads to what is perhaps the most counterintuitive consequence of Stein's phenomenon: When three or more unrelated parameters are measured, their total MSE can be reduced by using a combined estimator such as the James-Stein estimator; whereas when each parameter is estimated separately, the LS estimator is admissible. This quirk has caused some to sarcastically ask whether, in order to estimate the speed of light, one should jointly estimate tea consumption in Taiwan and hog weight in Montana. The response is that the James-Stein estimator always improves upon the MSE as defined in (2.2), i.e., the sum of the expected errors

[3] In fact, Stein assumed the variance σ^2 is unknown and estimated from the measurements. We ignore this point to simplify the presentation.

of each component. Therefore, the total MSE in measuring light speed, tea consumption and hog weight would improve by using the James-Stein estimator. However, any particular component (such as the speed of light) would improve for some parameter values, and deteriorate for others. Thus, although the James-Stein estimator dominates the LS estimator when three or more parameters are estimated, any single component does not dominate the respective component of the LS estimator [19].

The conclusion from this hypothetical example is that measurements should be combined if one is interested in minimizing their total MSE. For example, it is reasonable to combine channel tap measurements in a channel estimation scenario, as the goal is to minimize the total channel estimation error. Conversely, it is probably not reasonable to combine channel estimates of different users (in different locations), since no user would want their channel estimate to deteriorate in order to improve the average network performance.

In light of these counterintuitive results, it is not surprising that the James-Stein estimator drew considerable criticism from the statistics community in the 1960s and 1970s. For instance, many claimed that Stein's phenomenon is a result of some oddity of the MSE objective, or of the particular problem setting used. However, Brown [31] has shown that the LS estimator is inadmissible for a variety of risk functions. Stein's result has also been generalized to many other estimation settings, some of which are discussed in Subsection 2.3.3.

Like all new ideas, it took some time for the validity of Stein's phenomenon to gain credibility. Although the results of James and Stein are accepted today, they are still rarely applied to practical problems. Perhaps one reason for this is the lack of an intuitive explanation of the phenomenon. Apart from Stein's vague $\|\mathbf{y}\|^2$ argument described above, the only intuitive argument directly supporting the James-Stein estimator was provided by Efron and Morris [32]. They used an empirical Bayes approach to derive the James-Stein estimator, as follows. Suppose we are to estimate $\mathbf{x}$ in a Bayesian framework (Subsection 2.1.1), and suppose that $\mathbf{x}$ is known to have an i.i.d. Gaussian prior distribution with mean 0 and variance τ^2. The Bayesian estimator minimizing the MSE is then given by the shrinkage estimator

$$\hat{\mathbf{x}}_{\mathrm{B}} = \frac{\tau^2}{\sigma^2 + \tau^2}\mathbf{y}. \tag{2.18}$$

If τ^2 is unknown, it can be estimated from measurements, in what is called an empirical Bayes approach. This is done by conditioning on τ^2 and then estimating its value. When $\mathbf{y}$ is conditioned on τ^2, it is distributed as $N(\mathbf{0}, (\sigma^2 + \tau^2)\mathbf{I})$, so that $\|\mathbf{y}\|^2$ conditioned on τ^2 is distributed as $(\sigma^2 + \tau^2)\chi^2_m$. Using the fact that, for $m \geq 3$, the inverse moment of a χ^2_m variate is given by

$E\{1/\chi^2_m\} = 1/(m-2)$ [17, Section 6.3], we have

$$E\left\{1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right\} = \frac{\tau^2}{\sigma^2+\tau^2}. \tag{2.19}$$

Thus, $1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}$ is an unbiased estimate for the shrinkage factor of (2.18). Substituting this estimated shrinkage factor yields the James-Stein estimator (2.17).

As we shall see in Section 2.4, a similar empirical Bayes argument can be applied to the non-i.i.d. case, resulting in a generalized Tikhonov estimator. However, the resulting estimator no longer dominates the LS approach. To our knowledge, no researcher other than Efron and Morris has used empirical Bayes reasoning for justifying non-i.i.d. LS-dominating estimators.

The result of Efron and Morris does provide an intuitive derivation of the James-Stein estimator. However, the choice of the unbiased estimator (2.19) is somewhat arbitrary; the empirical Bayes approach generally makes use of a maximum likelihood estimate, rather than an unbiased estimate [2]. By differentiating the conditional pdf of $\mathbf{y}$ given τ^2, it is straightforward to show that the maximum likelihood estimate of τ^2 is given by $\frac{1}{p}\|\mathbf{y}\|^2 - \sigma^2$. Substituting this estimate into (2.18) yields the Stein estimator (2.16), rather than the James-Stein estimator.

One may also question the use of the normal prior of $\mathbf{x}$ as a basis for this derivation. Following the discussion of Subsection 2.1.1, we seek an estimator designed to work for deterministic parameters about which nothing is known a priori. The normal prior assumption, even when its variance is unknown, introduces a significant amount of additional information about the prior: not only the fact that it is likely close to zero, but also a description of the probabilities of obtaining larger values. In Chapter 4, we present an alternative approach for deriving Stein-type estimators, using the blind minimax concept.

2.3.3 Other LS-Dominating Estimators

Since the introduction of the James-Stein estimator, considerable research has gone into improving it and extending its use. The following is a review of the major contributions to this field.

The Positive-Part Estimator. The first improvement of the James-Stein estimator was provided by Baranchik [33], and became known as the positive-part estimator. Baranchik was bothered by the fact that the shrinkage factor of $\hat{\mathbf{x}}_{\text{JS}}$ (2.17) is sometimes negative, i.e., in some cases the James-Stein estimator inverts the sign of the LS estimate. However, even noisy measurements are still more likely to be correct than their inverse. Baranchik formalized this notion

in the following theorem.

Theorem 2.2 (Positive-Part Estimator). *Consider the i.i.d. estimation problem* (2.14)*, and let* $\hat{\mathbf{x}} = f(\mathbf{y})\,\mathbf{y}$ *be any estimator such that* $\Pr\{f(\mathbf{y}) < 0\} > 0$*. Then, the positive part estimator* $\hat{\mathbf{x}}_+ = f_+(\mathbf{y})\,\mathbf{y}$ *dominates* $\hat{\mathbf{x}}$*, where* $f_+(\mathbf{y}) = \max(f(\mathbf{y}), 0)$*.*

Thus, whenever an estimator inverts the sign of the LS estimator, its performance can be improved by replacing the negative shrinkage with zero. The positive-part James-Stein estimator then becomes

$$\hat{\mathbf{x}}_{\text{PJS}} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right)_+ \mathbf{y}, \tag{2.20}$$

where $(\cdot)_+ = \max(\cdot, 0)$. By Theorem 2.2, $\hat{\mathbf{x}}_{\text{PJS}}$ dominates $\hat{\mathbf{x}}_{\text{JS}}$.

The positive-part estimator demonstrates a weakness of the James-Stein estimator, but its solution is somewhat ad-hoc. One would expect a well-behaved estimator to inherently avoid negative shrinkage, rather than require an external mechanism for maintaining nonnegative shrinkage. Furthermore, the positive-part estimator creates an unwanted non-differentiability in the estimator. In Chapter 4, we present several estimators derived using the blind minimax technique, which inherently guarantee positive shrinkage.

Bock's Estimator. Thus far, the Stein phenomenon has been discussed under the assumption of i.i.d. measurements (2.14). Stein's results have since been extended to a variety of other scenarios. Among these, we are particularly interested in the non-i.i.d. linear estimation model (2.1). The most common estimator for the non-i.i.d. case was developed by Bock [15], who proposed the use of

$$\hat{\mathbf{x}}_{\text{Bock}} = \left(1 - \frac{d-2}{\hat{\mathbf{x}}_{\text{LS}}^* \mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H} \hat{\mathbf{x}}_{\text{LS}}}\right) \hat{\mathbf{x}}_{\text{LS}}. \tag{2.21}$$

Here d (referred to as the *effective dimension*) is defined as

$$d = \frac{\epsilon_0}{\epsilon_{\max}}, \tag{2.22}$$

where ϵ_0 is given by (2.5) and $\epsilon_{\max}$ is the largest eigenvalue of $(\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H})^{-1}$. The effective dimension d is a measure of the number of independent measurements in the system; for example, in the i.i.d. case d simply reduces to the number of parameters, m. Furthermore, $\hat{\mathbf{x}}_{\text{Bock}}$ reduces to $\hat{\mathbf{x}}_{\text{JS}}$ in the i.i.d. case; techniques exhibiting this property are referred to as extended James-Stein estimators. However, apart from the similarity to the James-Stein technique, no justification is provided for the particular choice of the shrinkage factor in (2.21).

Non-Shrinkage Estimators. Bock's estimator is a shrinkage estimator, i.e., it consists of a scalar multiplying the LS estimate. However, in the non-i.i.d. case, $\hat{\mathbf{x}}_{\mathrm{LS}}$ is distributed with mean $\mathbf{x}$ and variance $(\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}$, i.e., some elements of $\hat{\mathbf{x}}_{\mathrm{LS}}$ have higher variance than others. Thus, several researchers [14, 16] have proposed extended James-Stein estimators which are non-shrinkage in the non-i.i.d. case. The question is then whether high-variance eigenvalues of $\hat{\mathbf{x}}_{\mathrm{LS}}$ should be shrunk more than low-variance components, or vice versa.

Efron and Morris [14] consider the somewhat simpler case in which $\mathbf{H} = \mathbf{I}$ and $\mathbf{C}_{\mathbf{w}} = \mathrm{diag}(\sigma_1^2, \cdots \sigma_m^2)$. Thus, the noise elements are independent but not identically distributed. The proposed estimator shrinks high-variance measurements closer to zero. This is justified using empirical Bayes reasoning similar to the one presented in Subsection 2.3.2. However, the estimator can only be calculated by iteratively solving a set of nonlinear equations; furthermore, dominance of the LS estimator is not guaranteed.

By contrast, Berger [16,34] considers a class of estimators, the simplest of which is given by

$$\hat{\mathbf{x}}_{\mathrm{Berger}} = \hat{\mathbf{x}}_{\mathrm{LS}} - \frac{d-2}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\hat{\mathbf{x}}_{\mathrm{LS}}}\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}\hat{\mathbf{x}}_{\mathrm{LS}}, \tag{2.23}$$

where the effective dimension d is given by (2.22). This estimator is shown to dominate the LS estimator. However, $\hat{\mathbf{x}}_{\mathrm{Berger}}$ has the odd property that low-variance eigenvalues of $\hat{\mathbf{x}}_{\mathrm{LS}}$ are shrunk more than high-variance eigenvalues. This is counterintuitive since the LS estimator performs well precisely in those cases for which the variance is low; Berger [34, p. 367] remarks on this point, but provides no explanation for the choice of the estimator (2.23).

We are thus left with the ironic situation in which there is disagreement as to whether high-variance components should be shrunk more [14] or less [16]. This controversy demonstrates the fact that some estimators are justified solely on the fact that they dominate the LS estimator, without giving ample consideration to the selection of a shrinkage factor. We will return to this issue when discussing non-shrinkage blind minimax estimators in Section 4.2.

The Thompson Estimator. As we have seen, a variety of estimators based on the James-Stein approach have been constructed and shown to dominate the LS estimator. This certainly indicates a deficiency of the LS technique. However, it does not necessarily indicate that the James-Stein approach is paramount [35]. The James-Stein estimator is simply the first LS-dominating estimator discovered; it is possible that some totally different estimator performs even better.

Although there exist several improvements of the James-Stein estimator [16,29,30], they are often too complex for practical use. On the other hand, there is a wide and largely unexplored

class of LS-dominating estimators unrelated to the James-Stein approach. While these do not generally dominate the James-Stein estimator, it is quite possible that a simple LS-dominating estimator exists which performs better than the James-Stein estimator under most conditions.

One class of LS-dominating estimators for the i.i.d. case was discovered by Baranchik [13]. The Thompson estimator, given by

$$\hat{\mathbf{x}}_{\mathrm{Th}} = \left(\frac{\|\mathbf{y}\|^2}{\|\mathbf{y}\|^2 + m\sigma^2} \right) \mathbf{y}, \tag{2.24}$$

belongs to this class, and is of particular interest to us. Curiously, when Thompson proposed this estimator [12], he did not think that it might dominate the LS estimator; indeed, he wrote, "We do not hope [...] to best uniformly the [LS estimator], but to obtain an estimator which is better near the natural origin though possibly worse farther away." Thompson was apparently unaware of the work of James and Stein. Nevertheless, $\hat{\mathbf{x}}_{\mathrm{Th}}$ dominates the LS estimator when $m \geq 4$, and turns out to be a serious competitor of the James-Stein estimator. As we shall see in Chapter 4, Thompson's estimator can be derived within the blind minimax framework, in which it can also be extended to the non-i.i.d. case.

2.4 Tikhonov Regularization

Independently of the development of Stein-type estimators, many researchers in applied fields became aware of deficiencies of the LS estimator. A variety of more or less ad-hoc alternatives were proposed as a result of these experiences. Generally, these alternatives were not shown to dominate the LS estimator; rather, they were intended to improve estimation quality in specific scenarios, and were empirically observed to outperform the LS technique. Of these approaches, the most common is Tikhonov regularization [8], also referred to as ridge regression [9].

Tikhonov regularization is intended for ill-posed problems, i.e., problems in which $\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H}$ is nearly singular. The matrix $\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H}$ is guaranteed to be positive-definite (and hence invertible), since $\mathbf{H}$ is full-rank and $\mathbf{C}_{\mathbf{w}}$ is positive-definite (see Subsection 2.1.2). However, $\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H}$ may contain eigenvalues which are very close to zero. In such cases, the LS estimator (which depends on the term $(\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H})^{-1}$) causes severe amplification of measurement noise. In effect, an ill-posed setting is one in which the SNR of at least one parameter is extremely low; as we have seen in Subsection 2.3.2, the LS approach results in overestimation in such conditions.

Tikhonov viewed the LS estimator from a measurement error point of view (see Subsection 2.2.3). He proposed to combine the squared measurement error criterion with a restraint

on the norm of the estimate. The problem is then to obtain an estimator $\hat{\mathbf{x}}$ which minimizes[4]

$$\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|^2 + \alpha^2\|\hat{\mathbf{x}}\|^2, \tag{2.25}$$

where $\alpha > 0$ specifies the weight of the norm objective relative to the measurement error objective. This regularization parameter must be chosen empirically, and a variety of rules of thumb are used to select its value.

Differentiating (2.25) with respect to $\hat{\mathbf{x}}$ results in the Tikhonov regularization estimator

$$\hat{\mathbf{x}}_{\mathrm{T}} = \left(\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \alpha^2\mathbf{I}\right)^{-1}\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{y}. \tag{2.26}$$

The term $\alpha^2\mathbf{I}$ increases each of the eigenvalues of $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$ by α^2. This improves the conditioning of $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \alpha^2\mathbf{I}$, and its inverse can be accurately calculated.

Rather than increase each eigenvalue of $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$ equally, regularization can also be achieved by adding any positive-definite matrix. This results in the generalized Tikhonov regularization, given by

$$\hat{\mathbf{x}}_{\mathrm{T}} = \left(\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \alpha\mathbf{T}\right)^{-1}\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{y}, \tag{2.27}$$

where $\mathbf{T}$ is positive-definite. This approach can be derived by replacing the criterion (2.25) with

$$\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|^2 + \alpha^2\hat{\mathbf{x}}^*\mathbf{T}\hat{\mathbf{x}}, \tag{2.28}$$

where $\hat{\mathbf{x}}^*\mathbf{T}\hat{\mathbf{x}}$ can be viewed as a weighted norm of $\hat{\mathbf{x}}$.

The generalized Tikhonov regularization is often justified in a Bayesian framework. Suppose that the parameter vector $\mathbf{x}$ is known to be distributed normally with zero mean and covariance $\mathbf{T}^{-1}$. In this setting, the estimator minimizing the MSE (in the Bayesian sense) is known as the Wiener filter, and is given by (2.27), with $\alpha = 1$.

This Bayesian justification is reminiscent of the empirical Bayes approach used to derive the James-Stein estimator (Subsection 2.3.2). The similarity immediately suggests an empirical Bayes extension of Tikhonov regularization: One could estimate the value of α and substitute this result into (2.26). In Subsection 2.3.2, we demonstrated that a certain technique for estimating α results in the James-Stein estimator in the i.i.d. case. Applying this approach in the non-i.i.d. case would seem to result in a generalization of the James-Stein result, which will be referred to as "blind" Tikhonov regularization. Indeed, this approach is often used in practice for solving ill-posed problems, although it appears that the relation to the James-Stein approach has not previously been noted.

[4] As with the derivation of Subsection 2.2.2, we first whiten the noise using (2.11).

There are several methods for empirically estimating the parameters $\mathbf{T}$ and α from measurements. If nothing is known about the parameter $\mathbf{x}$, one possibility is to assume that the elements of $\mathbf{x}$ are i.i.d., and to estimate their variance σ_x^2. In (2.27), this implies $\mathbf{T} = \mathbf{I}$ and $\alpha = 1/\sigma_x^2$. Optimally, one would like to use the mean square value of the parameters x_i as an approximation of the variance σ_x^2. However, since $\mathbf{x}$ is unknown, σ_x^2 can be estimated as the mean square value of the elements of $\hat{\mathbf{x}}_{\mathrm{LS}}$; in other words, $\sigma_x^2 = \|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2/m$, where m is the length of the vector $\mathbf{x}$. This results in the estimator

$$\hat{\mathbf{x}}_{\mathrm{T}}^{(1)} = \left(\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \frac{m}{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2}\mathbf{I}\right)^{-1}\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{y}. \tag{2.29}$$

The derivation of $\hat{\mathbf{x}}_{\mathrm{T}}^{(1)}$ assumed that the parameters $\mathbf{x}$ are i.i.d. An alternative is to assume instead that the variance of $\mathbf{x}$ is proportional to the variance of the noise $\mathbf{w}$, which implies $\mathbf{T} = \mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$. In an analogy to the previous derivation, one may then estimate α as $m/\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}\hat{\mathbf{x}}_{\mathrm{LS}}$. Substituting these values into (2.27) results in the shrinkage estimator

$$\hat{\mathbf{x}}_{\mathrm{T}}^{(2)} = \frac{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}\hat{\mathbf{x}}_{\mathrm{LS}}}{m + \hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}\hat{\mathbf{x}}_{\mathrm{LS}}}\hat{\mathbf{x}}_{\mathrm{LS}}. \tag{2.30}$$

Unfortunately, the resulting estimators do not dominate the LS technique, i.e., in some cases lower MSE is achieved with the LS estimator than with either of the proposed estimators $\hat{\mathbf{x}}_{\mathrm{T}}^{(1)}$ and $\hat{\mathbf{x}}_{\mathrm{T}}^{(2)}$. This is demonstrated in Subsection 4.5.3, where the Tikhonov regularization is compared to other extensions of the James-Stein approach.

Chapter 3

Minimax Estimation

The results of this work are motivated by minimax estimators designed for a bounded parameter set. In the following, we define and derive these estimators, and present several well-known properties of minimax estimators, as well as some new results. These provide the basis for the approaches of Chapters 4 and 5.

3.1 Problem Statement

The deterministic estimation model presented in Subsection 2.1.2 does not assume any prior knowledge of the values of the parameter vector $\mathbf{x}$. In the minimax model, we still do not assume any statistical model for $\mathbf{x}$; however, we do assume that $\mathbf{x}$ lies within some known, bounded parameter set $\mathcal{S}$. For example, it may be known that the total power of all parameters is bounded, or that no parameter can exceed some universal maximum.

When $\mathbf{x}$ is known to lie in a given parameter set, one may seek an estimator to optimize performance for the worst possible value within the parameter set. Performance may be measured according to any given risk function $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$ (Subsection 2.1.2). An estimator achieving this requirement is called minimax. The following definition formalizes this concept.

Definition 3.1. Let $\mathcal{E}$ be a class of estimators, let $\mathcal{S}$ be a compact subset of the parameter space, and let $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$ be a risk function. An estimator $\hat{\mathbf{x}}_{\mathrm{M}} \in \mathcal{E}$ is said to be $\mathcal{E}$-minimax (over the set $\mathcal{S}$) if, for any other estimator $\hat{\mathbf{x}} \in \mathcal{E}$,

$$\sup_{\mathbf{x} \in \mathcal{S}} \epsilon(\hat{\mathbf{x}}_{\mathrm{M}}, \mathbf{x}) \leq \sup_{\mathbf{x} \in \mathcal{S}} \epsilon(\hat{\mathbf{x}}, \mathbf{x}). \tag{3.1}$$

Ideally, we would like to characterize a minimax estimator within the class of *all* estimators. However, in general, it is very difficult to find such estimators. Instead, in many cases one seeks

to minimize the worst-case risk among all *affine* estimators

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{G}\mathbf{y} + \mathbf{x}_0, \tag{3.2}$$

where $\mathbf{G} \in \mathbb{C}^{m\times n}$ and $\mathbf{x}_0 \in \mathbb{C}^m$ are constant. Alternatively, one may minimize the worst-case risk among all *linear* estimators

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{G}\mathbf{y}. \tag{3.3}$$

In many cases, we will be interested in a parameter set $\mathcal{S}$ which is symmetric about the origin, i.e.,

$$\mathbf{x} \in \mathcal{S} \quad \Rightarrow \quad -\mathbf{x} \in \mathcal{S}. \tag{3.4}$$

In such cases, one would expect a reasonable estimator to be symmetric about the origin as well. In particular, one would expect a linear minimax estimator to perform as well as any affine minimax estimator. This is, indeed, the case when the risk function is the MSE, as shown by the following proposition.

