CCIT Report #437
July 2003

Linear Minimax Regret Estimation with Bounded Data Uncertainties

Yonina C. Eldar*, A. Ben-Tal† and A. Nemirovski†

June 26, 2003

Abstract

We develop a new linear estimator for estimating an unknown vector of parameters $\mathbf{x}$ in a linear model, in the presence of bounded data uncertainties. The estimator is designed to minimize the worst-case *regret* over all bounded data vectors, namely the worst-case difference between the MSE attainable using a linear estimator that does not know the true parameters $\mathbf{x}$, and the optimal MSE attained using a linear estimator that knows $\mathbf{x}$. We demonstrate through several examples that the minimax regret estimator can significantly increase the performance over the conventional least-squares estimator, as well as several other least-squares alternatives.

1 Introduction

The problem of estimating a vector of unknown parameters $\mathbf{x}$ from noisy observations $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, where $\mathbf{H}$ is a known matrix and $\mathbf{w}$ is a noise vector, arises in many different fields in science and engineering, and consequently attracted much attention in the estimation literature.

If the unknown parameters $\mathbf{x}$ are assumed to be random variables with known second-order statistics, then the linear estimator minimizing the mean-squared error (MSE) is the well-known Wiener estimator [1, 2]. However, in many problems of practical interest there is no statistical information available on $\mathbf{x}$, so that $\mathbf{x}$ is treated as an unknown set of deterministic parameters. In this case, the MSE of an estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ depends explicitly on the unknown parameters $\mathbf{x}$, and therefore cannot be minimized directly.

Since the MSE between $\hat{\mathbf{x}}$ and $\mathbf{x}$ depends on $\mathbf{x}$, a common approach is to seek estimators that minimize some function of the data error $\hat{\mathbf{y}} - \mathbf{y}$, where $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}$ is the estimated data vector. The celebrated least-

*Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. E-mail: yonina@ee.technion.ac.il.

†MINERVA Optimization Center, Department of Industrial Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. E-mail: {morbt,nemirovs}@ie.technion.ac.il.

squares estimator, first studied by Gauss [3], seeks the estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ that minimizes the squared-norm of the data error $\|\hat{\mathbf{y}} - \mathbf{y}\|^2$. It is well known that the least-squares estimate is also the best linear unbiased estimator [4], *i.e.,* it has the smallest variance among all linear *unbiased* estimators. On the negative side, an unbiased estimator does not necessarily lead to a small MSE. In fact, it is well known that in many cases the least-squares estimator can result in a large MSE.

Various modifications of the least-squares estimator for the case in which the data model holds, so that $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ with $\mathbf{H}$ and $\mathbf{y}$ known exactly, and $\mathbf{x}$ represents deterministic unknown parameters, have been proposed. Among the alternatives are Tikhonov regularization [5], also known in the statistical literature as the ridge estimator [6], the shrunken estimator [7], and the covariance shaping least-squares estimator [8]. In general, these least-squares alternatives attempt to reduce the MSE in estimating $\mathbf{x}$ by allowing for a bias. Each of the estimators above can be shown to be a solution to an optimization problem which involves minimizing some function that depends on the *data error*.

In an estimation context, we typically would like to minimize the *estimation error*, rather than the data error. To this end we assume that $\mathbf{x}$ is known to satisfy a (possibly weighted) norm constraint, and then seek a robust estimator whose performance is reasonably good across all possible choices of the parameters $\mathbf{x}$, in the region of uncertainty. The most common approach for designing robust estimators is the minimax MSE approach, initiated by Huber [9, 10], in which we seek the estimator that minimizes the worst-case MSE in the region of uncertainty. This approach has been applied to a variety of different estimation problems in which the unknown parameters $\mathbf{x}$ are assumed to be random, but the statistics of $\mathbf{x}$ are not completely specified [11, 12, 13, 14, 15, 16, 17]. The minimax approach, in which the goal is to optimize the worst-case performance, is one of the major techniques for designing robust systems with respect to modelling uncertainties, and has been applied to many problems in detection and estimation [18, 19, 20].

Following the popular minimax approach, we may seek the linear estimator that minimizes the worst-case MSE over all possible values of $\mathbf{x}$ that satisfy a weighted norm constraint. The minimax estimator of this form is developed in [21], in which the case of uncertainties in the model matrix $\mathbf{H}$ is also considered.

Although the minimax approach has enjoyed widespread use in the design of robust methods for signal processing and communication [18, 20], its performance is often unsatisfactory. The main limitation of this approach is that it tends to be overly conservative since it optimizes the performance for the worst possible choice of unknowns. As we show in the context of concrete examples in Section 6, this can often lead to degraded performance.

To improve the performance of the minimax MSE estimator, we propose, in Section 3, a new approach to

linear estimation, in which we seek a linear estimator whose performance is as close as possible to that of the optimal linear estimator, *i.e.,* the one minimizing the MSE when $\mathbf{x}$ is assumed to be known. Specifically, we seek the estimator that minimizes the worst-case *regret*, which is the difference between the MSE of the linear estimator which does not know $\mathbf{x}$, and the smallest attainable MSE with a linear estimator that knows $\mathbf{x}$. Note that as we show in Section 3, since we are restricting ourselves to linear estimators, we cannot achieve zero MSE even in the case in which the parameters $\mathbf{x}$ are known. By considering the *difference* between the MSE and the optimal MSE rather than the MSE directly, we can counterbalance the conservative character of the minimax approach, as is evident in the examples we consider in Section 6.

The minimax regret concept has recently been used to develop a linear estimator for the unknowns $\mathbf{x}$ in the same linear model considered in this paper, but for the case that $\mathbf{x}$ is random with an unknown covariance matrix [17]. Similar competitive approaches have been used in a variety of other contexts, for example, universal source coding [22], hypothesis testing [23, 24], and prediction [25].

The paper is organized as follows. In Section 2, we provide an overview of our problem. In Section 3, we develop the form of the minimax regret estimator when the uncertainty region is defined by $\mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2$ for positive definite weighting matrices $\mathbf{T}$ that commute with $\mathbf{H}\mathbf{C}_w^{-1}\mathbf{H}$, where $\mathbf{C}_w$ is the noise covariance matrix. We then specialize the results to the case in which $\mathbf{T} = \mathbf{H}\mathbf{C}_w^{-1}\mathbf{H}$ in Section 4, and to the case in which $\mathbf{T} = \mathbf{I}$ in Section 5. In these special cases, we show that the minimax regret estimator can be derived as the solution to explicit, simple, and computationally tractable convex optimization problems. Section 6 presents several examples illustrating the performance advantage of the minimax regret estimator.

2 Problem Formulation and Main Results

We denote vectors in $\mathbb{C}^m$ by boldface lowercase letters and matrices in $\mathbb{C}^{n\times m}$ by boldface uppercase letters. $\mathbf{I}$ denotes the identity matrix of appropriate dimension, $(\cdot)^*$ denotes the Hermitian conjugate of the corresponding matrix, and $\hat{(\cdot)}$ denotes an estimated vector or matrix.

Consider the problem of estimating the unknown deterministic parameters $\mathbf{x}$ in the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{H}$ is a known $n \times m$ matrix with full rank m, and $\mathbf{w}$ is a zero-mean random vector with covariance $\mathbf{C}_w$. We assume that $\mathbf{x}$ is known to satisfy the weighted norm constraint $\|\mathbf{x}\|_{\mathbf{T}} \leq L$ for some positive definite covariance $\mathbf{T}$ and scalar $L > 0$, where $\|\mathbf{x}\|_{\mathbf{T}}^2 = \mathbf{x}^*\mathbf{T}\mathbf{x}$.

We estimate $\mathbf{x}$ using a linear estimator so that $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ for some $m \times n$ matrix $\mathbf{G}$. The variance of the estimator $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ is given by

$$V(\hat{\mathbf{x}}) = E(\|\hat{\mathbf{x}} - E(\hat{\mathbf{x}})\|^2) = \mathrm{Tr}(\mathbf{G}\mathbf{C}_w\mathbf{G}^*), \tag{2}$$

and the bias of the estimator is

$$B(\hat{\mathbf{x}}) = \mathbf{x} - E(\hat{\mathbf{x}}) = (\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{x}. \tag{3}$$

We would like to design an estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ to minimize the MSE, which is given by

$$E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) = V(\hat{\mathbf{x}}) + \|B(\hat{\mathbf{x}})\|^2 = \mathrm{Tr}(\mathbf{G}\mathbf{C}_w\mathbf{G}^*) + \mathbf{x}^*(\mathbf{I} - \mathbf{G}\mathbf{H})^*(\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{x}. \tag{4}$$

Since $B(\hat{\mathbf{x}})$ depends explicitly on the unknown parameters $\mathbf{x}$, we cannot choose an estimate to directly minimize the MSE (4).

A common approach is to restrict the estimator $\hat{\mathbf{x}}$ to be unbiased, so that $B(\hat{\mathbf{x}}) = \mathbf{0}$, and then seek the estimator of this form that minimizes the variance $V(\hat{\mathbf{x}})$, or the MSE. The resulting estimator is the (weighted) least-squares estimator [4], which is given by

$$\hat{\mathbf{x}} = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}. \tag{5}$$

If $\mathbf{w}$ is a zero-mean Gaussian random vector, then the least-squares estimator is also the minimum variance unbiased estimator, *i.e.,* it minimizes the variance from all linear and nonlinear *unbiased* estimators.

The least-squares estimator has a variety of optimality properties in the class of unbiased estimators. However, an unbiased estimator does not necessarily lead to a small MSE. To improve the performance over the least-squares estimator in the case in which the model (1) is assumed to hold perfectly, various modifications of the least-squares estimator have been proposed. These modifications attempt to reduce the MSE of the least-squares estimator by allowing for a bias. Among the alternatives are Tikhonov regularization [5, 6], the shrunken estimator [7], and the covariance shaping least-squares estimator [8]. In general, these least-squares alternatives attempt to reduce the MSE in estimating $\mathbf{x}$ by allowing for a bias. However, each of the estimators above is designed to optimize an objective which depends on the data error, and not directly on the MSE.

