# Constrained Linear Minimum MSE Estimation

Tomer Michaeli and Yonina C. Eldar, *Member, IEEE*

***Abstract*— We address the problem of linear minimum mean-squared error (LMMSE) estimation under constraints on the filter or the estimated signal. We develop a general formula that leads to closed form solutions for a wide class of constrained LMMSE problems. The results are applicable to both finite dimensional problems as well as to the Wiener filtering setup, in which infinitely-many measurements are available. Our approach generalizes previous known results such as the generalized Karhunen-Loeve transform (GKLT), the causal Wiener filter and more. As an application of our framework, we develop Wiener type filters under various restrictions, which allow for practical implementations.**

***Index Terms*— Estimation, Wiener filtering, Constrained Estimation.**

## I. Introduction

A COMMON problem in Bayesian estimation is to obtain an estimate of a random vector (r.v.) $\boldsymbol{x}$ based on a realization of another random vector $\boldsymbol{y}$ such that some error criterion is minimized [1]. The estimator $\widehat{\boldsymbol{x}} = \phi(\boldsymbol{y})$ assigns an estimated vector $\widehat{\boldsymbol{x}}$ to every possible realization of $\boldsymbol{y}$. Thus, constructing a Bayesian estimator amounts to a mapping from the space of measurement vectors to the space of signals based on the joint probability function of $\boldsymbol{x}$ and $\boldsymbol{y}$. One of the most commonly used error criteria is the mean-squared error (MSE), which is given by the expectation of the squared-norm error $E[\|\boldsymbol{x} - \phi(\boldsymbol{y})\|^2]$. It is well known that the estimator minimizing the MSE is the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$, denoted as $\phi_0(\boldsymbol{y}) = E[\boldsymbol{x}|\boldsymbol{y}]$.

The minimum MSE (MMSE) estimator, although seemingly simple, is not frequently used due to two main reasons. First, in many cases it is very hard to obtain an expression for $\phi_0$. Second, one often desires to constrain the estimator to belong to a certain class of mappings because of implementation reasons. One way to overcome the difficulties in computation and implementation of the MMSE method is to restrict the estimator to be linear. The linear MMSE estimator (LMMSE) minimizes the MSE among all linear functions. The LMMSE solution has a closed form that depends only on the second order statistics of $\boldsymbol{x}$ and $\boldsymbol{y}$ [2], quantities which may be easily estimated from a set of training data. Moreover, it is easy to implement in practical applications as the estimation procedure involves only matrix multiplication, i.e. $\widehat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y}$. The LMMSE approach also has a simple extension to the case where $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly wide-sense stationary (WSS) random processes, which is known as the Wiener filter [3].

The authors are with the department of electrical engineering, Technion–Israel Institute of Technology, Haifa, Israel (phone: +972-4-8293256, fax: +972-4-8295757, e-mail: tomermic@tx.technion.ac.il, yonina@ee.technion.ac.il).

This work was supported by the Israel Science Foundation under Grant 536/04.

In practical applications, there are situations that require that either the linear estimator itself or the signal at its output possess certain desired properties. These may stem for example from implementation limitations, efficiency of computation or the need to compress the data.

A prime example of constrained linear estimation is the famous work of Wiener [3] on causal LMMSE estimation and prediction of signals. Other restrictions on the Wiener filter include finite impulse response (FIR) [4], finite horizon and general restrictions on the support of the filter in the time domain [5]. Constrained LMMSE estimation arises in array processing applications as well (also termed vector Wiener filtering). In this context, it is often desired to reduce the dimensionality of the measurement vector process. This is analogous to restricting the rank of the estimator. Various approaches were devised in the past for this problem, some directed at minimizing the MSE and some ad hoc (see for example [6],[7] and references therein).

Reduced rank estimators are also at the heart of signal compression. A basic problem encountered in this field is the determination of a small set of vectors that allows the representation (via linear combinations) of a certain class of signals. These vectors are usually chosen to minimize the MSE between the original signal and its compact representation. Like in array processing, this approach can be expressed as a linear estimation problem with a rank constraint. The solution to this problem is known as principal component analysis (PCA) or the Karhunen-Loeve transform (KLT). One extension to this basic concept is the design of a linear compression transform that takes into account additive noise [8]. A more general setting was considered in [9] and [10] in which a compact representation is designed for the task of estimating a different signal (rather than the representation of the original signal itself).

There are applications in which the constraints on the estimator have a stochastic flavor. One example is the MMSE whitening technique which emerged recently [11] and found many applications in the fields of signal processing and communication. In this methodology, a linear transformation is designed such that, when applied to a r.v. $\boldsymbol{y}$, it produces the r.v. $\widehat{\boldsymbol{y}}$ that is as close as possible to $\boldsymbol{y}$ in an MSE sense and whose covariance matrix is diagonal. A generalization of this approach is the covariance shaping technique [12], in which the transform is designed to produce a r.v. with a predefined covariance matrix (not necessarily diagonal). A few of the applications of these approaches are improvement of least squares parameter estimation [13], multiuser detection [14] and matched filtering [15].

Another important class of restrictions emerges in setups where the estimator is only one block in a larger scheme. For example, in a relay communication system it is often desired to estimate a corrupted signal and then retransmit it under power or bandwidth limitations (see e.g. [16] and references therein). Such a setup requires the imposition of constraints on the spectrum of the estimated signal. In speech enhancement algorithms, the estimator is commonly used together with a voice activity detector. The purpose of this approach is to guaranty that the estimator, which is designed to operate on a noisy signal, does not work in noise-only periods. An alternative to this somewhat heuristic approach is to design an LMMSE estimator under a power or amplitude constraint on its output, when only noise is present at its input.

Some of the constrained linear estimation problems discussed above have a common structure: the constraint on the estimator is linear (as defined in Section III). Linear constraints can be treated using the orthogonality principle [1]. This principle leads, for example, to the Wiener-Hopf equations when enforcing a causality constraint on the Wiener filter. However, there seems to be no unifying approach to solving LMMSE estimation problems under general restrictions.

In this paper we show that the solution to all constrained linear estimation problems can be obtained by applying a certain transformation to the unconstrained LMMSE estimator. Specifically, the constrained estimator is a weighted projection of the unconstrained solution onto the constraint set. Motivated by this insight, we derive a connection between a weighted projection operator and the non-weighted projection, which leads to a simple formula that is applicable to almost all constrained LMMSE estimation problems. This then allows us to obtain closed form solutions to many constrained LMMSE problems, including all of the examples above and many more.

Our derivation treats the infinite dimensional Wiener filtering scenario and the finite dimensional setting in a unified way. The intention behind this choice of exposition is to uncover a certain intriguing structure that is common to both type of problems. Specifically, we show that the causal Wiener filter, KLT, reduced rank estimators, MMSE whitening, restrictions on the spectrum of the Wiener filter and many more constrained estimators are actually special cases of one simple formula.

The paper is organized as follows. A brief outline of the notations and mathematical techniques used in the paper is provided in Section II. In Section III we explain the basic ideas underlying the problem of constrained MMSE estimation in general and constrained LMMSE estimation in particular. A geometric viewpoint on constrained LMMSE estimation is presented in Section IV. This viewpoint motivates an alternative formulation of restrictions on the estimator, which leads to the main result, described in Section V. Using this result, in Section VI we derive closed form solutions to various finite dimensional linear estimation problems. Finally, in Section VII we address the problem of designing Wiener-type filters under various restrictions.

## II. Notation And Mathematical Preliminaries

### A. Notation and Definitions

Calligraphic letters are used to denote vector spaces, subspaces and sets of vectors. An inner product on a vector space is denoted by $\langle \cdot, \cdot \rangle$. The associated norm is defined by $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. The null space and range space of a matrix $\boldsymbol{A}$ are denoted by $\mathcal{N}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{A})$ respectively. The Moore-Penrose pseudo-inverse of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\dagger$ and the Hermitian conjugate is $\boldsymbol{A}^*$. If $\boldsymbol{A}$ is an infinite dimensional matrix then it can be viewed as an operator from $\ell_2$ to $\ell_2$, in which case $\boldsymbol{A}^\dagger$ exists if $\mathcal{R}(\boldsymbol{A})$ is closed and $\boldsymbol{A}$ is bounded (i.e. $\sup_{\|\boldsymbol{x}\| \le 1} \|\boldsymbol{A}\boldsymbol{x}\| < \infty$) [17]. The $j$'th component of a vector $\boldsymbol{v}$ is denoted $v_j$ and the $(i,j)$ entry of the matrix $\boldsymbol{A}$ is denoted $\boldsymbol{A}_{i,j}$. A positive (semi) definite matrix is written as $\boldsymbol{A} \succ (\succeq) 0$. Brackets are used for discrete time signals and capital letters for Fourier transforms (e.g. $Z(\omega) = \mathcal{F}\{z[n]\}$). Expectation is denoted by $E[\cdot]$ and conditional expectation is written as $E[\cdot|\cdot]$.

The space of matrices $\mathbb{C}^{M \times N}$ is equipped with the Forbenious norm, which is defined by $\|\boldsymbol{A}\|^2 = \mathrm{Tr}\{\boldsymbol{A}\boldsymbol{A}^*\}$. This definition holds also for matrices with infinitely many columns, i.e. when $M$ is finite and $N = \infty$. In this context, by writing $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ we mean that each row of $\boldsymbol{A}$ is in $\ell_2$. The $\boldsymbol{G}$–weighted norm of a matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N}$, where $\boldsymbol{G}$ is an $N \times N$ positive definite (PD) matrix, is defined as

$$\|\boldsymbol{A}\|_{\boldsymbol{G}}^2 = \mathrm{Tr}\{\boldsymbol{A}\boldsymbol{G}\boldsymbol{A}^*\}. \tag{1}$$

If $\boldsymbol{G}$ is positive semi-definite (PSD) then this is a semi-norm.

The auto covariance sequence of a WSS signal $y[n]$ is defined as $R_{yy}[n] = E[y[m]\, y^*[m-n]]$ and its spectrum is defined as $S_{yy}(\omega) = \mathcal{F}\{R_{yy}[n]\}$. Similarly, the cross covariance sequence associated with two jointly WSS signals $x[n]$ and $y[n]$ is $R_{xy}[n] = E[x[m]\, y^*[m-n]]$ and the cross spectrum is given by $S_{xy}(\omega) = \mathcal{F}\{R_{xy}[n]\}$.

### B. Projections onto Closed Sets

Throughout the paper we extensively use the concept of projections onto closed sets. The projection operator onto a closed set $\mathcal{A}$ in a Hilbert space $\mathcal{H}$ is defined by

$$P_{\mathcal{A}}(\boldsymbol{h}) = \arg\min_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{h} - \boldsymbol{a}\|, \tag{2}$$

where $\boldsymbol{h}$ is an arbitrary vector in $\mathcal{H}$ and $\|\cdot\|$ is the norm on $\mathcal{H}$. In the case $\mathcal{H} = \mathbb{C}^{M \times N}$, we use the notation $P_{\mathcal{A}}^{(\boldsymbol{G})}$ to denote a projection with respect to a $\boldsymbol{G}$–weighted norm, which we refer to as the $\boldsymbol{G}$–weighted projection onto $\mathcal{A}$. Specifically, the $\boldsymbol{G}$–weighted projection of a matrix $\boldsymbol{H} \in \mathbb{C}^{M \times N}$ onto a closed set $\mathcal{A} \subseteq \mathbb{C}^{M \times N}$ is defined by

$$P_{\mathcal{A}}^{(\boldsymbol{G})}(\boldsymbol{H}) = \arg\min_{\boldsymbol{A} \in \mathcal{A}} \|\boldsymbol{H} - \boldsymbol{A}\|_{\boldsymbol{G}}. \tag{3}$$

We shall use the concept of weighted projection also when $\boldsymbol{G}$ has a non-empty null space.

The fact that $\mathcal{A}$ is closed guarantees the existence of a solution to (3). However, generally the solution is not unique. A sufficient condition for the solution to be unique is that $\mathcal{A}$ is a convex set and $\boldsymbol{G} \succ 0$. In this paper, though, we take interest also in cases where at least one of these requirements is not met. To resolve this ambiguity, we shall seek among all possible solutions the one with minimum (Forbenious) norm.

Projections onto subspaces can be expressed in terms of frames. A set of vectors $\{\boldsymbol{h}_n\}$ in a Hilbert space $\mathcal{H}$ is called a frame for a subspace $\mathcal{A}$ if there exist constants $0 < A \leq B < \infty$ such that [18]

$$A \left\| \boldsymbol{a} \right\|^2 \leq \sum_n \left| \langle \boldsymbol{a}, \boldsymbol{h}_n \rangle \right|^2 \leq B \left\| \boldsymbol{a} \right\|^2, \quad \forall \boldsymbol{a} \in \mathcal{A}. \tag{4}$$

The projection of $\boldsymbol{h} \in \mathcal{H}$ onto $\mathcal{A}$ can be written as a linear combination of the frame vectors $P_{\mathcal{A}}\left(\boldsymbol{h}\right) = \sum_k c_k \boldsymbol{h}_k$, where the (possibly infinite) vector of coefficients $\boldsymbol{c}$ can be chosen as [18]:

$$\boldsymbol{c} = \boldsymbol{G}^{\dagger} \boldsymbol{v} \tag{5}$$

with matrix $\boldsymbol{G}$ and vector $\boldsymbol{v}$ defined by

$$\boldsymbol{G}_{j,k} = \langle \boldsymbol{h}_k, \boldsymbol{h}_j \rangle, \quad v_j = \langle \boldsymbol{h}, \boldsymbol{h}_j \rangle. \tag{6}$$

The matrix $\boldsymbol{G}$ is the Gram matrix associated with $\{\boldsymbol{h}_n\}$.

We shall use the following properties of frames:

1) The representation coefficients $\boldsymbol{c}$ of every vector $\boldsymbol{a} \in \mathcal{A}$ is a sequence in $\ell_2$ (i.e. if $\boldsymbol{a} = \sum_k c_k \boldsymbol{h}_k$ then $\boldsymbol{c} \in \ell_2$).
2) The mapping $\boldsymbol{T} : \ell_2 \to \mathcal{A}$ defined by $\boldsymbol{T}\boldsymbol{c} = \sum_k c_k \boldsymbol{h}_k$ is a bounded linear transformation with closed range, and as such, it transforms closed sets in $\ell_2$ into closed sets in $\mathcal{H}$.
3) The Gram matrix $\boldsymbol{G}$ of a frame is a bounded operator with closed range and thus $\boldsymbol{G}^{\dagger}$ exists.

## III. General Theory of Constrained Estimation

The problem of Bayesian estimation refers to the following setup. We wish to estimate the r.v. $\boldsymbol{x} \in \mathbb{C}^N$ based on a realization of another r.v. $\boldsymbol{y} \in \mathbb{C}^M$, where the estimate $\widehat{\boldsymbol{x}}$ is obtained by applying a function $\phi : \mathbb{C}^M \to \mathbb{C}^N$ to $\boldsymbol{y}$. In this paper, the error measure that is used to quantify the quality of the estimate is the MSE, which is defined as $E[\|\boldsymbol{x} - \phi\left(\boldsymbol{y}\right)\|^2]$. It is well known that if no restriction is posed on the estimator $\phi$ then the MSE is minimized by the conditional expectation $\phi_0\left(\boldsymbol{y}\right) = E\left[\boldsymbol{x} \left| \boldsymbol{y}\right.\right]$.

Constrained MMSE estimation refers to the problem of seeking the estimator $\phi$ that minimizes the MSE among a certain family of functions. Perhaps the most famous example of a constraint imposed on $\phi$ is that it be linear. This is known as the LMMSE estimator. Other constraints may be specified as a stochastic restriction, such as $E[\|\phi\left(\boldsymbol{y}\right)\|^2] \leq \varepsilon$, or a deterministic expression such as $\|\phi\left(\boldsymbol{y}\right)\| \leq \varepsilon$ or $\widehat{x}_k \geq 0$, $k = 1 \ldots M$ (for every realization).

In this paper we are interested in the problem of constrained linear MMSE estimation. A constrained LMMSE estimator minimizes the MSE among a certain class of linear transformations. Seeking an estimator in the form of a lower triangular or a low rank matrix are two such examples.

In the following subsections we present a rigorous treatment of the problem of constrained estimation. We begin with an overview on constrained nonlinear MMSE estimation, elaborating on the distinction between linear and nonlinear restrictions. We then discuss the problem of linear MMSE estimation, focusing on a unified treatment of the case where $\boldsymbol{y}$ is finite dimensional and the case where it is infinite dimensional. Finally, we explain the topic of constrained LMMSE estimation. The different types of constrained problems are shown in Fig. 1.

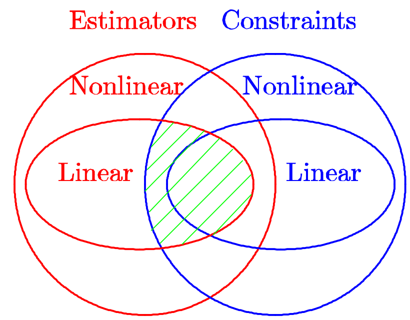


Fig. 1. Types of constrained estimation problems. In this paper we focus on linear estimators subject to nonlinear (and linear) constraints.