Proposition 3.1. *Let* $\hat{\mathbf{x}}_\mathrm{a} = \mathbf{G}\mathbf{y} + \mathbf{x}_0$ *be an affine minimax MSE estimator over the parameter set* $\mathcal{S}$*, and assume* $\mathcal{S}$ *satisfies* (3.4)*. Then,* $\hat{\mathbf{x}}_\mathrm{b} = \mathbf{G}\mathbf{y}$ *is also an affine minimax estimator.*

Proof. The MSE of an affine estimator is given by

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{x}} = \mathbf{G}\mathbf{y} + \mathbf{x}_0, \mathbf{x}) &= E\{\|\mathbf{G}\mathbf{H}\mathbf{x} + \mathbf{G}\mathbf{w} + \mathbf{x}_0 - \mathbf{x}\|^2\} \\ &= E\{\mathbf{w}^*\mathbf{G}^*\mathbf{G}\mathbf{w}\} + \|\mathbf{G}\mathbf{H}\mathbf{x} - \mathbf{x} + \mathbf{x}_0\|^2 \\ &= \mathrm{Tr}(\mathbf{G}\mathbf{C}_\mathbf{w}\mathbf{G}^*) + \|(\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{x} - \mathbf{x}_0\|^2.\end{aligned} \tag{3.5}$$

Assume by contradiction that $\hat{\mathbf{x}}_\mathrm{b} = \mathbf{G}\mathbf{y}$ is not minimax. We then have

$$\max_{\mathbf{x}\in\mathcal{S}} \mathrm{MSE}(\mathbf{G}\mathbf{y}, \mathbf{x}) > \max_{\mathbf{x}\in\mathcal{S}} \mathrm{MSE}(\mathbf{G}\mathbf{y} + \mathbf{x}_0, \mathbf{x}). \tag{3.6}$$

Using (3.5) and denoting $\mathbf{A} = \mathbf{I} - \mathbf{G}\mathbf{H}$, we obtain

$$\max_{\mathbf{x}\in\mathcal{S}} \|\mathbf{A}\mathbf{x}\|^2 > \max_{\mathbf{x}\in\mathcal{S}} \|\mathbf{A}\mathbf{x} - \mathbf{x}_0\|^2. \tag{3.7}$$

Thus, there exists $\mathbf{x}_1 \in \mathcal{S}$ such that

$$\forall \mathbf{x} \in \mathcal{S}, \quad \|\mathbf{A}\mathbf{x}_1\|^2 > \|\mathbf{A}\mathbf{x} - \mathbf{x}_0\|^2. \tag{3.8}$$

Since $\mathcal{S}$ is symmetric about the origin, (3.8) holds for both $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{x} = -\mathbf{x}_1$, yielding

$$\begin{aligned}\|\mathbf{A}\mathbf{x}_1\|^2 &> \|\mathbf{A}\mathbf{x}_1\|^2 + \|\mathbf{x}_0\|^2 - 2\mathbf{x}_0^*\mathbf{A}\mathbf{x}_1, \\ \|\mathbf{A}\mathbf{x}_1\|^2 &> \|\mathbf{A}\mathbf{x}_1\|^2 + \|\mathbf{x}_0\|^2 + 2\mathbf{x}_0^*\mathbf{A}\mathbf{x}_1.\end{aligned} \tag{3.9}$$

Combining these two equations, we obtain $\|\mathbf{x}_0\|^2 < 0$, which is a contradiction; therefore, $\hat{\mathbf{x}}_\mathrm{b}$ is minimax. □

Thus, when $\mathcal{S}$ is symmetric about the origin, an affine minimax MSE estimator may be found by optimizing over all linear estimators.

The restriction to affine estimators reduces the dimensionality of the problem by requiring the form (3.2). As a result, identification of an affine minimax estimator can often be stated as a convex optimization problem, and in many cases closed-form estimators can be derived. Some known minimax estimators are given in the next section. We will be particularly interested in minimax MSE estimators (i.e., estimators minimizing the worst-case MSE (2.2) within the set $\mathcal{S}$) and in minimax regret estimators (which minimize the worst-case regret (2.3)).

3.2 Examples of Minimax Estimators

In this section, we present several known results relating to minimax estimation [5–7]. These provide convenient closed forms for affine minimax estimators, for spherical and ellipsoidal parameter sets. The minimax estimators described in this section will be used in the derivations of Chapters 4 and 5.

3.2.1 Minimax MSE Estimators

We begin by studying the affine minimax MSE estimator for a spherical parameter set centered on the origin. A simple closed form for this estimator is given by the following theorem.

Theorem 3.2. *Consider the linear estimation model* (2.1)*, and let* $\mathcal{S} = \{\mathbf{x} : \|\mathbf{x}\|^2 \leq L^2\}$ *be a spherical parameter set centered on the origin. Then, the unique affine minimax MSE estimator for the set* $\mathcal{S}$ *is given by*

$$\hat{\mathbf{x}} = \frac{L^2}{L^2 + \epsilon_0}\hat{\mathbf{x}}_{\mathrm{LS}}, \tag{3.10}$$

where $\hat{\mathbf{x}}_{\mathrm{LS}}$ *is the LS estimator* (2.4) *and* ϵ_0 *is given by* (2.5)*. The maximum MSE obtained by this estimator over the set* $\mathcal{S}$ *is given by*

$$\left(\frac{L^2}{L^2 + \epsilon_0}\right)\epsilon_0. \tag{3.11}$$

Proof. The minimax estimator for the set $\mathcal{S}$ was shown in [6, Th. 1] to equal (3.10). We now show that the worst-case error of this estimator is given by (3.11). The bias (2.7) of $\hat{\mathbf{x}}$ is given by

$$\mathbf{b}(\hat{\mathbf{x}}, \mathbf{x}) = E\{\hat{\mathbf{x}} - \mathbf{x}\} = (\beta - 1)\mathbf{x}, \tag{3.12}$$

where $\beta = \frac{L^2}{L^2+\epsilon_0}$. Furthermore, the variance (2.8) of $\hat{\mathbf{x}}$ is given by

$$\mathrm{Var}(\hat{\mathbf{x}}) = \mathrm{Tr}(\mathbf{G}\mathbf{C}_{\mathbf{w}}\mathbf{G}^*) = \beta^2\epsilon_0. \tag{3.13}$$

Since the MSE is the sum of the variance and the squared norm of the bias (2.6), we have that for any $\mathbf{x} \in \mathcal{S}$,

$$\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = \beta^2 \epsilon_0 + (1-\beta)^2 \|\mathbf{x}\|^2. \tag{3.14}$$

Clearly, the maximum of (3.14) within $\mathcal{S}$ is obtained for $\|\mathbf{x}\|^2 = L^2$. The maximum MSE is therefore given by $\beta^2\epsilon_0 + (1-\beta)^2 L^2$, which is equivalent to (3.11). □

As shown by Theorem 3.2, the minimax estimator for a spherical parameter set is a shrinkage estimator, i.e., it consists of a scalar shrinkage factor multiplying the LS estimate. This is no longer the case when the parameter set is ellipsoidal, as demonstrated by the following theorem, a proof of which may be found in [6, Th. 1].

Theorem 3.3. *Consider the linear estimation model* (2.1)*, and let* $\mathcal{S} = \{\mathbf{x} : \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$ *be an ellipsoidal parameter set centered on the origin, for some positive-definite matrix* $\mathbf{T}$*. Suppose that* $\mathbf{T}$ *and* $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$ *have the same eigenvector matrix, so that* $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^*$ *and* $\mathbf{T} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^*$*, where* $\mathbf{V}$ *is unitary,* $\boldsymbol{\Sigma}$ *is diagonal with diagonal elements* $\sigma_1, \sigma_2, \ldots \sigma_m$*, and* $\boldsymbol{\Lambda}$ *is diagonal with diagonal elements* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$*. Then, the unique affine minimax MSE estimator for the set* $\mathcal{S}$ *is given by*

$$\hat{\mathbf{x}} = \mathbf{V}^* \operatorname{diag}(\mathbf{0}_k, \mathbf{1}_{m-k})\mathbf{V}(\mathbf{I} - \alpha\mathbf{T}^{1/2})\hat{\mathbf{x}}_{\mathrm{LS}}, \tag{3.15}$$

where

$$\alpha = \frac{\sum_{i=k+1}^{m} \frac{\lambda_i^{1/2}}{\sigma_i}}{L^2 + \sum_{i=k+1}^{m} \frac{\lambda_i}{\sigma_i}} \tag{3.16}$$

and k *is the smallest integer such that* $0 \leq k \leq m-1$ *and* $\alpha\lambda_{k+1}^{1/2} < 1$*. The maximum MSE obtained by this estimator over the set* $\mathcal{S}$ *is given by*

$$\sum_{i=k+1}^{m} \frac{1 - \alpha\lambda_i^{1/2}}{\sigma_i}. \tag{3.17}$$

A simple closed form of the minimax MSE estimator can be obtained for any ellipsoidal parameter set, if the parameter set is small enough to guarantee a certain SNR threshold. This is stated formally in the following theorem, the proof of which may be found in [28, Section IV-B.3].

Theorem 3.4. *Consider the linear estimation model* (2.1)*, and let* $\mathcal{S} = \{\mathbf{x} : \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$ *be an ellipsoidal parameter set centered on the origin, for some positive-definite matrix* $\mathbf{T}$*. Let* $\mathbf{Q} = \mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$*, and suppose*

$$\lambda_{\min}\left(\mathbf{Q}^{-1}(\mathbf{Q}\mathbf{T}^{-1}\mathbf{Q})^{1/2}\right) \geq \alpha \tag{3.18}$$

where

$$\alpha = \frac{\mathrm{Tr}\left((\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}^{-1})^{1/2}\right)}{L^2 + \mathrm{Tr}(\mathbf{Q}^{-1}\mathbf{T})}. \tag{3.19}$$

Then the unique affine minimax MSE estimator for the set $\mathcal{S}$ is given by

$$\hat{\mathbf{x}} = (\mathbf{I} - \alpha(\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}^{-1})^{1/2}\mathbf{Q})\mathbf{Q}^{-1}\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{y}. \tag{3.20}$$

3.2.2 Minimax Regret Estimators

As discussed in Subsection 2.1.2, the regret is useful when dealing with linear estimators, since it attempts to disregard errors resulting from the limitations of linear estimation. We will use this risk function in Chapter 5, which deals with the linear estimation setting. The following theorem asserts that minimax regret estimators can be calculated efficiently, for a large class of ellipsoidal parameter sets. The proof of this theorem may be found in [7, Th. 1]. Note that our notation differs somewhat from that of [7]; in particular, we define $\mathbf{r} = \mathbf{s}/L^2$.

Theorem 3.5. *Consider the linear estimation model* (2.1)*, and let $\mathcal{S} = \{\mathbf{x} : \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$ be an ellipsoidal parameter set centered on the origin, for some positive-definite matrix $\mathbf{T}$. Suppose that $\mathbf{T}$ and $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$ have the same eigenvector matrix, so that $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^*$ and $\mathbf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$. Here $\mathbf{V}$ is unitary, $\mathbf{\Sigma}$ is diagonal with diagonal elements σ_i, and $\mathbf{\Lambda}$ is diagonal with diagonal elements λ_i. Then, the linear minimax regret estimator is given by*

$$\hat{\mathbf{x}} = \mathbf{V}\mathbf{D}\mathbf{V}^*\hat{\mathbf{x}}_{\mathrm{LS}}, \tag{3.21}$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal elements $\mathbf{d} = (d_1, \ldots d_m)^T$ are the solution to the convex optimization problem

$$\min_{\tau,\mathbf{d}} \tau \tag{3.22}$$

$$s.t. \begin{cases} F_1(\mathbf{d}) \leq \tau & (a) \\ F_2(\mathbf{d}, \mathbf{r}) \leq \tau \quad \forall\, \mathbf{r} \in \mathcal{R}. & (b) \end{cases}$$

Here,

$$F_1(\mathbf{d}) = \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i}, \tag{3.23}$$

$$F_2(\mathbf{d}, \mathbf{r}) = \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} + L^2 \sum_{i=1}^{m} (1 - d_i)^2 r_i - \frac{L^2 \sum_{i=1}^{m} r_i}{1 + L^2 \sum_{i=1}^{m} \sigma_i r_i}, \tag{3.24}$$

and

$$\mathcal{R} = \left\{ \mathbf{r} :\ r_i \geq 0,\ \sum_{i=1}^{m} \lambda_i r_i = 1 \right\}. \tag{3.25}$$

In (3.22)*, the optimal value of τ is the worst-case regret.*

It is worth noting that, in many special cases, simpler forms have been derived for the minimax regret estimator. In particular, closed forms are available for the case $\mathbf{T} = \mathbf{I}$ [7, Th. 3] and for the case $\mathbf{T} = \mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H}$ [7, Th. 2].

3.2.3 Noncentral Estimation

In the previous subsections, several examples of affine minimax estimators were provided. All of these examples were based on parameter sets which are symmetrical around the origin, and hence the resulting minimax estimators were linear. These results can be extended to parameter sets centered on any constant point $\mathbf{x}_0$, as demonstrated by the following proposition.

Proposition 3.6. *Consider the linear estimation model* (2.1) *and any risk function* $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$*. Let* $\mathcal{S}$ *be a bounded parameter set, and let* $\mathcal{S} + \mathbf{x}_0$ *be a shifted parameter set given by*

$$\mathcal{S} + \mathbf{x}_0 = \{\mathbf{x} + \mathbf{x}_0 : \mathbf{x} \in \mathcal{S}\}, \tag{3.26}$$

for some constant $\mathbf{x}_0$*. Suppose* $\hat{\mathbf{x}}_{\mathrm{M}}(\mathbf{y})$ *is an affine minimax estimator over the set* $\mathcal{S}$ *and for the risk function* $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$*. Then, an affine minimax estimator for the set* $\mathcal{S} + \mathbf{x}_0$ *is given by*

$$\hat{\mathbf{x}}'_{\mathrm{M}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{M}}(\mathbf{y} - \mathbf{H}\mathbf{x}_0) + \mathbf{x}_0. \tag{3.27}$$

Proof. For any estimator $\hat{\mathbf{x}}$,

$$\sup_{\mathbf{x} \in \mathcal{S} + \mathbf{x}_0} \|\hat{\mathbf{x}} - \mathbf{x}\|^2 = \sup_{\mathbf{x} \in \mathcal{S} + \mathbf{x}_0} \|\hat{\mathbf{x}} - \mathbf{x}_0 - (\mathbf{x} - \mathbf{x}_0)\|^2 = \sup_{\boldsymbol{\xi} \in \mathcal{S}} \|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|^2, \tag{3.28}$$

where we defined $\boldsymbol{\xi} = \mathbf{x} - \mathbf{x}_0$ and $\hat{\boldsymbol{\xi}} = \hat{\mathbf{x}} - \mathbf{x}_0$. Hence, finding an estimator $\hat{\mathbf{x}}'_{\mathrm{M}}$ which minimizes the worst-case risk within $\mathcal{S} + \mathbf{x}_0$ is equivalent to finding an estimator $\hat{\boldsymbol{\xi}}$ which minimizes the worst-case risk within $\mathcal{S}$. However, since $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, we have

$$\mathbf{y} - \mathbf{H}\mathbf{x}_0 = \mathbf{H}\boldsymbol{\xi} + \mathbf{w}. \tag{3.29}$$

Thus, $\mathbf{y} - \mathbf{H}\mathbf{x}_0$ may be viewed as observations of the parameter vector $\boldsymbol{\xi}$ under the standard linear regression model (2.1); it follows that the estimator $\hat{\boldsymbol{\xi}}$ which minimizes the worst-case risk within $\mathcal{S}$ is given by

$$\hat{\boldsymbol{\xi}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{M}}(\mathbf{y} - \mathbf{H}\mathbf{x}_0), \tag{3.30}$$

and substituting $\hat{\mathbf{x}} = \hat{\boldsymbol{\xi}} + \mathbf{x}_0$ proves the proposition. □

As an example, consider the sphere $\mathcal{S} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\|^2 \leq L^2\}$. Use of Theorem 3.2 and Proposition 3.6 immediately shows that the minimax estimator in this case is given by

$$\hat{\mathbf{x}} = \frac{L^2}{L^2 + \epsilon_0}\hat{\mathbf{x}}_{\mathrm{LS}} + \frac{\epsilon_0}{L^2 + \epsilon_0}\mathbf{x}_0. \tag{3.31}$$

This result is a weighted average between the sphere center $\mathbf{x}_0$ and the LS estimate $\hat{\mathbf{x}}_{\mathrm{LS}}$. When the radius L tends to zero, the "prior knowledge" $\mathbf{x}_0$ has more weight. Contrariwise, when $L \to \infty$, the LS estimate takes precedence. A generalization of this averaging effect occurs for all ellipsoidal minimax MSE and minimax regret estimators, as shown by the following proposition.

Proposition 3.7. *Let $\mathcal{S} = \{\mathbf{x} : (\mathbf{x} - \mathbf{x}_0)^*\mathbf{T}(\mathbf{x} - \mathbf{x}_0) \leq L^2\}$ be an ellipsoidal parameter set, where $\mathbf{T}$ is any positive-definite matrix and $\mathbf{x}_0$ is any constant. Consider as a risk function either the MSE or the regret. Then, the unique affine minimax estimator is given by*

$$\hat{\mathbf{x}}_{\mathrm{M}} = \mathbf{B}\hat{\mathbf{x}}_{\mathrm{LS}} + (\mathbf{I} - \mathbf{B})\mathbf{x}_0, \tag{3.32}$$

where $\mathbf{B}$ is some $m \times m$ matrix.

Proof. Consider first the case $\mathbf{x}_0 = \mathbf{0}$. In this case, the set $\mathcal{S}$ is symmetrical about the origin, and hence $\hat{\mathbf{x}}^0_{\mathrm{M}}$ is linear. Furthermore, the risk depends on $\hat{\mathbf{x}}_{\mathrm{M}} = \mathbf{G}\mathbf{y}$ only through the forms $\mathbf{GH}$ and $\mathrm{Tr}(\mathbf{G}\mathbf{C}_{\mathbf{w}}\mathbf{G}^*)$. It has been shown [6, Prop. 1] that for such risk functions, the linear minimax estimator has the form $\hat{\mathbf{x}}^0_{\mathrm{M}} = \mathbf{B}\hat{\mathbf{x}}_{\mathrm{LS}}$, for some $m \times m$ matrix $\mathbf{B}$.

In the case $\mathbf{x}_0 \neq \mathbf{0}$, Proposition 3.6 ensures that an affine minimax estimator is given by

$$\hat{\mathbf{x}}'_{\mathrm{M}} = \mathbf{B}(\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}_0) + \mathbf{x}_0 = \mathbf{B}\hat{\mathbf{x}}_{\mathrm{LS}} + (\mathbf{I} - \mathbf{B})\mathbf{x}_0, \tag{3.33}$$

which completes the proof. □

3.3 Conditional Dominance

When the parameter vector $\mathbf{x}$ is known to belong to some bounded set $\mathcal{S}$, one would expect an affine minimax MSE estimator $\hat{\mathbf{x}}_{\mathrm{M}}$ to perform at least as well as the LS estimator $\hat{\mathbf{x}}_{\mathrm{LS}}$. In fact, as we demonstrate in the following theorem, the MSE performance of $\hat{\mathbf{x}}_{\mathrm{M}}$ is *better* than that of $\hat{\mathbf{x}}_{\mathrm{LS}}$, for all values of $\mathbf{x}$ in $\mathcal{S}$ [20].

Theorem 3.8. *Let $\mathcal{S}$ be a bounded parameter set, and let $\hat{\mathbf{x}}_{\mathrm{M}}$ be an affine minimax MSE estimator for the parameter set $\mathcal{S}$. Then, the MSE of $\hat{\mathbf{x}}_{\mathrm{M}}$ is lower than the MSE of the LS estimator* (2.4)*, for all $\mathbf{x} \in \mathcal{S}$.*

Since performance improvement is not guaranteed for all values of $\mathbf{x}$, this theorem does not prove dominance of $\hat{\mathbf{x}}_{\mathrm{M}}$ in the usual sense of Definition 2.1; in fact, neither of the estimators dominates the other. However, improvement is guaranteed as long as the prior information $\mathbf{x} \in \mathcal{S}$ holds. We refer to this state of affairs as "conditional dominance."

Proof of Theorem 3.8. For any bounded $\mathcal{S}$, there exists a finite r such that $\mathcal{S}$ is bounded within the sphere $\{\mathbf{x} : \|\mathbf{x}\| \leq r\}$. By Theorem 3.2, the affine minimax MSE estimator for this sphere is

$$\hat{\mathbf{x}}_r = \frac{r^2}{r^2 + \epsilon_0} \hat{\mathbf{x}}_{\mathrm{LS}}. \tag{3.34}$$

Furthermore, by Theorem 3.2, the maximum MSE obtained by this estimator within the set $\mathcal{S}$ is

$$\left(\frac{r^2}{r^2 + \epsilon_0} \right) \epsilon_0, \tag{3.35}$$

which is smaller than ϵ_0, the MSE of the LS estimator. From (3.1), it follows that

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{M}}, \mathbf{x}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_r, \mathbf{x}) < \epsilon_0 \quad \text{for all } \mathbf{x} \in \mathcal{S}, \tag{3.36}$$

which concludes the proof. □

Chapter 4

Blind Minimax Estimation

In this chapter, we present a framework for generating a wide class of low-complexity, LS-dominating estimators, which are constructed from a simple, intuitive principle, called the blind minimax approach [20–22]. This approach is used as a basis for selecting and generating estimators tailored for given estimation problems. Many blind minimax estimators (BMEs) reduce to Stein-type estimators or extended Stein-type estimators (Subsection 2.3.2). Thus, we show analytically that the proposed estimators *always* achieve lower MSE than the LS estimator. Throughout the chapter, the assumption of Gaussian noise will be adopted.