In this paper we consider an alternative method for developing optimal linear estimators. Specifically, we

develop estimators that minimize the worst-case regret, *i.e.,* the difference between the MSE of an estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ and the best possible MSE attainable using any estimator of the form $\hat{\mathbf{x}} = \mathbf{G}(\mathbf{x})\mathbf{y}$ where $\mathbf{x}$ is assumed to be known, so that $\mathbf{G}$ can depend explicitly on $\mathbf{x}$. As we show in Section 3, since we are restricting ourselves to linear estimators of the form $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$, even in the case in which the parameters $\mathbf{x}$ are known we cannot achieve zero MSE. The best possible MSE is illustrated schematically in Fig. 1. Instead of seeking an estimator to minimize the worst-case MSE, we therefore propose seeking an estimator to minimize the worst-case difference between its MSE and the best possible MSE, as illustrated in Fig. 1.

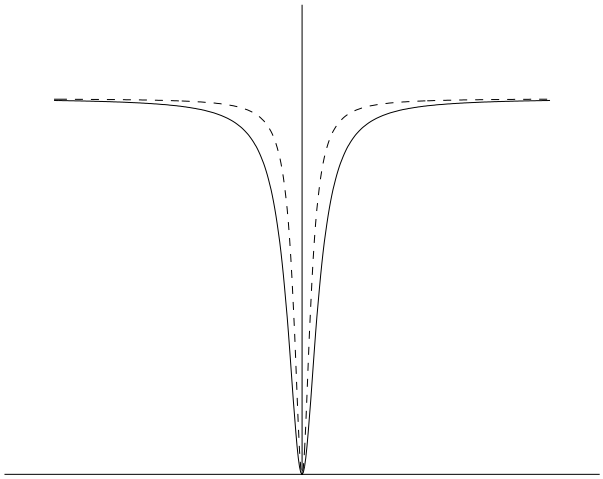


Figure 1: The line represents the best attainable MSE as a function of $\mathbf{x}$ when $\mathbf{x}$ is known, and the dashed line represents a desirable graph of MSE with small regret as a function of $\mathbf{x}$ using some linear estimator that does not depend on $\mathbf{x}$.

In Section 3 we develop the form of the estimator minimizing the worst-case regret for the case in which $\mathbf{T}$ commutes with $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$. We show that the minimax regret estimator can be described by m parameters, which are the solution to a convex optimization problem (Theorem 1). In Section 4 we consider the case in which $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$. As we show, when L is large enough with respect to m, the optimal minimax regret estimator is a shrunken estimator with a specific choice of shrinkage factor. For small values of L, the optimal estimator is given in terms of a single parameter, which is the solution to a nonlinear equation (Theorem 2). In Section 5 we consider the case in which $\mathbf{T} = \mathbf{I}$, and show that the minimax regret estimator can be determined by solving m convex optimization problems, each in 3 unknowns (Theorem 3). In Section 6 we demonstrate by examples, that the minimax regret estimator can significantly improve the performance over the traditional least-squares estimator. Furthermore, its performance is often better than that of the minimax estimator that minimizes the worst-case MSE [21], and the Wiener estimator which results from

assuming that $\mathbf{x}$ is a random vector with covariance $L^2\mathbf{I}$.

3 The Minimax Regret Estimator

The minimax regret estimator $\hat{\mathbf{x}}$ is designed to minimize the worst-case regret, where the regret $\mathcal{R}(\mathbf{x}, \mathbf{G})$ is defined as the difference between the MSE using an estimator $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ and the smallest possible MSE attainable with an estimator of the form $\hat{\mathbf{x}} = \mathbf{G}(\mathbf{x})\mathbf{y}$ when the parameters $\mathbf{x}$ are known, which we denote by MSE^o.

To develop an explicit expression for MSE^o we first determine the estimator $\hat{\mathbf{x}} = \mathbf{G}(\mathbf{x})\mathbf{y}$ that minimizes the MSE when $\mathbf{x}$ is known. To this end we differentiate[1] the MSE of (4) with respect to $\mathbf{G}$ and equate to 0, which results in

$$(\mathbf{G}(\mathbf{x})\mathbf{H} - \mathbf{I})\mathbf{x}\mathbf{x}^*\mathbf{H}^* + \mathbf{G}(\mathbf{x})\mathbf{C}_w = 0, \tag{6}$$

so that

$$\mathbf{G}(\mathbf{x}) = \mathbf{x}\mathbf{x}^*\mathbf{H}^*(\mathbf{C}_w + \mathbf{H}\mathbf{x}\mathbf{x}^*\mathbf{H}^*)^{-1}. \tag{7}$$

Using the Matrix Inversion Lemma [26] we can express $\mathbf{G}$ as

$$\mathbf{G}(\mathbf{x}) = \frac{1}{1 + \mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}}\mathbf{x}\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}. \tag{8}$$

Substituting $\mathbf{G}(\mathbf{x})$ back into (4), MSE^0 is given by

$$\text{MSE}^o = \frac{\mathbf{x}^*\mathbf{x}}{1 + \mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}}. \tag{9}$$

Since $\mathbf{x}$ is unknown, we cannot implement the optimal estimator (8). Instead we seek the estimator $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ that minimizes the worst-case regret $\mathcal{R}(\mathbf{x}, \mathbf{G})$, where

$$\mathcal{R}(\mathbf{x}, \mathbf{G}) = E(\|\mathbf{G}\mathbf{y} - \mathbf{x}\|^2) - \text{MSE}^o = \text{Tr}(\mathbf{G}\mathbf{C}_w\mathbf{G}^*) + \mathbf{x}^*(\mathbf{I} - \mathbf{G}\mathbf{H})^*(\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{x} - \frac{\mathbf{x}^*\mathbf{x}}{1 + \mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}}, \tag{10}$$

[1] We use the following derivatives: For any Hermitian $\mathbf{A}$,

$$\frac{\partial \text{Tr}(\mathbf{B}\mathbf{A}\mathbf{B}^*)}{\partial \mathbf{B}} = 2\mathbf{B}\mathbf{A},$$

and

$$\frac{\partial \mathbf{x}^*\mathbf{B}^*\mathbf{B}\mathbf{x}}{\partial \mathbf{B}} = 2\mathbf{B}\mathbf{x}\mathbf{x}^*.$$

subject to the constraint $\|\mathbf{x}\|_{\mathbf{T}} \leq L$. Thus we seek the matrix $\mathbf{G}$ that is the solution to the problem

$$\min_{\mathbf{G}} \max_{\mathbf{x}^*\mathbf{T}\mathbf{x} \leq L^2} \mathcal{R}(\mathbf{x}, \mathbf{G}). \tag{11}$$

For analytical tractability, we restrict our attention to weighting matrices $\mathbf{T}$ that commute with $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$, so that they can be jointly diagonalized. Thus, if $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$ has an eigendecomposition $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = \mathbf{V}\Sigma\mathbf{V}^*$ where $\mathbf{V}$ is a unitary matrix and Σ is a diagonal matrix, then $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^*$ for some diagonal matrix Λ.

Theorem 1 below establishes the general form of the solution to (11) for any $\mathbf{T}$ that commutes with $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$. In Sections 4 and 5 we use Theorem 1 to develop the solution for the case in which $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$ and $\mathbf{T} = \mathbf{I}$, respectively.

Theorem 1. *Let* $\mathbf{x}$ *denote the deterministic unknown parameters in the model* $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$*, where* $\mathbf{H}$ *is a known* $n \times m$ *matrix with rank* m*, and* $\mathbf{w}$ *is a zero-mean random vector with covariance* $\mathbf{C}_w$*. Let* $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = \mathbf{V}\Sigma\mathbf{V}^*$ *where* $\mathbf{V}$ *is a unitary matrix and* Σ *is an* $m \times m$ *diagonal matrix with diagonal elements* $\sigma_i > 0$ *and let* $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^*$ *where* Λ *is an* $m \times m$ *diagonal matrix with diagonal elements* $\lambda_i > 0$*. Then the solution to the problem*

$$\begin{aligned}&\min_{\hat{\mathbf{x}}=\mathbf{G}\mathbf{y}} \max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} \left\{ E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) - \min_{\hat{\mathbf{x}}=\mathbf{G}(\mathbf{x})\mathbf{y}} E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) \right\} = \\ &\quad = \min_{\mathbf{G}} \max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} \left\{ Tr(\mathbf{G}\mathbf{C}_w\mathbf{G}^*) + \mathbf{x}^*(\mathbf{I} - \mathbf{G}\mathbf{H})^*(\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{x} - \frac{\mathbf{x}^*\mathbf{x}}{1+\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}} \right\}\end{aligned}$$

has the form

$$\hat{\mathbf{x}} = \mathbf{V}\mathbf{D}\mathbf{V}^*(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y},$$

where $\mathbf{D}$ *is an* $m \times m$ *diagonal matrix with diagonal elements* d_i *which are the solution to the convex optimization problem*

$$\min_{\tau, d_i} \left\{ \tau : \begin{array}{c} \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \leq \tau \\ \max_{s_i \geq 0, \sum_i \lambda_i s_i = L^2} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \frac{\sum_{i=1}^m s_i}{1+\sum_{i=1}^m \sigma_i s_i} \right\} + \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \leq \tau \end{array} \right\}. \tag{12}$$

Proof. The proof of Theorem 1 is comprised of three parts. We first show that the optimal $\mathbf{G}$ minimizing the worst-case regret has the form

$$\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{V}^*(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}, \tag{13}$$

for some $m \times m$ matrix $\mathbf{D}$. We then show that $\mathbf{D}$ must be a diagonal matrix. Finally, we show that the diagonal elements of $\mathbf{D}$, denoted d_i, are the solution to (12).