### A. Constrained Nonlinear MMSE Estimation

To make the discussion on constrained estimation mathematically precise, we need to introduce a few definitions. Let $(\Omega, F, P)$ be a probability space of r.v.'s taking values in $\mathbb{C}^M$. The set of all finite variance r.v.'s is denoted by

$$\mathcal{L}_2^M = \left\{ \boldsymbol{x} \left| E\left[ \|\boldsymbol{x}\|^2 \right] < \infty \right. \right\}. \tag{7}$$

An inner product on $\mathcal{L}_2^M$ is defined by

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{L}_2^M} = E\left[\boldsymbol{v}^* \boldsymbol{u}\right]. \tag{8}$$

It can be shown that $\mathcal{L}_2^M$ is a Hilbert space given that two r.v.'s $\boldsymbol{u}$,$\boldsymbol{v}$ are considered identical if $\boldsymbol{u} = \boldsymbol{v}$ with probability 1 (w.p.1).

Let $\mathcal{H}_{\mathcal{Y}}$ be the subspace of $\mathcal{L}_2^M$ generated by applying all (Borel measurable) functions from $\mathbb{C}^N$ to $\mathbb{C}^M$ on the random vector $\boldsymbol{y}$:

$$\mathcal{H}_{\mathcal{Y}} = \left\{ \phi\left(\boldsymbol{y}\right) \left| \phi\colon \mathbb{C}^N \!\to \mathbb{C}^M,\, E\left[ \|\phi\left(\boldsymbol{y}\right)\|^2 \right] < \infty \right. \right\}. \tag{9}$$

The set $\mathcal{H}_{\mathcal{Y}}$ is the set of r.v.'s in $\mathcal{L}_2^M$ that can constitute an estimate of $\boldsymbol{x}$ based on $\boldsymbol{y}$. Among all r.v.'s in $\mathcal{H}_{\mathcal{Y}}$, the one that minimizes the MSE to $\boldsymbol{x}$ is the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$, denoted by

$$\widehat{\boldsymbol{x}}_0 = E\left[\boldsymbol{x} \left| \boldsymbol{y}\right.\right]. \tag{10}$$

Note that the MSE $E[\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_0\|^2]$ is simply the $\mathcal{L}_2^M$–distance $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_0\|^2_{\mathcal{L}_2^M}$ between $\boldsymbol{x}$ and $\widehat{\boldsymbol{x}}_0$. Since $\widehat{\boldsymbol{x}}_0$ is the minimizer of this distance among all r.v.'s in $\mathcal{H}_{\mathcal{Y}}$, it can be interpreted as the projection (in the $\mathcal{L}_2^M$ sense) of $\boldsymbol{x}$ onto the subspace $\mathcal{H}_{\mathcal{Y}}$.

The concept of constrained estimation refers to the problem of seeking the MMSE estimator $\widehat{\boldsymbol{x}} = \phi\left(\boldsymbol{y}\right)$ under the restriction that $\phi$ belongs to a certain family of mappings (a subset of all Borel measurable functions). This requirement narrows down the set of candidate r.v.'s in $\mathcal{L}_2^M$ from $\mathcal{H}_{\mathcal{Y}}$ to a subset $\mathcal{A} \subseteq \mathcal{H}_{\mathcal{Y}}$. We distinguish between two types of constraints that can be imposed on $\phi$: nonlinear and linear. Although the latter is a special case of the former, it deserves special attention as a vast majority of the known results in the field of constrained MMSE estimation fall into this category.

*1) Linear Constraints:* Let us examine the case where the estimator $\phi$ is constrained to belong to a class of mappings $\mathcal{M}$ which is a subspace, i.e. $\phi_1, \phi_2 \in \mathcal{M} \implies \alpha\phi_1 + \beta\phi_2 \in \mathcal{M}$. Restrictions of this type are referred to herein as linear constraints. For example, the restriction that the estimator be linear is a linear constraint (as any linear combination of linear functions is also a linear function). Similarly, the constraint that $\widehat{\boldsymbol{x}}$ lie in a subspace $\mathcal{W} \subseteq \mathbb{C}^M$ w.p.1 is a linear constraint as it corresponds to the set of functions $\phi$ whose range is contained in $\mathcal{W}$, which is clearly a subspace of mappings. On the other hand, restrictions of the type $E[\|\widehat{\boldsymbol{x}}\|^2] \leq \varepsilon$ or $\widehat{\boldsymbol{x}}_k \geq 0$ for every $k = 1 \ldots M$ are not linear. We emphasize that a linear constraint does not necessarily mean that $\phi$ is linear, for example confining the range of $\phi$ to be contained in a subspace $\mathcal{W}$ is a linear restriction but it allows the use of nonlinear estimators.

As stated earlier, when a constraint is imposed on the estimator $\phi$, it narrows down the set of candidate r.v.'s in $\mathcal{L}_2^M$ from $\mathcal{H}_\mathcal{Y}$ to a subset $\mathcal{A} \subseteq \mathcal{H}_\mathcal{Y}$. In the special case of linear constraints, the set $\mathcal{A}$ is a subspace of r.v.'s. If this subspace is closed, then the MMSE estimate $\widehat{\boldsymbol{x}}_\mathcal{A}$ among all r.v.'s in $\mathcal{A}$ exists and is given by the orthogonal projection (in the $\mathcal{L}_2^M$ sense) of $\boldsymbol{x}$ onto $\mathcal{A}$.

From the properties of projections in Hilbert spaces one immediately obtains the following characterization of the MMSE estimator under a linear constraint, known as the orthogonality principle. Suppose that $\mathcal{A}$ is a closed subspace of $\mathcal{H}_\mathcal{Y}$. Then $\phi_\mathcal{A}(\boldsymbol{y})$ is the MMSE estimate of $\boldsymbol{x}$ among all r.v.'s in $\mathcal{A}$ if and only if

$$E\left[\phi^*(\boldsymbol{y})(\boldsymbol{x} - \phi_\mathcal{A}(\boldsymbol{y}))\right] = 0, \quad \forall \phi(\boldsymbol{y}) \in \mathcal{A}. \tag{11}$$

Thus, the error $\boldsymbol{x} - \phi_\mathcal{A}(\boldsymbol{y})$ using the optimal estimate in $\mathcal{A}$ has to be orthogonal to any other r.v. in $\mathcal{A}$. For example, if $\mathcal{A}$ is the subspace of r.v.'s that are linear functions of $\boldsymbol{y}$ then the error of $\phi_\mathcal{A}(\boldsymbol{y})$, which is the LMMSE estimator, is orthogonal to *every linear transformation* of $\boldsymbol{y}$. Similarly, in the Wiener filtering setting, a causal filter $h[n]$ is the causal MMSE filter if and only if the error $x[m] - \{h * y\}[m]$ is orthogonal to $\{g * y\}[m]$ for every causal filter $g[n]$.

*2) Nonlinear Constraints:* While linear constraints are easy to handle, situations where the estimator $\phi$ is constrained to belong to a general class of mappings $\mathcal{M}$, which is not necessarily a subspace, are much more challenging. We call these type of restrictions nonlinear constraints. As before, if the corresponding set of r.v.'s $\mathcal{A}$ is a closed subset of $\mathcal{H}_\mathcal{Y}$ then there exists a r.v. $\widehat{\boldsymbol{x}}_\mathcal{A}$ in $\mathcal{A}$ that attains the minimum MSE and it is given by the projection of $\boldsymbol{x}$ onto $\mathcal{A}$. However, obtaining a closed form solution to this projection in $\mathcal{L}_2^M$ is usually impossible.

An interesting generalization of the orthogonality principle (11) to the case of convex (nonlinear) constraints was developed in [19] and is called the extended orthogonality principle. This principle states that if $\mathcal{A}$ is a closed convex set in $\mathcal{H}_\mathcal{Y}$ then $\phi_\mathcal{A}(\boldsymbol{y})$ is the MMSE estimate of $\boldsymbol{x}$ among all r.v.'s in $\mathcal{A}$ if and only if

$$E\left[(\phi_\mathcal{A}(\boldsymbol{y}) - \phi(\boldsymbol{y}))^*(\boldsymbol{x} - \phi_\mathcal{A}(\boldsymbol{y}))\right] \geq 0, \quad \forall \phi(\boldsymbol{y}) \in \mathcal{A}. \tag{12}$$

Note that in contrast to the orthogonality principle, condition (12) is an inequality and therefore does not lead to an equation, whose solution is the MMSE estimate in $\mathcal{A}$. In that sense, it is not constructive. Nevertheless, solutions to various constrained estimation problems were obtained in [19] using (12).

### *B. Linear MMSE Estimation*

The focus of this paper is the design of linear estimators under various constraints. Before giving examples of linear and nonlinear constraints on the LMMSE estimator, we first study its existence in the unconstrained setting. We take interest in two cases. The first is the finite dimensional setting, in which $\boldsymbol{x} \in \mathbb{C}^M$ and $\boldsymbol{y} \in \mathbb{C}^N$ where both $M$ and $N$ are finite. The second, is the Wiener filtering setup, in which the value of a WSS signal $x[m]$ is to be estimated at a certain time instance $m$ based on a realization of another WSS signal $\{y[n]\}_{n\in\mathbb{Z}}$. To treat this second situation in a unified way with the first, we associate an infinite dimensional vector $\boldsymbol{y}$ with the signal $\{y[n]\}_{n\in\mathbb{Z}}$, and a scalar variable $x$ with $x[m]$. Linear estimation can thus always be expressed as

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y}. \tag{13}$$

We emphasize that in both cases of interest, $\boldsymbol{x}$ is finite dimensional, i.e. $M < \infty$. In the Wiener filtering setting, $M = 1$ and $N = \infty$ and therefore the estimator $\boldsymbol{A}$ is an infinite dimensional row vector.

Each element of the vector $\widehat{\boldsymbol{x}}$ is a linear combination of the elements of $\boldsymbol{y}$. Hence, the MSE is minimized if we choose each element of $\widehat{\boldsymbol{x}}$ to be the projection in the $\mathcal{L}_2^1$ sense onto the subspace $\mathit{span}\{y_1, \ldots, y_N\}$. If $N < \infty$ then this is trivially a closed subspace and the projection exists, however for an infinite set of measurements this is not always the case. A sufficient condition for $\mathit{span}\{y_1, \ldots, y_N\}$ to be closed is that the random variables $\{y_k\}$ form a frame. To this end we introduce the following definition.

*Definition 1:* A sequence of random variables is called a *frame process* (FP) if they form a frame for a subspace in $\mathcal{L}_2^1$, where $\mathcal{L}_2^1$ is the set of finite-variance random variables (7).

As mentioned above, every finite sequence of random variables (i.e. a finite dimensional random vector) is a FP. In the case where $y[n]$ is a WSS random process, the conditions for it to be a FP may be specified in terms of its spectrum. This is established in the next theorem.

*Theorem 1:* A WSS process $y[n]$ is a frame process if and only if there exist constants $0 < A \leq B < \infty$ such that the spectrum of $y[n]$ satisfies

$$A \leq S_{yy}(\omega) \leq B, \quad \forall \omega \in \Omega_y, \tag{14}$$

where $\Omega_y$ is defined by

$$\Omega_y \triangleq \{\omega \,|S_{yy}(\omega) \neq 0\} \tag{15}$$

*Proof:* See Appendix I. ∎

If no restriction is posed on the matrix $\boldsymbol{A}$ in (13) and assuming that $\boldsymbol{y}$ is a FP, then the LMMSE estimator can be obtained using (5), leading to the known formula

$$\boldsymbol{A}_0 = \boldsymbol{R}_{\boldsymbol{xy}}\boldsymbol{R}_{\boldsymbol{yy}}^\dagger, \tag{16}$$

where $\boldsymbol{R_{xy}} = E[\boldsymbol{xy}^*]$ is the cross-covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$ and $\boldsymbol{R_{yy}} = E[\boldsymbol{yy}^*]$ is the auto-covariance of $\boldsymbol{y}$. The matrix $\boldsymbol{R_{yy}}$ is the Gram matrix associated with $\{y_k\}$, and since these are a FP, the pseudo inverse in (16) is well defined due to property 3 in subsection II-B. Besides guaranteeing the existence of the LMMSE estimator, the fact that we restrict the discussion to FPs also implies that each row of the matrix $\boldsymbol{A}_0$ is in $\ell_2$. This is implied by property 1 in subsection II-B. We emphasize that (16) applies also to the case of infinitely many measurements. In the Wiener filtering case, $\boldsymbol{R_{yy}}$ is an infinite dimensional matrix and $\boldsymbol{R_{xy}}$ and $\boldsymbol{A}_0$ are both infinite row vectors. Formula (16) then reduces to the well known Wiener filter, as shown in Section VII.

The treatment of finite and infinite dimensional problems in the same manner may seem unnatural at first as it involves infinite dimensional vectors and matrices. This route also calls for precautions, as we shall see in the sequel. The benefit, though, is in enabling a unified viewpoint on the subject of constrained LMMSE estimation. We shall see that this will allow us to reveal a certain common structure that exists in both constrained estimation problems putting, for example, the KLT and the causal Wiener filter under the same umbrella.

### C. Constrained LMMSE Estimation

When a constraint is imposed on the LMMSE estimator, the set of feasible matrices $\boldsymbol{A}$ becomes a subset of $\mathbb{C}^{M\times N}$. A linear constraint corresponds to a subspace in $\mathbb{C}^{M\times N}$. In the finite dimensional setting, examples are restricting $\boldsymbol{A}$ to be a Toeplitz or a lower triangular matrix. Linear constraints on the LMMSE estimator in the Wiener filtering problem include causal filters, FIR filters, and any other restriction on the support of the filter.

A nonlinear restriction on the LMMSE estimator is a general set of matrices in $\mathbb{C}^{M\times N}$ to which the estimator $\boldsymbol{A}$ is confined. Solutions to a few nonlinearly constrained LMMSE problems were obtained in the past. The restriction that the estimator be a low-rank matrix is one such example. This estimator is known as the generalized KLT (GKLT) [10]. In the special case where $\boldsymbol{x} = \boldsymbol{y}$, this estimator reduces to the KLT which is widely used in many applications. In the Wiener filtering setup, an amplitude restriction on the filter such as $|h[n]| \le \varepsilon$ or $|H(\omega)| \le \varepsilon$ is a nonlinear constraint. All these examples cannot be addressed using the orthogonality principle.

It is the aim of this paper to shed some light on the problem of LMMSE estimation under various restrictions. We tackle this problem by transforming the constraint on the estimator $\boldsymbol{A}$ onto a constraint on the covariance matrix $\boldsymbol{R}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}$ of the estimated vector. This representation then allows us to reveal the structure of the constrained estimator. We provide many examples where this approach leads to a closed form solution both for linear as well as nonlinear restrictions. Our emphasis is on the general framework rather than on a specific application. Therefore, a few of the examples that we present are well known results that have been extensively studied in the past. Here we show how they can be easily obtained using our approach.

## IV. Geometric Interpretation of Constrained Linear Estimation

In this section we introduce a geometric interpretation to the problem of constrained LMMSE estimation.

Assume that we wish to construct an estimator $\widehat{\boldsymbol{x}} = \boldsymbol{Ay}$ where the matrix $\boldsymbol{A}$ is restricted to possess a certain structure such as low rank, sparsity or bounded elements. Specifically, we wish to examine the situation where $\boldsymbol{A}$ is constrained to lie in a closed set $\mathcal{W} \subseteq \mathbb{C}^{M\times N}$. The solution to this constrained estimation problem amounts to applying a certain transformation to the unconstrained estimator $\boldsymbol{A}_0$ in (16), as described in the following theorem.

*Theorem 2 (constrained LMMSE):* Let $\boldsymbol{x} \in \mathbb{C}^M$ be a finite variance r.v. (i.e. $E[\|\boldsymbol{x}\|^2] < \infty$), let $\boldsymbol{y} \in \mathbb{C}^N$ be a FP where possibly $N = \infty$, and let $\mathcal{W}$ be a closed set in $\mathbb{C}^{M\times N}$. Then among all estimators of the form $\widehat{\boldsymbol{x}} = \boldsymbol{Ay}$ with $\boldsymbol{A} \in \mathcal{W}$, the matrix that minimizes the MSE is

$$\begin{aligned}\boldsymbol{A}_{\mathcal{W}} &= P_{\mathcal{W}}^{(\boldsymbol{R_{yy}})}(\boldsymbol{A}_0)\\ &= \arg\min_{\boldsymbol{A}\in\mathcal{W}} \|\boldsymbol{A}_0 - \boldsymbol{A}\|^2_{\boldsymbol{R_{yy}}}, \end{aligned} \tag{17}$$

where $\boldsymbol{A}_0 = \boldsymbol{R_{xy}}\boldsymbol{R}^\dagger_{\boldsymbol{yy}}$ is the unconstrained LMMSE estimator.