4.1 The Blind Minimax Approach

Blind minimax estimators are based on the minimax approach (Chapter 3). As we have seen (Theorem 3.8), for any bounded parameter set $\mathcal{S}$, the affine minimax estimator over $\mathcal{S}$ achieves lower MSE than the LS estimator, as long as $\mathbf{x} \in \mathcal{S}$. However, this improvement may be attributed to the fact that the minimax estimator presupposes knowledge of the set $\mathcal{S}$, while the LS approach makes no such assumptions. In this chapter, we develop a technique which dominates the LS estimator without assuming *any* prior information regarding the parameter set $\mathbf{x}$. We assume only that the linear model (2.1) holds, that the risk function is the MSE, and that the noise $\mathbf{w}$ is Gaussian. Under these conditions, we propose the following two-stage technique.

Definition 4.1. A blind minimax estimator (BME) is an estimator constructed as follows.

1. A parameter set $\mathcal{S}$ is estimated from the measurements;

2. A minimax estimator designed for $\mathcal{S}$ is used to estimate the parameter vector $\mathbf{x}$.

The resulting algorithm may be viewed as a simple estimator, independent of this two-stage construction process. Indeed, our LS-dominance proofs are independent of the method by which the estimators are generated. In particular, the dominance results do not depend on the parameter actually lying within the estimated parameter set. However, the blind minimax technique provides a framework whereby many different estimators can be generated, and provides insight into the mechanism by which these estimators outperform the LS estimator.

Blind minimax estimators differ in the method by which the parameter set $\mathcal{S}$ is estimated. In general, the LS estimate $\hat{\mathbf{x}}_{\text{LS}}$ is used in some way in this procedure. As a preliminary (and unsuccessful) example, suppose we define the set $\mathcal{S}$ as a sphere of some constant radius L centered on $\hat{\mathbf{x}}_{\text{LS}}$, i.e.,

$$\mathcal{S} = \{\mathbf{x} : \|\mathbf{x} - \hat{\mathbf{x}}_{\text{LS}}\|^2 \leq L^2\}. \tag{4.1}$$

The minimax estimator for a sphere centered on a constant point $\mathbf{x}_0$ is given by (3.31). The blind minimax estimator in this case is then obtained by substituting $\hat{\mathbf{x}}_{\text{LS}}$ for $\mathbf{x}_0$ in (3.31). Unfortunately, this estimator simply reduces to $\hat{\mathbf{x}}_{\text{LS}}$ itself, so that our preliminary example provides no advantage over the LS estimator. This is a result of the fact that a minimax estimator centered on $\mathbf{x}_0$ is a weighted average between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_{\text{LS}}$; thus, when $\hat{\mathbf{x}}_{\text{LS}}$ is substituted for $\mathbf{x}_0$, the LS estimate is obtained. Furthermore, as demonstrated by Proposition 3.7, a generalized average between $\hat{\mathbf{x}}_{\text{LS}}$ and $\mathbf{x}_0$ is obtained for ellipsoidal parameter sets as well. Clearly, then, the choice of a parameter set centered on $\hat{\mathbf{x}}_{\text{LS}}$ yields no improvement over the LS estimator. However, other blind minimax constructions prove to be very effective, as demonstrated in the following sections.

4.2 The Spherical Blind Minimax Estimator

As we have seen in Section 4.1, blind minimax estimators differ in the method by which the parameter set is estimated. In the following, we use a spherical parameter set centered on the origin,

$$\mathcal{S} = \{\mathbf{x} : \|\mathbf{x}\|^2 \leq L^2\}, \tag{4.2}$$

where the sphere radius L^2 is estimated from the measurements. (Later in this section, we will extend the results to sets centered on any constant point.) As we have seen in Theorem 3.2, for any constant value of L^2, the unique affine minimax estimator is defined by the closed form (3.10). The *spherical BME* (SBME) will have the same form, with L^2 estimated from measurements.

As an estimate of L^2, we seek a value as close as possible to $\|\mathbf{x}\|^2$: a smaller value would exclude the true vector $\mathbf{x}$ from the parameter set, while a larger value would yield an overly conservative estimator. Since $\mathbf{x}$ is unknown, a natural alternative is to use $\hat{\mathbf{x}}_{\mathrm{LS}}$ instead. Thus, we propose to estimate L^2 as $\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2$. The SBME is then given by

$$\hat{\mathbf{x}}_{\mathrm{SBM}} = \frac{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2}{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 + \epsilon_0}\hat{\mathbf{x}}_{\mathrm{LS}}. \tag{4.3}$$

In the i.i.d. case, the SBME reduces to the well-known Thompson estimator. Under suitable conditions, Thompson's technique is known to strictly dominate the LS estimator, meaning that it achieves lower MSE for all values of $\mathbf{x}$. However, the SBME is equally well-defined for the non-i.i.d. case. As we shall see, the SBME strictly dominates the LS estimator in the non-i.i.d. case, and can thus be viewed as a generalization of Thompson's results. In Section 4.4 we will demonstrate that the blind minimax approach can be used to derive generalizations of additional well-known methods, including Stein's estimator.

Up to this point, we have arbitrarily chosen the parameter set to be centered on the origin. The result was a weighted average between the LS estimate and the origin. The weight given to the LS estimate may be viewed as a restraint, which lessens the effect of measurement noise. As we shall see, the proposed BMEs outperform the LS estimator, illustrating the fact that the LS estimator is an overestimate. However, the choice of a parameter set centered on the origin is completely arbitrary; BMEs may be constructed around any constant center point $\mathbf{x}_0$ [17]. This would result in a weighted average between the LS estimator and $\mathbf{x}_0$, which may be useful if the parameter vector is expected to lie near a particular point. Thus, the "off-center" SBME is given by

$$\hat{\mathbf{x}} = \left(\frac{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2}{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 + \epsilon_0}\right)\hat{\mathbf{x}}_{\mathrm{LS}} + \left(\frac{\epsilon_0}{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 + \epsilon_0}\right)\mathbf{x}_0. \tag{4.4}$$

All dominance results continue to hold for the off-center estimators as well. In the sequel, we assume $\mathbf{x}_0 = \mathbf{0}$ merely for the sake of notational simplicity.

The following theorem demonstrates that the SBME is guaranteed to outperform the LS estimator in terms of MSE.

Theorem 4.1. *Suppose $d > 4$, where the effective dimension d is given by* (2.22)*. Then, the SBME strictly dominates the LS estimator.*

Note that the SBME is a special case of the estimator

$$\hat{\mathbf{x}}_b = \left(1 - \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^*\hat{\mathbf{x}}_{\mathrm{LS}}}\right)\hat{\mathbf{x}}_{\mathrm{LS}}, \tag{4.5}$$

in which $b = \epsilon_0$. Thus, rather than proving Theorem 4.1, we prove the following, more general proposition, which will also be used in Section 4.4.

Proposition 4.2. *Suppose $d > 4$, where the effective dimension d is given by* (2.22)*. Then, the estimator* (4.5) *dominates the LS estimator, for any $b \geq 0$.*

The proof of Proposition 4.2 makes use of the following result, known as Stein's lemma. In essence, Stein's lemma is a restatement of the law of integration by parts, formulated for a useful special case. The proof of a slightly more general version of Stein's lemma may be found in [2, Th. 1.5.15].

Lemma 4.3 (Stein). *Let $\hat{\mathbf{v}} \sim N_p(\mathbf{v}, \mathbf{I})$, and let $g(\hat{\mathbf{v}})$ be a differentiable function such that $E\left\{ \left| \frac{\partial g(\hat{\mathbf{v}})}{\partial \hat{v}_i} \right| \right\} < \infty$ for all i. Then,*

$$E\left\{ \frac{\partial g(\hat{\mathbf{v}})}{\partial \hat{v}_i} \right\} = -E\{ g(\hat{\mathbf{v}})(v_i - \hat{v}_i) \}. \tag{4.6}$$

Proof of Proposition 4.2. We will prove that $\hat{\mathbf{x}}_b$ strictly dominates $\hat{\mathbf{x}}_{\mathrm{LS}}$ for any $b \geq 0$. First, note that the MSE of $\hat{\mathbf{x}}_b$ is given by

$$\begin{aligned} \mathrm{MSE}(\hat{\mathbf{x}}_b, \mathbf{x}) &= E\{ \|\mathbf{x} - \hat{\mathbf{x}}_b\|^2 \} \\ &= E\left\{ \left((\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LS}}) + \frac{\epsilon_0 \hat{\mathbf{x}}_{\mathrm{LS}}}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \right)^* \left((\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LS}}) + \frac{\epsilon_0 \hat{\mathbf{x}}_{\mathrm{LS}}}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \right) \right\} \\ &= \epsilon_0 + E\left\{ \frac{\epsilon_0^2 \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}}{(b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}})^2} \right\} + 2E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \hat{\mathbf{x}}_{\mathrm{LS}}^* (\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LS}}) \right\}, \end{aligned} \tag{4.7}$$

where ϵ_0 is defined by (2.5). Define

$$\mathbf{Q} = \mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H}, \tag{4.8}$$

and denote the eigenvalue decomposition of $\mathbf{Q}$ by $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$, so that $\mathbf{V}$ is unitary and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots \lambda_m)$. Define $\hat{\mathbf{v}} = \mathbf{V}^* \mathbf{Q}^{1/2} \hat{\mathbf{x}}_{\mathrm{LS}}$ and $\mathbf{v} = \mathbf{V}^* \mathbf{Q}^{1/2} \mathbf{x}$. Note the following relations between $\hat{\mathbf{v}}$ and $\hat{\mathbf{x}}_{\mathrm{LS}}$:

$$\begin{aligned} \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}} &= \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}, \\ \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \mathbf{v} &= \hat{\mathbf{x}}_{\mathrm{LS}}^* \mathbf{x}, \\ \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-2} \hat{\mathbf{v}} &= \hat{\mathbf{x}}_{\mathrm{LS}}^* \mathbf{Q}^{-1} \hat{\mathbf{x}}_{\mathrm{LS}}. \end{aligned} \tag{4.9}$$

Using these properties, we now evaluate the third term in (4.7), obtaining

$$\begin{aligned} E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \hat{\mathbf{x}}_{\mathrm{LS}}^* (\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LS}}) \right\} &= E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}}} \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} (\mathbf{v} - \hat{\mathbf{v}}) \right\} \\ &= E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}}} \sum_{i=1}^{p} \lambda_i^{-1} \hat{v}_i (v_i - \hat{v}_i) \right\} \\ &= \epsilon_0 \sum_{i=1}^{p} \lambda_i^{-1} E\left\{ \frac{\hat{v}_i (v_i - \hat{v}_i)}{b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}}} \right\}. \end{aligned} \tag{4.10}$$

Let

$$g_i(\hat{\mathbf{v}}) \triangleq \frac{\hat{v}_i}{b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}}}, \tag{4.11}$$

and note that $\hat{\mathbf{v}}$ is distributed normally with mean $\mathbf{v}$ and covariance $\mathbf{I}$. We can thus apply Lemma 4.3 to obtain

$$\begin{aligned} E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \hat{\mathbf{x}}_{\mathrm{LS}}^* (\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{LS}}) \right\} &= -\epsilon_0 \sum_i \lambda_i^{-1} E\left\{ \frac{1}{b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}}} - 2 \frac{\lambda_i^{-1} \hat{v}_i^2}{(b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}})^2} \right\} \\ &= -\epsilon_0 E\left\{ \frac{\mathrm{Tr}(\mathbf{\Lambda}^{-1})}{b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}}} \right\} + 2\epsilon_0 E\left\{ \frac{\hat{\mathbf{v}}^* \mathbf{\Lambda}^{-2} \hat{\mathbf{v}}}{(b + \hat{\mathbf{v}}^* \mathbf{\Lambda}^{-1} \hat{\mathbf{v}})^2} \right\} \\ &= -\epsilon_0 E\left\{ \frac{\mathrm{Tr}(\mathbf{Q}^{-1})}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \right\} + 2\epsilon_0 E\left\{ \frac{\hat{\mathbf{x}}_{\mathrm{LS}}^* \mathbf{Q}^{-1} \hat{\mathbf{x}}_{\mathrm{LS}}}{(b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}})^2} \right\}. \end{aligned} \tag{4.12}$$

Substituting this result back into (4.7), we have

$$\begin{aligned} \mathrm{MSE}(\hat{\mathbf{x}}_b, \mathbf{x}) &= \epsilon_0 + E\left\{ \frac{\epsilon_0^2 \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}}{(b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}})^2} \right\} - 2\epsilon_0 E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \right\} + 4\epsilon_0 E\left\{ \frac{\hat{\mathbf{x}}_{\mathrm{LS}}^* \mathbf{Q}^{-1} \hat{\mathbf{x}}_{\mathrm{LS}}}{(b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}})^2} \right\} \\ &= \epsilon_0 + E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \left(\epsilon_0 \frac{\hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} - 2\epsilon_0 + 4 \frac{\hat{\mathbf{x}}_{\mathrm{LS}}^* \mathbf{Q}^{-1} \hat{\mathbf{x}}_{\mathrm{LS}}}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \right) \right\}. \end{aligned} \tag{4.13}$$

Since $b \geq 0$,

$$\begin{aligned} \mathrm{MSE}(\hat{\mathbf{x}}_b, \mathbf{x}) &\leq \epsilon_0 + E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \left(\frac{\epsilon_0 \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}}{\hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} - 2\epsilon_0 + 4 \frac{\hat{\mathbf{x}}_{\mathrm{LS}}^* \mathbf{Q}^{-1} \hat{\mathbf{x}}_{\mathrm{LS}}}{\hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \right) \right\} \\ &\leq \epsilon_0 + E\left\{ \frac{\epsilon_0}{b + \hat{\mathbf{x}}_{\mathrm{LS}}^* \hat{\mathbf{x}}_{\mathrm{LS}}} \left(-\epsilon_0 + 4\epsilon_{\max} \right) \right\}, \end{aligned} \tag{4.14}$$

where $\epsilon_{\max}$ is the largest eigenvalue of $\mathbf{Q}^{-1}$. Note that $d = \epsilon_0 / \epsilon_{\max}$. Furthermore, if $\epsilon_0 > 4\epsilon_{\max}$, then the expectation is taken over a strictly negative range, and hence $\mathrm{MSE}(\hat{\mathbf{x}}_b, \mathbf{x})$ is always lower than ϵ_0. As a result, $\hat{\mathbf{x}}_b$ strictly dominates $\hat{\mathbf{x}}_{\mathrm{LS}}$ when $d > 4$. □

As we have shown, in terms of MSE, the SBME outperforms the LS estimator, providing us with a first example of the power of blind minimax estimation. The SBME is a shrinkage estimator, i.e., it consists of the LS estimator multiplied by a gain factor smaller than one. The SBME thus illustrates the fact that the LS technique tends to be an overestimate, and shrinkage can improve its performance. However, in some applications, such as image reconstruction, a gain factor has no effect on the end result. In the next section, we use the blind minimax approach to develop a non-shrinkage estimator, which also dominates the LS estimator.

4.3 The Ellipsoidal Blind Minimax Estimator

Occasionally, one must use the MSE in place of a more accurate error measure, which may be overly complex or difficult to quantify. For example, in communication systems, one is

often interested in minimizing the bit error rate (BER). However, BER minimization techniques are generally too computationally expensive to be practical. In many situations, one chooses instead to minimize the MSE between the transmitted and reconstructed symbols [36]. This is done in the hope that low MSE provides low BER. However, MSE and BER are not always directly related. In the previous section, it was shown that SBMEs achieve lower MSE than the LS estimator. Yet in a binary communication system, only the signs of the estimated elements are used to evaluate the received bits. Since the SBMEs are shrinkage estimators, they yield the same sign as the LS estimator. Thus, in a binary communication system, the BER obtained by SBMEs is identical to the BER of the LS estimator.

Shrinkage estimators are therefore not applicable to *all* estimation problems. In this section, we develop a non-shrinkage estimator by considering ellipsoidal, rather than spherical, parameter sets as the basis for the blind minimax technique. In some situations, this estimator also outperforms the SBME in terms of MSE.

As we have seen in Subsection 2.3.3, not all elements of the least-squares estimate $\hat{\mathbf{x}}_{\mathrm{LS}}$ are equally trustworthy, and several researchers have proposed shrinking each measurement separately according to its variance. However, there has been disagreement as to whether high-variance components should be shrunk more or less, and little justification has been given to either choice.

The blind minimax approach provides a natural framework for resolving these disputes. To see this, note that $\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{x} + \mathbf{u}$, where $\mathbf{u} \sim N_m(\mathbf{0}, \mathbf{Q}^{-1})$, and $\mathbf{Q}$ is given by (4.8). The SBME was constructed by using $\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2$ as an estimate for $\|\mathbf{x}\|^2$. However, since the noise $\mathbf{u}$ is colored, it is sensible to first whiten the noise, obtaining

$$\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{Q}^{1/2}\mathbf{x} + \tilde{\mathbf{u}}, \tag{4.15}$$

where $\tilde{\mathbf{u}} \sim N_m(\mathbf{0}, \mathbf{I})$. One may then estimate $\|\mathbf{Q}^{1/2}\mathbf{x}\|^2$ as $\|\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}\|^2$, obtaining the *ellipsoidal BME* (EBME). Such an estimate can be readily incorporated into the blind minimax framework by using an ellipsoidal parameter set, $\mathcal{S} = \{\mathbf{x} : \mathbf{x}^*\mathbf{Q}\mathbf{x} \leq L^2\}$, rather than the spherical parameter set of the SBME; here, L^2 would be estimated as $\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}$. The ellipsoidal parameter set is elongated in directions of low noise, resulting in lower shrinkage for those directions, and thus resolving the aforementioned dilemma. In the i.i.d. case, $\mathbf{Q} = \mathbf{I}$, and the estimator reduces to the SBME.

Theorem 3.3 provides a closed form for minimax estimators of an ellipsoidal parameter set. By substituting the value of L^2 into this closed form, we obtain the following expression for the

EBME.

Proposition 4.4 (Closed-Form EBME). *Let* $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$ *be the eigenvalue decomposition of* $\mathbf{Q}$*, so that* $\mathbf{V}$ *is unitary,* $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots \lambda_m)$*, and* $\lambda_1 \geq \cdots \geq \lambda_m$*. The EBME is then given by*

$$\hat{\mathbf{x}}_{\text{EBM}} = \mathbf{V} \operatorname{diag}(\mathbf{0}_k, \mathbf{1}_{m-k}) \mathbf{V}^* \left(\mathbf{I} - \alpha \mathbf{Q}^{1/2} \right) \hat{\mathbf{x}}_{\text{LS}}, \qquad \text{for } \hat{\mathbf{x}}_{\text{LS}} \neq \mathbf{0} \tag{4.16a}$$

$$\hat{\mathbf{x}}_{\text{EBM}} = \mathbf{0}, \qquad \text{for } \hat{\mathbf{x}}_{\text{LS}} = \mathbf{0}. \tag{4.16b}$$

Here

$$\alpha = \frac{\sum_{i=k+1}^{m} \lambda_i^{-1/2}}{\hat{\mathbf{x}}_{\text{LS}}^* \mathbf{Q} \hat{\mathbf{x}}_{\text{LS}} + m - k}, \tag{4.17}$$

and k *is the smallest integer* $0 \leq k \leq m-1$ *such that*

$$\alpha < \lambda_{k+1}^{-1/2}. \tag{4.18}$$

Proof. In the case $\hat{\mathbf{x}}_{\text{LS}} = \mathbf{0}$, we are to find the linear minimax estimator for the set $\mathcal{S} = \{\mathbf{0}\}$. Clearly the linear minimax estimator in this case is $\hat{\mathbf{x}} = \mathbf{0}$. For all other values of $\hat{\mathbf{x}}_{\text{LS}}$, we are to find the linear minimax estimator for the set $\mathcal{S} = \{\mathbf{x} : \mathbf{x}^* \mathbf{Q} \mathbf{x} \leq L^2\}$, where $L^2 = \hat{\mathbf{x}}_{\text{LS}}^* \mathbf{Q} \hat{\mathbf{x}}_{\text{LS}} > 0$. Substituting this value of L^2 into Proposition 1 of [6] yields (4.16a). □

We note that it is always possible to find a value k which satisfies (4.18). In particular, for $k = m-1$, we have

$$\alpha = \frac{\lambda_m^{-1/2}}{\hat{\mathbf{x}}_{\text{LS}}^* \mathbf{Q} \hat{\mathbf{x}}_{\text{LS}} + 1}, \tag{4.19}$$

and since $\hat{\mathbf{x}}_{\text{LS}}^* \mathbf{Q} \hat{\mathbf{x}}_{\text{LS}} > 0$, this satisfies the requirement (4.18).

While the closed form of the EBME appears somewhat more intimidating than that of the SBME, the computational complexities of the two estimators are comparable. The major difference is the calculation of the value k, for which m divisions are required. Like the SBME, the EBME also dominates the LS estimator under suitable conditions, as shown in the following theorem.

Theorem 4.5. *Suppose* $\operatorname{Tr}(\mathbf{Q}^{-1/2}) > 4\epsilon_{\max}^{1/2}$*, where* $\epsilon_{\max}^{1/2}$ *is the largest eigenvalue of* $\mathbf{Q}^{-1/2}$*, and* $\mathbf{Q}$ *is given by* (4.8)*. Then, the EBME strictly dominates* $\hat{\mathbf{x}}_{\text{LS}}$*.*

The proof of Theorem 4.5 is based on an analogy between the diagonal matrix $\operatorname{diag}(\mathbf{0}_k, \mathbf{1}_{m-k})$ in (4.16a) and Baranchik's positive-part modification of the James-Stein estimator (see Subsection 2.3.3). Baranchik proposed using a shrinkage factor of 0 whenever the James-Stein estimator uses negative shrinkage, and showed that the resulting *positive-part estimator* dominates the James-Stein estimator. Although the EBME is not a shrinkage estimator, it resembles

Baranchik's modification. To see this, consider the estimator $\hat{\mathbf{x}}_0$ obtained by removing the term $\operatorname{diag}(\mathbf{0}_k, \mathbf{1}_{m-k})$ from (4.16a),

$$\begin{aligned}\hat{\mathbf{x}}_0 &= (\mathbf{I} - \alpha\mathbf{Q}^{1/2})\hat{\mathbf{x}}_{\mathrm{LS}} \\ &= \mathbf{V}\operatorname{diag}\left(1 - \alpha\lambda_1^{1/2}, \cdots 1 - \alpha\lambda_m^{1/2}\right)\mathbf{V}^*\hat{\mathbf{x}}_{\mathrm{LS}}.\end{aligned} \tag{4.20}$$

Since $\alpha \geq \lambda_i^{-1/2}$ for all $i \leq k$, this would introduce negative componentwise shrinkage for the first k eigenvectors of $\mathbf{V}$. Thus, the using the EBME inherently guarantees positive shrinkage, and does not require one to resort to post factum modifications such as those proposed by Baranchik. The following proposition, which is a generalization of Baranchik's result, asserts that the MSE can be reduced by eliminating this negative shrinkage.