We begin by showing that the optimal $\mathbf{G}$ has the form given by (13). To this end, note that the regret $\mathcal{R}(\mathbf{C}_x, \mathbf{G})$ of (10) depends on $\mathbf{G}$ only through $\mathbf{GH}$ and $\mathrm{Tr}(\mathbf{GC}_w\mathbf{G}^*)$. Now, for any choice of $\mathbf{G}$,

$$\mathrm{Tr}(\mathbf{GC}_w\mathbf{G}^*) = \mathrm{Tr}(\mathbf{GC}_w^{1/2}\mathbf{PC}_w^{1/2}\mathbf{G}^*) + \mathrm{Tr}(\mathbf{GC}_w^{1/2}(\mathbf{I}-\mathbf{P})\mathbf{C}_w^{1/2}\mathbf{G}^*) \geq \mathrm{Tr}(\mathbf{GC}_w^{1/2}\mathbf{PC}_w^{1/2}\mathbf{G}^*) \tag{14}$$

where

$$\mathbf{P} = \mathbf{C}_w^{-1/2}\mathbf{H}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1/2} \tag{15}$$

is the orthogonal projection onto the range space of $\mathbf{C}_w^{-1/2}\mathbf{H}$. In addition, $\mathbf{GH} = \mathbf{GC}_w^{1/2}\mathbf{PC}_w^{-1/2}\mathbf{H}$ since $\mathbf{PC}_w^{-1/2}\mathbf{H} = \mathbf{C}_w^{-1/2}\mathbf{H}$. Thus, to minimize $\mathrm{Tr}(\mathbf{GC}_w\mathbf{G}^*)$ it is sufficient to consider matrices $\mathbf{G}$ that satisfy

$$\mathbf{GC}_w^{1/2} = \mathbf{GC}_w^{1/2}\mathbf{P}. \tag{16}$$

Substituting (15) into (16), we have

$$\mathbf{G} = \mathbf{GC}_w^{1/2}\mathbf{PC}_w^{-1/2} = \mathbf{GH}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1} = \mathbf{B}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}, \tag{17}$$

for some $m \times m$ matrix $\mathbf{B}$. Denoting $\mathbf{B} = \mathbf{VDV}^*$, (17) reduces to (13).

We now show that $\mathbf{D}$ must be a diagonal matrix. Since $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = \mathbf{V}\Sigma\mathbf{V}^*$ we can express $\mathcal{R}(\mathbf{x}, \mathbf{G})$ as

$$\begin{aligned}
\mathcal{R}(\mathbf{x}, \mathbf{G}) &= \mathrm{Tr}(\mathbf{VD}^*\mathbf{DV}^*(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}) + \mathbf{x}^*(\mathbf{I}-\mathbf{VDV}^*)^*(\mathbf{I}-\mathbf{VDV}^*)\mathbf{x} - \frac{\mathbf{x}^*\mathbf{x}}{1+\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{Hx}} \\
&= \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\Sigma^{-1}) + \mathbf{x}^*\mathbf{V}(\mathbf{I}-\mathbf{D})^*(\mathbf{I}-\mathbf{D})\mathbf{V}^*\mathbf{x} - \frac{\mathbf{x}^*\mathbf{x}}{1+\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{Hx}} \\
&= \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\Sigma^{-1}) + \mathbf{z}^*(\mathbf{I}-\mathbf{D})^*(\mathbf{I}-\mathbf{D})\mathbf{z} - \frac{\mathbf{z}^*\mathbf{z}}{1+\mathbf{z}^*\Sigma\mathbf{z}},
\end{aligned} \tag{18}$$

where $\mathbf{z} = \mathbf{V}^*\mathbf{x}$. Combining (18) with

$$\mathbf{x}^*\mathbf{Tx} = \mathbf{x}^*\mathbf{V}\Lambda\mathbf{V}^*\mathbf{x} = \mathbf{z}^*\Lambda\mathbf{z}, \tag{19}$$

we conclude that the problem (11) reduces to finding $\mathbf{D}$ that minimizes

$$\mathcal{G}(\mathbf{D}) = \max_{\mathbf{z}^*\Lambda\mathbf{z}\leq L^2} \left\{ \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\Sigma^{-1}) + \mathbf{z}^*(\mathbf{I}-\mathbf{D})^*(\mathbf{I}-\mathbf{D})\mathbf{z} - \frac{\mathbf{z}^*\mathbf{z}}{1+\mathbf{z}^*\Sigma\mathbf{z}} \right\}. \tag{20}$$

Let $\mathbf{J}$ be a diagonal matrix with diagonal elements equal to ± 1. Then

$$\begin{aligned}
\mathcal{G}(\mathbf{JDJ}) &= \max_{\mathbf{z}^*\Lambda\mathbf{z}\leq L^2} \left\{ \mathrm{Tr}(\mathbf{D}^*\mathbf{DJ}\Sigma^{-1}\mathbf{J}) + \mathbf{z}^*(\mathbf{I}-\mathbf{JDJ})^*(\mathbf{I}-\mathbf{JDJ})\mathbf{z} - \frac{\mathbf{z}^*\mathbf{z}}{1+\mathbf{z}^*\Sigma\mathbf{z}} \right\} \\
&= \max_{\mathbf{z}^*\Lambda\mathbf{z}\leq L^2} \left\{ \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\Sigma^{-1}) + \mathbf{z}^*\mathbf{J}(\mathbf{I}-\mathbf{D})^*(\mathbf{I}-\mathbf{D})\mathbf{Jz} - \frac{\mathbf{z}^*\mathbf{z}}{1+\mathbf{z}^*\Sigma\mathbf{z}} \right\} \\
&= \max_{\mathbf{z}'^*\Lambda\mathbf{z}'\leq L^2} \left\{ \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\Sigma^{-1}) + \mathbf{z}'^*(\mathbf{I}-\mathbf{D})^*(\mathbf{I}-\mathbf{D})\mathbf{z}' - \frac{\mathbf{z}'^*\mathbf{z}'}{1+\mathbf{z}'^*\Sigma\mathbf{z}'} \right\} \\
&= \mathcal{G}(\mathbf{D}),
\end{aligned} \tag{21}$$

where $\mathbf{z}' = \mathbf{Jz}$ and we used the fact that $\mathbf{J}^2 = \mathbf{I}$ and for any diagonal matrix $\mathbf{M}$, $\mathbf{JMJ} = \mathbf{M}$. Therefore, if $\mathbf{D}$ minimizes $\mathcal{G}(\mathbf{D})$, then $\mathbf{JDJ}$ also minimizes $\mathcal{G}(\mathbf{D})$. Now, since the problem of minimizing $\mathcal{G}(\mathbf{D})$ is convex, the set of optimal solutions is also convex [27], which implies that if $\mathbf{JDJ}$ is optimal for any diagonal $\mathbf{J}$ with diagonal elements ± 1, then so is $\mathbf{D}' = (1/2^m)\sum_{\mathbf{J}} \mathbf{JDJ}$ where the summation is over all 2^m diagonal matrices $\mathbf{J}$ with diagonal elements ± 1. It is easy to see that $\mathbf{D}'$ has diagonal elements. Therefore, we have shown that there exists an optimal diagonal solution $\mathbf{D}$.

Denote by d_i the diagonal elements of $\mathbf{D}$, and further denote $s_i = |z_i|^2$ where z_i are the components of $\mathbf{z}$. Then we can express $\mathcal{G}(\mathbf{D})$ as

$$\begin{aligned}
\mathcal{G}(\mathbf{D}) &= \max_{s_i\geq 0, \sum_{i=1}^m \lambda_i s_i \leq L^2} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \frac{\sum_{i=1}^m s_i}{1+\sum_{i=1}^m \sigma_i s_i} \right\} + \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \\
&= \max_{\mathbf{s}\in\mathcal{I}} \Phi(\mathbf{s}) + \sum_{i=1}^m \frac{d_i^2}{\sigma_i},
\end{aligned} \tag{22}$$

where

$$\Phi(\mathbf{s}) = \sum_{i=1}^m (1-d_i)^2 s_i - \frac{\sum_{i=1}^m s_i}{1+\sum_{i=1}^m \sigma_i s_i}, \tag{23}$$

and $\mathcal{I}$ is the set of vectors $\mathbf{s} \in \mathcal{R}^m$ with components s_i such that $s_i \geq 0$ and $\sum_{i=1}^m \lambda_i s_i \leq L^2$, *i.e.,*

$$\mathcal{I} \stackrel{\triangle}{=} \left\{ \mathbf{s} \in \mathcal{R}^m \mid s_i \geq 0, \sum_{i=1}^m \lambda_i s_i \leq L^2 \right\}. \tag{24}$$

To complete the proof of Theorem 1 we rely on the following lemma.

Lemma 1. *Let*

$$\Phi(\mathbf{s}) = \sum_{i=1}^{m}(1-d_i)^2 s_i - \frac{\sum_{i=1}^{m} s_i}{1+\sum_{i=1}^{m}\sigma_i s_i}$$

for some given $\sigma_i, 1 \le i \le m$ *and* $d_i, 1 \le i \le m$. *If*

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}\in\mathcal{I}} \Phi(\mathbf{s}),$$

where $\mathcal{I}$ *is defined by (24), then* $\mathbf{s} = \mathbf{0}$ *or* $\sum_{i=1}^{m}\lambda_i s_i = L^2$.

Proof. Let $\mathcal{S}$ be the set of vectors $\mathbf{s}$ such that $\mathbf{s} = \mathbf{0}$ or $\sum_{i=1}^{m}\lambda_i s_i = L^2$. To establish the lemma we need to show that for any $\mathbf{s} \in \mathcal{I}$, $\Phi(\mathbf{s}) \le \Phi(\mathbf{s}')$ for some $\mathbf{s}' \in \mathcal{S}$.