*Proof:* First, the fact that $\{y_k\}$ is a frame guarantees that the closed set $\mathcal{W} \subseteq \mathbb{C}^{M\times N}$ corresponds to a closed set of r.v.'s in $\mathcal{H}_{\mathcal{Y}}$ (follows from property 2 in subsection II-B). Thus there exists a solution to this constrained estimation problem. Next, the MSE of the estimator (13) is given by

$$\begin{aligned}&E\left[\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|^2\right]\\ &= E\left[(\boldsymbol{x} - \boldsymbol{Ay})^*(\boldsymbol{x} - \boldsymbol{Ay})\right]\\ &= E\left[\mathrm{Tr}\left\{(\boldsymbol{x} - \boldsymbol{Ay})(\boldsymbol{x} - \boldsymbol{Ay})^*\right\}\right]\\ &= \mathrm{Tr}\{\boldsymbol{R_{xx}}\} - 2\,\mathrm{Re}\left\{\mathrm{Tr}\left\{\boldsymbol{R}^*_{\boldsymbol{xy}}\boldsymbol{A}\right\}\right\} + \mathrm{Tr}\left\{\boldsymbol{A}\boldsymbol{R}^*_{\boldsymbol{yy}}\boldsymbol{A}\right\}.\end{aligned} \tag{18}$$

Since the minimization of the MSE is over $\boldsymbol{A}$, we may choose to add the term $\mathrm{Tr}\left\{\boldsymbol{A}_0^*\boldsymbol{R_{yy}}\boldsymbol{A}_0\right\} - \mathrm{Tr}\{\boldsymbol{R_{xx}}\}$ to (18) as it does not depend on $\boldsymbol{A}$. Thus, instead of minimizing the MSE we can minimize

$$\begin{aligned}&\mathrm{Tr}\left\{\boldsymbol{A}_0^*\boldsymbol{R_{yy}}\boldsymbol{A}_0\right\} - 2\,\mathrm{Re}\left\{\mathrm{Tr}\left\{\boldsymbol{R}^*_{\boldsymbol{xy}}\boldsymbol{A}\right\}\right\} + \mathrm{Tr}\left\{\boldsymbol{A}\boldsymbol{R}^*_{\boldsymbol{yy}}\boldsymbol{A}\right\}\\ &= \mathrm{Tr}\left\{(\boldsymbol{A}_0 - \boldsymbol{A})\boldsymbol{R_{yy}}(\boldsymbol{A}_0 - \boldsymbol{A})^*\right\}\\ &= \|\boldsymbol{A}_0 - \boldsymbol{A}\|^2_{R_{\boldsymbol{yy}}}\end{aligned} \tag{19}$$

where we used the relations $\mathbf{A}_0 = \boldsymbol{R_{xy}}\boldsymbol{R}^\dagger_{\boldsymbol{yy}}$ and $\boldsymbol{R_{xy}}\boldsymbol{R}^\dagger_{\boldsymbol{yy}}\boldsymbol{R_{yy}} = \boldsymbol{R_{xy}}$ [20, lemma 2]. The constrained LMMSE problem is thus

$$\begin{aligned}&\arg\min_{\boldsymbol{A}} \|\boldsymbol{A}_0 - \boldsymbol{A}\|^2_{R_{\boldsymbol{yy}}}\\ &\text{s.t. } \boldsymbol{A} \in \mathcal{W}\end{aligned} \tag{20}$$

which means that the optimal $\boldsymbol{A}$ is an $\boldsymbol{R_{yy}}$–weighted projection of $\boldsymbol{A}_0$ onto $\mathcal{W}$. ■

Note that the relation between the unconstrained and constrained matrices is not a simple projection but rather an $\boldsymbol{R_{yy}}$–weighted projection. The methods coincide only if the elements of $\boldsymbol{y}$ are uncorrelated random variables with equal variance, in which case $\boldsymbol{R_{yy}} = \sigma_y^2\boldsymbol{I}$.

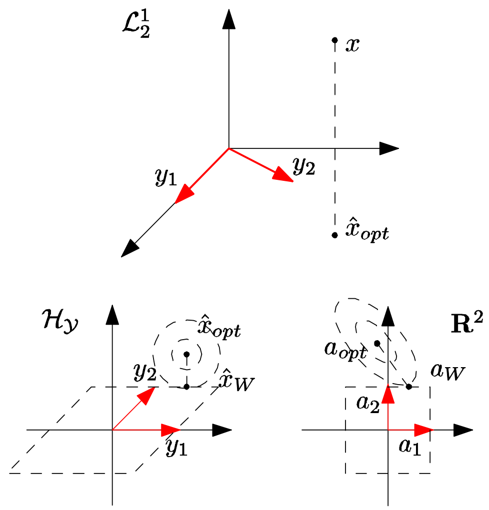


Fig. 2. The LMMSE estimate $\widehat{x}_{opt}$ is the projection of $x$, as a vector in $\mathcal{L}_1^2$, onto $span\{y_1, y_2\}$ (top). Among all estimates of the form $\widehat{x}_A = a_1y_1 + a_2y_2$ with $\|\boldsymbol{a}\|_\infty \leq \varepsilon$, the one that minimizes the MSE is the projection of $\widehat{x}_{opt}$, as a vector in $\mathcal{L}_2^1$, onto a parallelogram (bottom left). The optimal coefficients vector is a weighted projection of $\boldsymbol{a}_{opt} \in \mathbb{R}^2$ onto the box $\|\boldsymbol{a}\|_\infty \leq \varepsilon$.

To gain geometric intuition into this result let us consider the following example. Suppose that we wish to estimate the real random variable $x$ as a linear combination of the real random variables $y_1$ and $y_2$. Let us denote the unconstrained LMMSE estimator by $\widehat{x}_{opt} = a_1^{opt}y_1 + a_2^{opt}y_2$. The r.v. $\widehat{x}_{opt}$ is the projection of $x$ (in the $\mathcal{L}_2^1$ sense) onto $span\{y_1, y_2\}$. Now suppose that we wish to restrict the vector of coefficients to be bounded, i.e. $\|\boldsymbol{a}\|_\infty \leq \varepsilon$. When $y_1$ and $y_2$ are correlated (i.e. not orthogonal as vectors in $\mathcal{L}_2^1$), this box constraint on $\boldsymbol{a}$ translates into a parallelogram constraint in $span\{y_1, y_2\}$ as shown in Fig. 2. Projecting $x$ onto this parallelogram can be carried out in two stages – first projecting $x$ onto $span\{y_1, y_2\}$ and then projecting the result $\widehat{x}_{opt}$ onto the constraint set. This second stage is equivalent to finding the vector in the box $\|\boldsymbol{a}\|_\infty \leq \varepsilon$ that minimizes the weighted distance to $\boldsymbol{a}^{opt}$, as depicted in Fig. 2.

Theorem 2 gives a powerful tool for the design of linear estimators under various nonlinear constraints. To design a constrained linear estimator, one has to first compute the matrix $\boldsymbol{A}_0$ of the unconstrained LMMSE estimator using (16), and then project it on the constraint set $\mathcal{W}$ via an $\boldsymbol{R_{yy}}$–weighted projection, i.e. solve the optimization problem (20). Unfortunately, for a general closed set $\mathcal{W}$, obtaining a closed form solution for (20) is sometimes impossible. However, when the vector of measurements $\boldsymbol{y}$ is finite and $\mathcal{W}$ is convex, (20) can be solved numerically using any standard optimization method.

When there is an infinite number of measurements, such as in the Wiener filtering setup, it is impossible to numerically solve (20). Furthermore, even for finite dimensional problems, numerical procedures often preclude insight into the affect of various parameters on the solution. It is therefore desirable to obtain a closed form for $\boldsymbol{A}_\mathcal{W}$. In the next section we address a wide class of restrictions where this is possible.

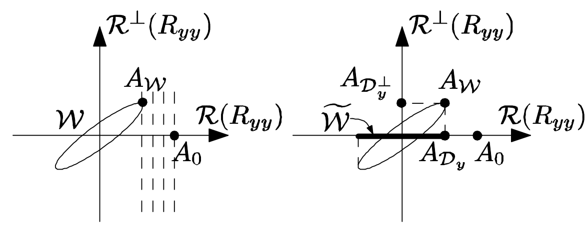


Fig. 3. Equidistant contours from $\boldsymbol{A}_0$ (according to the $\boldsymbol{Ryy}$–weighted norm) are vertical lines in this example (left). Therefore the optimal constrained estimator is the matrix in $\mathcal{W}$ whose first coordinate is closest to that of $\boldsymbol{A}_0$. This means that we can first seek the horizontal component of $\boldsymbol{A}$ that is closest to $\boldsymbol{A}_0$ and lies in the set $\widetilde{\mathcal{W}}$, which is the projection of $\mathcal{W}$ onto the horizontal axis (right). Then, a vertical component that leads to satisfaction of the constraint $\boldsymbol{A} \in \mathcal{W}$ should be added.

## V. An Alternative Viewpoint on Constrained LMMSE Estimation

To better understand constrained LMMSE estimation, notice that the MSE is not affected by the part of $\boldsymbol{A}$ that operates on vectors in $\mathcal{R}^\perp(\boldsymbol{R_{yy}})$. This can be seen in the weighted projection representation in Theorem 2, as demonstrated in Fig. 3 (left). This motivates a decomposition of $\boldsymbol{A}$ as

$$\boldsymbol{A} = P_{\mathcal{D}_y}(\boldsymbol{A}) + P_{\mathcal{D}_y^\perp}(\boldsymbol{A}) = \boldsymbol{A}_{\mathcal{D}_y} + \boldsymbol{A}_{\mathcal{D}_y^\perp}, \tag{21}$$

where $\mathcal{D}_y$ is the subspace defined by

$$\mathcal{D}_y \triangleq \left\{\boldsymbol{A} \in \mathbb{C}^{M\times N} \left| \mathcal{N}^\perp(\boldsymbol{A}) \subseteq \mathcal{R}(\boldsymbol{R_{yy}}) \right.\right\}. \tag{22}$$

The content of $\boldsymbol{A}$ along the subspace $\mathcal{D}_y$ (i.e. the component $\boldsymbol{A}_{\mathcal{D}_y}$) is the part of $\boldsymbol{A}$ that affects the MSE. A direct implication of this decomposition is that, when examining a restriction on $\boldsymbol{A}$, one should focus on the effective constraint imposed on $\boldsymbol{A}_{\mathcal{D}_y}$. The component $\boldsymbol{A}_{\mathcal{D}_y^\perp}$, then, can be chosen arbitrarily as long as $\boldsymbol{A}$ satisfies the constraint. More concretely, to minimize the MSE subject to $\boldsymbol{A} \in \mathcal{W}$, one can first seek the component $\boldsymbol{A}_{\mathcal{D}_y}$ that minimizes the MSE subject to $\boldsymbol{A}_{\mathcal{D}_y} \in \widetilde{\mathcal{W}}$, where $\widetilde{\mathcal{W}} \triangleq \left\{P_{\mathcal{D}_y}(\boldsymbol{A}) | \boldsymbol{A} \in \mathcal{W}\right\}$. Once the optimal $\boldsymbol{A}_{\mathcal{D}_y}$ is available, any $\boldsymbol{A}_{\mathcal{D}_y^\perp}$ such that $\boldsymbol{A} \in \mathcal{W}$ can be used. This approach is shown in Fig. 3 (right), in which $x \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^2$ and $\boldsymbol{Ryy} = \begin{pmatrix} 1 & 0 \end{pmatrix}^T \begin{pmatrix} 1 & 0 \end{pmatrix}$ so that $\boldsymbol{A} \in \mathbb{R}^{1\times 2}$, $\mathcal{R}(\boldsymbol{Ryy})$ coincides with the horizontal axis and the LMMSE estimator $\boldsymbol{A}_0$ lies on this axis as well.

In view of the above discussion, we now address the problem of restricting the component $\boldsymbol{A}_{\mathcal{D}_y}$ of the estimator $\boldsymbol{A}$. When a constraint is imposed on $\boldsymbol{A}_{\mathcal{D}_y}$ it directly affects the covariance $\boldsymbol{R}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}$ of the estimated vector. Therefore, instead of specifying restrictions on $\boldsymbol{A}_{\mathcal{D}_y}$ we may state the desired properties of $\boldsymbol{R}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}$. It is easily verified that *every restriction* on $\boldsymbol{A}_{\mathcal{D}_y}$ can be cast as a constraint on $\boldsymbol{R}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}$. As we show, the benefit of this approach is that in contrast to Theorem 2, the constrained LMMSE estimator in this case can be obtained by employing a *non-weighted* projection operator. These operators have a closed form for many constraint sets, hence allowing to obtain explicit solutions to a large variety of interesting problems.

We begin by noticing that due to (13), every factorization of $\boldsymbol{R_{yy}}$ induces a factorization of the covariance matrix of the estimated vector:

$$\boldsymbol{R}_{\widehat{x}\widehat{x}} = \boldsymbol{A}\boldsymbol{R_{yy}}\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{U_y}\boldsymbol{U_y}^*\boldsymbol{A}^* = \boldsymbol{U}_{\widehat{\boldsymbol{x}}}\boldsymbol{U}_{\widehat{\boldsymbol{x}}}^*, \tag{23}$$

where we defined

$$U_{\hat{x}} \triangleq A U_y. \tag{24}$$

We use the convention that $U_{\hat{x}}$ is an $M \times N$ matrix although $R_{\hat{x}\hat{x}}$ is $M \times M$. Since $\mathcal{R}(U_y) = \mathcal{R}(R_{yy})$, (24) can be written alternatively as $U_{\hat{x}} = A_{\mathcal{D}_y} U_y$.

Our purpose is to seek the LMMSE estimator under a constraint on $U_{\hat{x}}$. Clearly, once the factorization $U_y$ is chosen, a constraint on $U_{\hat{x}}$ translates into a constraint on $A_{\mathcal{D}_y}$ via (24). Note that by definition $U_{\hat{x}} \in \mathcal{D}_y$ and thus for a solution to exist, the constraint set must include at least one matrix in $\mathcal{D}_y$. More precisely, if we impose the restriction $U_{\hat{x}} \in \mathcal{V}$ then the effective constraint on $U_{\hat{x}}$ is actually $U_{\hat{x}} \in \mathcal{V} \cap \mathcal{D}_y$, and this intersection should not be empty.

The following theorem provides the solution to this constrained estimation problem.

*Theorem 3:* Let $x \in \mathbb{C}^M$ be a finite variance r.v. (i.e. $E[\|x\|^2] < \infty$) and let $y \in \mathbb{C}^N$ be a FP where possibly $N = \infty$. Suppose that $U_y$ is a factorization of the covariance matrix $R_{yy}$, i.e. $R_{yy} = U_y U_y^*$, and let $\mathcal{V}$ be a closed set in $\mathbb{C}^{M \times N}$ that contains at least one matrix $V$ with the property $\mathcal{N}^\perp(V) \subseteq \mathcal{R}(R_{yy})$. Then among all estimators of the form $\hat{x} = Ay$ with $U_{\hat{x}} = AU_y \in \mathcal{V}$, a matrix that minimizes the MSE is

$$A = P_{\mathcal{V} \cap \mathcal{D}_y}(A_0 U_y)\, U_y^\dagger, \tag{25}$$

where $A_0$ is the unconstrained LMMSE estimator (16) and $\mathcal{D}_y$ is the set defined in (22).

Before proving the theorem, we note that (25) has been proved in the past for several special cases, although presented differently (see e.g. [10]). Here we provide a rigorous proof which applies to the general case and includes also the infinite dimensional setting $N = \infty$. This will allow the solution of constrained Wiener filtering problems using the same tools, as we discuss in Section VII.

*Proof:* We first need to show that the pseudo-inverse operation in (25) is well defined. We prove in Appendix II that the fact that $y$ is a FP guarantees that $U_y$ is a bounded operator with closed range. This implies that $U_y^\dagger$ exists [17].

Next, for the projection $P_{\mathcal{V} \cap \mathcal{D}_y}$ to be well defined we need to verify that $\mathcal{V} \cap \mathcal{D}_y$ is a closed set. Since $R_{yy}$ is a Gram matrix corresponding to a frame (as $y$ is a FP), its range is closed, which implies that $\mathcal{D}_y$ in (22) is closed. The set $\mathcal{V}$ is closed by assumption and therefore the intersection $\mathcal{V} \cap \mathcal{D}_y$ is also closed.