Proposition 4.6 (Generalized Positive-Part Estimator). *Let $\hat{\mathbf{x}}$ be any estimator of the form $\hat{\mathbf{x}} = \mathbf{V}\mathbf{D}\mathbf{V}^*\hat{\mathbf{x}}_{\mathrm{LS}}$, where $\mathbf{D}$ is a diagonal matrix, whose diagonal elements d_i are functions of the random variable $\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}$. Suppose at least one of the elements d_i is negative with nonzero probability. Then, $\hat{\mathbf{x}}$ is dominated by the generalized positive-part estimator*

$$\hat{\mathbf{x}}_+ = \mathbf{V}\mathbf{D}_+\mathbf{V}^*\hat{\mathbf{x}}_{\mathrm{LS}}, \tag{4.21}$$

where $\mathbf{D}_+$ is a diagonal matrix with diagonal elements $d_{i+} = \max(0, d_i)$.

Proof. Our proof follows that of Baranchik [33]. We will show that $\mathrm{MSE}(\hat{\mathbf{x}}) - \mathrm{MSE}(\hat{\mathbf{x}}_+)$ is nonnegative for all $\mathbf{x}$, and positive for any value of $\mathbf{x}$ whose elements are all nonzero.

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{x}}) - \mathrm{MSE}(\hat{\mathbf{x}}_+) &= E\{\|\hat{\mathbf{x}} - \mathbf{x}\|^2\} - E\{\|\hat{\mathbf{x}}_+ - \mathbf{x}\|^2\} \\ &= E\{\|\hat{\mathbf{x}}\|^2 - \|\hat{\mathbf{x}}_+\|^2\} - 2E\{\hat{\mathbf{x}}^*\mathbf{x} - \hat{\mathbf{x}}_+^*\mathbf{x}\} \\ &= E\{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{V}(\mathbf{D}^2 - \mathbf{D}_+^2)\mathbf{V}^*\hat{\mathbf{x}}_{\mathrm{LS}}\} \\ &\quad - 2E\{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{V}(\mathbf{D} - \mathbf{D}_+)\mathbf{V}^*\mathbf{x}\}.\end{aligned} \tag{4.22}$$

Since $d_i^2 - d_{i+}^2 \geq 0$ for all i, the first term in (4.22) is nonnegative. Hence, to prove the proposition, it suffices to show that $E\{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{V}(\mathbf{D} - \mathbf{D}_+)\mathbf{V}^*\mathbf{x}\}$ is nonpositive for all $\mathbf{x}$, and negative for values $\mathbf{x}$ with nonzero elements.

To this end, define $\mathbf{z} = \mathbf{V}^*\mathbf{x}$ and $\hat{\mathbf{z}} = \mathbf{V}^*\hat{\mathbf{x}}_{\mathrm{LS}}$. We note that $\hat{\mathbf{z}} \sim N_m(\mathbf{z}, \mathbf{\Lambda}^{-1})$, so that the elements of $\hat{\mathbf{z}}$ are statistically independent. To calculate $E\{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{V}(\mathbf{D} - \mathbf{D}_+)\mathbf{V}^*\mathbf{x}\}$, we condition on $\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}$, obtaining

$$\begin{aligned}E\{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{V}(\mathbf{D} - \mathbf{D}_+)\mathbf{V}^*\mathbf{x}\} &= E\{E\{\hat{\mathbf{z}}^*(\mathbf{D} - \mathbf{D}_+)\mathbf{z}|\hat{\mathbf{z}}^*\mathbf{\Lambda}\hat{\mathbf{z}}\}\} \\ &= E\left\{\sum_{i=1}^m (d_i - d_{i+})E\{\hat{z}_i z_i|\hat{\mathbf{z}}^*\mathbf{\Lambda}\hat{\mathbf{z}}\}\right\},\end{aligned} \tag{4.23}$$

where we used the fact that $\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}} = \hat{\mathbf{z}}^{*}\mathbf{\Lambda}\hat{\mathbf{z}}$, and that d_i and d_{i+} are deterministic when conditioned on $\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}$. For each i, we further condition on $|\hat{z}_i|$, to obtain

$$\begin{aligned} E\{\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{V}(\mathbf{D}-\mathbf{D}_{+})\mathbf{V}^{*}\mathbf{x}\} &= E\left\{\sum_{i=1}^{m}(d_i-d_{i+})E\big\{\hat{z}_i z_i\big|\hat{\mathbf{z}}^{*}\mathbf{\Lambda}\hat{\mathbf{z}},|\hat{z}_i|\big\}\right\} \\ &= E\left\{\sum_{i=1}^{m}(d_i-d_{i+})|\hat{z}_i z_i|E\big\{\mathrm{sgn}(\hat{z}_i z_i)\big|\hat{\mathbf{z}}^{*}\mathbf{\Lambda}\hat{\mathbf{z}},|\hat{z}_i|\big\}\right\}. \end{aligned} \tag{4.24}$$

Since $\hat{z}_i \sim N(z_i,\lambda_i^{-1})$, we have that, for any $z_i \neq 0$,

$$\Pr\big\{\mathrm{sgn}(\hat{z}_i)=\mathrm{sgn}(z_i)\big|\hat{\mathbf{z}}^{*}\mathbf{\Lambda}\hat{\mathbf{z}},|\hat{z}_i|\big\} > \Pr\big\{\mathrm{sgn}(\hat{z}_i)\neq\mathrm{sgn}(z_i)\big|\hat{\mathbf{z}}^{*}\mathbf{\Lambda}\hat{\mathbf{z}},|\hat{z}_i|\big\}. \tag{4.25}$$

This is because, given $|\hat{z}_i|$, we have that either $\hat{z}_i = z_i$ or that $\hat{z}_i = -z_i$. From the pdf of $\hat{z}_i$ it is evident that the latter option has lower probability. It follows that $E\{\mathrm{sgn}(\hat{z}_i z_i)|\hat{\mathbf{z}}^{*}\mathbf{\Lambda}\hat{\mathbf{z}},|\hat{z}_i|\} \geq 0$, with strict inequality for $z_i \neq 0$. Therefore, all terms in (4.24) are nonnegative, except for $(d_i - d_{i+})$, which is nonpositive. As a result, (4.24) (and hence (4.22)) is nonpositive for all $\mathbf{x}$, so that the MSE of $\hat{\mathbf{x}}_{+}$ is never higher than that of $\hat{\mathbf{x}}$.

We must also show that, for some $\mathbf{x}$, (4.24) is strictly negative. To this end, we choose $\mathbf{x}$ for which all elements are nonzero; as a result, all terms in (4.24) are strictly positive with probability 1, except for $(d_i - d_{i+})$. The latter term is negative when $d_i < 0$ and zero otherwise. Since d_i is negative with nonzero probability for at least one value of i, we conclude that for the chosen value of $\mathbf{x}$, (4.24) is strictly negative, completing the proof of Proposition 4.6. □

This generalization of positive part estimation is now used to prove Theorem 4.5.

Proof of Theorem 4.5. In calculating the MSE of $\hat{\mathbf{x}}_{\mathrm{EBM}}$, we ignore the case $\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{0}$, since this case occurs with zero probability. We must therefore show that, for $\hat{\mathbf{x}}_{\mathrm{LS}} \neq \mathbf{0}$, the estimator (4.16a) strictly dominates the LS estimator. To do this, we show that $\hat{\mathbf{x}}_0$ of (4.20) strictly dominates the LS estimator. The result then follows from Proposition 4.6, since $\hat{\mathbf{x}}_{\mathrm{EBM}}$ is the positive part of $\hat{\mathbf{x}}_0$.

Denoting $s = \sum_{i=k+1}^{m}\lambda_i^{-1/2}$, the MSE of $\hat{\mathbf{x}}_0$ is given by

$$\begin{aligned} \mathrm{MSE} &= E\left\{\left\|\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{LS}}+\frac{s\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}}{\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k}\right\|^{2}\right\} \\ &= \epsilon_0 + E\left\{\frac{s^2\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}}{(\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k)^2}\right\} \\ &\quad + 2E\left\{\frac{s(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{LS}})^{*}\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}}{\hat{\mathbf{x}}_{\mathrm{LS}}^{*}\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k}\right\}. \end{aligned} \tag{4.26}$$

We now define $\hat{\mathbf{v}} = \mathbf{V}^*\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}$ and $\mathbf{v} = \mathbf{V}^*\mathbf{Q}^{1/2}\mathbf{x}$. Using this notation, the third term in (4.26) may be written as

$$\begin{aligned} A_3 &\triangleq E\left\{ \frac{s(\mathbf{x}-\hat{\mathbf{x}}_{\mathrm{LS}})^*\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k} \right\} \\ &= E\left\{ \frac{s(\mathbf{v}-\hat{\mathbf{v}})^*\mathbf{\Lambda}^{-1/2}\hat{\mathbf{v}}}{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k} \right\} \\ &= \sum_{i=1}^{m} \lambda_i^{-1/2} E\left\{ \frac{s(v_i-\hat{v}_i)\hat{v}_i}{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k} \right\}, \end{aligned} \tag{4.27}$$

where we have used the fact that $\hat{\mathbf{v}} \sim N_m(\mathbf{v}, \mathbf{I})$. Let

$$g_i(\hat{\mathbf{v}}) = \frac{s\hat{v}_i}{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k}, \tag{4.28}$$

noting that k is implicitly dependent on $\hat{\mathbf{v}}$, and that s is implicitly dependent on k. Thus, $g_i(\hat{\mathbf{v}})$ is discontinuous for some values of $\hat{\mathbf{v}}$, namely, those values for which $\alpha = \lambda_i^{-1/2}$. However, these values of $\hat{\mathbf{v}}$ occur with probability zero; for all other values, k (and hence s) are constant for sufficiently small changes in $\hat{\mathbf{v}}$. Thus,

$$\frac{\partial g_i(\hat{\mathbf{v}})}{\partial \hat{v}_i} = s\frac{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k-2\hat{v}_i^2}{(\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k)^2} \quad \text{with probability 1,} \tag{4.29}$$

and $E\left\{ \left| \frac{\partial g_i(\hat{\mathbf{v}})}{\partial \hat{v}_j} \right| \right\} < \infty$ for all j. Using Lemma 4.3, we have

$$E\left\{ \frac{s(v_i-\hat{v}_i)\hat{v}_i}{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k} \right\} = -E\left\{ s\frac{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k-2\hat{v}_i^2}{(\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k)^2} \right\}. \tag{4.30}$$

Combining (4.30) with (4.27), we obtain

$$\begin{aligned} A_3 &= -\sum_{i=1}^{m} \lambda_i^{-1/2} E\left\{ s\frac{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k-2\hat{v}_i^2}{(\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k)^2} \right\} \\ &= E\left\{ -\frac{s\,\mathrm{Tr}(\mathbf{Q}^{-1/2})}{\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k} + 2s\frac{\hat{\mathbf{v}}^*\mathbf{\Lambda}^{-1/2}\hat{\mathbf{v}}}{(\hat{\mathbf{v}}^*\hat{\mathbf{v}}+m-k)^2} \right\} \\ &= E\left\{ -\frac{s\,\mathrm{Tr}(\mathbf{Q}^{-1/2})}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k} + 2s\frac{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}}{(\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k)^2} \right\}. \end{aligned} \tag{4.31}$$

We note that $k < m$, so that $\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}} + m - k \geq 0$. Hence

$$\begin{aligned} A_3 &\leq E\left\{ \frac{s}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k} \left(-\mathrm{Tr}(\mathbf{Q}^{-1/2}) + 2\frac{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}^{1/2}\hat{\mathbf{x}}_{\mathrm{LS}}}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}} \right) \right\} \\ &\leq E\left\{ \frac{s}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k} \left(-\mathrm{Tr}(\mathbf{Q}^{-1/2}) + 2\epsilon_{\max}^{1/2} \right) \right\}, \end{aligned} \tag{4.32}$$

where $\epsilon_{\max}^{1/2}$ is the largest eigenvalue of $\mathbf{Q}^{-1/2}$. Substituting this result back into (4.26), and using the fact that $s \leq \mathrm{Tr}(\mathbf{Q}^{-1/2})$, yields

$$\mathrm{MSE} \leq \epsilon_0 + E\left\{ \frac{s\left(-\mathrm{Tr}(\mathbf{Q}^{-1/2}) + 4\epsilon_{\max}^{1/2}\right)}{\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}+m-k} \right\}. \tag{4.33}$$

If $\mathrm{Tr}(\mathbf{Q}^{-1/2}) > 4\epsilon_{\max}^{1/2}$, then the expectation above is negative, so that $\hat{\mathbf{x}}_0$ (and hence $\hat{\mathbf{x}}_{\mathrm{EBM}}$) strictly dominate the LS estimator. □

Thus far, we have presented two examples of blind minimax estimators which dominate the LS method. Both estimators were extensions of Thompson's technique to the non-i.i.d. case. In the next section, we demonstrate that other blind minimax estimators extend different LS-dominating techniques, notably Stein's estimator and Baranchik's positive-part improvement.

4.4 Relation to Stein-type Estimation

In Section 4.2, the SBME (4.3) was constructed by using $L^2 = \|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2$ as an estimate of $\|\mathbf{x}\|^2$. However, the fact that shrinkage estimators such as the SBME dominate the LS estimator indicates that $\hat{\mathbf{x}}_{\mathrm{LS}}$ is in fact an overestimate of $\mathbf{x}$. It is arguably more accurate to use a smaller estimate than $\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2$. In particular, it is readily shown that

$$E\{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2\} = \|\mathbf{x}\|^2 + \epsilon_0. \tag{4.34}$$

Hence, one may opt to use

$$L^2 = \|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 - \epsilon_0 \tag{4.35}$$

as an estimate of $\|\mathbf{x}\|^2$. It is important to note that such a value of L^2 cannot be used for minimax estimation, since L^2 is negative with nonzero probability; a parameter set with negative radius is undefined. However, substituting (4.35) into a minimax estimator, as per the blind minimax approach, can still lead to well-defined estimators. In particular, substituting (4.35) into the spherical minimax estimator (3.10) yields the *balanced BME*

$$\hat{\mathbf{x}}_{\mathrm{BBM}} = \left(1 - \frac{\epsilon_0}{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2}\right)\hat{\mathbf{x}}_{\mathrm{LS}}. \tag{4.36}$$

A striking property of the balanced BME is that it reduces to Stein's estimator [10] in the i.i.d. case. Both estimators are well-defined unless $\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{0}$, an event which has zero probability. Furthermore, the balanced BME extends Stein's estimator, in that it continues to dominate the LS estimator for the non-i.i.d. case, under suitable conditions. This is shown by the following theorem.

Theorem 4.7. *Suppose $\epsilon_0/\epsilon_{\max} > 4$, where ϵ_0 is given by* (2.5)*, $\epsilon_{\max}$ is the largest eigenvalue of $\mathbf{Q}^{-1}$, and $\mathbf{Q}$ is given by* (4.8)*. Then, the balanced BME* (4.36) *strictly dominates the LS estimator.*

Proof. The theorem follows by substituting $b = 0$ in Proposition 4.2. □

A well-known drawback of Stein's estimator is that it sometimes causes negative shrinkage, i.e., the shrinkage factor in (4.36) is negative with nonzero probability. This is known to increase the MSE [33]. From the blind minimax perspective, this negative shrinkage is a result of the fact that L^2 can become negative. Thus, it is natural to replace (4.35) with

$$L^2 = \left[\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 - \epsilon_0\right]_+ \tag{4.37}$$

where $[a]_+ = \max(a, 0)$. Substituting this value of L^2 into the spherical minimax estimator yields the *positive-part BME*, given by

$$\hat{\mathbf{x}}_{\mathrm{PBM}} = \left(\frac{[\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 - \epsilon_0]_+}{[\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 - \epsilon_0]_+ + \epsilon_0}\right)\hat{\mathbf{x}}_{\mathrm{LS}}. \tag{4.38}$$

Note that when $\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2 - \epsilon_0 < 0$, the estimator $\hat{\mathbf{x}}_{\mathrm{PBM}}$ equals $\mathbf{0}$; in all other cases, $\hat{\mathbf{x}}_{\mathrm{PBM}} = \hat{\mathbf{x}}_{\mathrm{BBM}}$. Thus, (4.38) may be written as

$$\hat{\mathbf{x}}_{\mathrm{PBM}} = \left[1 - \frac{\epsilon_0}{\|\hat{\mathbf{x}}_{\mathrm{LS}}\|^2}\right]_+ \hat{\mathbf{x}}_{\mathrm{LS}}. \tag{4.39}$$

In other words, $\hat{\mathbf{x}}_{\mathrm{PBM}}$ is the positive part of the balanced BME. Specifically, in the i.i.d. case, $\hat{\mathbf{x}}_{\mathrm{PBM}}$ is the positive-part correction of Stein's estimator. In the i.i.d. case, Baranchik [33] demonstrates that $\hat{\mathbf{x}}_{\mathrm{PBM}}$ dominates $\hat{\mathbf{x}}_{\mathrm{BBM}}$. An interesting question for further research is whether the dominance property holds in the non-i.i.d. case as well.

As we have seen, Stein's estimator can be viewed as a blind minimax estimator in which L^2 is estimated in an unbiased way, as in (4.35). The result is that L^2 is negative with nonzero probability; this adversely affects performance, by introducing negative shrinkage. Although L^2 can be corrected by zeroing out all negative values, as in (4.37), this results in a non-differentiable function, and improvement is not guaranteed for the non-i.i.d. case. Rather than introduce this correction mechanism, we propose the use of the direct estimates of L^2 presented in the previous sections. These estimate L^2 as a non-negative values, and thus inherently avoid negative shrinkage. In the next section, we compare the estimators of Sections 4.2 and 4.3 with other extended Stein estimators.

4.5 Numerical Results

Estimator performance generally depends on a number of operating conditions, including the effective dimension, the signal-to-noise ratio (SNR), the eigenvalues $\lambda_1, \ldots \lambda_m$ of $\mathbf{Q} = \mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H}$, and the value of the unknown parameter vector $\mathbf{x}$. Several computer simulations were implemented to test the effect of these conditions on performance. The simulations were also used

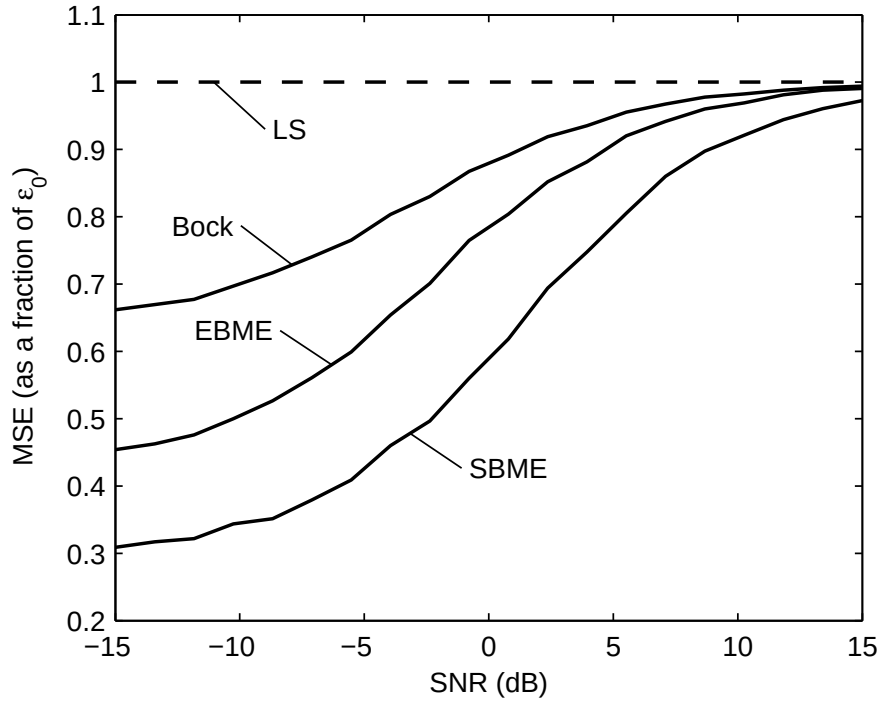

Figure 4.1: MSE vs. SNR for a typical operating condition: effective dimension 5.1, $m = n = 15$.

to compare the BMEs with commonly used estimators, including the LS estimator, Bock's estimator (2.21), and Tikhonov regularization (2.26).

4.5.1 Comparison with the LS Approach

The theorems of Sections 4.2 and 4.3 ensure that the BMEs achieve lower MSE than the LS estimator, but do not guarantee that this improvement is substantial. To measure this performance gain, we first chose a typical scenario, in which the number of parameters m and the number of measurements n were both 15. The system matrix $\mathbf{H}$ was chosen as $\mathrm{diag}(1, 1, 1, .5, .3, .2, .2, .2, .2, .1, .1, .1, .1, .05, .05)$, and the parameter vector was chosen randomly as a zero-mean, unit-variance i.i.d. Gaussian vector. The noise was i.i.d. with variance σ^2 chosen to achieve the desired SNR. Estimates of the MSE were calculated for a range of SNR values by generating 3000 random realizations of noise and parameter vectors per SNR value. The results are plotted in Figure 4.1.