Fix $\mathbf{s} \in \mathcal{I}$ such that $\mathbf{s} \neq \mathbf{0}$ and let $h(r) = \Phi(r\mathbf{s})$ be defined on the segment $[0, r_*]$, where r_* is the largest value of r for which $r\mathbf{s} \in \mathcal{I}$. Clearly, $r_*\mathbf{s} \in \mathcal{S}$ and $r_* \ge 1$. Since $h(1) = \Phi(\mathbf{s})$, $h(0) = \Phi(\mathbf{0})$ and $h(r_*) = \Phi(\mathbf{s}')$ where $\mathbf{s}' = r_*\mathbf{s} \in \mathcal{S}$, to prove that $\Phi(\mathbf{s}) \le \Phi(\mathbf{s}')$ for some $\mathbf{s}' \in \mathcal{S}$ it suffices to show that

$$h(1) \le \max(h(0), h(r_*)). \tag{25}$$

We now establish (25) by first showing that $h(r)$ is convex. It then follows that $h(r)$ obtains its maximum at one of its end points. Since $h(r)$ is defined on $[0, r_*]$ this implies that $h(r) \le \max(h(0), h(r_*))$ for any $r \in [0, r_*]$, and in particular for $r = 1$ which established (25). It remains to show that $h(r)$ is convex. Writing

$$h(r) = \alpha r - \frac{\beta r}{1+\gamma r}, \tag{26}$$

where $\alpha = \sum_{i=1}^{m}(1-d_i)^2 s_i \ge 0$, $\beta = \sum_{i=1}^{m} s_i > 0$ and $\gamma = \sum_{i=1}^{m}\sigma_i s_i > 0$, we can express $h(r)$ as

$$h(r) = \alpha r - \frac{\beta}{\gamma}\frac{1+\gamma r - 1}{1+\gamma r} = \alpha r + \frac{\beta}{\gamma}\frac{1}{1+\gamma r} - \frac{\beta}{\gamma}. \tag{27}$$

Since $1/r$ is convex in r, $h(r)$ is convex. □

From (22) and Lemma 1 it follows that finding $\mathbf{D}$ to minimize $\mathcal{G}(\mathbf{D})$ is equivalent to the problem of finding d_i to minimize

$$\max\left(\Phi(\mathbf{0}), \max_{s_i\ge 0, \sum_{i=1}^{m}\lambda_i s_i = L^2} \Phi(\mathbf{s})\right) + \sum_{i=1}^{m}\frac{d_i^2}{\sigma_i}. \tag{28}$$

Since $\Phi(\mathbf{0}) = 0$, this problem can be written as

$$\min_{\tau,d_i} \left\{ \tau : \begin{array}{c} \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \leq \tau \\ \max_{s_i \geq 0, \sum_i \lambda_i s_i = L^2} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \frac{\sum_{i=1}^m s_i}{1+\sum_{i=1}^m \sigma_i s_i} \right\} + \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \leq \tau \end{array} \right\}, \tag{29}$$

completing the proof of Theorem 1. □

Theorem 1 reduces the problem of minimizing the regret to the simpler optimization problem (12). As we show in Sections 4 and 5, for certain choices of $\mathbf{T}$, the problem can be further simplified, and in some cases a closed form solution for the minimax regret estimator exists. In Section 4 we consider the case in which the weighting $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w\mathbf{H}$, and in Section 5 we consider the case in which $\mathbf{T} = \mathbf{I}$. As we show, when L is large enough and $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w\mathbf{H}$, the minimax regret estimator of Theorem 1 reduces to a shrunken estimator with a shrinkage factor that depends only on L. For small values of L, the minimax regret estimator is a function of a single parameter, that is the solution to a nonlinear equation. In the case in which $\mathbf{T} = \mathbf{I}$ the minimax regret estimator depends on 3 parameters, which can be found by solving m convex optimization problems in 3 unknowns.

4 Minimax Regret Estimator With $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$

We now consider the case in which the weighting $\mathbf{T}$ is given by $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$, so that the eigenvalues λ_i of $\mathbf{T}$ are equal to σ_i. As we show, for large enough values of L with respect to m, the estimator minimizing the worst-case regret is a shrunken estimator with a shrinkage factor that depends only on the bound L.

From Theorem 1, the optimal $\mathbf{G}$ that minimizes the regret with $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$ is given by (13), where the diagonal elements d_i of $\mathbf{D}$ are the solution to the problem (Δ) given by

$$(\Delta): \min_{\tau,d_i} \left\{ \tau : \begin{array}{c} \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \leq \tau \\ \max_{s_i \geq 0, \sum_{i=1}^m \sigma_i s_i = L^2} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \frac{\sum_{i=1}^m s_i}{1+L^2} \right\} + \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \leq \tau \end{array} \right\}. \tag{30}$$

To develop a solution to (Δ), define the set $\mathcal{P}$ as

$$\mathcal{P} \stackrel{\triangle}{=} \left\{ \mathbf{s} \in \mathcal{R}^m \mid s_i \geq 0, \sum_{i=1}^m \sigma_i s_i = L^2 \right\}. \tag{31}$$

Then

$$\max_{\mathbf{s}\in\mathcal{P}}\left\{\sum_{i=1}^{m}(1-d_i)^2 s_i - \frac{\sum_{i=1}^{m} s_i}{1+L^2}\right\},$$

is a linear program[2] [28]. From linear programming duality theory it follows that

$$\max_{\mathbf{s}\in\mathcal{P}}\left\{\sum_{i=1}^{m}\left((1-d_i)^2 - \frac{1}{1+L^2}\right)s_i\right\} = \min_{y\in\mathcal{D}} L^2 y \tag{32}$$

where $\mathcal{D}$ is the set of scalars y for which

$$y \geq \frac{1}{\sigma_i}\left((1-d_i)^2 - \frac{1}{1+L^2}\right), \quad 1\leq i\leq m. \tag{33}$$

Thus the problem (Δ) can be written as

$$(\Gamma):\ \min_{\tau,d_i,y}\left\{\tau:\ \begin{array}{c}\sum_{i=1}^{m}\frac{d_i^2}{\sigma_i}\leq\tau\\ L^2y+\sum_{i=1}^{m}\frac{d_i^2}{\sigma_i}\leq\tau\\ \frac{1}{\sigma_i}\left((1-d_i)^2-\frac{1}{1+L^2}\right)\leq y,\quad 1\leq i\leq m\end{array}\right\}. \tag{34}$$

Since (Γ) is a convex optimization problem, from Lagrange duality theory [29] it follows that $A=\min\tau$ in the problem (Γ) is equal to the optimal value of the dual problem, namely,

$$A = \max_{\alpha,\beta,\gamma_i\geq 0}\min_{\tau,d_i,y}\mathcal{L}(\tau,d_i,y), \tag{35}$$

where the Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L}(\tau,d_i,y) = \tau+\alpha\left(\sum_{i=1}^{m}\frac{d_i^2}{\sigma_i}-\tau\right)+\beta\left(L^2y+\sum_{i=1}^{m}\frac{d_i^2}{\sigma_i}-\tau\right)+\sum_{i=1}^{m}\gamma_i\left[(1-d_i)^2-\frac{1}{1+L^2}-\sigma_i y\right]. \tag{36}$$

Substituting (36) into (35), we have

$$A = \\ = \max_{\alpha,\beta,\gamma_i\geq 0}\left\{-\frac{1}{1+L^2}\sum_{i=1}^{m}\gamma_i+\min_{\tau}\left\{(1-\alpha-\beta)\tau\right\}+\min_{y}\left\{\left(\beta L^2-\sum_{i=1}^{m}\sigma_i\gamma_i\right)y\right\}+\sum_{i=1}^{m}\min_{d_i}\left\{\frac{\alpha+\beta}{\sigma_i}d_i^2+\gamma_i(1-d_i)^2\right\}\right\}$$

[2] A linear program is a problem of the form $\min_{\mathbf{x}\in\mathcal{P}}\mathbf{c}^*\mathbf{x}$ for some given vector $\mathbf{c}$ where $\mathcal{P}$ is the set of vectors $\mathbf{x}$ satisfying $\mathbf{A}\mathbf{x}=\mathbf{b}$ for some given matrices $\mathbf{A}$ and $\mathbf{b}$ and $\mathbf{x}\geq 0$, where the inequality is to be understood as a component-wise inequality. From linear programming duality theory, $\min_{\mathbf{x}\in\mathcal{P}}\mathbf{c}^*\mathbf{x}=\max_{\mathbf{y}\in\mathcal{D}}\mathbf{b}^*\mathbf{y}$ where $\mathcal{D}$ is the set of vectors $\mathbf{y}$ for which $\mathbf{A}^*\mathbf{y}\leq\mathbf{c}$.

$$= \max_{\alpha,\beta,\gamma_i \geq 0} \left\{ -\frac{1}{1+L^2} \sum_{i=1}^{m} \gamma_i + \sum_{i=1}^{m} \frac{\gamma_i}{1+\sigma_i \gamma_i} : \alpha + \beta = 1, \beta L^2 = \sum_{i=1}^{m} \sigma_i \gamma_i \right\}$$
$$= \max_{\gamma_i \geq 0} \left\{ -\frac{1}{1+L^2} \sum_{i=1}^{m} \gamma_i + \sum_{i=1}^{m} \frac{\gamma_i}{1+\sigma_i \gamma_i} : \sum_{i=1}^{m} \sigma_i \gamma_i \leq L^2 \right\}, \tag{37}$$

where we used the fact that the optimal d_i are given by

$$d_i = \frac{\sigma_i \gamma_i}{1+\sigma_i \gamma_i}, \quad 1 \leq i \leq m. \tag{38}$$

The dual problem of (Γ) is therefore the problem

$$\max_{\gamma_i} \left\{ -\frac{1}{1+L^2} \sum_{i=1}^{m} \gamma_i + \sum_{i=1}^{m} \frac{\gamma_i}{1+\sigma_i \gamma_i} \right\} \tag{39}$$

subject to

$$\begin{aligned} &\textstyle\sum_{i=1}^{m} \sigma_i \gamma_i \leq L^2; \\ &\gamma_i \geq 0, \quad 1 \leq i \leq m. \end{aligned} \tag{40}$$

Once we find the dual optimal values γ_i, the optimal values d_i can be calculated using (38).