The constraint $U_{\hat{x}} \in \mathcal{V}$ is equivalent to $U_{\hat{x}} \in \mathcal{V} \cap \mathcal{D}_y$, as by definition $U_{\hat{x}} \in \mathcal{D}_y$. It follows from (24) that this corresponds to the restriction $A \in \mathcal{W}$, where $\mathcal{W}$ is defined by

$$\mathcal{W} = \{A \,|\, (AU_y) \in \mathcal{V} \cap \mathcal{D}_y\}. \tag{26}$$

From the proof of Theorem 2 we know that the optimal $A$ is the weighted projection

$$\begin{aligned} &\arg\min_{A \in \mathcal{W}} \operatorname{Tr}\left\{(A_0 - A) R_{yy} (A_0 - A)^*\right\} \\ &= \arg\min_{A \in \mathcal{W}} \operatorname{Tr}\left\{(A_0 U_y - AU_y)(A_0 U_y - AU_y)^*\right\}. \end{aligned} \tag{27}$$

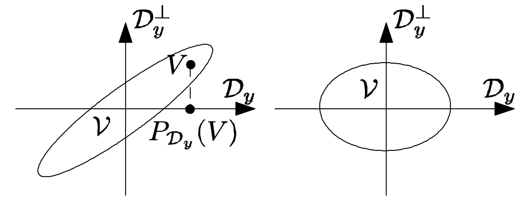


Fig. 4. In a non-$\mathcal{D}_y$-seperable constraint set $\mathcal{V}$, there exists at least one element whose projection onto $\mathcal{D}_y$ falls outside $\mathcal{V}$ (left). On the other hand, the projection onto $\mathcal{D}_y$ of any element in a $\mathcal{D}_y$-seperable set is also an element in $\mathcal{V}$ (right).

Since $AU_y = A_{\mathcal{D}_y} U_y$, both the set $\mathcal{W}$ in (26) and the objective in (27) depend solely on the component $A_{\mathcal{D}_y}$ of $A$. Therefore, $A_{\mathcal{D}_y}$ is the solution to

$$\arg\min_{A_{\mathcal{D}_y} \in \mathcal{W}} \operatorname{Tr}\left\{\left(A_0 U_y - A_{\mathcal{D}_y} U_y\right)\left(A_0 U_y - A_{\mathcal{D}_y} U_y\right)^*\right\}, \tag{28}$$

and $A_{\mathcal{D}_y^\perp}$ can be chosen arbitrarily.

A multiplication by $U_y$ from the right is a bijective transformation on $\mathcal{D}_y$, thus, making the change of variables $B = A_{\mathcal{D}_y} U_y$, the optimal $B$ is the solution of

$$\begin{aligned} B_{\mathcal{V}} &= \arg\min_{B \in \mathcal{V} \cap \mathcal{D}_y} \operatorname{Tr}\left\{(A_0 U_y - B)(A_0 U_y - B)^*\right\} \\ &= \arg\min_{B \in \mathcal{V} \cap \mathcal{D}_y} \|A_0 U_y - B\|^2 \\ &= P_{\mathcal{V} \cap \mathcal{D}_y}(A_0 U_y). \end{aligned} \tag{29}$$

This means that $B_{\mathcal{V}}$ is the projection of $A_0 U_y$ onto the set $\mathcal{V} \cap \mathcal{D}_y$. The optimal $A_{\mathcal{D}_y}$ then must satisfy $B_{\mathcal{V}} = A_{\mathcal{D}_y} U_y$ and $A_{\mathcal{D}_y^\perp}$ can be chosen to be the zero matrix. This corresponds to $A = B_{\mathcal{V}} U_y^\dagger$, which, among all possible solutions, has the minimal Forbenious norm. ■

Theorem 3 implies that every linear estimation problem with a constraint of the form $U_{\hat{x}} \in \mathcal{V}$ may be addressed within our framework, given that there exists a closed form expression for the projection of a matrix onto the set $\mathcal{V} \cap \mathcal{D}_y$. The choice of a specific factorization of $R_{yy}$, together with the choice of the feasible set $\mathcal{V}$, define the constraint posed on the estimator matrix $A \in \mathcal{W}$.

If $R_{yy}$ is non-singular then $\mathcal{D}_y = \mathbb{C}^{M \times N}$ and the projection $P_{\mathcal{V} \cap \mathcal{D}_y}$ in Theorem 3 simplifies to $P_{\mathcal{V}}$. When $R_{yy}$ is singular, there is still a broad family of constraints for which this property holds. These restrictions commonly arise in applications, as we demonstrate in Sections VI and VII.

*Definition 2:* A constraint set $\mathcal{V}$ is called separable with respect to $\mathcal{D}_y$, if for every matrix $V \in \mathcal{V}$ the relation $P_{\mathcal{D}_y}(V) \in \mathcal{V}$ holds. In other words, a $\mathcal{D}_y$-separable set is closed under projections onto $\mathcal{D}_y$, as demonstrated in Fig. 4.

In cases where the restriction posed on $U_{\hat{x}}$ is separable, the projection $P_{\mathcal{V} \cap \mathcal{D}_y}$ in Theorem 3 can be replaced by $P_{\mathcal{V}}$. This is stated formally in the next theorem.

*Theorem 4:* Consider the setup of Theorem 3, where $\mathcal{V}$ is a separable constraint. Then among all estimators of the form $\hat{x} = Ay$ with $U_{\hat{x}} = AU_y \in \mathcal{V}$, a matrix that minimizes the MSE is

$$A = P_{\mathcal{V}}(A_0 U_y)\, U_y^\dagger, \tag{30}$$

where $A_0$ is the unconstrained LMMSE estimator (16).

*Proof:* It is known from Theorem 3 that $\boldsymbol{A} = P_{\mathcal{V}\cap\mathcal{D}_y}\left(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\right)\boldsymbol{U}_{\boldsymbol{y}}^{\dagger}$. Thus, to prove the theorem we need to show that $P_{\mathcal{V}\cap\mathcal{D}_y}\left(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\right) = P_{\mathcal{V}}\left(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\right)$.

Let us denote $\boldsymbol{D} = \boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}$ and assume to the contrary that $P_{\mathcal{V}}\left(\boldsymbol{D}\right) \notin \mathcal{D}_y$. Since $\mathcal{D}_y$ is a closed subspace in $\mathbb{C}^{M\times N}$ and $\boldsymbol{D} \in \mathcal{D}_y$, the Pythagorean theorem implies that

$$\begin{aligned}&\left\|\boldsymbol{D} - P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right\|^2 \\ &= \left\|\boldsymbol{D} - P_{\mathcal{D}_y}\left(P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right)\right\|^2 + \left\|P_{\mathcal{D}_y}\left(P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right) - P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right\|^2 \\ &\geq \left\|\boldsymbol{D} - P_{\mathcal{D}_y}\left(P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right)\right\|^2. \end{aligned} \tag{31}$$

Now, since $P_{\mathcal{V}}\left(\boldsymbol{D}\right) \in \mathcal{V}$ and $\mathcal{V}$ is separable, we have that $P_{\mathcal{D}_y}\left(P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right) \in \mathcal{V}$ as well. By definition, the matrix $P_{\mathcal{D}_y}\left(P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right)$ also lies in $\mathcal{D}_y$ and therefore $P_{\mathcal{D}_y}\left(P_{\mathcal{V}}\left(\boldsymbol{D}\right)\right) \in \mathcal{V}\cap\mathcal{D}_y$. We have established, thus, that there exists a matrix in $\mathcal{V}\cap\mathcal{D}_y$ that is closer to $\boldsymbol{D}$ than $P_{\mathcal{V}}\left(\boldsymbol{D}\right)$, contradicting the assumption that $P_{\mathcal{V}}\left(\boldsymbol{D}\right) \notin \mathcal{D}_y$. ■

In the next sections we present examples for various constrained estimation problems that can be solved using Theorems 2, 3 and 4.

## VI. Finite Dimensional Constrained Estimation Problems

In this section we derive closed form solutions to some constrained linear estimation problems with $\boldsymbol{x} \in \mathbb{C}^M$ and $\boldsymbol{y} \in \mathbb{C}^N$, where $M, N < \infty$.

### A. Estimation with a Lower-Triangular Matrix

A familiar problem in constrained linear estimation is to impose that the estimator be causal. In the finite dimensional setting, causality means that every element of $\widehat{\boldsymbol{x}}$ is a function of the past and present elements of $\boldsymbol{y}$ only. This corresponds to the requirement that the matrix $\boldsymbol{A}$ be lower-triangular.

We assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are of the same dimensions, so that $\boldsymbol{A}$ is a square matrix. Moreover, it is assumed that $\boldsymbol{R}_{yy}$ is non-singular. To obtain a solution to this problem within our framework, we need to cast the constraint as a restriction on some factorization $\boldsymbol{U}_{\widehat{x}}$ of $\boldsymbol{R}_{\widehat{x}\widehat{x}}$. This is achieved by choosing $\boldsymbol{U}_y$ to be the Cholesky factorization of $\boldsymbol{R}_{yy}$ and imposing that $\boldsymbol{U}_{\widehat{x}} = \boldsymbol{A}\boldsymbol{U}_y$ be lower triangular. Since $\boldsymbol{U}_y$ is lower triangular by definition, this implies that $\boldsymbol{A}$ is lower triangular as well. Our constraint set is thus

$$\mathcal{V} = \left\{\boldsymbol{V}\,\middle|\,\boldsymbol{V} \in \mathbb{C}^{M\times M} \text{ is lower triangular}\right\}. \tag{32}$$

From Theorem 3, the solution to this problem is given by

$$\begin{aligned}\boldsymbol{A} &= P_{\mathcal{V}}\left(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\right)\boldsymbol{U}_{\boldsymbol{y}}^{-1} \\ &= \left\{\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\right\}_{+}\boldsymbol{U}_{\boldsymbol{y}}^{-1} \\ &= \left\{\boldsymbol{R}_{\boldsymbol{xy}}\boldsymbol{R}_{\boldsymbol{yy}}^{-1}\boldsymbol{U}_{\boldsymbol{y}}\right\}_{+}\boldsymbol{U}_{\boldsymbol{y}}^{-1} \\ &= \left\{\boldsymbol{R}_{\boldsymbol{xy}}\boldsymbol{U}_{\boldsymbol{y}}^{-*}\right\}_{+}\boldsymbol{U}_{\boldsymbol{y}}^{-1}, \end{aligned} \tag{33}$$

where the operator $\left\{\cdot\right\}_{+}$ sets the elements above the main diagonal to zero and the notation $\boldsymbol{U}_{\boldsymbol{y}}^{-*}$ stands for $\left(\boldsymbol{U}_{\boldsymbol{y}}^{-1}\right)^{*}$.

A causality constraint is actually a linear one, as $\mathcal{V}$ in (32) is a subspace in $\mathbb{C}^{M\times M}$. Therefore, (32) can also be obtained using the orthogonality principle (11). However, as demonstrated above, the utilization of Theorem 3 makes the derivation straightforward.

### B. Estimation with a Predefined Range Space

There are situations where the estimated vector $\widehat{\boldsymbol{x}}$ is obtained as a linear combination of a set of given vectors. These vectors may be several predefined templates, such as the first few basis functions in the DCT transform, as used e.g. in several image coding techniques [21]. Alternatively, these vectors may be templates learned in advance, such as in the Eigenfaces methodology [22]. Assuming that the given set of vectors span a subspace $\mathcal{W} \subseteq \mathbb{C}^M$, this situation corresponds to a constrained linear estimation problem with the constraint $\mathcal{R}\left(\boldsymbol{A}\right) \subseteq \mathcal{W}$. Since $\boldsymbol{U}_{\widehat{\boldsymbol{x}}} = \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{y}}$, we may alternatively cast the constraint as

$$\mathcal{R}\left(\boldsymbol{U}_{\widehat{\boldsymbol{x}}}\right) \subseteq \mathcal{W}. \tag{34}$$

Note that here there is no importance which factorization $\boldsymbol{U}_y$ is used.

Let us define the set $\mathcal{V}$ to be

$$\mathcal{V} = \left\{\boldsymbol{V}\,|\mathcal{R}\left(\boldsymbol{V}\right) \subseteq \mathcal{W}\right\}. \tag{35}$$

This set is separable with respect to $\mathcal{D}_y$. To see this, we note that the the projection of a $M\times N$ matrix $\boldsymbol{V}$ onto the set $\mathcal{D}_y$ can be shown to be given by $P_{\mathcal{D}_y}\left(\boldsymbol{V}\right) = \boldsymbol{V}\boldsymbol{P}_{\mathcal{R}\left(\boldsymbol{R}_{\boldsymbol{yy}}\right)}$, where $\boldsymbol{P}_{\mathcal{R}\left(\boldsymbol{R}_{\boldsymbol{yy}}\right)}$ is the $M\times M$ projection matrix onto the subspace $\mathcal{R}\left(\boldsymbol{R}_{\boldsymbol{yy}}\right)$. Therefore, if $\mathcal{R}\left(\boldsymbol{V}\right) \subseteq \mathcal{W}$ then clearly also $\mathcal{R}\left(P_{\mathcal{D}_y}\left(\boldsymbol{V}\right)\right) \subseteq \mathcal{W}$, implying that $\mathcal{V}$ is separable.

It is easily verified that the projection of a $M\times N$ matrix $\boldsymbol{A}$ onto the set $\mathcal{V}$ is given by $P_{\mathcal{V}}\left(\boldsymbol{A}\right) = \boldsymbol{P}_{\mathcal{W}}\boldsymbol{A}$, where $\boldsymbol{P}_{\mathcal{W}}$ is the $M\times M$ projection matrix onto the subspace $\mathcal{W}$. Therefore, from Theorem 4, the linear MMSE estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$ under constraint (34) is given by

$$\begin{aligned}\boldsymbol{A} &= P_{\mathcal{V}}\left(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\right)\boldsymbol{U}_{\boldsymbol{y}}^{\dagger} \\ &= \boldsymbol{P}_{\mathcal{W}}\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}}\boldsymbol{U}_{\boldsymbol{y}}^{\dagger} \\ &= \boldsymbol{P}_{\mathcal{W}}\boldsymbol{R}_{\boldsymbol{xy}}\boldsymbol{R}_{\boldsymbol{yy}}^{\dagger}\boldsymbol{U}_{\boldsymbol{y}}\boldsymbol{U}_{\boldsymbol{y}}^{\dagger} \\ &= \boldsymbol{P}_{\mathcal{W}}\boldsymbol{R}_{\boldsymbol{xy}}\boldsymbol{R}_{\boldsymbol{yy}}^{\dagger}, \end{aligned} \tag{36}$$

where the last equality follows from the fact that $\boldsymbol{U}_y\boldsymbol{U}_{\boldsymbol{y}}^{\dagger} = \boldsymbol{P}_{\mathcal{R}\left(\boldsymbol{R}_{\boldsymbol{yy}}\right)} = \boldsymbol{P}_{\mathcal{N}^{\perp}\left(\boldsymbol{R}_{\boldsymbol{yy}}^{\dagger}\right)}$. We conclude that in order to constrain $\widehat{\boldsymbol{x}}$ to belong to a subspace $\mathcal{W}$, one has to first compute the unconstrained linear estimate of $\boldsymbol{x}$, and then project it onto the subspace $\mathcal{W}$.

### C. Reduced Rank Estimator

The concept of designing an estimator with a predefined range space usually arises as means for compact representation of $\widehat{\boldsymbol{x}}$. Many times there is a need not only to represent $\widehat{\boldsymbol{x}}$ efficiently, but also to calculate $\widehat{\boldsymbol{x}}$ based on a compact representation of $\boldsymbol{y}$. Specifically, assuming that $\boldsymbol{x} \in \mathbb{C}^M$ and $\boldsymbol{y} \in \mathbb{C}^N$, we wish to obtain an estimate $\widehat{\boldsymbol{x}}$ given a representation of $\boldsymbol{y}$ with $K < N$ coefficients. Since we are concerned with linear estimators, this requirement implies that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^*$, where both $\boldsymbol{U}$ and $\boldsymbol{V}$ have $K$ columns. The vector $\boldsymbol{c} = \boldsymbol{V}^*\boldsymbol{y}$, then, is the compact representation of $\boldsymbol{y}$, and $\widehat{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{c}$ is the desired estimate. This constraint can be expressed as $\operatorname{rank}\left(\boldsymbol{A}\right) \leq K$. Note that as opposed to Subsection VI-B, where the range space of $\boldsymbol{A}$ was assumed to be given in advance, we now wish to find the optimal

$K$ dimensional range space and also the optimal $(N-K)$ dimensional null space.

A solution to the LMMSE estimator under a rank constraint was first derived in [2] and [9] for the case where $\boldsymbol{R}_{\boldsymbol{yy}}$ is nonsingular. It was later rediscovered in [8], where the authors treated the special case of additive independent noise $\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{n}$, naming it relative KLT (RKLT). Finally, it was developed in [10] (without knowledge of [2] and [9]) in the most general setting, without the assumption that $\boldsymbol{R}_{\boldsymbol{yy}}$ is nonsingular, under the name generalized KLT (GKLT). Here we show that the GKLT can be easily obtained using Theorem 4.