It is evident from this figure that substantial improvement in MSE can be achieved by using BMEs in place of the LS estimator: in some cases the MSE of the LS estimator is more than five times the MSE of the BMEs. The performance gain is particularly noticeable at low and moderate SNR. At infinite SNR, the LS estimator is known to be optimal in terms of MSE [1],

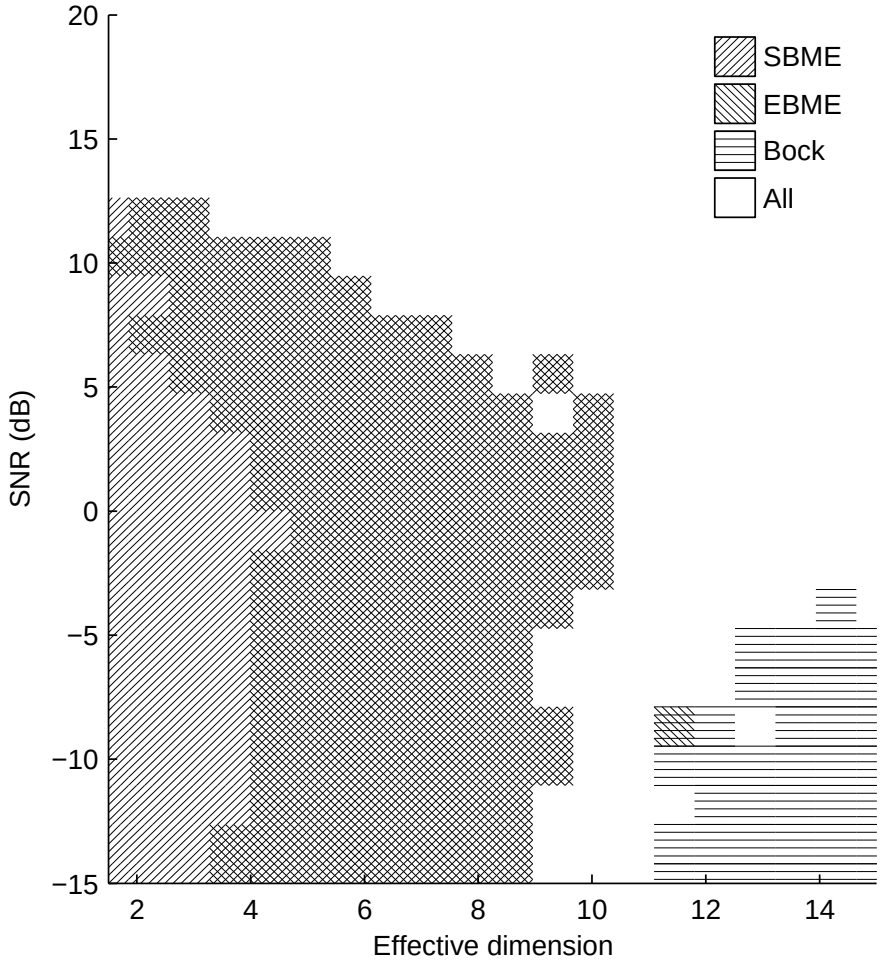

Figure 4.2: Estimator(s) achieving lowest MSE, among the five estimators tested ($m = n = 10$)

and all other estimators converge to the value of the LS estimate; as a result, performance gain is smaller at high SNR. However, significant improvement on the order of 10–20% may still be achieved even at SNR values of 5–10 dB.

4.5.2 Comparison with Bock's Estimator

Different settings call for the use of different estimators, as no single estimator is optimal under all operating conditions. To demonstrate this, estimator MSE was measured for various SNRs and effective dimensions. The parameter vector for this simulation was randomly chosen from an i.i.d. normal distribution. The number of measurements and the number of parameters were both equal to 10. The system matrix $\mathbf{H}$ was equal to $\mathbf{I}$, and the noise covariance matrix $\mathbf{C_w}$ was diagonal, with diagonal elements $c_1 = 1$, $c_2 = 0.5$, and $c_3 = \cdots = c_{10} = t$, where t was chosen to obtain the desired effective dimension. The simulation was repeated for 20 different

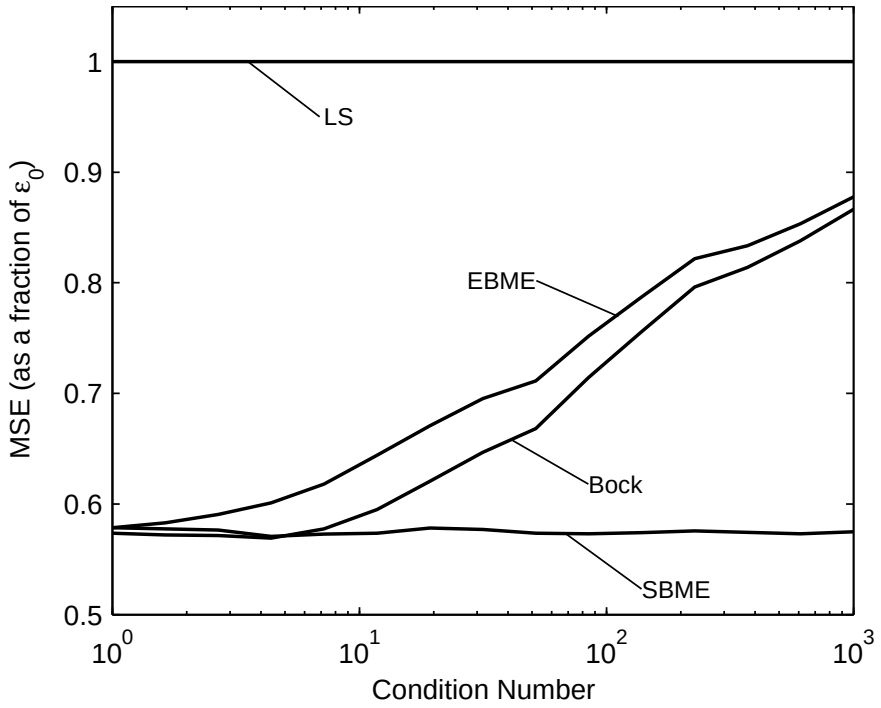

Figure 4.3: Estimator MSE vs. condition number; $m = n = 10$, SNR 0 dB

effective dimensions in the range 1.5 to 10, and for 20 different SNR values in the range –15 dB to 15 dB. For each operating condition, the MSE of each estimator was calculated. Figure 4.2 displays the estimator achieving lowest MSE among those tested; when two or more estimators achieve MSE within 5% of the lowest value, this is indicated by their combined pattern. The LS technique is outperformed by all others in this simulation, so it is not displayed in the figure.

Figure 4.2 demonstrates that the BMEs significantly outperform Bock's estimator under a very wide range of operating conditions. It is notable that the BMEs continue to outperform Bock's estimator and the LS estimator at effective dimensions of 2–4; the dominance results of Sections 4.2 and 4.3 only apply to effective dimensions above 4. This demonstrates the fact that, although sufficient conditions for dominance are provided by the aforementioned theorems, it may be possible to prove dominance for stronger conditions as well.

As noted previously, at high SNR, all estimators converge to the LS estimator and their performance is similar. However, for most operating conditions, the SBME is the optimal estimator among those tested; its improvement is significant particularly at low effective dimensions. This is indicative of a more subtle limitation of the EBME, namely, its sensitivity to the condition number[1] of $\mathbf{Q}$; this limitation is also present in Bock's estimator. The EBME makes

[1]The condition number of a matrix is defined as the ratio between its largest and smallest eigenvalues.

use of an ellipsoid of the form $\{\mathbf{x} : \mathbf{x}^*\mathbf{Q}\mathbf{x} \leq L^2\}$, which becomes eccentric when the condition number of $\mathbf{Q}$ is large. As a result, a slight increase in the measurements along a narrow axis of the ellipse greatly increases the radius in the wide axes. This causes the corresponding parameters to be estimated with negligible shrinkage, thus reducing the improvement over the LS estimator. Bock's estimator suffers from the same effect, since its shrinkage factor is also a function of $\hat{\mathbf{x}}_{\mathrm{LS}}^*\mathbf{Q}\hat{\mathbf{x}}_{\mathrm{LS}}$. Only the SBMEs, whose parameter set estimates are based on the value $\hat{\mathbf{x}}_{\mathrm{LS}}^*\hat{\mathbf{x}}_{\mathrm{LS}}$, continue to perform well for high condition numbers.

This effect is demonstrated in Fig. 4.3. Here, the settings are identical to those of Fig. 4.2, except that the SNR is constant and equals 0 dB, and the noise covariance matrix contains fourteen eigenvalues equal to 1, and an additional small eigenvalue whose value is modified to control the condition number. Thus, the condition number is changed with little influence on the effective dimension. As expected, the performance of the EBME and Bock's estimator deteriorates when high condition numbers are used, while the SBME is hardly affected by the change. Fortunately, both the effective dimension and the condition number depend only on the system matrix $\mathbf{H}$ and the noise covariance $\mathbf{C_w}$, which are known in advance; therefore, an informed choice may be made for any given estimation problem.

4.5.3 Comparison with Tikhonov Regularization

Several methods for obtaining nonlinear estimators derived from Tikhonov's technique were presented in Section 2.4. These include the estimators $\hat{\mathbf{x}}_{\mathrm{T}}^{(1)}$ of (2.29) and $\hat{\mathbf{x}}_{\mathrm{T}}^{(2)}$ of (2.30). The derivation of these "blind Tikhonov estimators" is similar to the empirical Bayes justification of the James-Stein estimator; thus, one could hope that they provide improvement over the LS estimator.

Unfortunately, the blind Tikhonov estimators do not dominate the LS estimator; as with the original Tikhonov regularization, they perform poorly for high SNR values. To illustrate this, we performed a simulation in which the MSE of various estimators were compared to those of $\hat{\mathbf{x}}_{\mathrm{T}}^{(1)}$ and $\hat{\mathbf{x}}_{\mathrm{T}}^{(2)}$. A setting identical to that of Figure 4.2 was used. To demonstrate the fact that the Tikhonov regularization does not dominate the LS estimator, it is sufficient to find a particular value of $\mathbf{x}$ for which the MSE is larger than ϵ_0. In Figure 4.4, we plot the MSE of the Tikhonov regularization for various values of $\mathbf{x}$ chosen in the direction of the maximum noise eigenvector, i.e., $\mathbf{x}$ is proportional to the maximum eigenvalue of $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}^{-1}$. For comparison, the MSE of the LS estimator and Bock's estimator, is also plotted. It is evident from this figure that the Tikhonov regularization is inadequate at high SNR, as it performs worse than

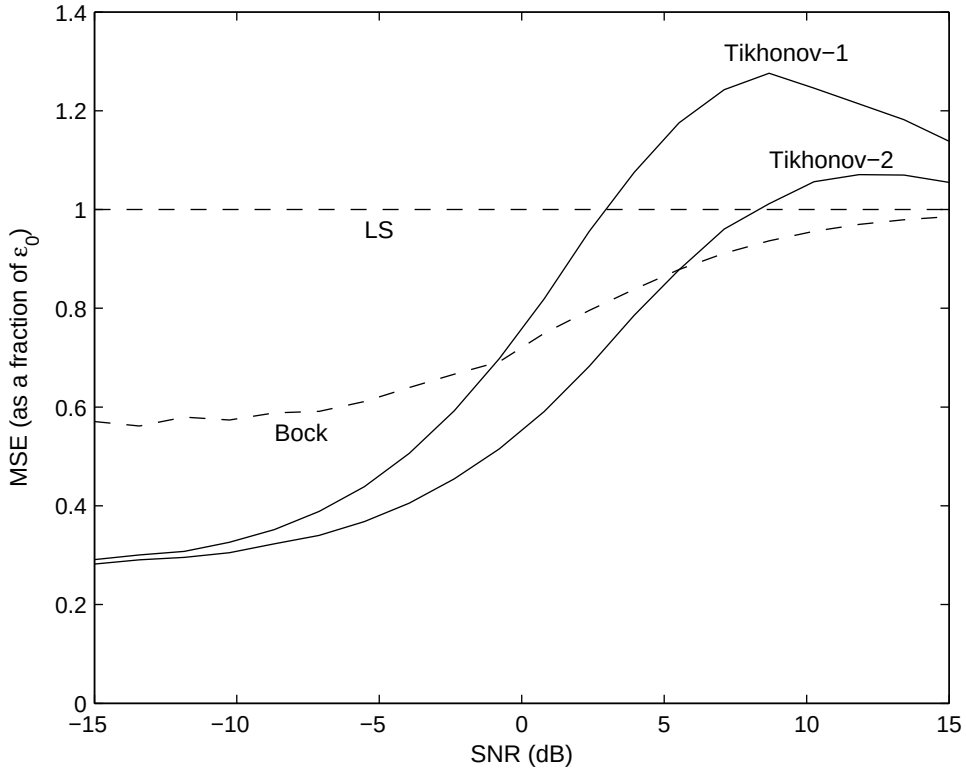

Figure 4.4: Tikhonov regularization does not dominate the LS estimator

the LS estimator. Both Tikhonov estimators converge to the LS approach at infinite SNR, but consistently obtain higher MSE than the LS estimator for SNR values about 0 dB. This makes them unattractive candidates for replacing the LS technique, as one would not want to improve low-SNR performance at the expense of poor performance for high SNR.

4.6 Discussion

The blind minimax approach is a general technique for using minimax estimators in situations for which no parameter set is known. We considered an application of this concept to the Gaussian linear regression model. Two novel estimators were proposed: an estimator based on a spherical parameter set, and one based on an ellipsoidal parameter set. In Sections 4.2 and 4.3, these estimators were shown to dominate the LS estimator. Thus, in *any* application which makes use of a LS estimator, the MSE performance can be improved by using a BME instead. Furthermore, in Section 4.4, we demonstrated that Stein's estimator, as well as its positive part modification, can be derived and generalized using the blind minimax framework.

It can readily be shown that the dominance condition of the SBME (Theorem 4.1) is weaker

than the dominance condition of the EBME (Theorem 4.5), i.e., the conditions for SBME dominance hold whenever the conditions for EBME dominance hold. The dominance condition of Bock's estimator (Subsection 2.3.3) is still weaker[2]. This would seem to indicate that Bock's estimator is superior to the proposed estimators. Yet the results of Section 4.5 demonstrate that the opposite is true: the BMEs usually outperform Bock's estimator — even in cases where their performance is not guaranteed by Theorems 4.1 and 4.5. Thus, while dominance theorems are useful in providing sufficient conditions for improving on the LS estimator, they are ill-suited for comparing LS-dominating estimators. This conclusion is significant since estimators are sometimes chosen by maximizing the range of conditions for which dominance is guaranteed. It seems that other analytical tools are required for comparing LS-dominating estimators. For example, it may be possible to prove that BMEs dominate Bock's estimator, for some problem settings.

The choice between the different BMEs is application-dependent. As explained in Section 4.5, the SBME outperforms the other estimators tested in most SNR ranges, and is particularly useful when dealing with system matrices have a high condition number. These results may be used to select an estimator depending on the problem setting, since the effective dimension and condition number may be calculated in advance from the problem setting.

A more fundamental difference is that the SBME and Bock's estimator are shrinkage estimators, while the EBME is not. Thus, in applications where the only goal is minimization of the MSE, the SBME may be preferred for its robustness and simplicity. For example, the SBME is an excellent estimator of system parameters, such as autoregression (AR) coefficients. However, in certain applications, MSE minimization is only a nominal goal which approximates some other error criterion. In some of these cases, a shrinkage estimator has no impact on the actual objective. For example, if the vector $\mathbf{x}$ is an image which is to be reconstructed, its subjective quality is not affected by multiplying the entire estimate by a scalar. Likewise, in a binary receiver, the sign of $\mathbf{x}$ must be determined, but the sign does not change when the estimate is shrunk. In such applications, the SBME (and Bock's estimator) have no effect on the final result, whereas the EBME can be used to improve performance. Other difficulties with shrinkage estimators were discussed in Subsection 2.3.3.

In this chapter, we have explored the idea of blind minimax estimation, whereby one uses

[2] A simple change to the SBME (adding -2 to the numerator) changes its dominance condition to that of Bock's estimator, without significantly affecting its performance. However, we have been unable to derive this modification using the blind minimax approach, and thus prefer the simpler form of the SBME used in the paper.

linear minimax estimators whose parameter set is itself estimated from measurements. This simple concept was examined in the setting of a linear system of measurements with colored Gaussian noise, where we have shown that the BMEs dominate the LS estimator. Hence, in any such problem, the proposed estimators can be used in place of the LS estimator, with a guaranteed performance gain. Apart from being useful in and of themselves, the proposed estimators support the underlying concept of blind minimax estimation. This concept can be applied to many other estimation problems, such as estimation with uncertain system matrices, estimation with non-Gaussian noise, and sequential estimation. Use of the blind minimax approach in such problems remains a topic for further study.

Stein's discovery of LS-dominating estimators, half a century ago, shocked the statistics community, and LS-dominating estimators are still rarely used in practice. It is our hope that the blind minimax concept will provide additional support for such estimators, both by supplying an intuitive understanding of Stein's phenomenon, and by providing a wide class of powerful new estimators.

Chapter 5

Maximum Set Estimation

The minimax approach, presented in Chapter 3, assumes that bounds on the parameter vector are known. These bounds have considerable impact on the obtained estimator; for example, if the parameter set is too small, then the estimator may receive values for which it was not designed, and the risk will be larger than expected. In Chapter 4, we proposed a method for estimating the parameter set from measurements, resulting in a class of nonlinear techniques referred to as the blind minimax approach. This approach is intended for situations in which nonlinear estimators are acceptable, and when no additional information about the parameters is available.

In this chapter, we limit discussion to linear estimators. However, instead of the minimax approach, which requires the unknown parameter to belong to a particular set, we assume that a maximum estimation error is required of the system. We seek to design an estimator which guarantees the required error for the widest range of conditions possible, an approach which follows the philosophy of information-gap decision theory [23,24]. The result is called a maximum set estimator [25,26].

The maximum set estimation strategy can be applied in several ways, depending on the uncertain system property. We present two different approaches as a demonstration of the power of this approach. In Section 5.1, we discuss the case in which the parameter set is uncertain; in this case, we seek the estimator which maximizes the parameter set for which error requirements are maintained. In Section 5.2, we consider the estimation problem when the noise covariance is known up to a constant, i.e., $E\{\mathbf{w}\mathbf{w}^*\} = \sigma^2\mathbf{C}_\mathbf{w}$, where σ^2 is unknown. In this case, we assume that $\mathbf{x}$ lies in a known parameter set $\mathcal{S}$, and find the estimator which guarantees a required estimation error for as large a range of noise levels σ^2 as possible.

5.1 Maximum Parameter Set Estimation

We begin our presentation of the maximum set approach in Subsection 5.1.1 with a simple example demonstrating the use of this technique. This example explores a special case of maximum parameter set estimation, which is later generalized and formalized in Subsection 5.1.2. Some consequences of this formalism are then derived in Subsection 5.1.3; these are later used to produce closed forms for many new types of estimators.

5.1.1 A Useful Special Case

Consider a minimax MSE estimator (Subsection 3.2.1), and suppose the parameter set $\mathcal{S}_L$ is an ellipsoid given by

$$\mathcal{S}_L = \{\mathbf{x} : \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\} \tag{5.1}$$

for some positive definite matrix $\mathbf{T}$. The matrix $\mathbf{T}$ defines relations between the uncertainty levels of the different parameters, while the value L indicates the overall size of the parameter set. For example, if all parameters have identical physical units and there is no reason to expect higher uncertainty in some of them, one may choose $\mathbf{T} = \mathbf{I}$. The matrix $\mathbf{T}$ can thus be chosen to represent knowledge about the system, even if no exact parameter bound is known. However, a suitable value of L is often difficult to determine. Even if a small amount of information about $\mathbf{x}$ is available, such as several past measurements, then these usually characterize typical values of $\mathbf{x}$, while L is meant to characterize the extreme or rare values of $\mathbf{x}$.

In some cases, it is our interest to find an estimator achieving "satisfactory" performance for as large a parameter set as possible. To this end, we assume that a maximum error ϵ_m is known; this is the maximum error allowed for satisfactory performance of the system. We aim to design an MPS estimator, for which satisfactory performance is achieved for as large a value of L as possible.

Formally, the parameter robustness $\hat{L}$ of an estimator $\hat{\mathbf{x}}$ is defined as the largest L for which performance is satisfactory,

$$\hat{L}(\hat{\mathbf{x}}) = \sup\left\{L :\ E\left\{\|\mathbf{x}-\hat{\mathbf{x}}\|^2\right\} \leq \epsilon_m,\ \forall \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\right\}. \tag{5.2}$$

A maximum parameter set (MPS) estimator $\hat{\mathbf{x}}_{\mathrm{PS}}$ is an estimator achieving maximal parameter robustness, i.e.,

$$\hat{L}(\hat{\mathbf{x}}_{\mathrm{PS}}) \geq \hat{L}(\hat{\mathbf{x}}), \quad \text{for any linear estimator } \hat{\mathbf{x}}. \tag{5.3}$$

Suppose we wish to find an MPS estimator for maximum error ϵ_m equal to ϵ_0 of (2.5), which is the MSE of the LS estimator. The LS estimator achieves this error regardless of the value of $\mathbf{x}$; thus, its parameter robustness is infinite when $\epsilon_m \geq \epsilon_0$. This implies that requiring a maximum error of ϵ_0 (or greater) yields the LS estimator as an MPS estimator. More interesting is the case $\epsilon_m < \epsilon_0$, for which the LS approach no longer achieves the required error, regardless of the value of $\mathbf{x}$. An MPS estimator $\hat{\mathbf{x}}$ for a given error level $\epsilon_m < \epsilon_0$ has finite robustness, but within the parameter set $\mathcal{S}_{\hat{L}(\hat{\mathbf{x}})}$, its worst-case error does not exceed ϵ_m. Thus, an MPS estimator outperforms the LS estimator for any $\mathbf{x} \in \mathcal{S}_{\hat{L}(\hat{\mathbf{x}})}$.