Since the problem of (39) subject to (40) is a convex optimization problem, we can find an optimal solution by forming the Lagrangian

$$\mathcal{L} = \frac{1}{1+L^2} \sum_{i=1}^{m} \gamma_i - \sum_{i=1}^{m} \frac{\gamma_i}{1+\sigma_i \gamma_i} + \rho \sum_{i=1}^{m} \sigma_i \gamma_i - \sum_{i=1}^{m} \zeta_i \gamma_i, \tag{41}$$

where from the Karush-Kuhn-Tucker conditions [28] we must have that $\rho, \zeta_i \geq 0$. The values γ_i are optimal if and only if they satisfy (40) and there exist $\rho, \zeta_i \geq 0$ such that

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \frac{1}{1+L^2} - \frac{1}{(\gamma_i \sigma_i + 1)^2} + \rho \sigma_i - \zeta_i = 0, \tag{42}$$

and the complementary slackness conditions are satisfied, namely,

$$\begin{aligned} &\rho \left(\textstyle\sum_{i=1}^{m} \sigma_i \gamma_i - L^2 \right) = 0; \\ &\zeta_i \gamma_i = 0. \end{aligned} \tag{43}$$

Suppose first that $\rho = 0$. If $\gamma_i = 0$ then from (42), $\zeta_i = -L^2/(1+L^2) < 0$, which contradicts the condition $\zeta_i \geq 0$. Therefore we must have that $\gamma_i > 0$, which implies from (43) that $\zeta_i = 0$. Substituting $\rho = \zeta_i = 0$ into (42),

$$\frac{1}{1+L^2} = \frac{1}{(\gamma_i \sigma_i + 1)^2}, \tag{44}$$

or,

$$\gamma_i = \frac{1}{\sigma_i}\left(\sqrt{1+L^2} - 1\right), \quad 1 \leq i \leq m. \tag{45}$$

From (38) we then have

$$d_i = 1 - \frac{1}{\sqrt{1+L^2}}, \quad 1 \leq i \leq m, \tag{46}$$

so that from Theorem 1 the optimal estimator in this case is

$$\hat{\mathbf{x}} = \left(1 - \frac{1}{\sqrt{1+L^2}}\right)(\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{y}. \tag{47}$$

To satisfy (40) we must have that

$$m\left(\sqrt{1+L^2} - 1\right) \leq L^2, \tag{48}$$

which is equivalent to

$$L^2 \geq (m-1)^2 - 1. \tag{49}$$

Next suppose that $\rho > 0$. Then the conditions (43), (40) and (42) become

$$\begin{gathered} \textstyle\sum_{i=1}^m \sigma_i \gamma_i = L^2; \\ \zeta_i \gamma_i = 0, \quad 1 \leq i \leq m; \\ \gamma_i \geq 0; \\ -\tfrac{1}{(1+\sigma_i \gamma_i)^2} + \tfrac{1}{1+L^2} + \sigma_i \rho - \zeta_i = 0. \end{gathered} \tag{50}$$

Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m > 0$. If $\gamma_j = 0$ for some $1 \leq j \leq m$, then $\gamma_i = 0, i \geq j$. For suppose that $\gamma_i > 0$ for some $i \geq j$. Then, from (50), $\zeta_i = 0$ and

$$\sigma_i \rho = \frac{1}{(1+\sigma_i \gamma_i)^2} - \frac{1}{1+L^2} < 1 - \frac{1}{1+L^2}. \tag{51}$$

On the other hand, since $\gamma_j = 0$, $\zeta_j \geq 0$ and from (50),

$$\sigma_j \rho \geq 1 - \frac{1}{1+L^2}, \tag{52}$$

which contradicts (51) because $\sigma_j \leq \sigma_i$. Thus we conclude that there exists a k such that $\gamma_i = 0$ for $i \leq k$, and $\gamma_i > 0$ for $i > k$.

Since $\gamma_i > 0$ for $i > k$, from (50), $\zeta_i = 0$ for $i > k$, and

$$\gamma_i = \frac{1}{\sigma_i}\left(\frac{1}{\sqrt{\rho\sigma_i + 1/(L^2+1)}} - 1\right), \quad i > k, \tag{53}$$

where ρ is chosen such that

$$\sum_{i=k+1}^{m} \sigma_i \gamma_i = L^2. \tag{54}$$

Note that if $\zeta_i = 0$, then from (51), $\rho\sigma_i + 1/(L^2+1) < 1$, so that γ_i defined by (53) satisfies $\gamma_i > 0$.

Define

$$\mathcal{G}(\rho, k) = \sum_{i=k+1}^{m} \sigma_i \gamma_i - L^2 = \sum_{i=k+1}^{m} \left(\frac{1}{\sqrt{\rho\sigma_i + 1/(L^2+1)}} - 1\right) - L^2. \tag{55}$$

It can be easily seen that $\mathcal{G}(\rho, k)$ is monotonically decreasing in k and ρ. In addition, $\mathcal{G}(\rho, k) \to -\infty$ as $\rho \to \infty$. Therefore, $\mathcal{G}(\rho, k) = 0$ for some ρ and k if and only if $\mathcal{G}(0, 0) > 0$, *i.e.,* if and only if

$$\mathcal{G}(0, 0) = m\left(\sqrt{1+L^2} - 1\right) > L^2. \tag{56}$$

Now, since $\gamma_i > 0$ for $i \geq k+1$, we have from (51) that

$$\rho < \frac{1}{\sigma_{k+1}} \frac{L^2}{L^2+1} \stackrel{\triangle}{=} \eta_{k+1}. \tag{57}$$

Similarly, since $\gamma_i = 0$ for $i \leq k$, we have from (52) that

$$\rho \geq \frac{1}{\sigma_k} \frac{L^2}{L^2+1} \stackrel{\triangle}{=} \eta_k. \tag{58}$$

Therefore, there exists a k such that $\mathcal{G}(\eta_k, k) \geq 0$ and $\mathcal{G}(\eta_{k+1}, k) < 0$. The optimal value of ρ is then given

by $\mathcal{G}(\rho,k)=0$, and from (53) and (38),

$$d_i=\begin{cases}0, & i\le k;\\ 1-\sqrt{\rho\sigma_i+1/(L^2+1)}, & i\ge k+1.\end{cases} \tag{59}$$

We summarize our results in the following theorem.

Theorem 2. *Let* $\mathbf{x}$ *denote the deterministic unknown parameters in the model* $\mathbf{y}=\mathbf{H}\mathbf{x}+\mathbf{w}$*, where* $\mathbf{H}$ *is a known* $n\times m$ *matrix with rank* m*, and* $\mathbf{w}$ *is a zero-mean random vector with covariance* $\mathbf{C}_w$*. Let* $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}=\mathbf{V}\Sigma\mathbf{V}^*$ *where* $\mathbf{V}$ *is a unitary matrix and* Σ *is an* $m\times m$ *diagonal matrix with diagonal elements* $\sigma_1\ge\ldots\ge\sigma_m>0$*. Then the solution to the problem*

$$\begin{aligned}&\min_{\hat{\mathbf{x}}=\mathbf{G}\mathbf{y}}\max_{\|\mathbf{x}\|_{\mathbf{T}}\le L}\left\{E(\|\hat{\mathbf{x}}-\mathbf{x}\|^2)-\min_{\hat{\mathbf{x}}=\mathbf{G}(\mathbf{x})\mathbf{y}}E(\|\hat{\mathbf{x}}-\mathbf{x}\|^2)\right\}=\\ &\quad=\min_{\mathbf{G}}\max_{\|\mathbf{x}\|_{\mathbf{T}}\le L}\left(Tr(\mathbf{G}\mathbf{C}_w\mathbf{G}^*)+\mathbf{x}^*(\mathbf{I}-\mathbf{G}\mathbf{H})^*(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{x}-\frac{\mathbf{x}^*\mathbf{x}}{1+\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}}\right)\end{aligned}$$

with $\mathbf{T}=\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$ *is given by*

$$\hat{\mathbf{x}}=\begin{cases}\left(1-\sqrt{\frac{1}{1+L^2}}\right)(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}, & L^2\ge(m-1)^2-1;\\ \mathbf{V}\mathbf{D}\mathbf{V}^*(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}, & L^2<(m-1)^2-1,\end{cases}$$

where $\mathbf{D}$ *is an* $m\times m$ *diagonal matrix with diagonal elements* d_i *that are given by*

$$d_i=\begin{cases}0, & i\le k;\\ 1-\sqrt{\rho\sigma_i+1/(L^2+1)}, & i\ge k+1.\end{cases}$$

Here k *is the unique value satisfying* $0\le k\le m-1$ *such that* $\mathcal{G}(\eta_k,k)\ge 0$ *and* $\mathcal{G}(\eta_{k+1},k)<0$ *with*

$$\eta_k=\frac{1}{\sigma_k}\frac{L^2}{L^2+1},$$

and $\mathcal{G}(\rho,k)$ *defined by (55), and* ρ *is the unique zero of* $\mathcal{G}(\rho,k)$ *in the interval* $[\eta_k,\eta_{k+1})$*.*

The minimax regret estimator of Theorem 2 for the case in which $L^2\ge(m-1)^2-1$ is a shrunken estimator proposed by Mayer and Willke [7], which is simply a scaled version of the least-squares estimator, with an optimal choice of shrinkage factor. We therefore conclude that this particular shrunken estimator has a strong optimality property: among all linear estimators of $\mathbf{x}$ such that $\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}\le L$, it minimizes

the worst-case regret.

As we expect intuitively, when $L \to \infty$, the minimax regret estimator $\hat{\mathbf{x}}$ of Theorem 2 reduces to the least-squares estimator (5). Indeed, when the weighted norm of $\mathbf{x}$ can be made arbitrarily large, the MSE, and therefore the regret, will also be arbitrarily large unless the bias is equal to zero. Therefore, in this limit, the worst-case regret is minimized by choosing an estimator with zero bias that minimizes the variance, which leads to the least-squares estimator.

5 Minimax Regret Estimator With $\mathbf{T} = \mathbf{I}$

We now consider the case in which the weighting matrix $\mathbf{T} = \mathbf{I}$. In this case, it follows from Theorem 1 that the optimal $\mathbf{G}$ that minimizes the worst-case regret has the form given by (13), where $\mathbf{D}$ is a diagonal matrix with diagonal elements d_i which are solution to the problem (Δ), defined as

$$(\Delta): \min_{\tau,d_i} \left\{ \tau : \begin{array}{c} \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \le \tau \\ \max_{s_i \ge 0, \sum_{i=1}^m s_i = L^2} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \frac{L^2}{1+\sum_{i=1}^m \sigma_i s_i} \right\} + \sum_{i=1}^m \frac{d_i^2}{\sigma_i} \le \tau \end{array} \right\}. \tag{60}$$

Here $\sigma_i > 0$ are the eigenvalues of $\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}$.