To treat this problem within our framework, we notice that any estimator $\boldsymbol{A}$ can be changed into $\boldsymbol{A}\boldsymbol{P}_{\mathcal{R}(\boldsymbol{R}_{\boldsymbol{yy}})}$ or equivalently into $\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{y}}\boldsymbol{U}_{\boldsymbol{y}}^{\dagger}$, without changing its MSE. Now, it is easily verified that $\operatorname{rank}(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{y}}\boldsymbol{U}_{\boldsymbol{y}}^{\dagger})=\operatorname{rank}(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{y}})$ and thus, we may alternatively pose a rank constraint on $\boldsymbol{U}_{\widehat{\boldsymbol{x}}}=\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{y}}$:

$$\operatorname{rank}(\boldsymbol{U}_{\widehat{\boldsymbol{x}}})\leq K. \tag{37}$$

Note that in cases where $\operatorname{rank}(\boldsymbol{U}_{\boldsymbol{y}})\leq K$, the unconstrained LMMSE estimator itself satisfies (37) and thus it is optimal. Our feasible set of factorizations $\boldsymbol{U}_{\widehat{\boldsymbol{x}}}$ is therefore

$$\mathcal{V}=\{\boldsymbol{V}\,|\operatorname{rank}(\boldsymbol{V})\leq K\}. \tag{38}$$

This set is separable as projecting a matrix onto $\mathcal{D}_y$ can only decrease its rank (i.e. $\operatorname{rank}(\boldsymbol{V})\leq K\Rightarrow\operatorname{rank}\left(P_{\mathcal{D}_y}(\boldsymbol{V})\right)\leq K$). The projection of a matrix onto the set $\mathcal{V}$ is obtained by setting its smallest singular values to zero, leaving only the largest $K$ unchanged. Therefore, using Theorem 3, the LMMSE estimator under a rank constraint is given by

$$\begin{aligned}\boldsymbol{A}&=P_{\mathcal{V}}(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}})\,\boldsymbol{U}_{\boldsymbol{y}}^{\dagger}\\&=\operatorname{trun}_K(\boldsymbol{A}_0\boldsymbol{U}_{\boldsymbol{y}})\,\boldsymbol{U}_{\boldsymbol{y}}^{\dagger}\\&=\operatorname{trun}_K\left(\boldsymbol{R}_{\boldsymbol{xy}}\left(\boldsymbol{U}_{\boldsymbol{y}}^{1/2}\right)^{\dagger}\right)\boldsymbol{U}_{\boldsymbol{y}}^{\dagger},\end{aligned} \tag{39}$$

where $\operatorname{trun}_K(\cdot)$ denotes the desired $K$–term truncation of the singular value decomposition representation of its argument.

In the special case $\boldsymbol{x}=\boldsymbol{y}$, (39) is the minimizer of $E[\|\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}\|^2]$ among all rank-$K$ matrices, which is the well known KLT.

### *D. Quadratic Constraint*

Consider the family of estimators of the form (13) with a matrix $\boldsymbol{A}$ that is constrained to have bounded weighted norm:

$$\operatorname{Tr}\{\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^*\}\leq\varepsilon \tag{40}$$

where $\varepsilon>0$ and $\boldsymbol{W}\succeq 0$. In this case, the solution can be obtained by directly solving (20). Defining the set $\mathcal{W}=\{\boldsymbol{A}\,|\operatorname{Tr}\{\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^*\}\leq\varepsilon\}$ and using the method of Lagrange multipliers, the solution to (20) can be easily shown to be $\boldsymbol{A}=\boldsymbol{A}_0\boldsymbol{R}_{yy}(\boldsymbol{R}_{yy}+\lambda\boldsymbol{W})^{\dagger}$, where $\lambda\geq 0$ is the minimal value for which (40) holds. Using the fact that $\boldsymbol{A}_0\boldsymbol{R}_{yy}=\boldsymbol{R}_{xy}$, the constrained LMMSE estimator can be expressed as

$$\boldsymbol{A}=\boldsymbol{R}_{\boldsymbol{xy}}(\boldsymbol{R}_{\boldsymbol{yy}}+\lambda\boldsymbol{W})^{\dagger}. \tag{41}$$

Comparing (41) with the unconstrained estimator (16), we see that imposing a quadratic constraint on the matrix $\boldsymbol{A}$ amounts to adding a regularization term to $\boldsymbol{R}_{yy}$. As a special case of (40) we may constrain the Forbenious norm of the matrix $\boldsymbol{A}$ by choosing $\boldsymbol{W}=\boldsymbol{I}$, which yields a Tikhonov-like regularization.

Constraint (40) may also be used to bound the variance of $\widehat{\boldsymbol{x}}$. To do so we use the fact that $\boldsymbol{R}_{\widehat{x}\widehat{x}}=\boldsymbol{A}\boldsymbol{R}_{yy}\boldsymbol{A}^*$, hence by setting $\boldsymbol{W}=\boldsymbol{R}_{yy}$ (40) becomes

$$\operatorname{Tr}\{\boldsymbol{R}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}\}\leq\varepsilon \tag{42}$$

and the estimator (41) reduces to

$$\boldsymbol{A}=c\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{\dagger}, \tag{43}$$

where $c$ is the largest value for which (42) holds.

A more general interpretation of (40) results from associating $\boldsymbol{W}$ with a covariance matrix of a random vector $\boldsymbol{z}$, i.e. $\boldsymbol{W}=\boldsymbol{R}_{\boldsymbol{zz}}$. The estimator (41) is thus designed to reject any existence of the interference $\boldsymbol{z}$ found at its input. Interestingly, the form of (41) is identical to an unconstrained estimator designed to estimate $\boldsymbol{x}$ from a noisy version of the measurements $\widetilde{\boldsymbol{y}}=\boldsymbol{y}+\sqrt{\lambda}\boldsymbol{z}$ instead of using the measurements themselves. Further discussion on the $\boldsymbol{z}$-resistant estimator is presented in Section VII in the context of Wiener filtering.

## VII. Constrained Wiener Filter

A challenging task in the context of constrained estimation is to impose restrictions on the Wiener filter. The difficulty in this type of problems stems from the fact that they correspond to infinite dimensional optimization problems. Thus, in general no numerical procedure can be carried out in order to obtain a solution. Perhaps the most important example for a constraint imposed on the Wiener filter is a causality restriction. Interestingly, a closed form solution for the MMSE filter under this constraint was derived in [3] and is commonly known as the causal Wiener filter. A causality constraint is a linear one and therefore can be studied using the orthogonality principle, which leads in this case to the Wiener-Hopf equations. However, here we are interested in obtaining closed form solutions also to non-linearly constrained filtering problems. In this section we derive an expression for the MMSE filter under certain nonlinear constraints by utilizing Theorem 3.

In order to derive an expression for a constrained Wiener filter, we first develop the unconstrained Wiener solution within our framework. We assume that $x[n]$ and $y[n]$ are complex valued jointly WSS signals. Our problem is to design a filter $h_m[n]$ that minimizes the MSE $E[|x[m]-\widehat{x}[m]|^2]$ at a certain time instance $m$, thus

$$\widehat{x}[m]=\sum_{k=-\infty}^{\infty}h_m[k]\,y[m-k]. \tag{44}$$

To express this as a standard LMMSE estimation problem, we use the following definitions:

$$\boldsymbol{A}=\begin{pmatrix}\cdots & h_m[-1] & h_m[0] & h_m[1] & \cdots\end{pmatrix}, \tag{45}$$

$$\boldsymbol{y}=\begin{pmatrix}\cdots & y[m+1] & y[m] & y[m-1] & \cdots\end{pmatrix}^T, \tag{46}$$

$$\boldsymbol{x}=x[m]. \tag{47}$$

The estimator $\boldsymbol{A}$ is therefore an infinite row vector, $\boldsymbol{y}$ is an infinite column vector and $\boldsymbol{x}$ is a scalar. Using this representation, the cross-covariance matrix $\boldsymbol{R_{xy}}$ is an infinite row vector corresponding to the cross covariance sequence $R_{xy}[n]$ of the processes $x[n]$ and $y[n]$, i.e.

$$(\boldsymbol{R_{xy}})_j = E\left[x[m]\, y^*[m-j]\right] = R_{xy}[j]. \tag{48}$$

Similarly, the auto-covariance $\boldsymbol{R_{yy}}$ is an infinite Toeplitz matrix which is related to the auto covariance sequence $R_{yy}[n]$, through

$$(\boldsymbol{R_{yy}})_{j,k} = E\left[y[m-j]\, y^*[m-k]\right] = R_{yy}[k-j]. \tag{49}$$

As discussed in Section III, linear estimation problems involving infinitely many measurements do not necessarily have a solution. To ensure the existence of a solution we assume that $y[n]$ is a FP. Specifically, we assume that $A \leq S_{yy}(\omega) \leq B$ for every $\omega \in \Omega_y$, where $\Omega_y \triangleq \{\omega \,|S_{yy}(\omega) \neq 0\}$ and $A$, $B$ are positive finite constants.

The LMMSE estimator is given by (16) $\boldsymbol{A}_0 = \boldsymbol{R_{xy}} \boldsymbol{R_{yy}^{\dagger}}$. Since both $\boldsymbol{R_{xy}}$ and $\boldsymbol{R_{yy}}$ are independent of $m$, the optimal filter is also independent of $m$. Using the fact that $\boldsymbol{R_{yy}}$ is an infinite Toeplitz matrix that corresponds to convolution with $R_{yy}[n]$ and $\mathbf{R_{xy}}$ is an infinite row vector that corresponds to the sequence $R_{xy}[n]$, it is easily verified that the frequency response of the optimal filter $H(\omega)$ can be chosen as

$$H(\omega) = \begin{cases} \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \omega \in \Omega_y \\ 0 & \omega \notin \Omega_y. \end{cases} \tag{50}$$

There are applications which demand that either the filter itself or the spectrum at the output of the filter posses certain properties. Many of these constraints can be handled using the theory of constrained linear estimation developed in Section IV, as demonstrated in the following subsections.

### *A. Constraints on the Output Spectrum*

When a constraint is imposed on the Wiener filter it affects the spectrum of the estimated signal $S_{\widehat{x}\widehat{x}}(\omega)$ at its output. Therefore, one possible way for specifying a restriction on the Wiener solution is to state the desired properties of $S_{\widehat{x}\widehat{x}}(\omega)$. Such an approach is more natural for example in applications where the signal $\widehat{x}[n]$ is to be transmitted under bandwidth or power restrictions, or to be coded efficiently. We will show in this subsection that constraints of this type may be addressed easily using theorems 3 and 4.

The estimated signal $\widehat{x}[n]$ is obtained by convolving the measurements with the Wiener filter, i.e. $\widehat{x}[n] = y[n] * h[n]$. This implies that the spectrum of $\widehat{x}[n]$ is related to the spectrum of $y[n]$ through the equation

$$S_{\widehat{x}\widehat{x}}(\omega) = S_{yy}(\omega)\, |H(\omega)|^2. \tag{51}$$

Let $U_y(\omega)$ be a factorization of $S_{yy}(\omega)$, i.e. $S_{yy}(\omega) = U_y(\omega)\, U_y^*(\omega)$. Note that we do not require spectral factorization in the sense that $u_y[n]$ be a causal sequence, but any decomposition of $S_{yy}(\omega)$ into a multiplication of two conjugate functions. For example $U_y(\omega) = \sqrt{S_{yy}(\omega)}$ is a valid choice. The factorization of $S_{yy}(\omega)$ induces a factorization of $S_{\widehat{x}\widehat{x}}(\omega)$ as

$$\begin{aligned} S_{\widehat{x}\widehat{x}}(\omega) &= (U_y(\omega)\, H(\omega))\, (U_y(\omega)\, H(\omega))^* \\ &= U_{\widehat{x}}(\omega)\, U_{\widehat{x}}^*(\omega), \end{aligned} \tag{52}$$

where we defined

$$U_{\widehat{x}}(\omega) \triangleq U_y(\omega)\, H(\omega). \tag{53}$$

Now, we wish to confine the factorization $U_{\widehat{x}}(\omega)$ to belong to a certain family of functions with desired properties. This requirement can be written as

$$U_{\widehat{x}}(\omega) \in \mathcal{V}, \tag{54}$$

where $\mathcal{V}$ is a closed set in $L_2[-\pi, \pi]$.

Two things must be noticed regarding the representation of restrictions in the form of (54). First, constraints on $U_{\widehat{x}}(\omega)$ only affect the content of $H(\omega)$ in $\operatorname{supp}\{S_{yy}(\omega)\}$. This is not a limitation, though, as $H(\omega)$ can be chosen arbitrarily for frequencies outside $\operatorname{supp}\{S_{yy}(\omega)\}$ without changing the MSE. Second, since $\operatorname{supp}\{U_{\widehat{x}}(\omega)\} \subseteq \operatorname{supp}\{S_{yy}(\omega)\}$ by definition, for a solution to exist we must require that $\mathcal{V}$ contain at least one function whose support is contained in $\operatorname{supp}\{S_{yy}(\omega)\}$.

The relation $S_{yy}(\omega) = U_y(\omega)\, U_y^*(\omega)$ can be written in the time domain as $R_{yy}[n] = u_y[n] * u_y^*[-n]$ or in matrix form as $\boldsymbol{R}_{yy} = \boldsymbol{U_y} \boldsymbol{U_y^*}$, where $\boldsymbol{R}_{yy}$ is the matrix given in (49) and $\boldsymbol{U_y}$ is the infinite Toeplitz matrix defined by $(\boldsymbol{U_y})_{j,k} = u_y[k-j]$. Combining this matrix representation with (53) and using the fact that $\boldsymbol{A}$ is the infinite row vector corresponding to the sequence $\{h[n]\}$, we see that (54) can be written as

$$\boldsymbol{A} \boldsymbol{U_y} \in \widetilde{\mathcal{V}}, \tag{55}$$

where $\widetilde{\mathcal{V}}$ denotes the closed set in $\ell_2$ in the time domain corresponding to $\mathcal{V}$ in (54), which was specified in the frequency domain. This matrix representation also allows to identify that the subspace $\mathcal{D}_y$ in (22) corresponds to the set of functions $\{D(\omega)\, |\operatorname{supp}\{D(\omega)\} \subseteq \operatorname{supp}\{S_{yy}(\omega)\}\}$.

The representation in (55) matches exactly the situation discussed in Theorem 3. Thus the MMSE filter under this constraint is $P_{\widetilde{\mathcal{V}} \cap \widetilde{\mathcal{D}}_y}(\boldsymbol{A}_0 \boldsymbol{U_y})\, \boldsymbol{U_y^{\dagger}}$, where $\boldsymbol{A}_0$ is the unconstrained Wiener filter and $\widetilde{\mathcal{D}}_y$ denotes the set of sequences in the time domain corresponding to $\mathcal{D}_y$, which was specified in the frequency domain. The frequency response of the optimal constrained filter is thus

$$\begin{aligned} H(\omega) &= \frac{1}{U_y(\omega)} P_{\mathcal{V} \cap \mathcal{D}_y}\left(U_y(\omega)\, \frac{S_{xy}(\omega)}{S_{yy}(\omega)}\right) \\ &= \frac{1}{U_y(\omega)} P_{\mathcal{V} \cap \mathcal{D}_y}\left(\frac{S_{xy}(\omega)}{U_y^*(\omega)}\right). \end{aligned} \tag{56}$$

Note that because of (50) and due to the pseudo-inverse operation, wherever a division by $U_y(\omega)$ appears the value of the fraction should be set to zero for frequencies where $U_y(\omega) = 0$. To simplify the exposition, this fact was omitted from (56) and will be omitted throughout the rest of the paper.

Our freedom to design different constraints for which (56) may be used is in choosing the factorization $U_y(\omega)$ (which affects $U_{\widehat{x}}(\omega)$ via (53)) and specifying the set $\mathcal{V}$.

In the Wiener filtering setup, a separable constraint is a set of functions $\mathcal{V}$ such that for every $V(\omega) \in \mathcal{V}$, setting $V(\omega) = 0$ for every $\omega \notin \operatorname{supp}\{S_{yy}(\omega)\}$ leads to a function that is also in $\mathcal{V}$. As explained in Section V, under constraints of this type the optimal filter becomes

$$H(\omega) = \frac{1}{U_y(\omega)} P_{\mathcal{V}}\left(\frac{S_{xy}(\omega)}{U_y^*(\omega)}\right). \tag{57}$$

In the following subsections we present a few examples, all of which have separable constraints.