In the remainder of this section, we show that an MPS estimator can be found by solving a quasiconvex optimization problem. An optimization problem is quasiconvex if its constraints are convex, and its objective function is quasiconvex; the function $f(\mathbf{z})$ is quasiconvex if the sublevel sets $\{\mathbf{z} : \ f(\mathbf{z}) \leq \alpha\}$ are convex for all α. Quasiconvex problems can be efficiently solved, for example, using bisection [37]. In addition, as we shall see in Subsection 5.1.4, in many special cases a closed form for an MPS estimator can be obtained, by exploring its relation to the minimax MSE estimator.

Theorem 5.1. *A linear maximum parameter set (MPS) estimator* $\hat{\mathbf{x}}_{\mathrm{PS}} = \mathbf{G}\mathbf{y}$ *satisfying* (5.3) *can be found by solving the following quasiconvex optimization problem:*

$$\min_{\mathbf{G},\lambda,y} y/\lambda \tag{5.4}$$
$$\text{s.t.} \begin{cases} \begin{bmatrix} y+\epsilon_m & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix} \succeq 0 \\ \begin{bmatrix} \lambda\mathbf{I} & \mathbf{T}^{-1/2}(\mathbf{I}-\mathbf{G}\mathbf{H})^* \\ (\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{T}^{-1/2} & \mathbf{I} \end{bmatrix} \succeq 0 \end{cases}$$

where $\mathbf{g}$ *is the vector obtained by stacking the columns of* $\mathbf{G}\mathbf{C}_{\mathbf{w}}^{1/2}$*. The parameter robustness* $\hat{L}$ *of this estimator is given by* $\sqrt{-y/\lambda}$ *for the optimal values of* y *and* λ*.*

Proof. We seek an estimator $\hat{\mathbf{x}}_{\mathrm{PS}}$ satisfying (5.3) with $\hat{L}(\hat{\mathbf{x}})$ defined by (5.2), which is equivalent to solving the optimization problem

$$\max_{\mathbf{G},L^2} L^2 \quad \text{s.t.} \ \ \epsilon_m \geq \max_{\|\mathbf{x}\|_{\mathbf{T}}^2 \leq L^2} E\left\{\|\mathbf{x}-\hat{\mathbf{x}}\|^2\right\}, \tag{5.5}$$

where $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$. Using (2.6)–(2.8), we find that for any given estimator $\hat{\mathbf{x}}$,

$$\max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} E\left\{\|\mathbf{x}-\hat{\mathbf{x}}\|^2\right\} = \mathrm{Tr}(\mathbf{G}\mathbf{C}_{\mathbf{w}}\mathbf{G}^*) + \max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} \|(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{x}\|^2. \tag{5.6}$$

However,

$$\max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} \|(\mathbf{I} - \mathbf{GH})\mathbf{x}\|^2 = \max_{\mathbf{z}^*\mathbf{z} \leq L^2} \|\mathbf{Pz}\|^2 = \lambda_{\max} L^2, \tag{5.7}$$

where $\mathbf{P} = (\mathbf{I} - \mathbf{GH})\mathbf{T}^{-1/2}$ and $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{P}^*\mathbf{P}$. We can express $\lambda_{\max}$ as the solution to the semidefinite problem

$$\min_{\lambda} \lambda \quad \text{s.t. } \mathbf{P}^*\mathbf{P} \preceq \lambda \mathbf{I}. \tag{5.8}$$

Consider the problem

$$\max_{\mathbf{G}, \lambda, L^2} L^2 \tag{5.9}$$

$$\text{s.t.} \begin{cases} \operatorname{Tr}(\mathbf{GC_w G}^*) + \lambda L^2 \leq \epsilon_m & \text{(a)} \\ \mathbf{P}^*\mathbf{P} \preceq \lambda \mathbf{I}. & \text{(b)} \end{cases}$$

We claim that the optimal solution to this problem always has $\lambda = \lambda_{\max}$. Suppose this were not the case, and $\lambda > \lambda_{\max}$ for the optimal solution. Then, λ can be decreased while still maintaining (5.9b). As a result, (5.9a) is no longer tight, so that L^2 can be increased, contradicting the assumption that λ was the optimal solution. Thus, the optimal solution for (5.9) always has $\lambda = \lambda_{\max}$, and therefore, by (5.6) and (5.7), the optimal solution of (5.9) satisfies

$$\max_{\|\mathbf{x}\|_{\mathbf{T}}^2 \leq L^2} E\{\|\mathbf{x} - \hat{\mathbf{x}}\|^2\} = \operatorname{Tr}(\mathbf{GC_w G}^*) + \lambda L^2, \tag{5.10}$$

so that (5.9) and (5.5) are equivalent.

Let $\mathbf{g}$ be the vector obtained by stacking the columns of $\mathbf{GC}_{\mathbf{w}}^{1/2}$. Using Schur's Lemma [38, p. 472], it is shown in [6] that (5.9a) and (5.9b) are equivalent to the following matrix inequalities:

$$\begin{bmatrix} \epsilon_m - \lambda L^2 & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix} \succeq 0, \tag{5.11a}$$

$$\begin{bmatrix} \lambda \mathbf{I} & \mathbf{P}^* \\ \mathbf{P} & \mathbf{I} \end{bmatrix} \succeq 0. \tag{5.11b}$$

Defining $r = -L^2$, (5.11a) becomes

$$r \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \succeq - \begin{bmatrix} \epsilon_m & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix}. \tag{5.12}$$

We now add a scalar optimization parameter y and note that the optimization problem is equivalent to

$$\min_{\mathbf{G},\lambda,r,y} r \tag{5.13}$$
$$\text{s.t.} \begin{cases} r\lambda \geq y \\ \begin{bmatrix} y & 0 \\ 0 & 0 \end{bmatrix} \succeq -\begin{bmatrix} \epsilon_m & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix} \\ \begin{bmatrix} \lambda\mathbf{I} & \mathbf{P}^* \\ \mathbf{P} & \mathbf{I} \end{bmatrix} \succeq 0. \end{cases}$$

It is evident that the optimal solution to this problem satisfies $r = y/\lambda$; substituting this into the above problem yields the required optimization problem (5.4). The objective function of (5.4) is quasiconvex, and all its constraints are convex, so that this is a quasiconvex optimization problem [37]. □

In the next subsection, we generalize the discussion to MPS estimators which optimize various error functions over different parameter sets. We also demonstrate a relation between MPS estimation and minimax estimation, which provides further insight into the idea of MPS estimation and yields an alternative method for finding MPS estimators. In particular, this leads to a closed form for an MPS estimator when the weighting matrix $\mathbf{T}$ commutes with $\mathbf{H}^*\mathbf{C}_\mathbf{w}^{-1}\mathbf{H}$, which occurs, for example, when $\mathbf{T} = \mathbf{I}$.

5.1.2 General Form of MPS Estimators

The example presented in the previous subsection is a special case of an MPS estimator, which can be generalized to include different error functions and parameter sets. We now provide definitions which construct the general form of these estimators.

Definition 5.1. The *system properties* required for the design of a maximum parameter set (MPS) estimator are the following:

1. A *risk function* $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$ which quantifies the degree to which an estimator $\hat{\mathbf{x}}$ misrepresents the specific value $\mathbf{x}$ (see Subsection 2.1.2). This function must be continuous.

2. A *maximum error* ϵ_m which defines the error value required for successful operation of the system. This is a deterministic real number which must be known to the designer. An MPS estimator seeks to maximize the range of values of $\mathbf{x}$ for which the maximum error is guaranteed.

3. A *class of parameter sets* $\{\mathcal{S}_L \subseteq \mathbb{C}^m\}_{L\geq 0}$ which define feasible values of $\mathbf{x}$ under varying parameter set bounds L. These sets obey three basic properties:

   (a) As L increases, more values of $\mathbf{x}$ become feasible, so that the sets $\mathcal{S}_L$ are nested:

   $$L_1 < L_2 \iff \mathcal{S}_{L_1} \subset \mathcal{S}_{L_2}. \tag{5.14}$$

   (b) The parameter sets are linearly expanding: For all $L_1, L_2 > 0$,

   $$\mathcal{S}_{L_1} = \frac{L_1}{L_2}\mathcal{S}_{L_2} = \left\{\mathbf{x} : \ \frac{L_2}{L_1}\mathbf{x} \in \mathcal{S}_{L_2}\right\}. \tag{5.15}$$

   This requirement implies that the parameter sets are centered on the origin, an assumption which we adopt without loss of generality.

   (c) The sets $\mathcal{S}_L$ are compact (i.e., closed and bounded). This requirement ensures the existence of a maximum error for every parameter set.

Most common bounds fulfill the requirements for the class of parameter sets above. The weighted norm $\mathcal{S}_L = \{\mathbf{x} : \ \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$ used in Subsection 5.1.1 is one example. Another example is the box bound, $\mathcal{S}_L = \{\mathbf{x} : \ |x_i| \leq Lb_i, \ \ \forall i\}$, where $b_i > 0$ are constants.

Definition 5.2. The *parameter robustness* $\hat{L}(\hat{\mathbf{x}})$ of an estimator $\hat{\mathbf{x}}$ (for particular system properties) is the largest parameter set bound L for which the maximum error is guaranteed, namely,

$$\hat{L}(\hat{\mathbf{x}}) = \sup\{L : \ \forall \mathbf{x} \in \mathcal{S}_L, \ \epsilon(\hat{\mathbf{x}}, \mathbf{x}) \leq \epsilon_m\}. \tag{5.16}$$

Definition 5.3. A *maximum parameter set (MPS) estimator* (among estimators of a given class $\mathcal{E}$) is an estimator $\hat{\mathbf{x}}_{\text{PS}}$ such that, for any $\hat{\mathbf{x}} \in \mathcal{E}$,

$$\hat{L}(\hat{\mathbf{x}}_{\text{PS}}) \geq \hat{L}(\hat{\mathbf{x}}). \tag{5.17}$$

Note that Definition 5.3 does not imply the unique existence of MPS estimators. In fact, for some choices of ϵ_m, many estimators with infinite robustness exist. However, we shall see that in many cases of interest, the MPS estimator exists and is unique.

The estimator presented in Subsection 5.1.1 is a special case of an MPS estimator, which makes use of a particular choice of the error function and of the class of parameter sets. Specifically, the MSE (2.2) is used as the risk function, ellipsoids (5.1) of increasing size and constant axis ratios are used as the nested parameter sets, and the estimator is restricted to being linear.

5.1.3 Relation to Minimax Estimation

An interesting and useful relation exists between the MPS estimator $\hat{\mathbf{x}}_{\mathrm{PS}}$ and the minimax estimator of Chapter 3. Put simply, the MPS estimator maximizes the parameter robustness L within a range defined by the known value of ϵ, while the minimax estimator minimizes the worst-case error ϵ within a range defined by the known value of L.

To formalize this relation, let $\{\mathcal{S}_L\}_{L\geq 0}$ be a class of parameter sets, and define the worst-case error function

$$e(L) = \max_{\mathbf{x}\in\mathcal{S}_L} \epsilon(\hat{\mathbf{x}}_{\mathrm{M}}(L), \mathbf{x}), \tag{5.18}$$

where $\hat{\mathbf{x}}_{\mathrm{M}}(L)$ is a minimax estimator for the parameter set $\mathcal{S}_L$. Clearly, $e(L)$ is non-decreasing, since enlarging the parameter set cannot decrease the worst-case error. This trade-off between parameter set size and worst-case error is applicable to MPS estimators as well. Indeed, if $e(L)$ is strictly increasing in L, there exists a one-to-one correspondence between the parameter set bound L and the worst-case error $e(L)$. In this case it is intuitive to expect a one-to-one correspondence between minimax and MPS estimators. Thus, we have the following theorem.

Theorem 5.2. *Consider a risk function $\epsilon(\hat{\mathbf{x}}, \mathbf{x})$ and a class of parameter sets $\{\mathcal{S}_L\}_{L\geq 0}$, as defined in Definition 5.1. Assume the worst-case error $e(L)$ of* (5.18) *is strictly increasing in L. For any L, an estimator $\hat{\mathbf{x}}$ is an MPS estimator with worst-case error $\epsilon_m = e(L)$ if, and only if, it is a minimax estimator over the parameter set $\mathcal{S}_L$.*

Proof. Suppose first that $\hat{\mathbf{x}}_{\mathrm{PS}}$ is an MPS estimator with worst-case error $\epsilon_m = e(L_0)$, for a given L_0. Let $L_1 = \hat{L}(\hat{\mathbf{x}}_{\mathrm{PS}})$ and notice that $L_1 \geq L_0$ (we shall show presently that $L_1 = L_0$). Assume by contradiction that $\hat{\mathbf{x}}_{\mathrm{PS}}$ is not a minimax estimator over $\mathcal{S}_{L_1}$. Then, by Definition 3.1, there exists an estimator $\hat{\mathbf{x}}_{\mathrm{M}}$ such that

$$\max_{\mathbf{x}\in\mathcal{S}_{L_1}} \epsilon(\hat{\mathbf{x}}_{\mathrm{M}}, \mathbf{x}) < \max_{\mathbf{x}\in\mathcal{S}_{L_1}} \epsilon(\hat{\mathbf{x}}_{\mathrm{PS}}, \mathbf{x}) \leq \epsilon_m. \tag{5.19}$$

By Definition 5.1, the parameter sets expand linearly, so that for sufficiently small $\alpha > 1$, each of the values in the parameter set $\mathcal{S}_{\alpha L_1}$ is arbitrarily close to some value in $\mathcal{S}_{L_1}$. Furthermore, by Definition 5.1, ϵ is continuous, so that sufficiently small changes in $\mathbf{x}$ yield arbitrarily small changes in $\epsilon(\hat{\mathbf{x}}_{\mathrm{PS}}, \mathbf{x})$. Hence, there exists a sufficiently small $\alpha > 1$ for which

$$\max_{\mathbf{x}\in\mathcal{S}_{\alpha L_1}} \epsilon(\hat{\mathbf{x}}_{\mathrm{M}}, \mathbf{x}) \leq \epsilon_m. \tag{5.20}$$

Thus the parameter robustness of $\hat{\mathbf{x}}_{\mathrm{M}}$ is at least $\alpha L_1 > L_1 = \hat{L}(\hat{\mathbf{x}}_{\mathrm{PS}})$, which contradicts the fact that $\hat{\mathbf{x}}_{\mathrm{PS}}$ is an MPS estimator. Hence, $\hat{\mathbf{x}}_{\mathrm{PS}}$ is a minimax estimator over $\mathcal{S}_{L_1}$, and its worst-case

error is $e(L_1)$. However, from (5.19), the worst-case error of $\hat{\mathbf{x}}_{\mathrm{PS}}$ is $e(L_0)$. Since $e(L)$ is strictly increasing, this implies $L_0 = L_1$. We conclude that $\hat{\mathbf{x}}_{\mathrm{PS}}$ is minimax over $\mathcal{S}_{L_0}$.

We now prove that a minimax estimator is an MPS estimator. For any L_0, let $\hat{\mathbf{x}}_{\mathrm{M}}(L_0)$ be a minimax estimator for the parameter set $\mathcal{S}_{L_0}$. Assume by contradiction that $\hat{\mathbf{x}}_{\mathrm{M}}(L_0)$ is not an MPS estimator for the maximum error $\epsilon_m = e(L_0)$. Then, there exists an $\hat{\mathbf{x}}_{\mathrm{PS}}$ with robustness $L_1 = \hat{L}(\hat{\mathbf{x}}_{\mathrm{PS}})$ such that $L_1 > \hat{L}(\hat{\mathbf{x}}_{\mathrm{M}}(L_0)) \geq L_0$. Therefore,

$$\max_{\mathbf{x}\in\mathcal{S}_{L_1}} \epsilon(\hat{\mathbf{x}}_{\mathrm{PS}}, \mathbf{x}) \leq \epsilon_m = e(L_0) < e(L_1). \tag{5.21}$$

However, by (5.18),

$$\max_{\mathbf{x}\in\mathcal{S}_{L_1}} \epsilon(\hat{\mathbf{x}}_{\mathrm{M}}(L_1), \mathbf{x}) = e(L_1). \tag{5.22}$$

Hence $\hat{\mathbf{x}}_{\mathrm{PS}}$ achieves a lower worst-case error over $\mathcal{S}_{L_1}$ than the minimax estimator of $\mathcal{S}_{L_1}$, which is a contradiction. We conclude that $\hat{\mathbf{x}}_{\mathrm{M}}(L_0)$ must be an MPS estimator. □

We have shown that when the worst-case error function $e(L)$ is strictly increasing in L, there is a one-to-one correspondence between minimax and MPS estimators. As we shall see in the following sections, $e(L)$ is indeed strictly increasing for many important cases, such as the MSE error function. However, this is not always the case. For instance, if the error function decreases with $\|\mathbf{x}\|$, then increasing the parameter set will not increase the worst-case error.

Theorem 5.2 can be used to efficiently find an MPS estimator using known minimax estimators. This is done using bisection on the worst-case error function $e(L)$: Since the function is strictly monotonic, a value of L yielding $e(L)$ which equals ϵ_m to any desired accuracy can efficiently be found. From Theorem 5.2, the minimax estimator $\hat{\mathbf{x}}_{\mathrm{M}}(L)$ equals the desired MPS estimator.

Similarities notwithstanding, minimax and MPS estimators differ qualitatively in the type of information on which their design is based. A minimax estimator requires that a bound on the uncertain parameter $\mathbf{x}$ be stated, while an MPS estimator requires knowledge of the maximum error under which the system still operates correctly. Thus, proper choice of an estimator should depend on the nature of the information available to the designer.

In the following subsections, we use Theorem 5.2 to develop MPS estimators for two cases of interest, the MSE estimator and the regret estimator.

5.1.4 Linear MSE Estimators

Consider the MPS estimation problem when the error function of interest is the MSE, and the estimator is restricted to being linear. In Theorem 5.3, we show that minimax and MPS criteria

for optimality are equivalent in these circumstances. This allows us to find an MPS estimator whenever an algorithm for finding a minimax estimator is known. In particular, Theorem 5.4 derives a closed form for the estimator when the uncertainty sets are spherical.

Theorem 5.3. *Suppose that the risk function of interest is the MSE* (2.2)*, let $\mathcal{E}$ be the class of linear estimators, and let $\{\mathcal{S}_L\}_{L\geq 0}$ be a class of parameter sets, as defined in Definition 5.1. An estimator $\hat{\mathbf{x}} \in \mathcal{E}$ is a linear minimax estimator over $\mathcal{S}_L$ if, and only if, it is a linear MPS estimator with maximum error ϵ_m equal to the worst-case error $e(L)$ of* (5.18)*.*

Proof. By Theorem 5.2, it is sufficient to show that $e(L)$ is strictly increasing. Let $\hat{\mathbf{x}}_{\mathrm{M}}(L) = \mathbf{G}_L\mathbf{y}$ be a linear minimax MSE estimator over the set $\mathcal{S}_L$. From (2.6)–(2.8), we have

$$e(L) = \mathrm{Tr}(\mathbf{G}_L\mathbf{C}_{\mathbf{w}}\mathbf{G}_L^*) + \max_{\mathbf{x}\in\mathcal{S}_L} \|(\mathbf{I} - \mathbf{G}_L\mathbf{H})\mathbf{x}\|^2. \tag{5.23}$$

By Theorem 3.8, the minimax MSE estimator achieves lower MSE than the LS estimator for any $\mathbf{x} \in \mathcal{S}_L$. Since the LS estimator achieves the lowest possible MSE among all unbiased estimators (Subsection 2.2.2), it follows that the minimax MSE estimator must be biased, i.e., $\mathbf{G}_L\mathbf{H} \neq \mathbf{I}$.

We now show that $\max_{\mathbf{x}\in\mathcal{S}_L} \|(\mathbf{I} - \mathbf{G}_L\mathbf{H})\mathbf{x}\|^2$ is obtained only on the boundary of $\mathcal{S}_L$. Let $\mathbf{x}_0 \in \mathcal{S}_L$ be a point which is *not* on the boundary. Then, there exists a sufficiently small sphere S, centered on $\mathbf{x}_0$, such that $S \subset \mathcal{S}_L$. In particular, S necessarily includes a point $(1+\delta)\mathbf{x}_0$ (for a sufficiently small $\delta > 0$). Since $\mathbf{G}_L\mathbf{H} \neq \mathbf{I}$, we have

$$\|(\mathbf{I} - \mathbf{G}_L\mathbf{H})(1+\delta)\mathbf{x}_0\|^2 > \|(\mathbf{I} - \mathbf{G}_L\mathbf{H})\mathbf{x}_0\|^2. \tag{5.24}$$

Thus, $\max_{\mathbf{x}\in\mathcal{S}_L} \|(\mathbf{I} - \mathbf{G}_L\mathbf{H})\mathbf{x}\|^2$ is not obtained at $\mathbf{x}_0$; rather, the maximum is obtained only on the boundary of $\mathcal{S}_L$. Therefore, by shrinking the parameter set, the worst-case error must decrease: for any $L < M$,

$$\max_{\mathbf{x}\in\mathcal{S}_L} E\left\{\|\hat{\mathbf{x}}_{\mathrm{M}}(M) - \mathbf{x}\|^2\right\} < \max_{\mathbf{x}\in\mathcal{S}_M} E\left\{\|\hat{\mathbf{x}}_{\mathrm{M}}(M) - \mathbf{x}\|^2\right\}. \tag{5.25}$$

However, since $\hat{\mathbf{x}}_{\mathrm{M}}(L)$ is a minimax MSE estimator for $\mathcal{S}_L$,

$$\max_{\mathbf{x}\in\mathcal{S}_L} E\left\{\|\hat{\mathbf{x}}_{\mathrm{M}}(L) - \mathbf{x}\|^2\right\} \leq \max_{\mathbf{x}\in\mathcal{S}_L} E\left\{\|\hat{\mathbf{x}}_{\mathrm{M}}(M) - \mathbf{x}\|^2\right\}. \tag{5.26}$$

Together with (5.25), this implies that $e(L) < e(M)$ for all $L < M$, which completes the proof. □

As we have seen, using the MSE as a risk function, the set of minimax estimators equals the set of MPS estimators for a given class of parameter sets. Thus, finding an MPS estimator

for a given maximum error ϵ_m becomes simply a matter of finding a minimax estimator whose worst-case error is ϵ_m. In particular, when a closed form is known for the set of minimax estimators and their worst-case errors, one can find a closed form for MPS estimators as well. This is the case for the class of ellipsoidal parameter sets, as demonstrated by the following theorem.