To develop a solution to (Δ), we note that

$$A = \max_{s_i \ge 0, \sum_{i=1}^m s_i = L^2} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \frac{L^2}{1+\sum_{i=1}^m \sigma_i s_i} \right\} = \max_{\rho, s_i \in \mathcal{P}} \left\{ \sum_{i=1}^m (1-d_i)^2 s_i - \rho \right\}, \tag{61}$$

where $\mathcal{P}$ is the set defined by

$$\mathcal{P} \stackrel{\triangle}{=} \left\{ \mathbf{s} \in \mathcal{R}^m, \rho \in \mathcal{R} \mid s_i \ge 0, \sum_{i=1}^m s_i = L^2, \rho \ge \frac{L^2}{1+\sum_{i=1}^m \sigma_i s_i} \right\}, \tag{62}$$

or, equivalently,

$$\mathcal{P} \stackrel{\triangle}{=} \left\{ \mathbf{s} \in \mathcal{R}^m, \rho \in \mathcal{R} \mid s_i \ge 0, \sum_{i=1}^m s_i = L^2, \rho \ge 0, 1 + \sum_{i=1}^m \sigma_i s_i - \frac{L^2}{\rho} \ge 0 \right\}. \tag{63}$$

Since $\mathcal{P}$ is a convex set, and the objective in (61) is linear, (61) is a convex optimization problem. From Lagrange duality theory [29] it then follows that A is equal to the optimal value of the dual problem, namely,

$$A = \min_{\mu \ge 0, \lambda} \max_{s_i \ge 0, \rho \ge 0} \mathcal{L}(s_i, \rho, \lambda, \mu), \tag{64}$$

where the Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L}(s_i, \rho, \lambda, \mu) = \sum_{i=1}^{m} (1-d_i)^2 s_i - \rho + \lambda \left(L^2 - \sum_{i=1}^{m} s_i \right) + \mu \left(1 + \sum_{i=1}^{m} \sigma_i s_i - \frac{L^2}{\rho} \right). \tag{65}$$

Substituting (65) into (64), we have

$$\begin{aligned} A &= \min_{\mu \geq 0, \lambda} \left\{ \lambda L^2 + \mu + \sum_{i=1}^{m} \max_{s_i \geq 0} \left\{ \left[(1-d_i)^2 - \lambda + \mu\sigma_i \right] s_i \right\} + \max_{\rho \geq 0} \left\{ -\rho - \frac{\mu L^2}{\rho} \right\} \right\} \\ &= \min_{\mu \geq 0, \lambda} \left\{ \lambda L^2 + \mu - 2L\sqrt{\mu} : (1-d_i)^2 + \mu\sigma_i \leq \lambda, \quad 1 \leq i \leq m \right\}, \end{aligned} \tag{66}$$

where we used the fact that the optimal ρ is $\rho = L\sqrt{\mu}$. The problem (Δ) of (60) can therefore be expressed as

$$(\Gamma): \min_{\tau, d_i, \mu, \lambda} \left\{ \tau : \begin{array}{c} \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} \leq \tau \\ -2L\sqrt{\mu} + \mu + L^2\lambda + \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} \leq \tau \\ (1-d_i)^2 + \sigma_i\mu \leq \lambda, \quad 1 \leq i \leq m \\ \mu \geq 0 \end{array} \right\}. \tag{67}$$

Since (Γ) is a convex optimization problem, we can find an optimal solution to Γ by forming the Lagrangian

$$\mathcal{L} = \tau + \alpha \left(\sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} - \tau \right) + \beta \left(-2L\sqrt{\mu} + \mu + L^2\lambda + \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} - \tau \right) + \sum_{i=1}^{m} \gamma_i \left((1-d_i)^2 + \sigma_i\mu - \lambda \right), \tag{68}$$

where from the Karush-Kuhn-Tucker conditions [28] we must have that $\alpha, \beta, \gamma_i \geq 0$. Differentiating $\mathcal{L}$ with respect to τ and equating to 0,

$$\alpha + \beta = 1. \tag{69}$$

Differentiating $\mathcal{L}$ with respect to d_i and equating to 0,

$$d_i = \frac{\gamma_i}{(1/\sigma_i)(\alpha+\beta) + \gamma_i} = \frac{\gamma_i}{1/\sigma_i + \gamma_i}, \tag{70}$$

from which we conclude that

$$0 \leq d_i \leq 1, \quad 1 \leq i \leq m. \tag{71}$$

From (67), d_i must satisfy

$$(1-d_i)^2 + \sigma_i\mu \leq \lambda. \tag{72}$$

Suppose that we have equality in (72) for some $1 \leq j \leq m$. Then to satisfy (71), we must have that

$$d_j = 1 - \sqrt{\lambda - \sigma_j \mu}, \tag{73}$$

and

$$\lambda - \sigma_i \mu \leq 1. \tag{74}$$

If for some j we have inequality in (72), so that

$$(1 - d_j)^2 + \sigma_j \mu < \lambda, \tag{75}$$

then by complementary slackness we must have that $\gamma_i = 0$, which from (70) implies that $d_j = 0$.

Let $\sigma_1 \geq \ldots \geq \sigma_m > 0$. Then by (70) we have that $d_1 \geq \ldots \geq d_m \geq 0$. Therefore, if $d_j = 0$ for some j, then $d_i = 0$ for all $i \geq j$.

It follows that at an optimal solution, there exists a $1 \leq k \leq m$ such that

$$\begin{gathered} d_i = 1 - \sqrt{\lambda - \sigma_i \mu}, \quad i \leq k; \\ \lambda \leq 1 + \sigma_k \mu; \\ d_i = 0, \quad i \geq k+1; \\ \text{if } k < m \text{ then } \lambda \geq 1 + \sigma_{k+1} \mu. \end{gathered} \tag{76}$$

We conclude that (Γ) of (67) can be solved by first solving m problems (Γ_k) with 3 unknowns each, where

$$(\Gamma_k): \min_{\tau,\mu,\lambda} \left\{ \tau : \begin{array}{c} \sum_{i=1}^{k} \frac{\left(1-\sqrt{\lambda-\sigma_i\mu}\right)^2}{\sigma_i} \leq \tau \\ -2L\sqrt{\mu} + \mu + L^2\lambda + \sum_{i=1}^{k} \frac{\left(1-\sqrt{\lambda-\sigma_i\mu}\right)^2}{\sigma_i} \leq \tau \\ \sigma_1 \mu \leq \lambda \leq 1 + \sigma_k \mu \\ \mu \geq 0 \\ \text{if } k < m \text{ then } \lambda \geq 1 + \sigma_{k+1}\mu \end{array} \right\}, \quad 1 \leq k \leq m, \tag{77}$$

and then choosing the value of k and the corresponding optimal values d_i given by (76), that result in the smallest possible value of τ.

Each problem (Γ_k) is a simple convex optimization problem involving 3 unknowns, and can therefore be solved very efficiently, for example, using the Ellipsoidal method (see, *e.g.,* [28, Ch. 5.2]).

Following similar steps to those taken in (37)–(40), we can derive the dual problem of (Γ) given by (67), which results in

$$\max_{\gamma_i,\beta}\left\{\sum_{i=1}^{m}\frac{\gamma_i}{1+\sigma_i\gamma_i}-\frac{L^2\beta^2}{\beta+\sum_{i=1}^{m}\sigma_i\gamma_i}\right\} \tag{78}$$

subject to

$$\begin{aligned}\textstyle\sum_{i=1}^{m}\gamma_i &= \beta L^2;\\ 0\le\beta&\le 1;\\ \gamma_i\ge 0,\quad 1\le i&\le m.\end{aligned} \tag{79}$$

Suppose now that $\sigma_i = 1, 1 \le i \le m$ so that $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = \mathbf{I}$. In this case the objective (78) can be written as

$$\max_{\gamma_i}\left\{\sum_{i=1}^{m}\frac{\gamma_i}{1+\gamma_i}-\frac{1}{1+L^2}\sum_{i=1}^{m}\gamma_i\right\} \tag{80}$$

where we used the fact that from (79), $\sum_{i=1}^{m}\gamma_i = \beta L^2$. Since β no longer appears in the objective, the constraints (79) can be expressed as

$$\begin{aligned}\textstyle\sum_{i=1}^{m}\gamma_i &\le L^2;\\ \gamma_i\ge 0,\quad 1\le i&\le m.\end{aligned} \tag{81}$$

As we expect, the resulting dual problem of (80) and (81) is equivalent to the dual problem of (39) and (40) derived in Section 4 for the case in which $\mathbf{T} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$, when substituting $\sigma_i = 1$. Indeed, if $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = \mathbf{I}$, then the weighting matrices considered in Section 4 and 5 are equal, so that the corresponding optimization problems must coincide.

We summarize our results in the following theorem.

Theorem 3. *Let* $\mathbf{x}$ *denote the deterministic unknown parameters in the model* $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$*, where* $\mathbf{H}$ *is a known* $n \times m$ *matrix with rank* m*, and* $\mathbf{w}$ *is a zero-mean random vector with covariance* $\mathbf{C}_w$*. Let* $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = \mathbf{V}\Sigma\mathbf{V}^*$ *where* $\mathbf{V}$ *is a unitary matrix and* Σ *is an* $m \times m$ *diagonal matrix with diagonal elements* $\sigma_1 \geq \ldots \geq \sigma_m > 0$*. Then the solution to the problem*

$$\begin{aligned}&\min_{\hat{\mathbf{x}}=\mathbf{G}\mathbf{y}} \max_{\|\mathbf{x}\|\leq L} \left\{ E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) - \min_{\hat{\mathbf{x}}=\mathbf{G}(\mathbf{x})\mathbf{y}} E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) \right\} = \\ &\quad = \min_{\mathbf{G}} \max_{\|\mathbf{x}\|\leq L} \left\{ Tr(\mathbf{G}\mathbf{C}_w\mathbf{G}^*) + \mathbf{x}^*(\mathbf{I} - \mathbf{G}\mathbf{H})^*(\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{x} - \frac{\mathbf{x}^*\mathbf{x}}{1+\mathbf{x}^*\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{x}} \right\}\end{aligned}$$

has the form

$$\hat{\mathbf{x}} = \mathbf{V}\mathbf{D}\mathbf{V}^*(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y},$$

where $\mathbf{D}$ *is an* $m \times m$ *diagonal matrix with diagonal elements* d_i *that are given by*

$$d_i = \begin{cases} 1 - \sqrt{\lambda - \sigma_i\mu}, & i \leq k; \\ 0, & i \geq k+1, \end{cases}$$

with $k = \arg\min \tau_i, \mu = \mu_k$ *and* $\lambda = \lambda_k$*. Here* τ_i, μ_i *and* λ_i *are the optimal solutions to the problem* (Γ_i) *given by (77).*

We now show that, as we expect intuitively, when $L \to \infty$, the minimax regret estimator $\hat{\mathbf{x}}$ of Theorem 3 reduces to the least-squares estimator (5). From (77) it follows that if $L \to \infty$ and $\lambda > 0$, then $-2L\sqrt{\mu} + \mu + L^2\lambda + \sum_{i=1}^{k} \left(1 - \sqrt{\lambda - \sigma_i\mu}\right)^2 / \sigma_i \to L^2\lambda$ which implies that $\tau \to \infty$. Therefore, to minimize τ we must have that $\lambda = 0$, which immediately implies that $\mu = 0$ since we must have that $\lambda \geq \sigma_1\mu$ and $\mu \geq 0$. In addition, since for $k < m$, $\lambda \geq 1 + \sigma_{k+1}\mu$, we must have that $k = m$. We conclude that for $L \to \infty$, $\lambda = \mu = 0$ and $k = m$, which from Theorem 3 implies that $\mathbf{D} = \mathbf{I}$, and $\hat{\mathbf{x}}$ reduces to the least-squares estimator.