### B. Causality Constraint

As a special case of (57), we may obtain the causal Wiener filter. Assuming that $S_{yy}(\omega) > 0$ at all frequencies, a causality constraint on $h[n]$ may be enforced by imposing that $u_{\widehat{x}}[n] = (h * u_y)[n]$ be a causal sequence, given that $u_y[n]$ is causal as well. Therefore, let us define $U_y(\omega)$ to be the spectral factorization of $S_{yy}(\omega)$, i.e. $U_y(\omega) = S^+(\omega)$ and $U_y^*(\omega) = S^-(\omega)$, where $S^+(\omega)$ is the Fourier transform of a causal series. The set of functions $\mathcal{V} \in L_2[-\pi, \pi]$ to which $U_{\widehat{x}}(\omega)$ must belong is then the set of all Fourier transforms of causal series

$$\mathcal{V} = \left\{V(\omega) \,\middle|\, \mathcal{F}^{-1}\{V(\omega)\}[n] = 0,\ n < 0\right\}. \tag{58}$$

Substituting this specific choice of $\mathcal{V}$ and $U_y(\omega)$ in (57), we obtain the known expression for the causal Wiener filter [3]

$$H(\omega) = \frac{1}{S^+(\omega)}\left\{\frac{S_{xy}(\omega)}{S^-(\omega)}\right\}_+. \tag{59}$$

The operator $\{\cdot\}_+$ corresponds to the projection operator $P_{\mathcal{V}}(\cdot)$, which simply sets the non-causal coefficients of its argument to zero in the time domain.

### C. Total Power Constraint

Suppose we enforce a power limitation on the estimated signal. To cast this as a constraint on $U_{\widehat{x}}(\omega)$, we can write

$$E\left[|\widehat{x}[n]|^2\right] = \int_{-\pi}^{\pi} S_{\widehat{x}\widehat{x}}(\omega)\, d\omega = \int_{-\pi}^{\pi} |U_{\widehat{x}}(\omega)|^2 d\omega \le \varepsilon. \tag{60}$$

Note that here the choice of $U_y(\omega)$ makes no difference. Therefore, the feasible set of $U_{\widehat{x}}(\omega)$ is a ball in $L_2[-\pi, \pi]$:

$$\mathcal{V} = \left\{V(\omega) \,\middle|\, \int_{-\pi}^{\pi} |V(\omega)|^2 d\omega \le \varepsilon\right\}. \tag{61}$$

The operator $P_{\mathcal{V}}(\cdot)$ in this case simply scales its argument to comply with the desired $L_2$-norm. Hence the optimal filter is

$$H(\omega) = \frac{1}{U_y(\omega)} c \frac{S_{xy}(\omega)}{U_y^*(\omega)} = c \frac{S_{xy}(\omega)}{S_{yy}(\omega)}, \tag{62}$$

where $c \le 1$ is the largest value for which (60) holds. We conclude that in order to optimally bound the power at the output of the filter all one needs to do is concatenate an attenuator after the unconstrained Wiener filter.

### D. Total Power Resistance to Interference

As a generalization of the power constraint (60), we may wish to design a filter that is resistant to a different input signal. Specifically, suppose that in a certain application one has a reason to suspect that an undesired signal $z[n]$ with spectrum $S_{zz}(\omega)$ might be fed to the filter. A robust approach is then to design a filter that minimizes the MSE between $x[n]$ and $y[n] * h[n]$ subject to the constraint that the power of the signal $z[n] * h[n]$ does not exceed a certain level. To address this constraint within our framework we need to express this as a restriction on $U_{\widehat{x}}(\omega)$. This can be done by writing

$$\begin{aligned} E\left[|z[n] * h[n]|^2\right] &= \int_{-\pi}^{\pi} |H(\omega)|^2 S_{zz}(\omega)\, d\omega \\ &= \int_{-\pi}^{\pi} |H(\omega)|^2 S_{yy}(\omega) \frac{S_{zz}(\omega)}{S_{yy}(\omega)} d\omega \\ &= \int_{-\pi}^{\pi} |U_{\widehat{x}}(\omega)|^2 \frac{S_{zz}(\omega)}{S_{yy}(\omega)} d\omega. \end{aligned} \tag{63}$$

The constraint set is thus a weighted $L_2$ ball:

$$\mathcal{V} = \left\{V(\omega) \,\middle|\, \int_{-\pi}^{\pi} |V(\omega)|^2 \frac{S_{zz}(\omega)}{S_{yy}(\omega)} d\omega \le \varepsilon\right\}. \tag{64}$$

Using (57) with the formula for projection onto a weighted $L_2$ ball (109), the optimal filter is

$$\begin{aligned} H(\omega) &= \frac{1}{U_y(\omega)} \frac{S_{xy}(\omega)}{U_y^*(\omega)} \frac{1}{1 + \lambda \frac{S_{zz}(\omega)}{S_{yy}(\omega)}} \\ &= \frac{S_{xy}(\omega)}{S_{yy}(\omega)} \frac{1}{1 + \lambda \frac{S_{zz}(\omega)}{S_{yy}(\omega)}}, \end{aligned} \tag{65}$$

where $\lambda \ge 0$ is the minimal value for which (64) holds.

The intuition behind the power $z$-resistant Wiener filter is simple. At frequencies for which $S_{zz}(\omega)$ is zero or very small with respect to $S_{yy}(\omega)$, the filter is identical to the unconstrained Wiener solution (50). This is because large spectral components found at those frequencies are more likely to come from the measurement signal $y[n]$ rather than the interference signal $z[n]$. On the other hand the filter strongly attenuates frequencies for which $S_{zz}(\omega)$ is large with respect to $S_{yy}(\omega)$. Any spectral content found at those frequencies is probably from the signal $z[n]$.

An interesting fact is that the power $z$-resistant Wiener filter is identical to the unconstrained Wiener solution designed to estimate $x[n]$ from the measurements $\widetilde{y}[n] = y[n] + \sqrt{\lambda} z[n]$, where $z[n]$ is a process independent of $x[n]$ and $y[n]$. Hence, the effect of constraining the filter to belong to the class defined in (63) can be understood as designing a filter to estimate $x[n]$ from a noisy version of the measurements instead of using the measurements themselves.

### E. An $L_2$ Constraint on $H(\omega)$

If the lower bound on $S_{yy}(\omega)$ in (14) is very small, then there may be frequencies at which $S_{yy}(\omega)$ is close to zero. This typically causes the impulse response of the filter to attain large values and have a slow decay, properties which may be undesirable in practical implementations. To overcome these difficulties, we may constrain the $\ell_2$ norm of the filter's impulse response. Equivalently, we can restrict the $L_2$ norm of $H(\omega)$. This is done by setting $S_{zz}(\omega) = 1$ in (63), which leads to the restriction

$$\int_{-\pi}^{\pi} |H(\omega)|^2 \, d\omega \leq \varepsilon. \tag{66}$$

The optimal filter is then given by a simple regularization (65):

$$H(\omega) = \frac{S_{xy}(\omega)}{S_{yy}(\omega) + \lambda}, \tag{67}$$

where $\lambda \geq 0$ is the minimal value for which (66) holds.

This modification of the Wiener filter barely affects frequencies at which $S_{yy}(\omega)$ is large. However it does prevent the denominator from being too small, which precludes the filter from exploding.

### F. Total Amplitude Constraint

The power limitation $\int_{-\pi}^{\pi} S_{\widehat{x}\widehat{x}}(\omega) \, d\omega \leq \varepsilon$ in (60) puts a heavy penalty on frequencies where $S_{\widehat{x}\widehat{x}}(\omega)$ is large. In many applications this is an over-pesimistic modeling of the underlying physical limitations. For example, one may be willing to trade a small frequency band where $S_{\widehat{x}\widehat{x}}(\omega)$ is very large for a large band with zero content. Such a behavior can be achieved by replacing the power limitation with an amplitude constraint $\int_{-\pi}^{\pi} \sqrt{S_{\widehat{x}\widehat{x}}(\omega)} d\omega \leq \varepsilon$. Similar to (60), we may express such a constraint in terms of $U_{\widehat{x}}(\omega)$ as

$$\int_{-\pi}^{\pi} \sqrt{S_{\widehat{x}\widehat{x}}(\omega)} d\omega = \int_{-\pi}^{\pi} |U_{\widehat{x}}(\omega)| \, d\omega \leq \varepsilon. \tag{68}$$

The corresponding set $\mathcal{V}$ is therefore the $L_1$-ball

$$\mathcal{V} = \left\{ V(\omega) \middle| \int_{-\pi}^{\pi} |V(\omega)| \, d\omega \leq \varepsilon \right\}. \tag{69}$$

The operator $P_{\mathcal{V}}(\cdot)$ associated with an $L_1$-ball (103) leads to the following filter

$$H(\omega) = \begin{cases} 0 & \frac{|S_{xy}(\omega)|}{\sqrt{S_{yy}(\omega)}} \leq \delta \\ \frac{S_{xy}(\omega)}{S_{yy}(\omega)} \left(1 - \delta \frac{\sqrt{S_{yy}(\omega)}}{|S_{xy}(\omega)|}\right) & \frac{|S_{xy}(\omega)|}{\sqrt{S_{yy}(\omega)}} > \delta \end{cases} \tag{70}$$

where $\delta \geq 0$ is the minimum value for which (68) is satisfied.

This filter can be viewed as the result of a applying a soft threshold operation on the unrestricted Wiener filter (50). Interestingly, the term $|S_{xy}(\omega)| \Big/ \sqrt{S_{yy}(\omega)}$ plays an important role in shaping the filter's frequency response. This term is related both to the amplitude $\sqrt{S_{\widehat{x}\widehat{x}}(\omega)}$ at the output of the Wiener solution (50) and to its MSE. Specifically, when the signal $y[n]$ is fed to the Wiener filter (50), $\sqrt{S_{\widehat{x}\widehat{x}}(\omega)} = |S_{xy}(\omega)| \Big/ \sqrt{S_{yy}(\omega)}$ and the MSE is given by $R_{xx}[0] - \int_{-\pi}^{\pi} |S_{xy}(\omega)|^2 / S_{yy}(\omega) \, d\omega$. This means that frequencies at which $|S_{xy}(\omega)|^2 / S_{yy}(\omega)$ is small have a minor contribution to the MSE. Therefore, in order to optimally bound the total amplitude at the output of the filter, one should discard exactly those frequency components. Note that at the rest of the frequencies, the constrained filter is a shrunk version of the unrestricted solution.

### G. Total Amplitude Resistance to Interference

Let us now return to the problem of designing a filter that is resistant to a different input signal $z[n]$ with spectrum $S_{zz}(\omega)$. As opposed to the power resistance constraint $\int_{-\pi}^{\pi} |H(\omega)|^2 S_{zz}(\omega) \, d\omega \leq \varepsilon$ imposed in (63), we now wish to bound the total amplitude of $h[n] * z[n]$. Specified in terms of $U_{\widehat{x}}(\omega)$, this constraint becomes

$$\begin{aligned} \int_{-\pi}^{\pi} |H(\omega)| \sqrt{S_{zz}(\omega)} d\omega &= \int_{-\pi}^{\pi} |H(\omega)| \sqrt{S_{yy}(\omega)} \sqrt{\frac{S_{zz}(\omega)}{S_{yy}(\omega)}} d\omega \\ &= \int_{-\pi}^{\pi} |U_{\widehat{x}}(\omega)| \sqrt{\frac{S_{zz}(\omega)}{S_{yy}(\omega)}} d\omega \leq \varepsilon. \end{aligned} \tag{71}$$

As discussed in the previous subsection, such a constraint allows to obtain a large frequency band where $|H(\omega)|^2 S_{zz}(\omega)$ is identically zero at the cost of a small band where it is very large. This essentially concentrates the affect of the interference into a small portion of the frequency axis $[-\pi, \pi]$. The feasible set $\mathcal{V}$ corresponding to (71) is a weighted $L_1$ ball

$$\mathcal{V} = \left\{ V(\omega) \middle| \int_{-\pi}^{\pi} |V(\omega)| \sqrt{\frac{S_{zz}(\omega)}{S_{yy}(\omega)}} d\omega \leq \varepsilon \right\}. \tag{72}$$

Using (57) with the formula for projection onto a weighted $L_1$ ball (103), the amplitude $z$-resistant filter is

$$H(\omega) = \begin{cases} 0 & \frac{|S_{xy}(\omega)|}{\sqrt{S_{zz}(\omega)}} \leq \delta \\ \frac{S_{xy}(\omega)}{S_{yy}(\omega)} \left(1 - \delta \frac{\sqrt{S_{zz}(\omega)}}{|S_{xy}(\omega)|}\right) & \frac{|S_{xy}(\omega)|}{\sqrt{S_{zz}(\omega)}} > \delta, \end{cases} \tag{73}$$

where $\delta \geq 0$ is the minimum value for which (71) holds.

The above filter is again related to the unconstrained filter via a soft threshold operation. However, in contrast to (70), here the frequency band of $H(\omega)$ is determined by the term $|S_{xy}(\omega)| \Big/ \sqrt{S_{zz}(\omega)}$. This ensures that frequencies at which the interference is most significant (i.e. wherever $S_{zz}(\omega)$ is large) are zeroed out.

### H. An $L_1$ Constraint on $H(\omega)$

There may be situations where an amplitude limitation is needed to be posed on the filter's frequency response rather than on the output spectrum. As we saw in the two preceding subsections, such a constraint leads to a filter with a small bandwidth. Filters of this type may be implemented e.g. by a filter-bank of band pass filters. In order to pose an $L_1$ restriction on $H(\omega)$ we simply substitute $S_{zz}(\omega) = 1$ in (71) to obtain

$$\int_{-\pi}^{\pi} |H(\omega)| \, d\omega \leq \varepsilon. \tag{74}$$

Making the same substitution in (73), the optimal $L_1$–constrained filter is

$$H(\omega) = \begin{cases} 0 & |S_{xy}(\omega)| \leq \delta \\ \frac{S_{xy}(\omega)}{S_{yy}(\omega)} \left(1 - \frac{\delta}{|S_{xy}(\omega)|}\right) & |S_{xy}(\omega)| > \delta. \end{cases} \tag{75}$$

We see that now, the set of frequencies that is discarded is determined by $|S_{xy}(\omega)|$. Interestingly, this filter has the same structure as (70) with $|S_{xy}(\omega)|$ here replacing $\sqrt{S_{\widehat{x}\widehat{x}}(\omega)}$.

### I. Per-Frequency Magnitude Constraint

There are applications that demand that the spectrum $S_{\widehat{x}\widehat{x}}(\omega)$ at the output of the filter be bounded for every $\omega$. Neither the total power constraint (60) nor the total amplitude one (68) can be of help in enforcing such a limitation. Suppose that we are interested in the restriction $S_{\widehat{x}\widehat{x}}(\omega) \leq B^2(\omega)$ for some given function $B(\omega) > 0$. This constraint may be written in terms of $U_{\widehat{x}}(\omega)$ as

$$S_{\widehat{x}\widehat{x}}(\omega) = |U_{\widehat{x}}(\omega)|^2 \leq B^2(\omega). \quad (76)$$

Our feasible set $\mathcal{V}$ is thus the weighted $L_\infty$ ball

$$\mathcal{V} = \{V(\omega)\,||V(\omega)| \leq B(\omega),\ \forall\omega \in [-\pi, \pi]\}. \quad (77)$$

The projection associated with this set (99) leads to the following filter:

$$H(\omega) = \begin{cases} \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \frac{|S_{xy}(\omega)|}{\sqrt{S_{yy}(\omega)}} \leq B(\omega) \\ \frac{B(\omega)}{\sqrt{S_{yy}(\omega)}}\frac{S_{xy}(\omega)}{|S_{xy}(\omega)|} & \frac{|S_{xy}(\omega)|}{\sqrt{S_{yy}(\omega)}} > B(\omega). \end{cases} \quad (78)$$

### J. Per-Frequency Resistance to Interference

Suppose that we wish to construct a filter that minimizes the MSE between $x[n]$ and $y[n] * h[n]$ and that, when fed the signal $z[n]$, the spectrum at its output is bounded for every frequency. Specifically, we wish to obtain $|H(\omega)|^2 S_{zz}(\omega) \leq B^2(\omega)$. Let us express this as a constraint on $U_{\widehat{x}}(\omega)$:

$$\begin{aligned} |H(\omega)|^2 S_{zz}(\omega) &= |H(\omega)|^2 S_{yy}(\omega)\frac{S_{zz}(\omega)}{S_{yy}(\omega)} \\ &= |U_{\widehat{x}}(\omega)|^2 \frac{S_{zz}(\omega)}{S_{yy}(\omega)} \leq B^2(\omega). \end{aligned} \quad (79)$$

This restriction corresponds to the constraint set

$$\mathcal{V} = \left\{V(\omega)\,\middle|\,|V(\omega)|\sqrt{\frac{S_{zz}(\omega)}{S_{yy}(\omega)}} \leq B(\omega),\ \forall\omega \in [-\pi, \pi]\right\}. \quad (80)$$

This too is a weighted $L_\infty$ ball the projection onto which leads to the filter

$$H(\omega) = \begin{cases} \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \frac{|S_{xy}(\omega)|}{S_{yy}(\omega)} \leq \frac{B(\omega)}{\sqrt{S_{zz}(\omega)}} \\ \frac{B(\omega)}{\sqrt{S_{zz}(\omega)}}\frac{S_{xy}(\omega)}{|S_{xy}(\omega)|} & \frac{|S_{xy}(\omega)|}{S_{yy}(\omega)} > \frac{B(\omega)}{\sqrt{S_{zz}(\omega)}}. \end{cases} \quad (81)$$

### K. Restriction on the Magnitude and Support of $H(\omega)$

Implementation considerations may sometimes dictate the use of filters whose magnitude is bounded by some value. We notice that the restriction $|H(\omega)| \leq B(\omega)$ is a special case of (79) with $S_{zz}(\omega) = 1$. Substituting $S_{zz}(\omega) = 1$ in (81) we get the optimal filter under a magnitude constraint:

$$H(\omega) = \begin{cases} \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \frac{|S_{xy}(\omega)|}{S_{yy}(\omega)} \leq B(\omega) \\ B(\omega)\frac{S_{xy}(\omega)}{|S_{xy}(\omega)|} & \frac{|S_{xy}(\omega)|}{S_{yy}(\omega)} > B(\omega). \end{cases} \quad (82)$$

Using (82) we may construct a Wiener filter with a predefined support in the frequency domain, say $\Omega_s \subseteq [-\pi, \pi]$. To do so we set $B(\omega) \to \infty$ for $\omega \in \Omega_s$ and $B(\omega) = 0$ otherwise. This choice of upper bound on $S_{\widehat{x}\widehat{x}}(\omega)$ is equivalent to demanding that $S_{yy}(\omega)|H(\omega)|^2$ vanishes outside the set $\Omega_s$. This constraint is satisfied only if $H(\omega) = 0$ at frequencies outside $\Omega_s$ for which $S_{yy}(\omega) > 0$. As for frequencies outside $\Omega_s$ where $S_{yy}(\omega) = 0$, we may choose to set $H(\omega) = 0$ anyway as it does not affect the MSE. Therefore we have that under the constraint

$$H(\omega) = 0, \quad \forall\omega \notin \Omega_s \quad (83)$$

the optimal filter is

$$H(\omega) = \begin{cases} \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \omega \in \Omega_s \\ 0 & \omega \notin \Omega_s. \end{cases} \quad (84)$$

### L. Generalized MMSE Whitening

The concept of whitening a signal arises in a variety of contexts in the fields of signal processing and communication (see [11] and references therein). In [11] a linear MMSE whitening approach was developed. The proposed scheme was to construct a filter that minimizes the MSE between its input and output subject to the constraint that the spectrum of the output signal is white. In other words, the filter is one that results in a white output $\widehat{y}[n]$ that is as close as possible to the input $y[n]$, in an MSE sense.