Theorem 5.4. *Consider the MSE risk and define the ellipsoidal parameter sets* $\mathcal{S}_L = \{\mathbf{x} : \ \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$. *Let* $\hat{\mathbf{x}}_{\mathrm{LS}}$ *be the LS estimator* (2.4)*, whose MSE* ϵ_0 *is given by* (2.5)*.*

(a) *Suppose* $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}$ *and* $\mathbf{T}$ *have the same unitary eigenvector matrix* $\mathbf{V}$*, so that* $\mathbf{H}^*\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^*$ *where* $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots \sigma_m)$*, and* $\mathbf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$ *where* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots \lambda_m)$ *with* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$*. An MPS estimator for a given maximum error* ϵ_m *is given by*

$$\hat{\mathbf{x}}_{\mathrm{PS}} = \begin{cases} \mathbf{V}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-k} \end{bmatrix}\mathbf{V}^*(\mathbf{I} - \alpha\mathbf{T}^{1/2})\hat{\mathbf{x}}_{\mathrm{LS}}, & \epsilon_m < \epsilon_0 \\ \hat{\mathbf{x}}_{\mathrm{LS}}, & \epsilon_m \geq \epsilon_0, \end{cases} \tag{5.27}$$

where

$$\alpha = \frac{\sum_{i=k+1}^{m} \frac{1}{\sigma_i} - \epsilon_m}{\sum_{i=k+1}^{m} \frac{\lambda_i^{1/2}}{\sigma_i}} \tag{5.28}$$

and

$$k = \min\left\{ i : \ \alpha\lambda_{i+1}^{1/2} < 1 \right\}. \tag{5.29}$$

(b) *Suppose* $\mathbf{T} = \mathbf{I}$*, i.e., the parameter sets are spherical. In this case, an MPS estimator is*

$$\hat{\mathbf{x}}_{\mathrm{PS}} = \begin{cases} \frac{\epsilon_m}{\epsilon_0}\hat{\mathbf{x}}_{\mathrm{LS}}, & \epsilon_m < \epsilon_0 \\ \hat{\mathbf{x}}_{\mathrm{LS}}, & \epsilon_m \geq \epsilon_0. \end{cases} \tag{5.30}$$

The parameter robustness of this estimator is given by

$$\hat{L}(\hat{\mathbf{x}}_{\mathrm{PS}}) = \begin{cases} \sqrt{\frac{\epsilon_0\epsilon_m}{\epsilon_0 - \epsilon_m}}, & \epsilon_m < \epsilon_0 \\ \infty, & \epsilon_m \geq \epsilon_0. \end{cases} \tag{5.31}$$

Proof. (a) We seek an estimator which guarantees an error not exceeding ϵ_m for as large a parameter set as possible. We begin with the case $\epsilon_m \geq \epsilon_0$. In this case, the allowed error is larger than ϵ_0, the MSE obtained by the LS estimator. Since the LS estimator guarantees this error for any value of $\mathbf{x}$, its parameter robustness is infinite; thus, $\hat{\mathbf{x}}_{\mathrm{LS}}$ is an MPS estimator for this trivial case.

We now consider the case $\epsilon_m < \epsilon_0$. By Theorem 5.3, an MPS estimator is also a minimax estimator. Furthermore, by Theorem 3.3, the minimax MSE estimator for a given parameter set $\mathcal{S}_L$ is given by (3.15), and its worst-case error is given by (3.17). We require a value of L for which the worst-case error equals ϵ_m. Equating (3.17) with ϵ_m and solving for L^2, we obtain the required estimator (5.27).

(b) The case $\mathbf{T} = \mathbf{I}$ is a special case of (a) in which $\mathbf{\Lambda} = \mathbf{I}$. Substituting $\lambda_i = 1$ in the MPS estimator obtained for (a), we observe that $\alpha < 1$ and thus $k = 0$. Furthermore,

$$\sum_{i=1}^{m} \frac{1}{\sigma_i} = \mathrm{Tr}(\mathbf{\Sigma}^{-1}) = \mathrm{Tr}\left((\mathbf{H}^* \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H})^{-1}\right) = \epsilon_0, \tag{5.32}$$

and thus

$$\alpha = \frac{\epsilon_0 - \epsilon_m}{\epsilon_0}. \tag{5.33}$$

Substituting these results into (5.27) yields the required estimator (5.30). We have already seen that the parameter robustness when $\epsilon_m \geq \epsilon_0$ is infinite. To find the parameter robustness when $\epsilon_m < \epsilon_0$, notice that (3.16) is now

$$\alpha = \frac{\epsilon_0}{L^2 + \epsilon_0}. \tag{5.34}$$

Combining this with (5.33) yields

$$L^2 = \frac{\epsilon_0 \epsilon_m}{\epsilon_0 - \epsilon_m}, \tag{5.35}$$

which is the required result (5.31). □

It is sometimes useful to find the actual MSE obtained by an MPS estimator for a particular value of $\mathbf{x}$. The MSE can be calculated using the matching minimax estimator. For example, it has been shown in (3.14) that the MSE of the minimax estimator for a spherical parameter set $\mathbf{x}^*\mathbf{x} \leq L^2$ is given by

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{M}}, \mathbf{x}) = \left(\frac{L^2}{L^2 + \epsilon_0}\right)^2 \epsilon_0 + \left(\frac{\epsilon_0}{L^2 + \epsilon_0}\right)^2 \|\mathbf{x}\|^2. \tag{5.36}$$

Substituting the value of L^2 from (5.31), we have

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{PS}}, \mathbf{x}) = \begin{cases} \frac{\epsilon_m^2}{\epsilon_0} + \left(\frac{\epsilon_0 - \epsilon_m}{\epsilon_0}\right)^2 \|\mathbf{x}\|^2, & \epsilon_m < \epsilon_0 \\ \epsilon_0, & \epsilon_m \geq \epsilon_0. \end{cases} \tag{5.37}$$

Thus, the MSE of the maximum spherical parameter set estimator is a linear function of $\|\mathbf{x}\|^2$. This result is useful for comparing the performance of the MPS estimator with other estimators, as we demonstrate in Section 5.3.

5.1.5 Linear Regret Estimators

We now present a different example of an MPS estimator, one which guarantees a worst-case regret (see Subsection 2.1.2). The regret is given by (2.3), and is defined as the difference between the MSE and the best MSE obtainable using a linear estimator $\hat{\mathbf{x}} = \mathbf{G}(\mathbf{x})\mathbf{y}$ which is a function of $\mathbf{x}$.

In this section, we limit our discussion to parameter sets of the form $\mathcal{S}_L = \{\mathbf{x} : \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$, where $\mathbf{T}$ is a Hermitian positive definite weighting matrix. For analytical tractability, we further restrict the discussion to the case where $\mathbf{T}$ and $\mathbf{H}^*\mathbf{C}_\mathbf{w}^{-1}\mathbf{H}$ have the same eigenvectors. We show that, under these assumptions, the linear MPS regret estimator is equivalent to the linear minimax regret estimator. It follows that the MPS estimator can be found as easily as the minimax estimator. In particular, closed-form solutions can be developed for some values of $\mathbf{T}$ and L, using the results of Subsection 3.2.2.

Theorem 5.5. *Suppose that the error function of interest is the regret* (2.3)*. Let $\mathcal{E}$ be the class of linear estimators, and let $\mathcal{S}_L = \{\mathbf{x} : \mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2\}$ be a class of parameter sets, where $\mathbf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$ is a Hermitian positive definite weighting matrix, $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_i > 0$, and $\mathbf{V}$ is an eigenvector matrix of $\mathbf{H}^*\mathbf{C}_\mathbf{w}^{-1}\mathbf{H}$. An estimator $\hat{\mathbf{x}} \in \mathcal{E}$ is a linear minimax regret estimator over $\mathcal{S}_L$ if, and only if, it is a linear MPS regret estimator with maximum error ϵ_m equal to the worst-case error $e(L)$ of* (5.18)*.*

Proof. By Theorem 5.2, it is sufficient to show that $e(L)$ is strictly increasing with L. By Theorem 3.5, the linear minimax regret estimator $\hat{\mathbf{x}}_\mathrm{M}(L)$ is the solution to the convex optimization problem (3.22). We will analyze this optimization problem to show that $e(L)$ is indeed strictly increasing with L.

We first show that (3.22b) is an active constraint in the optimization problem. Assume by contradiction that (3.22b) is inactive. Then, by the Karush-Kuhn-Tucker conditions for optimality [37, Sec. 5.5.3], (3.22) is equivalent to

$$\min_{\mathbf{d},\tau} \tau \quad \text{s.t. } \sum \frac{d_i^2}{\sigma_i} \leq \tau, \tag{5.38}$$

for which the optimal solution is $\mathbf{d} = \mathbf{0}, \tau = 0$. However, for any $\mathbf{r} \in \mathcal{S}$,

$$F_2(\mathbf{0}, \mathbf{r}) > 0 = \tau, \tag{5.39}$$

contradicting the fact that (3.22b) is inactive. Thus, for the optimal value of τ and $\mathbf{d}$, there exists at least one active $\mathbf{r} \in \mathcal{S}$ for which $F_2(\mathbf{d}, \mathbf{r}) = \tau$.

Next, define

$$g(\mathbf{d}, \mathbf{r}, L^2) \triangleq \sum (1 - d_i)^2 r_i - \frac{\sum r_i}{1 + L^2 \sum \sigma_i r_i}. \tag{5.40}$$

Let us study the behavior of $g(\mathbf{d}, \mathbf{r}, L^2)$ when L^2 is changed. Observe that

$$\frac{\partial g}{\partial L^2} = \frac{\sum \sigma_i r_i \sum r_i}{\left(1 + L^2 \sum \sigma_i r_i\right)^2} > 0. \tag{5.41}$$

Thus, $g(\mathbf{d}, \mathbf{r}, L^2)$ is strictly increasing with L. Therefore, if L is decreased, then $F_2(\mathbf{d}, \mathbf{r}) = F_1(\mathbf{d}) + L^2 g(\mathbf{d}, \mathbf{r}, L^2)$ is decreased for all active $\mathbf{r}$, and the constraint (3.22b) is relaxed, which implies that the optimal value of τ is also decreased. Since this value equals $e(L)$, we conclude that $e(L)$ is strictly monotonic in L, which completes the proof. □

5.2 Maximum Noise Level Estimation

In Section 5.1, we assumed that the noise covariance $E\{\mathbf{w}\mathbf{w}^*\}$ is known. In practice, this is rarely the case, and the covariance must itself often be estimated from measurements. In this section, we consider the case

$$E\{\mathbf{w}\mathbf{w}^*\} = \sigma^2 \mathbf{C}_\mathbf{w}, \tag{5.42}$$

for some unknown deterministic *noise level* σ^2, and some known covariance matrix $\mathbf{C}_\mathbf{w}$ [39]. For example, when the noise is i.i.d., $\mathbf{C}_\mathbf{w} = \mathbf{I}$ and σ^2 is the noise variance. The estimation techniques used so far require complete knowledge of the noise covariance. Thus, minimax or MPS approaches cannot be directly applied to this problem, unless the noise parameters are estimated from the measurements; this increases computational complexity and is potentially unreliable.

As an alternative approach, we propose to estimate $\mathbf{x}$ from the observations, while guaranteeing maximum error requirements, for as large a range of noise levels as possible. To this end, we assume that $\mathbf{x} \in \mathcal{S}$ for a known parameter set $\mathcal{S}$, and require a maximum error level ϵ_m. We seek the estimator which guarantees an error not exceeding ϵ_m for all $\mathbf{x} \in \mathcal{S}$, and for as large a noise level σ^2 as possible; this will be referred to as the maximum noise level (MNL) estimator. As we shall show, the MNL estimator is related to the minimax estimator, allowing us to efficiently find the MNL estimator whenever the minimax estimator is known.

Formally, we define an error function $\epsilon_{\sigma^2}(\hat{\mathbf{x}}, \mathbf{x})$, such as the MSE or the regret, and require some level of performance $\epsilon_{\sigma^2}(\hat{\mathbf{x}}, \mathbf{x}) \leq \epsilon_m$ to be satisfied over the entire range $\mathbf{x} \in \mathcal{S}$. We can now define a new type of maximum set estimator, in a manner analogous to the definition of the MPS in Subsection 5.1.2, as follows.

Definition 5.4. The *noise robustness* $\hat{\sigma}^2$ of an estimator $\hat{\mathbf{x}}$ is defined as the maximum σ^2 for which the performance requirement is satisfied,

$$\hat{\sigma}^2(\hat{\mathbf{x}}) = \sup \left\{ \sigma^2 : \max_{\mathbf{x} \in \mathcal{S}} \epsilon_{\sigma^2}(\hat{\mathbf{x}}, \mathbf{x}) \leq \epsilon_m \right\}. \tag{5.43}$$

Definition 5.5. The *maximum noise level (MNL) estimator* $\hat{\mathbf{x}}_{\mathrm{NL}}$ (among a class of estimators $\mathcal{E}$) is the estimator maximizing the noise robustness among all estimators in $\mathcal{E}$, for given $\mathcal{S}$, $\epsilon_{\sigma^2}(\hat{\mathbf{x}}, \mathbf{x})$ and ϵ_m:

$$\hat{\mathbf{x}}_{\mathrm{NL}} = \arg \max_{\hat{\mathbf{x}} \in \mathcal{E}} \hat{\sigma}^2(\hat{\mathbf{x}}). \tag{5.44}$$

We now show that, if the error function ϵ_{σ^2} is continuous in σ^2, then the MNL estimator is a minimax estimator. The error function is indeed continuous for many cases of interest, such as the MSE and the regret.

Theorem 5.6. *Suppose the error function ϵ of interest is continuous in σ^2. Then, the MNL estimator $\hat{\mathbf{x}}_{\mathrm{NL}}$ is a minimax estimator for the parameter set $\mathcal{S}$, with noise level $\sigma_1^2 = \hat{\sigma}^2(\hat{\mathbf{x}}_{\mathrm{NL}})$.*

Proof. Assume by contradiction that $\hat{\mathbf{x}}_{\mathrm{NL}}$ is not a minimax estimator. Then, there exists $\hat{\mathbf{x}}_{\mathrm{M}} \neq \hat{\mathbf{x}}_{\mathrm{NL}}$ such that

$$\max_{\mathbf{x} \in \mathcal{S}} \epsilon_{\sigma_1^2}(\hat{\mathbf{x}}_{\mathrm{M}}, \hat{\mathbf{x}}) < \max_{\mathbf{x} \in \mathcal{S}} \epsilon_{\sigma_1^2}(\hat{\mathbf{x}}_{\mathrm{NL}}, \mathbf{x}) \leq \epsilon_m. \tag{5.45}$$

However, since ϵ_{σ^2} is continuous in σ^2, a sufficiently small change in σ^2 causes an arbitrarily small change in ϵ_{σ^2}. Thus, there exists $\sigma_2^2 > \sigma_1^2$ such that

$$\max_{\mathbf{x} \in \mathcal{S}} \epsilon_{\sigma_2^2}(\hat{\mathbf{x}}_{\mathrm{M}}, \hat{\mathbf{x}}) \leq \epsilon_m. \tag{5.46}$$

Hence, $\hat{\sigma}^2(\hat{\mathbf{x}}_{\mathrm{M}}) \geq \sigma_2^2 > \sigma_1^2 = \hat{\sigma}^2(\hat{\mathbf{x}}_{\mathrm{NL}})$, contradicting the fact that $\hat{\mathbf{x}}_{\mathrm{NL}}$ is an MNL estimator. □

A consequence of this theorem is that an MNL estimator can be found if an algorithm for finding a minimax estimator is known. This can be performed efficiently using a line search, in which minimax estimators are calculated for various noise levels, until a minimax estimator whose worst-case error equals ϵ_m is found. Alternatively, as the following theorem demonstrates, a closed form for a linear MNL estimator can be identified when a closed form for the minimax estimator is known.

Theorem 5.7. *Let $\mathcal{S} = \{\mathbf{x} : \|\mathbf{x}\|^2 \leq L^2\}$ and let $\epsilon_{\sigma^2}(\hat{\mathbf{x}}, \mathbf{x})$ be the MSE. For a given maximum error ϵ_m, a linear maximum noise level (MNL) estimator is given by*

$$\hat{\mathbf{x}}_{\mathrm{NL}} = \begin{cases} \frac{L^2 - \epsilon_m}{L^2} \hat{\mathbf{x}}_{\mathrm{LS}}, & L^2 > \epsilon_m \\ \mathbf{0}, & L^2 \leq \epsilon_m, \end{cases} \tag{5.47}$$

where $\hat{\mathbf{x}}_{\mathrm{LS}}$ is the LS estimator (2.4).

Proof. We first consider the case $L^2 \leq \epsilon_m$. In this case, the performance requirements are extremely lax, and many estimators satisfy these requirements for *any* noise level. In particular, the estimator $\hat{\mathbf{x}} = \mathbf{0}$ has an MSE of $\|\mathbf{x}\|^2$, for which the worst case is $\max \|\mathbf{x}\|^2 = L^2 \leq \epsilon_m$; this is true regardless of the noise level. Thus, $\hat{\mathbf{x}} = \mathbf{0}$ is an MNL estimator (with infinite noise robustness) for the trivial case $L^2 \leq \epsilon_m$.

We now turn to the more interesting case $L^2 > \epsilon_m$. By Theorem 5.6, $\hat{\mathbf{x}}_{\mathrm{NL}}$ is a minimax estimator for some noise level σ^2. By Theorem 3.2, the unique minimax estimator over the parameter set $\mathcal{S}$, for a given noise level σ^2, is

$$\hat{\mathbf{x}}_{\mathrm{M}}(\sigma^2) = \frac{L^2}{L^2 + \sigma^2 \epsilon_0} \hat{\mathbf{x}}_{\mathrm{LS}}, \tag{5.48}$$

where ϵ_0 is given by (2.5). From (3.11), the worst-case error for this estimator within the set $\mathcal{S}$ is given by

$$\max_{\mathbf{x} \in \mathcal{S}} \epsilon_{\sigma^2}(\hat{\mathbf{x}}_{\mathrm{M}}(\sigma^2), \mathbf{x}) = \frac{L^2 \sigma^2 \epsilon_0}{L^2 + \sigma^2 \epsilon_0}. \tag{5.49}$$

The critical value of σ^2 for which this value exactly equals ϵ_m is given by

$$\sigma^2 = \frac{\epsilon_m L^2}{\epsilon_0 (L^2 - \epsilon_m)}. \tag{5.50}$$

Substituting this value of σ^2 into (5.48) yields the required estimator (5.47). □

It is instructive to compare the closed forms obtained for the MPS estimator (Theorem 5.4b) and the MNL estimator (Theorem 5.7), when spherical parameter sets are used. Both estimators take the form of a linear minimax MSE estimator for a spherical parameter set, and hence they are shrinkage estimators. They can thus be viewed as a compromise between the LS estimator and the zero estimator. However, for the MPS estimator, the shrinkage factor increases with the maximum allowed error ϵ_m; while for the MNL estimator, the shrinkage factor decreases with ϵ_m. The reason for this is as follows. When the allowed error is increased, an increase in either the parameter set or noise level is allowed. However, a larger parameter set is achieved by an estimator closer to the LS estimator (which provides constant error for all $\mathbf{x}$); while a larger noise level is achieved by an estimator closer to the zero estimator (which provides zero error, regardless of noise level, for the nominal value $\mathbf{x} = \mathbf{0}$). Thus, increasing the maximum allowed error has opposite effects, depending on whether the goal is to increase the robustness to uncertainty in the parameter set or in the noise level.

5.3 Application: Channel Estimation

As a demonstration of the maximum set estimation concept, we consider the problem of estimating the channel in a communication system. Specifically, we seek the impulse response of an unknown channel; to this end, we use a preamble (also called a training sequence), which is transmitted along with payload data. The received symbols are compared to the known preamble, and this information is used to obtain an estimate of the channel response. Knowledge of the channel response is required in many detection algorithms, for example, in maximum likelihood sequence estimation (MLSE) [40]. We will compare the standard LS approach to channel estimation with the MPS technique developed in Section 5.1.

Let $\mathbf{c} = (c_0, \dots c_{N_c-1})^T$ denote the unknown channel impulse response of known length N_c, and let

$$\mathbf{p} = (p_{-N_c+1}, p_{-N_c+2}, \dots p_0, \dots p_{N_p-N_c})^T \tag{5.51}$$

denote the known vector of preamble symbols of length N_p. The corresponding received symbols are given by

$$r_k = \sum_{l=0}^{N_c-1} c_l p_{k-l} + w_k, \quad k = 0, 1, \dots N_p - N_c, \tag{5.52}$$

where $\mathbf{w} = (w_0, \dots w_{N_p-N_c})^T$ is additive white noise with variance σ_w^2. Defining

$$\mathbf{H} = \begin{bmatrix} p_0 & p_{-1} & \cdots & p_{-N_c+1} \\ p_1 & p_0 & \cdots & p_{-N_c+2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N_p-N_c} & p_{N_p-N_c-1} & \cdots & p_{N_p-1} \end{bmatrix}, \tag{5.53}$$

we have

$$\mathbf{r} = \mathbf{H}\mathbf{c} + \mathbf{w}. \tag{5.54}$$

The classical approach to channel estimation using a preamble is least-squares estimation of the unknown, deterministic vector $\mathbf{c}$ from the measurements $\mathbf{r}$ [40–42]. The estimated channel in this case is

$$\hat{\mathbf{c}}_{\mathrm{LS}} = \mathbf{G}_{\mathrm{LS}}\mathbf{r} = (\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*\mathbf{r}. \tag{5.55}$$

This estimator minimizes the *measurement* error $\|\mathbf{r} - \mathbf{H}\mathbf{G}\mathbf{r}\|^2$. However, we are interested in minimizing the *estimation* error $\epsilon = E\{\|\mathbf{c} - \hat{\mathbf{c}}\|^2\}$, as the channel estimate is used for further processing (e.g., detection of payload data). For example, in [41], an increase in channel estimation error is assumed to be equivalent to an increase in noise level.