6 Examples

We now present some examples, illustrating the performance advantage of the minimax regret estimator.

We consider the problem of estimating a 2D image from noisy observations, which are obtained by blurring the image with a blurring kernel (a 2D filter), and adding random Gaussian noise. Specifically, we generate an image $x(z_1, z_2)$ which is the sum of m harmonic oscillations:

$$x(z_1, z_2) = \sum_{\ell=1}^{m} a_\ell \cos(\omega_{\ell,1} z_1 + \omega_{\ell,2} z_2 + \phi_\ell), \tag{82}$$

where

$$\omega_{\ell,i} = \frac{2\pi k_{\ell,i}}{n}, \tag{83}$$

and $k_{\ell,i} \in \mathbb{Z}^2$ are given parameters. Clearly, the image $x(z_1, z_2)$ is periodic with period n. Therefore, we can represent the image by a length-n^2 vector $\mathbf{x}$, with components $\{x(z_1, z_2) : 0 \le z_1, z_2 \le n-1\}$.

The observed image $y(z_1, z_2)$ is given by

$$y(z_1, z_2) = \sum_{\tau_1, \tau_2} H(\tau_1, \tau_2) x(z_1 - \tau_1 - d_1, z_2 - \tau_2 - d_2) + \sigma w(z_1, z_2), \quad 0 \le z_1, z_2 \le n-1, \tag{84}$$

where $H(z_1, z_2)$ is a blurring filter defined by

$$H(z_1, z_2) = \max\left(1 - \frac{\sqrt{z_1^2 + z_2^2}}{\rho}, 0\right), \tag{85}$$

for some parameter ρ, d_1 and d_2 are randomly chosen shifts, and $w(z_1, z_2)$ is an independent, zero-mean, Gaussian noise process so that for each z_1 and z_2, $w(z_1, z_2)$ is $\mathcal{N}(0, 1)$.

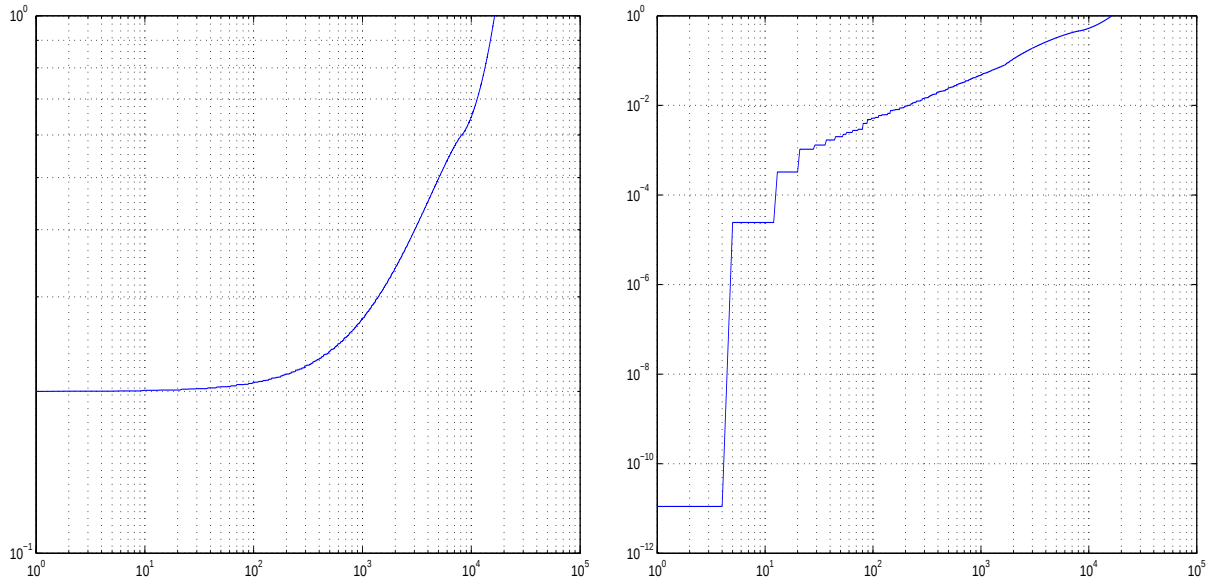
By defining the vectors $\mathbf{y}$ and $\mathbf{w}$ with components $y(z_1, z_2)$ and $w(z_1, z_2)$, respectively, and defining a matrix $\mathbf{H}$ with the appropriate elements $H(z_1, z_2)$, the observations $\mathbf{y}$ can be expressed in the form of a linear model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$.

To evaluate the performance of the minimax regret estimator, we consider 4 different data sets, with parameters given by Table 1. The filters used in all four simulations have, up to shifts, the same support $\{(0,0); (0,1); (0,-1); (1,0); (-1,0)\}$; however, the kernel used for the first data set is essentially different from the kernels used for data sets # 2–4, which are identical up to shifts of each other. The distributions of the singular values and the condition numbers of the kernels are shown in Fig. 2.

To estimate the image $x(z_1, z_2)$ from the noisy observations $y(z_1, z_2)$ we consider 4 different estimators:

Data Set	m	n	σ	$n \cdot \rho$	$(k_{\ell,1}, k_{\ell,2})$	a_ℓ	ϕ_ℓ
1	5	128	0.50	1.200000	(2,1)	0.9235	1.7870
					(1,2)	1.0155	2.9482
					(1,1)	0.8340	0.4070
					(2,2)	0.9329	6.2099
					(1,3)	0.7259	3.6618
2	5	128	0.50	1.414214	(1,1)	1.0681	2.1438
					(2,1)	0.8704	3.3557
					(1,2)	1.2027	4.5686
					(2,2)	1.0466	1.9433
					(3,2)	0.9449	5.2684
3	5	128	0.50	1.414214	(16,11)	0.5784	5.0572
					(4,6)	1.1408	5.7076
					(13,24)	0.6909	1.4570
					(6,19)	1.3439	1.5036
					(14,14)	0.6739	0.3126
4	5	128	0.50	1.414214	(97,52)	1.1649	1.0072
					(51,107)	1.3704	5.4843
					(101,41)	0.5099	1.4946
					(6,66)	0.6370	4.0579
					(123,39)	1.3188	6.0751

Table 1: Parameters for the 4 data sets.



Condition number 5.00 Condition number $9.00 \cdot 10^{10}$

Figure 2: Distribution of the singular values of the **H**-matrix for data set # 1 (left) and data sets # 2–4 (right).

The least-squares (LS) estimator of (5), the minimax regret (REG) estimator of Theorem 3, and two other estimators, the deterministic Wiener estimator (WNR), and the minimax estimator (MMX), which we now describe. We assume that $L = \|\mathbf{x}\|$ and the noise variance σ are known.

The least-squares estimator does not incorporate the prior knowledge on σ and the image norm $L = \|\mathbf{x}\|$. To develop an estimator that incorporates this knowledge, we may assume that $\mathbf{x}$ is a random vector with covariance $L^2\mathbf{I}$ independent of the noise $\mathbf{w}$, and design an MMSE Wiener filter matched to this covariance. The resulting estimator is [2]

$$\hat{\mathbf{x}} = \mathbf{C}_x\mathbf{H}^*(\mathbf{H}\mathbf{C}_x\mathbf{H}^* + \mathbf{C}_w)^{-1}\mathbf{y} = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} + \mathbf{C}_x^{-1})^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y} = \left(\mathbf{H}^*\mathbf{H} + \frac{\sigma^2}{L^2}\mathbf{I}\right)^{-1}\mathbf{H}^*\mathbf{y}. \tag{86}$$

The minimax estimator is developed in [21], and is designed to minimize the worst-case MSE over all possible values of $\mathbf{x}$, such that $\mathbf{x}^*\mathbf{x} \leq L^2$, *i.e.,* it is the solution to the problem

$$\min_{\hat{\mathbf{x}}=\mathbf{G}\mathbf{y}} \max_{\|\mathbf{x}\|\leq L} E\left(\|\hat{\mathbf{x}} - \mathbf{x}\|^2\right), \tag{87}$$

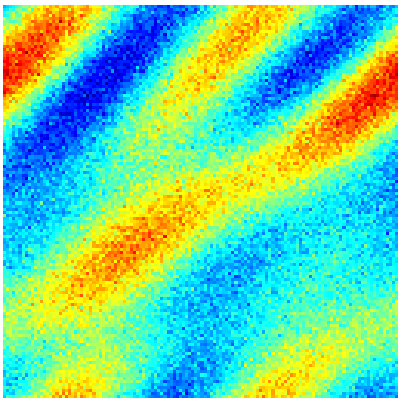
and is given by

$$\hat{\mathbf{x}} = \frac{L^2}{L^2 + \gamma_0}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}^*)^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}, \tag{88}$$

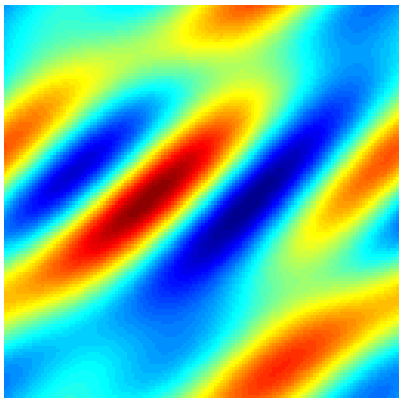
where $\gamma_0 = \mathrm{Tr}\left(\left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\right)$ is the variance of the least-squares estimator.