We wish to generalize this approach by constructing a filter that produces a white output that approximates the original signal $x[n]$ in an MSE sense. Recall that we assume that there is a positive lower bound on the spectrum of the input signal, i.e. $S_{yy}(\omega) \geq A > 0$. Here we also assume that the desired white spectrum satisfies $S_{\widehat{x}\widehat{x}}(\omega) = C \leq A$.

By setting $B^2(\omega) = C$ in (76) we get the constraint $S_{\widehat{x}\widehat{x}}(\omega) \leq C$ which will clearly be active for all frequencies. Therefore practically we pose the restriction

$$S_{\widehat{x}\widehat{x}}(\omega) = C.$$

The optimal filter is obtained by replacing $B^2(\omega)$ with $C$ in (78):

$$H(\omega) = \begin{cases} \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \frac{|S_{xy}(\omega)|^2}{S_{yy}(\omega)} \leq C \\ \sqrt{\frac{C}{S_{yy}(\omega)}}\frac{S_{xy}(\omega)}{|S_{xy}(\omega)|} & \frac{|S_{xy}(\omega)|^2}{S_{yy}(\omega)} > C. \end{cases} \quad (85)$$

Notice that by assumption

$$\begin{aligned} C &= S_{\widehat{x}\widehat{x}}(\omega) = |H(\omega)|^2 S_{yy}(\omega) = \\ &= \left|\frac{S_{xy}(\omega)}{S_{yy}(\omega)}\right|^2 S_{yy}(\omega) = \frac{|S_{xy}(\omega)|^2}{S_{yy}(\omega)} \end{aligned} \quad (86)$$

and hence the condition $|S_{xy}(\omega)|^2 / S_{yy}(\omega) \leq C$ in (85) is never satisfied. This simplifies the formula for the generalized whitening filter:

$$H(\omega) = \sqrt{\frac{C}{S_{yy}(\omega)}}\frac{S_{xy}(\omega)}{|S_{xy}(\omega)|} \quad (87)$$

If we set $x[n] = y[n]$ then we get the whitening filter derived in [11]: $H(\omega) = \sqrt{C/S_{yy}(\omega)}$. The departure of the generalized filter from the standard one is in the phase component. In order for $\widehat{x}[n]$ to be as close as possible in the MSE sense to the signal $x[n]$, the filter's phase should be identical to that of $S_{xy}(\omega)$.

### M. Bandwidth Constraint on the Output Spectrum

Consider the situation where the signal $\widehat{x}[n]$ is to be transmitted over a channel with bandwidth limitation or to be coded with a method that leaves only its strongest spectral components unchanged. These situations require that $S_{\widehat{x}\widehat{x}}(\omega)$ have a small support. As we saw previously, one way to comply with this requirement is to impose an $L_1$ constraint on $\sqrt{S_{\widehat{x}\widehat{x}}(\omega)}$. However such an approach is suboptimal in this setting as we would not like to pay a penalty proportional to $\sqrt{S_{\widehat{x}\widehat{x}}(\omega)}$ for frequencies at which $S_{\widehat{x}\widehat{x}}(\omega) > 0$. Rather, we would like our penalty to be proportional only to the size of the set $\Omega_{S_{\widehat{x}\widehat{x}}} \triangleq \{\omega \,|S_{\widehat{x}\widehat{x}}(\omega) > 0\}$. This requirement is translated into an $L_0$ constraint on $U_{\widehat{x}}(\omega)$ as

$$\int_{-\pi}^{\pi} \mathcal{X}_{\Omega_{S_{\widehat{x}\widehat{x}}}}(\omega)\, d\omega = \int_{-\pi}^{\pi} \mathcal{X}_{\Omega_{U_{\widehat{x}}}}(\omega)\, d\omega \le \varepsilon, \tag{88}$$

where $\Omega_{U_{\widehat{x}}} \triangleq \{\omega \,|U_{\widehat{x}}(\omega) > 0\}$ and $\mathcal{X}_{\Omega}(\omega)$ is the indicator function, which equals 1 for every $\omega \in \Omega$ and zero otherwise. Using (57) with the formula for projection onto an $L_0$ ball (114), the optimal filter is

$$H(\omega) = \begin{cases} 0 & \frac{|S_{xy}(\omega)|}{\sqrt{S_{yy}(\omega)}} \le \delta \\ \frac{S_{xy}(\omega)}{S_{yy}(\omega)} & \frac{|S_{xy}(\omega)|}{\sqrt{S_{yy}(\omega)}} > \delta \end{cases} \tag{89}$$

where $\delta$ is the smallest value for which the band restriction (88) is satisfied.

This filter is the result of a hard-threshold operation on the unrestricted Wiener filer (50). This filter highly resembles the filter (70), which is optimal under an $L_1$ restriction on $\sqrt{S_{\widehat{x}\widehat{x}}(\omega)}$. The difference is in frequencies at which $|S_{xy}(\omega)| > \delta\sqrt{S_{yy}(\omega)}$. In the $L_0$ approach, once we choose not to set to zero a certain frequency band, there is no additional penalty on the magnitude of the spectrum at this band. Therefore at those frequencies $H(\omega)$ in (89) coincides with the unrestricted Wiener filter.

### N. Summary

Besides providing tools for practitioners, the various examples presented in subsections VII-B-VII-M exhibit several interesting structures worth mentioning. We now summarize the results by dividing them into several categories.

The first categorization we consider is induced by the value of $p$ for a weighted $L_p$ restriction posed on the estimator. In subsections VII-C, VII-D and VII-E $L_2$-type constraints are addressed. The optimal filter in all these cases contains some kind of a regularization term in the denominator[1]. In subsections VII-F, VII-G and VII-H $L_1$-type restrictions are considered. In these examples, the result is related to the unconstrained filter via a soft threshold operation. Subsections VII-I, VII-J and VII-K treat a few variations of $L_\infty$ constraints. These impose a modification of the unconstrained solution through a clipping operation. Finally, subsection VII-M discusses an $L_0$ restriction, where a hard-threshold operation emerges.

Qualitatively, the $L_\infty$ and $L_0$ are two extremes, where the former affects only frequency components with large values and the latter affects only those with small magnitude. In this respect, the $L_1$ and $L_2$ restrictions occupy the middle of the scale as both affect the filter at all frequencies. Still, the $L_1$ constraint can be viewed as more similar in nature to $L_0$ as it causes several frequency bands to be zeroed out completely.

The second categorization we consider is by the designation of the constraint. In subsections VII-C, VII-F, VII-I and VII-M the restriction is imposed on the output of the filter. In all these examples the term $|S_{xy}(\omega)|^2 / S_{yy}(\omega)$ plays an important role. This quantity is no other than the spectrum at the output of the unconstrained Wiener filter. Specifically, when the signal $y[n]$ is fed to the Wiener filter (50), the spectrum at its output is $S_{\widehat{x}\widehat{x}}(\omega) = |S_{xy}(\omega)|^2 / S_{yy}(\omega)$. In the $L_0$ and $L_1$ cases (subsections VII-M and VII-F) frequencies for which this value is small are suppressed. In the $L_\infty$ case (subsection VII-I) frequencies for which it is large are clipped. Finally, the $L_2$ case (subsection VII-C) is not obtained as a threshold operation, however it can be viewed as operating in the same manner on small and large values of this term.

Resistance to interference is considered in subsections VII-D, VII-G and VII-J. These filters seem to share no structural resemblance except one obvious property. The larger $S_{zz}(\omega)$ is, the greater the tendency to suppress the magnitude of the filter.

In subsections VII-E, VII-H and VII-K the constraint is imposed on the filter itself. These examples may be useful when implementation considerations are involved. Thus, we point out a rule of thumb for the assessment of the affect of restrictions on the filter. The behavior of the $L_\infty$-constrained filter (subsection VII-K) is dictated by comparing the value of $|S_{xy}(\omega)| / S_{yy}(\omega)$ against a threshold. In the $L_1$ and $L_2$ methods (subsections VII-H and VII-E), on the other hand, the values of $|S_{xy}(\omega)|$ and $S_{yy}(\omega)$ respectively are compared to a threshold. As before, the $L_\infty$ approach induces a clipping operation, the $L_1$ method induces a soft threshold operation and in the $L_2$ case a soft decision technique is employed.

## VIII. Conclusion

We derived a general framework for treating linear MMSE estimation problems under nonlinear constraints. The approach is general in that it allows the treatment of both finitely-many as well as infinitely many measurements (as in the Wiener filtering setup). Using our proposed framework we obtained closed form solutions to many constrained linear estimation problems. In the finite dimensional case, these include estimation with a lower-triangular matrix, predefined range space and low-rank. In the finite dimensional case, formulas for various nonlinear constraints on the frequency response of the Wiener filter have been derived.

[1] In subsection VII-C the regularization is proportional to $S_{yy}(\omega)$.

## Appendix I
## Proof of Theorem 1

A necessary and sufficient condition for the set of r.v.'s $\{y[n]\}$ to form a frame for a subspace $\mathcal{A}$ in $\mathcal{L}_2^1$ is the existence of scalars $0 < A \le B < \infty$ such that condition (4) holds:

$$AE\left[|a|^2\right] \le \sum_{n\in\mathbb{Z}} |E[a^* y[n]]|^2 \le BE\left[|a|^2\right], \quad \forall a \in \mathcal{A}. \tag{90}$$

Every random variable $a$ in $\mathcal{A}$ can be written as a linear combination of $\{y[k]\}$, i.e. $a = \sum_{k\in\mathbb{Z}} h[k]\, y[k]$ with $h \in \ell_2$. Using this representation, the term $E\left[|a|^2\right]$ in (90) becomes

$$\begin{aligned} E\left[\left|\sum_{k\in\mathbb{Z}} h[k]\, y[k]\right|^2\right] &= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} h[k]\, h^*[l]\, E[y[k]\, y^*[l]] \\ &= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} h[k]\, h^*[l]\, R_{yy}[k-l] \\ &= \sum_{k\in\mathbb{Z}} h[k]\, (h^* * R_{yy})[k] \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(-\omega)|^2 S_{yy}(\omega)\, d\omega \end{aligned} \tag{91}$$

where the last equality follows from Parseval's theorem. Similarly, the term $\sum_{n\in\mathbb{Z}} |E[a^* y[n]]|^2$ in (90) can be written as

$$\begin{aligned} &\sum_{n\in\mathbb{Z}} \left| E\left[\sum_{k\in\mathbb{Z}} h^*[k]\, y^*[k]\, y[n]\right]\right|^2 \\ &= \sum_{n\in\mathbb{Z}} \left|\sum_{k\in\mathbb{Z}} h^*[k]\, R_{yy}[n-k]\right|^2 \\ &= \sum_{n\in\mathbb{Z}} |(h^* * R_{yy})[n]|^2 \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(-\omega)|^2 S_{yy}^2(\omega)\, d\omega. \end{aligned} \tag{92}$$

Using these two expressions, the condition for $\{y[k]\}$ to form a frame is the existence of scalars $0 < A \le B < \infty$ such that

$$\begin{aligned} A\int_{-\pi}^{\pi} |H(-\omega)|^2 S_{yy}(\omega)\, d\omega &\le \int_{-\pi}^{\pi} |H(-\omega)|^2 S_{yy}^2(\omega)\, d\omega \\ &\le B\int_{-\pi}^{\pi} |H(-\omega)|^2 S_{yy}(\omega)\, d\omega, \end{aligned} \tag{93}$$

for every $H \in L_2[-\pi,\pi]$. We see that the content of $H(-\omega)$ outside the set $\Omega_y$ defined in (15) has no importance from the viewpoint of (93). As for other frequencies, the maximal $A$ and minimal $B$ are $\min_{\omega\in\Omega_y} S_{yy}(\omega)$ and $\max_{\omega\in\Omega_y} S_{yy}(\omega)$ respectively. Hence, the r.v.'s $\{y[k]\}$ form a frame for a subspace in $\mathcal{L}_1^2$ if and only if there exist constants $0 < A \le B < \infty$ such that the spectrum of $y[n]$ satisfies $A \le S_{yy}(\omega) \le B$, $\omega \in \Omega_y$.

## Appendix II
## Closed Range Property

Suppose that $y[n]$ is a WSS FP and that $\boldsymbol{U_y}$ is a factorization of its covariance matrix, i.e. $\boldsymbol{R_{yy}} = \boldsymbol{U_y}\boldsymbol{U_y^*}$. We wish to show that $U_y$ is a bounded operator from $\ell_2$ to $\ell_2$ with closed range. We shall present a proof only for the case where $\boldsymbol{U_y}$ is a Toeplitz matrix as this is the only situation of interest in this paper (see Section VII). The proof can be extended to the case where $\boldsymbol{U_y}$ is not Toeplitz and also to general FP's (which are not necessarily WSS), however it is outside the scope of this paper. The arguments brought in this appendix can be also used to prove that $\boldsymbol{U_y^*}$ is a bounded operator with closed range.

We shall prove our claim by showing that $\boldsymbol{U_y} = \boldsymbol{U}\boldsymbol{R_{zz}}$ where $\boldsymbol{R_{zz}}$ is the covariance matrix of some WSS FP and $\boldsymbol{U}$ is a unitary matrix. The matrix $\boldsymbol{R_{zz}}$ is a Gram matrix corresponding to a frame and thus it is bounded and its range is closed. The concatenation of a unitary transform obviously does not affect these properties.

Since $\boldsymbol{U_y}$ is an infinite dimensional Toeplitz matrix, it corresponds to convolution with a sequence $u_y[n]$. Thus, the relation $\boldsymbol{R_{yy}} = \boldsymbol{U_y}\boldsymbol{U_y^*}$ can be written as $R_{yy}[n] = u_y[n] * u_y^*[-n]$, or in the frequency domain as $S_{yy}(\omega) = U_y(\omega)\, U_y^*(\omega)$. Let us write $U_y(\omega) = \sqrt{S_{yy}(\omega)}\, U(\omega)$, where $|U(\omega)| = 1$ is the phase component of $U_y(\omega)$ and $\sqrt{S_{yy}(\omega)}$ is its amplitude. Transforming back to the time domain, we see that the matrix $\boldsymbol{U_y}$ can be written as $\boldsymbol{U_y} = \boldsymbol{U}\boldsymbol{R_{zz}}$, where $\boldsymbol{R_{zz}}$ is a Toeplitz matrix corresponding to a sequence $R_{zz}[n]$ whose Fourier transform is $S_{zz}(\omega) = \sqrt{S_{yy}(\omega)}$, and $\boldsymbol{U}$ is a unitary Toeplitz matrix corresponding to a sequence $u[n]$ whose Fourier transform is $U(\omega)$.