Unfortunately, the channel estimation error ϵ is a function of the unknown channel $\mathbf{c}$, so direct minimization of ϵ is not possible. Were we to know that $\mathbf{c}$ lies within some bounded set $\mathcal{S}$, a minimax MSE approach would allow us to minimize the worst-case error among all possible channels within $\mathcal{S}$. However, we generally only have a vague understanding that channel dispersion is limited and that most of the energy in $\mathbf{c}$ lies in the first component.

On the other hand, the desired channel estimation error is a parameter with known implications for the system designer. In particular, the maximum channel estimation error may be treated as an added noise source [41]. In this case, the estimation error requirement is a design parameter; it is to be chosen together with other system properties such as receiver signal to interference plus noise ratio (SINR) requirements. We can use the MPS estimator to maximize the set of channels for which a required estimation error ϵ_m is achieved. Thus, we assume that the given maximum estimation error is critical for system operation, and should be guaranteed for as wide a range of channels as possible.

Let $\mathbf{c}^0 = (1, 0, \ldots 0)^T$ be a perfect (nondispersive) channel, and let $\mathbf{c}' = \mathbf{c} - \mathbf{c}^0$. We construct a simple class of parameter sets by defining

$$\mathcal{S}_L = \{\mathbf{c}' : \|\mathbf{c}'\| \leq L\}. \tag{5.56}$$

This model assumes that most of the channel energy is concentrated in the first tap, and that deviations from this nominal value are fairly uniform among the channel taps. More elaborate models may be constructed if additional information about the channel properties is known.

We seek an estimator guaranteeing estimation error of ϵ_m or less, for as large a parameter set as possible. From (5.54), we have

$$\mathbf{r} - \mathbf{H}\mathbf{c}^0 = \mathbf{H}\mathbf{c}' + \mathbf{w}. \tag{5.57}$$

By Theorem 5.4, the maximum error ϵ_m must first be compared with $\epsilon_0 = \mathrm{Tr}((\mathbf{H}^*\mathbf{H})^{-1})$, the MSE of the LS estimator. If $\epsilon_m \geq \epsilon_0$, then an error of ϵ_0 is allowable. Such an error is guaranteed by the LS estimator for *any* value of $\mathbf{c}$, so that the LS estimator has infinite parameter robustness in this case. However, if $\epsilon_m < \epsilon_0$, then an MPS estimator is given by

$$\hat{\mathbf{c}}'_{\mathrm{PS}} = \frac{\epsilon_m}{\epsilon_0}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*(\mathbf{r} - \mathbf{H}\mathbf{c}^0), \tag{5.58}$$

and thus

$$\hat{\mathbf{c}}_{\mathrm{PS}} = \frac{\epsilon_m}{\epsilon_0}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*\mathbf{r} + \left(1 - \frac{\epsilon_m}{\epsilon_0}\right)\mathbf{c}^0. \tag{5.59}$$

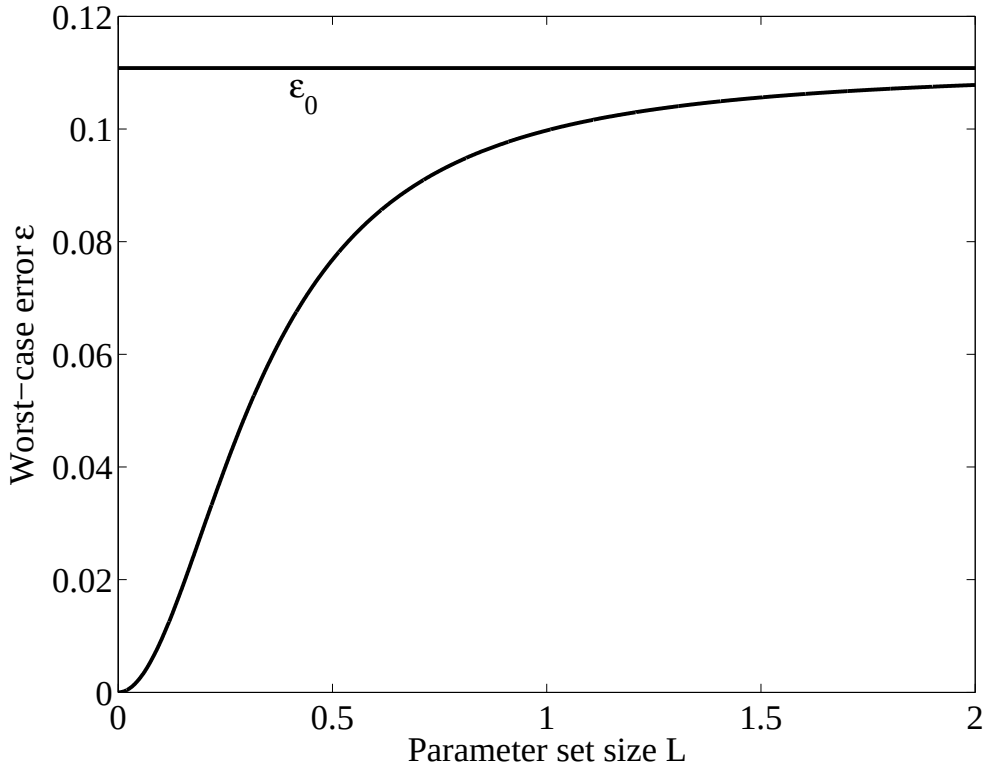

Figure 5.1: The worst-case error of various minimax MSE channel estimators

To compare the performance of the LS and MPS channel estimators, we consider the problem of estimating a 7-tap channel using the optimal 14-symbol BPSK preamble suggested in [41], and given by

$$[-1,-1,-1,+1,-1,-1,+1,-1,-1,-1,+1,+1,+1,-1]. \tag{5.60}$$

We assume that the noise variance is $\sigma_w^2 = 0.1$. The worst-case error of various minimax MSE estimators is given by (3.11) and plotted in Figure 5.1. By Theorem 5.4, all of these estimators are also MPS estimators. An engineer constructing a channel estimation system should use such a plot as a design tool, as it demonstrates the tradeoff between channel estimation error and the range of channels for which the error can be achieved.

Suppose we choose to design our system such that a channel estimation error of $\epsilon_m = \frac{2}{3}\epsilon_0$ is to be tolerated; this choice covers a reasonably-sized parameter set while substantially reducing the estimation error. We note that the choice of ϵ_m is accompanied by appropriate design steps, which will allow the receiver to handle the resulting estimation errors (for example, error correction capabilities suitable for such noise levels). The MSE (5.37) of the resulting MPS estimator is compared with the MSE ϵ_0 of the LS estimator in Figure 5.2.

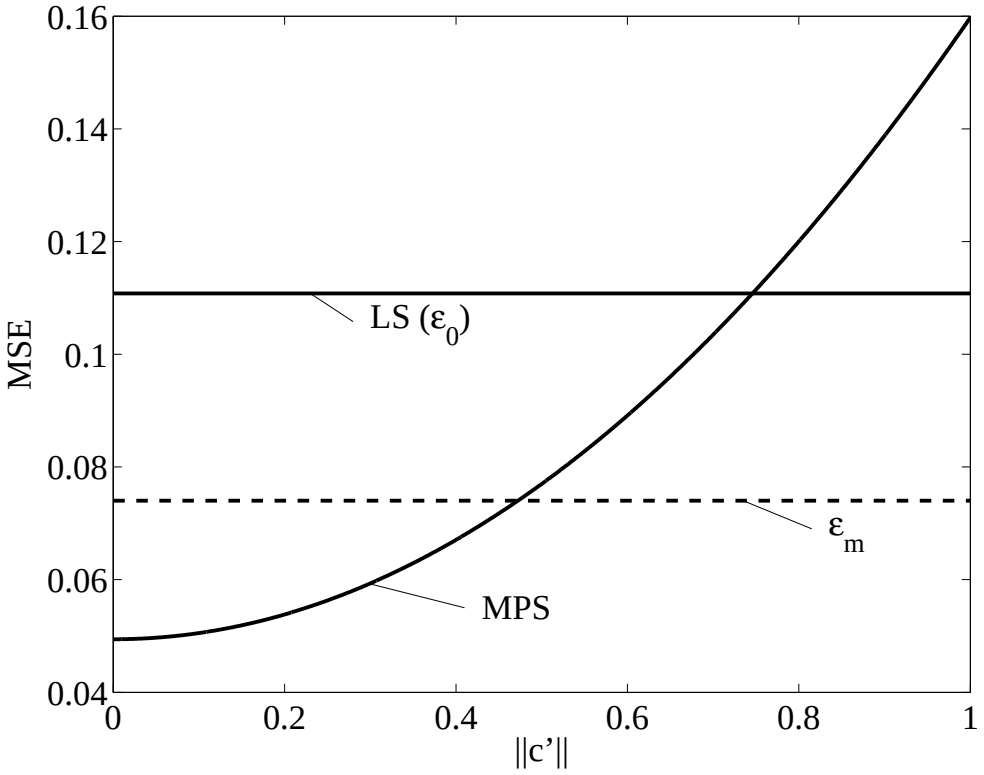

Figure 5.2: Channel estimation error of MPS and LS estimators for various channels

To verify that the reduced estimation error resulting in improved detection performance, a BPSK detection scenario was simulated [43]. A signal containing the 14 preamble symbols (5.60) and 100 random data symbols was generated. Channels were simulated by choosing each tap c_i ($1 \leq i \leq 7$) to be an independent Rayleigh-distributed variate with parameter $A\beta^i$, where $0 \leq \beta \leq 1$ is the channel dispersion factor, and A is chosen so that $E\{\|\mathbf{c}\|^2\} = 1$. Thus, $\beta = 0$ results in a nondispersive channel, while $\beta = 1$ indicates a channel for which the taps are identically distributed (maximum channel dispersion). The channel was estimated using both the LS and MPS estimators described above, and the resulting channel estimate was used for MLSE detection of the data symbols. The simulation was repeated to obtain an estimate of the bit error rate (BER). The results are presented in Figure 5.3. For comparison, a null estimator is also plotted; this "estimator" assumes a nondispersive channel, i.e., $\hat{\mathbf{c}} = \mathbf{c}^0$.

The MPS estimator is a compromise between the LS estimator and the null estimator: the LS estimator has modest estimation error requirements, but achieves them for all values of $\mathbf{c}$; the null estimator can be viewed as an estimator requiring zero estimation error, and achieves this requirement only for the nominal channel $\mathbf{c}^0$. MPS estimators provide a continuum of choices between these two extremes, allowing the designer to choose a point in the tradeoff between

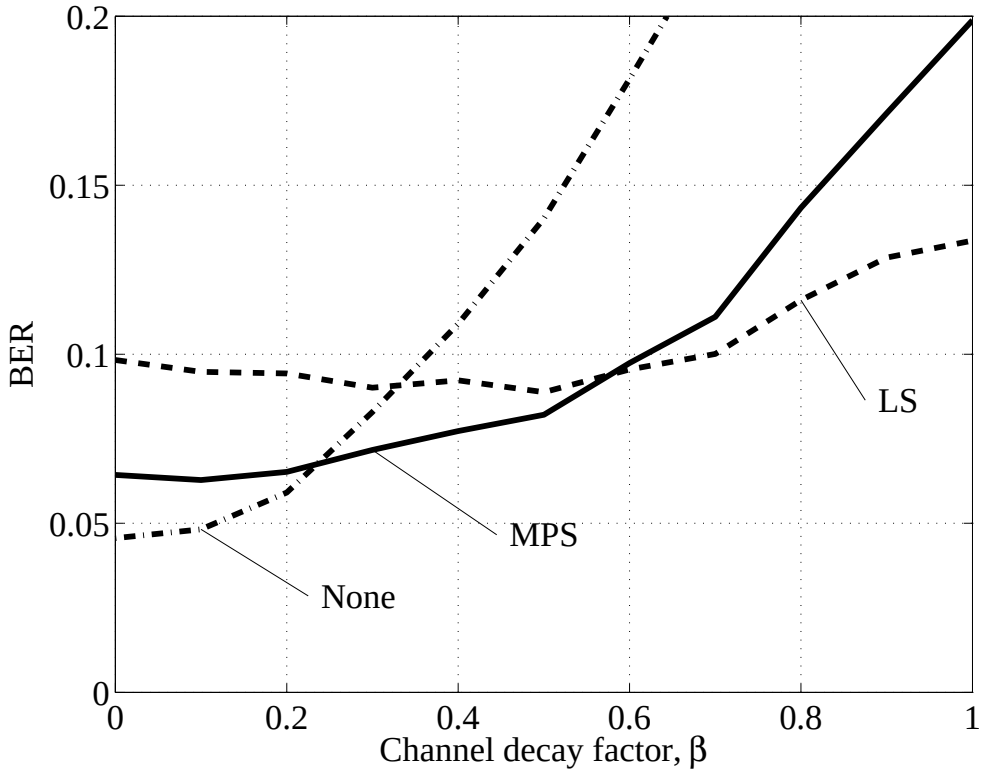

Figure 5.3: BER for various channels with the LS and MPS channel estimators

the estimation error requirement and the size of the parameter set for which the requirement is achieved. An appropriate choice of ϵ_m leads to an estimator which considerably outperforms the LS estimator for low- and moderate-dispersion channels, and fails only when channel taps are nearly identically distributed.

5.4 Discussion

In this chapter, we considered the problem of parameter estimation given a maximum allowed estimation error. This is appropriate for systems designed to function with a known and tolerable error margin, such as communication systems designed for a certain SNR level. We have developed estimators which guarantee the required estimation error for as wide a range of operating conditions as possible. The goal of this paper has been to show that estimators which make use of given estimation error requirements outperform classical approaches such as the LS estimator.

The maximum set estimation concept was first applied to find the largest parameter set $\mathcal{S}_L$ such that performance is guaranteed for any parameter $\mathbf{x}$ in $\mathcal{S}_L$. This results in the maximum

parameter set (MPS) estimator. Next, the maximum noise level (MNL) estimator was developed; this estimator maximizes the range of noise variances for which the required estimation error is guaranteed.

As we have seen, in many cases, the maximum set estimator is equivalent in form to a matching minimax estimator: the maximum set estimator for a given error requirement ϵ_m equals a minimax estimator whose worst-case error is ϵ_m. However, while minimax estimators assume a given bound on the parameter set, maximum set estimators assume a requirement on the obtained estimation error. Thus, these estimators are used under different circumstances, and their similarity in form merely serves as a mathematical tool for finding maximum set estimators based on known results for minimax estimators.

The maximum allowed error is often a function of system design parameters, and can be influenced by design decisions. In such cases, a plot of the worst-case error as a function of the size of the parameter set (as in Figure 5.1) can be used as a design tool. Such a plot can be interpreted in two complementary ways. It describes the worst-case error obtained if a minimax estimator is used with a given parameter set bound. However, it also defines the size of the parameter set obtained if an MPS estimator is used with a given maximum error. Thus, such a plot can be used to select a meaningful value for the maximum error, based on the tradeoff between estimation error and parameter set bound.

The choice of an appropriate estimator for a given problem depends on the data available to the designer. Knowledge of the second-order statistics of the parameters $\mathbf{x}$, for example, leads to the well-known Wiener estimator, which is optimal in an MSE sense. However, partial information can also be used to improve estimation performance. The maximum allowed estimation error is an example of added information which may be known to the designer, and as we have demonstrated, can often result in improved performance.

Bibliography

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[3] K. F. Gauss, "Theoria combinationis obsercationunt erronbus minimis obnoxiae," 1821.

[4] A.-M. Legendre, *Nouvelles méthodes pour la détermination des orbites des cometes*, 1806.

[5] M. S. Pinsker, "Optimal filtering of square-integrable signals in Gaussian noise," *Problems in Inform. Transmission*, vol. 16, pp. 120–133, 1980.

[6] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of bounded data uncertainties," *IEEE Trans. Signal Processing*, vol. 53, no. 1, pp. 168–181, Jan. 2005.

[7] ——, "Linear minimax regret estimation of deterministic parameters with bounded data uncertainties," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2177–2188, Aug. 2004.

[8] A. N. Tichonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: V. H. Winston, 1977.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[10] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate distribution," in *Proc. Third Berkeley Symp. Math. Statist. Prob.*, vol. 1, 1956, pp. 197–206.

[11] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, vol. 1, 1961, pp. 311–319.

[12] J. R. Thompson, "Some shrinkage techniques for estimating the mean," *J. Amer. Statist. Assoc.*, vol. 63, no. 321, pp. 113–122, Mar. 1968.

[13] A. J. Baranchik, "A family of minimax estimators of the mean of a multivariate normal distribution," *Ann. Math. Statist.*, vol. 41, no. 2, Apr. 1970.

[14] B. Efron and C. Morris, "Stein's estimation rule and its competitors: an empirical Bayes approach," *J. Amer. Statist. Assoc.*, vol. 68, pp. 117–130, 1973.

[15] M. E. Bock, "Minimax estimators of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 3, no. 1, pp. 209–218, Jan. 1975.

[16] J. O. Berger, "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss," *Ann. Statist.*, vol. 4, no. 1, pp. 223–226, Jan. 1976.

[17] E. Greenberg and C. E. Webster, Jr., *Advanced Econometrics*, 2nd ed. New York: Wiley, 1983.

[18] J. H. Manton, V. Krishnamurthy, and H. V. Poor, "James-Stein state filtering algorithms," *IEEE Trans. Signal Processing*, vol. 46, no. 9, pp. 2431–2447, Sept. 1998.

[19] B. Efron and C. Morris, "Combining possibly related estimation problems," *J. Roy. Statist. Soc. B*, vol. 35, no. 3, pp. 379–421, 1973.

[20] Z. Ben-Haim and Y. C. Eldar, "Minimax estimators dominating the least-squares estimator," in *Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP 2005)*, vol. IV, Philadelphia, PA, Mar. 2005, pp. 53–56.

[21] ——, "Blind minimax estimators: Improving on least-squares estimation," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP 2005)*, Bordeaux, France, July 2005.

[22] ——, "Blind minimax estimation," *IEEE Trans. Inform. Theory*, submitted.

[23] Y. Ben-Haim, "Set-models of information-gap uncertainty: axioms and an inference scheme," *J. Franklin Institute*, vol. 336, pp. 171–199, 2000.

[24] ——, *Information-Gap Decision Theory: Decisions under Severe Uncertainty*. San Diego, CA: Academic Press, 2001.

[25] Z. Ben-Haim and Y. C. Eldar, "Estimation with maximum error requirements," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI 2004)*, Tel-Aviv, Israel, Sept. 2004, pp. 416–419.

[26] ——, "Maximum set estimators with bounded estimation error," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 3172–3182, Aug. 2005.

[27] A. Cohen, "All admissible linear estimators of the mean vector," *Ann. Math. Statist.*, vol. 37, no. 2, pp. 458–463, Apr. 1966.

[28] Y. C. Eldar, "Comparing between estimation approaches: Admissible and dominating linear estimators," *IEEE Trans. Signal Processing*, to appear.

[29] Y. Maruyama, "A unified and broadened class of admissible minimax estimators of a multivariate normal mean," *J. Multivariate Analysis*, vol. 64, pp. 196–205, 1998.

[30] ——, "Minimax admissible estimation of a multivariate normal mean and improvement upon the James-Stein estimator," Ph.D. dissertation, University of Tokyo, 2000.

[31] L. D. Brown, "On the admissibility of invariant estimators of one or more location parameters," *Ann. Math. Statist.*, vol. 37, pp. 1087–1136, 1966.

[32] B. Efron and C. Morris, "Limiting the risk of Bayes and empirical Bayes estimators — Part II: The empirical Bayes case," *J. Amer. Statist. Assoc.*, vol. 67, no. 337, pp. 130–139, Mar. 1972.

[33] A. J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution," Department of Statistics, Stanford University, Stanford, CA, Tech. Rep. #51, 1964.

[34] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York, NY: Springer-Verlag, 1985.

[35] A. P. Dawid, "Comments on 'Combining possibly related estimation problems'," *J. Roy. Statist. Soc. B*, vol. 35, no. 3, pp. 409–410, 1973.

[36] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

[37] S. Boyd and L. Vanderberghe, *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press, 2004. [Online]. Available: http://www.stanford.edu/~boyd/cvxbook.html

[38] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge Univ. Press, 1985.

[39] A. Beck, A. Ben-Tal, and Y. C. Eldar, "Robust mean-squared error estimation of multiple signals in linear systems affected by model and noise uncertainties," *Math. Programming*, to appear.

[40] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1996.

[41] S. N. Crozier, D. D. Falconer, and S. A. Mahmoud, "Least sum of square errors (LSSE) channel estimation," *IEE Proc.-F*, vol. 138, no. 4, pp. 371–378, Aug. 1991.

[42] C. Fragouli, N. Al-Dhahir, and W. Turin, "Training-based channel estimation for multiple-antenna broadband transmissions," *IEEE Trans. Wireless Commun.*, vol. 2, no. 2, pp. 384–391, Mar. 2003.

[43] M. C. Jeruchim, P. Balaban, and K. S. Shanmugan, *Simulation of Communication Systems*, 2nd ed. New York: Plenum, 1992.