In Table 2 we report the relative error $\epsilon = \|\hat{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\|$ corresponding to the 4 estimators, for each of the 4 data sets. As can be seen in Fig. 2, for the first data set, where the matrix $\mathbf{H}$ is perfectly conditioned, all of the methods work reasonably well. In contrast, for data sets # 2-4 where $\mathbf{H}$ is poorly conditioned, the performance of the least-squares, minimax, and Wiener estimators are severely degraded. The surprising result is that even though the matrix is ill-conditioned, the minimax regret estimator works pretty well, as can be seen from the results of Table 2, as well as in Figs. 3 and 4 below.

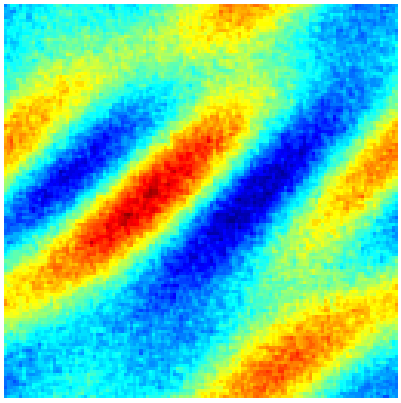
In Figs. 3 and 4 we plot the original image, the observations, and the estimated image for data sets 2 and 3. Since the error in the least-squares estimate is so large, we do not show the resulting image. In the images, the "more red" the image, the larger the signal value at that point.



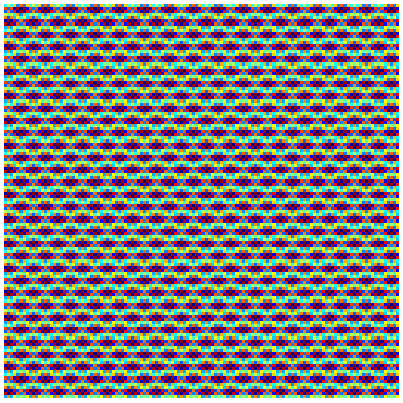
Observations, $\sigma = 0.50$



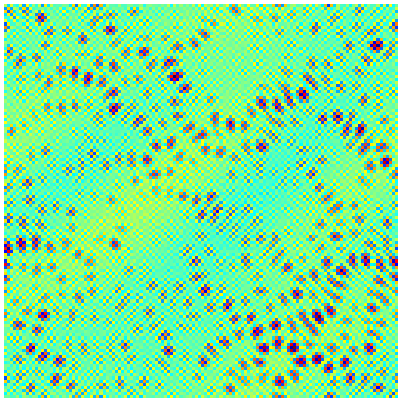
True signal



RGR, $\epsilon = 0.843$

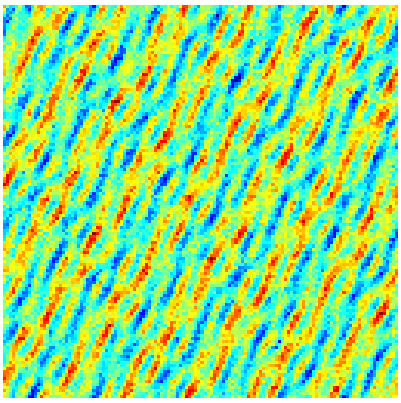


MMX, $\epsilon = 1.00$

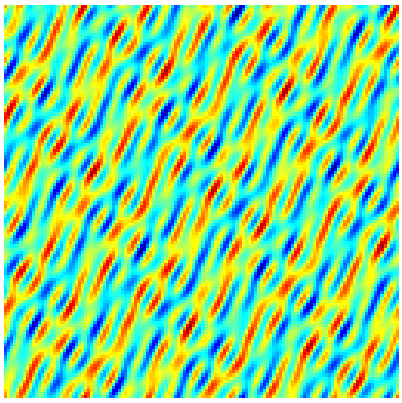


WNR, $\epsilon = 6.17$

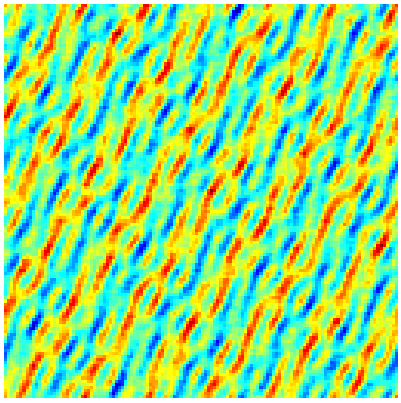
Figure 3: Data set # 2.



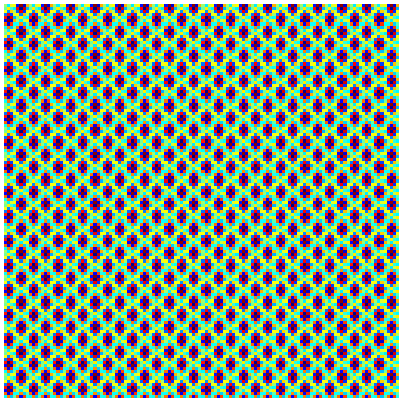
Observations, $\sigma = 0.50$



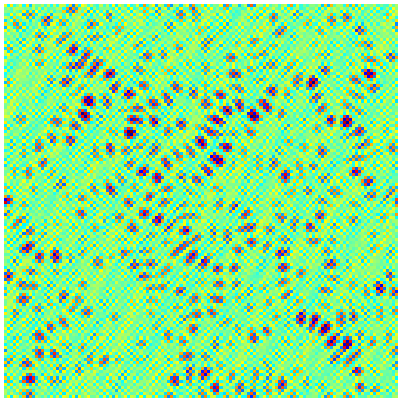
True signal



RGR, $\epsilon = 0.881$



MMX, $\epsilon = 1.00$



WNR, $\epsilon = 6.17$

Figure 4: Data set # 3.

Data	Estimator	Relative Error
1	LS	0.748
	MMX	0.599
	WNR	0.731
	RGR	0.599
2	LS	5.0e8
	MMX	1.00
	WNR	6.17
	RGR	0.843
3	LS	5.0e8
	MMX	1.00
	WNR	6.65
	RGR	0.881
4	LS	5.0e8
	MMX	1.00
	WNR	6.16
	RGR	0.969

Table 2: Relative error for the data sets of Table 1.

7 Conclusion

We considered the problem of estimating an unknown deterministic vector $\mathbf{x}$ in the linear model $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{w}$, where $\mathbf{x}$ is known to be bounded so that $\|\mathbf{x}\|_{\mathbf{T}} \leq L$ for some weighting matrix $\mathbf{T}$. We developed a new linear estimator based on minimizing the worst-case regret, which is the difference between the MSE of the estimator and the best possible MSE attainable with a linear estimator that knows $\mathbf{x}$. As we demonstrated, the minimax regret approach can significantly increase the performance over the traditional least-squares method, even in cases where the least-squares estimator as well as other linear estimators, turn out to be completely useless.

There are of course examples where the minimax regret, as well as all other *linear* estimators, will perform poorly, in which case one may need to consider nonlinear estimators.

In our development of the minimax regret, we assumed that $\mathbf{T}$ commutes with $\mathbf{H}\mathbf{C}_w^{-1}\mathbf{H}$. An interesting direction for future research is to develop the minimax regret estimator for more general classes of weighting matrices $\mathbf{T}$, as well as in the presence of uncertainties in $\mathbf{H}$.

8 Acknowledgment

The first author wishes to thank Prof. A. Singer for first suggesting to her the use of the regret as a figure of merit in the context of estimation.

References

[1] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, New York, NY: John Wiley & Sons, 1949.

[2] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York, NY: McGraw Hill, Inc., third edition, 1991.

[3] K. G. Gauss, *Theory of Motion of Heavenly Bodies*, New York, NY: Dover, 1963.

[4] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Upper Saddle River, NJ: Prentice Hall, Inc., 1993.

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

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

[7] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497–508, Aug. 1973.

[8] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Trans. Signal Processing*, vol. 51, pp. 686–697, Mar. 2003.

[9] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

[10] P. J. Huber, *Robust Statistics*, New York: NY, John Wiley & Sons, Inc., 1981.

[11] L. Breiman, "A note on minimax filtering," *Annals of Probability*, vol. 1, pp. 175–179, 1973.

[12] S. A. Kassam and T. L. Lim, "Robust Wiener filters," *J. Franklin Inst.*, vol. 304, pp. 171–185, Oct./Nov. 1977.

[13] H. V. Poor, "On robust Wiener filtering," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 521–526, June 1980.

[14] K. S. Vastola and H. V. Poor, "Robust Wiener-Kolmogorov theory," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 316–327, Mar. 1984.

[15] J. Franke, "Minimax-robust prediction of discrete time series," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 68, pp. 337–364, 1985.

[16] G. Moustakides and S. A. Kassam, "Minimax robust equalization for random signals through uncertain channels," in *Proc. 20th Annual Allerton Conf. Communication, Control and Computing*, Oct. 1982, pp. 945–954.

[17] Y. C. Eldar and N. Merhav, "A competitive minimax approach to robust estimation in linear models," submitted to *IEEE Trans. Signal Processing*, May 2003.

[18] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *IEEE Proc.*, vol. 73, pp. 433–481, Mar. 1985.

[19] S. Verdú and H. V. Poor, "On minimax robustness: A general approach and applications," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 328–340, Mar. 1984.

[20] S. A. Kassam and H. V. Poor, "Robust signal processing for communication systems," *IEEE Commun. Mag.*, vol. 21, pp. 20–28, 1983.

[21] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation with bounded data uncertainties," in preparation.

[22] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[23] M. Feder and N. Merhav, "Universal composite hypothesis testing: A competitve minimax approach," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1504–1517, June 2002.

[24] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2215–2229, Aug. 2002.

[25] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2124–2147, Oct. 1998.

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

[27] D. G. Luenberger, *Optimization by Vector Space Methods*, New York, NY: John Willey & Sons, 1968.

[28] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, MPS-SIAM Series on Optimization, 2001.

[29] D. P. Bertsekas, *Nonlinear Programming*, Belmont MA: Athena Scientific, second edition, 1999.