It remains to show that $\boldsymbol{R_{zz}}$ is indeed a covariance matrix of a WSS FP $z[n]$. But since $y[n]$ is a WSS FP, its spectrum $S_{yy}(\omega)$ satisfies (14) with bounds, say $A$ and $B$. This implies that $S_{zz}(\omega)$ satisfies (14) with bounds $\sqrt{A}$ and $\sqrt{B}$, and thus $z[n]$ is a WSS FP.

## Appendix III
## Projections onto $L_p$ Balls

Let $f : \mathbb{R} \to \mathbb{C}$ be a scalar function. The $L_p$ norm of $f$ is defined as

$$\|f\|_p = \begin{cases} \left(\int_{\mathbb{R}} |f(t)|^p\, dt\right)^{\frac{1}{p}} & 1 \le p < \infty \\ \sup\limits_{t\in\mathbb{R}} |f(t)| & p = \infty \end{cases} \tag{94}$$

The $L_0$ pseudo-norm of $f$ is defined by

$$\|f\|_0 = \int_{\mathbb{R}} \mathcal{X}_{\Omega_f}(t)\, dt \tag{95}$$

where $\Omega_f \triangleq \{t \,|\, |f(t)| \ne 0\}$. We are interested in obtaining a closed form solution to the problem of projecting in the $L_2$ sense a given function $g$ onto a weighted $L_p$ ball. Our problem is thus

$$\begin{aligned} &\arg\min_f \int_{\mathbb{R}} |g(t) - f(t)|^2\, dt \\ &\text{s.t. } f \in \mathcal{A} \end{aligned} \tag{96}$$

where $\mathcal{A} = \left\{f \,\middle|\, \|Wf\|_p \le \varepsilon\right\}$, $g \in L_2$ and $W$ is a diagonal operator defined by $(Wf)(t) = w(t)\, f(t)$. In what follows we give the solution to (96) for $p = 1, 2, \infty$ with a general PSD operator $W$ and also for $p = 0$ with $W$ being the identity operator. A sufficient condition for the set $\mathcal{A}$ to be closed is that the operator $W$ be bounded from above and also bounded from below on $\mathcal{N}^{\perp}(W)$ [17], hence it is assumed that $w(t)$ satisfies $A \leq w(t) \leq B$ wherever $w(t) \neq 0$, where $0 < A \leq B < \infty$.

We shall use the following property of projections onto closed convex sets. Suppose that $\mathcal{A}$ is a closed convex set in some Hilbert space $\mathcal{H}$, then $P_{\mathcal{A}}(h) \in \mathcal{A}$ is the projection of $h \in \mathcal{H}$ onto $\mathcal{A}$ if and only if [17]

$$\mathrm{Re}\left\{\langle h - P_{\mathcal{A}}(h), P_{\mathcal{A}}(h) - a\rangle\right\} \geq 0, \quad \forall a \in \mathcal{A}. \tag{97}$$

### *A. Projection onto a diagonaly-weighted $L_\infty$ ball*

*Lemma 1:* Let $\mathcal{A}$ be the weighted $l_\infty$ ball (box) defined by

$$\mathcal{A} = \left\{ f \in L_2 \left| \sup_{t\in\mathbb{R}} \{w(t)|f(t)|\} \leq \varepsilon \right.\right\}. \tag{98}$$

Then the solution of (96) is

$$f_{\mathcal{A}}(t) = \begin{cases} g(t) & |g(t)| \leq \frac{\varepsilon}{w(t)} \\ \frac{\varepsilon}{w(t)}\frac{g(t)}{|g(t)|} & |g(t)| > \frac{\varepsilon}{w(t)} \end{cases} \tag{99}$$

for every $t$ for which $w(t) \neq 0$ and $f_{\mathcal{A}}(t) = g(t)$ otherwise.

*Proof:* First, we notice that $f_{\mathcal{A}}$ in (99) is indeed in $\mathcal{A}$, i.e. $\sup_{t\in\mathbb{R}}\{w(t)|f(t)|\} \leq \varepsilon$ and also $f_{\mathcal{A}} \in L_2$ (as $g \in L_2$). It suffices to show that $f_{\mathcal{A}}$ in (99) satisfies the condition (97)

$$\mathrm{Re}\left\{\int (g(t) - f_{\mathcal{A}}(t))(f_{\mathcal{A}}(t) - a(t))^* dt\right\} \geq 0, \quad \forall a \in \mathcal{A}. \tag{100}$$

To prove (100), we show that $\mathrm{Re}\{(g(t) - f_{\mathcal{A}}(t))(f_{\mathcal{A}}(t) - a(t))^*\} \geq 0$ for every $t$. Clearly, if $t$ is such that $w(t) = 0$ then this term is zero as $g(t) - f_{\mathcal{A}}(t) = 0$. Now assume that $w(t) > 0$, then when $f_{\mathcal{A}}(t) = g(t)$ this term vanishes and otherwise,

$$\begin{aligned} &\mathrm{Re}\left\{\left(g(t) - \frac{\varepsilon}{w(t)}\frac{g(t)}{|g(t)|}\right)\left(\frac{\varepsilon}{w(t)}\frac{g^*(t)}{|g(t)|} - a^*(t)\right)\right\} = \\ &\frac{\varepsilon}{w(t)}|g(t)| - \frac{\varepsilon^2}{w^2(t)} - \left(1 - \frac{\varepsilon}{w(t)}\frac{1}{|g(t)|}\right)\mathrm{Re}\{a^*(t)g(t)\} \\ &\geq \frac{\varepsilon}{w(t)}|g(t)| - \frac{\varepsilon^2}{w^2(t)} - \left(1 - \frac{\varepsilon}{w(t)}\frac{1}{|g(t)|}\right)\frac{\varepsilon}{w(t)}|g(t)| \\ &= 0 \end{aligned} \tag{101}$$

where the inequality follows from the fact that since $a \in \mathcal{A}$ we have in particular also that $w(t)|a(t)| \leq \varepsilon$ for every $t$. Hence using the Cauchy-Schwartz inequality we obtain $\mathrm{Re}\{a^*(t)g(t)\} \leq |a(t)||g(t)| \leq \frac{\varepsilon}{w(t)}|g(t)|$. ■

### *B. Projection onto a diagonaly-weighted $L_1$ ball*

*Lemma 2:* Let $\mathcal{A}$ be the weighted $L_1$ ball defined by

$$\mathcal{A} = \left\{ f \left| \int w(t)|f(t)|dt \leq \varepsilon \right.\right\}. \tag{102}$$

Then the solution of (96) is

$$f_{\mathcal{A}}(t) = \begin{cases} 0 & |g(t)| \leq \delta w(t) \\ g(t) - \delta w(t)\frac{g(t)}{|g(t)|} & |g(t)| > \delta w(t), \end{cases} \tag{103}$$

where $\delta \geq 0$ is the minimum value for which $f_{\mathcal{A}} \in \mathcal{A}$.

*Proof:* We will show that with $f_{\mathcal{A}}$ given by (103),

$$\begin{aligned} &\mathrm{Re}\left\{\int f_{\mathcal{A}}^*(t)(g(t) - f_{\mathcal{A}}(t))dt\right\} \\ &\quad \geq \mathrm{Re}\left\{\int a^*(t)(g(t) - f_{\mathcal{A}}(t))dt\right\}, \quad \forall a \in \mathcal{A}. \end{aligned} \tag{104}$$

If $g \in \mathcal{A}$ then $\delta = 0$ and $f_{\mathcal{A}}(t) = g(t)$ so that (104) is satisfied. Now assume that $g \notin \mathcal{A}$. It is clear that in this case the constraint in (102) is active, i.e. $\int_{\mathbb{R}} w(t)|f(t)|dt = \varepsilon$. From (103) we see that

$$g(t) - f_{\mathcal{A}}(t) = \begin{cases} g(t) & |g(t)| \leq \delta w(t) \\ \delta w(t)\frac{g(t)}{|g(t)|} & |g(t)| > \delta w(t) \end{cases} \tag{105}$$

This means that if $t$ is such that $|g(t) - f_{\mathcal{A}}(t)| < \delta w(t)$ then $g(t) < \delta w(t)$, which implies that $f_{\mathcal{A}}(t) = 0$. For the rest of $\mathbb{R}$ we have $g(t) - f_{\mathcal{A}}(t) = \delta w(t)g(t)/|g(t)|$. Moreover, it is easily verified that for the latter case we also have $g(t)/|g(t)| = f_{\mathcal{A}}(t)/|f_{\mathcal{A}}(t)|$. Therefore we can write $g(t) - f_{\mathcal{A}}(t) = \delta w(t)f_{\mathcal{A}}(t)/|f_{\mathcal{A}}(t)|$ at these points. To summarize these two facts, either $f_{\mathcal{A}}(t) = 0$ or $g(t) - f_{\mathcal{A}}(t) = \delta w(t)f_{\mathcal{A}}(t)/|f_{\mathcal{A}}(t)|$, hence the term on the left-hand side of (104) reduces to

$$\begin{aligned} &\mathrm{Re}\left\{\int f_{\mathcal{A}}^*(t)(g(t) - f_{\mathcal{A}}(t))dt\right\} \\ &= \mathrm{Re}\left\{\int f_{\mathcal{A}}^*(t)\delta w(t)\frac{f_{\mathcal{A}}(t)}{|f_{\mathcal{A}}(t)|}dt\right\} \\ &= \delta\int w(t)|f_{\mathcal{A}}(t)|dt. \end{aligned} \tag{106}$$

As for the term on the right-hand side of (104), we have

$$\begin{aligned} &\mathrm{Re}\left\{\int a^*(t)(g(t) - f_{\mathcal{A}}(t))dt\right\} \\ &\leq \int |a(t)|(g(t) - f_{\mathcal{A}}(t))dt \\ &\leq \delta\int w(t)|a(t)|dt \\ &\leq \delta\int w(t)|f_{\mathcal{A}}[k]|dt, \end{aligned} \tag{107}$$

where the first inequality is a Cauchy-Schwartz one, the second inequality follows from the fact that $|g(t) - f_{\mathcal{A}}(t)| \leq \delta w(t)$ for every $t$, and the last inequality is a consequence of $a$ satisfying $\int w(t)|a(t)| \leq \varepsilon$ whereas $f_{\mathcal{A}}$ is known to satisfy $\int w(t)|f_{\mathcal{A}}(t)| = \varepsilon$. ■

### *C. Projection onto a diagonaly-weighted $L_2$ ball*

*Lemma 3:* Let $\mathcal{A}$ be the weighted $L_2$ ball defined by

$$\mathcal{A} = \left\{ f \left| \int w(t)|f(t)|^2 dt \leq \varepsilon \right.\right\} \tag{108}$$

Then the solution of (96) is

$$f_{\mathcal{A}}(t) = \frac{1}{1 + \lambda w(t)}g(t) \tag{109}$$

where $\lambda \geq 0$ is the minimum value for which $f_{\mathcal{A}} \in \mathcal{A}$.

*Proof:* First we notice that (108) poses no limitation on $f(t)$ at points $t$ for which $w(t) = 0$. Therefore, to minimize the $L_2$ distance to $g$ one must set $f_{\mathcal{A}}(t) = g(t)$ wherever $w(t) = 0$, which is indeed the case in (109). At the rest of the real line, $f_{\mathcal{A}}(t)$ must satisfy condition (97), which can be written as

$$\operatorname{Re}\left\{\int_{\Omega_w} f_{\mathcal{A}}^*(t)\, w(t)\, \frac{(g(t) - f_{\mathcal{A}}(t))}{w(t)} dt\right\} \geq \operatorname{Re}\left\{\int_{\Omega_w} a^*(t)\, w(t)\, \frac{(g(t) - f_{\mathcal{A}}(t))}{w(t)} dt\right\}, \quad \forall a \in \mathcal{A}, \tag{110}$$

where $\Omega_w \triangleq \{t\,|w(t) \neq 0\}$. Clearly, among all functions $a \in \mathcal{A}$, the one that maximizes the right hand side of (110) is proportional to $(g(t) - f_{\mathcal{A}}(t))/w(t)$. Hence, in order for (110) to be satisfied for every $a \in \mathcal{A}$, the function $f_{\mathcal{A}}$ must be chosen as

$$f_{\mathcal{A}}(t) = \alpha \frac{(g(t) - f_{\mathcal{A}}(t))}{w(t)}, \tag{111}$$

where $\alpha > 0$ is the maximal value for which $f_{\mathcal{A}} \in \mathcal{A}$. Now, solving for $f_{\mathcal{A}}(t)$ leads to

$$f_{\mathcal{A}}(t) = \frac{1}{1 + \alpha^{-1} w(t)} g(t), \tag{112}$$

which is identical to (109) with $\lambda = \alpha^{-1}$. ■

### D. Projection onto an $L_0$ ball

*Lemma 4:* Let $\mathcal{A}$ be the weighted $L_0$ ball defined by

$$\mathcal{A} = \left\{f \,\middle|\, \int \mathcal{X}_{\Omega_f}(t)\, dt \leq \varepsilon\right\}, \tag{113}$$

where $\Omega_f \triangleq \{t\,||f(t)| \neq 0\}$. Then the solution of (96) is

$$f_{\mathcal{A}}(t) = \begin{cases} 0 & |g(t)| \leq \delta \\ g(t) & |g(t)| > \delta \end{cases} \tag{114}$$

where $\delta \geq 0$ is the minimum value for which $f_{\mathcal{A}} \in \mathcal{A}$.

*Proof:* The $L_2$ distance between $g$ and any function $f \in \mathcal{A}$ can be written as

$$\int |g(t) - f(t)|^2\, dt = \int_{\Omega_f} |g(t) - f(t)|^2\, dt + \int_{\Omega_f^c} |g(t)|^2\, dt, \tag{115}$$

where $\Omega_f^c$ is the complementary of the set $\Omega_f$. This distance depends on two factors – the set $\Omega_f$ and the content of $f(t)$ in $\Omega_f$. For every choice of $\Omega_f$, (115) is minimized if we choose $f(t) = g(t)$ for every $t \in \Omega_f$ as this causes the first term to vanish. The set $\Omega_f$ should be chosen such that the second term is minimized. This is accomplished by setting $\Omega_f^c = \{t\,||g(t)| \leq \delta\}$ with $\delta$ being the smallest value for which $f_{\mathcal{A}} \in \mathcal{A}$. ■

## References


[1] A. N. Shiryaev, *Probability*. Springer, 1995.

[2] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Pub. Co, 1991.

[3] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. MIT Press, Cambridge, Mass. and Wiley, New York, 1949.

[4] A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of prediction," *Journal of Applied Physics*, vol. 21, pp. 654–655, 1950.

[5] B. Picinbono and M. Bouvet, "Constrained Wiener filtering," *IEEE Trans. Inform. Theory*, vol. 33, no. 1, pp. 160–166, 1987.

[6] J. S. Goldstein, I. S. Reed, and L. L. Scharf, "A multistage representation of the Wiener filter based onorthogonal projections," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2943–2959, 1998.

[7] R. Sharma and B. D. Van Veen, "Large modular structures for adaptive beamforming and the Gram-Schmidt preprocessor," *IEEE Trans. Signal Processing*, vol. 42, no. 2, pp. 448–451, 1994.

[8] Y. Yamashita and H. Ogawa, "Relative Karhunen-Loeve transform," *IEEE Trans. Signal Processing*, vol. 44, no. 2, pp. 371–378, 1996.

[9] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal Processing*, vol. 25, no. 2, pp. 113–133, 1991.

[10] Y. Hua and W. Liu, "Generalized Karhunen-Loeve transform," *Signal Processing Letters, IEEE*, vol. 5, no. 6, pp. 141–142, 1998.

[11] Y. C. Eldar and A. V. Oppenheim, "MMSE whitening and subspace whitening," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1846–1851, 2003.

[12] Y. C. Eldar, "Minimum mean-squared error covariance shaping," *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP'03)*, vol. 6, 2003.

[13] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 686–697, 2003.

[14] Y. C. Eldar and A. M. Chan, "An optimal whitening approach to linear multiuser detection," *IEEE Trans. Inform. Theory*, vol. 49, no. 9, pp. 2156–2171, 2003.

[15] Y. C. Eldar, A. V. Oppenheim, and D. Egnor, "Orthogonal and projected orthogonal matched filter detection," *Signal Processing*, vol. 84, pp. 677–693, 2002.

[16] J. N. Laneman and G. W. Wornell, "Energy-efficient antenna sharing and relaying for wireless networks," *Wireless Communications and Networking Conference (WCNC 2000)*, vol. 1, 2000.

[17] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*. John Wiley and Sons, Inc., 1974.

[18] O. Christensen, *An Introduction to Frames and Riesz Bases*. Birkhauser, 2002.

[19] T. Michaeli and Y. C. Eldar, "Minimum MSE Estimation with Convex Constraints," *Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP'07)*, vol. 3, 2007.

[20] A. P. Torokhti and P. G. Howlett, "An optimal filter of the second order," *IEEE Trans. Signal Processing*, vol. 49, no. 5, pp. 1044–1048, 2001.

[21] G. K. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[22] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'91), 1991.*, pp. 586–591, 1991.